Property prediction in metallic nanoclusters using machine-learning methodologies

Felipe V. Calderan, Lucas B. Pena, Gabriel Pinheiro, Karla F. Andriani,
Priscilla Felício-Sousa, Breno R. L. Galvão, Juarez L. F. Da Silva, Marcos G. Quiles

Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil

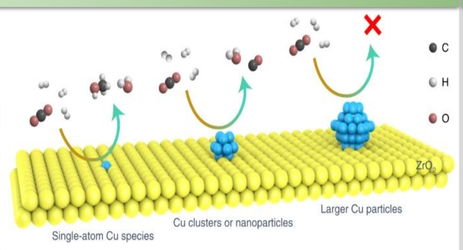
Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brazil

São Carlos Institute of Chemistry, University of São Paulo, São Carlos, SP, Brazil

Departament of Exact Sciences, State University of Santa Cruz, Ilhéus, BA, Brazil

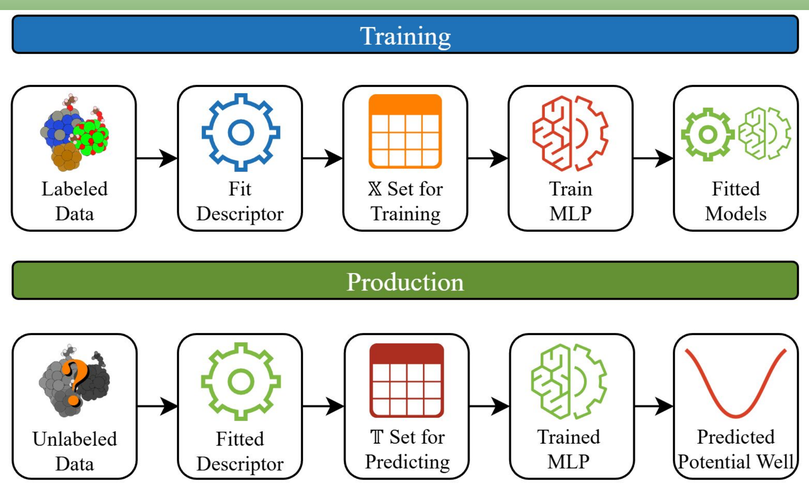
Motivation

Accurate interaction energies (ΔE_{tot}) are crucial for understanding catalytic mechanisms, yet using DFT calculations is costly to explore the vast configurational space of adsorption systems. Machine Learning (ML) offers a faster alternative, however lacking suitable chemical datasets. For reliable high-throughput screening, models must handle heterogeneous data while accounting for adsorption physics.



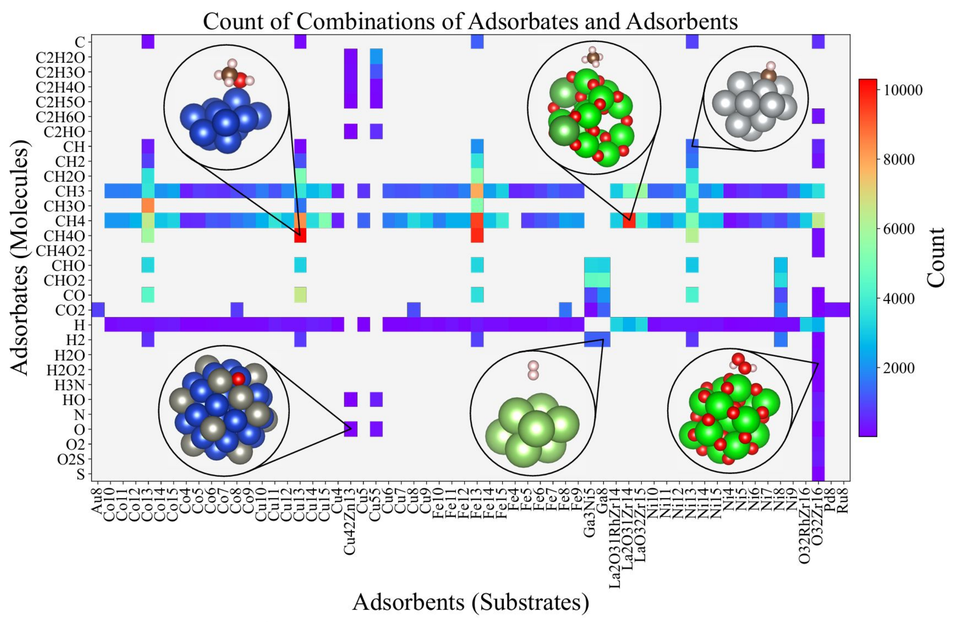
Catalysis products depend on (supported) cluster size. (ZHAO,2022)

Main Goals



- Develop a framework encompassing data parsing, featurization, model training, testing, and hyperparameter optimization.
- Train and evaluate ML regressors using multiple chemical descriptors to predict interaction energies in finite particle systems.
- Assess the reliability, transferability, and computational efficiency of the ML pipeline in comparison with conventional DFT calculations.

Dataset



The dataset was built from heterogeneous DFT optimizations employing the semilocal exchange-correlation PBE energy functional**[1]** with van der Waals corrections**[2]** in FHI-aims, being composed of 431,638 molecule-nanocluster configurations, as shown in the figure above. These are calculation frames from various works on nanoclusters adsorption systems with adsorbate species related to different catalytic pathways. The interaction energies were calculated as follows:

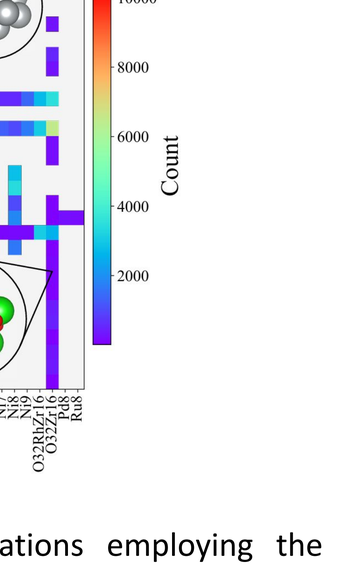
$$\Delta E_{\mathrm{tot}} = E_{\mathrm{system}} - (E_{\mathrm{cluster}} + E_{\mathrm{adsorbate}})$$

where samples with $\Delta E_{tot} > 0$ removed to keep physically valid adsorption states.

Descriptor Design - Site selection

Because adsorption energetics are highly local, the atoms near the binding region carry the most meaningful information. A good descriptor should capture this local environment, avoid irrelevant bulk atoms, and remain computationally efficient. To achieve this, we use the following procedure:

- For each substrate atom B_j, compute the minimum distance to any adsorbate atom A_i
- Sort all substrate atoms by increasing distance
- Choose the s nearest substrate atoms. These form the hypothetical adsorption site.



Descriptor Design - Product

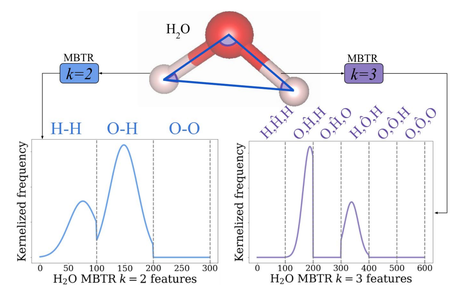
Using the same procedure to select the hypothetical adsorption site, we introduce two specialized adaptations of conventional descriptors.

Smooth Overlap of Atomic Positions (SOAP)

Once the adsorption site is identified, we compute the geometric center of both the adsorbate and the selected site atoms. SOAP**[3]** is then evaluated using these centers for initialization, applied to the full molecule but only the local site rather than the entire substrate. A subsequent correlation analysis removes redundant features. The result is an adsorption-focused descriptor, that we call Cut-SOAP, which is far more efficient than naïve implementations while maintaining competitive accuracy.

Local Many-Body Tensor Representation (LMBTR)

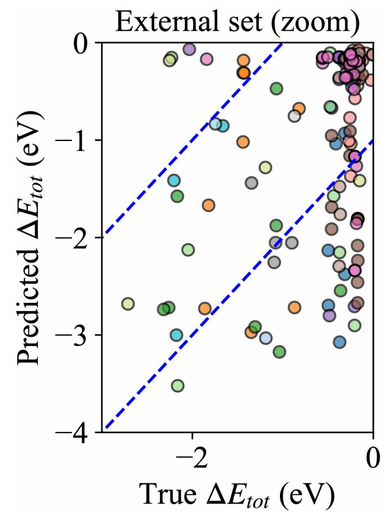
LMBTR**[4]** organizes structural information through a kernelization of atoms distances and angles organized into chemical classes. The descriptor is constructed considering only the adsorption site and adsorbent. Due to the diversity of chemical species in the dataset, the extensive final feature vector is reduced to principal components.



Results

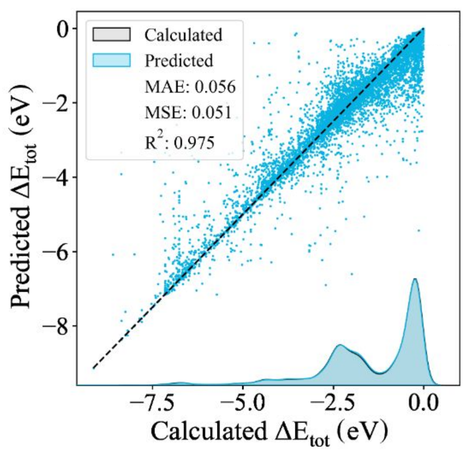
Cut-SOAP

Feature vector size was reduced from 75,000 to 1,835 with minimal information loss. Combined with a Multilayer Perceptron (MLP), it achieved a test-set average Mean Absolute Error (MAE) below 0.1 eV and an external-set average MAE below 1.0 eV. This performance drop on the external-set reflects the inherent difficulty of the task, given that the set contains adsorption geometries never encountered during training. Other predictors (e.g., XGBoost) performed similarly or worse. Across the evaluations, Cut-SOAP matched the accuracy of slower conventional descriptors.



LMBTR

The feature vector size is reduced from 18,000 to 1,800 features from the principal component analysis (PCA), maintaining 99% of explained variance. The trained random forest regressor (RFR) model achieves a MAE of 0.05 eV with comparable accuracy to DFT calculations. A similar result is achieved using the simpler descriptor Coulomb Matrix (CM) and the RFR, however the LMBTR greatly outperforms the CM using simpler, linear methods. When presented with new data, the model has a slightly worse MAE, being incapable of accurately determining energies for unseen nanocluster-adsorbate combinations.



Conclusions

- Local structural descriptors (Cut-SOAP, Cut-EVCM, LMBTR) are efficient in capturing relevant physics while controlling dimensionality.
- Large, real-world DFT datasets can indeed produce accurate ML models, despite not being designed for machine learning.
- ML models trained on optimized descriptors achieve high in-distribution accuracy, but transferability depends on structural representativeness.
- The pipeline enables rapid screening of adsorption energetics, a step toward:
  - accelerated catalyst discovery,
  - automated ΔE_{tot} exploration,
  - and large-scale materials design.

References

Figure: Zhao, H et al., The Role of Cu1–O3 Species in Single-Atom Cu/ZrO2 Catalyst for CO2 Hydrogenation. Nature Catalysis, 2022, 5, 818–831.
1 Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. Phys. Rev. Lett. 1996, 77, 3865–3868.
2 Tkatchenko, A.; Scheffler, M. Accurate Molecular van Der Waals Interactions From Ground-state Electron Density and Free-atom Reference Data. Phys. Rev. Lett. 2009, 102, 073005.
3 Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. npj Comput. Mater. 2018, 4, 37
4 Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. Mach. Learn.: Sci. Technol. 2022, 3 (4), 045017.

Acknowledgments

FAPESP
CAPES
anp Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
CINE CENTER FOR INNOVATION ON NEW ENERGIES
CIDAG Computational Intelligence and Data Analysis Group
QTNano Juarez L. F. Da Silva Group