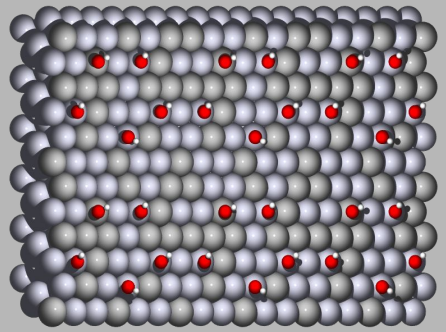
UNIVERSITY OF COPENHAGEN

Catalyzing Discovery: Machine Learning in Binary Alloy Catalysts

Mads, Mailde, Simon, Danielle

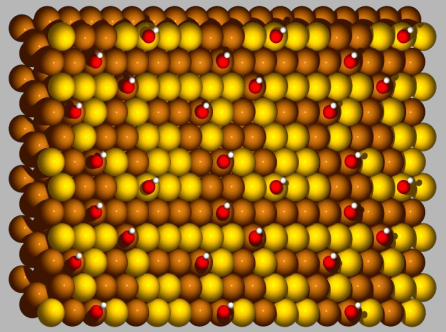
Why Binary Alloy Catalysts?

Binary alloy catalysts are special mixtures of two metals that speed up reactions, helping green energy technologies like clean fuel cells work more efficiently and sustainably.

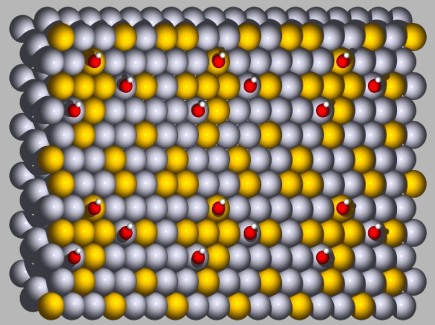


$Ag_xPt_{1-x}/Pt(111)$

$Au_xCu_{1-x}/Cu(111)$

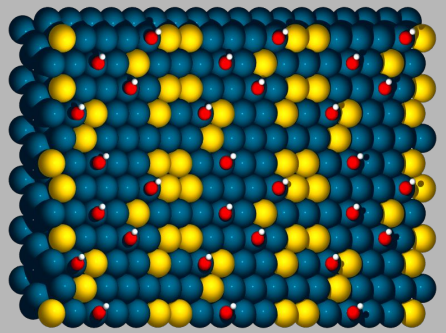


Guest composition ranging from 0 to 100% on surface: $0 \leq x \leq 1$.



$Au_xPt_{1-x}/Pt(111)$

$Au_xPd_{1-x}/Pd(111)$

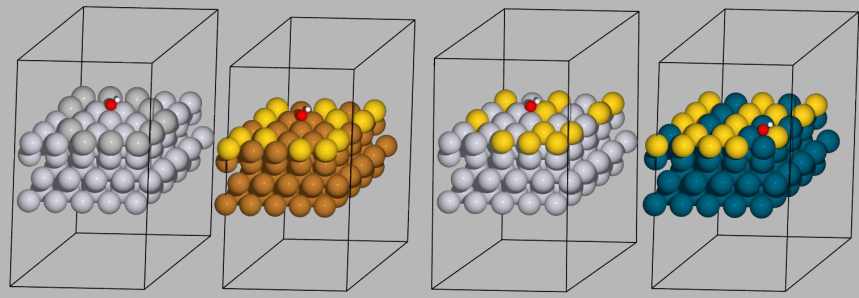


Data Description

The raw data stored in ADS_JOINED.db and SLAB_JOINED.db ASE databases consist of 2402 atomic-scale structures each: ADS_JOINED.db includes adsorption systems with adsorbates on surfaces, while SLAB_JOINED.db contains the matching clean slabs without adsorbates. The ASE database default info: atomic structure (positions, symbols), Calculator results for energies and forces, Metadata (unique id).

FEATURES GENERATED USING PYTHON

- **Elemental and Surface Information:** element_below_o, ads_type, ads_label, subsurface_label, surface_most_abundant, surface_least_abundant
- **Lattice and Geometric Parameters:** lattice_host, lattice_guest, Ad_ads, Ad_slab, ads_area_if_ontop, respective_slab_area
- **Neighbor Information:** neighbor_host_ads, neighbor_guess_ads, neighbor_host_slab, neighbor_guess_slab
- **Energetics:** ads_energy
- **Structural Shells:** second_shell_host, second_shell_guest
- **Entropy and Probabilities:** local_entropy, P_au, P_pt, P_ag, P_cu, P_pd

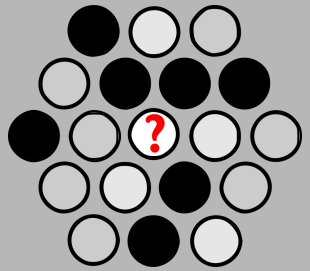


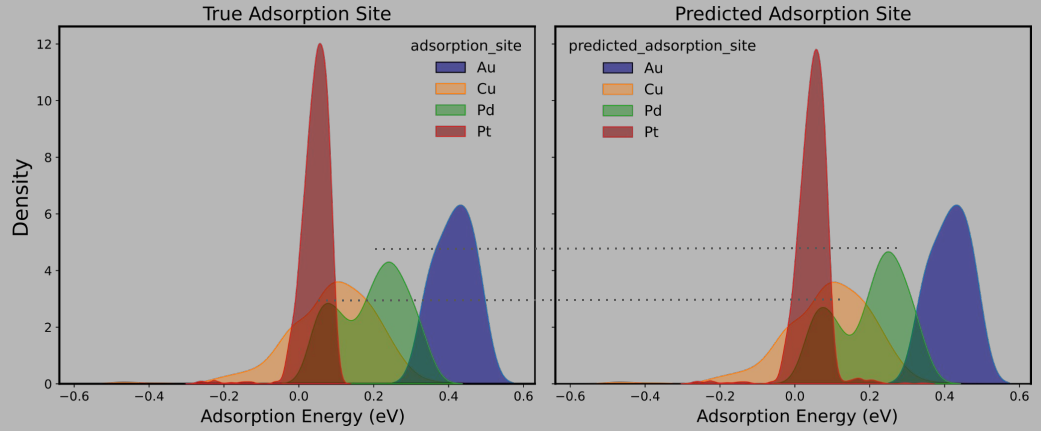
Examples of calculated structures

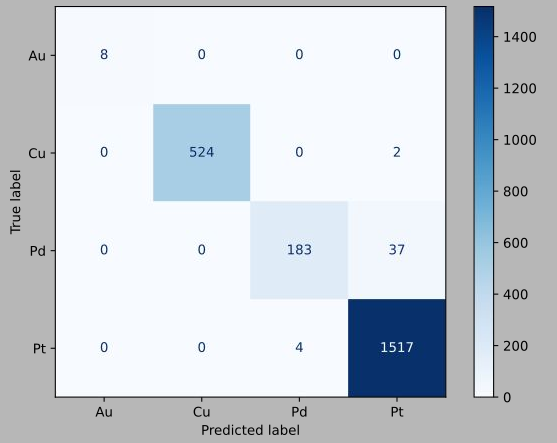
Data Source: Center for High Entropy Alloy Catalysis, University of Copenhagen
Data Availability: Partially available in *Journal of Catalysis*, Vol. 433 (2024), Article 115484; and Vol. 443 (2025), Article 115988.

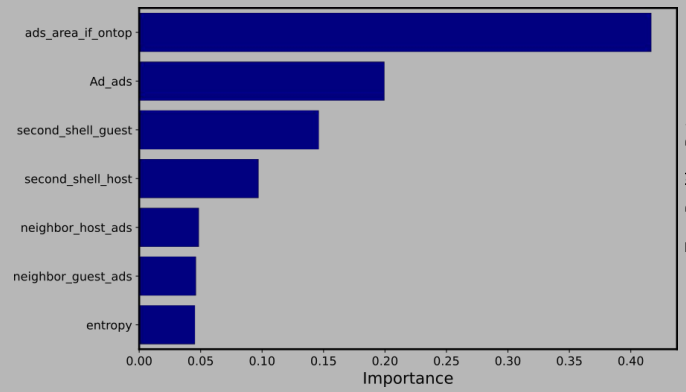
Guess the Adsorption Site

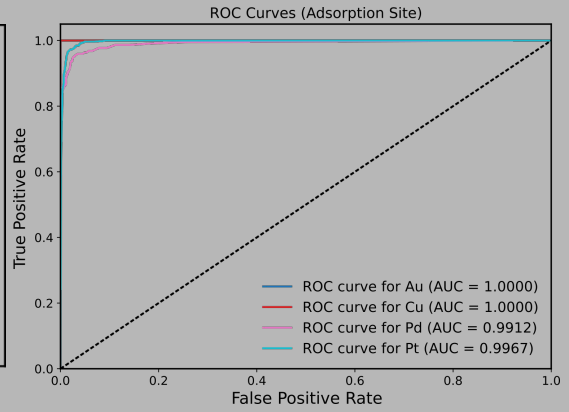
Random Forest classifier to predict the adsorption site given few surface features: Guess the Adsorption Site considering only 7 features: Ad_ads, ads_area_if_ontop, neighbor_host_ads, neighbor_guest_ads, second_shell_host, second_shell_guest, entropy. Without consider adsorption energy information, only surface structural features!







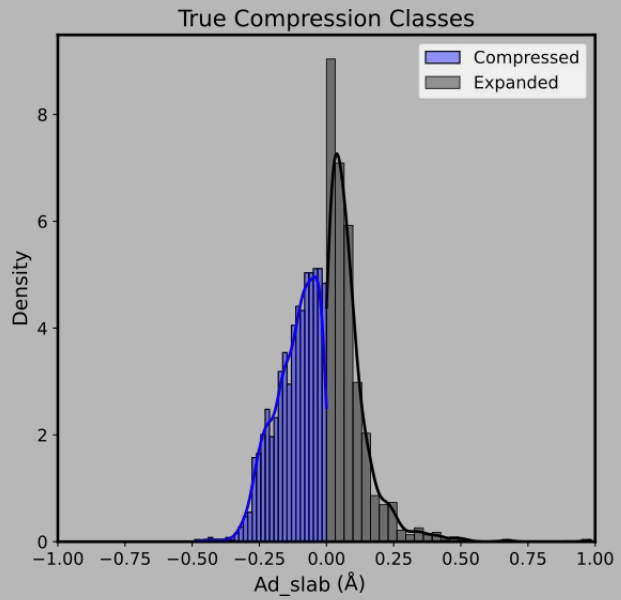


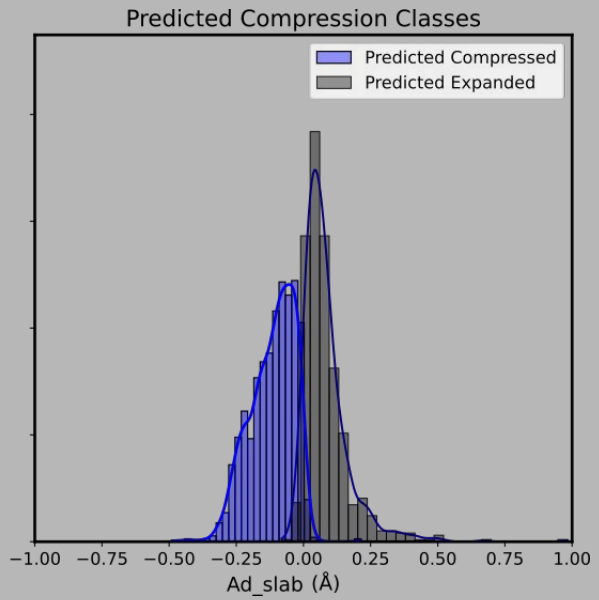


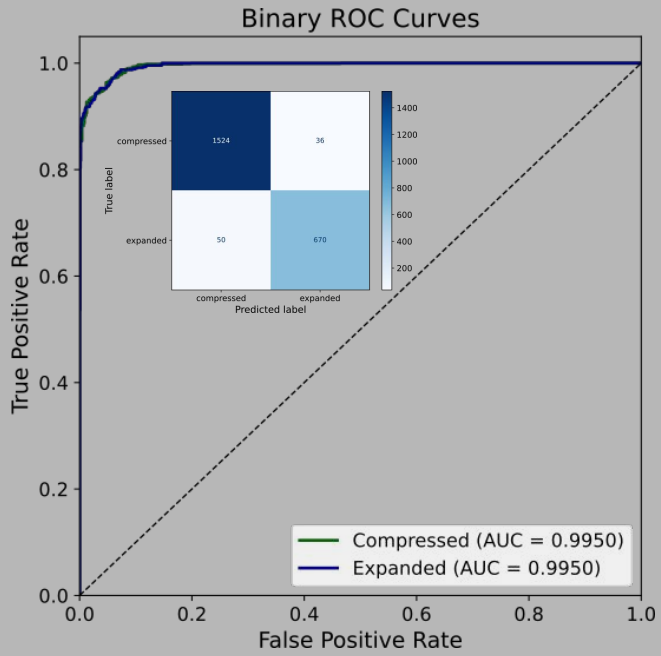
The distributions of the true and predicted adsorption sites are very similar. However, there are a few discrepancies in the classification of some Pd and Pt adsorption sites, which may be due to their similar properties and the comparable surface deformations they cause.

Classification of Compressed and Extended Classes

- Machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’**) of the adsorption site using **LGBMClassifier**.
- **Drops domain-specific and target-related columns** to avoid data leakage.
- Uses **Optuna** to tune LightGBM hyperparameters via 5-fold CV.

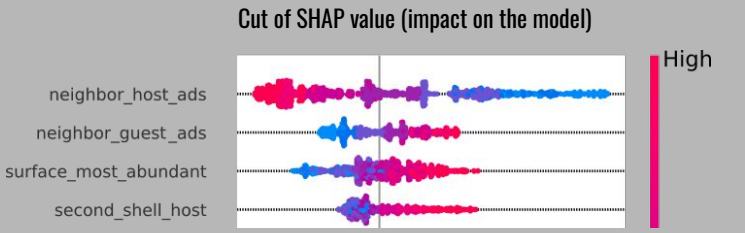


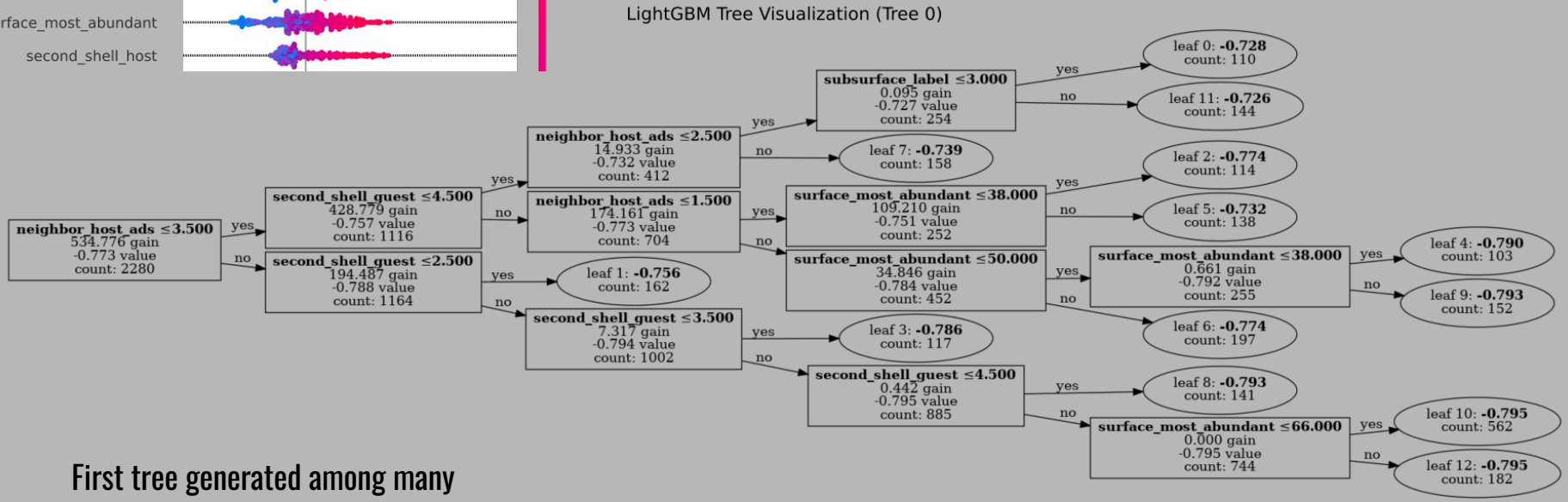




Classify the binary alloys as “compressed” or “expanded” relative to the non-alloyed.

CLASSIFICATION COMPRESSED AND EXTENDED CLASS





First tree generated among many

lgb.plot_tree(booster, tree_index=0, figsize=(20, 10), show_info=['split_gain', 'internal_value', 'internal_count', 'leaf_count'])

Graph Convolutional Network (GCN)

Geometric representation of data into graph

Reduced 5x5x4 + adsorbate atomic structure into 5x5 + substrate + adsorbate nodes

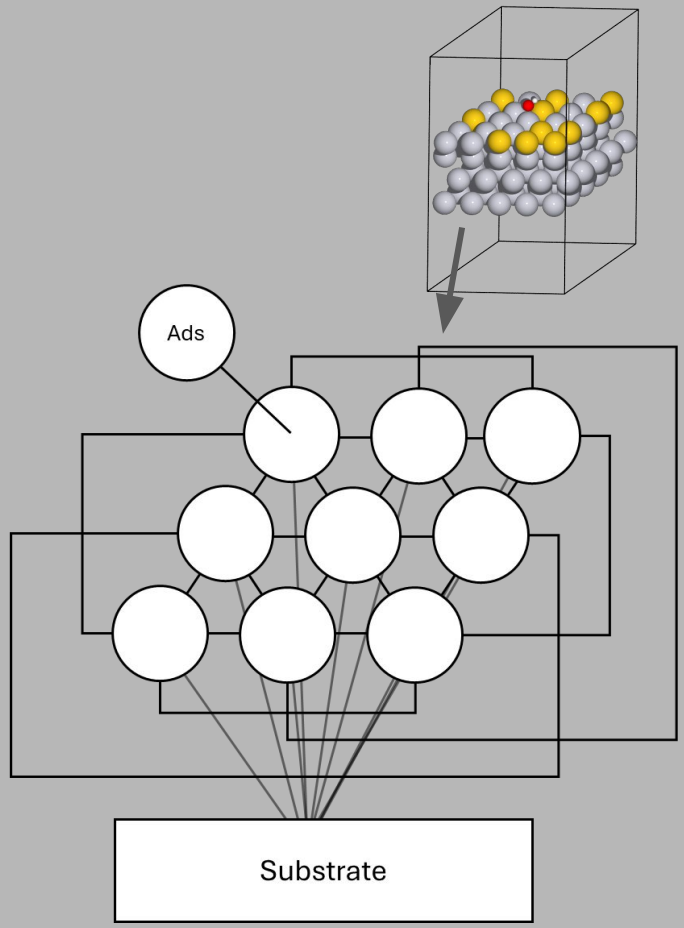
Nodes are connected to neighbor atoms

Each node has a set of features (7):

- One-hot-encoded element (Ag, Au, Cu, Pd, Pt, OH)
- Layer feature: Adsorbate=0, surface layer=1, and substrate=2

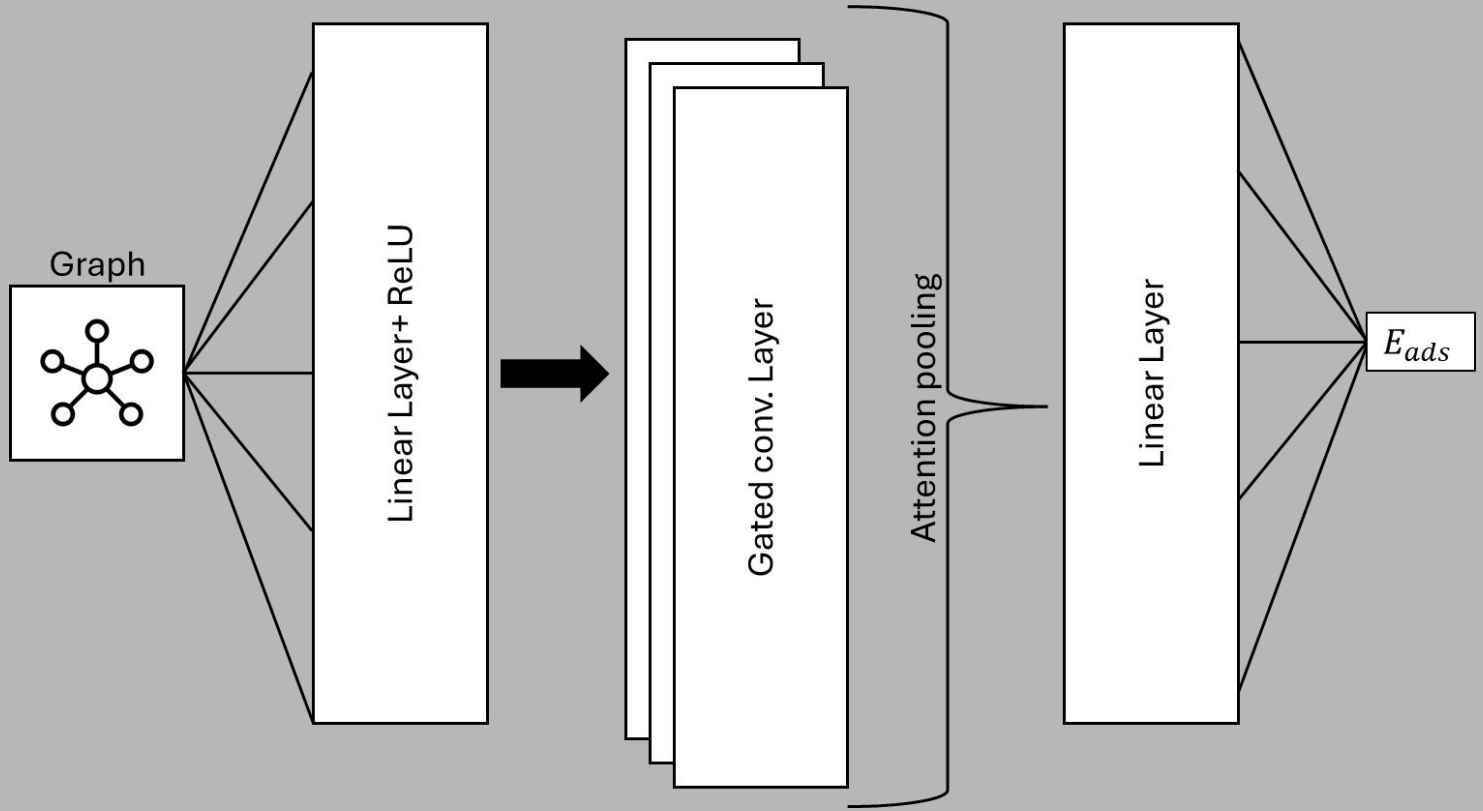
We only consider on-top adsorptions (OH has one connection)

Substrate connected to all surface layer atoms



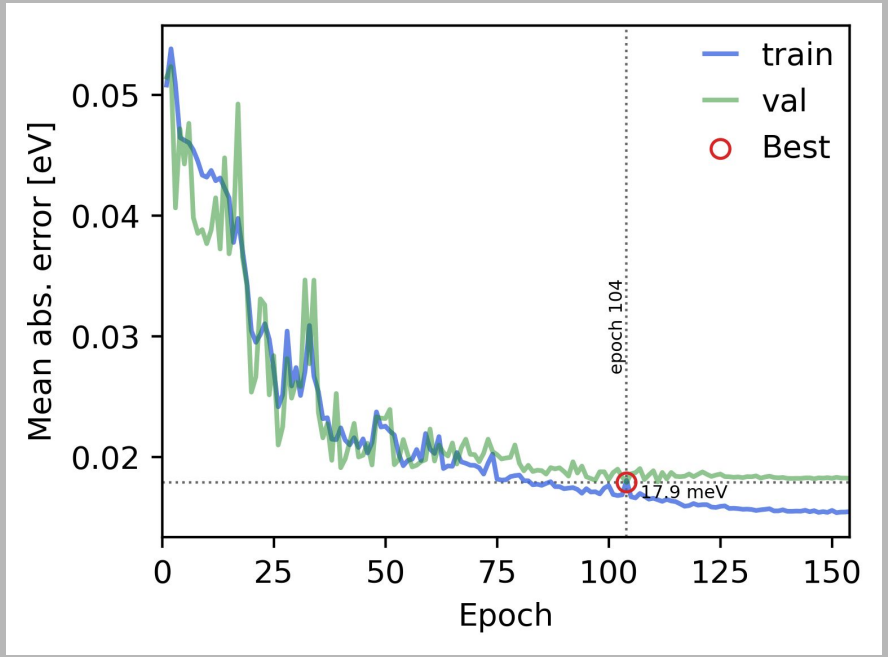
(3x3 as example)

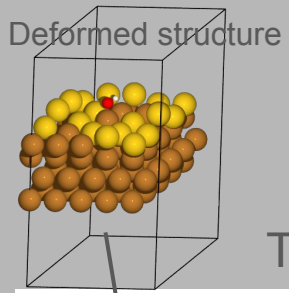
GCN Architecture



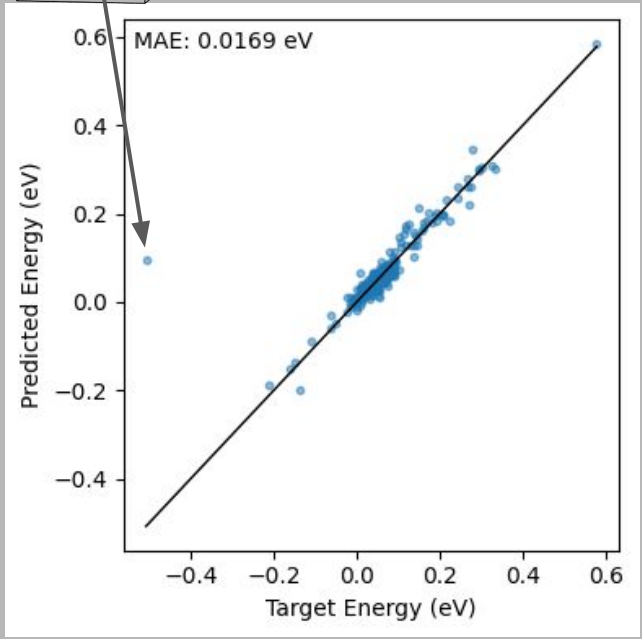
Test results (80/10/10 split)

- Using parameters from hyper parameter optimization





Test set



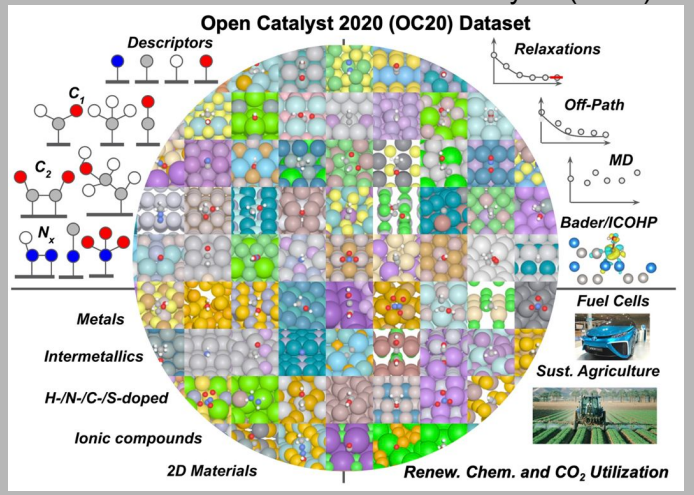
L. Chanussot et al. ACS Catalysis (2021)

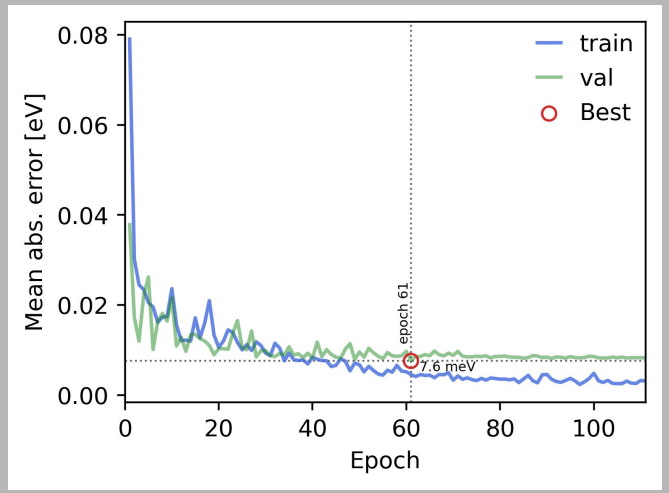
Finetune Pretrained OCP Model

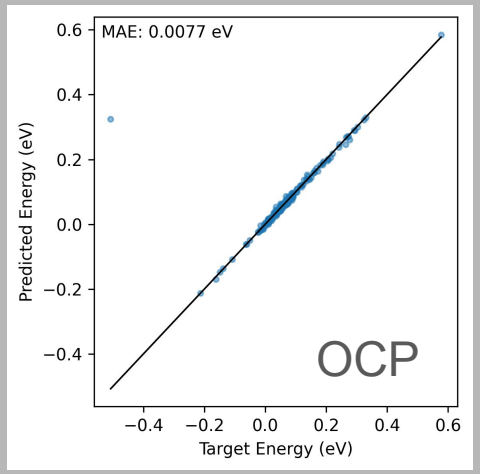
Equiformer V2 with 31 M parameters

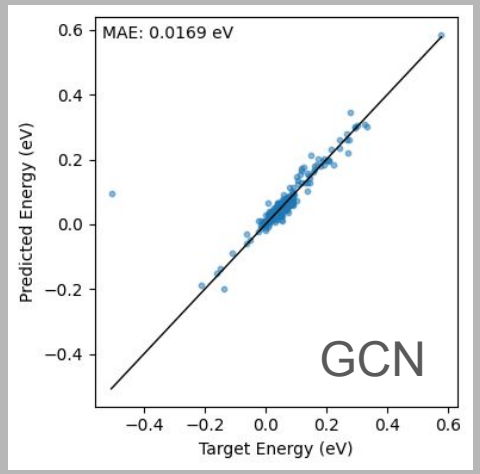
Open catalyst project (OCP) by Meta AI: Dataset contains 1,281,040 DFT relaxations

Ordered structures -> Finetune to our binary disordered structures









test set

Prediction of the *OH Adsorption Energy With LGBMRegressor

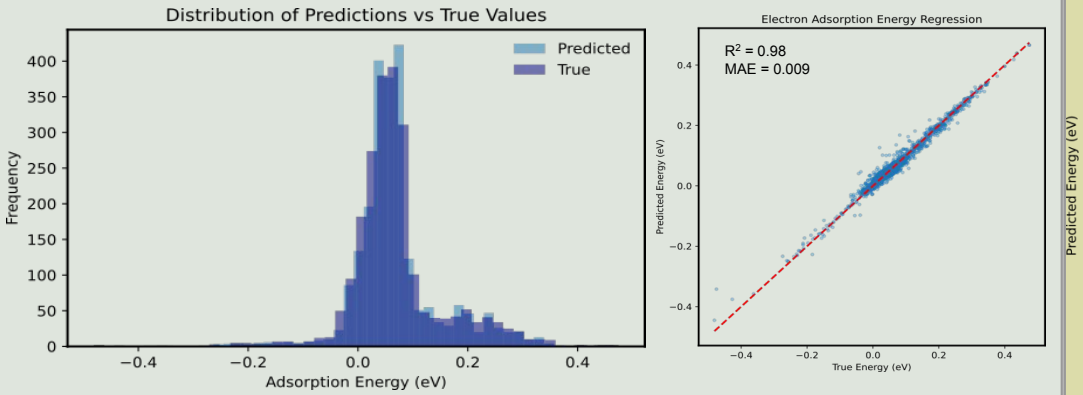
Cross Validation, k = 5

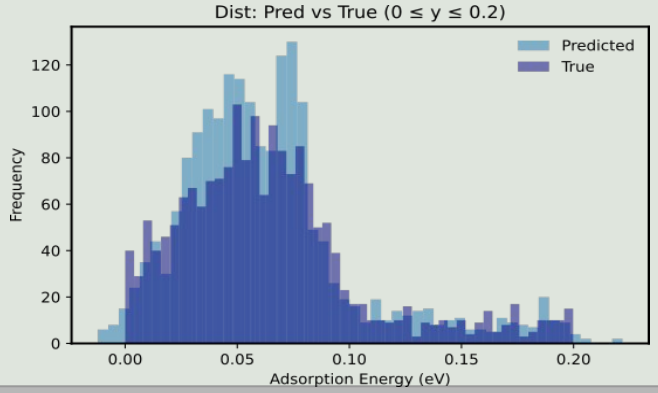
Features	MAE	RMSE	R^2
Catalyst features	0.011	0.019	0.95
Catalyst features + DFT structural features	0.014	0.025	0.92
Catalyst features + Feature engineering	0.012	0.020	0.95
Catalyst features + DFT features + Feature engineering	0.009	0.013	0.98

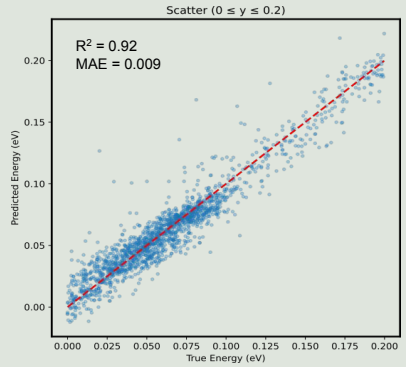
Prediction of the *OH Adsorption Energy With LGBMRegressor

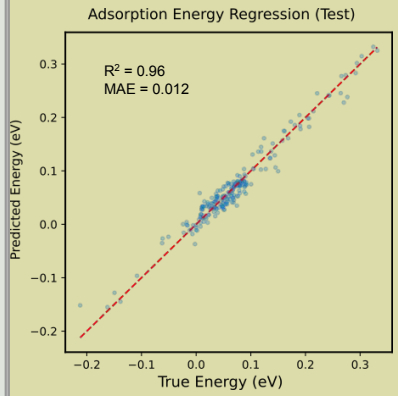
(Catalyst features + DFT features + Feature engineering)

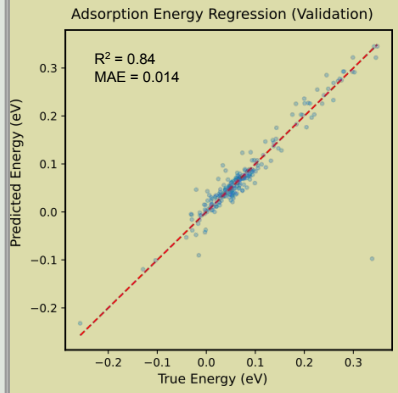
5-fold CV → Best optuna parameters → Train-Val-Test (80:10:10)











- **Feature Engineering**
- **Hyperparameter Tuning** with Optuna
- **Target**: ads_energy
- **Filters**: ads_type == 'ontop' and -0.5 ≤ ads_energy ≤ 0.5
- Applies arcsinh() transform to the target
- Analyzes performance on training data and in a focused range (0 ≤ ads_energy ≤ 0.2)
- Retrains with best Optuna parameters and compares validation and test results

Predicting Adsorption Energy of OH on Binary Alloy Catalysts

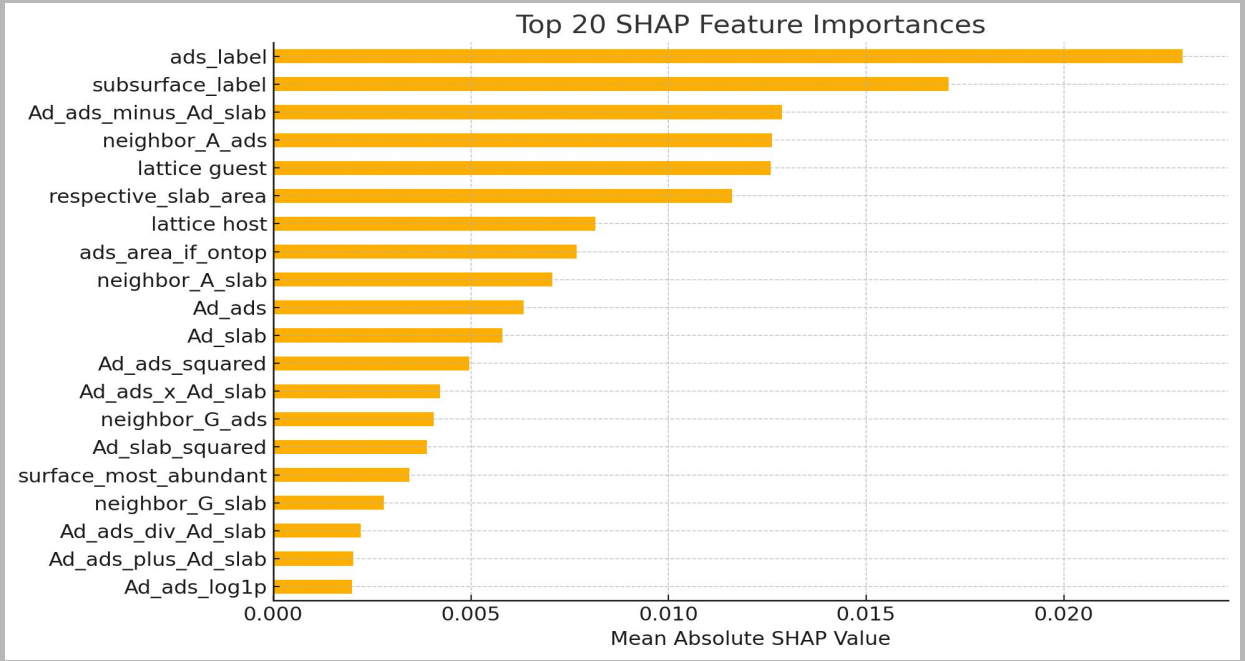
- **Purpose:** Accurate energy predictions help identify optimal alloy compositions for catalytic performance (e.g., in fuel cells).
- **Method:** Uses a regression model (LightGBM) trained on structured features extracted from catalyst surfaces.
- **Focused Feature Set:** Uses only `Ad_ads` and `Ad_slab` as core features, minimizing noise and overfitting.
- **Data Transformation:** Applies `arcsinh` transformation to the target to normalize extreme values.

Metric Comparison

Features	MAE	RMSE	R^2
Ad_slab + Ad_ads	0.016	0.036	0.93
Catalyst features + DFT structural features	0.014	0.025	0.92

- **Simpler Model:** Lacking Enough Geometric and Electronic Context
- **Lack of DFT Derived Structure Info:**
  - No `ads_type`, `element_below_o`, or slab surface geometry
  - Misses critical factors affecting adsorption energy (e.g., coordination, strain).
- **Higher Error:**
  - Higher **Relative MAE** and **RMSE** suggest the model is less precise across different scales.
- **R^2 Is Misleading:**
  - Although R^2 is slightly higher, other metrics reveal **poorer absolute accuracy**.

Feature Importance



- **`ads_label`** is the most influential feature, highlighting the critical role of the specific adsorbate in determining adsorption energy.
- **`subsurface_label`** contributes significantly, suggesting the atomic identity or composition just beneath the surface strongly impacts reactivity.
- **`Ad_ads_minus_Ad_slab`** and **`Ad_ads_x_Ad_slab`** rank highly — non-linear interactions between adsorbate and slab descriptors are highly predictive.

Predict OH Adsorption Energy on Binary Alloy Monolayers (Ontop, Bridge, Hollow)

- **Multi-site coverage:** While previous models were trained only on “ontop” adsorption sites, this model includes **hollow** (adsorbate sits in a three-fold coordinated pocket) and **bridge** (adsorbate spans two adjacent surface atoms) sites, which should allow for the model to learn geometry-dependent energy trends.

- **Broader coverage of coordination environments:** Adsorption energies can vary dramatically with local atomic coordination—bridge sites have two-fold coordination and hollow sites three-fold—so including all three site types ensures the model learns these distinct chemical interactions rather than overfitting to just one geometry.

- **Model enrichment through all site types:** While ontop adsorption is confirmed via spectroscopic transmission microscopy in this system, including bridge and hollow sites helps the model learn richer geometric and energetic patterns from non-top environments, improving its overall predictive capability.

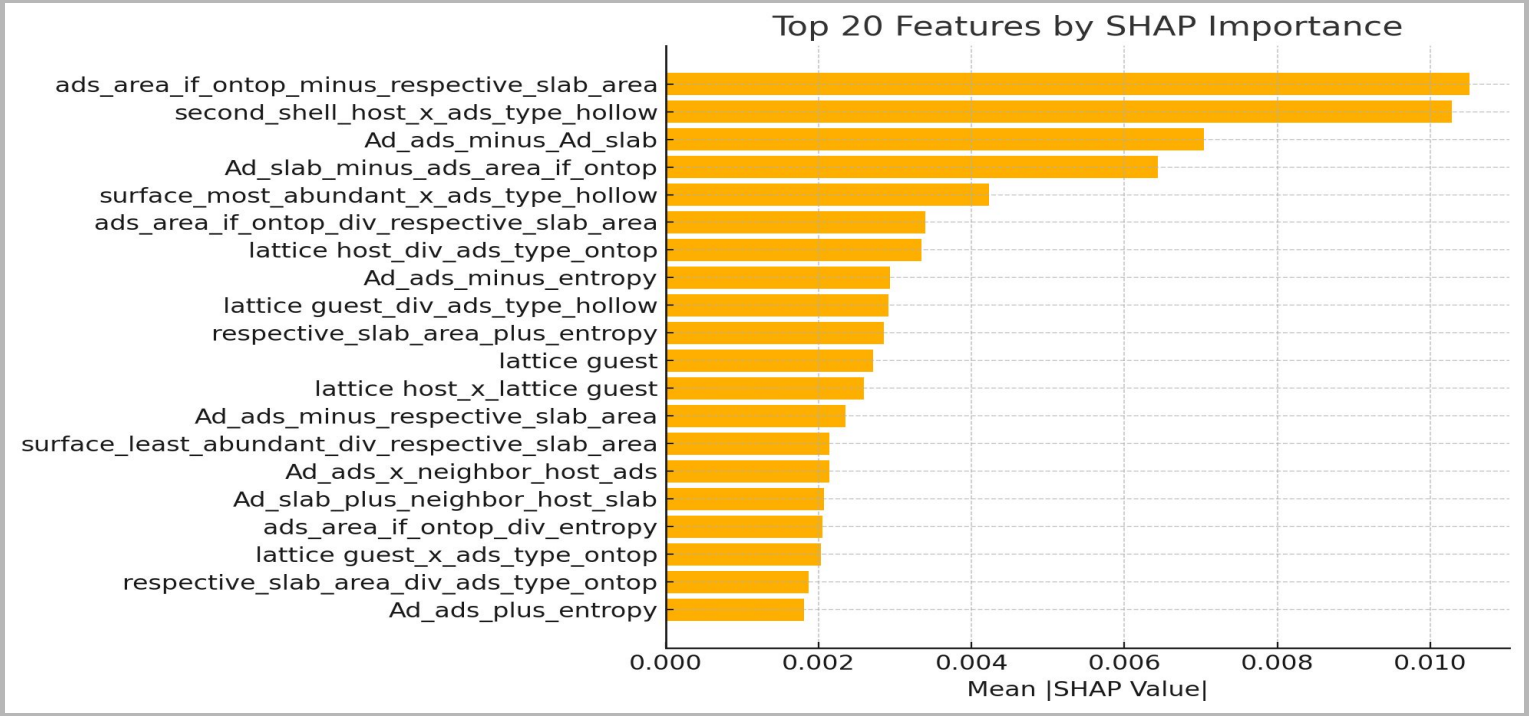
-

Performance Comparisons

Features	MAE	RMSE	R^2
Catalyst features	0.011	0.019	0.95
Catalyst features + DFT structural features	0.014	0.025	0.92
Multi-Site Model (ontop + bridge + hollow)	0.013	0.021	N/A
Catalyst features + Feature engineering	0.012	0.020	0.95

- Sits between Catalyst features + DFT structural features and Catalyst features + Feature engineering

Feature Importance



`ads_area_if_ontop_minus_respective_slab_area` is the most influential feature, highlighting how the net change in surface area upon adsorption on ontop sites critically determines adsorption energy.

`second_shell_host_x_ads_type_hollow` contributes significantly, suggesting that the chemical identity of the second-shell metal in hollow sites plays a crucial role in modulating binding strength.

`Ad_ads_minus_Ad_slab` and `ads_area_if_ontop_div_respective_slab_area` rank highly — non-linear difference and ratio metrics between adsorbate and slab geometries capture essential physical interactions predictive of adsorption energetics.

CLUSTERING

What do we expect to see and what do we hope to see

1. Slab configuration
2. Compressed or uncompressed surface
3. Something else

Multiple Models tried, Agglomerative Clustering worked the best

Elbow method used to determine the best number of clusters

Feature selection

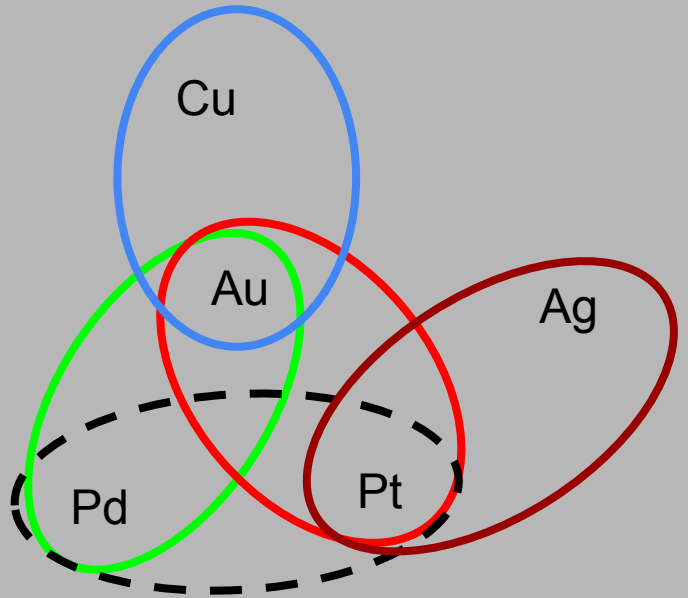
- Lattice parameter features
- Surface composition features
- Energy features

Many highly correlated features,
i.e number neighbouring_host_atoms and surface composition

Many slabs share multiple features

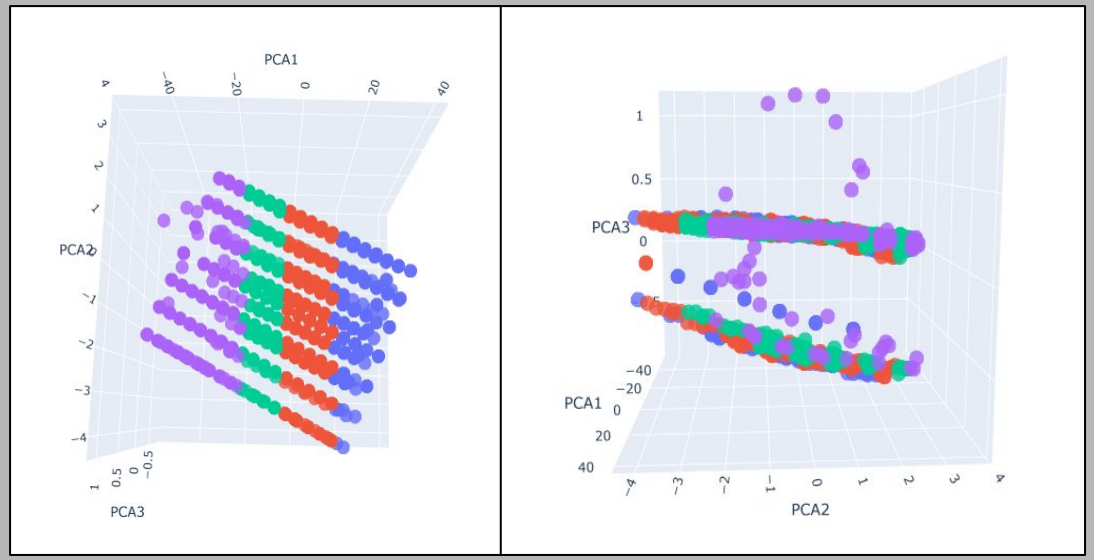
Manual feature selection to try and mitigate this

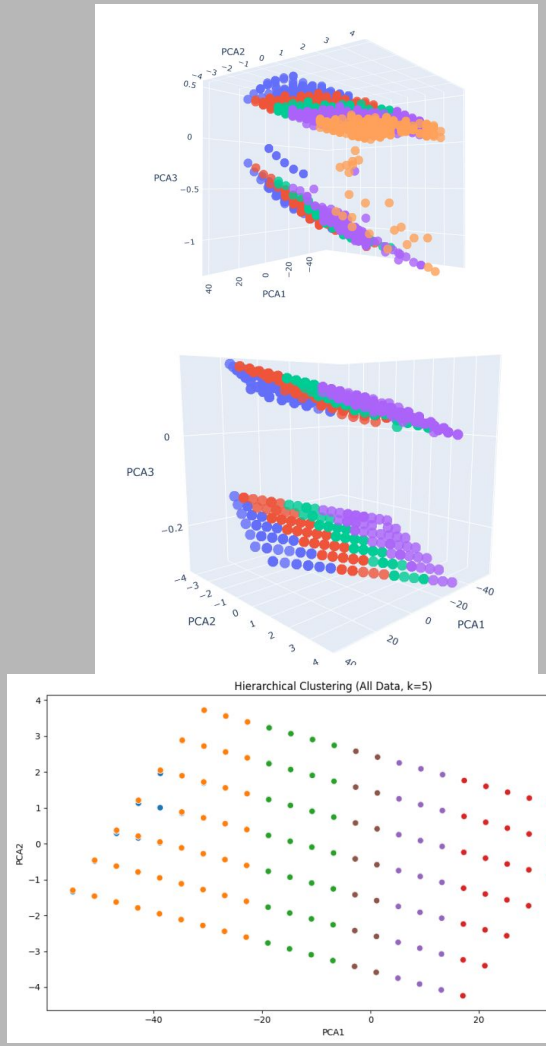
Lattice parameters in the features



It all the same

More or less every combination of features give the same result



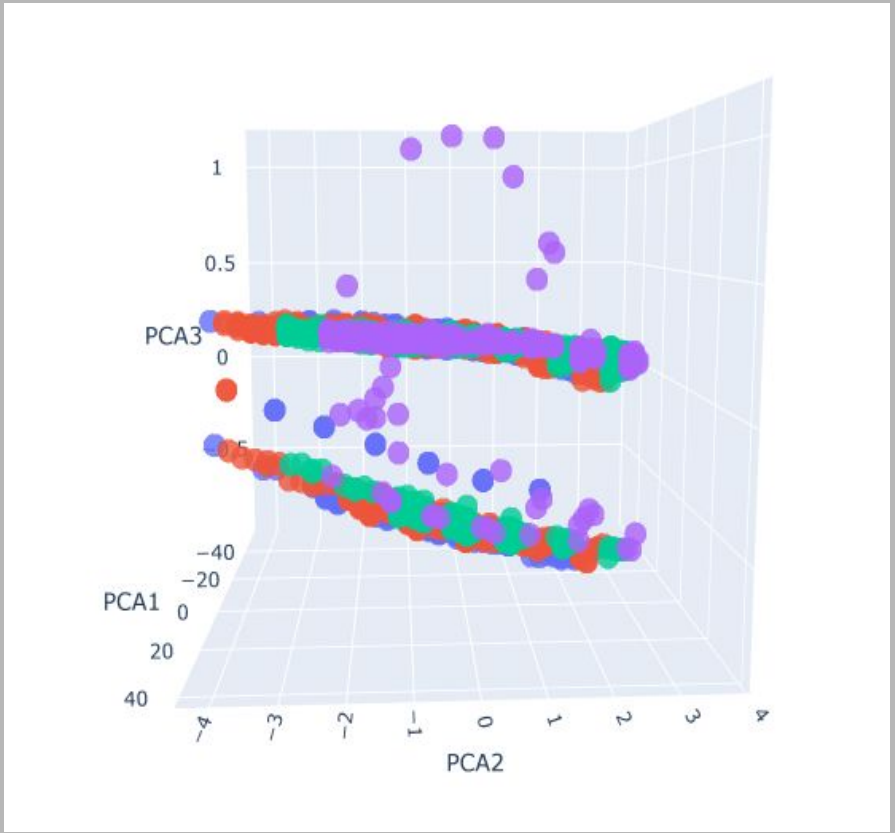


Two layers form in the Z-axis

It shows what the atom-size of the dominant element

With more different base slabs, a different pattern might have emerged

Outliers are related to strain



Make a new feature-set

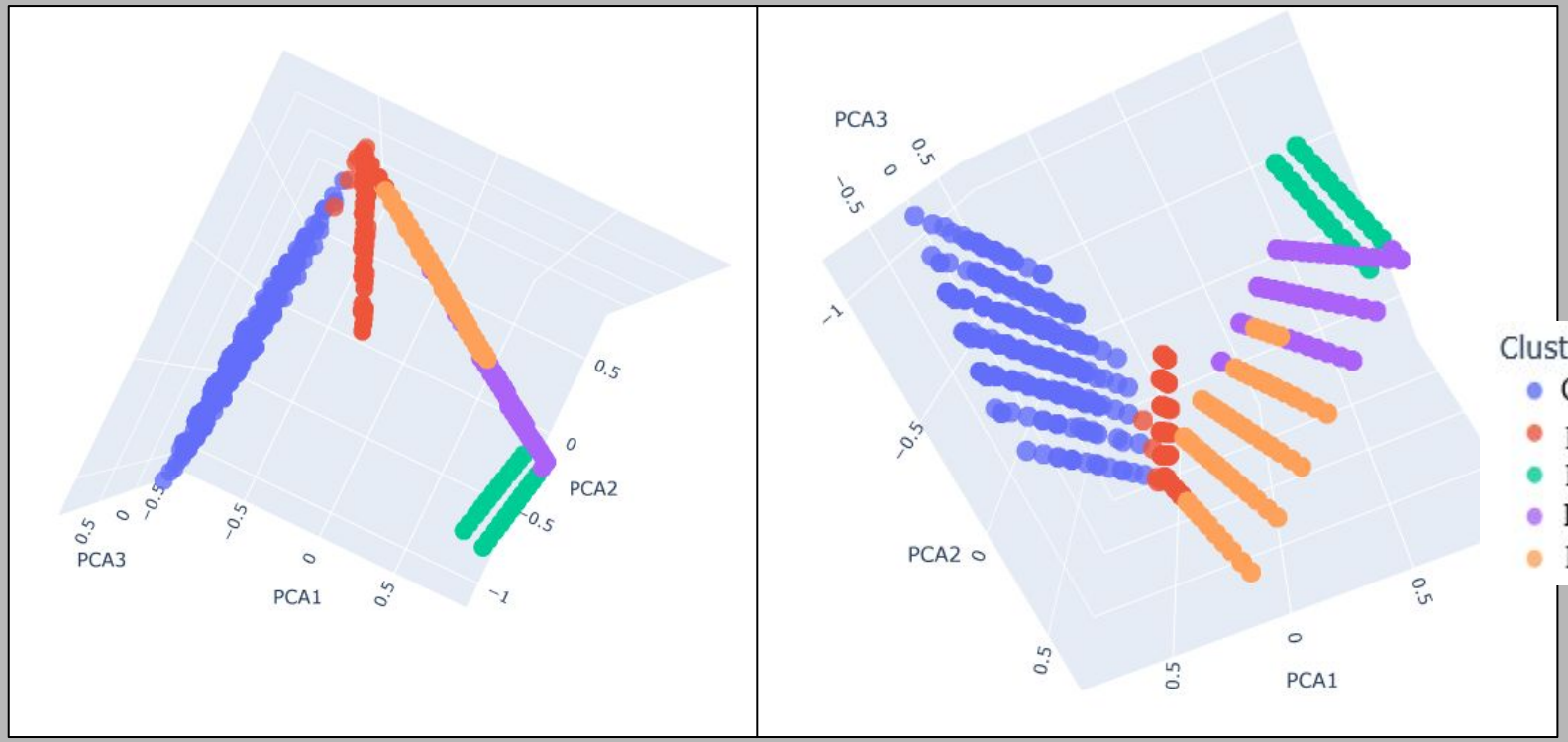
Try and make a new feature-set to describe the distribution of elements

Makes the features more distinct

Element 1	Element 2	Neighbour atoms 1	Neighbour atoms 2
Pd	Au	4	2
Pt	Ag	4	2

Pd	Au	Ag	Cu	Pt
0.66	0.33	0	0	0
0	0	0.33	0	0.66

Catagorise the catalyst surface configuration



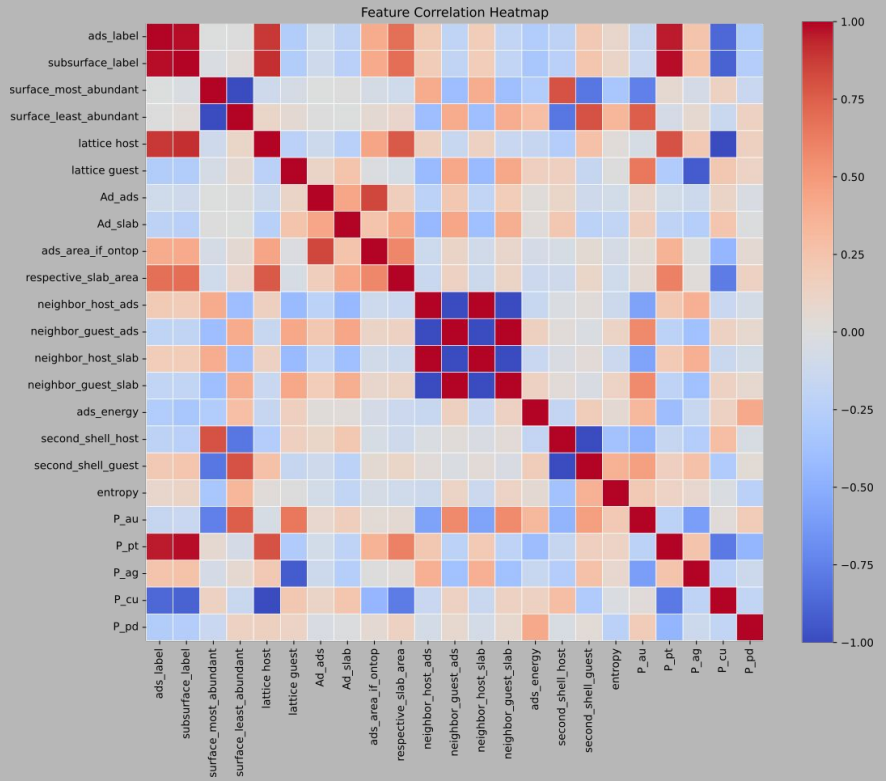
CONCLUSIONS

Clusters: The similarity of the features presented a challenge. A few key features where discernible from the clustering, however more distinct slabs are probably needed for good clustering

Graph models: Deformation is a challenge. Finetuning the pre-trained model performs best. More data might be needed to improve GCN performance.

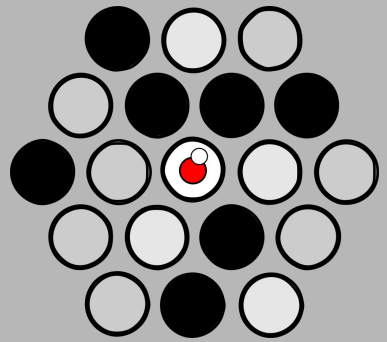
Tree-models: Deformation is also a challenge in tree-based methods, both for classification and regression. However, even with structural deformation, we could obtain relevant information about the catalysts providing very few information. This reveals that ML is a powerful method to reduce computational time of catalysts simulations.

APPENDIX-A
Tabular Data Features Description

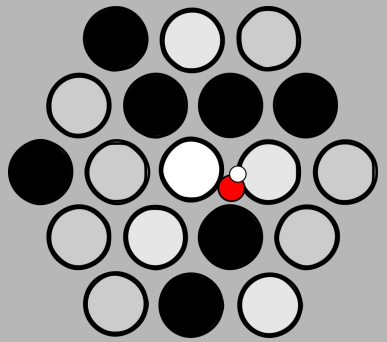


APPENDIX-A
Tabular Data Features Description

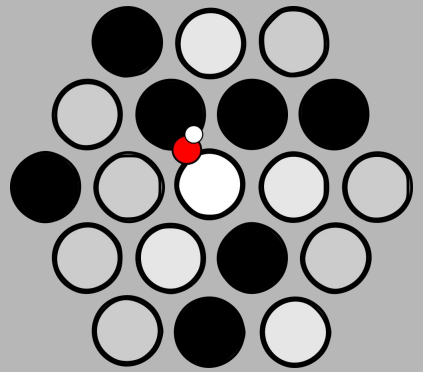
- Adsorption Site Description



Ontop sites



Hollow sites



Bridge sites

95% of data

5% of data

APPENDIX-A: List of Features Generated for Tabular Data and their description

- **element_below_o:** “Au”, “Cu”, “Pt”, “Ag” or “Pd” adsorption sites. The chemical element directly beneath the adsorbed oxygen (O).
- **ads_type:** Type of the adsorption site “ontop” and “non-ontop”, “non-ontop” refers to the adsorptions on “hollow” and “bridge”. A categorical indicator of the geometrical position.
- **Ads_label**: integer label for the adsorption site, element_labels = {'Cu': 1, 'Pd': 2, 'Ag': 3, 'Pt': 4, 'Au': 5}
- **subsurface_label**: Encodes the material composition below the surface.
- **surface_most_abundant**: Encodes the most abundant surface element.
- **surface_least_abundant**: Encodes the least abundant surface element.
- **lattice host**: Lattice constant (in Å) of the host structure.
- **lattice guest**: Lattice constant (in Å) of the guest material.
- **Ads_energy**: Adsorption energy of OH in eV — representing the binding strength of OH on the surface.
- **ads_area_if_ontop**: Area occupied by the adsorption site specifically if the OH adsorbate is on ontop site.
- **respective_slab_area**: Area occupied by the adsorption site when OH is not adsorbed.
- **neighbor_host_ads**: Count of neighboring host atoms near the adsorbate if OH is adsorbed.
- **neighbor_guest_ads**: Count of neighboring guest atoms near the adsorbate.
- **neighbor_host_slab**: Count of neighboring host atoms in the slab.
- **neighbor_guest_slab**: Count of neighboring guest atoms in the slab.
- **second_shell_host**: Number of host atoms in the second coordination shell (just outside the first six neighbors).
- **second_shell_guest**: Number of guest atoms in the second coordination shell.

APPENDIX-A: List of Features Generated for Tabular Data and their description

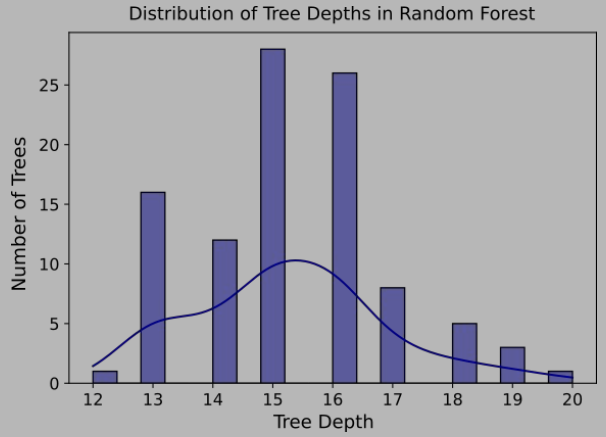
- **entropy:** local structural surface entropy (includes adsorption site, first shell and second shell surface atoms).
- **Ad_ads**: Difference in the adsorption site area between the alloyed catalyst *with* OH adsorbed and the non-alloyed catalyst. This captures changes due to both alloying and OH adsorption.
- **Ad_slab**: Difference in the adsorption site area between the alloyed catalyst *without* OH adsorbed and the non-alloyed catalyst. This captures changes due to alloying.
- **P_au:** Percentage of Au in the binary alloy catalyst.
- **P_pt:** Percentage of Pt in the binary alloy catalyst.
- **P_ag:** Percentage of Ag in the binary alloy catalyst.
- **P_cu:** Percentage of Cu in the binary alloy catalyst.
- **P_pd**: Percentage of Pd in the binary alloy catalyst.

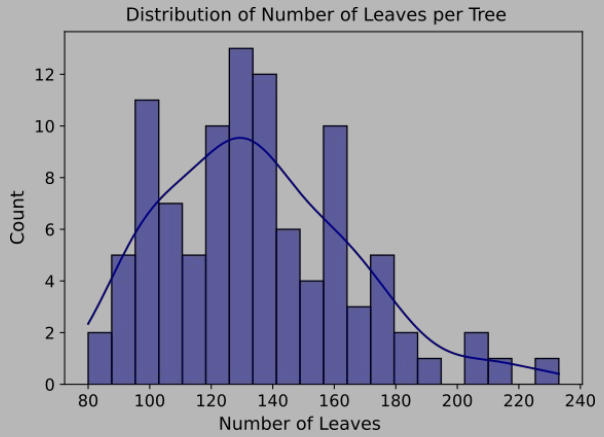
APPENDIX-B: Guess the Adsorption Site

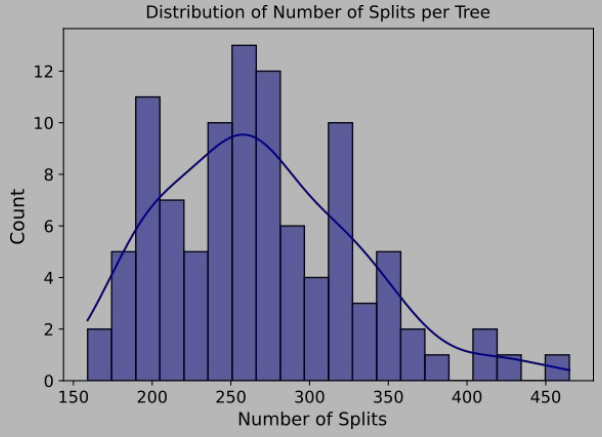
This code is a comprehensive pipeline for multi-class classification of adsorption sites ("Au", "Cu", "Pt" or "Pd") using a Random Forest Classifier.

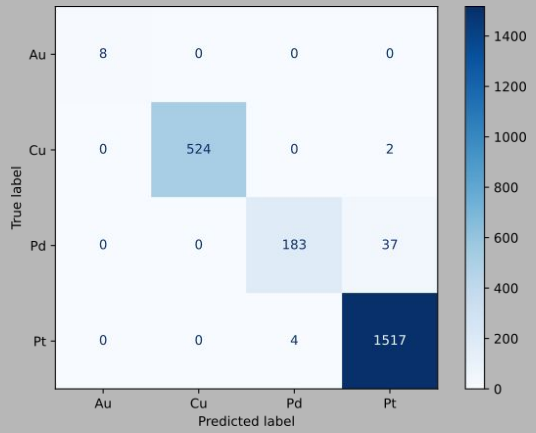
- Modeling: RandomForestClassifier, StratifiedKFold, cross_val_predict, etc.
- Reads a CSV file named "general_site_analysis_version4.csv".
- Filters data where ads_type == "ontop" and ads_energy is between -0.5 and 0.5.
- Target: element_below_o (encoded using LabelEncoder). Drops a predefined list of metadata/unwanted columns and keeps only 7 features.
- Hyperparameter Optimization (Optuna)
- 5-fold stratified CV.
- Optimization runs for 50 trials.
- Final Training and Evaluation using the best hyperparameters.
- Evaluation: metrics like f1_score, confusion_matrix, roc_curve, etc.
- Feature Importance: the top most important features.

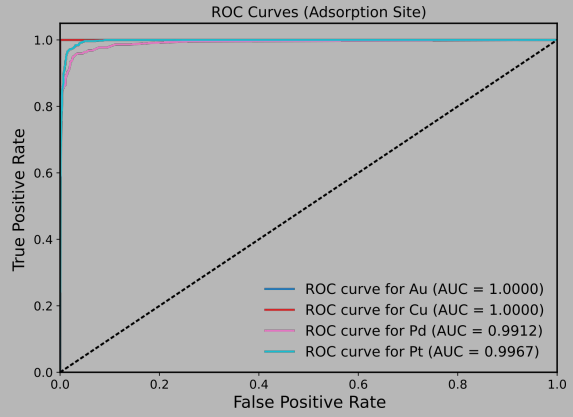
APPENDIX-B: Guess the Adsorption Site





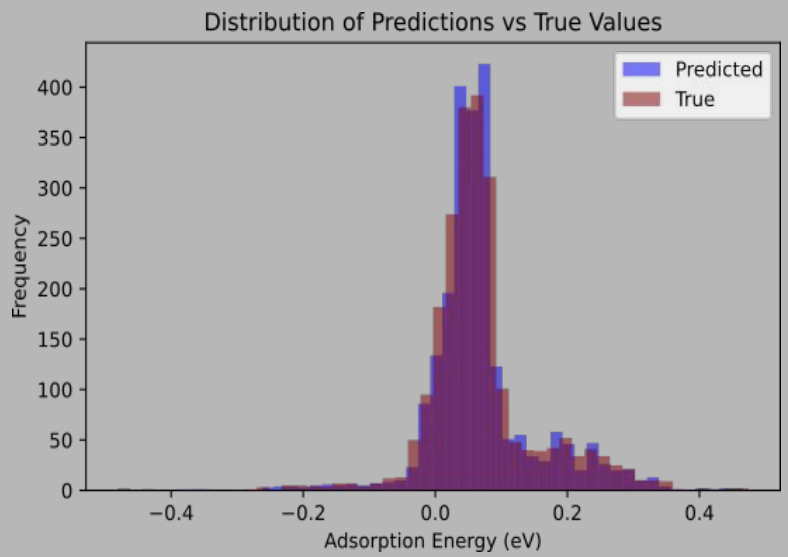


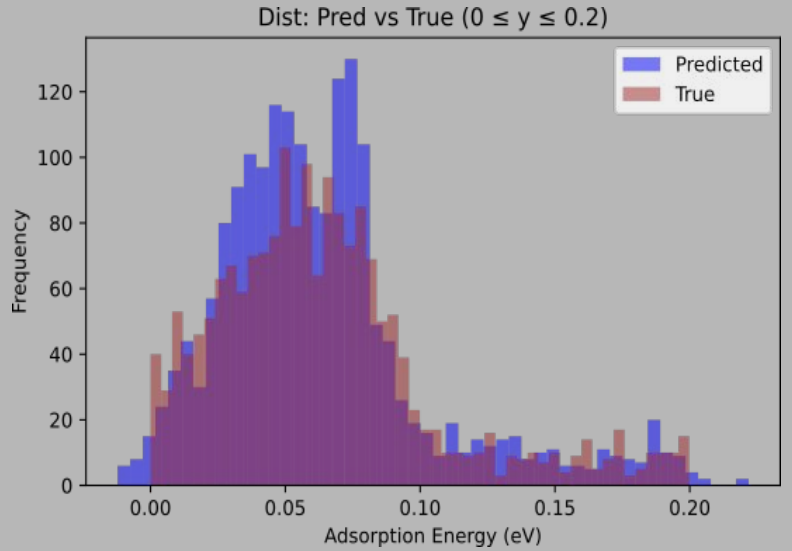




APPENDIX-C: LightGBM Regression

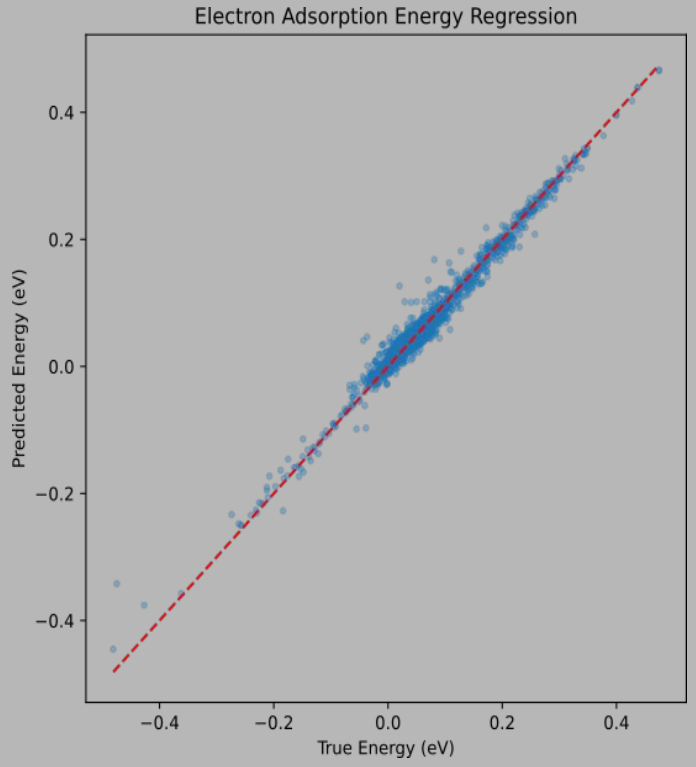
(Catalyst features + DFT features + Feature engineering)

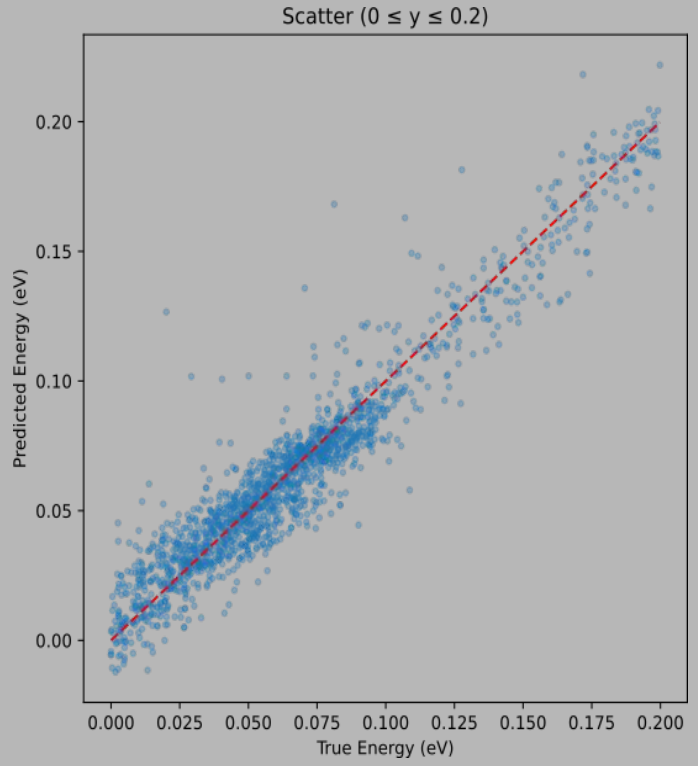




APPENDIX-C:LIGHTGBM REGRESSION

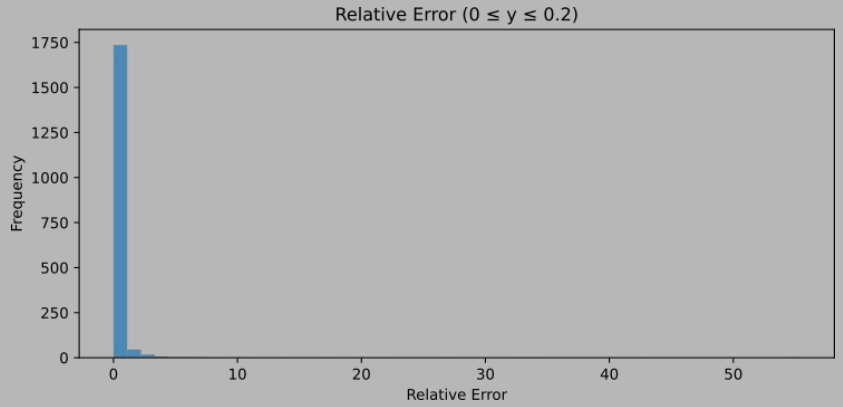
(Catalyst features + DFT features + Feature engineering)

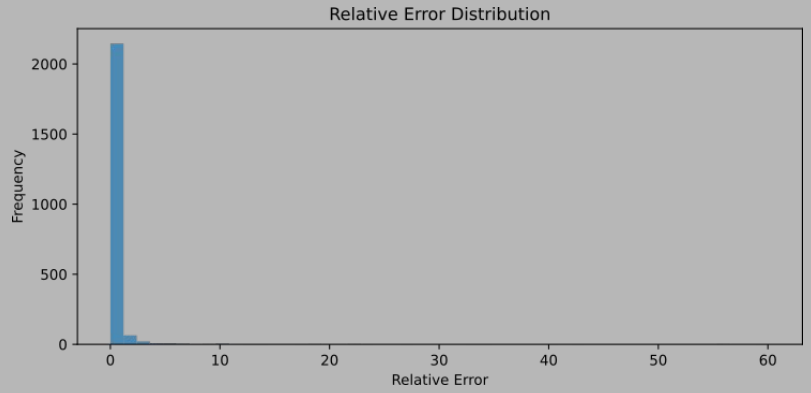




APPENDIX-C
LIGHTGBM REGRESSION

(Catalyst features + DFT features + Feature engineering)

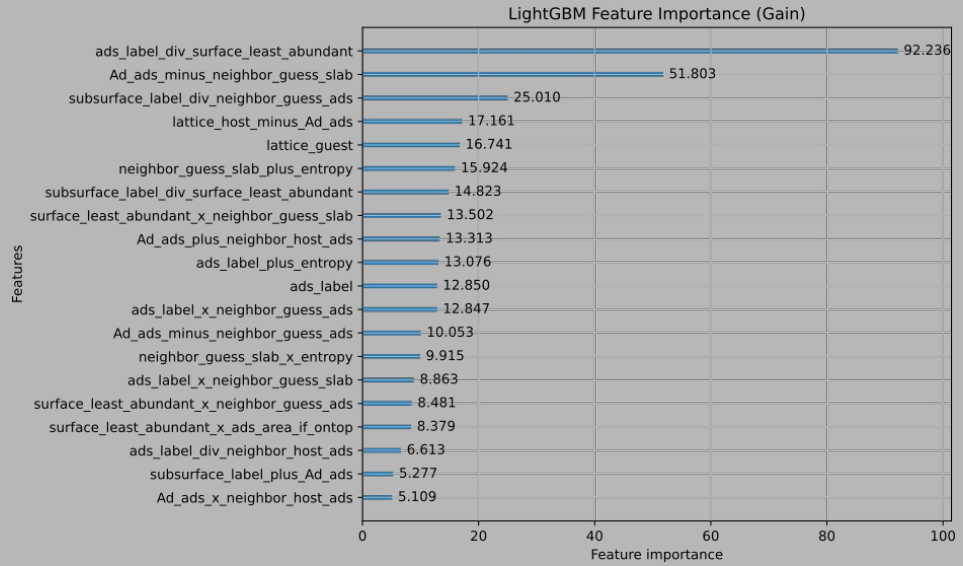


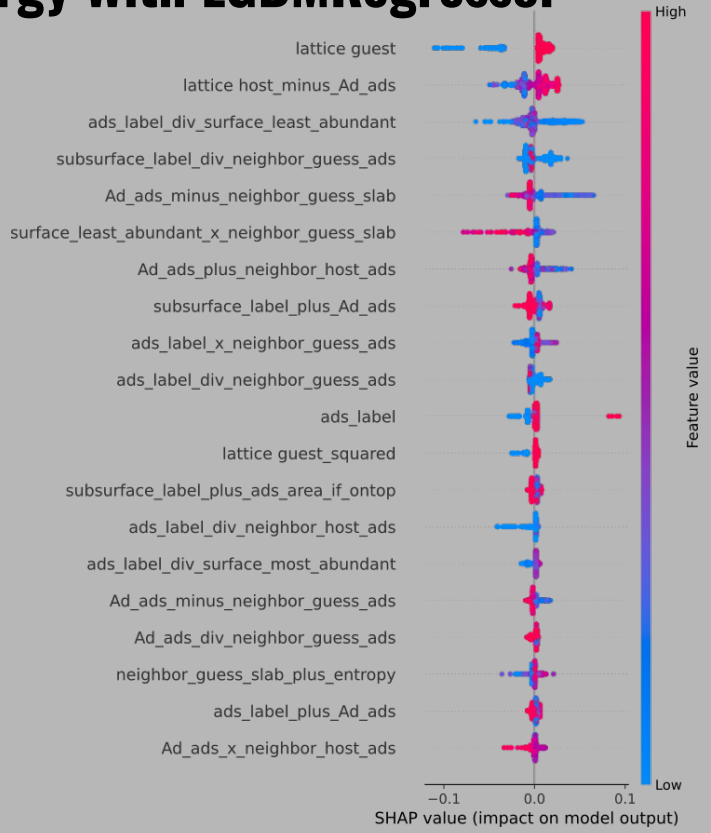


APPENDIX-C

Prediction of the *OH Adsorption Energy With LGBMRegressor

(Catalyst features + DFT features + Feature engineering)





APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is classified as **'compressed'** or **'expanded'** based on a the material features.
- Loads a CSV (general_site_analysis_version4.csv).
- Filters it to focus on "ontop" adsorptions with ads_energy in the range [-0.55, 0.55].
- Creates a new binary target label called compression_class from the Ad_slab column.
- Drops domain-specific or label-related columns (listed in truth_cols) from the feature set to avoid leakage.
- study.optimize(objective, n_trials=200, timeout=600)
- **Hyperparameter optimization** using Optuna
- Uses cross_val_predict() to evaluate the final model with the best parameters from Optuna.
- **Cross-validation** for robust model evaluation
- **Performance analysis** via classification metrics and ROC/AUC
- **Visualization** of feature importance.

APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is **'compressed'** or **'expanded'** based on a the material features.

RandomForestClassifier Summary

Best Hyperparameters

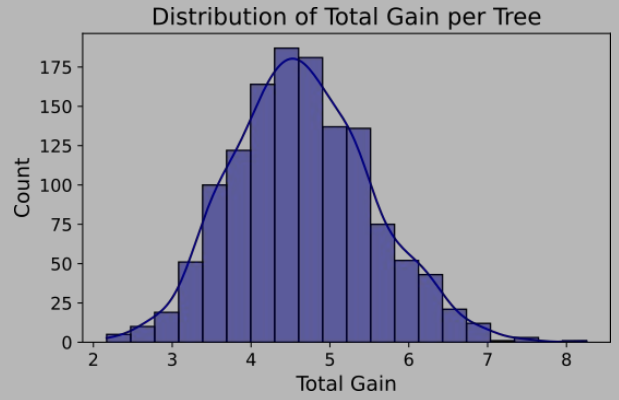
Parameter	Value
n_estimators	1320
max_depth	14
min_samples_split	10
max_features	sqrt

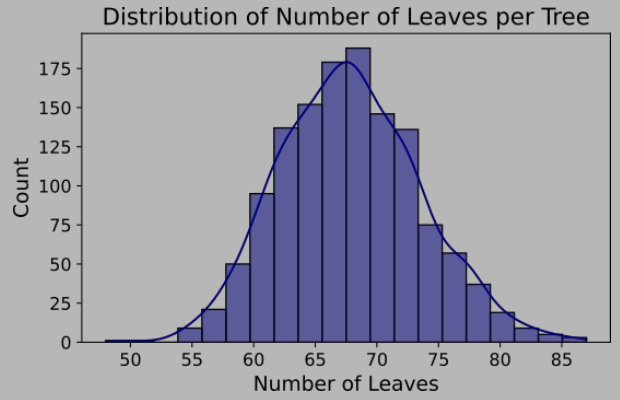
Classification Report

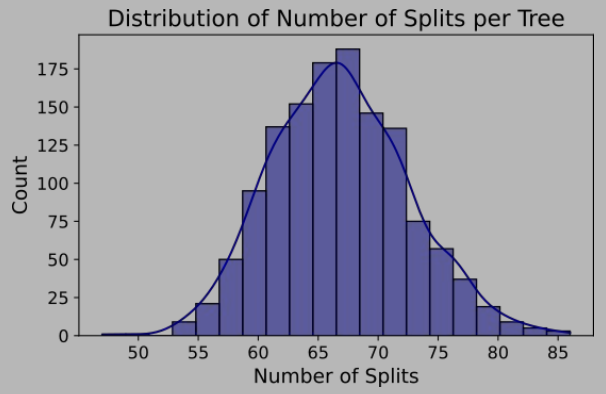
Class	Precision	Recall	F1-Score	Support
compressed	0.98	0.96	0.97	1560
expanded	0.92	0.95	0.93	720
accuracy			0.96	2280
macro avg	0.95	0.95	0.95	2280
weighted avg	0.96	0.96	0.96	2280

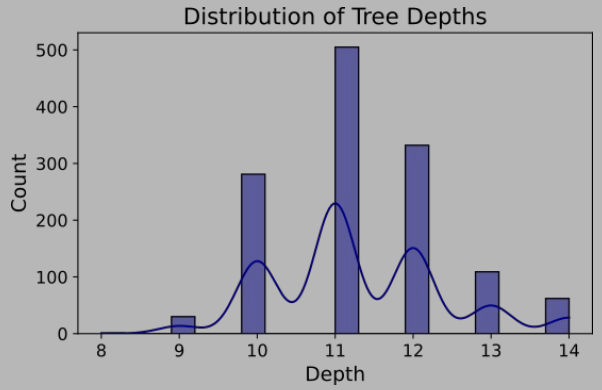
APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is classified as **'compressed'** or **'expanded'** based on a the material features.







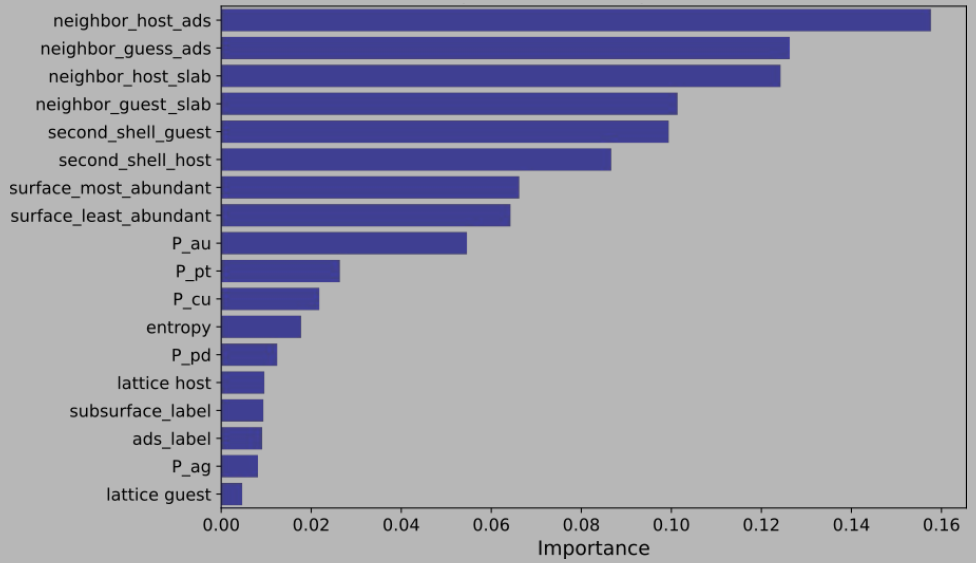


APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is classified as **'compressed'** or **'expanded'** based on a the material features.

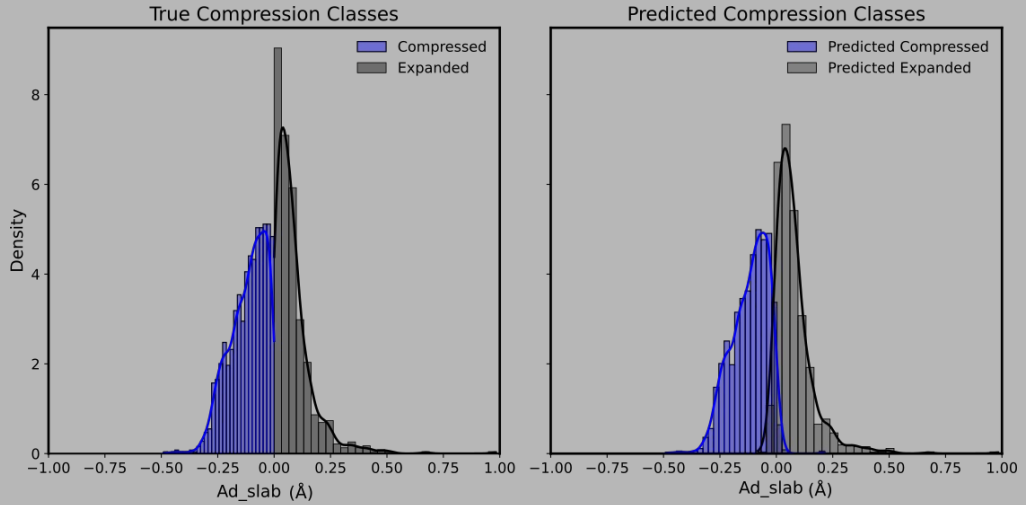
importances = pd.Series(best_model.feature_importances_,index=X.columns).sort_values(ascending=False)

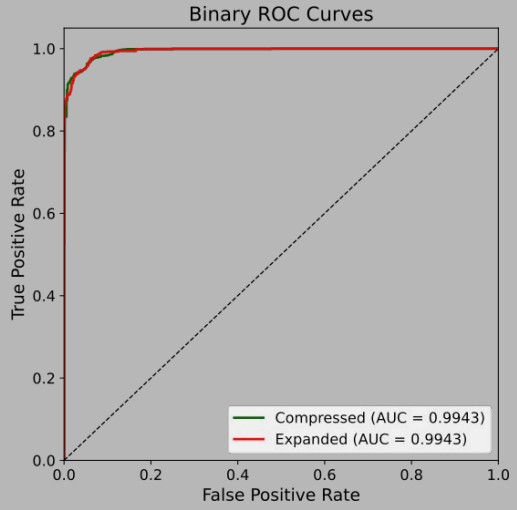
feature_importances_ returns the **Gini importance** values for each feature. creates a ranked list of features based on how much they contribute to reducing impurity across the forest

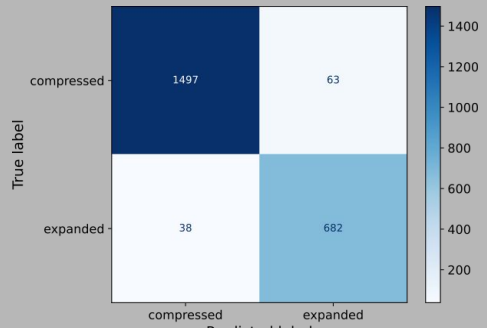


APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is classified as **'compressed'** or **'expanded'** based on a the material features.



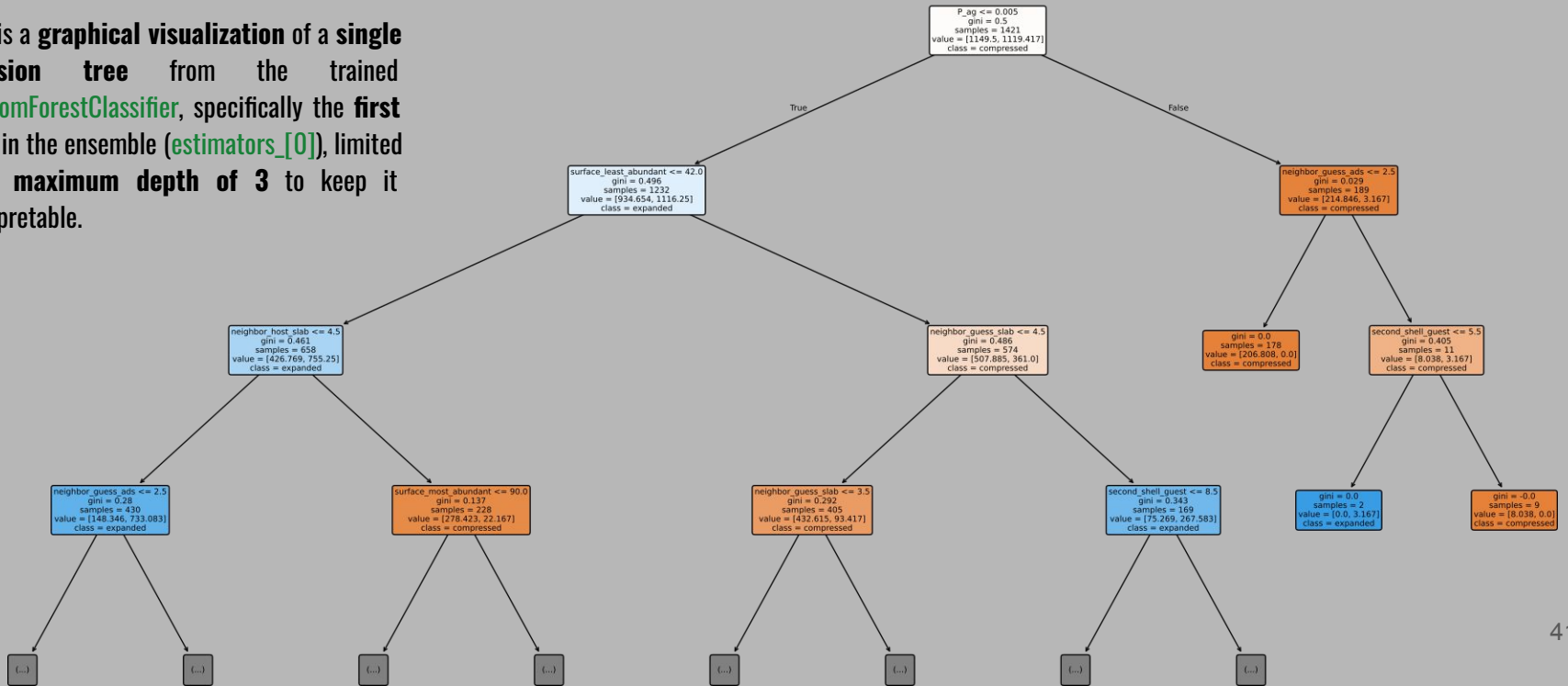




APPENDIX - D

- Appendix-C outlines the use of a **RandomForestClassifier** machine learning algorithm to predict whether an adsorption site is classified as **'compressed'** or **'expanded'** based on a the material features.

This is a **graphical visualization** of a **single decision tree** from the trained RandomForestClassifier, specifically the **first tree** in the ensemble (estimators_[0]), limited to a **maximum depth of 3** to keep it interpretable.



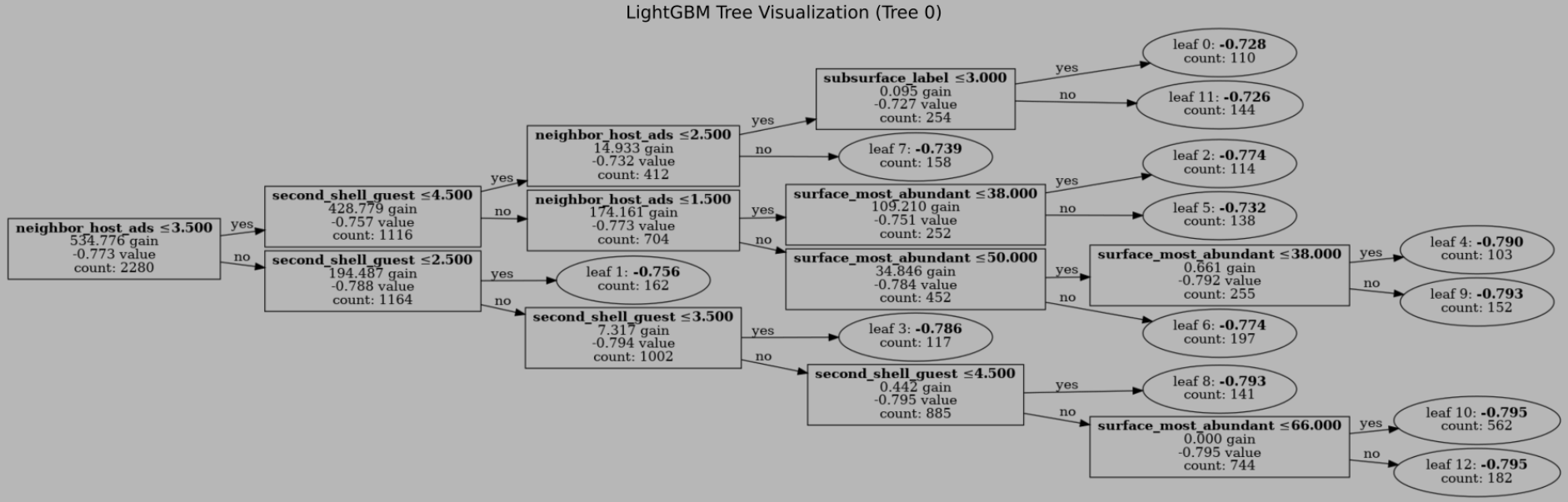
APPENDIX - E

- Appendix-D outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **LGBMClassifier**.

- **Loads a CSV** (general_site_analysis_version4.csv).
- **Filters** the data to focus on "ontop" adsorptions with ads_energy in the range **[-0.55, 0.55]**;
- **Creates a binary label** compression_class from the sign of the Ad_slab column.
- Encodes target using LabelEncoder.
- **Drops domain-specific and target-related columns** to avoid data leakage.
- Selects relevant features and binarizes target for ROC analysis.
- Uses **Optuna** to tune LightGBM hyperparameters via 5-fold CV, study.optimize(objective, n_trials=50, timeout=1200)
- Trains final model using the best parameters.
- Evaluates using: Classification report, Confusion matrix, ROC curves (per class)
- **Visualizations: feature importance** (LightGBM + SHAP), **class distributions** (true vs predicted), **Tree statistics** (depth, leaves, gain, splits)
- **Pairplot** of top features
- **Cumulative accuracy** vs number of trees
- Uses **SHAP** to visualize global feature importance (bar & beeswarm).

APPENDIX - E

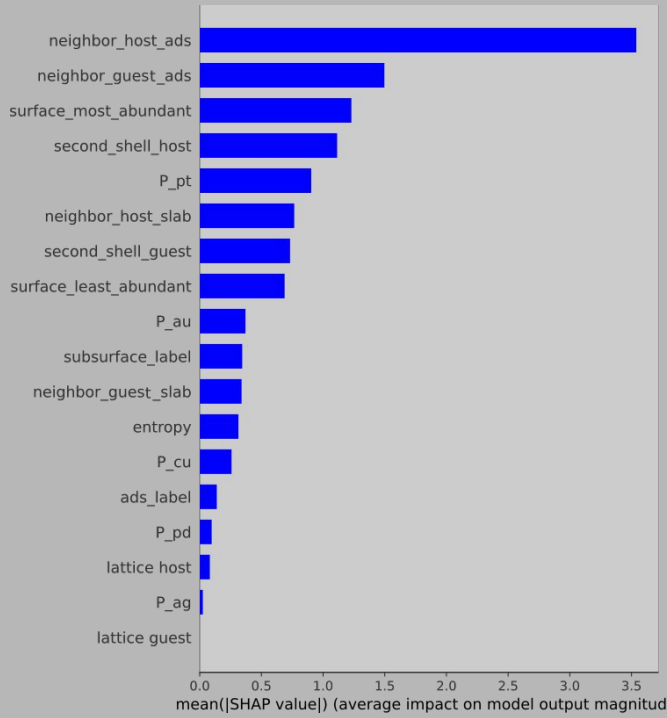
- Appendix-E outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **LGBMClassifier**.

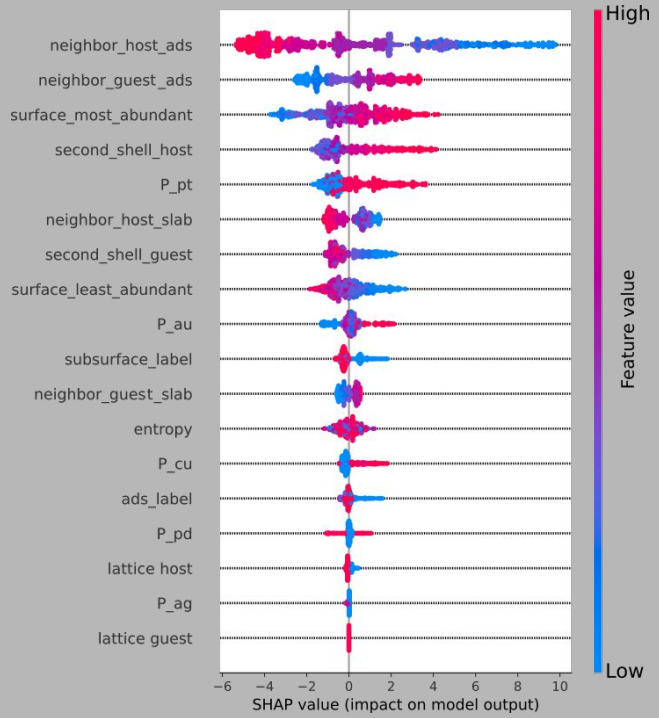


lgb.plot_tree(booster, tree_index=0, figsize=(20, 10), show_info=['split_gain', 'internal_value', 'internal_count', 'leaf_count'])

APPENDIX - E

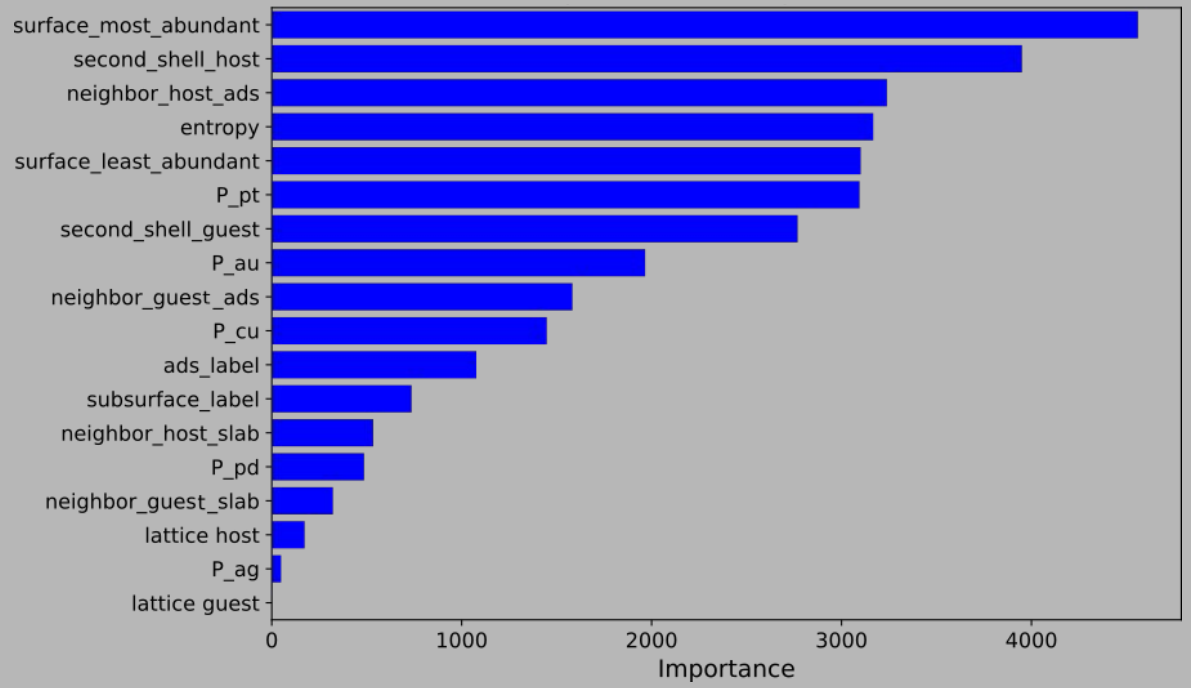
- Appendix-E outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **LightGBM**.





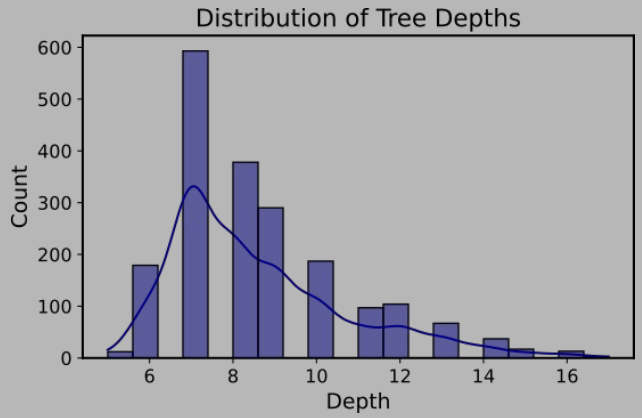
APPENDIX -E

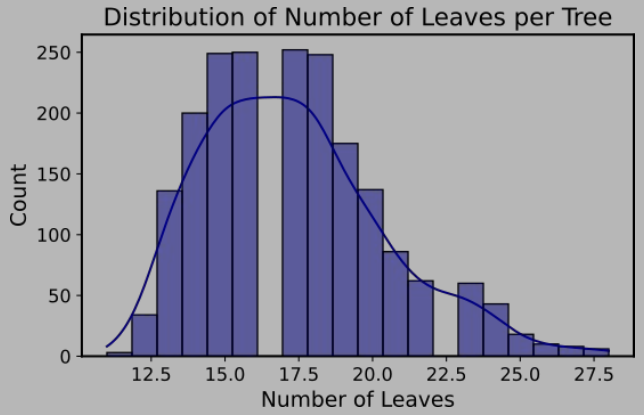
- Appendix-E outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **LGBMClassifier**.

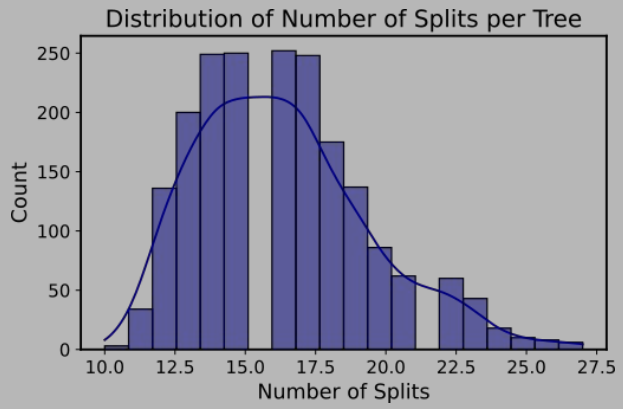


APPENDIX - E

- Appendix-E outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **LGBMClassifier**.

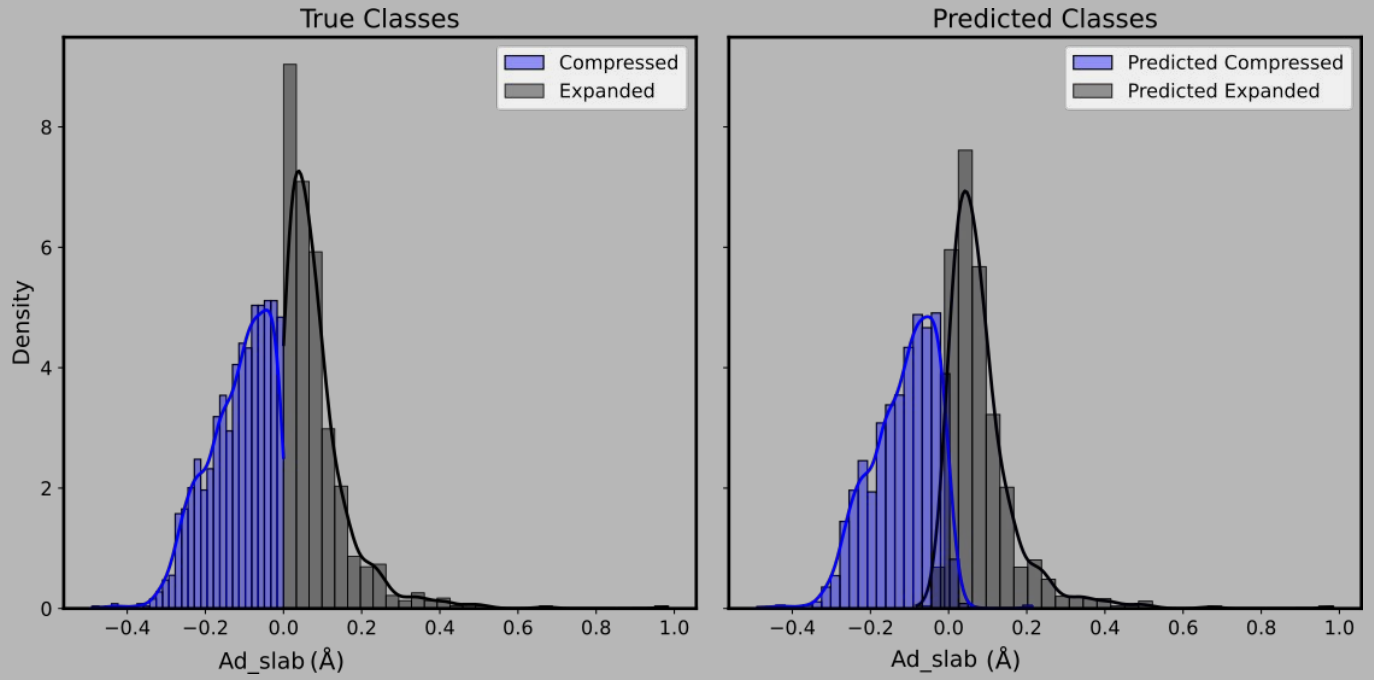






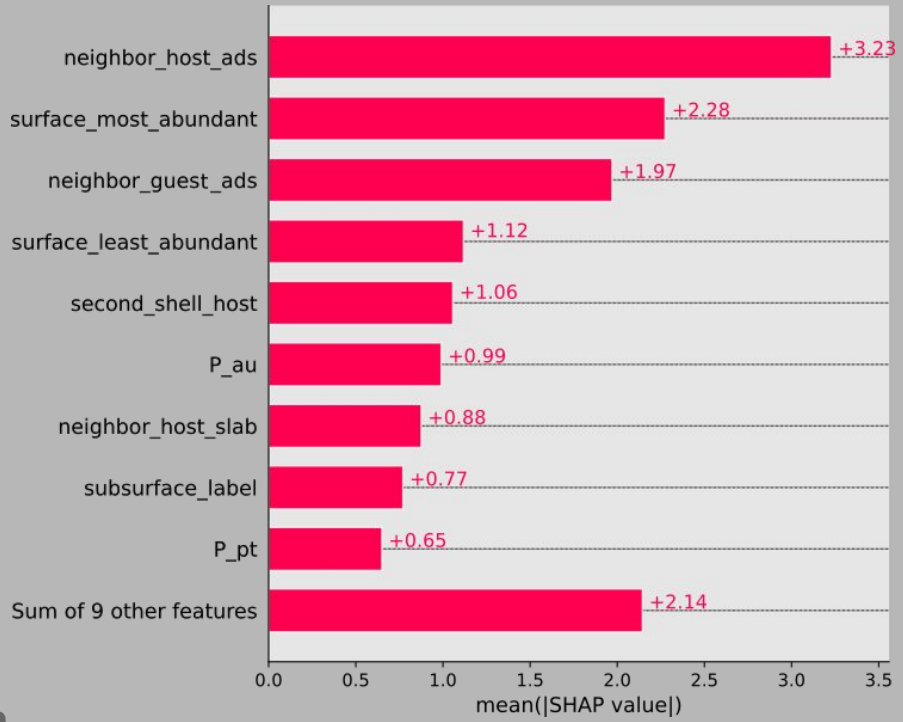
APPENDIX - F

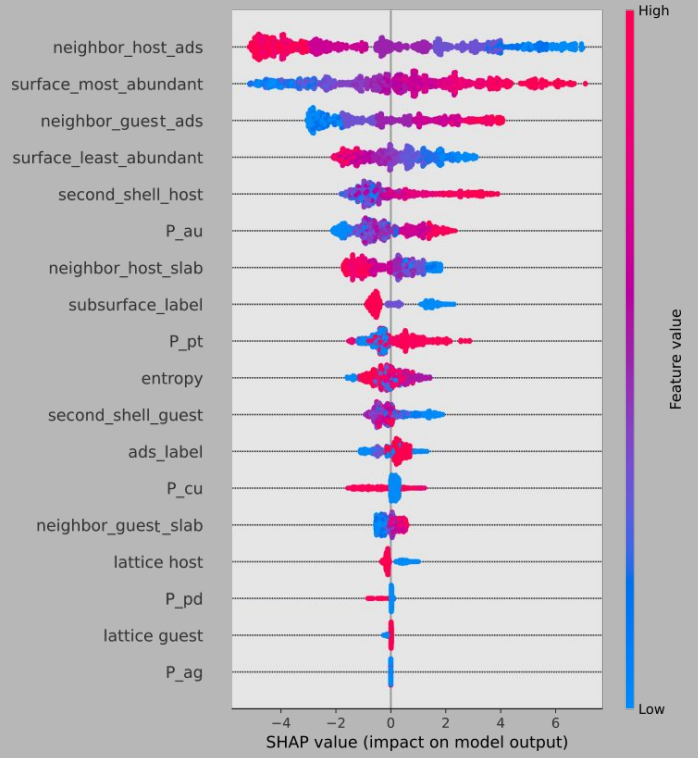
- Appendix-F outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **XBOOST (XGBClassifier)**.



APPENDIX - F

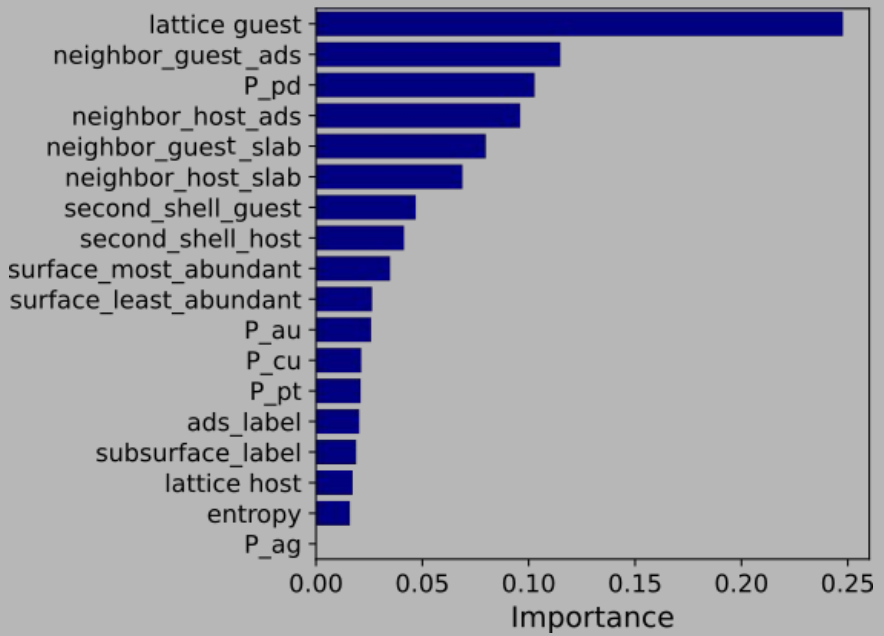
- Appendix-F outlines the machine learning algorithm for binary classification (**'compressed'** or **'expanded'** based on a the material features) using **XBOOST (XGBClassifier)**.





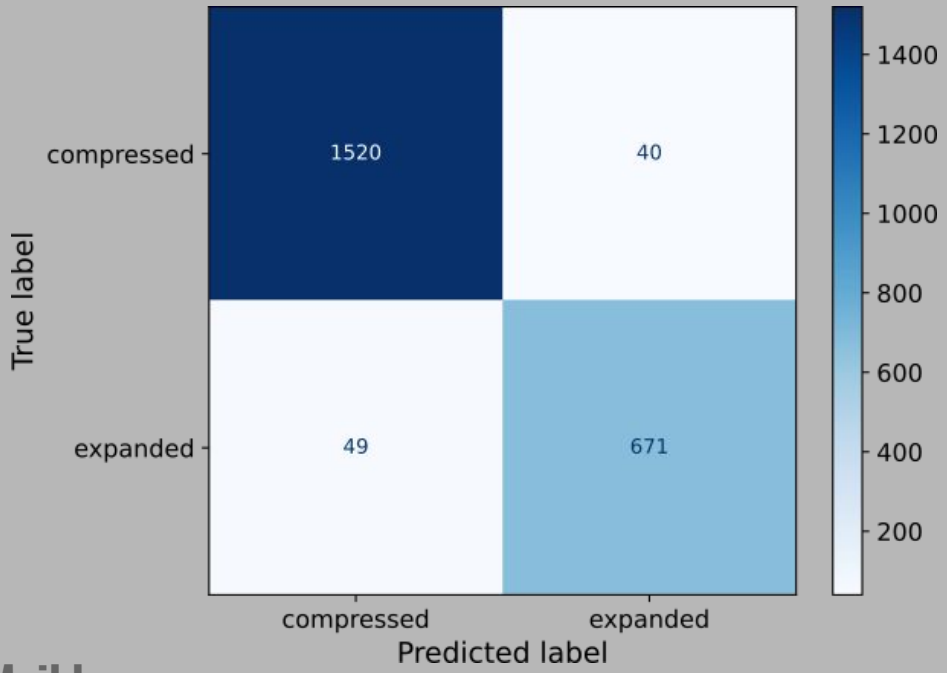
APPENDIX - F

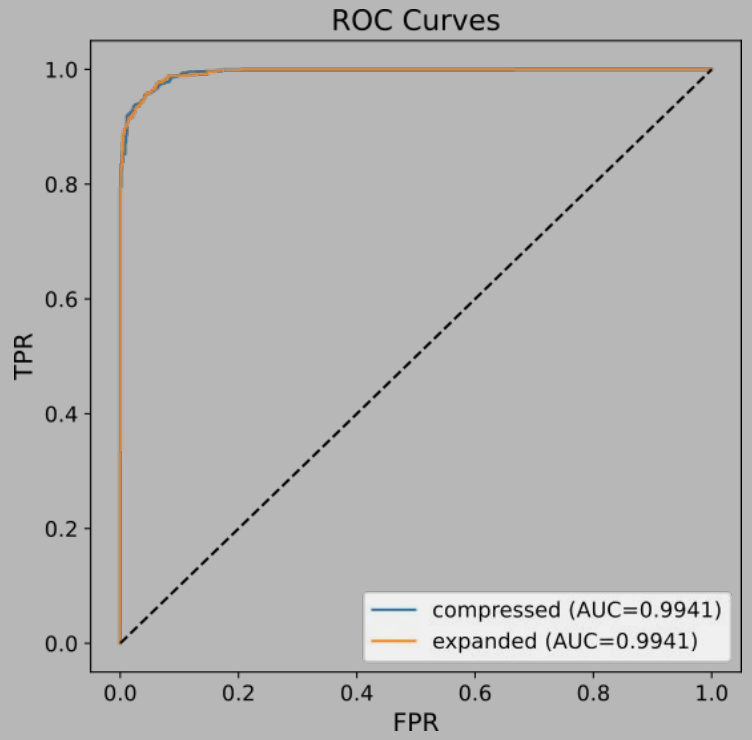
- Appendix-F outlines the machine learning algorithm for binary classification (**'compressed'** or **'expanded'** based on a the material features) using **XBOOST (XGBClassifier)**.



APPENDIX - F

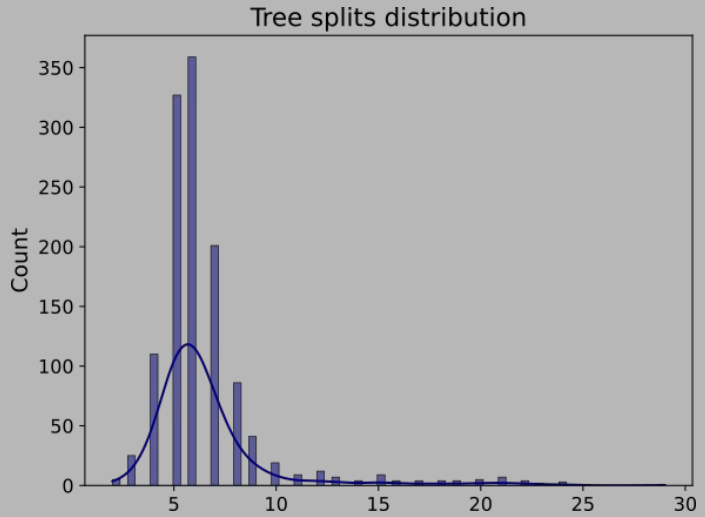
- Appendix-F outlines the machine learning algorithm for binary classification (**'compressed'** or **'expanded'** based on a the material features) using **XBOOST (XGBClassifier)**.

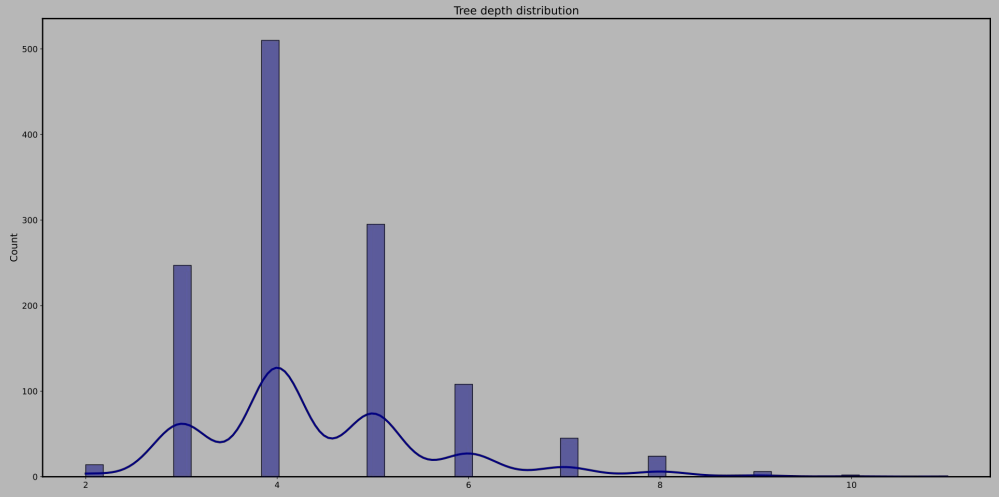




APPENDIX - F

- Appendix-F outlines the machine learning algorithm for binary classification (**‘compressed’** or **‘expanded’** based on a the material features) using **XBOOST (XGBClassifier)**.





Appendix G: Other graph motifs and added features

Site structure: Substructure of the 5x5 surface that includes the adsorption site, nearest neighbors and next-nearest neighbors.

Ekstra features: Added lattice parameter of atom (although correlated with element identity) and distance from adsorbate to assist in learning the hierarchy. Showed no improvement.

Also tried without the initial linear layer, showed no significant effect.

Appendix G: GCN Training Procedure

MSE loss function

MAE for validation

100 epoch early stopping patience, saving the model with best validation score

AdamW optimizer with ReduceLROnPlateau learning rate scheduler using factor of 0.8 (or 0.5) and patience of 10

Trained on L40s Nvidia GPU

Appendix G: GCN Hyperparameter Optimization

Using Tree-Structured Parzen Estimator (Bayesian opt. method) in optuna module

Dimension of hidden layers: 8-256

Number of gated conv. Layers: 2-5

Dropout: 0.0 to 0.5 in 0.1 steps

Batch size: 2^n, with n between 4 and 8

Weight decay (L2 reg.): 1e-6 to 1e-3 (drawn from log uniform)

Learning rate: 1e-4 to 5e-5 (drawn from log uniform)

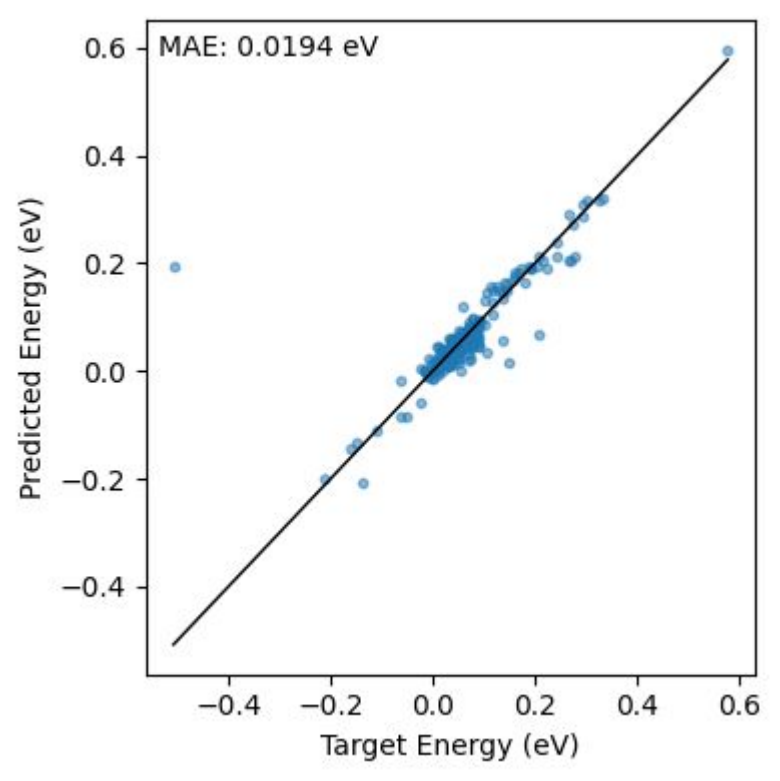
Appendix G: Hyperparameter Optimization best results from 100 trials

Model	Dim	Conv. layers	Dropout	Batch size	LR	W. dec	Mean MAE in 5-fold CV
Site	126	4	0.0	128	0.0012	0.00011	0.0153
Full surface	97	5	0.0	32	0.0004	3.8e-8	0.0140
Full surface ekstra feat.	128	4	0.0	64	0.0009	9.9e-6	0.0122
Full surface no linear in layer	187	5	0.0	16	0.0001	5.8e-6	0.0141

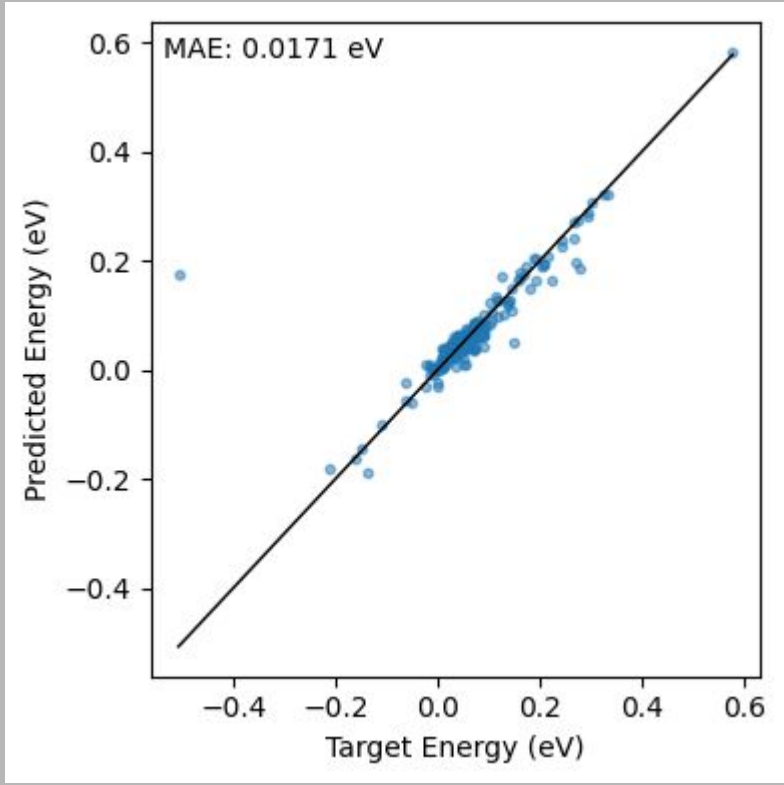
Trained with best found hyperparameters

Appendix G: Test parity plots (80/10/10)

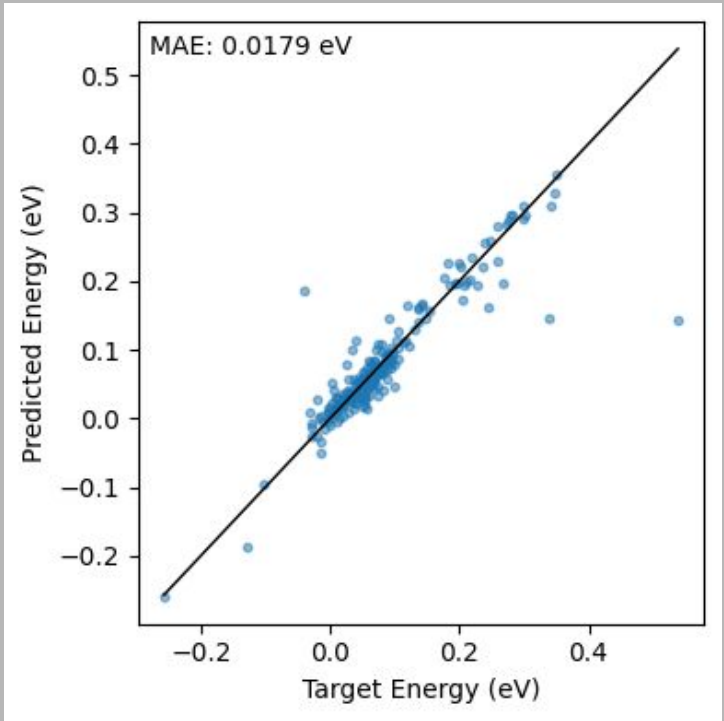
Site graph



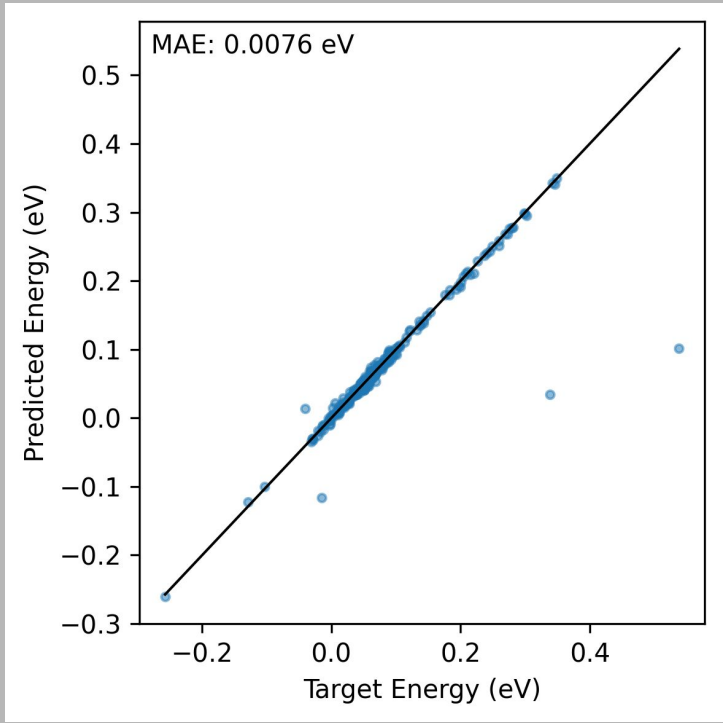
Full surface graph with extra features



Appendix G: GCN and OCP Validation Parity



GCN with full surface graph



OCP

Mads

Appendix H: Adsorption Energy Regression With Feature Engineering

Objective: Use LightGBM with engineered features (e.g., interaction terms, log/squared transforms) to predict *OH adsorption energy.

Data Source:

- Filtered DFT-calculated dataset (`general_site_analysis_version2.csv`), restricted to energies between -0.5 eV and 0.5 eV.

Key Features Used:

- `Ad_ads`, `Ad_slab`, `ads_label`, `subsurface_label`, `neighbor_*`, `lattice_`
- Engineered interactions like `Ad_ads_x_Ad_slab`, `Ad_ads_log1p`, etc.

Target Transformation::

- `arcsinh(ads_energy)` to reduce skew and stabilize training.

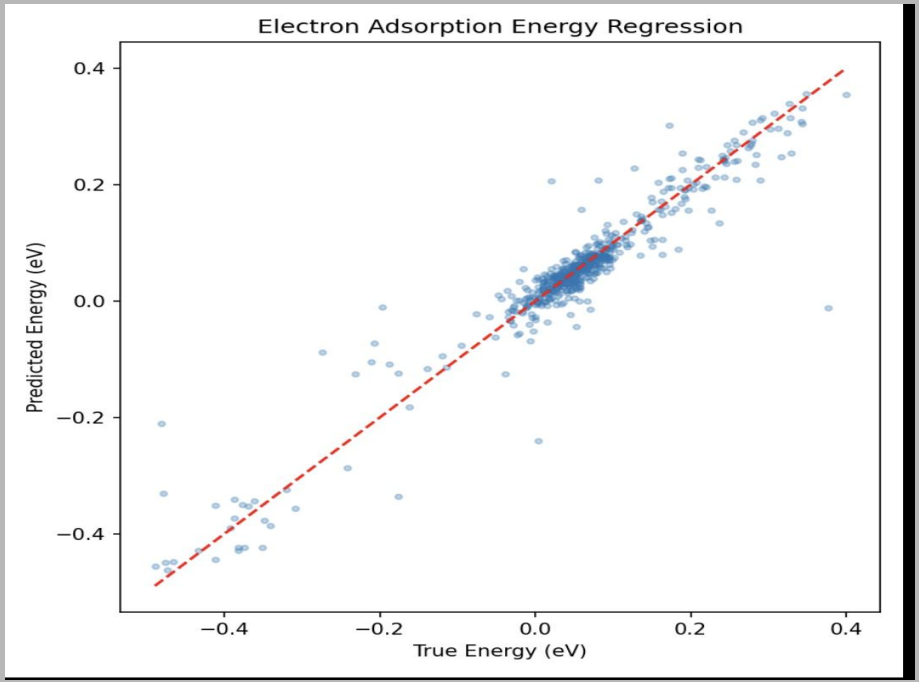
Danielle

Appendix H: Regression Metrics

Metric	Value
MAE	0.0165
RMSE	0.0336
R^2	0.9285
Relative MAE	1.4207

Danielle

Appendix H: Predicted vs. True OH Adsorption Energies



- Diagonal alignment indicates strong model accuracy
- Slight dispersion suggests feature coverage could still improve

Danielle

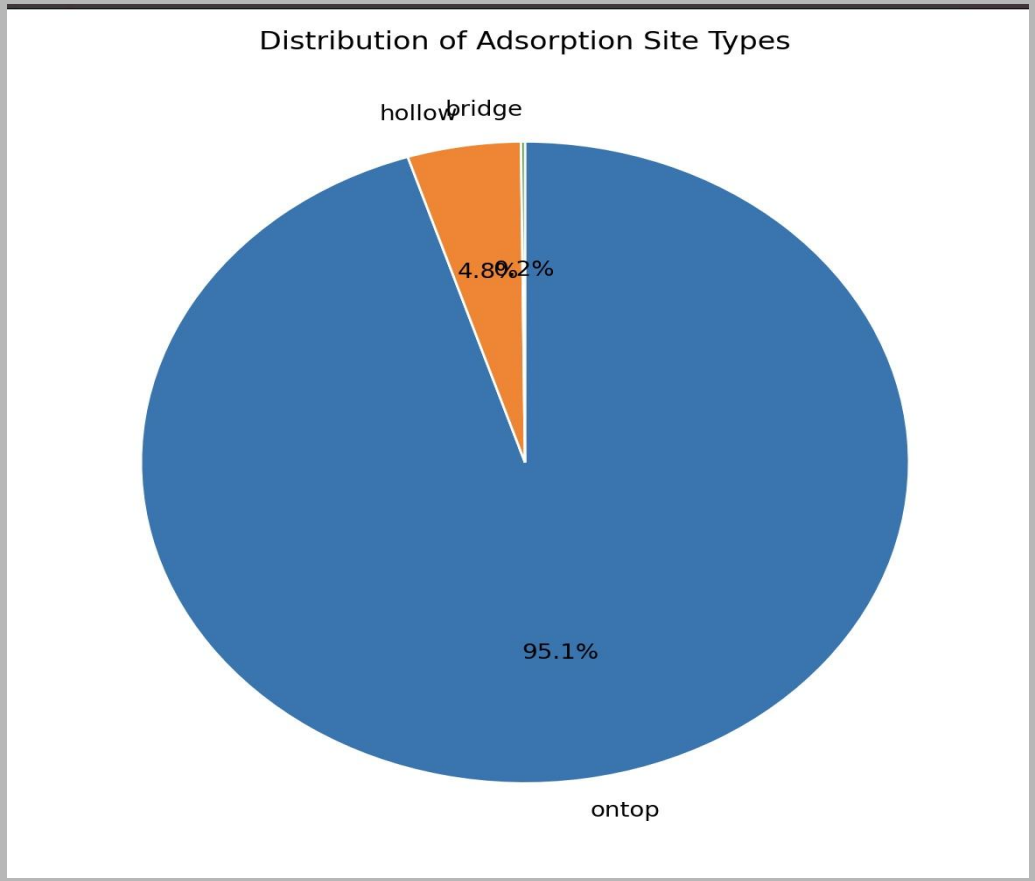
Appendix H: Final Model Performance (hollow + bridge)

- **Metrics on Test Set (80:10:10 split)** for the SHAP-pruned DART-CV ensemble:
  - MAE: 0.01346
  - RMSE: 0.02198
  - SMAPE: 28.90%
-

MAE	RMSE	SMAPE(%)
0.0134	0.0219	28.9

Danielle

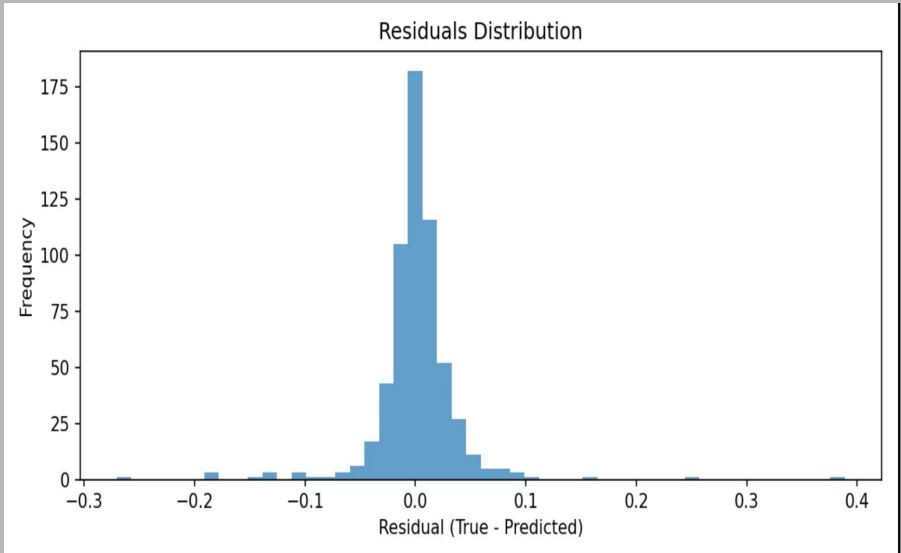
Appendix H: Multi-Site Coverage Visualization

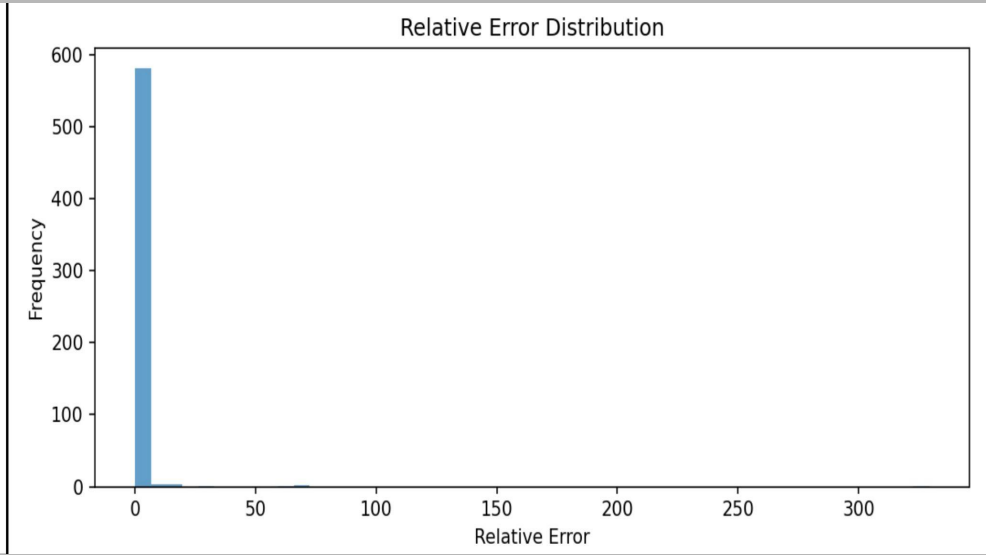


Adsorption site distribution across ontop, bridge, and hollow in the dataset.

Danielle

Appendix G: Residual & Relative Error Distributions





- Residuals mostly centered around 0
- Few high relative errors due to extreme adsorption energies

Appendix I: Clustering

Clustering was done using:

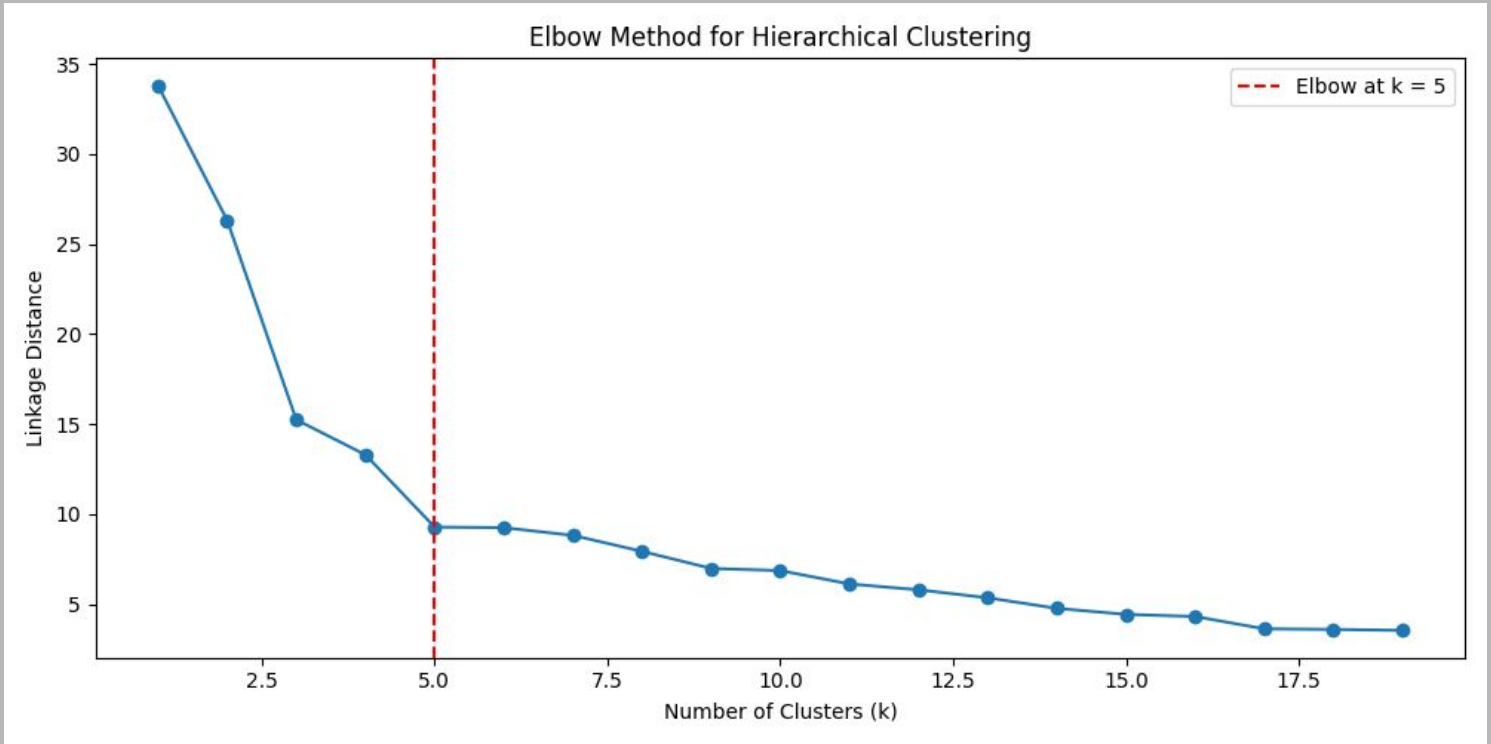
- Agglomerative clustering
- K-means
- Spectral Clustering
- Gmixture’

Data transformed using StandardScaler

Outliers removed with Mahalaobis distance

Agglomerative clustering performed the best overall, Gmixture and Spetral worked very poorly

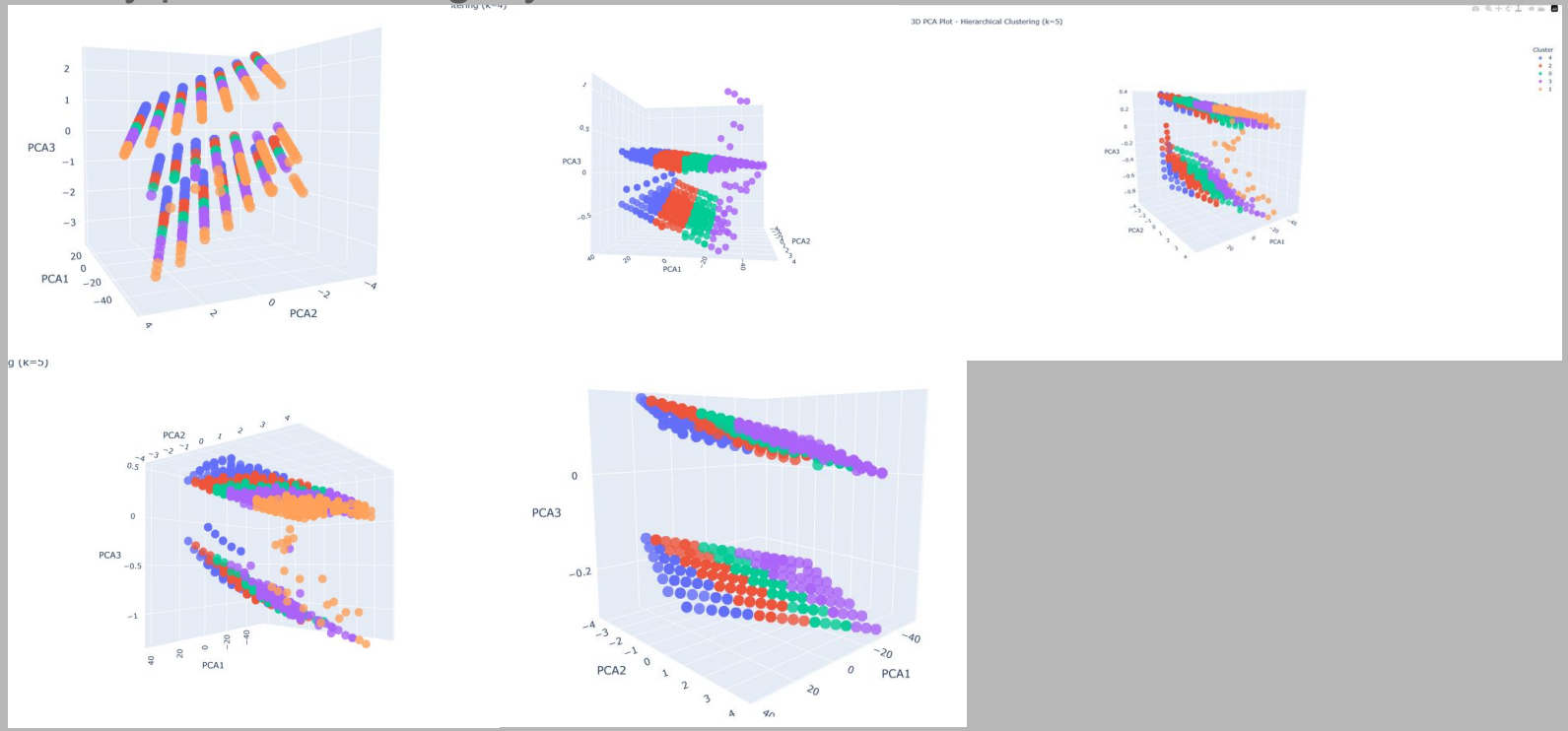
Multiple cluster sizes where tried, best cluster size was picked based on elbow method.



Features initially selected by removing features with high correlation

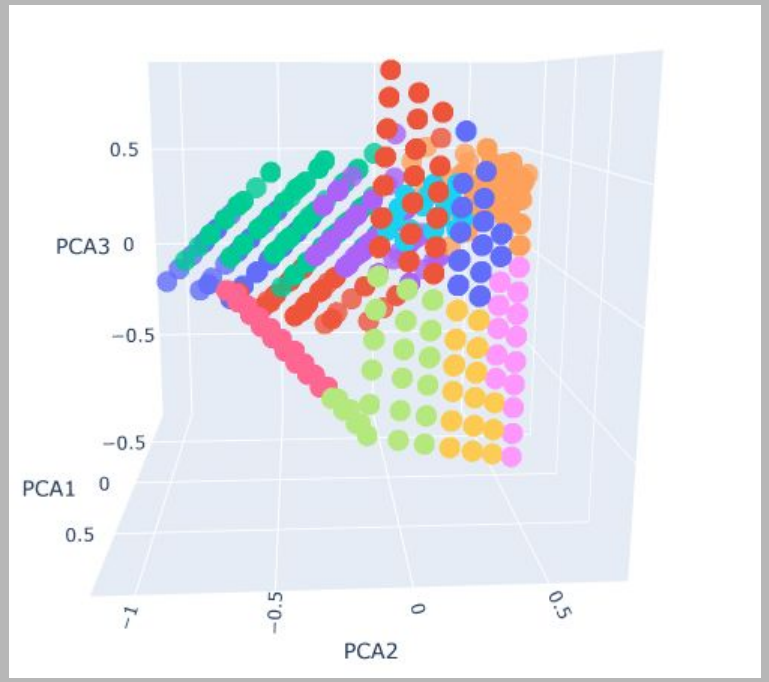
Later switched to manually selecting features based on knowledge of the features. Many different feature sets where tried (Basically every combination of features that didn’t include multiple correlated features)

A collage of how every set of features produces basically the same clustering in agglomerative clustering, except when using the new set of features created and presented in the presentation
Every picture is a slightly different set of the features

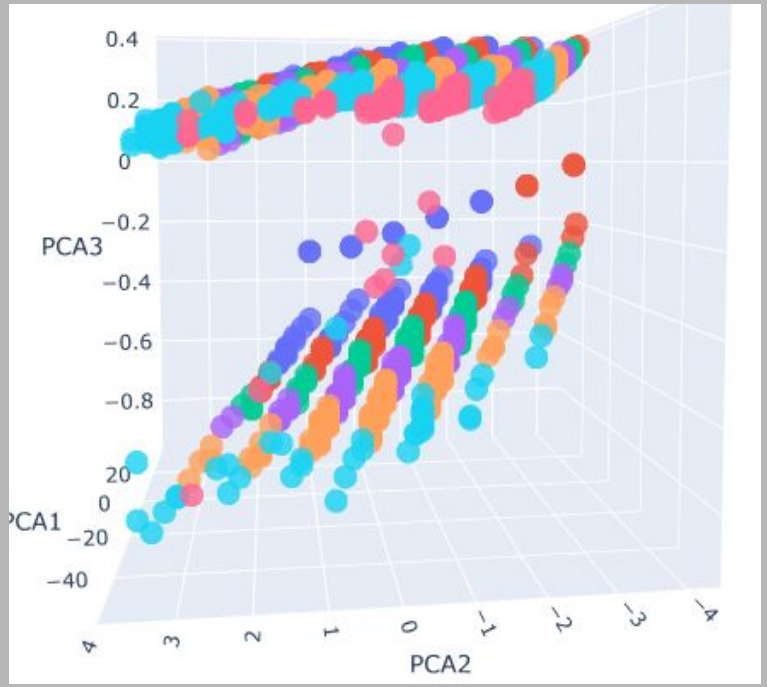


K-means looks very similar to agglomerative, but with more clusters

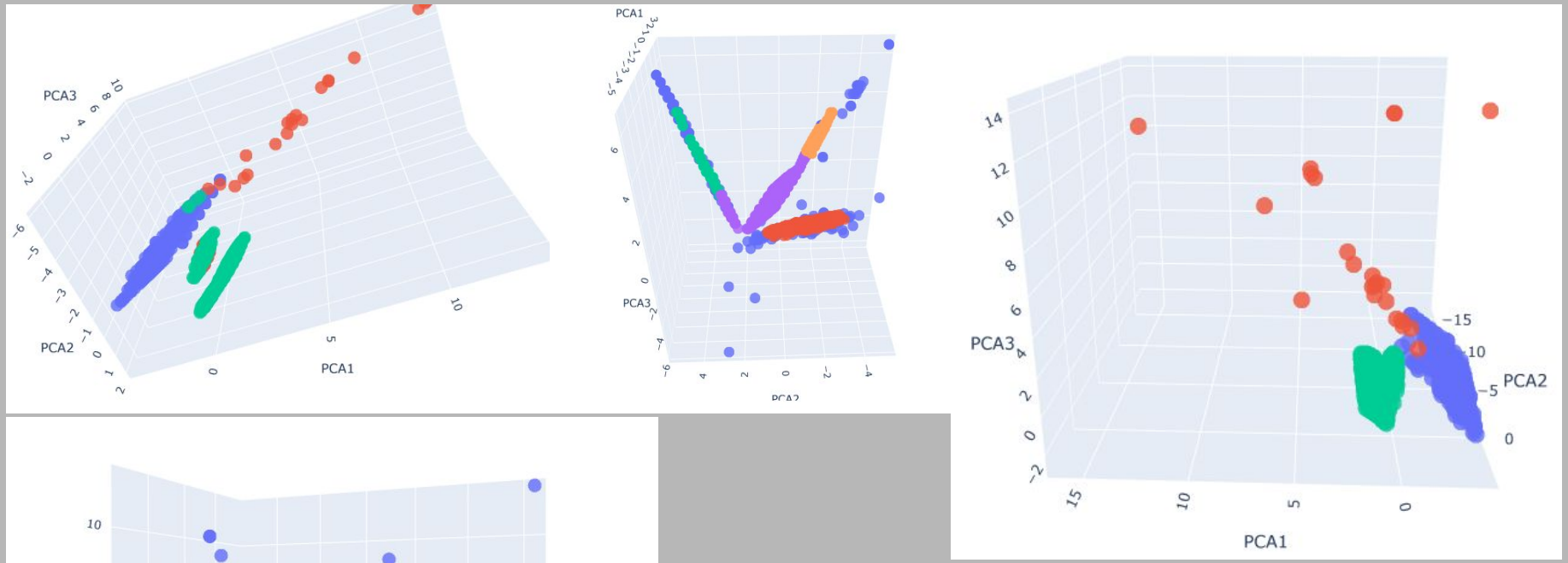
With the new feature-set



Without the new feature set



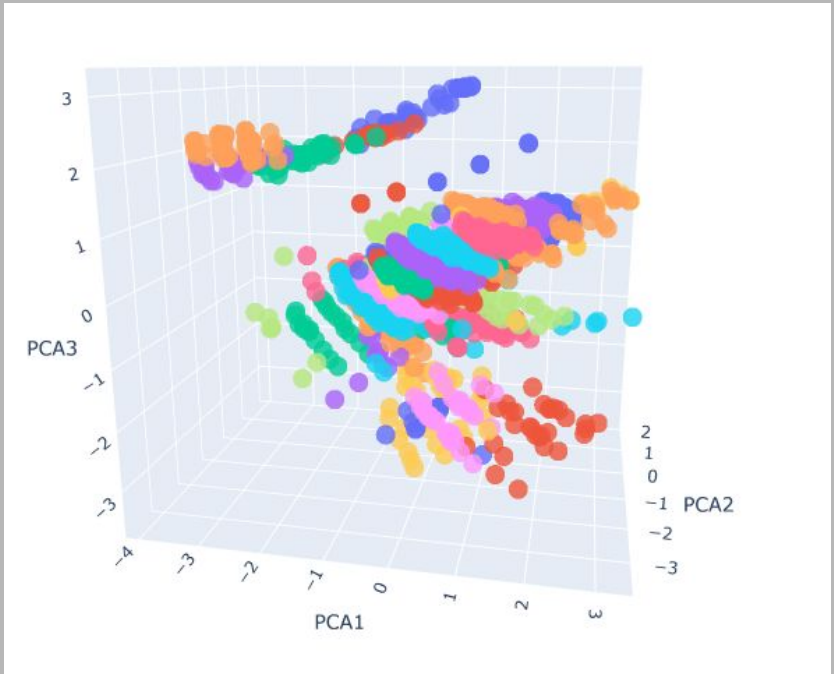
Spectral Clustering was the one model that produced a different configuration

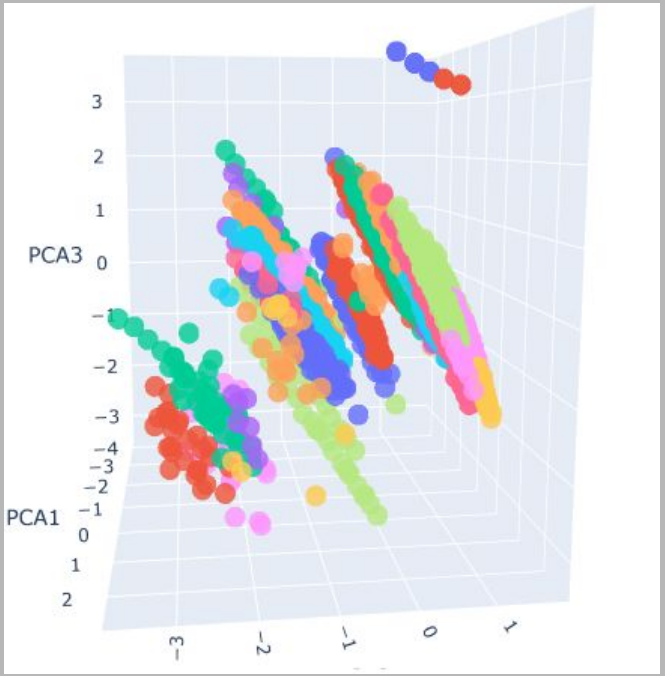


But it's not pretty, the different clusters are just different base slabs. It couldn't tell Pt and Pd apart most of the time

And Gmixture did something very unholy

But still 3-4 “planes” with outliers around





Project Statement

Danielle: Tree based models using a LightGBM regression pipeline—encoding ontop/bridge/hollow sites, crafting advanced feature transformations, running Optuna‑driven hyperparameter searches, ensembling models, and applying SHAP for interpretability. Additionally, contributed to code development, metric evaluation, SHAP-based feature importance interpretation, and model performance visualization.

Mads: Graph based models, GCN and OCP finetuning. Work included: Graph model. Data transformation into graphs. Hyperparameter optimization. Final training and testing. OCP data preprocessing and finetuning.

Mailde: Provided the data for training. Generated the tabular data from the initial ASE database. Performed Classification models with different tree-methods. Performed regression models with LightGBM considering different features combinations. Hyperparameter optimization. Final training and testing.

Simon: Graph convolutional network and clustering. Setup initial GCN model and converting data to graphs. Build multiple different clustering models using different base models are feature-sets.