CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday - October 27, 2023

Catalyst Property Prediction with CatBERTa: Unveiling Feature Exploration Strategies through Large Language Models

Janghoon Ock, Carnegie Mellon University

Efficient catalyst screening depends on predictive models for adsorption energy, which correlates with reactivity. Although Graph Neural Networks (GNNs) are leading machine learning tools for atomic system modeling, they necessitate accurate atomic coordinates to form graph representations. This can be challenging when trying to incorporate observable attributes. Addressing this, we introduce CatBERTa, a Transformer-based model designed to predict adsorption energy from textual inputs. Benefiting from a pretrained Transformer encoder, CatBERTa interprets human-readable text that integrates desired features. Analyzing attention scores reveals that CatBERTa prioritizes tokens concerning adsorbates, bulk composition, and their corresponding interacting atoms. Intriguingly, these interacting atoms emerge as compelling descriptors for adsorption configurations. Conversely, aspects like bond length and inherent atomic properties yield marginal predictive value. CatBERTa, by interpreting the text representation of initial structures, delivers a mean absolute error (MAE) of 0.75 eV—on par with conventional GNNs. Moreover, when CatBERTa's predicted energies are adjusted, systematic errors diminish by as much as 19.3% for chemically akin systems, outpacing the error reduction seen in GNNs. This highlights CatBERTa's potential to enhance energy difference predictions. Our work pioneers a textual approach to catalyst property prediction, bypassing the need for graphical depictions and unveiling nuanced feature-property linkages. This research underscores the viability of text as an alternative representation for atomic configurations.