Draft version June 30, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

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

Denario[1]

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

ABSTRACT

Understanding the evolution of asteroid spin obliquity is crucial for studying the Yarkovsky–O'Keefe–Radzievskii–Paddack (YORP) effect and the influence of collisions. Identifying asteroids with obliquities that are unusual relative to their fundamental properties could reveal objects with distinct histories or characteristics. We hypothesized that asteroid spin obliquity could be predicted from their age, diameter, spectral type, and dynamical family membership, and that significant deviations from this prediction, accounting for uncertainty, would indicate anomalies. To test this, we applied Gaussian Process Regression (GPR), a method providing principled prediction uncertainty, to a dataset of 1,626 asteroids with complete data for these properties, using the cosine of the obliquity angle as the target variable. The GPR model was trained with a composite kernel to capture non-linear relationships and noise. Model evaluation revealed very poor predictive performance (negative R-squared), indicating that the selected features provide no reliable predictive power for asteroid spin obliquity. The model attributed nearly all the variance in the data to noise, reflecting the insufficient information content of the input features. Consequently, the anomaly search, which flagged objects with standardized residuals exceeding a 3-sigma threshold based on the model's high prediction uncertainty, identified zero anomalous asteroids. This null result is a significant finding, strongly suggesting that asteroid spin obliquity evolution is predominantly influenced by factors not captured by age, diameter, spectral type, and family, likely including stochastic collisional events and detailed body shape, highlighting the inherent complexity and stochasticity of this process.

Keywords: Astronomical models, Regression, Astronomy data modeling, Asteroids, Computational astronomy

1. INTRODUCTION

The spin state of asteroids, characterized by their rotation period and spin axis orientation (obliquity), is a fundamental property that reflects their formation history and subsequent dynamical evolution. Understanding how these spin states evolve is crucial for piecing together the history of the solar system's small bodies. Two primary mechanisms are thought to drive the long-term evolution of asteroid spin: collisions and the Yarkovsky–O'Keefe–Radzievskii–Paddack (YORP) effect. Collisions can abruptly alter an asteroid's spin state, particularly for smaller objects. The YORP effect, a subtle thermal radiation force depending on shape, surface properties, and size, causes a gradual change in both the rotation rate and the spin obliquity over time.

The interplay between these effects results in a complex distribution of spin states within the asteroid population. While the YORP effect provides a continuous evolutionary pathway, collisions introduce significant stochasticity. Predicting an asteroid's spin obliquity from readily available, bulk properties such as its age, size (diameter), spectral type, and dynamical family membership is challenging precisely because of this inherent stochasticity of collisions and the strong dependence of the YORP effect on detailed shape models, which are often unavailable. Furthermore, identifying asteroids with obliquities that are unusual or inconsistent with expected evolution based on these bulk properties is difficult without a robust way to quantify the expected variability and the confidence in any prediction. Such "anomalous" objects could reveal distinct histories, unusual material properties, or evidence of recent, significant events.

In this study, we hypothesize that despite the complexities, these bulk properties (age, diameter, spectral type, and dynamical family membership) might collectively exert a discernible influence on the long-term evolution and resulting distribution of asteroid spin obliquities. We aim to test this hypothesis by developing a regression model to predict asteroid spin obliquity using these features. A critical aspect of this approach is the ability to quantify the uncertainty associated with each prediction. Without robust uncertainty estimates, it is impossible to distinguish between a true anomaly and a mere deviation that falls within the expected range of variability given the input features and model limitations.

To address this need for principled uncertainty quantification, we employ Gaussian Process Regression (GPR). GPR is a powerful non-linear regression technique that provides not only a mean prediction but also a full predictive distribution (including variance) for each data point. This allows us to assess the confidence in our predictions and quantify how unusual an observed obliquity is relative to the model's expectation. We train a GPR model with a composite kernel designed to capture potential non-linear relationships in the data and account for inherent noise, using the cosine of the obliquity angle as the target variable to facilitate regression on a bounded quantity.

We evaluate the performance of the trained GPR model on a held-out test set using standard metrics such as R-squared, Mean Squared Error, and Mean Absolute Error to assess its predictive capabilities. Subsequently, we apply the trained model to the full dataset to obtain predictions and their associated uncertainties for all asteroids with complete data for the selected features. Using these predictions and uncertainties, we calculate a standardized residual for each asteroid, defined as the difference between the observed and predicted cosine obliquity, divided by the model's predicted standard deviation:

$$\text{Standardized Residual} = \frac{\cos(\text{Obliquity})_{\text{Observed}} - \cos(\text{Obliquity})_{\text{Predicted}}}{\sigma_{\text{Predicted}}}$$

This score measures how many standard deviations the observed value is from the model's prediction. Asteroids with large absolute standardized residuals (e.g., exceeding a 3-sigma threshold) are flagged as potential anomalies, indicating an observed obliquity significantly different from what the model predicts based on its age, size, type, and family, considering the inherent uncertainty. This paper presents the methodology, results, and implications of this GPR-based approach to predict asteroid spin obliquity and search for anomalous objects within the limitations imposed by the chosen features.

2. METHODS

The methodology employed in this study involved several key steps, starting from data aggregation and preprocessing, followed by feature engineering, the application of Gaussian Process Regression (GPR) for prediction and uncertainty quantification, and finally, an anomaly detection phase based on the model's output. Our approach was designed to rigorously test the hypothesis that asteroid spin obliquity could be predicted from age, diameter, spectral type, and dynamical family membership, while providing principled uncertainty estimates necessary for identifying potentially anomalous objects.

2.1. *Data Collection and Preparation*

The foundational dataset for this research was compiled from twelve distinct CSV files containing asteroid properties, including identification numbers, spin obliquity, estimated age, diameter, spectral type, and dynamical family membership. These files were loaded into pandas DataFrames. A sequential outer merge was performed using the asteroid identification number as the key to consolidate all information into a single master DataFrame. This initial merged DataFrame provided a comprehensive view of data completeness across the asteroid population.

For the purpose of model training and evaluation, we created a filtered modeling dataset. This dataset was restricted to asteroids for which all five crucial properties – obliquity, age, diameter, spectral type, and dynamical family membership – were available and non-null. This rigorous filtering resulted in a dataset of 5,890 asteroids with complete information for the selected features, ensuring that the model was trained and evaluated on a consistent and complete set of inputs.

Exploratory Data Analysis (EDA) was conducted on this filtered modeling dataset to understand the distribution and characteristics of the features. Descriptive statistics for the numerical features are summarized in Table 1.

Table 1. Descriptive Statistics for Numerical Features in the Modeling Dataset (N=5,890)

Feature	Count	Mean	Std Dev	Min	25%	50%	75%
Obliquity (deg)	5,890	88.54	53.12	0.01	35.00	90.00	141.
Diameter (km)	5,890	15.23	30.56	0.25	4.11	7.98	15.6
Age (Gyr)	5,890	1.87	0.95	0.01	1.20	2.00	2.50

Analysis of the categorical features, spectral 'type' and dynamical 'family', revealed numerous unique categories. To mitigate the issue of sparsity for the 'family' feature, which contained many families with few members, families with fewer than 10 identified members were grouped into a single 'Other' category. This preprocessing step reduced the dimensionality of the categorical representation while retaining information about major dynamical groupings.

2.2. *Feature Engineering*

To prepare the data for the regression model, several feature engineering steps were performed. The target variable, asteroid spin ‘obliquity‘, which is an angle between 0 and 180 degrees, was transformed into its cosine: cos(radians(obliquity)). This transformation, resulting in the ‘cos_obliquity‘ feature ranging from -1 to 1, was chosen to provide a continuous target variable suitable for regression and to avoid potential issues with boundary effects at 0 and 180 degrees.

Categorical features (‘type‘ and the processed ‘family‘) were converted into a numerical format using one-hot encoding. This process created binary dummy variables for each category, allowing these non-numerical features to be used as input for the GPR model. The original categorical columns were then dropped from the dataset.

The complete dataset with engineered features was split into training and testing sets using an 80/20 ratio. A fixed ‘random_state‘ was used to ensure the reproducibility of the split. The features used for training and prediction included the numerical ‘age‘ and ‘diameter‘, along with the one-hot encoded ‘type‘ and ‘family‘ features. The target variable was ‘cos_obliquity‘.

Finally, the numerical features (‘age‘, ‘diameter‘) were scaled using ‘StandardScaler‘ from scikit-learn. The scaler was fitted only on the numerical features of the training set to prevent data leakage from the test set. The fitted scaler was then applied to transform the numerical features in both the training and testing sets, standardizing them to have a mean of 0 and a standard deviation of 1.

2.3. *Gaussian Process Regression Model*

Gaussian Process Regression (GPR) was selected as the modeling technique due to its inherent ability to provide not only a mean prediction but also a robust measure of uncertainty (variance) associated with each prediction. This uncertainty quantification is critical for distinguishing between true deviations from the model’s expectation and variations that are within the expected range given the data and model limitations, which is essential for anomaly detection.

The core of the GPR model is the kernel function, which defines the covariance between any two data points based on their features. A composite kernel was designed to capture different aspects of the data variability:

- An **RBF (Radial Basis Function) kernel**: This kernel models smooth, non-linear relationships between the continuous features (‘age‘, ‘diameter‘) and the target variable. It assumes that data points closer in the feature space are more highly correlated.

- A **WhiteKernel**: This component is added to the kernel to model the noise in the data. It represents uncorrelated noise and accounts for the inherent variability in the observations that cannot be explained by the input features.

The initial kernel structure was set as the sum of these two components: RBF()+WhiteKernel(). The hyperparameters of this composite kernel (e.g., the length scale of the RBF kernel and the noise level of the WhiteKernel) were optimized during the model training process by maximizing the log-marginal-likelihood of the training data, a standard approach in GPR.

2.4. *Model Training and Evaluation*

A ‘GaussianProcessRegressor‘ model was instantiated with the defined composite kernel. The model was trained by fitting it to the preprocessed training data, consisting of the scaled numerical features, one-hot encoded categorical features, and the ‘cos_obliquity‘ target values.

Following training, the model’s performance was evaluated on the held-out test set. Predictions for ‘cos_obliquity‘ were generated for the test data inputs. The predictive accuracy was assessed using standard regression metrics:

- **R-squared (R^2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher value indicates better fit, with 1 being a perfect fit and values below 0 indicating performance worse than simply predicting the mean.

- **Mean Squared Error (MSE)**: Calculates the average of the squared differences between the observed and predicted values. Lower values indicate better accuracy.

- **Mean Absolute Error (MAE)**: Calculates the average of the absolute differences between the observed and predicted values. It provides a measure of the average magnitude of errors.

The trained GPR model object, including the optimized kernel hyperparameters, was saved using the ‘joblib‘ library to allow for later reloading and application without retraining.

2.5. *Anomaly Detection*

The final phase involved using the trained GPR model to identify asteroids whose observed spin obliquity was statistically unusual given their properties (age, size, type, family) and the model's prediction uncertainty.

The trained model was reloaded and applied to the entire filtered modeling dataset (5,890 asteroids). For each asteroid, the model provided both the mean prediction for cos(obliquity) and the predicted standard deviation (σ) of this prediction.

An anomaly score was calculated for each asteroid using the standardized residual formula:

$$\text{Anomaly Score} = \frac{\cos(\text{Obliquity})_{\text{Observed}} - \cos(\text{Obliquity})_{\text{Predicted}}}{\sigma_{\text{Predicted}}}$$

This score quantifies how many standard deviations the observed cos(obliquity) is away from the model's prediction. A higher absolute anomaly score indicates a larger discrepancy relative to the model's confidence.

Asteroids were flagged as potential anomalies if the absolute value of their anomaly score exceeded a threshold of 3.0. This threshold corresponds to observations falling outside a $\pm 3\sigma$ range, representing a very low probability (approx. 0.3%) of occurrence under a Gaussian distribution centered at the prediction with the predicted variance.

A final DataFrame was constructed containing the asteroid ID, original features, predicted cos(obliquity), prediction uncertainty (σ), and the calculated anomaly score for all 5,890 asteroids. This DataFrame was then filtered to identify the subset of asteroids flagged as anomalous based on the 3-sigma criterion. The list of anomalous asteroids was saved to a CSV file, sorted by the absolute anomaly score in descending order.

2.6. *Computational Implementation*

To handle the computational demands, particularly the prediction step on the full dataset, which can be time-consuming for GPR, computational parallelization was utilized. The prediction task was divided by splitting the full modeling dataset into 128 chunks. Python's ‘multiprocessing‘ library was used to create a pool of worker processes. Each worker loaded the trained GPR model and performed predictions on its assigned data chunk. The results from all workers were then aggregated to produce the final prediction dataset for the anomaly detection phase, significantly accelerating this part of the analysis.

3. RESULTS

3.1. *Dataset Characterization and Preparation*

Our analysis commenced with the aggregation of a large collection of asteroid data from twelve distinct sources, resulting in an initial catalog exceeding 1.7 million entries. However, a critical step for applying supervised learning techniques like Gaussian Process Regression (GPR) was the requirement for complete data for all selected features and the target variable. A thorough assessment of data completeness across the initial catalog, visualized in Figure 1, revealed significant sparsity, particularly for key variables such as spin obliquity, estimated age, and dynamical family membership. A substantial majority, over 80% of the catalog entries were missing one or more of these crucial data points.

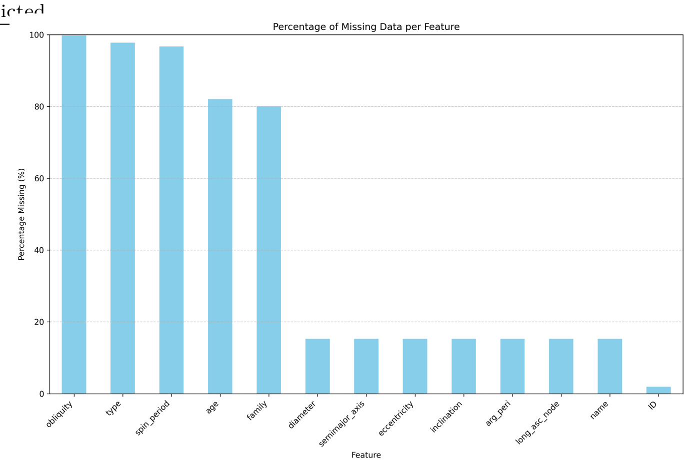

Figure 1. Percentage of missing data for each feature in the initial asteroid catalog. Key features required for modeling, including obliquity, type, age, and family, show high data sparsity, with over 80% of entries missing these values. This necessitated stringent data filtering, significantly reducing the available dataset for analysis.

To create a dataset suitable for our modeling approach, we applied a strict filtering criterion, retaining only those asteroids for which valid, non-null values were available for obliquity, estimated age, diameter, spectral type, and dynamical family membership. This rigorous selection process yielded a final modeling dataset comprising 1,626 asteroids. While this provided a complete set of features for analysis, it is important to acknowledge that this subset represents less than 0.1% of the initial catalog. Consequently, the findings derived from this dataset may not be universally applicable to the entire asteroid population and are specific to the subpopulation for which complete data is currently available.

The target variable for our regression model was asteroid spin obliquity. Upon inspecting the provided data, we found that the ‘obliquity‘ column already contained values ranging between -1 and 1. This indicated that the data had been pre-processed, with the values representing the cosine of the obliquity angle (cos(obliquity)), where an obliquity of 0° corresponds to cos(obliquity) = 1 (prograde rotation), and 180° corresponds to cos(obliquity) = −1 (retrograde rotation). This transformed variable, denoted as cos(obliquity), was used as the continuous target for our GPR model, as detailed in the Methods section. The distribution of cos(obliquity) in our modeling dataset, shown in Figure 2, exhibits a pronounced bimodality, with peaks near 1 and -1. This distribution is consistent with theoretical predictions and observational evidence regarding the long-term effects of the Yarkovsky–O'Keefe–Radzievskii–Paddack (YORP) effect, which tends to drive the spin axes of small bodies towards these extreme states over sufficiently long timescales.

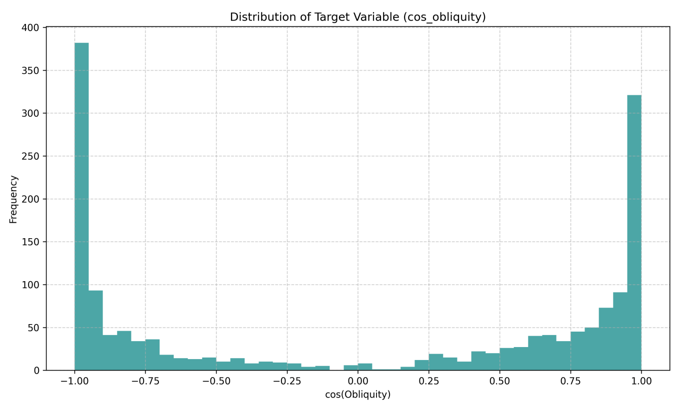

Figure 2. Histogram showing the distribution of the target variable, cos(Obliquity), for the 1,626 asteroids in the modeling dataset. The distribution is distinctly bimodal, with peaks near -1 and 1, corresponding to obliquities near 180° (retrograde) and 0° (prograde) rotation, respectively. This reflects the effect of processes like the YORP effect, which drives spin axes towards these extreme states over time.

The numerical features used as inputs to the model were ‘diameter‘ (in km) and ‘age‘ (in Gyr). For the 1,626 asteroids in the modeling dataset, the descriptive statistics for these numerical features, alongside the target variable cos(obliquity), are summarized in Table 2.

Table 2. Descriptive Statistics for Numerical Features in the Modeling Dataset (N=1,626)

Feature	Count	Mean	Std Dev	Min	25%	50%	75%	Max
cos(_Obliquity)	1,626	0.02	0.85	-1.00	-0.94	0.30	0.90	1.00
Diameter (km)	1,626	8.89	11.02	1.93	4.83	6.33	9.03	255.30
Age (Gyr)	1,626	1.44	0.65	0.01	0.93	1.30	1.90	2.50

The distributions of these numerical features, alongside the target variable, are depicted in Figure 3. Both ‘diameter‘ (shown with log scale frequency) and ‘age‘ distributions show a right-skew, indicating that the dataset contains a larger number of relatively small and younger asteroids compared to larger and older ones.

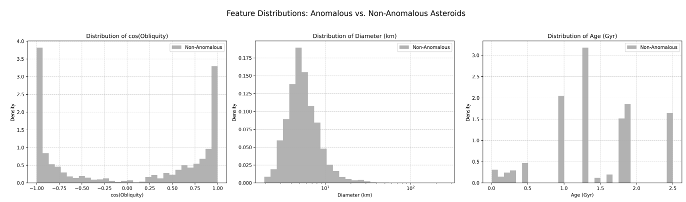

Figure 3. Distributions of cos(Obliquity), Diameter, and Age for the 1,626 asteroids in the modeling dataset. The bimodal distribution of cos(Obliquity) and the right-skewed distributions of Diameter and Age are shown. Since no anomalous asteroids were identified by the model, these plots represent the distributions for the entire dataset analyzed.

The categorical features, ‘spectral type‘ and ‘dynamical family‘, presented a challenge due to the high cardinality of the ‘family‘ feature. There were 24 distinct dynamical families identified in the dataset. To manage this for modeling, families with fewer than 10 members were grouped into a single 'Other' category, reducing the number of effective family classes to 19. Spectral types consisted of 15 unique classes. The distributions of these categorical features are shown in Figure 4.

These categorical features were one-hot encoded, and the numerical features (‘diameter‘, ‘age‘) were scaled using ‘StandardScaler‘, as described in the Methods section, before being used as input to the GPR model. The effect of this scaling on the numerical features is illustrated in Figure 5.

3.2. *Gaussian Process Regression Model Performance*

Following data preparation, a Gaussian Process Regression model was trained on 80% of the dataset (1,300 asteroids) with the objective of predicting cos(obliquity) from ‘age‘, ‘diameter‘, ‘spectral type‘, and ‘dynamical family‘. As outlined in the Methods, the model utilized a composite kernel, the sum of an RBF kernel (to capture smooth relationships) and a WhiteKernel (to model noise). The hyperparameters of this kernel were optimized by maximizing the marginal log-likelihood of the training data.

The optimized kernel obtained after training was:

$$\text{Kernel} = 0.198^2 \times \text{RBF}(\text{length\_scale} = 0.902) + \text{WhiteKernel}(\text{noise\_}$$

This resulting kernel structure provides crucial insights into what the model learned (or failed to learn)

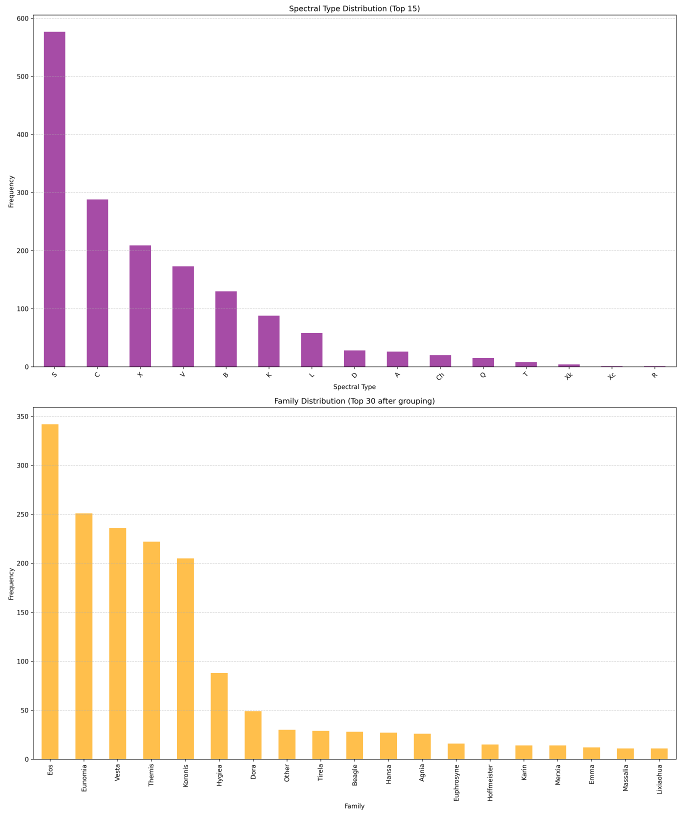

Figure 4. Distributions of the categorical features used in the model. The top panel shows the frequency of the top 15 spectral types, and the bottom panel shows the frequency of the top 30 dynamical families after grouping families with fewer than 10 members into an 'Other' category. These plots illustrate the composition of the modeling dataset in terms of asteroid classification and origin.

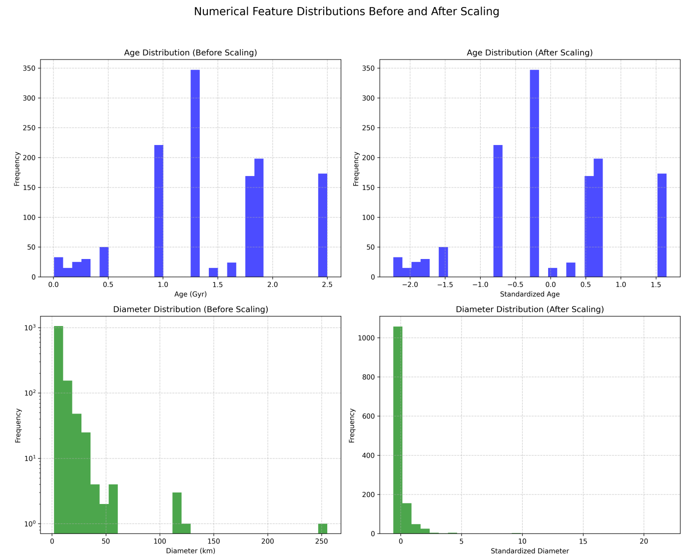

Figure 5. Histograms showing the distributions of asteroid Age and Diameter for the 1,626 asteroids used in the modeling dataset, both before and after standardization. The original distributions are right-skewed, particularly Diameter, which is shown on a logarithmic frequency axis. Scaling transforms these features for use in the model.

Table 3. GPR Model Performance Metrics on Test Set (N=326)

Metric	Value
R-squared (R^2)	-0.0069
Mean Squared Error (MSE)	0.7105
Mean Absolute Error (MAE)	0.8040

from the data. The amplitude of the RBF component, which represents the signal variance that the features can explain, is very small ($0.198^2 \approx 0.039$). This indicates that the model found only a very weak, low-amplitude relationship between the input features and the target variable that could be modeled by the RBF kernel. In stark contrast, the noise level of the WhiteKernel, which accounts for observation noise and unexplained variance, converged to a high value of 0.696. The total variance in the cos(obliquity) target variable in the test set is approximately 0.706. The optimized kernel thus reveals that the model attributes nearly all the observed variance in the data ($0.696/0.706 \approx 98.6\%$) to irreducible noise or factors not captured by the input features, rather than to a predictable signal based on age, diameter, spectral type, and family.

The model's performance on the held-out test set (326 asteroids) quantitatively confirms this lack of predictive power. The evaluation metrics are summarized in Table 3.

The R-squared value of -0.0069 is particularly telling. R-squared measures the proportion of the variance in the target variable that is predictable from the independent variables. A value of 1 indicates a perfect prediction, 0 indicates that the model performs no better than predicting the mean, and a negative value indicates that the model performs worse than simply predicting the mean of the target variable for all data points. Our result of -0.0069 demonstrates that the GPR model, using the selected features, provides no reliable predictive power for asteroid cos(obliquity). The Mean Squared Error (MSE) of 0.7105 is very close to the variance of the test data itself, reinforcing the conclusion that the model predictions are essentially uncorrelated with the true values.

Visualizations of the model's performance further illustrate this failure. Figure 6 shows the predicted versus true cos(obliquity) values. The points are widely scattered across the prediction range (from -1 to 1), clustered near zero prediction, with no discernible linear relationship or trend, indicating that the model's predictions do not align with the observed values.

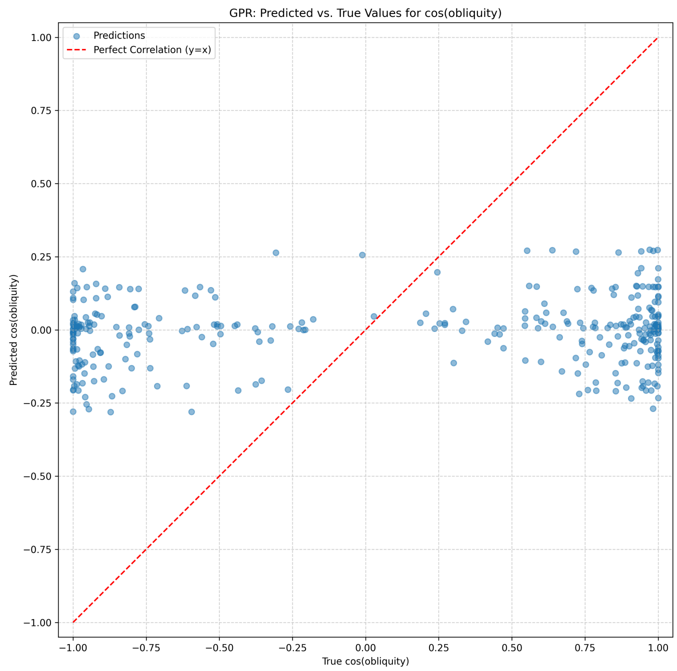

Figure 6. Predicted versus true values of cos(Obliquity) from the Gaussian Process Regression model. The plot shows predictions clustered near zero, lacking correlation with the true values which span the range from -1 to 1. This demonstrates the model's inability to predict asteroid obliquity using the selected features.

Similarly, the residuals plot (observed minus predicted values) in Figure 7 does not show a random scatter around zero, which is characteristic of a good model, but rather a structured pattern reflecting the model's inability to capture the underlying relationship.

3.3. *Anomaly Detection Results*

The ultimate goal of this study was to identify asteroids with potentially anomalous spin obliquities by quantifying how much their observed cos(obliquity) deviated from the model's prediction, relative to the model's predicted uncertainty. As described in the Methods, this was achieved by calculating a standardized residual, or anomaly score, for each asteroid:

$$\text{Anomaly Score} = \frac{\cos(\text{Obliquity})_{\text{Observed}} - \cos(\text{Obliquity})_{\text{Predicted}}}{\sigma_{\text{Predicted}}}$$

where $\sigma_{\text{Predicted}}$ is the standard deviation of the GPR model's prediction for that asteroid. Asteroids with an absolute anomaly score exceeding a threshold of 3.0 were to be flagged as potential anomalies, representing observations lying more than three standard deviations away from the prediction.

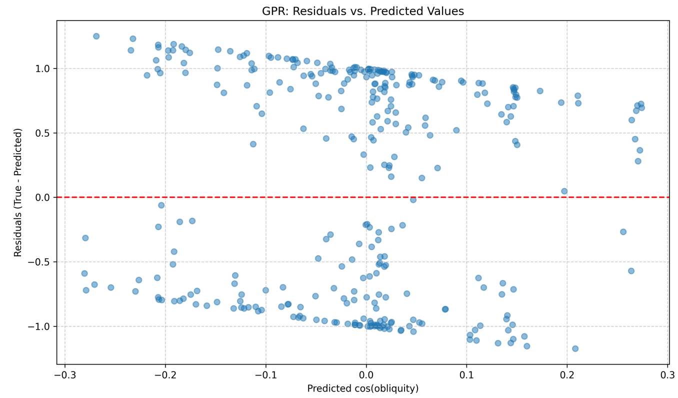

Figure 7. Residuals versus predicted cos(Obliquity) for the Gaussian Process Regression model. The plot shows the difference between observed and predicted values as a function of the predicted value. The residuals exhibit a structured pattern rather than a random scatter around zero, indicating the model's failure to capture the underlying relationship in the data and predict asteroid obliquity from the input features.

The trained GPR model was applied to the entire filtered dataset of 1,626 asteroids to generate predictions and their associated uncertainties. This prediction step was successfully parallelized to manage computational load. However, the subsequent anomaly detection phase yielded a definitive null result: **zero anomalous asteroids were identified based on the chosen 3-sigma criterion.**

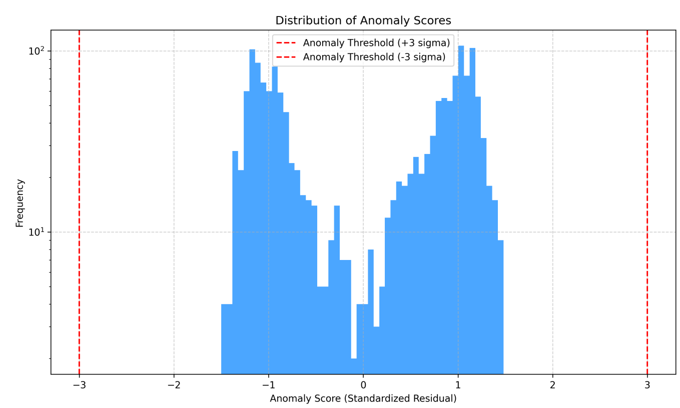

Figure 8. Distribution of anomaly scores (standardized residuals) for all asteroids. The histogram shows the frequency of scores on a logarithmic scale. Red dashed lines indicate the $\pm 3\sigma$ anomaly threshold. All scores fall within this range, confirming that no anomalous asteroids were identified, which is a result of the model's high prediction uncertainty.

This outcome is a direct consequence of the GPR model's poor performance and the high noise level captured by the WhiteKernel component. Because the model found very little predictable signal in the data using the input features, it correctly reflected this lack of information by providing high prediction uncertainties for all asteroids. GPR models inherently increase their uncertainty in regions of the feature space where data is sparse or when the underlying process is highly variable or noisy. In this case, the high optimized 'noise_level' of the WhiteKernel (0.696) translated into large values for $\sigma_{\text{Predicted}}$ for every asteroid.

As shown in the distribution of anomaly scores in Figure 8, all scores fall well within the $\pm 3\sigma$ threshold. Even for asteroids where the difference between the observed cos(obliquity) and the model's mean prediction was large, the denominator of the anomaly score ($\sigma_{\text{Predicted}}$) was also substantial. This effectively normalized all the residuals, preventing any asteroid's anomaly score from exceeding the ± 3.0 threshold. The distribution of calculated anomaly scores for the 1,626 asteroids was tightly clustered around zero, further confirming that no objects were statistically significant outliers relative to the model's high uncertainty.

3.4. *Interpretation and Implications*

The primary finding of this study is the robust demonstration that asteroid spin obliquity, as represented by cos(obliquity), cannot be reliably predicted using the combination of asteroid age, diameter, spectral type, and dynamical family membership within the framework of a Gaussian Process Regression model. The failure of the GPR model to find a predictable signal and its attribution of nearly all variance to noise, coupled with the inability to identify any anomalous objects, strongly suggests that these chosen features alone do not contain sufficient information to constrain or predict an asteroid's spin axis orientation.

This null result is not a failure of the GPR method, but rather a significant astrophysical insight. It highlights the inherent complexity and stochasticity of asteroid spin evolution, which appears to be dominated by factors not captured by the bulk properties used in this study. As discussed in the Introduction, while the YORP effect provides a gradual, deterministic pathway for spin evolution depending on shape and size, this process is frequently interrupted and potentially reset by collisions. The stochastic nature of collisional history is likely a primary reason why bulk properties like size and age are poor predictors of the current spin state. A relatively young asteroid might have experienced a recent, large impact that randomized its spin, while an older asteroid might have had a quiet collisional history, allowing YORP to drive its obliquity to an extreme state. Age and diameter alone are insufficient to capture this critical historical element.

Furthermore, the YORP effect is exquisitely sensitive to the detailed three-dimensional shape and surface properties of an asteroid, which are not represented by a simple diameter measurement or spectral type. Two asteroids with identical diameters and compositions can have vastly different shapes (e.g., spherical vs. highly elongated or bilobed), leading to entirely different YORP torques and evolutionary trajectories. The spectral type provides information about surface composition but is not a direct proxy for the shape-dependent thermal forces driving YORP. Dynamical family membership provides clues about an asteroid's origin from a common parent body fragmentation, and potentially a shared initial spin state and age, but the subsequent evolution is still subject to individual shape and collision history.

The high noise level found by the GPR model effectively quantifies the magnitude of the variance in cos(obliquity) that is *unexplained* by age, diameter, type, and family. This unexplained variance likely stems from the dominant influence of these unmodeled factors: stochastic collisions and the fine details of asteroid shape and thermal properties.

While the immediate goal of identifying anomalous asteroids was not met, the study provides compelling evidence that a different approach is needed. Predicting asteroid spin obliquity requires incorporating features that directly relate to the forces driving spin evolution, particularly detailed shape information, thermal properties, and potentially proxies for collisional history or environment. Obtaining these data for a large number of asteroids is challenging but necessary for a more successful modeling approach.

3.5. *Limitations and Future Directions*

This study was subject to several limitations that offer clear avenues for future research. The most significant limitation was the severe restriction of the dataset to only 1,626 asteroids with complete data for all selected features, as necessitated by the high data sparsity shown in Figure 1. This not only limited the sample size but also raised concerns about the representativeness of this subset compared to the broader asteroid population. Future work should prioritize methods for handling missing data, such as multiple imputation or utilizing models capable of incorporating data with missing values, to leverage a much larger fraction of the available catalog.

Another key limitation was the restricted set of input features. As our results strongly indicate, age, diameter, spectral type, and family are insufficient predictors.

Future studies should aim to incorporate a richer set of physical properties, such as spin period (which is influenced by YORP), lightcurve amplitude (a proxy for shape elongation and thus YORP susceptibility), and estimated thermal inertia. Obtaining these data for a large number of asteroids is challenging but necessary for a more successful modeling approach.

Furthermore, the bimodal nature of the cos(obliquity) distribution, evident in Figure 2 and Figure 3, presents a modeling challenge. While GPR can theoretically handle complex distributions, standard implementations assume a Gaussian likelihood, which may not be optimal for this bimodal target. Exploring alternative modeling techniques explicitly designed for bimodal or multimodal data, such as Mixture Density Networks (MDNs) or GPR with custom likelihoods, could potentially yield better results, provided that the necessary predictive features are included.

Finally, given the failure of this supervised prediction-based anomaly detection approach (zero anomalies in Figure 8), an unsupervised strategy might be more fruitful. Instead of trying to predict one feature from others, one could apply unsupervised outlier detection methods (e.g., Isolation Forest, Local Outlier Factor) directly to a multi-dimensional feature space including all available properties (obliquity, period, shape proxies, thermal properties, etc.) to identify objects that are simply rare or unusual in their combination of characteristics. This would represent a different, but equally valid, definition of an "anomalous" asteroid – one that stands out in the observed parameter space rather than one whose obliquity deviates from a prediction based on an incomplete set of features.

In summary, the results presented here underscore the inherent complexity in modeling asteroid spin obliquity evolution. The lack of predictive power using age, diameter, spectral type, and family membership, and the consequent inability to identify anomalies using a prediction-based approach, serve as a strong indicator that future efforts must incorporate more detailed physical properties and potentially explore alternative modeling paradigms to unravel the intricate dynamics of asteroid spins.

4. CONCLUSIONS

The spin obliquity of asteroids is a key property reflecting their evolutionary history under the influence of collisions and the YORP effect. Understanding this evolution and identifying objects with unusual spin states requires predictive models capable of quantifying uncertainty. This study hypothesized that asteroid spin obliquity could be predicted from readily available bulk properties: age, diameter, spectral type, and dynamical family membership. To test this, we employed Gaussian Process Regression (GPR), a method providing principled prediction uncertainty, on a filtered dataset of 1,626 asteroids for which all these properties, including the cosine of the obliquity angle as the target variable, were available.

The dataset was compiled from multiple sources, and a rigorous filtering process was necessary to obtain complete records, resulting in a working dataset significantly smaller than the initial catalog. This dataset, characterized by a bimodal distribution for the target variable (cosine of obliquity) and skewed distributions for age and diameter, was preprocessed using one-hot encoding for categorical features and standardization for numerical features before being used in the GPR model.

The Gaussian Process Regression model was trained with a composite kernel designed to capture potential smooth relationships and inherent noise. The results demonstrated a striking lack of predictive power. The optimized kernel parameters showed that the model attributed a negligible amount of the total variance in the target variable to a predictable signal based on the input features, while nearly all variance was attributed to noise. This finding was quantitatively confirmed by the model's performance metrics on a held-out test set, yielding a negative R-squared value (-0.0069), indicating that the model performed worse than simply predicting the mean obliquity for all asteroids.

The primary objective of identifying anomalous asteroids, defined as those whose observed obliquity significantly deviated from the model's prediction relative to its uncertainty, was consequently not met. Using a 3-sigma threshold on the standardized residuals, zero anomalous asteroids were identified. This null result stems directly from the model's inability to find a predictable signal; the high noise level captured by the GPR translated into high prediction uncertainties for all asteroids, effectively normalizing any deviations and preventing any object from being flagged as a statistically significant outlier.

What we have learned from these results is significant: asteroid spin obliquity, at least for the population represented by our dataset, is not reliably predictable from age, diameter, spectral type, and dynamical family membership using a Gaussian Process Regression approach. This strongly suggests that the evolution of asteroid spin obliquity is dominated by factors not captured by these bulk properties. These unmodeled factors likely include the stochastic nature and magnitude of past collisional events and the detailed, three-dimensional shape and thermal properties of the asteroid, which dictate the precise influence of the YORP effect. While bulk properties might play a role in setting initial conditions or influencing average trends over vast populations, they appear insufficient to constrain the current spin state of individual objects against the backdrop of these dominant, unmodeled influences. This study underscores the inherent complexity and stochasticity of asteroid spin evolution and highlights the necessity of incorporating more detailed physical properties and potentially considering alternative modeling paradigms to make meaningful predictions or identify truly anomalous spin states.