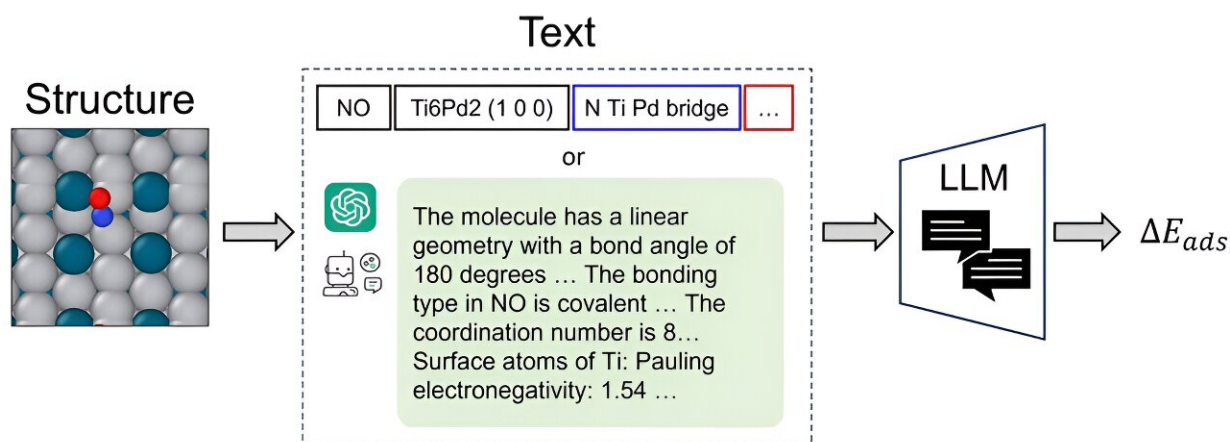# In a world run by catalysts, why is optimizing them still so tough?

January 3 2024, by Kaitlyn Landram

Graphical abstract. Credit: *ACS Catalysis* (2023). DOI: 10.1021/acscatal.3c04956

We depend on catalysts to turn our milk into yogurt, to produce Post-It notes from paper pulp, and to unlock renewable energy sources like biofuels. Finding optimal catalyst materials for specific reactions requires laborious experiments and computationally intensive quantum chemistry calculations.

Oftentimes, scientists turn to graph neural networks (GNNs) to capture and predict the structural intricacy of atomic systems, an efficient system only after the meticulous conversion of 3D atomic structures into precise spatial coordinates on the graph is complete.

CatBERTa, an energy prediction Transformer model, was developed by researchers in Carnegie Mellon University's College of Engineering as an approach to tackle molecular property prediction using machine learning.

"This is the first approach using a large language model (LLM) for this task, so we are opening up a new avenue for modeling," said Janghoon Ock, Ph.D. candidate in Amir Barati Farimani's lab.

A key differentiator is the model's ability to directly employ text (natural language) without any preprocessing to predict the properties of the adsorbate-catalyst system. This method is notably beneficial as it remains easily interpretable by humans, allowing researchers to integrate observable features into their data seamlessly.

Additionally, applying the transformer model in their research offers substantial insights. The self-attention scores, particularly, are crucial in enhancing their comprehension of interpretability within this framework.

"I can't say that it will be an alternative to state-of-the-art GNNs, but maybe we can use this as a complementary approach," said Ock. "As they say, 'The more the merrier.'"

The model delivers predictive accuracy comparable to that achieved by earlier versions of GNNs. Notably, CatBERTa was more successful when trained on limited-size data sets. Additionally, CatBERTa has surpassed the error cancellation abilities of existing GNNs.

The team focused on adsorption energy but said that the approach can be extended to other properties, such as the HOMO-LUMO gap and stabilities related to adsorbate-catalyst systems, given an apt dataset.

By integrating the capabilities of extensive language models with the demands of catalyst discovery, the team aims to streamline the process of effective catalyst screening. Ock is working to improve the accuracy of the model.

The findings are published in the journal *ACS Catalysis*.

More information: Janghoon Ock et al, Catalyst Energy Prediction with CatBERTa: Unveiling Feature Exploration Strategies through Large Language Models, *ACS Catalysis* (2023). DOI: 10.1021/acscatal.3c04956

Provided by Carnegie Mellon University Mechanical Engineering