**Original data baseline - consistency and complexity analysis**

To understand and evaluate the results, it essential to understand the original data and utilize the original data as a baseline for the results. The methods to evaluate the original data is to analyze the complexity and structural and variable consistency. This will ensure and verify that the original data in fact possess the desired features and structures to a close to perfect degree. The work and results in this section can be found in all notebooks that start with 8 and notebook 1.

The data that can be used as a baseline is the full original dataset containing 38 348 merger trees, the dataset containing 8 161 merger trees with six branches and the samples of generated trees from Robles et al. (2022) containing 35 000 merger trees with five to ten branches.

Consistency datasets:

|  |  |
| --- | --- |
| **dataset** | **Consistency rate** |
| **Full dataset** | 97.51 % |
| **Six branches** | 98.17 % |
| **Generated merger trees Robles et al. (2022)** | 0.0 % |

An important aspect of consistency is to understand what inconsistencies occur and the frequency of these inconsistencies, to understand the issues with the generated data. Here is an overview of the inconsistencies, showing how many of the inconsistent images are inconsistent because of the given inconsistency:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **dataset** | **Zero index** | **Dist zero in main branch** | **Gap between branches** | **Gap within branches** | **Last descendants exist** |
| **Full dataset** | 0.0 % | 0.0 % | 68.52 % | 35.15 % | 0.0 % |
| **Six branches** | 0.0 % | 0.0 % | 47.65 % | 53.03 % | 0.0 % |
| **Generated merger trees Robles et al. (2022)** | 100.0 % | 100.0 % | 68.55 % | 99.89 % | 0.0 % |

% of the inconsistent trees with 2 or more inconsistencies

|  |  |
| --- | --- |
| **dataset** | **% of inconsistent trees with 2+ inconsistencies** |
| **Full dataset** | 3.37 % |
| **Six branches** | 0.67 % |
| **Generated merger trees Robles et al. (2022)** | 100.0 % |

**Training data**

It is noteworthy that the original dataset only has two kinds of inconsistencies, namely gaps within a branch and gaps between branches. One question is why these inconsistent merger trees are in the data set to begin with, and whether or not they should be removed from the dataset. Additionally, filtering the dataset to trees with only six branches, we observe that the ratio of merger trees with gaps between branches drop. This is expected since the number of branches drop, hence the possibility to have gaps between branches decrease. However, the ratio of gaps within branches increases when only using merger trees with six branches.

Furthermore, the percentage of the inconsistent trees in the original dataset with two or more inconsistencies is fairly low. Only 3.37% of the 2.49 % of inconsistent images have two or more inconsistencies. Simply put: 0.09% of all original training merger trees have two or more inconsistencies, which is very low.

The original training data is not perfect, but around 98% consistency is a tolerable and reasonable ratio of consistent merger trees for training data. 2% inconsistent trees in the training provide the opportunity to remove the inconsistent trees and have a training data set consisting of only 100% consistent merger trees.

The full training data have an average of 8.11 branches per tree, ranging from 5 to 10 branches in a merger tree. The majority of the trees have 5 to 8 branches, but 12.8% of the trees have 9 branches and 10.6% of the trees have 10 branches. The average length of the subbranches in the original dataset is 6.58

Knowing the properties of the training data in terms of consistency and complexity gives a good understanding of what to expect for the generated merger trees.

**Generated merger trees Robles et al. (2022)**

The generated merger trees from Robles et al. (2022) have 0 consistent merger trees. All generated merger trees have the zero-location inconsistency and distance being nonzero in the main branch inconsistency. Additionally, almost all generated merger trees (99.89%) have the gap within a branch inconsistency. The only positive remark regarding consistency and the generated merger trees from Robles et al. (2022) is that the last descendent consistency is consistent for all generated merger trees.

Given the quantity and volume of inconsistencies, it is no surprise, it is actually quite obvious that all generated merger trees have two or more inconsistencies. Basically, all generated merger trees have zero-location, distance in main branch not zero and gap within branch inconsistencies. This is not a solid result or baseline in terms of structural consistency, however, the generated trees from Robles et al. (2022) performs very well in terms of complexity.

The average number of branches in the generated trees are 8.72, which is 0.6 larger than the original training data. The number of branches ranges from 5 to 10, with only 59 out of 38 500 trees having 5 or 6 branches. That means 99.85% of the generated trees have 7 or more branches and a total of 30.92% of the generated trees have 10 branches. Adding the fact that the average length of the subbranches being 10.43, means the complexity of the generated trees are extremely good, even better than the original dataset. But clearly, this comes at the cost of structural consistency.

**Overall**

Two baselines for generated images have been set. There is a clear tradeoff between consistency and complexity in the generated data compared to the original data. The original data is very structural consistent and have a decent complexity. However, the generated merger trees from Robles et al. (2022) are incredibly complex but are structural inconsistent. This tradeoff is going to be observed and analyzed in other parts of the generative task, and is going to be a central aspect of the evaluation and model adaption.