Draft version July 1, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

The Spatial Architecture of the Main Asteroid Belt: Size, Composition, and Dynamical Gradients

Denario[1]

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

ABSTRACT

The asteroid belt's structure provides a window into its formation and long-term evolution. To understand how dynamical processes have shaped this population, we mapped the joint distribution of asteroid size and composition with orbital elements (semimajor axis, eccentricity, inclination). Using a dataset of 35,623 main-belt asteroids with measured properties, we applied a suite of statistical and machine learning techniques, including one- and two-dimensional binning, Kernel Density Estimation, unsupervised clustering (DBSCAN, Gaussian Mixture Models), and predictive modeling (regression and classification). Our analysis reveals profound structural gradients: asteroid size systematically increases with increasing semimajor axis, and a stark compositional zoning transitions from S-type dominated populations in the inner belt to C-type dominated populations in the outer belt. Kernel Density Estimation highlights the fine-scale density variations in orbital space, while clustering successfully identifies distinct dynamical groups, many corresponding to known asteroid families, each exhibiting characteristic size and compositional distributions. Predictive modeling demonstrates that while orbital location predicts population-level trends, it provides limited predictive power for the properties of individual asteroids, emphasizing the role of stochastic processes like collisions. Furthermore, analysis of mean-motion resonance regions reveals they act as dynamic filters, preferentially depleting smaller asteroids and altering the local compositional mix, consistent with the influence of size-dependent non-gravitational forces such as the Yarkovsky effect. This comprehensive mapping provides a detailed view of the asteroid belt's architecture, illustrating how primordial conditions, collisional evolution, and dynamical sculpting have jointly shaped its present-day configuration.

Keywords: Semimajor axis, Astronomical techniques, Regression, Asteroid dynamics, Small Solar System bodies

1. INTRODUCTION

The main asteroid belt, located between the orbits of Mars and Jupiter, represents a crucial remnant population from the epoch of planet formation. Its structure and composition preserve information about the physical and dynamical conditions of the early solar system and the subsequent evolutionary processes that have shaped this population over billions of years. Understanding the distribution of asteroids in orbital space with respect to their intrinsic properties provides a fundamental window into these processes, bridging the gap between initial formation locations and their present-day configurations.

Deciphering the intricate architecture of the main asteroid belt is a complex challenge. The observed distribution of asteroids is a cumulative outcome of several intertwined factors: the initial mass and compositional gradients within the protoplanetary disk, the gravitational influence of giant planets (particularly Jupiter), ongoing collisional fragmentation and aggregation, non-gravitational forces such as the Yarkovsky effect that cause orbital drift, and the long-term effects of mean-motion and secular resonances which can sculpt the distribution by removing or concentrating objects in specific orbital regions. Disentangling the relative contributions of these diverse processes requires a detailed, quantitative characterization of the asteroid population across its full orbital state space.

Asteroid size and composition are two key properties that serve as tracers of both primordial conditions and evolutionary pathways. Larger asteroids are generally more resilient to collisional disruption and less susceptible to orbital changes induced by the size-dependent Yarkovsky effect. Compositional types, typically determined from spectroscopic observations, are believed to broadly reflect the temperature and material gradients present in the protoplanetary disk during their forma-

tion, although subsequent dynamical mixing has significantly altered the initial distribution. While general trends, such as the prominent compositional gradient transitioning from S-type dominated populations in the inner belt to C-type dominated populations in the outer belt, have been established, a comprehensive and quantitative mapping of the *joint* distribution of asteroid size and composition as a function of orbital elements (semimajor axis, eccentricity, and inclination) is essential to reveal finer structures and rigorously test theoretical models of solar system evolution. The difficulty lies not only in the complexity of the underlying physics but also in analyzing a large, potentially incomplete, and heterogeneous astronomical dataset to extract statistically significant patterns amidst inherent observational biases and stochastic effects.

In this paper, we address this challenge by conducting a detailed, data-driven analysis of the spatial architecture of the main asteroid belt. Leveraging a large dataset of over 35,000 main-belt asteroids with measured orbital elements, estimated diameters, and compositional classifications, we systematically map how asteroid size distribution and spectral types vary across the multi-dimensional orbital parameter space. Our approach employs a suite of state-of-the-art statistical and machine learning techniques, including one- and two-dimensional binning, Kernel Density Estimation to visualize density variations, unsupervised clustering (DBSCAN, Gaussian Mixture Models) to identify distinct dynamical groups, and predictive modeling (regression and classification) to quantify the relationship between orbital location and asteroid properties. These methods allow us to move beyond simple univariate or pairwise trends and explore the intricate correlations between physical properties and the full orbital state.

The insights derived from this comprehensive mapping are then interpreted in the context of known dynamical processes and theories of asteroid belt evolution. By examining the properties of asteroids within and adjacent to prominent resonant regions, analyzing the characteristics of identified dynamical families, and quantifying the gradients across major compositional zones, we seek to understand how processes such as resonant depletion, collisional evolution, and non-gravitational forces have sculpted the asteroid belt into its present-day configuration. The robustness of our findings is assessed through the application of multiple analytical techniques and the statistical significance of the observed patterns across our large sample. This work provides a detailed, quantitative map of the asteroid belt's size, compositional, and dynamical architecture, offering a foundation for refining our understanding of the solar system's formation and evolutionary history.

2. METHODS

2.1. *Data Acquisition and Preparation*

The dataset used in this study was compiled from multiple sources, representing a comprehensive collection of physical properties and orbital elements for main-belt asteroids. The raw data were provided in separate CSV files, each containing a specific property linked by a unique asteroid identifier. These files included data on asteroid name, diameter, semimajor axis, eccentricity, inclination, argument of periapsis, longitude of ascending node, spin period, obliquity, spectral type, asteroid family membership, and estimated age.

Initial data processing involved loading each CSV file into a pandas DataFrame, ensuring the asteroid identifier column was correctly recognized. A master DataFrame was then constructed by sequentially merging these individual DataFrames using an outer join on the asteroid identifier. This approach preserved all asteroids present in any of the source files, allowing for subsequent analysis of data availability for different properties. The resulting merged DataFrame served as the primary working dataset.

Following the merge, an initial cleaning step was performed. Data types for each column were verified and corrected where necessary, ensuring numerical properties were stored as floating-point numbers and categorical properties as strings or objects. Common placeholders for missing values (e.g., -999, 'NaN' strings) were identified and converted to the standard 'numpy.nan' representation.

2.2. *Exploratory Data Analysis and Data Refinement*

Exploratory Data Analysis (EDA) was conducted to understand the distributions of variables, identify patterns of missing data, and detect potential outliers. For continuous variables (diameter, semimajor axis, eccentricity, inclination, argument of periapsis, longitude of ascending node, spin period, obliquity, age), descriptive statistics (count, mean, median, standard deviation, min, max, quartiles) were computed. Histograms and box plots were generated to visualize their distributions and identify skewness or the presence of outliers. The number and percentage of missing values were recorded for each variable.

For categorical variables (spectral type, family name), value counts were calculated to understand the frequency of different categories. Bar charts were used for visualization, and the proportion of missing values was determined.

Based on the EDA, a strategy for handling missing data was implemented. For the core analysis focusing on the joint distribution of size, composition, and orbital elements, we required complete data for diameter, semimajor axis, eccentricity, inclination, and spectral type. Records with missing values in any of these five key variables were removed from the dataset using listwise deletion, creating a refined DataFrame, 'df_analysis', specifically for this analysis. The number of records removed during this filtering step was documented. Auxiliary variables with missing data were handled on a case-by-case basis in specific sub-analyses where they were relevant.

Outlier treatment focused on ensuring the dataset represented the main asteroid belt population. Asteroids with orbital elements (semimajor axis, eccentricity, inclination) falling outside typical main-belt bounds (specifically, semimajor axis between 2.0 AU and 3.5 AU, eccentricity less than 0.4, and inclination less than 30 degrees) were excluded from 'df_analysis', while noting their characteristics for potential future study. Diameter values were examined, and while extreme values were noted, they were generally retained after a log-transformation. Asteroid diameters were log-transformed ($log\_diameter = log\_10$(diameter_km)) for analysis involving size, as asteroid size distributions are often better represented on a logarithmic scale.

Bivariate analysis included generating scatter plot matrices for the key continuous variables (semimajor axis, eccentricity, inclination, and log_diameter) to visualize pairwise correlations. The Pearson correlation matrix was calculated for these variables. Relationships between orbital elements and spectral types were explored using grouped box plots (e.g., distribution of semimajor axis for each spectral type).

For subsequent modeling tasks, spectral types were grouped. Minor spectral types occurring in less than 1% of the population were consolidated into an 'Other' category to manage the number of classes for classification models.

2.3. *Analysis of Spatial Distributions and Gradients*

To quantitatively map the distribution of asteroid size and composition across orbital space, the orbital parameter space was discretized using bins. Semimajor axis was binned with a width of 0.1 AU, eccentricity with a width of 0.05, and inclination with a width of 2.5 degrees. These bin widths were selected based on the data density observed in the EDA to ensure a sufficient number of asteroids in most bins for meaningful statistical analysis.

2.3.1. *Diameter Gradients*

Within each defined bin of semimajor axis, the mean, median, and standard deviation of both 'log_diameter' and the raw diameter were calculated. The number of asteroids in each bin was also recorded. These statistics were plotted as a function of semimajor axis to visualize the size gradient. The statistical significance of the variation in mean/median diameter across semimajor axis bins was assessed using appropriate tests (e.g., ANOVA or Kruskal-Wallis test, depending on the normality of the diameter distribution within bins).

This analysis was extended to two-dimensional bins combining semimajor axis with eccentricity and semimajor axis with inclination. For these 2D bins, the average 'log_diameter' was calculated and visualized using heatmaps, where color intensity represented the average size in that region of orbital space.

2.3.2. *Spectral Type Gradients*

For each semimajor axis bin, the proportion of major spectral types (specifically focusing on the most common types like C, S, M, V, and the 'Other' category) was calculated. These proportions were plotted as stacked bar charts or line plots against semimajor axis to illustrate the compositional gradient. Chi-squared tests were performed to determine if the distribution of spectral types varied significantly across semimajor axis bins.

Similar to the diameter analysis, compositional gradients were also explored in 2D bins (semimajor axis vs. eccentricity, semimajor axis vs. inclination). Heatmaps were generated to visualize these distributions, with color representing the dominant spectral type or the proportion of a specific type within each 2D bin.

2.4. *Advanced Statistical and Machine Learning Analyses*

More sophisticated techniques were employed to reveal fine-scale structures and quantify relationships within the asteroid belt's architecture.

2.4.1. *Kernel Density Estimation*

Kernel Density Estimation (KDE) was applied to pairs of orbital parameters (semimajor axis vs. eccentricity, semimajor axis vs. inclination, eccentricity vs. inclination) to visualize the density distribution of asteroids in these 2D orbital planes, providing a smoother representation than simple binning. The 'sklearn_neighbors.KernelDensity' implementation was used. Bandwidth optimization for the KDE models was performed using cross-validation techniques (specifically, employing 'GridSearchCV' to search for the optimal bandwidth parameter). Contour plots were generated from the fitted KDE models to illustrate density variations. These density plots were then inte-

grated with information about asteroid properties by overlaying contours colored according to the average 'log_diameter' or dominant spectral type in the corresponding orbital regions. Alternatively, separate KDEs were computed for major spectral types to directly compare their spatial distributions.

2.4.2. *Clustering in Orbital Parameter Space*

Unsupervised clustering algorithms were applied to the 3D orbital parameter space defined by semimajor axis, eccentricity, and inclination to identify dynamical groupings. Prior to clustering, these orbital elements were standardized (mean-centered and scaled to unit variance) to ensure that distance calculations were not dominated by the variable with the largest range.

Two clustering algorithms were used:

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** This algorithm was chosen for its ability to find clusters of arbitrary shape and identify outliers as noise. The key parameters, 'eps' (maximum distance between two samples for one to be considered as in the neighborhood of the other) and 'min_samples' (number of samples in a neighborhood for a point to be considered as a core point), were explored. Heuristics, such as analyzing the k-distance graph, guided the selection of appropriate 'eps' values.
- **Gaussian Mixture Models (GMM):** GMM models assume the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. This approach provides probabilistic assignments of asteroids to clusters. The optimal number of Gaussian components was determined by fitting models with varying numbers of components and evaluating them using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

For the clusters identified by both methods, the mean orbital parameters were calculated. The distribution of 'log_diameter' (mean, median, standard deviation) and the proportions of different spectral types were analyzed within each cluster to characterize their physical and compositional properties. The identified clusters were compared with known asteroid families (using the provided family membership data) to assess the extent to which the clustering recovered known dynamical groups. Cluster assignments, summary statistics per cluster, and visualizations (e.g., 3D scatter plots of orbital elements colored by cluster) were saved.

2.4.3. *Regression Modeling*

Predictive modeling was employed to quantify the relationship between orbital location and asteroid properties.

- **Predicting Diameter:** Regression models were developed to predict 'log_diameter' using orbital elements (semimajor axis, eccentricity, inclination) as features. Initial exploration used Multiple Linear Regression. Generalized Additive Models (GAMs) were also fitted to capture potential nonlinear relationships between orbital elements and 'log_diameter'. More complex ensemble methods, specifically Gradient Boosting Regressors (e.g., XGBoost, LightGBM), were trained for improved predictive performance. Hyperparameter tuning for these models was performed using cross-validation techniques (e.g., 'GridSearchCV', 'RandomizedSearchCV') on a training set. Model performance was evaluated on a held-out test set using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Feature importance scores or partial dependence plots were analyzed to interpret the contribution of each orbital element to the predicted diameter.
- **Predicting Spectral Type:** Classification models were built to predict the spectral type of an asteroid based on its orbital elements. This analysis focused on discriminating between the major spectral classes. Multinomial Logistic Regression and ensemble classifiers like Random Forest and Gradient Boosting Classifiers were trained. Hyperparameter tuning using cross-validation was applied to the ensemble models. Model evaluation on a held-out test set utilized metrics including accuracy, precision, recall, F1-score (macro-averaged for handling class imbalance), and confusion matrices.

Trained models, evaluation results, and diagnostic plots were saved for further analysis and interpretation.

2.5. *Linking to Dynamical Processes*

To interpret the observed patterns in the context of asteroid belt evolution, the spatial distributions and properties derived from the analyses were examined in relation to known dynamical features. Major mean-motion resonances with Jupiter (e.g., the 3:1, 5:2, 7:3, 2:1 Kirkwood gaps) and significant secular resonances were identified based on locations reported in the astrophysical literature. Windows around the semimajor axis values corresponding to these resonances were defined. The

binned statistics, KDE results, and cluster properties were then specifically analyzed within and immediately adjacent to these resonance regions. Comparisons were made between the asteroid density, average diameter, and spectral type distributions inside resonances versus in the surrounding non-resonant regions to assess the influence of resonant dynamics on the asteroid population structure.

2.6. *Computational Implementation*

All data processing and analysis steps were implemented in Python using standard scientific libraries including pandas for data manipulation, numpy for numerical operations, matplotlib and seaborn for visualization, scipy for statistical functions, and scikit-learn for machine learning algorithms. For computationally intensive tasks, particularly hyperparameter tuning of machine learning models and cross-validation, parallel processing was utilized across multiple CPU cores to accelerate computation. Code was structured modularly, and intermediate results and plots were saved in an organized directory structure.

3. RESULTS

3.1. *Global properties and bivariate correlations of the main-belt population*

The initial data collection process yielded a large dataset, which was subsequently refined to a high-quality sample of 35,623 main-belt asteroids. This refined sample includes complete information for diameter, semimajor axis, eccentricity, inclination, and spectral type, and is restricted to the typical main-belt orbital bounds ($2.0 < a < 3.5$ AU, $e < 0.4$, $i < 30°$). This stringent filtering ensures that the subsequent analysis focuses on the primary population of the main asteroid belt, minimizing potential biases from incomplete data or non-main-belt objects.

The fundamental properties of this population exhibit distinct distributions. As shown in Figure 1, asteroid diameters follow an approximately log-normal distribution, a characteristic feature of populations shaped by collisional processes. This finding supports the use of the logarithm of diameter ($\log_{10}$(diameter_km)) in subsequent analyses, as it yields a more symmetric distribution. The orbital parameters are, by definition, confined to the main-belt region, with their distributions reflecting this selection. The compositional makeup of the sample is dominated by S-type (silicaceous) asteroids, which comprise 44.8% of the population, followed by C-type (carbonaceous) at 17.8%, and X-type at 12.0%, as detailed in Figure 2. Other significant groups include V-type (7.0%) and B-type (4.6%) asteroids.

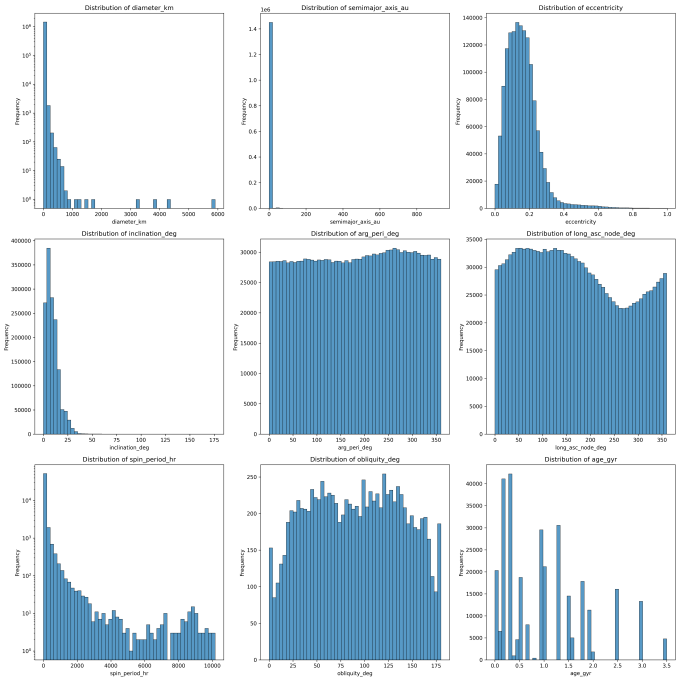

Figure 1. Univariate distributions of key physical and orbital parameters for the asteroid sample. The diameter distribution (top left) is approximately log-normal, dominated by smaller bodies. The shapes of the semimajor axis, eccentricity, and inclination distributions (top row) are defined by the sample's confinement to the main belt.

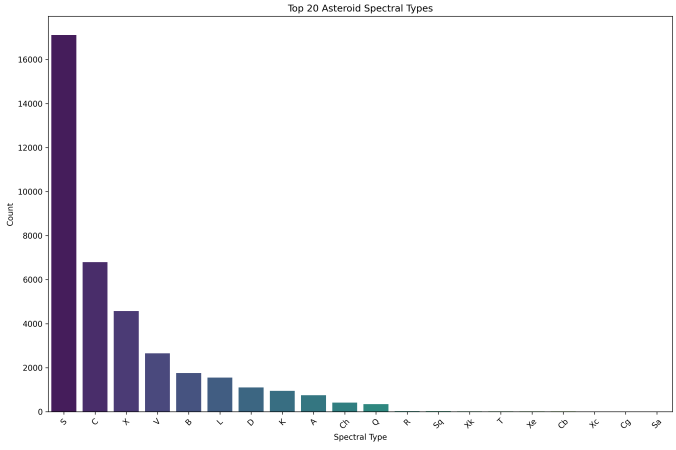

Figure 2. Distribution of the 20 most frequent spectral types within the main-belt asteroid sample. The population is heavily skewed, with S-type (silicaceous) asteroids comprising the largest fraction, followed by C-type (carbonaceous) and X-type asteroids.

To explore the relationships between these properties, we performed a bivariate analysis, summarized in Figure 3. The diagonal panels of this plot reaffirm the univariate distributions shown previously, while the off-diagonal scatter plots reveal the belt's large-scale

structure. A moderately strong positive Pearson correlation coefficient of $r = 0.427$ is observed between $\log_{10}$(diameter_km) and semimajor axis, indicating a systematic tendency for larger asteroids to be located further from the Sun. A weaker but statistically significant positive correlation ($r = 0.251$) exists between semimajor axis and inclination, suggesting that asteroids in the outer belt tend to occupy more inclined orbits. The scatter plots also clearly show the voids corresponding to the Kirkwood gaps, which will be analyzed in detail later.

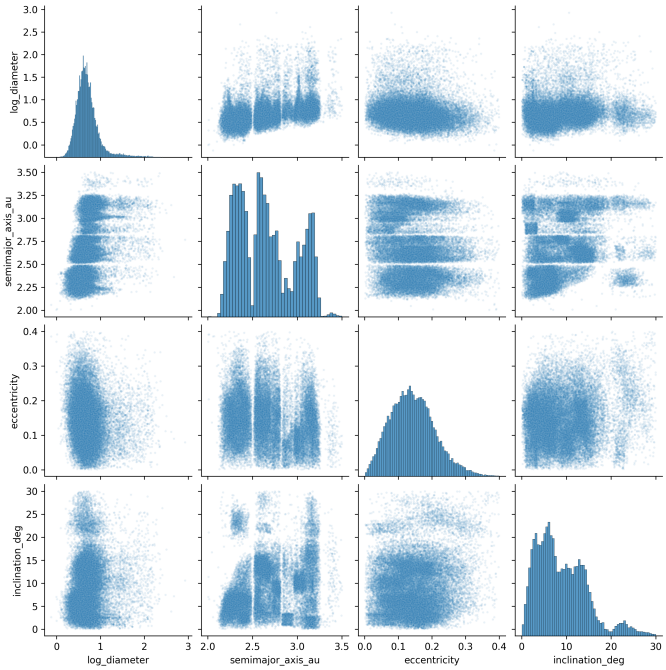

Figure 3. Bivariate analysis of key physical and orbital parameters for 35,623 main-belt asteroids. Diagonal panels show the univariate distribution of each parameter (logarithmic diameter, semimajor axis, eccentricity, and inclination). Off-diagonal scatter plots reveal the belt's structure, highlighting a positive correlation between asteroid size and semimajor axis, and the clear voids of the Kirkwood gaps.

A preliminary investigation into how composition relates to orbital location reveals a clear compositional zoning across the main belt. As shown in Figure 4, S-type asteroids are concentrated in the inner belt (low semimajor axis), while C- and X-types are predominantly found in the outer belt. V-type asteroids exhibit a distinct, compact distribution in both semimajor axis and inclination, characteristic of a collisional family. These initial findings highlight the presence of large-scale structural gradients that warrant more detailed investigation.

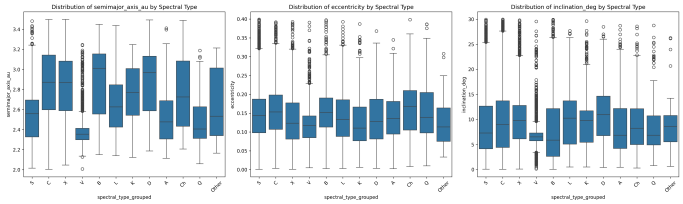

Figure 4. Distributions of orbital parameters—semimajor axis (left), eccentricity (middle), and inclination (right)—for the major asteroid spectral types. The plots reveal a clear compositional zoning across the main belt. S-type asteroids are concentrated in the inner belt (low semimajor axis), while C- and X-types are predominantly found in the outer belt. V-type asteroids exhibit a distinct, compact distribution in both semimajor axis and inclination, characteristic of a collisional family.

3.2. *Spatial gradients of asteroid size and composition*

To systematically map the asteroid belt's architecture, we analyzed the distribution of asteroid properties within discrete bins of orbital elements, as described in the methods. This approach allows for a quantitative assessment of how size and composition vary across orbital space.

3.2.1. *One-dimensional gradients along the semimajor axis*

Analysis of asteroid properties binned by semimajor axis reveals two profound and statistically significant gradients that are visualized in Figure 5. The first is a clear trend of increasing asteroid size with increasing semimajor axis. A Kruskal-Wallis H-test confirmed that the differences in $\log_{10}$(diameter_km) across semimajor axis bins are highly significant ($p \approx 0$). Quantitatively, the mean $\log_{10}$(diameter_km) increases from approximately 0.56 (∼3.6 km geometric mean diameter) in the inner belt (2.1-2.4 AU) to over 0.88 (∼7.6 km) in the outer belt (3.1-3.2 AU). This may reflect variations in protoplanetary disk conditions or be a consequence of more intense collisional evolution in the denser inner belt, which grinds bodies down to smaller sizes.

The second major gradient, also shown in Figure 5, is a striking compositional zoning. A Chi-squared test confirmed a highly significant variation in spectral type distribution with orbital distance ($p \approx 0$). The inner belt is overwhelmingly dominated by S-types (∼75% at 2.2 AU), whose proportion steadily declines to less than 10% by 3.2 AU. Conversely, C-types are rare in the inner belt (< 5% at 2.2 AU) but become the dominant class in the outer belt (∼47% at 3.2 AU). This is widely interpreted as a relic of the protoplanetary disk's thermal structure, with rocky S-types forming inside the "snow line" and primitive, carbonaceous C-types forming beyond it. The distinct peak of V-type asteroids around

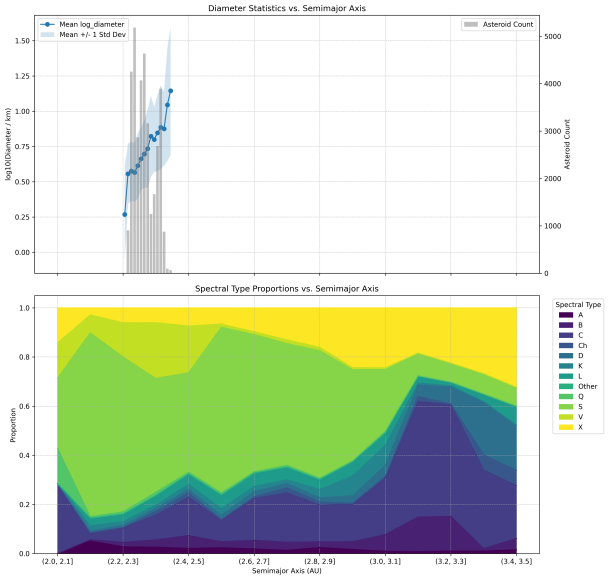

Figure 5. Physical and compositional gradients across the main asteroid belt as a function of semimajor axis. The top panel shows that the mean asteroid diameter (blue line) systematically increases towards the outer belt. The grey bars indicate the asteroid number distribution, peaking in the inner belt. The bottom panel reveals a strong compositional zoning: silicaceous S-type asteroids (light green) dominate the inner belt, while carbonaceous C-type asteroids (dark blue) become prevalent in the outer belt. These concurrent gradients reflect fundamental differences in formation and collisional evolution across the belt.

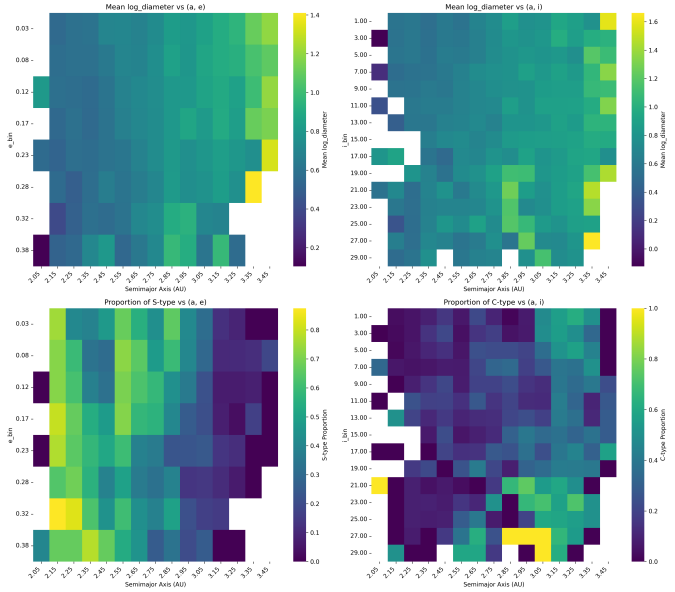

Figure 6. Two-dimensional distribution of asteroid physical properties across orbital space. The heatmaps show mean logarithmic diameter (top row), S-type proportion (bottom left), and C-type proportion (bottom right) as a function of semimajor axis versus eccentricity (left column) and inclination (right column). These visualizations reveal the main belt's large-scale structure: a clear trend of increasing asteroid size with semimajor axis, and a pronounced compositional zoning with S-types concentrated in the inner belt and C-types in the outer belt.

2.3-2.4 AU is also clearly visible, corresponding to the Vesta family.

3.2.2. *Two-dimensional distribution in orbital space*

Extending the analysis to two-dimensional bins (semimajor axis vs. eccentricity, and semimajor axis vs. inclination) provides a more detailed view of these gradients. The heatmaps in Figure 6 confirm the primary trend of increasing size with semimajor axis across all eccentricities and inclinations (top row). These maps also vividly display the compositional zoning (bottom row). S-types are predominantly located at low semimajor axes across all inclinations. C-types dominate the outer belt and show a tendency to be more prevalent at higher inclinations within that region. These visualizations highlight how the belt's fundamental gradients manifest across the full range of orbital parameters.

3.3. *Advanced structural and predictive modeling*

Moving beyond simple binning, we employed advanced statistical and machine learning techniques to uncover finer structures and quantify the relationship between orbital location and physical properties.

3.3.1. *Kernel density estimation of orbital space*

Kernel Density Estimation (KDE) provides a smooth, continuous representation of asteroid density in orbital parameter space. The resulting density maps, shown in Figure 7, reveal that the highest concentrations of asteroids are found in the inner-to-central main belt (2.2 to 2.8 AU) at low eccentricities ($e < 0.2$) and inclinations ($i < 10°$). When colored by size, the points on the KDE plots show that these dense regions are populated by small-to-intermediate-sized bodies. In contrast, the largest asteroids appear more sparsely distributed, preferentially occupying the less dense, outer regions of the belt, consistent with the observed size gradient.

3.3.2. *Unsupervised clustering: identifying dynamical groupings*

To objectively identify dynamically coherent groups, we applied unsupervised clustering algorithms to the three standardized orbital parameters. First, we used the density-based algorithm DBSCAN. The crucial neighborhood radius parameter, ϵ, was determined by examining the k-distance plot (Figure 8), where the 'elbow' indicates the optimal threshold distinguishing dense clusters from the sparser background. The DB-

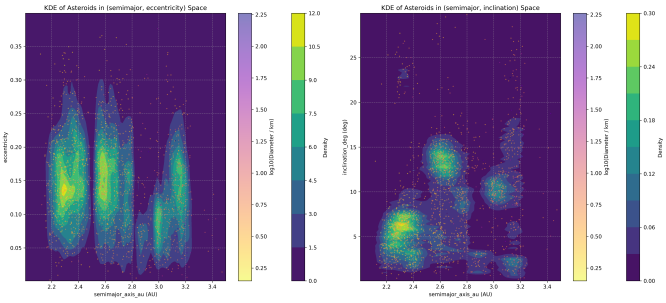

Figure 7. Kernel Density Estimation (KDE) of main-belt asteroids in semimajor axis–eccentricity (left) and semimajor axis–inclination (right) space. Color contours show asteroid density, while individual points are colored by their logarithmic diameter. The plots reveal that the highest density regions in the inner-central belt are populated by small-to-intermediate-sized bodies. In contrast, the largest asteroids (brighter points) are sparsely distributed, preferentially occupying less dense regions in the outer belt.

SCAN analysis successfully identified 38 distinct, dense clusters, with 14.25% of asteroids classified as 'noise'. This result supports the view of the belt as numerous compact families embedded within a diffuse background population.

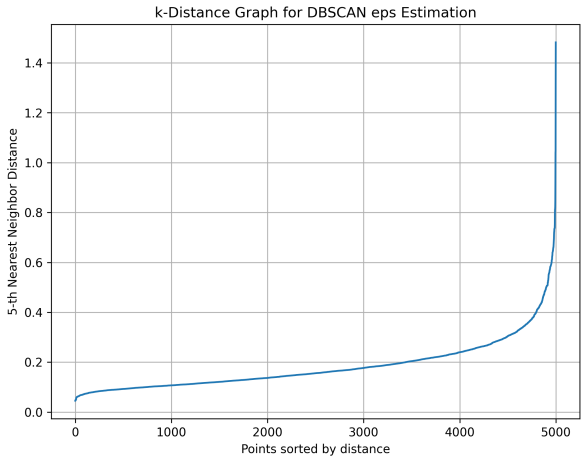

Figure 8. Determination of the neighborhood radius (ϵ) for the DBSCAN clustering algorithm. The plot shows the distance to the 5th nearest neighbor for each asteroid in the orbital parameter space, with points sorted by this distance. The distinct 'elbow' in the curve indicates the optimal distance threshold that distinguishes dense clusters, corresponding to asteroid families, from the sparser background population.

As a complementary approach, we used Gaussian Mixture Models (GMM), which model the data as a mixture of Gaussian distributions. Based on the Bayesian Information Criterion (BIC), an optimal number of 10 components was identified. These 10 clusters, visualized in 3D orbital space in Figure 9, align remarkably well with known dynamical regions and prominent asteroid families. For instance, Cluster 7 (mean $a = 2.30$ AU) is enriched in V-types (25%), corresponding to the Vesta family. Outer belt components like Cluster 6 (mean $a = 3.14$ AU) and Cluster 9 (mean $a = 3.15$ AU) are dominated by C-types (55% and 43%, respectively), likely encompassing the Themis and Hygiea families. The GMM analysis effectively partitions the belt into meaningful dynamical units, each with distinct compositional and size signatures.

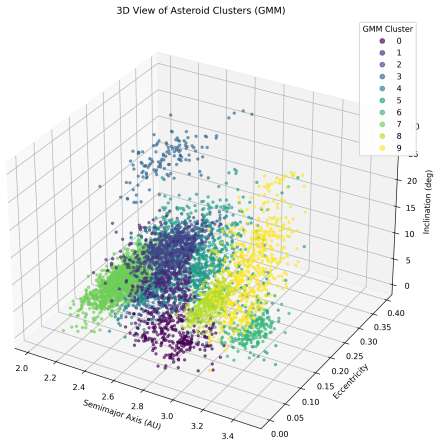

Figure 9. Gaussian Mixture Model (GMM) clustering of asteroids in 3D orbital space (semimajor axis, eccentricity, inclination). The 10 identified clusters, represented by different colors, reveal the main belt's structure as a complex tapestry of distinct, overlapping dynamical groups. These data-driven clusters correspond well with known asteroid families and compositional zones, highlighting the belt's intricate structure shaped by formation history and subsequent evolution.

3.3.3. *Predictive modeling: the limits of location*

To quantify how well an asteroid's orbital location determines its physical properties, we built two machine learning models. First, an XGBoost regression model trained to predict $\log_{10}$(diameter_km) from orbital elements achieved a coefficient of determination (R^2) of only 0.22. As shown in the feature importance analysis (Figure 10, left), all three orbital parameters contribute to the prediction. However, the low R^2 value is a critical finding: while a clear statistical trend exists at the population level, an individual asteroid's specific size is

poorly predicted by its orbit, highlighting the dominant role of its unique, stochastic collisional history.

Second, a Random Forest classification model attempting to predict spectral type from orbital elements achieved a modest overall accuracy of 53%. Feature importance analysis (Figure 10, right) shows that semimajor axis is overwhelmingly the most influential predictor. The model performed reasonably well for spatially segregated types like S-types (F1-score of 0.70), but poorly for less common or intermixed types (macro-averaged F1-score of 0.16). This indicates that while orbital location provides strong probabilistic clues about composition, it is not a definitive predictor due to significant dynamical mixing.

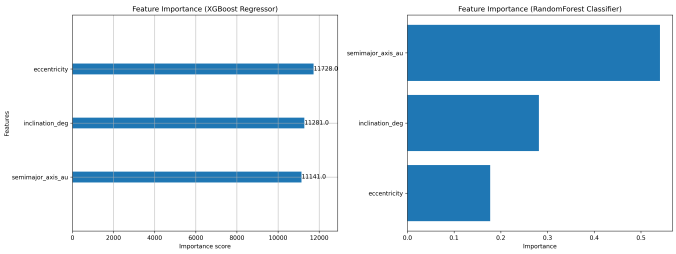

Figure 10. Relative importance of orbital parameters in predicting asteroid physical properties. *(Left)* For predicting asteroid diameter using an XGBoost regression model, all three orbital parameters—eccentricity, inclination, and semimajor axis—contribute almost equally. *(Right)* For predicting spectral type using a Random Forest classification model, semimajor axis is overwhelmingly the most important feature. This contrast highlights that while an asteroid's composition is strongly linked to its location, its size is poorly constrained by orbital parameters alone, reflecting the stochastic nature of its collisional history.

3.4. *The influence of mean-motion resonances*

The Kirkwood gaps, depleted zones associated with mean-motion resonances with Jupiter, provide a natural laboratory for studying dynamical sculpting. As expected, asteroid density drops sharply within these gaps (Figure 11, top panel). More revealingly, a Kruskal-Wallis test showed a highly significant difference ($p \approx 10^{-174}$) in the size distribution between asteroids inside, adjacent to, and far from the gaps. As seen in Figure 11, the mean $\log_{10}$(diameter_km) is higher for asteroids in regions 'Adjacent to Gaps' (0.739) and 'Inside Gaps' (0.722) compared to the 'Background' population (0.670). This indicates that smaller asteroids are preferentially cleared from resonant regions, a size-dependent filtering consistent with the Yarkovsky effect, which causes smaller bodies to drift more rapidly into unstable resonant orbits.

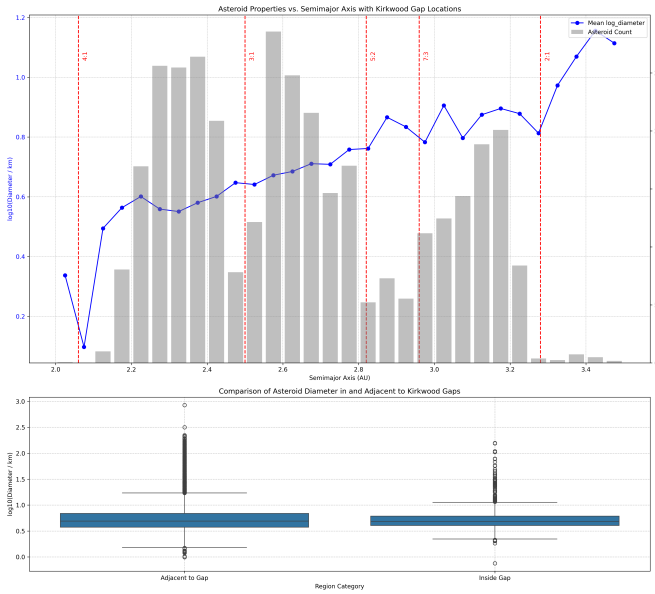

Figure 11. The influence of Kirkwood gaps on asteroid size and spatial distribution. Top: Asteroid count (bars) and mean log-diameter (line) as a function of semimajor axis, with major mean-motion resonances indicated. Bottom: Comparison of the size distribution for asteroids located inside versus adjacent to these gaps. The figure demonstrates that while asteroid density is significantly depleted within the gaps, the asteroids that remain in these dynamically unstable regions are, on average, larger, suggesting a size-dependent clearing mechanism that preferentially removes smaller bodies.

Furthermore, the compositional mix is also significantly altered near resonances (Chi-squared test, $p \approx 10^{-129}$). The proportion of X-type asteroids is significantly enhanced 'Inside Gaps' (18.2%) compared to the background (9.8%), while S-types are slightly depleted. This suggests that the dynamical clearing mechanisms are not compositionally neutral, possibly because different spectral types have different physical properties (e.g., thermal inertia) that affect their response to non-gravitational forces like the Yarkovsky effect.

3.5. *Synthesis*

This comprehensive analysis provides a detailed, data-driven view of the main asteroid belt's spatial architecture, revealing it as a complex structure shaped by multiple evolutionary processes. The most prominent features are the large-scale gradients in composition and size along the semimajor axis, reflecting initial conditions in the protoplanetary disk and subsequent collisional evolution. Superimposed on these gradients are finer structures created by gravitational dynamics, such as the sculpting effects of mean-motion resonances and the localized concentrations corresponding

to asteroid families. The analysis of resonant regions highlights their role as dynamic filters that preferentially remove smaller asteroids, likely mediated by size-dependent non-gravitational forces, and also subtly alter the local compositional mix. While population-level trends are clear and statistically significant, the limited predictive power of orbital location for the properties of individual asteroids underscores the importance of stochastic events, particularly collisions, in determining the specific characteristics of each body. The observed architecture is thus the integrated outcome of primordial formation, ongoing collisional grinding, and the long-term dynamical sculpting by planetary resonances and non-gravitational forces.

4. CONCLUSIONS

The spatial architecture of the main asteroid belt, as revealed by this comprehensive data-driven analysis, is a complex tapestry woven from primordial conditions, subsequent collisional evolution, and long-term dynamical sculpting. This study leveraged a large dataset of 35,623 main-belt asteroids with measured size, composition, and orbital elements, applying a suite of statistical and machine learning techniques to map the joint distribution of these properties in orbital space and understand the underlying physical processes.

The problem addressed is the intricate nature of the asteroid belt's structure, which holds clues to its formation and evolution. Disentangling the contributions of initial formation gradients, giant planet migration, collisions, and non-gravitational forces requires a detailed, quantitative characterization of the population. This paper aimed to provide such a characterization by systematically analyzing the distribution of asteroid size and composition across the orbital parameter space (semimajor axis, eccentricity, inclination).

Using methods including one- and two-dimensional binning, Kernel Density Estimation, unsupervised clustering (DBSCAN, Gaussian Mixture Models), and predictive modeling (regression, classification), we analyzed a refined dataset of main-belt asteroids. This allowed us to move beyond simple correlations and explore the multi-dimensional relationships between physical properties and orbital dynamics. The analysis of mean-motion resonance regions provided specific insights into the effects of dynamical filtering.

The results reveal profound structural gradients and fine-scale features. We confirmed a highly statistically significant trend of increasing asteroid size with increasing semimajor axis, suggesting either primordial differences in accretion or differential collisional processing across the belt. A stark compositional zoning was clearly mapped, showing a transition from S-type dominated populations in the inner belt to C-type dominated populations in the outer belt, a strong indicator of the thermal gradient in the protoplanetary disk during planetesimal formation. Kernel Density Estimation provided a smooth visualization of density variations, highlighting regions of high asteroid concentration and less populated zones. Unsupervised clustering successfully identified numerous distinct dynamical groups, many corresponding to known asteroid families, each exhibiting characteristic size and compositional distributions that reflect their parent body properties and formation history. Predictive modeling demonstrated that while orbital location predicts population-level trends, its ability to predict the size or spectral type of an individual asteroid is limited, underscoring the dominant role of stochastic events like collisions in shaping individual asteroid properties. Analysis of mean-motion resonance regions showed they are significantly depleted of smaller asteroids, consistent with size-dependent orbital drift mechanisms like the Yarkovsky effect, and also exhibit altered compositional mixes compared to the surrounding regions, suggesting non-uniform removal efficiencies across spectral types.

From these results, we have learned that the main asteroid belt's present-day architecture is the cumulative outcome of multiple interacting processes. The broad compositional gradient reflects initial conditions, while the size gradient likely results from a combination of initial conditions and subsequent collisional grinding, which is more efficient in the inner, denser regions. Gravitational dynamics, particularly planetary resonances, act as powerful filters, sculpting the distribution by preferentially removing objects based on their size (via non-gravitational forces) and potentially composition. The identification of numerous dynamical families through clustering highlights the ongoing importance of catastrophic collisions in creating localized, coherent groups. The limited predictability of individual asteroid properties from their current orbit emphasizes that each asteroid's history, particularly its collisional past, is a crucial determinant of its final state. This study provides a detailed quantitative map that serves as a foundation for refining models of asteroid belt formation, evolution, and the interplay between physical properties and dynamical processes over solar system history.