Training Data Report

training_data

2025-09-18

Overview

This report summarizes the training table used for classification and physics validation. **Rows:** 155,042; **Cols:** 91; **Full rows:** 130,042; **Partial rows:** 25,000. The partial subset is explicitly flagged via `meta_partial` and should be retained for stratification or targeted imputation.

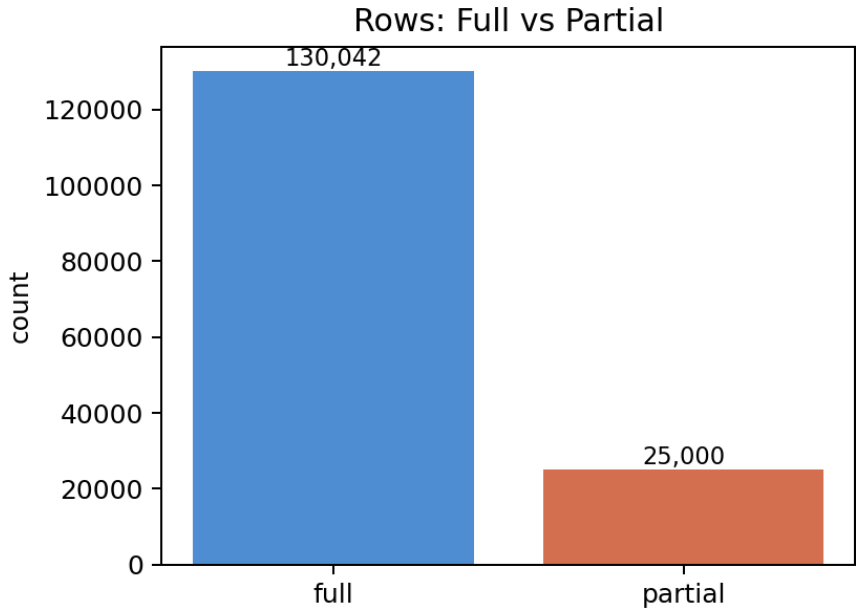


Figure 1: Full vs. Partial rows.

Group breakdown

The dataset intentionally emphasizes resonances, with smaller random and TTV cohorts for contrast.

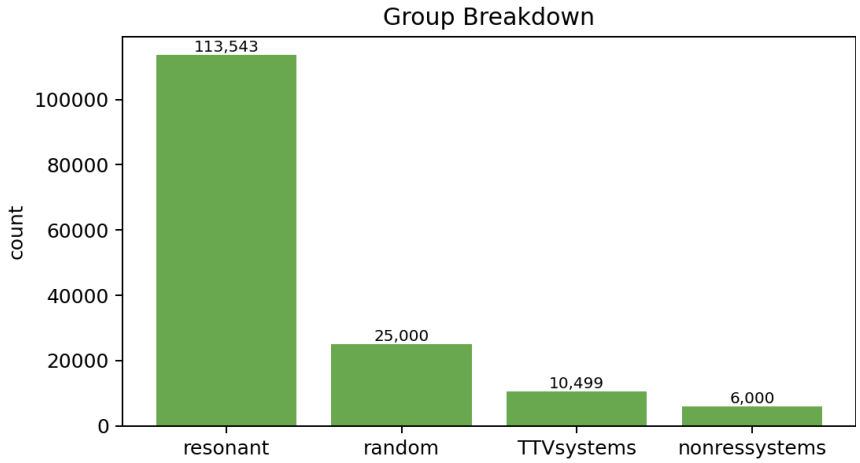


Figure 2: Group counts by `system_id`.

Core distributions (full rows)

Heavy tails are visible in kinetic/energetic families, motivating log/rank transforms and robust scaling.

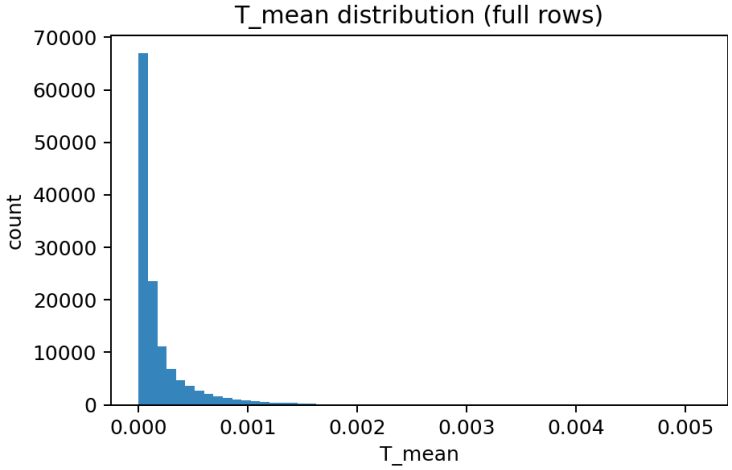


Figure 3: `T_mean`: histogram over full rows.

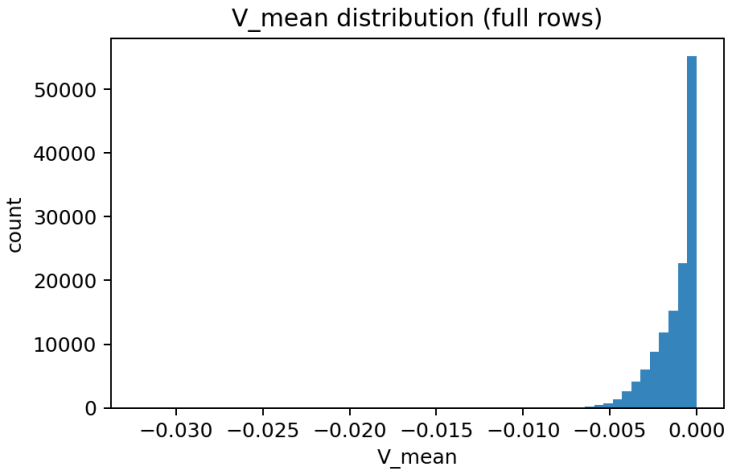


Figure 4: `V_mean`: histogram over full rows.

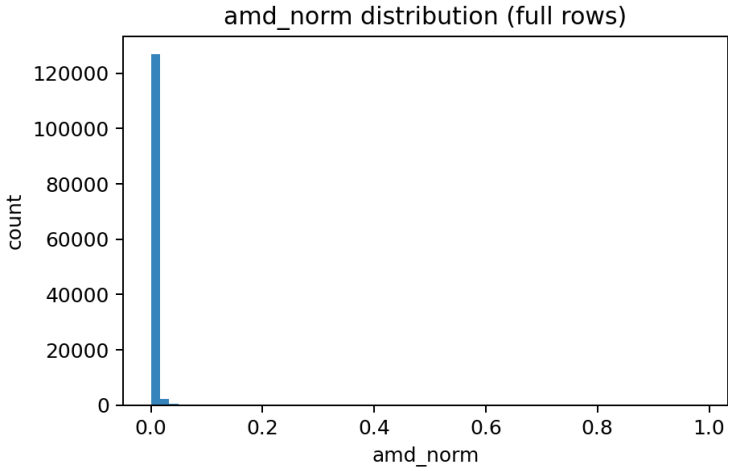


Figure 5: `amd_norm`: histogram over full rows.

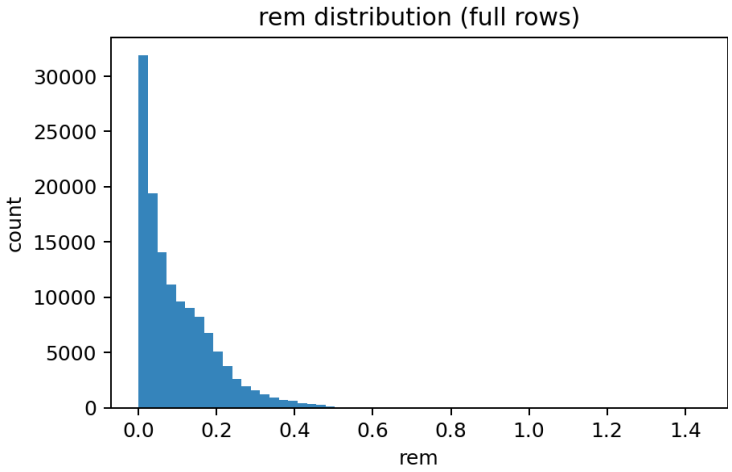
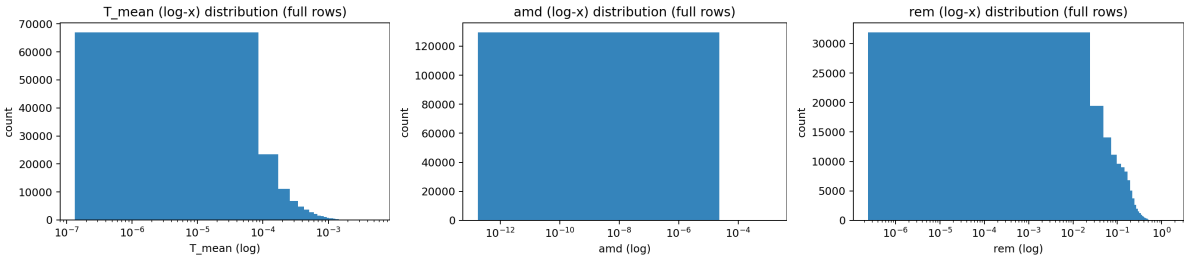


Figure 6: `rem`: histogram over full rows.

Log-scale variants



Resonances

A period-ratio panel with canonical resonance lines confirms the expected structure in the resonant cohort.

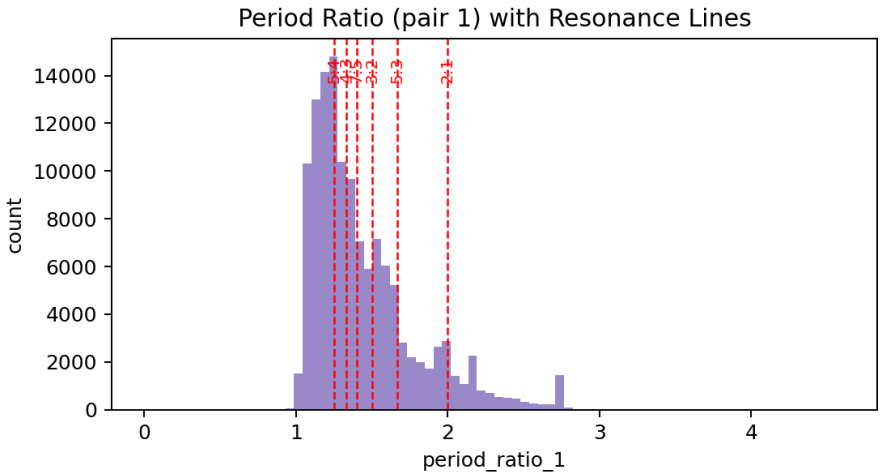


Figure 7: Period ratio (pair 1) with resonance overlays.

By group

Boxplots show dispersion across cohorts; `TTVsystems` skew large as expected, while `random` behaves as a low-variance baseline.

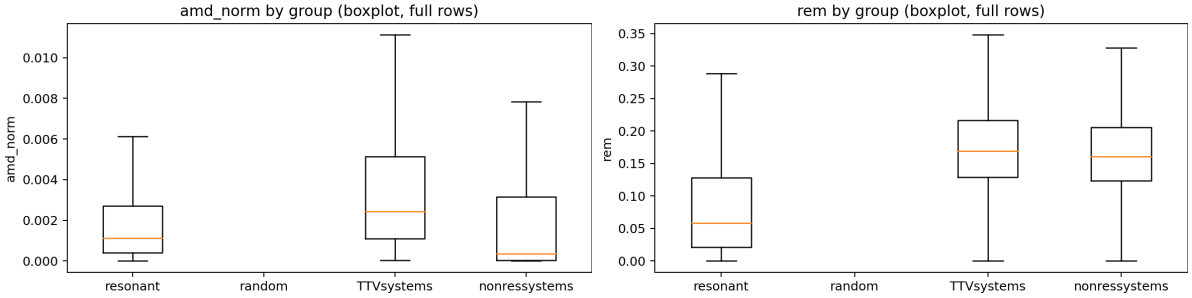


Figure 8: Group-wise distributions.

Relationships

Two representative relationships used during feature vetting.

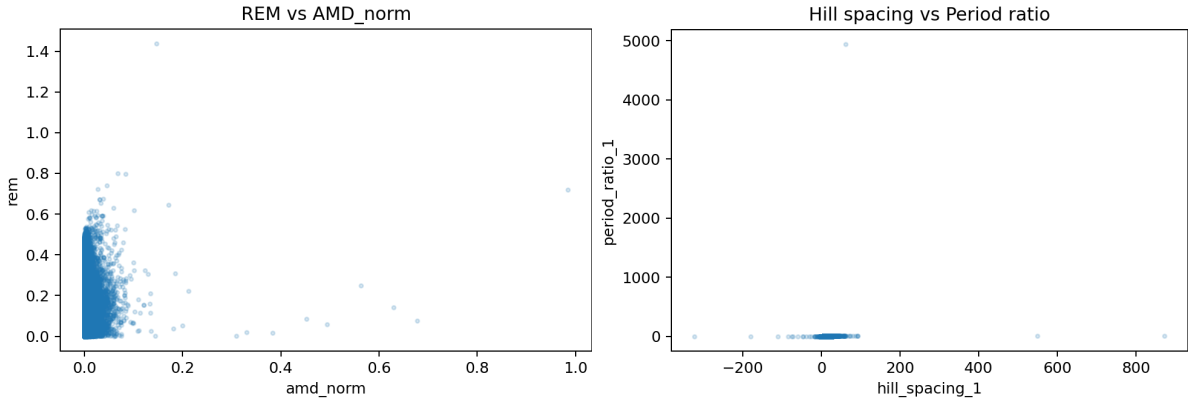


Figure 9: Left: `amd_norm` vs. `rem`. Right: Hill spacing vs. period ratio.

Correlations

The core correlation view shows coherent blocks for kinetic/energetic features and moderate off-diagonal structure with chaos/geometry.

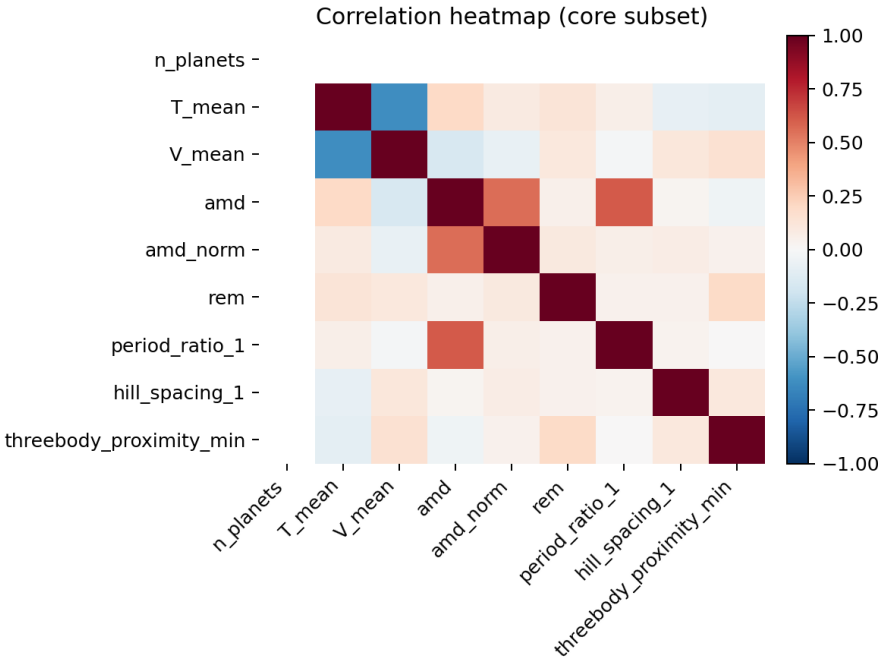


Figure 10: Correlation heatmap (core subset).