ML-SAPIE: An Autonomous Workflow Bridging High-Throughput DFT and Machine Learning for Surface Interface Discovery

Tabut Mary[1,3], Trevizam Dorini Thiago[2], Salzemann Caroline[3], Calatayud Monica[3]

[1]Sorbonne Université, CNRS, Laboratoire de Chimie Théorique, LCT, 4 Place Jussieu, F-75005 Paris, France.
[2]Institute of Chemistry, State University of Campinas, Campinas, São Paulo, Brazil
[3]Sorbonne Université, MONARIS, CNRS-UMR 8233, 4 Place Jussieu, F-75005 Paris, France.
mary.tabut@sorbonne-universite.fr

The discovery of stable molecule–surface interfaces is a key bottleneck in heterogeneous catalysis, energy storage, and electronic materials design. While Density Functional Theory (DFT) provides reliable atomic-scale accuracy, the exhaustive exploration of surface–adsorbate configurational space remains computationally challenging, limiting large-scale data-driven materials innovation[1]. However, recent advances in the application of machine learning algorithms for the prediction of novel structures are driving a new and exciting era in data-driven materials design[2]. Here, we present the ML-SAPIE code, a Python package for Machine Learning–driven Surface Adsorption Prediction, Interpretation & Exploration, designed to bridge the gap between high-throughput computational materials modelling and predictive machine learning. Built on the AiiDA WorkGraph framework[3], ML-SAPIE establishes an autonomous pipeline that integrates bulk optimization, surface construction[4], adsorption configuration generation, and large-scale DFT relaxations using VASP. The resulting database is explored and analysed in order to store physics- and chemistry- based descriptors that capture local atomic environments and electronic fingerprints to link it to the global stability of each systems. The extracted descriptors are then used to train machine learning models for predictive stability assessment, enabling rapid screening and energetic predictions of new surface-molecule systems with DFT accuracy. Beyond property prediction, ML-SAPIE supports reverse interface engineering by generatively exploring chemical space and proposing stable configurations prior to explicit quantum evaluation. As a proof of concept, we investigate cysteine adsorption on gold surfaces. The framework autonomously recovers known stable binding modes[5] while identifying previously unreported minima, demonstrating its capacity for both validation and discovery. By combining automated quantum simulations, data-driven modelling, and scalable machine learning algorithms, ML-SAPIE contributes to the development of intelligent materials workflows and advances machine learning-enabled materials discovery for catalytic and energy applications.

References

1. Hammer, B., & Nørskov, J. K., Advances in catalysis, 2000, Vol. 45, pp. 71-129.
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A., Nature, 2018, *559*(7715), 547-555.
3. Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Pizzi, G., Scientific data, 2020, vol. 7, no 1, p. 300.
4. Dorini, T.T., & San-Miguel, M.A., *Applied Surface Science*, 2025, 164350.
5. Tabut, M., Stishenko, P. V., & Calatayud, M., Surface Science, 2025, *757*, 122740.

Figures

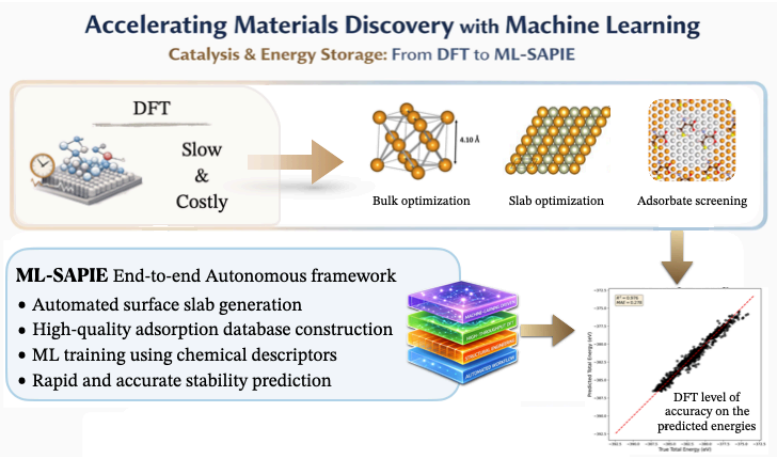

Figure 1. Computational pipeline of the ML-SAPIE code: From bulk optimization to ML-based energy prediction.