

Tractable Generative Modelling of Cosmological Numerical Simulations

Amit Parag¹ ^a and Vaishak Belle^{1,2}  ^b

¹The University of Edinburgh, U.K.

²Alan Turing Institute, U.K.

Keywords: Cosmological Simulations, Generative Models, Sum-Product Networks.

Abstract: Cosmological simulations aim to understand the matter distribution in the universe by employing either semi-analytic methods or hydrodynamical models of matter distribution. These simulations describe the evolution of baryonic structures within dark matter potential wells, where dark matter is modeled as a self-gravitating, collisionless system. Despite advances in reducing computational costs, these simulations still require millions of CPU hours to achieve stable solutions. This raises the question: can generative models predict galaxy properties from a partial history of their dynamical evolution? Tractable probabilistic models, such as sum-product networks, enable efficient computation of conditional probabilities, allowing conditional marginals to be computed in time linear to the model size. In this work, we investigate the application of sum-product networks to compactly represent and learn distributions for predictions in concordance cosmology. Using the Eagle suite of cosmological hydrodynamical simulations, we demonstrate that these graphical models can effectively reproduce mock galaxy catalogs, capturing the relationship between baryonic and dark matter with promising accuracy.

1 INTRODUCTION


Probabilistic models, like Bayesian and Markov networks, are key in statistical machine learning for expressing dependencies compactly. However, inference in such models is intractable, making learning difficult (Koller et al., 2009). Tractable learning, which enables efficient probabilistic querying, addresses this issue, with early work focusing on low tree-width models (Bach and Jordan, 2002) and later efforts utilizing local structures (Chavira and Darwiche, 2008), leading to arithmetic circuits (ACs) that support exact inference in polynomial time.


Sum-product networks (SPNs) (Poon and Domingos, 2011) are notable instances of ACs with a recursive structure, where an SPN is a weighted sum of products of SPNs, and its leaves represent tractable distributions (e.g., univariate Bernoulli). SPNs are often seen as tractable deep architectures due to their hierarchical nature. While deep learning models face challenges in structure learning (Bengio et al., 2009), SPNs inherently provide a reliable structure learning framework. Although SPNs can be manually specified, weight learning and adherence to conditions like

completeness and decomposability make automated structure learning preferable.

Since their introduction, various structure learning methods for SPNs and related models have emerged, such as LearnSPN (Gens and Domingos, 2013), OSL (Hsu et al., 2017), and ID-SPN (Liang et al., 2017). Extensions like RAT-SPNs (Peharz et al., 2020), deep tractable models (Vergari et al., 2021), and hybrid SPN frameworks (Rahman and Gogate, 2014) further demonstrate their adaptability. Although related models may differ in properties and features (Liang et al., 2017), SPNs remain attractive for generative modeling due to their simplicity and scalability.

In this work, we study how SPNs can be used to model a novel and challenging problem related to the evolution of galaxies, particularly in the context of concordance cosmology with hydrodynamical simulations. Since actual experiments in cosmology are not feasible, simulations are crucial for testing cosmological theories, properties, and parameters. Numerical simulations are indispensable due to the inability to derive analytic solutions for gravitationally interacting particles, which form the basis of all cosmological simulations. Dark matter, modeled as particles interacting solely through gravity, plays a pivotal role in these simulations. When combined with baryonic

^a  <https://orcid.org/0009-0001-6597-3976>

^b  <https://orcid.org/0000-0001-5573-8465>

physics, numerical simulations validate cosmological models and provide a general picture of structure formation in the universe.

Simulations attempt to describe the cosmological structure by modeling galaxies inside dark matter halos, which involves making unverifiable assumptions (Somerville and Davé, 2015) and incurs high computational costs, as shown by the millions of CPU hours required by simulations like Eagle and Illustris (Schaye et al., 2014; Llinares, 2017). Several algorithms aim to reduce computation times while improving accuracy (Llinares, 2017; Gheller et al., 2015). Machine learning has been used in cosmology for particle tracing and classification (Guest et al., 2018), but its potential for enhancing cosmological research is still underexplored. This work takes initial steps toward using recent advances in tractable probabilistic models for generative modeling in cosmology. The focus of the paper is on applying these models rather than extending existing algorithms. As demonstrated, this involves understanding concordance cosmology and overcoming challenges in feature selection. Ultimately, we hope our work inspires further interdisciplinary research, enabling machine learning to tackle significant cosmological problems.

2 BACKGROUND

In this section, we proceed by first by giving a fairly informal picture of the cosmological model, before turning to the equations driving the computational task. We then discuss works where machine learning has been used to model structure formation.

2.1 Concordance Cosmology

The concordance cosmological model, the Λ CDM model, is based on the Copernican principles of isotropy and homogeneity (Ryden, 2016). This model assumes that observers in any location in the universe cannot be the central observers, implying that the universe is isotropic and homogeneous on the largest scales. These properties, however, only hold on cosmological scales where the differences between baryonic features are smoothed over.

Under these assumptions, the current cosmological model describes a universe with rotational and translational symmetry, dominated by dark matter. Dark energy comprises the majority of the universe's energy, while dark matter plays a secondary role. Baryonic matter, forming structures like galaxies, is a small fraction of the universe. The Λ CDM model (Bertschinger, 1994) describes the curvature of space-

time using the Robertson-Walker metric, and the evolution of the universe follows the Friedmann equations (Mörtsell, 2016).

The concordance cosmology posits that the large-scale structure of baryonic matter today originated from seeds in dark matter halos. Galaxies form within these halos, with their morphology largely determined by the properties of the surrounding halo, merger history, and feedback effects. The presence of a dark matter halo is essential for galaxy formation, with massive subhalos at the center of halos containing the central galaxies. Modeling the universe involves painting a temporal picture of the cosmic web.

There are broadly two flavors of simulations that model matter distribution in the universe (Dolag et al., 2008): semi-analytic simulations and hydrodynamical simulations. The predictive power of both of these approaches is in agreement with actual observations. We refer the reader to (Benson et al., 2001) 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 (Somerville and Davé, 2015) 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*.

2.2 Learning Structure Formation

Using machine learning algorithms to model structure formation has inevitably resulted in varying degrees of efficacy. Algorithms like k -nearest neighbors and support vector machines used in (Xu et al., 2013) have conclusively shown that machine learning galaxy-halo relation is not unsuccessful. The work was further extended in (Kamdar et al., 2016), (Agarwal et al., 2018), and (Cavuoti et al., 2018) by including other discriminative or ensemble algorithms like decision trees and/or random forests. Recent advances in deep learning have shown significant potential in this domain. For instance, (Villaescusa-Navarro et al., 2021) introduced the CAMELS project, leveraging convolutional neural networks (CNNs) and variational autoencoders (VAEs) to predict galaxy properties. Similarly, (Lucie-Smith et al., 2023) employed graph neural networks (GNNs) to model complex galaxy-halo interactions. (Ho et al., 2022) explored the application of foundation models in astrophysics, focusing on tasks like large-scale structure modeling. Additionally, (Heitmann et al., 2021) emphasized

the use of emulators for precision cosmology, and (Kobayashi et al., 2022) introduced machine learning frameworks for predicting baryonic effects on the matter power spectrum. Works like (Davies et al., 2021) further demonstrate how simulations combined with machine learning can enhance our understanding of galaxy morphologies, while (Tamosiunas et al., 2023) applied semi-supervised learning to improve galaxy classification in limited data scenarios.

However, focusing on the algorithmic aspects of the task is equally important since the choice of algorithms usually involves 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) (Poon and Domingos, 2011), to estimate a generative model for the data such that a mock catalog of galaxies can be built. The added advantage of using SPNs is that they guarantee that inference will always be in time linear in the model size.

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 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 1, 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.

Table 1: Features of dark and baryonic matter.

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, V_{max}	Maxima of the circular velocity curve of a galaxy

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, as well as a reporting of the results in discussed in a subsequent section.

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 (Butz et al., 2017), while guaranteeing inference in time linear in tree width of the network. We generate the dataset using the results of the Eagle suite of smoothed particle hydrodynamical simulations (Schaye et al., 2014).

In the interest of space, we do not go into the details of SPNs, and refer interested readers to (Poon

and Domingos, 2011). These data structures allow the modes and marginals of a probability distribution to be computed efficiently. (See (Kisa et al., 2014) for other data structures with such properties.) Moreover, the size, shape and the weights of the network can also be learned from the data, either discriminatively or generatively (Gens and Domingos, 2012; Gens and Domingos, 2013). In this work, 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 (Molina et al., 2017; Molina et al., 2018; Rashwan et al., 2016; Hsu et al., 2017; Bueff et al., 2018), SPNs can also be learned online with Gaussian and other distributions, which might be useful for future work on modelling physical phenomena via generative models.

4 EMPIRICAL EVALUATIONS

Simulation Overview. The suite of Eagle simulations (Schaye et al., 2014) uses a modified version of Gadget3 hydrodynamical code, last described in (Springel, 2005), 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 (Planck et al., 2014), 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 $m_g = 2.26 \times 10^5 M_\odot$ while intermediate resolution simulations have a higher initial baryonic particle mass, $m_g = 1.81 \times 10^6 M_\odot$, 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 $1.81 \times 10^6 M_\odot$ and the same amount of dark matter particles with $9.70 \times 10^6 M_\odot$.

Substructures, including galaxies, in Eagle simulations were identified using the SUBFIND algorithm (developed in (Springel et al., 2001)). First, halos were detected with the Friends-of-Friends (FOF) algorithm (More et al., 2011) on dark matter particles, with a linking length of 0.2 times the mean interpar-

ticle separation. Gas and star particles were assigned to the same halo as their nearest dark matter particles. Next, SUBFIND identified substructure candidates by locating overdense regions within halos, defined by saddle points in the density distribution. Finally, gravitationally unbound particles were removed, and the remaining substructures were classified 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. In general, many galaxy properties are unreliable below a stellar mass of $10^9 M_\odot$. Thus, we only select central galaxies with halo mass above $10^{10} M_\odot$. For a comprehensive discussion on the parameters of the simulation, we refer readers to (Schaye et al., 2014).

4.1 Feature Engineering

Modeling galaxy formation involves complex physical processes like supernovae, accretion shocks, and stellar winds, which require approximations and partial differential equations. Probabilistic machine learning can help model the interactions between baryonic and dark matter, capturing their joint dynamical evolution. However, feature selection remains challenging due to the high-dimensional nature of cosmological datasets and the influence of non-linear processes.

Feature selection in cosmology relies heavily on domain knowledge, particularly because most of the universe's energy and matter is dark. A simplified galaxy model includes four components: dark matter halo, stellar halo, central black hole, and stellar bulge. The virial radius, which represents the galaxy's size, is often used heuristically but may not always be accurate, especially for galaxies undergoing tidal stress or collisions.

A generative model of dark and baryonic matter is crucial for understanding their distribution and the mapping between them. For instance, predicting baryonic content based on a halo's merger history is challenging, particularly considering the peak of star formation at redshift 1–2.

The baryonic features we model as random variables are:

- Black hole mass: Modeled in simulations with feedback from active galactic nuclei, where black holes grow through mergers and accretion.
- Stellar mass: Determined within a 30 kpc aperture, aligning with observed data from the Galaxy

And Mass Assembly (Baldry et al., 2012) and SDSS (Li and White, 2010).

- Velocity dispersion: Calculated as $\sqrt{2E_k/3M}$, with E_k being kinetic energy and M the stellar mass within the aperture.
- Maximum circular velocity: Derived using $v_c(r) = \sqrt{\frac{G_N M(<r)}{r}}$, where $M(<r)$ is the enclosed mass at radius r .

Dark matter features include:

- Halo mass (M_{200}) and radius (R_{200}), defined at the virial radius.
- Halo group mass, referring to the total mass of dark matter subhalos within a group, as identified by the SUBFIND and FOF algorithms.

4.2 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 Figure 1, of every galaxy.

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.

In the *second approach*, applied to Datasets 2 and 4 where halo history was not provided, SPN generates models of matter distribution from snapshots. These datasets, created from the fiducial and dark matter-only runs, focus on galaxy-halo systems and halos between redshifts 3.5 and 1.7. This redshift range was chosen to model the universe more accurately, as star formation peaked during this period. The generative model trained on this data was tested by querying

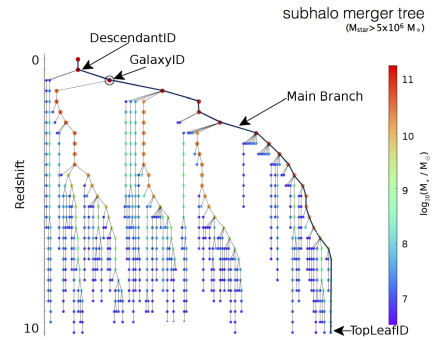


Figure 1: Merger history of a galaxy with stellar mass, $M_{\text{star}} > 10^5 M_{\odot}$. Figure from the Eagle Database (McAlpine et al., 2016). A galaxy’s present state results from mergers over billions of years. Redshift 0 represents the present, while redshift 10 corresponds to a 12 Gigayear lookback time. The galaxy’s merger history follows a main progenitor branch, shown as a thick black line in the figure. The Descendant ID represents a galaxy at a specific time, while the TopLeafID indicates the first progenitor along the main branch. All other branches are indicated with a thin line. The merger history is traced along the main progenitor branch.

properties at redshift 0 to evaluate how well SPN captures the matter distribution. The use of dark matter-only simulations helps assess how well SPNs approximate N-body calculations, given the lack of an analytic solution and the reliance on energy and momentum conservation for convergence.

5 ANALYSIS

In this section, we present and discuss the results obtained when applying the algorithm to the Eagle data. Using dark matter internal halo properties as inputs, we predict the following baryonic features: black hole mass, stellar mass, velocity dispersion, and V_{max} . These attributes result from billions of years of evolution through dissipative, nonlinear baryonic processes. While large-scale structure formation follows the Λ CDM model, smaller scales are vastly more complex.

Once the SPN captures the joint distribution over all the variables at its root node, it can be queried for conditional and marginal likelihoods of any random variable, such as stellar mass or central black hole mass. The SPN’s root node can also generate synthetic datasets following the learned joint distribution. Our trained SPN has 143 edges with 144 nodes in 20 layers. The network has 19 sum nodes, 40 product nodes and 85 leaf nodes each modeling a univariate Bernoulli distribution. The SPN took 149 seconds to be learned.

Tables 2 and 3 show minimal difference in errors

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

between the two approaches. However, the computation time to learn the joint distribution is much shorter when only snapshots are provided. Merger histories do not significantly enhance model richness. Tables 4 and 5 present similar results for dark matter properties.

The results show that SPNs can recreate mock catalogs with properties similar to those from hydrodynamic codes. Baryonic properties, which are mass-dependent, are predicted accurately, with the central black hole and stellar mass linearly related to halo mass M_{200} , and velocity dispersion and V_{\max} governed by mass and radius. The predicted and true distributions for stellar mass and central black hole mass match closely.

A key observation is that progenitor history does not improve prediction accuracy, even with increased computation time. For example, the mean squared errors for stellar mass are nearly identical with and without progenitor history, but computation time increases with progenitor history.

This trend also holds for dark matter-only runs, where errors in subhalo number and halo group mass are similar with or without progenitor history, though training with progenitor history takes longer.

Overall, the results are surprising given the complexity of numerical simulations. While our model cannot replace numerical simulations, it provides a useful tool for exploring the galaxy-halo connection and the impact of different simulation physics, as seen in semi-analytic modeling.

6 DISCUSSION AND CONCLUSIONS

We conducted an empirical study to explore the relationship between dark matter halos and their enclosed galaxies using a Sum-Product Network (SPN), a prob-

abilistic graphical model, in the context of a large cosmological hydrodynamic simulation. Our study demonstrates that SPNs offer significant computational savings, making predictions in minutes compared to the millions of CPU hours required by hydrodynamical simulations. The model accurately predicts baryonic properties like stellar mass and central black hole mass, with strong R^2 and Pearson correlation metrics. Additionally, SPNs generate synthetic datasets, enabling further exploration of galaxy-halo relationships. Comparing different approaches, we found that using random snapshots instead of progenitor histories does not greatly affect accuracy. SPNs' hierarchical structure and probabilistic nature provide enhanced interpretability over other machine learning models.

However, the model is insensitive to progenitor histories, questioning its ability to capture complex baryonic feedback. Its phenomenological nature limits its ability to simulate processes like AGN feedback or star formation. The need for extensive domain knowledge and the finite resolution of the Eagle simulations also limit its generalizability. Future work could combine SPNs with physics-based constraints, test on other simulations or observational data, and develop methods to better model temporal dependencies. Despite these challenges, SPNs show promise for tractable generative modeling in cosmology, offering efficiency and accuracy while complementing traditional simulations.

The aim of this work was to assess how dark matter properties can inform the evolutionary properties of galaxies, rather than replicating a numerically identical population. The results suggest SPNs can effectively mimic galaxy evolution in a hydrodynamic context, with runtimes in the order of minutes versus the millions of hours required by simulations. This highlights the potential of probabilistic models in analyzing complex physical phenomena and their role

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

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

in testing machine learning models. Future work will explore advanced algorithms to further integrate machine learning in cosmology.

ACKNOWLEDGEMENTS

This work is partly supported by the EPSRC grant *Towards Explainable and Robust Statistical AI: A Symbolic Approach*

REFERENCES

- Agarwal, S., Davé, R., and Bassett, B. A. (2018). Painting galaxies into dark matter haloes using machine learning. *Monthly Notices of the Royal Astronomical Society*, 478(3):3410–3422.
- Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.
- Baldry, I., Driver, S. P., Loveday, J., Taylor, E., Kelvin, L., Liske, J., Norberg, P., Robotham, A., Brough, S., Hopkins, A. M., et al. (2012). Galaxy and mass assembly (gamma): the galaxy stellar mass function at z_i 0.06. *Monthly Notices of the Royal Astronomical Society*, 421(1):621–634.
- Bengio, Y. et al. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.
- Benson, A., Pearce, F., Frenk, C., Baugh, C., and Jenkins, A. (2001). A comparison of semi-analytic and smoothed particle hydrodynamics galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 320(2):261–280.
- Bertschinger, E. (1994). Cosmic structure formation. *Physica D: Nonlinear Phenomena*, 77(1):354 – 379. Special Issue Originating from the 13th Annual International Conference of the Center for Nonlinear Studies Los Alamos, NM, USA.
- Bueff, A., Speichert, S., and Belle, V. (2018). Tractable querying and learning in hybrid domains via sum-product networks. *CoRR*, abs/1807.05464.
- Butz, C. J., Oliveira, J. S., and dos Santos, A. E. (2017). On learning the structure of sum-product networks. *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–8.
- Cavuoti, S., Brescia, M., Riccio, G., Longo, G., et al. (2018). Stellar formation rates in galaxies using machine learning models. *arXiv preprint arXiv:1805.06338*.
- Chavira, M. and Darwiche, A. (2008). On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6-7):772–799.
- Davies, E. et al. (2021). Galsim: Combining simulations and machine learning for galaxy morphology analysis. *Astronomy and Computing*, 35:100453.
- Dolag, K., Borgani, S., Schindler, S., Diaferio, A., and Bykov, A. M. (2008). Simulation techniques for cosmological simulations. *Space science reviews*, 134(1-4):229–268.
- Gens, R. and Domingos, P. (2012). Discriminative learning of sum-product networks. *Advances in Neural Information Processing Systems*, pages 3239–3247.
- Gens, R. and Domingos, P. (2013). Learning the structure of sum-product networks. *International conference on machine learning*, pages 873–880.
- Gheller, C., Wang, P., Vazza, F., and Teyssier, R. (2015). Numerical cosmology on the gpu with enzo and ramses. In *Journal of Physics: Conference Series*, volume 640, page 012058. IOP Publishing.
- Guest, D., Cranmer, K., and Whiteson, D. (2018). Deep learning and its application to lhc physics. *Annual Review of Nuclear and Particle Science*, 68:161–181.
- Heitmann, K. et al. (2021). Precision cosmology with machine learning emulators. *Astrophysical Journal*, 909:122.
- Ho, S. et al. (2022). Astrophysics with foundation models: Prospects and challenges. *Nature Astronomy*, 6:489–495.
- Hsu, W., Kalra, A., and Poupart, P. (2017). Online structure learning for sum-product networks with gaussian leaves. *CoRR*, abs/1701.05265.

- Kamdar, H., Turk, M., and Brunner, R. (2016). Machine learning and cosmological simulations. *American Astronomical Society Meeting Abstracts# 227*, 227.
- Kisa, D., Van den Broeck, G., Choi, A., and Darwiche, A. (2014). Probabilistic sentential decision diagrams. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Kobayashi, K. et al. (2022). Predicting baryonic effects on the matter power spectrum using machine learning. *Monthly Notices of the Royal Astronomical Society*, 511:3453–3463.
- Koller, D., Friedman, N., and Bach, F. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
- Li, C. and White, S. D. (2010). Autocorrelations of stellar light and mass in the low-redshift universe. *Monthly Notices of the Royal Astronomical Society*, 407(1):515–519.
- Liang, Y., Bekker, J., and Van den Broeck, G. (2017). Learning the structure of probabilistic sentential decision diagrams. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Llinares, C. (2017). The shrinking domain framework i: a new, faster, more efficient approach to cosmological simulations. *arXiv preprint arXiv:1709.04703*.
- Lucie-Smith, J. et al. (2023). Learning galaxy-halo relationships with graph neural networks. *Monthly Notices of the Royal Astronomical Society*, 519:501–516.
- McAlpine, S., Helly, J. C., Schaller, M., Trayford, J. W., Qu, Y., Furlong, M., Bower, R. G., Crain, R. A., Schaye, J., Theuns, T., et al. (2016). The eagle simulations of galaxy formation: Public release of halo and galaxy catalogues. *Astronomy and Computing*, 15:72–89.
- Molina, A., Natarajan, S., and Kersting, K. (2017). Poisson sum-product networks: A deep architecture for tractable multivariate poisson distributions. *AAAI*, pages 2357–2363.
- Molina, A., Vergari, A., Di Mauro, N., Natarajan, S., Esposito, F., and Kersting, K. (2018). Mixed sum-product networks: A deep architecture for hybrid domains. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- More, S., Kravtsov, A. V., Dalal, N., and Gottlöber, S. (2011). The overdensity and masses of the friends-of-friends halos and universality of halo mass function. *The Astrophysical Journal Supplement Series*, 195(1):4.
- Mörtsell, E. (2016). Cosmological histories from the friedmann equation: The universe as a particle. *European Journal of Physics*, 37(5):055603.
- Peharz, R., Vergari, A., Stelzner, K., Molina, A., de Campos, C. P., and Kersting, K. (2020). Randomly assembled tractable probabilistic models. *Journal of Machine Learning Research*, 21(148):1–60.
- Planck, Ade, P., Aghanim, N., Armitage-Caplan, C., et al. (2014). Planck 2013 results. xvi. cosmological parameters. *Astron. Astrophys.*, 571:A16.
- Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690.
- Rahman, T. and Gogate, V. (2014). Hybrid probabilistic models with tractable inference. *Artificial Intelligence*, 266:196–225.
- Rashwan, A., Zhao, H., and Poupart, P. (2016). Online and distributed bayesian moment matching for parameter learning in sum-product networks. In *Artificial Intelligence and Statistics*, pages 1469–1477.
- Ryden, B. (2016). *Introduction to cosmology*. Cambridge University Press.
- Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I., Helly, J. C., et al. (2014). The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554.
- Somerville, R. S. and Davé, R. (2015). Physical models of galaxy formation in a cosmological framework. *Annual Review of Astronomy and Astrophysics*, 53:51–113.
- Springel, V. (2005). The cosmological simulation code gadget-2. *Monthly notices of the royal astronomical society*, 364(4):1105–1134.
- Springel, V., White, S., Tormen, G., and Kauffmann, G. (2001). Populating a cluster of galaxies-i. results at [formmu2] z= 0, *mras* 328 (dec., 2001) 726–750. *arXiv preprint astro-ph/0012055*.
- Tamosiunas, A. et al. (2023). Semi-supervised learning for galaxy classification with limited labeled data. *Astronomy & Astrophysics*, 672:A1.
- Vergari, A., Mauro, N. D., Esposito, F., and Peharz, R. (2021). Compositional generative models with tractable inference. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*.
- Villaescusa-Navarro, F. et al. (2021). The camels project: Machine learning cosmological and astrophysical constraints from galaxy catalogues. *Astrophysical Journal*, 915:71.
- Xu, X., Ho, S., Trac, H., Schneider, J., Poczos, B., and Ntampaka, M. (2013). A first look at creating mock catalogs with machine learning techniques. *The Astrophysical Journal*, 772(2):147.