Skepthical review: *The Limited Predictability of Asteroid Spin Obliquity from Age, Size, Type, and Family: A Gaussian Process Regression Study*

Summary

The manuscript asks whether asteroid spin-axis obliquity (modeled as **cos**(obliquity)) can be predicted from a limited set of bulk properties (age, diameter, spectral type, dynamical family), and whether large prediction residuals can identify anomalous objects. The authors compile a large merged catalog from multiple CSV sources ($\sim \mathbf{1.7} \times \mathbf{10^6}$ entries), apply completeness filtering to obtain a much smaller modeling sample, one-hot encode categorical features and standardize continuous inputs (Sec. 2.1–2.2), then train a Gaussian Process Regressor (RBF + WhiteKernel) on an 80/20 split to predict **cos**(obliquity) and use the GP predictive uncertainty to compute standardized-residual anomaly scores (Sec. 2.3–2.6).

As currently reported, predictive performance is effectively null (test $\boldsymbol{R^2}$ slightly negative; MSE comparable to the variance of the target), the optimized kernel attributes nearly all variance to the white-noise term (Sec. 3.2), and no objects exceed the $\boldsymbol{\pm 3\sigma}$ anomaly threshold (Sec. 3.3). The paper is valuable as an honestly reported negative result and a cautionary finding about predictability from coarse bulk features, but several internal inconsistencies (sample size and target definition), incomplete model specification (kernel/hyperparameters; categorical handling), and missing baselines/calibration checks mean it is not yet clear whether the null result is intrinsic to the feature set or partly driven by avoidable modeling/implementation choices (mixed categorical/continuous inputs, bimodal target under a Gaussian likelihood, inflated predictive variances affecting anomaly scoring).

Strengths

- Clear scientific question connecting predictability of spin obliquity to physical drivers and to an anomaly-detection use case (Introduction; Sec. 3.4).
- Appropriate choice of a probabilistic model (GPR) that naturally yields predictive uncertainties relevant to anomaly detection (Sec. 2.3–2.5).
- Transparent reporting and interpretation of a null result, with discussion of plausible physical reasons (YORP, collisions) and limitations (Sec. 3.2–3.5; Conclusion).
- Useful exploratory characterization of the target distribution and class imbalance/bimodality (Sec. 3.1; Figures 2–3).
- Several figures (notably residual/anomaly-related plots) are conceptually well chosen for communicating model behavior and the absence of flagged outliers (Figures 7–8).

Major issues

1. **Contradictory sample sizes across Methods/Results and downstream analyses undermine interpretability and reproducibility (Sec. 2.1, Sec. 2.5, Sec. 3.1–3.3; Tables 1–2; figure captions).** The manuscript alternates between $\boldsymbol{N = 5,!890}$ and $\boldsymbol{N = 1,!626}$ as the "complete/filtered/modeling" dataset, and Sec. 2.5 describes anomaly detection on $\mathbf{5,!890}$ while Sec. 3.3 and figures appear to use $\mathbf{1,!626}$. This also affects the implied train/test sizes and the computational narrative in Sec. 2.6.

   Recommendation: Provide a single authoritative data-flow description with $\boldsymbol{N}$s at each stage: (1) initial merged rows and unique asteroids; (2) after each join/deduplication rule; (3) after requiring each feature (obliquity target, age, diameter, type, family); (4) after any additional cuts/grouping that lead to the final modeling table. Include a compact flow table/diagram and ensure Sec. 2.1, Sec. 2.5, Sec. 3.1–3.3, Tables 1–2, and all figure captions use consistent $\boldsymbol{N}$s and clearly state which $\boldsymbol{N}$ is used for training/testing and which $\boldsymbol{N}$ is used for "full-sample" prediction/anomaly scoring.

2. **Target definition/handling is internally inconsistent (Sec. 2.1–2.2 vs. Sec. 3.1).** Methods describe converting obliquity angles (degrees) to **cos**(**radians**(obliquity)), while Results state the ingested "obliquity" column already lies in $[-1, 1]$, implying it is already **cos**(obliquity). Without resolving this, readers cannot know whether the cosine transform was applied, applied twice, or merely described.

   Recommendation: Unify target terminology and preprocessing: explicitly name the raw column(s) and units (e.g., obliquity $\boldsymbol{\cos(\textit{obliquity})}$" (dimensionless) are never conflated.

3. **Model specification is incomplete and partially corrupted in presentation, limiting reproducibility and making some interpretations unreliable (Sec. 2.3–2.4; Sec. 3.1–3.2).** Table 2 appears corrupted (kernel string replacing numeric Max values), the final kernel is not written cleanly in the text, and key training settings are missing (hyperparameter bounds, restarts, normalize_y, alpha, optimizer, random_state). The amplitude/noise interpretation also contains an algebraic inconsistency (Sec. 3.2: "0.1982 ≈ 0.039").

   Recommendation: Fix Table 2 to contain only data summary statistics and move kernel details to Sec. 3.2. In Sec. 2.3–2.4 and Sec. 3.2, report the exact final kernel as printed by the implementation (e.g., ConstantKernel × RBF + WhiteKernel with numeric values), and list training settings: bounds, n_restarts_optimizer, optimizer, alpha, normalize_y, convergence/termination, and random_state. Correct the amplitude statement by explicitly defining whether the printed scalar is an amplitude or variance parameter (and avoid squaring unless the kernel form warrants it).

4. **Kernel choice/feature geometry may be mismatched to the mixed continuous + high-dimensional one-hot categorical feature space (Sec. 2.2–2.3; Sec. 3.2).** If a single Euclidean-distance RBF is applied to a concatenated vector containing sparse one-hot family/type indicators, distances can become dominated by categorical mismatches, effectively making most points "far," inflating uncertainty and encouraging the optimizer to push variance into the WhiteKernel. As written, it is unclear whether the null result reflects astrophysical irreducibility or this representational mismatch.

   Recommendation: Clarify exactly what feature matrix is fed to the GP (continuous + one-hot together, and whether any scaling is applied to one-hot columns). Add at least one robustness experiment better aligned with mixed data types, e.g.: (a) an additive kernel with separate blocks (RBF on standardized [age, diameter] + linear/DotProduct on one-hot or a dedicated categorical kernel); and/or (b) ARD length scales (per-dimension) rather than a single length scale. Report whether conclusions (R^2, residual structure, noise dominance) change; if they do not, the null result is much more convincing.

5. **Missing baselines and robustness checks make the "no predictability" conclusion under-supported (Sec. 3.2–3.4).** Only a single GPR configuration and a single 80/20 split are reported. With $N \approx 1{,}!626$ (or $5{,}!890$; currently unclear), performance estimates can vary, and it is essential to show the GP is not underperforming trivial alternatives.

   Recommendation: Add baseline models in Sec. 3.2–3.4: (1) constant-mean predictor; (2) linear/ridge regression on the same preprocessing; and optionally (3) a nonparametric baseline (random forest or gradient boosting). Report R^2/MSE/MAE side-by-side. Replace or supplement the single split with repeated K-fold cross-validation (or repeated train/test splits) and provide mean ± std of metrics. This will clarify whether the null result is model-independent for the chosen feature set.

6. **Bimodal target distribution and Gaussian-likelihood regression: the modeling framework may be misspecified for the main structure in the data (Sec. 3.1; Sec. 3.5).** The target appears strongly bimodal near ± 1, suggesting a latent "prograde vs retrograde" (or multi-regime) structure. A Gaussian-likelihood regressor will often collapse toward intermediate means (≈ 0) with large uncertainties, which can resemble the observed behavior and should be treated as a central modeling concern rather than only a future direction.

   Recommendation: Quantify bimodality briefly (e.g., peak locations/heights or a simple 2-component mixture fit) and explicitly connect it to expected behavior of Gaussian-likelihood regression. Add one complementary experiment: classify $\mathbf{sign}(\mathbf{cos}(\text{obliquity}))$ (or a 3-class version: near $+1$ / near -1 / intermediate) and report accuracy/AUC relative to a majority-class baseline. Even if predictability remains poor, this reframes the null result in a way that matches the target's structure.

7. **Anomaly detection is tightly coupled to (possibly inflated/miscalibrated) predictive uncertainty; "zero anomalies" may be methodological rather than astrophysical (Sec. 2.5; Sec. 3.3).** Using $z = (y - \mu)/\sigma$ with σ taken from GaussianProcessRegressor(return_std=True) will mechanically suppress $|z|$ when the model attributes most variance to noise. The paper does not specify whether σ includes observation noise vs latent function uncertainty, nor does it assess calibration (coverage / standardized residual distribution).

   Recommendation: In Sec. 2.5, define σ precisely (predictive std including WhiteKernel noise vs latent mean uncertainty) and state the exact code path used. Add calibration diagnostics in Sec. 3.3: (1) histogram of standardized residuals vs $N(0,1)$; (2) empirical coverage of nominal 68\%/95\% predictive intervals. Report sensitivity to threshold choice (2σ, 2.5σ, 3σ) and also consider an outlier criterion less sensitive to variance inflation (e.g., leave-one-out negative log predictive density / low predictive probability). Reframe the "no anomalies" conclusion as conditional on calibration and the chosen anomaly criterion.

8. **Data provenance and physical meaning of key inputs—especially "age"—are insufficiently defined, limiting scientific interpretability (Sec. 2.1–2.2; Sec. 3.5).** It is unclear whether age is a family-level attribute (shared by many objects) or an object-specific estimate, what method produced it, typical uncertainties, and how conflicting entries across sources were resolved. Similar concerns apply to obliquity derivation consistency across catalogs and to diameter uncertainties.

   Recommendation: Add a concise provenance subsection (Sec. 2.1 or Appendix): for each feature ($\mathbf{cos}$(obliquity), age, diameter, type, family), list the source catalog(s), unit conventions, typical uncertainties (if known), and reconciliation rules for duplicates/conflicts. Explicitly state whether age varies within families or is assigned per family, and discuss the implications (effective degrees of freedom; potential leakage/collinearity with family).

Minor issues

1. Residual analysis is discussed qualitatively without simple quantitative diagnostics, and the residual "structure" may largely reflect mean-collapse toward 0 under bimodality (Sec. 3.2; Figures 6–7).

   Recommendation: Report basic residual statistics (mean, std, skew), correlation of residuals with age/diameter, and explicitly comment on whether predictions concentrate near 0 (mean predictor-like behavior). If feasible, include a binned residual-vs-feature plot and/or a histogram of predicted means $\mu(x)$.

2. Family/type grouping and class imbalance are not fully quantified and appear inconsistent across text/figures (Sec. 2.2; Sec. 3.1; Figure 4). The manuscript references 24 vs 19 families and "top 30" in places; the "Other" threshold (< 10) is stated but not enumerated.

   Recommendation: List (in text or Appendix) the retained family/type categories after grouping, the number merged into "Other," and counts per category. Ensure Figure 4 caption and Sec. 3.1 use consistent numbers and specify whether distributions are computed on the full modeling dataset or a split.

3. Selection effects and representativeness are only briefly acknowledged despite the modeling set being a tiny fraction of the initial merged catalog (Sec. 3.1; Sec. 3.5).

   Recommendation: Add a short quantitative comparison between the final modeling subset and the broader catalog on at least one readily available property (e.g., diameter distribution, H magnitude if present, or orbital element ranges) to illustrate bias. Clearly scope conclusions to the well-characterized subset.

4. Figure clarity and metadata: Figure 1 and Figure 4 are hard to read and have ambiguous labels/denominators; several figures omit the N shown and do not always clarify whether "obliquity" means angle or cosine (Figures 1, 4, 7, 8).

   Recommendation: Increase export resolution / use vector formats; enlarge fonts; make denominators explicit (N_{total} for missingness plots); label the target consistently as $\cos(\text{obliquity})$. Add N and key metrics (R^2/RMSE) in captions or panels for Figures 7–8.

5. Train/test split and preprocessing details are incomplete (Sec. 2.2–2.4). It is not explicit that scaling/encoding are fit on the training set only and then applied to test/full data; random_state/shuffle/stratification are not stated.

   Recommendation: State the exact split settings (random_state, shuffle, stratification if any). Confirm scaler and encoder are fit on training data only and reused for test/full prediction, and report the final feature dimensionality after one-hot encoding.

6. The 3σ threshold justification is generic and not tied to dataset size or desired false positive rate (Sec. 2.5; Sec. 3.3).

   Recommendation: Briefly justify 3σ relative to N (expected false positives under calibration) and/or scientific cost of false positives. Include sensitivity counts for alternative thresholds.

7. Code/data availability is not stated, which is particularly important given the multi-source compilation and filtering (Sec. 2; Conclusion).

   Recommendation: Add a Data/Code Availability statement. If raw catalogs cannot be redistributed, provide derived artifacts that enable reproduction (final list of modeled asteroids with features, preprocessing scripts, and exact environment versions).

Very minor issues

1. Typographical/formatting issues and inconsistent notation (e.g., $\cos(\text{Obliquity})$ vs $\cos(\text{obliquity})$; "Predicence" typo; broken words at line breaks; inconsistent capitalization; stray punctuation in headings) reduce polish (various sections, notably Sec. 3.1–3.3).

   Recommendation: Proofread and standardize naming across text, tables, and figures; correct typos ("Predicted"), remove stray underscores, fix broken hyphenation/line-break artifacts, and ensure section headings are consistently formatted.

2. Figure referencing/caption ordering and axis-scale notes are occasionally unclear (e.g., figures introduced out of order; log-scaling not always stated).

   Recommendation: Ensure figures are referenced in numerical order, captions state any log scales/binning, and axis labels match the terminology used in Sec. 2–3.

Key statements and references

- • **The bimodal distribution of $\cos(\text{obliquity})$ in the modeling dataset, with peaks near 1 and -1 corresponding to prograde and retrograde rotation, is consistent with theoretical predictions and observational evidence that the Yarkovsky–O'Keefe–Radzievskii–Paddack (YORP) effect drives asteroid spin axes toward these extreme states over sufficiently long timescales.**
- *Reference(s):* Yarkovsky-O'Keefe-Radzievskii-Paddack (YORP) effect
- • **The optimized Gaussian Process kernel shows that the RBF signal variance is very small ($0.198^2 \approx 0.039$) while the WhiteKernel noise level converges to 0.696, and given that the total variance of $\cos(\text{obliquity})$ in the test set is ≈ 0.706, the model attributes about 98.6% ($0.696/0.706$) of the variance to irreducible noise or factors not captured by age, diameter, spectral type, and family.**
- *Reference(s):* Gaussian Process Regression (GPR)
- • **The Gaussian Process Regression model's performance on the held-out test set ($N = 326$) yields an R-squared of -0.0069, a Mean Squared Error of 0.7105, and a Mean Absolute Error of 0.8040, quantitatively demonstrating that predictions of $\cos(\text{obliquity})$ from age, diameter, spectral type, and dynamical family are essentially no better (and slightly worse) than predicting the mean.**
- *Reference(s):* Gaussian Process Regression (GPR)
- • **Using the standardized residual anomaly score with a $\pm 3\sigma$ threshold and the GPR-predicted uncertainties, zero asteroids in the $1,626$-object dataset exceed the 3-sigma criterion, because the high optimized noise level (0.696) leads to large $\sigma_{\text{predicted}}$ values that normalize even large residuals below the anomaly threshold.**
- *Reference(s):* Gaussian Process Regression (GPR)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains limited explicit mathematics: (i) a cosine transform of obliquity for regression, (ii) a standardized-residual anomaly score, and (iii) a stated optimized Gaussian-process kernel and qualitative variance/noise interpretation. No formal GP derivations are provided, so the audit focuses on definition consistency, algebraic coherence of the kernel interpretation, and the internal consistency of symbols and transformations.

Checked items

1. ⚠ **Cosine target transform definition** (Sec. 2.2, p.3)
    - **Claim:** Obliquity angle in $[0, 180]$ degrees is transformed to $\cos(\text{radians}(\text{obliquity}))$ to create a continuous regression target in $[-1, 1]$.
    - **Checks:** definition consistency, range/sanity check
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** Obliquity is an angle measured in degrees in the dataset at this stage., The cosine is applied exactly once to obtain the target.
    - **Notes:** The transform itself is mathematically sound and yields a target in $[-1, 1]$. However, Results (Sec. 3.1, p.4–5) later states the provided 'obliquity' column already lies in $[-1, 1]$ and represents $\cos(\text{obliquity})$, contradicting the assumption that obliquity is in degrees here. Without a clear statement of raw column units and preprocessing, it is unclear whether the cosine transform was applied appropriately or redundantly.
2. ✔ **Standardized residual / anomaly score formula** (Sec. 1–2 intro text on p.2; Sec. 2.5, p.4; Sec. 3.3, p.7)
    - **Claim:** Anomaly score is (observed $\cos(\text{obliquity})$ − predicted $\cos(\text{obliquity})$) $/\sigma_{\text{Predicted}}$.
    - **Checks:** algebra, dimensional/units, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** $\sigma_{\text{Predicted}}$ is a standard deviation associated with the prediction for that asteroid., Observed and predicted quantities are on the same scale (both cosines).
    - **Notes:** As written, the score is dimensionless and consistent with a z-score. Internal consistency requires observed and predicted values to both be $\cos(\text{obliquity})$, which the paper intends, but see the separate inconsistency about whether the dataset column is already cosine-transformed.
3. ✔ **3-sigma threshold interpretation** (Sec. 2.5, p.4)
    - **Claim:** Using $|\text{score}| > 3$ corresponds to a very low probability (approx. 0.3%) under a Gaussian distribution centered at the prediction with the predicted variance.
    - **Checks:** conceptual/probabilistic consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Anomaly score is approximately standard normal under the null (Gaussian predictive distribution)., Threshold is two-sided (± 3).
    - **Notes:** The analytic mapping between a $\pm 3\sigma$ rule and an 'about 0.3%' two-sided tail probability is reasonable. The paper could specify two-sided explicitly, but the statement is not internally inconsistent.
4. ✔ **Optimized kernel expression (form)** (Sec. 3.2, p.5)
    - **Claim:** The optimized kernel is $\text{Kernel} = 0.1982 \times \text{RBF}(\text{length}_\text{s}\text{cale} = 0.902) + \text{WhiteKernel}(\text{noise}_\text{l}\text{evel} = 0.696)$.
    - **Checks:** notation consistency, structural correctness
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** The kernel is a sum of a scaled RBF component and a white-noise component as written.
    - **Notes:** The kernel form (signal + noise) is syntactically consistent with the described composite kernel. The audit does not verify software-specific parameter semantics; it checks only internal symbolic coherence.
5. ✖ **Kernel amplitude narrative vs kernel expression** (Sec. 3.2, p.6)
    - **Claim:** The RBF amplitude is very small: ($0.1982 \approx 0.039$).
    - **Checks:** algebra/numerical identity used in narrative
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** The amplitude referenced corresponds to the 0.1982 factor shown multiplying the RBF.
    - **Notes:** The stated approximation '$0.1982 \approx 0.039$' is algebraically incorrect; 0.039 is approximately the square of 0.1982, not approximately equal to it. If the intent was to refer to a variance term (square), the kernel expression or explanation must be adjusted to match.
6. ⚠ **Noise fraction / variance attribution logic** (Sec. 3.2, p.6)
    - **Claim:** Because WhiteKernel noise_level is high relative to total variance, the model attributes nearly all variance to noise rather than signal.

- **Checks:** definition consistency, conceptual consistency
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** noise_level is on the same variance scale as the target variance being compared., The comparison being made is conceptually meaningful for 'variance explained' interpretation.
- **Notes:** The qualitative conclusion 'noise dominates' is plausible given the displayed parameters, but the paper does not provide the underlying GP predictive variance decomposition needed to justify a precise 'percentage of variance' attribution from kernel hyperparameters alone. A brief analytic statement relating kernel components to marginal variance (and what is meant by 'total variance') is missing.

7. ✖ **Dataset size used in Methods vs Results** (Sec. 2.1–2.5, p.2–4 vs Sec. 3.1, p.4–5)
    - **Claim:** The complete-data modeling dataset has $N = 5,!890$ (Methods) and later $N = 1,!626$ (Results).
    - **Checks:** definition consistency, cross-section consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Both refer to the same modeling dataset used for training/testing/anomaly detection.
    - **Notes:** This is a direct internal contradiction. It also impacts stated train/test sizes (e.g., 80/20 split counts) and downstream analytic claims linked to dataset characterization.
8. ✖ **Target variable description vs Table 1 obliquity units** (Table 1, p.2 vs Sec. 3.1, p.4–5)
    - **Claim:** Table 1 summarizes 'Obliquity (deg)' while later the paper describes 'obliquity' as already being $\cos(\text{obliquity})$ in $[-1, 1]$.
    - **Checks:** units consistency, notation consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** The same column name 'obliquity' is being referenced throughout.
    - **Notes:** The unit-bearing summary in Table 1 conflicts with the later assertion that the column already stores cosine values. This must be reconciled by distinguishing raw vs transformed columns and ensuring the tables correspond to the stated modeling dataset.

Limitations

- The paper provides no step-by-step mathematical derivations of Gaussian Process Regression (posterior mean/variance, marginal likelihood) or explicit formulas linking kernel hyperparameters to explained variance; this limits verification to consistency checks of definitions and stated expressions.
- Some claims mix narrative numeric comparisons (e.g., variance attribution percentages) with kernel parameters; without explicit analytic definitions of those quantities within the paper, these checks remain partially uncertain even when the qualitative direction is clear.
- The audit uses only the content present in the provided PDF text/images; any missing appendices, code, or supplementary material are not considered.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Out of 12 numerical/consistency checks, 11 passed and 1 failed. The only failure is a cross-section inconsistency in the reported modeling dataset size ($5,!890$ in Methods vs $1,!626$ in Results/Table 2). Other internal arithmetic/ratio statements (e.g., $0.1982^2 \approx 0.039$; $0.696/0.706 \approx 98.6\%$; MSE 0.7105 close to variance ~ 0.706; 80/20 split counts; percentage-of-catalog claims) are numerically consistent within stated/heuristic tolerances.

Checked items

1. ✔ **C1** (Page 1 (Abstract) vs Page 4–5 (Results §3.1))
    - **Claim:** Abstract states the dataset used had $1,!626$ asteroids with complete data; Results §3.1 states the final modeling dataset comprises $1,!626$ asteroids.
    - **Checks:** repeated_constant_match
    - **Verdict:** PASS
    - **Notes:** Exact match check. Exact match expected.
2. ✖ **C2** (Page 2 (Methods §2.1) vs Page 4–5 (Results §3.1 & Table 2))
    - **Claim:** Methods §2.1 reports a filtered modeling dataset size of $5,!890$ asteroids with complete data, but Results §3.1 reports final modeling dataset of $1,!626$ asteroids (and Table 2 uses $N = 1,!626$).
    - **Checks:** cross_section_inconsistency_flag
    - **Verdict:** FAIL
    - **Notes:** Cross-section consistency: values should match unless explained. This is a logical consistency check; values differ materially.
3. ✔ **C3** (Page 6 (Results §3.2, kernel discussion))

- **Claim:** Paper states: 'The amplitude of the RBF component ... is very small ($0.1982 \approx 0.039$).'.
- **Checks:** numeric_recomputation
- **Verdict:** PASS
- **Notes:** Computed square compared to claimed approximation. Given approximate symbol, allow small rounding differences.

4. ✔ **C4** (Page 6 (Results §3.2, variance attribution))
   - **Claim:** Paper states: total variance in $\cos(\text{obliquity})$ test set ≈ 0.706; noise level $= 0.696$; fraction $0.696/0.706 \approx 98.6\%$.
   - **Checks:** ratio_and_percentage
   - **Verdict:** PASS
   - **Notes:** Computed percentage compared to claimed value. Text uses ≈; allow rounding.
5. ✔ **C5** (Page 6 (Table 3 and surrounding text))
   - **Claim:** Table 3 reports test-set MSE $= 0.7105$ and text states MSE is very close to the variance of the test data itself (~ 0.706).
   - **Checks:** difference_close_to_claim
   - **Verdict:** PASS
   - **Notes:** Heuristic closeness check using abs_tol as threshold. Heuristic tolerance for 'very close' in this context.
6. ✔ **C6** (Page 5–6 (Results §3.2) vs Page 6 (Table 3))
   - **Claim:** Results §3.2 states test set size is 326 asteroids; Table 3 header shows ($N = 326$).
   - **Checks:** repeated_constant_match
   - **Verdict:** PASS
   - **Notes:** Exact match check. Exact match expected.
7. ✔ **C7** (Page 5 (Results §3.2))
   - **Claim:** Paper states: model trained on 80% of dataset ($1,!300$ asteroids) and test set is 326 asteroids from total $1,!626$.
   - **Checks:** split_consistency
   - **Verdict:** PASS
   - **Notes:** Checked ($N_{\text{train}} + N_{\text{test}} \approx N_{\text{total}}$) and fractions near 80/20. Integer rounding may cause $+/-1$ discrepancy; ratio near 0.8.
8. ✔ **C8** (Page 4 (Results §3.1))
   - **Claim:** Paper states: final modeling dataset of $1,!626$ asteroids is less than 0.1% of the initial catalog, which exceeded 1.7 million entries.
   - **Checks:** percentage_upper_bound
   - **Verdict:** PASS
   - **Notes:** Inequality check pct $<$ claimed upper percent; uses conservative minimum initial N. Inequality check; using 1.7M makes the percent largest, so if it still $< 0.1\%$ the claim holds.
9. ✔ **C9** (Page 4 (Results §3.1, missingness claim))
   - **Claim:** Paper states: 'A substantial majority, over 80% of the catalog entries were missing one or more of these crucial data points.' With final complete $N = 1,!626$ and initial catalog > 1.7M.
   - **Checks:** implied_missingness_lower_bound
   - **Verdict:** PASS
   - **Notes:** Lower-bound check: missing fraction based on complete count and minimum initial size. Using 1.7M as minimum initial size yields the smallest missing fraction bound; if still $> 80\%$, claim supported.
10. ✔ **C10** (Page 5 (Results §3.1, family grouping counts))
    - **Claim:** Paper states: 24 distinct dynamical families; after grouping families with fewer than 10 members into “Other”, number of effective family classes becomes 19.
    - **Checks:** arithmetic_difference
    - **Verdict:** PASS
    - **Notes:** Computed grouped = distinct − effective; expected 5. Simple arithmetic identity.
11. ✔ **C11** (Page 4 (Methods §2.6))
    - **Claim:** Prediction task was divided into 128 chunks for parallel processing.
    - **Checks:** integer_parameter_sanity_check
    - **Verdict:** PASS
    - **Notes:** Sanity check for chunking parameter (not a proof of correctness). Sanity check to ensure chunking isn't nonsensical; not a proof of correct implementation.
12. ✔ **C12** (Page 4 (Methods §2.5) and Page 7–8 (Results §3.3))

- **Claim:** Asteroids flagged as anomalies if $|\mathbf{score}| > \mathbf{3.0}$; Results state zero anomalous asteroids were identified based on **3**-sigma criterion.
- **Checks:** threshold_definition_consistency
- **Verdict:** PASS
- **Notes:** Cross-reference check on extracted threshold and reported anomaly count. Exact match expected for threshold and 'zero' count.

Limitations

- Only parsed text provided from the PDF was used; no underlying CSV data, code, or supplementary materials were available to recompute dataset-derived statistics or metrics.
- Numerical values embedded only in figures (bar heights, histogram bin counts) were not extracted, per instruction to avoid plot-pixel/image value extraction.
- Some checks are limited to logical/consistency validations (e.g., repeated N, thresholds) rather than full recomputation, because intermediate data needed for recomputation is not present in the PDF.