**Evaluation**

Evaluating generative models is a challenging task given there are no gold labels. One solution to evaluating the generated merger trees would be to evaluate a SAM of galaxy formation that is populated and trained with the generated merger trees and compare these results with other SAMs [2]. However, given that SAMs and galaxy formation is not the main aspect of the thesis and that my knowledge regarding these topics are extremely limited, this would be extremely hard.

As stated in paper 2 and 3, generated trees should resemble the real merger trees that are present in the training data [2]. That means the generated merger trees should not have reappearing haloes after a merger, meaning to there should be no gap in a branch (column), if there is no progenitor present, this should be represented by a zero in all three channels (variables), the last row should contain one halo only, which should be in the main branch [2], and other structural aspects in a generated merger tree such as no gap between branches. These structural continuities are important since SAMs require consistent and well-constructed merger trees to dependable predictions [2].

One central aspect of the generated merger trees are the three variables and how they are generated. Mass is arguably the most important variable of the merger tree, since it describes the mass assembly history of a haloes [2]. One assumption of SAMs is that in a branch, progenitor mass usually increases monotonically, or preserve its mass. But this is not a requirement since the expected behavior could be changed given other variables behavior [124], [139], [133], [134], [130] and since the training data does not strictly follow that requirement as will be shown. The mass monotonicity requirement might be hard to evaluate, especially given other variables and their effect on the mass behavior. This is true for all three variables; however, it will be used as a consistency measure along with other methods to evaluate the generated merger trees.

The choice and preferred evaluation metric of paper 2 and 3 is the Kolmogorov-Smirnov (KS) test, which compare the real and generated distributions in terms of similarities. In terms of utilizing the KS test, Paper 2 focuses on “the distance between merging progenitors for several snapshots before the merger takes place” [2], while paper 3 apply the KS test mainly on the mass variable to verify that the generated mass values are close to the original mass distribution, which implies that the generated mass assembly history are generated somewhat correctly [3].

Lastly, as mentioned in paper 2 and 3, complexity of the generated merger trees is also important. The reason for this is that the generated trees should resemble the original data which consists of trees with 5 to 10 branches, and an average subbranch length of over 6. Additionally, being able to construct merger trees with a large number of branches and long branches is great for applications of the generated trees, and it proves that the model has managed to capture the complexity in the original data. Complexity in terms of merger trees mainly mean two things: the number of branches and the length of subbranches. A decent argument for complexity as an evaluation a counterexample: If a model manages to create consistent trees with a great KS test statistic, but the tree is only a main branch, that means the model haven’t been able to capture and learn the patterns, structures and complexity of the training data.

To evaluate the generated images, four main methods will be utilized: A structure consistency check function will be used to examine whether the correct structure is captured in the generated images, a complexity check of the number of branches generated and the length of the subbranches, a variable consistency check which look at the behavior of the variables within a given tree and branch, and lastly the KS test which can tell something about the distribution of the variables of the generated images compared to the distribution of the original data and whether it is likely that they come from the same distribution or not. With these four methods of evaluation, it should be possible to evaluate the generated merger trees in terms of structure, complexity, variables and distribution and compare them to the original data. Additionally, with these four evaluation methods, it will be possible, if the results allow it, to understand, analyze and compare differences in different results. For example, a consistency vs. complexity tradeoff which will be examined in the “Original data baseline - consistency and complexity analysis” section. Details and explanations regarding each evaluation method will be presented in their individual subsection.