MalTree · Tracing Malware Evolution from Embeddings at Scale

ICML 2026
Seoul, South Korea

PAPER CODE

Akash Amalan · Georgios Smaragdakis · Tom J. Viering Delft University of Technology

MalTree applies phylogenetics over structural, behavioral, and image features to reconstruct malware lineage automatically, matching real VirusTotal emergence dates **87% of the time.**

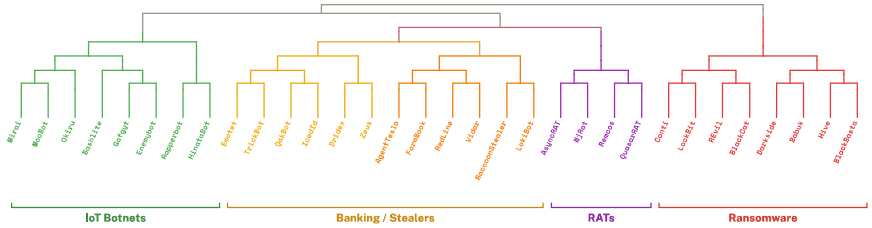
Why trace malware evolution?

- **Detection is reactive.** Signature, feature, and deep-learning detectors all train on known samples and decay as malware mutates through packing, polymorphism, and now LLM-generated code.
- **Malware moves in lineages.** Variants rarely appear in isolation; they inherit code, share builder kits, and mutate iteratively, forming a co-evolving ecosystem that mirrors biological evolution.
- **Phylogenetics is the missing lens.** The mutation, selection, and inheritance used to reconstruct viral ancestry can map malware families, yet prior work used tiny sample sets, single feature types, and never checked trees against real timelines.
- **Our goal.** Shift analysis from sample-by-sample classification to lineage-aware evolutionary modeling, validated against when strains actually emerged.

What MalTree does

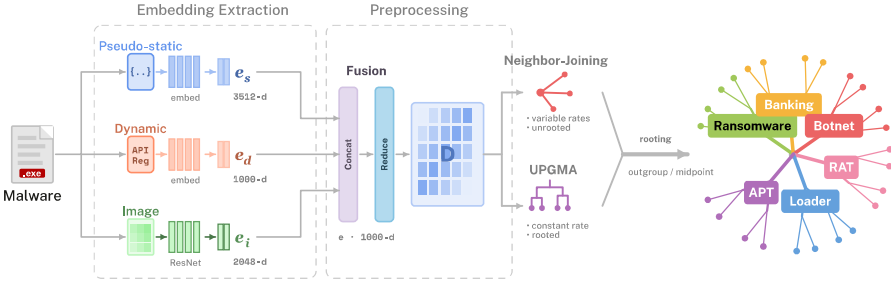
- **Phylogenetics at scale.** Validated trees from **103,883 samples** across **538 families**, the largest such analysis to date (NJ in 3 days, UPGMA in 11 h on 20 cores).
- **Temporal validation, the missing test.** The first framework to check inferred branching against real VirusTotal first-seen dates, confirming the trees capture evolution, not just feature similarity.
- **Insights it unlocks.** Families drift at **very high rates**, and we recover many documented lineages such as mirai.

Malware from a biological perspective



Like organisms, malware families share ancestry: representative families cluster into IoT botnets, banking/stealers, RATs, and ransomware, joined at a common root — the evolutionary structure MalTree reconstructs.

The MalTree pipeline



Multi-modal embeddings (pseudo-static e_s, dynamic e_d, image e_i) are concatenated and reduced; pairwise distances yield matrix D, which Neighbor-Joining or UPGMA turns into the family-level tree.

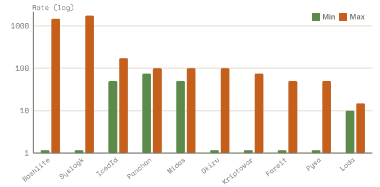
3 Key results

Temporal consistency (year)

Method	Def.	Outgrp	Mid.
NJ	0.811	**0.871**	0.631
UPGMA	0.860	0.601	0.562

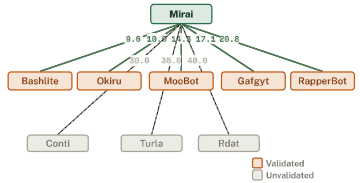
NJ with outgroup rooting is best; random ordering ≈ 0.5. Full tree, 103,883 samples, before outlier removal. After removal the temporal consistency is **0.885.**

Embedding drift varies by family



Distance / year (log scale). Bashlite and Syslogk drift >10× faster than slow families, violating UPGMA's clock assumption and motivating NJ.

Mirai lineage recovered



Symbol-set Jaccard

	Mirai	Okiru	Bashl.	Gafgyt	MooBot
Mirai	–	**0.82**	0.58	0.07	0.04
Okiru	0.82	–	0.60	0.06	0.05
Bashlite	0.58	0.60	–	0.17	0.04
Gafgyt	0.07	0.06	0.17	–	0.05
MooBot	0.04	0.05	0.04	0.05	–

Five documented descendants recovered (orange, low w); three weak links flagged (gray). Mirai–Okiru 0.82 confirms code reuse; low Gafgyt/MooBot reflects static linking, not absent lineage.

1 How good are the embeddings?

Embedding	Accuracy (%)
Image e_i	94.91 ± 0.16
Combined e	**93.69 ± 0.19**
Pseudo-static e_s	87.51 ± 0.48
Dynamic e_d	87.23 ± 0.19
Random guessing	19.36 ± 0.17

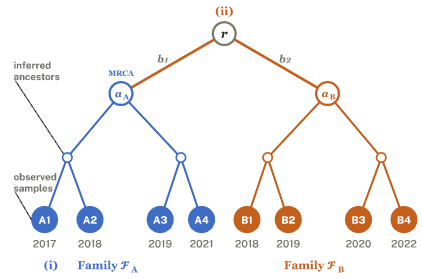
High, linearly separable accuracy confirms the embeddings carry genuine family structure, suited to distance-based phylogenetics.

What each view captures

- **Image e_i** · 2048-d — byte-level memory texture.
- **Pseudo-static e_s** · 3512-d — static features of malware dump.
- **Dynamic e_d** · 1000-d — behavioral features through sandbox.

2 How to interpret & validate

Each **leaf is an observed malware sample**; **internal nodes are inferred ancestors** (not observed, reconstructed from the data), and the root is their oldest common ancestor. Branch length measures divergence. VirusTotal first-submission dates are an independent signal, used to validate the tree.



Notation

- leaf = observed sample
- internal node = inferred ancestor
  MRCA = most recent common ancestor
- root = oldest ancestor
- b = branch length
- $d_{\mathcal{T}}(u,v)$ = Σ branch lengths on the path
- L_i = $d_{\mathcal{T}}$(leaf, ancestor)
- year = VirusTotal first-submission date

(i) Intra-family

For two samples sharing MRCA a (e.g. A_1, A_4 share a_A), the shallower leaf is the more ancestral one:

$$L_i < L_j \implies t(s_i) < t(s_j)$$

Local clock. The comparison is restricted to immediate siblings (shared author and toolchain), so it needs only local consistency, not a constant rate across families.

Temporal consistency score. For each sibling pair, check whether tree order matches date order; the score is the fraction of consistent pairs:

$$\langle L_i < L_j \wedge t_i < t_j \rangle \vee \langle L_i > L_j \wedge t_i > t_j \rangle$$

Best score 0.871 (NJ, outgroup rooting; random order ≈ 0.5).

(ii) Inter-family

The shared ancestor of two families is a *deep* node (root r). Single leaf pairs are noisy, so families are compared as aggregates by median depth:

$$\tilde{d}_{\mathcal{A}} = \text{median}\{\, d_{\mathcal{T}}(s, r) : s \in \mathcal{F}_{\mathcal{A}} \,\}$$

A shorter median depth means earlier-diverging. A directed, weighted edge records the order:

$$\tilde{d}_A < \tilde{d}_B \implies \mathcal{F}_A \to \mathcal{F}_B, \quad w = \tilde{d}_A$$

Keep only each node's minimum-weight outgoing edge to expose the primary lineage.

Outliers beyond 1.5×IQR of the family median are removed first.