Bridging Epochs: Machine Learning in the Era of Big Data Cosmology

by Sneh J. Pandya

B.S. in Physics, University of Illinois at Urbana-Champaign

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

March 31, 2026

Dissertation Committee

Prof. Jonathan Blazek, Co-chair

Prof. James Halverson, Co-chair

Dr. Aleksandra Ćiprijanović

Prof. Fabian Ruehle

Prof. Robin Walters

To my parents, whose shoulders I've stood upon all my life, allowing me to see so far.

Acknowledgments

I have peers who found physics at a very young age and who have dreamt about this achievement ever since. Physics did not find me until much later in life. A simple high school physics experiment studying trajectories in 2D kinematics set me on my own trajectory studying physics at the University of Illinois in Urbana-Champaign (UIUC), and eventually to pursuing my Ph.D. in physics here at Northeastern. Along the way I was fortunate to spend some time at Fermilab and also travel to conferences across the country to share my work. This thesis serves as a snapshot of this journey.

It was difficult to make the financial and lifestyle sacrifices that come with pursuing this degree, including moving to a city far away from my friends and family. It was difficult to fail my qualifying exam the first time after months of studying and repeat that process again for my second try, which I passed. It was difficult to have the first paper I submitted in graduate school rejected, which I then revised and resubmitted for acceptance. I did not overcome these difficulties alone, let alone reach this point in my career.

I begin by thanking my high school physics teacher, Mr. Eric T. Johnson, for cultivating my interest in physics and taking the time to entertain both my many questions about physics and my many grievances with mathematics. I am also deeply grateful for my undergraduate experience at UIUC, including the many professors who helped establish my foundation in the subject and the friends with whom I spent countless late nights staring at problem sets in confusion. A special thank you to Xin and Joshua, who gave me my first research project and opened my eyes to the possibilities of applying machine learning to astrophysics and cosmology.

I have had many great friends, collaborators, and mentors at Northeastern, some of whom are my committee members. I was fortunate to be taught by Robin and Fabian in geometric deep learning and quantum field theory, respectively, and remain grateful for their continued guidance. And to Aleksandra: thank you for being my surrogate advisor and adopting me into your scientific family during my time at Fermilab. I also want to acknowledge my peers in the cosmology[1] and high-energy theoretical physics[2] offices, whom I frequently interrupted during the work day[3], rock climbed with[4], and came to for advice[5]. I especially want to thank Emily, Gabriel, Hamza, Ilana, and Nick for their close friendship over the last five years. Though we have dispersed in many ways, in physics subfields and geographic locations, I owe much of my sanity during the first two years of the program to you all. Being a physicist who walks across disciplines has given me the privilege of learning from so many people — from biophysicists and string theorists to observational cosmologists and computer scientists. It is my hope to one day transfer $\geq \epsilon$ of this wisdom to others.

To my Ph.D. advisors, Jim and Jonathan: your friendship has meant as much to me as your scientific guidance. It has been an honor to learn from and collaborate with you over the past five years. Thank you for giving me the direction I needed to earn this degree, but also the freedom to pursue my own interests. You were two of my biggest supporters and, simultaneously, two of my biggest critics—which is to say you did your jobs as Ph.D. advisors exceptionally well. You helped

[1] Eric, Nick (Van Alfen), Sayan, Zepei

[2] Aaron, Audrey, Christian, Luigi, Sam, Yidi

[3] all of the above

[4] Jacob, Keiichiro, Li

[5] all of the above, and Ning

me carve out a unique scientific path at the intersection of statistics, machine learning, and physics, a path I could not have envisioned on my own. I will miss knocking on your doors and derailing your respective trains of thought. I will also miss our many bikes-and-brews, even more impulsive coffee runs, and of course the free donuts.

On a more personal note, I owe a great deal to the many non-physicists in my life. I have been blessed with amazing friends who have supported and encouraged me over the past ten years, from high school physics through college and my Ph.D. I want to especially thank Ellie, for all of her encouragement and support over the past year. Lastly, to my family—words can hardly describe how much I owe this accomplishment to all of you. To my siblings, Mili, Didi, and Sagar: you have been there for me at the highs and lows of this journey. Thank you for resisting my many "I think I'm going to drop out" moments, and for believing in my abilities when I struggled to do so myself; I do not believe I would have reached this moment without you. And to my parents, who took a leap of faith when immigrating to the United States many decades ago: thank you for also taking a leap of faith with me—not only allowing me, but encouraging me, to pursue what I love—despite my being the first in our family to pursue a graduate degree in the physical sciences. You have provided for me my whole life, including the education that led me to find my passion. I have seen so far thanks to you.

"This black board is meant for music, not physics.
Of course, physics is also music — but of a different type.
It is the music of Nature."

Pran Nath
***Quantum Mechanics Lecture*, 2021**

"Artificial Intelligence is hard because there
is no such thing as a free lunch. [. . .]
Physics is simple because there is no such
thing as free lunch."

Daniel A. Roberts
***Why is AI Hard and Physics Simple?*, 2021**

"It is stupid to claim that birds are better than frogs because they see
farther, or that frogs are better than birds because they see deeper.
The world [. . .] is both broad and deep, and we need birds and frogs
working together to explore it."

Freeman Dyson
***Birds and Frogs*, 2009**

M.C. Escher
Hand with Reflecting Sphere, 1935

Abstract of Dissertation

Artificial Intelligence (AI) and differentiable programming are becoming increasingly central to modern cosmology, enabling scalable inference in high-dimensional, physically constrained models. This dissertation develops robust, fully differentiable forward models spanning multiple cosmological epochs—from early-Universe dynamics to galaxy formation and observational systematics—to accelerate inference. We first investigate Beyond the Standard Model (BSM) cosmology through the theory of cosmological stasis, in which extended epochs of constant cosmological abundances emerge. Using a fully differentiable Boltzmann solver that enables gradient-based optimization and accelerated Bayesian inference conditioned on stasis, we identify a new exponential model that generically produces longer stasis epochs than previous power-law constructions. We then develop differentiable models and neural-network-based emulators for galaxy intrinsic alignments, a key systematic in weak-lensing analyses for Stage IV surveys, within the halo occupation distribution framework, enabling end-to-end gradient-based inference from halo occupation and alignment parameters to the resulting galaxy field and its statistics. Finally, we introduce a symmetry-aware domain adaptation framework based on optimal transport to improve robustness under distributional shifts, allowing AI models trained on simulations or individual surveys to generalize reliably across observational domains. Together, these methods demonstrate how differentiable simulations and domain-aware machine learning can be integrated to enable robust and efficient inference in the era of Stage IV cosmology.

Table of Contents

List of Figures

List of Tables

List of Acronyms

2PCF Two-Point Correlation Function. A statistical measure quantifying spatial correlations between pairs of objects as a function of separation.

AI Artificial Intelligence. The capability of computer systems to perform tasks that typically require human intelligence, such as reasoning, learning, and decision-making.

BAO Baryonic Acoustic Oscillation. Oscillatory features imprinted in the matter distribution by sound waves in the pre-recombination photon-baryon plasma, providing a standard ruler for cosmological distance measurements.

BBN Big Bang Nucleosynthesis. The production of light elements in the early universe during the first few minutes after the Big Bang.

BNAF Block Neural Autoregressive Flow. A normalizing flow architecture that models probability distributions via autoregressive transformations.

BSM Beyond the Standard Model. Theoretical frameworks (such as string theory) extending the Standard Model of particle physics to address phenomena it cannot explain.

CDF Cumulative Distribution Function. A function that gives the probability that a random variable takes a value less than or equal to a specified value.

CMB Cosmic Microwave Background. The thermal radiation left over from the early universe, observable today as nearly uniform microwave radiation across the sky.

CNN Convolutional Neural Network. A neural network architecture designed to process data with spatial structure using convolutional layers.

DA Domain Adaptation. A class of methods that improve model performance under distributional shifts between training and deployment data.

DESI Dark Energy Spectroscopic Instrument. A spectroscopic survey designed to measure the expansion history of the universe.

DL Deep Learning. A subfield of machine learning based on neural networks with multiple hidden layers that learn hierarchical representations from data.

ELBO Evidence Lower Bound. A variational objective function that provides a lower bound on the log marginal likelihood in Bayesian inference.

ENN Equivariant Neural Network. A neural network architecture designed to preserve equivariance of its outputs under specified symmetry transformations of the inputs.

FLRW Friedmann–Lemaître–Robertson–Walker. A class of homogeneous and isotropic solutions to Einstein's field equations used to model the expanding universe.

GR General Relativity. Einstein's theory of gravitation describing gravity as the curvature of space-time caused by mass and energy.

GZ Galaxy Zoo. A citizen science project providing morphological classifications of galaxies.

HMC Hamiltonian Monte Carlo. A Markov Chain Monte Carlo method that uses Hamiltonian dynamics to efficiently sample (via gradients) from high-dimensional probability distributions.

HOD Halo Occupation Distribution. A statistical model describing how galaxies populate dark matter halos as a function of halo properties.

IA Intrinsic Alignment. The correlation of galaxy shapes induced by local tidal fields rather than gravitational lensing.

JS Jensen–Shannon. A symmetrized and smoothed version of the Kullback–Leibler divergence used to compare probability distributions.

MCMC Markov Chain Monte Carlo. A class of algorithms for sampling from probability distributions using Markov chains.

ML Machine Learning. Algorithms and statistical models that enable computers to perform tasks by learning patterns from data without explicit programming.

MLP Multilayer Perceptron. A fully connected feedforward neural network composed of multiple hidden layers.

MMD Maximum Mean Discrepancy. A kernel-based statistical distance used to compare probability distributions.

MRE Matter–Radiation Equality. The cosmological epoch at which the energy densities of matter and radiation are equal.

MVE Mean–Variance Estimation. A probabilistic modeling approach in which a model predicts both the mean and variance of a target distribution, parameterizing it as a Gaussian.

NFW Navarro–Frenk–White. A widely used analytic density profile describing the radial structure of dark matter halos.

NN Neural Network. A computational model composed of interconnected layers of nodes that learn representations from data through optimization.

NRMSE Normalized Root Mean Square Error. A normalized error metric measuring prediction accuracy relative to the scale of the data.

NUTS No-U-Turn Sampler. An adaptive variant of Hamiltonian Monte Carlo that automatically tunes trajectory lengths.

ODE Ordinary Differential Equation. An equation involving functions of a single independent variable and their derivatives.

OOD Out-of-Distribution. Data drawn from a distribution that differs from the training distribution of a machine learning model.

OT Optimal Transport. A mathematical framework for comparing probability distributions by minimizing the cost of transporting probability mass from one distribution to another.

PDF Probability Distribution Function. A function that describes the relative likelihood of a continuous random variable taking a given value.

PSF Point Spread Function. The response of an imaging system to a point source, characterizing image blurring.

RMSE Root Mean Square Error. A standard metric quantifying the average magnitude of prediction errors.

SBI Simulation-Based Inference. A class of methods that use forward simulations and neural networks to learn an implicit likelihood, posterior, or parameter–data mapping, enabling parameter inference when the likelihood function is intractable or unavailable.

SCC Spearman Correlation Coefficient. A nonparametric measure of rank correlation between two variables.

SDSS Sloan Digital Sky Survey. A major astronomical survey that has mapped large portions of the sky in multiple wavelengths.

SIDDA Sinkhorn Dynamic Domain Adaptation. A domain adaptation framework using dynamic Sinkhorn divergences to align latent representations.

SMAPE Symmetric Mean Absolute Percentage Error. A scale-invariant error metric commonly used for regression evaluation.

SVI Stochastic Variational Inference. A scalable variational inference method wherein a distribution from a chosen variational family is optimized towards the posterior by using gradient-based optimization to maximize the evidence lower bound.

WL Weak Lensing. The subtle distortion of distant galaxy images caused by the gravitational deflection of light by intervening matter distributions.

Chapter 1

Basic Principles of Cosmology and Artificial Intelligence

> *Sober minds have always known that it pays to be patient before pronouncing judgment on ideas as lofty as those necessary to understand our Universe.*
>
> — Scott Dodelson

Cosmology is among the oldest sciences. Humanity has sought to model the cosmos since the earliest days of civilization, seeking to connect the movement of the heavens with phenomena on Earth. Yet it is only within the last century that we have come to understand the dynamics of the Universe—and our place within it—at a level of scientific rigor. Our perception of the Universe's scale has undergone a dramatic transformation over that time: from the revelation in the 1920s that other galaxies existed (kpc[1] scale), to mapping the cosmic web across tens to hundreds of megaparsecs (Mpc), to modern surveys that now chart structure out to gigaparsec (Gpc) scales. Moreover, we can now describe the Universe and all its complexity with a few free parameters via the ΛCDM cosmological model, which are constrained to within a few-percent accuracy. Though we are in no position to pronounce absolute judgment on all aspects of cosmology, we have waited patiently for millennia to arrive at an inflection point—*one we are currently living through*—where such a judgment begins to feel within reach.

At the turn of the 20th century, the Universe was believed to be static and eternal; Einstein himself stood as a prominent proponent of this view, despite its roots in theoretical preference rather than empirical evidence. With Edwin Hubble's discovery of the expanding Universe,

[1] 1kpc $\approx 3 \cdot 10^{19}$ meters

a Pandora's box in cosmology was unleashed. Humanity's oldest—and perhaps most comfortable—assumptions about the Universe were overturned. Namely, we recognized that the Universe no longer *always was* nor *always will* be. This revolution was catalyzed by theoretical contributions from Albert Einstein, Georges Lemaître, and Alexander Friedmann, alongside technological advances that enabled increasingly precise astronomical instrumentation. Foundational discoveries such as cosmic expansion [10], late-time acceleration [11, 12], the Cosmic Microwave Background (CMB) [13], and gravitational lensing [14] have played a critical role in guiding theoretical progress, transforming cosmology into a precision science. Moreover, our model rests upon only a handful of assumptions—namely, that General Relativity (GR) constitutes a valid theory of gravity at macroscopic scales, and that the Universe is homogeneous and isotropic on sufficiently large scales. As a testament to its recency, we note that the taxonomy of *cosmology* has not yet reflected its precision (why do we not call it *cosmonomy*?[2]).

This experience is not unique to cosmology. Modern computer science, driven by equally modern advances in computer hardware, has undergone a parallel transformation in both scope and capability. Advances in computational physics have enabled simulations of the Universe at its largest scales, incorporating phenomenology from single-atom interactions to the formation of cosmic structure. Simultaneous advances in machine learning over the past 30 years have propelled Artificial Intelligence (AI) systems from proof-of-concept digit classifiers to indispensable tools throughout the scientific pipeline as of 2026, ranging from hypothesis generation to methodology and analysis (see [16] for an example spanning all of the above). At the cusp of the Stage IV data revolution spurred by surveys such as *Rubin* [17], *Roman* [18], *Euclid* [19], and DESI [20], this thesis provides an overview of theoretical, computational, and AI works and contributions that offer a glimpse into the next century of cosmology—and the questions we may soon be able to answer.

Our contributions span domains from Beyond the Standard Model (BSM) early-universe cosmology to systematics in modern Weak Lensing (WL) surveys, leveraging AI methods as well as differentiable simulations. We also include contributions in pure AI, specifically investigating how to design and train Neural Network (NN) architectures that exhibit enhanced robustness to Out-of-Distribution (OOD) shifts. Our approach leverages symmetries inherent in the data and introduces novel training algorithms that explicitly incorporate these structural properties into the learning process. To this end, the following sections provide the reader with a baseline foundation in cosmology and AI. We begin with an overview of the ΛCDM concordance cosmological model,

[2] See [15] for a discussion of this.

outlining the history of the Universe's evolution and key epochs from cosmic inflation to dark energy domination in the modern day. We will have a particular focus in more recent times, as this is the regime of WL and our focus on galaxy Intrinsic Alignment (IA). We will then provide a basic foundation in AI methods, including simple NNs (the Multilayer Perceptron (MLP)), CNNs, and equivariant models. We conclude by covering the basics of NN training.

1.1 The Standard Model of Cosmology (ΛCDM)

In the following sections we provide a pedagogical introduction to ΛCDM and modern cosmology. We draw inspiration from a number of texts [21], reviews [22], and theses [23].

1.1.1 Foundations of the Standard Model

The ΛCDM model rests on two foundational pillars: Einstein's theory of General Relativity and the *cosmological principle*, which asserts that the Universe is homogeneous and isotropic on sufficiently large scales. In more detail, the principle is the following:

> *The cosmological principle is usually stated formally as 'Viewed on a sufficiently large scale, the properties of the universe are the same for all observers.' This amounts to the strongly philosophical statement that the part of the universe which we can see is a fair sample, and that the same physical laws apply throughout. In essence, this in a sense says that the universe is knowable and is playing fair with scientists.*
>
> — William Keel

Keel's remark highlights an unfortunate circumstance of doing cosmology: since we cannot feasibly perform measurements over the entire observable Universe, we must assume that a sufficiently large sample is representative of the whole through these properties. Stated more plainly, we assume that the Universe looks roughly the same at every point (homogeneity), and that it looks the same in every direction (isotropy). Of course, we do not observe this to be true in daily life ($\ll 1$pc), and the cosmological principle only applies at the largest scales. It is additionally only a statement about space; the Universe is neither homogeneous nor isotropic in time.

While the Universe clearly exhibits structure on scales of galaxies and galaxy clusters, observations of the CMB and large-scale galaxy surveys confirm that on scales exceeding roughly $100\,\text{Mpc}$, the matter distribution of the Universe becomes statistically uniform. This is reflected in the CMB temperature, which is remarkably uniform at $T_{\text{CMB}} = 2.7255 \pm 0.0006\,\text{K}$ [24], with fluctuations at the level of $\delta T/T_{\text{CMB}} \sim 10^{-5}$. The small scale of the fluctuations is indicative

of a smooth Universe at the time that the CMB was emitted. Under these assumptions, there are precisely three spatial geometries consistent with the cosmological principle: flat Euclidean space $\mathbb{R}^3$, the positively curved three-sphere S^3, and the negatively curved hyperboloid H^3. To describe these uniformly, we introduce the comoving radial coordinate χ and define:

$$S_k(\chi) = \begin{cases} \frac{1}{\sqrt{|k|}} \sin(\sqrt{k}\chi) & k > 0 \text{ (spherical)} \\ \chi & k = 0 \text{ (Euclidean)} \\ \frac{1}{\sqrt{|k|}} \sinh(k\chi) & k < 0 \text{ (hyperbolic).} \end{cases} \tag{1.1}$$

In this Universe, one then has the Friedmann–Lemaître–Robertson–Walker (FLRW) metric:

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t) \left[\mathrm{d}\chi^2 + S_k^2(\chi) \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\phi^2\right)\right], \tag{1.2}$$

where $a(t)$ is the dimensionless scale factor and (χ, θ, ϕ) are comoving spherical coordinates. Comoving coordinates are defined such that objects at rest in their local reference frame maintain fixed coordinate positions as the Universe expands; these can be converted to proper distances by multiplying by $a(t)$. For bookkeeping, we use the $(+,-,-,-)$ metric signature, normalize the scale factor such that $a(t_0) = 1$ today, and work in natural units with $c = 1$ throughout.

We can proceed to understand the dynamics of empty space by solving Einstein's field equations, stated as

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.3}$$

where $R_{\mu\nu}$ is the Ricci tensor, R the Ricci scalar, $g_{\mu\nu}$ is the metric tensor, Λ is the cosmological constant, G is Newton's gravitational constant, and $T_{\mu\nu}$ is the stress-energy tensor describing the matter and energy content of spacetime. We model the energy content of the Universe as perfect fluids with energy density ρ and pressure p, related via

$$p = w\rho, \tag{1.4}$$

where the parameter w depends on the type of energy (e.g. matter, radiation, etc.). Further enforcing the FLRW metric, we arrive at two solutions to Einstein's field equations, known as the Friedmann equations [25]:

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{R^2 a^2}, \tag{1.5}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p). \tag{1.6}$$

The Friedmann equations define the dynamics of a homogeneous and isotropic Universe dependent on the type of energy density in the Universe. We have taken the liberty to define the Hubble rate, $H \equiv \dot{a}/a$, which describes the expansion rate of the scale factor. If $\dot{a} > 0$, the distance between any two points is increasing and the Universe is expanding, as is the case with our own. The present-day value is $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ [26], and the inverse $H_0^{-1} \approx 14\,\mathrm{Gyr}$ provides a rough estimate of the age of the Universe. For nearby galaxies at a distance d, the recession velocity v due to expansion is approximately $v = H_0 d$, as was originally established by Hubble [27] and was seminal evidence for the expanding Universe.

We also define the critical density as the energy density required for a flat universe:

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G}. \tag{1.7}$$

The dimensionless density parameters are:

$$\Omega_i \equiv \frac{\rho_i}{\rho_{\text{crit}}}. \tag{1.8}$$

It is at times more convenient to define the *Hubble Parameter* as a function of redshift z:

$$H(z) = \frac{\dot{a}(z)}{a(z)}. \tag{1.9}$$

The redshift effect describes that light traveling through an expanding universe is stretched due to the expansion. As a result, a photon emitted with wavelength λ_1 at time t_1 is observed with wavelength λ_0 at a later time t_0, where

$$\frac{\lambda_0}{\lambda_1} = \frac{a(t_0)}{a(t_1)} = \frac{1}{a(t_1)}. \tag{1.10}$$

The redshift parameter z is defined as the fractional increase in wavelength:

$$1 + z = \frac{1}{a(t_1)}. \tag{1.11}$$

Today we sit at $z = 0$. When $z = 1$, the Universe was half its current size; when $z = 2$, it was one-third. In the standard hot Big Bang picture, the universe began at $t = 0$ and $z = \infty$. The redshift is directly measurable from spectroscopy or photometry of galaxies.

1.1.2 Types of Energy Densities

We now describe the primary components of the cosmic energy budget. ΛCDM allows for three primary energy densities: matter, radiation, and dark energy. The matter energy density ρ_M further

contains baryonic matter (stars, planets, and everything else that is "us"), and dark matter which certainly exists [28–30] but does not couple to the photon (hence being "dark"). The radiation energy density ρ_r includes all relativistically moving particle species. This includes all photons ρ_γ and some neutrinos ρ_ν. Despite having mass, neutrinos can contribute to ρ_r when the temperature T_ν is much larger than the mass m_ν, $T_\nu \gg m_\nu$. The dark energy density ρ_Λ is well described by a cosmological constant (as of 2026) and remains approximately constant in time, driving the late-time accelerated expansion of the Universe.

These energy components and their relative abundances have direct consequences on the Universe's phenomenology and evolution. In an expanding universe, energy conservation is expressed through the continuity equation:

$$\dot{\rho} + 3H(\rho + p) = 0. \tag{1.12}$$

Assuming a constant w, this integrates to:

$$\rho(a) = \rho_0 a^{-3(1+w)}, \tag{1.13}$$

where ρ_0 denotes the present-day density. In a flat universe ($k = 0$) dominated by a single component, the Friedmann equation yields the solution:

$$a(t) \propto t^{2/[3(1+w)]}. \tag{1.14}$$

We thus see that the evolution of the scale factor is *power-law* dependent on the type of energy density via w. In the following sections, we discuss these components in more detail.

Radiation

Radiation specifically refers to relativistic particle species for which the kinetic energy greatly exceeds the rest mass energy, i.e., $p \gg m$. In this limit, the energy-momentum relation $E^2 = p^2 + m^2$ simplifies to $E \approx p$, implying that such particles travel close to the speed of light. For an isotropic gas of relativistic particles, kinetic theory yields the pressure

$$P = \frac{1}{3}\rho, \tag{1.15}$$

which corresponds to an equation of state parameter $w_r = 1/3$. From Equation (1.13), the radiation energy density scales as

$$\rho_r \propto a^{-4}, \tag{1.16}$$

further implying a scale factor dependence of $a(t) \propto t^{1/2}$. Notably, this differs from the naive a^{-3} dependence expected from number density dilution alone. The additional factor of a^{-1} arises from cosmological redshift, which reduces photon energy as the universe expands.

The radiation energy density of the Universe is primarily dominated by the CMB. Since photon temperature scales as $T \propto a^{-1}$, the early Universe was significantly hotter. More generally, this relationship can be expressed in terms of redshift:

$$T(z) = T_0(1 + z). \tag{1.17}$$

Other relativistic species include neutrinos, which decoupled from the primordial plasma at $T \sim 1\,\text{MeV}$ and now constitute a cosmic neutrino background. Because they decoupled before electron-positron annihilation, neutrinos did not receive the entropy transferred to photons during that event, resulting in a lower present-day temperature of $T_\nu \approx 1.95\,\text{K}$ [31]. Today, the total radiation density parameter is $\Omega_r \approx 9 \times 10^{-5}$ [3]—negligible relative to matter and dark energy, but dominant during the early stages of cosmic evolution.

Matter

Matter refers to non-relativistic species, where $p \ll m$ and the energy is dominated by rest mass, $E \approx m$. For such species, the velocities are small compared to the speed of light, and the pressure is negligible compared to the energy density corresponding to $w = 0$. The matter energy density scales as

$$\rho_m \propto a^{-3}, \tag{1.18}$$

reflecting a simple dilution proportional to volume increasing. This also implies that in a matter-dominated universe, $a(t) \propto t^{2/3}$. We note that this scaling is less drastic than that of radiation, indicating that a radiation-dominated Universe would preferentially evolve into matter-domination, as we will later see.

The matter content of the Universe further divides into two components with dramatically different properties:

Baryonic Matter ($\Omega_b \approx 0.05$): This encompasses all ordinary matter composed of protons, neutrons, and electrons, which comprises stars, planets, gas, and dust. Despite being the matter we directly interact with in day-to-day life, baryons constitute only about 5% of the total energy density. The baryon density is precisely constrained by Big Bang Nucleosynthesis (BBN) and CMB observations [3].

Dark Matter ($\Omega_c \approx 0.27$): The dominant matter component does not couple to the photon, revealing itself only through gravitational effects. Evidence for dark matter comes from multiple independent sources: galaxy rotation curves that remain flat at large radii rather than declining as expected from visible matter alone [29], the velocity dispersion of galaxies in clusters [28], gravitational lensing measurements [32], and the pattern of CMB anisotropies. Particularly compelling is the Bullet Cluster, where gravitational lensing reveals that the mass distribution (dominated by dark matter) is spatially offset from the X-ray emitting gas, providing direct evidence for collisionless dark matter [32]. To be consistent with observations, dark matter must further be non-relativistic ("cold") at the time of Matter–Radiation Equality (MRE). Hot dark matter (relativistic species like neutrinos) would free-stream out of overdense regions, erasing structure on small scales, which is disfavored by observational evidence of the large scale structure of the Universe. As of 2026, the particle nature of dark matter remains a mystery, and a primary source of scientific investigation in both particle physics and cosmology. The total matter density parameter is $\Omega_m = \Omega_b + \Omega_c \approx 0.32$ [3].

Dark Energy (Λ?): The most exotic component of the cosmic energy budget is dark energy, which drives the observed accelerated expansion of the Universe. The simplest model is a cosmological constant Λ, equivalent to a fluid with equation of state

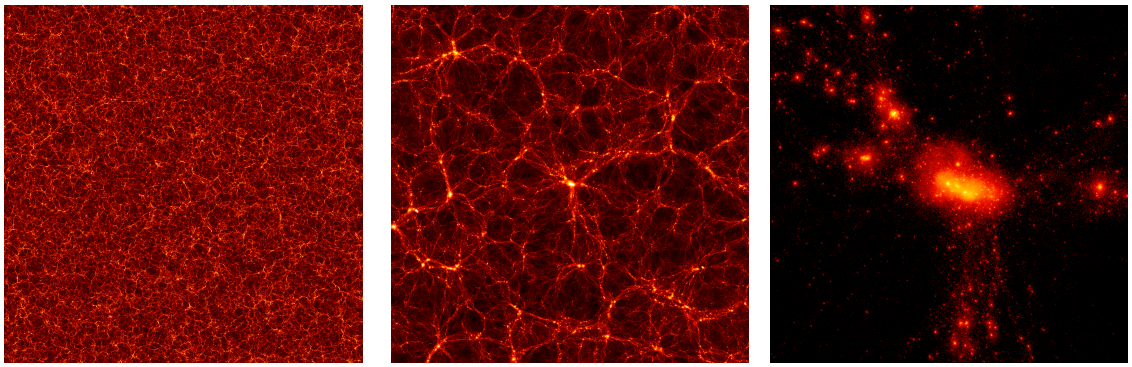
$$p = -\rho, \tag{1.19}$$

corresponding to $w = -1$. From the continuity equation (1.12), a fluid with $\rho + p = 0$ has $\dot{\rho} = 0$, meaning the energy density remains constant as the Universe expands. This is peculiar, and dissimilar from contributions coming from matter or radiation. The cosmological constant thus contributes

$$\rho_\Lambda = \frac{\Lambda}{8\pi G} = \text{const.} \tag{1.20}$$

Because ρ_Λ remains constant while ρ_m and ρ_r dilute with time, dark energy inevitably comes to dominate at late times. The measured value $\Omega_\Lambda \simeq 0.68$ [3] indicates that the Universe has recently entered a dark-energy–dominated phase, in which the expansion rate is accelerating. It is also believed that the cosmological constant is indeed constant (i.e. not time-dependent). We do, however, note that recent results from DESI [33] indicate that a time-dependent form of dark energy $\Lambda(t)$ may more appropriately describe observations better than a time-independent Λ.

For a constant vacuum energy density ρ_Λ, the Friedmann equations admit an asymptotic de Sitter solution with equation of state $w = -1$ and scale factor evolution $a(t) \propto e^{H_\Lambda t}$, where



(a) 2 Gpc/h (b) 250 Mpc/h (c) 20 Mpc/h

Figure 1.1: Simulation snapshots from the AbacusSummit N-body simulation suite [1] with the physical scale decreasing from left to right. The luminous regions indicate dark matter halos and subhalos, while voids are shown to be darker. The snapshots are taken at $z = 0.1$ and are 10 Mpc/h deep. We see a visual indication of statistical homogeneity and isotropy at the largest scale, which is less apparent at 250 Mpc/h and no longer valid at 20 Mpc/h .

$H_\Lambda^2 = 8\pi G\rho_\Lambda/3$. The cosmological constant is theoretically well motivated within GR, since any constant shift in the stress–energy tensor gravitates and is observationally indistinguishable from vacuum energy. With these constituents defined, we can return the Friedmann equations and study the geometry of our Universe with hindsight of the abundances of individual components. With the curvature parameterized as $\Omega_k = -k/(R^2H^2a^2)$, the Friedmann equation becomes:

$$\frac{H^2}{H_0^2} = \frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda. \tag{1.21}$$

By construction, $\Omega_r + \Omega_m + \Omega_k + \Omega_\Lambda = 1$. Observations from the CMB and BAO indicate $|\Omega_k| \lesssim 10^{-3}$, consistent with spatial flatness [3].

1.1.3 The Cosmic Web and Large Scale Structure

While the cosmological principle asserts homogeneity and isotropy on scales exceeding ~ 100 Mpc, the Universe on smaller scales is richly structured. Galaxies are not distributed uniformly, but instead trace a complex network known as the *cosmic web*, consisting of interconnected filaments, sheets, and nodes surrounding underdense regions called voids [34] as illustrated for various scales in Figure 1.1.[3] This structure can be traced to the primordial density perturbations, wherein regions

[3]One may note a striking resemblance between the cosmic web and biological neural networks as a testament to the beauty of Nature's self-similarity.

with slight overdensities attracted surrounding matter, growing denser over time, while underdense regions gradually became empty. The result is a highly anisotropic matter distribution that we see on small scales.

The nodes of the cosmic web are gravitationally bound structures called *dark matter halos*, which form through hierarchical clustering. The existence of these halos is directly required for the formation of galaxies like our own. Dark matter halos are also not smooth; they contain substantial numbers of *subhalos*—smaller halos that have fallen into a larger host but retain their substructure. Mapping the topography of dark matter is of central importance in cosmology; however, we are left with using visible matter, which is much less abundant, as a tracer for this underlying field.

Baryonic matter falls into these dark matter halos, where it cools, condenses, and forms galaxies. More complicated astrophysical processes further source the production of stars, planets, supernovae, and active galactic nuclei within these galaxies. Modeling this phenomenology analytically is generally intractable due to the complexity and vast range of astrophysical scales involved. Moreover, while the Friedmann equations and linear perturbation theory successfully describe the early Universe and large-scale dynamics, they break down in the nonlinear regime, where $\delta \gtrsim 1$. On scales below $\sim 10\,h^{-1}\,\mathrm{Mpc}$, gravitational collapse becomes nonlinear, baryonic physics dominates, and analytic predictions fail. Numerical simulations have therefore become indispensable for understanding structure formation, and several complementary methodologies have emerged.

N-body simulations such as MILLENNIUM [35], BOLSHOI [36], and ABACUSSUMMIT [37] follow the gravitational dynamics of dark matter, employing $N \gtrsim 10^{10}$ particles to resolve halos across a wide mass range while capturing $\mathcal{O}(\mathrm{Gpc}^3)$ volumes encompassing the cosmic web. Hydrodynamic simulations additionally incorporate baryonic physics: operating on smaller volumes of $\mathcal{O}(100\,\mathrm{Mpc}^3)$, but resolving the internal structure of galaxies down to $\sim$ kpc scales and including various feedback processes. Suites such as TNG [38], EAGLE [39], and SIMBA [40] have achieved remarkable success in reproducing observed galaxy populations, morphologies, and scaling relations. Despite their impressive capabilities, hydrodynamic simulations can be costly, and additionally disagree considerably depending on the subgrid physical models used [41]. Complementing these physics-forward approaches, Halo Occupation Distribution (HOD) models [6, 42, 43] provide an empirical framework that statistically populates dark matter halos with galaxies according to parametrized prescriptions calibrated to observations. Given a halo catalog from an N-body simulation, HOD methods assign central and satellite galaxies based on halo mass, enabling rapid generation of mock galaxy catalogs. While incorporating less phenomenology than hydrodynamic simulations, HOD models are computationally efficient and effectively capture the galaxy-halo con-

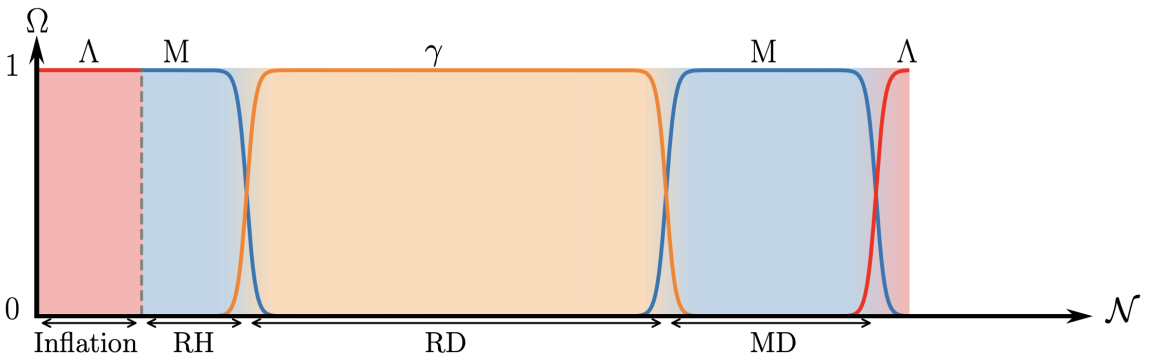


Figure 1.2: Schematic diagram of the Universe's evolution in e-folds $\mathcal{N}$, from Inflation ($t \approx 10^{-32}\,\mathrm{s}$), reheating ($t \approx 10^{-32}\,\mathrm{s}$), radiation domination ($t \approx 1\,\mathrm{s}$), matter domination ($t \approx 5 \times 10^4\,\mathrm{yr}$), and the current era of dark energy domination ($t \approx 11\,\mathrm{Gyr}$). Figure taken from [2].

nection relevant for large-scale clustering statistics. These simulation methodologies underpin modern observational cosmology.

1.2 The History of the Universe

With these theoretical ingredients in mind, we can now proceed to work backwards and study the history of the Universe, extrapolating back to $t = 0$. We call this the Big Bang, which marks the beginning of our Universe that subsequently evolved according to the laws of physics. This history is quite rich, and almost entirely deduced from assuming a homogeneous and isotropic universe that obeys the Friedmann equations. Several aspects, however, were guided as much by theory as by observation. For example, the idea of a finite Universe with a definitive beginning was not considered standard until the CMB, which was direct evidence for a hot big bang model [13]. We will proceed to discuss several important epochs of the Universe's evolution, including inflation, the hot big bang, radiation domination, matter domination, and the current era, which is dark energy dominated. A schematic timeline of this evolution is visualized in Figure 1.2.

1.2.1 Inflation

The Big Bang model successfully describes the evolution of the Universe from the first fraction of a second onward, but taken at face value it suffers from severe fine-tuning problems. The *horizon problem* arises from causality. The particle horizon, which is the maximum distance light could have traveled since the Big Bang, was much smaller at early times than the scales we observe in

the CMB. Regions separated by more than a few degrees on the sky were never in causal contact, yet they share the same temperature to one part in 10^5 as we observe in the CMB. How, then, did the CMB become so nearly uniform? In addition, the *flatness problem* concerns the curvature term in the Friedmann equation. We observe $|\Omega_k| \lesssim 10^{-3}$ today, but Ω_k grows relative to matter and radiation as the Universe expands. Extrapolating backward, $|\Omega_k|$ must have been tuned to better than one part in 10^{60} at the Planck time, posing another fine-tuning problem.

The inflationary paradigm, proposed independently by Guth [44] and Linde [45], resolves these issues by postulating a period of accelerated expansion in the very early Universe, driven by a scalar field ϕ called the inflaton [46]. In the simplest models, the inflaton evolves in a potential $V(\phi)$ according to the Klein-Gordon equation in an expanding FLRW background:

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0. \tag{1.22}$$

The slow-roll approximation assumes that the potential is sufficiently flat such that $\ddot{\phi} \ll 3H\dot{\phi}$ and $\dot{\phi}^2 \ll V(\phi)$, allowing the field to slowly roll down its potential while driving quasi-exponential expansion [47]. These conditions are quantified by the slow-roll parameters

$$\epsilon_V = \frac{M_{\rm Pl}^2}{2}\left(\frac{V'}{V}\right)^2, \quad \eta_V = M_{\rm Pl}^2\frac{V''}{V}, \tag{1.23}$$

where $M_{\rm Pl}$ is the reduced Planck mass. Inflation occurs when $\epsilon_V \ll 1$ and $|\eta_V| \ll 1$. This phase seeds the density perturbations we observe in the CMB and provides the initial conditions for structure formation. Inflation ends when the slow-roll conditions are violated and the inflaton field oscillates about the minimum of its potential. These oscillations decay, transferring the inflaton's energy to Standard Model particles through a process called reheating, which initiates the hot Big Bang.

1.2.2 Big Bang Nucleosynthesis

At temperatures $T \gtrsim 1\,\text{MeV}$ (corresponding to $t \lesssim 1\,\text{s}$), the Universe was a hot plasma of protons, neutrons, electrons, positrons, neutrinos, and photons in thermal equilibrium. Weak interactions maintained chemical equilibrium between neutrons and protons. As the Universe cooled, the weak interaction rate dropped below the expansion rate H, and neutrons "froze out" at $T \sim 0.8\,\text{MeV}$ [48]. Though unstable, the Universe was still hot enough for nuclear fusion to occur.

BBN proceeded over $t \approx 3$–$20\,\text{min}$, synthesizing light elements [48]. Nearly all neutrons ended up in helium-4, the most tightly bound light nucleus. BBN produced primordial abundances

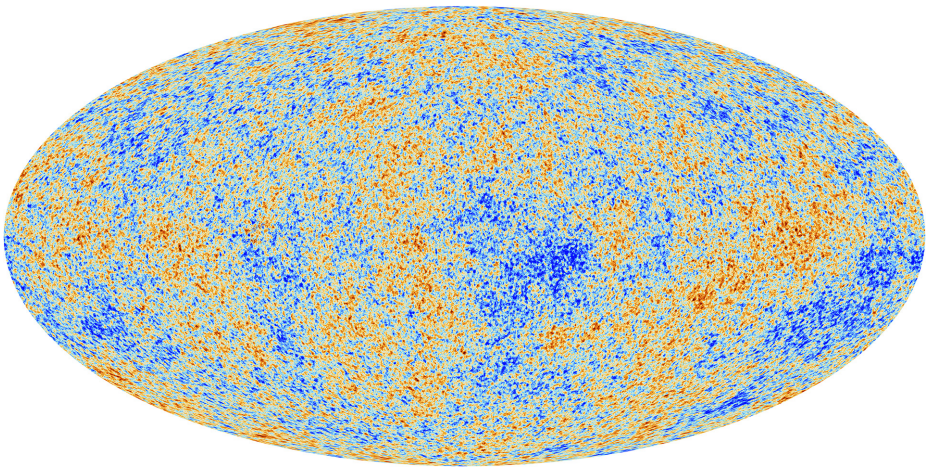


Figure 1.3: The cosmic microwave background as observed by the *Planck* satellite [3]. The color scale represents temperature fluctuations of order $\delta T/T \sim 10^{-5}$ about the mean temperature $T_{\text{CMB}} = 2.7255\,\text{K}$. These anisotropies encode information about the primordial density perturbations seeded during inflation, which subsequently grew under gravitational instability to form the large-scale structure we observe today.

of approximately 75% hydrogen and 25% helium-4 by mass, with trace amounts of deuterium ($\sim 10^{-5}$), helium-3 ($\sim 10^{-5}$), and lithium-7 ($\sim 10^{-10}$) (all values obtained from [49]). These predictions depend sensitively on the baryon-to-photon ratio $\eta = n_b/n_\gamma \approx 6 \times 10^{-10}$, which is independently constrained by CMB observations. The agreement between predicted and observed primordial abundances provides one of the strongest confirmations of the hot Big Bang model.

Importantly, much of the dynamics of the Universe are constrained from BBN onward. This does not preclude, however, any changes to the Universe's chronology pre-BBN that may arise from BSM physics. In Chapter 2 of this thesis, we study cosmological stasis, wherein large towers of states (such as those predicted in string theory) decay coherently such that the total matter abundance of the Universe remains constant despite cosmological expansion. An epoch of stasis can have phenomenological implications on structure formation, and can therefore be tested observationally.

1.2.3 Radiation Domination

Following BBN, the Universe entered an extended period of radiation domination, lasting from $t \sim 20\,\text{min}$ until $t \sim 50{,}000\,\text{yr}$. The scale factor evolved as $a(t) \propto t^{1/2}$, corresponding to a decelerating expansion with Hubble parameter $H = 1/(2t)$. Initially, electrons and positrons remained in thermal equilibrium with photons through electromagnetic interactions. As the temperature dropped

below $T \sim 0.5\,\text{MeV}$, electron-positron pairs annihilated, transferring their entropy to the thermal bath. Throughout radiation domination, dark matter and baryonic matter remained subdominant but grew in relative importance. The transition from radiation to matter domination occurred at MRE, when $\rho_m = \rho_r$, at a redshift $z_{\text{eq}} \approx 3400$ [3]. This epoch is crucial for structure formation: density perturbations could only begin growing significantly after matter domination, as radiation pressure suppressed gravitational collapse during radiation domination.

1.2.4 Matter Domination

After MRE, the Universe transitioned to matter domination, during which non-relativistic matter (both dark matter and baryons) dominated the expansion dynamics. This epoch persisted from $z \sim 3400$ until dark energy began to dominate at $z \sim 0.3$ (corresponding to $t \sim 10\,\text{Gyr}$). A pivotal event during matter domination was recombination, occurring at $z \sim 1100$ (or $t \sim 380{,}000\,\text{yr}$) when the temperature dropped to $T \sim 0.3\,\text{eV}$. At this point, protons and electrons combined to form neutral hydrogen. Shortly thereafter, photons decoupled from matter producing the CMB radiation we observe today, as shown in Figure 1.3, providing a snapshot of the Universe at this decoupling. The tiny temperature anisotropies in the CMB ($\Delta T/T \sim 10^{-5}$) encode information about the primordial density perturbations seeded during inflation.

Following recombination, dark matter overdensities, which had been growing since MRE, continued to collapse gravitationally and formed the first dark matter halos. Baryons, now decoupled from radiation pressure, fell into these gravitational wells. Eventually, the first stars and galaxies formed during the epoch of reionization ($z \sim 6$–20), when their ionizing radiation reheated and reionized the intergalactic medium. This marked the beginning of cosmic structure as we observe it today, with microscopic physics interacting with gravity to produce structure at consecutively larger scales. This initiates the era of galaxies, galaxy clusters, and the cosmic web.

1.2.5 Late Times and Structure Formation

The late-time Universe ($z \lesssim 2$) is characterized by the emergence of large-scale structure and the transition to dark energy domination. This epoch is probed by Stage IV surveys and is the primary focus of observational cosmology. In 1998, observations of Type Ia supernovae revealed that the expansion of the Universe is accelerating [12, 50]. These standardizable candles appeared fainter than expected, implying larger distances than a decelerating universe would produce. Combined with CMB [30] and BAO measurements [51, 52], this established the ΛCDM concordance model.

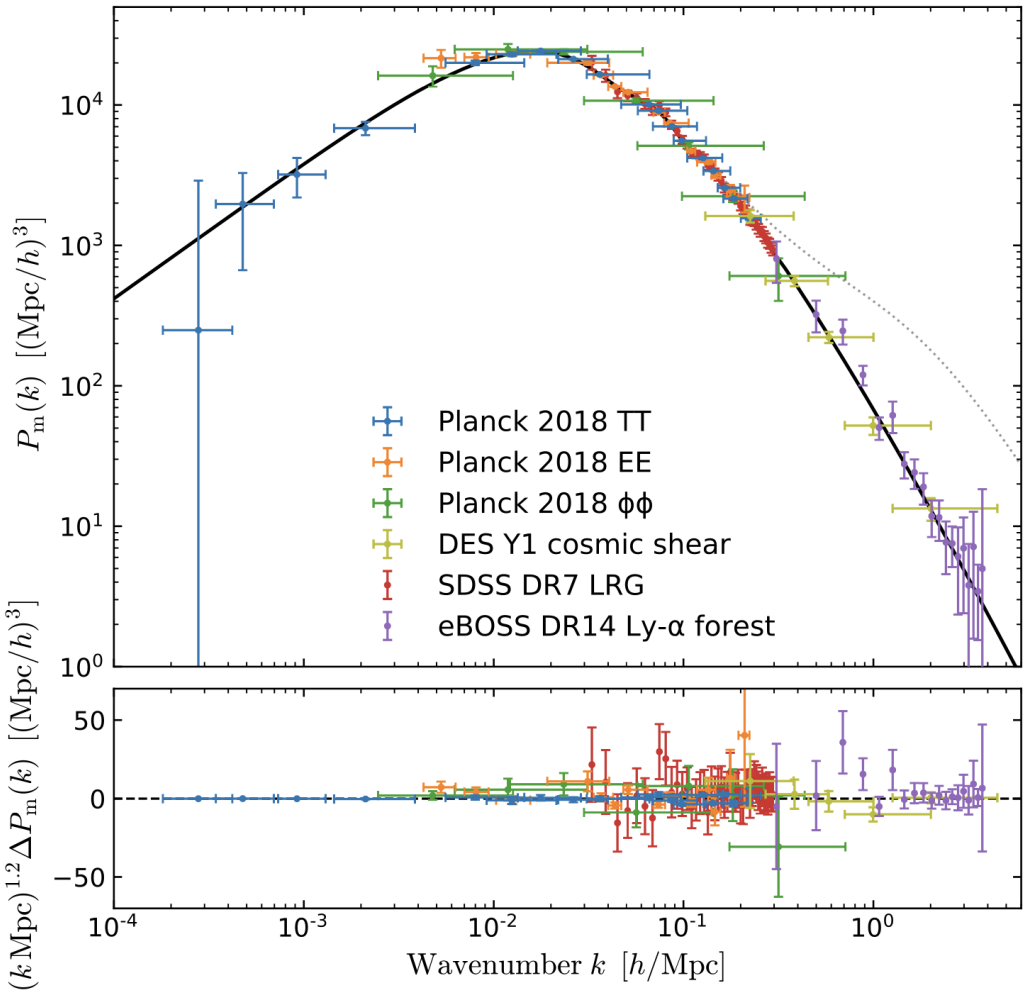


Figure 1.4: The linear matter power spectrum $P(k)$ at $z = 0$. The spectrum rises as $P(k) \propto k^{n_s}$ on large scales (small k), turns over near the scale corresponding to matter-radiation equality ($k_{\rm eq} \sim 0.01\,h\,{\rm Mpc}^{-1}$), and falls as $P(k) \propto k^{n_s-4}$ on small scales due to the suppression of growth during radiation domination. The baryon acoustic oscillations appear as small wiggles at $k \gtrsim 0.05\,h\,{\rm Mpc}^{-1}$, imprinting the sound horizon scale at recombination. On scales below $k \sim 0.1\,h\,{\rm Mpc}^{-1}$, nonlinear gravitational evolution enhances power beyond the linear prediction shown here. Image taken from [4].

The "Λ" was chosen as the acceleration equation (1.6) shows that acceleration requires $\rho + 3p < 0$, which is satisfied for $w < -1/3$. A cosmological constant with $w = -1$ easily satisfies this condition, and indeed has been in remarkable agreement with observations for nearly three decades [53].

While the Friedmann equation describes the homogeneous background, structure formation requires understanding perturbations. In the linear regime ($\delta \equiv \delta\rho/\bar{\rho} \ll 1$), density perturbations in pressureless matter obey

$$\ddot{\delta} + 2H\dot{\delta} - 4\pi G\bar{\rho}\delta = 0. \tag{1.24}$$

The solutions of this equation govern the time evolution of structure in this regime. It is also interesting to note that there is no spatial derivatives or position-dependence present, meaning that the initial conditions of the large scale structure was codified much earlier. We can then express overdensities at later times as a function of the primordial perturbations (δ_0),

$$\delta(\chi, a) = D(a)\delta_0(\chi). \tag{1.25}$$

During matter domination, structures grow linearly with the scale factor, $D(a) \propto a$. This structure growth eventually ends with the advent of dark energy domination.

An invaluable statistic for analyzing large-scale clustering is the matter power spectrum $P(k)$, shown in Figure 1.4, which quantifies the amplitude of density fluctuations as a function of spatial scale. In Fourier space, the density contrast $\delta(\mathbf{x}) = (\rho(\mathbf{x}) - \bar{\rho})/\bar{\rho}$ has Fourier transform $\tilde{\delta}(\mathbf{k})$, and the power spectrum is defined through the two-point correlation function

$$\langle\tilde{\delta}(\mathbf{k})\tilde{\delta}^*(\mathbf{k}')\rangle = (2\pi)^3\delta_D(\mathbf{k} - \mathbf{k}')P(k). \tag{1.26}$$

Larger values of $P(k)$ indicate stronger clustering at the corresponding spatial scale $\lambda \sim 2\pi/k$. The overall shape of the power spectrum encodes key physical processes. On large scales, $P(k)$ retains the nearly scale-invariant primordial form $P(k) \propto k^{n_s}$ generated during inflation [30]. On smaller scales, perturbations that entered the horizon during radiation domination experienced suppressed growth, as radiation pressure prevented gravitational collapse. This introduces a characteristic peak at the scale corresponding to matter-radiation equality, beyond which $P(k)$ decreases with increasing k. Additional features include the Baryonic Acoustic Oscillation (BAO)—subtle oscillatory imprints from sound waves in the pre-recombination photon-baryon plasma—which serve as a standard ruler for cosmological distance measurements [51]. The amplitude of matter fluctuations is conventionally characterized by $\sigma_8 \approx 0.81$ [30], defined as the root-mean-square density fluctuation in spheres of radius $8\,h^{-1}$ Mpc.

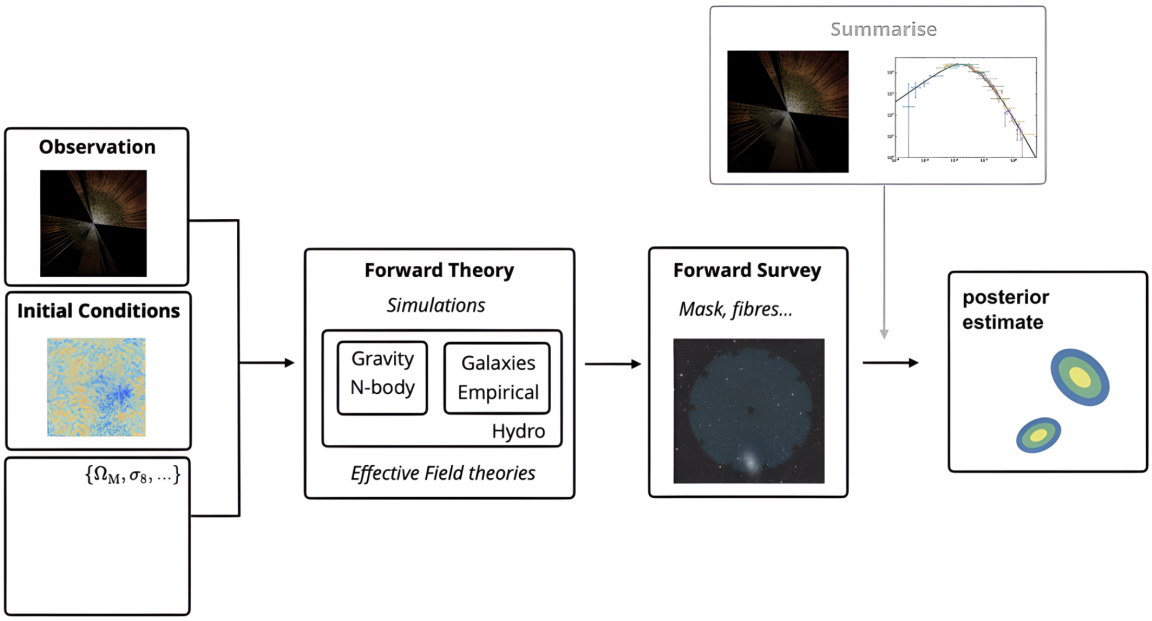


Figure 1.5: Schematic of the forward modeling approach to cosmological inference. Cosmological parameters $\{\Omega_m, \sigma_8, \ldots\}$ and initial conditions are propagated through a forward model comprising gravitational dynamics (N-body simulations or effective field theories), galaxy formation physics (hydrodynamical simulations or empirical models such as HOD), and survey-specific effects to produce mock observations. These are compressed into summary statistics (e.g., the matter power spectrum) and compared against real observations to obtain posterior constraints on cosmological parameters. The computational expense of this pipeline—particularly the simulation and summary statistic stages—motivates the development of emulators and differentiable forward models discussed in this thesis. Figure courtesy of Carolina Cuesta-Lazaro.

1.3 Bayesian Inference in Cosmology

Equipped with a framework for cosmology, we now turn to statistics. To be a cosmologist is to be Bayesian; we do not have the luxury of performing repeated experiments across multiple Universes to reach statistical precision, as the *frequentist* paradigm would require. Constrained to a single Universe, and only a limited set of observational probes within it, we adopt a Bayesian framework wherein beliefs about model parameters are continuously updated as new data become available. The goal of cosmological inference is thus to constrain model parameters $\boldsymbol{\theta}$ given observed data $\mathbf{d}$.

In the modern setting, this pipeline begins with an initial guess for the Universe's cosmological parameters (e.g. Ω_m, σ_8, etc.) as well as information for the primordial density field

(e.g. from the CMB). This information is forward-modeled through simulation-based and analytic calculations and then compressed into summary statistics such as $P(k)$ or two-point correlation functions. Comparing these predictions to observations yields a posterior over cosmological parameters (Figure 1.5). This pipeline is visualized in Figure 1.5. Bayes' theorem provides the foundation for this analysis:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{d})}, \tag{1.27}$$

where $p(\boldsymbol{\theta}|\mathbf{d})$ is the *posterior* distribution, $p(\mathbf{d}|\boldsymbol{\theta})$ is the *likelihood*, and $p(\boldsymbol{\theta})$ is the *prior*. The denominator $p(\mathbf{d})$ is known as the *Bayesian evidence* and is the reason Bayes' theorem cannot typically be solved analytically, as it requires integration over the entire parameter space ($\boldsymbol{\theta}$):

$$p(\mathbf{d}) = \int p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{1.28}$$

In principle, the posterior contains all information about the parameters extractable from the data. In practice, evaluating and exploring this posterior presents computational challenges.

1.3.1 Two-Point Statistics and the 3×2pt Framework

Modern WL surveys extract cosmological information through correlation functions of galaxy positions and shapes. The standard approach combines three two-point correlations: cosmic shear (shape-shape), galaxy clustering (position-position), and galaxy-galaxy lensing (position-shape). This combination, known as the 3×2pt analysis, has become the workhorse of modern WL survey cosmology [54–56].

The appeal of two-point statistics lies in their well-understood theoretical foundation. For a Gaussian random field, two-point correlations capture all statistical information. The primordial density field emerging from inflation is indeed nearly Gaussian, with non-Gaussianities constrained to be small by CMB observations. This motivates a Gaussian likelihood for the data vector:

$$\ln p(\mathbf{d}|\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \mathbf{C}^{-1} (\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta})) + \text{const}, \tag{1.29}$$

where $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the theoretical prediction and $\mathbf{C}$ is the covariance matrix. However, gravitational evolution fundamentally breaks the Gaussian assumption. On scales below $\sim 10\,h^{-1}$ Mpc, mode coupling induces significant non-Gaussianity in the matter distribution. The WL convergence field inherits this non-Gaussianity, therefore requiring the entire hierarchy of moments that two-point statistics cannot access for a full description. Studies have shown that higher-order statistics—including the bispectrum [57, 58], peak counts [59, 60], Minkowski functionals [61], and the convergence probability distribution function [62]—can improve constraints on Ω_m and σ_8 considerably

compared to power spectrum analysis alone [63]. The 3×2pt framework thus provides a conservative, well-validated baseline, but leaves substantial cosmological information to be leveraged.

1.3.2 The Computational Challenge of MCMC

Traditional approaches sample the posterior using Markov Chain Monte Carlo (MCMC) methods, typically the Metropolis-Hastings algorithm or variants thereof [64, 65]. MCMC constructs a Markov chain whose stationary distribution is the target posterior; after sufficient iterations, samples from the chain approximate draws from the posterior. A typical cosmological analysis requires 10^4–10^5 chain elements to achieve convergence and adequate sampling of parameter degeneracies [66]. The computational cost of MCMC scales unfavorably with both the dimensionality of the parameter space and the expense of likelihood evaluation. A Stage III 3×2pt analysis involves $\mathcal{O}(20\text{–}30)$ parameters: cosmological parameters, IA parameters, galaxy bias parameters, and nuisance parameters [67]. Each likelihood evaluation requires computing theoretical predictions across the full data vector. This calculation that can take seconds to minutes, depending on the required accuracy. A single MCMC analysis thus requires hours to days of wall-clock time on modern hardware [68].

The situation becomes increasingly dire for Stage IV surveys. The parameter space expands to $\mathcal{O}(40\text{–}100)$ dimensions as additional systematic effects must be marginalized over, while the precision requirements on theoretical predictions tighten. Joint analyses combining multiple surveys or probes push dimensionality even higher; a recent forecast for combined Stage IV 3×2pt analyses found parameter spaces of 157–159 dimensions, for which traditional nested sampling would require a projected 12 years of compute time on 48 CPU cores [69]. This demands more efficient inference algorithm alternatives for Stage IV cosmology.

1.3.3 Gradient-Based Sampling and Differentiable Pipelines

Inference over high-dimensional spaces can be made more manageable if one knows where to look. To this end, gradient-based sampling methods offer a path forward for more efficient inference. Hamiltonian Monte Carlo (HMC) simulates Hamiltonian dynamics on the parameter space, using the gradient of the log-posterior to flow toward regions of high probability while maintaining detailed balance [70, 71]. This guided exploration dramatically improves acceptance rates compared to random-walk proposals. A more detailed explanation of HMC can be found in Chapter 3.

HMC demonstrates superior scaling in high dimensions. This is also enjoyed by sampling techniques such as Stochastic Variational Inference (SVI), wherein one introduces a learnable, parametric distribution that is optimized towards the posterior via gradient-based optimization. Recent cosmological applications have shown HMC outperforming traditional samplers for problems with tens to hundreds of parameters, with particularly strong gains when the posterior exhibits complex geometry or parameter degeneracies [69, 72]. We observe something similar in our own applications within galaxy IA modeling, as detailed in Chapter 3.

The prerequisite for gradient-based sampling is a differentiable forward model that gives one access to $\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{d}|\boldsymbol{\theta})$. This requirement has spurred development of differentiable cosmological pipelines, from Boltzmann codes to power spectrum emulators to summary statistic calculators [72]. This also includes the use of NNs, which can act as surrogates for the likelihood or posterior directly, through frameworks like Simulation-Based Inference (SBI) [73]. When combined with automatic differentiation frameworks, these pipelines enable differentiable sampling with minimal additional implementation cost.

1.3.4 Toward Scalable Inference for Stage IV Surveys

The inference challenges facing Stage IV cosmology motivate the methodological developments presented in this thesis. Traditional MCMC methods, while well-understood and extensively validated, cannot scale to the high-dimensional parameter spaces and precision requirements of next-generation surveys. Three complementary approaches offer paths forward:

- **Emulators** replace expensive forward model evaluations with fast neural network surrogates, enabling orders-of-magnitude speedups in likelihood computation. The IAEMU emulator developed in Chapter 3 exemplifies this approach, providing $\sim 10{,}000\times$ acceleration for intrinsic alignment correlation function predictions. For practical use, these emulators must be robust to perturbations in the data, as well as have the capability to perform across simulations and cosmological parameters. In other words, they must generalize.

- **Differentiable forward models** enable gradient-based sampling methods that scale favorably to high dimensions. The DIFFHOD-IA framework in Chapter 3 demonstrates how differentiable implementations of galaxy-halo connection models unlock HMC sampling for intrinsic alignment inference. The differentiable Boltzmann solver in Chapter 2 demonstrates how one can use differentiable models in purely theoretical spaces to study interesting phenomenology. Because the physics is explicitly encoded, these models are typically less sensitive to

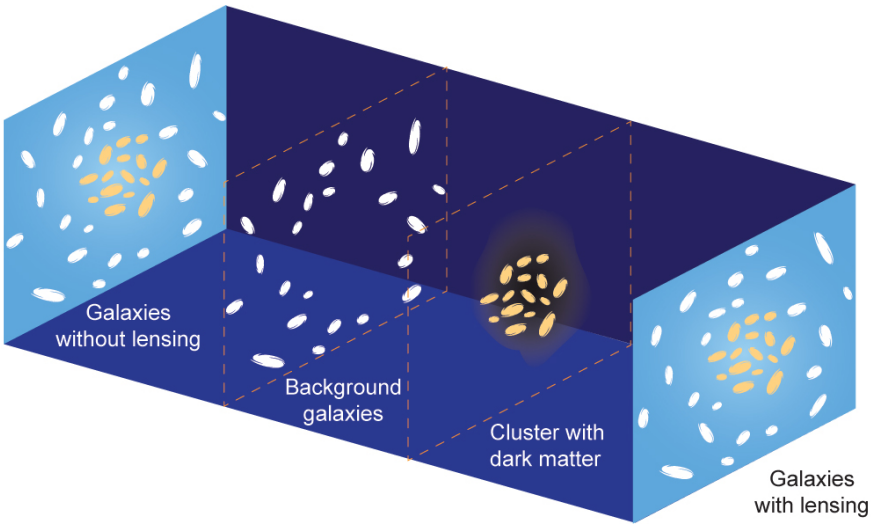


Figure 1.6: Schematic illustration of weak gravitational lensing. Photons emitted from the background galaxies are deflected by the intervening dark matter distribution, causing the lensed galaxy images.

shifts in the underlying data distribution than purely data-driven surrogates. Differentiable simulations, however, generally remain slower than NN surrogates for forward modeling.

- **Simulation-based inference** [73] bypasses the need for explicit likelihoods entirely, enabling inference with arbitrary summary statistics and naturally handling non-Gaussian information. While not the focus of this thesis, SBI methods interface naturally with the emulators and differentiable models we develop.

Together, these methods represent the emerging paradigm for cosmological inference in the Stage IV era. The following section introduces the deep learning foundations underlying these approaches.

1.4 Weak Lensing and Intrinsic Alignments

Weak gravitational lensing has emerged as one of the most powerful probes of the matter distribution in the Universe [74, 75]. Unlike galaxy surveys or the CMB, weak lensing directly probes the total matter distribution (baryonic and dark matter), making it sensitive to both the geometry of the Universe and the growth of structure across cosmic time. As light from distant galaxies travels across the Universe, it is deflected by the gravitational potential of intervening matter, inducing small but measurable distortions in the observed shapes of these galaxies. This is visualized in

Figure 1.6. If we can accurately model how galaxy shapes are distorted, we can infer cosmological parameters. Moreover, since most galaxy shapes are distorted in this way, WL analyses have the ability to employ large catalogs of galaxies for statistical precision.

This, however, introduces its own set of challenges. The shape distortions can be incredibly "weak", requiring robust statistics. These shape measurements also have to be extraordinarily accurate, and are subject to astrophysical and survey design systematics which can be considerable. These challenges require that WL measurements account for both systematics and statistical uncertainties. In this thesis we will address the first challenge via accurate modeling of galaxy IA, and the second problem via modern techniques which enable tractable high-dimensional inference.

1.4.1 Gravitational Lensing Basics

The lensing effect is characterized by the convergence field κ, which describes isotropic magnification, and the shear $\gamma = \gamma_1 + i\gamma_2$, which describes anisotropic stretching of images. The shear is a spin-2 quantity, transforming under rotations by angle ϕ as $\gamma \to \gamma e^{2i\phi}$. For weak lensing ($\kappa, |\gamma| \ll 1$), the observed ellipticity of a galaxy is approximately

$$\epsilon^{\rm obs} \approx \epsilon^{\rm int} + \gamma, \tag{1.30}$$

where $\epsilon^{\rm int}$ is the intrinsic (unlensed) ellipticity. This can be measured by averaging over many galaxies. The key assumption enabling this measurement is that intrinsic ellipticities are randomly oriented, so that $\langle \epsilon^{\rm int} \rangle = 0$ and the average observed ellipticity yields an unbiased estimate of the shear.

The cosmic shear power spectrum for sources in redshift bins i and j is related to the matter power spectrum via the Limber approximation [76, 77]

$$C_\ell^{ij} = \int_0^{\chi_H} d\chi \, \frac{W^i(\chi) W^j(\chi)}{\chi^2} \, P_\delta \left(k = \frac{\ell + 1/2}{\chi}, z(\chi) \right), \tag{1.31}$$

where we have assumed a flat cosmology. Above, $W^i(\chi)$ is the lensing kernel for sources in bin i and χ_H is the comoving distance to the horizon. Cosmic shear is thus sensitive to both the geometry of the Universe (through the distance-redshift relation) and the growth of structure (through $P_\delta(k, z)$), constraining the combination $S_8 \equiv \sigma_8 (\Omega_m / 0.3)^{0.5}$. Modern surveys including DES [54], KiDS [55], and HSC [56] have measured cosmic shear with increasing precision, while Stage IV surveys will reduce statistical uncertainties to the sub-percent level [17–19].

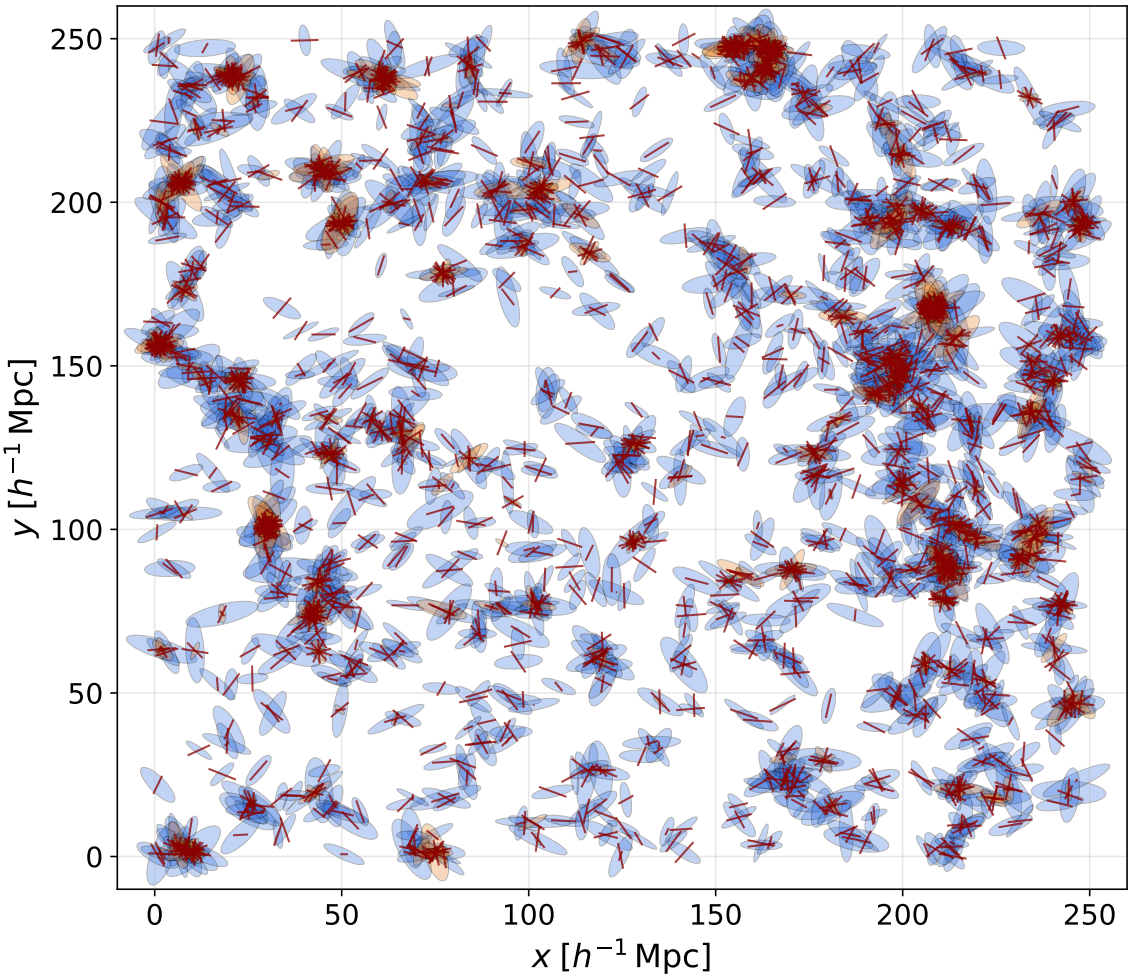


Figure 1.7: An example galaxy field from the DIFFHOD-IA simulation (discussed in Chapter 3) representative of a sample from TNG300 conducted on the BOLSHOI-PLANCK simulation. Red line segments denote galaxy positions, with their orientations reflecting the projected galaxy orientation in the plane of the sky and their lengths proportional to the magnitude of the projected orientation vector. Blue ellipsoids indicate host dark matter halos and orange ellipsoids indicate subhalos, with ellipsoid sizes proportional to halo mass. Intrinsic alignments of galaxies can be qualitatively observed by inspecting more massive (sub)halos hosting multiple galaxies, and are quantified more robustly through galaxy orientation correlation statistics.

1.4.2 Intrinsic Alignments

The assumption that intrinsic ellipticities are randomly oriented is violated if galaxies preferentially align with their environment. This IA effect arises because galaxy shapes are determined by the same large-scale tidal fields that source the lensing signal [78, 79]. Two primary physical mechanisms generate IA: tidal alignment, where elliptical galaxies inherit the shape of their parent dark matter halo which was set by the local tidal field [80, 81], and tidal torquing, where disk galaxies acquire angular momentum from tidal interactions during formation [82, 83]. In Figure 1.7, we show a $1\,\mathrm{Mpc}/h$ slice of a galaxy field (including IA), wherein a relationship between dark matter halo orientations and galaxy orientations can be deduced.

IA contaminates weak lensing measurements through correlations that mimic or obscure the true signal. The observed shape correlation function receives contributions [84]

$$\xi^{\mathrm{obs}}(\theta) = \xi^{GG}(\theta) + \xi^{GI}(\theta) + \xi^{II}(\theta), \tag{1.32}$$

where ξ^{GG} is the true lensing signal we seek to measure. The II term correlates the intrinsic shapes of galaxies that formed in the same tidal environment. This correlation is strongest for galaxy pairs at similar redshifts. The GI term cross-correlates the intrinsic shape of a foreground galaxy with the lensing shear experienced by a background galaxy.

Current analyses employ phenomenological models such as the nonlinear linear alignment (NLA) model [85], with free parameters that are marginalized over in cosmological inference [54, 55]. More sophisticated approaches include the tidal alignment and tidal torquing (TATT) model [86]. See [87] for a full review of the IA formalism and terminology. For Stage IV surveys, IA modeling must achieve percent-level accuracy to avoid biasing cosmological constraints, motivating the development of simulation-calibrated models. The correlation functions relevant for IA modeling are central to Chapter 3 of this thesis.

1.5 Deep Learning, Neural Networks, and All That

The idea of allowing machines to *learn* predates modern Machine Learning (ML) and AI. Linear regression—where, in the simplest case, a two-parameter model is iteratively optimized to fit data—is conceptually not far removed from contemporary learning systems. Modern Deep Learning (DL) is built on large collections of simple computational units, or "neurons", each applying a linear transformation followed by a non-linear activation function. These systems have had a profound impact across the sciences, particularly in cosmology, and form a central focus of this thesis.

We will also see that their learning paradigms can inspire the augmentation of existing simulation frameworks with analogous adaptive capabilities.

The computational challenges of traditional inference methods, combined with the need to model complex astrophysical systematics like IA, motivate the use of modern ML techniques. While ML methods are diverse, modern practice often centers on training NNs. In particular, we focus on the use of NNs as emulators for simulations, to approximate posteriors via sampling techniques like SVI, and for other analysis tasks, such as galaxy morphology classification. In the latter setting, our contributions focus on symmetry-informed architectural design and training techniques that allow trained NNs to generalize across multiple datasets.

1.5.1 Why Neural Networks?

NNs are parametric function approximators composed of layers of neurons. The universal approximation theorem guarantees that sufficiently expressive networks can approximate any continuous function on a compact domain to arbitrary precision. The intrinsic expressivity of these networks is encoded in their *weights*, which can be updated according to data and are therefore *trainable*. These weights can be arranged in different ways, thereby defining the NN *architecture*. Importantly, this result is not restricted to scalar-valued functions, but extends to general mappings between finite-dimensional spaces, $\mathbb{R}^n \to \mathbb{R}^m$, providing the theoretical foundation for the widespread use of NNs across diverse application domains. Moreover, such mappings need not be deterministic nor limited to point estimates; we will therefore discuss *probabilistic* ML and its applicability in cosmology.

Despite this theoretical guarantee, training NNs remains an empirical practice. Furthermore, NNs are typically not robust to perturbations in the data and often fail when applied to OOD tasks [88]. Consequently, one does not automatically benefit from the universal approximation theorem by remaining agnostic about architectural choices. Inductive biases, of which physicists know many, have proven exceptionally useful in designing architectures that are better-suited for certain types of problem by leveraging available symmetries in the data. We will therefore devote time to motivating different architectures, including so-called "equivariant" NNs, which are instilled with symmetry inductive biases. These architectures are studied in 4 in the context of constructing generalizable models that can be transferred across different datasets.

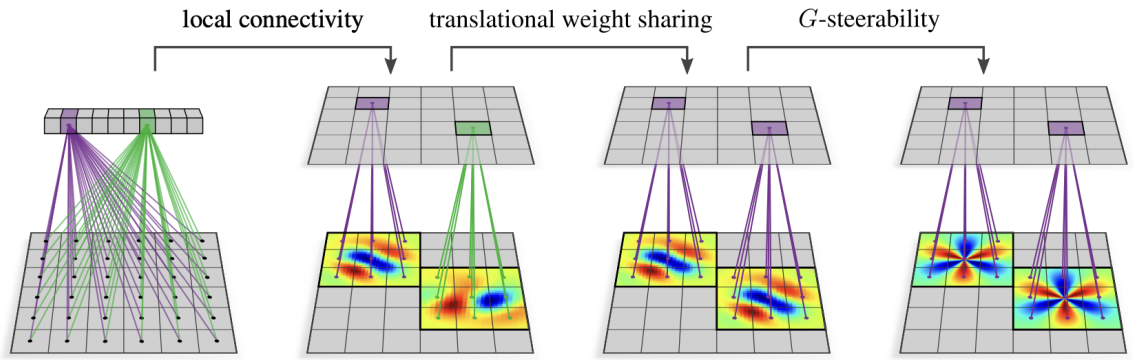


Figure 1.8: Progression of neural network architectures with increasing symmetry constraints. *From left to right:* A fully-connected MLP connects every input to every output, respecting no spatial structure. Imposing local connectivity restricts connections to spatial neighborhoods but uses different weights at each location. Convolutional networks (CNNs) share weights across spatial positions, achieving translation equivariance. Steerable CNNs further constrain the filter kernels to transform predictably under a symmetry group G (e.g., rotations), achieving equivariance to the full affine group $\mathrm{Aff}(G)$. Each additional constraint reduces the parameter space and improves generalization when the corresponding symmetry is present in the data. Image adapted from [5].

1.5.2 Architectures

The Multilayer Perceptron

The canonical NN, modeled after biological NNs, is the MLP [89]. It is also known as a fully-connected (dense) network, is the simplest feedforward architecture. An MLP maps input $\mathbf{x} \in \mathbb{R}^{n^{(0)}}$ to output $\mathbf{z}^{(L)} \in \mathbb{R}^{n^{(L)}}$ through successive transformations:

$$z_i^{(\ell)} = \sigma\left(h_i^{(\ell)}\right), \quad h_i^{(\ell)} = \sum_{j=1}^{n^{(\ell-1)}} W_{ij}^{(\ell)} z_j^{(\ell-1)} + b_i^{(\ell)}, \quad \ell = 1, \ldots, L, \tag{1.33}$$

where $z_i^{(0)} = x_i$, $\mathbf{z}^{(L)}$ is the network output, $W_{ij}^{(\ell)}$ are the weight matrices, $b_i^{(\ell)}$ are the bias vectors, $h_i^{(\ell)}$ is the preactivation at layer ℓ, and $\sigma(\cdot)$ is a nonlinear activation function. Upon inspection, we see that this corresponds to a wide and deep network of linear regression up to the inclusion of σ. The activation function introduces nonlinearity; without it, the composition of linear transformations would remain linear and NNs would not be able to represent nontrivial functions. Common choices for σ include the rectified linear unit $\sigma(z) = \max(0, z)$ (ReLU) [90].

Convolutional Neural Networks

While the MLP is well-suited for vectorized data, its construction does not respect symmetries of structured data. For example, a data vector that is an image would have to be flattened upon passing through an MLP, removing much of the local information of the image. For structured data such as images, the Convolutional Neural Network (CNN) [91] exploits spatial structure through two key architectural principles that distinguish it from the MLP: *locality* and *translation invariance*. The CNN architecture replaces the matrix multiplication in (1.33) with discrete convolution:

$$h_i^{(\ell)} = (W^{(\ell)} * z^{(\ell-1)})_i + b_i^{(\ell)}, \tag{1.34}$$

where $*$ denotes convolution and $W^{(\ell)}$ is a learned filter kernel of finite support. Locality is enforced via nearest neighbor influences in the sliding kernel function.

Locality is enforced as the output depends only on a local neighborhood of inputs through the finite filter kernel. Translation invariance is enforced similarly, as the same filter $W^{(\ell)}$ is applied at every spatial position, making learned features *equivariant* to translations. Together, these inductive biases dramatically reduce the parameter count compared to MLPs while making CNNs more suited to structured data. Successive convolutional layers learn hierarchical features, from edges and textures to objects and scenes. For cosmological applications, CNNs have been applied

to galaxy morphology classification [92], strong lens detection [93], and convergence map analysis [94], to name a few.

Equivariant Neural Networks

Taking inspiration from mathematical convolution, CNNs employ inductive biases to aid learning. That is, CNNs implicitly enforce notions of symmetry that we know exist in images (translation invariance). This reasoning can be extended to high order symmetries via Equivariant Neural Network (ENN)s [5, 95, 96], which can be equivariant to higher-order symmetries.

ENNs generalize convolution to other groups like rotations and reflections. A function $\phi : X \to Y$ is *equivariant* with respect to a group G if:

$$\phi(T_g x) = T'_g \phi(x), \quad \forall g \in G, \tag{1.35}$$

where T_g and T'_g are group actions on the input and output spaces. This construction includes both discrete and continuous groups, with nuances in how the two are handled. In the case of rotational symmetries, it is natural to consider equivariance to the full group of rotations as well as its subgroups, $C_N \subset SO(2)$, and the corresponding case of reflections and rotations, $D_N \subset O(2)$. For cosmological data, relevant symmetries include the above, as well as statistical isotropy of convergence maps, and Euclidean symmetry of point cloud halo catalogs.

By encoding these symmetries architecturally, ENNs reduce the parameter space, are more data efficient, and improve generalization. This is particularly valuable when training data is limited or when the test distribution differs from training. Chapter 4 explores equivariant architectures combined with domain adaptation methods for improved robustness under dataset shift.

1.5.3 Training

Recall that NNs are parametric function approximators, $f_\theta(x)$. The NN parameters $\boldsymbol{\theta}$ are optimized to minimize a loss function $\mathcal{L}$ measuring prediction error on training data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$. The choice of loss function depends on the task. For regression, the mean squared error is common:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\|^2, \tag{1.36}$$

while for classification, the cross-entropy loss is typically used:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log f_{\boldsymbol{\theta}}(\mathbf{x}_i)_k, \tag{1.37}$$

where K is the number of classes and y_{ik} are one-hot encoded labels.

Optimization proceeds via gradient descent, iteratively updating parameters in the direction of steepest descent:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t), \tag{1.38}$$

where η is the learning rate. Gradients are computed efficiently through backpropagation, which applies the chain rule recursively through the computational graph of the network. This graph in practice is constructed through ML frameworks like PyTorch [97] and Jax [98]. Stochastic gradient descent (SGD) additionally uses random mini-batches of data rather than the full dataset, enabling scalability to large datasets. Modern optimizers like Adam [99] adaptively adjust learning rates for each parameter based on first and second moment estimates of the gradients, often accelerating convergence. Regularization techniques such as weight decay [100], skip connections for gradient flow [101], dropout [102], and early stopping [103] help prevent overfitting.

Despite these advances, training deep networks remains challenging. Loss landscapes are non-convex with many local minima, and hyperparameter choices can significantly impact performance. Nevertheless, a robust empirical observation across diverse domains is that scaling model size—increasing depth, width, and training data—consistently improves performance. This observation has yielded the current era of *scaling laws* [104], providing a reliable path to better models through architectural scale. As of 2026, these scaling relationships continue to hold across increasingly large systems.

1.6 Outline of Contributions

This thesis contains four chapters, each roughly modeled after publications I've written in my Ph.D. In total, the following chapters summarize four first-author publications [8, 105–107], and one second-author publication [108]. It also incorporates material from two first-author extended abstracts [92, 109] submitted to NeurIPS and ICLR. The paper contributions per chapter are outlined as follows.

Chapter 2: A Machine-Learned Model of Cosmological Stasis

Paper 1: ***On the Generality and Persistence of Cosmological Stasis***

> **Summary:** This paper explores the phenomenon of cosmological stasis, a phase of cosmic evolution in which the relative abundances of matter and radiation remain constant despite expansion. We study stasis in the full $2N$–dimensional space of decay rates and

abundances that feed the Boltzmann equations governing cosmic dynamics. We construct a differentiable Boltzmann solver and use gradient-based analyses to find configurations that maximize the number of stasis e–folds, identifying log–uniform priors that extend earlier power–law models. We demonstrate that random draws from these distributions regularly exhibit extended stasis, and incorporate these as priors in a Bayesian analysis using SVI with normalizing flows to model the posterior. A new exponential model of stasis is identified, shown to exactly solve the stasis equations, act as an attractor, and scale the number of e–folds linearly with species number—qualitatively different from previous power–law models. The paper also discusses implications for string–theoretic conjectures such as the emergent string hypothesis and the string axiverse.

Contributions: First author. I wrote all simulation and optimization code, conducted all experiments and analyses, and made all plots. J.H. wrote Sections V.B and V.C, as well as assisted with paper writing and editing.

> James Halverson and Sneh Pandya. "Generality and Persistence of Cosmological Stasis." *Physical Review D*, Vol. 110, No. 7 (2024), DOI: 10.1103/PhysRevD.110.075041.

Chapter 3: Neural Network Emulators and Differentiable Modeling for Galaxy Intrinsic Alignments

Paper 2: IAEMU*: Learning Galaxy Intrinsic Alignment Correlations*

Summary: The IAEMU paper introduces a neural network–based emulator for galaxy intrinsic alignment correlations. IAEMU is trained to jointly predict the position–position (ξ), position–orientation (ω), and orientation–orientation (η) correlation functions, along with associated aleatoric and epistemic uncertainties, from HOD parameters using mock catalogs. Unlike traditional analytic or simulation-based methods, IAEMU directly models these statistics, offering orders-of-magnitude speed-ups (approximately $10{,}000\times$ for GPU inference) compared to conventional approaches and enabling both rapid forward modeling and accelerated inverse inference via gradient-based sampling. The emulator generalizes beyond its training HOD mocks to non-HOD signals drawn from hydrodynamical simulations, demonstrating robustness and broad applicability for Stage IV cosmological surveys.

Contributions: First author. I developed the IAEMU architecture, wrote all training and evaluation code, and conducted all model analyses and validation experiments. I produced all figures and documented results. My coauthors generated training and testing data, provided input on the simulations and modeling, assisted with simulation comparisons, and contributed to paper writing and editing.

Sneh Pandya, Yuanyuan Yang, Nicholas Van Alfen, Jonathan Blazek, and Robin Walters. "IAEmu: Learning Galaxy Intrinsic Alignment Correlations." *The Open Journal of Astrophysics*, Vol. 8 (2025), DOI: 10.33232/001c.151749.

Paper 3: ***On Soft Clustering For Correlation Estimators***

Summary: This paper investigates soft clustering methods for estimating correlation functions in large-scale structure analyses, focusing on improving statistical efficiency and robustness in the presence of complex sample selections and survey geometries. We develop a framework that blends hard assignment approaches with soft clustering probabilities to better capture the underlying distribution of galaxies when estimating two-point and higher-order correlations. Through a combination of theoretical considerations and empirical evaluations on mock and observational datasets, we demonstrate that soft clustering estimators can reduce variance and systematic biases compared to traditional estimators, especially in regimes where sample sparsity and selection effects are significant. The results suggest that incorporating soft clustering into correlation estimation routines can enhance the fidelity of large-scale structure inferences for current and future cosmological surveys.

Contributions: Second author. I implemented and conducted some IAEMU -related experiments for this work, including preparing the related text. I also advised on other experiments, and assisted in paper writing and editing.

Edward Berman, Sneh Pandya, Jacqueline McCleary, Marko Shuntov, Caitlin Casey, Nicole Drakos, Andreas Faisst, Steven Gillman, Ghassem Gozaliasl, Natalie Hogg, Jeyhan Kartaltepe, Anton Koekemoer, Wilfried Mercier, and Diana Scognamiglio. "On Soft Clustering For Correlation Estimators." *The Open Journal of Astrophysics*, Vol. 8 (2025), DOI: 10.33232/001c.144313.

Paper 4: ***Differentiable Stochastic Halo Occupation Distribution with Galaxy Intrinsic Alignments***

Summary: This paper introduces DIFFHOD-IA, a differentiable implementation of HALOTOOLS-IA. We extend the differentiable HOD methodology of [110] to the IA implementation of [111] by differentiably sampling from the Dimroth-Watson distribution which governs galaxy misalignments. We further extend this framework to the summary statistic level, including contributions from [112] to differentiably calculate correlation functions. We validate the accuracy of DIFFHOD-IA and showcase expedited inference with HMC.

Contributions: First author. I developed the code, ran experiments, analyzed results, and wrote the paper. My coauthor reviewed the draft and provided useful guidance during the development of the paper.

> Sneh Pandya and Jonathan Blazek. "Stochastic Differentiable Halo Occupation Distribution with Galaxy Intrinsic Alignments" *Under review at The Open Journal of Astrophysics*, 2026, arXiv: 2602.04977.

Chapter 4: Symmetries and Domain Adaptation for Neural Network Generalization

Paper 5: *SIDDA: SInkhorn Dynamic Domain Adaptation for Image Classification with Equivariant Neural Networks*

Summary: This paper introduces SInkhorn Dynamic Domain Adaptation (SIDDA), a domain adaptation training algorithm designed to improve neural network generalization under covariate shifts by leveraging the Sinkhorn divergence to align latent feature distributions with minimal hyperparameter tuning. SIDDA dynamically adjusts the entropic regularization and the balance between classification and domain adaptation loss during training, reducing the need for manual tuning and computational overhead. The method is shown to improve classification accuracy (up to 40% on unlabeled target data) and calibration metrics (expected calibration error and Brier score) across multiple simulated and real datasets, including shapes, handwritten digits, and astronomical observations, and achieves stronger performance when paired with equivariant neural networks, owing to their built-in symmetry constraints. SIDDA's compatibility with a range of architectures and its automated approach to domain alignment make it a versatile tool for multi-dataset studies.

Contributions: First author. I developed the SIDDA algorithm, implemented and optimized the training and domain adaptation code, generated experimental datasets, and conducted all training, evaluation, and analysis. I also produced all figures and documented results. My coauthors contributed domain expertise, assisted with experimental design, and helped refine the manuscript.

> Sneh Pandya, Purvik Patel, Brian D. Nord, Mike Walmsley, and Aleksandra Ćiprijanović. "SIDDA: SInkhorn Dynamic Domain Adaptation for Image Classification with Equivariant Neural Networks." *Machine Learning: Science and Technology*, 2025, DOI: 10.1088/2632-2153/adf701.

Chapter 2

A Machine-Learned Model of Cosmological Stasis

One of the hallmark characteristics of an expanding universe is the time-evolution of components that contribute to the overall energy density of the Universe. It has been introduced [2] that this assumption is not always true, and in fact the Universe could experience extended phases of cosmological "stasis" in which the cosmological abundances of matter, radiation, and/or vacuum energy can remain steady over extended e-folds of cosmological evolution, facilitated by various physical mechanisms driving energy pumps which oppose the effects of cosmological expansion [113]. These phenomena can arise naturally from a variety of BSM physical theories, for instance in those that predict towers of unstable states that inevitably decay.

There are a number of different flavors of stasis. In the original formulation [2], stasis is achieved when a tower of ϕ_ℓ matter states dominate the energy density but then hierarchically decays into radiation (hereby referred to as $M \to \gamma$ stasis), whereas [113] further introduced mechanisms of vacuum energy to matter ($\Lambda \to M$) and vacuum energy to radiation ($\Lambda \to \gamma$) stasis in a similar context, as well as studies of the dynamics of "triple stasis" with simultaneous $\Lambda \to M \to \gamma$ stasis. Such alternative cosmological histories can have observational consequences [114, 115] depending crucially on the flavor, timescale, and duration of stasis. Though in many cases these provide interesting possibilities for the evolution of our Universe, stasis may also constrain ultraviolet-complete theories such as string theory if it arises after nucleosynthesis. We will comment on stasis in the context of Kaluza-Klein towers, string towers, and the string axiverse.

In this chapter, we present an analysis of stasis to complement [2], focusing on $M \to \gamma$

stasis. This flavor of stasis depends crucially on the abundances $\Omega_\ell^{(0)}$ and decay rates Γ_ℓ of the N particle species that matter dominate prior to stasis. Whereas [2] laid out a general theory of strict stasis and derived many analytic results in an eight-dimensional power law model, including attractor behavior, we focus on understanding stasis in a model-agnostic manner on the full $2N$-dimensional space of rates and abundances. Such an analysis seems to require numerics, for which we employ a variety of machine learning tools (to optimize stasis e-folds, not cosmological viability; see Section 2.6), but lets us probe the generality of stasis and understand aspects of its duration. The numerical methods uncover a new exponential model of stasis, which leads to extended periods of stasis.

This chapter is organized as follows. First, we review essentials of $M \to \gamma$ stasis from [2]. In Section 2.2 we develop a differentiable Boltzmann solver that facilitates a number of numerical analyses involving stasis. Performing gradient ascent to maximize stasis, we see the emergence of exponential models that motivate log-uniform statistics. In Section 2.3 we study random stasis where rates and abundances are drawn from both these log-uniform distributions and also power law distributions. Stasis occurs quite generally, with longer duration in the log-uniform case. In Section 2.4 we use both types of distributions as priors for stasis-conditioned posteriors. The posteriors in this Bayesian analysis are modeled using a neural network known as a normalizing flow, which are optimized using stochastic variational inference. Posterior samples lead to more robust stasis, referring to a longer duration of stasis epochs, and again prefer an exponential model. In Section 2.5 we study the exponential model directly, demonstrating that it leads to parametrically-in-N longer periods of stasis than power law models, and discuss potential interfaces with string theory.

2.1 Matter-Radiation Stasis

For a tower of states ϕ_ℓ where $\ell \in [0, 1, 2, ..., N]$, let ρ_ℓ denote a corresponding energy density and Ω_ℓ a corresponding abundance. For general types of stasis, ϕ_ℓ can either be a tower of massive scalar fields contributing to the total matter abundance Ω_M of the Universe, or a tower of vacuum energy fields contributing to Ω_Λ. Below, we will focus on the $M \to \gamma$ stasis formalism. Recall that for any energy density ρ_i, Ω_i is related via

$$\Omega_i \equiv \frac{8\pi G}{3H^2}\rho_i \ , \tag{2.1}$$

where H is the Hubble parameter and G is Newton's gravitational constant.

Differentiating, the time evolution of Ω_i is then

$$\frac{d\Omega_i}{dt} = \frac{8\pi G}{3}\left(\frac{1}{H^2}\frac{d\rho_i}{dt} - 2\frac{\rho_i}{H^3}\frac{dH}{dt}\right) . \tag{2.2}$$

This set of Boltzmann equations is dependent on the time-evolution of individual ρ_i as well as H. Using the Friedmann equation for dH/dt in a FLRW Universe, we obtain

$$\frac{dH}{dt} = -H^2 - \frac{4\pi G}{3}\left(\sum_\ell \rho_\ell + 3\sum_\ell p_\ell\right) , \tag{2.3}$$

where in practice this is expressed in terms of the equation-of-state parameter $w \equiv p_\ell/\rho_\ell$ for a component. Simplifying further, we arrive at

$$\frac{dH}{dt} = -\frac{1}{2}H^2(4 - \Omega_M) , \tag{2.4}$$

having invoked $w_\gamma = 1/3$ and $w_M = 0$ in equation 2.3. Integrating both sides, we arrive at

$$H(t) = \frac{2}{4 - \langle\Omega_M\rangle}\left(\frac{1}{t - t^{(0)}}\right) , \tag{2.5}$$

where we have used the approximation that $H^{(0)}(t - t^{(0)}) \gg 1$ and $\langle\Omega_M\rangle$ is the time-averaged matter abundance defined as

$$\langle\Omega_M\rangle = \frac{1}{t - t^{(0)}}\int_{t^{(0)}}^{t} dt'\Omega_M(t') . \tag{2.6}$$

During stasis, $d\langle\Omega_M\rangle/dt = 0$, in which $\langle\Omega_M\rangle = \overline{\Omega}_M$, the asymptotic stasis abundance.

In $M \to \gamma$ stasis, the decays of individual matter species are what source the production of radiation in this universe. Stasis epochs necessarily require time-dependent energy densities, whose equations of motion in this case are given by

$$\frac{d\rho_\ell}{dt} = -3H\rho_\ell - \Gamma_\ell\rho_\ell . \tag{2.7}$$

Returning to equation 2.2 and substituting in the result of equation 2.4 and 2.7, we arrive at the set of N Ordinary Differential Equation (ODE)s which govern the time evolution of individual Ω_ℓ

$$\frac{d\Omega_\ell}{dt} = H\Omega_\ell\,(1 - \Omega_M) - \Gamma_\ell\Omega_\ell . \tag{2.8}$$

Equation 2.8, in combination with the ODE for the Hubble parameter directly gives the dynamics for our system. As our Universe only contains matter and radiation, $\Omega_M + \Omega_\gamma = 1$ at all times and the dynamics for radiation are easily obtained by recognizing that $d\Omega_M/dt = -d\Omega_\gamma/dt$.

For $M \to \gamma$ stasis, the universe is beginning in a matter dominated state, $\Omega_M(t^{(0)}) = 1$. With time-evolution, the individual matter species gradually redshift or begin decaying into radiation; the effects of both must balance to have $d\Omega_M/dt = 0$ during stasis. This will happen for all times when the decays of ϕ_ℓ are exactly counterbalanced by Hubble expansion. Further, all individual Ω_ℓ's during stasis must cooperate to produce an asymptotic abundance $\overline{\Omega}_M$. These combined form the two necessary and sufficient conditions for an extended period of stasis:

$$\sum_\ell \Gamma_\ell \Omega_\ell = H(\Omega_M - \Omega_M^2) \tag{2.9}$$

$$\sum_\ell \Omega_\ell(t) = \overline{\Omega}_M \ . \tag{2.10}$$

Equation 2.9 as written is actually a condition for *eternal* stasis, which of course cannot be physical for some finite tower of states. However, we can illuminate how the stasis epoch ends by operating under the assumption of eternal stasis. Let us assume we are in a period of stasis where Ω_M has achieved its $\overline{\Omega}_M$ stasis value. With these conditions, we can study the explicit time dependence for many of the quantities of interest. Beginning with the Hubble parameter in equation 2.3, the solution is

$$H(t) = \left(\frac{2}{4 - \overline{\Omega}_M}\right)\frac{1}{t} \ , \tag{2.11}$$

which further implies that the scale factor grows as

$$a(t) = a_* \left(\frac{t}{t_*}\right)^{2/(4-\overline{\Omega}_M)} \tag{2.12}$$

for a fiducial time t_*. It further follows from equation 2.7 that

$$\rho_\ell(t) = \rho_\ell^* \left(\frac{t}{t_*}\right)^{-6/(4-\overline{\Omega}_M)} e^{-\Gamma_\ell(t-t_*)} \ , \tag{2.13}$$

which in turn implies that

$$\Omega_\ell(t) = \Omega_\ell^* \left(\frac{t}{t_*}\right)^{2-6/(4-\overline{\Omega}_M)} e^{-\Gamma_\ell(t-t_*)} \ . \tag{2.14}$$

The result of equation 2.11 when inserted into equation 2.9 while assuming a period of stasis gives

$$\sum_\ell \Gamma_\ell \Omega_\ell = \frac{2\overline{\Omega}_M(1 - \overline{\Omega}_M)}{4 - \overline{\Omega}_M}\frac{1}{t} \ , \tag{2.15}$$

exhibiting a power law dependence for t, which cannot be true for all t. Thus, this will not yield an eternal stasis epoch, but a stasis epoch which is terminated when all species decays have concluded.

The models studied in [2, 113] consider a spectrum of decay widths $\{\Gamma_\ell\}$ and abundances $\{\Omega_\ell\}$ motivated by a variety of BSM models which follow a power law scaling

$$\Gamma_\ell = \Gamma_0 \left(\frac{m_\ell}{m_0}\right)^\gamma, \quad \Omega_\ell^{(0)} = \Omega_0^{(0)} \left(\frac{m_\ell}{m_0}\right)^\alpha \tag{2.16}$$

where the mass spectrum takes the form

$$m_\ell = m_0 + (\Delta m)\ell^\delta \tag{2.17}$$

and $\Omega_0^{(0)}$ is a normalization factor enforcing $\Omega_M(t^{(0)}) = 1$. The parameters α, γ, and δ are further restricted to the following range:

$$-\frac{1}{\delta} < \alpha \leq \frac{\gamma}{2} - \frac{1}{\delta}\,. \tag{2.18}$$

This therefore defines a $8D$ model parameterized by

$$\{\alpha, \gamma, \delta, m_0, \Delta m, \Gamma_0, \Omega_0^{(0)}, t^{(0)}\} \tag{2.19}$$

which is crucially a subset of the full $2N$-dimensional input parameter space that we seek to study.

It is important to note that missing from this list of parameters is the initial value of the Hubble constant, $H^{(0)}$. This model of stasis has exhibited *global attractor* properties that were extensively studied in [2]. The initial timescale for ϕ_ℓ decays is dictated by the ratio $\Gamma_{N-1}/H^{(0)}$. When this ratio is small, the decays begin slowly after the starting time $t^{(0)}$, and the so-called "edge effects" in [2] are mild. Conversely, when $\Gamma_{N-1}/H^{(0)} \gg 1$, particle decays begin almost immediately and there are severe edge effects. These edge effects are indeed necessary for a stasis state to both begin and end, as the condition in equation 2.15 cannot be true when the decay process is just beginning or has concluded. However, due to the global attractor nature of stasis it is possible to achieve the same configuration of stasis, with the exception that the stasis state is approached from *below* rather than above.

2.2 Maximizing Stasis with Differentiable Simulations

The duration of stasis may be determined by solving the $N+1$ Boltzmann equations on a $(2N+1)$-dimensional parameter space of decay rates, abundances, and the initial value of the Hubble parameter. This problem is difficult due to its high dimensional nature, and in general we would also like to be able to differentiate the numerical solution to the Boltzmann equations to understand how the duration of stasis responds to variations in the rates and abundances. A numerical Boltzmann

solver may be thought of as a type of simulator, and we seek to differentiate through the entire simulation process.

Differentiable simulators are part of a growing trend in ML applications within the sciences, motivated by the emergence of more powerful and robust simulations. They have accelerated scientific analyses from molecular dynamics [116] and biology to physics and cosmology [117]. Further, with the advent of neural-network based probabilistic modeling, differentiable simulations are essential for implementing techniques such as stochastic variational inference [118] and Hamiltonian Monte Carlo [119]. Differentiable simulations thus serve as powerful scientific tools for which to do both purely data-driven and probabilistic modeling studies.

For N species, our task is to solve a set of $N + 1$ coupled ODEs as given in equations 2.4 and 2.8 and compute the number of e-folds of a potential stasis epoch. We utilize `diffrax` [120], a `jax`-based [98] library that provides numerical differential equation solvers that preserve gradients and also allow backpropagation through our solutions to the Boltzmann equations. `jax` is a differentiable high-performance numerical computing library which utilizes just-in-time (`JIT`) compilation and vectorized computations, earning popularity in the sciences for its efficiency and modularity. The set of ODEs is stiff, meaning certain numerical techniques will be unstable without a sufficiently small resolution for the step-size. We use the `Kvaerno5` [121] solver which is a 5th order explicit singly diagonal implicit Runge-Kutta method suited for stiff ODEs, with an absolute and relative tolerance atol $=$ rtol $= 10^{-8}$ in the step size. We terminate the solver at $t = t_{max}$ when $\Omega_M(t) = 10^{-4}$, indicating that the universe has passed into a radiation-dominated epoch. Backpropagation through a differential equation can also be very memory expensive. We use `RecursiveCheckpointAdjoint` [122, 123] which utilizes a binomial checkpointing scheme to preserve memory usage during backpropagation.

We seek to utilize the gradients in our solver to guide the parameter space towards stasis configurations. It is therefore not enough to solve the Boltzmann equations and be able to backpropagate through them, we also define a *differentiable* algorithm to compute the stasis duration and asymptotic abundance from a given Ω_M curve. To this extent, we must first define a numerical notion of numerical stasis that can be applied to numerical Ω_M curves.

2.2.1 ϵ-Stasis

Stasis as defined up to this point is a strict condition on the time-evolution of a cosmological component. In [2], the model introduced is constructed so that this can be achieved exactly. However,

from a numerical perspective, configurations that yield small deviations from exact stasis can nevertheless yield significant alterations to a cosmology. A more general definition of stasis is an epoch in which Ω_M is "flat" enough for "long" enough. This requires a parametric notion of stasis, and we therefore introduce an ϵ-tolerance on stasis according to two different definitions. First, we develop a notion of stasis that allows us to create a differentiable (to enable the differentiable simulator) and accurate *stasis finder* algorithm to isolate epochs of stasis. Then we present a more intuitive notion of stasis that is utilized in all the presentations of our results.

We begin with the notion of stasis that admits a differentiable stasis finder algorithm. Recall that stasis is a phenomenon induced by the cooperative behavior of individual Ω_ℓ. A period of stasis is therefore computed by analyzing the total abundance $\Omega_M(t) = \sum_\ell \Omega_\ell(t)$. For a given $\Omega_M(t)$ curve, we must compute the asymptote around which stasis occurs, and further isolate the duration of a stasis epoch. To do this in a differentiable way, we introduce an exponential weighting to compute a "flatness score" that yields the stasis duration in t

$$t_s(\Omega_M, \overline{\Omega}_M) = \left[\sum_i \exp\left(-\frac{1}{\sigma}|\Omega_{M,i+1} - \Omega_{M,i}| \ -\frac{1}{\delta}|\Omega_{M,i} - \overline{\Omega}_M|\right)\right] \times \Theta((0.99 - \overline{\Omega}_M)(\overline{\Omega}_M - 0.01)) \ , \tag{2.20}$$

where Θ is the Heaviside function, enforcing that only mixed-component cosmologies are considered, and the index i is that of the i^{th} time-step in the solution $\Omega_M(t)$. A tolerance $\sigma = 0.02$ and $\delta = 0.09$, which roughly corresponds to a window-tolerance of ± 0.1 about the asymptotic stasis abundance, was found to work best for optimization.

Upon termination of the solver, we use the solution for $H(t)$ to calculate the total number of e-folds of the simulation via

$$\mathcal{N}_{max} = \int_{t^{(0)}}^{t_{\text{max}}} H(t)\, dt \ , \tag{2.21}$$

where it is assumed that $a(t^{(0)}) = 1$. This integral is evaluated numerically using the composite trapezoidal rule (i.e. `jax.numpy.trapezoid`). It is then straightforward to compute the number of e-folds of stasis $\mathcal{N}$ via normalization with respect to the total number of e-folds,

$$\mathcal{N} = \mathcal{N}_{max} \cdot \frac{t_s}{t_{max}} \ . \tag{2.22}$$

This normalization step allows us to isolate the contribution of stasis epoch relative to the entire simulation duration.

Intuitively, the stasis finder measures flatness by considering both how similar consecutive in $\Omega_{M,i}$ values are and how close these values are to $\overline{\Omega}_M$. It does so by creating a score that rewards

sequences where values are close to each other (indicating flatness) and close to the target value (indicating relevance), adjusted by the parameters ϵ and δ to fine-tune sensitivity and scaling.

This differentiable calculation, however, requires that $\overline{\Omega}_M$ is known. This is retrieved from a separate, differentiable abundance-finder algorithm. The algorithm first marks Ω_M values that are not matter dominated or radiation dominated (i.e. $0.01 < \Omega_M < 0.99$). It then determines links between consecutive valid values that are within a specified tolerance ϵ. It proceeds to count these links for each time step and identifies the range of indices with the maximum number of links, indicating a period of stasis. Finally, it filters the valid $\Omega_{\rm M}$ values within this range and computes $\overline{\Omega}_M$ as the median of these values, representing the typical abundance during the stasis period. This algorithm, of course, also yields a prediction for $\mathcal{N}$, but we find by examining solutions that it is inaccurate and does not function as well as equation 2.20.

A pitfall of the differentiable stasis-finder in equation 2.20 is that the sum of flatness scores considers *all* parts of $\Omega_M(t)$ that are near the asymptotic abundance, as opposed to those *only* within the stasis period. As such, the differentiable finder can be biased in e-folds; however, we will see that it still properly serves the purpose of a guiding the parameter space towards stasis configurations. For these reasons, a more accurate, non-differentiable stasis finder which uses a sliding-window algorithm to isolate the longest stasis period is used in all presented results. This sliding window stasis finder is configured for a 10%-tolerance, e.g. the number of e-folds of stasis is determined by considering a window of ± 0.1 around $\overline{\Omega}_M$ as valid. All the results presented in this chapter use this notion of stasis.

2.2.2 Maximizing Stasis

We have amassed the necessary ingredients to use the differentiable simulation to optimize for stasis. Using the simulation $\mathcal{S}(\theta = \{\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)}\})$, which returns $\mathcal{N}$ and $\overline{\Omega}_M$ after solving the Boltzmann equations and its gradients ∇_{Γ_ℓ} and $\nabla_{\Omega_\ell^{(0)}}$, we can employ gradient ascent to optimize on stasis e-folds. We begin with a single vector of samples of $\Omega_\ell^{(0)}$ and Γ_ℓ for a given N, initialized according to draws from any appropriate distribution. The Boltzmann equation for Ω_ℓ is inherently dimensionless; the only dimensionful parameters are $H^{(0)}$ and Γ_ℓ. We will take all Γ_ℓ to be in units of Planck mass M_p and no greater than this scale, e.g., $\max(\Gamma_\ell) \leq 1\ M_p$. Similarly, we will restrict decay rates to be no smaller than those corresponding to the current age of the Universe, $\min(\Gamma_\ell) \geq 10^{-62}\ M_p$.

We consider both power law and standard uniform distributions for initializing Γ_ℓ and

$\Omega_\ell^{(0)}$. The power law distribution is representative of the original model of stasis in the limit that $\Delta m/m_0 \gg 1$, while the standard uniform respects minimal physical constraints (i.e. positive-definiteness and $\max(\Gamma_\ell) \leq 1\ M_p$) on the parameters. Recall that $\Omega_\ell^{(0)}$ are normalized upon entering the simulation such that $\Omega_M(t^{(0)}) = 1$. To generate samples from a power law distribution with a relative scaling ℓ^β, we draw from a Pareto distribution with a shape parameter $\alpha_p = 1/\beta$. These samples are then inverted to represent draws from a power law distribution. Henceforth, we denote power law distribution samples as $X \sim \ell^\beta$, with the understanding that they are generated as the inverse of samples $X \sim \text{Pareto}(\alpha_p = 1/\beta)$.

After sampling but before entering the simulation, we sort the Γ_ℓ, which is without loss of generality since the ℓ subscript amounts to a species definition. However, we also often sort the $\Omega_\ell^{(0)}$, so that increasingly large rates correspond to increasingly large abundances. This introduces a non-trivial physics-motivated correlation between the parameters, and we henceforth refer to such samples as "sort-correlated." When we speak of identically and independently draw (i.i.d.) parameters, we mean prior to sort-correlation, unless otherwise stated.

Sort-correlation may be performed with any off-the-shelf sorting algorithm, including but not limited to a simple `jax.numpy.sort`. As Γ_ℓ and $\Omega_\ell^{(0)}$ are updated with gradient ascent, we would like to continue to enforce that their spectra are sort-correlated. Doing so crucially requires that they are sorted smoothly and differentiably, to not interrupt the flow of gradients during optimization. For this reason, we implement a custom differentiable bitonic sorting algorithm inspired by [124]. This algorithm works by recursively dividing an array into smaller sub-arrays, sorting them, and then merging them using compare-and-swap operations, the latter of which is modified to be differentiable. This sorting technique is also used when doing SVI experiments.

One necessary constraint to recognize is that simply optimizing on stasis e-folds will encourage matter-dominated cosmologies, as species' time spent decaying detracts from time spent redshifting, which is what contributes to the overall stasis duration. As such, a stasis optimization condition f is designed to enforce a mixed component cosmology

$$\begin{aligned} f(\theta) = \mathcal{N}(\theta) - \alpha \left[(\overline{\Omega}_M(\theta) - l)^2 + (\overline{\Omega}_M(\theta) - u)^2\right] \\ \times \left(1 - \Theta(\overline{\Omega}_M(\theta) - l)\Theta(u - \overline{\Omega}_M(\theta))\right) \ , \end{aligned} \tag{2.23}$$

where we recognize that $\mathcal{N}$ and $\overline{\Omega}_M$ are outputs of $\mathcal{S}(\theta)$. Above, $l = 0.2$ and $u = 0.8$, defining bounds on the allowed matter abundance, and α specifies the regularization strength depending on the experiment, ensuring the constraint has a meaningful effect on the optimization landscape. The Heaviside function indicates that the penalty is only applied when $\overline{\Omega}_M$ is outside the allowed

abundance window. It is now clear to see why the ϵ-stasis finder algorithm needs to be differentiable in both e-folds and matter abundances, as the gradients of $\overline{\Omega}_M$ are necessary to study optimization of stasis.

When performing gradient ascent on stasis, the sequential gradient updates for Γ_ℓ and $\Omega_\ell^{(0)}$ are computed as

$$\Omega_{\ell,i+1}^{(0)} = \Omega_{\ell,i}^{(0)} + \eta(t)\nabla_{\Omega_{\ell,i}^{(0)}} f(\theta) \tag{2.24}$$

$$\Gamma_{\ell,i+1} = \Gamma_{\ell,i} + \eta(t)\nabla_{\Gamma_{\ell,i}} f(\theta) \tag{2.25}$$

for a given step i, where $\eta(t)$ is a (potentially) time-dependent learning rate. In subsequent experiments, a decay-factor γ is applied at epoch t' such that the learning rate has the functional form

$$\eta(t) = \begin{cases} \eta_0, & \text{for } t < t' \\ \gamma \cdot \eta_0, & \text{for } t \geq t' \ . \end{cases} \tag{2.26}$$

Optimization is subject to an early-stopping criterion if $f(\theta)$ does not improve over a specified number of epochs ξ or if a `NaN` is encountered during optimization.

A detailed algorithm of optimizing stasis with gradient ascent is shown in Algorithm 1. In short, the algorithm 1) generates initial samples for rates and abundances; 2) sort-correlates them; 3) solves the Boltzmann equations; 4) uses the differentiable stasis-finder to compute the number of stasis e-folds; 5) compute gradients through the solution; 6) updates the parameters according to gradient ascent on stasis. This yields the parameters for the next iteration of the pipeline.

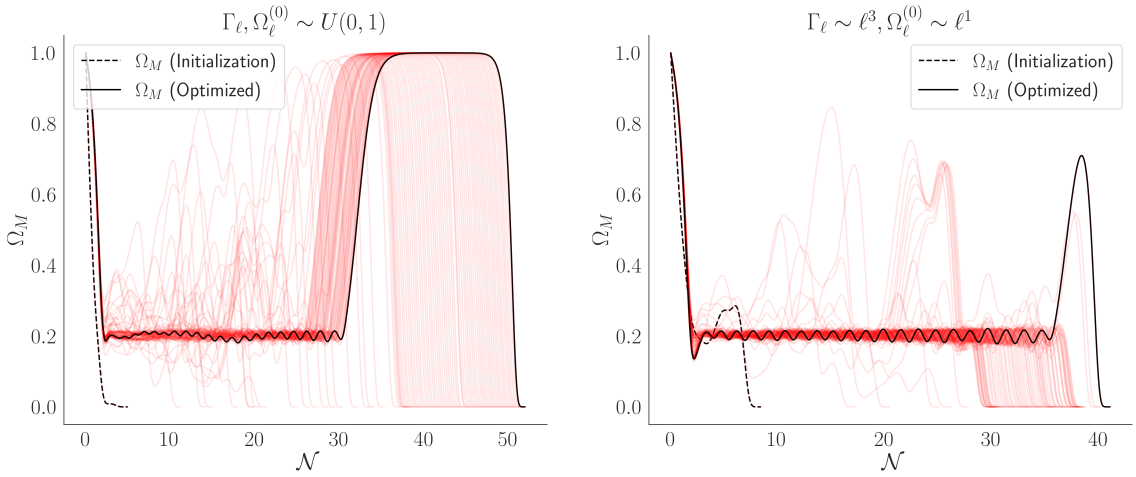
We study the outcome of gradient ascent optimization for initializations from a power law distribution corresponding to $\Gamma_\ell \sim \ell^3$ and $\Omega_\ell^{(0)} \sim \ell^1$ and a uniform initialization where $\Gamma_\ell, \Omega_\ell^{(0)} \sim \text{Uniform}(0,1)$. We conduct the experiments for $N = 50$ species and with $\Gamma_{N-1}/H^{(0)} = 0.1$. We optimize both for a total of 50000 epochs with an early stopping threshold of $\xi = 2000$ epochs. An initial learning rate of $\eta = 0.01$ is used.

Example gradient ascent trajectories for the uniform initialization are shown in Figure 2.1(a), where the optimization has settled on $\overline{\Omega}_M = 0.2$ and has achieved 27 e-folds of stasis. The change from initialized to optimized parameters is seen in going from the black-dashed line at initialization, through increasingly dark red intermediate trajectories as gradient ascent progresses, converging to the solid black line. From the trajectories, it can be seen that the optimization was relatively noisy. Indeed, from a numerical perspective what is essential for a robust epoch of stasis is large hierarchies in Γ_ℓ and smaller hierarchies in $\Omega_\ell^{(0)}$, as is manifest for the power law model

Algorithm 1 Gradient Ascent on Stasis

Require: $\theta = \{\Gamma_\ell, \Omega_\ell^{(0)}\}$ simulation parameters, $\mathcal{N}$ stasis e-folds, α penalty coefficient, l lower bound on matter abundance, u upper bound on matter abundance, $\eta(t)$ learning rate, ξ early-stopping threshold

Initialize: Γ_ℓ, $\Omega_\ell^{(0)}$ via sampling and sort-correlate their spectra
while not converged **do**
  $\mathcal{N}, \overline{\Omega}_M \leftarrow \mathcal{S}(\theta)$
  $f(\theta) \leftarrow \mathcal{N}(\theta) - \alpha \left[(\overline{\Omega}_M(\theta) - l)^2 + (\overline{\Omega}_M(\theta) - u)^2\right] \times \big(1 - \Theta(\overline{\Omega}_M(\theta) - l)\Theta(u - \overline{\Omega}_M(\theta))\big)$
  $\nabla_{\Omega_\ell^{(0)}} f(\theta), \nabla_{\Gamma_\ell} f(\theta)$ ▷ Compute gradients
  $\Omega_{\ell,i+1}^{(0)} \leftarrow \Omega_{\ell,i}^{(0)} + \eta(t) \nabla_{\Omega_{\ell,i}^{(0)}} f(\theta)$ ▷ Update abundances
  $\Gamma_{\ell,i+1} \leftarrow \Gamma_{\ell,i} + \eta(t) \nabla_{\Gamma_{\ell,i}} f(\theta)$ ▷ Update decay rates
  $\Omega_1^{(0)} \leq \Omega_2^{(0)} \leq \ldots \leq \Omega_N^{(0)}$ ▷ Differentiably sort $\Omega_\ell^{(0)}$
  $\Gamma_1 \leq \Gamma_2 \leq \ldots \leq \Gamma_N$ ▷ Differentiably sort Γ_ℓ
  $\Gamma_\ell \leftarrow \text{clip}(\Gamma_\ell, 0, 1)$ ▷ Ensure Γ_ℓ physical
  if $f(\theta)$ does not improve over ξ epochs or `NaN` is encountered **then**
    break ▷ Early-stopping criterion
  end if
end while



(a) Uniform initialization

(b) Power law initialization

Figure 2.1: Gradient ascent trajectories for $N = 50$ species and $\Gamma_{N-1}/H^{(0)} = 0.1$, optimized subject to the condition in equation 2.23 for 50,000 epochs with $\alpha = 10$. (a) Uniform initialization produces relatively noisy intermediate trajectories (red lines). (b) Power law initialization benefits from decompressed Γ_ℓ and $\Omega_\ell^{(0)}$ spectra, resulting in less noisy trajectories and a more robust epoch of stasis.

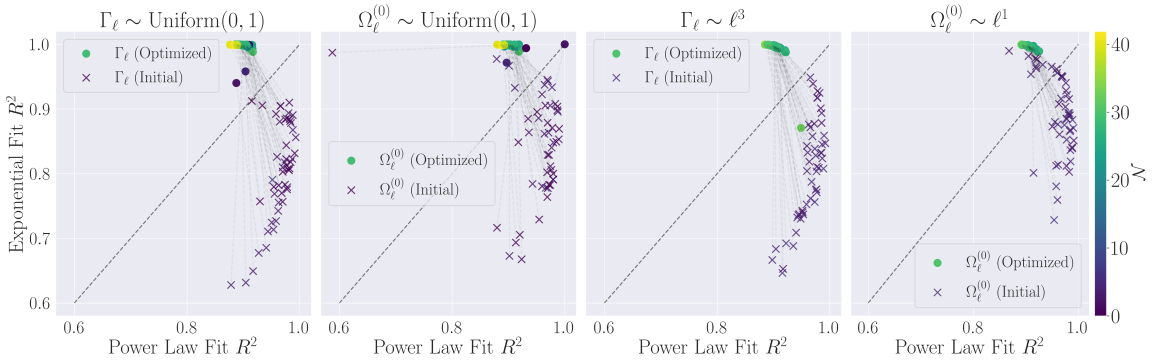


Figure 2.2: Experiments showing a preference for the exponential model upon optimizing stasis with gradients for uniform and power law initializations for $N = 50$ and $\Gamma_{N-1}/H^{(0)} = 0.1$. Gradient ascent was subject to the constraint $0.2 < \overline{\Omega}_M < 0.8$. Optimization was done for 50 random initializations for 50,000 epochs with early-stopping. An initial learning rate $\eta_0 = 0.01$ was used with a $\gamma = 0.1$ multiplicative decay at epoch $t' = 10,000$. A clear bias towards an exponential model is shown for optimal Γ_ℓ and $\Omega_\ell^{(0)}$, even when initialized with a power law distribution similar to the original model of stasis. It is also seen that the drift shifts across the 1:1 line dividing exponential and power law confidence equality. In some instances, the initialized parameters are shown to already be a good exponential fit due to the compression of the relative abundance and decay spectra. Even under such conditions, the drift towards a more exponential model and away from a power law model is evident.

of stasis and not characteristic of draws from a uniform distribution. Conceptually, this means that large changes are required to achieve stasis.

When initializing parameters as a power law, the benefit of immediate hierarchies in the species spectra is apparent in Figure 2.1(b). The optimization is more stable, as shown by the well-behaved intermediate values in red, and results in a longer epoch of stasis lasting 34 e-folds. It is then interesting to wonder: does a power law initialization still result in a power law model after following gradients, or does it change qualitatively? Similarly, does the uniform initialization change qualitatively under gradient ascent?

We can further deploy the differentiable simulation to study the model dependence of the optimized Γ_ℓ and $\Omega_\ell^{(0)}$. We optimize 50 random initializations of $\theta = \{\Gamma_\ell, \Omega_\ell^{(0)}\}$ with the differentiable simulation initialized according to a uniform distribution, $\Gamma_\ell, \Omega_\ell^{(0)} \sim U(0,1)$, and power law distributions, $\Gamma_\ell \sim \ell^3$ and $\Omega_\ell^{(0)} \sim \ell^1$. We conduct our experiments for $N = 50$ and $\Gamma_{N-1}/H^{(0)} = 0.1$, the results of which are shown in Figure 2.2.

We can quantify the model dependence of Γ_ℓ and $\Omega_\ell^{(0)}$ to a power law model with linear fit in log-log space when looking at the functional dependence of Γ_ℓ and $\Omega_\ell^{(0)}$ with ℓ. Similarly, a linear fit in semi-log space is indicative of an exponential dependence with ℓ. We use the coefficient of determination (R^2 score) for comparisons of model dependence across distributions and across scales, as it enjoys a scale invariance while still encoding information of fit residuals. The scale invariance is crucial, as the relative scale of optimized Γ_ℓ and $\Omega_\ell^{(0)}$ are a priori different than from their respective initializations. Traditional metrics such as mean squared error are sensitive to this, and can yield misleading results.

In Figure 2.2 we see that at initialization there is a clear preference towards a power law model over exponential, as both $\Omega_\ell^{(0)}$ and Γ_ℓ generally exhibit $R^2 > 0.9$, while exponential fits feature $R^2 > 0.6$. This is expected, since neither of the initialization distributions were exponential. After optimization, there is a clear drift towards perfect exponential fit ($R^2 = 1$) across all experiments, indicating a clear preference towards an exponential model of stasis when optimizing on stasis e-folds, even when being initialized with a power law distribution that is reflective of the existing model of stasis. It is also seen in Figure 2.2 that the optimized values are much more abundant in stasis e-folds than their respective initializations.

This result suggests a new model of stasis for which

$$\Omega_\ell^{(0)} \propto e^{\alpha \ell} \qquad \Gamma_\ell \propto e^{\gamma \ell}, \tag{2.27}$$

an exponential model which is qualitatively different than the power law model introduced in [2]. It also opens the door for additional physical mechanisms that can result in particle spectra that can induce a stasis state. We will comment on those physics models in Section 2.5. We emphasize that this result has very little model bias: no trained neural networks or strong prior beliefs that restrict the effective parameter dimension were used in arriving at this result. The only assumption is a prior on the full parameter space from which the initial parameter are drawn, and then we simply differentiably simulate the stasis phenomenon conditioned on increasing stasis. On statistical grounds, these results suggest a new *distribution* that these parameters follow; that is, a log-uniform distribution, a simple distribution for which samples are uniformly distributed in orders of magnitude. We will further study this model of stasis, and more generally the stasis parameter space from a distributional point of view, using neural networks in a Bayesian inference setting. To that end, these results suggest a new prior distribution.

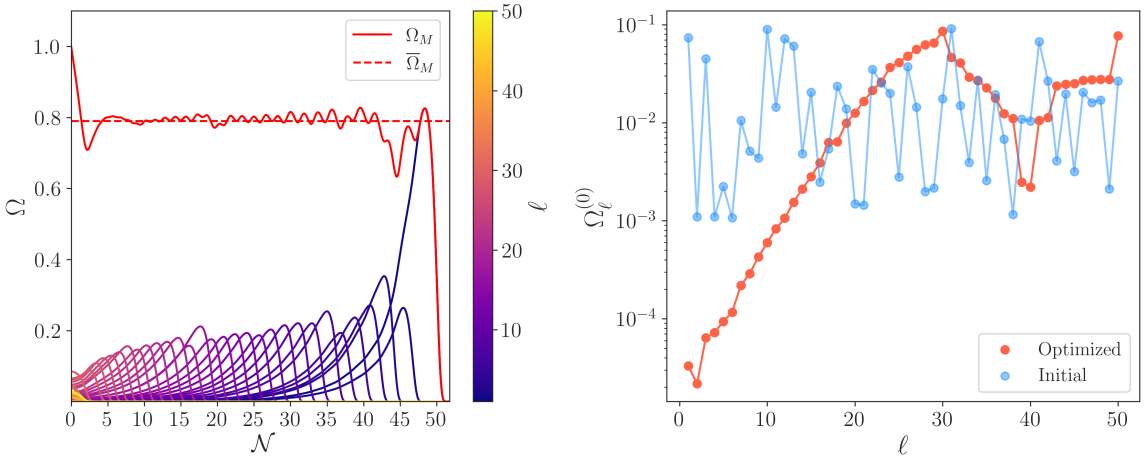


Figure 2.3: Example stasis epoch and gradient ascent trajectories for unsorted abundances with $N = 50$ and $\Gamma_{N-1}/H^{(0)} = 0.1$. The initializations chosen were $\Gamma_\ell \sim \text{Log-U}(10^{-62}, 10^0)$ and $\Omega_\ell^{(0)} \sim \text{Log-U}(10^{-2}, 10^0)$. Γ_ℓ samples were sorted before optimization via gradient ascent, which is without loss of generality as ℓ is a species definition. $\Omega_\ell^{(0)}$ are left unsorted. We see that the gradients learn to correlate lower ℓ species which contribute to the overall stasis duration, resulting in a period of stasis lasting ~ 43 e-folds at an abundance of $\overline{\Omega}_M = 0.80$. We see in the right panel that these contributing species approximately follow an exponential, characterized by a linear dependence on ℓ on a semi-log plot. For high ℓ species, which decay at early times to set the stasis abundance $\overline{\Omega}_M$, the algorithm does not learn to sort them. Indeed, $\sim 40\%$ of abundances do not monotonically increase with ℓ.

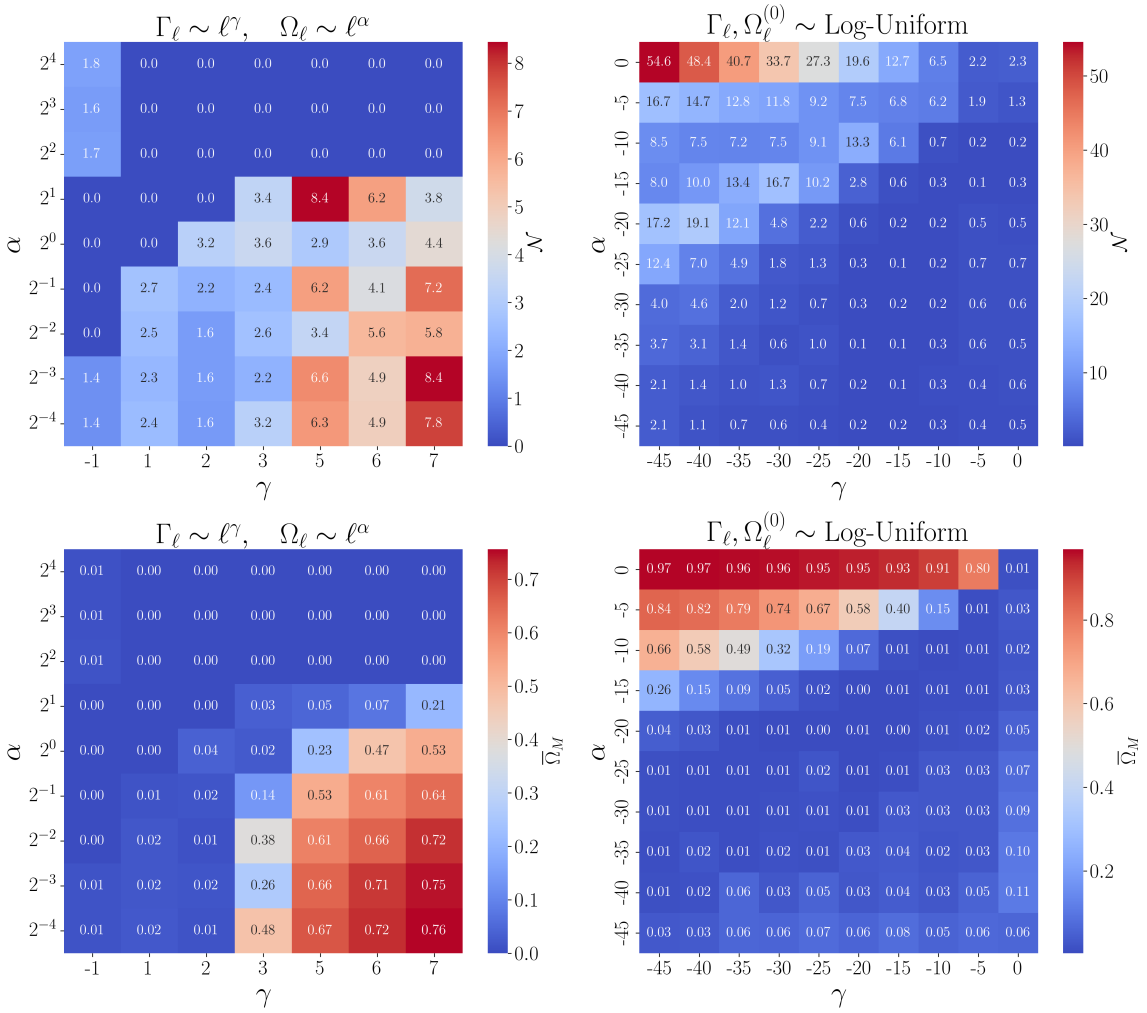


Figure 2.4: (**Left**) Stasis configurations for $N = 100$ species initialized with power law prior draws across 100 realizations for $\Omega_\ell^{(0)}$ and Γ_ℓ, which correspond to the model of stasis introduced in [2]. $\Omega_\ell^{(0)}$ and Γ_ℓ are additionally sort-correlated before entering the simulation. The (non-differentiable) sliding-window stasis finder was used with a 10%-tolerance. We see that the maximum number of e-folds for this prior distribution is $\mathcal{N} \sim 8.4$ e-folds for a more matter-dominated cosmology. (**Right**) Stasis configurations for $N = 100$ species initialized with exponential prior draws across 10 realizations, which correspond to sorted samples from a log-uniform distribution. The axes values correspond to $\Omega_\ell^{(0)} \sim$ Log-U$(10^\alpha, 10^0)$ and $\Gamma_\ell \sim$ Log-U$(10^\gamma, 10^0)$, with the values chosen to illustrate the transition in $\mathcal{N}$ and $\overline{\Omega}_M$. The maximum number of e-folds for this prior is $\mathcal{N} \sim 55.1$ e-folds, a noticeably longer stasis duration than the power law distribution and completely matter-dominated. Both distributions also feature a disallowed region in $\Omega_\ell^{(0)}$, in which the abundance spectrum becomes sufficiently stressed that stasis is not possible.

2.2.3 Stasis and Unsorted Abundances

We have so far exclusively operated under the setting of sort-correlated Γ_ℓ and $\Omega_\ell^{(0)}$. It is nonetheless interesting to study the emergence of a stasis epoch without respecting strict physical correlations between the parameters. As discussed, we may sort the Γ_ℓ without loss of generality, but in this section we keep $\Omega_\ell^{(0)}$ unsorted. We proceed by initializing the decay rates as $\Gamma_\ell \sim \text{Log-Uniform}(10^{-62}, 10^0)$ and sorting the spectrum, and $\Omega_\ell^{(0)} \sim \text{Log-Uniform}(10^{-2}, 10^0)$ for $N = 50$ species, evolving the parameters according to the gradient update rules given in equations 2.24 for 50,000 epochs with early-stopping.

The result of this experiment is shown in Figure 2.3. This analysis has indeed shown that *robust stasis epochs are possible without strictly correlated decay rates and abundances*, even though the species contributing to stasis have abundances increasing with decay rates. This realization would have been difficult to make with analytical modeling!

We see in Figure 2.3 that the randomly distributed abundances upon initialization become roughly monotonic up to $\ell \sim 30$, which are exactly the species in the left panel that balance to produce the stasis epoch lasting 43 e-folds. With the spectrum viewed on a semi-log plot in the right panel of Figure 2.3, it is clear to see that the contributing species follow an exponential model. The species $\ell > 30$ are those that decay at early times, setting the $\overline{\Omega}_M$ value. By just following gradients, it was deduced that these species did not in fact need to respect strict correlations with Γ_ℓ to produce stasis.

2.3 Random Stasis in Physics-Motivated Distributions

The results of simply following gradients in the differentiable simulation motivates an exponential model, as opposed to the original power law model of [2]. In a statistical language, this motivates the analysis of log-uniform priors in addition to the power law priors. Power law models were argued to be physically relevant in [2], and we will argue in Section 2.5 that exponential models are also physically motivated. In this section, we study stasis in random draws from these physically-motivated priors, and argue that they are also statistically well-motivated in Bayesian statistics. First we will provide a statistical perspective on the priors, and then study stasis.

2.3.1 Scale-Invariant Priors and Decay Rates

A prior on a vector of random variables $\boldsymbol{\theta}$ is a density that represents some knowledge of the system or beliefs about the way it behaves, before further evidence is introduced. There exists a particular class of priors known as "uninformative priors," and a subclass of such priors that are scale-agnostic. One of these scale-invariant priors is known as the Jeffreys prior. Mathematically, it is defined to exhibit the scaling

$$p(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})} \tag{2.28}$$

for its density function, where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information metric defined as

$$\mathcal{I}(\boldsymbol{\theta})_{ij} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(\boldsymbol{X}|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(\boldsymbol{X}|\boldsymbol{\theta})\right)\middle|\boldsymbol{\theta}\right] \tag{2.29}$$

for a given likelihood function of parameters $f(\boldsymbol{X}|\boldsymbol{\theta})$, which is the probability density of a set of observed data $\boldsymbol{X}$ given parameters $\boldsymbol{\theta}$. We see that the Jeffreys prior is the volume measure with respect to the Fisher information metric. It is therefore diffeomorphism invariant and has the significant advantage that it does not depend on the choice of coordinates for model parameters.

We will now proceed to derive an appropriate Jeffreys prior for Γ_ℓ. The time until decay τ for a particle given a decay rate Γ is governed by the probability density function (PDF)

$$f(\tau|\Gamma) = \Gamma\, e^{-\Gamma\tau}\ . \tag{2.30}$$

Under conditions in which a likelihood function is twice differentiable and has a vanishing expectation value of the gradient of the log-likelihood, a simpler definition of the Fisher information can be used. By inspection, we see that $f(\tau|\Gamma)$ is twice differentiable with respect to Γ. It only remains to check that

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\Gamma}\log f(\tau|\Gamma)\right] &= \mathbb{E}\left[\frac{1}{\Gamma} - \tau\right] \\ &= \frac{1}{\Gamma} - \frac{1}{\Gamma} \\ &= 0,\end{aligned} \tag{2.31}$$

where we have invoked that $\mathbb{E}[\tau] = 1/\Gamma$, as τ follows an exponential distribution with respect to Γ. With these conditions satisfied, we can use a simpler form of the Fisher information

$$\mathcal{I}(\boldsymbol{\theta})_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X};\boldsymbol{\theta})\middle|\boldsymbol{\theta}\right] \tag{2.32}$$

in deriving the Jeffreys prior for Γ_ℓ.

The likelihood of observing a set of particles for a period of time $\boldsymbol{\tau} = \{\tau_1, \tau_2, ..., \tau_N\}$ before they decay with i.i.d decay rates $\boldsymbol{\Gamma} = \{\Gamma_1, \Gamma_2, ..., \Gamma_N\}$ is given by

$$p(\boldsymbol{\tau}|\boldsymbol{\Gamma}) = \prod_{i=1}^{N} f(\tau_i|\Gamma_i) = \prod_{i=1}^{N} \Gamma_i \, e^{-\Gamma_i \tau_i} \, . \tag{2.33}$$

Exploiting properties of logarithms, we arrive at

$$\log p(\boldsymbol{\tau}|\boldsymbol{\Gamma}) = \sum_{i=1}^{N} \left(\log \Gamma_i - \Gamma_i \tau_i\right), \tag{2.34}$$

which allows one to compute the Fisher information matrix $I(\boldsymbol{\Gamma})_{ij}$ as

$$I(\boldsymbol{\Gamma})_{ij} = -\delta_{ij} \mathbb{E}_{\boldsymbol{\tau}} \left[\frac{\partial^2 \log p(\boldsymbol{\tau}|\boldsymbol{\Gamma})}{\partial \Gamma_i \partial \Gamma_j} \right] \Rightarrow \tag{2.35}$$

$$I(\boldsymbol{\Gamma})_{ii} = -\mathbb{E}_{\boldsymbol{\tau}} \left[-\frac{1}{\Gamma_i^2} \right] = \frac{1}{\Gamma_i^2}, \tag{2.36}$$

where δ_{ij} is the Kronecker delta, enforcing that the Γ_i's are i.i.d. distributed. Therefore, we see that a Jeffreys prior for Γ_ℓ must obey

$$p(\boldsymbol{\Gamma}) \propto \sqrt{\det I(\boldsymbol{\Gamma})} = \prod_{i=1}^{N} \frac{1}{\Gamma_i}. \tag{2.37}$$

Such a property is exactly true for a log-uniform prior, as defined by its PDF

$$f(x) = \frac{1}{x \log\left(\frac{b}{a}\right)}, \quad \text{for } a \leq x \leq b; \; a > 0 \tag{2.38}$$

where the interval $[a, b]$ is known as the *support* and the term $\log(b/a)$ appears after normalizing (2.37).

It is remarkable that the distribution on Γ_ℓ motivated by flow on Γ_ℓ using the differentiable simulation aligns exactly with the Jeffreys prior for decay rates! We have thus established that such a distribution for decay rates is worthy of study, on both numerical and statistical grounds, with physics motivation in Section 2.5.

2.3.2 Random Stasis from Physical Priors

It is now natural to study in detail how much stasis can emerge for Γ_ℓ and $\Omega_\ell^{(0)}$ spectra simply drawn from power law and log-uniform priors. We study the mean e-folds of stasis that emerge

across several configurations of the log-uniform and power law prior, which are related to a new exponential model of stasis and the original model, respectively.

We initialize the power law priors according to draws $\Gamma_\ell \sim \ell^\gamma$ and $\Omega_\ell^{(0)} \sim \ell^\alpha$, where $\gamma \in [-1, 7]$ and $\alpha = 2^k$ for $k \in [-4, 4]$. The procedure for generating power law draws is given in 2.2.2. Similarly, the exponential model prior is initialized according to $\Gamma_\ell \sim$ Log-Uniform$(10^\gamma, 10^0)$ and $\Omega_\ell^{(0)} \sim$ Log-Uniform$(10^\alpha, 10^0)$ where $\gamma, \alpha = 5k$ for $k \in [-9, 0]$. These samples are sort-correlated before entering the simulation. Experiments are run across 100 realizations for $N = 100$ and with $\Gamma_{N-1}/H^{(0)} = 0.1$ while simultaneously enforcing $\max(\Gamma_\ell) = 1\ M_p$.

We see in the top panel of Figure 2.4 that the power law prior achieves a maximum mean of $\mathcal{N} = 7.8$ e-folds of stasis for $\alpha = 2^{-3}$ and $\gamma = 7$. Indeed, we see that stretched spectra in Γ_ℓ and compressed spectra in $\Omega_\ell^{(0)}$ result in more matter-dominated cosmologies with longer epochs of stasis. Similarly, the exponential model prior achieves a maximum mean of $\mathcal{N} = 55$ e-folds for $\gamma = -45$ and $\alpha = 0$. It is important to note that after normalization, as $\Omega_\ell^{(0)}(t^{(0)}) = 1$, this configuration corresponds to a constant mass spectrum where $\Omega_\ell^{(0)} = 1/N$. This is indeed also a configuration in which the cosmology is matter dominated, as seen in the bottom panel of Figure 2.4. All configurations in Figure 2.4 were computed using the sliding-window algorithm for computing ϵ-Stasis, with $\epsilon = 0.1$.

The lower bounds of α and γ in Figure 2.4 are chosen to illustrate an effective cutoff in $\Omega_\ell^{(0)}$ in which the species completely decouple from the stasis phenomena. That is, their abundance spectra are so stretched that the individual species achieve their peak and decay such that any possible stasis epoch would occur with little to no matter in the Universe. In other words, in order for to achieve stasis with a non-trivial mixture of components, the prior on abundances should allow for an effective number of species to be abundant enough at their peaks such that there is a possibility of a stasis epoch with $\overline{\Omega}_M > 0$. This is seen clearly in both choices of priors, with the cutoff occurring for $\alpha \leq -30$ and $\alpha \geq 4$.

It is remarkable that stasis can emerge in some robust configurations with random (albeit sort-correlated) abundances and decay rates! We emphasize that there is no optimization or gradient information being used here; this is simply the result of sampling from distributions and sort-correlating the Γ_ℓ and $\Omega_\ell^{(0)}$ spectra before entering the simulation. Further, it is interesting to see how persistent in e-folds an exponential model of stasis can be compared to a power law; yielding close to 60 e-folds in certain cases with just $N = 100$ species.

2.4 Stasis-Conditioned Bayesian Posteriors

We have so far operated in a data-driven setting without any explicit statistical modeling. In this section, we wish to perform a Bayesian analysis on the complete $2N$-dimensional space of $\Omega_\ell^{(0)}$ and Γ_ℓ. Probabilistic inference on such a high-dimensional space using traditional techniques such as MCMC would be computationally infeasible. Nonetheless, in the current age of machine learning, we can leverage techniques that exploit gradient information to make such tasks more accessible. To this end, we employ SVI, a gradient-based inference technique that approximates complex probability distributions which are often intractable.

We would like to utilize SVI on the task of searching the full input parameter space of $\theta = \{\Gamma_\ell, \Omega_\ell^{(0)}\}$ (keeping in mind that we fix the ratio $\Gamma_{N-1}/H^{(0)}$) by modeling the *posterior* distribution $p(\theta|\mathcal{N})$ conditioned on optimizing $\mathcal{N}$. At first glance, this seems like a tall order. The parameter space in question is $2N$-dimensional; it must be searched, optimized on, and result in an easily-sampled posterior. The crux of what makes this feasible is in the *expressivity* of neural networks, combined with the *information* of gradients. In a modern setting, SVI can transform Bayesian inference into a neural-network based optimization problem, thereby accelerating statistical modeling in high-dimensional spaces. We begin with a review of the essentials of Bayesian and stochastic variational inference, including the evidence lower bound, and then apply these techniques in the context of stasis.

2.4.1 The Evidence Lower Bound

We proceed to define the fundamental object in SVI, which defines the objective function that is regressed on. All Bayesian inference problems begin with Bayes' Theorem, from which we have

$$p(\theta|\mathcal{N}) = \frac{p(\mathcal{N}|\theta)p(\theta)}{p(\mathcal{N})} \ , \tag{2.39}$$

where $p(\mathcal{N}|\theta)$ is the likelihood and $p(\theta)$ is the prior distribution on θ. In a traditional setting, where there are observations to guide the posterior towards, the likelihood encODEs the probability of observed data given input parameters. It is sought to be maximized. For our purposes, the likelihood is computed via the simulation $\mathcal{S}(\theta)$ which outputs $\mathcal{N}$ e-folds of stasis. Since solving Boltzmann equations is deterministic, the simulation likelihood is exactly

$$p(\mathcal{N}|\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)}) = \delta(\mathcal{N} - \mathcal{S}(\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)})) \ , \tag{2.40}$$

where δ is the Dirac-delta function. We will see that the simulation likelihood is disadvantageous as far as introducing information that aids in searching the full parameter space. For this reason, we

adopt a simple optimization likelihood with the definition

$$p(\mathcal{N}|\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)}) \propto e^{\kappa \cdot \mathcal{N}} \ , \tag{2.41}$$

which can be potentially subject to a matter abundance constraint similar to that used in equation 2.23. We emphasize the use of $\propto$ as this optimization likelihood is, of course, not a normalized probability density. Further, when implementing SVI, we find that the use of a numerical pre-factor κ to increase the "strength" of the likelihood term is beneficial. This surrogate (unnormalized) likelihood serves the utility of encouraging the posterior distribution to explore parameter space that yields stasis; the exact motivation behind this optimization likelihood will be expanded upon in section 2.4.3. The denominator of equation 2.39, $p(\mathcal{N}) = \int p(\mathcal{N}|\theta)p(\theta)d\theta$ is known as the *Bayesian evidence* or *marginal likelihood* and is in general computationally intractable, as it requires an integral over the entire parameter space.

The crux of SVI is bypassing the direct application of Bayes' theorem by introducing a variational family $q_\phi(\theta|\mathcal{N})$ with variational parameters ϕ, which should be distinguished from the parameters θ of the statistical model. The variational family can range from anything as simple as a standard Gaussian distribution with a trainable mean and variance parameter, to a complex neural network modeling a density on θ with millions of trainable parameters ϕ. In this work, we will choose the latter.

We would like to maximize the similarity between the variational family and the true posterior by minimizing the Kullback-Leibler (KL) divergence between the two distributions, generally defined as

$$D_{KL}(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \tag{2.42}$$

for two continuous distributions P and Q. The KL divergence is a type of statistical "distance", measuring how different the two distributions are. It is always positive semi-definite, satisfying $D_{KL}(P||Q) \geq 0$; however, it is not a metric in a formal sense as it is not a symmetric quantity (i.e. $D_{KL}(P||Q) \neq D_{KL}(Q||P)$).

Proceeding with our derivation, after some algebra and substitutions of Bayes' Theorem we arrive at

$$\begin{aligned} D_{KL}(q_\phi(\theta)||p(\theta|\mathcal{N})) = & -\mathbb{E}_{q_\phi(\theta)}\left[\log p(\mathcal{N}|\theta) - \log q_\phi(\theta)\right. \\ & \left. + \log p(\theta)\right] + \log p(\mathcal{N}) \ , \end{aligned} \tag{2.43}$$

where the first term, known as the Evidence Lower Bound (ELBO), is the only term on that RHS that depends on the variational parameters. This is a fundamental object in SVI. It is easy to interpret

the purpose of the ELBO when written as

$$\text{ELBO}(\phi, \theta, \mathcal{N}) = \mathbb{E}_{q_\phi(\theta)} \left[\log p(\mathcal{N}|\theta) - D_{KL} \left(q_\phi(\theta) \| p(\theta) \right) \right] \ , \tag{2.44}$$

where the first term is the expected log-likelihood, introducing information from the likelihood (simulation) to the posterior, and the second term is the KL-divergence between the variational distribution and prior, which ensures the posterior distribution has knowledge of the prior distribution, as required by Bayes' theorem. We are then free to rewrite equation (2.43) as

$$D_{KL}(q_\phi(\theta) || p(\theta|\mathcal{N})) = -\text{ELBO}(\phi, \theta, \mathcal{N}) + \log p(\mathcal{N}) \ . \tag{2.45}$$

Thus, it is clear to see that minimizing the KL divergence is equivalent to maximizing the ELBO. Moreover, since the KL divergence is positive-definite, the ELBO is the lower bound of the evidence. If $q_\phi(\theta)$ approximates $p(\theta|\mathcal{N})$ well, computing the intractable evidence is no longer needed.

We see that the ELBO is a function of both variational family parameters ϕ and simulation parameters θ. It is a loss function that the variational family is regressing on, so its gradients must also be accessible. What must be computed to optimize $q_\phi(\theta)$ is the following:

$$\nabla_\phi \text{ELBO}(\phi, \theta, \mathcal{N}) = \left\{ \underbrace{-\mathbb{E}_{q_\phi(\theta)} \left[\nabla_\phi \log q_\phi(\theta) \right]}_{\text{Variational Distribution}}, \underbrace{\mathbb{E}_{q_\phi(\theta)} \left[\nabla_\phi \log p(\mathcal{N}|\theta) \right]}_{\text{Likelihood Term}} \right\} , \tag{2.46}$$

where the expectation value is taken over $q_\phi(\theta)$. To backpropagate through the likelihood term in updating the variational family parameters, we use the chain rule

$$\nabla_\phi \log p(\mathcal{N}|\theta) = \nabla_\theta \log p(\mathcal{N}|\theta) \cdot \nabla_\phi \theta \tag{2.47}$$

where we see the first term requires the gradients with respect to θ:

$$\nabla_\theta \log p(\mathcal{N}|\theta) = \left\{ \underbrace{\frac{\partial \log p(\mathcal{N}|\theta)}{\partial \Gamma_\ell}}_{\text{Decay Rates}}, \underbrace{\frac{\partial \log p(\mathcal{N}|\theta)}{\partial \Omega_\ell^{(0)}}}_{\text{Abundances}} \right\} . \tag{2.48}$$

One can recognize from the optimization likelihood equation 2.41 that $\log p(\mathcal{N}|\theta) = \mathcal{N}$, which is precisely the output of the differentiable simulation $S(\theta)$. It is thus clear to see mathematically why gradients are necessary when implementing SVI — in order to backpropagate through the likelihood to update ϕ, the gradients with respect to θ must be known.

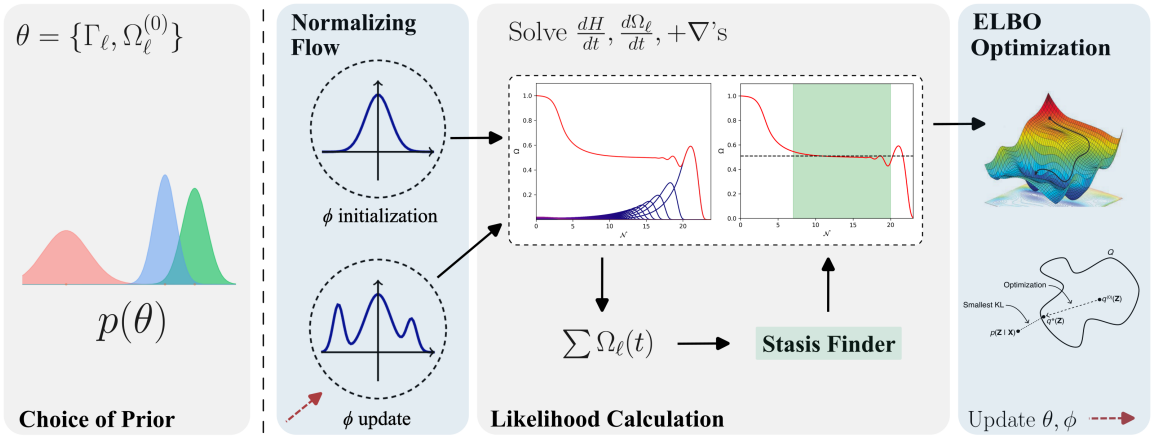


Figure 2.5: Stochastic variational inference pipeline. For a given experiment, a Bayesian prior in parameters is chosen which acts as a form of regularization in the ELBO. During training, parameters are sampled from the variational family $q_\phi(\theta)$, in this work chosen to be a BNAF. Samples are differentiably sorted before entering the stasis simulation, which solves the set of $N + 1$ coupled Boltzmann equations using `diffrax` to preserve the flow of gradients. The following $\Omega_M(t)$ curve is passed into the differentiable stasis finder to isolate the stasis e-folds $\mathcal{N}$ and the asymptotic matter abundance $\overline{\Omega}_M$. The stasis value is used in the likelihood calculation which is factored into the ELBO loss, which is used to iteratively optimize $q_\phi(\theta)$.

2.4.2 Normalizing Flows

There are several options one can consider for constructing a variational family in SVI. One of the most expressive options is the normalizing flow (NF) [125]. This method operates by transforming samples z from a simple probability density (e.g., Gaussian or Uniform) $q_0(z)$ into a complex posterior density $q_\phi(x)$. Here, x represents the transformed variables in the target distribution, achieved through a series of invertible transformations. The invertibility is crucial because it allows for both forward and inverse mappings between the base distribution and the complex target distribution, generally enabling the calculation of probability densities.

NFs are a class of bijective neural networks with trainable parameters ϕ, where $q_\phi(x) : \mathbb{R}^d \to \mathbb{R}^d$. To learn arbitrarily complex invertible functions, NFs are constructed as the composition of a series of N invertible and bijective functions f:

$$q_\phi(x) = f_N \circ f_{N-1} \circ \cdots \circ f_1(z) \ , \tag{2.49}$$

leveraging the fact that the composition of a set of invertible functions is itself still invertible.

The mapping between two density functions $q_\phi(x)$ and $q_0(z)$ is related via the absolute value of Jacobian of the transformation

$$q_\phi(x) = q_0(z) \cdot \left| \det \left(\mathbf{J}_{q_\phi(z)} \right) \right| \ . \tag{2.50}$$

Despite its simple functional form, the Jacobian computation can be prohibitively expensive for high-dimensional data or more complex architectures, where it is computed as

$$\left| \det \left(\mathbf{J}_{q_\phi(z)} \right) \right| = \prod_{i=1}^{N} \left| \det \left(\frac{\partial f_i(z_{i-1})}{\partial z_i} \right) \right| \ . \tag{2.51}$$

This matrix can in general be dense, yielding this computation $\mathcal{O}(N \cdot d^3)$ expensive.

There are a variety of methods and architectures one can choose when constructing a NF. To bypass the Jacobian computation expense, we implement the BNAF [126], a variation of the neural autoregressive flow (NAF) [127]. BNAFs employ a NN architecture where the transformations are applied in blocks, each one generating one dimension of the output at a time, conditioned on the previously generated dimensions. The autoregressive property and block structure of BNAFs leads to more stable training and mitigated computational expense, which are essential for implementing SVI for high-dimensional problems.

The crux of NAFs that makes them so efficient is that one can construct $q_\phi(x)$ such that its Jacobian is lower-triangular, and thus its determinant that must be computed in equation 2.50 is

a simple product:

$$\left|\det\left(\mathbf{J}_{q_\phi(z)}\right)\right| = \prod_{i=1}^{d} \frac{\partial f_i(z_{i-1})}{\partial z_i} \ . \tag{2.52}$$

The Jacobian is then computed with backpropagation which is $\mathcal{O}(N \cdot d)$ expensive, a clear advantage over the non-autoregressive NF. The typical NAF architecture is a set of functions $f^{(i)}$, where each $f^{(i)}$ can be decomposed into "conditioners" $c^{(i)}$ and invertible "transformers" $t^{(i)}$

$$f_\phi(x_{<i}) = t_\phi^{(i)}(x_i, c_\phi^{(i)}(x_{<i})) \ . \tag{2.53}$$

Despite this structure satisfying all the basic needs of a flow, we see that the number of parameters scales quadratically. To bypass this expense, the BNAF structure models each $t_\phi^{(i)}$ directly as an NN with no accompanying conditioner. These are all necessary ingredients for a high-dimensional problem like stasis; however, this comes at the expense of some functionality. Despite being theoretically invertible, accessing the inverse of the BNAF, which can be used to compute posterior samples' probability densities, is not currently computationally feasible.

2.4.3 Searching for Stasis Theories with SVI

We have now gathered all the essential components to conduct SVI on the stasis simulation, and will use it to sample configurations of rates and abundances that yield stasis, from which we will be able to understand trends of stasis configurations. For this, we employ `numpyro` [128], a probabilistic programming library that leverages `jax` for automatic differentiation in the implementation of SVI. Together with `diffrax`, these packages can utilize GPU computation, significantly speeding up the inference process.

We now revisit our choice of the surrogate optimization likelihood in equation 2.41 instead of the simulator likelihood in equation 2.40. The likelihood function is theoretically a normalized PDF that encodes the probability of the observed data given the model parameters. However, in this theoretical particle cosmology setting, there are no direct observations; instead, we aim to understand the input parameter space that produces epochs of stasis. This raises the question: how can SVI be made appropriate here?

The answer lies in the definition of the optimization likelihood given in equation 2.41. In a typical Bayesian inference setting, the likelihood function captures the probability of the observed data given the model parameters. This can often be interpreted in terms of residuals between simulation outputs and observed data. *Minimizing* this residual corresponds to *maximizing*

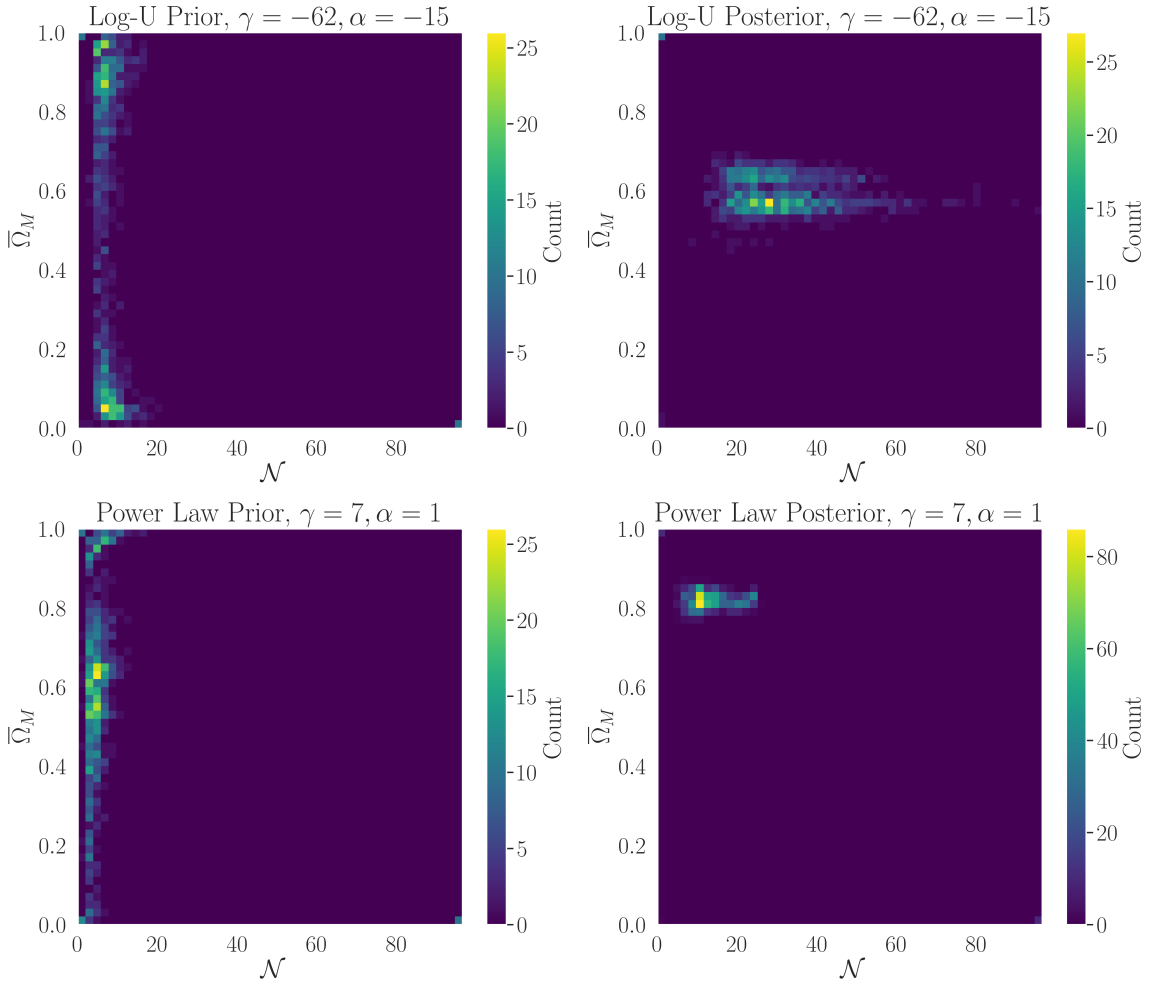


Figure 2.6: Prior and posterior comparison for choices of power law prior with $\Gamma_\ell \sim \ell^7$ and $\Omega_\ell^{(0)} \sim \ell$ and log-uniform prior with choices of $\gamma = -62$ and $\alpha = -15$ for $N = 50$. Each individual heat map is a depiction of 1000 samples and their stasis configuration in $(\mathcal{N}, \overline{\Omega}_M)$ space. For both choices of priors, there is a higher degree of stasis in the posterior in both mean and maximum value. The power law posterior features a mean stasis value of 13.59 e-folds and maximum of 25.24 e-folds, while the log-uniform posterior has a mean stasis value of 31.06 e-folds and maximum of 96.5 e-folds. Over 1000 samples, both priors had $< 1\%$ of samples achieve a stasis epoch of more than 10 e-folds, while the power law and exponential posteriors have 76% and 99%, respectively. While SVI is able to find non-trivial distributions of Γ_ℓ and $\Omega_\ell^{(0)}$ that results in epochs of stasis, it is clear to see the effect of the prior regularization in optimization in the large discrepancy between posterior configurations.

Algorithm 2 Searching Stasis with SVI

Require: ϕ variational parameters, $\theta = \{\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)}\}$ simulation parameters, $p(\Gamma_\ell)$ prior over decay rates, $p(\Omega_\ell^{(0)})$ prior over initial abundances, $\Gamma_{N-1}/H^{(0)}$ fixed early time decays relative to Hubble constant, ξ early-stopping threshold

while not converged **do**
  $z_0 \sim q_0(z)$ ▷ Sample initial latent variables
  $(\Gamma_\ell, \Omega_\ell^{(0)}) \leftarrow \mathrm{q}_\phi(z_0)$ ▷ Normalizing Flow
  $\Omega_1^{(0)} \leq \Omega_2^{(0)} \leq \ldots \leq \Omega_N^{(0)}$ ▷ Differentiably sort $\Omega_\ell^{(0)}$
  $\Gamma_1 \leq \Gamma_2 \leq \ldots \leq \Gamma_N$ ▷ Differentiably sort Γ_ℓ
  $\mathcal{N} \leftarrow \mathcal{S}(\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)})$ ▷ Run simulation
  $p(\mathcal{N}|\Gamma_\ell, \Omega_\ell^{(0)}, H^{(0)})$ ▷ Compute surrogate likelihood
  $\mathcal{L} \leftarrow \mathrm{ELBO}(\phi, \Gamma_\ell, \Omega_\ell^{(0)}, \mathcal{N})$ ▷ Compute ELBO
  $\nabla_\phi \propto \{\nabla_{\Omega_\ell^{(0)}}\mathcal{L}, \nabla_{\Gamma_\ell}\mathcal{L}\}$ ▷ Backpropagate through stasis simulation
  $\Gamma_\ell \leftarrow \mathrm{clip}(\Gamma_\ell, 0, 1)$ ▷ Ensure Γ_ℓ physical
  $\Gamma_{N-1}/H^{(0)} \leftarrow 0.1$ ▷ Enforce initial decays time scale
  $\Delta\phi \propto -\nabla_\phi \mathcal{L}(\Gamma_\ell, \Omega_\ell^{(0)})$ ▷ Update ϕ
  if $\mathcal{L}$ does not improve over ξ epochs or `NaN` loss **then**
    break ▷ Early-stopping criterion
  end if
end while

the log-likelihood which enters the ELBO, and therefore results in a concrete optimization objective. However, instead of following this traditional approach, we adopt a utility-oriented perspective of the likelihood, treating it purely as an optimization objective. The simulation likelihood given in equation 2.40 is unsuitable for optimization as it is singular and lacks a well-defined gradient. For this reason, we adopt the optimization likelihood.

A schematic of the SVI pipeline is provided in Figure 2.5, highlighting the four major components of the SVI pipeline: the prior distribution, the variational family, the simulation $\mathcal{S}(\theta)$ and stasis-finder used in the likelihood calculation, and gradient update. A more detailed description of how we employ SVI in this setting is given in Algorithm 2, where we are again reminded the importance of preserving differentiability. A methodology such as SVI requires that one be differentiable *at every step*, from differentiably sorting $\Omega_\ell^{(0)}$ and Γ_ℓ, to solving the Boltzmann equations while preserving their gradients, and calculating the stasis duration and matter abundance in a differentiable way. This leads to an uninterrupted flow of gradients through the pipeline.

Conducting these numerical analyses under the Bayesian machine learning paradigm offers both the benefits of expressivity of neural networks with the statistical rigor of Bayesian analysis. We have introduced an alternative formulation of SVI that allows one to operate without direct observations, yet still exploit the properties and benefits of SVI. This approach generally enables the study of otherwise prohibitively difficult high-dimensional parameter spaces, and can be adapted to any physical system that can be simulated.

2.4.4 Stasis Results with SVI

The analyses of the previous sections exhibited a flow in the space of parameters that preferred an exponential model of stasis. Additionally, it was shown that an exponential model had manifestly more robust stasis epochs than the power law prior corresponding to the model in [2]. Using both sets of priors, it is then natural to consider whether such flows persist when using neural networks to do inference on the full parameter space.

We run SVI with both priors (power law and log-uniform) for $N = 50$ and $\Gamma_{N-1}/H^{(0)} = 0.1$. We denote the power law prior samples as being drawn from $\Omega_\ell^{(0)} \sim \ell^\alpha$ and $\Gamma_\ell \sim \ell^\gamma$. Similarly, we denote the sort-correlated log-uniform samples as being drawn from $\Omega_\ell^{(0)} \sim \text{Log-U}(10^\alpha, 10^0)$ and $\Gamma_\ell \sim \text{Log-U}(10^\gamma, 10^0)$. A numerical prefactor of $\kappa = 10$ was used in the surrogate likelihood. We utilize a BNAF with two hidden layers and a hidden layer width of 8 neurons. A batch size of 10 was used for all experiments, corresponding to 10 log-likelihood evaluations per training

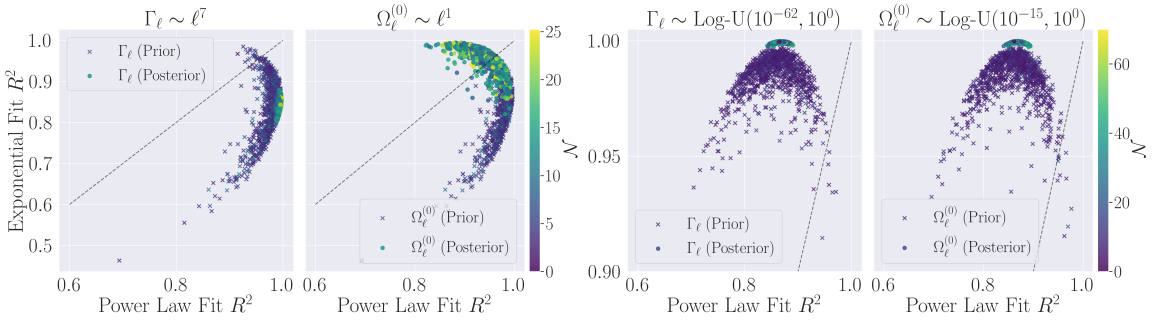


Figure 2.7: Optimizing stasis with SVI and the ELBO loss (equation 2.44) for log-uniform (α = -15, γ = -62) and power law (α = 1, γ = 7) initializations for $N = 50$ and $\Gamma_{N-1}/H^{(0)} = 0.1$. A BNAF with two hidden layers and a hidden layer width of 8 neurons was trained for 2000 epochs with Adam optimizer and early-stopping. A batch size of 10 was used in training. (**Left**) A comparison of model fit for 1000 power law prior and posterior samples. A flow toward a strictly more power law posterior is seen for Γ_ℓ, with $\Omega_\ell^{(0)}$ becoming more exponential in the posterior, with a comparable mean R^2 score for both power law and exponential fits in the posterior. (**Right**) A comparison of model fit for 1000 exponential prior and posterior samples. Posteriors for both Γ_ℓ and $\Omega_\ell^{(0)}$ become strictly more exponential. This posterior additionally has a much larger mean stasis e-folds and maximum stasis e-folds than the power law prior.

step. We use Adam [99] optimizer with an initial learning rate of $\eta = .01$ for 2000 epochs of training with an early-stopping threshold of $\xi = 200$ epochs. All training was done on a single NVIDIA A100-80GB GPU. We find that the power law prior was susceptible to a mode collapse of either completely matter dominated or radiation dominated stasis configurations using the likelihood defined in equation 2.41. To address this, an additional constraint on the likelihood was imposed, similar to the regularization condition when just using gradients in the differentiable simulation, shown in equation 2.23. The experiments with the log-uniform prior did not exhibit this mode collapse.

We find that SVI is able to learn distributions of Γ_ℓ and $\Omega_\ell^{(0)}$ that result in epochs of stasis for both choices of priors, as shown in the $\mathcal{N} - \overline{\Omega}_M$ space heat maps for 1000 posterior samples in Figure 2.6. Samples from the power law prior exhibit a mean stasis value of 4.61 e-folds and maximum of 16.51 e-folds, which after optimization is 13.59 e-folds and 25.24 e-folds in the posterior. The posterior/prior discrepancy of the exponential prior experiment is much more drastic, with prior samples exhibiting a mean of 7.46 e-folds and maximum of 19.38 e-folds, which is 31.06 e-folds and 96.49 e-folds in the posterior. While the SVI optimization has worked in both scenarios, it is clear by the discrepancies in posterior configurations the effect that the choice of prior has in the optimization, as shown in equation 2.44. This was indeed expected from the numerical study of prior configurations shown in Figure 2.4; it is therefore interesting to study the model dependence of posterior samples and their flow from model dependence in their priors.

The flow in the space of parameters of Γ_ℓ and $\Omega_\ell^{(0)}$ for both experiments is shown in Figure 2.7. For the power law prior, it is seen that Γ_ℓ and $\Omega_\ell^{(0)}$ are more confidently fit with a power law model, as expected. After training, there is a clear drift towards a more power law model in Γ_ℓ. Interestingly, it is only $\Omega_\ell^{(0)}$ that displays a drift towards an exponential model with the $\Omega_\ell^{(0)}$ posterior featuring a mean fit value of $R^2 = 0.95$ for both power law and exponential, with the mean power law R^2 value decreasing from $R^2 = 0.99$ in the prior and the mean exponential fit R^2 value increasing from $R^2 = 0.87$ in the prior. This flow of parameters is contrary to the result in which just the differentiable simulation was used (Figure 2.2), where both Γ_ℓ and $\Omega_\ell^{(0)}$ drift towards an exponential. The posterior in this case is a qualitative hybrid of both models of stasis, and even in such conditions the emergence of stasis epochs is possible. For the log-uniform prior, both Γ_ℓ and $\Omega_\ell^{(0)}$ posteriors display clear preference towards an exponential model with a mean score of $R^2 = 1$ for an exponential fit. The mean power law fit R^2 score is invariant after optimization, but it is seen in Figure 2.7 that the scatter has significantly decreased.

These results highlight that in a properly Bayesian setting, both a power law and expo-

nential model for Γ_ℓ and $\Omega_\ell^{(0)}$ are valid solutions when conditioning on producing epochs of stasis. The drift towards an exponential model of $\Omega_\ell^{(0)}$ even when having a power law prior can indicate that completely power law distributions of Γ_ℓ and $\Omega_\ell^{(0)}$ are more sparse in the full parameter space. Additionally, the discrepancy in the *statistics* of stasis in the two posteriors illuminate a sharp difference in the persistence of stasis epochs for the two types of models – a mean difference of 18 e-folds.

It could be criticized that the flow of the posterior towards an exponential model is purely due to the choice of prior. We see that this is not strictly the case as with the $\Omega_\ell^{(0)}$ parameter flow for the power law prior experiment, in which the parameter flow goes distinctly *against* the prior regularization in the ELBO. Thus, the prior plays an important role in the ELBO, but SVI can lead to results that are qualitatively at odds with the prior, due to the condition intrinsic to the posterior.

2.5 Models of Stasis

Thus far, we have remained as model-agnostic about stasis as possible, choosing only the prior distribution on rates and abundances for both our gradient ascent and Bayesian inference analyses. In both analyses, optimization was chosen to optimize stasis, with a clear preference for exponential models over power law models. In this section we wish to study an exact exponential model both analytically and numerically, compare it to exact power law models, and provide a preliminary analysis of stasis in the String axiverse; more thorough analysis will be reserved for future work. By “exact" in both contexts, we mean a model with fixed formulae for Γ_ℓ and $\Omega_\ell^{(0)}$ and no noise, as opposed to the intrinsically noisy draws of previous sections.

2.5.1 An Exact Exponential Model of Stasis

Motivated by our numerical results, we introduce an exponential model of stasis that yields parametrically more stasis than power law models, as a function of the number of species. We will show that such a model obeys the dynamical equations of stasis and will result in robust, finite epochs of stasis. We additionally derive algebraic relations with exact predictions for $\mathcal{N}$ and $\overline{\Omega}_M$ from our model. We begin my parameterizing this model as

$$\Gamma_\ell = \Gamma_N e^{\gamma(\ell-N)}, \qquad \Omega_\ell^{(0)} = \Omega_N^{(0)} e^{\alpha(\ell-N)} \tag{2.54}$$

for a spectrum of N species indexed by ℓ and general scaling factors α, γ, and Γ_N. As in the original model of stasis, α and γ control the hierarchies in abundances and rates, this time with

exponential dependence. We are also reminded that $\Omega_\ell^{(0)}(t^{(0)}) = 1$, which requires the overall normalization factor $\Omega_N^{(0)} = \left[\sum_{\ell=1}^{N} e^{\alpha(\ell-N)}\right]^{-1}$. However, in contrast with the model in [2], where the fundamental object was the mass spectrum m_ℓ, we choose to parameterize this model of stasis directly with the decay rates and abundances. We are thus ready to define the five parameter model inspired by an exponential:

$$\{\alpha, \gamma, \Gamma_N, \Omega_N^{(0)}, t^{(0)}\} \,, \tag{2.55}$$

with which shall proceed to show obeys the dynamical equations required of a stasis epoch.

We begin by reminding ourselves that a model of stasis must satisfy the constraint equations given in equations 2.10 and 2.15. Focusing on the time evolution of the total $\Omega_\ell^{(0)}$, we get

$$\sum_\ell \Omega_\ell(t) = \Omega_\ell^{(0)} h(t^{(0)}, t) \sum_{\ell=1}^{N} e^{\alpha(\ell-N)} e^{-\Gamma_N (t-t^{(0)}) e^{\gamma(\ell-N)}} \,, \tag{2.56}$$

where we have used the result from equation 2.14, but replaced the time dependence with a more general net gravitational redshift factor $h(t^{(0)}, t)$ such that there is no assumption that we are in a period of stasis. $h(t^{(0)}, t)$ invariant across all species (i.e. is independent of ℓ); a more in-depth discussion of this h-factor can be found in [2]. Moving to the $N \to \infty$ limit, we simplify this calculation by transforming the sum into an integral

$$\sum_{\ell=1}^{N} e^{\alpha(\ell-N)} e^{-\Gamma_N (t-t^{(0)}) e^{\gamma(\ell-N)}} \to \int_0^\infty e^{\alpha(\ell-N+1)} e^{-\Gamma_N (t-t^{(0)}) e^{\gamma(\ell-N+1)}} d\ell \,, \tag{2.57}$$

which can be computed by invoking the Gamma function identity $\Gamma(k)/b^k = \int_0^\infty z^{k-1} e^{-bz} dz$, yielding

$$\sum_\ell \Omega_\ell(t) = \frac{\Omega_N^{(0)}}{\gamma} \Gamma\left(\frac{\alpha}{\gamma}\right) h(t^{(0)}, t) \times \left[\Gamma_N (t - t^{(0)})\right]^{-\alpha/\gamma} \,.$$

We can solve for the quantity $\sum \Gamma_\ell \Omega_\ell(t)$ similarly, resulting in

$$\sum_\ell \Gamma_\ell \Omega_\ell(t) = \frac{\Omega_N^{(0)} \Gamma_N}{\gamma} \Gamma\left(\frac{\alpha}{\gamma} + 1\right) h(t^{(0)}, t) \times \left[\Gamma_N (t - t^{(0)})\right]^{-(\alpha/\gamma+1)} \,.$$

It is lastly straightforward to compute the ratio of these conditions, where we invoke the Gamma function identity $\Gamma(z+1)/\Gamma(z) = z$,

$$\frac{\sum_\ell \Gamma_\ell \Omega_\ell(t)}{\sum_\ell \Omega_\ell(t)} = \frac{\alpha}{\gamma} \left(\frac{1}{(t - t^{(0)})}\right) \,. \tag{2.58}$$

We see that this result is power law in the difference $(t-t^{(0)})$, dissimilar to the t^{-1} dependence from the stasis dynamical equations in equations 2.10 and 2.15. Indeed, this discrepancy also arises in the power law model of stasis. Qualitatively, this means that the stasis epoch must emerge some time after the initial species production time, when $t \gg t^{(0)}$ and hence $(t-t^{(0)})^{-1} \approx t^{-1}$. This is when the edge effects, most apparent when $\Gamma_N/H^{(0)} \gg 1$, have died away. This further indicates that our model produces *finite* epochs of stasis, with a natural beginning of the stasis epochs emerging due to the presence of edge effects, and an end when all species have decayed.

With this result in hand, we can proceed to compare coefficients in 2.15 and 2.58, in which we find

$$\frac{\alpha}{\gamma} = \frac{2(1-\overline{\Omega}_M)}{4-\overline{\Omega}_M} \,. \tag{2.59}$$

We can further invert this to obtain the prediction for $\overline{\Omega}_M$ during stasis in terms of our model parameters, resulting in

$$\overline{\Omega}_M = \frac{2(\gamma-2\alpha)}{2\gamma-\alpha} \,, \tag{2.60}$$

from which we can also find a parameter restriction for α and γ by enforcing $0 < \overline{\Omega}_M < 1$:

$$0 < \alpha < \frac{\gamma}{2} \,. \tag{2.61}$$

With this result, we can identify the model parameters that result in MRE. Solving equation 2.59 for $\overline{\Omega}_M = 1/2$, we get the condition

$$\frac{\alpha}{\gamma} = \frac{2}{7} \tag{2.62}$$

for MRE. We will study this result numerically and further compare it to the power law model of MRE.

It remains now to check that the constraints for $\sum_\ell \Omega_\ell(t)$ (equation 2.10) and $\sum_\ell \Gamma_\ell \Omega_\ell(t)$ (equation 2.15) are individually satisfied. To that end, without loss of generality we can simply show that the condition in equation 2.10 is satisfied, as we have already checked the condition of their ratio. We begin by operating under $t \gg t^{(0)}$, for which we have

$$\begin{aligned} h(t^{(0)},t) &= h(t^{(0)},t_*)h(t_*,t) \\ &= h(t^{(0)},t_*)\left(\frac{t}{t_*}\right)^{2-6/(4-\overline{\Omega}_M)} \,, \end{aligned} \tag{2.63}$$

where we have inserted the factor for $h(t_*,t)$ from equation 2.14, and $t_* \gg t^{(0)}$ is some fiducial time much later than the initial species production. Inserting this into equation 2.58, we arrive at

$$\sum_\ell \Omega_\ell(t) = \frac{\Omega_N^{(0)}}{\gamma}\Gamma\left(\frac{\alpha}{\gamma}\right) h(t^{(0)},t_*)\left(\frac{t}{t_*}\right)^{2-6/(4-\overline{\Omega}_M)} \times \left[\Gamma_N(t-t^{(0)})\right]^{-\alpha/\gamma} \,,$$

from which we see using the result in equation 2.59 and considering that $t \gg t^{(0)}$, the time dependence falls out, ensuring that we are in a period of stasis in which $\Omega_M \equiv \sum_\ell \Omega_\ell(t)$. Our exponential model of stasis is thus consistent with the conditions for stasis as defined from their corresponding dynamical equations.

We are now in a position to estimate exactly the number of e-folds of stasis expected from a configuration of our model, given by:

$$\begin{aligned} \mathcal{N} \equiv \log\left[\frac{a(t=\tau_1)}{a(t=\tau_N)}\right] &= \frac{2}{4-\overline{\Omega}_M}\log\left(\frac{\Gamma_N}{\Gamma_1}\right) \\ &= \frac{2\gamma}{4-\overline{\Omega}_M}\left(N-1\right) \ , \end{aligned} \tag{2.64}$$

where we have used the result for the time evolution of $a(t)$ in equation 2.12. We see that the exponential model of stasis yields parametrically more e-folds than the power law model, which exhibited $\mathcal{N} \sim \log(N)$ scaling. A numerical comparison of e-fold data and the theoretical prediction for $\gamma = 1$, $\alpha = 2/7$, and $\Gamma_N = 0.01$ yielding a stasis epoch with MRE, is shown in Figure 2.8. Further, see Figure 2.9 for a comparison of stasis plots with $N = 300$ species in the exponential and power law model. As expected, there is excellent agreement between data and theory in the limit $N \to \infty$. The Figure also illustrates that more than 60 e-folds of MRE can be achieved with just 100 species in the exponential model. This result shows that a relatively small number of species is needed to achieve inflation-level e-folding, which may have important phenomenological implications.

We additionally compare e-fold scaling with the power law model of stasis in Figure 2.8, where $\alpha = \delta = 1$ and γ is being varied. The Figure demonstrates the qualitative difference in e-folds for stasis epochs between the two models, even with strong power law scaling of $\Gamma_\ell \sim \ell^7$ corresponding to MRE. Specifically, the exponential model achieves the same e-folds of stasis with just $N \sim 50$ species, compared to the 400 species required for $\Gamma_\ell \sim \ell^7$. The logarithmic scaling in the power law model and the linear scaling at high N for the exponential model are both evident. Therefore, while both models produce epochs of stasis, it is clear that the exponential model results in longer epochs of stasis.

We lastly recall that in [2], it was shown that a stasis state is a *global attractor* of the dynamical system. This was demonstrated to be true for the power law model, and we can proceed to show minimally that it holds for the exponential model as well. In doing so, it is sufficient to see

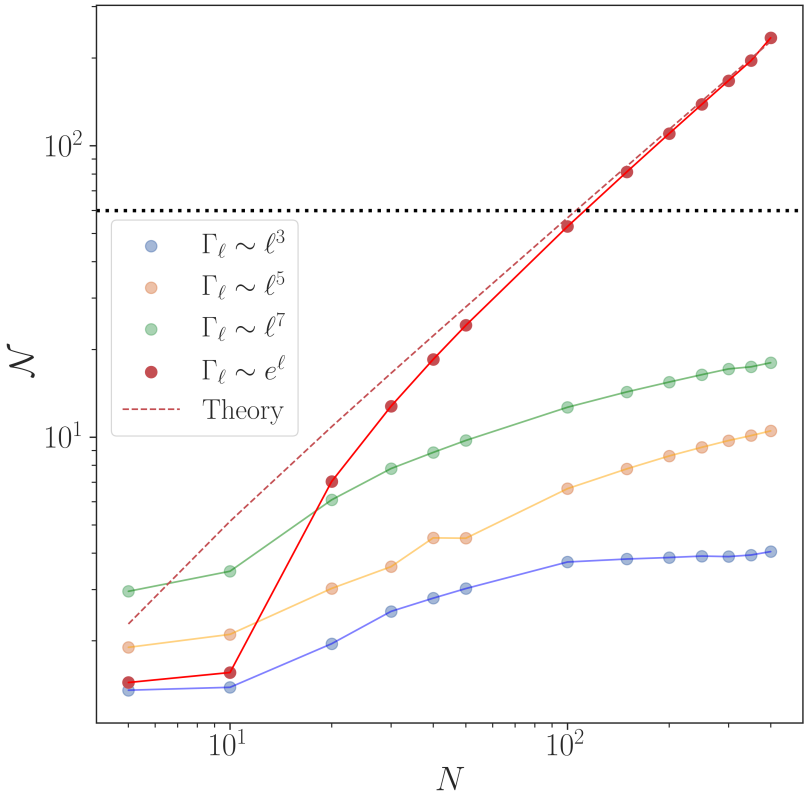


Figure 2.8: Comparison between exponential model of stasis with $\gamma = 1$, $\alpha = 2/7$, and $\Gamma_N = 0.01$ yielding matter-radiation equality, and the power law model of stasis with $\alpha = \delta = 1$. The theoretical prediction of e-fold scaling for the exponential model from equation 2.64 is shown in the red dashed line, in which we see that in the $N \to \infty$ limit there is exact agreement between numerical data and the theoretical prediction. MRE for the power law model with $\gamma = 7$ is shown in green, with the power law model exhibiting logarithmic scaling with N and the exponential model exhibiting linear scaling with N as $N \to \infty$. It is also seen that more than 60 e-folds of MRE (black dashed line) is achieved with just $N \sim 100$ species for the exponential model.

that for the exponential model

$$\sum_{\ell} \Gamma_\ell \Omega_\ell = \left(\frac{\alpha}{\gamma}\right) \frac{\Omega_M}{t - t^{(0)}} = \left[\frac{2(1-\overline{\Omega}_M)}{4-\overline{\Omega}_M}\right] \frac{\Omega_M}{t - t^{(0)}} \,, \tag{2.65}$$

where we see that this is exactly the result of equation 4.1 in [2] up to a constant factor. It then follows that the subsequent analyses showing that the stasis state is an attractor also apply to our exponential model of stasis.

2.5.2 Stasis and the String Axiverse

Axions arise readily in string theory via the dimensional reduction of higher-form gauge fields; see, e.g., [129, 130]. Notably, they need not be the QCD axion or axions that couple to any dark gauge sector. Accordingly, string theory axions can be relevant in a number of phenomenological or cosmological roles, including as the inflaton [46, 131–133], as the reheaton [134], as the QCD axion [129, 135–137], as particles that couple to photons [138, 139], and more. The largest number of axions in a known Type IIB / F-theory compactification is 181,820 [140] in a single geometry, whereas [141, 142] and [143, 144] give rise to large ensembles of geometries that typically have thousands or hundreds of axions, respectively.

Could axions give rise to stasis? Features that enable robust periods of stasis include a spectrum of N particles that may decay to radiation and non-trivial hierarchies in the decay rates. The former is clearly satisfied in many corners of the string landscape, but in lieu of a detailed analysis we would like to comment on an essential feature that affects rates: in many string constructions, axion masses are generated non-perturbatively and the ℓ^{th} axion masses appears schematically as

$$m_\ell \propto e^{-c\,T_\ell} \tag{2.66}$$

where T_ℓ is the volume of the cycle wrapped by the string instanton that generates the mass. Decays to radiation often occur via dimension five operators, in which case we have

$$\Gamma_\ell \propto e^{-3c\,T_\ell} \qquad \Omega_\ell^{(0)} \propto e^{-\alpha c\,T_\ell}, \tag{2.67}$$

where α depends on the production mechanism as discussed in [2]. Though the rates and abundances have power law dependence on masses, the exponential dependence of the non-perturbative mass causes it to depend exponentially on the cycle volumes, which are themselves sensitive to the details

of moduli stabilization. However, due to the non-perturbative effects there is a clear opportunity for exponential hierarchies, and in fact in string theoretic constructions of the models similar to the Standard Model, the exponential dependence on $T_{\rm QCD}$ sets $\Lambda_{\rm QCD}$ exponentially below the Planck scale. Furthermore, if moduli stabilization distributes T_ℓ uniformly, then the rates and abundances are log-uniform and can achieve many e-folds of stasis, as we have shown.

2.5.3 Stasis and the Emergent String Conjecture

The Emergent String Conjecture [145] is a generalization of the Swampland Distance Conjecture [146] in which a tower of either Kaluza-Klein or string states

$$m_\ell(\phi) = m_\ell(\phi_0)e^{-\lambda d(\phi,\phi_0)} \tag{2.68}$$

becomes exponentially light upon going a distance $d(\phi, \phi_0)$ from ϕ_0 toward ϕ near the boundary of moduli space. The towers still have power law masses, and therefore the number of e-folds of stasis satisfies

$$\mathcal{N} \propto \log(N), \tag{2.69}$$

but the exponential suppression of the lowest mass state yields an exponentially large number of states below a fixed cutoff. We therefore have

$$\mathcal{N} \propto \frac{\lambda d(\phi, \phi_0)}{\delta}, \tag{2.70}$$

where $m_\ell \propto \ell^\delta$. We therefore expect stasis to become increasingly important as the boundary of moduli space is approached, potentially leading to experimental constraints or observational consequences.

2.6 Summary & Discussion

In this chapter, we have studied $M \to \gamma$ cosmological stasis. In this scenario, a tower of matter states initially dominates the energy density of the Universe, but subsequent decays to radiation balance against Hubble expansion, leading to a cosmological epoch with approximately constant matter and radiation abundances.

A central result of this chapter is a new exponential model of stasis that was motivated by our numerical approach, searching the full $2N$-dimensional parameter space of decay rates Γ_ℓ and initial abundances $\Omega_\ell^{(0)}$ using gradients from a differentiable Boltzmann solver, as well as stochastic

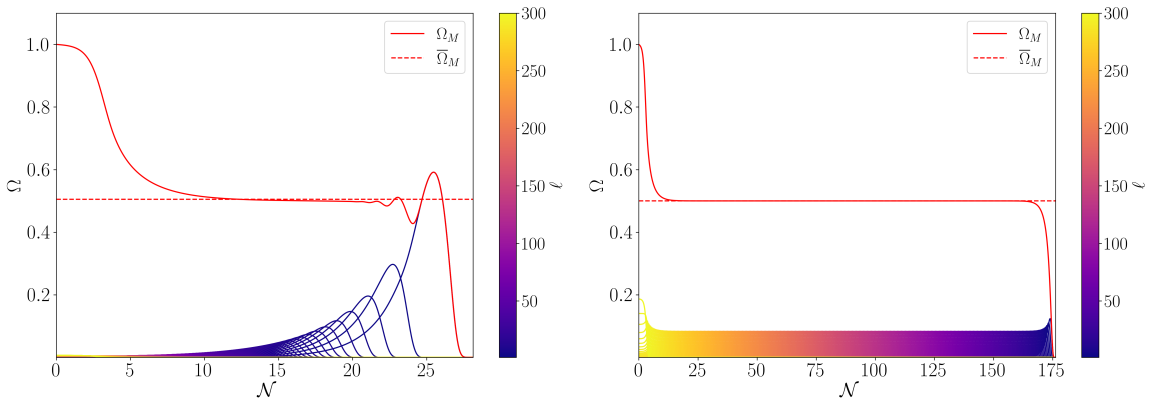


Figure 2.9: Comparison of MRE between the power law (left) and exponential (right) models of stasis, both for $N = 300$ species. MRE in the power law model from [2] corresponds to $\alpha = \delta = 1$ and $\gamma = 7$ and with $\Gamma_{N-1}/H^{(0)} = 0.01$. MRE in the exponential model corresponds to $\gamma = 1$, $\alpha = 2/7$, and $\Gamma_N/H^{(0)} = 0.01$. We see that the power law model achieves a stasis configuration of ~ 17 e-folds with 300 species, while the exponential model achieves ~ 165 e-folds, demonstrating the qualitative difference in e-folds between the two models, attributable to different scaling of e-folds with N.

variational inference with neural networks. This model leads to a longer duration of stasis in e-folds relative to a power law model and could be motivated by non-perturbative effects in string theory. We elaborate on and review our main results in what follows.

A stasis epoch could have significant cosmological implications [2], potentially affecting dark matter production, large-scale structure formation, and the overall understanding of cosmological evolution and the age of the Universe. Of particular interest for $M \to \gamma$ stasis is the possibility of stasis epochs occurring after a matter-dominated epoch at the end of inflation or after a later matter-dominated epoch prior to nucleosynthesis. In the first case, the decay of the inflaton can instead source ϕ_ℓ states. The hierarchical decays of those states during stasis would then be the source of inflationary reheating, with the Universe entering radiation domination with the conclusion of the decays. In the second case, the tower of matter fields can lead to an early matter dominated epoch prior to BBN, but well after inflation, the decays of which give rise to stasis. There are a number of potential phenomenological implications of such an epoch, but a stasis state is an attractor regardless of the arena in which ϕ_ℓ are produced.

In this work, we have taken a model-agnostic data-driven approach to studying theories of stasis, employing analytic methods only in the final stages. Such an analysis is complementary

to a model-driven approach, as e.g. in the initial model of stasis that was inspired by Kaluza-Klein excitations, motivating a mass spectrum that follows a power law in the species index. Together, the model-driven and model-agnostic approaches demonstrate that stasis is a very general phenomenon.

Our methodology began with constructing a differentiable Boltzmann solver, which is the crux of our approach. We maximized the number of e-folds of stasis by following gradients in our differentiable simulation with model-agnostic (aside from the prior) random initializations of Γ_ℓ and $\Omega_\ell^{(0)}$. Such stasis-maximizing trajectories through the space of rates and abundances motivated the study of log-uniform distributions, which we compared to power law distributions by taking random samples and solving the Boltzmann equations. This comparison provided an initial indication of the significant discrepancy in the duration of stasis epochs between the two models. With distributional priors established, we employed SVI with normalizing flows for a Bayesian analysis of the full, high-dimensional parameter space; there are some essential differences from standard SVI that we discuss in the text. From the posteriors, we again saw that stasis maximization generally prefers log-uniform distributed $\Omega_\ell^{(0)}$ and Γ_ℓ, corresponding to a model that is exponential in the species index ℓ. This motivates a hybrid power law / exponential model beyond the strict power law or exponential models we have studied, further indicating the generality of stasis.

The culmination of our numerical analyses is the analytic exponential model of stasis studied in Section 2.5. Although our study of stasis does not assume specific physical mechanisms to source the ϕ_ℓ, we find in Section 2.5.2 that such a model of stasis could be naturally realized within the string axiverse or near the boundaries of moduli space in accord with the emergent string conjecture. This new model of stasis is also an attractor.

In addition to the analytic model it inspired, our numerical methodology allows us to conclude with the following observations:

- Deviations from strict stasis (e.g., ϵ-stasis that could have small oscillations), which can still be phenomenologically relevant, indicate that $\Omega_\ell^{(0)}$ and Γ_ℓ drawn from certain families of distributions can result in epochs of stasis.

- Stasis epochs can arise without strictly correlated rates and abundances, with the species participating in stasis having abundances increasing with decay rates.

- There is a numerical preference for an exponential model when maximizing stasis; this was observed in both gradient ascent and SVI experiments.

- An exponential model of rates and abundances yields more e-folds of stasis than the power law model, with both analytics and numerics demonstrating that $\mathcal{N} \propto N$ in the exponential model whereas $\mathcal{N} \propto \log(N)$ in the power law model.

In our opinion, these model-agnostic numerical analyses together with model-driven analytic considerations together make stasis a very compelling phenomenon from a theoretical perspective. The generality of the alternative cosmological histories implies that in complex cosmologies, such as those in string theory, one should not make assumptions about the history and instead one must simply solve the Boltzmann equations.

In future work, it would be interesting to study other flavors of stasis, notably $\Lambda \to \gamma$, $\Lambda \to M$, and triple stasis with $\Lambda \to M \to \gamma$, using the numerical techniques presented in this chapter. To that end, extending the differentiable methodology to include other energy pumps, such as the overdamped/underdamped transition described in [113] used to model $\Lambda \to M$, would be essential. From a numerical perspective, these transitions can be challenging to implement due to their instantaneous (i.e., non-differentiable) nature; however, one can consider a family of numerical approximations (e.g., `tanh`, `sigmoid`) to model such transitions in a differentiable way.

It would also be interesting to change the optimization objective in future work. All of our numerical analyses were aimed at ensuring robust periods of stasis by optimizing the number of stasis e-folds $\mathcal{N}$. Of course, this optimization is different from being physically optimal, which depends crucially on theory priors, as well as phenomenological and cosmological viability. The latter considerations motivate different optimization objectives that could easily be implemented by adapting our open-source code, provided that the objective is implemented in a differentiable way. Doing so would open up new avenues for different physical studies of stasis.

We have seen that studies of theories with high-dimensional parameter spaces using differentiable simulators, despite being arduous work, can be essential to understanding the physics. Specifically, we have demonstrated the power of gradients from these simulators to direct flows in parameter space, both with and without neural networks. The methodology presented here is flexible and easily adaptable. Indeed, these techniques, in conjunction with neural networks, can provide profound physical insights, motivate new physical models, and have far-reaching implications in various scientific fields, including those beyond particle cosmology. Having explored an early-universe phenomenon, we now turn to a much later cosmological epoch – the era of galaxy formation – and show how machine learning can aid in modeling astrophysical effects (like intrinsic alignments) in large-scale structure (Chapter 3).

Chapter 3

Neural Network Emulators and Differentiable Modeling for Galaxy Intrinsic Alignments

3.1 Introduction

The spatial distribution and intrinsic shapes of galaxies encode fundamental information about the formation and evolution of cosmic structure. Galaxy clustering, quantified through correlation functions, has long served as a primary probe of cosmological parameters and the galaxy-halo connection [147, 148]. More recently, the IA of galaxy shapes have emerged as both a significant systematic for weak gravitational lensing measurements and a cosmological signal in their own right [78, 79, 86, 149]. As Stage IV surveys such as the Vera C. Rubin Observatory Legacy Survey of Space and Time [17], *Euclid* [19], and the Nancy Grace Roman Space Telescope [18] come online, precise modeling of both galaxy clustering and IA is essential for extracting unbiased cosmological constraints from weak lensing data.

IA modeling has traditionally relied on analytic approaches, such as perturbation theory [e.g. 87, 150–157]. However, these analytic models often struggle to accurately capture nonlinear effects. Simulation-based approaches offer a complementary path forward. Fully nonlinear scales can be described with a halo model [42], which provides a framework for connecting galaxies to their host dark matter halos without explicit modeling of baryonic physics. Simulation-based approaches that account for both gravitational and baryonic effects have provided profound insights

into cosmological evolution [e.g. 158–161]. These methods have the potential to better capture IA effects compared to purely analytic models. Magnetohydrodynamic simulations, often referred to as "hydro" simulations, incorporate baryonic effects but exhibit significant variance in their predictions depending on the simulation suite and the sub-grid physics models employed at sub-parsec scales. This variance includes disagreements over IA effects measured in different simulation suites [e.g. 162–164]. However, the volumes required for modern cosmological analyses are prohibitively large for hydrodynamical simulations. Given these challenges, gravity-only N-body simulations, such as *AbacusSummit* [1] and *Quijote* [165], provide a cost-effective and general alternative. These simulations avoid the need to specify sub-parsec-scale baryonic physics, but inherently lack galaxy formation and evolution processes. To bridge this gap, various HOD models have been employed to populate halos from N-body simulations with galaxies. This approach has been extended to include galaxy populations that exhibit correlated alignments [e.g. 111, 166, 167]. At the scales required for cosmological inference, emulators built on these N-body and HOD frameworks are essential for rapidly generating accurate mock galaxy catalogs across large parameter spaces.

The HOD framework provides a powerful statistical description of the galaxy-halo connection [6, 43, 168–170]. By specifying the probability that a halo of given mass hosts a certain number of galaxies, HOD models can populate dark matter halos from N-body simulations with realistic galaxy distributions. This approach has been widely successful in modeling galaxy clustering across a range of scales and redshifts [171–173]. Extensions to the basic HOD framework have incorporated galaxy properties such as color, luminosity, and stellar mass [171, 174], as well as secondary halo properties beyond mass [175, 176]. Importantly, a growing body of observational evidence from photometric, spectroscopic, and narrowband surveys has established that IA is strongly correlated with these same galaxy properties [42, 79, 177–182], suggesting that these HOD frameworks are naturally well-suited to model these dependencies compared to analytic approaches such as NLA and TATT.

[183] first introduced a halo model for IA, while subsequent work developed simulation-based approaches that assign galaxy orientations by sampling from distributions that encode alignment with halo shapes or the local tidal field [79, 184]. The HALOTOOLS package [185] provides a flexible framework for HOD modeling, which was extended by [111] to include IA modeling via the Dimroth-Watson distribution. This HALOTOOLS-IA framework parameterizes the alignment strength of central and satellite galaxies separately, enabling joint modeling of galaxy clustering and orientation statistics.

Despite its utility, HOD modeling can still be computationally demanding, as it requires

the generation of extensive galaxy catalogs from halo catalogs and the application of estimators to extract IA and clustering signals. To reduce computational cost and time associated with IA modeling, surrogate modeling and emulation based on existing simulations represents a promising avenue of research. One approach that can offer both precision and efficiency is DL, by training NN surrogates to accelerate numerical simulations. NNs have seen many different scientific applications, including in cosmology (see [186] for a review), with the availability of large datasets and powerful GPU-driven computation. They not only provide new insights, but also have the potential to accelerate numerous analyses when deployed on GPUs. The benefits of NN-based surrogate models are not exclusive to forward modeling, as the differentiability of such models can also be exploited in accelerating inverse problems using differentiable sampling techniques.

A key limitation of the HOD is its reliance on stochastic sampling procedures that are not differentiable, restricting Bayesian inference with HOD simulations to methods like Markov Chain Monte Carlo (MCMC). This precludes the use of gradient-based optimization and inference methods like Hamiltonian Monte Carlo (HMC; Duane et al. 70, Neal 71), which have proven highly efficient in machine learning and increasingly in cosmological applications (see [186] for a review). As Stage IV survey analyses begin to require inference over larger parameter spaces to account for systematics, MCMC-based inference becomes increasingly computationally demanding in high-dimensional settings. For instance, the Dark Energy Survey Year 6 result included inference over 50 nuisance parameters, in addition to cosmological parameters [187]. In [69], it was estimated that a Stage IV full 3x2 analysis would require up to 12 years of compute time on 48 CPU cores with MCMC-like samplingok, , which is contrasted with 8 days on GPU with differentiable sampling techniques. Consequently, there is a growing need for more computationally efficient (differentiable) methods that enable tractable inference over many free parameters.

[110] introduced DIFFHOD, a differentiable implementation of the HOD framework that employs continuous relaxations of discrete sampling distributions. By utilizing the Gumbel-Softmax trick [188, 189] for Bernoulli and Poisson distributions, DIFFHOD enables end-to-end gradient flow from HOD parameters through to galaxy catalogs. This approach was shown to accelerate parameter inference by orders of magnitude compared to traditional likelihood-free methods. Separately, [112] developed differentiable estimators for galaxy clustering statistics, enabling gradients to flow through Two-Point Correlation Function (2PCF) measurements.

In this chapter, we present two complementary approaches to efficient IA modeling: IAEMU, a NN-based emulator that predicts galaxy position and shape correlations from HOD parameters, and DIFFHOD-IA, a fully differentiable implementation of the HOD model with galaxy

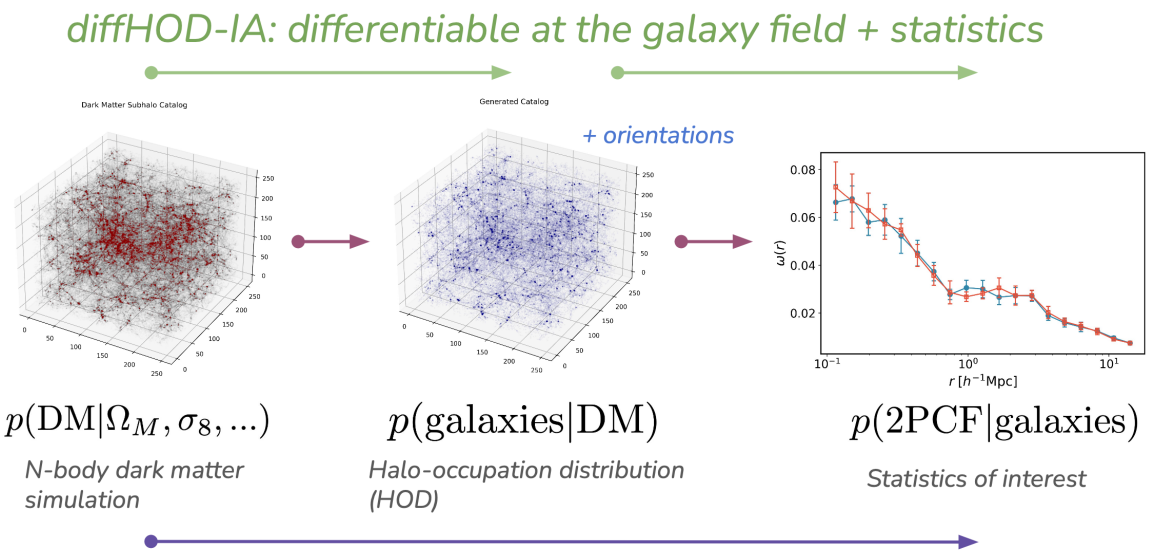


Figure 3.1: A graphic illustration of this chapter's contributions. IAEMU is a NN-based surrogate model that directly maps from HOD and IA parameters to the galaxy clustering statistic $\xi(r)$, galaxy position-orientation statistic $\omega(r)$, and galaxy orientation-orientation statistic $\eta(r)$. IAEMU bypasses the galaxy catalog generation step. DIFFHOD-IA is an end-to-end differentiable model of HALOTOOLS-IA, with tractable gradients to the galaxy catalog generation step as well as the aforementioned statistics.

IA. IAEMU is the first NN-based model to directly predict galaxy position and shape correlations from HOD parameters, eliminating the need to rerun simulations or generate explicit galaxy catalogs. This approach offers a significant speed advantage over traditional HOD-based modeling. Additionally, IAEMU successfully models galaxy shape statistics, whose stochasticity is dominated by galaxy shape noise, as discussed in [111]. IAEMU successfully captures the mean behavior of these noisier statistics, which would otherwise require multiple realizations of the underlying HOD. It also estimates galaxy shape noise (aleatoric uncertainty) and quantifies its own epistemic uncertainty – reflecting uncertainty in the predicted correlation amplitudes – primarily due to limited training data. IAEMU's uncertainty estimates enable one to assess the reliability of these predictions and further enable error propagation in modeling pipelines that incorporate IAEMU.

The DIFFHOD-IA framework extends the differentiable HOD methodology to include the orientation-dependent statistics crucial for weak gravitational lensing analyses. Our implementation combines the [6] HOD formulation with the intrinsic alignment model of [111], enabling end-to-end automatic differentiation from HOD and IA parameters through to the galaxy field. We additionally extend this framework to differentiably model correlation functions, including galaxy clustering and IA statistics. Unlike emulator-based approaches, DIFFHOD-IA provides differentiability at the catalog level, enabling its integration into field-level inference pipelines and extension to arbitrary summary statistics for next-generation weak lensing analyses.

3.1.1 Related Work

Several previous works have constructed simulation-based emulators for cosmological statistics, with a focus on matter or galaxy density. [190] constructed Gaussian process-based emulators based on the AEMULUS Project's N-body simulations for nonlinear galaxy clustering. [191] similarly used a Gaussian process-based emulator, HOD modeling, and the MIRA-TITAN Suite of N-body simulations to predict galaxy correlation functions, building on earlier work from the same group [192]. The BACCO simulation project ([193], [194]) built NN emulators to include nonlinear and baryonic effects from simulations. These projects emulate various cosmological statistics from simulations, but do not include IA. [195], [196], and [197] developed generative models trained on the TNG100 simulation [159] to emulate IA in hydrodynamic simulations, but these models do not emulate statistics. The work presented in this chapter is the first to emulate galaxy-IA correlation statistics using simulated galaxy catalogs, and the first to extend differentiable HOD modeling to include intrinsic alignments.

3.1.2 Chapter Organization

The structure of this chapter is as follows. Section 3.2 reviews the HALOTOOLS-IA formulation for HOD modeling with intrinsic alignments. Section 3.3 presents the IAEMU neural network emulator, including dataset generation, architecture, training, performance evaluation, and uncertainty quantification. Section 3.4 describes the DIFFHOD-IA differentiable implementation, including relaxed sampling procedures, differentiable Dimroth-Watson sampling, and differentiable correlation function estimators. Section 3.5 validates both approaches against the reference HALOTOOLS-IA implementation and verifies gradient accuracy. Section 3.6 demonstrates science applications including gradient-based optimization and Hamiltonian Monte Carlo inference. We summarize and discuss future directions in Section 3.7.

3.2 Halo Occupation Distribution with Intrinsic Alignments

We begin by reviewing the HALOTOOLS-IA formalism [111, 198], which builds upon the HOD formulation from [6]. Given a catalog of dark matter halos, we generate a galaxy catalog using an HOD model. This model consists of several interconnected components: (1) an occupation component, which populates halos with galaxies, (2) a phase space component, which determines the spatial distribution of galaxies within halos, and (3) an alignment component, which models galaxy intrinsic alignments. The `halotools` [199] package constructs these HOD-based galaxy catalogs following this framework. Specifically, it employs the halo model [200, 201] along with alignment models introduced in [111], providing a flexible approach for generating mock galaxy catalogs while simultaneously tracking intrinsic alignments. We refer to this extension of `halotools`, which incorporates IA information, as HALOTOOLS-IA [198]. This structure enables the rapid generation of multiple galaxy catalogs using consistent occupation, phase space, and alignment parameters. Depending on the chosen HOD parameters, a given halo may or may not host a central galaxy – the most massive galaxy residing at the halo's center. Additionally, halos may contain satellite galaxies, which are distributed throughout the halo.

The [6] formulation populates halos only using their mass. In total, the HALOTOOLS-IA model has seven free parameters, with five of them governing the halo occupation and two of them governing the galaxy IA. These parameters are as follows:

$$\{\underbrace{\log M_{\rm min}, \sigma_{\log M}, \log M_0, \log M_1, \alpha}_{\rm HOD}, \underbrace{\mu_{\rm cen}, \mu_{\rm sat}}_{\rm IA}\}. \tag{3.1}$$

The first two HOD parameters, $\log M_{\rm min}$ and $\sigma_{\log M}$, are related to the occupation of central galaxies, while $\log M_0$, $\log M_1$, and α govern satellite galaxy occupation. More detailed descriptions of the HOD parameters can be found in [6]. The IA parameterization is a two-parameter family that statistically describes the alignment strength of central and satellite galaxies with respect to their host halos, as defined in [111].

3.2.1 Central Occupation

The central galaxy occupation is characterized by the minimum halo mass required to host a central galaxy, and the width of the transition around this threshold. These are described by the parameters $\log M_{\rm min}$ and $\sigma_{\log M}$. The expected number of central galaxies in a halo of mass M is given by:

$$\langle N_{\rm cen}(M)\rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\log M}}\right)\right], \tag{3.2}$$

where erf denotes the error function. Since the presence of a central galaxy is a binary outcome, central galaxies are assigned to host halos via a Bernoulli distribution:

$$N_{\rm cen} \sim {\rm Bernoulli}(p = \langle N_{\rm cen}(M)\rangle).$$

3.2.2 Satellite Occupation

Satellite galaxy occupation follows a power-law model, governed by the parameters $\log M_0$, $\log M_1$, and α. The expected number of satellite galaxies in a halo of mass M is defined as:

$$\langle N_{\rm sat}(M)\rangle = \langle N_{\rm cen}(M)\rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \tag{3.3}$$

indicating that satellites are only present in halos that already host a central galaxy. Because multiple satellite galaxies can inhabit a single halo, they are sampled via a Poisson distribution:

$$N_{\rm sat} \sim {\rm Poisson}(\lambda = \langle N_{\rm sat}(M)\rangle). \tag{3.4}$$

We utilize a `SubhaloPhaseSpace` model for the IA in which satellite galaxies are always placed at the center of subhalos. These are preferentially placed in more massive subhalos first. In the event that the number of satellites exceeds the number of available subhalos, the satellite occupation resorts to a `NFWPhaseSpace` mode, where satellites are spatially distributed by sampling from a Navarro–Frenk–White (NFW) distribution [202].

3.2.3 Galaxy Intrinsic Alignments

In the [111] formulation, both galaxies and (sub)halos are modeled as triaxial homologous ellipsoids, meaning the orientations of the halos and galaxies can be described entirely from their axes. All galaxies are originally aligned to be parallel to their host halos. The misalignment angle of the galaxies between their host halos, $\theta_{\rm MA}$, is governed by sampling from a Dimroth-Watson distribution [203]. The Dimroth-Watson distribution is chosen to model galaxy alignments as it provides a maximum entropy distribution over a sphere, while accounting for the spin-2 symmetry of galaxy orientations. The Probability Distribution Function (PDF) for the Dimroth-Watson $P(\theta, \phi)$ is defined as

$$P(\theta, \phi) = \frac{B(\kappa)}{2\pi} e^{-\kappa \cos^2(\theta)} \sin(\theta)\, d\theta\, d\phi, \tag{3.5}$$

for polar angle $\theta = \theta_{\rm MA}$ and azimuthal angle ϕ. The normalization factor is given by

$$B(\kappa) = \frac{1}{2} \int_0^1 e^{-\kappa t^2}\, dt. \tag{3.6}$$

In this formulation, ϕ is sampled from a uniform distribution.

The fundamental parameter governing galaxy and (sub)halo alignment is κ. It is convenient to reparameterize this as

$$\mu = \frac{-2\tan^{-1}(\kappa)}{\pi}, \tag{3.7}$$

such that $\mu = \pm 1$ corresponds to perfect (mis)alignment, and $\mu = 0$ corresponds to random alignments. Both central and satellite galaxies are assigned their own alignment strengths, $\mu_{\rm cen}$ and $\mu_{\rm sat}$, respectively. HALOTOOLS-IA includes two separate alignment formulations: *subhalo alignment*, where satellites are oriented with respect to their host subhalo, and *radial alignment*, where satellites are oriented with respect to the host halo radial vector. The radial alignment model also admits two alignment strength possibilities: constant alignment strength, where it is the same for all galaxies, and distance-dependent alignment strength, where it has a power-law dependence on the distance to the central galaxy. We implement all cases in the DIFFHOD-IA code; in this chapter, all experiments utilize the radial alignment model with constant alignment strength.

3.2.4 Correlation Function Estimators

To measure the correlations in galaxy catalogs, HALOTOOLS-IA uses estimators for the position-position ($\xi(r)$), position-orientation ($\omega(r)$), and orientation-orientation ($\eta(r)$) correlations. The

$\xi(r)$ correlation is defined as

$$\xi(r) = \left\langle \frac{n(r)}{\bar{n}(r)} \right\rangle - 1, \tag{3.8}$$

where $n(r)$ is the number of galaxies separated by distance r, and $\bar{n}(r)$ is the expected number of galaxies separated by distance r for a random distribution. This equation is simpler than the Landy-Szalay estimator [204] and may be suboptimal in some cases [205], since it omits the random-pair corrections that reduce variance and account for survey geometry that are present in Landy-Szalay, making it more sensitive to noise and boundary effects. However, due to the periodic nature of the simulation box, HALOTOOLS-IA can use analytical randoms, mitigating much of this suboptimality. This estimator is also computationally faster and is sufficient for HOD models.

The $\omega(r)$ correlation is defined as

$$\omega(r) = \langle |\hat{e}(\mathbf{x}) \cdot \hat{r}|^2 \rangle - \frac{1}{3}\,, \tag{3.9}$$

and quantifies how the orientation of a galaxy at a position $\mathbf{x}$ is aligned with the positions of other galaxies at a distance $\mathbf{r}$. If $\omega(r)$ is positive, the orientation tends to align with the direction to nearby galaxies; if negative, it tends to be perpendicular. Similarly, the $\eta(r)$ correlation is defined as

$$\eta(r) = \langle |\hat{e}(\mathbf{x}) \cdot \hat{e}(\mathbf{x}+\mathbf{r})|^2 \rangle - \frac{1}{3}, \tag{3.10}$$

and measures how similarly two galaxies at positions $\mathbf{x}$ and $\mathbf{x}+\mathbf{r}$ are oriented. A positive $\eta(r)$ indicates that the orientations tend to be aligned, while a negative value means they tend to be perpendicular. For both $\omega(r)$ and $\eta(r)$, $\mathbf{x}$ is the position vector of a given galaxy, $\mathbf{r}$ is the separation vector between two galaxies, $\hat{r}$ is the unit vector of the separation vector $\mathbf{r}$, and $\hat{e}$ is the galaxy orientation unit vector that specifies the intrinsic orientation of each galaxy's major axis. The factor of $1/3$ in these equations accounts for the fact that

$$\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} \cos^2\theta \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\phi = \frac{1}{3}\,, \tag{3.11}$$

where integrating $\cos^2\theta$ over a sphere corresponds to the case of random alignments.

In this chapter, correlation functions are measured for simulated galaxies across 20 r bins, evenly spaced in logarithmic scale, between a minimum separation of $0.1\,h^{-1}\mathrm{Mpc}$ and a maximum separation of $16\,h^{-1}\mathrm{Mpc}$. In future work, the maximum range of this correlation could be extended. However, for this dataset, we chose this maximum separation because the number of galaxies n increases with r, and the computational cost of measuring correlations scales as $\mathcal{O}(n)\log(n)$. In general, galaxies at $r \leq 1\,h^{-1}\mathrm{Mpc}$ are considered to be in the "1-halo regime" (galaxies within the same halo) and galaxies outside this range are in the "2-halo regime" (galaxies residing in separate halos).

Table 3.1: BOLSHOI-PLANCK simulation parameters.

Particle Mass ($h^{-1}M_\odot$)	$\Omega_{m,0}$	σ_8	n_s	h	$L_{\rm box}$ (h^{-1}Mpc)	z
$\sim 10^8$	0.30711	0.82	0.96	0.70	250	0

3.3 IAEMU: A Neural Network Emulator

In this section, we introduce IAEMU, a neural network-based emulator designed to predict the galaxy position-position ($\xi(r)$), position-orientation ($\omega(r)$), and orientation-orientation ($\eta(r)$) correlation functions, and their associated uncertainties, using halo occupation distribution (HOD)-based mock galaxy catalogs. Compared to the simulated catalogs, IAEMU exhibits an approximately 3% average error for $\xi(r)$ and 5% for $\omega(r)$, while capturing the stochasticity in $\eta(r)$, avoiding overfitting this inherently noisier statistic. Importantly, the emulator also provides aleatoric and epistemic uncertainties, which when analyzed jointly, can help identify regions in parameter space where IAEMU's predictions may be less reliable. Furthermore, we demonstrate the model's generalization to a non-HOD based signal by fitting alignment parameters from the TNG300 hydrodynamical simulations. Since IAEMU is an NN, it enables approximately a $10{,}000\times$ speed-up in mapping HOD parameters to correlation functions when deployed on a GPU, compared to conventional CPU resources. This substantial acceleration also facilitates solving inverse problems more efficiently by supporting gradient-based sampling algorithms.

3.3.1 Dataset Generation

To train IAEMU, we generate galaxy catalogs using HALOTOOLS-IA, incorporating seven HOD and IA parameters derived from an existing dark matter halo catalog that is consistent with a realistic cosmology. We use dark matter catalogs from the BOLSHOI-PLANCK simulations, which are available directly through HALOTOOLS-IA for this purpose [206]. The BOLSHOI-PLANCK simulation is a 250 h^{-1}Mpc box, with correlations in this work measured up to 16 h^{-1}Mpc. Simulation parameters for BOLSHOI-PLANCK can be found in Table 3.1. We populate halos with galaxies following occupation equations from [6]. Further details of how we employ these occupation methods as well as a discussion of the phase space and alignment models are given in Section 3.2.

The five occupation parameters are: $\log M_{\rm min}$, $\log M_0$, $\log M_1$, α, and $\sigma_{\log M}$. To choose physically plausible values for the five occupation parameters used by these two models, we select the best-fit HOD parameter values for the Sloan Digital Sky Survey (SDSS) sample from Table 1 of

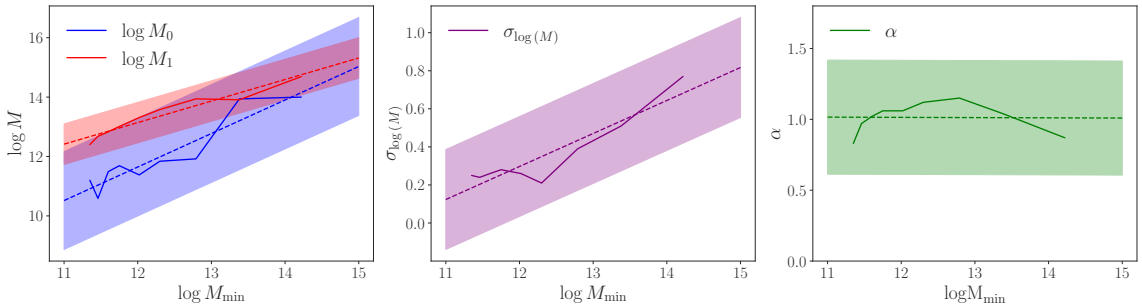


Figure 3.2: Ranges of HOD parameters used in generating the training data from HALOTOOLS-IA. We generate uniform random values for the four occupation parameters, excluding $\log M_{\rm min}$. These values are based on a linear relationship with $\log M_{\rm min}$, serving as a central line. The range for random values extends $4 \cdot$ RMSE surrounding this line. To clarify the visualization, $\sigma_{\log(M)}$ is displayed separately from other mass variables. Each panel presents published data from [6] as a solid line, while the dotted line of the same color illustrates the linear fit to $\log M_{\rm min}$, with the shaded area indicating the range for uniform random value selection for each parameter. Not shown here are the two alignment parameters, $\mu_{\rm cen}$ and $\mu_{\rm sat}$, which both vary uniformly on the range $[-1, 1]$ with no relation to these five occupation parameters.

[6]. The parameters $\log M_{\rm min}$, $\log M_0$, and $\log M_1$ control the relationship between dark matter halo masses and the likelihood of hosting central and satellite galaxies in the HOD model. Specifically, $\log M_{\rm min}$ defines the minimum halo mass required to host a central galaxy, $\log M_0$ sets the mass scale associated with the suppression of the satellite galaxy occupation, and $\log M_1$ determines the amplitude of the satellite occupation profile. The number of galaxies in a given catalog ranges from 10^5 to 10^6, with the average number decreasing with larger $\log M_{\rm min}$. The parameter α describes the asymptotic slope of satellite occupation at high halo masses, while $\sigma_{\log M}$ characterizes the width of the transition between halos that do and do not host central galaxies. Figure 3.2 shows the regions from which four of the five occupation model parameters are drawn.

The two alignment parameters, $\mu_{\rm cen}$ and $\mu_{\rm sat}$, govern the shape of the Dimroth-Watson distribution from which galaxy misalignments are sampled, as introduced in [111]. More specifically, an alignment parameter value of 0 corresponds to a uniform distribution in $\cos(\theta)$, where θ is the galaxy misalignment angle, indicating randomly oriented galaxies. Values approaching 1 indicate perfectly aligned galaxies, while values approaching -1 correspond to perpendicular alignments.

To generate training data, we generate evenly spaced values of $\log M_{\rm min}$ within the range

$[11, 15]$, covering the typical halo mass scales that host galaxies. For each of these points, we draw a value for each of the other four occupation parameters uniformly from a region $\pm 4 \cdot \mathrm{RMSE}$ around the linear fit to $\log M_{\mathrm{min}}$, where Root Mean Square Error (RMSE) refers to the root mean squared error between the fiducial SDSS values of each occupation parameter from [6] and their corresponding values predicted by the linear fit. The two alignment parameters, μ_{cen} and μ_{sat}, are each sampled uniformly on $[-1, 1]$. Earlier iterations of the dataset employed Latin hypercube sampling to fully sample the parameter space; however, this frequently resulted in galaxy samples that were unrealistic. For this reason, we restricted the sampling to values lying near the empirical trends of the SDSS-based HOD fits, which ensures that the resulting correlations remain physically plausible while still spanning a representative range of parameter space.

Despite the constrained sampling strategy, it was observed that certain input configurations could lead to an absence of galaxy pairs in specific separation bins, resulting in `NaN` values in the correlation functions. This issue frequently arises at small scales, or when the values of $\log M_{\mathrm{min}}$ are sufficiently large, making the halos that host galaxies rare. To address this, corresponding galaxy catalogs were removed. Additionally, as a further screening measure, we impose a restriction on input configurations that yield ξ/ξ_{DM} values exceeding 100, as these are deemed unphysical.

With these input parameter values, we generate galaxy catalogs using HALOTOOLS-IA and measure the three correlations described in Section 3.2.4. As HOD modeling is inherently stochastic, we generate 10 realizations of a galaxy catalog for each given set of input values for training. The multiple realizations can enable IAEMU to distinguish the signal from the shape noise of the data, and they later serve to quantify the performance of IAEMU for the noisier correlations. We note that the sample variance does also contribute to the variance in the correlations, but it is always subdominant to the shape noise (see Appendix D of [111] and Equation A3 of [207]). Thus, we can capture most of the statistical variance by re-aligning galaxies through these extra realizations. The final dataset has $110,526$ parameter choices, with 10 realizations, for a total of $1,105,260$ entries. These are split into a 70% train, 10% validation, and 20% test set with unique input parameters in each subset. The training data was generated using a combination of 2.4 GHz Intel E5-2680 CPUs and 2.1 GHz Intel Xeon Platinum 8176 CPUs. The simulations were parallelized across 150 cores, split evenly to allow simultaneous calculation of the correlation functions.

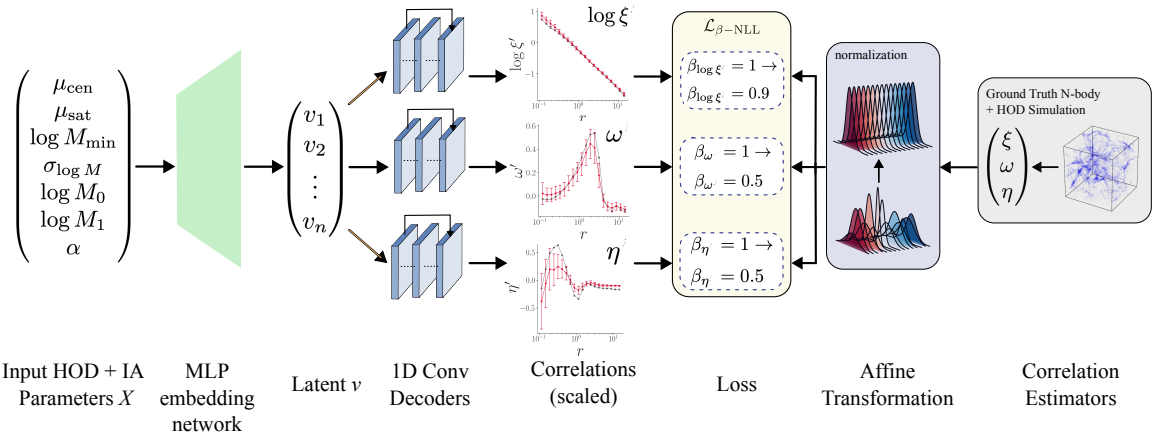


Figure 3.3: Model Pipeline. The HOD input model parameters are normalized before entering the 7-layer deep multilayer perceptron (MLP) embedding network. The embedding network expands the dimensionality of the input before a bottleneck latent space that transitions to the decoder stage, which features seven 1D convolutional layers which learn the individual local correlations present in the output correlation functions, $\log \xi$, $\omega(r)$, and $\eta(r)$. Both the embedding network and decoder feature residual connections to aid the convergence of IAEMU during training. IAEMU is trained using the β-NLL loss [7] with a 100 epoch warm-up period corresponding to mean-squared-error optimization before re-introducing aleatoric uncertainties into the optimization. The generated correlation functions are then re-scaled back to their original values. A detailed description of the model training procedure is given in [8]. N-body simulation visualization in the right panel is from [9].

3.3.2 Model Architecture

The objective is to construct an NN that replicates the mapping between HOD simulations and correlation function estimators. Specifically, the NN will take a 7-dimensional input vector of galaxy HOD and IA parameters, as described in Section 3.2 and illustrated in Figure 3.3, and predict the correlation functions $\xi(r)$, $\omega(r)$, and $\eta(r)$ across 20 r bins. We represent each correlation function by a vector recording the value for 20 evenly spaced values of r. Additionally, the model directly outputs predictions of the aleatoric uncertainties σ^{aleo} on the correlation amplitudes. This allows us to capture the stochastic nature of HOD modeling through a Mean–Variance Estimation (MVE) training procedure [208]. Separately, we use Monte Carlo dropout to track the epistemic uncertainties σ^{epi} inherent in the NN model. These can arise from limited training data or architecture misspecification. Both types of uncertainties are useful for analyzing $\omega(r)$ and $\eta(r)$ performance, which are inherently noisier statistics due to the significant effects of galaxy shape noise in correlations [209].

Mathematically, the task mapping is a function

$$f_\phi \colon \mathbb{R}^7 \to \mathbb{R}^{2\times 20} \times \mathbb{R}^{2\times 20} \times \mathbb{R}^{2\times 20}$$

where f_ϕ maps the input X to a set of mean and aleatoric uncertainty pairs:

$$X \mapsto (\underbrace{[\mu_\xi, \sigma_\xi^{\text{aleo}}]}_{\in\mathbb{R}^{2\times 20}}, \quad \underbrace{[\mu_\omega, \sigma_\omega^{\text{aleo}}]}_{\in\mathbb{R}^{2\times 20}}, \quad \underbrace{[\mu_\eta, \sigma_\eta^{\text{aleo}}]}_{\in\mathbb{R}^{2\times 20}}).$$

We implement f_ϕ as an NN called IAEMU using `PyTorch` [210]. The IAEMU architecture includes a fully connected embedding network and three 1D convolutional NN decoder heads, trained using a multitask learning approach as shown in Figure 3.3.

The embedding network contains five fully connected linear layers, each followed by batch normalization and `LeakyReLU` activation [211]. Residual connections link the second and third layers using a linear projection to match dimensions, and the third and fourth layers by directly adding the layer outputs, which improves information flow and gradient stability [101]. The embedding network increases the size of the input vector $X \in \mathbb{R}^7$ layer-by-layer to a 256-dimensional latent feature, which is then mapped through a final bottleneck layer to a 128-dimensional latent vector v. To mitigate overfitting, we incorporate dropout [212] into the IAEMU architecture during training. Additionally, as detailed later, we leverage dropout to estimate the epistemic uncertainty associated with the model's parameters using the Monte Carlo dropout technique [213].

The decoders each contain seven 1D convolutional decoder layers. Each decoder first takes the output of the embedding network, a feature vector v of size 128, and maps it into an expanded feature space. This expanded feature vector is then reshaped to create a multichannel 1D feature map, enabling the decoder to utilize 1D convolution to spatially transform the latent representation. Each layer has batch normalization, `LeakyReLU` activation, and dropout. Residual connections are introduced by adding the output of the second convolutional layer to the output of the third layer and by adding the output of the fifth layer to the output of the sixth layer. Each decoder gradually downsamples the latent representation v and finally outputs a 2-channel 1D signal as a tensor of shape 2×20, where the 2 channels represent the correlation amplitudes and variances of the correlation function, respectively. To ensure variances are strictly positive, they are passed through a `softplus` activation in the output layer. The IAEMU design serves a dual purpose: it facilitates vector-to-sequence conversion through the convolution of encoded representations from the embedding network and, within the multitask framework, enables separate forward paths to isolate features unique to each individual correlation estimator.

3.3.3 Training

We now describe the training procedure for IAEMU. We normalize each feature within the 7-dimensional input vector $X \in \mathbb{R}^7$ such that the overall distribution of each component of X has a mean of 0 and unit variance. That is, each individual feature x (i.e., a single component of X) undergoes the transformation:

$$z = \frac{x - \mu_x}{\sigma_x} \,, \tag{3.12}$$

where μ_x and σ_x are the mean and standard deviation of the respective feature across the entire training dataset. This is known as z-score standardization; it is an *affine* transformation and is thus easily invertible.

We are interested in predicting three sequences, each of length 20, corresponding to the correlation functions $\xi(r)$, $\omega(r)$, and $\eta(r)$ for $0.1\ h^{-1}\text{Mpc} < r < 16\ h^{-1}\text{Mpc}$. Since these correlations exhibit different magnitudes and characteristics, each correlation function is also standardized separately for the training of IAEMU. This ensures that each correlation is scaled to have a mean of 0 and unit variance across all bins. Without this, the loss landscape would be unevenly influenced by the differing magnitudes of the correlation functions. For example, $\xi(r)$ can exhibit strong correlations at low values of r, reaching amplitudes on the order of 10^4 or higher. In contrast, $\omega(r)$ and $\eta(r)$ exhibit amplitudes several orders of magnitude smaller than $\xi(r)$, and can also frequently

take on negative values. Applying separate standardization to each correlation function ensures that all three contribute equally to the loss landscape during training. Since $\xi(r)$ can vary over several orders of magnitude, we take its logarithm before z-score standardization. This transformation reduces skewness and can help mitigate the dominance of high-magnitude correlations in the standardization process. We thus denote the IAEMU predicted correlations as $\log\hat{\xi}(r)$, $\hat{\omega}(r)$, and $\hat{\eta}(r)$. This standardization additionally applies to the IAEMU predicted aleatoric uncertainties: $\widehat{\sigma^{\text{aleo}}_{\log\xi}}$, $\widehat{\sigma^{\text{aleo}}_{\omega}}$, and $\widehat{\sigma^{\text{aleo}}_{\eta}}$, as well as to the epistemic uncertainties: $\widehat{\sigma^{\text{epi}}_{\log\xi}}$, $\widehat{\sigma^{\text{epi}}_{\omega}}$, and $\widehat{\sigma^{\text{epi}}_{\eta}}$. All presented results are for rescaled correlations and uncertainties, with the rescaling transformations given in Section 3.3.4.

To predict the mean and variance of the values of the correlation function, we use the β-NLL loss from [7], which is defined as

$$\mathcal{L}_{\beta\text{-NLL}} = \mathbb{E}_{X,Y}\left[\hat{\sigma}^{2\beta}(X)\left(\frac{1}{2}\log\hat{\sigma}^2(X) + \frac{(Y-\hat{\mu}(X))^2}{2\hat{\sigma}^2(X)} + C\right)\right] \tag{3.13}$$

This is similar to Gaussian-NLL loss [208], defined

$$\mathcal{L}_{\text{NLL}} = \mathbb{E}_{X,Y}\left[\frac{1}{2}\log\hat{\sigma}^2(X) + \frac{(Y-\hat{\mu}(X))^2}{2\hat{\sigma}^2(X)} + C\right] , \tag{3.14}$$

where X denotes the input data vector, $\hat{\mu}(X)$ and $\hat{\sigma}^2(X)$ the model predictions at an individual bin, Y the ground truth label, and C a normalization constant. The numerator of the second term in Equation 3.14 is the typical mean-squared-error (MSE) loss, used when the model only outputs a point estimate approximating the mean of the distribution. One drawback of the Gaussian NLL loss is that the model can become stuck in local minima in the loss landscape during training. This results in a prediction with an incorrect mean and high variance. However, by adjusting β appropriately, this risk can be reduced. The utility of the β-NLL loss can be seen in the gradients:

$$\nabla_{\hat{\mu}}\mathcal{L}_{\beta\text{-NLL}}(\theta) = \mathbb{E}_{X,Y}\left[\frac{\hat{\mu}(X)-Y}{\hat{\sigma}^{(2-2\beta)}(X)}\right] , \tag{3.15}$$

$$\nabla_{\hat{\sigma}^2}\mathcal{L}_{\beta\text{-NLL}}(\theta) = \mathbb{E}_{X,Y}\left[\frac{\hat{\sigma}^2(X)-(Y-\hat{\mu}(X))^2}{2\hat{\sigma}^{(4-2\beta)}(X)}\right] . \tag{3.16}$$

The β parameter allows one to interpolate between Gaussian-NLL in the limit that $\beta \to 0$, and standard MSE in the limit that $\beta \to 1$. This loss has the benefit of allowing one to encode the contribution of the mean prediction to the loss, to discourage local minima with poor mean predictions and large variances. It was empirically found in [7] that a value of $\beta = 0.5$ generally performs best. However, we explore different values of β and introduce a warm-up period of ℓ'

epochs to enable individualized training for each correlation. The total loss function during training at epoch ℓ is:

$$\mathcal{L}(\ell;\theta) = \begin{cases} \mathcal{L}^{\xi}_{\beta\text{-NLL}}(\theta, 1.0) + \mathcal{L}^{\omega}_{\beta\text{-NLL}}(\theta, 1.0) + \mathcal{L}^{\eta}_{\beta\text{-NLL}}(\theta, 1.0) & \text{if } \ell < \ell' \\ \mathcal{L}^{\xi}_{\beta\text{-NLL}}(\theta, 0.9) + \mathcal{L}^{\omega}_{\beta\text{-NLL}}(\theta, 0.5) + \mathcal{L}^{\eta}_{\beta\text{-NLL}}(\theta, 0.5) & \text{if } \ell \geq \ell' \end{cases} \tag{3.17}$$

where we set $\beta_\xi = 0.9$ after the warm-up as this is a higher-signal correlation.

We train the IAEMU for a maximum of 500 epochs with a 100-epoch warm-up period and early stopping. We also employ gradient clipping for numerical stability, as the training of MVE networks can suffer from instability. The use of residual connections and a shallower embedding network than the decoder is to stabilize convergence during training. We employ various techniques to further aid the convergence of the model. Following the recommendations in [214], we initialize all variance output-neurons to have a bias of zero which results in a constant variance prediction across all bins at initialization, ensuring that no bins are biased towards large variances.

The choice of the β parameter in the β-NLL loss dictates the strength in which the loss interpolates between standard MSE and Gaussian-NLL loss. The optimal value of β will also not be the same for each correlation that is predicted. We implement a warm-up period during training with $\beta = 1.0$ for all correlations to maximize regression on the means before transitioning to a value of $\beta_\xi = 0.9$ and $\beta_\omega = \beta_\eta = 0.5$ for the remainder of training. The value of β_ξ was chosen as $\xi(r)$ correlations exhibit a very high signal-to-noise ratio, so the aleatoric uncertainties on these correlations are generally not significant or of interest.

We use the `AdamW` optimizer [100] with a training batch size of 128 and a step learning rate scheduler (10% decay at 167-epoch intervals with a starting learning rate of 0.01). Additional L2-regularization via a weight decay factor of 10^{-4} is used in the optimizer. All training was done on two NVIDIA A100-80GB GPUs. During training, IAEMU is validated every 5 epochs with an early stopping patience of 100 epochs based on the validation criteria. The validation criteria for saving the model is a linear combination of MSE and Gaussian-NLL losses computed for each correlation ξ, ω, and η. The total MSE and NLL losses are calculated as the sum of each correlation, and the averaged validation losses are computed over the validation dataset. The final combined validation loss $\mathcal{L}_{\text{val}}$ is defined as:

$$\mathcal{L}_{\text{val}} = \alpha \cdot \mathcal{L}_{\text{MSE}} + (1 - \alpha) \cdot \mathcal{L}_{\text{NLL}},$$

where $\alpha = 0.7$ determines the weighting between MSE and NLL, guiding model selection based on this combined criterion.

3.3.4 Correlation Rescaling

As IAEMU outputs standardized correlations, it is crucial to properly rescale the model-predicted amplitudes, aleatoric uncertainties $\widehat{\sigma^{\text{aleo}}}$, and epistemic uncertainties $\widehat{\sigma^{\text{epi}}}$ for analysis. For $\xi(r)$, we denote $\hat{\xi}$ as the (standardized) model prediction, $\bar{\xi}$ refers to the $\xi(r)$ correlations from the training dataset used for calculating statistics, and $\widehat{\sigma^{\text{aleo}}_{\xi}}$ and $\widehat{\sigma^{\text{epi}}_{\xi}}$ are the IAEMU-predicted aleatoric and epistemic uncertainties for $\xi(r)$. The reverse transformation is as follows:

$$\xi = \exp\left(\log\hat{\xi}\cdot\sigma_{\log\bar{\xi}} + \mu_{\log\bar{\xi}}\right).$$

As a result of taking the $\log$ of $\xi(r)$ for training, the rescaled aleatoric and epistemic uncertainties are only symmetric in log space. The corresponding transformation of uncertainties from log to linear space follows from the log normal moments:

$$\sigma^{\text{aleo}}_{\xi} = \sqrt{\left(e^{(\sigma^{\text{aleo}}_{\log\xi})^2} - 1\right) e^{2(\log\hat{\xi}\,\sigma_{\log\bar{\xi}} + \mu_{\log\bar{\xi}}) + (\sigma^{\text{aleo}}_{\log\xi})^2}},$$
$$\sigma^{\text{epi}}_{\xi} = \exp\big(\log\hat{\xi}\,\sigma_{\log\bar{\xi}} + \mu_{\log\bar{\xi}}\big)\cdot\sigma_{\log\bar{\xi}}\cdot\widehat{\sigma^{\text{epi}}_{\log\xi}}.$$

Here the aleatoric term follows from the exact log normal variance, while the epistemic term is obtained by linear error propagation through the exponential mapping.

The galaxy shape correlations $\omega(r)$ and $\eta(r)$ had no $\log$-scaling and therefore have a simpler inversion procedure:

$$\omega = \omega'\cdot\sigma_{\bar{\omega}} + \mu_{\bar{\omega}}, \quad \sigma^{\text{aleo}}_{\omega} = \sigma_{\bar{\omega}}\cdot\widehat{\sigma^{\text{aleo}}_{\omega}}, \quad \sigma^{\text{epi}}_{\omega} = \sigma_{\bar{\omega}}\cdot\widehat{\sigma^{\text{epi}}_{\omega}}$$
$$\eta = \eta'\cdot\sigma_{\bar{\eta}} + \mu_{\bar{\eta}}, \quad \sigma^{\text{aleo}}_{\eta} = \sigma_{\bar{\eta}}\cdot\widehat{\sigma^{\text{aleo}}_{\eta}}, \quad \sigma^{\text{epi}}_{\eta} = \sigma_{\bar{\eta}}\cdot\widehat{\sigma^{\text{epi}}_{\eta}}.$$

3.3.5 Performance

We evaluate the model on the 20% in-distribution but held-out test set, as summarized in Figure 3.4. All test-set predictions are mean predictions averaged over 50 forward passes (i.e., predictions with Monte Carlo Dropout) of IAEMU, so that an epistemic uncertainty on predictions can be retrieved. Reported metrics are evaluated on the correlations in their original domain; they are not computed in the standardized or log-transformed domain employed during IAEMU training.

In reporting metrics, we exclude outlier examples in the $\omega(r)$ and $\eta(r)$ correlations where IAEMU shows a strong Spearman correlation coefficient (> 0.5) with the ground truth or its predicted amplitude is within 1σ of the true uncertainty of the data for the majority of the bins, but

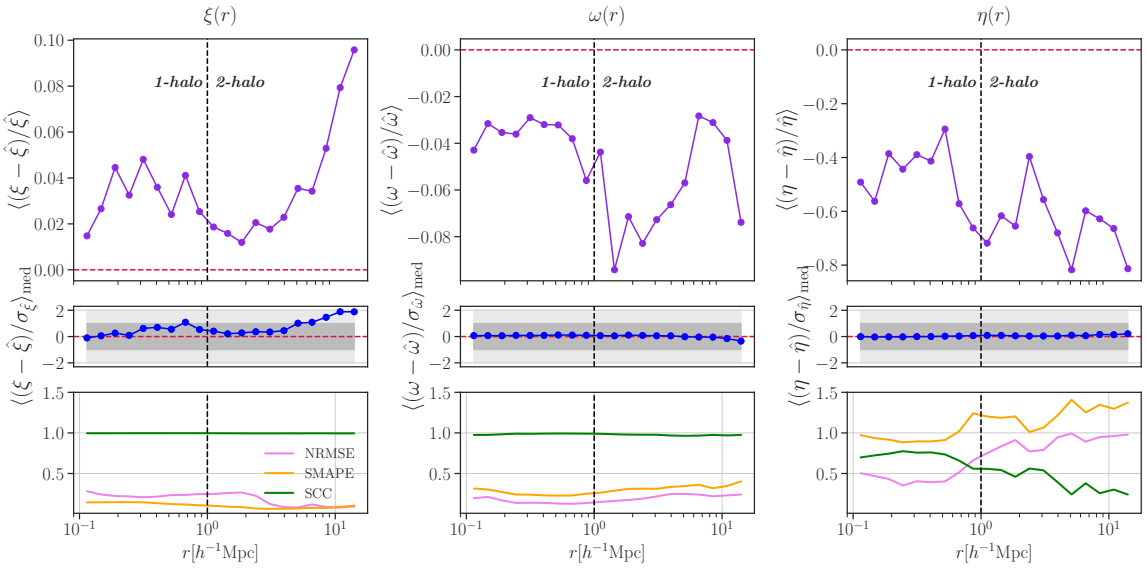


Figure 3.4: **Top:** Average fractional error for the position-position ($\xi(r)$), position-orientation ($\omega(r)$), and orientation-orientation ($\eta(r)$) correlation function predictions in the test set shown in purple. **Middle:** Median residuals of the test set predictions, expressed in units of the standard deviation of the ground truth data, $\hat{\sigma}$, obtained from 10 realizations used to construct the dataset shown in blue. **Bottom:** Per-bin Spearman correlation coefficient (SCC, green), normalized root-mean-square error (NRMSE, pink), and symmetric mean absolute percentage error (SMAPE, orange) for the correlation functions. A black vertical dashed line is included in all plots to indicate the transition in r between the 1-halo and 2-halo regimes. It is seen that $\xi(r)$ features a 3% error, on average, and $\omega(r)$ features a 5% error. Though exhibiting a larger fractional error, $\eta(r)$ predictions are on average strictly within 1σ of the true uncertainty. This similarly holds for $\omega(r)$, and $\xi(r)$ exhibits a bias at large r, reflecting the higher fractional error. Both $\xi(r)$ and $\omega(r)$ exhibit large SCC values and low NRMSE and SMAPE values across all bins, indicating good performance. For $\eta(r)$, the SCC value at low r (SCC ≥ 0.5) indicates a strong correlation between IAEMU predictions and the ground truth. This gradually decreases at the onset of the 2-halo regime, with the NRMSE and SMAPE performance decreasing as well.

the fractional error exceeds a factor of 100 and 450, respectively. These extreme values arise from the lack of a high-signal-to-noise ground truth and occur when the true correlation amplitude is small and changes sign, conditions under which the fractional error can become arbitrarily large even when predictions accurately follow the underlying signal. They are therefore artifacts of the small amplitude and noise of the data, rather than indicators of genuine model inaccuracies. The thresholds were chosen empirically to exclude only these pathological cases, affecting fewer than $\lesssim 1\%$ of test-set predictions. Including these outliers does not impact the overall assessment of performance, but significantly inflates the mean fractional error, introducing numerical instability. Notably, the median fractional error remains stable—within approximately 0.1%—for both $\omega(r)$ and $\eta(r)$, irrespective of whether these extreme cases are retained. Even with this mitigation, many instances remain in the test set where the performance of IAEMU is visually suitable, but features large fractional error due to these numerical artifacts.

For the position-position ($\xi(r)$) correlation, the mean fractional error per bin (top panel) reaches a maximum of 10%, with IAEMU achieving an average error of 3.2% for $\xi(r)$. The $\xi(r)$ performance is biased high at large r, where the $\xi(r)$ correlation amplitudes are small ($||\xi|| \ll 1$) and approach zero. This bias may arise in part from the standardization for training, which can disproportionately emphasize regions with larger amplitudes or compress the dynamic range at small values, leading to systematic bias. Additionally, as IAEMU naturally predicts $\log \xi$, small residuals near zero can appear large when transformed back to linear space.

For $\omega(r)$, the accuracy drops at the onset of the 2-halo regime, with an average model error across all bins of 4.9% and a similar maximum of $\sim 10\%$. We also find that the median fractional error across all bins is less than 10% for 66% of test-set predictions. This is approaching the accuracy for IA modeling likely required for Stage IV surveys [215]. The mean fractional error for orientation-orientation ($\eta(r)$) is significantly higher, averaging 54%. However, it is important to note that the ground truth $\omega(r)$ and $\eta(r)$ correlations – even after averaging over 10 realizations of the dataset – are generally noisy and can often fluctuate between positive and negative values. Fractional error can thus be misleading in this case due to the absence of high-signal ground truth values for comparison, and due to the correlation amplitudes being close to zero and frequently changing sign. We also studied the r-weighted mean fractional errors ((5.5%, 5.2%, and 67.4%) compared to their unweighted counterparts (3.2%, 4.9%, and 53.8%), which reveals a $\sim 2\%$ difference for $\xi(r)$, a similar performance for $\omega(r)$, and a larger difference for $\eta(r)$. This is in line with the observed bias for $\xi(r)$ at large r. In the case of $\eta(r)$, we emphasize that fractional error for $\eta(r)$ should not be interpreted in isolation as a gauge of model accuracy.

With this in mind, we show in the middle panel of Figure 3.4 the median residual in units of the dataset's true aleatoric uncertainty $\hat{\sigma}$. From this metric, it is observed that despite the large fractional error in η, the predictions of IAEMU remain strictly within 1σ of the ground truth correlations across all bins. This trend also holds for $\omega(r)$. For $\xi(r)$, the residual is computed in log space in the 2-halo regime to more consistently represent the bias with how IAEMU was trained, and to avoid exaggerated deviations caused by exponentiating small correlation values. Despite the large stochasticity of $\omega(r)$ and $\eta(r)$, this indicates that IAEMU has learned to capture the mean behavior and not overfit to the noise fluctuations in these correlations. This provides the added benefit of capturing the "cosmic mean" of the correlations directly with IAEMU, which would otherwise require running multiple realizations of the underlying HOD. This can also be frequently seen in example IAEMU predictions for $\omega(r)$ and $\eta(r)$.

Metrics

We further evaluate the performance of IAEMU using three key metrics: the Spearman Correlation Coefficient (SCC), which measures the rank correlation between predicted and true values; the Normalized Root Mean Square Error (NRMSE); and the Symmetric Mean Absolute Percentage Error (SMAPE). The SCC, which ranges between 0 and 1, is particularly useful for assessing rank-based correlations and is well-suited for analyzing sequence data. The NRMSE is defined as:

$$\text{NRMSE} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2}} \tag{3.18}$$

where y_i represents the ground truth value, and $\hat{y}_i$ denotes the corresponding prediction by IAEMU. This metric provides an indication of prediction accuracy, but can be sensitive to outliers, due to dependence on squared error. To quantify relative percentage error, we use the SMAPE, which is defined

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \epsilon} \tag{3.19}$$

where $\epsilon = 10^{-8}$ is introduced to prevent division by zero. The SMAPE is generally more robust to outliers compared to the NRMSE, but it tends to be more sensitive to small values (i.e., small correlation amplitudes). These three metrics are selected due to their scale-invariant properties, which are essential for comparing IAEMU's performance across the varying scales of ξ, ω, and η. An SCC value of 1 indicates a perfect correlation between IAEMU and the ground truth data, while lower values of NRMSE and SMAPE reflect better predictive performance. The reported metrics

are averaged over correlations in the test set. Together, these metrics provide a comprehensive assessment of IAEMU's performance.

For $\xi(r)$, we find an SCC value of 0.99 when averaged across all bins, and a value of 0.98 for $\omega(r)$ as seen in Figure 3.4. This indicates a very strong correlation between the IAEMU predictions and the underlying data. For $\eta(r)$, the average SCC across all bins is 0.55, with the SCC ≈ 0.75 at low r, but it is around 0.5 after entering the 2-halo regime, which still reflects a moderate correlation between the data and model. At larger r, the SCC decreases, indicating a weak correlation. It is important to note that the SCC can be strongly affected by stochasticity and the low amplitude of the data, particularly when the amplitudes approach zero, as is the case frequently for $\eta(r)$.

For $\xi(r)$ and $\omega(r)$, the NRMSE averaged across all bins is 0.19, as shown in the bottom panel of Figure 3.4. The corresponding SMAPE values averaged across all bins are SMAPE(ξ) $=$ 0.10 and SMAPE(ω) $= 0.30$. The relatively low NRMSE values indicate that, on average, the predictions closely follow the ground truth across the full range of data. However, the higher SMAPE for $\omega(r)$ compared to $\xi(r)$ suggests that the relative error is more pronounced for $\omega(r)$, potentially due to the generally smaller correlation amplitudes in $\omega(r)$. This implies that while the absolute prediction error remains comparable, the percentage error is exacerbated by the lower magnitude of the true values in $\omega(r)$. A similar trend is observed for $\eta(r)$, where both the NRMSE (0.69) and SMAPE (1.11) are significantly larger. These higher values indicate that IAEMU's predictions for $\eta(r)$ exhibit larger absolute and relative deviations from the ground truth. This could be attributed to a decrease in performance, increased variability, or a broader dynamic range in $\eta(r)$, which naturally poses greater challenges for accurate predictions.

Limitations

IAEMU's predictions for $\eta(r)$ are less accurate compared to $\xi(r)$ and $\omega(r)$, which perform well across metrics considered. While IAEMU successfully captures the correct scaling of $\eta(r)$ across all bins, its accuracy for $\omega(r)$ and $\eta(r)$ is primarily limited by the stochastic nature of these correlations, even when trained on multiple realizations and evaluated on their means. The averaged ground truth correlations still exhibit fluctuations that are indicative of noise due to the relatively small volume considered for the simulations. This hinders the evaluation of IAEMU's performance as well as training; however, as also demonstrated in the middle panel of Figure 3.4, IAEMU reliably captures the underlying mean behavior despite the presence of noise. The bias at large r for $\xi(r)$ can likely be

attributed to the use of $\log$ and the standardization procedure for training. In a multitask framework, standardization can potentially be avoided by using trainable loss coefficients [216].

Efficiency

We emphasize the stark difference in speed for obtaining correlations given input HOD parameters using IAEMU versus HALOTOOLS-IA. IAEMU performs inference on a batch of size $32,768$ in 1.02 seconds on a single NVIDIA A100-80GB GPU, while the HOD, when run in parallel on 150 CPU cores for the same parameters, takes approximately 3 hours. This constitutes an approximate factor of 10^4 improvement in runtime. On a single CPU core, this would constitute an improvement of roughly 10^6. While a direct comparison between a GPU and multiple CPU cores is inherently challenging due to differences in hardware architectures and parallelization capabilities, this comparison highlights the practical advantage of IAEMU in terms of computational efficiency for large-scale inference tasks with typical hardware availability. Additionally, IAEMU's compatibility with differentiable sampling algorithms allows for rapid posterior estimation, further showcasing its efficiency in inverse modeling applications.

3.3.6 Aleatoric and Epistemic Uncertainty

Due to the high stochasticity of correlations like $\omega(r)$ and $\eta(r)$, IAEMU was designed to produce *distributions* on its outputs, tracking multiple types of uncertainty, thereby enabling confidence assessment in its predictions. Aleatoric uncertainty represents the intrinsic variability in the data, in this case representing variance in the correlations due to galaxy shape noise and sample variance, as studied in [111]. The aleatoric uncertainties of $\omega(r)$ and $\eta(r)$ can thus be reduced through a larger simulation box size (resulting in more galaxies) and through multiple realizations of the same volume. Shape noise dominates over sample variance in the HOD model predictions [111], making multiple realizations important for retrieving accurate correlation functions. Epistemic uncertainties are uncertainties inherent to a model and can be large when an architecture is ill-suited for a task, or when a model is not trained on sufficient data [217]. Aleatoric uncertainties are directly output from IAEMU through its design and training procedure. Epistemic uncertainties are obtained via the Monte Carlo dropout technique [213], where dropout is used during inference across multiple forward passes. This introduces stochasticity into IAEMU's predictions, and the resulting variance in the outputs represents the epistemic uncertainty (see [217] for a review on distinguishing between aleatoric and epistemic uncertainty).

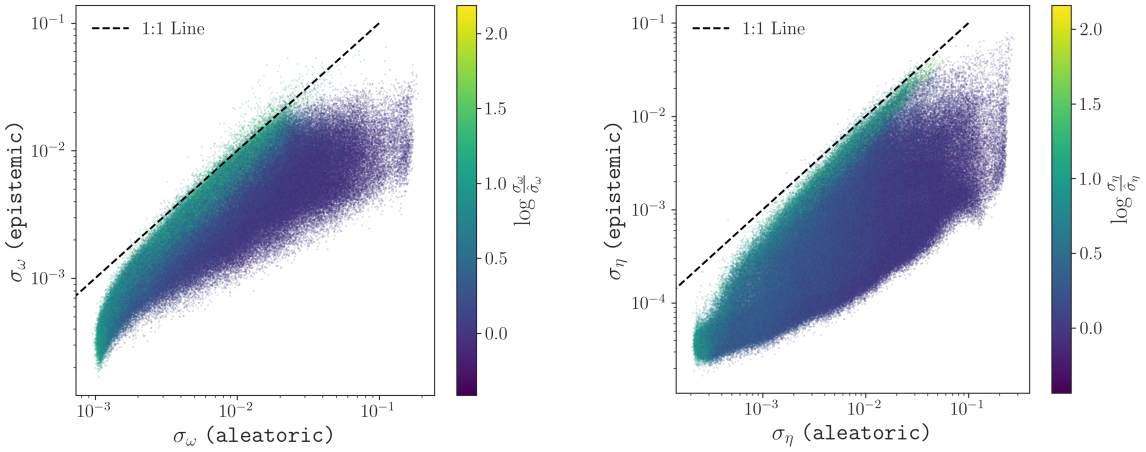


Figure 3.5: Aleatoric vs. epistemic uncertainty comparison for $\omega(r)$ and $\eta(r)$ with uncertainty bias. For test-set predictions, we analyze the total spread of aleatoric uncertainties of the data predicted by IAEMU and epistemic uncertainties due to the stochasticity of IAEMU. The coloring corresponds to the log-residual between IAEMU predicted aleatoric uncertainties and (true) aleatoric uncertainties from HALOTOOLS-IA produced from the 10 realizations used in producing the dataset. It is seen that the epistemic uncertainty is generally smaller than the aleatoric uncertainty, due to the majority of the scatter falling below the 1:1 line in aleatoric-epistemic uncertainty space. A general bias of 0.42 dex for $\omega(r)$ and 0.24 dex for $\eta(r)$ is observed between the true and predicted aleatoric uncertainties, with IAEMU uncertainty estimates being biased high. This is exacerbated near the 1:1 line, in which the epistemic uncertainty of IAEMU is comparable to the predicted aleatoric uncertainty.

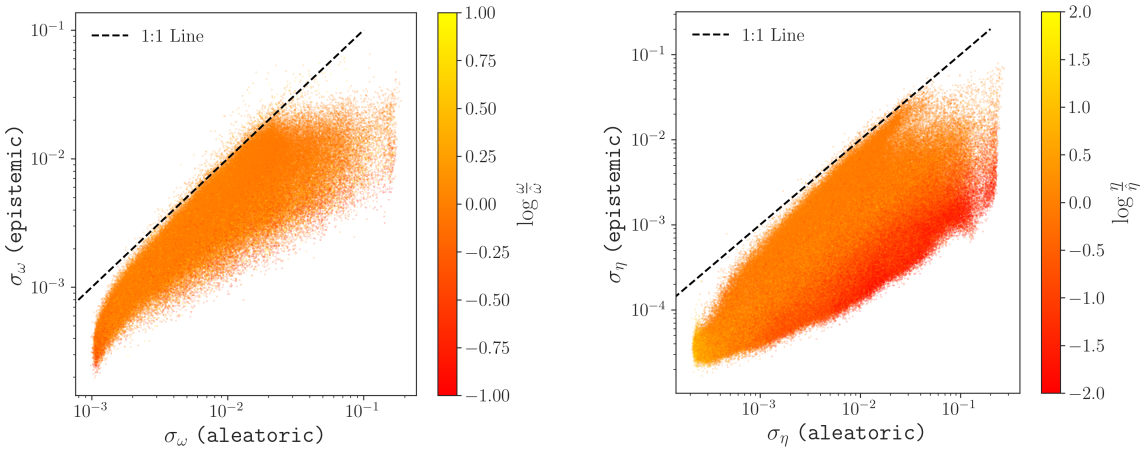


Figure 3.6: Aleatoric vs. epistemic uncertainty comparison for $\omega(r)$ and $\eta(r)$ with correlation amplitude bias. For test-set predictions, we analyze the total spread of aleatoric uncertainties of the data predicted by IAEMU and epistemic uncertainties due to the stochasticity of IAEMU. The coloring corresponds to the log-residual between IAEMU predicted correlation amplitudes and (mean) ground truth amplitudes from HALOTOOLS-IA produced from the 10 realizations used in producing the dataset. It is seen that there is no clear correlations between residuals in the amplitudes and IAEMU aleatoric and epistemic uncertainties in the case of $\omega(r)$. For $\eta(r)$, it is seen that the sharpest log-residual occurs for predictions in the region where the IAEMU aleatoric uncertainty is $\approx$ 2 dex larger than the associated epistemic uncertainties. This can be an instance of IAEMU overfitting, wherein the intrinsic uncertainty of the model on the correlation amplitude is negligible compared to the correlation's own uncertainty.

Figure 3.5 compares the aleatoric and epistemic uncertainties from IAEMU with the true aleatoric uncertainty from HALOTOOLS-IA across 10 realizations of the simulation. Figure 3.5 shows that epistemic uncertainties are generally smaller than aleatoric uncertainties, as indicated by the majority of scatter points falling below the 1:1 line. This suggests that IAEMU's architecture is sufficiently expressive for this task, and that it was not data-limited during training despite the stochasticity of these correlations. However, a median bias of 0.34 dex for $\omega(r)$ and 0.18 dex for $\eta(r)$ for aleatoric uncertainties when compared to the true aleatoric uncertainties is observed, suggesting that IAEMU is not perfectly calibrated for aleatoric uncertainties. This residual is particularly pronounced near the 1:1 line, wherein IAEMU's epistemic uncertainty predictions are comparable to the aleatoric uncertainty predictions. That is, IAEMU tends to overestimate aleatoric uncertainties in regions where the correlation amplitudes are less certain. Nevertheless, the shape noise estimates from IAEMU can provide valuable covariance information for Monte Carlo inference [108], significantly improving posterior constraints compared to inference without covariance information.

We also study the relationship between aleatoric and epistemic uncertainties in terms of the residuals in the correlation amplitudes, as shown in Figure 3.6. For $\omega(r)$, we observe a trend where the largest errors in the correlation amplitudes occur in the regime where the epistemic uncertainties are 1 dex smaller than the predicted aleatoric uncertainties. This trend is more pronounced for $\eta(r)$, where the highest errors occur when the aleatoric uncertainties are 2 dex larger than IAEMU's epistemic uncertainties, as seen in Figure 3.5. This may indicate an overconfidence for IAEMU predictions of $\eta(r)$ in this regime; however, it is also clear in comparing Figure 3.5 and Figure 3.6 that this regime is where the $\eta(r)$ correlations are the noisiest. In other words, this occurs when the galaxy shape noise is most prominent. It is thus expected that the residual on IAEMU predictions would be exaggerated due to IAEMU not overfitting to the shape noise. Nonetheless, this regime is also where IAEMU aleatoric uncertainty predictions are the most accurate.

These insights lead to the following conclusions about the performance of IAEMU for $\omega(r)$ and $\eta(r)$, and provide a useful diagnostic for gauging its accuracy in the absence of HALOTOOLS-IA ground truth data:

- IAEMU is not significantly limited by data, as evidenced by the scale of its epistemic uncertainties compared to aleatoric uncertainties for $\omega(r)$ and $\eta(r)$.
- IAEMU residuals on correlation amplitudes are largest when both the true and predicted aleatoric uncertainties of IA correlations are large, which is typically an artifact of IAEMU

learning the mean behavior of these noisier statistics.

- IAEMU tends to overestimate aleatoric uncertainties for both $\omega(r)$ and $\eta(r)$ in regimes where they are comparable to the epistemic uncertainties. This occurs when the model is most uncertain. IAEMU correlation amplitudes are still accurate in this regime, as shown in Figure 3.6.

- IAEMU aleatoric uncertainty predictions are most accurate for regions of parameter space that yield the noisiest correlations. This can be attributed to stronger gradient information with larger variance magnitudes, as seen in Equation 3.14.

In practice, one may consider both the aleatoric and epistemic uncertainties predicted by IAEMU to assess the quality of its predictions in the absence of an underlying HALOTOOLS-IA ground truth. Despite the observed bias, aleatoric uncertainty remains valuable for covariance estimation (e.g., accounting for shape noise) when performing parameter inference with IAEMU [108]. Post-hoc calibration methods, such as those discussed in [218], can help correct for these biases in parameter inference. These methods calibrate uncertainty estimates after training, ranging from non-parametric approaches like histogram binning and isotonic regression, to simple parametric schemes such as Platt scaling [219] and its extensions (e.g. temperature scaling and Dirichlet calibration). Even when the primary concern is the correlation amplitudes, the relationship between IAEMU's epistemic and aleatoric uncertainties provides valuable insight into the reliability of the predictions, as illustrated in Figures 3.5 and 3.6.

3.4 diffHOD-IA: A Differentiable Implementation

While IAEMU provides rapid emulation of correlation functions at the summary statistic level, it cannot be extended to arbitrary observables without retraining. To address this limitation and enable field-level inference, we developed DIFFHOD-IA, a fully differentiable implementation of a halo occupation distribution (HOD) model that incorporates galaxy intrinsic alignments (IA). Building upon the DIFFHOD framework of [110], we extend differentiable galaxy population modeling to include the orientation-dependent statistics crucial for weak gravitational lensing analyses.

In particular, DIFFHOD-IA allows an end-to-end differentiable mapping for:

$$\text{HOD} + \text{IA parameters} \rightarrow \text{misaligned galaxy field} \rightarrow \text{summary statistics}$$

under the HOD model of [6] and IA model of [111]. Our contributions also include differentiable modeling of 2PCFs within this framework, following the work of [112]. Differentiably computing summary statistics is a priori nontrivial, because common cosmological summaries such as the 2PCF rely on inherently discrete operations like galaxy pair counting. We will showcase the utility of DIFFHOD-IA with both 2PCFs and other objectives. Importantly, the differentiability up to the galaxy field level enables the modular extension to any differentiably computed summary statistic for a variety of applications.

For galaxy clustering, our differentiable HOD methodology closely follows the DIFFHOD methodology of [110], which we will summarize below. We extend this framework with a differentiable procedure for Dimroth-Watson sampling via inverse Cumulative Distribution Function (CDF) sampling. As we adopt their methodology closely, we refrain from extensively benchmarking the HOD component of DIFFHOD-IA, and instead focus our analyses on the IA modeling and its accuracy compared to HALOTOOLS-IA.

3.4.1 Differentiable Sampling

Sampling from distributions is not an operation whose gradients can be tracked, as an individual sample z does not encode parametric information about the distribution which it was sampled from. A common approach for backpropagating through distributions is via the *reparameterization trick*, as is extensively used in variational autoencoders [220]. In this procedure, the random variable is expressed as a deterministic function of both the distribution parameters and a source of parameter-free noise. For example, for a normally distributed variable $z \sim \mathcal{N}(\mu, \sigma^2)$, one instead samples $\epsilon \sim \mathcal{N}(0, 1)$ and rewrites z as:

$$z = \mu + \sigma \cdot \epsilon, \tag{3.20}$$

which allows gradients to be backpropagated through μ and σ.

Discrete distributions, such as the Bernoulli or Poisson distribution, assign binary values (e.g., galaxy/no galaxy) to their random variables. As a result, it is a priori difficult to differentiably sample from such distributions. One approach to this challenge is the Gumbel-Softmax trick [188], which defines a continuous relaxation of a discrete distribution, allowing gradients to be backpropagated via the reparameterization trick. Specifically, for a categorical distribution with class probabilities $\{\pi_1, \ldots, \pi_k\}$, a sample z is typically drawn as a one-hot vector using:

$$z \sim \text{Categorical}(\pi_1, \ldots, \pi_k), \tag{3.21}$$

where $z_i = 1$ indicates the selected category. However, this process is non-differentiable due to the discrete argmax operation implicit in categorical sampling and the inherently stochastic nature of sampling. To enable differentiability, one introduces the reparameterization trick by injecting independent Gumbel noise $g_i \sim \mathrm{Gumbel}(0, 1)$ – a choice that follows from the Gumbel-Max trick for sampling from categorical distributions [189] – and replacing the argmax with a differentiable $\mathrm{softmax}$ approximation. The relaxed sample y_i is then given by:

$$y_i = \frac{\exp\left((\log \pi_i + g_i)/\tau\right)}{\sum_{j=1}^{k} \exp\left((\log \pi_j + g_j)/\tau\right)}, \tag{3.22}$$

where $\tau > 0$ is a temperature parameter that controls the degree of approximation. In the limit $\tau \to 0$, the softmax converges to a hard (non-differentiable) categorical sample; for larger τ, the distribution becomes smoother and more uniform. The trade off is in the gradients, where small τ values result in a large variance of gradients, while large τ results in a smaller variance.

3.4.2 Differentiable Central Occupation

Central occupation sampling is defined by a Bernoulli distribution, as described in Section 3.2.1. We must differentiably sample

$$N_{\mathrm{cen}} \sim \mathrm{Bernoulli}(p = \langle N_{\mathrm{cen}}(M \mid M_{\mathrm{min}}, \sigma_{\log M})\rangle). \tag{3.23}$$

To this end, we utilize the Gumbel-Softmax trick in defining the Relaxed Bernoulli distribution

$$N_{\mathrm{cen}} = \frac{1}{1 + \exp\left(-\left(\log\left(\frac{p}{1-p}\right) + \epsilon\right)/\tau\right)}, \tag{3.24}$$

with $\epsilon \sim \mathrm{Logistic}(0, 1)$. We adopt a temperature value of $\tau = 0.1$, consistent with the full analysis of the occupation accuracy dependent on values of τ in [110].

3.4.3 Differentiable Satellite Occupation

Satellite occupation requires sampling from a Poisson distribution as

$$N_{\mathrm{sat}} \sim \mathrm{Poisson}(\lambda = \langle N_{\mathrm{sat}}(M \mid M_0, M_1, \alpha)\rangle). \tag{3.25}$$

In [110], it was proposed to treat each potential satellite galaxy assignment as an independently sampled Bernoulli distribution with probability $p = \lambda / N_{\mathrm{max}}$, where N_{max} is the number of trials and

λ is the Poisson rate. The resulting statistics are then Binomial distributed, which has an identical mean to a Poisson distribution. The two distributions only differ in the variance, where

$$\mathrm{Var}(N_{\mathrm{sat}}^{\mathrm{Pois.}}) = \langle N_{\mathrm{sat}} \rangle \tag{3.26}$$

$$\mathrm{Var}(N_{\mathrm{sat}}^{\mathrm{Bin.}}) = \langle N_{\mathrm{sat}} \rangle \left(1 - \frac{\langle N_{\mathrm{sat}} \rangle}{N_{\mathrm{max}}}\right). \tag{3.27}$$

In the limit $N_{\mathrm{max}} \to \infty$, the two distributions are identical in their first two moments. In this work, we use a fiducial value of $N_{\mathrm{max}} = 48$ for all experiments, in line with the results of [110].

We have now simplified satellite sampling into independent Bernoulli sampling identical to Equation 3.24 with $p = \langle N_{\mathrm{sat}} \rangle / N_{\mathrm{max}}$. In this work, all experiments use `SubhaloPhaseSpace`, wherein satellite galaxies are deterministically placed at the centers of subhalos, prioritized by subhalo mass. This prioritization is implemented differentiably by applying a softmax over the subhalos within each host halo, with logits determined by subhalo *rank*, i.e., the ordering of subhalos by decreasing mass within each host halo. Formally, the rank-weighted soft assignment q_i for subhalo i is computed as:

$$q_i = \frac{\exp\left(-\mathrm{rank}_i / t_{\mathrm{rank}}\right)}{\sum_{j \in \mathcal{H}_h} \exp\left(-\mathrm{rank}_j / t_{\mathrm{rank}}\right)}, \tag{3.28}$$

where $\mathrm{rank}_i \in \{0, 1, 2, \dots\}$ is the indexed position of subhalo i within host halo h, $\mathcal{H}_h$ is the set of all subhalos associated with that host, and $t_{\mathrm{rank}} = 0.5$ is a temperature parameter controlling the sharpness of the prioritization. These normalized weights q_i are then scaled by the expected number of satellites $\langle N_{\mathrm{sat}} \rangle$ to yield the relaxed per-subhalo satellite probabilities.

3.4.4 Differentiable NFW Sampling

In the event that $N_{\mathrm{sat}} > |\mathcal{H}_h|$, we assign the locations of all remaining satellite galaxies according to samples from a NFW distribution. We generate satellite positions via an inverse CDF sampling procedure, which we will also use for IA sampling in Section 3.4.5. Our treatment follows the approach of [110], but rather than using their closed-form approximation based on the Lambert W function, we employ a numerically stable Newton iteration to invert the CDF. Both approaches are differentiable and similarly accurate. We begin with the CDF of the NFW profile:

$$P(< r) = \frac{\ln(1 + cr/r_{\mathrm{vir}}) - cr/r_{\mathrm{vir}}/(1 + cr/r_{\mathrm{vir}})}{\ln(1 + c) - c/(1 + c)}, \tag{3.29}$$

where c is the concentration parameter of the host halo and r_{vir} is the virial radius of the host halo. We obtain position samples by drawing $u \sim \mathrm{Uniform}(0, 1)$ and solving $P(< r) = u$ for r using Newton's method for six iterations.

3.4.5 Differentiable Galaxy Intrinsic Alignments

The central and satellite galaxy misalignments are modeled by the Dimroth-Watson distribution, defined by its PDF in Equation 3.5. To differentiably sample from it, we perform inverse CDF sampling, analogous to the NFW sampling procedure described in Section 3.4.4. The original (unnormalized) marginal density over the misalignment polar angle $\theta \in [0, \pi]$ is given by:

$$p(\theta) \propto \exp(-\kappa \cos^2 \theta)\, \sin\theta. \tag{3.30}$$

Changing variables to $t = \cos\theta \in [-1, 1]$, the density becomes:

$$p(t) = \frac{1}{Z(\kappa)} \exp(-\kappa t^2), \qquad t \in [-1, 1], \tag{3.31}$$

with normalization constant

$$Z(\kappa) = \int_{-1}^{1} \exp(-\kappa t^2)\, dt. \tag{3.32}$$

To draw differentiable samples, we define the CDF

$$F(t \mid \kappa) = \frac{1}{Z(\kappa)} \int_{-1}^{t} \exp(-\kappa s^2)\, ds. \tag{3.33}$$

This integral admits closed-form expressions that depend on the sign of κ, yielding the following piecewise form of the CDF:

$$F(t \mid \kappa) = \begin{cases} \frac{1}{2}\left[1 + \dfrac{\mathrm{erf}(\sqrt{\kappa} t)}{\mathrm{erf}(\sqrt{\kappa})}\right], & \kappa > 0 \\ \frac{t+1}{2}, & \kappa = 0 \\ \frac{1}{2}\left[1 + \dfrac{\mathrm{erfi}(\sqrt{-\kappa} t)}{\mathrm{erfi}(\sqrt{-\kappa})}\right], & \kappa < 0 \end{cases} \tag{3.34}$$

As we are interested in solving $t = F^{-1}(u \mid \kappa)$, we must invert the expression in Equation 3.34 for t:

$$\cos(\theta) = \begin{cases} \frac{1}{\sqrt{\kappa}}\, \mathrm{erf}^{-1}\left[\mathrm{erf}(\sqrt{\kappa}) \cdot u\right], & \kappa > 0 \\ u, & \kappa = 0 \\ \frac{1}{\sqrt{-\kappa}}\, \mathrm{erfi}^{-1}\left[\mathrm{erfi}(\sqrt{-\kappa}) \cdot u\right], & \kappa < 0. \end{cases} \tag{3.35}$$

To generate samples, we invert the CDF by solving $t = F^{-1}(u \mid \kappa)$ for $u \sim \mathrm{Uniform}(0, 1)$. For convenience, we rescale the uniform random variable via the transformation $u' = 2u - 1$, which maps $u \in (0, 1)$ to $u' \in (-1, 1)$; we drop the prime in what follows. To construct the full 3D

orientation vector $\boldsymbol{n}$, we draw $u \sim \text{Uniform}(-1, 1)$ to obtain $t = \cos\theta$ using Equation 3.35, and independently sample $\phi \sim \text{Uniform}(0, 2\pi)$. These are converted to Cartesian coordinates as:

$$\boldsymbol{n} = \begin{bmatrix} \sqrt{1-t^2}\cos\phi \\ \sqrt{1-t^2}\sin\phi \\ t \end{bmatrix}. \tag{3.36}$$

In practice, the inverse CDF is solved via Newton's method. The resulting galaxy orientation vectors are fully differentiable with respect to the IA parameters.

3.4.6 Differentiable Correlation Functions

Typical cosmological analyses employ low dimensional summary statistics computed over the full galaxy field. We proceed by outlining a prescription to differentiably calculate 2PCFs, enabling an end-to-end differentiable pipeline from halo occupation and IA parameters to summary statistics. We find that computing the IA statistics is directly differentiable with respect to the IA parameters using discrete galaxy catalogs; computing $\xi(r)$ requires generating catalogs where the galaxies have a weight proportional to their occupation probability, as done in [112]. We note that the IA statistics $\omega(r)$ and $\eta(r)$ also depend on the HOD parameters through the galaxy positions and number counts; however, for the experiments in this work, we operate in a fixed HOD setting to validate the differentiable IA implementation. All correlation measurements use 20 logarithmically-spaced bins from $r = 0.1\ h^{-1}$Mpc to $r = 16\ h^{-1}$Mpc. We present a numerical comparison of the differentiable correlations compared with their non-differentiable counterparts in Section 3.5.1.

Weighted Galaxy Catalogs

2PCF calculations require galaxy pair counting, which is a discrete operation and not easily made differentiable. To compute 2PCFs differentiably, we follow the methodology of [112]. This requires treating galaxies as existing with a probability p, as opposed to discrete objects with weight 1 (exists) or 0 (does not exist).

For central galaxies, the weight w_i^{cen} corresponds to the mean central occupation probability from Equation 3.2:

$$w_i^{\text{cen}} = \langle N_{\text{cen}}(M) \rangle, \tag{3.37}$$

evaluated at the host halo mass M. For satellite galaxies placed in subhalos, the weight is computed as follows:

$$w_i^{\text{sat}} = q_i \cdot \langle N_{\text{sat}}(M) \rangle, \tag{3.38}$$

where q_i is the softmax weight prioritizing massive subhalos, defined in Equation 3.28. These weights encode the probability that each galaxy exists in the catalog, enabling gradient flow through the HOD parameters to clustering statistics like $\xi(r)$. Orientation-dependent statistics $\omega(r)$ and $\eta(r)$ receive gradients through the IA parameters.

Galaxy Clustering Statistics

$\xi(r)$ quantifies the excess probability of finding galaxy pairs at separation r relative to a uniform random distribution:

$$\xi(r) = \frac{DD(r)}{RR(r)} - 1, \tag{3.39}$$

where $DD(r)$ is the (weighted) count of galaxy pairs separated by distance r, and $RR(r)$ is the expected pair count if galaxies were drawn from a uniform distribution. $\xi(r)$ depends only on the HOD parameters that govern galaxy occupation; the IA parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$ do not affect $\xi(r)$, as they influence only galaxy orientations.

We pre-compute all galaxy neighbor pairs (i, j) with separations $|\mathbf{r}_{ij}| < r_{\max}$ using a KD-tree with periodic boundary conditions. The weighted pair count in each bin is computed as:

$$DD(r_k) = \sum_{(i,j)\in\mathcal{B}_k} w_i\, w_j, \tag{3.40}$$

where w_i and w_j are the galaxy occupation weights and $\mathcal{B}_k$ denotes pairs with separations falling in bin k. For the expected pair count, we use the analytic expectation:

$$RR(r_k) = \left[\left(\sum_i w_i\right)^2 - \sum_i w_i^2\right] \frac{V_k}{V_{\rm box}}, \tag{3.41}$$

where $V_k = \frac{4\pi}{3}(r_{k+1}^3 - r_k^3)$ is the shell volume and $V_{\rm box}$ is the simulation volume. The term in brackets represents the effective number of distinct galaxy pairs for the weighted galaxy catalog.

While galaxy positions are fixed in `SubhaloPhaseSpace`, the weights w_i are differentiable functions of the HOD parameters through $\langle N_{\rm cen}\rangle$ and $\langle N_{\rm sat}\rangle$. Gradients thus flow through the weight computation:

$$\frac{\partial \xi}{\partial \beta} = \sum_i \frac{\partial \xi}{\partial w_i}\frac{\partial w_i}{\partial \beta}, \tag{3.42}$$

where $\beta \in \{\log M_{\min}, \sigma_{\log M}, \log M_0, \log M_1, \alpha\}$. Satellites placed via `NFWPhaseSpace` fallback are assigned weights proportional to $\langle N_{\rm sat}\rangle$ of their host halo, preserving gradient flow. This enables gradient-based inference of HOD parameters from galaxy clustering measurements.

Intrinsic Alignment Statistics

There are two IA statistics of interest: $\omega(r)$ and $\eta(r)$, as defined in Equations 3.9 and 3.10. As with $\xi(r)$, we pre-compute all neighbor pairs (i, j) with separations $|\mathbf{r}_{ij}| < r_{\max}$ using a KD-tree with periodic boundary conditions. For each pair, we compute the separation vector with periodic wrapping. For $\omega(r)$, we compute the position-orientation alignment:

$$a_{ij}^{\omega} = (\hat{e}_i \cdot \hat{r}_{ij})^2 \,. \tag{3.43}$$

For $\eta(r)$, we compute the orientation-orientation alignment:

$$a_{ij}^{\eta} = (\hat{e}_i \cdot \hat{e}_j)^2 \,, \tag{3.44}$$

which measures the alignment between the orientation vectors of galaxies i and j.

Pairs are assigned to radial bins, and the correlation functions are estimated as:

$$\omega(r_k) = \frac{\sum_{(i,j)\in\mathcal{B}_k} a_{ij}^{\omega}}{|\mathcal{B}_k|} - \frac{1}{3} \tag{3.45}$$

and

$$\eta(r_k) = \frac{\sum_{(i,j)\in\mathcal{B}_k} a_{ij}^{\eta}}{|\mathcal{B}_k|} - \frac{1}{3}, \tag{3.46}$$

where $\mathcal{B}_k$ denotes the set of pairs in bin k and $|\mathcal{B}_k|$ is the pair count. Crucially, since galaxy positions are fixed, the pair counts $|\mathcal{B}_k|$ are constants with respect to the IA parameters. Gradients flow exclusively through the orientation vectors $\hat{e}_i$, which depend on μ_{cen} and μ_{sat} via the differentiable Dimroth-Watson sampling described in Section 3.4.5.

3.5 Validation

We validate both IAEMU and DIFFHOD-IA against the reference HALOTOOLS-IA implementation using the Bolshoi-Planck simulation. We additionally study the differentiability with respect to both the HOD and IA parameters and compare the gradient quality with finite difference methods. We do not extensively benchmark the HOD implementation, as the implementation follows that of [110].

We construct a mock observable galaxy catalog and use this fiducial catalog as a reference in the following sections. This catalog is constructed using the parameters:

$$\log M_{\min} = 12.54,\ \sigma_{\log M} = 0.26,\ \log M_0 = 12.68$$
$$\log M_1 = 13.48,\ \alpha = 1.0,\ \mu_{\text{cen}} = 0.79,\ \mu_{\text{sat}} = 0.30.$$

The HOD parameters were determined by fitting the halo occupation to match a galaxy catalog from the TNG300 [159] simulation for a stellar mass M_* cutoff of $\log M_* \geq 10.5$. This HOD is run on the Bolshoi-Planck dark matter catalog at $z = 0$, whose simulation parameters can be found in Table 3.1. This is the same HOD configuration used in [8, 111]. The IA parameters correspond to best fit parameters to the TNG300 galaxy catalog from HALOTOOLS-IA in [8].

3.5.1 Comparison of diffHOD-IA to halotools-IA

We compare the performance between DIFFHOD-IA and HALOTOOLS-IA using visualizations of the galaxy field density, comparison of 1-pt statistics (i.e., galaxy number counts, $N_{\rm gal}$), and relevant 2-pt statistics. This is shown in Figure 3.7. This comparison provides a comprehensive metric for the performance of DIFFHOD-IA.

We find excellent agreement between DIFFHOD-IA and HALOTOOLS-IA for the fiducial galaxy catalog used. The galaxy field densities produced by the two simulations are visually consistent, as shown in the top left and center panels of Figure 3.7. As galaxy placements are deterministic due to the use of `SubhaloPhaseSpace`, this is representative of agreement in the per-halo $\langle N_{\rm cen} \rangle$ and $\langle N_{\rm sat} \rangle$ between the two implementations. In both cases, approximately 0.17% of the remaining satellites were occupied according to `NFWPhaseSpace`.

To go beyond a visual comparison, this is further confirmed upon examining the histogram of $N_{\rm gal}$ across 100 realizations in the top right panel of Figure 3.7. The DIFFHOD-IA galaxy number counts are 29772 ± 109, and the HALOTOOLS-IA number counts are 29729 ± 110, illustrating excellent agreement. Minor differences in the mean values may stem from the fact that the seed mapping between DIFFHOD-IA and HALOTOOLS-IA is not one-to-one, which can lead to small deviations over a finite number of realizations.

In the bottom panels of Figure 3.7, we compare $\xi(r)$, $\omega(r)$, and $\eta(r)$ between the two implementations. This is using the (non-differentiable) HALOTOOLS-IA implementation for the correlation estimators, to highlight any potential differences at the 2PCF level coming solely from the galaxy occupation and IA. We find generally excellent agreement between DIFFHOD-IA and HALOTOOLS-IA across all three statistics. This even includes the IA statistic $\eta(r)$, which is much noisier than $\omega(r)$ due to galaxy shape noise. We additionally see comparable error bars between the implementations across 100 realizations. This is confirmed upon inspecting the fractional error of the correlations, from which we find a mean bias of 0.28% for $\xi(r)$, 0.16% for $\omega(r)$, and 1.77% for $\eta(r)$. We benchmark the accuracy of the differentiable correlation estimators as outlined in Section

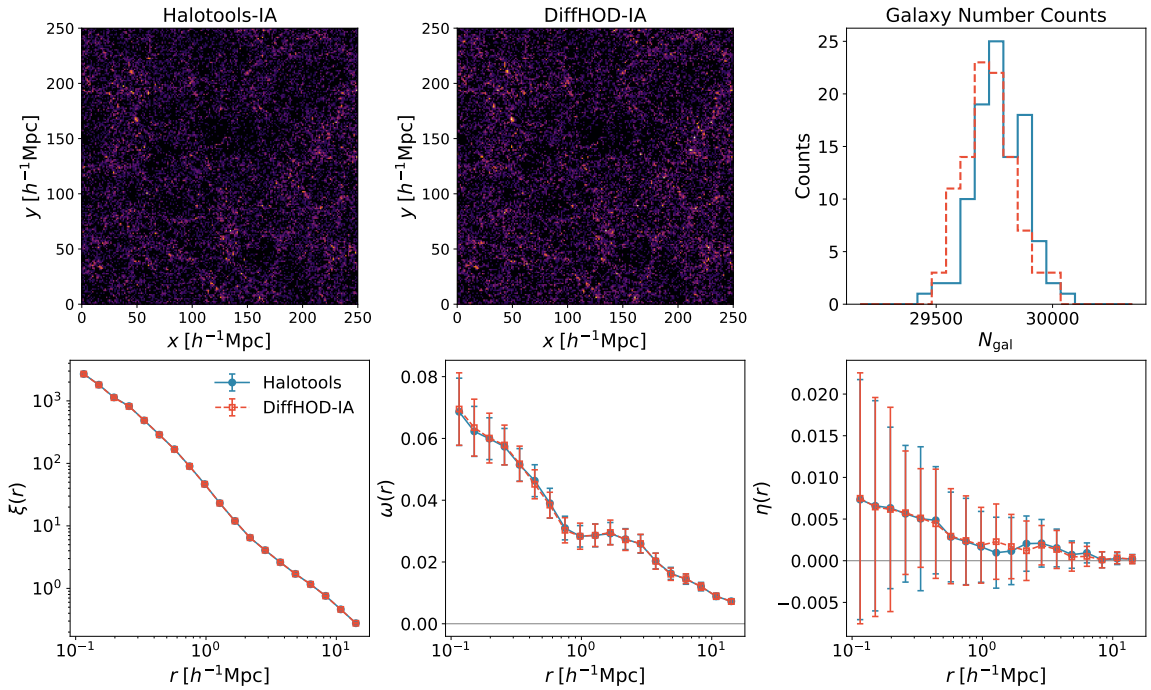


Figure 3.7: Validation of DIFFHOD-IA against the reference HALOTOOLS-IA implementation for the TNG300 fiducial HOD. **Top left and center:** Projected galaxy density fields across the simulation volume along the line of sight. Both implementations produce visually indistinguishable large-scale structure. **Top right:** Distribution of galaxy number counts $N_{\rm gal}$ across 100 realizations using identical random seeds. Both implementations produce consistent galaxy number densities with similar scatter. **Bottom left:** Galaxy position-position correlation function $\xi(r)$ averaged over 100 realizations, with error bars indicating the standard deviation across realizations. The two implementations show excellent agreement across all scales. **Bottom center and right:** Galaxy position-orientation correlation function $\omega(r)$ and orientation-orientation correlation function $\eta(r)$. These correlations show strong agreement between the two implementations across all scales. $\eta(r)$ exhibits larger statistical noise and error bars due to the effects of galaxy shape noise. Despite the noise, the two implementations remain consistent within uncertainties.

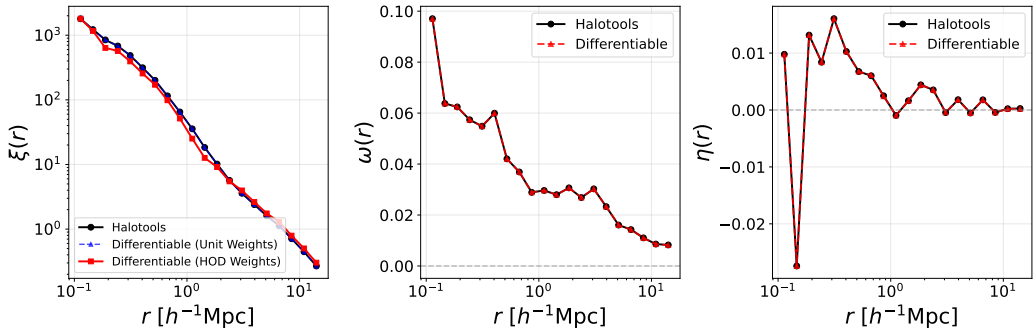


Figure 3.8: Validation of differentiable correlation function estimators against `halotools` reference implementations for the same galaxy catalog. **Left:** Galaxy position-position correlation function $\xi(r)$. The black circles show the (non-differentiable) measurement from `halotools`, while red squares show the differentiable estimator using HOD-derived occupation probability weights. **Center:** Galaxy position-orientation correlation function $\omega(r)$, comparing `halotools` (black) with our differentiable estimator (red). **Right:** Galaxy orientation-orientation correlation function $\eta(r)$, comparing `halotools` (black) with our differentiable estimator (red). Both $\omega(r)$ and $\eta(r)$ show excellent agreement between implementations, with $\eta(r)$ exhibiting larger statistical fluctuations due to galaxy shape noise. The differentiable estimators enable gradient-based inference: $\xi(r)$ gradients flow through galaxy occupation weights from HOD parameters, while $\omega(r)$ and $\eta(r)$ gradients flow through orientation vectors from IA parameters.

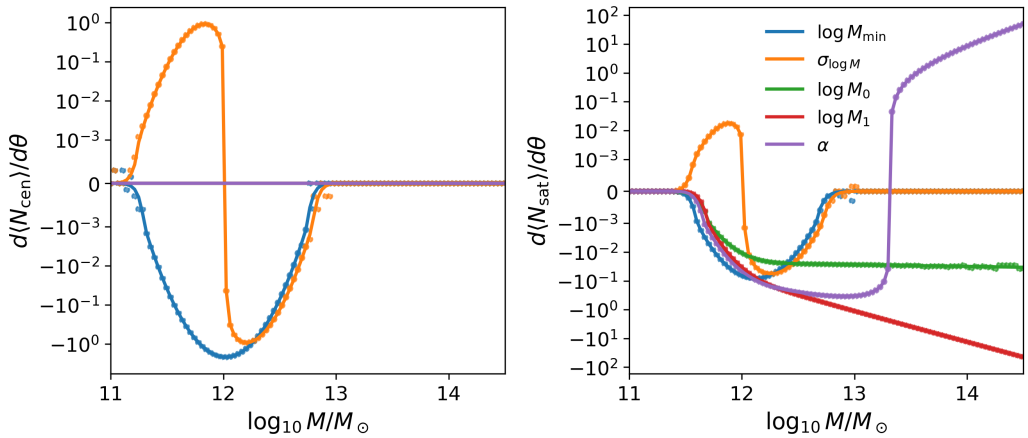


Figure 3.9: Gradients of the halo occupation distribution functions with respect to HOD parameters as a function of halo mass. **Left panel:** Gradients of the mean central galaxy occupation $\langle N_{\rm cen}\rangle$ with respect to $\log M_{\rm min}$ (blue) and $\sigma_{\log M}$ (orange). The gradients are largest in the transition region around $\log_{10} M/M_\odot \approx 12$ where the occupation probability transitions from 0 to 1, and vanish at high masses where $\langle N_{\rm cen}\rangle$ saturates to unity. **Right panel:** Gradients of the mean satellite galaxy occupation $\langle N_{\rm sat}\rangle$ with respect to all five HOD parameters: $\log M_{\rm min}$, $\sigma_{\log M}$, $\log M_0$, $\log M_1$, and α. In both panels, solid lines show gradients computed via automatic differentiation and dotted points show finite difference estimates, demonstrating excellent agreement. The IA parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$ do not affect galaxy number counts and have zero gradient everywhere. These gradients enable efficient gradient-based inference of HOD parameters from galaxy clustering observations.

3.4.6 in Appendix

We next benchmark the accuracy of the differentiable correlation estimators as outlined in Section 3.4.6. This is shown in Figure 3.8. There is exact agreement between the DIFFHOD-IA and HALOTOOLS-IA estimators for $\omega(r)$, $\eta(r)$, and $\xi(r)$ in the case of unit galaxy weights. In the case of HOD weights, which is necessary for full differentiability of $\xi(r)$ with respect to HOD parameters, the two correlations more noticeably differ, but still exhibit good agreement.

3.5.2 HOD Gradients in diffHOD-IA

We first study gradients of the 1-pt statistics of the galaxy catalog, $\langle N_{\rm cen}\rangle$ and $\langle N_{\rm sat}\rangle$, for various halo masses evaluated at the fiducial HOD values. We compare the autodifferentiation (autodiff) values from DIFFHOD-IA with finite-difference methods. The IA parameters, $\mu_{\rm cen}$ and $\mu_{\rm sat}$, are excluded in this analysis as they do not affect the galaxy number counts. The results of this analysis

are shown in Figure 3.9.

We see in the left panel of Figure 3.9 that the only HOD parameters with non-zero gradients for $\langle N_{\rm cen} \rangle$ are $\log M_{\rm min}$ and $\sigma_{\log M}$, which is in agreement with the analytic expression for $\langle N_{\rm cen} \rangle$ given in Equation 3.2. The finite difference gradients are additionally plotted as scatter, exhibiting excellent agreement with the autodiff gradients. The gradients vanish for $\log_{10} M/M_\odot \gtrsim 13$, at which point all available halos in the catalog are populated with a central galaxy. Similarly, at $\log_{10} M/M_\odot \lesssim 11$, the halos are not massive enough to frequently host central galaxies, so the gradients become small. The gradients for both $\sigma_{\log M}$ and $\log M_{\rm min}$ are approximately unity (up to a sign) at $\log_{10} M/M_\odot \approx 12$, where the occupation probability for the fiducial HOD transitions from 0 to 1.

In the right panel of Figure 3.9, we see that all five HOD parameters have nonzero gradients for $\langle N_{\rm sat} \rangle$, as expected from the analytic expression given in Equation 3.3. The magnitudes of the gradients vary across several orders of magnitude as halo mass increases, which is an important diagnostic for specifying learning rates when using the gradients in a gradient-based optimization pipeline. We again see excellent agreement between the DIFFHOD-IA autodiff gradients and the finite differences. These results are in excellent agreement with a similar analysis shown in [110].

3.5.3 Intrinsic Alignment Gradients in diffHOD-IA

We next examine the differentiability of the IA model in DIFFHOD-IA by validating the accuracy and gradients of the Dimroth–Watson distribution sampling procedure. The Dimroth–Watson distribution is parameterized by the alignment strength $\mu \in [-1, 1]$, which controls the shape of the Dimroth–Watson distribution from which galaxy misalignment angles are sampled. We compare the differentiable Dimroth–Watson samples with typical samples for a range of misalignment angles in Figure 3.10. The DIFFHOD-IA data was generated using 500000 samples and 6 Newton iterations for the inverse CDF sampling procedure. For each sample, we compute the misalignment angle θ between the sampled galaxy orientation and a reference alignment direction.

The left panel of Figure 3.10 shows the empirical probability distributions $P(\cos\theta)$ constructed from the samples (solid lines) alongside the analytic Dimroth–Watson PDF (dashed lines). The histograms are computed with 100 bins spanning $\cos\theta \in [-1, 1]$ and normalized to unit area. The close agreement across all values of μ validates our implementation of the inverse-CDF sampling procedure for sampling from the Dimroth–Watson. As expected, positive μ produces alignment with probability mass concentrated near $\cos\theta = \pm 1$ (corresponding to $\theta \approx 0^\circ$ or 180°), while

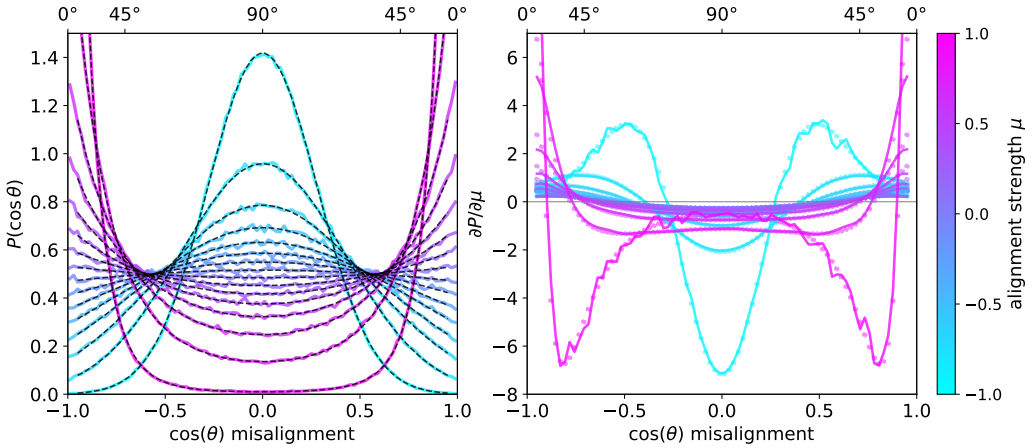


Figure 3.10: Validation of differentiable sampling from the Dimroth-Watson distribution for galaxy-halo misalignment angles. **Left panel:** Probability distribution $P(\cos\theta)$ of misalignment angles for varying alignment strength μ. The top axes show the corresponding misalignment angle θ in degrees. Solid lines show histograms from samples drawn using our differentiable inverse-CDF sampler; dashed lines show the analytic Dimroth-Watson PDF. The close agreement validates the differentiable sampling implementation. Positive μ (magenta) produces alignment with probability concentrated at $\cos\theta = \pm 1$, while negative μ (cyan) produces anti-alignment peaked at $\cos\theta = 0$. **Right panel:** Gradient of the probability distribution with respect to the alignment parameter, $\partial P/\partial\mu$. Dashed lines show analytic gradients derived from the PDF formula; scatter points show finite-difference gradients of the analytic Dimroth-Watson PDF; solid lines show gradients computed via automatic differentiation through the sampling procedure, where a Gaussian kernel density estimate is used to obtain a smooth density from discrete samples. A discrepancy is seen for the autodiff computed gradients for $\mu \approx 0$.

negative μ produces anti-alignment with probability peaked at $\cos\theta = 0$ ($\theta \approx 90°$).

To validate the gradients, we compute $\partial P(\cos\theta)/\partial\mu$ in three ways. First, we use autodiff through the DIFFHOD-IA sampling procedure. We draw samples for each μ value, construct a differentiable Gaussian kernel density estimate (KDE) to obtain a smooth density function from the discrete samples, and compute gradients via autodiff. The KDE is necessary to convert the discrete histogram into a continuous, differentiable function. We additionally compute analytic and finite-difference gradients of the exact Dimroth–Watson PDF with respect to μ, which can be derived from Equation 3.5 for the analytic case. The right panel of Figure 3.10 shows generally good agreement between the autodiff gradients (solid lines), the analytic gradients (dashed lines), and finite-difference gradients (scatter). We note that in the case of $\mu \approx 0$, the gradient similarly tends to zero as the Dimroth–Watson distribution becomes uniform. The gradient can thus be sensitive to Monte Carlo noise in this regime.

3.6 Applications

We demonstrate the utility of both IAEMU and DIFFHOD-IA with gradient-based optimization experiments. These experiments illustrate how gradients in the simulation can be used to obtain parameter estimates and posteriors over parameters of interest from mock observational data. All DIFFHOD-IA experimental results presented use a value of $N_{\max} = 48$, $\tau = 0.1$, and the subhalo phase space model with a NFW fallback. A constant radial alignment strength model is used for the IA. Our emphasis will be on the IA parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$, as several experiments illustrating the differentiability of the HOD are included in [110]. We construct moment-matching and differentiable correlation function optimization objectives. In addition, we demonstrate accelerated inference with both DIFFHOD-IA and IAEMU using HMC, compared to HALOTOOLS-IA with MCMC.

3.6.1 Moment-matching Objective

Setup. We consider the task of inferring the IA parameters $\boldsymbol{\mu} = (\mu_{\rm cen}, \mu_{\rm sat})$ from a target galaxy catalog observable, restricting for simplicity to a single HOD configuration given by the fiducial TNG300 model used previously. Importantly, fixing the HOD does not fix the realized galaxy catalog: different random seeds produce catalogs with varying numbers of galaxies, requiring an optimization procedure that is robust to stochastic catalog realizations. To this end, we consider a

moment matching optimization procedure, matching the misalignment angle distributions between the generated and target galaxy catalogs via their first two moments. This does not rely on a one-to-one correspondence between galaxies, and additionally requires no knowledge of the underlying PDF governing the galaxy misalignments.

Optimization. For a given parameter vector $\boldsymbol{\mu}$ and simulation seed s, we generate the mock galaxy catalog and compute the per-galaxy alignment statistic $t^2 = (\mathbf{n} \cdot \mathbf{u})^2$, where $\mathbf{n}$ is the galaxy orientation and $\mathbf{u}$ is the reference axis (host halo major axis for centrals, radial direction for satellites). We use t^2 rather than t because the Dimroth–Watson distribution is symmetric about $t = 0$, making $\langle t \rangle = 0$ for all μ. For each population (centrals and satellites separately), we compute the mean $\langle t^2 \rangle$ and variance $\mathrm{Var}(t^2)$ across all galaxies, which serve as our summary statistics.

Our loss function matches these first and second moments between simulated and observed catalogs across seeds s, treating central and satellite galaxy populations separately

$$\mathcal{L}(\boldsymbol{\mu}; s) = \alpha_{\text{cen}} \mathcal{L}_{\text{cen}} + \mathcal{L}_{\text{sat}}, \tag{3.47}$$

where

$$\begin{aligned} \mathcal{L}_{\text{cen}} &= \left(\langle t^2 \rangle_{\text{sim}} - \langle t^2 \rangle_{\text{obs}}\right)^2 \\ &\quad + w_{\text{cen},\sigma^2} \left(\mathrm{Var}[t^2]_{\text{sim}} - \mathrm{Var}[t^2]_{\text{obs}}\right)^2, \end{aligned} \tag{3.48}$$

$$\begin{aligned} \mathcal{L}_{\text{sat}} &= \left(\langle t^2 \rangle_{\text{sim}} - \langle t^2 \rangle_{\text{obs}}\right)^2 \\ &\quad + w_{\text{sat},\sigma^2} \left(\mathrm{Var}[t^2]_{\text{sim}} - \mathrm{Var}[t^2]_{\text{obs}}\right)^2, \end{aligned} \tag{3.49}$$

where it is implied that the t^2 statistics are computed over central galaxies only in $\mathcal{L}_{\text{cen}}$ and satellites only in $\mathcal{L}_{\text{sat}}$. We use weights $w_{\text{cen},\sigma^2} = w_{\text{sat},\sigma^2} = 0.5$, and $\alpha_{\text{cen}} = 2.0$ to up-weight the smaller central galaxy population. To mitigate the Monte Carlo noise in Dimroth–Watson sampling, we average the loss over $N_s = 3$ random seeds:

$$\bar{\mathcal{L}}(\boldsymbol{\mu}) = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}(\boldsymbol{\mu}; s_i). \tag{3.50}$$

This construction matches the variance $\mathrm{Var}[t^2]$ within each catalog to capture galaxy shape noise, while averaging the loss over multiple simulated realizations stabilizes the gradient signal. The seed averaging is performed over a fixed set of seeds; importantly, the target catalog uses a different seed than the optimization seeds. We compare experiments using one and three seeds in Figure 3.11.

We note that for the target catalog generated at fixed input parameters, the observed alignment angles represent a single Monte Carlo draw from the underlying Dimroth–Watson distribution,

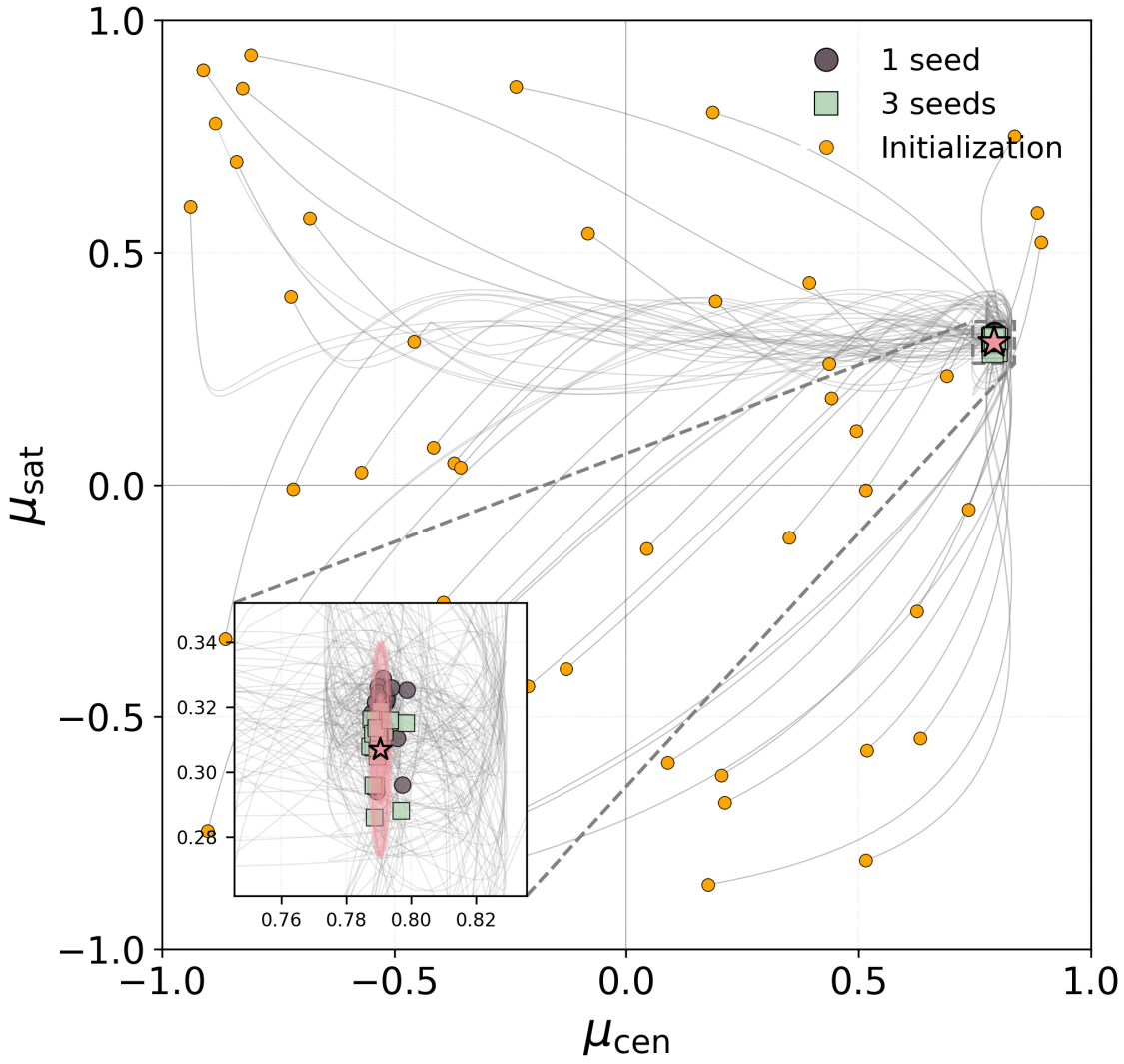


Figure 3.11: Gradient-based recovery of IA parameters from 50 random initializations using moment-matching optimization. The target parameters ($\mu_{\rm cen} = 0.7905$, $\mu_{\rm sat} = 0.307$, pink star) represent the best-fit μ values to the TNG300 HOD configuration, with the empirical uncertainty shown as pink contours. Optimization trajectories are shown as gray lines connecting initial positions to converged solutions. Brown circles denote optimizations using a single HOD realization seed per gradient step, while green squares show results when averaging over three seeds. The inset panel (lower left) shows a zoomed view of the convergence region, revealing tight clustering of final parameter estimates around the true values. Both single-seed and three-seed strategies successfully recover the target parameters across diverse initializations.

inducing an effective $\boldsymbol{\mu}$ that may differ slightly from the nominal input values. To determine the effective $\boldsymbol{\mu}$ values for the target catalog, we independently match the observed $\langle t^2 \rangle$ to the theoretical Dimroth–Watson expectation for the central and satellite populations. This yields best-fit values of $\mu_{\rm cen} = 0.7905$ and $\mu_{\rm sat} = 0.307$, which differ slightly from the input values $\mu_{\rm cen} = 0.79$ and $\mu_{\rm sat} = 0.30$.

We estimate the uncertainties in $\mu_{\rm cen}$ and $\mu_{\rm sat}$ by quantifying the variance in effective μ values across independent catalog realizations. To this end, we generate 50 independent HOD realizations at the fiducial IA parameters using different random seeds. For each realization, we compute $\langle t^2 \rangle$ separately for centrals and satellites, as done in the optimization, and determine the effective μ by matching to the theoretical Dimroth–Watson expectation via numerical inversion. The scatter in these effective μ values across realizations quantifies the uncertainty in the effective μ for a single catalog realization, capturing the combined variance from HOD sampling and orientation sampling. Importantly, this also defines the expected region within which our optimization should converge. This yields uncertainties $\sigma(\mu_{\rm cen}) = 0.001$ and $\sigma(\mu_{\rm sat}) = 0.012$, with a correlation coefficient of $\rho = 0.04$. These uncertainties are visualized in pink in Figure 3.11.

We note that the roughly one order of magnitude larger uncertainty for $\mu_{\rm sat}$ compared to $\mu_{\rm cen}$ is primarily due to the smaller alignment strength of $\mu_{\rm sat} = 0.30$ relative to $\mu_{\rm cen} = 0.79$ At lower μ, the Dimroth–Watson distribution is less sharply peaked and galaxy orientation samples have a higher entropy, producing greater variance in μ inferred from a finite catalog. This larger uncertainty is reflected in the scatter of converged $\mu_{\rm sat}$ values seen in Figure 3.11.

Results. Figure 3.11 shows the results of gradient-based optimization of IA parameters using the moment matching objective. We perform 50 independent optimization runs from random initializations uniformly sampled from $\mu \in [-1.0, 1.0]$, each run for 100 optimization steps. The trajectories in Figure 3.11 show the parameter space flow from initializations to the fiducial TNG300 IA values, with final points filtered by best loss. We see that the moment-matching loss over the full galaxy field successfully optimizes the IA parameters toward the ground truth values for the HOD across all initializations and both seed-averaging strategies. In the inset panel, we observe that the 3-seed optimization exhibits slightly tighter convergence to the true values. Quantitatively, the 1-seed optimization results in final values $\mu_{\rm cen} = 0.791 \pm 0.002$ and $\mu_{\rm sat} = 0.320 \pm 0.007$, while the 3-seed optimization achieves $\mu_{\rm cen} = 0.790 \pm 0.002$ and $\mu_{\rm sat} = 0.309 \pm 0.006$.

The scatter in converged $\mu_{\rm cen}$ values is slightly larger than the computed target uncertainty, likely reflecting optimization noise and finite optimization time. Conversely, the scatter in converged $\mu_{\rm sat}$ values is comparable to the target uncertainty, and the slight bias toward higher values may

arise from an asymmetry of the loss landscape near the true value. Nevertheless, all optimization runs converge to within a few percent of the true values, demonstrating robust parameter recovery. Although the IA parameters in this simplified setting could be recovered via a direct curve fit, this experiment illustrates how differentiability enables efficient parameter recovery using only one-point alignment statistics. This is especially relevant when the analytic form of the PDF is not known; in this case, we have captured properties of the true distribution as characterized by its first two moments.

3.6.2 Correlation Function Objective

Setup. We now turn to optimization with a more traditional summary statistic with 2PCFs. In this section, we will optimize the IA parameters such that the generated catalog $\omega(r)$ correlation matches that of the fiducial TNG300 correlation function, denoted $\hat{\omega}(r)$. $\hat{\omega}(r)$ is computed by averaging over 20 independent orientation samples at the fiducial IA parameters. The per-bin variance of $\hat{\omega}(r)$ is estimated from the scatter across these realizations, which reduces galaxy shape noise in the optimization target. Unlike a single realization of galaxy orientations, which can result in effective $\boldsymbol{\mu}$ values that differ from the input, this averaging ensures that the optimal $\boldsymbol{\mu}$ parameters for $\hat{\omega}(r)$ converge to the true input values. We choose to optimize on this statistic over $\eta(r)$, as $\omega(r)$ is less contaminated with galaxy shape noise due to it being a cross correlation with galaxy positions. We cannot use $\xi(r)$ as it has no dependence on $\boldsymbol{\mu}$; a full analysis that includes the HOD parameters would jointly use $\xi(r)$, $\omega(r)$, and $\eta(r)$.

Optimization. We construct a weighted mean-squared-error loss between the predicted and target correlation functions:

$$\mathcal{L}(\boldsymbol{\mu}) = \sum_k W_k \left[\omega(r_k) - \hat{\omega}(r_k)\right]^2, \tag{3.51}$$

where $\hat{\omega}(r_k)$ is the averaged target correlation computed from the fiducial TNG300 catalog and W_k are inverse-variance weights, determined by estimating the variance of $\hat{\omega}(r_k)$ across multiple orientation realizations at fixed positions. We generate 20 independent orientation samples at reference IA parameter values and compute:

$$W_k = \frac{1}{\sigma_\omega^2(r_k) + \epsilon}, \tag{3.52}$$

where $\sigma_\omega^2(r_k)$ is the empirical variance in bin k and $\epsilon = 10^{-6}$ provides numerical stability to prevent division by zero in bins with very low variance. The weights are normalized such that $\sum_k W_k = 1$. This inverse-variance weighting approximates the covariance due to galaxy shape noise, which is in

general dominant over the sample variance at the scales being considered [111]. While the jackknife covariances that account for both sample variance and galaxy shape noise from the TNG300 data itself are available following the procedure of [111], we instead use a diagonal variance estimate obtained by averaging over many DIFFHOD-IA orientation realizations (for fixed galaxy positions), which serves as a proxy for the shape noise and is considerably more computationally efficient.

Results. Figure 3.12 shows the results of gradient-based optimization of IA parameters using the differentiable $\omega(r)$ objective. We perform 50 independent optimization runs from random initializations uniformly sampled from $\mu \in [-1.0, 1.0]$, each run for 2000 optimization steps. The optimization successfully recovers the target parameters across all initializations, with mean recovered values $\mu_{\rm cen} = 0.791 \pm 0.003$ and $\mu_{\rm sat} = 0.303 \pm 0.009$, compared to target values of $\mu_{\rm cen} = 0.79$ and $\mu_{\rm sat} = 0.30$. All converged solutions lie well within the expected statistical uncertainty, shown in pink in Figure 3.12.

To estimate the uncertainty, we generate 50 independent HOD catalog realizations at the true IA parameters with different random seeds. For each realization, we compute $\omega(r)$ averaged over 20 orientation samples and find the maximum likelihood IA parameters via grid search, minimizing

$$\chi^2 = (\omega - \hat{\omega})^T \mathbf{C}^{-1} (\omega - \hat{\omega}), \tag{3.53}$$

where $\mathbf{C}$ is the covariance matrix of $\omega(r)$ estimated from 50 independent orientation samples at the fiducial parameters. We additionally correct for the finite number of realizations by incorporating the Hartlap factor [221] when inverting the covariance matrix. The scatter in recovered parameters across these 50 catalogs provides an empirical estimate of the covariance, yielding uncertainties $\sigma(\mu_{\rm cen}) = 0.008$ and $\sigma(\mu_{\rm sat}) = 0.026$, with correlation coefficient $\rho = -0.68$. This anti-correlation aligns with theoretical expectations and was also seen experimentally in [108]. The tighter clustering of converged values reflects the fact that all optimization runs fit the same catalog to the same target, whereas the uncertainty contours capture the full variance including HOD stochasticity.

3.6.3 Hamiltonian Monte Carlo

One of the key advantages of both IAEMU and DIFFHOD-IA is their differentiability, which enables the use of gradient-based sampling algorithms. We now demonstrate Hamiltonian Monte Carlo (HMC) inference that achieves substantial speedups over traditional Markov Chain Monte Carlo (MCMC) approaches.

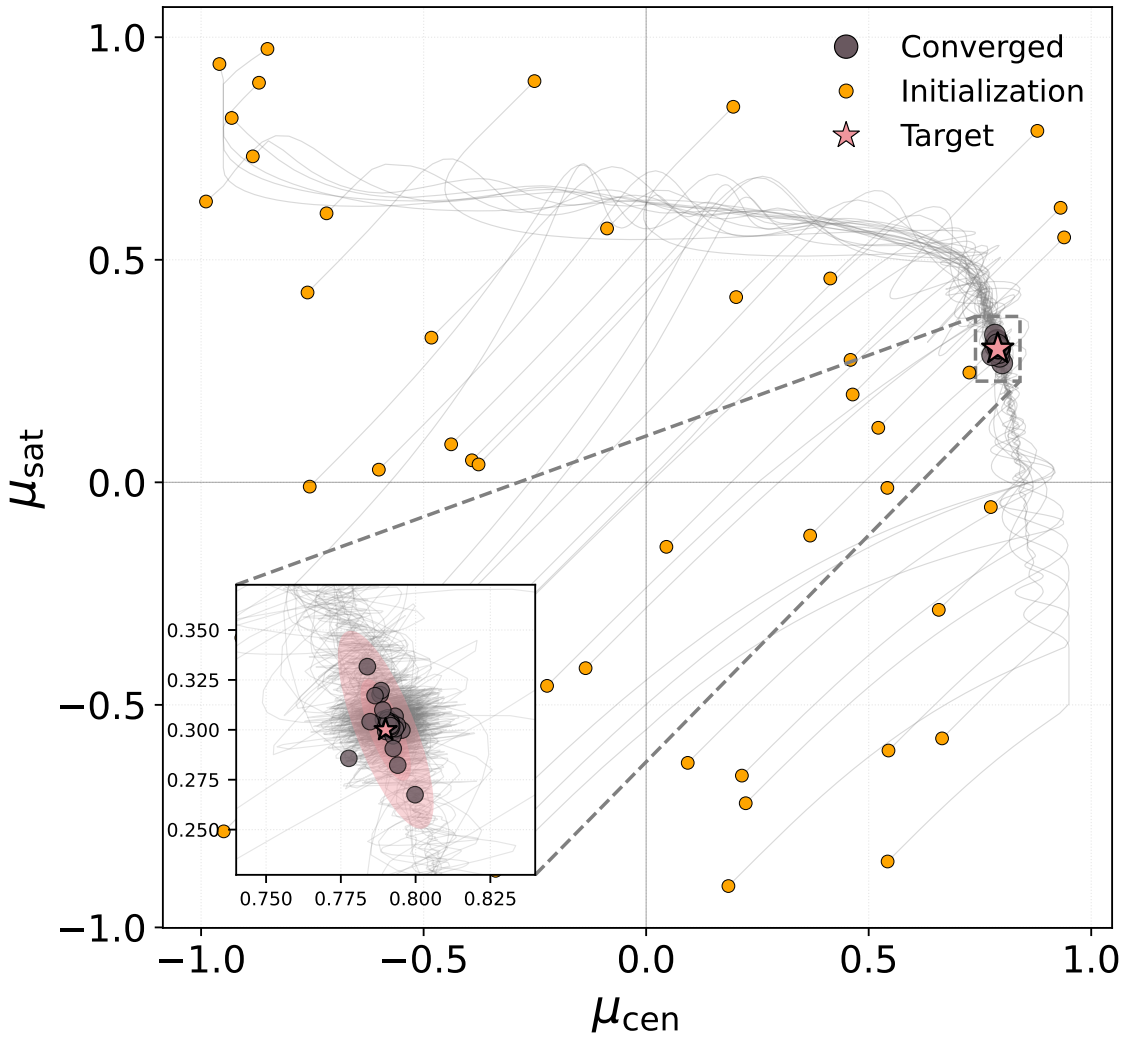


Figure 3.12: Gradient-based recovery of IA parameters from 50 random initializations using correlation function matching optimization. The target parameters ($\mu_{\rm cen} = 0.79$, $\mu_{\rm sat} = 0.30$, pink star) represent the fiducial TNG300 HOD configuration. Optimization trajectories are shown as gray lines connecting initial positions to converged solutions. Gray circles denote converged values. The inset panel (lower left) shows a zoomed view of the convergence region, revealing tight clustering of final parameter estimates around the true values.

Background on HMC

HMC is a variant of the Metropolis-Hastings algorithm, where Hamiltonian dynamics are simulated using a time-reversible, volume-preserving numerical integrator to propose transitions to new points in the state space. We use HMC to sample from a posterior distribution over the inputs x, given trained NN parameters θ and observations $\mathcal{D}$. This is described by

$$p(x|\mathcal{D},\theta) \propto p(\mathcal{D}|x,\theta)p(x) \ , \tag{3.54}$$

Equation 3.54 is a form of Bayes' Theorem, where $p(\mathcal{D}|x,\theta)$ is the likelihood function and $p(x)$ is the prior distribution on x. HMC achieves this by forward modeling the dynamics of a governing Hamiltonian H:

$$H = T + U = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} - \ln p(x|\mathcal{D},\theta) \tag{3.55}$$

where T is the kinetic energy with mass matrix M and momentum $\mathbf{p}$, which controls the exploration in parameter space, and $-\ln p(x|\mathcal{D},\theta)$ takes the role of the potential energy U. The time-evolution of x and p is accordingly governed by Hamilton's equations:

$$\frac{dx}{dt} = \frac{\partial H}{\partial \mathbf{p}}, \qquad \frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial x}. \tag{3.56}$$

HMC thus arrives at the posterior distribution over the inputs by sequentially evolving the dynamical variables according to Hamiltonian dynamics; this of course corresponds to minimizing the potential energy, which maximizes the log probability. As seen in equation 3.56, Hamilton's equations require gradients with respect to H, specifically $-\nabla_x \ln p(x|\mathcal{D},\theta)$. Decomposing this with chain rule,

$$\begin{aligned}
\nabla_x \ln p(x|\mathcal{D},\theta) &= \nabla_x \ln p(\mathcal{D}|x,\theta) + \nabla_x \ln p(x) \\
&\propto \nabla_x \ln p(f_\theta(x)|\mathcal{D}) \\
&= \nabla_{f_\theta(x)} \ln p(f_\theta(x)|\mathcal{D}) \cdot \nabla_x f_\theta(x) \ ,
\end{aligned}$$

where in the second line we recognize that the likelihood is implicitly a function of the outputs of the forward model, $f_\theta(x)$, explicitly denoting its dependence on parameters θ. Due to the differentiability of both IAEMU and DIFFHOD-IA, both are compatible with HMC. We thus see how differentiability through the forward model is leveraged in this algorithm.

Comparison of Methods

Setup. To complement our previous correlation function matching gradient descent experiment, we now turn to HMC to illustrate Monte Carlo inference capabilities with DIFFHOD-IA. We use the

Table 3.2: Comparison of $\mu_{\rm cen}$ and $\mu_{\rm sat}$ posteriors from the Monte Carlo experiments. DIFFHOD-IA and IAEMU posteriors used HMC, while HALOTOOLS-IA posteriors used MCMC.

Method	$\mu_{\rm cen}$	$\mu_{\rm sat}$
HALOTOOLS-IA	0.793 ± 0.017	0.294 ± 0.056
IAEMU	0.799 ± 0.022	0.317 ± 0.049
DIFFHOD-IA	0.802 ± 0.016	0.318 ± 0.048

same $\hat{\omega}(r)$ mock observation from the previous section, this time using the jackknife covariance estimate from TNG300. For this analysis, we use a fixed HOD and only consider $\mu_{\rm sat}$ and $\mu_{\rm cen}$ as free parameters.

We use the uninformative priors:

$$\mu_{\rm cen}, \mu_{\rm sat} \sim \text{Uniform}(-1, 1), \tag{3.57}$$

and employ the `numpyro` [128] implementation of a No-U-Turn Sampler (NUTS) [222]). We use four chains with 1500 steps (500 burn-in). This exact analysis was conducted in [8], allowing a direct comparison for DIFFHOD-IA with the ground truth HALOTOOLS-IA, as well as HALOTOOLS-IA emulator IAEMU [8]. HMC on DIFFHOD-IA took roughly 5 minutes to converge on a single NVIDIA A100 GPU, while HMC with IAEMU converged in approximately one minute on the same GPU. As HALOTOOLS-IA is not differentiable, a comparable MCMC analysis required up to a full day across 150 CPU cores.

Results. Posteriors for DIFFHOD-IA, HALOTOOLS-IA, and IAEMU are shown in Figure 3.13. There is in general excellent overlap between the three posteriors, illustrating that DIFFHOD-IA offers similar inference capabilities as HALOTOOLS-IA with much faster convergence with HMC. A slight bias is seen in the case of $\mu_{\rm sat}$ for both DIFFHOD-IA and IAEMU compared to HALOTOOLS-IA, which can potentially be attributed to different seeds in the target galaxy catalog. Nonetheless, DIFFHOD-IA is within 0.4σ agreement with HALOTOOLS-IA. Exact posterior values are given in Table 3.2.

3.6.4 Out-of-Distribution Generalization with IAEMU

Previously, [111] showed that HALOTOOLS-IA is expressive enough to model the IA signal derived from The TNG300 suite of hydrodynamical simulations, which incorporate more complex physics, including baryonic effects. This constitutes an OOD shift over the joint distribution of inputs and

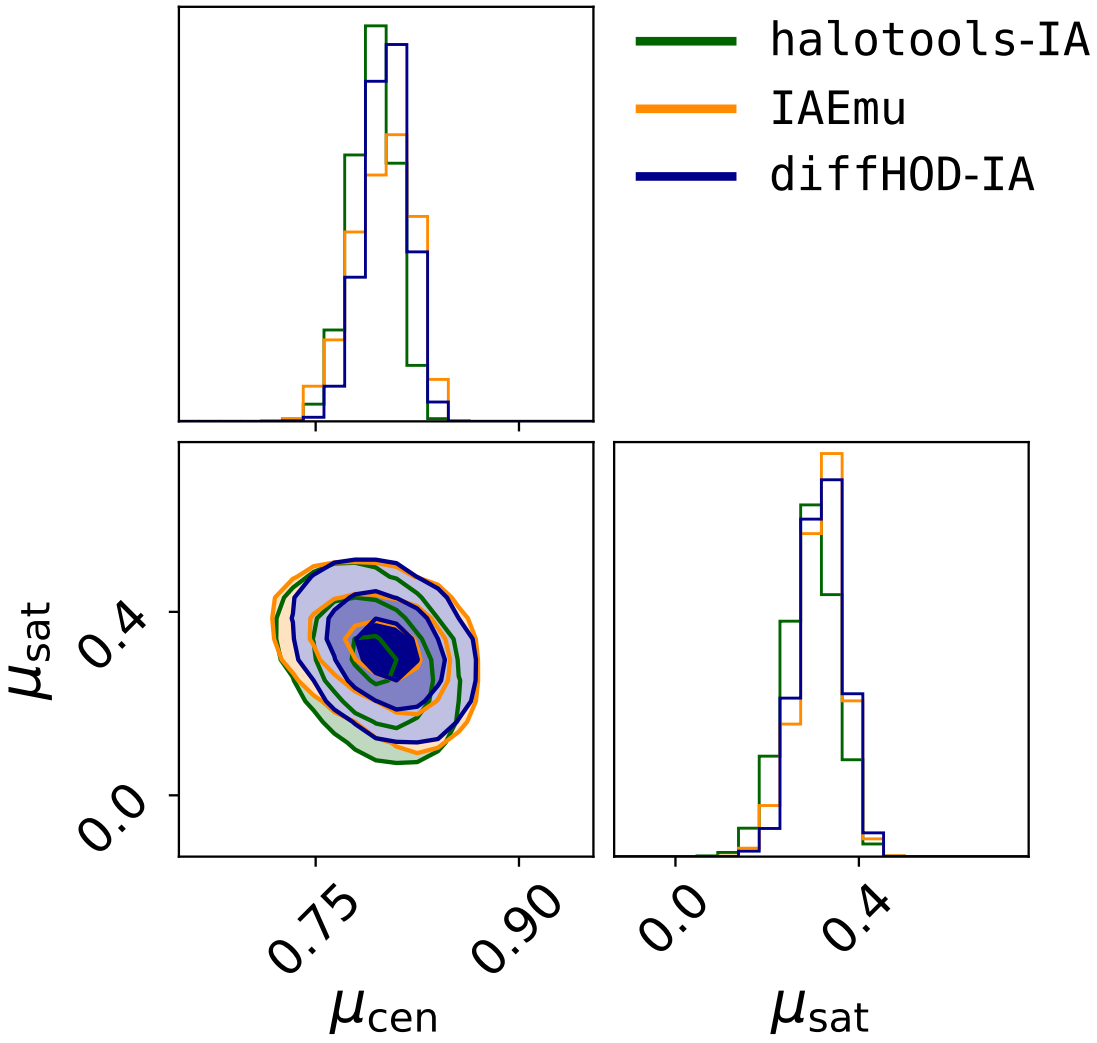


Figure 3.13: μ_{cen} and μ_{sat} posteriors for the fiducial TNG300 catalog derived using DIFFHOD-IA with HMC (blue), IAEMU with HMC (orange), and HALOTOOLS-IA with MCMC (green). DIFFHOD-IA is in excellent agreement with HALOTOOLS-IA, while exhibiting substantially faster inference convergence with HMC.

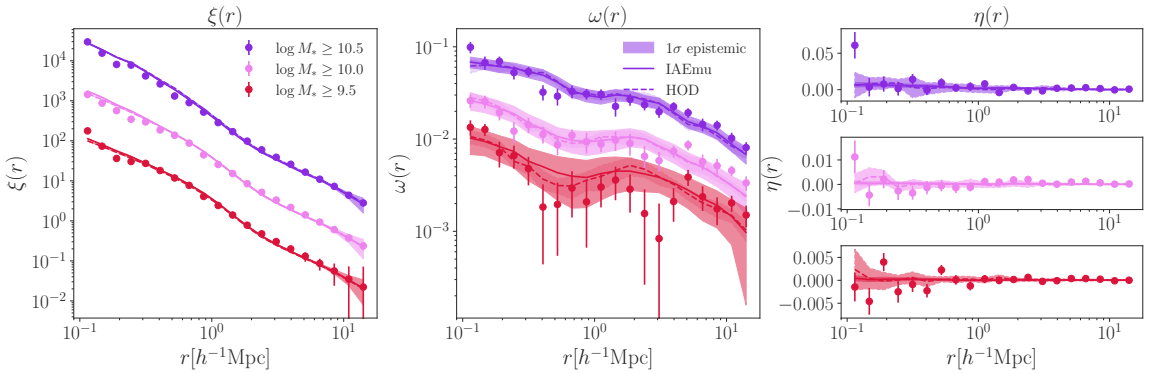
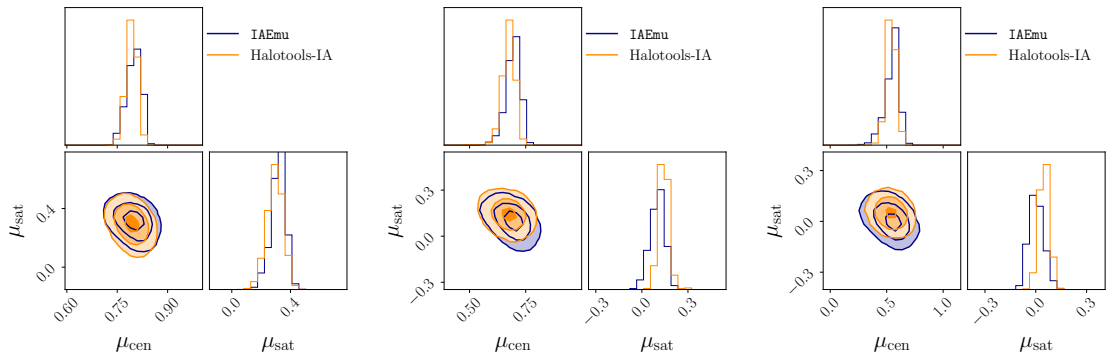


Figure 3.14: The 2PCFs for IA, fitted to observations from the TNG300 simulation, using both HALOTOOLS-IA and IAEMU. The correlations are measured across three mass threshold samples, as denoted in the left panel legend. Purple corresponds to most massive sample, pink for intermediate, and red for least massive. True correlations are shown as scatter points and HOD and IAEMU fits shown as lines. These 2PCFs correspond to the posterior mean values of $\mu_{\rm cen}$ and $\mu_{\rm sat}$, as shown in Figure 3.15. Error bars for TNG300 are obtained via jackknife resampling, while the 1σ epistemic uncertainty for IAEMU is estimated from 50 forward passes using the Monte Carlo dropout technique. The 1σ uncertainty band for HALOTOOLS-IA reflects variations from random realizations of the model. **Left:** Position-position correlation function $\xi(r)$ with the upper and lower curves offset by 1 dex for visual clarity, showing that IAEMU can model galaxy bias. **Middle:** Position-orientation correlation function $\omega(r)$. **Right:** Orientation-orientation correlation function $\eta(r)$.

outputs from an HOD that IAEMU was trained on. In this section, we investigate whether IAEMU exhibits a similar modeling capability as HALOTOOLS-IA, and can thus be robust to OOD shifts for inverse modeling. To this end, we select the best-fit occupation model parameters that reproduce the HOD of TNG300, as described in [111], and determine the posterior distributions on $\mu_{\rm cen}$ and $\mu_{\rm sat}$ that fit the signal. This ensures that halos with comparable masses are populated with a comparable number of galaxies as in TNG300, leaving galaxy alignment as the major factor affecting how similar correlations from the two samples are. This experiment therefore enables us to investigate potential biases between IAEMU and HALOTOOLS-IA in the alignment parameter input space when modeling IA for an OOD sample.

To perform parameter inference, we leverage the differentiability of IAEMU to attain efficient posterior estimates. We employ HMC with NUTS, using 2000 warm-up steps and an initial learning rate of 0.005, collecting 4000 posterior samples for analysis. All posteriors resulted in an effective sample size greater than 1000, and all HMC experiments were executed on a single GPU and converged in roughly one minute. For comparison, the MCMC implementation in [111] utilized parallelization across 150 CPU cores and required up to a full day due to computational constraints. This highlights a near $2000\times$ speed-up for IAEMU-HMC relative to HALOTOOLS-IA-MCMC on the tested hardware. While this is somewhat lower than the acceleration achieved in forward modeling with IAEMU compared to HALOTOOLS-IA, it remains a substantial improvement. The reduced gain is anticipated: HMC necessitates backpropagation through IAEMU, roughly doubling the computational load compared to a forward pass [223], along with added overhead from NUTS numerical integration. Additionally, the sequential nature of HMC does not allow for parallelization. Nevertheless, its rapid convergence demonstrates its efficiency over traditional, parallelized MCMC approaches. We emphasize that a detailed convergence comparison was not performed, and that the parallelized MCMC yielded approximately $75,000$ posterior samples, in contrast to the 4000 obtained via IAEMU. Hence, the reported speed-up metrics should be interpreted as indicative benchmarks rather than definitive measurements.

The correlation function predictions from IAEMU with posterior means for $\mu_{\rm cen}$ and $\mu_{\rm sat}$ are shown in Figure 3.14, in which we see that there is generally good agreement with IAEMU predictions compared to those from HALOTOOLS-IA for all correlations. The correlation function, $\xi(r)$, is also shown to illustrate the agreement between HALOTOOLS-IA and IAEMU for galaxy clustering statistics; however, it does not depend on $\mu_{\rm cen}$ or $\mu_{\rm sat}$. We also see that the quality of fit for both HALOTOOLS-IA and IAEMU decreases with decreasing stellar mass, which was also observed in [111]. This provides some indication that a constant alignment strength parameterization is not



(a) $\log M_* \geq 10.5$ (b) $\log M_* \geq 10.0$ (c) $\log M_* \geq 9.5$

Figure 3.15: Optimal parameter values for central alignment strength ($\mu_{\rm cen}$) and satellite alignment strength ($\mu_{\rm sat}$) fit to $\omega(r)$ observations from TNG300 with three distinct mass cutoffs for halos included in the underlying HOD model. Posterior contours for HALOTOOLS-IA and IAEMU are shown with 4000 posterior samples each, with HALOTOOLS-IA contours in orange and IAEMU contours in blue. Posteriors for HALOTOOLS-IA were obtained via MCMC using 75 walkers running in parallel for 23 hours on CPU, resulting in up to 1300 steps per walker, or as few as about 450 steps per walker for slower runs. Posteriors for IAEMU were retrieved using NUTS, a variant of the HMC algorithm, with 2000 warm-up steps around a minute on a single GPU. IAEMU posteriors exhibit a better than 0.4σ overlap with posteriors from HALOTOOLS-IA, indicating that IAEMU can generalize to OOD shifts for inverse modeling. Exact posterior summaries for comparison can be found in Table 3.3. **Left:** Sample 1 IAEMU posteriors with optimal values $\mu_{\rm cen} = 0.81$ and $\mu_{\rm sat} = 0.35$. **Middle:** Sample 2 IAEMU posteriors with optimal values $\mu_{\rm cen} = 0.70$ and $\mu_{\rm sat} = 0.14$. **Right:** Sample 3 IAEMU posteriors with optimal values $\mu_{\rm cen} = 0.52$ and $\mu_{\rm sat} = 0.01$.

sufficient to extend this HOD-based model to hydrodynamic simulations in the low-mass regime, which could potentially be addressed with a mass-dependent alignment parameterization.

The corner plots in Figure 3.15 show the joint $(\mu_{\rm cen}, \mu_{\rm sat})$ posteriors for three separate stellar mass thresholds M_* for both IAEMU and HALOTOOLS-IA. Sample 1 corresponds to $\log(M_*) > 10.5$, Sample 2 to $\log(M_*) > 10.0$, and Sample 3 to $\log(M_*) > 9.5$. The HOD parameter fits corresponding to these mass cutoffs can be found in [111]. We confirm two trends also observed in [111]: central alignment strength is larger than satellite alignment strength, and the alignment strength monotonically increases with the stellar mass threshold. We find a greater than 0.4σ agreement between MCMC with HALOTOOLS-IA and HMC with IAEMU for all samples. The strongest discrepancy is in the posterior variance for Sample 3, which is the noisiest set

Table 3.3: Posterior values for IAEMU and HALOTOOLS-IA fit on TNG300.

Sample	**Mass Cutoff**	**Posterior**	μ_{cen}	μ_{sat}
1	$\log M_* > 10.5$	IAEMU	$0.80^{+0.02}_{-0.03}$	$0.32^{+0.04}_{-0.05}$
		HALOTOOLS-IA	$0.79^{+0.02}_{-0.02}$	$0.30^{+0.05}_{-0.06}$
2	$\log M_* > 10.0$	IAEMU	$0.69^{+0.03}_{-0.03}$	$0.11^{+0.04}_{-0.05}$
		HALOTOOLS-IA	$0.68^{+0.03}_{-0.03}$	$0.14^{+0.03}_{-0.03}$
3	$\log M_* > 9.5$	IAEMU	$0.56^{+0.04}_{-0.07}$	$0.00^{+0.05}_{-0.04}$
		HALOTOOLS-IA	$0.54^{+0.04}_{-0.04}$	$0.05^{+0.03}_{-0.03}$

of TNG300 data and also has the largest IAEMU epistemic uncertainty, as seen in Figure 3.14. This reflects the discussion in Section 3.3.6, wherein it was seen that the epistemic uncertainty of IAEMU is correlated with the true and predicted aleatoric uncertainty. Exact values for HALOTOOLS-IA and IAEMU posterior summary statistics are shown in Table 3.3.

3.6.5 External Usage of IAEMU

We proceed to summarize an external application of IAEMU that demonstrates the utility of its differentiability. This experiment was presented in [108] and is included here to illustrate how the emulator enables inference workflows that would otherwise be difficult to implement using traditional correlation function estimators. The goal of the experiment is to identify regions of parameter space that minimize the IA contamination. As it is understood what configurations of μ_{cen} and μ_{sat} generally minimize the IA correlations $\omega(r)$ and $\eta(r)$, this experiment tests for potential biases in IAEMU. In contrast to the inference task considered earlier, the objective here is to infer parameters that drive the IA signal toward zero. The proceeding analysis focuses on $\omega(r)$ as done previously.

The differentiability of IAEMU allows direct access to gradients of $\omega(r)$ and $\eta(r)$ with respect to the input parameters. This is quantified in Figure 3.16, which shows the Jacobian of the predicted IA correlations with respect to the seven model parameters. The gradients of $\omega(r)$ are both larger in magnitude and more structured across radial bins than those of $\eta(r)$, and can be evaluated efficiently. We additionally see stronger gradient signal with respect to the IA parameters

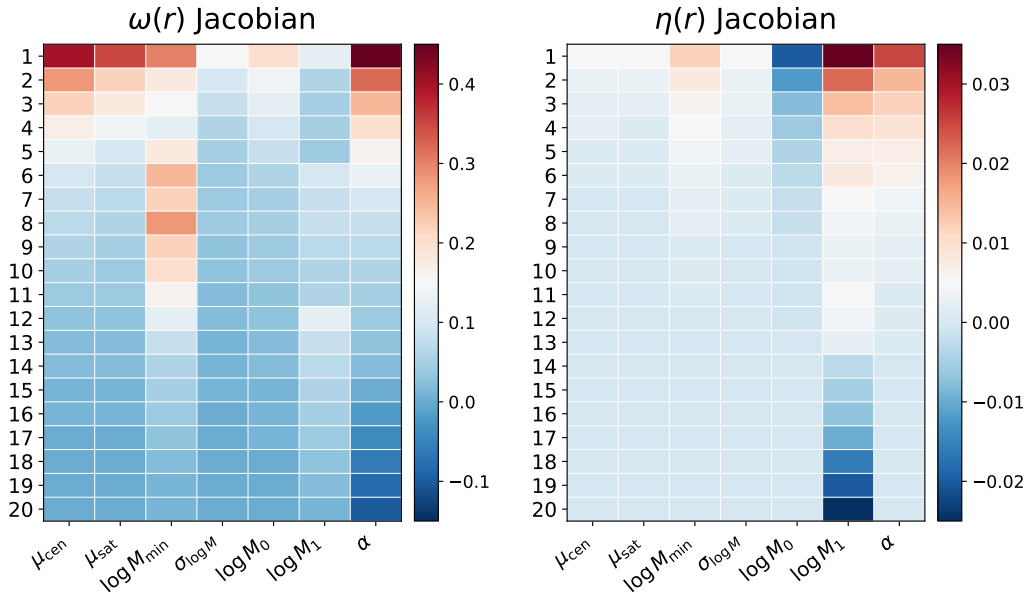
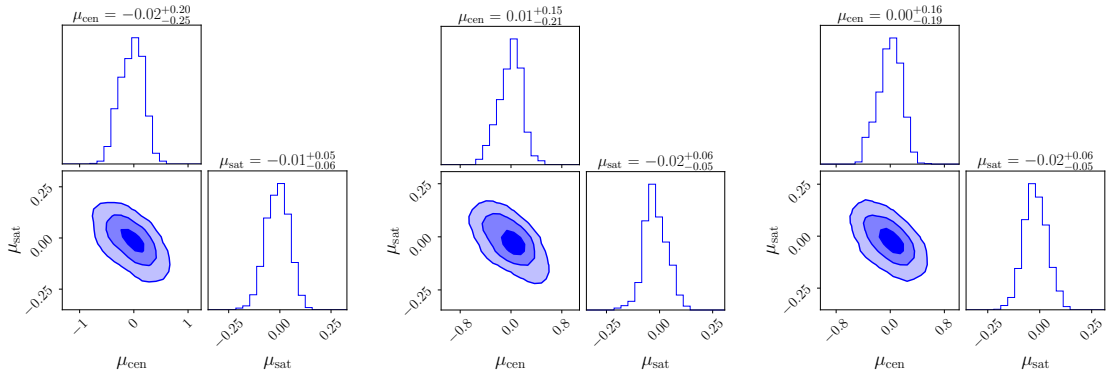


Figure 3.16: Jacobian matrices of the predicted IA correlation functions $\omega(r)$ (left) and $\eta(r)$ (right) with respect to the seven HOD and IA model parameters, evaluated at the fiducial parameter values. Each row corresponds to a radial bin and each column to an input parameter. The $\omega(r)$ Jacobian exhibits larger and more structured gradients, particularly with respect to $\mu_{\rm cen}$, $\mu_{\rm sat}$, and $\log M_{\rm min}$, while the $\eta(r)$ Jacobian is dominated by sensitivity to $\log M_1$ and α at small scales and is generally weaker in magnitude due to higher shape noise. The sign reversal in α for $\omega(r)$ at large radial bins reflects the transition from the one-halo to the two-halo regime.



(a) $\log M_* \geq 10.5$ (b) $\log M_* \geq 10.0$ (c) $\log M_* \geq 9.5$

Figure 3.17: Posterior distributions of the IA parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$ obtained by minimizing the $\omega(r)$ correlation using HMC with NUTS, for three stellar mass thresholds: $\log M_* \geq 10.5$ (left), $\log M_* \geq 10.0$ (center), and $\log M_* \geq 9.5$ (right). The HOD parameters are fixed to fiducial values from TNG300, while uniform priors of $[-1, 1]$ are imposed on both IA parameters. In all cases the posteriors peak near zero, consistent with the expectation that vanishing alignment strengths minimize the IA signal. An anti-correlated degeneracy between $\mu_{\rm cen}$ and $\mu_{\rm sat}$ is visible in the joint contours, reflecting the physical cancellation between opposing central and satellite alignments. The posteriors tighten with decreasing stellar mass threshold as the larger galaxy samples reduce shape noise.

alone, which motivates the use of $\omega(r)$ as a higher-fidelity summary statistic. $\eta(r)$ is more strongly affected by shape noise, so this aligns with expectations.

To isolate regions of parameter space that minimize $\omega(r)$, we again employ HMC with NUTS. The HOD parameters are fixed to fiducial values corresponding to the three stellar mass thresholds drawn from TNG300 and Table C1 of [111]. Uniform priors in the range $[-1, 1]$ are imposed on the IA parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$. The likelihood is taken to be a multivariate normal distribution centered at zero correlation, with a covariance given by the aleatoric uncertainty predicted by IAEMU. In this way, the inference favors regions of parameter space that yield minimal IA signal while accounting for shape noise. The resulting posterior distributions are shown in Figure 3.17 for the three different stellar mass cuts. In all cases, the posterior peaks near $\mu_{\rm cen} = \mu_{\rm sat} = 0$, indicating that vanishing alignment strengths minimize the $\omega(r)$ correlation, as expected. However, a clear degeneracy is present along directions where $\mu_{\rm cen}$ and $\mu_{\rm sat}$ take opposite signs. Physically, this corresponds to scenarios in which perpendicular satellite alignments partially cancel the parallel

alignments of central galaxies, resulting in a suppressed net alignment signal.

3.7 Summary & Discussion

In this chapter, we have presented two complementary approaches for efficient and differentiable modeling of galaxy intrinsic alignments within the halo occupation distribution framework: IAEMU, a neural network-based emulator, and DIFFHOD-IA, a fully differentiable HOD implementation with intrinsic alignments.

3.7.1 IAEMU Summary

We developed IAEMU, a neural network-based surrogate model designed to predict galaxy intrinsic alignment correlations derived from halo occupation distribution modeling. IAEMU eliminates the need to generate full galaxy catalogs and computes correlation functions using traditional HOD pipelines, which are computationally expensive. On a single GPU, IAEMU achieves a $\times\mathbf{10^4}$ speed-up in wall-clock time compared to HALOTOOLS-IA run on a moderately parallelized CPU setup representative of typical resources (e.g., $\sim$150 cores). When comparing single GPU to single CPU performance, this corresponds to an approximate $10^6\times$ speed-up. This substantial acceleration enables efficient forward modeling and significantly expedites inverse modeling tasks. The differentiable nature of IAEMU facilitates the use of gradient-based inference methods such as Hamiltonian Monte Carlo (HMC), which are otherwise infeasible with HALOTOOLS-IA. Although the speed-up in individual evaluations is dramatic, the end-to-end improvement in sampling-based inference compared to parallelized MCMC is somewhat lower, due to the additional computational overhead from gradient evaluations and the inherently sequential nature of HMC arising from numerical trajectory integration. Nevertheless, HMC achieves significantly faster convergence than parallelized MCMC, making it a far more efficient option overall for inverse modeling despite the reduced relative speed-up.

IAEMU was also designed to account for both aleatoric and epistemic uncertainties, corresponding to the uncertainty inherent in the data and the model, respectively. This enables confidence assessments for IAEMU predictions in the absence of ground truth data, as well as provides covariance information for inverse modeling with IAEMU. To isolate aleatoric uncertainty, we trained IAEMU using a mean-variance estimation framework under the assumption of Gaussian-distributed outputs, optimized with the β-negative-log-likelihood loss function. For epistemic uncertainty,

we employed the Monte Carlo dropout technique, which randomly nullifies certain nodes within IAEMU during inference, introducing stochasticity into the model predictions. We find that analyzing these distinct sources of uncertainty provides valuable insight into the strengths and weaknesses of IAEMU, offering a practical method for diagnosing the quality of emulator predictions and motivating future improvements. We further show the benefits of accelerated parameter inference (i.e., inverse problems) using gradient-based sampling techniques with IAEMU, exploiting the fact that NNs are differentiable models.

IAEMU achieves an average error of approximately 3% in emulating position-position and 5% in position-orientation galaxy IA correlations. Although the orientation-orientation correlation $\eta(r)$ is inherently noisier and thus more difficult to quantify performance for, IAEMU's predictions for $\eta(r)$ on average remained within 1σ of the true aleatoric uncertainty of the data when evaluated on the test set. This indicates that IAEMU still successfully captures the average behavior of this correlation without overfitting to the shape noise, which would otherwise require multiple realizations of HALOTOOLS-IA. IAEMU also generally exhibits strong SCC values with the data across all three correlations, indicating that despite the large fractional errors, NRMSE, and SMAPE in the case of $\eta(r)$, IAEMU captures the overall shape of the correlations well.

Finally, we found that IAEMU has comparable performance to HALOTOOLS-IA when used to fit the alignment parameters $\mu_{\rm cen}$ and $\mu_{\rm sat}$ to IA correlation measurements from the TNG300 hydrodynamic simulation, in a manner similar to the robustness test originally performed in [111]. This demonstrates IAEMU's robustness to OOD shifts for inverse problems. Specifically, we observe a better than 0.4σ agreement in the $\mu_{\rm cen}$ and $\mu_{\rm sat}$ posteriors across three separate mass regimes between IAEMU, fit using HMC, and HALOTOOLS-IA, fit with Markov Chain Monte Carlo (MCMC). A significant advantage was the improvement in computational efficiency; while HALOTOOLS-IA with MCMC required approximately one day on a cluster CPU, IAEMU completed the inverse problem in less than a minute on a single GPU. This constitutes a nearly $2000\times$ speed up over MCMC with HALOTOOLS-IA, demonstrating that the efficiency benefits of neural network surrogate models extend beyond forward modeling.

Limitations. IAEMU is an emulator designed to predict 2PCFs based on HOD modeling. Consequently, one inherent limitation of IAEMU is its reliance on phenomenological HOD parameterizations, and IAEMU thus inherits any limitations currently present in HALOTOOLS-IA. Moreover, since IAEMU was trained solely on HODs conducted on the BOLSHOI-PLANCK N-body simulation, it does not currently factor in cosmology dependence, which is an avenue of future work.

3.7.2 diffHOD-IA Summary

We have developed DIFFHOD-IA, a differentiable HOD framework that includes galaxy IA modeling. Our HOD implementation closely follows that of [110], which is extended to the IA modeling of [111] via inverse Dimroth–Watson CDF sampling. We additionally draw inspiration from [112] to extend the differentiability of DIFFHOD-IA to include IA correlation functions. The DIFFHOD-IA code is publicly available.

Speed. The utility of DIFFHOD-IA over HALOTOOLS-IA stems from its differentiability, which enables the use of gradient-based algorithms to optimize the input model parameters. For forward modeling, DIFFHOD-IA runs in comparable time to HALOTOOLS-IA (seconds per catalog on CPU), with no significant speedup on GPU. For forward modeling, IAEMU is substantially faster, exhibiting an approximately $10000\times$ speedup over HALOTOOLS-IA and DIFFHOD-IA. The benefits of differentiability are clearer in inverse modeling. Our HMC analysis converged in approximately 5 minutes on a single GPU, compared to approximately one day across 150 CPU cores for HALOTOOLS-IA with MCMC.

The utility of DIFFHOD-IA over IAEMU comes from its differentiability at the catalog level. IAEMU models the correlations $\xi(r)$, $\omega(r)$, and $\eta(r)$, bypassing the galaxy catalog generation step. This limits its predictive abilities to only 2PCFs, whereas DIFFHOD-IA can be extended to differentiably model any summary statistic. This also requires that any changes to the HOD formulation require retraining IAEMU, while these changes, if necessary, can be made directly within the DIFFHOD-IA simulation. In addition, IAEMU may not generalize well to different dark matter catalogs (e.g., different cosmologies), whereas this is not a concern with DIFFHOD-IA.

DIFFHOD-IA is written in JAX [98], enabling several processing and vectorization speedups via `jax.jit` and `jax.vmap`. As vectorization on GPU requires static array sizes, modeling up to the 2PCF level can be easily vectorized. This does not include generating galaxy catalogs, as the sizes of galaxy catalogs vary across HOD instances and random seeds. For potentially different galaxy-halo connections, or with using weighted-galaxy catalogs, static galaxy catalogs can be generated in parallel.

Results. We benchmarked the accuracy of the sampling procedure and gradients of DIFFHOD-IA, comparing autodiff computed gradients with analytic and finite-differences, and comparing the differentiable Dimroth Watson samples with samples from the true distribution. DIFFHOD-IA agreed with HALOTOOLS-IA across all tests: including galaxy number counts, and the $\xi(r)$, $\omega(r)$, and $\eta(r)$ statistics for the fiducial galaxy catalog.

For a fixed HOD, we utilized gradients in DIFFHOD-IA to retrieve IA parameters $\boldsymbol{\mu}$ from a mock observable galaxy catalog and correlations corresponding to the TNG300 simulation. We tested this using gradient-based optimization with a moment matching and correlation function loss. We further leveraged the differentiability of DIFFHOD-IA in a Hamiltonian Monte Carlo pipeline, showing excellent agreement with HALOTOOLS-IA and IAEMU.

3.7.3 Future Directions

Future extensions of this work could proceed in several directions. The current implementation uses the [6] HOD formulation, but the differentiable framework readily accommodates more sophisticated HOD models that include assembly bias [175] or environment-dependent effects [224]. In addition, HALOTOOLS-IA specifies galaxy misalignments solely according to galaxy orientations, and does not currently include galaxy shapes or ellipticities. DIFFHOD-IA can readily be extended with shape information in accordance with future versions of HALOTOOLS-IA. Similarly, while we have focused on the radial alignment model with constant alignment strength, the distance-dependent alignment strength model implemented in DIFFHOD-IA could be explored for galaxies with radially-varying alignment properties. More generally, this can be extended to different IA parameters *per galaxy*, as opposed to catalog-wide definitions.

The differentiability also allows inserting DIFFHOD-IA into larger differentiable inference pipelines. For example, DIFFHOD-IA can be integrated into simulation pipelines that incorporate a differentiable particle mesh solver such as JAXPM [225], along with a differentiable halo finder like JFOF [110]. This would enable end-to-end gradient flow from cosmological, HOD, and IA parameters to the galaxy field.

The differentiable correlation function framework could also be extended to 2D-projected statistics, which are more directly comparable to observational weak lensing measurements [226]. Of equal interest is the extension to higher order statistics and field-level inference methods which would impose tighter constraints on parameters. A joint inference over both HOD and IA parameters represents a natural application, using the differentiable estimators for $\xi(r)$ and $\omega(r)$. In the present work, fixing the HOD avoids degeneracies between HOD and IA parameters that arise at the 2PCF level — for instance, the satellite fraction governed by $\log M_1$ and α directly affects the relative contributions of centrals and satellites to $\omega(r)$, which could partially mimic changes in $\mu_{\rm sat}$. Such joint analyses over the full 7-dimensional parameter space would require careful treatment of these degeneracies and are reserved for future work.

There remain several functional improvements required for IAEMU to be fully deployable in cosmological analyses. At present, IAEMU predicts only 2PCFs and is not configured to model corresponding one-point statistics such as galaxy number counts. Furthermore, the 2PCFs computed in HALOTOOLS-IA only include galaxy orientations and not their full shapes, which are essential for a complete IA analysis. In addition, emerging simulation suites are beginning to forego explicit host-subhalo distinctions in favor of halo-core models [227]. As a result, HALOTOOLS-IA would need to be updated to incorporate halo-core models, and IAEMU would accordingly need to be retrained. Incorporating these improvements is ongoing work in both HALOTOOLS-IA and future iterations of IAEMU. Lastly, IAEMU also has the capacity to perform joint inference over both HOD and IA parameters; however, this was not explored in Section 3.6.4. The central purpose of that experiment was reproducing the results of [111], which focused on varying the IA parameters. Performing a joint inference over both HOD and IA parameters will be explored in future iterations of IAEMU.

In future versions of the emulator, we will improve upon the IAEMU pipeline by exploring different architectural, data, and modeling choices. IAEMU, as presented here, operates in a traditional supervised learning regime, where the model learns a direct, deterministic mapping between the HOD parameters and correlations. Although we introduce stochasticity for uncertainty quantification via MC dropout and MVE, a more natural probabilistic approach could be achieved through conditional diffusion generative modeling, where the model learns a probabilistic mapping via a denoising process on the data, or through flow-based architectures as in [228]. These models can also have their internal representations restricted by known symmetries in the data, enhancing their effectiveness in physical settings like this [see 229, for an example of $SO(3)$-equivariant diffusion applied to IA], which naturally lends such techniques to field-level modeling of the full galaxy catalog. Field level emulation can also expand IAEMU summary statistics outside the 2PCFs modeled here. The denoising training paradigm when applied to cosmology has thus far exhibited promising results in enhancing the resolution of existing simulations [230] as well as functioning as surrogate models [231].

A field-level emulator for HALOTOOLS-IA could incorporate elements of NN-based modeling together with differentiable components, such as differentiable HOD models [110]. Both NNs and differentiable simulations share a differentiable structure, allowing their components to be integrated within the same modeling framework. While non-NN-based differentiable methods provide useful tools for inverse modeling, they can be computationally demanding and are often complemented by faster NN-based surrogates.

Simulation-based modeling has opened a new set of opportunities to better understand galaxy intrinsic alignments, complementing earlier analytic and semi-analytic efforts. However, these new techniques have incurred additional computational expense. In this chapter, we have shown a compelling case in which accuracy and efficiency can be achieved with both NN-based emulators and differentiable simulations for galaxy intrinsic alignments from HOD simulations. These developments provide a foundation for efficient, gradient-based joint inference of HOD and IA parameters in current and future weak lensing surveys, and offer efficient and promising surrogate models for halo-based galaxy bias and IA modeling with the potential to expedite model validation in Stage IV weak lensing surveys.

Chapter 4

Symmetries and Domain Adaptation for Neural Network Generalization

Deep NNs excel at extracting complex features from data, making them a powerful tool for a wide range of tasks, including classification, regression, and anomaly detection. Unfortunately, some extracted features can be very dataset-specific, which makes it challenging for NN models to generalize to data that differs from the training data, even when the differences are subtle. For instance, a significant drop in performance occurs when the input distribution changes between the training and test datasets, despite the conditional distribution of the labels given the inputs remaining the same — a scenario commonly referred to as a "covariate shift" [232, 233].

Generalization allows models to perform well across diverse data domains, ranging from subtle variations in input distributions to entirely different datasets or environments. Differences between training and testing data can be due to data collection or quality [234], distortions [235], image corruptions [236], or even single-pixel level differences, which can cause the NN to give inaccurate predictions [237]. Generalization capabilities, in turn, aid the efficiency and applicability of NNs in both science and industry, as they would otherwise need to be continually retrained on new data. For example, in astronomy, a generalized model trained on data from one telescope should accurately predict properties of data from another telescope that has different noise characteristics or resolution, significantly accelerating the process of identifying or characterizing celestial objects across surveys.

Domain Adaptation (DA) is a group of methods that aim to improve the generalization capabilities of NNs by enabling the NN to learn features in the data that persist across domains [238,

239]. It is often applied to problems where one has access to labeled data from a "source" domain, but would also like the model to perform well on unlabeled "target" domain data. A large group of distance-based DA methods tackles the covariate shift problem by minimizing some distance metric between internal NN latent representations (distributions) of the source and target data. This, in turn, forces the NN to extract mainly domain-invariant features, which makes both latent data distributions well-aligned.

Some well-known distance-based DA methods use Maximum Mean Discrepancy (MMD) [240], correlation alignment (CORAL) [241], contrastive domain discrepancy (CDD) [242], the Kullback-Leibler (KL) divergence [243], or the Wasserstein distance [244], which is derived from the Optimal Transport (OT) [245] theory for measuring distance between probability distributions. The theory is aimed at solving the OT problem, which includes determining the minimal cost of transporting a probability mass from one distribution to another, where the cost is defined by a chosen metric that measures the effort required to move it. OT thus provides a principled way to quantify the dissimilarity between distributions, capturing both the geometry of the space and the magnitude of the differences.

In the sciences, NNs have made significant progress — from mapping the structure of biological proteins [246] to the cosmos [247]. In astronomy, astrophysics, and cosmology, the success is mainly due to the emergence of large datasets and high-fidelity simulations. Stage IV projects, such as the Vera Rubin Observatory Legacy Survey of Space and Time (LSST) [248], the Nancy Grace Roman Telescope [18], and the Euclid mission [19], will yield an unprecedented amount of data that analysis pipelines must analyze efficiently. Simultaneously, there are many simulations of the Universe from sub-parsec to gigaparsec scales with magneto-hydrodynamic simulation suites, such as IllustrisTNG [249] and CAMELS [250], that can be used to prepare pipelines for the analysis of real data.

Since simulation-trained models often exhibit a substantial drop in performance when applied to real data, there has been extensive work in implementing DA for astronomical applications: for galaxy morphology classification [251, 252], classification of supernovae and identification of Mars landforms, [253], inference of cosmology [254], inference of strong gravitational lensing parameters [255, 256], constraining star formation histories of galaxies [257], and gravitational lens finding [258]. In astronomy, atmospheric distortion, telescope noise, Point Spread Function (PSF) blurring, and data processing errors commonly affect data quality and contribute to the domain shift between simulated and real data. Most DA applications in astronomy and beyond require extensive hyperparameter tuning that is highly sensitive to the dataset. The goal of this work is to address this

shortcoming.

Domain shift problems are common in other areas of science and industry, with DA methods being applied to a variety of tasks such as medical image analysis [259], classification of remote sensing data [260], geospatial semantic segmentation [261], autonomous driving in the presence of changing weather conditions [262], material discovery [263], etc. In such domains, shifts can occur across image modalities—for instance, between different wavelengths such as optical and infrared, or, in medical imaging, between MRI and CT scans of the same tissue.

Most image-based problems in the sciences are addressed using CNNs. At the same time, ENNs are gaining popularity due to their ability to explicitly encode symmetry information present in the data, which is often explicitly known — e.g., $\mathrm{SL}(2, \mathbb{C})$ in particle physics [264] and $\mathrm{SE}(3)$ for rigid body motion [265]. ENNs have also been shown to achieve state-of-the-art performance on many tasks, such as 3D object classification and alignment [266], dynamical systems modeling, representation learning in graph autoencoders and predicting molecular properties [267], classification and segmentation tasks [268], and accounting for function-preserving scaling symmetries that arise from activation functions [269]. They have also been shown to possess inherent robustness to symmetric transformations and noise perturbations due to their restricted feature learning [92, 270, 271]. This robustness has been observed to increase with the group order N for the cyclic group C_N and dihedral group D_N, which are subgroups of $\mathrm{SO}(2)$ and $\mathrm{O}(2)$, respectively. Still, in the presence of covariate shifts, even ENNs can exhibit a drop in performance [92].

In this chapter, we focus on the Sinkhorn divergence [272], a symmetrized variant of regularized OT distances. We introduce Sinkhorn Dynamic Domain Adaptation (SIDDA), a more automated training algorithm for DA that minimizes the need for hyperparameter tuning. To achieve this, we leverage active scaling of (1) the entropic regularization of the OT plan, and (2) the weighting of classification and DA loss (i.e., for addressing the covariate shift in the datasets) terms, during training. To demonstrate the efficacy and broad scope of applications of our proposed DA method, in this work, we use several datasets of varying complexity: simple simulated datasets, well-known benchmark datasets used in the computer science community, real observational galaxy datasets, and real remote sensing datasets. We also study the robustness of ENNs and the improved efficacy of SIDDA when used in conjunction with ENNs compared to typical CNNs.

The chapter is organized as follows. In Section 4.1, we describe existing DA methods and their shortcomings, motivating the use of OT-based distances to facilitate DA. We describe the core methodology of SIDDA, and motivate the use of ENNs as inherently robust architectures. In Section 4.2, we describe the construction of all simulated and real datasets we use in our study. In Section

4.3, we describe the network architectures used, training procedures, metrics for calibration, and metrics for interpreting NN latent spaces. In Section 4.4, we summarize our results and conclude in Section 4.5.

4.1 Methods

The major components of this work are DA and equivariance, which we combine to create a more efficient and robust NN classifier. Within DA, we implement the Sinkhorn divergence, a symmetrized and regularized variant of OT distances that offers considerable improvement in DA over traditional methods. We construct a training program that constantly adjusts the loss landscape and regularization strength of the Sinkhorn plan, offering optimal domain alignment with minimal hyperparameter tuning.

4.1.1 Domain Adaptation

DA comprises a set of techniques aimed at aligning the latent distributions of NNs in the presence of covariate shifts in data. Typically, DA operates in settings where one can access labeled source images $x \in \mathbf{X_s} \subseteq \mathbb{R}^{m \times m}$, and unlabeled target images $x^* \in \mathbf{X_t} \subseteq \mathbb{R}^{m \times m}$ from $\mathbf{X}_s$ and $\mathbf{X}_t$ source and target data domains, where m denotes the number of pixels in each dimension (height and width) of the image.

Consider the latent vectors $z \in \mathbf{Z}_s \subseteq \mathbb{R}^l$ and $z^* \in \mathbf{Z}_t \subseteq \mathbb{R}^l$, where $\mathbf{Z}_s$ and $\mathbf{Z}_t$ denote the latent spaces of the source and target domains, respectively, and l represents the dimension of the latent vectors (i.e., the width of the corresponding neural network layer). Latent distributions refer to the probability distributions over these latent vectors, and during training, DA minimizes a statistical distance measure between them. DA is incorporated through an additional loss term, $\mathcal{L}_{\text{DA}}$, alongside the standard task loss (e.g., cross-entropy for classification, $\mathcal{L}_{\text{CE}}$), to promote alignment between the two latent distributions. In this work, $\mathcal{L}_{\text{DA}}$ (the "DA loss") is the loss due to covariate shifts. The total loss function is then:

$$\mathcal{L} \propto \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DA}} \ . \tag{4.1}$$

In practice, a delicate balance between the two terms must be achieved to ensure proper alignment.

There are numerous DA methods, each with its own strengths and limitations. One commonly used approach is MMD [240, 273], where the distance between the means of the latent

embeddings from the source and target domains serves as the DA loss function. In DA, comparisons are often made between distributions that are not explicitly known but can be sampled. MMD can be combined with kernel methods, which map probability distributions into a high-dimensional reproducing kernel Hilbert space (RKHS) [273], providing a more flexible method for comparing distributions. This approach allows for analyzing distributions through well-defined operations in the RKHS, even when the original distributions are not well-defined. The MMD between two probability distributions μ and ν — representing distributions over latent vectors z and z^*, respectively — is

$$\text{MMD}(\mu,\nu) = \left(\mathbb{E}_{z,z'\sim\mu}\left[k(z,z')\right] + \mathbb{E}_{z^*,z^{*'}\sim\nu}\left[k(z^*,z^{*'})\right] - 2\mathbb{E}_{z\sim\mu,z^*\sim\nu}\left[k(z,z^*)\right] \right)^{1/2},$$

where k represents the kernel function, and z, z', and z^*, $z^{*'}$ are individual samples from latent distributions μ and ν, respectively.

Despite its utility, MMD has several theoretical and implementation-related shortcomings. First, its efficacy is highly sensitive to the choice of k. In typical applications, the Gaussian kernel $k(z,z^*) = \exp\left(-\frac{||z-z^*||^2}{2\epsilon^2}\right)$ is used with kernel bandwidth ϵ. Other kernel options include the linear kernel $k(z,z^*) = z^T z^*$, the Laplacian kernel $k(z,z^*) = \exp\left(-\frac{||z-z^*||}{2\epsilon}\right)$, and others. Most kernels generally belong to a one-parameter family (e.g., ϵ for Gaussian and Laplacian kernels) and must be carefully tuned, or complex linear combinations of kernels with many different parameter values must be used. The specific choice of kernel depends heavily on the nature of the problem. That is, MMD can exhibit bias with small sample sizes and often struggles with domain alignment when dealing with high-dimensional distributions [274, 275].

4.1.2 Optimal Transport and The Sinkhorn Divergence

OT distances and their symmetrized variants, such as Sinkhorn divergences, offer an alternative to MMD. Traditionally, computing OT is prohibitively expensive [276]. Entropic regularization, OT_σ, [277] provides a more efficient method for estimating OT distances. The regularized OT is defined as

$$\text{OT}_\sigma(\mu,\nu) = \min_{\gamma\in U(\mu,\nu)} \left(\sum_{i,j} \gamma_{ij} d(z_i, z_j^*)^p + \sigma H(\gamma) \right), \tag{4.2}$$

where $d(z_i, z_j^*)^p$ is the distance between source feature z_i and target feature z_j^*. When $p = 1$, this distance becomes the Earth Mover's distance [278], and when $p = 2$, it becomes the quadratic

Wasserstein distance. The transport plan $\gamma \in U(\mu, \nu)$ is a joint probability distribution between μ and ν, where the set of admissible transport plans $U(\mu, \nu)$ is defined by the marginal constraints:

$$\sum_j \gamma_{ij} = \mu_i, \quad \sum_i \gamma_{ij} = \nu_j. \tag{4.3}$$

The entropy $H(\gamma) = -\sum_{i,j} \gamma_{ij} \log \gamma_{ij}$ regularizes the transport plan γ, and σ controls the regularization strength. One limitation of OT_σ is that $\text{OT}_\sigma(\mu, \mu) \neq 0$, implying a non-zero cost even when transporting a distribution to itself, leading to bias in the measure.

To correct this bias, the Sinkhorn divergence $S_\sigma(\mu, \nu)$, defined as

$$S_\sigma(\mu, \nu) = \text{OT}_\sigma(\mu, \nu) - \frac{1}{2}\text{OT}_\sigma(\mu, \mu) - \frac{1}{2}\text{OT}_\sigma(\nu, \nu), \tag{4.4}$$

can compensate for the bias in OT_σ [279]. As $\sigma \to 0$, $S_\sigma(\mu, \nu)$ converges to the true optimal transport OT_0, and as $\sigma \to \infty$, it interpolates towards MMD loss [279]. For small values of σ, an unbiased transport plan that still enjoys the benefits of OT-based distances can be constructed.

4.1.3 Dynamic Sinkhorn Divergences for Domain Adaptation

For this work, $\mathcal{L}_{\text{DA}}$ in Equation 4.1 is specifically the Sinkhorn divergence $S_\sigma(\mu, \nu)$. However, a careful balance between $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{DA}}$ must be achieved to optimize the classification task while simultaneously maximizing domain alignment.

Finding the best weights for each of the loss terms can be very challenging and time-consuming. Furthermore, a single choice of weights might not be the best choice throughout the whole training procedure. To manage this balance, we employ dynamic weighting of the losses by introducing two trainable parameters, η_1 and η_2, which dynamically adjust the contributions of the loss terms for each task throughout training. These parameters ensure that no single loss term dominates the optimization process, allowing the loss landscape to be optimally adjusted for both tasks. Drawing inspiration from [216], we use the following for the total loss function:

$$\mathcal{L} = \frac{1}{2\eta_1^2}\mathcal{L}_{\text{CE}} + \frac{1}{2\eta_2^2}\mathcal{L}_{\text{DA}} + \log(|\eta_1\eta_2|)\ , \tag{4.5}$$

where η_1 and η_2 are trainable scalars, and their values are jointly learned with the model weights during training. At the beginning of training, both η_1 and η_2 are initialized to a value of one and are subsequently updated during training. The inclusion of the term $\log(|\eta_1\eta_2|)$ acts as a regularization to prevent η_1 and η_2 from collapsing to unstable values, such as zero. As $\eta_i \to 0$, the corresponding loss term is more heavily weighted. To ensure that no single component dominates, we impose the

additional constraint $\eta_2/\eta_1 \geq 0.25$. In general, the DA term must not dominate over the classification loss, which the above inequality enforces. For our implementation, we found that this threshold worked best and stabilized training, but such a cutoff may not always be optimal.

In [216], the two weight terms η_1 and η_2 were introduced for the dynamic weighting of the losses. These terms explicitly minimize the regression uncertainty associated with each loss term, as their model outputs a Gaussian distribution with variance η_i^2 for each task. In the case of classification, their weight terms become $1/\eta_i^2$. Since uncertainties are not one of the network outputs in our case, the exact written form of loss weights is not important and the extra factor of two can very well be absorbed into the trainable weight parameter.

The level of regularization σ in $S_\sigma(\mu, \nu)$ is another critical hyperparameter [279]. When σ is too small, the transport plan approaches the true Wasserstein distance, substantially increasing the computational cost of domain alignment. In this regime, the Sinkhorn iterations may also fail to fully converge, potentially introducing biases similar to those observed in OT_σ. Conversely, if σ is too large, the regularization interpolates toward MMD, removing the unique benefits of using $S_\sigma(\mu, \nu)$. To address this, we adopt a unique, dynamic regularization per epoch of training ℓ, σ_ℓ, where the transport plan is continually updated. We compute σ_ℓ iteratively as:

$$\sigma_\ell = \max\left(0.05 \cdot \max_{i,j} \|z_i - z_j^*\|_2,\ 0.01\right) . \tag{4.6}$$

In this formulation, σ_ℓ is dynamically adjusted based on the maximum pairwise distance $D_{ij} = ||z_i - z_j^*||_2$ between the source and target latent distributions. We additionally set scaling based on the appropriate measures on the unit square or cube, which justifies the prefactor of 0.05 in Equation 4.6 [279]. This is further stabilized through the layer normalization of the latent vectors prior to computing D_{ij}. This stabilization discourages outliers from disproportionately affecting the computation of σ_ℓ. Finally, we impose a lower bound of $\sigma_\ell \geq 0.01$ to mitigate numerical instabilities and excessive computation as S_σ approaches the unregularized Wasserstein distance.

Batches of size n of latent vectors from source and target data, denoted z_n and z_n^*, are retrieved through a forward pass of a single combined batch of the source and target data, $\mathbf{X} = [x_n, x_n^*]$, through the NN. The NN outputs a combined batch of latent vectors, denoted as $\mathbf{Z}$, which is subsequently separated into two subsets: one corresponding to the source domain and the other to the target domain. This is particularly important for NNs, which utilize batch normalization [280]. If z_n and z_n^* were passed separately, the batch statistics would be computed independently, leading to inconsistent normalization as the z_n batch statistics will not incorporate z_n^* and vice versa.

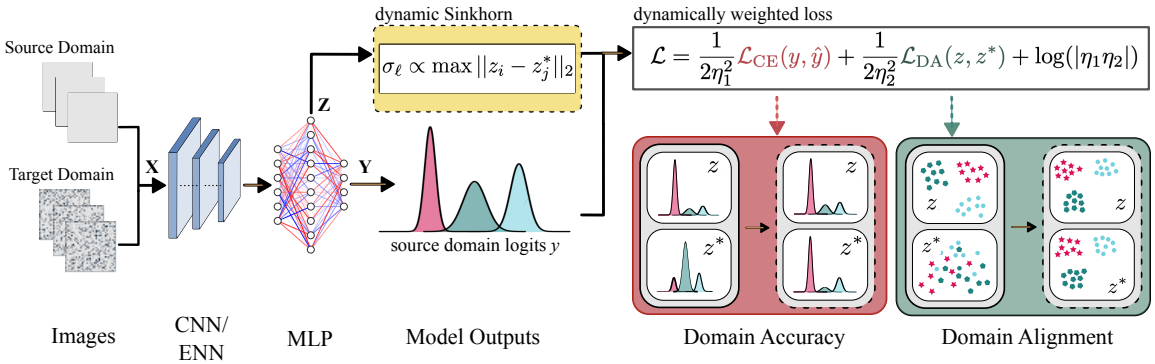


Figure 4.1: SIDDA pipeline. The source and target domain batches of size n, x_n and x_n^*, are first concatenated into a single batch $\mathbf{X}$ before being passed into the model. After passing through the convolutional layers, the neural network produces a combined batch of latent vectors, $\mathbf{Z}$, extracted from the final linear layer. This layer is positioned just before the output layer, which generates the class probabilities, $\mathbf{Y}$. Both $\mathbf{Z}$ and $\mathbf{Y}$ are split into separate batches for the source and target domains, resulting in z_n (source) and z_n^* (target) from $\mathbf{Z}$, and y_n (source) and y_n^* (target) from $\mathbf{Y}$, respectively. Only the source y_n are used in training, as there are typically no target domain labels. Both z_n and z_n^* are used to compute σ_ℓ, a parameter that iteratively updates the regularization of the Sinkhorn plan in $\mathcal{L}_{\mathrm{DA}}$. This process aligns the latent distributions of the source and target domains. This loss contribution is appropriately weighted with the classification loss, $\mathcal{L}_{\mathrm{CE}}$, using a dynamic weighting of the tasks. The result of training using SIDDA is improved classification accuracy in both domains due to the aligned latent distributions, which can be visualized using non-linear clustering algorithms on the NN latent distributions.

We name this combined approach, which is dynamically adjusting a balance between cross-entropy and DA loss terms, and likewise dynamically adjusting the regularization of the Sinkhorn plan that facilitates the DA, SIDDA. SIDDA not only facilitates effective alignment between source and target domains, as we will demonstrate, but also reduces the need for extensive hyperparameter tuning — a common challenge in DA implementations [281]. As a result, this method provides a more automatic and reliable DA implementation with minimal computational overhead, leveraging existing resources that allow efficient computation of Sinkhorn divergences [282]. We implement this technique using the `geomloss` library [283], which provides GPU implementations for Sinkhorn divergences and compatibility with `PyTorch` [210]. A pipeline illustrating the dynamic DA approach during training is given in Figure 4.1.

4.1.4 The Jensen-Shannon Distance

Recent advancements in DA theory have introduced the Jensen–Shannon (JS) divergence [284] as a fundamental tool for understanding the inherent limitations of DA [285]. The JS divergence is a symmetrized statistical distance metric. For two latent distributions μ and ν, the JS divergence D_{JS} is defined as

$$D_{\text{JS}}(\mu\|\nu) = \frac{1}{2}D_{\text{KL}}(\mu\|\tau) + \frac{1}{2}D_{\text{KL}}(\nu\|\tau) \ , \tag{4.7}$$

where D_{KL} denotes the KL divergence [243], defined as

$$D_{\text{KL}}(\mu \ || \ \nu) = \int_{-\infty}^{\infty} p(z) \log\left(\frac{p(z)}{q(z)}\right) dz \ . \tag{4.8}$$

Here, p and q denote probability densities of the latent features z in the source and target distributions μ and ν. $\tau = \frac{1}{2}(\mu+\nu)$ represents the mixture distribution of μ and ν. The JS divergence offers two key advantages over D_{KL}: it is symmetric (since, in general, $D_{\text{KL}}(\mu\|\nu) \neq D_{\text{KL}}(\nu\|\mu)$), and it is always finite. Additionally, the square root $\sqrt{D_{\text{JS}}}$ defines a metric known as the Jensen-Shannon distance.

For DA applications, there exists a lower bound on the target domain loss [285]:

$$\mathcal{L}_t(z^*) \geq \mathcal{L}_s(z) - \sqrt{D_{\text{JS}}(\mu\|\nu)} \ , \tag{4.9}$$

where $\mathcal{L}_s$ is the source domain loss, $\mathcal{L}_t$ is the target domain loss, and $\sqrt{D_{\text{JS}}(\mu\|\nu)}$ is the JS distance between the source and target latent distributions μ and ν, respectively. This bound emphasizes that perfect alignment between source and target domains is fundamentally constrained by two

components: $\mathcal{L}_s$ and the JS distance between the source and target distributions. A smaller JS distance implies that the feature distributions of the source and target domains are closely aligned, which lowers the bound on $\mathcal{L}_t$ and enables better transferability of the learned model. Conversely, a larger JS distance indicates a greater discrepancy, limiting the potential for minimizing target domain loss through adaptation alone.

The similarity between the source and the target latent distributions μ and ν is inherently influenced by the feature extraction capabilities of the neural network. DA methods aim to align features in the latent distribution; however, these features are ultimately limited by the network architecture. As a toy example, consider the case of image classification, where the architecture is a MLP. Many image classification tasks exhibit translation invariance, inherent in CNNs, but not in MLPs. DA on this task with CNNs will likely be more successful than with MLPs, as the translation invariance of the CNN further restricts the allowable features, and thus, the cost of alignment will be smaller. In particular, we define "robust" features as those that respect the underlying data symmetries. More specifically, they yield similar classification probabilities under isometries that preserve the symmetries of the images. If these symmetries persist in both the source and target domains, which is typically true except for extreme symmetry-breaking perturbations, then the cost of aligning robust features will be less than features learned from symmetry-agnostic architectures. Consequently, it is reasonable to expect that ENNs endowed with appropriate higher-order symmetries will exhibit greater robustness and achieve more precise DA alignment than CNNs, because the latent distributions of ENNs are inherently more constrained. Furthermore, when the assumption of underlying symmetries holds in both the source and target domains, as is typical in many DA applications, this advantage of ENNs becomes even more pronounced.

4.2 Data

We evaluate the performance of our method on three simulated datasets and two real datasets: (1) a single-channel dataset of shapes consisting of lines, circles, and rectangles; (2) a single-channel dataset resembling astronomical objects, including stars, spirals, and elliptical galaxies; (3) the multichannel MNIST-M dataset [286]; (4) the Galaxy Zoo (GZ) Evo dataset of galaxies observed by two different optical telescopes [287]; and (5) the Multi-Modal Remote Sensing Scene Classification (MRSSC2) dataset that contains optical and synthetic aperture radar (SAR) imaging [288]. The MRSSC2 dataset enables us to test performance in the presence of more severe covariate shifts caused by different wavelengths of the imaging data. The shapes and astronomical

objects datasets are constructed using `DeepBench` [289]. All datasets used in our experiments can be found on Zenodo.

4.2.1 Covariate Shifts

We use images from three simulated datasets, shown in Figure 4.2, to study the performance on induced covariate shifts between the source and target domains. For all of our simulated datasets, we introduce fixed levels of Poisson noise in the target domain. Additionally, for MNIST-M, we also study the effects of PSF blurring in the target domain. By studying these two distinct covariate shifts, we evaluate the robustness of our method on covariate shifts relevant to data in realistic settings, particularly in the context of astrophysics and cosmology.

This is implemented for an image I with grid values (ζ, ξ) and channels c as

$$I_{\text{Poisson}}(\zeta, \xi, c; S) = I(\zeta, \xi, c) + P\left(\frac{\langle I \rangle}{S} - \langle I \rangle\right) , \tag{4.10}$$

where S is the signal-to-noise ratio, and P denotes the Poisson distribution with rate parameter $\lambda = \frac{\langle I \rangle}{S} - \langle I \rangle$. We incur PSF noise in each image channel by convolving the images with a Gaussian kernel G of kernel width ϵ:

$$I_{\text{PSF}}(\zeta, \xi) = (I * G)(\zeta, \xi), \tag{4.11}$$

where

$$G(\zeta, \xi) = \frac{1}{2\pi\epsilon^2} \exp\left(-\frac{\zeta^2 + \xi^2}{2\epsilon^2}\right) . \tag{4.12}$$

4.2.2 Simulated Images

For the shapes dataset, we use `DeepBench` [289], an open-source library for generating simulated datasets, and randomly construct rectangles, lines, and circles with varying radii (for circles), heights and widths (for rectangles), and lengths (for lines). The object positions, orientations, and thicknesses are also randomly assigned to introduce variance in the dataset. Poisson noise in the images is normalized with respect to the original image signal. We set a signal-to-noise ratio $S = 0.05$. Example images from the dataset can be seen in the left two panels of Figure 4.2.

We also use `DeepBench` for simulating astronomical objects. We generate astronomical objects resembling spiral galaxies, elliptical galaxies, and stars. For spiral galaxies, we randomly assign the centroid, winding number, and pitch to ensure morphological variation. The pitch is the angle indicating how tightly the arms are wound, while the winding number measures the total

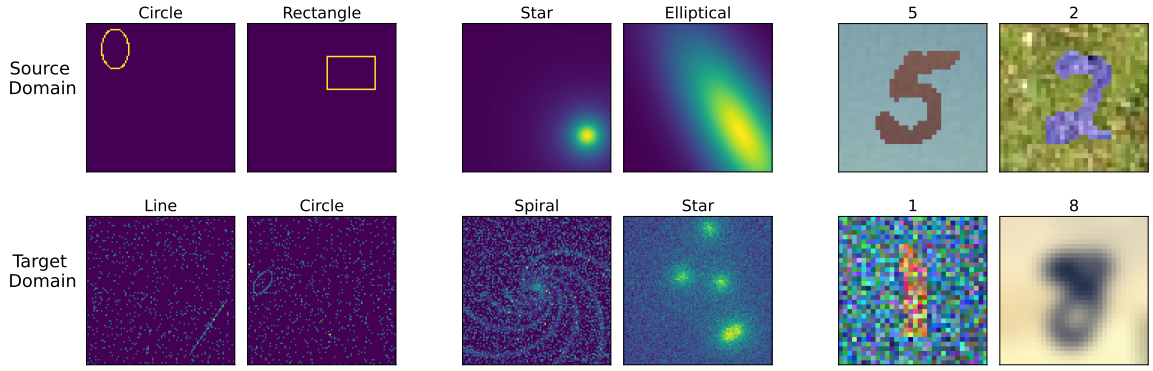


Figure 4.2: Example images for simulated datasets in the source domain (top row) and the target domain (bottom row) with corresponding labels. **Left Panels:** Shapes dataset, featuring lines, rectangles, and circles, simulated with `DeepBench`. This dataset includes variations in object positions and orientations, with Poisson noise added and normalized relative to the image signal in the target domain. **Middle Panels:** Astronomical objects dataset, generated using `DeepBench`. Parameters for spiral and elliptical galaxies were randomly sampled to determine morphology and position, while stars were generated similarly, with the number of stars as an additional parameter. Target domain images include additional Poisson noise. **Right Panels:** MNIST-M dataset with simulated Poisson noise (bottom left) and PSF blurring (bottom right) in the target domain.

number of arm rotations from the center to the galaxy's edge. For elliptical galaxies, we vary the amplitude, radius, ellipticity, Sérsic index [290], and rotation, as well as the centroid location. The amplitude sets the brightness level, ellipticity describes the degree of deviation from a circle, the Sérsic index controls light concentration (higher values indicate more central concentration), and the rotation defines the orientation angle of the galaxy's major axis. Lastly, we apply similar variations to generate stars, with the number of stars in each image uniformly distributed in the range $[0, 10]$. We use a fixed Poisson noise level of $S = 0.2$ to generate noisy target domain images. This level of noise was chosen as it allows for a sufficient decrease in the target domain performance for models without DA. Example images are shown in the middle two panels of Figure 4.2. Both of these datasets contain 12,000 training images (with 20% being used for validation) and 3,000 test images in each domain. The images are square with 100 pixels on each side, and they are single-channel: each sample image has dimensions $100 \times 100 \times 1$.

MNIST-M [286] is a dataset that combines the handwritten digits of MNIST [291] with randomly extracted color photos from BSDS500 [292] as background images. The original dataset contains 59,001 training images and 90,001 test images, out of which we use a balanced subset of 15,000 training (with 20% set aside for validation) and 5,000 testing images in each domain. Since this is a three-channel dataset, images have a dimension of $32 \times 32 \times 3$. We then create two types of target domain covariate shifts: 1) we set a signal-to-noise ratio of $S = 0.05$ for Poisson noise, and 2) a kernel width of $\epsilon = 2$ for PSF blurring. Example images are shown in the right two panels of Figure 4.2.

Both the images and the induced covariate shifts are simulated and do not capture all the complexities of real-world noise. Nevertheless, these datasets provide valuable benchmarks for challenges commonly encountered in astronomical and cosmological contexts, where DA methods can substantially enhance the robustness of neural network-based image classification pipelines.

4.2.3 Real-Sky Galaxy Image Dataset

We use the GZ Evo dataset [287] to test cross-domain robustness in a more realistic scenario, where the covariate shift is present due to differences between images from two different astronomical surveys. These differences are due to different levels of observational noise, PSF blurring, pixel scale, as well as differences in populations of observable astronomical objects (how distant or how faint a resolved object can be). GZ is a citizen science project that labels galaxy images through online participation. GZ Evo combines labeled image datasets across several surveys

Figure 4.3: **Top Panel:** Example source domain images from the GZ Evo dataset with corresponding labels. Images are from GZ2 data observed by SDSS. **Bottom Panel:** Example target domain images from the GZ Evo dataset with the same labels. Images are from GZ DESI (combined observations from the DESI Imaging Surveys).

and iterations of GZ. Within GZ Evo, we use the GZ2 Dataset from the Sloan Digital Sky Survey (SDSS) [293] as the source, and a GZ Dark Energy Spectroscopic Instrument (DESI) dataset that combines observations from the DESI Imaging Surveys (DECals, MzLS, BASS, DES) [294, 295] as the target. Older GZ SDSS data contains objects up to magnitude 17 in the r band, and redshifts below 0.25, while newer GZ DESI includes fainter objects up to magnitude 19 and more distant objects with redshifts below 0.4. Additionally, these surveys are different in the amount of observational noise and PSF blurring. Finally, GZ SDSS images include 3-filter gri images, while GZ DESI includes 3-filter grz images.

The original GZ Evo dataset contains $664,219$ images, each labeled according to vote counts (e.g., "8 of 10 volunteers answered Spiral, and 7 of 10 answered Bar"). GZ Evo also offers an aggregated version where the vote counts are converted into distinct classes, which is convenient for developing machine learning models. $239,408$ galaxies could be confidently assigned a distinct class based on the original vote counts. Of this sample, $82,185$ corresponded to GZ DESI and $95,703$ to GZ SDSS. The galaxy dataset used in this work contains a random sample of $40,000$ (32,000 training images with 20% set aside for validation, and 8,000 testing images) images in each of the domains, across six distinct classes: "smooth-round", "smooth-cigar", "unbarred spiral", "edge-on-disk", "barred spiral", and "featured" galaxies. The "featured" galaxy class corresponds to any galaxy without a clear spiral structure or a visible bar, but which is also not completely smooth. The original galaxy images had dimensions $428 \times 428 \times 3$ and were subsequently downsampled to $100 \times 100 \times 3$ for more efficient training. Example images across all classes between GZ SDSS and

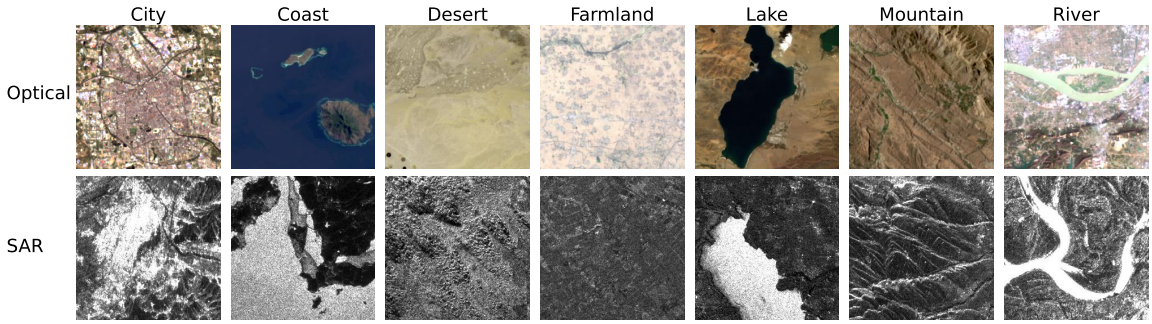


Figure 4.4: **Top Panel:** Example source domain optical images from the MRSSC2 dataset with corresponding labels. **Bottom Panel:** Example target domain SAR images from the MRSSC2 dataset with the same labels.

GZ DESI are shown in Figure 4.3.

4.2.4 Remote Sensing Scene Classification Dataset

Finally, we test the efficacy of SIDDA on a more extreme covariate shift problem, where the source and target domain images are coming from two different wavelength ranges. Such applications are relevant in various scenarios, including medical imaging, remote sensing applications, and astronomy.

We use the MRSSC2 dataset, which contains multi-channel images across four wavelengths: optical, short wavelength infrared, thermal infrared, and SAR (microwave/radio wavelengths) [288]. Optical images show the true surface state, short wavelength infrared is more sensitive to soil moisture, thermal infrared reflects the surface temperature state, and SAR reflects the degree of surface backscattering. This can lead to substantial differences between images observed at different wavelengths. Together, the four imaging modalities (i.e., wavelengths) contain a total of 26710 images split across seven scenes, including "city," "farmland," "mountain," "desert," "coast," "lake," and "river."

We used 4924 optical images (source domain) and 4809 SAR images (target domain) for training, which were split into 80% training and 20% validation. For testing, there are 1231 source domain images and 1203 target domain images. Example images across all classes between optical and SAR are shown in Figure 4.4. We chose these two datasets as the differences between images visually appeared the most extreme, which should lead to the most extreme covariate shift between them. Finally, the original images have dimensions $256 \times 256 \times 3$; we downsampled them to $100 \times 100 \times 3$ for more efficient training.

4.3 Network Architectures and Experiments

We evaluate our method on two sets of NNs: (1) CNNs constructed in `PyTorch` and (2) ENNs constructed in `escnn`, a PyTorch-based library for easy construction of ENNs. Details of architectures and training are presented in [106]. Most of our experiments use an ENN equivariant to D_4, following the results in our previous work [92], as this level of equivariance is beneficial but not overly computationally expensive to train. We also study the performance of our method as a function of group order for the dihedral group. The code used in this work is available on our GitHub.

4.3.1 Equivariant Neural Networks

The efficacy of DA is limited by the feature extraction capabilities of the NN and its performance on the source domain. For image classification tasks, CNNs are useful due to their translation invariance and locality. There are, however, often additional symmetries inherent in the data, such as rotational and reflection invariance, that can be leveraged to enhance feature extraction and improve performance.

ENNs are a subclass of CNNs that can exploit higher-order symmetries besides the typical translation equivariance of CNNs [95, 96]. Of interest for 2D images are symmetries of the Euclidean group $E(2)$ — in particular, the 2D special orthogonal group $\mathrm{SO}(2)$ and the orthogonal group $\mathrm{O}(2)$ and its associated subgroups. These (sub)groups allow ENNs to inherit symmetries of the circle and N-gon, respectively. As $\mathrm{SO}(2)$ and $\mathrm{O}(2)$ are continuous, they contain an infinite number of irreducible representations and have associated challenges when constructing architectures. For this reason, the discrete subgroup of $\mathrm{O}(2)$, the dihedral group D_N, is used in this work, which is straightforward to construct using open-source software, such as `escnn` [296, 297].

For image data (particularly astronomical data), rotational symmetry (with or without reflections) is typically inherent. For instance, galaxy morphologies are often invariant under rotations because there is no preferred reference frame in the Universe. Learning these symmetries can be induced in typical CNNs through data augmentation during training [298, 299]. However, the output feature map f_{out} with grid values (ζ, ξ) from the convolution with kernel K with grid values (i, j),

$$f_{\text{out}}(\zeta, \xi) = \sum_{i,j} K(i, j) \cdot f(\zeta + i, \xi + j), \tag{4.13}$$

is inherently only translation-equivariant. In contrast, group convolution [95, 300] in ENNs can be

equivariant to any arbitrary group G:

$$f_{\text{out}}(g) = \sum_{h \in G} K(h^{-1}g) \cdot f(h) \tag{4.14}$$

for $g, h \in G$: g and h correspond to the transformation for which the output feature map $f_{\text{out}}(g)$ is computed. For example, if G is the group of 2D rotations $\text{SO}(2)$, g and h represent specific rotation angles. ENNs construct specialized filters K that are equivariant to the desired symmetries. Therefore, they learn transformation-invariant features that preserve the underlying symmetries in the data throughout training.

Each NN operation — convolution, activation, pooling, and dropout — must be equivariant [102]. Our ENN architectures have three convolutional blocks, each containing a group convolution (`R2Conv`), batch normalization (`InnerBatchNorm`), ReLU activation, max pooling (`MaxPoolPointwise2D`), and dropout (`PointwiseDropout`). The global group equivariance for the network is defined before the first convolutional layer. Following the convolutional layers, a group pooling operation aggregates over all the symmetry channels, an essential step for constructing invariant representations in the latent distribution as discussed in [301]. Finally, the ENN includes two linear layers that downsample the learned representation to the designated number of output classes for classification.

The latent vector of the NN is extracted before the last linear layer, with no dropout in the linear layers. The latent vectors for all our networks undergo a layer normalization [302], which serves to stabilize the latent distributions by standardizing the feature distributions across each sample, ensuring consistent scaling and preventing the activations from drifting to extreme values. We found that the layer norm is useful when employing SIDDA during training for more stable computations of σ_ℓ as given in Equation 4.6.

The architectures used in our experiments are based on the D_N group, which exhibits reflection symmetry. In the majority of our experiments, we set $N = 4$, as it provides notable benefits without imposing significant computational overhead during training. Our CNN has the same architecture, except the network components are not equivariant.

4.3.2 Training

We train all networks typically ($\mathcal{L} = \mathcal{L}_{\text{CE}}$) and with SIDDA (Equation 4.5) and study performance differences in the source and target domain for the two techniques. In all experiments, we use the AdamW optimizer [303] with an initial learning rate of 10^{-2}, a weight decay of 10^{-3},

Algorithm 3 SIDDA optimization step

Require: x: Source domain inputs, $\hat{y}$: Source domain labels, x^*: Target domain inputs, n: batch size in each of the domains, η_1, η_2: Dynamic weighting parameters

while not converged **do**
  $\mathbf{X} \leftarrow [x_n,\, x_n^*]$ ▷ Concatenate inputs
  $\mathbf{Y},\, \mathbf{Z} \leftarrow \text{model}(\mathbf{X})$ ▷ Compute logits and latent features
  $z_n,\, z_n^* \leftarrow \mathbf{Z}$ ▷ Split features into source and target
  $y_n,\, y_n^* \leftarrow \mathbf{Y}$ ▷ Split predictions into source and target
  $D_{ij} \leftarrow \|z_i - z_j^*\|_2, \forall\, i, j \in \{1, \ldots, n\}$ ▷ Compute pairwise distances
  $\sigma_\ell \leftarrow \max(0.05 \times \max_{i,j} D_{ij},\, 0.01)$ ▷ Compute dynamic blur parameter
  $\mathcal{L}_{\text{DA}} \leftarrow \text{Sinkhorn}(z_n,\, z_n^*,\, \sigma_\ell)$ ▷ Compute Sinkhorn loss
  $\mathcal{L}_{\text{CE}} \leftarrow \text{CrossEntropy}(y_n,\, \hat{y}_n)$ ▷ Compute classification loss
  $\mathcal{L} \leftarrow \frac{1}{2\eta_1^2}\mathcal{L}_{\text{CE}} + \frac{1}{2\eta_2^2}\mathcal{L}_{\text{DA}} + \log(|\eta_1\eta_2|)$ ▷ Compute total loss
  loss.backward() ▷ Compute gradients
  clip_grad_norm_(model.parameters(), 10.0) ▷ Gradient clipping
  $\eta_1 \leftarrow \max(\eta_1,\, \text{1e-3})$ ▷ Clip η_1
  $\eta_2 \leftarrow \max(\eta_2,\, 0.25 \times \eta_1)$ ▷ Clip η_2
  optimizer.step() ▷ Update model parameters
end while

and a batch size of 128. A multiplicative learning rate decay of 0.1 is applied twice sequentially during training to stabilize convergence.

For all experiments, we use data augmentation comprising random rotations, flips, and affine translations. For the CNN, the augmentation instills approximate equivariance to encourage the model to learn rotation-invariant features. This is done to prevent predisposing the ENNs from performing better, but it also encourages faster convergence for the CNN. This later allows a more fruitful comparison between the latent distributions between ENNs, which have inherent rotation invariance to discrete rotations, and the CNNs, which have approximate invariance [301].

For experiments with DA, an initial warm-up phase is implemented, during which only classification tasks are trained (using only $\mathcal{L}_{\text{CE}}$). A similar 0.1 multiplicative learning rate decay as in the case of experiments without DA is also used. Model-saving criteria are based on the following: for experiments without DA, we use the best validation loss on the source domain for classification; for DA experiments, we use the sum of the validation classification loss on the source domain and validation DA loss. The warm-up phase is intended to predispose the model to be at an ideal location in the loss landscape before starting DA. Empirically, we found this beneficial. The duration of the warm-up phase was tuned for each experiment, ensuring that it was long enough that the models were performant on the source domain but short enough that there was no overfitting. It was also found that ENNs required a shorter warm-up phase than CNNs. For example, for the shapes dataset, a warm-up of five and ten epochs were used for the D_4 and the CNN models, respectively. For the GZ Evo dataset, the warm-up phase was 20 and 30 epochs for the D_4 and the CNN models, respectively. All models were trained for a maximum of 50 epochs for the shapes and astronomical objects datasets, whereas training was extended to a maximum of 100 epochs for the MNIST-M and GZ Evo datasets. All training was done on one NVIDIA A100-80GB GPU, with the most complex experiment requiring about one hour to train. Algorithm 3 presents the forward pass for training with SIDDA.

4.3.3 Calibration

Despite the impressive predictive capability of NNs in classification tasks across various fields, many real-world applications of NN-based classifiers also consider the confidence of each output class. Specifically, many NN-based classifiers can be uncalibrated, wherein the predicted class probabilities can frequently misrepresent the true class likelihood and lead to under- or overconfident predictions. In data-sensitive or safety-sensitive settings—such as medicine and

biology—proper model calibration is essential for deploying NN-based classifiers [304]. Similarly, in cosmology, simulation-based inference (SBI) pipelines that rely on trained classifiers must ensure proper calibration to guarantee that the inferred likelihood ratios or posterior probabilities are accurate and trustworthy [305].

Calibration techniques vary from regularization during training (either through architectural choices or additional loss terms) to post hoc methods that scale predicted probabilities (see [306] for a review). DA-based methods, however, have traditionally not been considered in the realm of regularization methods for model calibration. We will show that including DA with SIDDA not only improves accuracy (see Sections 4.4.1 and 4.4.2), but also calibration.

We evaluate model calibration using the Brier score and the Expected Calibration Error (ECE). The Brier score is the mean prediction error over all the classes:

$$\text{Brier Score} = \frac{1}{C}\sum_{i=1}^{C}(y_i - \delta_{i\hat{y}})^2, \tag{4.15}$$

where y_i is the NN-predicted score for each class $i \in C$, where C is the number of classes, $\hat{y}$ is the true class label, and $\delta_{i\hat{y}}$ is the Kronecker delta. Thus, a lower Brier Score is indicative of better calibration.

The ECE is the weighted average of the absolute difference between accuracy and confidence over V equally spaced confidence bins. For each bin B_v, where $v \in \{1, \ldots, V\}$, the accuracy and confidence are calculated based on the predictions within that bin. The ECE is defined as

$$\text{ECE} = \sum_{v=1}^{V}\frac{|B_v|}{W}\left|\text{acc}(B_v) - \text{conf}(B_v)\right|, \tag{4.16}$$

where $|B_v|$ is the number of samples in bin v, W is the total number of samples, $\text{acc}(B_v)$ is the average accuracy in bin B_v, and $\text{conf}(B_v)$ is the confidence (average estimated probability) in bin B_v. The ECE is thus a measure of how much model confidence aligns with the true distribution of classes, and a lower ECE indicates better calibration.

4.3.4 Neural Network Latent Distributions

The distributions over the source and target latent encodings, z and z^*, are the fundamental objects used in DA techniques. Probing the latent distributions can give crucial insights into the success and failure points of DA. The dimensionality of latent distributions is typically too large for visualization and analysis (256 in experiments used in this work). Therefore, dimensionality

reduction techniques are often employed before visualizing the latent distributions. Techniques like t-SNE and UMAP [307, 308] use local metrics, like pairwise distances or nearest-neighbor graphs to preserve the structure of the data at small scales while embedding it into a lower-dimensional space. However, local metrics, and therefore these techniques, are limited because they primarily focus on preserving relationships within small neighborhoods of the data, often at the expense of capturing global structures or long-range dependencies that are critical for understanding the overall geometry or topology of the dataset. In contrast, the isomap [309] is a non-linear dimensionality reduction technique that estimates the global geometry of a latent vector manifold by using information of the nearest neighbors for each point in the latent space.

In this work, we use isomaps to visualize latent distributions, and we use the mean Silhouette score to quantify the inter-class (between clusters) and intra-class (within a cluster) distances and evaluate the quality of the clustering. The silhouette score is

$$s = \frac{1}{Q}\sum_{i=1}^{Q}\frac{b(i)-a(i)}{\max(a(i),b(i))}, \tag{4.17}$$

where Q is the number of points; $a(i)$ is the intra-cluster distance, which is the mean distance between the i-th data point and all other points within the same cluster; and $b(i)$ is the inter-cluster distance, the mean distance between the i-th data point and points in the nearest neighboring cluster. The Silhouette score is in the range $[-1, 1]$, where values close to one indicate well-clustered data, values near zero indicate overlapping clusters, and negative values indicate that the data point may be assigned to the wrong cluster.

4.4 Results

All results are computed from three trained NNs, each with a different random seed for initializing weights. For each set of three trained networks, we estimate 1σ uncertainties on our diagnostic metrics. We refer to a model trained without DA (i.e., only with cross-entropy loss) as "<model>" — e.g., "CNN" or "D_4". In contrast, we refer to a model trained with SIDDA as "<model>-DA" — e.g., "CNN-DA" or "D_4-DA".

4.4.1 Simulated Datasets

The test set accuracies for the CNN, D_4, CNN-DA, and D_4-DA models on the shapes, astronomical objects, MNIST-M, and GZ Evo datasets are shown in Table 4.1. For the shapes and

Table 4.1: Classification accuracies for different model configurations on all datasets.

Dataset	**Metric**	**CNN**	**CNN-DA**	D_4	D_4**-DA**
Shapes	Source Acc. (%)	99.80 ± 0.04	$\mathbf{99.82 \pm 0.12}$	99.90 ± 0.04	$\mathbf{99.92 \pm 0.02}$
	Target Acc. (%)	50.47 ± 8.39	$\mathbf{78.20 \pm 1.73}$	64.76 ± 3.42	$\mathbf{99.71 \pm 0.06}$
Astro. Objects	Source Acc. (%)	$\mathbf{99.34 \pm 0.21}$	95.32 ± 1.57	$\mathbf{99.98 \pm 0.02}$	99.89 ± 0.50
	Target Acc. (%)	50.81 ± 2.89	$\mathbf{91.33 \pm 1.41}$	66.41 ± 2.07	$\mathbf{97.19 \pm 0.51}$
MNIST-M (Noise)	Source Acc. (%)	$\mathbf{95.64 \pm 0.12}$	95.31 ± 0.09	97.30 ± 0.30	$\mathbf{97.45 \pm 0.02}$
	Target Acc. (%)	68.32 ± 2.72	$\mathbf{76.24 \pm 1.12}$	70.31 ± 0.96	$\mathbf{87.55 \pm 0.16}$
MNIST-M (PSF)	Source Acc. (%)	95.64 ± 0.12	$\mathbf{95.66 \pm 0.13}$	97.30 ± 0.30	$\mathbf{97.95 \pm 0.10}$
	Target Acc. (%)	75.00 ± 1.44	$\mathbf{85.68 \pm 1.66}$	77.85 ± 1.31	$\mathbf{93.00 \pm 1.14}$
Galaxy Zoo Evo	Source Acc. (%)	81.49 ± 0.32	$\mathbf{81.57 \pm 0.82}$	86.65 ± 0.31	$\mathbf{87.58 \pm 0.06}$
	Target Acc. (%)	70.65 ± 2.26	$\mathbf{77.54 \pm 0.62}$	79.48 ± 1.52	$\mathbf{83.13 \pm 0.53}$
MRSSC2	Source Acc. (%)	$\mathbf{76.14 \pm 1.63}$	71.27 ± 0.85	$\mathbf{88.71 \pm 0.48}$	88.06 ± 0.57
	Target Acc. (%)	31.28 ± 3.78	$\mathbf{36.80 \pm 1.58}$	45.30 ± 4.32	$\mathbf{48.10 \pm 1.56}$

simulated astronomical objects dataset, both models achieve near-perfect accuracy (above 99%) on the source domain without DA, but classification accuracy is much lower on the target domain for both (between 50% and 66%). With the inclusion of DA, the largest increase in target domain accuracy of $\approx 40\%$ is for the astronomical objects datasets with CNN-DA. Despite this, the D_4 model significantly outperforms the CNN in the target domain, in the case of both datasets. With DA, the D_4-DA model achieves greater than 97% accuracy in both the source and target domains. While the CNN-DA shows a substantial improvement in target domain performance, the gap between the CNN-DA and D_4-DA remains considerable, with a difference of 21% in target domain accuracy for the shapes dataset.

We test generalization capabilities in the presence of Poisson noise and PSF blurring on the MNIST-M dataset. A similar trend emerges: the D_4 model is significantly more robust against both types of noise compared to the CNN. As in previous cases, with DA, neither model achieves the same performance as on the source domain (approximately 95% for CNN-DA and 97% for D_4-DA). However, the D_4-DA model demonstrates a greater potential for alignment than CNN-DA, achieving 93% accuracy in the target domain during the PSF blurring experiments.

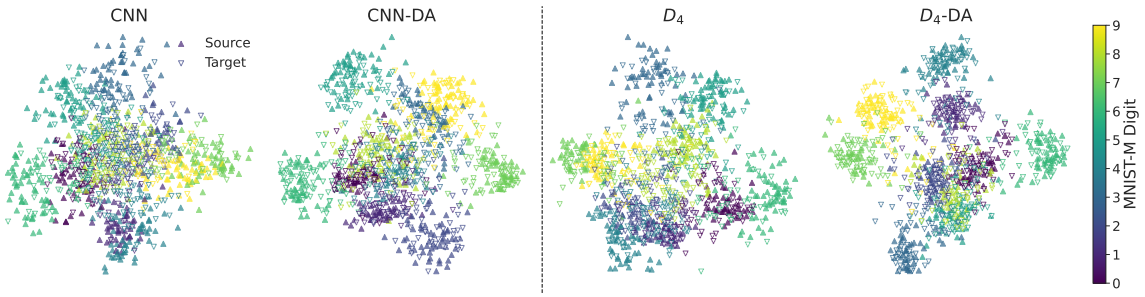


Figure 4.5: MNIST-M (Noise) latent distributions visualized with isomaps. Source (solid) and target (hollow) latent distributions are plotted atop each other to visualize latent distribution misalignment. The inclusion of DA clearly improves the alignment of source and target latent distributions for both CNN and D_4 models. It is also seen that the latent distribution of the D_4 is more clustered than the CNN, and even more so when D_4-DA and CNN-DA are compared. The improved clustering and separation of classes in the latent space is suggestive of improved feature learning.

The results show that SIDDA improves the source domain performance across most datasets for both CNN and D_4 models, except for the astronomical objects dataset for both models, and for the MNIST-M (noise) dataset for the CNN-DA model. In most cases, this drop in the source domain performance is below 1% (except in the case of astronomical objects classified with the CNN-DA model), which is acceptable given that the inclusion of DA substantially improves the accuracy on the unlabeled target domain. Additionally, the D_4-DA models converged to more similar performance across different seeds, as reflected by the small uncertainties in classification accuracy compared to CNN-DA.

We visualize the latent distributions of the CNN, D_4, CNN-DA, and D_4-DA in Figure 4.5 for the MNIST-M dataset with Poisson noise with the source (solid triangles) and target domain (hollow triangles) overlapped using isomaps [309]. Without DA, there is significant overlap between different classes, particularly in the middle of the figure for both models, though it is more apparent for the CNN. This is reflected in the CNN and D_4 Silhouette scores (Table 4.2), indicating similar levels of misclassified points, as seen in the target domain accuracy of approximately 68% for CNN and 70% for D_4. With the inclusion of DA, the same classes from the source and target latent distributions for both models are better aligned. Furthermore, with DA, there is increased class separation, especially along the periphery, but there is still some overlap in the middle of the diagram (particularly for CNN-DA). Compared to the CNN latent distribution, the class separation for the D_4 latent distribution is much larger due to the equivariance in the D_4 model, which is

reflected in the target domain Silhouette scores increasing more significantly for the D_4-DA when compared to the CNN-DA.

Model	**Source**	**Target**
CNN	0.1930	-0.0539
CNN-DA	0.3360	0.1150
D_4	0.2744	-0.0247
D_4-DA	**0.4023**	**0.1983**

Table 4.2: Silhouette scores for CNN, D_4, CNN-DA, and D_4-DA on MNIST-M (Noise).

We lastly study the evolution of the trainable loss coefficients η_1 and η_2, as well as the regularization strength of the Sinkhorn plan σ_ℓ during training. We show these parameters for the CNN-DA model trained on MNIST-M (Noise) in Figure 4.6; the overall behavior was typically seen in all model training. After the $\mathcal{L}_{\text{CE}}$-only warm-up phase, η_2^{-2} (blue) quickly becomes greater than η_1^{-2} (red), providing a larger weight for $\mathcal{L}_{\text{DA}}$. As training progresses, both η_1^{-2} and η_2^{-2} increase commensurately, indicating that a constant parameterization of loss coefficients throughout training is indeed not optimal, and that model convergence was aided by treating the coefficients as trainable parameters. The model convergence is also stable across seeds used during training, as indicated by the small shaded regions for η_1 and η_2 in the figure. Upon looking at the loss curves, $\mathcal{L}_{\text{DA}}$ (without being weighted by η_2^{-2}) is roughly an order of magnitude smaller than $\mathcal{L}_{\text{CE}}$. After subsequent weighting of both terms with $1/2\eta_1^2$ and $1/2\eta_2^2$, the relative contribution of $\mathcal{L}_{\text{DA}}$ is still smaller than $\mathcal{L}_{\text{CE}}$; however, they are now roughly the same order of magnitude. This allows the model to still prioritize the primary learning task (classification), while actively aligning the latent distributions with $\mathcal{L}_{\text{DA}}$. This was enforced by the clipping procedure for η_2 as shown in Algorithm 3, where the ratio must satisfy $\eta_2/\eta_1 \geq 0.25$.

We additionally study the evolution of η_1^{-2} and η_2^{-2} for the same experiment without enforcing the ratio for η_2/η_1. These are shown in the more transparent lines with the same corresponding colors in Figure 4.6. We see that without clipping, η_2^{-2} generally undergoes a sharp increase at the very beginning. Both η values show a larger variance across seeds, with η_2 exhibiting significant fluctuations, as indicated by the more transparent blue shaded area in Figure 4.6. By the end of training, the unclipped η values both converge to similar values as their clipped counterparts. This indicates that the clipping enforced in SIDDA stabilizes the evolution of these trainable loss coefficients.

We additionally see in Figure 4.6 that σ_ℓ gradually decreases during training as the NN latent distributions continue to become more closely aligned. With SIDDA, S_σ initially interpolates more closely to MMD at the start of training, while by the end, it approaches the minimum allowed $S_{0.01}$, as shown in Equation 4.6.

4.4.2 Galaxy Zoo Evo Dataset

We aim to address a common problem in real astronomical applications: the scenario where one has access to lower-quality, older observations with labels, but no labels exist for the newer, higher-quality target dataset from a more recent astronomical survey. To test our model in this situation, we use galaxy datasets from GZ Evo [287]. Namely, we use GZ2 from SDSS as the source domain and GZ DESI as the target domain. This dataset is considerably larger than the previous simulated datasets. The results for these experiments are summarized in Table 4.1. Similar to the results for the simulated datasets, the D_4 model fully outperforms the CNN, with a $\approx 9\%$ higher accuracy in the target domain. With DA, both CNN-DA and D_4-DA models have higher accuracy in both the source and target domains. The performance difference between the two kinds of models is more moderate for the GZ Evo dataset than for the simulated datasets. This may be attributable to the GZ Evo data's larger size, greater morphological complexity, and the use of data augmentation during training. In the large data limit, the inclusion of data augmentation causes CNNs to become approximately equivariant [298, 299]. Nevertheless, the D_4-DA achieves $\sim 6\%$ higher accuracy in the target domain compared to the CNN-DA model.

The accuracy of all models in this experiment was lower than in the experiments with the simulated data, with no model achieving greater than $\approx 88\%$ accuracy in either the source or target domain. This dataset is significantly larger than the others, so a deeper or larger model with additional training aids, such as residual connections, would likely yield better performance [101, 310]. Our goal is to study the efficacy of SIDDA, and not to achieve state-of-the-art performance on this dataset, so we did not experiment with a more complex model.

4.4.3 Robustness with Group Order

Next, we study the robustness of ENNs, both with and without SIDDA, and with increasing orders of the dihedral group D_N, with $N \in \{1, 2, 4, 8\}$ on MNIST-M with Poisson noise in the target domain. The robustness of ENNs as a function of group order was previously studied in [92] for generalization to Poisson noise in images in the task of galaxy morphology classification. That

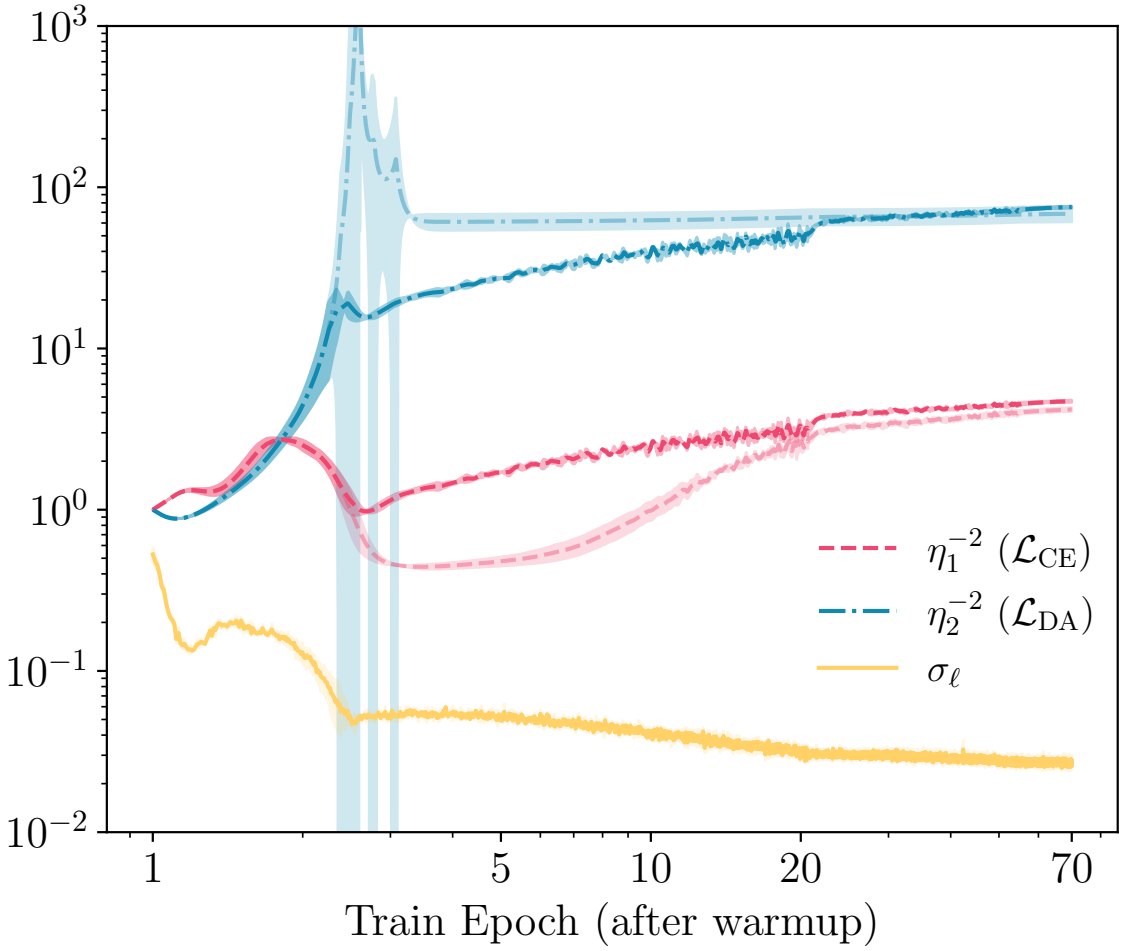


Figure 4.6: Evolution of the trainable coefficients, η_1^{-2} and η_2^{-2}, and the Sinkhorn plan regularization strength σ_ℓ for CNN-DA trained on MNIST-M (Noise) after the $\mathcal{L}_{\mathrm{CE}}$-only warm-up period. The shaded regions correspond to 1σ uncertainties from three training runs initialized with varying random seeds. With parameter clipping, as indicated in Algorithm 3, the time-evolution of both η's is more stable, as indicated by the darker lines and their lower variance (i.e., narrower shaded regions). Without parameter clipping (more transparent η curves), the evolution of both η parameters is more unstable with different random seeds, as indicated by the larger shaded regions and sharp increase of η_2^{-2} around epoch three. Still, both η's (with and without clipping) arrive at similar final values. It is also seen that the regularization strength of the Sinkhorn plan σ_ℓ continually decreases during training, as the NN latent spaces gradually become more aligned. This corresponds to the Sinkhorn plan more closely behaving as MMD at the beginning of training, while approaching the minimum allowed $S_{0.01}$ by the end.

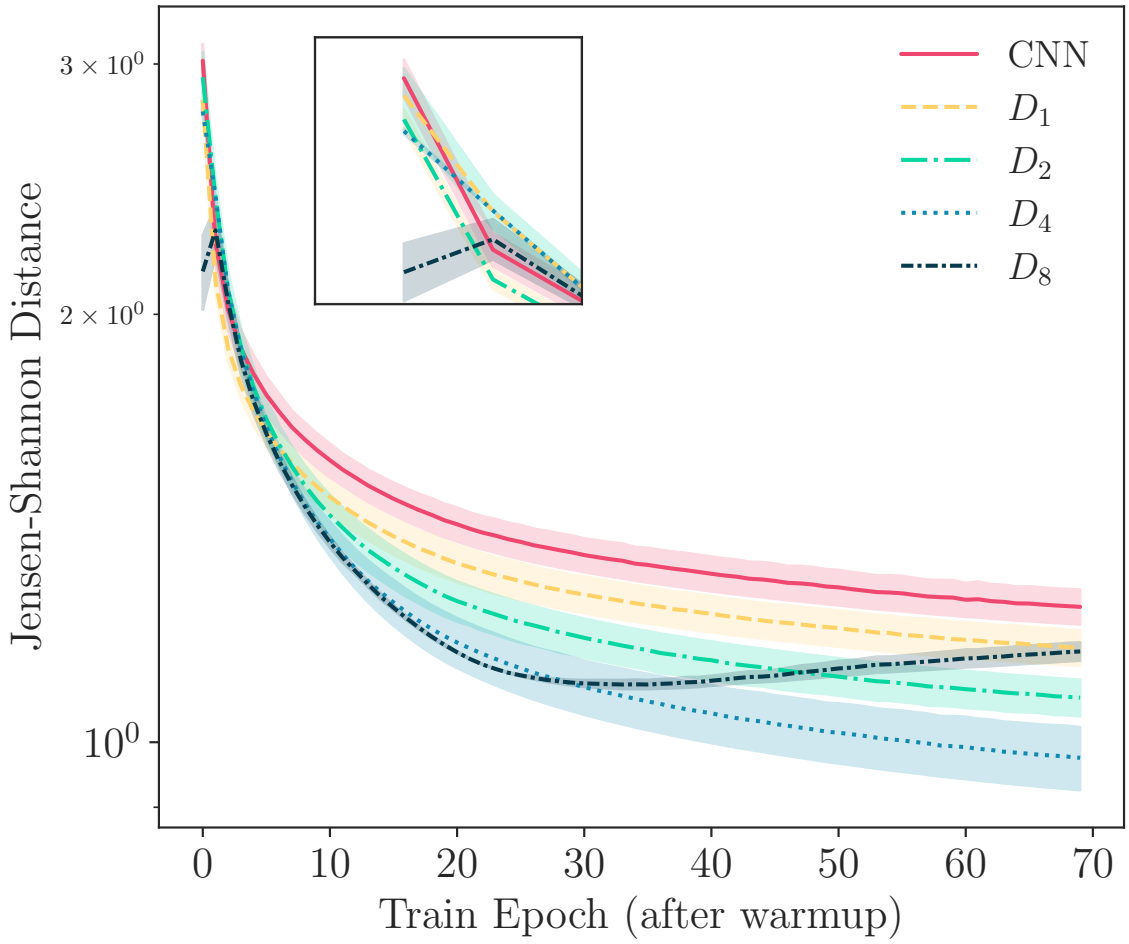


Figure 4.7: Jensen-Shannon (JS) distances for CNN-DA and D_N-DA ($N \in \{1, 2, 4, 8\}$) models trained on MNIST-M (Noise). Shaded regions correspond to 1σ uncertainties from three training runs initialized with varying random seeds. All models underwent a 30-epoch warm-up phase without DA, and the JS distance is shown for epochs thereafter, which is where SIDDA is used. Compared to the CNN, all D_N models exhibit a lower JS distance between source and target domains after the warm-up phase, which can be attributed to the fact that the equivariance constraint encourages distributional similarity between the source and target latent distributions. We see that the D_N models also achieve more perfect alignment with the introduction of DA, as shown by the lower JS distances by the end of the training. This behavior correlates with the group order, except for D_8-DA, which achieved its best model much earlier in training and began overfitting.

Table 4.3: Performance results for different orders of the dihedral group D_N, without and with DA.

Group	**Source Domain**	**Target Domain**
D_1	$96.03 \pm 0.18\%$	$63.69 \pm 0.61\%$
D_2	$97.06 \pm 0.048\%$	$67.88 \pm 2.5\%$
D_4	$97.30 \pm 0.28\%$	$70.30 \pm 0.97\%$
D_8	$97.42 \pm 0.10\%$	$71.67 \pm 0.28\%$
D_1-DA	$95.40 \pm 0.084\%$	$75.35 \pm 0.71\%$
D_2-DA	$97.10 \pm 0.23\%$	$84.98 \pm 1.9\%$
D_4-DA	$97.50 \pm 0.074\%$	$87.70 \pm 0.26\%$
D_8-DA	$\mathbf{97.69 \pm 0.081\%}$	$\mathbf{88.96 \pm 0.32\%}$

work showed that robustness generally increased with group order, but high group-order models tended to overfit or have lower accuracy as a result of the equivariance constraint becoming too strong (see Figure 2 in [92]). Additionally, despite the robustness, there was no significant overlap in the latent distribution of the NNs when introduced to covariate shifts in the data. In the current work, we add to that previous study by examining the effect of SIDDA on the robustness of ENNs across different group orders with the same data setup.

We follow a training procedure similar to that outlined in Section 4.3.2 for all networks. Models are still saved based on the best validation loss, and we show results from the best-performing models on the held-out test set. Source and target domain classification accuracies are shown in Table 4.3. Both source and target accuracies increase with group order, with and without DA. There is also a slight increase in source domain accuracy when using DA, except for D_1, which contains a single reflection and therefore exhibits trivial equivariance.

We also track the evolution of the mean JS distance between source and target latent distributions for all models during training (after the initial warm-up phase has passed). Assuming perfect feature learning and optimal performance on the source domain, minimizing the JS distance between source and target domain latent distributions corresponds to optimal performance on the target domain [285]. As shown in Figure 4.7, after the initial 30-epoch warm-up phase, the JS distance decreases as the group order increases. Among the models, the D_8-DA model exhibits the lowest JS distance at the end of the warm-up, while the CNN-DA exhibits the highest. This supports the claim that the constrained latent distribution of ENNs leads to significantly better alignment between source and target domains, as made in Section 4.1.4.

As training progresses with DA, the JS distance generally decreases with increasing group order, except for the D_8 model. This model begins to overfit 30 epochs after the warm-up phase concludes, as indicated by the JS distance in Figure 4.7. This was also confirmed upon inspection of the validation loss. All other models reach their highest validation loss after epoch 90. Despite this overfitting, the best D_8-DA model still achieves the highest source and target domain accuracy, as shown in Table 4.3.

The overfitting observed in the D_8 model can be attributed to the stronger equivariance constraint (i.e., more weight sharing), which may limit the model's expressivity when the covariate shifts in the target domain do not fully respect the underlying symmetry. That is, the allowable space of features that respects the stronger equivariance will be inherently smaller than those respected by more lenient equivariance or the typical translation equivariance in CNNs, considering that ENNs assume perfect symmetries in the data. In the presence of perturbations, which is a typical case in the target domain for DA applications, this rarely holds. Solutions to relaxing the equivariance constraint while still enjoying the benefits of symmetry constraints have been extensively studied in other works [311, 312].

4.4.4 Model Calibration

For all datasets, we observe that the D_4 and D_4-DA models result in lower ECE and Brier scores across most experiments in both the source and target domains when compared to the CNN and CNN-DA. Across all experiments, the largest improvement is observed for the D_4-DA model applied to the shapes target domain data, where both the ECE and Brier score are reduced by more than an order of magnitude compared to the regular D_4 model. Furthermore, we see that in most experiments, the classification accuracies as given in Table 4.1 are correlated with the calibration metrics in Table 4.4 in both the source and target domains. That is, as classification accuracy in either the source or target domain increases, model calibration improves.

With the inclusion of DA, there is an improvement in the calibration scores for both data domains and for all models for the MNIST-M datasets, as well as GZ Evo data. Between these datasets, the largest improvement is for the D_4-DA model on the MNIST-M (PSF) dataset in the target domain, exhibiting an approximate factor of two reduction for both the ECE and Brier score. However, at the same time, we observe a decrease in the source domain calibration for all DA experiments in the shapes and astronomical objects datasets, except for D_4-DA in the shapes dataset.

Table 4.4: Calibration metrics (expected calibration error and Brier score) for different model configurations.

Metric	CNN	CNN-DA	D_4	D_4-DA
Shapes				
Source ECE	$\mathbf{0.011 \pm 0.001}$	0.013 ± 0.001	0.011 ± 0.002	$\mathbf{0.0074 \pm 0.0003}$
Source Brier	$\mathbf{0.000734 \pm 0.000090}$	0.00112 ± 0.00020	0.000814 ± 0.000200	$\mathbf{0.000349 \pm 0.000034}$
Target ECE	0.35 ± 0.04	$\mathbf{0.29 \pm 0.004}$	0.20 ± 0.03	$\mathbf{0.013 \pm 0.002}$
Target Brier	0.110 ± 0.010	$\mathbf{0.0925 \pm 0.002}$	0.0564 ± 0.009	$\mathbf{0.0015 \pm 0.0003}$
Astronomical Objects				
Source ECE	$\mathbf{0.041 \pm 0.010}$	0.075 ± 0.020	$\mathbf{0.00695 \pm 0.00030}$	0.00899 ± 0.00090
Source Brier	$\mathbf{0.00798 \pm 0.003}$	0.0220 ± 0.006	$\mathbf{0.000132 \pm 0.000031}$	0.000746 ± 0.000300
Target ECE	0.17 ± 0.04	$\mathbf{0.142 \pm 0.010}$	0.294 ± 0.020	$\mathbf{0.053 \pm 0.008}$
Target Brier	0.0440 ± 0.010	$\mathbf{0.0420 \pm 0.006}$	0.0804 ± 0.009	$\mathbf{0.0150 \pm 0.003}$
MNIST-M (Noise)				
Source ECE	0.161 ± 0.003	$\mathbf{0.126 \pm 0.004}$	0.114 ± 0.002	$\mathbf{0.0790 \pm 0.002}$
Source Brier	0.00991 ± 0.00023	$\mathbf{0.00880 \pm 0.00030}$	0.00610 ± 0.00030	$\mathbf{0.00481 \pm 0.00010}$
Target ECE	0.409 ± 0.013	$\mathbf{0.355 \pm 0.009}$	0.390 ± 0.020	$\mathbf{0.250 \pm 0.009}$
Target Brier	0.0450 ± 0.003	$\mathbf{0.0370 \pm 0.002}$	0.0410 ± 0.0008	$\mathbf{0.0210 \pm 0.0031}$
MNIST-M (PSF)				
Source ECE	0.161 ± 0.003	$\mathbf{0.124 \pm 0.004}$	0.114 ± 0.002	$\mathbf{0.0750 \pm 0.002}$
Source Brier	0.00991 ± 0.00023	$\mathbf{0.00850 \pm 0.00040}$	0.00610 ± 0.00030	$\mathbf{0.00400 \pm 0.00020}$
Target ECE	0.384 ± 0.013	$\mathbf{0.272 \pm 0.012}$	0.340 ± 0.020	$\mathbf{0.181 \pm 0.001}$
Target Brier	0.0340 ± 0.001	$\mathbf{0.0230 \pm 0.001}$	0.0270 ± 0.003	$\mathbf{0.0130 \pm 0.00007}$
Galaxy Zoo Evo				
Source ECE	0.283 ± 0.0019	$\mathbf{0.264 \pm 0.0044}$	0.2322 ± 0.00086	$\mathbf{0.206 \pm 0.0011}$
Source Brier	0.0453 ± 0.00031	$\mathbf{0.0439 \pm 0.0010}$	0.0341 ± 0.00059	$\mathbf{0.0319 \pm 0.00014}$
Target ECE	0.324 ± 0.0078	$\mathbf{0.301 \pm 0.0018}$	0.271 ± 0.0042	$\mathbf{0.241 \pm 0.0026}$
Target Brier	0.0538 ± 0.0015	$\mathbf{0.051 \pm 0.00029}$	0.0411 ± 0.0012	$\mathbf{0.0382 \pm 0.00048}$
MRSSC2				
Source ECE	$\mathbf{0.331 \pm 0.00600}$	0.407 ± 0.0121	$\mathbf{0.250 \pm 0.0152}$	0.304 ± 0.0159
Source Brier	$\mathbf{0.0409 \pm 0.00113}$	0.0615 ± 0.00247	$\mathbf{0.0243 \pm 0.00178}$	0.0328 ± 0.00312
Target ECE	$\mathbf{0.422 \pm 0.00760}$	0.446 ± 0.00418	$\mathbf{0.407 \pm 0.0212}$	0.452 ± 0.0453
Target Brier	$\mathbf{0.0584 \pm 0.00248}$	0.0975 ± 0.0000711	$\mathbf{0.0505 \pm 0.00510}$	0.0656 ± 0.0141

This indicates that in the source domain, at least for simpler datasets, the inclusion of DA can potentially worsen the calibrations of the models. This decrease in performance is not always apparent in the (uncalibrated) accuracies, as shown in Table 4.1. For instance, the source accuracies with and without DA are very similar (within the margin of error) in all cases, except for the CNN-DA model applied to the astronomical objects dataset, where the source domain accuracy drops around 4% with the inclusion of DA. Nonetheless, in the target domain, we observe strict improvements in both accuracy and calibration across all experiments with the inclusion of DA with SIDDA, as shown in Table 4.1 and Table 4.4.

4.4.5 Method Comparisons

Next, we use the MNIST-M (Noise) dataset to compare SIDDA to MMD and the Wasserstein distance in terms of accuracy, calibration, and computational efficiency (see Table 4.5). The latter two methods are distance-based and represent limiting cases of the Sinkhorn divergence. We use a Gaussian MMD with a fixed kernel width of $\epsilon = 0.05$ and approximate the true Wasserstein distance as the Sinkhorn divergence in the limit that $\sigma \to 0$, with a fixed regularization of $\sigma = 10^{-12}$. All models were trained for 100 epochs using identical training hyperparameters (see [106]), model saving criteria, and hardware (NVIDIA A100 GPU).

SIDDA outperforms both MMD and the Wasserstein distance in terms of target domain accuracy, ECE, and the Brier score, achieving the best overall performance. Specifically, it yields an approximate 1% improvement over the Wasserstein distance and about 15% over MMD in target domain accuracy. The Wasserstein D_4 model achieves a 97.56% accuracy on the source domain, almost within error bars of the SIDDA-D_4 model, which has a 97.45% accuracy.

SIDDA achieves the best calibration among all methods, including on the source domain, despite the Wasserstein distance performing slightly better in that setting. Overall, models based on the Wasserstein distance demonstrate the next best calibration, followed by those using MMD, as evidenced by the ECE and Brier scores. In all cases, the D_4 models achieve better calibration than the CNN models.

The wall clock time of SIDDA is similar to that of MMD; both models require < 10 minutes to train. The Wasserstein distance models require much more time — e.g., 24 minutes for the Wasserstein-D_4 model. This can be attributed to the slower convergence for Sinkhorn iterations as the entropic regularization approaches zero ($\sigma \to 0$).

Table 4.5: Performance comparison of SIDDA with Gaussian MMD and Wasserstein distance DA methods on MNIST-M (Noise).

Method	**Model**	**Accuracy (%)**		**Expected Calibration Error (ECE)**		**Brier Score**		**Time (min)**
		Source	**Target**	**Source**	**Target**	**Source**	**Target**	
SIDDA	CNN	95.31 ± 0.09	76.24 ± 1.12	$0.126 \pm .004$	$0.355 \pm .009$	$0.00880 \pm .00030$	0.0370 ± 0.002	8
	$\boldsymbol{D_4}$	97.45 ± 0.02	$\mathbf{87.55 \pm 0.16}$	$\mathbf{0.0790 \pm .002}$	$\mathbf{0.250 \pm 0.009}$	$\mathbf{0.00481 \pm 0.00010}$	$\mathbf{0.0210 \pm 0.0031}$	10
MMD	CNN	95.60 ± 0.18	68.90 ± 1.52	0.164 ± 0.00686	0.409 ± 0.00778	0.0102 ± 0.000523	0.0459 ± 0.00172	5
	$\boldsymbol{D_4}$	97.25 ± 0.04	72.94 ± 1.97	0.122 ± 0.00134	0.382 ± 0.000297	0.00656 ± 0.000112	0.0390 ± 0.000462	7
Wasserstein	CNN	94.83 ± 0.19	75.54 ± 1.13	0.143 ± 0.00466	0.371 ± 0.00955	0.0100 ± 0.000456	0.0389 ± 0.00152	19
	$\boldsymbol{D_4}$	$\mathbf{97.56 \pm 0.05}$	86.62 ± 0.24	0.0938 ± 0.00242	0.273 ± 0.00693	0.00552 ± 0.000141	0.0232 ± 0.000752	24

4.4.6 Comparison with Fixed Loss Coefficients

Table 4.6: Source and target domain accuracies for different loss formulations for the CNN-DA model on MNIST-M (Noise).

Loss Formulation	Source Acc. (%)	Target Acc. (%)
$\mathcal{L}_{\text{train}}$	95.31 ± 0.09	$\mathbf{76.24 \pm 1.12}$
$\mathcal{L}_{\text{C,D}}$	95.35 ± 0.13	72.35 ± 1.08
$\mathcal{L}_{\text{C,10D}}$	94.96 ± 0.18	74.88 ± 1.00
$\mathcal{L}_{\text{10C,D}}$	$\mathbf{95.61 \pm 0.28}$	71.69 ± 0.92

Next, we compare SIDDA with trainable loss coefficients η_i (which in this experiment we will denote as $\mathcal{L}_{\text{train}}$) with fixed loss coefficients by examining three losses, each associated with a different fixed coefficient. Respectively, these put equal emphasis on each loss, favor the DA loss, and favor the classification loss:

$$\mathcal{L}_{\text{C,D}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DA}} \tag{4.18}$$

$$\mathcal{L}_{\text{C,10D}} = \mathcal{L}_{\text{CE}} + 10 \cdot \mathcal{L}_{\text{DA}} \tag{4.19}$$

$$\mathcal{L}_{\text{10C,D}} = 10 \cdot \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DA}}. \tag{4.20}$$

For this set of experiments, we train the CNN models on MNIST-M (Noise). All models were trained with hyperparameters as detailed in [106], were saved according to the best validation loss, and the best model results are communicated in Table 4.6.

Table 4.6 compares SIDDA ($\mathcal{L}_{\text{train}}$) with the previously described loss formulations. The highest source domain accuracy is achieved with $\mathcal{L}_{\text{10C,D}}$, whereas the highest target domain accu-

racy corresponds to $\mathcal{L}_{\text{train}}$. Still, $\mathcal{L}_{\text{train}}$ source domain accuracy is in 1σ agreement with that achieved with $\mathcal{L}_{\text{10C,D}}$ and exhibits a smaller error bar, suggesting more stable training. $\mathcal{L}_{\text{train}}$ also achieves better performance in the target domain when compared with $\mathcal{L}_{\text{C,D}}$, and with the DA-favored loss ($\mathcal{L}_{\text{C,10D}}$). As evidenced by the high performance in both source and target domains, SIDDA ($\mathcal{L}_{\text{train}}$) overall performs better than any of the models with fixed loss coefficients. All models exhibit an approximate 1% uncertainty in target domain accuracy, while SIDDA has a slightly larger uncertainty in the target domain; this may be due to the trainable nature of the coefficients. Additional model initializations and experiments are necessary to draw more definitive conclusions regarding the stability of dynamic loss weighting.

The optimal choice of fixed loss coefficients will change for each dataset, problem, and type of loss. A poor choice of fixed loss coefficients can lead to models that perform poorly in the target domain when the DA loss is too small, or completely fail to learn the main objective when the DA loss is too large and prevents minimization of the main task loss. Additionally, suboptimal performance in both domains can occur if the DA loss is introduced at the wrong time during training. Our findings demonstrate that the approach introduced by Kendall et al. [216] allows one to avoid manual hyperparameter tuning for loss coefficients in DA tasks, while still achieving strong predictive performance on both source and target domains. This is particularly important for problems with very large datasets or complex models, where longer training times may be required.

4.4.7 Application to Severe Covariate Shifts

We have thus far demonstrated the efficacy of SIDDA across a range of datasets and covariate shifts—arising from factors such as varying noise levels, blurring, or differences in image quality (e.g., images captured by different telescopes within the same wavelength range). More substantial covariate shifts can arise when the source and target datasets include images captured at different wavelengths. We study the efficacy of SIDDA on this type of covariate shift using the MRSSC2 dataset. We use optical observations as the source domain and SAR (which operates in cm wavelengths, i.e., microwave/radio range) images as the target domain.

As shown in Table 4.1, the efficacy of SIDDA on the MRSSC2 dataset is considerably lower than the other datasets studied. In both the CNN-DA and D_4-DA models, the target domain performance increases (up to $\sim 5\%$). However, in both cases, the target domain performance is considerably worse than the source domain, even with the inclusion of SIDDA. Furthermore, even this small increase in target domain performance with DA causes a decrease in source domain

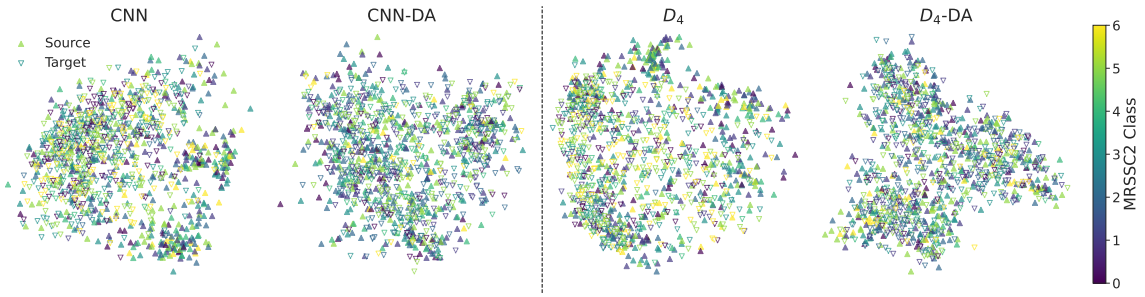


Figure 4.8: MRSSC2 latent distributions, visualized using isomaps, with the source domain shown as solid markers and the target domain as hollow markers. Both the CNN and D_4 models exhibit substantially less clustering in the latent space—compared to the MNIST-M (Noise) dataset shown in Figure 4.5—which may account for the modest performance gains achieved by DA on this dataset.

accuracy. The source domain performance of the CNN-DA model is less than the CNN model; similarly, the D_4-DA model performance is lower than the D_4 model, but the accuracy drop is much lower (less than 1%).

We use isomaps to visualize the latent distributions of the CNN, D_4, CNN-DA, and D_4-DA models in Figure 4.8 for the MRSSC2 dataset, with the source (solid triangles) and target domain (hollow triangles). Without DA, there is no well-defined clustering in any of the latent distributions, which can be indicative of poor feature learning. There is also significant misalignment between the source and target latent distributions. With DA, the misalignment decreases, but the latent space representations of different classes are not as defined as previously seen with well-performing SIDDA models (see MNIST-M (Noise) latent distributions in Figure 4.5). The poor feature learning even without DA may explain the modest performance improvement with DA on the MRSSC2 dataset, as shown in Table 4.1.

In addition, the inclusion of SIDDA does not improve model calibration for both models on the MRSSC2 dataset, as seen in Table 4.4. Even in the target domain, where SIDDA somewhat increased classification accuracy, the calibration decreases. This is in contrast with what was observed with our other datasets, where the domain alignment was successful, and the ECE and Brier scores generally improved for the CNN-DA and D_4-DA models.

The sub-optimal performance on the MRSSC2 dataset might be the result of larger differences between classes when observed at different wavelengths. In this case, features extracted from the source domain images may be much less applicable to the target domain, which causes a considerably larger difference between source and target domain accuracy for the CNN and D_4

models (trained without DA) compared to other datasets studied. These feature learning differences can then make domain alignment of latent representations near the end of the architecture — as done here — unfeasible.

For distance-based DA approaches like SIDDA, this sub-optimality can potentially be addressed by intermediate domain alignment, similar to Deep Adaptation Networks (DANs) [313] and Joint Adaptation Networks (JANs) [314], and as indicated by results presented in Liu et al. [288]. This would then impose domain alignment at all stages of the architecture, from the early convolutional layers, which perform general feature learning, to the later convolutional and linear layers, which learn more detailed and task-specific features. Additionally, one could approach more extreme domain shift problems with adversarial methods such as Domain Adversarial Neural Networks (DANNs) [286], which are more flexible and avoid explicit calculation of any distance metric. Finally, using more complex networks could enable better feature learning, leading to better clustering of different classes compared to what we see in Figure 4.8. This would enable easier domain alignment, potentially improving the efficacy of SIDDA when applied to the MRSSC2 dataset.

4.5 Summary & Discussion

In this work, we introduced SIDDA — a semi-supervised DA approach that leverages principled methods for optimal alignment of NN latent spaces and training with multiple loss terms. This is in contrast to most DA applications, which face the challenge of requiring extensive hyperparameter tuning, making training NNs time-consuming, expensive, or unfeasible. SIDDA is an "out-of-the-box" DA method that is applicable in various domains, and it requires labeled training data only in the source domain, not in the target domain.

Our method relies on the Sinkhorn divergence, a symmetrized variant of regularized OT distances that corrects for the bias in OT_σ. In particular, our method dynamically adjusts the strength of regularization on a per-epoch basis, dependent on the pairwise distance between the source and target latent distribution vectors. In addition, we use dynamic weighting of the cross-entropy and DA loss terms on a per-epoch basis, which ensures a balance in training and improves predictive performance in both the source and target domains.

We test our method on shapes and astronomical objects datasets simulated using `DeepBench`, the MNIST-M dataset, a dataset of real galaxy images from Galaxy Zoo, and the MRSSC2 remote sensing dataset. These experiments encompass covariate shifts induced by Poisson noise, PSF

blurring, differences between real optical astronomical surveys, and differences between images observed at different wavelengths.

We draw the following conclusions about SIDDA:

- SIDDA requires minimal hyperparameter tuning to achieve a considerable increase in target domain performance, and can outperform traditional fixed-hyperparameter use cases, as shown in Table 4.6;
- SIDDA combines the performance of the Wasserstein distance with the efficiency of MMD, shown through comparisons presented in Table 4.5;
- SIDDA is compatible with various models, including CNNs and ENNs, which are equivariant to different groups. Its efficacy is more pronounced when paired with ENNs, offering in some cases nearly 50% better target domain accuracy when compared to CNNs trained without DA (Table 4.1);
- We find that SIDDA is effective across ENNs with varying group-order equivariance, and its performance improves as the degree of equivariance increases (Table 4.3 and Figure 4.7);
- SIDDA can improve the source and target domain performance of NNs, and its benefits are concretely seen when analyzing the clustering and alignment of NN latent spaces (Table 4.2 and Figure 4.5);
- Though not constructed as such, SIDDA can inherently improve the calibration of trained NN-based classifiers (Table 4.4);
- SIDDA does not incur any considerable computational expense when training models, as it builds upon existing, efficient coding frameworks. All models in this work were trained on one GPU in typically less than an hour.

There are multiple opportunities for further development of SIDDA. First, the metric used for adjusting the dynamic Sinkhorn plan S_σ relies on the pairwise norm between entries in the latent space. Other notions of distance to adjust the regularization of S_σ can be considered. Second, in this work, we implemented a manual clipping of the η_i terms in the loss function (Equation 4.5) to ensure that the DA loss does not overpower the classification loss. Other levels of clipping or regularization of the loss should be explored to determine the optimal balance between the two loss terms. Third, all experiments performed here utilize a fixed architecture that incorporates several features of NNs

now considered standard, including dropout, pooling, and batch normalization. Notably missing are residual connections, which have been found to aid the convergence of many NNs. It would be interesting to study the efficacy of SIDDA with deeper, more complex networks, which could be particularly beneficial in cases of severe domain shift, such as in the MRSSC2 dataset, as well as ENNs equivariant to the continuous analog of groups studied here (i.e., $\mathrm{O}(2)$). Lastly, our method works with fixed classes and cannot operate when the classes between the source and target domains are not the same. A potential extension of SIDDA could involve making it compatible with a flexible number of classes in both the source and target domains, drawing inspiration from the DeepAstroUDA method [252].

For more extreme covariate shifts, a multi-layer approach to SIDDA as described in Section 4.4.7 could result in better domain alignment. In this setting, one could impose domain alignment after intermediate convolutional layers with a similar dynamic scaling of σ_ℓ as used here. These can then be aggregated into a single DA loss term, or be treated separately as additional loss terms $\mathcal{L}_i$ paired with trainable coefficients η_i. This is a promising area of future work.

The problem of generalization in classification tasks in NNs can be primarily considered a problem of robust feature learning (e.g., architectural choice) and domain alignment. The most successful domain adaptation methods should leverage principled choices for both aspects. ENNs are natural candidates for robust feature learning because their feature learning capabilities can be inherently constrained to symmetries of the data. However, most existing methods for domain adaptation implicitly require many empirical choices or hyperparameter tuning. In this work, we have combined these aspects in introducing SIDDA, which leverages a dynamic parameterization for OT-based DA hyperparameters during training, and works particularly well when paired with ENNs. Our future work will be in refining this approach and further developing more automated DA algorithms.

Chapter 5

Conclusion

In this dissertation, we have developed and demonstrated a range of machine-learning-assisted methods that bridge different epochs of cosmology, from the early Universe's cosmology to the challenges of modern astronomical surveys. Our results highlight how integrating differentiable simulations with AI techniques can enhance inference across the cosmic timeline. Here, we summarize our key findings and the broader implications of this work.

In Chapter 2, we investigated the phenomenon of cosmological stasis, an early-universe epoch during which the relative abundances of matter and radiation remain nearly constant despite cosmic expansion. We constructed a fully differentiable Boltzmann solver to simulate the dynamics of multi-component cosmologies and employed gradient-based methods (including automatic differentiation) to explore this high-dimensional parameter space. By optimizing the initial conditions (decay rates Γ_ℓ and abundances $\Omega_\ell^{(0)}$ of particle species) for maximal stasis duration, we discovered a new exponential solution that produces significantly longer stasis epochs than previously known power-law models. Notably, our results show that for an ensemble of N decaying species, an exponential hierarchy of decay rates can yield a stasis lasting $\sim N$ e-folds, vastly outlasting the $\sim \log N$ scaling of the power-law constructions.

We validated this finding via both gradient-based analyses and Bayesian posterior sampling, using SVI with normalizing flows to handle the high-dimensional inference. The exponential stasis model was found to act as an attractor in parameter space, similar to previous power-law constructions. This model was also found to be robust, with random draws from suitably chosen log-uniform priors exhibiting extended epochs of approximate stasis. Physically, these results broaden our understanding of exotic early-universe physics, with implications for existing BSM frameworks such as string theory. This insight may inform theoretical frameworks such as the string axiverse or

emergent string scenarios, where many light scalar fields are present.

More broadly, this work lays a foundation for the use of numerical and gradient-based optimization techniques in theoretical contexts, where analytic calculations have historically prevailed. The contributions of this chapter were not possible without these modern analysis tools. Moreover, we illustrated the utility of applying Bayesian inference in settings where it is not traditionally invoked—namely, theoretical work without direct observational data.

In Chapter 3, we shifted focus to modeling challenges in more recent cosmological times. Specifically, we accelerated existing simulation pipelines that model the IA of galaxies and their impact on upcoming weak lensing surveys. We developed two complementary approaches to improve modeling and inference of IA within the HOD framework. First, we introduced IAEMU, a neural network emulator that directly predicts essential two-point statistics for IA modeling: the galaxy clustering correlation function $\xi(r)$, the position–orientation correlation $\omega(r)$, and the orientation–orientation correlation $\eta(r)$, given a set of HOD and IA parameters. Moreover, IAEMU tracks both aleatoric and epistemic uncertainty, providing uncertainty estimates in emulator predictions and covariance information for Monte Carlo analyses. By training IAEMU on a large suite of mock catalogs generated using HALOTOOLS-IA, we showed that it can reproduce these correlation functions at percent-level precision while also providing an estimate of its predictive uncertainty. Crucially, IAEMU offers a $\sim 10^4\times$ speed-up in computing these statistics (on GPU) compared to HALOTOOLS-IA on CPU. Its differentiability also enables accelerated inference via HMC, yielding a $\sim 2000\times$ speed-up compared to HALOTOOLS-IA.

Second, we developed a differentiable, stochastic implementation of the HOD with IA (DIFFHOD-IA). To this end, we made each step of the galaxy alignment modeling differentiable. This includes differentiably drawing central and satellite galaxy populations and assigning their orientations using a differentiable form of the Dimroth–Watson distribution for misalignment angles. This enables end-to-end gradient-based inference, from HOD and IA parameters to the galaxy field and associated summary statistics. We demonstrated this by applying HMC to the IA parameter inference problem, exhibiting a similar inference speed-up as with IAEMU. Both IAEMU and DIFFHOD-IA were validated against each other and against HALOTOOLS-IA, exhibiting agreement with the underlying simulations in both forward-modeling and inference tasks. These advances have direct implications for upcoming surveys, in particular providing tools to marginalize over or constrain IA effects without significant computational cost.

In Chapter 4, we addressed the broader machine-learning challenge of generalization, which is equally important for astrophysics and cosmology. Generalization ensures that mod-

els trained on one dataset (or simulation) remain accurate when applied to a different but related dataset. This phenomenon is ubiquitous, and poses one of the most outstanding challenges for ML applications within the sciences. Our contributions toward generalization examined the effects of symmetry-based inductive biases in NN architectures (the degree of equivariance) and DA training techniques that transfer across datasets.

We introduced SIDDA (Sinkhorn Dynamic Domain Adaptation), a novel training framework that leverages optimal transport distances between NN representations to improve out-of-distribution robustness. The SIDDA algorithm dynamically balances two objectives during training: (1) the primary learning objective on labeled source data (e.g., classification), and (2) the Sinkhorn divergence between source and target feature distributions, which acts to minimize the domain discrepancy. By dynamically adjusting the weight of the loss terms and its entropy regularization during training, SIDDA largely removes the need for manual tuning of hyperparameters. We tested SIDDA on several benchmark tasks, including simple image classification problems where the "source" and "target" domains differed by various systematics (noise, PSF distortions, etc.), as well as on an astronomical dataset of galaxy images with fundamentally different image qualities. Our results showed that SIDDA consistently improved target-domain performance. For instance, on an unlabeled set of real galaxy morphology observations, SIDDA achieved an approximate 13% improvement in classification accuracy relative to a baseline without DA, while also significantly lowering calibration error, indicating more reliable uncertainty estimates.

Moreover, we found that SIDDA is most effective when combined with $E(2)$-equivariant neural networks, providing the strongest overall generalization performance. This is in addition to observing an increased performance of $E(2)$-equivariant models over traditional NNs without DA training. By additionally leveraging these higher-order symmetries (e.g., rotational invariance of galaxy morphologies), one can mitigate differences between simulations and observations at the feature-learning level. Combined with sophisticated DA techniques like SIDDA, this provides a prescription for generalization via informed decisions in both architectural design and training. Chapter 4's contribution is thus more generally towards robust AI, but it is highly relevant for the era of big data cosmology, where models must handle multi-survey, heterogeneous datasets without frequent re-training.

The success on both toy and real-data scenarios suggests that our approach can be applied to a range of problems. In the context of cosmology, methods like SIDDA could be used to adapt models trained on simulated data (which often serve as a source domain) to real observational data (the target domain). For example, SIDDA can be applied in adapting a network trained on simulated

galaxy images to work robustly on actual telescope images with noise and systematics. It can also be used in generalizing models trained on legacy datasets to new observations.

Looking ahead, there are several exciting directions that build on this work. On the theoretical side, the stasis optimization framework could be applied to other cosmological scenarios, including alternative realizations of stasis. In ongoing work, we are applying a similar methodology to differentiably study bubble nucleation models within eternal inflation. In observational cosmology, an exciting follow-up would be a diffusion-based emulator at the galaxy field level, which is also robust to different cosmologies. IAEMU, as it currently stands, is specific to a limited set of cosmological parameters, and while DIFFHOD-IA is robust to different cosmologies, its forward modeling expense is similar to that of HALOTOOLS-IA. Developing a diffusion-based emulator that is robust across cosmologies would represent a tangible step toward field-level inference for galaxy IA.

On the machine learning and data side, an important future direction is extending the SIDDA framework to learning objectives outside of classification, including regression and SBI. Extending symmetry-based approaches to include approximately equivariant networks is another promising direction, as shown in Chapter 4, where overly strong equivariance constraints can limit NN expressivity and degrade performance. These approaches can be applied in tandem to interpolate between subgrid physical models in different hydrodynamic simulations, for example. Our study of IA in this thesis was limited to N-body and HOD simulations, as the variance in subgrid physics makes IA modeling across hydrodynamic simulations difficult. One could imagine equivariant graph neural networks for cosmological graphs or hydro simulations, which, when paired with dynamic domain alignment, could allow models trained on one simulation suite to generalize to others with different subgrid physics.

In conclusion, this thesis illustrates how modern ML techniques, when thoughtfully integrated with domain-specific physics knowledge, can substantially advance the way we extract information from cosmological data. We have shown examples across the cosmic timeline: from early-universe theory, where differentiable simulations and ML-augmented Bayesian inference help uncover new phenomena; to galaxy formation and large-scale structure, where emulators and differentiable models make previously intractable inference tasks feasible; and finally to the practical challenges of astronomy in the big-data era, where symmetry considerations and novel training algorithms like SIDDA ensure that our AI models remain reliable even as we move between simulations, surveys, and instruments. Each of these pieces reflects the common theme of using ML as a tool integrated with physical insight to improve both our efficiency and understanding in cosmology.

Bibliography

[1] Nina A Maksimova, Lehman H Garrison, Daniel J Eisenstein, Boryana Hadzhiyska, Sownak Bose, and Thomas P Satterthwaite. <scp>abacussummit</scp>: a massive set of high-accuracy, high-resolution n-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):4017–4037, September 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab2484. URL http://dx.doi.org/10.1093/mnras/stab2484.

[2] Keith R. Dienes, Lucien Heurtier, Fei Huang, Doojin Kim, Tim M. P. Tait, and Brooks Thomas. Stasis in an expanding universe: A recipe for stable mixed-component cosmological eras. *Physical Review D*, 105(2), jan 2022. doi: 10.1103/physrevd.105.023530. URL https://doi.org/10.1103%2Fphysrevd.105.023530.

[3] Planck Collaboration, N. Aghanim, et al. Planck 2018 results. VI. Cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020. doi: 10.1051/0004-6361/201833910.

[4] Solène Chabanier, Marius Millea, and Nathalie Palanque-Delabrouille. Matter power spectrum: from ly alpha forest to cmb scales. *Monthly Notices of the Royal Astronomical Society*, 489(2):2247–2253, August 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2310. URL http://dx.doi.org/10.1093/mnras/stz2310.

[5] Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns, 2021.

[6] Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy evolution from halo occupation distribution modeling of DEEP2 and SDSS galaxy clustering. *The Astrophysical Journal*, 667 (2):760–779, oct 2007. doi: 10.1086/521074. URL https://doi.org/10.1086%2F521074.

[7] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls

of heteroscedastic uncertainty estimation with probabilistic neural networks, 2022. URL https://arxiv.org/abs/2203.09168.

[8] Sneh Pandya, Yuanyuan Yang, Nicholas Van Alfen, Jonathan Blazek, and Robin Walters. Iaemu: Learning galaxy intrinsic alignment correlations. *The Open Journal of Astrophysics*, 8, December 2025. ISSN 2565-6120. doi: 10.33232/001c.151749. URL http://dx.doi.org/10.33232/001c.151749.

[9] Nathanaël Perraudin, Ankit Srivastava, Aurelien Lucchi, Tomasz Kacprzak, Thomas Hofmann, and Alexandre Réfrégier. Cosmological n-body simulations: a challenge for scalable generative models. *Computational Astrophysics and Cosmology*, 6, 12 2019. doi: 10.1186/s40668-019-0032-1.

[10] Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929. doi: 10.1073/pnas.15.3.168. URL https://www.pnas.org/doi/abs/10.1073/pnas.15.3.168.

[11] S. Perlmutter, G. Aldering, M. Della Valle, S. Deustua, R. S. Ellis, S. Fabbro, A. Fruchter, G. Goldhaber, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, R. A. Knop, C. Lidman, R. G. McMahon, P. Nugent, R. Pain, N. Panagia, C. R. Pennypacker, P. Ruiz-Lapuente, B. Schaefer, and N. Walton. Discovery of a supernova explosion at half the age of the universe. *Nature*, 391(6662):51–54, January 1998. ISSN 1476-4687. doi: 10.1038/34124. URL http://dx.doi.org/10.1038/34124.

[12] Adam G. Riess et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116(3):1009–1038, 1998. doi: 10.1086/300499.

[13] Arno A Penzias and Robert W Wilson. A measurement of excess antenna temperature at 4080 mc/s. *The Astrophysical Journal*, 142:419–421, 1965. doi: 10.1086/148307.

[14] Frank W Dyson, Arthur S Eddington, and Charles Davidson. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of may 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 220(571-581):291–333, 1920. doi: 10.1098/rsta.1920.0009.

[15] Emmanuel N. Saridakis. From cosmology to cosmonomy, 2025. URL https://arxiv.org/abs/2512.20416.

[16] Matthew D. Schwartz. Resummation of the c-parameter sudakov shoulder using effective field theory, 2026. URL https://arxiv.org/abs/2601.02484.

[17] Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Gar-

rett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang,

Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. LSST: From Science Drivers to Reference Design and Anticipated Data Products. ApJ, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

[18] Rachel Akeson, Lee Armus, Etienne Bachelet, Vanessa Bailey, Lisa Bartusek, Andrea Bellini, Dominic Benford, David Bennett, Aparna Bhattacharya, Ralph Bohlin, Martha Boyer, Valerio Bozza, Geoffrey Bryden, Sebastiano Calchi Novati, Kenneth Carpenter, Stefano Casertano, Ami Choi, David Content, Pratika Dayal, Alan Dressler, Olivier Doré, S. Michael Fall, Xiaohui Fan, Xiao Fang, Alexei Filippenko, Steven Finkelstein, Ryan Foley, Steven Furlanetto, Jason Kalirai, B. Scott Gaudi, Karoline Gilbert, Julien Girard, Kevin Grady, Jenny Greene, Puragra Guhathakurta, Chen Heinrich, Shoubaneh Hemmati, David Hendel, Calen Henderson, Thomas Henning, Christopher Hirata, Shirley Ho, Eric Huff, Anne Hutter, Rolf Jansen, Saurabh Jha, Samson Johnson, David Jones, Jeremy Kasdin, Patrick Kelly, Robert Kirshner, Anton Koekemoer, Jeffrey Kruk, Nikole Lewis, Bruce Macintosh, Piero Madau, Sangeeta Malhotra, Kaisey Mandel, Elena Massara, Daniel Masters, Julie McEnery, Kristen McQuinn, Peter Melchior, Mark Melton, Bertrand Mennesson, Molly Peeples, Matthew Penny, Saul Perlmutter, Alice Pisani, Andrés Plazas, Radek Poleski, Marc Postman, Clément Ranc, Bernard Rauscher, Armin Rest, Aki Roberge, Brant Robertson, Steven Rodney, James Rhoads, Jason Rhodes, Russell Ryan Jr. au2, Kailash Sahu, David Sand, Dan Scolnic, Anil Seth, Yossi Shvartzvald, Karelle Siellez, Arfon Smith, David Spergel, Keivan Stassun, Rachel Street, Louis-Gregory Strolger, Alexander Szalay, John Trauger, M. A. Troxel, Margaret Turnbull, Roeland van der Marel, Anja von der Linden, Yun Wang, David Weinberg, Benjamin Williams, Rogier Windhorst, Edward Wollack, Hao-Yi Wu, Jennifer Yee, and Neil Zimmerman. The wide field infrared survey telescope: 100 hubbles for the 2020s, 2019. URL https://arxiv.org/abs/1902.05569.

[19] R. Scaramella, J. Amiaux, Y. Mellier, C. Burigana, C. S. Carvalho, J.-C. Cuillandre, A. Da Silva, A. Derosa, J. Dinis, E. Maiorano, M. Maris, I. Tereno, R. Laureijs, T. Boenke, G. Buenadicha, X. Dupac, L. M. Gaspar Venancio, P. Gómez-Álvarez, J. Hoar, J. Lorenzo Alvarez, G. D. Racca, G. Saavedra-Criado, J. Schwartz, R. Vavrek, M. Schirmer, H. Aussel, R. Azzollini, V. F. Cardone, M. Cropper, A. Ealet, B. Garilli, W. Gillard, B. R. Granett, L. Guzzo, H. Hoekstra, K. Jahnke, T. Kitching, T. Maciaszek, M. Meneghetti, L. Miller, R. Nakajima, S. M. Niemi, F. Pasian, W. J. Percival, S. Pottinger, M. Sauvage, M. Scodeg-

gio, S. Wachter, A. Zacchei, N. Aghanim, A. Amara, T. Auphan, N. Auricchio, S. Awan, A. Balestra, R. Bender, C. Bodendorf, D. Bonino, E. Branchini, S. Brau-Nogue, M. Brescia, G. P. Candini, V. Capobianco, C. Carbone, R. G. Carlberg, J. Carretero, R. Casas, F. J. Castander, M. Castellano, S. Cavuoti, A. Cimatti, R. Cledassou, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, A. Costille, F. Courbin, H. Degaudenzi, M. Douspis, F. Dubath, C. A. J. Duncan, S. Dusini, S. Farrens, S. Ferriol, P. Fosalba, N. Fourmanoit, M. Frailis, E. Franceschi, P. Franzetti, M. Fumana, B. Gillis, C. Giocoli, A. Grazian, F. Grupp, S. V. H. Haugan, W. Holmes, F. Hormuth, P. Hudelot, S. Kermiche, A. Kiessling, M. Kilbinger, R. Kohley, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, O. Lahav, S. Ligori, P. B. Lilje, I. Lloro, O. Mansutti, O. Marggraf, K. Markovic, F. Marulli, R. Massey, S. Maurogordato, M. Melchior, E. Merlin, G. Meylan, J. J. Mohr, M. Moresco, B. Morin, L. Moscardini, E. Munari, R. C. Nichol, C. Padilla, S. Paltani, J. Peacock, K. Pedersen, V. Pettorino, S. Pires, M. Poncet, L. Popa, L. Pozzetti, F. Raison, R. Rebolo, J. Rhodes, H.-W. Rix, M. Roncarelli, E. Rossetti, R. Saglia, P. Schneider, T. Schrabback, A. Secroun, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, J. Skottfelt, L. Stanco, J. L. Starck, P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, H. I. Teplitz, R. Toledo-Moreo, F. Torradeflot, M. Trifoglio, E. A. Valentijn, L. Valenziano, G. A. Verdoes Kleijn, Y. Wang, N. Welikala, J. Weller, M. Wetzstein, G. Zamorani, J. Zoubian, S. Andreon, M. Baldi, S. Bardelli, A. Boucaud, S. Camera, D. Di Ferdinando, G. Fabbian, R. Farinelli, S. Galeotta, J. Graciá-Carpio, D. Maino, E. Medinaceli, S. Mei, C. Neissner, G. Polenta, A. Renzi, E. Romelli, C. Rosset, F. Sureau, M. Tenti, T. Vassallo, E. Zucca, C. Baccigalupi, A. Balaguera-Antolínez, P. Battaglia, A. Biviano, S. Borgani, E. Bozzo, R. Cabanac, A. Cappi, S. Casas, G. Castignani, C. Colodro-Conde, J. Coupon, H. M. Courtois, J. Cuby, S. de la Torre, S. Desai, H. Dole, M. Fabricius, M. Farina, P. G. Ferreira, F. Finelli, P. Flose-Reimberg, S. Fotopoulou, K. Ganga, G. Gozaliasl, I. M. Hook, E. Keihanen, C. C. Kirkpatrick, P. Liebing, V. Lindholm, G. Mainetti, M. Martinelli, N. Martinet, M. Maturi, H. J. McCracken, R. B. Metcalf, G. Morgante, J. Nightingale, A. Nucita, L. Patrizii, D. Potter, G. Riccio, A. G. Sánchez, D. Sapone, J. A. Schewtschenko, M. Schultheis, V. Scottez, R. Teyssier, I. Tutusaus, J. Valiviita, M. Viel, W. Vriend, and L. Whittaker. Euclid preparation: I. the euclid wide survey. *Astronomy &; Astrophysics*, 662:A112, June 2022. ISSN 1432-0746. doi: 10.1051/0004-6361/202141938. URL http://dx.doi.org/10.1051/0004-6361/202141938.

[20] Michael E. Levi, Lori E. Allen, Anand Raichoor, Charles Baltay, Segev BenZvi, Flo-

rian Beutler, Adam Bolton, Francisco J. Castander, Chia-Hsun Chuang, Andrew Cooper, Jean-Gabriel Cuby, Arjun Dey, Daniel Eisenstein, Xiaohui Fan, Brenna Flaugher, Carlos Frenk, Alma X. Gonzalez-Morales, Or Graur, Julien Guy, Salman Habib, Klaus Honscheid, Stephanie Juneau, Jean-Paul Kneib, Ofer Lahav, Dustin Lang, Alexie Leauthaud, Betta Lusso, Axel de la Macorra, Marc Manera, Paul Martini, Shude Mao, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Will J. Percival, Carlos Allende Prieto, Constance M. Rockosi, Vanina Ruhlmann-Kleider, David Schlegel, Hee-Jong Seo, Yong-Seon Song, Greg Tarle, Risa Wechsler, David Weinberg, Christophe Yeche, and Ying Zu. The dark energy spectroscopic instrument (desi), 2019. URL https://arxiv.org/abs/1907.10688.

[21] Scott Dodelson and Fabian Schmidt. *Modern Cosmology*. Academic Press, 2nd edition, 2020. ISBN 978-0128159484.

[22] David Tong. Lectures on cosmology. University of Cambridge lecture notes, Department of Applied Mathematics and Theoretical Physics, 2019. URL https://www.damtp.cam.ac.uk/user/tong/cosmo/cosmo.pdf. Cosmology lecture notes PDF.

[23] S. Samuroff. *Systematic Biases in Weak Lensing Cosmology with the Dark Energy Survey*. PhD thesis, University of Manchester, September 2017. 4 September 2017.

[24] D. J. Fixsen. The Temperature of the Cosmic Microwave Background. *The Astrophysical Journal*, 707:916–920, 2009. doi: 10.1088/0004-637X/707/2/916.

[25] A. Friedman. On the Curvature of space. *Z. Phys.*, 10:377–386, 1922. doi: 10.1007/BF01332580.

[26] Wendy L. Freedman, Barry F. Madore, Taylor J. Hoyt, In Sung Jang, Abigail J. Lee, and Kayla A. Owens. Status report on the chicago-carnegie hubble program (cchp): Measurement of the hubble constant using the hubble and james webb space telescopes. *The Astrophysical Journal*, 985(2):203, may 2025. doi: 10.3847/1538-4357/adce78. URL https://doi.org/10.3847/1538-4357/adce78.

[27] Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929. doi: 10.1073/pnas.15.3.168.

[28] F. Zwicky. Die rotverschiebung von extragalaktischen nebeln. *Helvetica Physica Acta*, 6: 110–127, 1933. English translation: Gen. Relativ. Gravit. 41, 207 (2009).

[29] Vera C. Rubin and Jr. Ford, W. Kent. Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *The Astrophysical Journal*, 159:379–403, 1970. doi: 10.1086/150317.

[30] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, et al. Planck 2018 results. vi. cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020. doi: 10.1051/0004-6361/201833910.

[31] Steven Weinberg. *Cosmology*. Oxford University Press, 2008. ISBN 978-0198526827.

[32] Douglas Clowe, Maruša Bradač, Anthony H. Gonzalez, Maxim Markevitch, Scott W. Randall, Christine Jones, and Dennis Zaritsky. A direct empirical proof of the existence of dark matter. *The Astrophysical Journal*, 648(2):L109–L113, August 2006. ISSN 1538-4357. doi: 10.1086/508162. URL http://dx.doi.org/10.1086/508162.

[33] A.G. Adame, J. Aguilar, S. Ahlen, S. Alam, D.M. Alexander, M. Alvarez, O. Alves, A. Anand, U. Andrade, E. Armengaud, S. Avila, A. Aviles, H. Awan, B. Bahr-Kalus, S. Bailey, C. Baltay, A. Bault, J. Behera, S. BenZvi, A. Bera, F. Beutler, D. Bianchi, C. Blake, R. Blum, S. Brieden, A. Brodzeller, D. Brooks, E. Buckley-Geer, E. Burtin, R. Calderon, R. Canning, A. Carnero Rosell, R. Cereskaite, J.L. Cervantes-Cota, S. Chabanier, E. Chaussidon, J. Chaves-Montero, S. Chen, X. Chen, T. Claybaugh, S. Cole, A. Cuceu, T.M. Davis, K. Dawson, A. de la Macorra, A. de Mattia, N. Deiosso, A. Dey, B. Dey, Z. Ding, P. Doel, J. Edelstein, S. Eftekharzadeh, D.J. Eisenstein, A. Elliott, P. Fagrelius, K. Fanning, S. Ferraro, J. Ereza, N. Findlay, B. Flaugher, A. Font-Ribera, D. Forero-Sánchez, J.E. Forero-Romero, C.S. Frenk, C. Garcia-Quintero, E. Gaztañaga, H. Gil-Marín, S.Gontcho A. Gontcho, A.X. Gonzalez-Morales, V. Gonzalez-Perez, C. Gordon, D. Green, D. Gruen, R. Gsponer, G. Gutierrez, J. Guy, B. Hadzhiyska, C. Hahn, M.M.S. Hanif, H.K. Herrera-Alcantar, K. Honscheid, C. Howlett, D. Huterer, V. Iršič, M. Ishak, S. Juneau, N.G. Karaçaylı, R. Kehoe, S. Kent, D. Kirkby, A. Kremin, A. Krolewski, Y. Lai, T.-W. Lan, M. Landriau, D. Lang, J. Lasker, J.M. Le Goff, L. Le Guillou, A. Leauthaud, M.E. Levi, T.S. Li, E. Linder, K. Lodha, C. Magneville, M. Manera, D. Margala, P. Martini, M. Maus, P. McDonald, L. Medina-Varela, A. Meisner, J. Mena-Fernández, R. Miquel,

J. Moon, S. Moore, J. Moustakas, E. Mueller, A. Muñoz-Gutiérrez, A.D. Myers, S. Nadathur, L. Napolitano, R. Neveux, J.A. Newman, N.M. Nguyen, J. Nie, G. Niz, H.E. Noriega, N. Padmanabhan, E. Paillas, N. Palanque-Delabrouille, J. Pan, S. Penmetsa, W.J. Percival, M.M. Pieri, M. Pinon, C. Poppett, A. Porredon, F. Prada, A. Pérez-Fernández, I. Pérez-Ràfols, D. Rabinowitz, A. Raichoor, C. Ramírez-Pérez, S. Ramirez-Solano, M. Rashkovetskyi, C. Ravoux, M. Rezaie, J. Rich, A. Rocher, C. Rockosi, N.A. Roe, A. Rosado-Marin, A.J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, L. Samushia, E. Sanchez, C. Saulder, E.F. Schlafly, D. Schlegel, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, J. Silber, A. Slosar, A. Smith, D. Sprayberry, T. Tan, G. Tarlé, P. Taylor, S. Trusov, L.A. Ureña-López, R. Vaisakh, D. Valcin, F. Valdes, M. Vargas-Magaña, L. Verde, M. Walther, B. Wang, M.S. Wang, B.A. Weaver, N. Weaverdyck, R.H. Wechsler, D.H. Weinberg, M. White, J. Yu, Y. Yu, S. Yuan, C. Yèche, E.A. Zaborowski, P. Zarrouk, H. Zhang, C. Zhao, R. Zhao, R. Zhou, T. Zhuang, and H. Zou. Desi 2024 vi: cosmological constraints from the measurements of baryon acoustic oscillations. *Journal of Cosmology and Astroparticle Physics*, 2025 (02):021, February 2025. ISSN 1475-7516. doi: 10.1088/1475-7516/2025/02/021. URL http://dx.doi.org/10.1088/1475-7516/2025/02/021.

[34] J. Richard Bond, Lev Kofman, and Dmitry Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380(6575):603–606, April 1996. ISSN 1476-4687. doi: 10.1038/380603a0. URL http://dx.doi.org/10.1038/380603a0.

[35] Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, John A. Peacock, Shaun Cole, Peter Thomas, Hugh Couchman, August Evrard, Joerg Colberg, and Frazer Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435: 629–636, 2005. doi: 10.1038/nature03597.

[36] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. Dark matter halos in the standard cosmological model: Results from the bolshoi simulation. *The Astrophysical Journal*, 740:102, 2011. doi: 10.1088/0004-637X/740/2/102.

[37] Nina A. Maksimova, Lehman H. Garrison, Daniel J. Eisenstein, Boryana Hadzhiyska, Sownak Bose, and Thomas P. Satterthwaite. Abacussummit: a massive set of high-accuracy, high-resolution n-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508:4017–4037, 2021. doi: 10.1093/mnras/stab2484.

[38] 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*, 475:648–675, 2018. doi: 10.1093/mnras/stx3112.

[39] Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446:521–554, 2015. doi: 10.1093/mnras/stu2058.

[40] Romeel Davé, Daniel Anglés-Alcázar, Desika Narayanan, Qi Li, Mika H. Rafieferantsoa, and Sarah Appleby. Simba: Cosmological simulations with black hole growth and feedback. *Monthly Notices of the Royal Astronomical Society*, 486:2827–2849, 2019. doi: 10.1093/mnras/stz937.

[41] Yongseok Jo, Shy Genel, Anirvan Sengupta, Benjamin Wandelt, Rachel Somerville, and Francisco Villaescusa-Navarro. Towards robustness across cosmological simulation models tng, simba, astrid, and eagle, 2025. URL https://arxiv.org/abs/2502.13239.

[42] Maria Cristina Fortuna, Henk Hoekstra, Benjamin Joachimi, Harry Johnston, Nora Elisa Chisari, Christos Georgiou, and Constance Mahony. The halo model as a versatile tool to predict intrinsic alignments. MNRAS, 501(2):2983–3002, February 2021. doi: 10.1093/mnras/staa3802.

[43] Zheng Zheng, Andreas A. Berlind, David H. Weinberg, Andrew J. Benson, Carlton M. Baugh, Shaun Cole, Romeel Davé, Carlos S. Frenk, Neal Katz, and Cedric G. Lacey. Theoretical models of the halo occupation distribution: Separating central and satellite galaxies. *The Astrophysical Journal*, 633:791–809, 2005. doi: 10.1086/466510.

[44] Alan H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. *Physical Review D*, 23:347–356, 1981. doi: 10.1103/PhysRevD.23.347.

[45] A. D. Linde. A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Physics Letters B*, 108: 389–393, 1982. doi: 10.1016/0370-2693(82)91219-9.

[46] Jihn E. Kim, Hans Peter Nilles, and Marco Peloso. Completing natural inflation. *JCAP*, 01: 005, 2005. doi: 10.1088/1475-7516/2005/01/005.

[47] Andrew R. Liddle and David H. Lyth. *Cosmological Inflation and Large-Scale Structure*. Cambridge University Press, 2000. ISBN 978-0521575980.

[48] Richard H. Cyburt, Brian D. Fields, Keith A. Olive, and Tsung-Han Yeh. Big bang nucleosynthesis: Present status. *Rev. Mod. Phys.*, 88:015004, Feb 2016. doi: 10.1103/RevModPhys.88.015004. URL https://link.aps.org/doi/10.1103/RevModPhys.88.015004.

[49] Particle Data Group. 24. big bang nucleosynthesis. Review of Particle Physics, 2023. URL https://pdg.lbl.gov/2023/reviews/rpp2023-rev-bbang-nucleosynthesis.pdf. Primordial abundances of D, ^{3}He, ^{4}He, and ^{7}Li.

[50] S. Perlmutter et al. Measurements of ω and λ from 42 high-redshift supernovae. *The Astrophysical Journal*, 517(2):565–586, 1999. doi: 10.1086/307221.

[51] Daniel J. Eisenstein et al. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *The Astrophysical Journal*, 633:560–574, 2005. doi: 10.1086/466512.

[52] Will J. Percival, Shaun Cole, Daniel J. Eisenstein, Robert C. Nichol, John A. Peacock, Adrian C. Pope, and Alexander S. Szalay. Measuring the baryon acoustic oscillation scale using the sloan digital sky survey and 2df galaxy redshift survey. *Monthly Notices of the Royal Astronomical Society*, 381(3):1053–1066, 2007. doi: 10.1111/j.1365-2966.2007.12268.x.

[53] David H. Weinberg, Michael J. Mortonson, Daniel J. Eisenstein, Christopher Hirata, Adam G. Riess, and Eduardo Rozo. Observational probes of cosmic acceleration. *Physics Reports*, 530 (2):87–255, 2013. doi: 10.1016/j.physrep.2013.05.001.

[54] DES Collaboration, T. M. C. Abbott, et al. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105:023520, 2022. doi: 10.1103/PhysRevD.105.023520.

[55] Marika Asgari et al. KiDS-1000 cosmology: Cosmic shear constraints and comparison between two point statistics. *Astronomy & Astrophysics*, 645:A104, 2021. doi: 10.1051/0004-6361/202039070.

[56] Chiaki Hikage et al. Cosmology from cosmic shear power spectra with Subaru Hyper Suprime-Cam first-year data. *Publications of the Astronomical Society of Japan*, 71(2):43, 2019. doi: 10.1093/pasj/psz010.

[57] Masahiro Takada and Bhuvnesh Jain. The three-point correlation function in cosmology. *Monthly Notices of the Royal Astronomical Society*, 348(3):897–915, 2004. doi: 10.1111/j.1365-2966.2004.07410.x.

[58] Liping Fu, Martin Kilbinger, Thomas Erben, Catherine Heymans, Hendrik Hildebrandt, Henk Hoekstra, Thomas D. Kitching, Yannick Mellier, Lance Miller, Elisabetta Semboloni, Patrick Simon, Ludovic Van Waerbeke, Jean Coupon, Joachim Harnois-Déraps, Michael J. Hudson, Konrad Kuijken, Barnaby Rowe, Tim Schrabback, Sanaz Vafaei, and Malin Velander. Cfhtlens: baryon acoustic oscillations in the angular galaxy clustering and in cosmic shear out to z 1. *Monthly Notices of the Royal Astronomical Society*, 441(3):2725–2743, 2014. doi: 10.1093/mnras/stu754.

[59] Jörg P. Dietrich and Jan Hartlap. Cosmology with the shear-peak statistics. *Monthly Notices of the Royal Astronomical Society*, 402(2):1049–1058, 2010. doi: 10.1111/j.1365-2966.2009.15948.x.

[60] Nicolas Martinet, Peter Schneider, Hendrik Hildebrandt, HuanYuan Shan, Marika Asgari, Jörg P. Dietrich, Joachim Harnois-Déraps, Thomas Erben, Reiko Nakajima, Massimo Viola, and Ami Choi. Kids-450: cosmological constraints from weak lensing peak statistics - i. inference from analytical prediction of high signal-to-noise ratio convergence peaks. *Monthly Notices of the Royal Astronomical Society*, 474(1):712–730, 2018. doi: 10.1093/mnras/stx2793.

[61] Jan M. Kratochvil, Zoltán Haiman, and Morgan May. Probing cosmology with weak lensing minkowski functionals. *Physical Review D*, 85(10):103513, 2012. doi: 10.1103/PhysRevD.85.103513.

[62] Andrea Petri, Zoltán Haiman, and Morgan May. Sample variance in weak lensing: how many simulations are required? *Physical Review D*, 91(10):103511, 2015. doi: 10.1103/PhysRevD.91.103511.

[63] Euclid Collaboration, V. Ajani, M. Baldi, A. Barthelemy, et al. Euclid preparation. xxviii. forecasts for ten different higher-order weak lensing statistics. *Astronomy & Astrophysics*, 675:A120, 2023. doi: 10.1051/0004-6361/202346017.

[64] Antony Lewis and Sarah Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Physical Review D*, 66(10):103511, 2002. doi: 10.1103/PhysRevD.66.103511.

[65] Luis E. Padilla, Luis O. Tellez, Luis A. Escamilla, and Jose Alberto Vazquez. Cosmological parameter inference with Bayesian statistics. *Universe*, 7(7):213, 2021. doi: 10.3390/universe7070213.

[66] Andrew R. Liddle. Statistical methods for cosmological parameter selection and estimation. *Annual Review of Nuclear and Particle Science*, 59:95–114, 2009. doi: 10.1146/annurev.nucl.010909.083706.

[67] Catherine Heymans et al. Kids-1000 cosmology: Multi-probe weak gravitational lensing and spectroscopic galaxy clustering constraints. *Astronomy & Astrophysics*, 646:A140, 2021. doi: 10.1051/0004-6361/202039063.

[68] J. Akeret, S. Seehars, A. Amara, A. Refregier, and A. Csillaghy. CosmoHammer: Cosmological parameter estimation with the MCMC Hammer. *Astronomy and Computing*, 2:27–39, 2013. doi: 10.1016/j.ascom.2013.06.003.

[69] Davide Piras, Alicja Polanska, Alessio Spurio Mancini, Matthew A. Price, and Jason D. McEwen. The future of cosmological likelihood-based inference: accelerated high-dimensional parameter estimation and model comparison. *The Open Journal of Astrophysics*, 7, 2024. doi: 10.33232/001c.123368.

[70] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

[71] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman and Hall/CRC, 2011.

[72] J. Ruiz-Zapatero, D. Alonso, C. García-García, A. Nicola, A. Mootoovaloo, J. M. Sullivan, M. Bonici, and P. G. Ferreira. LimberJack.jl: auto-differentiable methods for angular power spectra analyses. *The Open Journal of Astrophysics*, 7, 2024. doi: 10.21105/astro.2310.08306.

[73] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117.

[74] Matthias Bartelmann and Peter Schneider. Weak gravitational lensing. *Physics Reports*, 340:291–472, 2001. doi: 10.1016/S0370-1573(00)00082-X.

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

[76] D. Nelson Limber. The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field. *The Astrophysical Journal*, 117:134, 1953. doi: 10.1086/145672.

[77] Marilena LoVerde and Niayesh Afshordi. Extended Limber approximation. *Physical Review D*, 78:123506, 2008. doi: 10.1103/PhysRevD.78.123506.

[78] 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, 2015. doi: 10.1016/j.physrep.2014.11.001.

[79] 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, July 2015. ISSN 1572-9672. doi: 10.1007/s11214-015-0177-4. URL http://dx.doi.org/10.1007/s11214-015-0177-4.

[80] Paolo Catelan, Marc Kamionkowski, and Roger D. Blandford. Intrinsic and extrinsic galaxy alignment. *Monthly Notices of the Royal Astronomical Society*, 320:L7–L13, 2001. doi: 10.1046/j.1365-8711.2001.04105.x.

[81] Christopher M. Hirata and Uroš Seljak. Intrinsic alignment-lensing interference as a contaminant of cosmic shear. *Physical Review D*, 70:063526, 2004. doi: 10.1103/PhysRevD.70.063526.

[82] P. J. E. Peebles. Origin of the Angular Momentum of Galaxies. *The Astrophysical Journal*, 155:393, 1969. doi: 10.1086/149876.

[83] Simon D. M. White. Angular momentum growth in protogalaxies. *The Astrophysical Journal*, 286:38–41, 1984. doi: 10.1086/162573.

[84] C. M. Hirata and U. Seljak. Intrinsic alignment-lensing interference as a contaminant of cosmic shear. Phys. Rev. D, 70(6):063526–+, September 2004. doi: 10.1103/PhysRevD.70.063526.

[85] Sarah Bridle and Lindsay King. Dark energy constraints from cosmic shear power spectra: impact of intrinsic alignments on photometric redshift requirements. *New Journal of Physics*, 9:444, 2007. doi: 10.1088/1367-2630/9/12/444.

[86] Jonathan A. Blazek, Niall MacCrann, M. A. Troxel, and Xiao Fang. Beyond linear galaxy alignments. *Physical Review D*, 100:103506, 2019. doi: 10.1103/PhysRevD.100.103506.

[87] Claire Lamman, Eleni Tsaprazi, Jingjing Shi, Nikolina Niko Šarčević, Susan Pyne, Elisa Legnani, and Tassia Ferreira. The ia guide: A breakdown of intrinsic alignment formalisms. *The Open Journal of Astrophysics*, 7, February 2024. ISSN 2565-6120. doi: 10.21105/astro.2309.08605. URL http://dx.doi.org/10.21105/astro.2309.08605.

[88] Aleksandra Ćiprijanović, Diana Kafkes, Gregory Snyder, F Javier Sánchez, Gabriel Nathan Perdue, Kevin Pedro, Brian Nord, Sandeep Madireddy, and Stefan M Wild. Deepadversaries: examining the robustness of deep learning models for galaxy morphology classification. *Machine Learning: Science and Technology*, 3(3):035007, 2022.

[89] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

[90] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

[91] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665.

[92] Sneh Pandya, Purvik Patel, Franc O, and Jonathan Blazek. E(2) equivariant neural networks for robust galaxy morphology classification, 2023. URL https://arxiv.org/abs/2311.01500.

[93] C. Schaefer, M. Geiger, T. Kuntzer, and J.-P. Kneib. Deep convolutional neural networks as strong gravitational lens detectors. *Astronomy and; Astrophysics*, 611:A2, March 2018. ISSN 1432-0746. doi: 10.1051/0004-6361/201731201. URL http://dx.doi.org/10.1051/0004-6361/201731201.

[94] Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490(2):1843–1860, September 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2610. URL http://dx.doi.org/10.1093/mnras/stz2610.

[95] Taco S. Cohen and Max Welling. Group equivariant convolutional networks, 2016. URL https://arxiv.org/abs/1602.07576.

[96] Taco S. Cohen and Max Welling. Steerable cnns, 2016. URL https://arxiv.org/abs/1612.08498.

[97] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[98] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. https://github.com/jax-ml/jax, 2018. Version 0.3.13.

[99] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[100] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[101] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[102] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[103] Lutz Prechelt. Early stopping-but when? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pages 55–69. Springer, 1996. ISBN 3-540-65311-2. URL http://dblp.uni-trier.de/db/conf/nips/nips1996.html#Prechelt96.

[104] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[105] James Halverson and Sneh Pandya. Generality and persistence of cosmological stasis. *Physical Review D*, 110(7), October 2024. ISSN 2470-0029. doi: 10.1103/physrevd.110.075041. URL http://dx.doi.org/10.1103/PhysRevD.110.075041.

[106] Sneh Pandya, Purvik Patel, Brian D Nord, Mike Walmsley, and Aleksandra Ciprijanovic. Sidda: Sinkhorn dynamic domain adaptation for image classification with equivariant neural networks. *Machine Learning: Science and Technology*, August 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/adf701. URL http://dx.doi.org/10.1088/2632-2153/adf701.

[107] Sneh Pandya and Jonathan Blazek. Differentiable stochastic halo occupation distribution with galaxy intrinsic alignments, 2026. URL https://arxiv.org/abs/2602.04977.

[108] Edward Berman, Sneh Pandya, Jacqueline McCleary, Marko Shuntov, Caitlin Casey, Nicole Drakos, Andreas Faisst, Steven Gillman, Ghassem Gozaliasl, Natalie Hogg, Jeyhan Kartaltepe, Anton Koekemoer, Wilfried Mercier, Diana Scognamiglio, COSMOS-Web, :, and The JWST Cosmic Origins Survey. On soft clustering for correlation estimators: Model uncertainty, differentiability, and surrogates, 2025. URL https://arxiv.org/abs/2504.06174.

[109] Sneh Pandya, Yuanyuan Yang, Nicholas Van Alfen, Jonathan Blazek, and Robin Walters. Learning galaxy intrinsic alignment correlations, 2024. URL https://arxiv.org/abs/2404.13702.

[110] Benjamin Horowitz, ChangHoon Hahn, Francois Lanusse, Chirag Modi, and Simone Ferraro. Differentiable stochastic halo occupation distribution, 2022. URL https://arxiv.org/abs/2211.03852.

[111] Nicholas Van Alfen, Duncan Campbell, Jonathan Blazek, C. Danielle Leonard, Francois Lanusse, Andrew Hearin, Rachel Mandelbaum, and The LSST Dark Energy Science Collaboration. An empirical model for intrinsic alignments: Insights from cosmological simulations, 2024.

[112] Andrew P. Hearin, Nesar Ramachandra, Matthew R. Becker, and Joseph DeRose. Differentiable predictions for large scale structure with shamnet. *The Open Journal of Astrophysics*, 5(1), February 2022. ISSN 2565-6120. doi: 10.21105/astro.2112.08423. URL http://dx.doi.org/10.21105/astro.2112.08423.

[113] Keith R. Dienes, Lucien Heurtier, Fei Huang, Tim M. P. Tait, and Brooks Thomas. Stasis, stasis, triple stasis, 2023.

[114] Keith R. Dienes, Lucien Heurtier, Fei Huang, Doojin Kim, Tim M. P. Tait, and Brooks Thomas. Primordial black holes place the universe in stasis, 2023. URL https://arxiv.org/abs/2212.01369.

[115] Keith R. Dienes, Lucien Heurtier, Fei Huang, Tim M. P. Tait, and Brooks Thomas. Cosmological stasis from dynamical scalars: Tracking solutions and the possibility of a stasis-induced inflation, 2024. URL https://arxiv.org/abs/2406.06830.

[116] Samuel S. Schoenholz and Ekin D. Cubuk. Jax m.d. a framework for differentiable physics. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/file/83d3d4b6c9579515e1679aca8cbc8033-Paper.pdf.

[117] Jean-Eric Campagne, François Lanusse, Joe Zuntz, Alexandre Boucaud, Santiago Casas, Minas Karamanis, David Kirkby, Denise Lanzieri, Austin Peel, and Yin Li. Jax-cosmo: An end-to-end differentiable and gpu accelerated cosmology library. *The Open Journal of Astrophysics*, 6, April 2023. ISSN 2565-6120. doi: 10.21105/astro.2302.05163. URL http://dx.doi.org/10.21105/astro.2302.05163.

[118] Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference, 2013. URL https://arxiv.org/abs/1206.7051.

[119] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

[120] Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

[121] Anne Kværnø. Singly diagonally implicit runge–kutta methods with an explicit first stage. *BIT Numerical Mathematics*, 44(3):489–502, 2004.

[122] Philipp Stumm and Andrea Walther. New algorithms for optimal online checkpointing. *SIAM Journal on Scientific Computing*, 32(2):836–854, 2010. doi: 10.1137/080742439.

[123] Qiqi Wang, Parviz Moin, and Gianluca Iaccarino. Minimal repetition dynamic checkpointing algorithm for unsteady adjoint calculation. *SIAM Journal on Scientific Computing*, 31(4): 2549–2567, 2009. doi: 10.1137/080727890.

[124] Felix Petersen, Christian Borgelt, Hilde Kuehne, and Oliver Deussen. Differentiable sorting networks for scalable sorting and ranking supervision, 2021. URL https://arxiv.org/abs/2105.04019.

[125] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016. URL https://arxiv.org/abs/1505.05770.

[126] Nicola De Cao, Ivan Titov, and Wilker Aziz. Block neural autoregressive flow, 2019. URL https://arxiv.org/abs/1904.04676.

[127] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows, 2018. URL https://arxiv.org/abs/1804.00779.

[128] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

[129] Peter Svrcek and Edward Witten. Axions In String Theory. *JHEP*, 06:051, 2006. doi: 10.1088/1126-6708/2006/06/051.

[130] Asimina Arvanitaki, Savas Dimopoulos, Sergei Dubovsky, Nemanja Kaloper, and John March-Russell. String Axiverse. *Phys. Rev. D*, 81:123530, 2010. doi: 10.1103/PhysRevD.81.123530.

[131] S. Dimopoulos, S. Kachru, J. McGreevy, and Jay G. Wacker. N-flation. *JCAP*, 08:003, 2008. doi: 10.1088/1475-7516/2008/08/003.

[132] Eva Silverstein and Alexander Westphal. Monodromy in the CMB: Gravity Waves and String Inflation. *Phys. Rev. D*, 78:106003, 2008. doi: 10.1103/PhysRevD.78.106003.

[133] Liam McAllister, Eva Silverstein, and Alexander Westphal. Gravity Waves and Linear Inflation from Axion Monodromy. *Phys. Rev. D*, 82:046003, 2010. doi: 10.1103/PhysRevD.82.046003.

[134] James Halverson, Cody Long, Brent Nelson, and Gustavo Salinas. Axion reheating in the string landscape. *Phys. Rev. D*, 99(8):086014, 2019. doi: 10.1103/PhysRevD.99.086014.

[135] Joseph P. Conlon. The QCD axion and moduli stabilisation. *JHEP*, 05:078, 2006. doi: 10.1088/1126-6708/2006/05/078.

[136] Mehmet Demirtas, Naomi Gendler, Cody Long, Liam McAllister, and Jakob Moritz. PQ axiverse. *JHEP*, 06:092, 2023. doi: 10.1007/JHEP06(2023)092.

[137] Naomi Gendler and David J. E. Marsh. Qcd axion dark matter in string theory: Haloscopes and helioscopes as probes of the landscape, 2024. URL https://arxiv.org/abs/2407.07143.

[138] James Halverson, Cody Long, Brent Nelson, and Gustavo Salinas. Towards string theory expectations for photon couplings to axionlike particles. *Phys. Rev. D*, 100(10):106010, 2019. doi: 10.1103/PhysRevD.100.106010.

[139] Naomi Gendler, David J. E. Marsh, Liam McAllister, and Jakob Moritz. Glimmers from the axiverse, 2023. URL https://arxiv.org/abs/2309.13145.

[140] Yi-Nan Wang. On the elliptic calabi-yau fourfold with maximal h1, 1. *Journal of High Energy Physics*, 2020(5):1–32, 2020.

[141] James Halverson, Cody Long, and Benjamin Sung. Algorithmic universality in F-theory compactifications. *Phys. Rev. D*, 96(12):126006, 2017. doi: 10.1103/PhysRevD.96.126006.

[142] Washington Taylor and Yi-Nan Wang. Scanning the skeleton of the 4D F-theory landscape. *JHEP*, 01:111, 2018. doi: 10.1007/JHEP01(2018)111.

[143] Maximilian Kreuzer and Harald Skarke. Complete classification of reflexive polyhedra in four-dimensions. *Adv. Theor. Math. Phys.*, 4:1209–1230, 2000. doi: 10.4310/ATMP.2000.v4.n6.a2.

[144] Mehmet Demirtas, Cody Long, Liam McAllister, and Mike Stillman. The Kreuzer-Skarke Axiverse. *JHEP*, 04:138, 2020. doi: 10.1007/JHEP04(2020)138.

[145] Seung-Joo Lee, Wolfgang Lerche, and Timo Weigand. Emergent strings from infinite distance limits. *Journal of High Energy Physics*, 2022(2):1–105, 2022.

[146] Hirosi Ooguri and Cumrun Vafa. On the geometry of the string landscape and the swampland. *Nuclear physics B*, 766(1-3):21–33, 2007.

[147] P. J. E. Peebles. *The Large-Scale Structure of the Universe*. Princeton University Press, 1980.

[148] M. Davis and P. J. E. Peebles. A survey of galaxy redshifts. V. The two-point position and velocity correlations. ApJ, 267:465–482, April 1983. doi: 10.1086/160884.

[149] Alina Kiessling, Marcello Cacciato, Benjamin Joachimi, Donnacha Kirk, Thomas D. Kitching, Adrienne Leonard, Rachel Mandelbaum, Björn Malte Schäfer, Cristóbal Sifón, Michael L. Brown, and Anais Rassat. Galaxy alignments: Theory, modelling and simulations. *Space Science Reviews*, 193(1–4):67–136, September 2015. ISSN 1572-

9672. doi: 10.1007/s11214-015-0203-6. URL http://dx.doi.org/10.1007/s11214-015-0203-6.

[150] S. Bridle and L. King. Dark energy constraints from cosmic shear power spectra: impact of intrinsic alignments on photometric redshift requirements. *New Journal of Physics*, 9:444, December 2007. doi: 10.1088/1367-2630/9/12/444.

[151] Jonathan Blazek, Zvonimir Vlah, and Uroš Seljak. Tidal alignment of galaxies. *Journal of Cosmology and Astroparticle Physics*, 2015(08):015, aug 2015. doi: 10.1088/1475-7516/2015/08/015. URL https://dx.doi.org/10.1088/1475-7516/2015/08/015.

[152] Jonathan A. Blazek, Niall MacCrann, M. A. Troxel, and Xiao Fang. Beyond linear galaxy alignments. *Phys. Rev. D*, 100:103506, Nov 2019. doi: 10.1103/PhysRevD.100.103506. URL https://link.aps.org/doi/10.1103/PhysRevD.100.103506.

[153] Zvonimir Vlah, Nora Elisa Chisari, and Fabian Schmidt. An EFT description of galaxy intrinsic alignments. J. Cosmology Astropart. Phys., 2020(1):025, January 2020. doi: 10.1088/1475-7516/2020/01/025.

[154] Zvonimir Vlah, Nora Elisa Chisari, and Fabian Schmidt. Galaxy shape statistics in the effective field theory. J. Cosmology Astropart. Phys., 2021(5):061, May 2021. doi: 10.1088/1475-7516/2021/05/061.

[155] Francisco Maion, Raul E Angulo, Thomas Bakx, Nora Elisa Chisari, Toshiki Kurita, and Marcos Pellejero-Ibáñez. Hymalaia: a hybrid lagrangian model for intrinsic alignments. *Monthly Notices of the Royal Astronomical Society*, 531(2):2684–2700, May 2024. ISSN 1365-2966. doi: 10.1093/mnras/stae1331. URL http://dx.doi.org/10.1093/mnras/stae1331.

[156] Thomas Bakx, Toshiki Kurita, Nora Elisa Chisari, Zvonimir Vlah, and Fabian Schmidt. Effective field theory of intrinsic alignments at one loop order: a comparison to dark matter simulations. J. Cosmology Astropart. Phys., 2023(10):005, October 2023. doi: 10.1088/1475-7516/2023/10/005.

[157] Shi-Fan Chen and Nickolas Kokron. A lagrangian theory for galaxy shape statistics. *Journal of Cosmology and Astroparticle Physics*, 2024(01):027, jan 2024. doi: 10.1088/1475-7516/2024/01/027. URL https://doi.org/10.1088/1475-7516/2024/01/027.

[158] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, July 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abf7ba. URL http://dx.doi.org/10.3847/1538-4357/abf7ba.

[159] 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, 2021.

[160] A. Pillepich, V. Springel, D. Nelson, S. Genel, J. Naiman, R. Pakmor, L. Hernquist, P. Torrey, M. Vogelsberger, R. Weinberger, and F. Marinacci. Simulating galaxy formation with the IllustrisTNG model. MNRAS, 473:4077–4106, January 2018. doi: 10.1093/mnras/stx2656.

[161] Ana Maria Delgado, Boryana Hadzhiyska, Sownak Bose, Volker Springel, Lars Hernquist, Monica Barrera, Rüdiger Pakmor, Fulvio Ferlito, Rahul Kannan, César Hernández-Aguayo, Simon D. M. White, and Carlos Frenk. The MillenniumTNG project: intrinsic alignments of galaxies and haloes. MNRAS, 523(4):5899–5914, August 2023. doi: 10.1093/mnras/stad1781.

[162] A. Tenneti, R. Mandelbaum, and T. Di Matteo. Intrinsic alignments of disc and elliptical galaxies in the MassiveBlack-II and Illustris simulations. MNRAS, 462:2668–2680, November 2016. doi: 10.1093/mnras/stw1823.

[163] S. Samuroff, R. Mandelbaum, and J. Blazek. Advances in constraining intrinsic alignment models with hydrodynamic simulations. MNRAS, 508(1):637–664, November 2021. doi: 10.1093/mnras/stab2520.

[164] M. L. van Heukelum, D. Neumann, M. García Escobar, N. E. Chisari, and H. Hoekstra. Intrinsic alignments of galaxies in multiple projections. *Astronomy and; Astrophysics*, 706: A148, February 2026. ISSN 1432-0746. doi: 10.1051/0004-6361/202557167. URL http://dx.doi.org/10.1051/0004-6361/202557167.

[165] Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah, Tom Charnock, Elena Giusarma, Yin Li, Erwan Allys, Antoine Brochard, Cora Uhlemann, Chi-Ting Chiang, Siyu He, Alice Pisani, Andrej Obuljen, Yu Feng, Emanuele Castorina, Gabriella Contardo, Christina D. Kreisch, Andrina Nicola, Justin Alsing, Roman Scoccimarro, Licia Verde, Matteo Viel, Shirley Ho, Stephane Mallat, Benjamin Wandelt, and David N. Spergel. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250(1):2, August 2020. ISSN 1538-4365. doi: 10.3847/1538-4365/ab9d82. URL http://dx.doi.org/10.3847/1538-4365/ab9d82.

[166] B. Joachimi, E. Semboloni, S. Hilbert, P. E. Bett, J. Hartlap, H. Hoekstra, and P. Schneider. Intrinsic galaxy shapes and alignments – II. Modelling the intrinsic alignment contamination of weak lensing surveys. *Monthly Notices of the Royal Astronomical Society*, 436(1):819–838, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1618. URL https://doi.org/10.1093/mnras/stt1618.

[167] Kai Hoffmann, Lucas F. Secco, Jonathan Blazek, Martin Crocce, Pau Tallada-Crespí, Simon Samuroff, Judit Prat, Jorge Carretero, Pablo Fosalba, Enrique Gaztañaga, Francisco J. Castander, and DES Collaboration. Modeling intrinsic galaxy alignment in the MICE simulation. Phys. Rev. D, 106(12):123510, December 2022. doi: 10.1103/PhysRevD.106.123510.

[168] J. A. Peacock and R. E. Smith. Halo occupation numbers and galaxy bias. *Monthly Notices of the Royal Astronomical Society*, 318(4):1144–1156, November 2000. ISSN 1365-2966. doi: 10.1046/j.1365-8711.2000.03779.x. URL http://dx.doi.org/10.1046/j.1365-8711.2000.03779.x.

[169] U. Seljak. Analytic model for galaxy and dark matter clustering. *Monthly Notices of the Royal Astronomical Society*, 318(1):203–213, October 2000. ISSN 1365-2966. doi: 10.1046/j.1365-8711.2000.03715.x. URL http://dx.doi.org/10.1046/j.1365-8711.2000.03715.x.

[170] Andreas A. Berlind and David H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ, 575(2):587–616, August 2002. doi: 10.1086/341469.

[171] Idit Zehavi, Zheng Zheng, David H. Weinberg, Michael R. Blanton, Neta A. Bahcall, Andreas A. Berlind, Jon Brinkmann, Joshua A. Frieman, James E. Gunn, Robert H. Lupton, Robert C. Nichol, Will J. Percival, Donald P. Schneider, Ramin A. Skibba, Michael A. Strauss, Max Tegmark, and Donald G. York. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. ApJ, 736(1):59, July 2011. doi: 10.1088/0004-637X/736/1/59.

[172] J. Coupon, M. Kilbinger, H. J. McCracken, O. Ilbert, S. Arnouts, Y. Mellier, U. Abbas, S. de la Torre, Y. Goranova, P. Hudelot, J.-P. Kneib, and O. Le Fèvre. Galaxy clustering in the cfhtls-wide: the changing relationship between galaxies and haloes since z 1.2. *Astronomy and Astrophysics*, 542:A5, May 2012. ISSN 1432-0746. doi: 10.1051/0004-6361/201117625. URL http://dx.doi.org/10.1051/0004-6361/201117625.

[173] John K. Parejko, Tomomi Sunayama, Nikhil Padmanabhan, David A. Wake, Andreas A. Berlind, Dmitry Bizyaev, Michael Blanton, Adam S. Bolton, Frank van den Bosch, Jon Brinkmann, Joel R. Brownstein, Luiz Alberto Nicolaci da Costa, Daniel J. Eisenstein, Hong Guo, Eyal Kazin, Marcio Maia, Elena Malanushenko, Claudia Maraston, Cameron K. McBride, Robert C. Nichol, Daniel J. Oravetz, Kaike Pan, Will J. Percival, Francisco Prada, Ashley J. Ross, Nicholas P. Ross, David J. Schlegel, Don Schneider, Audrey E. Simmons, Ramin Skibba, Jeremy Tinker, Rita Tojeiro, Benjamin A. Weaver, Andrew Wetzel, Martin White, David H. Weinberg, Daniel Thomas, Idit Zehavi, and Zheng Zheng. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: the low-redshift sample. MNRAS, 429(1):98–112, February 2013. doi: 10.1093/mnras/sts314.

[174] Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy evolution from halo occupation distribution modeling of deep2 and sdss galaxy clustering. *The Astrophysical Journal*, 667(2):760–779, October 2007. ISSN 1538-4357. doi: 10.1086/521074. URL http://dx.doi.org/10.1086/521074.

[175] Andrew P. Hearin, Andrew R. Zentner, Frank C. van den Bosch, Duncan Campbell, and Erik Tollerud. Introducing decorated hods: modelling assembly bias in the galaxy–halo connection. *Monthly Notices of the Royal Astronomical Society*, 460(3):2552–2570, May 2016. ISSN 1365-2966. doi: 10.1093/mnras/stw840. URL http://dx.doi.org/10.1093/mnras/stw840.

[176] Risa H. Wechsler and Jeremy L. Tinker. The connection between galaxies and their dark matter halos. *Annual Review of Astronomy and Astrophysics*, 56(1):435–487, September 2018. ISSN 1545-4282. doi: 10.1146/annurev-astro-081817-051756. URL http://dx.doi.org/10.1146/annurev-astro-081817-051756.

[177] Harry Johnston, Christos Georgiou, Benjamin Joachimi, Henk Hoekstra, Nora Elisa Chisari, Daniel Farrow, Maria Cristina Fortuna, Catherine Heymans, Shahab Joudaki, Konrad Kuijken, and Angus Wright. KiDS+GAMA: Intrinsic alignment model constraints for current and future weak lensing cosmology. A&A, 624:A30, April 2019. doi: 10.1051/0004-6361/201834714.

[178] Maria Cristina Fortuna, Andrej Dvornik, Henk Hoekstra, Benjamin Joachimi, Christos Georgiou, Benjamin Giblin, Catherine Heymans, Hendrik Hildebrandt, Arun Kannawadi, Konrad Kuijken, and Angus H. Wright. KiDS-1000: Weak lensing and intrinsic alignment around luminous red galaxies. *arXiv e-prints*, art. arXiv:2409.15416, September 2024.

[179] S. Samuroff, R. Mandelbaum, J. Blazek, A. Campos, N. MacCrann, G. Zacharegkas, A. Amon, J. Prat, S. Singh, J. Elvin-Poole, A. J. Ross, A. Alarcon, E. Baxter, K. Bechtol, M. R. Becker, G. M. Bernstein, A. Carnero Rosell, M. Carrasco Kind, R. Cawthon, C. Chang, R. Chen, A. Choi, M. Crocce, C. Davis, J. DeRose, S. Dodelson, C. Doux, A. Drlica-Wagner, K. Eckert, S. Everett, A. Ferté, M. Gatti, G. Giannini, D. Gruen, R. A. Gruendl, I. Harrison, K. Herner, E. M. Huff, M. Jarvis, N. Kuropatkin, P. F. Leget, P. Lemos, J. McCullough, J. Myles, A. Navarro-Alsina, S. Pandey, A. Porredon, M. Raveri, M. Rodriguez-Monroy, R. P. Rollins, A. Roodman, G. Rossi, E. S. Rykoff, C. Sánchez, L. F. Secco, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. A. Troxel, I. Tutusaus, N. Weaverdyck, B. Yanny, B. Yin, Y. Zhang, and J. Zuntz. The Dark Energy Survey Year 3 and eBOSS: constraining galaxy intrinsic alignments across luminosity and colour space. MNRAS, 524(2):2195–2223, September 2023. doi: 10.1093/mnras/stad2013.

[180] Christos Georgiou, Nora Elisa Chisari, Maciej Bilicki, Francesco La Barbera, Nicola R. Napolitano, Nivya Roy, and Crescenzo Tortora. Intrinsic galaxy alignments in the KiDS-1000 bright sample: dependence on colour, luminosity, morphology, and galaxy scale. *arXiv e-prints*, art. arXiv:2502.09452, February 2025.

[181] J. Siegel, J. McCullough, A. Amon, C. Lamman, N. Jeffrey, B. Joachimi, H. Hoekstra, S. Heydenreich, A. J. Ross, J. Aguilar, S. Ahlen, D. Bianchi, C. Blake, D. Brooks, F. J.

Castander, T. Claybaugh, A. de la Macorra, J. DeRose, P. Doel, N. Emas, S. Ferraro, A. Font-Ribera, J. E. Forero-Romero, E. Gaztañaga, S. Gontcho A Gontcho, G. Gutierrez, K. Honscheid, M. Ishak, S. Joudaki, R. Kehoe, D. Kirkby, T. Kisner, A. Krolewski, O. Lahav, A. Lambert, M. Landriau, L. Le Guillou, M. E. Levi, M. Manera, A. Meisner, R. Miquel, J. Moustakas, S. Nadathur, J. A. Newman, G. Niz, N. Palanque-Delabrouille, W. J. Percival, A. Porredon, F. Prada, I. Pérez-Ràfols, G. Rossi, E. Sanchez, C. Saulder, D. Schlegel, M. Schubnell, A. Semenaite, J. Silber, D. Sprayberry, Z. Sun, G. Tarlé, B. A. Weaver, R. Zhou, and H. Zou. Intrinsic alignment demographics for next-generation lensing: Revealing galaxy property trends with DESI Y1 direct measurements. *arXiv e-prints*, art. arXiv:2507.11530, July 2025.

[182] D. Navarro-Gironés, M. Crocce, E. Gaztañaga, A. Wittje, M. Siudek, H. Hoekstra, H. Hildebrandt, B. Joachimi, R. Paviot, C. M. Baugh, J. Carretero, R. Casas, F. J. Castander, M. Eriksen, E. Fernandez, P. Fosalba, J. García-Bellido, R. Miquel, C. Padilla, P. Renard, E. Sánchez, S. Serrano, I. Sevilla-Noarbe, and P. Tallada-Crespí. The PAU Survey: Measuring intrinsic galaxy alignments in deep wide fields as a function of colour, luminosity, stellar mass and redshift. *arXiv e-prints*, art. arXiv:2505.15470, May 2025.

[183] Michael D. Schneider and Sarah Bridle. A halo model for intrinsic alignments of galaxy ellipticities. *Monthly Notices of the Royal Astronomical Society*, 402(4):2127–2139, March 2010. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.15956.x. URL http://dx.doi.org/10.1111/j.1365-2966.2009.15956.x.

[184] B. Joachimi, R. Mandelbaum, F. B. Abdalla, and S. L. Bridle. Constraints on intrinsic alignment contamination of weak lensing surveys using the megaz-lrg sample. *Astronomy and Astrophysics*, 527:A26, January 2011. ISSN 1432-0746. doi: 10.1051/0004-6361/201015621. URL http://dx.doi.org/10.1051/0004-6361/201015621.

[185] Andrew P. Hearin, Duncan Campbell, Erik Tollerud, Peter Behroozi, Benedikt Diemer, Nathan J. Goldbaum, Elise Jennings, Alexie Leauthaud, Yao-Yuan Mao, Surhud More, John Parejko, Manodeep Sinha, Brigitta Sipöcz, and Andrew Zentner. Forward modeling of large-scale structure: An open-source approach with halotools. *The Astronomical Journal*, 154(5):190, October 2017. ISSN 1538-3881. doi: 10.3847/1538-3881/aa859f. URL http://dx.doi.org/10.3847/1538-3881/aa859f.

[186] Cora Dvorkin, Siddharth Mishra-Sharma, Brian Nord, V. Ashley Villar, Camille Avestruz, Keith Bechtol, Aleksandra Ćiprijanović, Andrew J. Connolly, Lehman H. Garrison, Gautham Narayan, and Francisco Villaescusa-Navarro. Machine learning and cosmology, 2022.

[187] DES Collaboration, T. M. C. Abbott, M. Adamow, M. Aguena, A. Alarcon, S. S. Allam, O. Alves, A. Amon, D. Anbajagane, F. Andrade-Oliveira, S. Avila, D. Bacon, E. J. Baxter, J. Beas-Gonzalez, K. Bechtol, M. R. Becker, G. M. Bernstein, E. Bertin, J. Blazek, S. Bocquet, D. Brooks, D. Brout, H. Camacho, G. Camacho-Ciurana, R. Camilleri, G. Campailla, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, P. Carrilho, F. J. Castander, R. Cawthon, C. Chang, A. Choi, J. M. Coloma-Nadal, M. Costanzi, M. Crocce, W. d'Assignies, L. N. da Costa, M. E. da Silva Pereira, T. M. Davis, J. De Vicente, J. DeRose, H. T. Diehl, S. Dodelson, P. Doel, C. Doux, A. Drlica-Wagner, T. F. Eifler, J. Elvin-Poole, J. Estrada, S. Everett, A. E. Evrard, J. Fang, A. Farahi, A. Ferté, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, G. Giannini, P. Giles, K. Glazebrook, M. Gorsuch, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, I. Harrison, W. G. Hartley, E. Henning, K. Herner, S. R. Hinton, D. L. Hollowood, K. Honscheid, E. M. Huff, D. Huterer, B. Jain, D. J. James, M. Jarvis, N. Jeffrey, T. Jeltema, T. Kacprzak, S. Kent, A. Kovacs, E. Krause, R. Kron, K. Kuehn, O. Lahav, S. Lee, E. Legnani, C. Lidman, H. Lin, N. MacCrann, M. Manera, T. Manning, J. L. Marshall, S. Mau, J. McCullough, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, J. Muir, J. Myles, R. C. Nichol, B. Nord, J. H. O'Donnell, R. L. C. Ogando, A. Palmese, M. Paterno, J. Peoples, W. J. Percival, D. Petravick, A. Pieres, A. A. Plazas Malagón, A. Porredon, A. Pourtsidou, J. Prat, C. Preston, M. Raveri, W. Riquelme, M. Rodriguez-Monroy, P. Rogozenski, A. K. Romer, A. Roodman, R. Rosenfeld, A. J. Ross, E. Rozo, E. S. Rykoff, S. Samuroff, C. Sánchez, E. Sanchez, D. Sanchez Cid, T. Schutt, I. Sevilla-Noarbe, E. Sheldon, N. Sherman, T. Shin, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, M. Tabbutt, G. Tarle, D. Thomas, C. To, A. Tong, L. Toribio San Cipriano, M. A. Troxel, M. Tsedrik, D. L. Tucker, V. Vikram, A. R. Walker, N. Weaverdyck, R. H. Wechsler, D. H. Weinberg, J. Weller, V. Wetzell, A. Whyley, R. D. Wilkinson, P. Wiseman, H. Y. Wu, M. Yamamoto, B. Yanny, B. Yin, G. Zacharegkas, Y. Zhang, and J. Zuntz. Dark energy survey year 6 results: Cosmological constraints from galaxy clustering and weak lensing, 2026. URL https://arxiv.org/abs/2601.14559.

[188] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-

softmax, 2017. URL https://arxiv.org/abs/1611.01144.

[189] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations*, 2017.

[190] Zhongxu Zhai, Jeremy L. Tinker, Matthew R. Becker, Joseph DeRose, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Risa H. Wechsler. The aemulus project. iii. emulation of the galaxy correlation function. *The Astrophysical Journal*, 874(1):95, March 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/ab0d7b. URL http://dx.doi.org/10.3847/1538-4357/ab0d7b.

[191] Juliana Kwan, Shun Saito, Alexie Leauthaud, Katrin Heitmann, Salman Habib, Nicholas Frontiere, Hong Guo, Song Huang, Adrian Pope, and Sergio Rodriguéz-Torres. Galaxy clustering in the mira-titan universe. i. emulators for the redshift space galaxy correlation function and galaxy–galaxy lensing. *The Astrophysical Journal*, 952(1):80, July 2023. ISSN 1538-4357. doi: 10.3847/1538-4357/acd92f. URL http://dx.doi.org/10.3847/1538-4357/acd92f.

[192] Earl Lawrence, Katrin Heitmann, Martin White, David Higdon, Christian Wagner, Salman Habib, and Brian Williams. The Coyote Universe. III. Simulation Suite and Precision Emulator for the Nonlinear Matter Power Spectrum. ApJ, 713(2):1322–1331, April 2010. doi: 10.1088/0004-637X/713/2/1322.

[193] Giovanni Aricò, Raul E. Angulo, and Matteo Zennaro. Accelerating Large-Scale-Structure data analyses by emulating Boltzmann solvers and Lagrangian Perturbation Theory. *arXiv e-prints*, art. arXiv:2104.14568, April 2021. doi: 10.48550/arXiv.2104.14568.

[194] Giovanni Aricò, Raul E. Angulo, Sergio Contreras, Lurdes Ondaro-Mallea, Marcos Pellejero-Ibañez, and Matteo Zennaro. The BACCO simulation project: a baryonification emulator with neural networks. MNRAS, 506(3):4070–4082, September 2021. doi: 10.1093/mnras/stab1911.

[195] 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, August 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac2083. URL http://dx.doi.org/10.1093/mnras/stac2083.

[196] Yesukhei Jagvaral, François Lanusse, and Rachel Mandelbaum. Geometric deep learning for galaxy-halo connection: a case study for galaxy intrinsic alignments. *Monthly Notices of the Royal Astronomical Society*, 542(3):2560–2571, 04 2025. ISSN 0035-8711. doi: 10.1093/mnras/staf592. URL https://doi.org/10.1093/mnras/staf592.

[197] Yesukhei Jagvaral, Francois Lanusse, and Rachel Mandelbaum. Unified framework for diffusion generative models in SO(3): applications in computer vision and astrophysics. *arXiv e-prints*, art. arXiv:2312.11707, December 2023. doi: 10.48550/arXiv.2312.11707.

[198] Nicholas Van Alfen, Duncan Campbell, Andrew Hearin, and Jonathan Blazek. Halotools: A new release adding intrinsic alignments to halo-based methods. *Journal of Open Source Software*, 10(107):7421, 2025. doi: 10.21105/joss.07421. URL https://doi.org/10.21105/joss.07421.

[199] Andrew P. Hearin, Duncan Campbell, Erik Tollerud, Peter Behroozi, Benedikt Diemer, Nathan J. Goldbaum, Elise Jennings, Alexie Leauthaud, Yao-Yuan Mao, Surhud More, John Parejko, Manodeep Sinha, Brigitta Sipöcz, and Andrew Zentner. Forward modeling of large-scale structure: An open-source approach with halotools. *The Astronomical Journal*, 154 (5):190, oct 2017. doi: 10.3847/1538-3881/aa859f. URL https://dx.doi.org/10.3847/1538-3881/aa859f.

[200] Asantha Cooray and Ravi Sheth. Halo models of large scale structure. Phys. Rep., 372(1): 1–129, December 2002. doi: 10.1016/S0370-1573(02)00276-4.

[201] Marika Asgari, Alexander J. Mead, and Catherine Heymans. The halo model for cosmology: a pedagogical review. *The Open Journal of Astrophysics*, 6:39, November 2023. doi: 10.21105/astro.2303.08752.

[202] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The structure of cold dark matter halos. *The Astrophysical Journal*, 462:563, May 1996. ISSN 1538-4357. doi: 10.1086/177173. URL http://dx.doi.org/10.1086/177173.

[203] G. S. Watson. Equatorial distributions on a sphere. *j-BIOMETRIKA*, 52(1/2):193–201, June 1965. ISSN 0006-3444 (print), 1464-3510 (electronic). doi: https://doi.org/10.2307/2333824. URL http://www.jstor.org/stable/2333824.

[204] Stephen D. Landy and Alexander S. Szalay. Bias and Variance of Angular Correlation Functions. ApJ, 412:64, July 1993. doi: 10.1086/172900.

[205] Sukhdeep Singh, Rachel Mandelbaum, Uroš Seljak, Anže Slosar, and Jose Vazquez Gonzalez. Galaxy–galaxy lensing estimators and their covariance properties. *Monthly Notices of the Royal Astronomical Society*, 471(4):3827–3844, 07 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx1828. URL https://doi.org/10.1093/mnras/stx1828.

[206] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. ApJ, 740(2):102, October 2011. doi: 10.1088/0004-637X/740/2/102.

[207] Nora Elisa Chisari and Cora Dvorkin. Cosmological information in the intrinsic alignments of luminous red galaxies. *Journal of Cosmology and Astroparticle Physics*, 2013(12): 029–029, December 2013. ISSN 1475-7516. doi: 10.1088/1475-7516/2013/12/029. URL http://dx.doi.org/10.1088/1475-7516/2013/12/029.

[208] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. doi: 10.1109/ICNN.1994.374138.

[209] G. M. Bernstein and M. Jarvis. Shapes and shears, stars and smears: Optimal measurements for weak lensing. *The Astronomical Journal*, 123(2):583, feb 2002. doi: 10.1086/338085. URL https://dx.doi.org/10.1086/338085.

[210] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

[211] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.

[212] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[213] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[214] Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of mean variance estimation neural networks, 2023. URL https://arxiv.org/abs/2302.08875.

[215] Anya Paopiamsap, Natalia Porqueres, David Alonso, Joachim Harnois-Deraps, and C. Danielle Leonard. Accuracy requirements on intrinsic alignments for stage-iv cosmic shear. *The Open Journal of Astrophysics*, 7, May 2024. ISSN 2565-6120. doi: 10.33232/001c.117419. URL http://dx.doi.org/10.33232/001c.117419.

[216] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018. URL https://arxiv.org/abs/1705.07115.

[217] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://dx.doi.org/10.1007/s10994-021-05946-3.

[218] Daniela Grandón and Elena Sellentin. Differentiable predictions for large scale structure with shamnet. *The Open Journal of Astrophysics*, 5(1), July 2022. ISSN 2565-6120. doi: 10.21105/astro.2205.11587. URL http://dx.doi.org/10.21105/astro.2205.11587.

[219] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599.

[220] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

[221] J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix. *Astronomy and; Astrophysics*, 464(1):399–404, December 2006. ISSN 1432-0746. doi: 10.1051/0004-6361:20066170. URL http://dx.doi.org/10.1051/0004-6361:20066170.

[222] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, 2011. URL https://arxiv.org/abs/1111.4246.

[223] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[224] Sihan Yuan, Daniel J Eisenstein, and Lehman H Garrison. Exploring the squeezed three-point galaxy correlation function with generalized halo occupation distribution models. *Monthly Notices of the Royal Astronomical Society*, 478(2):2019–2033, April 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty1089. URL http://dx.doi.org/10.1093/mnras/sty1089.

[225] Yin Li, Libin Lu, Chirag Modi, Drew Jamieson, Yucheng Zhang, Yu Feng, Wenda Zhou, Ngai Pok Kwan, François Lanusse, and Leslie Greengard. pmwd: A differentiable cosmological particle-mesh n-body library, 2022. URL https://arxiv.org/abs/2211.09958.

[226] Nicholas Van Alfen, Jonathan Blazek, and Andrew Hearin. Coherent 2D and 3D simulations of intrinsic alignments. *in prep.*, 2025.

[227] Isabele Souza Vitório, Michael Buehlmann, Eve Kovacs, Patricia Larsen, Nicholas Frontiere, and Katrin Heitmann. Exploring the core-galaxy connection, 2025. URL https://arxiv.org/abs/2407.00268.

[228] Shivam Pandey, Chirag Modi, Benjamin D. Wandelt, Deaglan J. Bartlett, Adrian E. Bayer, Greg L. Bryan, Matthew Ho, Guilhem Lavaux, T. Lucas Makinen, and Francisco Villaescusa-Navarro. Charm: Creating halos with auto-regressive multi-stage networks, 2024. URL https://arxiv.org/abs/2409.09124.

[229] Yesukhei Jagvaral, Francois Lanusse, and Rachel Mandelbaum. DIFFUSION GENERATIVE MODELS ON SO(3), 2023. URL https://openreview.net/forum?id=jHA-yCyBGb.

[230] Andreas Schanz, Florian List, and Oliver Hahn. Stochastic super-resolution of cosmological simulations with denoising diffusion models, 2023. URL https://arxiv.org/abs/2310.06929.

[231] Nayantara Mudur, Carolina Cuesta-Lazaro, and Douglas P. Finkbeiner. Diffusion-hmc: Parameter inference with diffusion model driven hamiltonian monte carlo, 2024. URL https://arxiv.org/abs/2405.05255.

[232] Abolfazl Farahani, Sahar Voghoei, Khaled M. Rasheed, and Hamid Reza Arabnia. A brief review of domain adaptation. *ArXiv*, abs/2010.03978, 2020. URL https://api.semanticscholar.org/CorpusID:222209143.

[233] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives. *arXiv e-prints*, art. arXiv:2208.07422, August 2022. doi: 10.48550/arXiv.2208.07422.

[234] Samuel F. Dodge and Lina Karam. Understanding how image quality affects deep neural networks. *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016. URL https://api.semanticscholar.org/CorpusID:12047850.

[235] Milind S. Gide, Samuel F. Dodge, and Lina J. Karam. The effect of distortions on the prediction of visual attention, 2016. URL https://arxiv.org/abs/1604.03882.

[236] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise, 2019. URL https://arxiv.org/abs/1901.10513.

[237] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, October 2019. ISSN 1941-0026. doi: 10.1109/tevc.2019.2890858. URL http://dx.doi.org/10.1109/TEVC.2019.2890858.

[238] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, October 2018. ISSN 0925-2312. doi: 10.1016/j.neucom.2018.05.083. URL http://dx.doi.org/10.1016/j.neucom.2018.05.083.

[239] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), July 2020. ISSN 2157-6904. doi: 10.1145/3400066. URL https://doi.org/10.1145/3400066.

[240] Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A kernel method for the two-sample problem, 2008. URL https://arxiv.org/abs/0805.2368.

[241] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *ArXiv*, abs/1612.01939, 2016. URL https://api.semanticscholar.org/CorpusID:10084602.

[242] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, 2019. URL https://api.semanticscholar.org/CorpusID:57572938.

[243] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214%2Faoms%2F1177729694.

[244] Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 2018. URL https://api.semanticscholar.org/CorpusID:88523547.

[245] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1853–1865, 2014. URL https://api.semanticscholar.org/CorpusID:13347901.

[246] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie,

Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

[247] N. Jeffrey, L. Whiteway, M. Gatti, J. Williamson, et al. Dark Energy Survey Year 3 results: likelihood-free, simulation-based wCDM inference with neural compression of weak-lensing map statistics. *Monthly Notices of the Royal Astronomical Society*, 535(1):1–21, 2024. doi: 10.1093/mnras/stae2338.

[248] Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gan-

gler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole

Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang, Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, March 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/ab042c. URL http://dx.doi.org/10.3847/1538-4357/ab042c.

[249] 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, 2021. URL https://arxiv.org/abs/1812.05609.

[250] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, July 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abf7ba. URL http://dx.doi.org/10.3847/1538-4357/abf7ba.

[251] Aleksandra Ćiprijanović, Diana Kafkes, Gregory Snyder, F. Javier Sánchez, Gabriel Nathan Perdue, Kevin Pedro, Brian Nord, Sandeep Madireddy, and Stefan M. Wild. DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification. *Machine Learning: Science and Technology*, 3(3):035007, September 2022. doi: 10.1088/2632-2153/ac7f1a.

[252] A. Ćiprijanović, A. Lewis, K. Pedro, S. Madireddy, B. Nord, G. N. Perdue, and S. M. Wild. DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy mor-

phology classification and anomaly detection. *Machine Learning: Science and Technology*, 4(2):025013, June 2023. doi: 10.1088/2632-2153/acca5f.

[253] Ricardo Vilalta, Kinjal Dhar Gupta, Dainis Boumber, and Mikhail M. Meskhi. A general approach to domain adaptation with applications in astronomy. *Publications of the Astronomical Society of the Pacific*, 131(1004):108008, September 2019. ISSN 1538-3873. doi: 10.1088/1538-3873/aaf1fc. URL http://dx.doi.org/10.1088/1538-3873/aaf1fc.

[254] Andrea Roncoli, Aleksandra Ćiprijanović, Maggie Voetberg, Francisco Villaescusa-Navarro, and Brian Nord. Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets. *arXiv e-prints*, art. arXiv:2311.01588, November 2023. doi: 10.48550/arXiv.2311.01588.

[255] Paxson Swierc, Megan Zhao, Aleksandra Ćiprijanović, and Brian Nord. Domain Adaptation for Measurements of Strong Gravitational Lenses. *arXiv e-prints*, art. arXiv:2311.17238, November 2023. doi: 10.48550/arXiv.2311.17238.

[256] Shrihan Agarwal, Aleksandra Ćiprijanović, and Brian D. Nord. Neural Network Prediction of Strong Lensing Systems with Domain Adaptation and Uncertainty Quantification. *arXiv e-prints*, art. arXiv:2411.03334, October 2024. doi: 10.48550/arXiv.2411.03334.

[257] Sankalp Gilda, Antoine de Mathelin, Sabine Bellstedt, and Guillaume Richard. Unsupervised Domain Adaptation for Constraining Star Formation Histories. *Astronomy*, 3(3):189–207, July 2024. doi: 10.3390/astronomy3030012.

[258] Hanna Parul, Sergei Gleyzer, Pranath Reddy, and Michael W. Toomey. Domain adaptation in application to gravitational lens finding. *arXiv e-prints*, art. arXiv:2410.01203, October 2024. doi: 10.48550/arXiv.2410.01203.

[259] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69:1173–1185, 2021. URL https://api.semanticscholar.org/CorpusID:231951465.

[260] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016. doi: 10.1109/MGRS.2016.2548504.

[261] Johannes Jakubik, Michal Muszynski, Michael Vössing, Niklas Kühl, and Thomas Brunschwiler. Unsupervised Domain Adaptation for Geospatial Semantic Segmentation. In *AGU Fall Meeting Abstracts*, volume 2023, pages IN53A–07, December 2023.

[262] Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. Domain adaptation based enhanced detection for autonomous driving in foggy and rainy weather. *ArXiv*, abs/2307.09676, 2023. URL https://api.semanticscholar.org/CorpusID:271403939.

[263] Rui Ding, Jianguo Liu, Kang Hua, Xuebin Wang, Xiaoben Zhang, Minhua Shao, Yuxin Chen, and Junhong Chen. Leveraging data mining, active learning, and domain adaptation for efficient discovery of advanced oxygen evolution electrocatalysts. *Science Advances*, 11(14):eadr9038, 2025. doi: 10.1126/sciadv.adr9038. URL https://www.science.org/doi/abs/10.1126/sciadv.adr9038.

[264] Alexander Bogatskiy, Brandon Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics, 2020. URL https://arxiv.org/abs/2006.04780.

[265] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020. URL https://arxiv.org/abs/2006.10503.

[266] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns, 2018. URL https://arxiv.org/abs/1711.06721.

[267] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

[268] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas Guibas. Vector neurons: A general framework for so(3)-equivariant networks, 2021. URL https://arxiv.org/abs/2104.12229.

[269] Ioannis Kalogeropoulos, Giorgos Bouritsas, and Yannis Panagakis. Scale equivariant graph metanetworks, 2024. URL https://arxiv.org/abs/2406.10685.

[270] Srinath Bulusu, Matteo Favoni, Andreas Ipp, David I. Müller, and Daniel Schuh. Equivariance and generalization in neural networks. *EPJ Web of Conferences*, 258:09001, 2022. ISSN 2100-014X. doi: 10.1051/epjconf/202225809001. URL http://dx.doi.org/10.1051/epjconf/202225809001.

[271] Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Bin Shao, and Tie-Yan Liu. Se(3) equivariant graph neural networks with complete local frames, 2022. URL https://arxiv.org/abs/2110.14811.

[272] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, 2018. URL https://arxiv.org/abs/1705.09634.

[273] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

[274] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

[275] Sashank J. Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions, 2014. URL https://arxiv.org/abs/1406.2083.

[276] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

[277] Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance, 2018. URL https://arxiv.org/abs/1610.06447.

[278] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.

[279] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018. URL https://arxiv.org/abs/1810.08278.

[280] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

[281] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density, 2021. URL https://arxiv.org/abs/2108.10860.

[282] Jean Feydy, Joan Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *Advances in Neural Information Processing Systems*, 33, 2020.

[283] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

[284] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

[285] Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. A novel domain adaptation theory with jensen–shannon divergence. *Knowledge-Based Systems*, 257:109808, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109808. URL https://www.sciencedirect.com/science/article/pii/S0950705122009200.

[286] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016. URL https://arxiv.org/abs/1505.07818.

[287] Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J.

Lintott, Kameswara Mantha, Devina Mohan, David O'Ryan, and Inigo V. Slijepevic. Scaling laws for galaxy images, 2024. URL https://arxiv.org/abs/2404.02973.

[288] Kang Liu, Jian Yang, and Shengyang Li. Remote-Sensing Cross-Domain Scene Classification: A Dataset and Benchmark. *Remote Sensing*, 14(18):4635, September 2022. doi: 10.3390/rs14184635.

[289] M. Voetberg, Ashia Livaudais, Becky Nevin, Omari Paul, and Brian Nord. Deepbench: A simulation package for physical benchmarking data. *JOSS*, 6 2023. doi: 10.2172/1989920. URL https://www.osti.gov/biblio/1989920.

[290] J. L. Sersic. Photometry of southern galaxies: NGC 5128. *The Observatory*, 78:24–29, February 1958.

[291] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[292] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.161. URL http://dx.doi.org/10.1109/TPAMI.2010.161.

[293] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1458.

[294] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L Masters, Vihang Mehta, Brooke D Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2093.

[295] Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M M Scaife, Chris Lintott, Karen L Masters, James M Dawson, Hugh Dickinson, Lucy Fortson, Izzy L Garland, Kameswara Mantha, David O'Ryan, Jürgen Popp, Brooke Simmons, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys. *Monthly Notices of the Royal Astronomical Society*, 526(3): 4768–4786, 09 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad2919.

[296] Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WE4qe9xlnQw.

[297] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1911.08251.

[298] Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization, 2020. URL https://arxiv.org/abs/1806.03852.

[299] Rui Wang, Robin Walters, and Rose Yu. Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting, 2022. URL https://arxiv.org/abs/2206.09450.

[300] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups, 2018. URL https://arxiv.org/abs/1802.03690.

[301] Andreas Abildtrup Hansen, Anna Calissano, and Aasa Feragen. Interpreting equivariant representations, 2024. URL https://arxiv.org/abs/2401.12588.

[302] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

[303] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[304] Jacob Carse, Andres Alvarez Olmo, and Stephen McKenna. Calibration of deep medical image classifiers: An empirical comparison using dermatology and histopathology

datasets. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, page 89–99, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-16748-5. doi: 10.1007/978-3-031-16749-2_9. URL https://doi.org/10.1007/978-3-031-16749-2_9.

[305] Alex Cole, Benjamin K. Miller, Samuel J. Witte, Maxwell X. Cai, Meiert W. Grootes, Francesco Nattino, and Christoph Weniger. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *Journal of Cosmology and Astroparticle Physics*, 2022(09):004, September 2022. ISSN 1475-7516. doi: 10.1088/1475-7516/2022/09/004. URL http://dx.doi.org/10.1088/1475-7516/2022/09/004.

[306] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art, 2024. URL https://arxiv.org/abs/2308.01222.

[307] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[308] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[309] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. ISSN 1095-9203. doi: 10.1126/science.290.5500.2319. URL http://dx.doi.org/10.1126/science.290.5500.2319.

[310] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt, 2019. URL https://arxiv.org/abs/1912.02292.

[311] Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics, 2022. URL https://arxiv.org/abs/2201.11969.

[312] Gamaleldin F. Elsayed, Prajit Ramachandran, Jonathon Shlens, and Simon Kornblith. Revisiting spatial invariance with low-rank local connectivity, 2020. URL https://arxiv.org/abs/2002.02959.

[313] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks, 2015. URL https://arxiv.org/abs/1502.02791.

[314] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks, 2017. URL https://arxiv.org/abs/1605.06636.

Appendix A

Other Notable Quotes

"Let's say you have like three protons . . . you can consider that [to be] a very small galaxy."

Jonathan Blazek
***In conversation*, 2025**

"What is this, a frat party!?"

James Halverson
***In response to finding a contraction in a manuscript*, 2024**