**Background**

Evolution of galaxy formations are complex non-linear problems with a wide range of physical phenomena like dark matter gravitational collapse in haloes down to star formation, metal enrichment, large scale structure formation and other very different phenomena. [1], [2] [106]. Despite the success of galaxy formation theory [121], developing a complete theory of galaxy formation and evolution is a challenging and intimidating task that require a lot of information [104] since the properties we observe are the result of multiple complex cosmological phenomenon and mechanisms that have occurred throughout the galaxies history (assembly history) [105]. The best solution to these challenges is science fiction inspired: creating and simulating real universe and observing them. However, this approach is unrealistic, at least as of now. The best solution available today is to compare observations with results from supercomputer simulations that test the theoretical ideas [104]. To do so require a lot of information regarding galaxies and the universe, which is what the “Evolution and assembly of galaxies and their environment” (EAGLE) simulation suite [97], [96] released by Virgo Consortium provides. The Eagle simulation suite is a set of cosmological hydrodynamical simulations [2] and provide a powerful resource to understand galaxies and their evolution. This is because the EAGLE project provides key observational data regarding the structure, relationships, and information of galaxy formations. This simulation suite includes a wide range of galaxy information regarding their characteristics and properties and can therefore be used to study and compare assembly histories of galaxies and is therefore a helpful tool to understand and solve the complexity of galaxy formation problems [1]. However, the EAGLE simulation suite only provides valuable data and information regarding the complex non-linear problems related to galaxy formations.

To actually solve such problems, the literature suggests two strategies: cosmological hydrodynamical simulations [104], [141], [140], [139] (Carlberg et al., 1990; Katz et al., 1992), and semi-analyic models (SAMs) (White & Frenk, 1991). Hydrodynamical simulations manages to solve the problems related to fluid dynamics and gravity at the same time, but it comes at the cost of being computational expensive, especially for high-resolution simulations [2]. Therefore, the less computational expensive option of the two, SAMs have become a popular choice. SAMs manages to explain galaxy formation and evolution in a straightforward manner, while being able to produce greater quantity of simulations than the more computational expensive alternative that is hydrodynamical simulations [2] [106] [130]. Furthermore, SAMs are extremely flexible in terms of utilizing different method to explore physical phenomena [139], [138], [137], [136], [135], [134], [133], [132], [131], [130], [124] (see e.g. Cole, 1991; Lacey & Silk, 1991; White & Frenk, 1991; Kauffmann & White, 1993; Kauffmann et al., 1993; 1999; Cole et al., 1994; 2000; Bower et al., 2006; Somerville et al., 2008; Lagos et al., 2018) [2]. The main disadvantage of SAMs is that their processes and results include a greater degree of approximation than other methods mentioned. However, the degree and scope of the effect and impact of this are not determined yet and comparison between SAMs and hydrodynamical simulations have illustrated a good concurrence [106] [130].

The backbone of galaxy formation SAMs is the mass assembly history of haloes, which is a necessity to simulate consistent evolutions of galaxies [2], [118], [125]. Halo merger trees provide exactly that information, hence consistent and robust merger trees are required to make theoretically reliable predictions from SAMs [118]. The idea of using merger trees in SAMs was introduced by Kauffmann et al (1993). Halo merger trees are a data structure that describes hierarchical formation history. In galaxy formation theory, halo mergers trees trace progenitors back in time and captures the growth and merges of dark matter haloes within a galaxy. [2]

There are multiple ways to construct merger trees, the most popular method is based on high resolution dark matter only (N-body) simulations. This method manages to produce realistic evolutionary history of haloes and can produce multiple merger trees at the same time. This, however, comes with the cost of being computationally expensive and require long runs.[135], [132], [129], [128], [127], [126], [125], (Kauffmann et al., 1999; Hatton et al., 2003; De Lucia et al., 2004; Croton et al., 2006; Bower et al., 2006; Guo et al., 2011; Lee et al., 2014). [2]

Another simple but fairly effective method to produce halo merger trees that is based on Monte Carlo simulations (Kauffmann & White 1993; Kauffmann et al. 1993; Cole et al. 1994) and extended Press-Schechter formalism (Bond et al. 1991). This method can only construct one merger tree at a time which often is inconsistent with simulations (Jiang & van den Bosch 2014). Even though the downside of this method is well known, it is still described as especially reliable for hierarchical universe mass history descriptions and efficient both in terms of speed and quality results with high mass resolution. [133], [123], [122], [122], [121], [120] (Lacey & Cole 1993; Somerville & Primack 1999; Cole et al. 2000; Somerville et al. 2008; Benson & Bower 2010; Ricciardelli & Franceschini 2010).

Other methods like cosmological N-body simulations rely heavily on complex clustering algorithms to find haloes and build trees and have mass resolution limit which can make it challenging to identify substructures for halo finders while being computationally expensive. [119], [118], [117], [116], [115], [114] (Knebe et al. 2011; Avila et al. 2014; Gómez et al. 2022; Muldrew et al. 2011; Onions et al. 2013; Elahi et al. 2013).

In other words, there exists multiple methods for merger tree constructions that produces decent results, but these methods are computationally intensive, rely on complex algorithms, produce inconsistent constructions, or can only produce one tree at a time. Recently, however, machine learning techniques, more precisely, deep learning algorithms have been utilized and experimented with in astrophysics [109], [107], [110], [111], [112], [113]. Deep learning algorithms have the ability and capacity to process large dataset and extract patterns, features, structures and relationships from the data.[2] Within the deep learning framework, there are specifically one method which are designed to and manages to capture hierarchical features, which is similar to those of merger trees: convolutional neural networks (CNNs) [108], [113] [ (ho et al. 2019). CNNs have been used to multiple astrophysical tasks such as morphologically classify galaxies [112], [111], [110], [109] (Dieleman et al. 2015; Kim & Brunner 2017; Barchi et al. 2020; Cavanagh et al. 2021), segment large-scale structures in the universe [108] (Aragon-Calvo 2019), measure dynamical mass of galaxy clusters [113] (ho et al. 2019), and map N-body and hydrodynamical simulations [107] (Wadekar et al. 2021).[3]

The introduction of machine learning and deep learning in the astrophysical field have opened up for new methods, for example utilizing these methods in galaxy formation theory, like making SAMs of galaxy formation more robust and reliable, so that they can be able to simulate large upcoming galaxy surveys, which will make it possible to compare theory with observations faster [3]. This is exactly what [2] have done. They propose a new method for halo merger tree construction using machine learning methods, more specifically, generative machine learning models. Their proposed method utilizes large volume simulations to generate new halo merger trees that are consistent with simulations at a reasonable computational cost. [2] The generated trees should be similar to, have the same characteristics and be consistent in terms of structure and patterns that the training data contain. The results are generated merger trees, which are represented as matrices of tree variables, similar to trees from the EAGLE simulation suite (Schaye et al., 2015; Crain et al., 2015).

The three variables [2] chosen to use in their model are the most basic input variables required by SAMs, which is the mass of the progenitors, the distance to the main branch and the progenitor type (which have become a more fundamental part of structure formation studies, merger tree production and SAMs seesing [116]) [133], [106], [105], [104], (Cole et al. 2000; Benson 2012; Croton et al. 2016; Cora et al. 2018). The variables are represented in matrix form, then these matrices are “merged” together to construct an image like structure with three channels, where each channel represent one of the variables.

The machine learning architecture [2] proposed is a neural network-based model known as a generative adversarial network (GAN), more specifically, a deep convolutional GAN (DCGAN) [103], [102] (Dosovitskiy et al., 2015; Radford et al., 2015) containing an encoder-decoder architecture [101]. The goal of the GAN is to learn a compressed representation of the training data, then generate new data by simulating samples in the compressed representation space.

The GAN model in [2] manages to reconstruct training images and generate new consistent, robust, and high-resolution halo merger trees. These generated trees have the same statistical structures and patterns as the training data, which means they can be utilized in similar ways as the training halo merger trees, e.g. in SAMs.

While [2] focuses on developing a new vanilla framework, [3] extend the proposed model and solution and improve the results. Additionally, [3] also focuses on how the merger trees from different halo merger tree finder/builder algorithms behaves/performs with the new machine learning based generation/simulation.