Skepthical review: *Unveiling the Intrinsic Structure of the Asteroid Belt: Correcting for Observational Selection Bias in Physical and Compositional Properties*

Summary

The manuscript proposes a survey-agnostic framework to correct observational selection in asteroid physical/compositional property studies by explicitly modeling the probability that a "sparse" property is observed. From a merged catalog of ∼1.45M asteroids with near-universal orbital elements (and purportedly diameter) but sparse measurements of spectral type, spin period, obliquity, family membership, and family age, the authors define binary availability indicators (has_Property) and train separate propensity models (Random Forest classifiers, with logistic-regression baselines) to estimate $\hat{p}_i = P(\text{Observed} = 1 \mid Z_i)$ from diameter and orbital elements (Secs. 2.1–2.3). They assess discrimination via AUC-ROC/AUC-PR and show confusion matrices and calibration plots (Sec. 3.1), then invert propensities into inverse-probability weights (IPW) with a high-percentile cap to control extremes (Sec. 2.4). These weights are used to re-estimate "bias-corrected" distributions and relationships (spin period, obliquity, family age; type and family frequencies; diameter-by-type; age-by-family) (Secs. 3.2–3.4). Reported corrections are generally modest but interpretable (e.g., slightly higher C-type fraction, somewhat smaller intrinsic diameters within types, slightly faster spin-period distribution, and reduced dominance of bright/highly studied families) (Sec. 3.3).

Conceptually, applying propensity modeling/IPW to asteroid-catalog missingness at this scale is valuable and timely. However, several elements required for the corrected estimates to be scientifically interpretable are currently underdeveloped: (i) the missing-data identification assumptions (MAR/ignorability and positivity) are not stated and are plausibly violated for key endpoints; (ii) the "universal" predictors—especially diameter and family-related variables—raise endogeneity/circularity concerns; (iii) calibration and overlap diagnostics tailored to IPW are not quantified; (iv) sensitivity to truncation/model choice and formal uncertainty quantification for the corrected estimates are absent; and (v) key implementation details (data provenance, preprocessing, splits, hyperparameters, estimator definitions) are insufficient for reproduction. Addressing these points would substantially strengthen the paper's credibility and long-term usefulness.

Strengths

- Targets an important and widely recognized problem in asteroid population inference —non-random availability of physical characterization—and frames a practical, scalable correction strategy (Introduction, Secs. 2–4).

- Builds a very large merged dataset (∼1.45M objects) and adopts a modular workflow separating (a) selection/propensity modeling from (b) downstream weighted descriptive inference (Secs. 2.1–2.5).
- Uses appropriate discrimination metrics for imbalanced outcomes (AUC-ROC and AUC-PR) and includes visual diagnostics (confusion matrices and calibration curves) rather than reporting a single headline score (Sec. 3.1, Figs. 1–10).
- Applies inverse probability weighting in a consistent spirit across multiple scientific questions (type fractions, spin distributions, family representation), illustrating where selection effects matter and where results appear more robust (Secs. 3.2–3.4).
- The narrative motivation—that selection in follow-up characterization can distort apparent population properties—is clear and the examples are relevant to current asteroid survey practice (Introduction, Sec. 4).

Major issues

1. **Identification assumptions for IPW (conditional ignorability/MAR and positivity) are not explicitly stated, justified, or stress-tested, yet they are essential for interpreting the weighted results as unbiased population estimates (Introduction, Secs. 2.3–2.5, Secs. 3.2–3.5, Sec. 4).** The current propensity models condition only on orbital elements and diameter, but many plausible drivers of whether a property is measured are not represented: apparent brightness/H, observing geometry/phase angle coverage, survey footprint/strategy, discovery circumstances, follow-up targeting (e.g., unusual colors/lightcurve amplitude), NEOness/MOID, number of observations/arc length, and epoch. For endpoints like spectral type and spin period, it is plausible that the property value (or proxies correlated with it) influences follow-up even after conditioning on $(a, e, i, \Omega, \omega, D)$, which would violate MAR and bias IPW "corrections."

   Recommendation: Add a dedicated "Assumptions and identification" subsection (end of Sec. 2.4 or start of Sec. 2.5) that: (i) states the estimand (distribution over the master catalog) and the required assumptions (e.g., $\mathbf{Observed}_X \perp X \mid Z$ and positivity $0 < P(\mathbf{Observed}_X = 1|Z) < 1$); (ii) argues property-by-property why the assumption is plausible or likely violated; and (iii) discusses likely directions of bias under violations. Where feasible, expand Z with more proximate observability/targeting covariates available at scale (e.g., absolute magnitude H, perihelion distance q, MOID/NEO flag, number of astrometric observations/arc length/oppositions, discovery epoch, survey/catalog-of-origin flags). Report in Sec. 3.1 whether discrimination/calibration improve and, crucially, whether key weighted conclusions in Secs. 3.2–3.4 are stable.

2. **Using diameter as a "universal" predictor is potentially problematic and may induce endogeneity or hidden selection, undermining the propensity model and its interpretation (Sec. 2.1–2.2, Sec. 3.1).** In many catalogs, diameter is not directly observed for all asteroids; it is inferred using thermal IR surveys

(with their own selection) or brightness plus assumed/albedo-dependent modeling. If diameter is missing for a nontrivial subset, or if its construction is correlated with the same follow-up processes that generate the sparse properties, it may not be an appropriate conditioning variable and can leak selection information in ways that are hard to interpret.

Recommendation: In Sec. 2.1–2.2, document the provenance of Diameter_km (source(s), method of derivation, coverage fraction, and how conflicts are resolved). Report the fraction of the 1.45M catalog with diameter and the fraction used in each propensity model. Add an ablation/sensitivity study (Sec. 3.1 and/or Appendix): fit propensity models (a) with diameter, (b) without diameter, and (c) with H (and optionally both H+diameter) where available; compare calibration/overlap and show the effect on at least the headline corrected quantities (e.g., C/S fractions, median spin period, top family frequencies in Sec. 3.2–3.3). If diameter is near-universal only because of imputation/assumptions, state this clearly and discuss implications in Sec. 4.

3. **Calibration and overlap (positivity) diagnostics are not quantified, despite being central to IPW stability and credibility (Secs. 2.3.4–2.4, Sec. 3.1, Figs. 6–11).** Visual calibration curves alone are insufficient because small probability miscalibration in the low-propensity tail can dominate weighted estimates. The manuscript also does not report overlap diagnostics (propensity distributions among observed, minimum/percentiles), effective sample size (ESS) after weighting, or how much inference is driven by a small subset of extreme weights.

   Recommendation: In Sec. 3.1, add quantitative calibration metrics on a held-out test set for each propensity model (Brier score and either ECE or calibration slope/intercept; optionally log-loss). Include tail-focused diagnostics (e.g., calibration restricted to $\hat{p} < 0.01$ and $\hat{p} < 0.001$, with bin counts). Add overlap/positivity diagnostics per property: summary of $\hat{p}$ among observed objects, weight quantiles, ESS (e.g., $(\sum w)^2 / \sum w^2$), and the fraction of total weight carried by the top 1% of objects. If post-hoc calibration (Platt/isotonic) is used, specify it in Sec. 2.3.4 and show metric improvements.

4. **Sensitivity to truncation/stabilization choices and to propensity-model specification is not demonstrated, yet the reported debiasing can be highly sensitive to these design decisions (Sec. 2.4, Secs. 3.2–3.4).** The current text also contains ambiguity about "raw," "truncated," and "stabilized" weights (Sec. 2.4, Sec. 3.1, Fig. 11 captions), making it unclear which weights underpin each downstream table/figure.

   Recommendation: In Sec. 2.4, provide explicit mathematical definitions for each weight variant used (e.g., raw $w_i = 1/\hat{p}_i$; truncated $\tilde{w}_i = \min(w_i, c)$; stabilized $w_i^{\text{stab}} = \Pr(\text{Observed} = 1)/\hat{p}_i$ if applicable) and state unambiguously which variant is used in each analysis (Secs. 3.2–3.4; Tables 2–6). Then perform a sensitivity analysis: recompute a small set of headline results under multiple caps (e.g.,

95th/97.5th/99th/99.5th percentiles) and at least one alternative well-calibrated model class (e.g., logistic regression with calibration, or gradient boosting). Summarize the variation in Sec. 3.5 and temper claims in Sec. 4 accordingly.

5. **Uncertainty quantification is absent for the bias-corrected estimates, preventing assessment of which reported shifts are robust versus within statistical/model uncertainty (Secs. 3.2–3.4, Tables 2–6, Sec. 3.5).** IPW typically increases variance, and here weights are also estimated, introducing additional uncertainty. Without confidence intervals (CIs) or standard errors (SEs), statements about changes (e.g., modest changes in taxonomic fractions or spin-period summaries) cannot be evaluated for practical/statistical significance.

   Recommendation: Augment Secs. 3.2–3.4 and Tables 2–6 with uncertainty for key quantities (weighted means/medians/proportions and selected cross-feature summaries). Prefer a two-stage bootstrap that resamples asteroids and refits the propensity model (or, if too expensive, a hybrid bootstrap that reuses fitted propensities but resamples outcomes with weights, clearly labeled as partial). Report 95% CIs for headline shifts (e.g., C/S fractions, median spin period, top-N family fractions). In Sec. 3.5, explicitly distinguish results that remain meaningfully shifted under CIs/sensitivity analyses from those that do not.

6. **FamilyName and Age_Gyr are treated like object-level "observed/not observed" endpoints, but their construction and selection mechanisms differ qualitatively from spectral/spin/obliquity and may be circular with the predictors (Sec. 2.1, Secs. 3.3.2–3.4).** Family identification is typically derived from (proper) orbital elements; thus predicting "has_FamilyName" from orbital elements may be partly tautological rather than a model of observational selection. Age estimates are often assigned at the family level (not per object), so interpreting IPW-weighted per-object age distributions requires careful definition of the estimand (object-weighted vs family-weighted) and selection process (propensity for families to have ages, not just objects).

   Recommendation: In Sec. 2.1, define precisely: (i) what "FamilyName" represents (membership in any family vs background; which catalog; whether multiple memberships exist and how resolved), and (ii) what "Age_Gyr" represents (per-family constant vs per-object estimate; source and reconciliation). In Sec. 2.3, justify whether modeling "has_FamilyName" is intended as an observational selection process or merely a completeness-of-family-classification process; if the latter, explicitly reframe it. For ages, consider a hierarchical alternative: model (a) propensity to be assigned to a family, and (b) propensity for a family to have an age estimate; then report both object-weighted and family-weighted age summaries (Sec. 3.4). If a hierarchical treatment is out of scope, narrow claims about "intrinsic age distributions" (Sec. 4) and clearly state the estimand being estimated.

7. **Key implementation and estimator details remain under-specified, limiting reproducibility and making it difficult to verify that downstream weighted quantities are well-defined (Secs. 2.1–2.6, Sec. 2.5, Sec. 3.1).** This includes: data provenance/versioning; exact preprocessing (log transforms, handling of zeros/negatives, clipping $\hat{p}$ away from 0); train/validation/test splitting and seeds; Random Forest hyperparameter search spaces and final settings; whether/why features were scaled for RF; and explicit formulas for weighted estimators (Hájek vs Horvitz–Thompson, weighted quantiles/median, weighted correlation).

   Recommendation: Expand Secs. 2.1–2.6 to include: (i) catalog sources, query dates, and selection criteria (numbered/multi-opposition, filtering); (ii) precise preprocessing steps with formulas (e.g., log or log1p with ϵ; angular encoding; missing-value handling) and confirmation that any scaling is fit on training data only; (iii) split strategy (shared vs per-target splits, stratification, random seeds); (iv) hyperparameter ranges, CV setup, and chosen hyperparameters per model (table in main text or Appendix); and (v) explicit definitions of all weighted estimators used in Sec. 2.5 (and which software functions implement them). Also add explicit safeguards: clip $\hat{p} \in [\epsilon, 1-\epsilon]$ prior to inversion and document ϵ. Provide code and/or a DOI-linked repository in Sec. 2.6 where feasible.

Minor issues

1. Missingness rates and missingness patterns are not summarized, limiting readers' ability to understand the severity/structure of the selection problem before modeling (Secs. 2.1–2.3, Sec. 3.1).

   Recommendation: Add a concise descriptive table/figure (new Sec. 2.1.1 or early Sec. 3.1) reporting, for each sparse property, the count and fraction observed, plus simple comparisons of key predictors (e.g., diameter and semi-major axis) between observed vs unobserved subsets.

2. Model-performance reporting overemphasizes threshold-based metrics (precision/recall/F1, confusion matrices) that are not central for propensity estimation, and decision thresholds/splits are often unclear (Sec. 3.1, Table 1, Figs. 1–5).

   Recommendation: In Sec. 3.1, clearly state the threshold used for confusion matrices (or omit thresholded matrices from the main text) and emphasize proper scoring/calibration metrics (Brier/log-loss/ECE) and overlap/ESS diagnostics. If keeping confusion matrices, add normalized (percentage) versions and explicitly label test-set evaluation.

3. Angular orbital elements (e.g., Ω, ω) are treated as linear predictors; this can introduce discontinuities near $0/360^\circ$ and spurious splits, potentially affecting propensity estimates (Sec. 2.2–2.3).

Recommendation: Either encode angular variables using sine/cosine pairs (preferred) or justify why linear treatment is acceptable. If changed, report whether calibration/diagnostics materially improve (Sec. 3.1).

4. Independent per-property propensity models and IPWs are used, but cross-feature analyses implicitly condition on multiple properties being observed; correlated missingness could bias such joint analyses (Secs. 2.3–2.5, Sec. 3.3.2, Sec. 3.4).

   Recommendation: Clarify in Sec. 2.5 how weights are applied when analyzing relationships that require multiple observed properties (e.g., type-by-diameter, age-by-family). Add a robustness check in Sec. 3.3/3.4 (or Appendix): compare results across subsets defined by the availability of another property, or fit a simple joint selection model for one key pair (e.g., spectral type and spin period).

5. Validation of the corrected distributions is largely internal; confidence would increase with external benchmarks or negative-control checks (Sec. 3.5, Sec. 4).

   Recommendation: Add at least one validation exercise: (i) compare corrected taxonomic fractions or gradients against an external dataset/subsample with better-characterized selection (cite relevant surveys), and/or (ii) include a negative-control outcome expected to be minimally affected by follow-up targeting conditional on Z. Summarize implications in Sec. 3.5.

6. Figures (especially Figs. 1–5 and 11–12) are hard to read at typical print size; confusion matrices show raw counts only; several plots omit units, sample sizes, or normalization details (Sec. 3.1–3.3).

   Recommendation: Increase resolution/font sizes, provide normalized confusion matrices (row/column %), and add units/normalization (count vs density) and sample sizes in captions. Ensure figures are introduced in order and do not interrupt sentences (Sec. 3 formatting).

7. Astrophysical contextualization and related work on selection functions/completeness corrections are present but not yet systematic (Introduction, Sec. 3.5, Sec. 4).

   Recommendation: Add a short related-work subsection (e.g., Sec. 1.1) on asteroid selection/completeness corrections and position this work relative to survey simulators and hierarchical Bayesian selection-function approaches. In Sec. 4, temper broad generality claims and state explicitly when this IPW approach is expected to work well versus when explicit survey modeling is preferred.

Very minor issues

1. Inconsistent notation and terminology for propensities and weights (e.g., multiple ways of writing $P(\mathbf{observed})$; ambiguity between 'truncated' and 'stabilized'), and inconsistent variable naming across text/figures (Secs. 2.4–2.5, Sec. 3.1, Fig. 11).

Recommendation: Add a brief notation block defining Z_i, $\hat{p}_i$, and each weight variant, and standardize variable names (e.g., monospace `SpinPeriod_hr`, `SpectralType`, `FamilyName`) throughout.

2. Minor LaTeX/HTML artifacts (e.g., >, <), heading-format inconsistencies (Markdown-style hashes), and small typographical issues appear in several places (Secs. 2–3).

   Recommendation: Proofread to remove encoding artifacts, standardize heading styles/numbering to journal format, and fix minor typos and spacing.

3. Acronyms (IPW, AUC-ROC, AUC-PR) are not always defined at first use and are sometimes inconsistently formatted (Secs. 2.3–2.5, Sec. 3.1).

   Recommendation: Define all acronyms at first use and use consistent hyphenation/capitalization throughout.

4. Some keywords and claims appear broader than the manuscript's actual scope (Abstract/keywords, Sec. 4), and language occasionally implies stronger claims ("intrinsic truth," "significant") than warranted without uncertainty/sensitivity analysis.

   Recommendation: Revise keywords to match content, and soften language to "bias-corrected under the stated selection model" unless supported by the added uncertainty and robustness analyses.

Key statements and references

- **✖ The observed symmetric, nearly uniform distribution of asteroid spin-axis obliquities around 91° is interpreted as a robust representation of the true population and is described as consistent with a collisionally evolved system in which spin axes have been randomized over time.**
- *Reference(s):* Denario [11]
- *Justification:* Denario [11] studies noise and continuous norm emergence in agent-based social simulations. It contains no analysis of asteroid spin-axis obliquities, no distribution around **91°**, and no discussion of collisional randomization of spin axes. Therefore the statement is not supported.

- **✖ The bias-corrected increase in the relative abundance of C-type (carbonaceous) asteroids from 17.6% to 18.3% of the typed population is said to reinforce the view of a solar system with a more pronounced compositional gradient, in which carbonaceous bodies are intrinsically more common than raw observations suggest.**

- *Reference(s):* Denario [11]
- *Justification:* Denario [11] studies noise and norm emergence in agent-based societies and does not discuss asteroid taxonomy, C-type abundances, bias-correction, or solar system compositional gradients. No figures or text mention **17.6%** or **18.3%** values or

carbonaceous bodies. Therefore, the statement is not supported.

- ✖ **The finding that C-type asteroids, which are generally darker (lower albedo) and more prevalent in the outer main belt, are underrepresented in spectroscopic surveys is linked to the broader discussion that observational biases hinder accurate inference of the asteroid belt's composition.**
- *Reference(s):* Denario [11]
- *Justification:* Denario [11] analyzes noise in agent-based models of social norm emergence and does not discuss asteroids, albedo, main-belt distributions, or spectroscopic survey biases. Thus it provides no support for the statement.
- ✖ **The appearance of the Tirela family in the weighted top 10 of asteroid families, despite its absence from the unweighted list, is highlighted as emphasizing the significant impact of observational biases on our understanding of the relative abundance of different asteroid families, as previously discussed in the introduction.**
- *Reference(s):* Denario [11]
- *Justification:* Denario [11] analyzes noise in the emergence of social norms in agent-based simulations and does not discuss asteroid families, weighted vs. unweighted rankings, the Tirela family, or observational biases in asteroid family abundance.

Mathematical consistency audit

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

Maths relevance: light

The paper is primarily methodological and empirical, with limited explicit mathematics. The central analytic component is inverse probability weighting (IPW) based on a learned observation-probability model. Several downstream weighted estimators (means, quantiles, correlations, category frequencies) are invoked but not written as formulas, which limits symbolic verification of internal consistency beyond basic IPW definitions.

Checked items

1. ✔ **IPW weight definition** (Sec. 1 (end of p.2) and Sec. 2.4 (p.4))
    - **Claim:** Define inverse probability weight for asteroid i and property X as $w_i = 1/P(\text{observed } X \mid \text{orbital elements, diameter})$ (or model-predicted probability).
    - **Checks:** symbol/notation consistency, well-definedness
    - **Verdict:** PASS; confidence: high; impact: critical
    - **Assumptions/inputs:** Model outputs a valid probability $\hat{p}_i$ in $(0, 1]$, Inference uses only the observed subset for property X

- **Notes:** The definition $w_i = 1/\hat{p}_i$ is internally consistent with later descriptions of up-weighting low-probability observed cases.

2. ✔ **Restriction to observed subset for a given property** (Sec. 2.4 (p.4))
    - **Claim:** For each sparse property, compute weights only for asteroids where that property is observed, then use those weights to correct distributions/relationships.
    - **Checks:** logic of estimator setup
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Observed subset is identified via has_Property indicator, Weighting targets full-catalog estimands via IPW
    - **Notes:** Conceptually consistent with standard IPW usage (reweight observed cases to represent the full population). However, the exact estimators used downstream are not written, so only the setup can be checked.

3. ✔ **Weight 'stabilization' via 99th-percentile cap** (Sec. 2.4 (p.4))
    - **Claim:** To mitigate extremely large weights, cap weights at the 99th percentile of raw weights.
    - **Checks:** definition consistency, well-definedness
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Raw weights are finite and a percentile is computable, Capping threshold is defined per property
    - **Notes:** As a truncation rule, this is well-defined if raw weights are finite. The paper later uses separate terms ('truncated' vs 'stabilized') without defining them distinctly (see next item).

4. ✖ **Raw vs truncated vs stabilized weights terminology** (Sec. 2.4 (p.4) vs Sec. 3.1 and Fig. 11 caption (pp.5, 9))
    - **Claim:** The analysis uses 'stabilized inverse probability weights' and shows distributions of 'raw, truncated, and stabilized' weights.
    - **Checks:** definition consistency, traceability of quantities
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Distinct weight variants exist and are defined
    - **Notes:** Sec. 2.4 equates 'stabilization' with capping at the 99th percentile (a truncation). Later text/caption implies three separate versions (raw, truncated, stabilized) but does not provide a separate mathematical definition of 'stabilized' beyond capping. This prevents determining which weights produced Tables 2–6 and Fig. 12.

5. ⚠ **Log transformation definition** (Sec. 2.2 (p.3))
    - **Claim:** Apply $\log(x)$ transform to skewed predictors such as Diameter_km and SemimajorAxis_AU.

- **Checks:** domain/well-definedness
- **Verdict:** UNCERTAIN; confidence: medium; impact: minor
- **Assumptions/inputs:** Transformed variables satisfy $x > 0$ for all rows used, Handling of missing/zero values is specified
- **Notes:** $\log(x)$ is undefined at $x \leq 0$. The paper does not state any handling for zeros or nonpositive diameters/axes (or whether such values are guaranteed absent). Clarify preprocessing (filtering or adding epsilon).

6. ⚠ **Inversion requires nonzero observation probabilities** (Secs. 1–2.4 (pp.2–4))

   - **Claim:** Compute w_i = 1 / P_\text{observed_PropertyX_i} for each observed asteroid.
   - **Checks:** well-definedness, edge-case sanity
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Model never outputs exactly 0 probability for any observed case, If it can, probabilities are clipped
   - **Notes:** If the classifier can output $\hat{p}_i = 0$ (or numerically 0), weights become infinite and percentile capping is ill-posed. The text discusses 'very small probabilities' but does not specify clipping away from 0.

7. ✔ **Weighted category 'corrected counts' as sum of weights** (Sec. 2.5.2 (p.4))

   - **Claim:** Corrected count for a category equals the sum of IPW weights over observed asteroids in that category.
   - **Checks:** algebraic form, estimand consistency
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Weights correspond to inverse inclusion/observation probabilities for the property, Target is a total over the full catalog (Horvitz–Thompson-style)
   - **Notes:** Summing w_i within category is internally consistent with an IPW/HT estimator for population totals, assuming weights are defined as inverse observation probabilities for having the category label measured.

8. ⚠ **Weighted percentages/proportions for categories** (Sec. 3.3 and Table 3 (pp.7–9))

   - **Claim:** Report weighted (%) for spectral types based on IPW-corrected frequencies.
   - **Checks:** normalization consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Weighted percent = (sum of weights in category)/(sum of weights over all typed asteroids) × 100, or some other stated normalization

- **Notes:** The paper does not state the normalization used to convert weighted counts into percentages (especially important if weights are capped/truncated). Without an explicit formula, the reported weighted (%) cannot be symbolically verified as properly normalized.

9. ⚠ **Weighted descriptive statistics (mean/median/std/quartiles)** (Sec. 2.5.1 (p.4) and Table 2 (pp.6–7))

    - **Claim:** Compute weighted mean/median/std/quartiles for observed numerical properties using IPW weights.
    - **Checks:** definition completeness, well-definedness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** A specific definition of weighted quantiles/median is chosen, Mean/std computed with either normalized or unnormalized weights
    - **Notes:** No formulas are provided for weighted variance/std or weighted quantiles. Different standard definitions exist and can yield different results (analytic point). Provide explicit estimator definitions to make the math checkable.

10. ⚠ **Weighted Pearson correlation** (Sec. 2.5.3 (p.4))

    - **Claim:** Recalculate Pearson correlation coefficients using weights when at least one property is sparse.
    - **Checks:** definition consistency, missing-case logic
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor
    - **Assumptions/inputs:** Definition of weighted covariance and correlation is specified, Handling of two sparse variables (which weights?) is defined
    - **Notes:** The paper does not define the weighted correlation formula nor the rule for choosing weights when correlating two variables (each with potentially different observation processes/weights). This is analytically under-specified.

11. ✔ **Within-family weighted mean age invariance explanation** (Sec. 3.4.2 (p.10) and Table 6 (p.11))

    - **Claim:** If Age_Gyr is assigned at the family level (constant for all members), then weighting does not change the mean age within that family.
    - **Checks:** algebraic invariance/sanity check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** All asteroids in a given family share exactly the same Age_Gyr value in the analysis subset, Weights are positive and finite
    - **Notes:** If $y_i = c$ for all i in the group, then weighted mean $\sum w_i y_i / \sum w_i = c$ regardless of weights, so the stated reasoning is internally consistent.

Limitations

- The provided PDF text contains very few explicit equations; many key analytic steps are described verbally without formal estimator definitions, limiting the scope of symbolic verification.
- Figures are referenced but their underlying computations (e.g., how stabilized weights differ from truncated weights) are not mathematically defined in the text available here.
- No appendix or formal notation section is present to disambiguate weighting/normalization choices for medians/quantiles, correlations, and percentages.

Numerical results audit

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

All 19 automated numeric checks passed. Verified items include dataset size narrative consistency, train/test split arithmetic, k-fold parameter sanity, recomputed F1-scores from precision/recall (Table 1), AUC-ROC range statements vs tabulated values, table-to-text rounding consistency for descriptive statistics (Table 2), percentage-point change arithmetic for spectral types (Table 3), family-percentage rounding consistency (Table 4), bias-corrected vs unweighted diameter inequalities (Table 5), exact equality of mean ages by family (Table 6), and monotonic ordering/bounds checks for descriptive statistics.

Checked items

1. ✔ **CAND-001** (Page 2, Section 2.1 (Data Acquisition and Preparation))
    - **Claim:** The study obtained data for 1,452,682 unique asteroids.
    - **Checks:** cross-section consistency (repeated constant)
    - **Verdict:** PASS
    - **Notes:** Interpreted 'over 1.4 million' as strict lower bound; 1,452,682 satisfies.
2. ✔ **CAND-002** (Page 3, Section 2.3.3 (Training, Tuning, and Validation))
    - **Claim:** Dataset of 1,452,682 asteroids is split into training (80%) and test (20%).
    - **Checks:** parts vs total (train/test split arithmetic)
    - **Verdict:** PASS
    - **Notes:** Fractions sum to 1.0 and a consistent integer split exists (e.g., 1,162,145 train + 290,537 test = 1,452,682).
3. ✔ **CAND-003** (Page 3, Section 2.3.3 (Training, Tuning, and Validation))
    - **Claim:** k-fold cross-validation with $k = 5$.
    - **Checks:** parameter self-consistency

- **Verdict:** PASS
- **Notes:** $k = 5$ is a positive integer > 1.

4. ✔ **CAND-004** (Page 6, Table 1 (Performance Metrics of Selection Models on Test Data))
    - **Claim:** For each model, F1-score should equal $2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$ using the tabulated precision and recall.
    - **Checks:** metric recomputation (F1 from precision/recall)
    - **Verdict:** PASS
    - **Notes:** All five rows match recomputed F1 within tolerance; worst absolute difference in SpinPeriod.

5. ✔ **CAND-005** (Page 5-6, Section 3.1 + Table 1)
    - **Claim:** AUC-ROC scores range from 0.8615 (FamilyName) to 0.9852 (SpectralType).
    - **Checks:** min/max consistency (range statement vs table)
    - **Verdict:** PASS
    - **Notes:** Claimed endpoints equal min/max over the five tabulated AUC-ROC values.

6. ✔ **CAND-006** (Page 5-6 (Abstract/Intro mentions 0.86-0.99) vs Page 6 Table 1)
    - **Claim:** AUC-ROC scores are reported as (0.86-0.99) in narrative; Table 1 shows 0.8615-0.9852.
    - **Checks:** narrative range vs exact values
    - **Verdict:** PASS
    - **Notes:** Table min/max fall within narrative range and round to 2 decimals as 0.86 and 0.99.

7. ✔ **CAND-007** (Page 6-7, Section 3.2.1 + Table 2)
    - **Claim:** Weighted median SpinPeriod_hr decreases from 8.89 hours to 8.45 hours.
    - **Checks:** table-to-text consistency (rounding)
    - **Verdict:** PASS
    - **Notes:** Table medians round to the narrative values and weighted median is smaller than unweighted.

8. ✔ **CAND-008** (Page 7, Section 3.2.2 + Table 2)
    - **Claim:** Obliquity mean and median remain centered around **91** degrees; unweighted and weighted nearly identical.
    - **Checks:** difference check (small shift claim)
    - **Verdict:** PASS

- **Notes:** Mean abs diff ≈ **0.0002°**; median abs diff ≈ **0.0817°**; all values within **90–92°** and under the stated **0.1°** threshold.

9. ✔ **CAND-009** (Page 7, Section 3.2.3 + Table 2)
   - **Claim:** Mean age decreases from **1.14** Gyr to **1.06** Gyr; median remains **0.93** Gyr.
   - **Checks:** table-to-text consistency (rounding + equality)
   - **Verdict:** PASS
   - **Notes:** Means match narrative after 2-decimal rounding; weighted/unweighted medians both equal **0.93**.
10. ✔ **CAND-010** (Page 9, Table 3 (Top Spectral Types))
    - **Claim:** Change (%) column equals Weighted (%) minus Unweighted (%).
    - **Checks:** difference recomputation (table arithmetic)
    - **Verdict:** PASS
    - **Notes:** All five rows match to within floating-point noise; worst abs diff in S row.
11. ✔ **CAND-011** (Page 7-9, Section 3.3.1 + Table 3)
    - **Claim:** C-type increases from **17.6%** to **18.3%** of the typed population.
    - **Checks:** table-to-text consistency (rounding)
    - **Verdict:** PASS
    - **Notes:** **17.64%** and **18.27%** round to 1 decimal as **17.6%** and **18.3%**; direction is an increase.
12. ✔ **CAND-012** (Page 9, Table 3 (Top Spectral Types))
    - **Claim:** Unweighted (%) and Weighted (%) for listed spectral types sum to a plausible partial total (top-5 types only).
    - **Checks:** partial sum check (sanity)
    - **Verdict:** PASS
    - **Notes:** Top-5 sums are **86.25%** (unweighted) and **86.28%** (weighted), both $\leq$ 100%.
13. ✔ **CAND-013** (Page 11, Table 4 (Major Asteroid Families))
    - **Claim:** Family proportions changes described in text match Table 4 (e.g., Vesta **11.8%** to **10.3%**; Eos **6.1%** to **6.5%**; Hungaria **5.7%** to **4.4%**).
    - **Checks:** table-to-text consistency (rounding)
    - **Verdict:** PASS
    - **Notes:** Rounding Table 4 to 1 decimal reproduces the narrative numbers and change directions (Vesta down, Eos up, Hungaria down).
14. ✔ **CAND-014** (Page 11, Table 4 (Major Asteroid Families))

- **Claim:** Family percentages listed (unweighted and weighted) do not exceed 100% when summed (note: table is partial list).
- **Checks:** partial sum check (sanity)
- **Verdict:** PASS
- **Notes:** Summed listed families: **47.56%** (unweighted) and **45.11%** (weighted), both $\leq$ **100%**.

15. ✔ **CAND-015** (Page 11, Table 5 (Mean Diameter by Spectral Type) + Page 10, Section 3.4.1)
    - **Claim:** Bias-corrected mean diameter is smaller than unweighted mean diameter for all listed spectral types; example: S-type **5.70** km vs **5.42** km.
    - **Checks:** inequality check (weighted vs unweighted)
    - **Verdict:** PASS
    - **Notes:** For B,C,S,V,X, weighted mean diameter is smaller than unweighted; S-type example matches 2-decimal rounding.
16. ✔ **CAND-016** (Page 11, Table 6 (Mean Age by Family Name) + Page 10, Section 3.4.2)
    - **Claim:** For listed families, unweighted and weighted mean ages are identical.
    - **Checks:** equality check (table internal consistency)
    - **Verdict:** PASS
    - **Notes:** All listed families match exactly as printed (difference 0.0).
17. ✔ **CAND-017** (Page 7, Table 2 (SpinPeriod_hr stats))
    - **Claim:** Quartiles and extrema are ordered correctly: Min $\leq$ 25% $\leq$ Median $\leq$ 75% $\leq$ Max (for each unweighted/weighted numerical property).
    - **Checks:** monotonic order check (descriptive stats sanity)
    - **Verdict:** PASS
    - **Notes:** Both unweighted and weighted SpinPeriod sequences are monotone increasing from min through max.
18. ✔ **CAND-018** (Page 7, Table 2 (Obliquity_deg stats))
    - **Claim:** Obliquity_deg bounds are within physical range **0-180** degrees, and ordered correctly.
    - **Checks:** range + monotonic order check
    - **Verdict:** PASS
    - **Notes:** All listed obliquity stats lie within $[\mathbf{0}, \mathbf{180}]$ and satisfy min$\leq$q1$\leq$median$\leq$q3$\leq$max for both unweighted and weighted.
19. ✔ **CAND-019** (Page 7, Table 2 (Age_Gyr stats))
    - **Claim:** Age_Gyr bounds and quartiles are ordered correctly and within stated min/max (**0.0020** to **3.5000**).

- **Checks:** monotonic order + bounds check
- **Verdict:** PASS
- **Notes:** Unweighted and weighted age sequences are ordered correctly and all quantiles fall within their min/max bounds.

Limitations

- Only the provided PDF text was used; figures were not numerically parsed, so any claims supported only by plots (confusion matrices, calibration plots, IPW histograms) could not be verified.
- Several narrative numeric claims are rounded; checks therefore rely on rounding rules (e.g., to 1–2 decimals) rather than exact equality.
- No raw dataset, intermediate tables (e.g., full spectral-type or family distributions), or model outputs are included, limiting verification to arithmetic relationships among reported summary statistics.