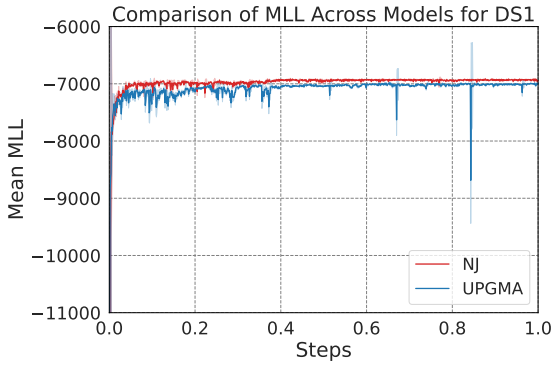

Figure R1: Comparison of different tree construction functions NJ and UPGMA on the DS1 dataset.

Table R1: Topological comparison of three tree diversity indices on DS1, DS2, and DS3 datasets. Higher values of Simpson's Diversity Index and the number of topologies accounting for the top 95% cumulative frequency indicate better diversity, while a lower frequency of the most frequent topology reflects a balanced distribution

Dataset	Statistics	MrBayes	GeoPhy	Ours
DS1	Diversity Index	0.87	0.36	0.89
	Top Frequency	0.27	0.80	0.008
	Top 95% Frequency	42	11	149
DS2	Diversity Index	0.89	0.68	0.93
	Top Frequency	0.27	0.55	0.11
	Top 95% Frequency	208	58	362
DS3	Diversity Index	0.98	0.99	0.99
	Top Frequency	0.02	0.02	0.001
	Top 95% Frequency	753	553	1163

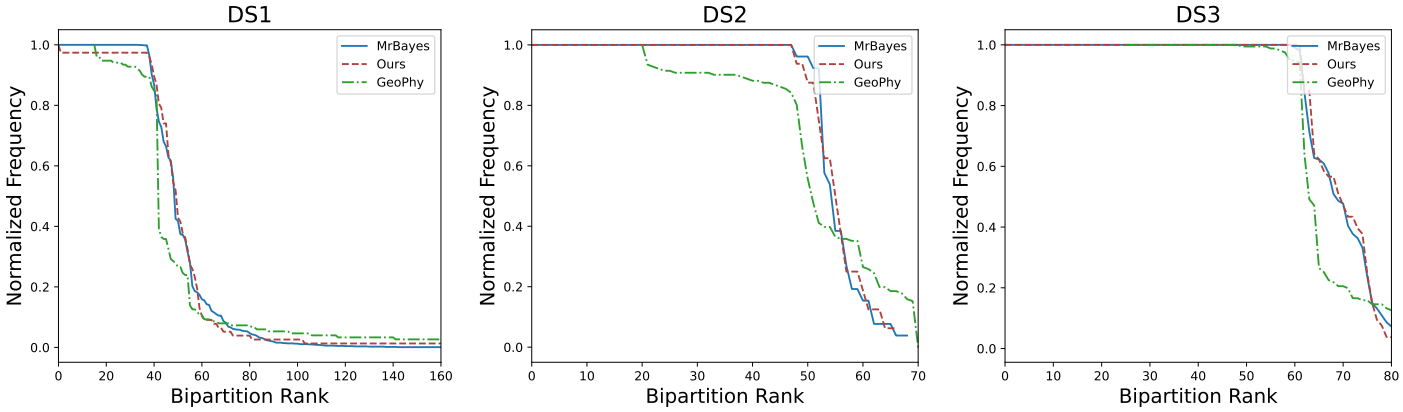

Figure R2: Comparative Bipartition Frequency Distribution in Tree Topologies for DS1, DS2, and DS3. **The closer the two curves are, the better,** suggesting that PhyloGen is highly consistent with the gold standard MrBayes approach than Geophy.

Table R2: Runtime comparison with VI algorithms on DS1 dataset.

	MrBayes	GeoPhy	PhyloGFN	ARTree	Ours
Runtime (H)	22h46m	8h10m	20h40m	62h21min	6h53min
Memory (MB)	–	1450.93	2341.80	2040.50	1051.11