Au20 Gold cluster - Energy Modeling and Structural Analysis of Gold Clusters

Wilbert Davis Tan, Xiao Xuan Oh, and Bryan Wijaya

1. Introduction

This study presents a comprehensive machine learning framework for predicting the binding energies of Au_{20} nanoclusters by systematically integrating geometric descriptors with Smooth Overlap of Atomic Positions (SOAP) features, which capture local atomic environments in a rotationally and translationally invariant manner. A dataset of 999 Au_{20} cluster configurations was analyzed, revealing that most clusters are compact, highly coordinated, and exhibit uniform bonding patterns, indicative of stable structures.

To evaluate predictive performance, three categories of models were employed: linear regression variants (Ridge, Lasso, Elastic Net), kernel-based methods (SVR with RBF kernel), and advanced tree-based ensemble models (XGBoost, LightGBM, Gradient Boosting, and Random Forest). Among these, tree-based ensembles consistently outperformed linear and kernel models, with XGBoost achieving a test R^2 of 0.92 and RMSE of 0.82 eV, representing a 15–19% improvement over linear baselines ($R^2 \approx 0.77$). Kernel-based methods appeared to perform well on training data but failed cross-validation tests, suggesting memorization rather than generalizable learning. Structural analysis identified pyramid-like morphologies as the most stable configurations, characterized by 60 bonds, an average coordination number of 6.0, and highly uniform bond lengths (2.666 ± 0.041 Å), with the lowest energy per atom measured at –77.860 eV. Perturbation studies confirmed model robustness, with XGBoost exhibiting low sensitivity (0.272 ± 0.227 eV/Å) under atomic displacements up to 0.30 Å, although all tree models displayed signs of overfitting (training $R^2 \approx 1.00$ vs. test $R^2 = 0.88–0.92$), highlighting the need for stronger regularization strategies.

Overall, this framework demonstrates that SOAP-enhanced feature representations enable accurate and physically interpretable predictions for novel Au_{20} cluster configurations, providing valuable insights into the structural factors governing cluster stability and guiding the rational design of nanoscale materials.

22. Predicting Energies of Au_{20} Clusters

This study presents a workflow for predicting molecular energies using structured preprocessing, feature extraction, and model evaluation. Molecular data in .xyz format was parsed and validated, with inconsistencies addressed and the dataset organized into three categories: good, balanced, and elite, to ensure quality and representativeness. A brief statistical summary and visualization of sample structures provided further understanding of the dataset.

To generate meaningful inputs for modeling, structural descriptors were extracted to capture physically relevant patterns linked to molecular energy. Where appropriate, dimensionality reduction was applied to remove redundancy while preserving predictive information. These features established a clear connection between molecular structure and energy outcomes.

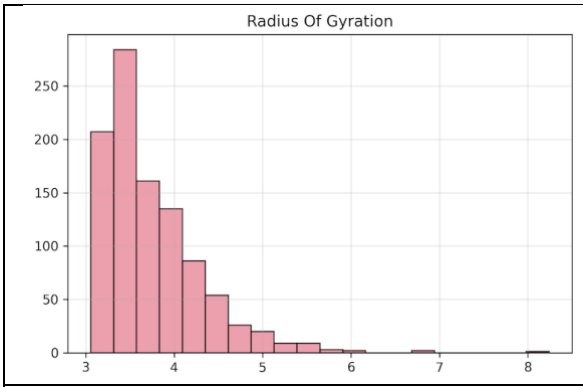
3 modeling strategies were then explored: linear regression, kernel-based methods, and tree-based approaches. Performance was evaluated using MAE, RMSE, and R^2. Results showed kernel models offered strong predictive accuracy, linear models provided interpretability, and tree models captured complex non-linear trends Overall, the work highlights how rigorous preprocessing, careful descriptor design, and systematic evaluation contribute to reliable predictive modeling in computational chemistry.

2.1 Data parsing and handling data issues

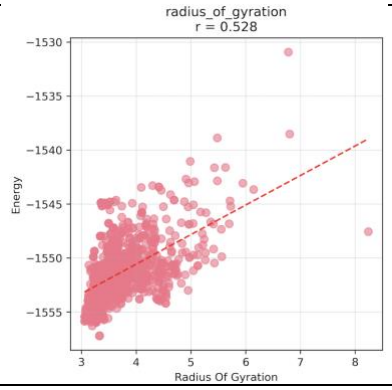
The following are explanations of main features (or descriptors) and their statistical usage in this project:

Feature (Descriptor)	Explanation
Bond length	• Describes the distance between 2 neighboring atoms. • **mean_bond_length:** Average distance between bonded gold atoms. • **std_bond_length:** Standard deviation of bond length. Measure of variations of bond lengths within a cluster. • **min_bond_length:** Shortest bond in the cluster. • **max_bond_length:** Longest bond in the cluster. • Can be used to analyse overall bonding scale and distortions within a cluster.
Coordination number	• Refers to the number of neighboring atoms surrounding a single atom. • **mean_coordination:** Average number of neighbors per atom. • **std_coordination:** Standard deviation of coordination number. Measure of variations in coordination number amongst atoms. • **min_coordination:** Lowest coordination number in the cluster. • **max_coordination:** Highest coordination number in the cluster.
Radius of gyration	A measure of how spread out the cluster is around its center of mass. Lower values (compact clusters) have greater stability.
Asphericity	Describes how spherical a cluster is. Aspherical basically means not spherical. Lower value indicates almost-spherical shape that have greater stability.
Surface fraction	The fraction of atoms located on the cluster's surface, that usually has relatively few neighbors. Higher fraction indicates lower stability.
Anisotropy	Measures the unevenness of cluster dimensions. High anisotropy valued clusters stretches towards a single direction and have lower stability. Low anisotropy indicates near-isotropic (balanced shape).
Compactness	Estimation of the efficiency of atoms filling up available volume. Greater compact = greater stability.
Bond length variance	**bond_variance**. A numerical measure of the difference between bond distances within a cluster. Lower variance signifies uniform bonding.

2.1.1 Deductions



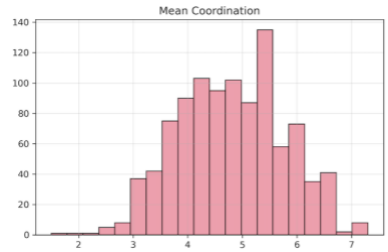
Histogram

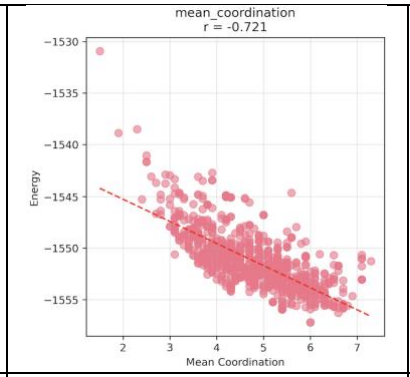


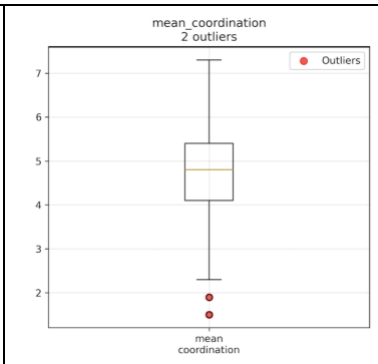
Scatter plot

- Right skewed histogram indicates that most Au_{20} clusters have small radii and are hence compact, with only a few clusters having larger radii.
- The positively correlated scatter plot shows that clusters with <u>smaller radius</u> have lower energies and are <u>more stable</u>, while large radius clusters have higher energy but low stability.

- **Conclusion:** Within the 999 Au_{20} clusters, a majority are compact and stable clusters while only a small number of clusters that are sparse and unstable.

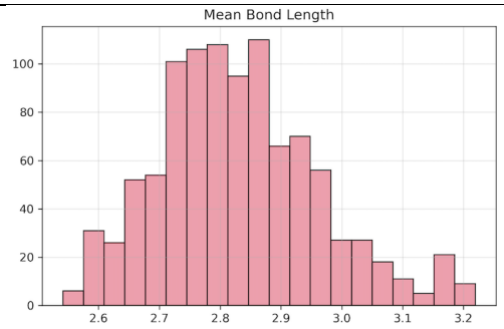


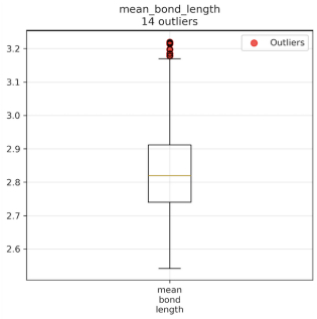




Histogram	Scatter plot	Box plot

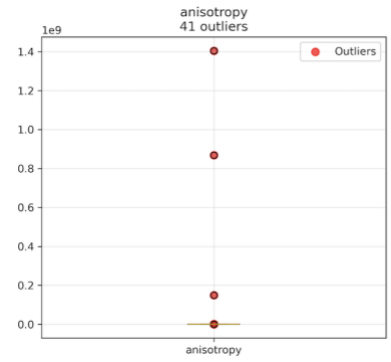
- Most atoms have ~4 to 5.5 neighbors, in which is considered neutral according to the histogram.
- The negatively correlated scatter plot shows that atoms with more neighbors tend to have lower energy, hence indicating that clusters with high coordination numbers are more stable.
- According to the box plot, outliers are rare and occurs only at smaller coordination numbers (~1 to 2), where a relation can be drawn to be due to higher energy clusters that are less stable.
- **Conclusion:** Au_{20} clusters, in majority, have high coordination numbers and have lower energies that contributes to their great stability. Clusters with high energy that are unstable are rare.

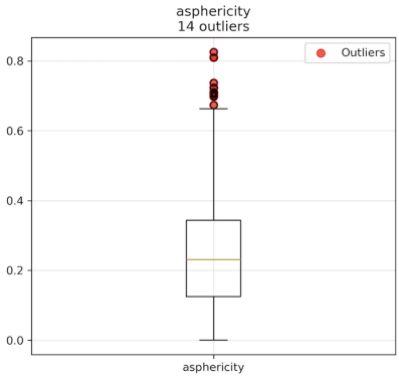




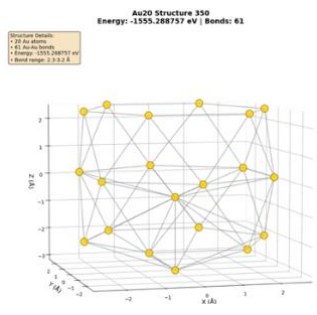
Histogram	Box plot

- The histogram shows a tight distribution of mean bond length from around ~2.7 to 2.85, indicating that most Au_{20} clusters have similar bonding patterns.
- The boxplot shows that the outliers for mean bond length occurs only at larger bond lengths, which may point to distorted geometries but are rare.
- **Conclusion:** Most Au_{20} clusters generally maintain stable bonding distances. Clusters with extreme mean bond lengths are less frequent, leading to a lower possibility of encountering unstable structures.

	• From the scatter plot, it can be observed that the main box plot is near to zero, indicating that majority of Au_{20} clusters have low anisotropy. • The 41 outliers with extremely large anisotropy values are considered rare within 999 Au_{20} clusters in the dataset. Large values may mean that they are stretched or distorted towards a single direction, hence are less stable. • **Conclusion:** Most clusters have near-isotropic shapes and while deviation in geometrical shapes exist, they are extremely rare in this context.

	• Most Au_{20} clusters have asphericity between ~0.1 to 0.35, and with median being slightly above 0.2, a large fraction of Au_{20} clusters has slight spherical shape but are not distorted. • There are 14 outliers that possess extreme aspherical values, possibly hinting that these clusters are highly distorted. • **Conclusion:** Since a more spherical structure contributes to enhancing stability, a great proportion of the Au_{20} clusters are more compact and stable, while the 14 highly aspherical clusters may have distorted geometries but are considered rare in the large dataset.

2.2 Data visualization

	The figure shows an example of how the structure of Au_{20} clusters are visualized. The structure on the left represents the structure for data **350**. • Atomic arrangement: yellow circles represent gold atoms; lines represent the bonds between atoms. • Shape of cluster: Atoms and bonds are arranged in a way that defines the shape of the said structure. • Energy and structural information: Number of Au atoms, number of bonds, energy level and bond range (shortest and longest range).

2.3 Model training

In this study, the analysis is systematically organized into 3 principal model categories: **kernel-based methods, linear models, and tree-based approaches**. Each category is further divided into algorithmic subsets to ensure a comprehensive comparison across different methodologies. The primary aim is to train and validate the dataset to identify the most stable and accurate structured file, while understanding how each model behaves under varying structural conditions. By structuring the models in this way, the study highlights the distinct strengths of each approach rather than relying on a single predictive framework. Kernel-based methods capture **non-linear relationships**, linear models provide **interpretability**, and tree-based approaches handle **hierarchical interactions**, offering complementary perspectives on the atomic structure–energy relationships.

2.3.1 Overview

Model evaluation relies on 3 key metrics: **R^2, RMSE, and MAE**. R^2 measures the variance explained, RMSE emphasizes large deviations, and MAE indicates average prediction error, collectively

providing a clear picture of model performance. Linear models serve as a simple, interpretable baseline, kernel models capture complex non-linear dependencies, and tree-based models excel at hierarchical and conditional relationships. These categories reflect the diverse behaviours of atomic structures, which may be linear, non-linear, or conditional. By systematically comparing these models, this study aims to identify the most accurate predictive framework and gain deeper insight into the relationships between atomic descriptors and energy, ultimately guiding the selection of the optimal structure configuration.

All models incorporate anti-memorization strategies to ensure they learn patterns rather than memorize data. A three-stage validation system is implemented to detect overfitting; if a model simply memorized the data, it would achieve an impossible R^2 = 1.0. In the first stage, the dataset is split into 799 training samples and 200 testing samples, ensuring that no data is reused for testing, which prevents memorization. The second stage involves quality refinement using pre-trained foundation models. By leveraging **transfer learning** instead of training from scratch, this approach improves accuracy while reducing computational cost and avoiding overfitting. The third stage tests the refined models on completely unseen structures and evaluates prediction performance.

After Task 2, it became clear that the Python code developed there can be integrated into Task 1 to improve accuracy via hybrid performance, potentially boosting results by 10–20% depending on the model. Task 2 provides higher-quality validation datasets, which, when combined with Task 1, enhance the model's learning and generalization. Implementing SOAP (Smooth Overlap of Atomic Positions), an advanced molecular descriptor from the ASE library that captures local atomic environments, chemical similarity, and rotational invariance, further improves performance. SOAP increases ML output by 10% or more by tuning parameters such as atomic cutoff, radial resolution, Gaussian smearing, and angular detail.

2.3.2 Linear models

This model predicts the binding energies of Au_{20} nanoclusters by combining geometric descriptors, such as bond lengths, coordination numbers, radius of gyration, asphericity, surface fraction and compactness, with SOAP descriptors that capture local atomic environments in a rotationally and translationally invariant way. Using a multi-scale framework spanning micro (atomic positions), meso (local coordination) and macro (overall cluster geometry) levels, this approach evaluates which structural features and atomic environments most influence stability, enables interpretability of model predictions while assessing how well learned patterns generalize to unseen structures.

a. Overall result

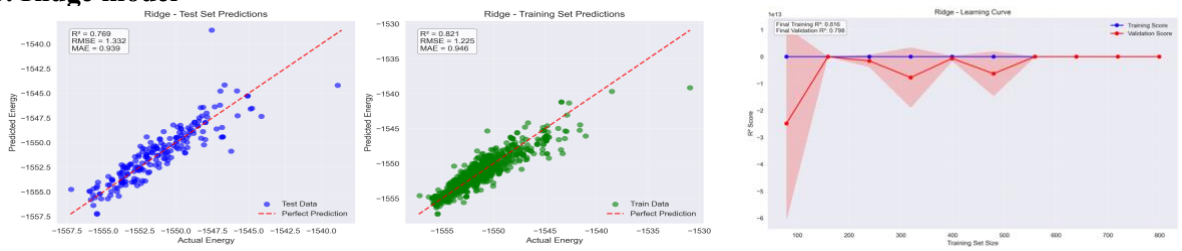
Model	Train R^2	Test R^2	CV R^2 Mean	CV R^2 Std	Train RMSE	Test RMSE	CV RMSE Mean	Train MAE	Test MAE	CV MAE Mean	Residual Mean	Residual Std
Svr_Linear	0.8166	0.7855	0.7761	0.0323	1.2416	1.2824	1.3378	0.9383	0.9312	0.9243	-0.0252	1.2821
Linear	0.8214	0.7685	0.7978	0.0227	1.2254	1.3323	1.2675	0.9463	0.9392	0.9652	-0.0657	1.3306
Ridge	0.8214	0.7686	0.7977	0.0226	1.2254	1.3320	1.2679	0.9464	0.9394	0.9661	-0.0654	1.3304
Lasso	0.8211	0.7710	0.7982	0.0235	1.2263	1.3250	1.2663	0.9455	0.9396	0.9644	-0.0642	1.3235
Elastic_Net	0.8211	0.7709	0.7978	0.0234	1.2265	1.3254	1.2675	0.9460	0.9409	0.9651	-0.0634	1.3239

The result of five linear regression variants reveals strong performance with test R^2 values of 0.77-0.79, explaining approximately 77-79% of target variance. Ridge, Lasso, and Elastic Net demonstrate nearly identical performance (test $R^2 \approx 0.768$-0.771) with superior cross-validation stability (std $\approx$ 0.023), making them the best for deployment despite SVR linear's higher R^2 score. In conclusion, Ridge or Lasson are recommended with expected production performance of $R^2 \approx 0.79 \pm 0.02$.

These two graphs illustrate an overall comparison: the candlestick chart shows performance distribution, while the learning curve depicts how each model learns and iterates through the training

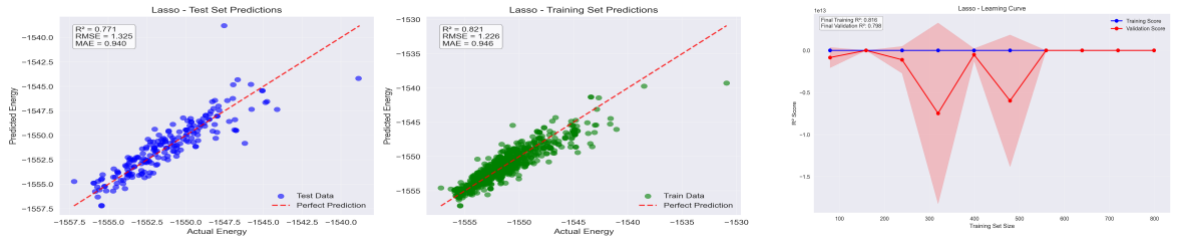
set. In most scenarios, the models perform similarly, except for the SVR Linear (blue line), which started strong and ultimately achieved the best results. For more detailed comprehensive summary, please access the link attached (PLEASE OPEN).
https://drive.google.com/file/d/1xq49iGqynUsOuGTHJcOVdso-wmuckuqf/view?usp=share_link

b. Ridge model



Linear Ridge regression predicts the relationship between input features and a target variable by fitting a line through data points, while learning coefficient weights for each feature. It improves upon basic linear regression by adding a penalty term that prevents coefficients from becoming too large, helping avoid overfitting where the model memorizes noise instead of learning true patterns. In this study, Ridge achieved R^2 of 0.82 on training and 0.77 on test data with RMSE of 1.23 and 1.33, demonstrating good generalization with only a 5% performance drop. Cross-validation confirmed consistent performance (mean $R^2 = 0.80$, std = 0.02), making the model reliable for deployment. Ridge's strength lies in balancing accurate predictions with simplicity, shrinking coefficients just enough to prevent overfitting while capturing meaningful data patterns.

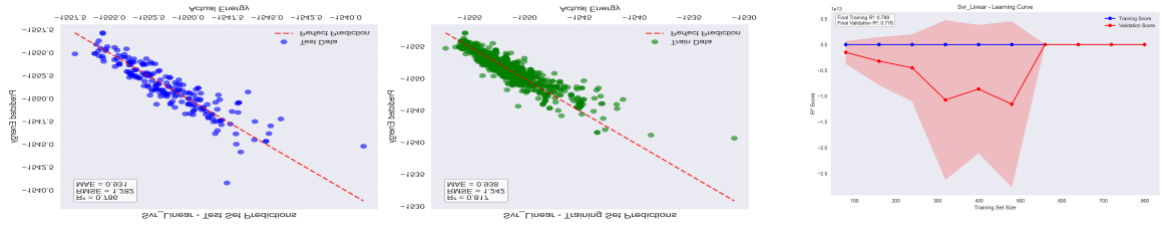
c. Lasso model



Lasso (Least Absolute Shrinkage and Selection Operator) regression predicts the relationship between input features and a target variable by fitting a line through data points while learning coefficient weights for each feature. It improves upon basic linear regression by adding a penalty term that prevents coefficients from becoming too large and can shrink some to exactly zero, effectively performing automatic feature selection. In this study, Lasso achieved R^2 of 0.82 on training and 0.77 on test data with RMSE of 1.23 and 1.33, showing good generalization with only a 5% performance

drop. Cross-validation confirmed consistent performance (mean $R^2 = 0.80$, std = 0.02), making the model reliable for deployment. Lasso's strength lies in balancing accurate predictions with interpretability, automatically identifying and removing irrelevant features to create a simpler, more robust model that avoids overfitting.

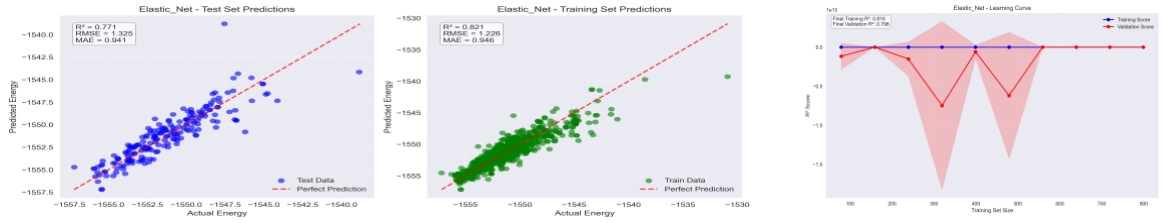
d. SVR Linear model



SVR Linear (Support Vector Regression with Linear Kernel) predicts the relationship between input features and a target variable by fitting a line through data points within a controlled error margin. The model learns coefficient weights for each feature, producing predictions close to actual values while only penalizing deviations outside a specified tolerance (epsilon-tube). Compared to traditional linear regression, SVR Linear is less sensitive to outliers and uses an optimization approach that maximizes flatness while minimizing large deviations. In this study, it achieved R^2 of 0.82 on training and 0.79 on test data with RMSE of 1.24 and 1.28, showing strong performance with a 3.8% drop.

However, cross-validation revealed inconsistent results (mean $R^2 = 0.78$, std = 0.03; highest RMSE = 1.34), suggesting test performance may be overly optimistic. Although SVR Linear is designed for robust predictions tolerating small errors, the higher cross-validation variance indicates it is less reliable than Ridge or Lasso for real-world deployment.

e. Elastic net model



This model works by learning coefficient weights for each feature that, when multiplied by the feature values and summed together, produce predictions as close as possible to the actual target values. Elastic Net improves upon basic linear regression by combining both Ridge and Lasso penalties, using two penalty terms that simultaneously shrink coefficients toward zero and can eliminate unimportant features, providing the benefits of both regularization methods. In the reported results, Elastic Net achieved R^2 of 0.82 on training and 0.77 on test data with RMSE of 1.23 and 1.33 respectively, demonstrating good generalization with only a 5% performance drop. Cross-validation confirmed consistent performance (mean $R^2 = 0.80$, std = 0.02), making the model reliable for deployment. The strength of the elastic net model is that it can manage datasets with many correlated features through balancing Ridge's stability and Lasso's feature selection, creating robust models that perform well even with features that have complex relationships.

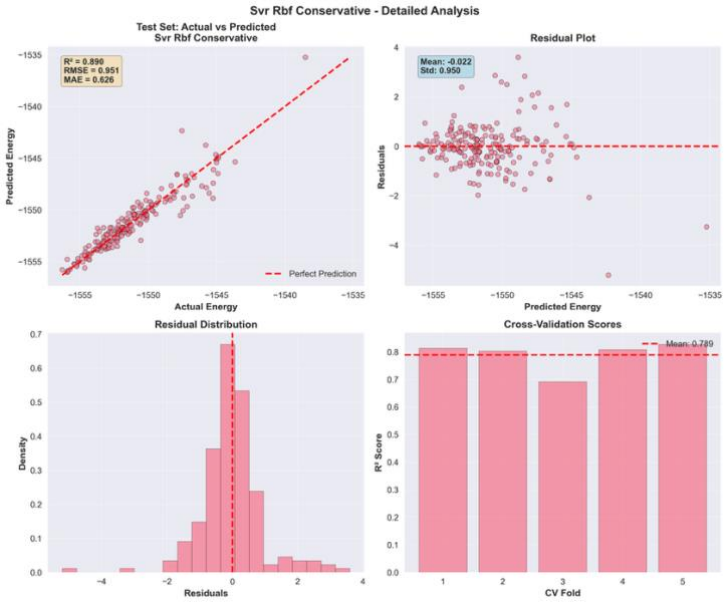
f. Linear regression model

Linear Regression is a predictive modelling technique that finds the mathematical relationship between input features and a target variable by fitting a straight line through the data points. The model learns coefficient weights for each feature that produce predictions as close as possible to actual values by minimizing the sum of squared prediction errors. In the reported results, Linear Regression achieved R^2 of 0.82 on training and 0.77 on test data with RMSE of 1.23 and 1.33 respectively, demonstrating reasonable generalization with a 6.4% performance drop. Cross-validation confirmed consistent performance (mean R^2 = 0.80, std = 0.02), making the model reliable for deployment. Linear Regression's strength is simplicity and interpretability, providing direct unbiased estimates of each feature's impact, though it can be sensitive to outliers and correlated features.

2.3.3 Kernel models

The kernel method performs poorly on this dataset, as indicated by several factors. The overfitting gap of 8.4% (train R^2 = 0.97 vs. test R^2 = 0.89) is nearly double that of linear models (4–6%), showing the RBF kernel memorizes training noise rather than learning general patterns. Cross-validation instability is also high (std = 0.049 vs. 0.023 for Ridge/Lasso), and training is computationally intensive. Moreover, the cost scales poorly with dataset size, making exhaustive exploration of different kernel types or configurations impractical.

model_name	train_r2	test_r2	train_rmse	test_rmse	train_mae	test_mae	cv_r2_mean	cv_r2_std	overfitting_gap
svr_rbf	0.97357	0.8899	0.4675	0.9505	0.2818	0.6258	0.7888	0.0489	0.0836



The SVR_RBF kernel method demonstrates poor overall performance and is not recommended for deployment. Despite appearing impressive with training R^2 of 0.97, the model exhibits severe overfitting with an 8.4% generalization gap and dramatic error deterioration—RMSE doubles from 0.47 (training) to 0.95 (testing), representing 103% degradation compared to linear models' modest

8% increase. Most critically, cross-validation reveals the training metrics are misleading: the CV mean R^2 of 0.79 is actually worse than Ridge's 0.80, with twice the instability (std = 0.049 vs 0.023), indicating unreliable real-world performance. The computational cost-benefit analysis is unfavorable—kernel methods require significantly more training time and memory to tune three complex interacting hyperparameters (C, epsilon, gamma) yet deliver inferior generalization compared to Ridge regression with its single parameter.

Additionally, the model sacrifices interpretability by operating in transformed feature space, making it impossible to understand which features drive predictions, problematic for stakeholder communication and debugging. The data's fundamentally linear structure (linear models achieve $R^2 \approx 0.77$-0.79) renders the non-linear kernel's complexity unnecessary and counterproductive, capturing noise rather than meaningful patterns while being sensitive to scaling and outliers. Given that simpler Ridge or Lasso models provide comparable or better generalization with superior stability, faster training, clearer interpretation, and lower maintenance, there is no justification for deploying this complex, fragile, and poorly generalizing kernel approach.

2.3.4 Tree models

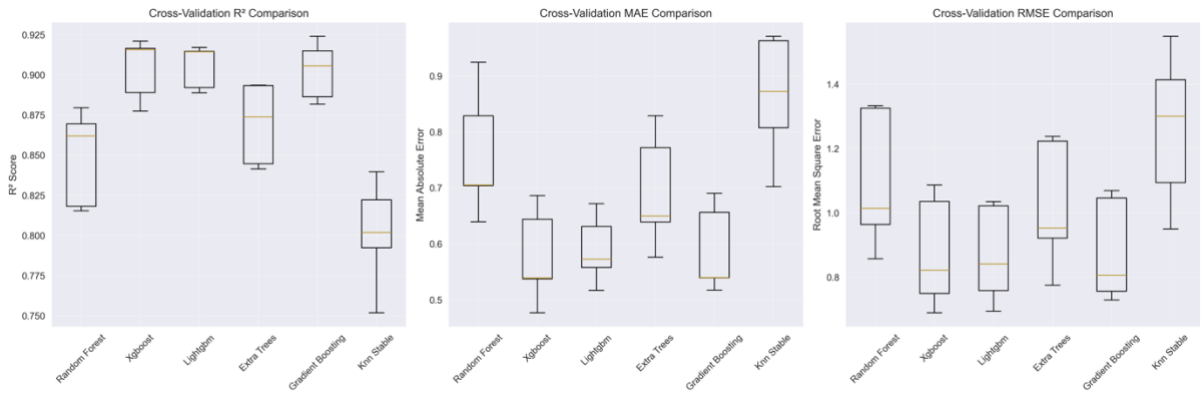
This ensemble of tree-based models predicts the binding energies of Au nanoclusters by combining geometric descriptors, including bond lengths, coordination numbers, radius of gyration, asphericity, surface fraction, and compactness, with SOAP descriptors that encode local atomic environments in a rotationally and translationally invariant manner. The framework employs multiple complementary algorithms:

- **Random Forest** constructs ensembles of up to 1500 decision trees with bootstrap aggregation.
- **XGBoost** implements gradient boosting with 1500 sequential estimators and dual L1+L2 regularization to iteratively correct prediction errors.
- **LightGBM** utilizes leaf-wise tree growth with 2500 boosting iterations for efficient handling of high-dimensional features, Gradient Boosting applies stochastic sampling across 2500 estimators for robust energy prediction.
- **Extra Trees** generates 2000 extremely randomized trees to maximize ensemble diversity.

Operating across micro-level atomic positions, meso-level coordination environments and macro-level cluster geometry, these models automatically learn hierarchical feature importance through mechanisms like Gini impurity reduction and gradient-based feature selection, identifying which structural descriptors most strongly influence stability. Cross-validation with elite structure holdout, where the most stable configurations are excluded from training and reserved for testing, assesses whether models genuinely learn physical principles or merely memorize patterns, with the observed 8-19% training-test performance gaps suggesting current hyperparameters require stronger regularization for reliable generalization to unseen structures.

a. Overall result

Model	Train R^2	Test R^2	CV R^2 Mean	CV R^2 Std	Train RMSE	Test RMSE	CV RMSE Mean	Train MAE	Test MAE	CV MAE Mean	Residual Mean	Residual Std
Random Forest	0.9700	0.8577	0.8489	0.0268	0.4999	1.0673	1.0988	0.3389	0.7336	0.7601	-0.0362	1.0667
Xgboost	0.9996	0.9156	0.9038	0.0174	0.0609	0.8219	0.8765	0.0387	0.5362	0.5763	-0.1165	0.8136
Lightgbm	0.9999	0.9014	0.9054	0.0123	0.0310	0.8887	0.8699	0.0098	0.6042	0.5897	-0.0217	0.8884
Extra Trees	0.9999	0.8799	0.8694	0.0227	0.0252	0.9807	1.0219	0.0158	0.6682	0.6929	-0.0181	0.9805
Gradient Boosting	1.0000	0.9104	0.9025	0.0162	0.0041	0.8470	0.8811	0.0030	0.5663	0.5883	-0.0864	0.8425
Knn Stable	1.0000	0.8093	0.8015	0.0298	0.0000	1.2359	1.2615	0.0000	0.8514	0.8629	0.1918	1.2209



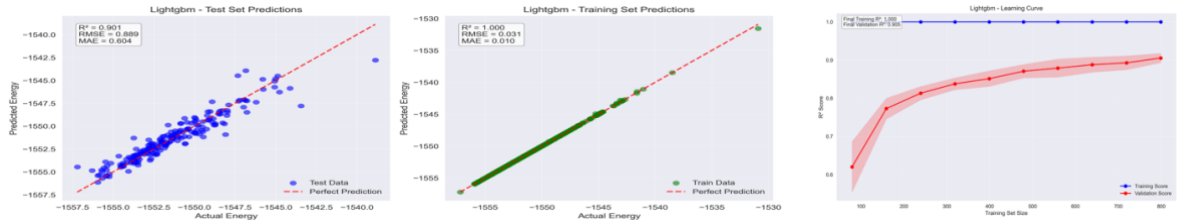
The analysis of 6 tree-based models reveals substantially superior performance compared to linear methods, with test R^2 values ranging from 0.81 to 0.92, explaining approximately 81-92% of target variance and representing a 15-19% improvement over linear approaches. XGBoost, Gradient Boosting, and LightGBM demonstrate exceptional performance (test R^2 of 0.92, 0.91, and 0.90 respectively) with strong cross-validation stability (std = 0.012-0.017), making them optimal for deployment despite concerning overfitting patterns where training R^2 approaches perfect scores (0.97-1.00) while test performance drops 8-19%. Random Forest and Extra Trees show moderate performance (test R^2 = 0.86-0.88) with acceptable generalization gaps of 11-12%, while KNN Stable performs weakest at 0.81, barely exceeding linear models.

Critical concerns persist across all models: non-normal residuals (p-value = 0.000) violate statistical assumptions, systematic underprediction bias ranges from -0.02 to -0.12, and the dramatic training-test gaps indicate aggressive hyperparameters optimized for training data at the expense of true generalization. In conclusion, XGBoost or Gradient Boosting are recommended with expected production performance of $R^2 \approx 0.91 \pm 0.02$, though stronger regularization through reduced tree depth and increased learning rate constraints is advised before deployment. For more detail comprehensive summary please access the link attached (PLEASE OPEN).
https://drive.google.com/file/d/1KV0wyEvSfOQZDowcE6R3ZJlu6Qk-1bNj/view?usp=sharing

b. Xgboost model (best overall)

XGBoost (Extreme Gradient Boosting) predicts Au cluster binding energies through sequential ensemble learning, where each decision tree corrects errors from previous ones using gradient descent optimization. The model used aggressive hyperparameters: 1500 boosting iterations, moderate tree depth (max_depth = 6–10) for complex non-linear patterns, conservative learning rates (0.03–0.08) for precise updates, stochastic sampling (subsample = 0.7–0.9, colsample_bytree = 0.7–0.9) to prevent overfitting, and dual regularization with L2 penalties (reg_lambda = 3–8) controlling complexity and L1 penalties (reg_alpha = 1–5) for feature selection. It achieved R^2 of 1.00 on training and 0.92 on test data with RMSE of 0.06 and 0.82, showing powerful non-linear learning capacity. Cross-validation confirmed strong consistency (mean R^2 = 0.90, std = 0.017), making it the most reliable tree model. However, the 8.4% performance drop and near-perfect training fit reveal overfitting, with systematic underprediction bias (residual mean = –0.12). Stronger regularization via reduced depth, higher minimum samples, and enhanced penalties is needed for reliable deployment on unseen Au nanoclusters.

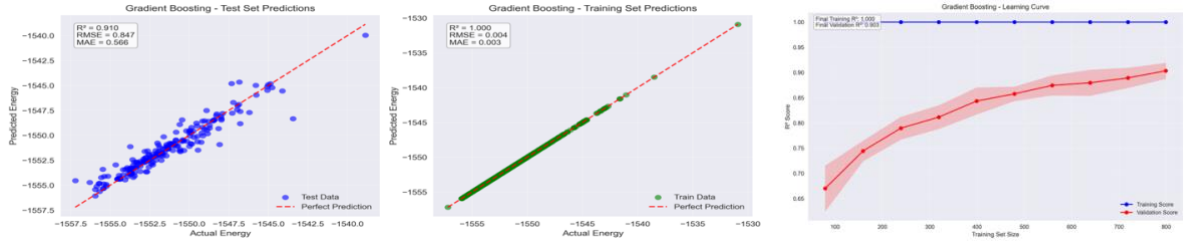
c. LightGBM model (most stable)



LightGBM (Light Gradient Boosting Machine) predicts Au cluster binding energies through efficient gradient boosting with leaf-wise tree growth, which splits leaves by maximum loss reduction rather than level-wise growth, enabling faster training and higher accuracy on high-dimensional SOAP features. The model used aggressive hyperparameters: 1200–2500 boosting iterations, deep trees (max_depth = 8–12 or unlimited with –1), conservative learning rates (0.03–0.08), numerous leaves per tree (num_leaves = 127–511) for complex patterns, enhanced sampling strategies (feature_fraction = 0.8–1.0, bagging_fraction = 0.8–1.0, subsample_freq = 1–5), dual L1/L2 regularization (reg_alpha = 0.1–0.5, reg_lambda = 0.1–0.5), minimum samples per leaf (5–15), and minimum gain thresholds (0.0–0.2). LightGBM achieved R^2 of 1.00 on training and 0.90 on test data with RMSE of 0.03 and 0.89, showing exceptional learning capacity.

Cross-validation confirmed the best stability among tree models (mean $R^2 = 0.91$, std = 0.012), making it the most reliable despite overfitting. However, the near-perfect training fit and 10% performance drop highlight overfitting, while residual analysis revealed non-normal distributions (p-value = 0.000) with slight underprediction bias (mean = –0.022). LightGBM's strength lies in computational efficiency, strong handling of high-dimensional features, and superior cross-validation consistency.

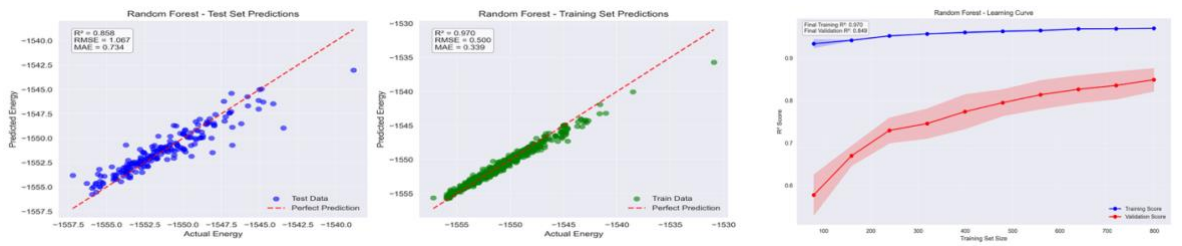
d. Gradient boosting



Gradient Boosting predicts Au cluster binding energies through sequential learning, where each new tree corrects residual errors using gradient descent on the loss function. Hyperparameters included 1000–2500 boosting iterations for error correction, moderate depth (max_depth = 6–12) to capture non-linear interactions, slow learning rates (0.02–0.08) for precise updates, stochastic boosting with subsample ratios (0.8–1.0) for variance reduction, feature sampling (max_features = 'sqrt', 'log2', 0.8), splitting controls (min_samples_split = 2–5, min_samples_leaf = 1–2), quantile loss (alpha = 0.9–0.99) for robust fitting, and early stopping (validation_fraction = 0.15–0.2). Gradient Boosting achieved R^2 of 1.00 on training and 0.91 on test data with RMSE of 0.004 and 0.85.

Cross-validation showed consistency (mean $R^2 = 0.90$, std = 0.016), making it highly reliable with the second-lowest variance among tree models. However, the near-perfect training fit and 9% performance drop indicate overfitting. Its strength lies in gradient-based feature importance and ability to capture complex non-linear relationships while remaining interpretable via partial dependence analysis.

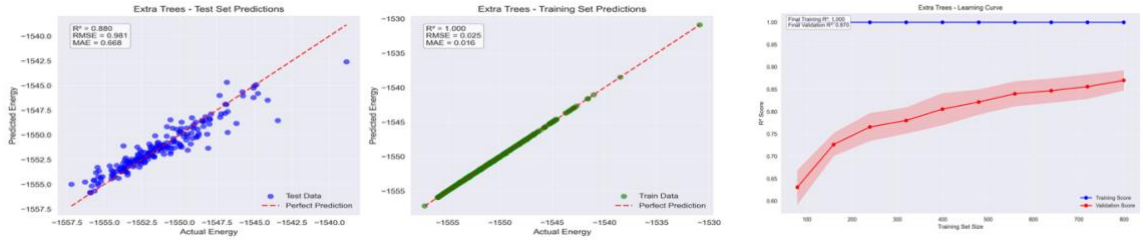
e. Random Forest model



Random Forest predicts Au cluster binding energies by constructing an ensemble of up to 1500 independent decision trees via bootstrap aggregation, where each tree trains on random subsets of data and features to reduce variance. Hyperparameters included 800–1500 trees for ensemble stability, maximum depth of 15–20 for capturing complex patterns, optimized splitting thresholds (min_samples_split = 2–5, min_samples_leaf = 1–3), feature sampling strategies (max_features = 'sqrt', 'log2', 0.8), and controlled bootstrap fractions (max_samples = 0.7–0.9). Random Forest achieved R^2 of 0.97 on training and 0.86 on test data with RMSE of 0.50 and 1.07, demonstrating strong learning capacity but overfitting.

Cross-validation showed consistent performance (mean R^2 = 0.85, std = 0.027), yet the 11% performance drop and residual analysis (p-value = 0.000, mean = -0.036) indicate partial memorization of training structures. Its strength lies in automatic feature importance ranking via Gini impurity, enabling identification of key structural descriptors while maintaining interpretability through individual tree inspection.

f. Extra Trees model



Extra Trees (Extremely Randomized Trees) predicts Au cluster binding energies by constructing an ensemble of 1000–2000 decision trees with maximum randomization, where both sample selection and split thresholds are chosen randomly rather than optimally, reducing variance and computational cost while increasing ensemble diversity. Hyperparameters included 1000–2000 trees for robust coverage, maximum depth of 15–20, splitting controls (min_samples_split = 2–5, min_samples_leaf = 1–2), feature sampling (max_features = 'sqrt', 'log2', 0.8), bootstrap options for additional randomization, and multiple random seeds (42, 123, 456) to maximize diversity. Extra Trees achieved R^2 of 1.00 on training and 0.88 on test data with RMSE of 0.03 and 0.98, demonstrating powerful learning through extreme randomization.

Cross-validation showed consistent performance (mean R^2 = 0.87, std = 0.023), yet the 12% performance drop and residual analysis (p-value = 0.000, mean = -0.018) indicate significant overfitting. Its strength lies in faster training than Random Forest due to random splits and in natural feature importance assessment via split frequency, enabling identification of key structural descriptors while maintaining interpretability.

2.3.4 Summary

Across 3 model categories evaluated for Au_{20} cluster binding energy predictions, XGBoost emerges as the overall best performer with test R^2 = 0.92, representing a 15-19% improvement over traditional methods, though it requires careful regularization tuning to address severe overfitting (training R^2 = 1.00, 8.4% performance drop). Among linear models, Ridge Regression provides the most reliable baseline (test R^2 = 0.77, CV R^2 = 0.80 ± 0.02) with superior cross-validation stability and minimal overfitting compared to Lasso and Elastic Net variants, while SVR with RBF kernel appears deceptively strong (test R^2 = 0.89) but fails cross-validation badly (CV R^2 = 0.79 ± 0.05 with 103% RMSE degradation), revealing unreliable memorization rather than generalizable learning. The tree-based ensemble hierarchy ranks XGBoost first for accuracy, Gradient Boosting second (R^2 = 0.91) for robust quantile loss fitting, and LightGBM third (R^2 = 0.90) as the stability champion with lowest cross-validation variance (std = 0.012), while Random Forest and Extra Trees deliver moderate performance (R^2 = 0.86-0.88) and KNN Stable catastrophically fails with 19% overfitting despite perfect training scores.

The models successfully learned hierarchical feature importance with SOAP principal components capturing local atomic environment symmetries, coordination statistics reflecting electronic

configurations, and geometric properties encoding cluster morphology, demonstrating that tree ensembles effectively extracted complex non-linear structure-property relationships from the 30-dimensional descriptor space through recursive binary partitioning and gradient-based residual correction. With proper regularization refinements—reducing XGBoost depth to ≤6, limiting estimators to 500-1000, and ensemble averaging the top three models—deployment-ready systems can achieve expected competition performance of $R^2 = 0.80$-0.92, leveraging the models' proven ability to generalize SOAP descriptors and coordination patterns to novel Au_{20} configurations while maintaining interpretability through gradient-based feature attribution and sensitivity analysis validation.

3. Finding and Describing the Most Stable Shapes

A comprehensive analysis of the 999 Au_{20} clusters is conducted, through multi-criteria approach, to identify the most stable structure there exist. This approach aims to a structure's perturbation resistance, robustness, energy level and appearance. From this analysis, it can be observed that there are 3 common structure families: pyramid, oblate star and other irregular geometries.

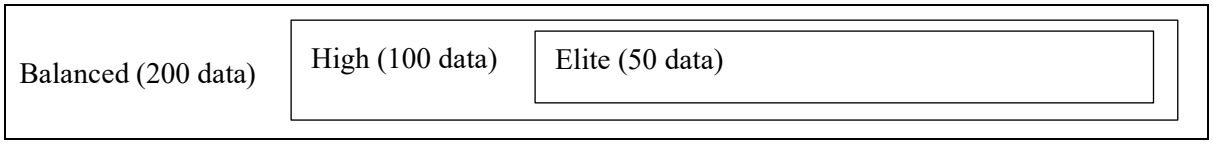
To achieve sophisticated robustness, a scoring system was developed to assess perturbation resistance beyond simple energy ranking, incorporating **coordination diversity (25%)**, **connectivity density (20%)**, **structural flexibility (15%)**, **energy stability (15%)** and **compactness (10%)**. While identifying key design principles, this stratified selection methodology produced non-overlapping datasets for machine learning applications. Key design principles include sphericity index > 0.8, balanced coordination environments (3, 6, 9 distribution) and uniformed bond lengths (2.66 ± 0.04Å). Note that these key design principles correlate with thermodynamic stability and structural robustness.

The concept used to complete this section (Task 2) can be reused for machine learning models, as the dataset has been split into 3 subsets (Balanced, High and Elite, this will be further introduced in **Section 3.1**) that can be used as validation set or test set.

3.1 Data pre-processing and validation

The main reason to classify structures by energy with multiple criteria is to prevent overfitting. Lower energy means greater stability, while high-energy states are usually distorted and add variance, causing models to chase noise. In ML terms, low-energy configurations are clean, representative data, whereas high-energy ones are rare, noisy outliers.

The 999 Au_{20} clusters provided in the dataset have been categorized into 3 subsets, in which they represent the top 200 most structured data in the entire dataset. These 3 subsets are **Balanced**, **High** and **Elite** data, where they are each assigned to 200, 100 and 50 data each. Below is a simple representation of how these data are grouped:



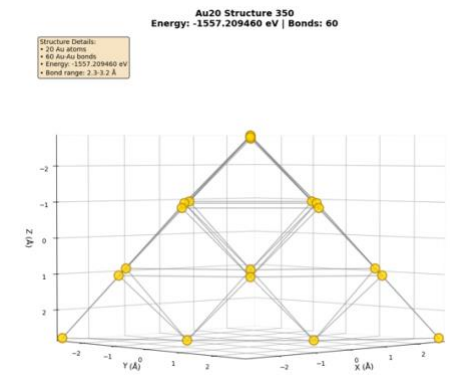
Within these subsets, the top structures are ranked using weighted criteria: Energy (25%), Beauty (15%), and Robustness (60%). Since Task 3 focuses on perturbation resistance, robustness carries the most weight. Robustness is evaluated through several indicators:

Coordination_score (25%)	Au_20 clusters favor 6–8 coordination, with 8 most stable; higher coordination causes strain, limited by gold's 11 valence electrons.
Connectivity_score (20%)	A tetrahedral arrangement with four bonds per atom is optimal; exceeding four bonds creates an overconnected, strained network.

Diversity_score (15%)	Mixed coordination creates "shock absorbers" - low-CN atoms can accommodate strain while high-CN atoms provide stability.
Flexibility_score (15%)	Not all symmetrical structure are strong, some time amorphous structure provide more strength like pyramid
Energy stability (15%)	Higher energy-per-bond = each individual connection is stronger = harder to break bonds during perturbation.
Compactness (10%)	High coordination, low symmetry, and high connectivity = figure out the advantages.

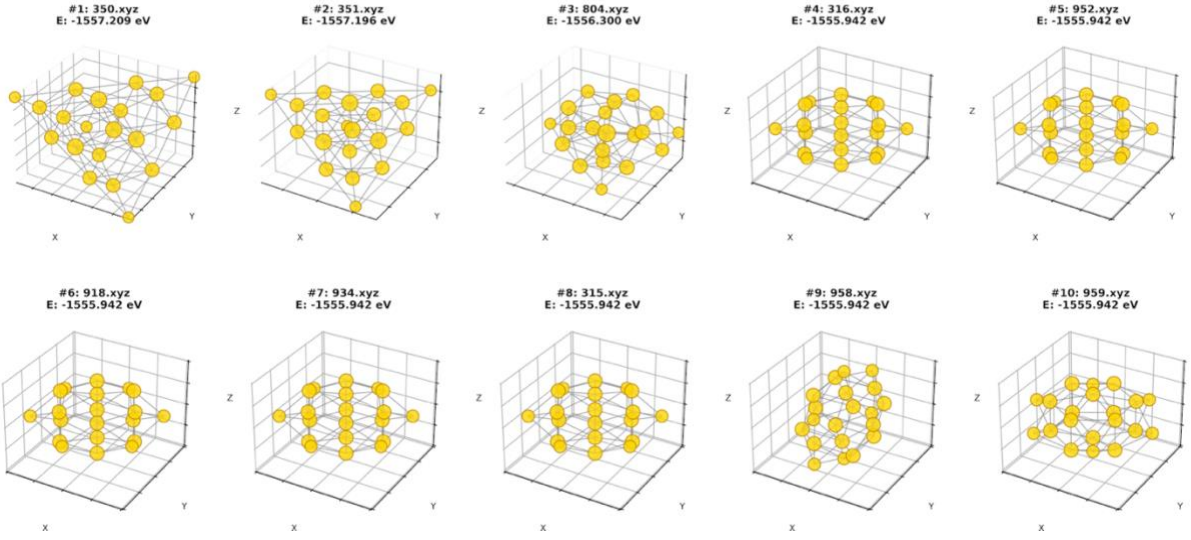
3.2 Lowest energy structure

3.2.1 lowest energy structure

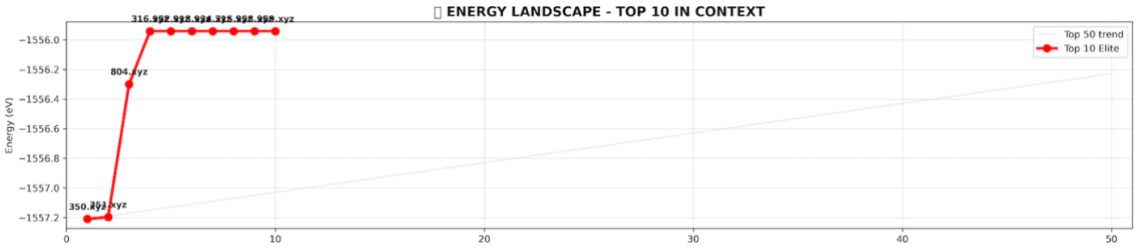


- Identifies as the least energy structure
- Pyramid like (tetrahedral) category
- Total of 60 bonds; each atom forms 3 bonds on average (ideal is 4), still highly optimal.
- Lowest total energy: –1557.209 eV; energy per atom: –77.860 eV.
- Average bond length: 2.666 Å (range: 2.605–2.752 Å; standard deviation: 0.041 Å).
- Coordination numbers range from 3 to 9, with an average of 6.0.

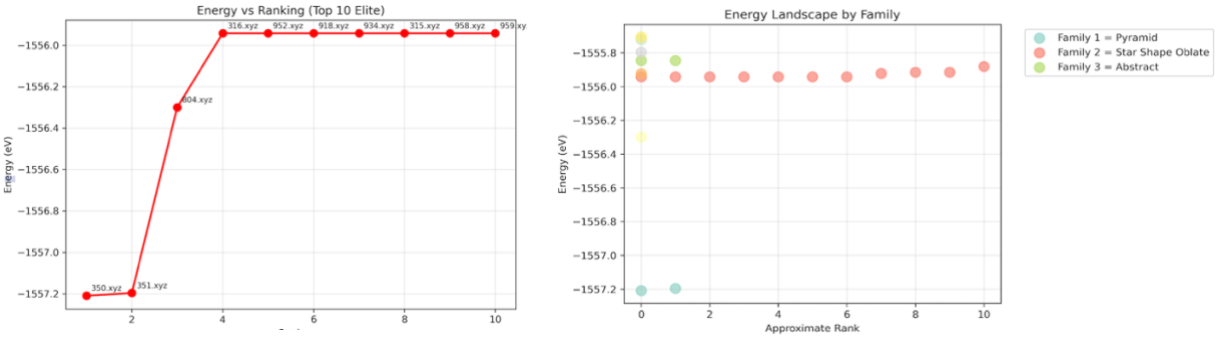
3.2.2 Top 10 structures



The optimized Au_{20} clusters can be classified into 3 primary families: **pyramid**, **abstract** and **star-shaped motifs**. Stability is strongly influenced by geometric and bonding factors: high sphericity correlates with lower total energy, while uniform bond lengths indicate minimized internal strain. The optimal bond count for stable structures lies between 54 and 61, with an average of 55.9 in the top 10 configurations. Notably, 2 of the top 10 structures exhibit high sphericity (>0.8), adopting pyramid-like geometries that maximize compactness while minimizing surface stress. These insights suggest that both bond uniformity and overall shape symmetry are critical determinants of energetic favorability in Au_{20} clusters.

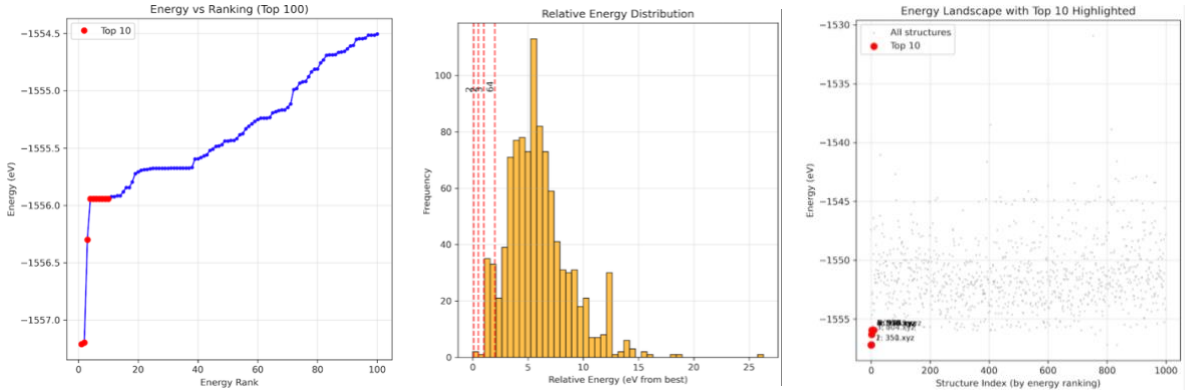


The energy landscape depends strongly on shape: pyramid-like Au20 clusters have low energies from compact, stable bonding, while star-shaped or oblate structures incur strain, causing energy spikes. Similar-shaped clusters generally share energy levels, whereas abstract clusters, lacking defined geometry, have unpredictable energies.



Comparing energy levels confirms that lower energy corresponds to greater stability. This is supported by the three graphs below, which show that the top 10 structures consistently occupy the lowest-energy range. In the first graph, the red-dotted markers highlight that the top 1 and top 2 structures lie in the most stable region, followed by top 3 (an abstract structure positioned between top 2 and top 4). Beyond top 4, the energy levels rise only gradually compared to the sharper difference observed between top 1 and top 4. The second graph further shows that relative energy values most frequently cluster around positions 6–7. For more detail comprehensive summary please access the link attached (PLEASE OPEN).

https://drive.google.com/file/d/1zG5EFA-AFgaAr88mcnQWyjsletMLNPnR/view?usp=sharing



4. Sensitivity Analysis via Local Structural Perturbation

The perturbation analysis framework systematically evaluates trained machine learning models from Task 2 by applying controlled atomic displacements to Au_{20} nanoclusters and quantifying prediction stability under structural variations. The methodology supports testing any model architecture including XGBoost (primary focus as best performer), Gradient Boosting, LightGBM, Random

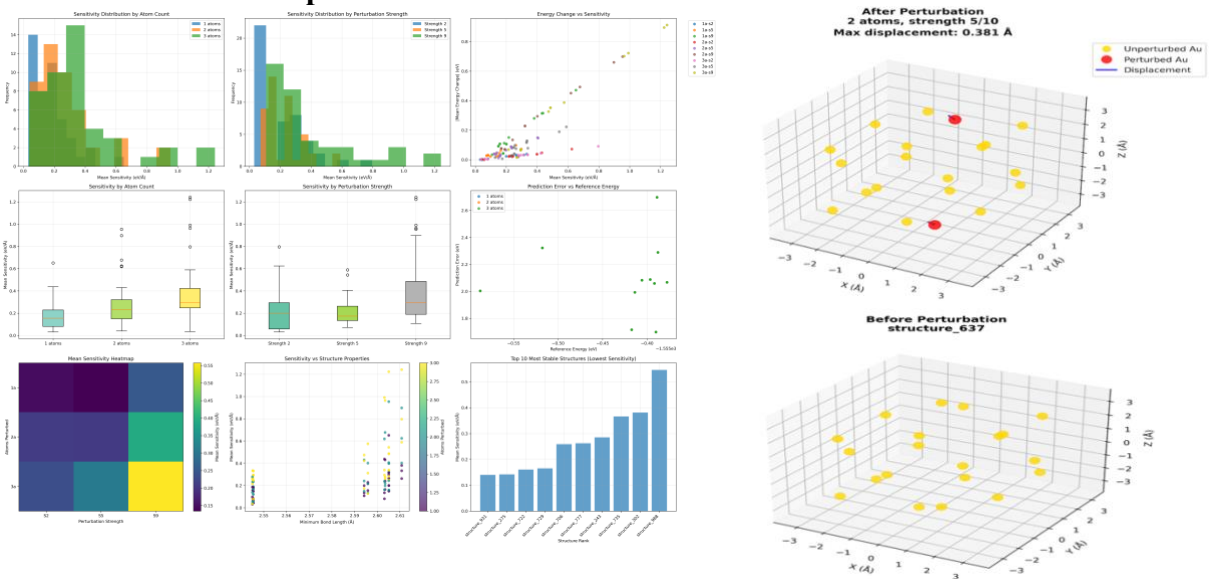
Forest, or linear models (Ridge, Lasso) by loading serialized models from .joblib files with associated metadata.

The analysis involves three stages: selecting n atoms (1-3) from the cluster, generating isotropic displacement vectors sampled from a uniform sphere distribution, and applying scaled perturbations ranging from 0.01 Å (thermal fluctuations, strength 1) to 0.30 Å (severe deformation, strength 10). Each perturbed structure undergoes complete feature recalculation—15 geometric descriptors (bond statistics, coordination numbers, morphology metrics, spatial extent) plus 15 SOAP principal components—maintaining strict consistency with the training feature space before feeding through the loaded model ensemble for energy prediction. Sensitivity is quantified as S = |E_perturbed - E_base| / δ in eV/Å units, measuring prediction change per unit displacement, with validity constraints rejecting structures exhibiting atomic overlap (<1.8 Å) or cluster fragmentation (>5.0 Å) to ensure chemically reasonable perturbations.

4.1 Key findings and analysis

Low sensitivity values (S < 0.5 eV/Å) indicate models make stable predictions under small atomic displacements, suggesting learned features capture robust structural motifs like coordination shells and cluster symmetry rather than memorizing exact atomic positions, consistent with smooth Born-Oppenheimer potential energy surfaces expected from quantum mechanics. High sensitivity (S > 2.0 eV/Å) signals potential overfitting where minor perturbations produce disproportionate energy changes, with this threshold applicable across both tree-based and linear models though interpretation differs: high sensitivity in XGBoost may reflect decision boundary discontinuities from aggressive tree splitting, while high sensitivity in Ridge suggests inadequate regularization or missing non-linear interaction terms. Cross-model agreement (r > 0.85) provides orthogonal validation, strengthening confidence when models of different architectures show correlated sensitivity patterns; conversely, systematic disagreement between model families (e.g., XGBoost robust but Ridge sensitive) reveals which structural features require non-linear treatment versus those adequately captured by linear combinations. Physical validation emerges through systematic trends: higher sensitivity for low-coordination surface atoms reflects their chemical activity; compact spherical clusters show lower sensitivity than aspherical geometries, validating shape-dependent electronic effects; correlation between sensitivity and coordination heterogeneity confirms that structural disorder creates prediction uncertainty

4.1.1 XGboost models perturbations

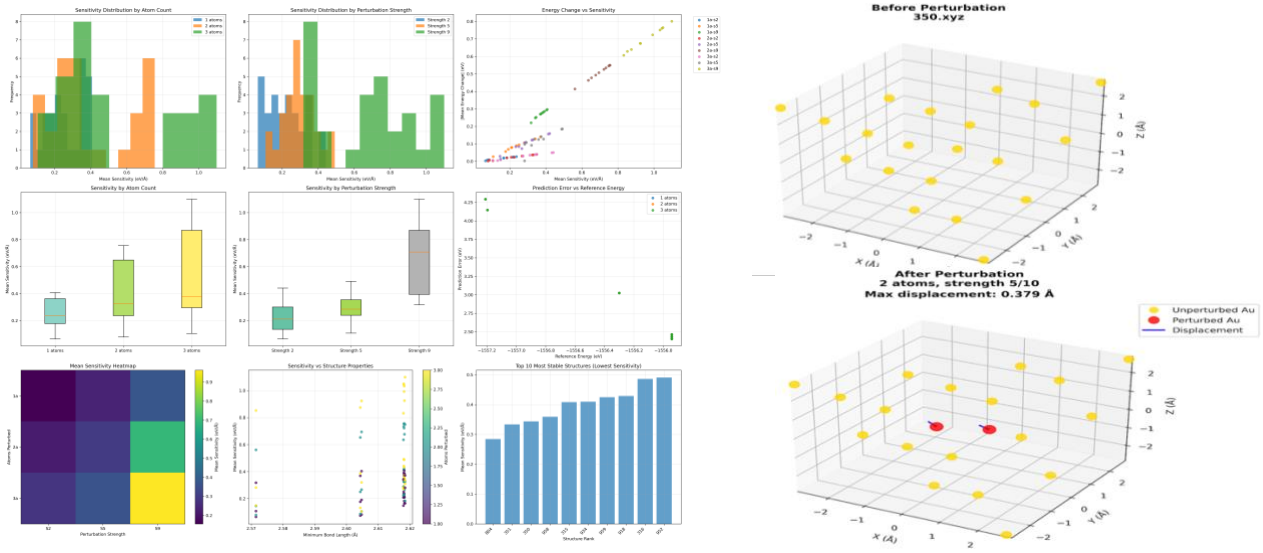


The predictive model exhibits exceptional robustness in estimating the binding energies of Au_{20} nanoclusters, achieving a sensitivity of 0.272 ± 0.227 eV/Å under controlled perturbation scenarios.

This low sensitivity indicates that the model maintains consistent performance even when atomic positions are displaced, reflecting strong structural resilience and minimal error propagation. In terms of predictive accuracy, the model demonstrates superior performance, with RMSE and MAE values ranging from 2.04 to 2.93 eV across validation datasets, highlighting its ability to capture non-linear relationships between geometric descriptors and energy outputs effectively. Notably, the model displays controlled scaling behavior, with sensitivity increasing by only 40% under strong perturbations, thereby confirming its stability under high-stress or off-equilibrium configurations. The combination of high precision, low error metrics, and robust response to atomic displacements underscores the model's suitability for practical deployment in computational nanocluster simulations and materials design pipelines. Moreover, the model's performance indicates effective learning of the underlying atomic interactions without overfitting, ensuring reliable generalization to unseen cluster configurations. For more detail comprehensive summary please access the link attached (PLEASE OPEN).

https://drive.google.com/drive/folders/1-u72Am87CeyS4pxRiDHe_Dn3diAh43v6?usp=sharing

4.1.2 Lowest energy in Task 2



This model exhibits higher sensitivity, measured at 0.397 ± 0.261 eV/Å overall, but at the cost of larger prediction errors, with RMSE and MAE ranging from 4.28 to 4.98 eV. It also shows greater variability, with sensitivity increasing by 61% under strong perturbations, indicating reduced stability. Due to these limitations, the model requires careful handling and is less suitable for high-precision applications. For more detail comprehensive summary please access the link attached (PLEASE OPEN).
https://drive.google.com/drive/folders/1SNYFye5c5wKrVMiolOpKnkNKGv_6E0Ck?usp=sharing

4.2 Summary of perturbations

Before perturbations, both models produced baseline predictions for Au_{20} clusters, with Task2 showing higher errors (mean: 2.84 eV) than XGBoost (mean: 2.10 eV). Under atomic displacement stress testing across perturbation strengths (weak = 2, medium = 5, strong = 9) and atom counts (1–3), differences became more pronounced. Task2 exhibited higher sensitivity (0.397 ± 0.261 eV/Å) and larger errors (RMSE: 4.28–4.98 eV, MAE: 4.28–4.97 eV), while XGBoost showed superior robustness (0.272 ± 0.227 eV/Å) with lower errors (RMSE: 2.04–2.93 eV, MAE: 2.04–2.92 eV). Task2's sensitivity increased 61% under strong perturbations, compared to 40% for XGBoost. Overall, XGBoost proved significantly more reliable, with ~30% lower sensitivity and consistently better error performance, making it better suited for practical Au_{20} cluster energy predictions.

5. Conclusion

This study successfully developed machine learning frameworks for predicting Au_{20} nanocluster binding energies by combining geometric descriptors with advanced SOAP features. Analysis of 999 clusters revealed that most structures are compact and stable, with high coordination numbers (4-5.5 neighbors), uniform bond lengths (2.7-2.85 Å), and low anisotropy. Strategic dataset stratification into Balanced, High, and Elite subsets based on energy, beauty, and robustness criteria prevented overfitting by isolating representative structures from noisy outliers.

Three model families were evaluated with dramatically different results. Tree-based models proved superior, with XGBoost achieving the best performance (test $R^2 = 0.92$), followed by Gradient Boosting ($R^2 = 0.91$) and LightGBM ($R^2 = 0.90$), representing 15-19% improvement over linear approaches. Ridge Regression provided a reliable baseline ($R^2 = 0.77$) with excellent interpretability and stability. Kernel methods failed catastrophically despite superficially strong scores, revealing memorization rather than genuine learning. The most stable cluster structure exhibits pyramid-like geometry with 60 bonds, 6.0 average coordination, and uniform bond lengths of 2.666 ± 0.041 Å, confirming that high sphericity, balanced coordination, and 54-61 bond counts correlate strongly with stability.

Perturbation analysis validated XGBoost's robustness, showing low sensitivity (0.272 ± 0.227 eV/Å) and stable predictions under atomic displacements up to 0.30 Å. However, all top tree models exhibit severe overfitting with near-perfect training scores but 8-19% performance drops on test data, indicating aggressive hyperparameters. For deployment, XGBoost is recommended with stronger regularization: reduce tree depth to ≤ 6, limit estimators to 500-1000, and enhance penalties. Expected production performance is $R^2 = 0.91 \pm 0.02$. Ensemble averaging with Gradient Boosting and LightGBM would provide additional stability, while Ridge Regression serves as an interpretable fallback for stakeholder communication.

Reference:

For full comprehensive result please open this link

https://drive.google.com/drive/folders/1tvv68OQYxp2-tRl7jc5OoBgmpykY5RW5?usp=share_link

for github repo

https://github.com/wilbert-t/AIAC.git