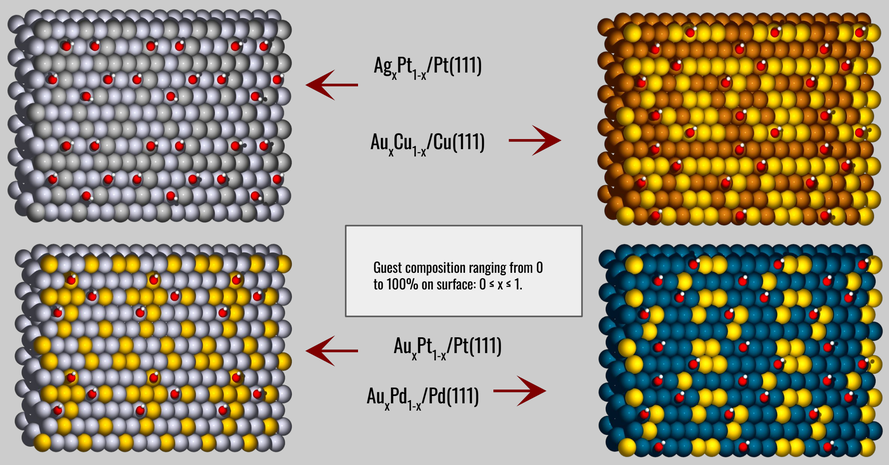
# Machine Learning for Predicting Catalyst Properties in Binary Alloys

**Mailde S. Ozório**
Postdoc, mdso@chem.ku.dk, Department of Chemistry, University of Copenhagen

## Why study binary alloys?

Binary alloys of transition metals hold significant potential as catalytic materials for the cathodic oxygen reduction reaction (ORR) in polymer-electrolyte-membrane fuel cells. However, their catalytic application requires a better understanding of the key factors influencing ORR performance. To address this gap, we present a machine learning framework trained on high-throughput DFT data that not only predicts OH adsorption energies with high accuracy but also predicts compressive/tensile strain of the adsorption site and label the adsorption sites element. The findings reinforce the lattice mismatch as a critical factor controlling ORR activity in monolayer binary alloys, in line with recent studies in the *Journal of Catalysis*.[1,2]

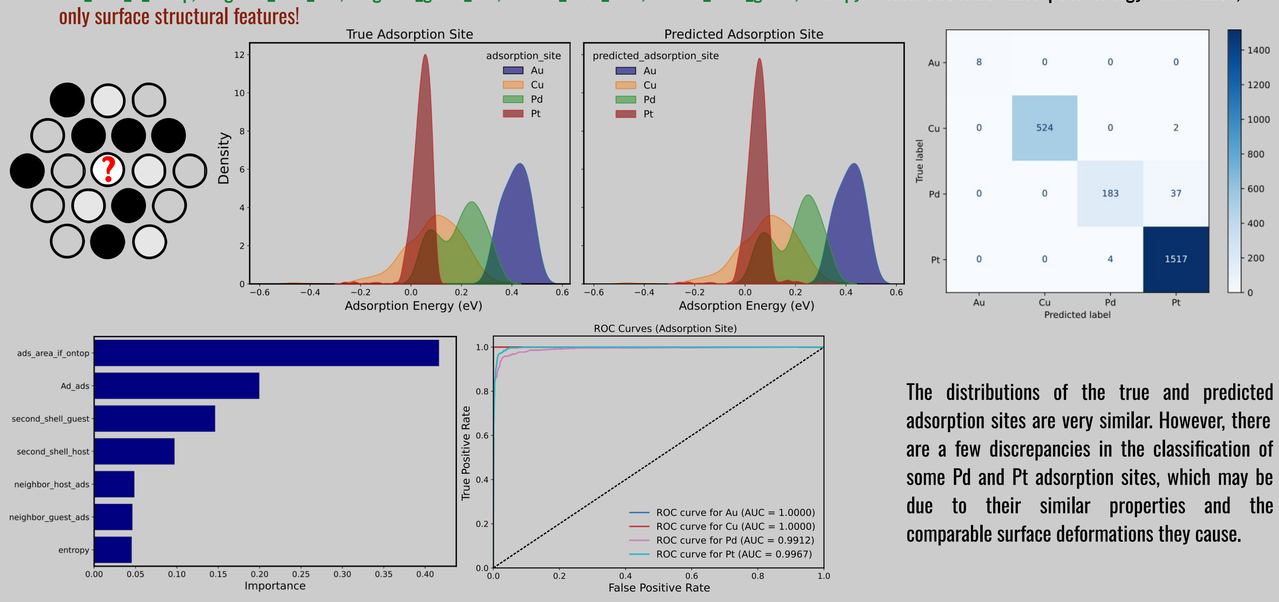
## Guess the Adsorption Site




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).

- **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

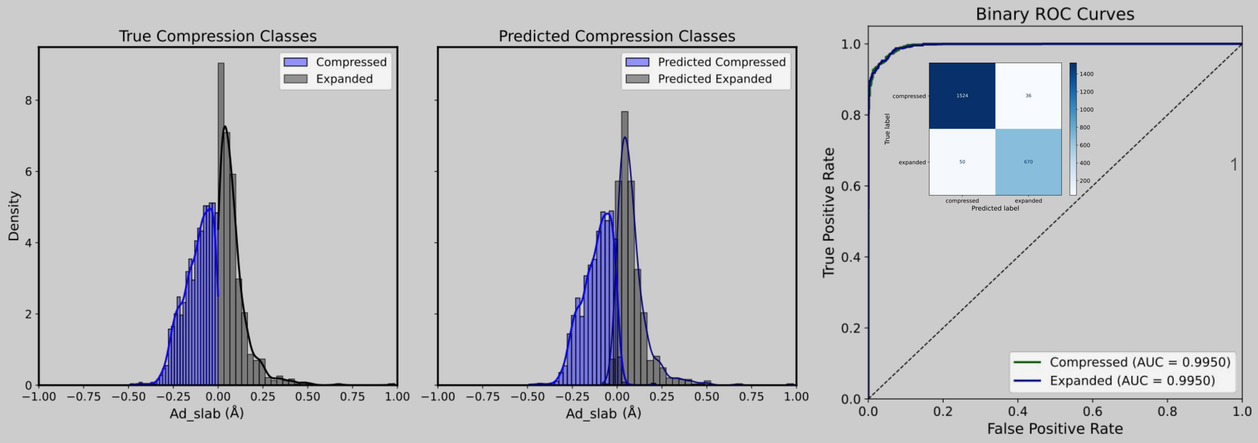
**Random Forest classifier to predict the adsorption site given few surface features:** Guess the Adsorption Site considering only 7 features: Ad_ads, ads_area_if_ontop, neighbor_host_ads, neighbor_guest_ads, second_shell_host, second_shell_guest, entropy. Without consider adsorption energy information,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



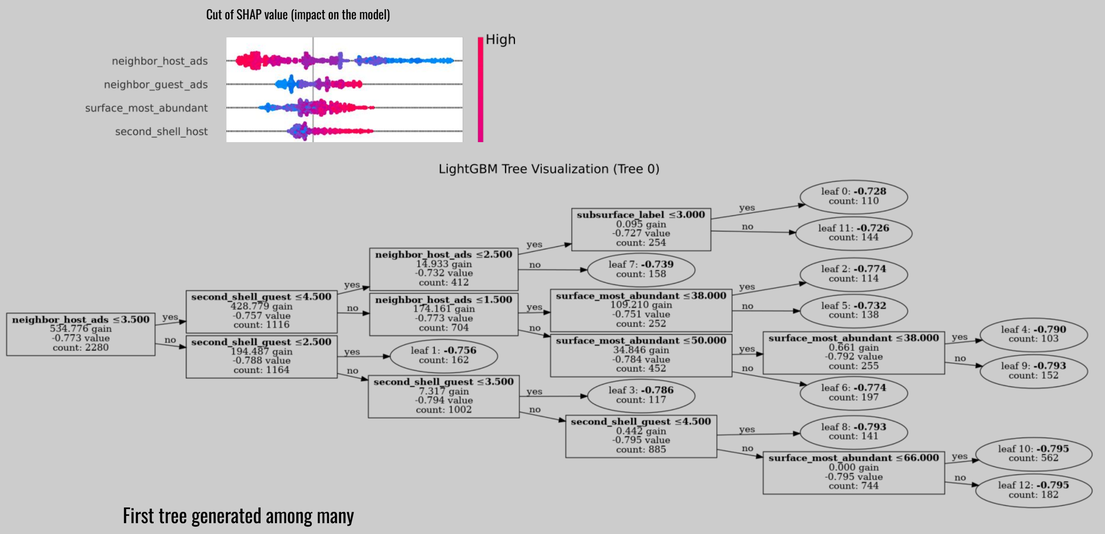

The distributions of the true and predicted adsorption sites are very similar. However, there are a few discrepancies in the classification of some Pd and Pt adsorption sites, which may be due to their similar properties and the comparable surface deformations they cause.

## 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.



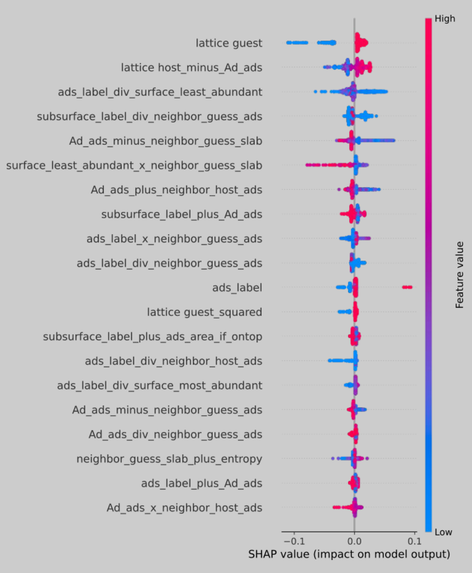

Classify the binary alloys as "compressed" or "expanded" relative to the non-alloyed.

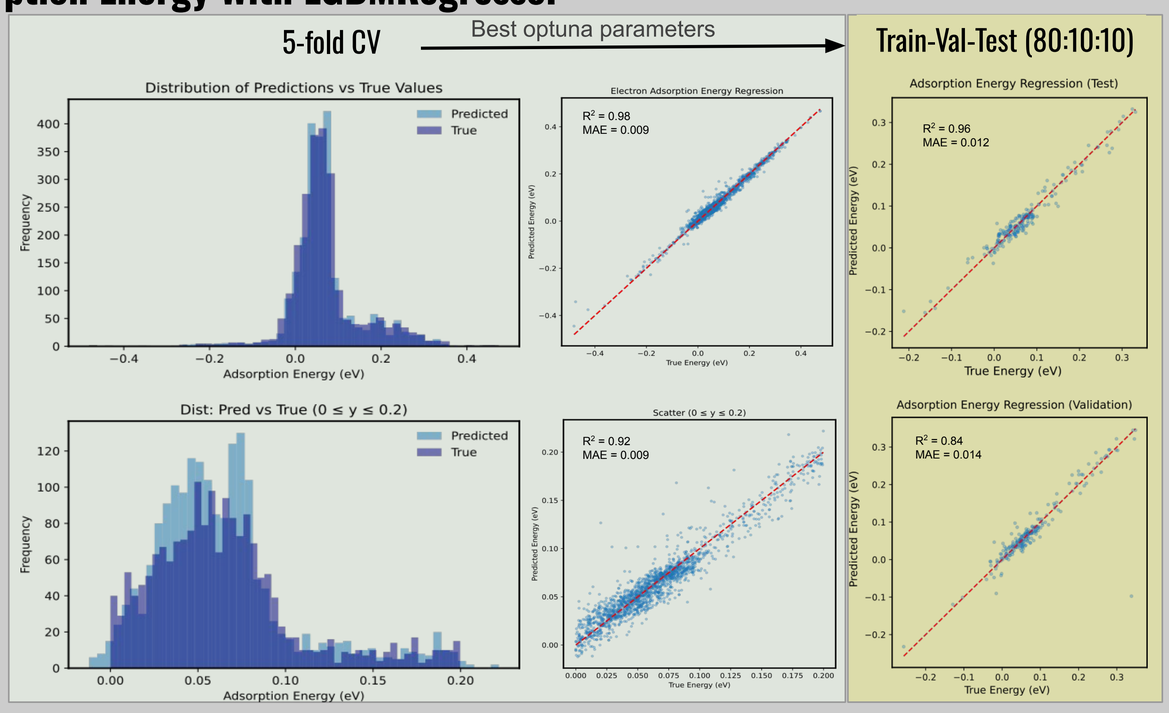



First tree generated among many

## Prediction of the *OH Adsorption Energy With LGBMRegressor

| 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 |




- **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

**CONCLUSIONS**: In conclusion, this study establishes an interpretable machine learning framework for predicting ORR activity on monolayer binary alloys. The models accurately classified adsorption sites and distinguished compressed from expanded adsorption states, while regression models reliably predicted OH adsorption energies. Notably, the guest lattice parameter emerged as a key descriptor for the adsorption energy prediction, underscoring its central role in ORR catalysis. Together, these findings highlight the utility of targeted feature engineering and data-driven modeling in guiding the rational design of efficient alloy catalysts.

CAPeX
Danmarks Grundforskningsfond Danish National Research Foundation
Ministry of Higher Education and Science Denmark

**References**

[1] M. S. Ozório, M. F. Nygaard, A. S. Petersen, R. J. Behm, J. Rossmeisl, J. Catal., 433, 115484, (2024).
[2] M. S. Ozório, M. F. Nygaard, J. Rossmeisl, J. Catal., 443, 115988, (2025).
[3] S. Brimaud, A. K. Engstfeld, O. B. Alves, R. J. Behm, J. Electroanal. Chem., 716, 71– 79 (2014)
[4] S. Beckord, S. Brimaud, R. J. Behm, J. Electroanal. Chem., 819, 401– 409 (2018).