**My data**

**Three variables**

Given the arguments and results from the papers and the results of my data analysis, there are choices related to data that has to be made to continue. Even though the mass assembly history is the most important feature of a merger tree, hence the most significant variable to generate, the results from paper 2 and 3 suggests that including other variables improve the results. Including the distance variable affects the mass generation positively, and the same is true for the progenitor type variable [3]. However, the KS test indicate that including all three variables have the best performance in terms of mass reconstruction and generation, which is the most significant part of the generation [2]. This is backed up further by paper 3 which shows by a comprehensive KS test result analysis that including all three variables yield the best results in terms of mass generation [3].

Given that both paper 2 and 3 obtain the best results with three variables, there is a reasonable amount of evidence backing up that using three variables is the best choice for such a generation task. Therefore, the data I will utilize in the generation process will contain all three variables.

**Number of branches**

Paper 2 only used trees with six branches and obtained good results. My data analysis shows that given the number of branches, there are most merger trees with six branches. Since paper 2 utilize merger trees with six branches only, and that in the training data there are most merger trees with six branches, it is reasonable to argue that if such an approach is desirable, using merger trees with six branches is the most practical choice. This is because of data availability and size, and connection to reproduction and results from paper 2. Therefore, one dataset that will be experimented with will only contain merger trees with 6 branches.

Paper 3, however, have a different approach to the number of branches. First of all, paper 3 utilize merger trees with at least 6 branches, since they want the generated merger trees to be able to capture and reproduce complex tree structures [3].

Furthermore, paper 3 test the number of branches used when training their model thoroughly. Given the available data, they test merger trees with up to 19 branches [3]. Their method is to use merger trees with 6 – x branches, where x ranges from 10 to 19. Overall, the results suggest that generating merger trees with many branches (above 15) is harder than generating merger trees with less branches, this is due to lack of data points of larger trees and that larger trees are more complex [3]. Using merger trees with 6 – 10 branches yield decent and consistent results for all experiments conducted by paper 3 and are especially good for generating merger trees with a smaller number of branches. Due to these results, its availability and training speed the full dataset that will be considered from in this paper will be using merger trees with 6 – 10 branches.

**Data processing**

In terms of data processing, my data analysis shows that the subhalo variable is best of when left alone without any data processing. The mass variable will be normalized with the standard min max normalization. And the distance variable will be log scaled then normalized.

However, there will also be experiments with the original dataset that is not processed in any way. The reasoning behind this choice is because paper 2 and 3 doesn’t mention anything regarding data normalization and processing, which implies that the papers didn’t do anything with the data. If that is true and they managed to get good results generation well-constructed merger trees, then it is worth trying that approach.

**Dataset used from this point on:**

These are the following datasets that will be considered and experimented with from this point on:

* Original dataset
  + Three variables
  + 6 - 10 branches
  + No normalization
  + No preprocessing
* 6 branches original dataset
  + Three variables
  + 6 branches only
  + No normalization
  + No preprocessing
* Normalized full dataset
  + Three variables
  + 6 - 10 branches
  + Normalization of distance and mass variables
* Normalized 6 branches dataset
  + Three variables
  + 6 branches
  + Normalization of distance and mass variables

**Activation function**

Lastly, as discussed in the data analysis, I will make a different choice than the papers regarding activation function and use the relu. However, I will also experiment with the sigmoid activation since that is what the papers do.