Tractable Generative Modelling of Cosmological Numerical Simulations

Amit Parag, Vaishak Belle

Tractable Generative Modelling of Cosmological Numerical Simulations

Amit Parag[a], Vaishak Belle[a,b,*]

[a] *University of Edinburgh, Edinburgh, UK*
[b] *Alan Turing Institute, London, UK.*

Abstract

Cosmological simulations aim to understand the matter distribution in the universe either by following a semi-analytic approach or by formulating a hydrodynamical version of the matter distribution. Both approaches describe the evolution of baryonic structures inside potential wells created by dark matter, while the dark matter itself is modelled as self gravitating collision-less system. While major advancements have been made to reduce the computational costs, these simulations still take millions of CPU hours to converge to a stable set of solutions. This naturally leads to the question: can generative models predict the properties of a galaxy, given a partial history of its dynamical evolution?

Given that computing conditional probabilities is intractable in general, tractable probabilistic models such as sum-product networks have emerged, where conditional marginals can be computed in time linear in the size of the model. In this work, we investigate how sum-product networks can be used to compactly represent and learn distributions for prediction in concordance cosmology. Our results study the extent to which they can infer the relation between baryonic matter and dark matter. We test the algorithm on the Eagle suite of cosmological hydrodynamical simulations to show that graphical models can satisfactorily reproduce mock catalogs of galaxies.

Keywords: Probabilistic Modelling, Cosmological Simulations, Dark Matter, Tractable Graphical Model

1. Introduction

Probabilistic representations, such as Bayesian and Markov networks, are fundamental to statistical machine learning. The attractiveness of such networks is that they can express probabilistic dependencies in a compact manner. However, owing to the intractability of inference, learning also becomes challenging, since learning typically uses inference as a subroutine [21], and moreover, even if such a representation is learned, prediction will suffer because inference has to be approximated. Tractable learning is a powerful new paradigm that attempts to learn representations that support efficient probabilistic querying. Much of the initial work focused on low tree-width models, [2], but later, building on properties such as local structure [10], numerous proposals based on arithmetic circuits (ACs) emerged. These circuit learners can also represent high tree-width models and enable exact inference for a range of queries in time polynomial in the circuit size. Sum-product networks (SPNs) [32], for example, are instances of ACs with an elegant recursive structure; essentially, an SPN is a weighted sum of products of SPNs, and the base case is a leaf node denoting a tractable probability distribution (e.g., a univariate Bernoulli distribution). In so much as deep learning models can be understood as graphical models with multiple hidden variables, SPNs can be seen as a tractable deep architecture. Of course, learning the architecture of standard deep models is very challenging [4], and in contrast, SPNs, by their very design, offer a reliable structure learning paradigm. While it is possible to specify SPNs by hand, weight learning is additionally required to obtain a probability distribution, but also the specification of SPNs has to obey conditions of completeness and decomposability, all of which makes structure learning an obvious choice. Since SPNs were introduced, a number of structure learning frameworks have been developed for those and related data structures, e.g., [14, 18, 23]. (Note that these related structures are not necessarily equivalent in terms of

*Corresponding author. Vaishak Belle is supported by a Royal Society University Research Fellowship.
Email addresses: aparag@laas.fr (Amit Parag), vaishak@ed.ac.uk (Vaishak Belle)

its features and properties; see works such as [23] for discussions. We will focus on SPNs here owing to their simple specification for generative modelling.)

In this work, we study how SPNs can be used to model a novel and challenging problem related to the evolution of galaxies; in particular, we consider the concordance cosmology in the backdrop of hydrodynamical simulations. In essence, in cosmology, there is no scope of actual experiments and therefore, the need to simulate universes to test various cosmological parameters, galactic properties and theories of universes require computationally efficient methods. One of the most effective tools to study astrophysical phenomena are numerical simulations. This is primarily due to the inability to arrive at analytic solutions for gravitationally interacting particles which immediately leads to an investigation of the computational methods. The mutual gravitational interactions between large sets of particles forms the basis of all simulations. This is easy to see: in simulations, dark matter is modelled as particles that can only interact with each other through gravity and since all matter must form in wells of dark matter, n-body simulations of billions of particles are of paramount importance. When juxtaposed with suitable numerical techniques that permit a realistic treatment of baryonic physics, this allows numerical simulations to be used as a tool for validating cosmological models and provide a general picture of structure formation in the universe.

Numerical simulations attempt to describe the cosmological picture by painting galaxies inside dark matter halos. This task, however, is extremely difficult: first, it involves making certain assumptions which cannot be substantiated, [38] and second, the computation process is prohibitively expensive as indicated by the millions of hours of computation times on CPU for simulations such as the Eagle simulations or the Illustris simulations [36, 24]. Various algorithms have been proposed to reduce the exorbitant computations time and increase the accuracy of the process [24, 15]. Scalable machine learning architectures are used in a variety of particle physics and cosmology experiments but their use has been restricted to particle tracing and classification [16]. We believe the role of machine learning can be enhanced extensively in cosmological settings, and in this work, we take first steps towards that goal. In particular, given the recent advances in tractable probabilistic models, we study the question of how those advances could be leveraged in service of yielding a generative model. Thus, we reiterate that the thrust of this paper is purely on the applicability of tractable learning for the identified cosmological domain, and not about extending existing algorithms. As we demonstrate below, this requires a non-trivial understanding of the mathematical concepts surrounding concordance cosmology, and raises considerable challenges in feature selection and engineering. In the long term, we hope our results here will lead to more interdisciplinary research, and enable machine learning and probabilistic logical learning to tackle the deep, existing problems in cosmology.

2. Problem Statement

In this section, we will formulate the problem statement, first by giving a fairly informal picture of the cosmological model, before turning to the equations driving the computational task. (As we expect both physicists and machine learning practitioners to be readers of the article, our exposition will generally follow the style of introducing the overall background at an intuitive level, but then describing the core concepts using mathematical notation.)

Preliminaries. Of the many pictures that can be painted about origins of universe, the widely celebrated one is the Big Bang theory. The universe began with a bang, a *hot big bang*, so much so that the temperatures of the particles within 10^{-33}s of Big Bang far surpasses the particles produced in current high energy physics experiments. Inflation soon took over and morphed space by a factor of 10^{26} over time of the order 10^{-32}s and so began epochs where the finest grains of the universe combined, decoupled and synthesized to establish the order that was to follow. The order that followed is usually described by one particular parametrization of the Big Bang theory, the ΛCDM model.

Concordance Cosmology. The concordance cosmological model is the ΛCDM model of the universe, founded on the Copernican principles of isotropy and homogeneity [34]. To build the concordance model of the universe, we start by assuming that observers in any location in the universe can never be the principle observers of the universe. This simple assumption immediately leads to the proposition that the universe is isotropic and homogeneous.

However a few things need to be kept in mind. Any theory is only as good as the initial assumptions made to form that theory. The properties of isotropicity and homogeneity are only apparent on the largest scales of the universe. As an example, consider the solar system. If we imagine a sphere engulfing the solar system, then the criteria of isotropicity and homogeneity are not met. The sun is located in a specific point in the sphere and on observing the

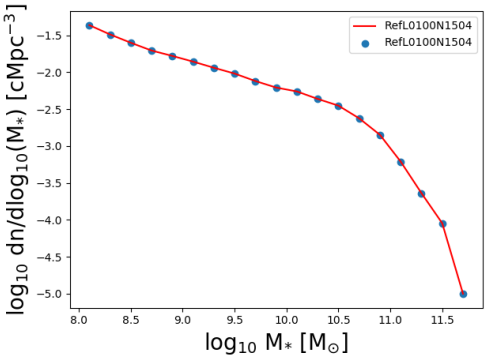

Figure 1: The galaxy stellar mass function (GSMF). The evolution of the stellar mass density of the universe provides a good overview of the growth of stellar mass in the simulation. The GSMF can generally be described approximately by a Schechter function [37], i.e. a power law and an exponential break which starts at a characteristic mass. The figure describes the relation between galaxy stellar mass density and redshift based on data taken from Eagle simulations. The Eagle simulation is described in §4.1

sphere of solar system, we can immediately conclude that the sphere is not homogeneous. Planets in solar system are not alike and the sun outshines them all. Similarly, if we take into account the local group of galaxies – Andromeda, Milky Way, Large and Small Magellanic Clouds with the remaining fifty dwarf satellite galaxies – we observe that Andromeda and Milky Way are the two most pronounced galaxies in terms of mass, luminosity and size while the remaining can be classified as dwarf satellites. Only at largest scales of multiple superclusters of galaxies when differences between baryonic features like star formation rate, size, luminosity can be smoothed over, we assume the Copernican principle to hold.

Under the assumptions of isotropy and homogeneity, the currently accepted cosmological model describes a universe with rotational and translational symmetry, where the matter content is dominated by dark matter. In the energy context of the universe, dark energy accounts for the overwhelming majority of the energy in the universe while dark matter plays the second fiddle. Baryonic content, out of which structures like galaxies form, accounts for a very small portion of the universe. While the nature and the physical properties of dark energy and dark matter are not fully established, the prevalent theory of cosmological structure formation assumes that the central aspect of galaxy formation is the hierarchical assembly of dark matter halos with the properties of dark energy being not necessarily germane to galaxy formation [35]. This is the ΛCDM cosmological model [6] of the universe where the curvature of spacetime is described by the Robertson-Walker metric. If we further hypothesize that the universe is filled with perfect fluid, then the evolution of the universe follows the Friedmann equations as outlined in [30]:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\varepsilon + \frac{\Lambda}{3} - \frac{\kappa c^2}{a^2 R^2} \tag{1}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\varepsilon + 3P) + \frac{\Lambda}{3}. \tag{2}$$

where G is the universal gravitational constant, c is the speed of light, ε and P are the energy density and pressure of the mass-energy in the universe. The scale factor $a(t)$ models an expanding universe. Conventionally, $a(t_0) = 1$ where t_0 is present day. Λ represents the cosmological constant.

The concordance cosmology states that the present day large scale structure of baryonic matter can be traced back to its seeds embedded within dark matter halos in a very young universe, with the morphology of a galaxy, at any point in cosmic time, predominantly dependent on the properties of the surrounding halo, merger history of that galaxy and feedback effects. As is evident from Figures 1 and 2, galaxies rarely evolve or are found in complete seclusion, but a necessary condition for formation of galaxies is the presence of a dark matter halo. Each dark matter halo typically contains a few subhalos inside which baryonic matter collapses to form galaxies. The most massive subhalo lies at the center of potential of the halo and contains the central galaxy of the halo. Therefore, modeling the universe for a better understanding is essentially painting a temporal picture of the cosmic web.

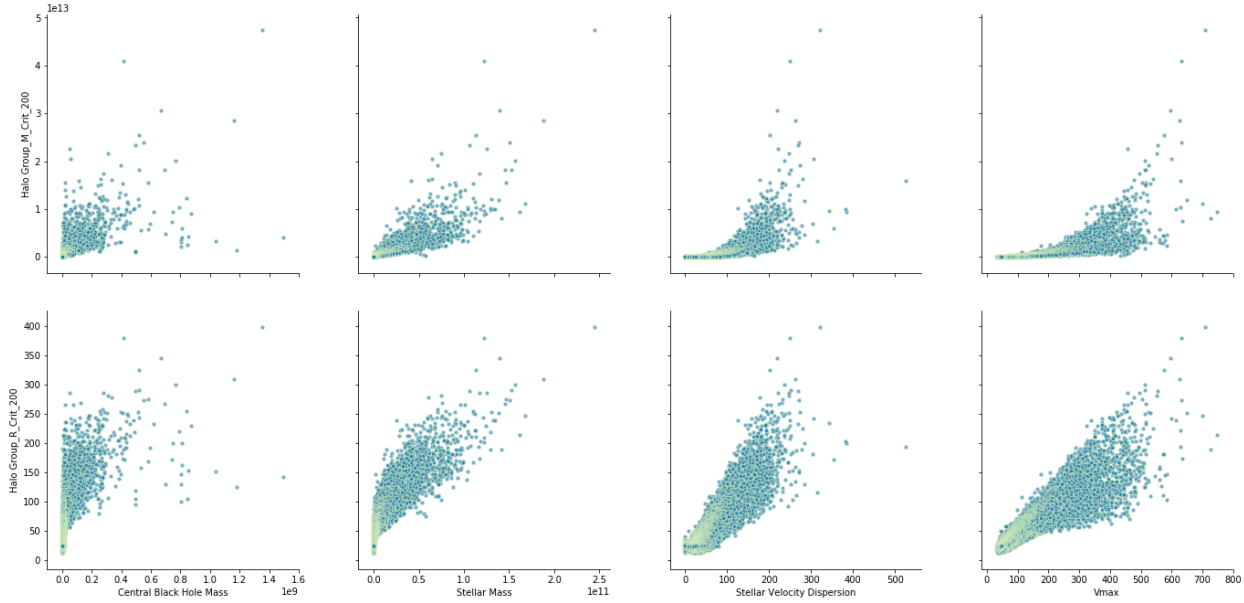

Figure 2: Dependence of baryonic features on dark matter. In fact, as we can see, the baryonic features of a galaxy are indeed dependent on the dark matter features of the galaxy. We provide a concise description of the features in Table 1 and present a more detailed analysis of the features in §4.2

There are broadly two flavors of simulations that model matter distribution in the universe [12]: semi-analytic simulations and hydrodynamical simulations. There is, however, no reason for supposing the superiority of hydrodynamics over semi-analytic modeling. The predictive power of both of these approaches is in agreement with actual observations. We refer the reader to [5] for a comparison between semi-analytic methods and hydrodynamical modelling. In particular, physical processes critical to galaxy formation and evolution such as core collapse supernovae, accretion shocks, stellar winds, involve multiple sets of partial differential equations [38] such that modeling structure formation through either approach becomes extremely difficult. The already intractable complexity of this problem is further compounded by the addition of approximations of physical phenomena which cannot be derived *ab initio*.

Related Work. Using machine learning algorithms to model structure formation has inevitably resulted in a difference of efficacy. Algorithms like k-nearest neighbors and support vector machines used in [41] have conclusively shown that machine learning galaxy-halo relation is not unsuccessful. The work was further extended in [19], [1] and [9] by including other discriminative or ensemble algorithms like decision trees and/or random forests. However, focusing on the algorithmic aspects of the task is equally important since the choice of algorithms usually involve some trade offs between scalability and accuracy while certain algorithms like decision trees are prone to overfitting. To our knowledge, *tractable graphical models* have never been applied to this problem. Our contribution in this paper is to apply a deep architecture with probabilistic semantics, sum product networks (SPNs) [32], to estimate a generative model for the data such that a mock catalog of galaxies can be build. The added advantage of using SPNs is that it guarantees that inference will always be in time linear in the size of the model [13].

3. Method

Making machines *learn* to recognize halos and their corresponding baryonic content broadly involves two steps.

The first step is finding features which are good representatives of a halo-galaxy system and indicate a strong correlation between the potential well of host dark matter halo and the galaxy inside it. This is usually followed up by providing a merger history of the galaxy-halo system to the machine. The choice of the depth of history to be provided is generally the prerogative of the machine learning practitioner. However, this choice comes with a few caveats. Since galaxy clusters generally formed in a very early universe, their merger histories usually cover billions

Dark Matter Features	
Feature	Description
Halo Group Mass	Aggregate Group Mass of all subhalos within a larger halo
Mass Critical 200, M_{200}	Defines the mass of a halo
Radius Critical 200, R_{200}	The Radius that bounds Mass Critical 200
Number of Subhalos	Representative of the number of smaller subhalos that make up a larger halo

Baryonic Matter Features	
Feature	Description
Black Hole Mass	The mass of the central black hole in a halo
Stellar Mass	Representative of the stellar content of a galaxy
Velocity Dispersion	Provides a measure of velocity of a galaxy
Maximum of Circular Velocity, Vmax	Maxima of the circular velocity curve of a galaxy

Table 1: A concise description of the features of dark and baryonic matter.

of years and involve thousands of progenitors. Providing a description of all the progenitors of any galaxy is simply an impractical task. A good way to approach this choice is by constraining the number of progenitors of a galaxy (subhalo) and providing their corresponding properties only for a subset of the cosmic time. This is done keeping in mind that even though a subhalo may have thousands of progenitors and continuously morphs through multiple collisions and accretions, only a few of its progenitors play an overwhelming role in its overall shape and so only these few progenitors are sufficient to indicate the overall *lineage* of the subhalo. A partial merger history is choosing how far to travel along the main branch of a galaxy. As shown in Figure 3, the morphology of a galaxy at some redshift, is the result of evolution along many branches, but its protogalaxies along the main branch can adequately trace the history of a galaxy.

An alternate approach is to provide the algorithm with only a few random snapshots of the universe corresponding to different look-back times. In this approach, merger history need not be provided. The algorithm learns the underlying generative model which can be subsequently used to *infer* the morphology of galaxies at different redshifts. The motivation behind this is the drastic reduction in the dimensionality of the dataset. This then reduces the computation time.

In this paper, we find the set of progenitors of a galaxy along the main branch between redshift 0 and 0.5 sufficient for our purposes. We provide progenitor history in our first approach. In our second approach to model the relation between dark and baryonic matter, we do not provide progenitor history at all. The dataset construction is described in detail in §4.3. We compare and contrast the results of our approaches in §5.

The added advantage of using a graphical model is the greater interpretability. SPNs augment this by allowing probabilistic semantics even when there are no conditional dependencies present [8], while guaranteeing inference in time linear in tree width of the network. In the next section, we briefly review SPNs. We generate the dataset using the results of the Eagle suite of smoothed particle hydrodynamical simulations [36].

3.1. Tractable Probabilistic Graphical Models

Computing the modes and marginals of a probability distribution efficiently while learning the underlying distribution accurately is the main challenge of probabilistic modeling. This is more often than not done by exploiting the inherent relations present in the dataset to learn a joint distribution. However, inference remains the primary bottleneck in probabilistic modeling; in general, inference in Bayesian networks is intractable [11, 10].

3.1.1. Sum Product Networks

A graphical model represents a probability distribution over a d dimensional vector x, where $x \in X$, as a product of factors ϕ_k such that $P(X = x) = \frac{1}{Z}\Pi_k \phi_k(x_k)$, where Z is the partition function and ϕ_k is a function over a subset x_k : $k = 1, 2, 3, \ldots, d$.

While the representation is expressive and compact, practical realizations of graphical models often fail to take advantage of their generative nature due to the computation cost of computing an intractable partition function, Z, which itself is made up of an exponential number of terms, $Z = \Sigma_{x \in X} \Pi_k \phi_k(x_k)$. This immediately results in intractable inference queries.

Sum Product Networks ameliorate this problem by imposing certain restriction on the nodes in the graph. In its simplest description, SPNs are directed acylic graphs with alternate layers of sums and products. SPNs can be recursively defined as follows [32]:

- A (tractable) univariate distribution is an SPN
- A product of SPNs over disjoint variables is a SPN
- A weighted sum of SPNs over the same set of variables is a SPN.

Formally:
Let $\mathcal{X} = \{X_1, \cdots X_n\}$ be a set of variables. A Sum-Product Network (SPN) defined over $\mathcal{X}$ is a rooted directed acyclic graph. The leaves are indicators $[X_p = \cdot]$, where the indicator function [.] has value 1 when its argument is true, 0 otherwise. The internal nodes are sum nodes and product nodes. Each edge (i, j) from sum node i has a non-negative weight w_{ij}. The value of a sum node is $\sum_{j \in Ch(i)} w_{ij} v_j$, where $Ch(i)$ is the children of i and v_j is the value of its j^{th} child. The value of a product node is the product of the values of its children. The value of an SPN is the value of its root.

We use S to denote an SPN as a function of the leaves. Let x be an instantiation of the indicator variables, that is, a full state. Let e be the evidence (partial instantiation). For a given node i, we use S_i to denote the sub-SPN rooted at i. Also, we use x_p^a to mean $[X_p = a]$ is true, and use $\bar{x}_p^a$ to mean its negation. We are interested in learning SPNs with the following properties:

- **Validity**: A SPN is *valid* if and only if it always correctly computes the probability of evidence: $S(e) = \Phi_S(e)$, where $\Phi_S(e)$ is the unnormalized probability of e.
- **Consistency**: A SPN is *consistent* if and only if for every product node i, there is no variable X_p that has indicator x_p^a as one leaf of the sub-SPN S_i and indicator x_p^b with $b \neq a$ as another leaf.
- **Completeness**: A SPN is *complete* if and only if all children of the same sum node have the same scope. (The scope of an SPN is the set of variables in $\vec{X}$ that the indicators of an SPN are defined on)
- **Decomposability**: A SPN is *decomposable* if and only if the children of every product node have disjoint scopes.

To ensure the tractability at the root node, the descendant sum nodes are forced to have children nodes with identical scope, where the scope of a node is defined as the set of variables encoded in that node, while the descendant product nodes must have children with disjoint scopes. It is easy to see that a sum node is a symbolic mixture model made from the decomposable product nodes with leaf nodes at the bottom attached to univariate distributions. The parameters of the networks are the priors over the mixture components in the sum node. This allows the computation of the partition function to remain tractable, which in turn, allows exact inference.

3.2. Structure Learning

Typically, the network can be learned in two ways. The first method involves initializing a random SPN and then estimating the parameters (the weights) of the network from the data either discriminatively as shown in [13], or generatively as described in [14]. For the former case, stochastic gradient descent is usually applied to estimate the parameters of the network. The pervasive gradient diffusion problem in layered networks [17] can be overcome by doing hard expectation-maximization (EM). For the generative case, expectation-maximization is generally employed to learn the weights. Once the weights have been learned, unnecessary edges and nodes can be pruned away.

The size, shape and the weights of the network can also be learned from the data itself. To build a valid SPN such that the conditions of decomposability and completeness are satisfied, instances and variables are recursively split

and partitioned. Sum nodes are assigned to the clusters of instances using any of the cluster estimation algorithms, whereas independence tests are usually employed to partition variables and assign them to product nodes. Leaf nodes are assigned to univariate distributions. We learn generatively with the leaf nodes of our SPN explicitly encoded to contain univariate Bernoulli distributions. This is however not a strict requirement: as shown in works such as [27, 28, 33, 18, 7], SPNs can also be learned online with Gaussian, Poisson and other distributions.

4. Empirical Evaluations

4.1. Simulation overview

The suite of Eagle simulations [36] uses a modified version of Gadget3 hydrodynamical code, last described in [39], to evolve resolution elements in boxes of size 12, 25, 50 and 100 comoving mega parsecs (cMpc) on a side. The cosmology employed in the simulations is consistent with the results of [31], where $\Omega_\Lambda = 0.693$, $\Omega_m = 0.307$, $\Omega_b = 0.04825$, $\sigma_8 = 0.8288$, $n_s = 0.9611$, $h = 0.677$, where, $\Omega_\Lambda, \Omega_m, \Omega_b, \sigma_8, n_s, h$ stand for the contributions to matter/energy content of the universe from cosmological constant, matter, baryons respectively, h is the dimensionless Hubble parameter, n_s is the spectral index of the primordial power spectrum while σ_8 is the rms amplitude of the linear mass fluctuations. High resolution simulations correspond to simulations with an initial baryonic particle mass of m_g = 2.26 x 10^5 $M_\odot$ while intermediate resolution simulations have a higher initial baryonic particle mass, $m_g = 1.81 * 10^6$ $M_\odot$, where $M_\odot$ is 1 solar mass.

The key run of the simulations, which we use in this paper, the Fiducial Ref-L0100N1504 simulation is an intermediate resolution simulation with periodic box with a volume of $(100\text{cMpc})^3$, initially containing 1504^3 gas particles, with an initial mass of $1.81 * 10^6$ $M_\odot$ and the same amount of dark matter particles with $9.70 * 10^6$ $M_\odot$.

Substructures, like galaxies, in Eagle simulations were identified using the the standard SUBFIND algorithm (developed in [40]). Galaxies were defined as gravitationally bound subhalos and identified using three steps. First, halos were identified by running the Friends-of-Friends algorithm (FOF) [29] on the dark matter particles with linking length 0.2 times the mean interparticle separation. Gas and star particles are assigned to the same, if any, halo as their nearest dark matter particles. Second, SUBFIND defines substructure candidates by identifying overdense regions within the halos bounded by saddle points in the density distribution. Finally, particles that are not gravitationally bound to the substructure are removed and the resulting substructures are referred to as galaxies.

The simulations themselves have a finite resolution and are generally not reliable on lower mass range of satellite galaxies and dwarf halos; the physics on lower scales is more influenced by feedback effects and stellar winds which are poorly understood and have no analytic solutions. The properties of low mass galaxies are not entirely reliable due to the finite resolution of the simulations. In general, many galaxy properties are unreliable below a stellar mass of 10^9 $M_\odot$. Due to this, we only select central galaxies with halo mass above 10^{10} $M_\odot$. For a more comprehensive discussion on the parameters of the simulation, we refer the readers to [36].

4.2. Feature Engineering

Physical processes critical to galaxy formation and evolution such as core collapse supernovae, accretion shocks, stellar winds, involve multiple sets of partial differential equations [38] such that modeling structure formation becomes extremely difficult. The already intractable complexity of this problem is further compounded by the addition of approximations of physical phenomena which cannot be derived *ab initio*. This is where the dependence between baryonic matter and dark matter can be exploited in a probabilistic machine learning setting to generate mock catalogs of galaxies. If baryonic matter and dark matter are modelled as random variables, then a joint distribution over these random variables can give a very good indication of the dynamical co-evolution of the universe. But this again is a complex task since the effectiveness of any machine learning algorithm dramatically increases with a good choice of features. This crucial aspect of probabilistic modeling plays a more pronounced role in rich, high dimensional datasets such as cosmological simulations since these simulations essentially take snapshots of the state of particles, the state of any particle itself is an aggregate of multiple non-linear couplings between various physical processes. The choice of features to make up the input space of machine learning algorithms rests solely on the domain knowledge. The domain here is the universe, a universe where majority of the energy content is completely dark and the properties of the overwhelming majority of matter is a mystery and the features of the remaining matter are only finitely resolvable.

To better illustrate the difficulty of the problem, let us consider a simple toy example: assume that a galaxy is a system with just four components: dark matter halo, stellar halo, central black hole and stellar bulge. Suppose we further fix the location of the central black hole of a galaxy as the starting point of a galaxy. In this simplistic scenario, the answer to the question: *what is the size of a galaxy or equivalently, where does a galaxy end?*, is reduced to heuristics. To answer this question, we normally assume that the galaxy with all its components are in a state of equilibrium. This leads to definition of a virialized state of that system such that the virial radius of the galaxy can act as a representative of the galaxy size. But, can virial radius, established under the assumptions of lack of perturbations, be a good indication of the size of a galaxy like Milky Way, which at the moment is undergoing tidal stress and is on a collision course with Andromeda Galaxy? It is obvious that making a generative model of dark and baryonic matter so as to understand the overall matter distribution in the universe and using it to learn the mapping between dark and baryonic matter is not a trivial task. To reiterate the point let us consider another example. If we take into account that star formation rate peaked at redshift 1∼2, see [25] for stages of evolution in cosmology, while the first stars were born just 200 million years after the Big Bang, then an interesting formulation of the task is defining the word *partial*. If we are to predict the the baryonic content of a halo using at any redshift using the halo merger history, then modeling the dynamical evolution of the galaxy depends, in no small measure, on the kind of history provided.

Overall, the baryonic features we model as random variables are: mass of central super massive black hole, stellar mass, velocity dispersion and the maximum of circular velocity of the galaxy, which are described below, followed by a short description of dark matter features.

4.2.1. Black Hole Mass

The Eagle simulations implement a quasar mode active galactic nuclei feedback to model black hole with black hole seeds only planted in halos with masses more than 10^{10} $M_{\odot}h^{-1}$, that do not already contain a black hole. The black hole grows through mergers and accretion.

4.2.2. Stellar Mass

The stellar mass of the central galaxies is chosen to be the mass contained within a 3 dimensional aperture centred on the minimum of potential of that galaxy. In this paper, we use stellar mass within an aperture of size 30 proper kiloparsecs to describe the stellar content of the galaxy. As shown in Figure 1, the galaxy stellar mass function produced in the simulations corresponds well with the observations from Galaxy And Mass Assembly [3] and SDSS [22] surveys.

4.2.3. Velocity Dispersion

The Eagle simulations model the velocity dispersion of the stars as $\sqrt{2E_k/3M}$, where the kinetic energy, E_k and mass M, is calculated for all the stars within a spherical 30 physical kiloparsecs aperture centered on the galaxy's center of potential.

4.2.4. Maximum of Circular Velocity

The maximum value of the circular velocity (hereafter V*max*) is derived through $v_c(r) \equiv \sqrt{\frac{G_N M(<r)}{r}}$, where $M(< r)$ is the mass enclosed within a sphere of radius r.

4.2.5. Mass Critical 200

The halo mass, M_{200}, is defined as the total mass contained within the virial radius R_{200},

4.2.6. Radius Critical 200

The corresponding radius term for Mass Critical 200 is Radius Critical 200 (R_{200}). Formally, it is the physical radius within which the mean density is 200 times the critical density of the universe.

4.2.7. Halo Group Mass

The halo features, Radius Critical 200 (R_{200}) and Mass Critical 200 (M_{200}), we use were derived by placing a sphere at the minimum of gravitational potential centered on the central galaxy. The halo group mass is the aggregate mass of all the dark matter subhalos within a group while the number of subhalos refers to the number of galaxies inside a halo. As previously noted in §4.1, the (sub)halos within a group were identified through SUBFIND and FOF algorithms. The halo group mass is the aggregate mass of all subhalos within any group.

4.3. Dataset Construction

Since our method involves two different approaches, we construct four datasets by querying both the fiducial and dark matter-only models in the database for the properties of sub-halos (galaxies) with their corresponding dark matter halos and halos only.

The first approach, where we provide a merger history, corresponds to Dataset 1 and Dataset 3.

With *Dataset 1*, we provide SPNs with a selection of properties of the central galaxy at zero redshift in each halo along with a description of their corresponding central subhalo merger history from redshift 0 to redshift 0.50. This is equivalent to providing the halo history for approximately the last 5 billion years. The merger tree was traversed only along the main branch, see Fig.3, of every galaxy.

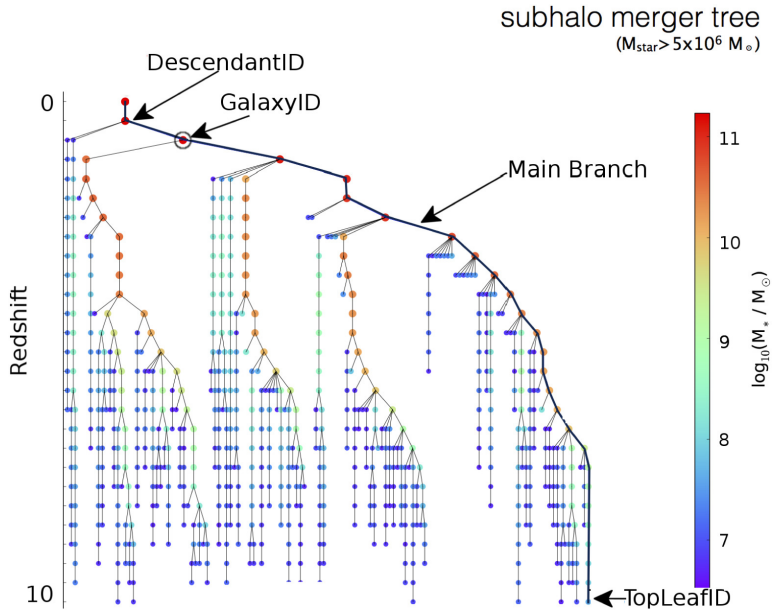

Figure 3: Merger history of a galaxy with stellar mass, $M_{star} > 10^5 M_{\odot}$. Figure taken from the Eagle Database [26]. As described before, a galaxy at the present time, is the result of mergers of multiple galaxies over the course of billions of years. Redshift 0 corresponds to the present day, while redshift 10 corresponds to a lookback time of 12 Gigayears. The merger history of a galaxy usually follows a main progenitor branch. The Eagle simulations model the merger history as shown in the figure, where the Descendant ID represent a galaxy at a specific point in time, while the TopLeafID represents the first progenitor of that galaxy along the main branch. The main progenitor branch is indicated with a thick black line, all other branches with a thin line. To get the merger history, we only traverse along the main progenitor branch.

The galactic properties we model are the mass of its central black hole, stellar mass, velocity dispersion of the stars and the maximum of the circular velocity rotation curve of the galaxy.

Dataset 3 was generated in a similar way through the Dark Matter-Only snapshots in the Eagle simulations. In Dataset 3, we only use halo properties and halo merger histories, from redshift 0 to redshift 0.50, as inputs and query for properties redshift 0. The common factors in *Dataset 1* and *Dataset 3* are the halo properties and merger histories.

Through the ***second approach***, corresponding to Dataset 2 and Dataset 4 where halo history was not provided at all, SPN builds generative models of matter distribution out of snapshots.

Dataset 2 and *Dataset 4* were generated again from the fiducial and the dark matter-only run. These datasets contain the same properties of galaxy-halo systems and only halos as in the first approach, but between redshifts 3.5

to 1.7 respectively. We chose this particular redshift range so that our dataset can better model the universe since the peak of star formation rate is generally supposed to have occurred within this regime [25]. The generative model created by training the machine on the values between these particular redshifts were tested by querying for the values of the corresponding properties at redshift 0. We wanted to see how well the SPN picks up on the underlying matter distribution of the simulated universe, when given only the state of cosmic web at different redshifts. The ulterior motive for using the dark matter-only runs was to see how well the algorithm approximates the N-body calculations. The dark matter-only simulations use gravity as the governing law for evolution of billions of particles. Since there is no analytic solution for this case and convergence is usually established through the use of energy or momentum conservation laws, so we can, to some extent, see how well SPNs can approximate the convergence.

5. Analysis

In this section, we present and discuss the results that were obtained when we applied the algorithm to the Eagle data. Using the dark matter internal halo properties as our inputs, the following baryonic features were are predicted: black hole mass, stellar mass of the galaxies, velocity dispersion and V*max*. These attributes are the result of evolution over billions of years through dissipative, nonlinear baryonic processes. As discussed earlier, the overall picture of large scale structure formation is the ΛCDM model; but, on smaller scales, the details are incredibly rich and vastly more complicated.

In Fig 4 we first show the performance of SPNs in reproducing the simulated properties of the galaxies in Eagle and subsequently follow it up by discussing the implications of our results for the halo-galaxy connection. We use the following statistics to quantify the effectiveness of SPNs in predicting the galaxy properties.

- First, we use the standard mean squared error (MSE) metric, which is defined as:

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{i=N_{test}-1} \left(X^i_{test} - X^i_{predicted}\right)^2 \tag{3}$$

  Here, X^i_{test} is the i^{th} value of the test set, $X^i_{predicted}$ is the i^{th} value of the predicted set and N_{test} is the size of the test set.

- The Pearson correlation coefficient is perhaps the most widely used measure for linear relationships between two normal distributed variables and thus often just called "correlation coefficient". The Pearson correlation coefficient measures the linear relationship between two variables. Usually, the Pearson coefficient is obtained via a Least-Squares fit and a value of 1 represents a perfect positive relationship, -1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables.

  For two variables X and Y:

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y} \tag{4}$$

  with cov(X, Y) as the covariance between X and Y and σ_x and σ_y the standard deviations of X and Y respectively.

  The estimate:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{5}$$

  where x_i, y_i, denote the i^{th} element of the vector X and Y.

  $\overline{x}, \overline{y}$ are the respective means of X and Y.

- The Coefficient of Determination is defined as:

$$R^2 = 1 - \frac{\sum_i (X^i_{test} - X^i_{predicted})^2}{\sum_i (X^i_{test} - X_{mean,train})^2} \tag{6}$$

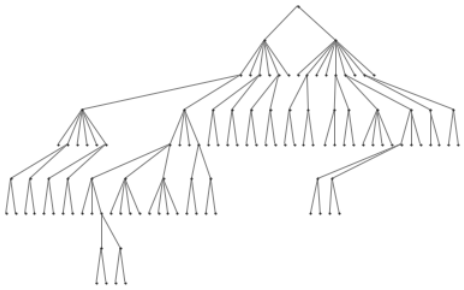

Figure 4: The structure of the SPN for *Dataset 1*. The SPN has 143 edges with 144 nodes in 20 layers. The network has 19 sum nodes, 40 product nodes and 85 leaf nodes. Each leaf node contains a univariate Bernoulli distribution. The single node at the top is the root sum node. The structure took 149 seconds to be learned

Once the SPN has captured the joint distribution over all the variables in the dataset at its root node, it can be easily queried for conditional and marginal likelihoods for any random variable like stellar mass or mass of central black hole. The root node in SPN can also be sampled to generate synthetic datasets that follow the learned joint distribution.

Tables 2 and 3 clearly show that there is not much difference in the errors between the two approaches. The computation time taken to learn the joint distribution however, is much less when just snapshots are provided. So merger histories of halos do not really play much role in building richer models. Tables 4 and 5 are analogous, but for dark matter properties only.

The results shown in Tables 2, 3, 4, 5 demonstrate that SPNs are able to recreate mock catalogs with properties strikingly similar to those produced by extensive hydrodynamic codes. The baryonic properties which are heavily

Table 2: *Dataset 1*: The structure of SPN for this dataset was learned in 847.6 seconds. Progenitor history was provided.

Feature	MSE	R^2	Accuracy Score	PearsonR
Central Black Hole Mass	0.041714	0.464182	0.958286	0.743506
Stellar Mass	0.019964	0.732150	0.980036	0.870518
Velocity Dispersion	0.118812	0.464540	0.881188	0.727086
V*max*	0.065533	0.680239	0.934467	0.837789

Table 3: *Dataset 2*: The structure of SPN for this dataset was learned in 144.7 seconds. Only random snapshots were provided.

Feature	MSE	R^2	Accuracy Score	PearsonR
Central Black Hole Mass	0.039717	0.469593	0.960283	0.735701
Stellar Mass	0.019542	0.727792	0.980458	0.867607
Velocity Dispersion	0.107178	0.512861	0.892822	0.751796
V*max*	0.055159	0.728921	0.944841	0.863211

Table 4: *Dataset 3*: The structure of SPN for this dataset was learned in 1890.15 seconds. Dark Matter Only run with halo history.

Feature	MSE	R^2	Accuracy Score	PearsonR
Number of Subhalos	0.053304	0.442150	0.946696	0.701084
Halo group Mass	0.014672	0.799276	0.985328	0.905383
M_{200}	0.005449	0.929433	0.994551	0.965560
R_{200}	0.012702	0.938336	0.987298	0.969134

dependent on mass are predicted extremely well. This is the direct consequence of gravity being the most dominant force at large scale. The mass of the central black hole and the stellar content of a galaxy, given by stellar mass, is linearly dependent on the mass of the halo, M_{200} around it. Velocity Dispersion and V*max* are implicitly governed by kinetic and potential energies which are in turn, dependent on mass and radius. The predicted and true distributions are almost identical in the case of stellar mass and mass of central black hole.

A striking feature is the inability of progenitor history to increase the accuracy of the predictions, even at the cost of increased computation time. As we can see from Tables 2 and 3, the mean squared errors for stellar mass of a galaxy when the progenitor history is provided is 0.019964, while the mean squared error when no progenitor history is provided is 0.019542. The contrast becomes more pronounced when we compare the computation time taken by our algorithm. For *Dataset 1* and *Dataset 2*, the time taken for the same computation is 846 seconds and 144 seconds. This clearly shows that even though large scale structure can only grow hierarchically through mergers over cosmic time, progenitor history does not play any significant role in predictions.

The same is true for dark matter-only runs. Table 4 gives the result for dark matter-only simulation with progenitor history of halos provided to SPNs, while Table 5 delineates the result when halo history is not provided. As we can see in the mean squared errors for number of subhalos and halo group mass, there is not even an appreciable difference in errors, while learning the structure of SPN to model the joint distribution takes drastically more time with *Dataset 3*, relatively to the computation time taken by *Dataset 4*.

Overall, we get somewhat surprising results. Numerical simulations evolve many gaseous interactions on an *ad hoc* basis and the baryonic physics is vastly complicated. We did not really expect the algorithm to pick up so well on the galaxy halo relation. However it is important to note that our model is purely a phenomenological one. Unlike hydrodynamics, machine learning does not presume a relation between dark matter halos and the galaxies in it. This implies that machine learning can never be used as a replacement for numerical simulations, instead it can be used as a tool to study galaxy-halo connection and explore the influence of different simulation physics, like the one employed in semi-analytic modelling, to explore structure formation in the universe.

6. Conclusions

We performed an empirical study of the relation between dark matter halo and the corresponding galaxies it encloses through the use of a tractable probabilistic graphical model, sum product network, in the backdrop of one of the largest hydrodynamic simulations of cosmology.

The core underlying physics is the dependence of baryonic matter like stars and galaxies on dark matter. Dark matter itself cannot be seen by any form of instruments. We only infer the presence of dark matter. Dark matter itself must be modelled through simulations and the properties of dark matter must be defined through heuristics. This is where simulations are necessary to generate the dataset. Without the simulation, we will not have any dataset to begin with. But on the other hand, a full simulation will take hours to complete. To this end, we showed that given the data from a partial simulation, SPNs can provide a generative model for full simulation. So, one possible idea for future work is to let a simulation run for a small amount of time and then train SPNs to generate newer data.

The goal of this project was not to construct a numerically identical population of galaxies, but to explore how much information can be extracted from dark matter properties about the eventual evolutionary properties of galaxies. The conclusion seems to indicate that SPNs can clearly mimic the evolution of galaxies in a hydrodynamic setting. Furthermore, the runtime of SPN is of the order of minutes, in sharp contrast to millions of hours spend by numerical simulations. Challenges for the future include using more advanced algorithms to fully explore the extent to which

Table 5: *Dataset 4*: The structure of SPN for this dataset was learned in 149.5 seconds. Dark Matter Only run with just snapshots.

Feature	MSE	R^2	Accuracy Score	PearsonR
Number of Subhalos	0.051744	0.464274	0.948256	0.714493
Halo group Mass	0.015195	0.793547	0.974805	0.914319
M_{200}	0.004964	0.933783	0.994036	0.963175
R_{200}	0.010756	0.947684	0.989244	0.973857

machine learning can be assimilated in cosmology. Potential applications of such an extended framework include a new approach to obtaining a halo mass function, which can be directly tested against existing fitting formulae adopted by analytic approaches. A second interesting avenue is the use of machine learning to compare and contrast different cosmologies. If we train machines to learn galaxy-halo connection based on the concordance model, then the trained model can be tested to see how well it can explain baryonic features generated by other parametrizations of the Big Bang theory. A further ambitious project lies in using machine learning to constrain cosmological parameters using weak lensing data or deep sky surveys. In this regard, tractable probabilistic models such as probabilistic sentential decision diagrams (e.g [20]) that permit the structure learning process to incorporate symbolic constraints may be particularly useful.

[1] S. Agarwal, R. Davé, and B. A. Bassett. Painting galaxies into dark matter haloes using machine learning. *Monthly Notices of the Royal Astronomical Society*, 478(3):3410–3422, 2018.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[3] I. Baldry, S. P. Driver, J. Loveday, E. Taylor, L. Kelvin, J. Liske, P. Norberg, A. Robotham, S. Brough, A. M. Hopkins, et al. Galaxy and mass assembly (gama): the galaxy stellar mass function at z¡ 0.06. *Monthly Notices of the Royal Astronomical Society*, 421(1):621–634, 2012.

[4] Y. Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[5] A. Benson, F. Pearce, C. Frenk, C. Baugh, and A. Jenkins. A comparison of semi-analytic and smoothed particle hydrodynamics galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 320(2):261–280, 2001.

[6] E. Bertschinger. Cosmic structure formation. *Physica D: Nonlinear Phenomena*, 77(1):354 – 379, 1994. Special Issue Originating from the 13th Annual International Conference of the Center for Nonlinear Studies Los Alamos, NM, USA.

[7] A. Bueff, S. Speichert, and V. Belle. Tractable querying and learning in hybrid domains via sum-product networks. *CoRR*, abs/1807.05464, 2018.

[8] C. J. Butz, J. S. Oliveira, and A. E. dos Santos. On learning the structure of sum-product networks. *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–8, 2017.

[9] S. Cavuoti, M. Brescia, G. Riccio, G. Longo, et al. Stellar formation rates in galaxies using machine learning models. *arXiv preprint arXiv:1805.06338*, 2018.

[10] M. Chavira and A. Darwiche. On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6-7):772–799, 2008.

[11] A. Darwiche and J. D. Park. Complexity results and approximation strategies for map explanations. *Journal Of Artificial Intelligence Research*, 2006.

[12] K. Dolag, S. Borgani, S. Schindler, A. Diaferio, and A. M. Bykov. Simulation techniques for cosmological simulations. *Space science reviews*, 134(1-4):229–268, 2008.

[13] R. Gens and P. Domingos. Discriminative learning of sum-product networks. *Advances in Neural Information Processing Systems*, pages 3239–3247, 2012.

[14] R. Gens and P. Domingos. Learning the structure of sum-product networks. *International conference on machine learning*, pages 873–880, 2013.

[15] C. Gheller, P. Wang, F. Vazza, and R. Teyssier. Numerical cosmology on the gpu with enzo and ramses. In *Journal of Physics: Conference Series*, volume 640, page 012058. IOP Publishing, 2015.

[16] D. Guest, K. Cranmer, and D. Whiteson. Deep learning and its application to lhc physics. *Annual Review of Nuclear and Particle Science*, 68:161–181, 2018.

[17] S. S. Haykin et al. *Neural networks and learning machines/Simon Haykin.* New York: Prentice Hall,, 2009.

[18] W. Hsu, A. Kalra, and P. Poupart. Online structure learning for sum-product networks with gaussian leaves. *CoRR*, abs/1701.05265, 2017.

[19] H. Kamdar, M. Turk, and R. Brunner. Machine learning and cosmological simulations. *American Astronomical Society Meeting Abstracts# 227*, 227, 2016.

[20] D. Kisa, G. Van den Broeck, A. Choi, and A. Darwiche. Probabilistic sentential decision diagrams. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2014.

[21] D. Koller, N. Friedman, and F. Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[22] C. Li and S. D. White. Autocorrelations of stellar light and mass in the low-redshift universe. *Monthly Notices of the Royal Astronomical Society*, 407(1):515–519, 2010.

[23] Y. Liang, J. Bekker, and G. Van den Broeck. Learning the structure of probabilistic sentential decision diagrams. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[24] C. Llinares. The shrinking domain framework i: a new, faster, more efficient approach to cosmological simulations. *arXiv preprint arXiv:1709.04703*, 2017.

[25] P. Madau and M. Dickinson. Cosmic star-formation history. *Annual Review of Astronomy and Astrophysics*, 52:415–486, 2014.

[26] S. McAlpine, J. C. Helly, M. Schaller, J. W. Trayford, Y. Qu, M. Furlong, R. G. Bower, R. A. Crain, J. Schaye, T. Theuns, et al. The eagle simulations of galaxy formation: Public release of halo and galaxy catalogues. *Astronomy and Computing*, 15:72–89, 2016.

[27] A. Molina, S. Natarajan, and K. Kersting. Poisson sum-product networks: A deep architecture for tractable multivariate poisson distributions. *AAAI*, pages 2357–2363, 2017.

[28] A. Molina, A. Vergari, N. Di Mauro, S. Natarajan, F. Esposito, and K. Kersting. Mixed sum-product networks: A deep architecture for hybrid domains. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[29] S. More, A. V. Kravtsov, N. Dalal, and S. Gottlöber. The overdensity and masses of the friends-of-friends halos and universality of halo mass function. *The Astrophysical Journal Supplement Series*, 195(1):4, 2011.

[30] E. Mörtsell. Cosmological histories from the friedmann equation: The universe as a particle. *European Journal of Physics*, 37(5):055603, 2016.

[31] Planck, P. Ade, N. Aghanim, C. Armitage-Caplan, et al. Planck 2013 results. xvi. cosmological parameters. *Astron. Astrophys*, 571:A16, 2014.

[32] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690, 2011.

[33] A. Rashwan, H. Zhao, and P. Poupart. Online and distributed bayesian moment matching for parameter learning in sum-product networks. In *Artificial Intelligence and Statistics*, pages 1469–1477, 2016.

[34] B. Ryden. *Introduction to cosmology*. Cambridge University Press, 2016.

[35] J. Salcido, R. G. Bower, L. A. Barnes, G. F. Lewis, P. J. Elahi, T. Theuns, M. Schaller, R. A. Crain, and J. Schaye. The impact of dark energy on galaxy formation. what does the future of our universe hold? *Monthly Notices of the Royal Astronomical Society*, 477(3):3744–3759, 2018.

[36] J. Schaye, R. A. Crain, R. G. Bower, M. Furlong, M. Schaller, T. Theuns, C. Dalla Vecchia, C. S. Frenk, I. McCarthy, J. C. Helly, et al. The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, 2014.

[37] P. Schechter. An analytic expression for the luminosity function for galaxies. *The Astrophysical Journal*, 203:297–306, 1976.

[38] R. S. Somerville and R. Davé. Physical models of galaxy formation in a cosmological framework. *Annual Review of Astronomy and Astrophysics*, 53:51–113, 2015.

[39] V. Springel. The cosmological simulation code gadget-2. *Monthly notices of the royal astronomical society*, 364(4):1105–1134, 2005.

[40] V. Springel, S. White, G. Tormen, and G. Kauffmann. Populating a cluster of galaxies-i. results at [formmu2] z= 0, mnras 328 (dec., 2001) 726–750. *arXiv preprint astro-ph/0012055*, 2001.

[41] X. Xu, S. Ho, H. Trac, J. Schneider, B. Poczos, and M. Ntampaka. A first look at creating mock catalogs with machine learning techniques. *The Astrophysical Journal*, 772(2):147, 2013.