Skepthical review: *Identifying Anomalous Asteroids via Predictive Modeling of Physical and Spin Properties based on Orbit and Age*

Summary

The manuscript proposes a supervised, residual-based anomaly detection pipeline to identify asteroids whose physical/spin properties (diameter, spin period, obliquity) deviate from values predicted from orbital elements (a, e, i), estimated age, and taxonomic type (Sec. 1–2.4). The authors merge $\sim$ **12** catalogs into a master table ($>$ **1.7**M rows; Sec. 2.1, 3.1), create three feature-complete task datasets (diameter/spin/obliquity), train Gaussian Process Regression (GPR) and neural network (MLP) regressors (Sec. 2.3, 3.2), and define anomaly scores via standardized residuals (GPR: $(y-\mu)/\sigma$; NN: global z-scored residuals; Sec. 2.4, 3.3.1). Using a fixed $|S| > \mathbf{3}$ threshold across model–property pairs, the paper reports **1,138** unique anomalous asteroids, dominated by diameter-driven flags, and interprets them as predominantly large objects on dynamically "quiet" main-belt-like orbits with extreme spin periods, potentially connected to primordial survivors and/or unusual collisional/YORP evolution (Sec. 3.3.2–3.4, Conclusions).

The overall idea—defining "expected" properties conditional on orbit/age and highlighting outliers—is promising and potentially useful. However, several core methodological gaps currently make it difficult to distinguish astrophysically meaningful anomalies from (i) selection effects due to extreme sparsity and heterogeneous measurement quality, (ii) model inadequacy/overfitting (especially for spin period and obliquity), and (iii) inconsistent population definitions across targets when summarizing the final anomalous set (e.g., Table 1 denominators). Strengthening out-of-sample scoring, quantitative model evaluation, uncertainty calibration, and bias/eligibility accounting would substantially improve the reliability and interpretability of the anomaly catalog and the physical claims.

Strengths

- Well-motivated framing: using conditional predictive models (orbit/age $\rightarrow$ properties) to define outliers is conceptually clear and relevant for asteroid population studies (Introduction, Sec. 1; Methods Sec. 2.4).
- Ambitious data consolidation effort: merging $\sim$ **12** sources into a large master catalog with a described cleaning/feature pipeline (Sec. 2.1–2.2, 3.1) is valuable and could be reusable.
- Using two complementary model families (GPR and NN/MLP) is a sensible design choice; the GPR formulation naturally exposes predictive uncertainty that can be used for anomaly scoring (Sec. 2.3.1, 2.4; Sec. 3.2.1).

- Feature preprocessing is mostly sensible and clearly sequenced (log transforms, standardization, one-hot encoding of type; Sec. 2.2), and the manuscript reports GPR kernel hyperparameters and qualitative diagnostics (Sec. 3.2.1).
- The paper provides an initial characterization of the flagged set (Table 1; Fig. 21; Sec. 3.3.2) and connects observations to plausible physical mechanisms (collisions/YORP; Sec. 3.4), generating testable hypotheses.
- The anomaly-score definitions are explicit and mostly consistent across Methods and Results (Sec. 2.4; Sec. 3.3.1), which helps readability.

Major issues

1. **Ambiguous and potentially inconsistent definition of the scored population ("non-anomalous" vs "not eligible") and mixing across target-specific datasets undermines Table 1 and the interpretation of "1,138 unique anomalous asteroids" (Sec. 3.1, 3.3.1–3.3.2; Table 1).** Each target uses a different feature-complete subset (e.g., $\sim 10,340$ diameter vs $\sim 6,396$ spin vs $\sim 1,626$ obliquity), yet Results compare anomalies to very large "non-anomalous" counts that appear to include objects that were never scored/eligible. This makes differences in size/orbit potentially reflect eligibility/measurement availability rather than anomaly status.

   Recommendation: In Sec. 3.1 and Sec. 3.3.1–3.3.2, define eligibility and denominators unambiguously for each target/model: (i) N_{eligible} (feature-complete, scored), (ii) N_{flagged} within that eligible set, and (iii) anomaly rate per target/model. When summarizing properties (Table 1, Fig. 21), compare flagged objects to an appropriate control group drawn from the same eligible dataset ("scored and not flagged"), not to the full master catalog. If you also want a unified cross-target list (the 1,138 unique objects), include a flow table: master → eligible per task → flagged per task → union, and clearly state which statistics use which subset.

2. **Anomaly scoring appears to use in-sample predictions (training+test combined) and global residual statistics, risking biased residual distributions and distorted anomaly counts—especially for flexible NNs and the overfit-looking obliquity GP (Sec. 2.3–2.4, Sec. 3.2, 3.3.1).** Computing z-scores from residuals that include training points can artificially shrink residual variance and change which points exceed $|S| > 3$.

   Recommendation: Revise Sec. 2.4/3.3.1 so anomaly scores are computed from strictly out-of-sample predictions: e.g., k-fold cross-validation with out-of-fold predictions for every eligible object, or a held-out test-only analysis (with the caveat that it reduces coverage). For NN scoring, compute mean/std of residuals from out-of-fold residuals only. Report (and optionally plot) training vs test residual distributions to show the magnitude of the bias avoided.

3. **Lack of quantitative predictive-skill evaluation and uncertainty calibration prevents assessing whether residuals reflect conditional outliers or simply model inadequacy—most acute for spin period and obliquity (Sec. 3.2.1–3.2.2).** The manuscript relies heavily on qualitative plots and kernel parameters, while acknowledging weak/mean-regressing behavior for spin and pathological/overfit behavior for obliquity; in such regimes, "anomalies" can reduce to extremes of the marginal distribution (spin) or be suppressed by misestimated uncertainty (obliquity).

   Recommendation: Augment Sec. 3.2 with standard out-of-sample metrics for each model–target pair (RMSE/MAE in physical units and/or log-units, R^2, correlation). Include simple baselines (e.g., ridge/linear regression, random forest) to contextualize whether GPR/MLP add value. For GPR, add calibration/coverage checks: fraction of truths within nominal $1\sigma/2\sigma$ intervals. Use these results in Sec. 3.3–3.4 to gate interpretation: if a target's model has near-zero conditional skill, treat its "anomalies" as low-confidence and separate them from the main physical conclusions.

4. **The fixed $|S| > 3$ threshold is not statistically justified given (i) large sample sizes, (ii) heavy-tailed/heteroscedastic residuals, and (iii) multiple testing across 3 properties $\times$ 2 model classes (Sec. 2.4, 3.3.1).** For NN scores, global z-scoring ignores heteroscedasticity (visible in residual structure for diameter), and outliers can inflate the residual std, changing the effective threshold.

   Recommendation: In Sec. 3.3.1, empirically characterize standardized-residual distributions per model/target (histograms + QQ plots; tail behavior; heteroscedasticity vs predicted value). Provide a sensitivity analysis of anomaly counts and key Table 1 statistics versus threshold (e.g., 2.5/3/3.5/4). Consider a multiple-testing-aware framing (expected false positives under a null; FDR control), or at minimum report expected vs observed $> 3\sigma$ exceedances under a standard-normal assumption. For NNs, consider heteroscedastic-aware alternatives: bin-wise residual scaling, a second-stage variance model, quantile regression, or conformal prediction intervals, so "$|S| > 3$" has clearer meaning.

5. **Selection effects, missing-not-at-random measurement processes, and heterogeneous uncertainties across catalogs are likely to dominate both training and the anomalous set, but are not analyzed systematically (Sec. 2.1–2.2, 3.1, 3.3.2, 3.4).** The usable training sizes ($\sim$ 10k diameter; $\sim$ 6k spin; $\sim$ 1.6k obliquity) are tiny relative to the 1.7M master list; measured spin/obliquity and even taxonomic type are strongly biased toward larger/brighter/better-observed objects. The finding that anomalies are "much larger" may therefore largely reflect measurement availability and/or model extrapolation rather than a distinct physical population.

   Recommendation: Add a dedicated subsection (Sec. 3.1 or new Sec. 3.1.1) quantifying completeness and selection: for each property (diameter/spin/obliquity/age/type), report fraction available in the master set; show distributions of (a, e, i), H/magnitude

proxy if available, and diameter for (i) master, (ii) eligible per task, and (iii) flagged anomalies. If possible, summarize typical measurement uncertainties and any quality flags per source, and test whether anomalies are overrepresented in specific catalogs/surveys or low-quality subsets (Sec. 3.3.2). Temper Sec. 3.4/Conclusions to explicitly state that anomalies are defined within biased, inhomogeneous measured subsets.

6. **Obliquity modeling/definition is internally unclear and appears numerically/pathologically fit (Sec. 2.2–2.4; Figures 9 and 19; Sec. 3.2.1, 3.3.1).** Plots suggest the obliquity target spans $\sim [-1, 1]$, inconsistent with an angle in degrees/radians, and the optimized GP kernel shows extremely small length scale and near-zero noise ($1e{-}10$), suggestive of overfitting and/or optimizer boundary behavior. This makes both predictions and anomaly scoring for obliquity unreliable.

   Recommendation: First, define the obliquity target precisely in Sec. 2.2 (units, range; if it is $\cos(\epsilon)$ or a normalized quantity, rename it accordingly and update interpretation and formulas). Second, report GP hyperparameter bounds, optimizer restarts, and whether parameters hit bounds (Sec. 2.3.1, 3.2.1). Add regularization: enforce a noise floor, adjust bounds, consider Matern kernels, and evaluate via cross-validation. Given the tiny sample ($1,626$; Sec. 3.1), consider presenting obliquity as exploratory/methodological only, or clearly flag obliquity anomalies as highly tentative (Sec. 3.3–3.4).

7. **Physical interpretation currently outpaces the demonstrated inference: claims that the anomalous set is a distinct/primordial population are plausible but not uniquely supported, and alternative explanations (model extrapolation at large diameters; known large bodies/family parents; catalog systematics) are not ruled out (Abstract; Sec. 3.3.2–3.4; Conclusions).** In particular, predicting diameter from orbit/age has limited causal grounding; "diameter anomalies" may reflect dataset composition and extrapolation failures rather than unusual physics.

   Recommendation: Reframe conclusions to clearly separate empirical statements (conditional model residual outliers; large/low-e/i among flagged objects) from origin hypotheses (primordial survivors). Add targeted cross-checks in Sec. 3.3.2: (i) stratify or match-control by diameter (and/or brightness proxy) to see whether anomaly status adds information beyond being large; (ii) inspect model behavior at the large-diameter end (are these just underpredicted extremes due to training imbalance?); (iii) cross-match the top anomalies with known large asteroids/dwarf planets/family parent bodies and note whether flags are expected; (iv) provide a small table of exemplar anomalies (IDs, measured vs predicted, score, data sources) and validate a few against literature.

8. **Family membership is excluded as a predictor due to high cardinality (Sec. 2.2), but family context is relevant for both "age" interpretation and for whether anomalies simply trace family-specific trends or parent bodies (Sec. 3.3.2, 3.4).** Without quantitative post hoc analysis, it is unclear whether the flagged set is dominated by a few families/background populations.

   Recommendation: In Sec. 3.3.2/3.4, quantify family membership in anomalies vs matched controls: over/under-representation of major families, fraction of "background," and whether anomalies cluster in specific families or dynamical regions. If feasible, include a coarse family encoding (largest-N families + 'other', or embedding/target encoding) as an ablation in Sec. 3.2 and report the impact on predictive skill and anomaly lists.

Minor issues

1. Catalog provenance and scientific meaning of key fields are under-specified (Sec. 2.1–2.2): "12 CSVs" is not enough for reproducibility, and it is unclear which surveys/compilations provide diameter/spin/obliquity/type and what their typical uncertainties/quality flags are.

   Recommendation: In Sec. 2.1 (and/or Appendix/Supplement), list each source catalog with citation/version/date, field definitions, and known uncertainty/quality indicators used. Summarize how conflicts between sources are resolved (e.g., multiple diameters). Provide a released mapping of asteroid IDs across catalogs if possible.

2. The "age" predictor is central to the narrative but remains vague (Sec. 1, 2.1–2.2, 3.3.2, 3.4): it is unclear whether this is family age, collisional age, or another proxy; coverage and uncertainties are not described, yet age is used to support evolutionary/primordial interpretations.

   Recommendation: Define age precisely in Sec. 2.1/2.2: source, physical meaning, applicability (which objects have ages), and typical uncertainties/systematics. In Sec. 3.4, discuss how age uncertainty propagates into model predictions and anomaly interpretation.

3. Log-transform handling is not fully specified and appears inconsistent with reported zeros (Table 1 reports diameter min = **0.00** km while **log(diameter)** is used; Sec. 2.2, 3.1, Table 1).

   Recommendation: Explicitly state positivity requirements and handling for log-transformed variables (filtering nonpositive values vs using $\log(x + \epsilon)$, and the ϵ used). Clarify whether the **0.00** in Table 1 is rounding, missing-value encoding, or a value outside the modeling subset.

4. Train/test split and NN tuning details are insufficient for reproducibility (Sec. 2.3.2): the text describes possible tools rather than what was used; split random seeds/stratification are not stated; final architectures are not reported.

   Recommendation: In Sec. 2.3/2.3.2, specify the exact split protocol (random_state, number of repeats), whether any stratification was used, the tuning framework and search budget, and the final NN architectures/hyperparameters per target (table preferred).

5. Several plots and tables lack essential quantitative annotations (units, transformations, sample sizes, data split shown), and overplotting obscures patterns (multiple figures in Sec. 3.1–3.3). Some key interpretive claims are made visually without effect sizes or tests.

   Recommendation: Add units/log-base/transform labels and N per panel in captions; indicate whether plots show train/test/out-of-fold predictions. Use hexbin/density for predicted-vs-true plots, and include metrics (R^2/RMSE) on-figure. Where comparing anomalous vs control distributions (Sec. 3.3.2), report effect sizes and simple tests (KS/Mann–Whitney) alongside plots.

6. The NN anomaly score uses a single global residual standard deviation even when residual variance is evidently feature-/scale-dependent (heteroscedastic), which can bias which value ranges are flagged (Sec. 3.2.2, 3.3.1).

   Recommendation: Adopt a heteroscedastic-aware NN scoring approach (bin-wise scaling by predicted value, or a model for $\sigma(x)$, or conformal intervals). Report how the anomaly list changes relative to the current global-z approach (Sec. 3.3.1).

7. Some internal explanatory text around anomaly counts/mechanisms is inconsistent with the stated scoring formula (e.g., discussion of why obliquity anomalies are rare vs the role of σ in S_{GPR}; Sec. 3.3.1).

   Recommendation: Align explanations explicitly with $S_{\mathrm{GPR,i}} = (y - \mu)/\sigma$ and clarify whether rarity of anomalies arises from large σ (uncertainty inflation), small residuals, in-sample prediction effects, or thresholding choices.

Very minor issues

1. Typographical/formatting inconsistencies in headings, variable naming (spin_period vs spin period), and equation typesetting (e.g., “1e - 10”, occasional malformed subscripts like “SGP R,i”) occur throughout Sec. 2–3 and figures/captions.

   Recommendation: Standardize section hierarchy, variable nomenclature, and math typesetting (e.g., $S_{\mathrm{GPR,i}}$, $S_{\mathrm{NN,i}}$; “$1e{-}10$”). Perform a focused proofreading pass for consistency.

2. Obliquity coverage fraction is reported inconsistently relative to the cleaned dataset size (Sec. 3.1 vs stated percentages).

   Recommendation: Reconcile the percentage by clearly stating the denominator (master vs filtered/typed/age-available subset) and whether the figure refers to raw presence or post-cleaning eligibility.

3. Potential numerical instability in GPR anomaly scoring if σ_i becomes extremely small is not addressed (Sec. 2.4, 3.3.1), particularly given the near-zero GP noise reported for obliquity.

   Recommendation: State and implement a safeguard (e.g., $\sigma_i \leftarrow \mathbf{max}(\sigma_i, \epsilon)$) and report that σ_i is bounded away from zero (or provide the observed $\mathbf{min}(\sigma_i)$) for each target.

4. Minor count/denominator discrepancies are mentioned but not explained (e.g., small differences between unique anomalies and per-feature anomaly counts; Sec. 3.3.2/Table 1 context).

   Recommendation: Add brief notes in Table 1 captions and Sec. 3.3.2 explaining that per-feature counts can be smaller due to missingness/eligibility, and ensure all denominators are explicitly stated.

Key statements and references

- **✖ Non-gravitational torques from the YORP effect can substantially modify asteroid spin periods and obliquities over time, but this process is highly sensitive to an asteroid's detailed shape, surface thermal properties, and internal structure, so these spin-state evolutions cannot be reliably inferred from orbital elements and age alone [11].**
- *Reference(s):* [11]
- *Justification:* [11] studies human–chatbot interaction and self-disclosure effects of agent embodiment; it does not discuss asteroids, YORP, non-gravitational torques, or spin-state evolution. Therefore, the statement is not supported by [11].

- **✖ Thermal-radiation–driven orbital evolution via the Yarkovsky effect and spin-state evolution via the YORP effect are key non-gravitational processes that, together with collisions and gravitational perturbations, shape the present-day distribution of asteroid sizes, spins, and obliquities over Solar System timescales [11].**

- *Reference(s):* [11]
- *Justification:* [11] examines human–chatbot interaction and self-disclosure; it contains no discussion of asteroids, Yarkovsky/YORP effects, or Solar System dynamical processes. Therefore the statement is not supported by [11].

- ✖ **Because YORP torques scale inversely with size, they become much less efficient for large asteroids (typically with diameters greater than about 40 km), so the spin states of such bodies are expected to be far less modified by YORP than those of smaller asteroids and may instead preserve primordial or collisionally reset rotation states [11].**
- *Reference(s):* [11]
- *Justification:* [11] investigates human–chatbot interaction and information disclosure; it contains no discussion of asteroids, YORP torques, size-scaling, or spin-state evolution. Therefore the statement is not supported by [11].
- ✖ **Large asteroids on low-inclination, low-eccentricity orbits in the main belt are interpreted in prior work as dynamically 'cold' objects that are less affected by Yarkovsky-driven orbital drift and catastrophic disruption, making them plausible surviving primordial planetesimals from the early Solar System [11].**
- *Reference(s):* [11]
- *Justification:* [11] investigates human–chatbot interaction and self-disclosure; it contains no discussion of asteroids, orbital dynamics, the Yarkovsky effect, catastrophic disruption, or primordial planetesimals. Therefore the statement is not supported by [11].
- ✖ **Previous studies have shown that the combination of large size, dynamically stable main-belt orbits, and atypical spin states can signal a distinct evolutionary pathway, such as survival as primordial planetesimals or outcomes of unusual collisional or internal processes, motivating the use of anomaly-detection methods to systematically identify such objects [11].**
- *Reference(s):* [11]
- *Justification:* [11] focuses on human–chatbot interaction and self-disclosure effects of agent embodiment. It does not discuss asteroids, main-belt orbital dynamics, spin states, primordial planetesimals, or anomaly-detection methods in astronomy. Therefore the statement is not supported by [11].

Mathematical consistency audit

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

Maths relevance: light

The paper contains a small number of central analytic definitions (log transforms, standardization, residuals, and anomaly scores) rather than extended derivations. The main internal-consistency risks are variable-definition clarity (especially obliquity) and the domain requirements of logarithms given a reported zero minimum diameter.

Checked items

1. ⚠ **Log transforms applied to diameter/spin_period/semimajor_axis** (Sec. 2.2, p.3; reiterated Sec. 3.1, p.5)
    - **Claim:** Apply natural logarithm transforms $\log(x)$ to diameter, spin_period, and semimajor_axis to reduce skewness; obliquity not transformed.
    - **Checks:** domain/definition consistency, notation consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
    - **Assumptions/inputs:** Each log-transformed variable is strictly positive on the rows used for modeling, Any zeros/negatives are removed or otherwise handled before applying log
    - **Notes:** Table 1 (p.11) reports diameter min = 0.00 km for the non-anomalous population, which conflicts with an unqualified $\log(\text{diameter})$ transform ($\log(0)$ undefined). The paper does not specify whether zeros are filtered from the modeling datasets, treated as missing, or shifted by an offset before logging.
2. ✔ **Standardization (z-scoring) of predictors** (Sec. 2.2, p.3; Sec. 3.1, p.5)
    - **Claim:** Numerical predictor variables are standardized to mean 0 and std 1 using training-set statistics, then applied to test data.
    - **Checks:** definition consistency, pipeline logic consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Standardization is applied only to predictors, not necessarily to targets, Scaler fit uses only training partition per task
    - **Notes:** The described order (log-transform then StandardScaler for predictors; one-hot encode categorical type) is internally coherent and matches later references to "standardized units" (e.g., kernel length_scale interpretation in Sec. 3.2.1).
3. ✔ **GPR standardized residual anomaly score** (Sec. 2.4, p.4; Sec. 3.3.1, p.9–10)
    - **Claim:** Define $S_{\mathrm{GPR,i}} = (y_i - \mu_i)/\sigma_i$, where μ_i and σ_i are the GPR predictive mean and predictive standard deviation, and y_i is the observed (possibly log-transformed) target.
    - **Checks:** algebra/definition correctness, symbol consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** $\sigma_i > 0$ for all evaluated points, y_i and μ_i are on the same scale (log space for logged targets)
    - **Notes:** Formula is algebraically correct for a standardized residual and is consistently described in both Methods and Results. The paper does not discuss $\sigma_i \to 0$ edge cases (handled separately in another item).
4. ✔ **NN residual definition** (Sec. 2.4, p.4)

- **Claim:** Define the NN residual $r_{\mathrm{NN,i}} = y_i - \hat{y}_i$.
- **Checks:** algebra/definition correctness
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** y_i and $\hat{y}_i$ are on the same scale (log space for logged targets)
- **Notes:** Residual definition is standard and consistent with later residual plots labeled "True − Predicted" (e.g., Figures 16, 18, 20).

5. ✔ **NN z-score anomaly score** (Sec. 2.4, p.4; Sec. 3.3.1, p.10)

   - **Claim:** Define $S_{\mathrm{NN,i}} = (r_{\mathrm{NN,i}} - \mathbf{mean}(r_{\mathrm{NN}}))/\,\mathbf{std}(r_{\mathrm{NN}})$; equivalently ($y_i - \hat{y}i - \operatorname{mean}(r)$.}))/\operatorname{std}(r_{\rm NN
   - **Checks:** algebraic equivalence, definition consistency
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** $\mathbf{std}(r_{\mathrm{NN}}) > 0$ for the residual set over which it is computed, mean/std are computed over a clearly defined population (train, test, or full cleaned dataset)
   - **Notes:** The two written forms are algebraically equivalent. The exact residual population used to compute mean/std is not stated (full cleaned set vs. held-out), which affects interpretation but not algebra.

6. ✔ **Threshold rule $|S_i| > 3$** (Sec. 2.4, p.4; Sec. 3.3.1, p.10)

   - **Claim:** Flag anomalies when the absolute anomaly score exceeds **3**, interpreted as greater than three standard deviations from prediction.
   - **Checks:** dimensional consistency, logic consistency with definitions
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Scores are approximately standardized (unitless), Using a common threshold across properties is meaningful after scoring standardization
   - **Notes:** Both S_{GPR} and S_{NN} are dimensionless by construction, so a common numeric threshold is dimensionally consistent. The statistical interpretation differs between heteroscedastic σ_i (GPR) and global $\mathbf{std}(r_{\mathrm{NN}})$ (NN), but that is a modeling choice rather than an internal algebra error.

7. ⚠ **Count of anomaly flags (six per asteroid) vs property-level wording** (Sec. 2.4, p.4; Sec. 3.3.1, p.10)

   - **Claim:** The process generates six boolean anomaly flags per asteroid (3 properties × 2 model types).
   - **Checks:** definition consistency, counting/logic consistency
   - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
   - **Assumptions/inputs:** Each model-property combination is evaluated separately, Flags are not collapsed across models unless explicitly stated

- **Notes:** Methods clearly state per-model-per-property flags (six flags). Results contain a sentence suggesting a property-level OR across models ("flagged … if … from either the GPR or NN model for that property exceeded a threshold of 3") but then still states six flags and reports per-model anomaly counts. This is likely wording ambiguity but should be clarified.

8. ✖ **Obliquity target scale consistency** (Sec. 2.2, p.3; Sec. 3.2.1 (Figures 9–11), p.7–8; Sec. 3.2.2 (Figures 19–20), p.10)
    - **Claim:** Obliquity is modeled without transformation and used directly in predictions and residuals.
    - **Checks:** symbol/definition consistency, units/range sanity check
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Obliquity is consistently defined across dataset, plots, models, and anomaly scoring, Units/range are specified or inferable
    - **Notes:** The text treats the target as "obliquity" (an angular quantity) but plots show true/predicted values approximately in $[-1, 1]$, suggesting a normalized/cosine-like representation rather than an angle. No definition is provided to reconcile this. This undermines the precise meaning of y_i in $S_{\mathrm{GPR,i}}$ and $S_{\mathrm{NN,i}}$ for obliquity.

9. ✖ **Interpretation of low obliquity anomaly yield vs scoring definition** (Sec. 3.3.1, p.10)
    - **Claim:** Near-absence of obliquity anomalies is partly due to GPR overfitting leaving minimal residual variance to identify outliers.
    - **Checks:** consistency between verbal reasoning and equations
    - **Verdict:** FAIL; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Anomaly scoring uses $S_{\mathrm{GPR,i}} = (y_i - \mu_i)/\sigma_i$ as defined, Model behavior described (overfitting) should map to changes in μ_i, σ_i , and residuals
    - **Notes:** Given $S_{\mathrm{GPR,i}}$ divides by σ_i, fewer threshold exceedances align with larger σ_i (larger predictive uncertainty), not "minimal residual variance." Also, earlier the paper notes high predictive uncertainty for obliquity (Figure 10, p.7), which is the mathematically direct explanation for small $|S|$, whereas "minimal residual variance" is not aligned with the defined scoring mechanism.

10. ⚠ **Division by σ_i stability in GPR scores** (Sec. 2.4, p.4; kernel discussion Sec. 3.2.1, p.6)
    - **Claim:** Compute $S_{\mathrm{GPR,i}} = (y_i - \mu_i)/\sigma_i$ using the GPR predictive standard deviation.
    - **Checks:** well-posedness/edge-case analysis
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor

- **Assumptions/inputs:** σ_i is never zero or numerically negligible, Kernel settings (e.g., WhiteKernel noise $\sim 1e{-}10$) do not create degenerate predictive variances on evaluated points
- **Notes:** The scoring formula is well-defined only when $\sigma_i > 0$. The obliquity kernel includes a near-zero noise term ($1e{-}10$, p.6), which can, depending on data geometry, produce extremely small σ_i for points effectively repeated/very near training points. The paper does not state any safeguard (ϵ-flooring) or confirm σ_i is bounded away from zero.

Limitations

- The paper provides few explicit equations and no step-by-step derivations; most content is methodological description, so the audit focuses on definition consistency, algebraic equivalence, and domain/range issues.
- Figures are referenced for variable ranges (e.g., obliquity), but the underlying data schema is not included; conclusions about inconsistencies rely on what is visible in the plotted axes and captions.
- Kernel expressions and training objectives (e.g., log-marginal likelihood maximization) are described at a high level without mathematical detail; this audit does not (and cannot) verify omitted derivations.

Numerical results audit

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

Executed **21** numeric consistency checks: **20** PASS, **1** FAIL. The only failed check concerns the stated obliquity coverage percentage versus the cleaned obliquity dataset size and the master-size lower bound. Other checks (dataset split arithmetic, anomaly-count bounds, ratio statement, kernel noise vs. stated predictive standard deviation, and Table 1 quantile ordering/sanity) are internally consistent.

Checked items

1. ✔ **C1_dataset_split_diameter** (p.3 §2.3 ("split into an 80% training set and a 20% testing set"), p.4 §3.1 ("10,340 asteroids for the diameter prediction task"))
    - **Claim:** Diameter task dataset size is $\mathbf{10,340}$ and was split 80/20 into training/testing.
    - **Checks:** integer_split_consistency
    - **Verdict:** PASS
    - **Notes:** Computed $N_{\text{train}} = 8272$ and $N_{\text{test}} = 2068$; matches $0.2N_{\text{total}}$ exactly.
2. ✔ **C2_dataset_split_spin** (p.3 §2.3, p.4 §3.1 ("6,396 for the spin_period task"))
    - **Claim:** Spin_period task dataset size is $\mathbf{6,396}$ and was split 80/20 into training/testing.

- **Checks:** integer_split_consistency
- **Verdict:** PASS
- **Notes:** Chose $N_{\text{train}} = 5117$ and $N_{\text{test}} = 1279$; N_{test} differs from $0.2N_{\text{total}}$ by 0.2 due to integer rounding.

3. ✔ **C3_dataset_split_obliquity** (p.3 §2.3, p.4 §3.1 ("1,626 for the obliquity task"))
   - **Claim:** Obliquity task dataset size is $1,626$ and was split 80/20 into training/testing.
   - **Checks:** integer_split_consistency
   - **Verdict:** PASS
   - **Notes:** Chose $N_{\text{train}} = 1301$ and $N_{\text{test}} = 325$; N_{test} differs from $0.2N_{\text{total}}$ by 0.2 due to integer rounding.

4. ✔ **C4_master_size_vs_nonanomalous_count** (p.4 §3.1 ("over 1.7 million asteroid entries"), p.11 Table 1 (Non-Anomalous diameter count $1,462,706$; Anomalous diameter count $1,138$))
   - **Claim:** Master dataset has 'over 1.7 million' entries; Table 1 provides counts for diameter populations that sum to $1,463,844$ entries with diameter stats.
   - **Checks:** sum_and_bound_check
   - **Verdict:** PASS
   - **Notes:** Verified exact sum $1,462,706 + 1,138 = 1,463,844$ and that this is below the stated master-size lower bound ($1,700,000$).

5. ✔ **C5_anomalies_unique_vs_flags_lower_bound** (p.10 §3.3.1 ("1,138 unique asteroids flagged as anomalous"; flags: $973, 989, 136, 141, 3, 0$))
   - **Claim:** A total of $1,138$ unique anomalous asteroids were flagged; per-model/property anomaly counts are listed.
   - **Checks:** set_union_bounds
   - **Verdict:** PASS
   - **Notes:** Union bounds hold: unique=$1,138$ is $\geq$ max individual count (989) and $\leq$ sum of counts ($2,242$).

6. ✔ **C6_spin_period_anomalous_count_mismatch** (p.10 §3.3.1 (spin_period anomalies listed), p.11 Table 1 (spin_period anomalous count $1,136$; diameter anomalous count $1,138$))
   - **Claim:** Table 1 reports spin_period anomalous population count $1,136$, while overall unique anomalous asteroids count is $1,138$.
   - **Checks:** cross_table_count_consistency
   - **Verdict:** PASS
   - **Notes:** Computed implied missing spin_period among anomalous objects: $1,138 - 1,136 = 2$.

7. ✔ **C7_spin_period_coverage_from_table** (p.4 §3.1 ("spin_period data were present for only approximately 3.2% of entries"), p.11 Table 1 (spin_period Non-Anomalous count $\mathbf{62,767}$; spin_period Anomalous count $\mathbf{1,136}$; master size 'over $\mathbf{1.7}$ million'))
    - **Claim:** Spin period coverage is $\sim \mathbf{3.2\%}$ of master entries; Table 1 implies $\mathbf{62,767 + 1,136 = 63,903}$ entries with spin_period in the comparison.
    - **Checks:** percentage_recompute_with_bounds
    - **Verdict:** PASS
    - **Notes:** Upper-bound fraction using master lower bound: $\mathbf{63,903/1,700,000 = 0.03759}$ $(\mathbf{3.759\%})$, within the loose relative tolerance vs. the $\sim \mathbf{3.2\%}$ claim given the 'over $\mathbf{1.7}$ million' denominator.
8. ✖ **C8_obliquity_coverage_from_cleaned_vs_master_claim** (p.4 §3.1 ("obliquity data for a mere 0.2%"; cleaned obliquity dataset size $\mathbf{1,626}$; master 'over $\mathbf{1.7}$ million'))
    - **Claim:** Obliquity presence is $\sim \mathbf{0.2\%}$ of master; cleaned obliquity modeling dataset has $\mathbf{1,626}$ entries after requiring predictors too.
    - **Checks:** percentage_bound_check
    - **Verdict:** FAIL
    - **Notes:** Computed $\mathbf{1,626/1,700,000 = 0.00095647}$ $(\mathbf{0.0956\%})$, which is $\mathbf{0.478\times}$ the stated $\mathbf{0.2\%}$ (relative difference $\approx \mathbf{0.522}$), exceeding the allowed relative tolerance.
9. ✔ **C9_over_30_times_claim** (p.10 §3.3.2 ("mean diameter ... $\mathbf{63.81}$ km ... over $\mathbf{30}$ times larger than ... $\mathbf{1.96}$ km"), p.11 Table 1)
    - **Claim:** Mean anomalous diameter $\mathbf{63.81}$ km is over $\mathbf{30\times}$ mean non-anomalous diameter $\mathbf{1.96}$ km.
    - **Checks:** ratio_check
    - **Verdict:** PASS
    - **Notes:** Computed ratio $\mathbf{63.81/1.96 = 32.5561}$, which satisfies the 'over $\mathbf{30\times}$ ' claim.
10. ✔ **C10_table1_quartile_ordering_diameter_anom** (p.11 Table 1 (diameter anomalous row))
    - **Claim:** For anomalous diameter, min $\leq \mathbf{25\%} \leq \mathbf{50\%} \leq \mathbf{75\%} \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $\mathbf{2.15 \leq 35.40 \leq 53.20 \leq 117.45 \leq 498.10}$.
11. ✔ **C11_table1_quartile_ordering_diameter_nonanom** (p.11 Table 1 (diameter non-anomalous row))

- **Claim:** For non-anomalous diameter, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
- **Checks:** quantile_monotonicity
- **Verdict:** PASS
- **Notes:** Ordering holds: $0.00 \leq 0.74 \leq 1.10 \leq 1.64 \leq 5909.55$.

12. ✔ **C12_table1_quartile_ordering_spin_anom** (p.11 Table 1 (spin_period anomalous row))
    - **Claim:** For anomalous spin_period, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $2.53 \leq 7.03 \leq 14.21 \leq 16.51 \leq 9567.39$.
13. ✔ **C13_table1_quartile_ordering_spin_nonanom** (p.11 Table 1 (spin_period non-anomalous row))
    - **Claim:** For non-anomalous spin_period, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $0.04 \leq 5.16 \leq 8.96 \leq 28.89 \leq 10167.60$.
14. ✔ **C14_table1_quartile_ordering_inclination_anom** (p.11 Table 1 (inclination anomalous row))
    - **Claim:** For anomalous inclination, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $0.24 \leq 1.01 \leq 1.51 \leq 2.21 \leq 34.92$.
15. ✔ **C15_table1_quartile_ordering_inclination_nonanom** (p.11 Table 1 (inclination non-anomalous row))
    - **Claim:** For non-anomalous inclination, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $0.01 \leq 4.33 \leq 7.87 \leq 12.70 \leq 175.98$.
16. ✔ **C16_table1_quartile_ordering_eccentricity_anom** (p.11 Table 1 (eccentricity anomalous row))
    - **Claim:** For anomalous eccentricity, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.

- **Checks:** quantile_monotonicity
- **Verdict:** PASS
- **Notes:** Ordering holds: $0.01 \leq 0.06 \leq 0.13 \leq 0.17 \leq 0.27$.

17. ✔ **C17_table1_quartile_ordering_eccentricity_nonanom** (p.11 Table 1 (eccentricity non-anomalous row))
    - **Claim:** For non-anomalous eccentricity, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $0.00 \leq 0.09 \leq 0.15 \leq 0.20 \leq 0.997$.
18. ✔ **C18_table1_quartile_ordering_age_anom** (p.11 Table 1 (age anomalous row))
    - **Claim:** For anomalous age, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds (including equality at min=$q25$): $0.01 \leq 0.01 \leq 1.30 \leq 2.50 \leq 3.50$.
19. ✔ **C19_table1_quartile_ordering_age_nonanom** (p.11 Table 1 (age non-anomalous row))
    - **Claim:** For non-anomalous age, min $\leq 25\% \leq 50\% \leq 75\% \leq$ max should hold.
    - **Checks:** quantile_monotonicity
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $0.00 \leq 0.30 \leq 0.93 \leq 1.50 \leq 3.50$.
20. ✔ **C20_kernel_uncertainty_peak_vs_noise_level** (p.6-7 §3.2.1 (spin_period kernel noise_level= 2.07); Fig 8 caption says predictive std ~ 1.44 corresponding to model's noise level)
    - **Claim:** Spin_period GPR: WhiteKernel(noise_level = 2.07); Fig.8 caption: predictive standard deviation peaked around ~ 1.44 corresponding to the model's noise level.
    - **Checks:** sqrt_consistency
    - **Verdict:** PASS
    - **Notes:** $\sqrt{2.07} = 1.43875$ matches the stated ~ 1.44 within tight tolerance.
21. ✔ **C21_nonanom_diameter_max_vs_mean_sanity** (p.11 Table 1 (diameter non-anomalous row))
    - **Claim:** Non-anomalous diameter has mean 1.96 km and max 5909.55 km; verify max $\geq$ mean and max $\geq$ quartiles, etc.

- **Checks:** basic_sanity_bounds
- **Verdict:** PASS
- **Notes:** Sanity inequalities hold (max≥mean, max≥ $q75$, and min≤mean≤ max).

Limitations

- Audit is restricted to the provided PDF text; no underlying CSV data, code outputs, or per-object flags/IDs are available to recompute many reported statistics.
- Several key quantities are described approximately (e.g., "over **1.7** million", "approximately **3.2%**"), limiting strict consistency checks to bounds/ plausibility rather than exact equality.
- Image-based figures are not used for numeric extraction; only numbers explicitly written in the PDF are considered.
- Exact percentages of entries with spin_period and obliquity in the master dataset cannot be verified because the PDF provides only approximate percentages and a lower-bound master size rather than exact counts.
- Neural-network validation losses cannot be recomputed without training logs/data splits/model outputs.
- The exact union/overlap structure behind the anomaly-flag counts cannot be verified without per-asteroid flags/IDs.
- Quantitative values implied only by histogram/plot shapes in figures cannot be verified without image-based numeric extraction.