# Evaluation of ASTRAL-Pro's Gene Tree Tagging Accuracy

Saharsh Maloo

May 11, 2025

## 1 Introduction

Species trees represent the evolutionary relationships among species, while gene trees trace the evolutionary history of specific genes across species. Due to events such as Gene Duplication and Loss (GDL), gene trees often differ from the underlying species tree, making accurate species tree inference a central challenge in compuatational biology. Reconstructing species trees is essential for understanding evolutionary history, interpreting gene function, and supporting biodiversity and conservation efforts.

Astral-Pro is a widely used algorithm designed to infer species trees in the presence of gene duplication and loss. However, its accuracy has not been systematically evaluated under a broad range of evolutionary conditions, particularly across varying duplication/loss rates and levels of incomplete lineage sorting (ILS). This thesis evaluates the accuracy of ASTRAL-Pro by simulating gene trees and comparing the algorithm's infered species trees to the ground truth.

Understanding where ASTRAL-Pro succeeds or fails in species tree inference can help guide its application in biological research and inform future improvements to the algorithm.

## 2 Background

Gene tree discordance and incomplete lineage sorting (ILS) are two common factors that hinder accurate species tree inference in phylogenetics. Gene duplication occurs when an existing gene creates a copy of itself, resulting in multiple gene copies within the genome. In contrast, gene loss refers to the deletion of a gene from the genome, which can obscure the evolutionary signal across species.

ASTRAL (Accurate Species TRee ALgorithm) is a widely used tool for estimating unrooted species trees from unrooted gene trees. ASTRAL-Pro (ASTRAL for PaRalogs and Orthologs) extends this approach by incorporating multi-copy genes, enabling it to account for gene duplication and loss (GDL).

These tools are essential for reconstructing evolutionary relationships that can inform areas such as drug discovery, conservation biology, and disease research.

SimPhy is a simulation framework that generates phylogenomic data under user-defined evolutionary scenarios, including ILS, GDL, horizontal gene transfer (HGT), and gene conversion (GC). It produces species trees, locus trees, gene trees, and mapping files (.mapsl and .maplg), which can be used to generate a known ground truth. This ground truth enables rigorous benchmarking of species tree inference algorithms like ASTRAL-Pro.

# 3 Experimental Study

## 3.1 Overview

To evaluate the accuracy of ASTRAL-Pro's species tree inference algorithm, gene trees and species trees were simulated using SimPhy along with corresponding ground truth mappings. These mappings served to construct a reference dataset. The simulated gene trees were processed by ASTRAL-Pro, which annotated duplication nodes. Edges between non-duplication nodes were subsequently contracted to emphasize speciation events. Accuracy was assessed by computing Robinson-Foulds distances and tree-edge similarity scores on the resulting trees.

## 3.2 Simulation Design

SimPhy v1.0.2 was used to simulate gene trees, species trees, and associated evolutionary events under varying duplication-loss (DL) rates (0.0, 1e-10, 2e-10, 5e-10) and two population sizes (1e7, 5e7). Each of the eight configurations was replicated 10 times, yielding 80 datasets. For each replicate, 10 species trees were generated (-rs 10), each with 1000 locus trees (-rl F:1000) and one gene tree per locus tree (-rg 1). Duplication and loss rates were set using -lb F:$rate and -ld F:lb. Additional parameters included a fixed species branch length of 1.8e-9 (-sb) and a fixed species tree height of 1.8000003375e9 (-st). Ground truth mappings output by SimPhy enabled direct comparison to species trees inferred by ASTRAL-Pro.

## 3.3 Ground-Truth Annotation

Gene trees were annotated using SimPhy's .maplg and .mapsl files by linking each gene tree node to its corresponding locus and species tree nodes. Events at the species tree level were used to determine whether a node represented a duplication or speciation. Nodes identified as duplications were labeled with "D," producing ground-truth annotated gene trees for each replicate.

### 3.4 ASTRAL-Pro Tagging

To prepare gene trees for duplication tagging with ASTRAL-Pro, mapping files linking gene tree leaves to their corresponding species were first constructed. This was accomplished by extracting species identifiers from the prefixes of each leaf name. These mappings were then provided to ASTRAL-Pro using the -a option, and the software was executed with the -T flag to infer and annotate gene trees. The resulting trees included internal node labels indicating duplication events, which were used in downstream analyses.

### 3.5 Tree Edge Contraction

To reduce noise stemming from variation in speciation node resolution and to better assess the accuracy of duplication node tagging, a tree contraction step was applied. Specifically, any edge connecting a parent and child node both labeled as speciation (i.e., unlabeled or non-duplication nodes) was contracted. This process simplified the tree structure by collapsing regions not relevant to duplication inference.

### 3.6 Tree Distance Metrics

To evaluate the accuracy of inferred gene trees, three tree distance metrics were calculated: Robinson-Foulds (RF) distance, edge similarity, and reverse edge similarity. The RF distance was used due to its widespread adoption in phylogenetic analysis as a standard measure of topological difference, quantifying the number of bipartitions that differ between two trees.

Edge similarity was included to assess the proportion of shared internal edges between the inferred and ground-truth trees, providing a more interpretable metric for overlap in tree structure. Reverse edge similarity, which calculates the proportion of edges in the ground-truth tree recovered by the inferred tree, was used to emphasize recall of true evolutionary relationships.

## 4 Results

### 4.1 Overview

The analysis evaluates how varying duplication-loss rates and population sizes affect ASTRAL-Pro's inference accuracy. Tree comparison metrics were examined across replicates to assess trends in topological similarity and duplication node recovery.

### 4.2 General Accuracy of ASTRAL-Pro

ASTRAL-Pro demonstrated high overall accuracy, with Robinson-Foulds distances consistently below 0.2 and edge similarity exceeding 98% across all tested parameter settings. These results highlight the algorithm's effectiveness as a species tree inference tool.

All Metrics vs Population Size: Shows that larger psize consistently leads to worse NRF and lower edge similarity.

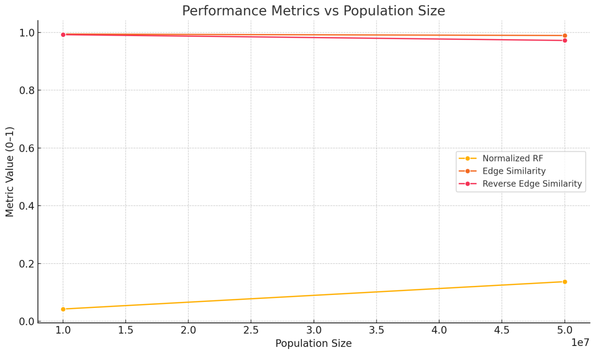

Figure 1: Performance Metrics Vs Population Size

Boxplot (NRF by psize): Distribution of error widens as psize increases.

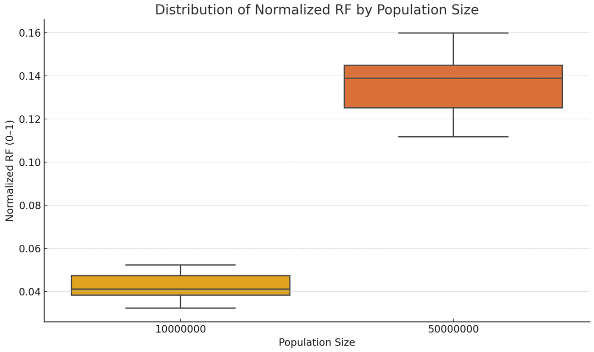

Figure 2: Distribution Of Normalized RF By Population Size

Boxplot (Edge Similarity by rate): Slight spread with higher variance at higher rates.

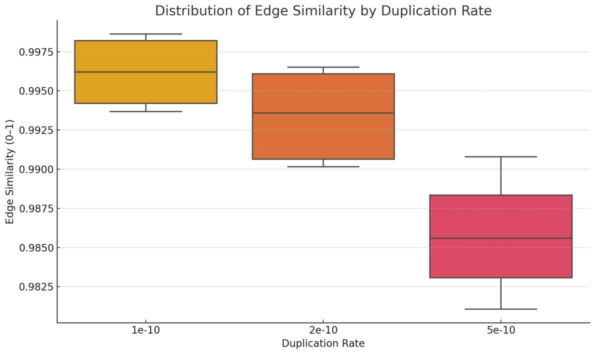

Figure 3: Distribution Of Edge Similarity By Duplication Rate

## 4.3 Normalized Error Metrics vs Duplication Rate

Error rates generally increased with higher duplication-loss rates across all metrics, except for normalized Robinson-Foulds (RF) distance under the larger population size condition. This exception may reflect ASTRAL-Pro's emphasis on recovering overall tree topology rather than accurately labeling duplication events.

Line Plot (NRF vs Rate): Topological error increases with duplication rate for population size of 1e7 but decreases for population size of 5e7.

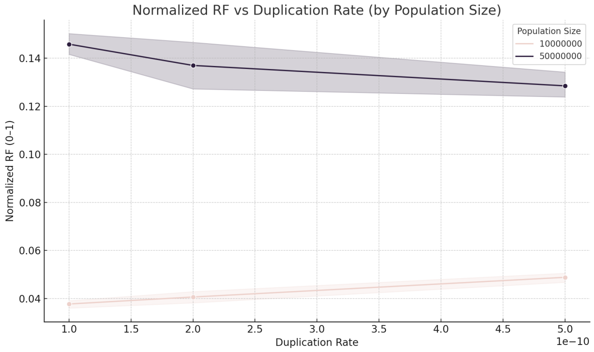

Figure 4: Line Plot Showing Average Normalized Robinson Foulds Distance across Duplication Rates

Line Plot (Edge Similarity vs Rate): Edge recovery slightly drops as duplication increases.

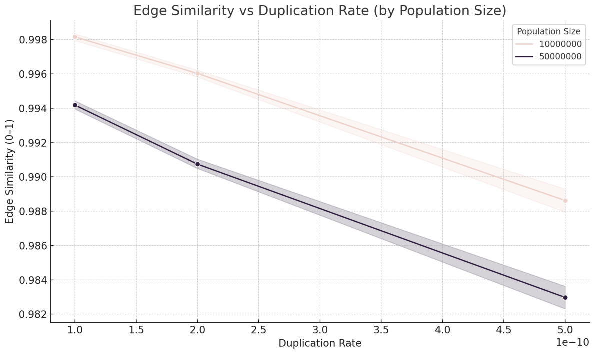

Figure 5: Line Plot Showing Average Edge Similarity across Duplication Rates

Line Plot (Reverse Edge Similarity vs Rate): Similar downward trend, especially at high duplication rates

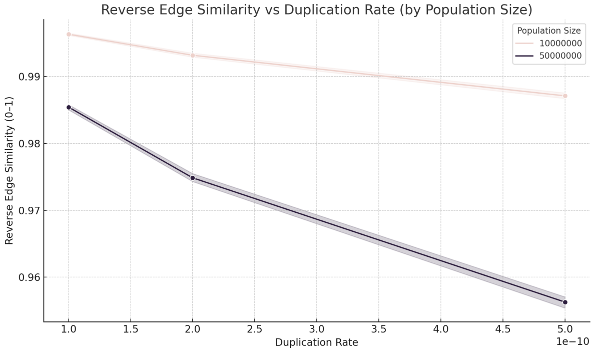

Figure 6: Line Plot Showing Average Reverse Edge Similarity across Duplication Rates

## 4.4 Heatmaps of NRF and Edge Similarity

Heatmap (NRF): Visualizes which combinations hurt inference most.

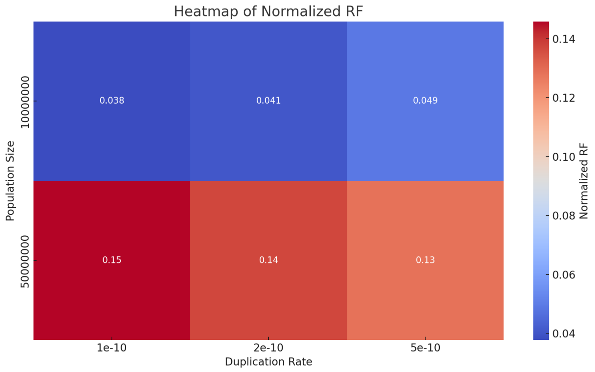

Figure 7: Heatmap Of Normalized RF

Heatmap (Edge Similarity): Highlights high-performing zones (top-left) and degradation zones (bottom-right).s

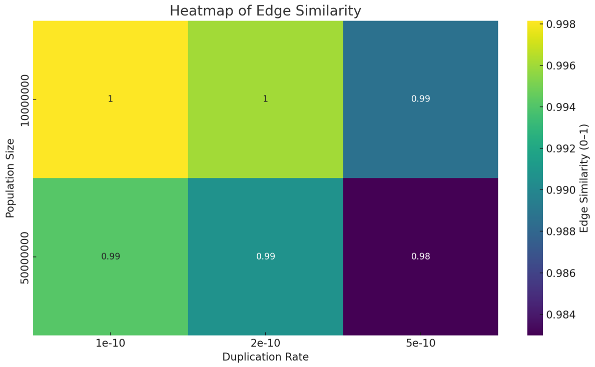

Figure 8: Heatmap Of Edge Similarity

# 5 Discussion and Conclusion

## 5.1 Inference Accuracy and General Performance

ASTRAL-Pro demonstrated high overall accuracy across a variety of simulation settings. Normalized Robinson-Foulds (NRF) distances remained below 0.2, and edge similarity exceeded 98% in most cases, highlighting the tool's effectiveness in recovering tree topology and internal edge structure. These results support ASTRAL-Pro's reliability as a species tree inference method, particularly in settings with low evolutionary complexity.

## 5.2 Impact of Population Size

Increased population size led to higher error rates across all metrics except for NRF under high duplication rates. The rise in error is likely due to greater incomplete lineage sorting (ILS), which introduces more discordance between gene and species trees. As shown in the NRF distribution plots, variability in accuracy also increased with population size, indicating reduced consistency in inference under higher ILS conditions.

## 5.3 Effect of Duplication-Loss Rate

Performance declined with higher duplication-loss rates, as seen in the downward trends of edge similarity and reverse edge similarity metrics. A notable exception was observed in the NRF values for larger population sizes, which slightly decreased with increasing duplication rates—suggesting that ASTRAL-Pro may prioritize recovering tree topology over precise event labeling in complex scenarios.

## 5.4 Interpreting Heatmap Trends

Heatmaps of NRF and edge similarity revealed that inference accuracy degrades under combinations of high duplication-loss rates and large population sizes. Conversely, the best performance occurred in simulations with both low duplication rates and small population sizes. These visualizations reinforce the importance of understanding biological context when interpreting ASTRAL-Pro results.

## 5.5 Conclusion and Future Work

Overall, ASTRAL-Pro is a robust and reliable tool for species tree inference, especially in biologically simple conditions. However, its performance is influenced by evolutionary parameters such as duplication rate and population size. Future work may focus on enhancing duplication node labeling and improving inference under high-ILS conditions. Further validation on empirical datasets will also help assess the practical applicability of these findings.

# 6 Supplementary Information

## 6.1 Software and Dataset Availability

All Software and Scripts can be found in this **GitHub**

## 6.2 Key Software Versions

- **SimPhy:** v1.0.2 (compiled for Linux 64 bit executables)
- **ASTRAL-Pro:** v1.20.3.6
- **Python Libraries:** `treeswift`, `ete3`, `pandas`, `seaborn`, `matplotlib`

## 6.3 Hardware Environment

All experiments were conducted on a Lenovo ThinkPad equipped with an Intel Core i7 processor and 16 GB of RAM. Processing approximately 80,000 gene trees across all simulation conditions required an estimated 20 minutes of compute time.