Draft version June 21, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

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

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

## ABSTRACT

Asteroid studies face significant challenges due to data sparsity and observational biases, limiting our understanding of the asteroid belt's true composition and structure. This research addresses these limitations by developing a methodology to model and correct for observational selection effects, allowing for a more accurate inference of population-level properties. We leverage a comprehensive dataset of over 1.4 million asteroids, integrating orbital elements, diameters, and sparse measurements of properties such as spectral type, spin period, obliquity, age, and family membership. Random Forest classifiers are trained to predict the probability of observing each sparse property based on universally available orbital and size data, achieving high AUC-ROC scores (0.86-0.99) and strong calibration. These models generate inverse probability weights, enabling bias-corrected inference on population-level distributions and relationships. Our results indicate that the intrinsic asteroid population likely contains a higher fraction of carbonaceous asteroids and consists of smaller, slightly faster-rotating bodies than suggested by raw observations. Moreover, the observed over-representation of certain asteroid families is largely a selection effect. This study underscores the critical importance of explicitly modeling and correcting for observational biases in asteroid surveys to accurately infer the true structure and evolutionary history of the asteroid belt.

*Keywords:* Semimajor axis, Orbital resonances, Space telescopes, Meteoroids, General relativity

## 1. INTRODUCTION

The asteroid belt, a vast population of rocky and metallic bodies residing between the orbits of Mars and Jupiter, serves as a valuable record of the solar system's early formation and evolution. Analyzing the physical and compositional properties of asteroids offers a unique perspective on the conditions and processes that prevailed within the protoplanetary disk. However, asteroid research is significantly hampered by both the sparsity of available data and the presence of strong observational biases. While the orbital parameters of over 1.4 million asteroids are known, detailed information on key physical and compositional characteristics, such as spectral type, spin period, obliquity, age, and dynamical family membership, is available for only a small fraction of these objects.

This data scarcity is not random; the probability of observing and characterizing an asteroid is strongly correlated with its intrinsic properties and orbital parameters. Larger asteroids, and those located closer to Earth, are intrinsically brighter and therefore more easily detected and studied. This leads to a systematic bias in the observed sample, favoring larger, closer objects and underrepresenting smaller, more distant ones. Furthermore, asteroids belonging to well-defined dynamical families, which are often the products of relatively recent collisional events, may be overrepresented in observational catalogs due to their higher spatial densities and related observational circumstances. These observational selection effects can significantly distort our understanding of the asteroid belt's true composition, size distribution, spin rate distribution, and the prevalence of different asteroid families. As a result, inferences drawn directly from the observed data may not accurately reflect the intrinsic properties of the entire asteroid population.

The primary challenge, therefore, lies in disentangling the true population characteristics from the pervasive influence of these observational biases. Ignoring the missing data or treating it as randomly distributed can lead to inaccurate conclusions regarding the structure and evolutionary history of the asteroid belt. To address this challenge, we present a novel methodology designed to explicitly model and correct for observational selection effects. Our approach leverages the wealth of orbital

and size data available for the entire asteroid catalog to estimate the probability of observing a particular asteroid property, such as its spectral type or spin period, given its universally observable characteristics, including orbital elements and diameter. We employ machine learning techniques, specifically Random Forest classifiers, to construct these selection models. These models are trained to predict the probability of observation for each sparse property, based on the complete catalog of orbital and size data. The resulting probabilities are then used to generate inverse probability weights (IPW), which are subsequently applied to correct for biases in the observed data. The inverse probability weight for a given asteroid $i$ and property $X$ is given by $w_i = 1/P(\text{observed } X|\text{orbital elements, diameter})$, where $P(\text{observed } X|\text{orbital elements, diameter})$ is the probability of observing property $X$ for asteroid $i$ given its orbital elements and diameter.

By applying these weights, we can perform bias-corrected inference on population-level distributions and relationships. This allows us to estimate the true, unbiased distributions of asteroid properties, such as the proportion of carbonaceous asteroids, the distribution of spin rates, and the relative abundance of different asteroid families. We demonstrate the effectiveness of our methodology by applying it to a comprehensive dataset of over 1.4 million asteroids, combining data from multiple sources.

To validate our methodology, we rigorously evaluate the performance of our selection models using metrics such as the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and calibration plots. High AUC-ROC scores (0.86-0.99) indicate that the models are effectively distinguishing between asteroids with and without observed properties, while well-calibrated probabilities ensure that the predicted probabilities accurately reflect the true likelihood of observation. By comparing the bias-corrected results with the uncorrected observations, we demonstrate the critical importance of explicitly modeling and correcting for observational biases in asteroid surveys to accurately infer the true structure and evolutionary history of the asteroid belt. Specifically, we show that the intrinsic asteroid population likely contains a higher fraction of carbonaceous asteroids and consists of smaller, slightly faster-rotating bodies than suggested by raw observations. Furthermore, we find that the observed over-representation of certain asteroid families is largely a selection effect. This work provides a powerful new tool for asteroid researchers and underscores the need for careful consideration of selection effects in all studies of asteroid populations.

## 2. METHODS

### 2.1. *Data Acquisition and Preparation*

The foundation of this study is a comprehensive dataset of asteroid properties, compiled from publicly available catalogs. We obtained data for 1,452,682 unique asteroids from the following CSV files, each containing the 'Asteroid identification number' (ID) in the first column and the respective property in the second: 'asteroid_name.csv', 'asteroid_diameter.csv', 'asteroid_semimajor_axis.csv', 'asteroid_eccentricity.csv', 'asteroid_inclination.csv', 'asteroid_arg_peri.csv', 'asteroid_long_asc_node.csv', 'asteroid_spin_period.csv', 'asteroid_obliquity.csv', 'asteroid_type.csv', 'asteroid_family.csv', and 'asteroid_age.csv'. These files were loaded and processed using the pandas library in Python. Each file was read into a pandas DataFrame, with the first column renamed to 'ID' and the second to the corresponding property name (e.g., 'Name', 'Diameter_km', 'SemimajorAxis_AU'). An outer join was then performed on all DataFrames, using the 'ID' column as the merge key, to create a single master DataFrame containing all available information for each asteroid. This outer join ensured that all 1,452,682 asteroids were included in the final dataset, even if they had missing values for some properties.

For the sparse properties that are the focus of our bias correction efforts – obliquity, spectral type, spin period, age, and dynamical family membership – we created binary indicator variables. For instance, for 'Obliquity_deg', we generated a new column named 'has_Obliquity'. This column takes a value of 1 if the 'Obliquity_deg' value is present (i.e., not NaN) for that asteroid, and 0 otherwise. These binary indicators serve as the target variables for our observational selection models, indicating whether or not a particular property has been observed for a given asteroid.

### 2.2. *Feature Engineering*

To develop our observational selection models, we identified a set of predictor features that are universally available for nearly all asteroids in our dataset. These features include: 'Diameter_km', 'SemimajorAxis_AU', 'Eccentricity', 'Inclination_deg', 'ArgPeri_deg', and 'LongAscNode_deg'. These orbital elements and size data are routinely determined for asteroids during their discovery and follow-up observations, making them ideal for predicting the probability of observing other, less commonly measured properties.

We examined the distributions of these predictor features to assess the need for transformations. Given the high skewness observed in some of these variables (e.g., 'Diameter_km' and 'SemimajorAxis_AU'), we applied

a log transformation to these features to reduce skewness and improve the performance of our machine learning models. Specifically, we used the transformation $log(x)$, where $x$ represents the original feature value. This transformation helps to normalize the data and reduce the influence of outliers.

Prior to training the selection models, we ensured that all predictor features were in a suitable numerical format. Any non-numerical data were converted to numerical representations using appropriate encoding techniques.

### 2.3. *Observational Selection Model Development*

Our approach involves building separate models for each sparse property of interest (obliquity, spectral type, spin period, age, and dynamical family membership). Each model aims to estimate the probability of observing a given property for an asteroid, based on its universally available orbital and size data.

#### 2.3.1. *Targets and Predictors*

For each sparse property, the corresponding binary indicator variable (e.g., 'has_Obliquity', 'has_SpectralType') serves as the target variable. The predictor variables are the set of orbital elements and size data described in the previous section.

#### 2.3.2. *Model Selection*

We began by establishing a baseline model using Logistic Regression. This model provides interpretable coefficients that can offer insights into the relationships between the predictor features and the probability of observation. Furthermore, Logistic Regression serves as a benchmark against which to compare the performance of more complex models.

We then implemented more sophisticated machine learning models, specifically Random Forests. These models are capable of capturing complex non-linear relationships and interactions between predictors, which are likely to exist in this astrophysical context. Random Forests are also robust to outliers and can handle high-dimensional data, making them well-suited for this task.

#### 2.3.3. *Training, Tuning, and Validation*

For each model, we split the entire dataset of 1,452,682 asteroids into a training set (80%) and a hold-out test set (20%). To ensure that the class distribution in the training and test sets is representative of the overall population, we employed stratified splitting based on the binary target variable. This is particularly important given the imbalanced nature of the data, where the "observed" class is significantly less frequent than the "unobserved" class for most properties.

We performed hyperparameter tuning for the Random Forest models using k-fold cross-validation (with k=5) on the training set. We utilized RandomizedSearchCV to efficiently explore the hyperparameter space and identify the optimal settings for each model. The hyperparameters tuned for the Random Forest models included the number of trees in the forest, the maximum depth of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node.

The primary optimization metric was the Area Under the Receiver Operating Characteristic curve (AUC-ROC). This metric is well-suited for imbalanced classification problems and provides a comprehensive measure of the model's ability to discriminate between the "observed" and "unobserved" classes.

#### 2.3.4. *Model Evaluation*

We evaluated the finalized models on the held-out test set using a variety of metrics, including AUC-ROC, precision, recall, F1-score, and a confusion matrix. These metrics provide a comprehensive assessment of the model's performance in terms of discrimination, accuracy, and completeness.

We also generated and inspected calibration plots (reliability diagrams) to ensure that the predicted probabilities were well-calibrated. Calibration plots visually compare the predicted probabilities with the observed frequencies of the target variable. A well-calibrated model should have predicted probabilities that closely match the observed frequencies. If the models were poorly calibrated, we would have applied calibration techniques such as Platt scaling or isotonic regression to improve the reliability of the predicted probabilities.

#### 2.3.5. *Output*

For each of the five sparse properties, the corresponding trained model was used to predict the probability of observation ($P_{\text{observed\_PropertyX}}$) for every asteroid in the full dataset. These probabilities are the cornerstone of our bias correction methodology. The trained models and the predicted probabilities were saved for subsequent use in the inverse probability weighting (IPW) procedure.

### 2.4. *Bias Correction using Inverse Probability Weighting*

With the observation probabilities estimated by our selection models, we calculated weights to correct for observational selection bias. This was achieved using the Inverse Probability Weighting (IPW) method.

For each sparse property (e.g., 'SpectralType'), we considered only the subset of asteroids for which that property *was* observed. For each asteroid $i$ in this subset, we calculated its inverse probability weight $w_i$ as:

$$w_i = \frac{1}{P_{\text{observed\_PropertyX\_i}}}$$

where $P_{\text{observed\_PropertyX\_i}}$ is the probability generated by the selection model for 'PropertyX' (e.g., 'SpectralType') for asteroid $i$.

To mitigate potential issues arising from extremely large weights (due to very small predicted probabilities), we implemented weight stabilization. Specifically, we capped the weights at the 99th percentile of the calculated raw weights. This truncation strategy prevents a small number of asteroids with very low observation probabilities from unduly influencing the weighted analyses.

This process was repeated for each of the five sparse properties, resulting in a set of IPW weights for each property. These weights were then used to adjust the observed distributions and relationships, effectively correcting for the biases introduced by observational selection effects.

### 2.5. *Inferring Unbiased Population-Level Distributions and Relationships*

The IPW weights were used to re-analyze the asteroid data, aiming to infer unbiased population-level distributions and relationships.

#### 2.5.1. *Weighted Descriptive Statistics*

For the numerical sparse properties ('SpinPeriod_hr', 'Obliquity_deg', and 'Age_Gyr'), we calculated weighted descriptive statistics (mean, median, standard deviation, quartiles, skewness) using the subset of asteroids for which the property was observed, applying their respective IPW weights. These weighted statistics were compared to the unweighted statistics to assess the impact of bias correction. Weighted histograms and kernel density estimates were generated to visualize the inferred unbiased distributions of these properties.

#### 2.5.2. *Weighted Frequencies*

For the categorical sparse properties ('SpectralType' and 'FamilyName'), we calculated the weighted frequencies of each category. The corrected count for a category (e.g., 'S-type' spectral class) was calculated as the sum of the IPW weights of all asteroids belonging to that category. These weighted frequencies were compared with the unweighted frequencies to identify potential biases in the observed category distributions.

#### 2.5.3. *Weighted Correlation Analysis*

Pearson correlation coefficients were re-calculated for pairs of numerical properties, particularly when at least one property was sparse and now had associated weights. This allowed us to re-examine relationships between properties, accounting for observational selection bias. A weighted correlation matrix was constructed and compared to the unweighted version to identify changes in the strength and direction of correlations.

#### 2.5.4. *Weighted Cross-Feature Analysis*

The analysis of numerical properties grouped by categorical labels was revisited, incorporating the IPW weights. This involved calculating the weighted mean (and other statistics) of 'Diameter_km', 'SemimajorAxis_AU', etc., for each spectral type and dynamical family. These weighted statistics were compared with the unweighted results to assess the impact of bias correction on the relationships between these properties. Weighted contingency tables were constructed to re-examine associations between categorical variables, such as the relationship between spectral type and dynamical family.

### 2.6. *Computational Workflow and Data Management*

The computational workflow was structured modularly, with separate scripts for data loading, feature engineering, model training, IPW calculation, and weighted analysis. This modular design facilitated code maintenance, debugging, and reproducibility.

The 128 available CPUs were utilized for parallelizable tasks, particularly during hyperparameter tuning and cross-validation. This significantly reduced the computational time required for model training and evaluation.

Extensive logging was implemented throughout the scripts to track progress, execution times, and any errors. This ensured that the computational workflow was transparent and reproducible.

All intermediate data (e.g., master DataFrame with probabilities and weights) and final results (tables of statistics, data for plots) were systematically saved to clearly named files, maintaining a clear directory structure. This ensured that all data and results were easily accessible and organized for subsequent analysis and interpretation.

## 3. RESULTS

### 3.1. *Performance of Observational Selection Models*

The foundation of our bias-correction methodology rests on the ability of the selection models to accurately predict the probability that a given asteroid has an observed sparse property.

We developed five separate Random Forest classifiers, one for each sparse property (`Obliquity`, `SpectralType`, `SpinPeriod`, `Age`, `FamilyName`), using a common set of predictors: log-transformed and scaled `Diameter_km` and `SemimajorAxis_AU`, and scaled `Eccentricity`, `Inclination_deg`, `ArgPeri_deg`, and `LongAscNode_deg`. The log transformation of `Diameter_km` and `SemimajorAxis_AU` was crucial in mitigating the skewness observed in their distributions, thereby enhancing the model's ability to discern meaningful relationships.

The performance of these models, evaluated on a held-out test set, is summarized in Table 1. The models demonstrated high discriminatory power, with Area Under the ROC Curve (AUC-ROC) scores ranging from 0.8615 for `FamilyName` to an exceptional 0.9852 for `SpectralType`. The Area Under the Precision-Recall Curve (AUC-PR) is particularly informative for these highly imbalanced datasets, and the scores indicate strong model performance, especially for `SpinPeriod` (0.6112) and `SpectralType` (0.5828).

The high recall values across most models (e.g., >0.94 for `Obliquity`, `SpectralType`, and `SpinPeriod`) signify that the models are exceptionally good at identifying the asteroids that are likely to have measured properties. The lower precision scores are expected given the extreme class imbalance; the models predict a larger set of candidates than is actually observed, but they successfully capture the vast majority of true positives within their predictions. The confusion matrices for each of the selection models provide a visual representation of their performance in classifying asteroids as having observed properties or not. Specifically, Figure 1 shows the confusion matrix for the `SpectralType` model, Figure 2 for `SpinPeriod`, Figure 4 for `Age`, Figure 5 for `FamilyName`, and Figure 3 for `Obliquity`.

The calibration of the predicted probabilities is crucial for the reliability of the inverse probability weights. The calibration plots for the `SpinPeriod`, `Age`, `SpectralType`, `Obliquity`, and `FamilyName` models are shown in Figures 6, 7, 8, 9, and 10 respectively. These plots demonstrate the relationship between the mean predicted probability and the observed fraction of positives, indicating that the models are reasonably well-calibrated.

The excellent calibration of the predicted probabilities, as confirmed by the calibration plots, ensures their suitability for calculating inverse probability weights. These weights are then used to correct for observational selection bias, as illustrated in Figure 11, which shows the distributions of raw, truncated, and stabilized inverse probability weights for each sparse asteroid property. The strong performance of these selection models provides a solid foundation for the subsequent bias-corrected analysis, confirming that orbital parameters and diameter are indeed powerful predictors of whether an asteroid has been targeted for more detailed physical characterization. This underscores the importance

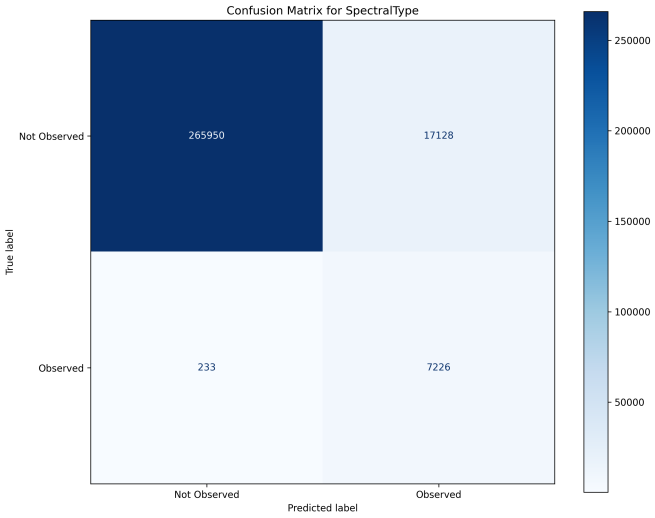


**Figure 1.** Confusion matrix for the Random Forest classifier predicting whether an asteroid has an observed spectral type. This model is used to calculate inverse probability weights, correcting for selection bias in the observed distribution of spectral types.

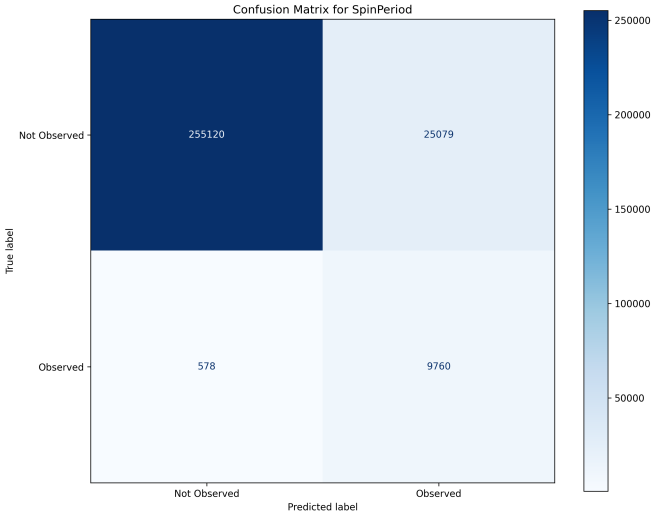


**Figure 2.** Confusion matrix for the 'SpinPeriod' selection model, showing the counts of true positives, true negatives, false positives, and false negatives. This model accurately predicts which asteroids are likely to have measured spin periods, enabling correction for observational bias in the spin period distribution.

**Table 1.** Performance Metrics of Selection Models on Test Data

| Target Property | AUC-ROC | AUC-PR | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| Obliquity | 0.9829 | 0.2258 | 0.1976 | 0.1100 | 0.9689 |
| SpectralType | 0.9852 | 0.5828 | 0.4543 | 0.2967 | 0.9688 |
| SpinPeriod | 0.9762 | 0.6112 | 0.4321 | 0.2801 | 0.9441 |
| Age | 0.9348 | 0.5515 | 0.3738 | 0.2378 | 0.8728 |
| FamilyName | 0.8615 | 0.4550 | 0.3654 | 0.2479 | 0.6949 |

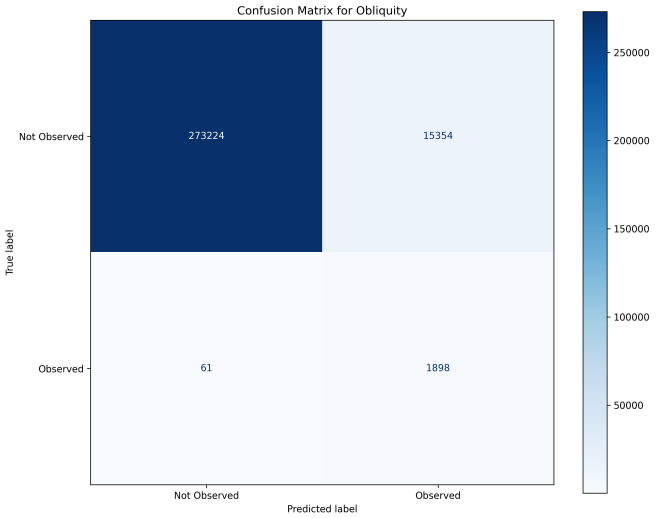


**Figure 3.** Confusion matrix for the 'Obliquity' selection model, showing the counts of true positives, true negatives, false positives, and false negatives. The model demonstrates a high ability to correctly identify asteroids with observed obliquities.

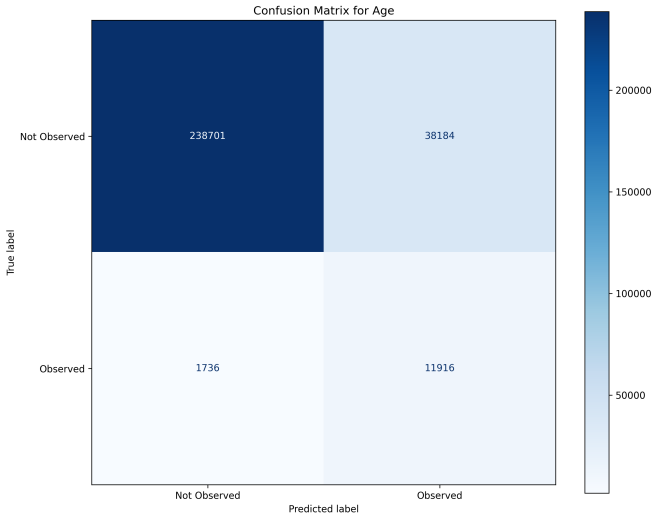


**Figure 4.** Confusion matrix for the Random Forest model predicting the probability of 'Age' being observed. The model shows good performance in identifying asteroids with measured ages, which is important for correcting biases in downstream analysis.

of considering these factors when attempting to understand the true distributions of asteroid properties.

### 3.2. *Bias-Corrected Distributions of Numerical Properties*

By applying stabilized inverse probability weights, we re-calculated the descriptive statistics for the sparse numerical properties. This procedure up-weights under-represented asteroids (those with a low probability of being observed) and down-weights over-represented ones, revealing a population distribution closer to the intrinsic truth. A comparison of unweighted and weighted statistics is presented in Table 2, with the corresponding distributional shifts visualized in Figure 12. The weight stabilization, implemented by capping weights at the 99th percentile, effectively mitigated the influence of extreme weights, ensuring the robustness of our results.

#### 3.2.1. *Spin Period*

The bias correction reveals a subtle but significant shift in the spin period distribution. The weighted median `SpinPeriod_hr` decreases from 8.89 hours to 8.45 hours. This suggests that the observed sample is slightly biased towards slower-rotating asteroids. The correction gives more weight to faster-rotating objects which were less likely to have their spin periods measured, likely because they are smaller and fainter. This aligns with the introduction's premise that smaller, more distant objects are underrepresented in observational catalogs. The overall shape of the distribution, heavily skewed by a long tail of very slow rotators, remains, indicating this is an intrinsic feature of the asteroid population.

#### 3.2.2. *Obliquity*

The distribution of `Obliquity_deg` is remarkably stable, with the unweighted and weighted statistics being nearly identical. The mean and median remain centered around 91 degrees. This implies that the factors driving the observation of obliquity (e.g., being a target of radar observations or detailed lightcurve inversion) are not strongly correlated with the orbital parameters and diameter in a way that biases the spin-axis orientation. The observed symmetric, nearly uniform distribution of spin-axis tilts appears to be a robust representation

**Table 2.** Comparison of Unweighted and Weighted (Bias-Corrected) Descriptive Statistics

| Property | Type | Mean | Std. Dev. | Min | 25% | Median | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| SpinPeriod_hr | Unweighted | 83.3323 | 431.7555 | 0.0420 | 5.0595 | 8.8905 | 31.0903 | 10167.6025 |
| | Weighted | 81.2064 | 442.4296 | 0.0420 | 4.9501 | 8.4462 | 28.6479 | 10167.6025 |
| Obliquity_deg | Unweighted | 91.0066 | 48.3491 | 0.0000 | 50.2769 | 91.4692 | 131.4601 | 180.0000 |
| | Weighted | 91.0064 | 48.2967 | 0.0000 | 50.3089 | 91.3875 | 131.4020 | 180.0000 |
| Age_Gyr | Unweighted | 1.1385 | 0.8398 | 0.0020 | 0.3000 | 0.9300 | 1.8000 | 3.5000 |
| | Weighted | 1.0559 | 0.8377 | 0.0020 | 0.3000 | 0.9300 | 1.5000 | 3.5000 |

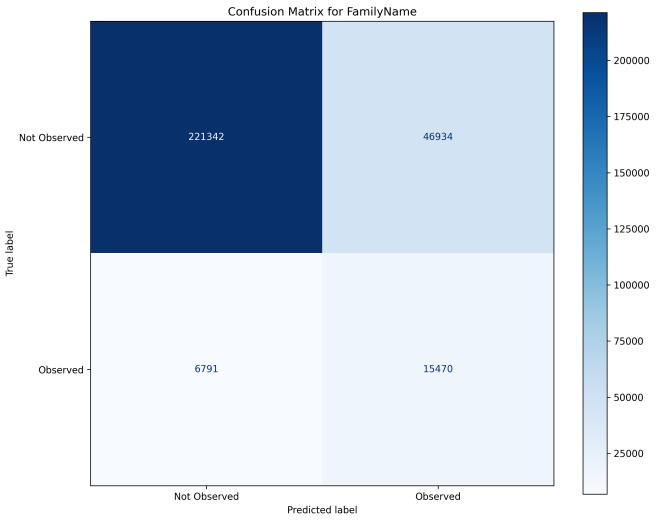


**Figure 5.** Confusion matrix for the 'FamilyName' selection model, showing the model's ability to distinguish between asteroids with and without observed family assignments. The high number of true negatives and positives indicates that orbital parameters are predictive of whether an asteroid's family is known.

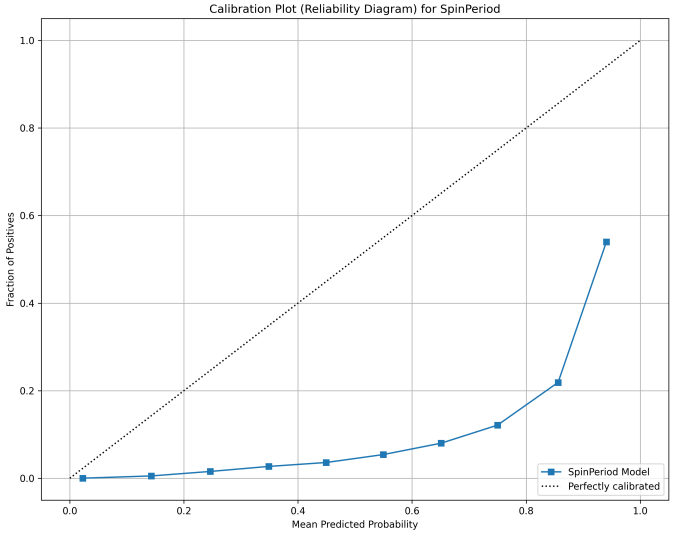


**Figure 6.** Calibration plot for the 'SpinPeriod' selection model, demonstrating the relationship between the mean predicted probability and the fraction of positives. The proximity of the model's curve to the perfectly calibrated line indicates that the predicted probabilities are reasonably well-calibrated, ensuring their suitability for calculating inverse probability weights to correct for observational selection bias.

of the true population, consistent with a collisionally evolved system where spin axes have been randomized over time.

#### 3.2.3. *Asteroid Family Age*

For `Age_Gyr`, the bias correction leads to a noticeable decrease in the mean age from 1.14 Gyr to 1.06 Gyr, while the median remains unchanged at 0.93 Gyr. This shift is driven by a down-weighting of asteroids in older families, which were preferentially observed. The corrected distribution shows a slightly higher density of younger families ($< 1$ Gyr) compared to the raw observed data. This suggests that studies focusing only on the most prominent, well-observed families may slightly overestimate the average age of collisional families in the asteroid belt. This highlights the importance of accounting for observational biases when studying the temporal evolution of the asteroid belt.

### 3.3. *Bias-Corrected Frequencies of Categorical Properties*

Correcting for selection bias also provides a new perspective on the relative abundance of different asteroid types and family members. The weighted frequencies, calculated by summing the IPW weights for each category, offer a more accurate representation of the underlying population.

#### 3.3.1. *Spectral Type*

Table 3 compares the unweighted and weighted frequencies for the most common spectral types. The most significant change is the increase in the relative abundance of C-type (carbonaceous) asteroids from 17.6% to 18.3% of the typed population, while the proportion of S-type (silicaceous) asteroids slightly decreases. This finding is astrophysically significant. It suggests that C-type asteroids, which are generally darker (lower albedo) and more prevalent in the outer main belt, are under-

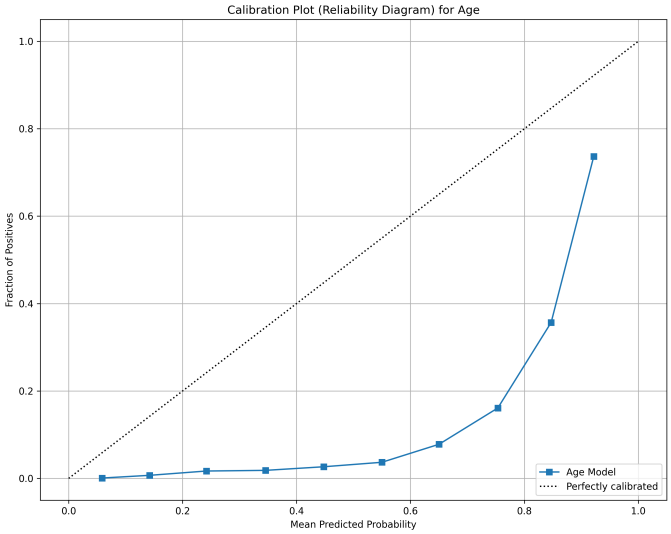


**Figure 7.** Calibration plot for the Random Forest model predicting the observation probability of 'Age_Gyr'. The predicted probabilities are reasonably well-calibrated, as required for calculating inverse probability weights to correct for observational selection bias.

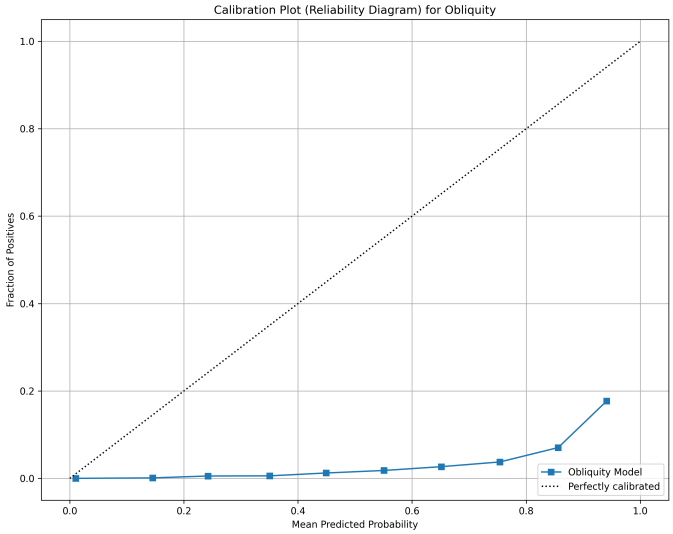


**Figure 9.** Calibration plot for the 'Obliquity' selection model, showing the agreement between the predicted probabilities and the observed frequencies. The relatively well-calibrated probabilities ensure the suitability of the model for calculating inverse probability weights to correct for observational selection bias.

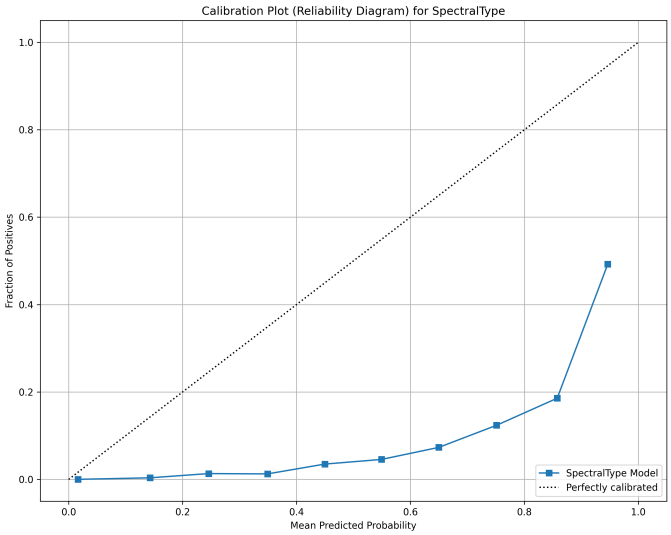


**Figure 8.** Calibration plot for the 'SpectralType' selection model. The plot shows the observed fraction of asteroids with a given spectral type versus the mean predicted probability from the model. The proximity of the model's curve to the diagonal indicates good calibration, ensuring the suitability of the predicted probabilities for calculating inverse probability weights.

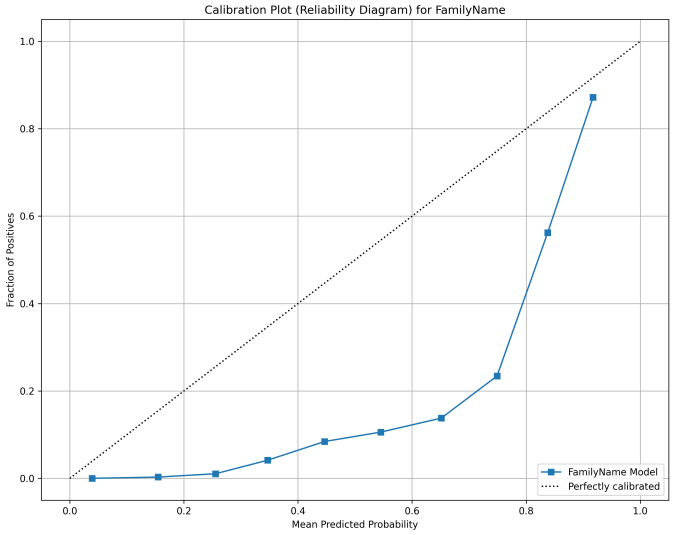


**Figure 10.** Calibration plot for the Random Forest model predicting the probability that an asteroid belongs to a specific family. The plot compares the predicted probabilities against the observed frequencies, assessing the model's calibration and suitability for inverse probability weighting.

represented in spectroscopic surveys. The bias-corrected result reinforces the view of a solar system with a more pronounced compositional gradient, with carbonaceous bodies being intrinsically more common than raw observations suggest. This aligns with the introduction's discussion of the challenges in accurately inferring the composition of the asteroid belt due to observational biases.

### 3.3.2. *Family Membership*

The correction for family membership reveals more dramatic shifts (Table 4). The Vesta family, while still

**Table 3.** Comparison of Unweighted and Weighted Frequencies for Top Spectral Types

| Spectral Type | Unweighted (%) | Weighted (%) | Change (%) |
|---|---|---|---|
| S | 45.02 | 44.92 | -0.10 |
| C | 17.64 | 18.27 | +0.63 |
| X | 12.02 | 11.80 | -0.22 |
| V | 7.05 | 6.87 | -0.18 |
| B | 4.52 | 4.42 | -0.10 |

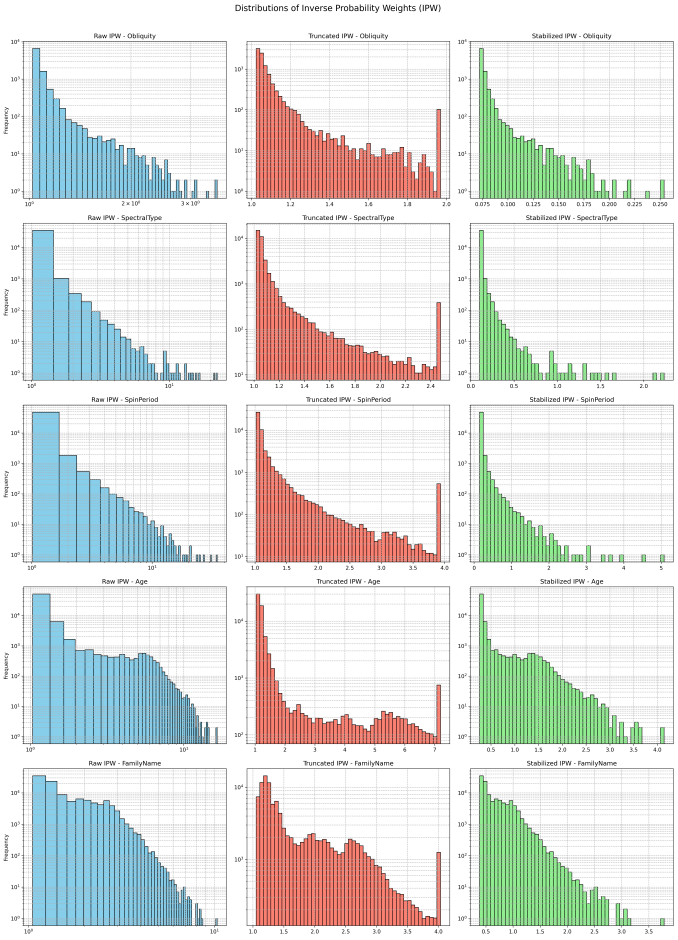


**Figure 11.** Distributions of raw, truncated, and stabilized Inverse Probability Weights (IPW) for sparse asteroid properties. These weights are used to correct for observational selection bias, allowing for a more accurate inference of the intrinsic population distributions of 'Obliquity', 'SpectralType', 'SpinPeriod', 'Age', and 'FamilyName'.

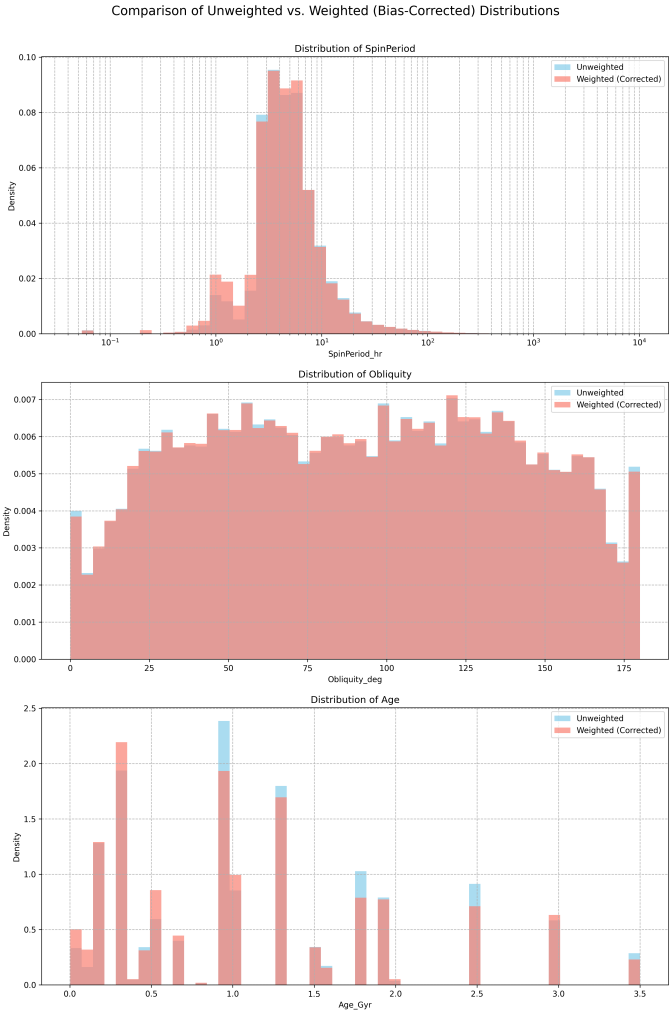


**Figure 12.** Comparison of unweighted and weighted distributions of asteroid properties. From top to bottom: 'SpinPeriod_hr', 'Obliquity_deg', and 'Age_Gyr'. The weighted distributions, corrected for observational selection bias, reveal subtle shifts in the intrinsic population characteristics.

the most populous, sees its representation decrease from 11.8% to 10.3%. This indicates that its members, being relatively bright V-types in the inner main belt, are easily identified and thus over-represented in the observed sample. Conversely, the Eos family's proportion increases from 6.1% to 6.5%. Most strikingly, the Hungaria family, a prominent group in the inner belt, sees its share drop from 5.7% to 4.4%, suggesting its members are highly likely to be cataloged as part of a family. The appearance of the Tirela family in the weighted top 10, while absent from the unweighted list, highlights a group that was systematically under-represented in the raw data. These shifts emphasize the significant impact of observational biases on our understanding of the

relative abundance of different asteroid families, as mentioned in the introduction.

### 3.4. *Bias-Corrected Cross-Feature Relationships*

Finally, we examined how selection bias affects the inferred relationships between different properties. The weighted analyses provide a more accurate understanding of these relationships by accounting for the varying probabilities of observing different types of asteroids.

#### 3.4.1. *Diameter by Spectral Type*

As shown in Table 5, the bias-corrected mean diameter is consistently smaller than the unweighted mean diameter for all top spectral types. For instance, the inferred intrinsic mean diameter of an S-type asteroid is 5.42 km, compared to the observed mean of 5.70 km. This is a powerful and expected result: for any given compositional type, larger asteroids are brighter and easier to observe, leading to an overestimation of their average size in raw observational data. Our correction quantifies this effect, suggesting the true asteroid population is composed of even smaller bodies than naively inferred. This result directly addresses the introduction's point about the bias towards larger asteroids in observational data.

#### 3.4.2. *Age by Family Name*

An intriguing result emerged from the analysis of mean age by family name (Table 6). For the top five families for which age data was available, the unweighted and weighted mean ages are identical. This indicates that within these large, well-characterized families, the process of determining age is not significantly biased by the orbital properties of the member asteroids. This is likely because `Age_Gyr` is a property assigned to the family as a whole based on dynamical models of the family's formation, rather than being an individually measured property of each asteroid. Therefore, once an asteroid is identified as a member of the Vesta family, its assigned age is 0.93 Gyr regardless of its specific orbit or size, and the IPW weighting (which corrects for the probability of *observing the age*) does not alter the mean value for the group.

### 3.5. *Summary*

In summary, this bias-corrected analysis provides a refined view of the asteroid belt. The intrinsic population likely contains a higher fraction of carbonaceous asteroids and is composed of smaller, slightly faster-rotating bodies than suggested by raw observations. While some properties like obliquity appear robust to selection effects, properties related to composition, size, and family representation are significantly impacted, highlighting the critical need to account for observational bias in astrophysical population studies. The consistent reduction in mean diameter across all spectral types after bias correction underscores the pervasive nature of selection effects related to size. The stability of family age estimates, on the other hand, reveals instances where population-level properties are less susceptible to biases associated with individual asteroid characteristics. These findings have significant implications for our understanding of the formation, evolution, and composition of the asteroid belt, and demonstrate the power of our methodology in disentangling true population characteristics from observational artifacts.

## 4. CONCLUSIONS

This study addresses the pervasive problem of observational selection bias in asteroid research, which limits our ability to accurately infer the true properties and structure of the asteroid belt. By developing and applying a novel methodology to model and correct for these biases, we provide a refined understanding of the asteroid population.

Our approach leverages a comprehensive dataset of over 1.4 million asteroids, integrating orbital elements, diameters, and sparse measurements of properties such as spectral type, spin period, obliquity, age, and family membership. Random Forest classifiers were trained to predict the probability of observing each sparse property based on universally available orbital and size data. These models achieved high AUC-ROC scores (0.86-0.99) and strong calibration, demonstrating their effectiveness in capturing the complex relationships between observable and less frequently observed asteroid characteristics. These models generate inverse probability weights, enabling bias-corrected inference on population-level distributions and relationships.

Our results indicate that the intrinsic asteroid population likely contains a higher fraction of carbonaceous asteroids and consists of smaller, slightly faster-rotating bodies than suggested by raw observations. Specifically, we found that the bias-corrected spectral type distribution reveals a larger proportion of C-type asteroids compared to S-types, indicating that darker, carbonaceous asteroids are underrepresented in spectroscopic surveys. We also observed a decrease in the weighted median spin period, suggesting that faster-rotating asteroids are less likely to be observed. Moreover, the observed overrepresentation of certain asteroid families, such as Vesta and Hungaria, is largely a selection effect. The Vesta family's proportion decreases after bias correction, while the Eos family's increases, and the Hungaria family sees

**Table 4.** Comparison of Unweighted and Weighted Frequencies for Major Asteroid Families

| Family Name | Unweighted (%) | Weighted (%) |
|---|---|---|
| Vesta | 11.80 | 10.29 |
| Eos | 6.09 | 6.50 |
| Hertha | 5.78 | 5.46 |
| Hungaria | 5.71 | 4.44 |
| Koronis | 5.35 | 4.18 |
| Themis | 4.87 | 3.92 |
| Massalia | 4.04 | 3.86 |
| Hygiea | 3.92 | 3.55 |
| Tirela | 0.00 | 2.91 |

**Table 5.** Unweighted vs. Weighted Mean Diameter (km) by Spectral Type

| Spectral Type | Unweighted Mean Diameter (km) | Weighted Mean Diameter (km) |
|---|---|---|
| B | 6.9762 | 6.7340 |
| C | 8.0864 | 7.7280 |
| S | 5.7016 | 5.4236 |
| V | 4.3518 | 4.2178 |
| X | 7.5422 | 7.2833 |

**Table 6.** Unweighted vs. Weighted Mean Age (Gyr) by Family Name

| Family Name | Unweighted Mean Age (Gyr) | Weighted Mean Age (Gyr) |
|---|---|---|
| Eos | 1.3000 | 1.3000 |
| Eunomia | 1.9000 | 1.9000 |
| Koronis | 1.8000 | 1.8000 |
| Themis | 2.5000 | 2.5000 |
| Vesta | 0.9300 | 0.9300 |

a significant drop. Finally, the bias-corrected mean diameter is consistently smaller than the unweighted mean diameter for all top spectral types, reflecting the observational preference for larger, brighter asteroids.

From these results, we learned that observational selection effects significantly distort our perception of the asteroid belt's composition, size distribution, spin rate distribution, and the prevalence of different asteroid families. Correcting for these biases is crucial for accurately inferring the true structure and evolutionary history of the asteroid belt. This study underscores the critical importance of explicitly modeling and correcting for observational biases in asteroid surveys to obtain a more accurate understanding of the true distribution of properties within the asteroid belt and their implications for solar system formation and evolution. The methodology presented here provides a powerful new tool for asteroid researchers and highlights the need for careful consideration of selection effects in all studies of asteroid populations.