ACS Catalysis

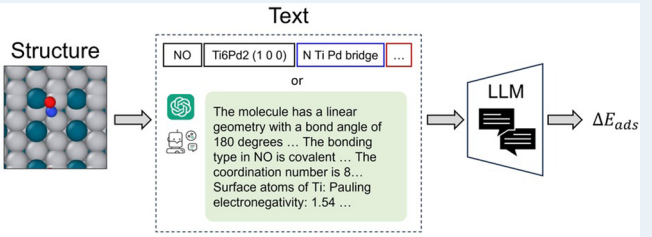
pubs.acs.org/acscatalysis
Research Article

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

Janghoon Ock, Chakradhar Guntuboina, and Amir Barati Farimani*

Cite This: ACS Catal. 2023, 13, 16032−16044
Read Online

ABSTRACT: Efficient catalyst screening necessitates predictive models for adsorption energy, which is a key descriptor of reactivity. Prevailing methods, notably graph neural networks (GNNs), demand precise atomic coordinates for constructing graph representations, while the integration of observable attributes remains challenging. This research introduces CatBERTa, an energy prediction Transformer model that uses textual inputs. Built on a Transformer encoder pretrained for language modeling purposes, CatBERTa processes human-interpretable text, incorporating target features. Attention score analysis reveals CatBERTa's focus on tokens related to adsorbates, bulk composition, and their interacting atoms. Moreover, interacting atoms emerge as effective descriptors for adsorption configurations, while factors such as the bond length and atomic properties of these atoms offer limited predictive contributions. In predicting the adsorption energy from textual representations of initial structures, CatBERTa exhibits a precision comparable to that of conventional GNNs. Notably, in subsets recognized for their high accuracy with GNNs, CatBERTa consistently achieves a mean absolute error of 0.35 eV. Furthermore, the subtraction of the CatBERTa-predicted energies effectively cancels out their systematic errors by as much as 19.3% for chemically similar systems, surpassing the error reduction observed in GNNs. This outcome highlights its potential to enhance the accuracy of the energy difference predictions. This research establishes a fundamental framework for text-based catalyst property prediction without relying on graph representations while also unveiling intricate feature−property relationships.



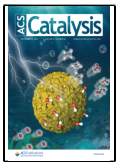
KEYWORDS: *computational catalysis, catalyst screening, renewable energy, machine learning, Transformer, large language model*

INTRODUCTION

The search for optimal catalyst materials for specific reactions poses a significant challenge in the development of sustainable chemical processes. Traditional avenues of exploration have involved laborious experiments or computationally intensive quantum chemistry calculations, exemplified by density-functional theory (DFT) simulations.[1,2] Nevertheless, the requirement to assess a vast array of systems makes the catalyst screening for optimality more challenging.[3,4] This is attributed to the fact that a singular bulk catalyst can exhibit a range of surface orientations.[3,5,6] Additionally, adsorbates have the potential to bind to numerous distinct adsorption sites on these surfaces, with varying orientations.[7] As such, relying solely on DFT calculations proves inadequate for swiftly assessing the vast array of potential adsorbate−catalyst combinations because of their time and resource demands. In response to these challenges, an increasingly prevalent approach involves harnessing the capabilities of machine-learning (ML) methodologies to expedite the prediction of catalyst properties.[8−11]

In the field of molecular property prediction, graph neural networks (GNNs) have emerged as a promising ML approach. They center on the graphical representation of molecular systems,[9,10] where atoms are depicted as nodes and bonds as edges, forming the primary input. Particularly in the context of catalysis research, these graph representations are generated by converting the structures of adsorbate−catalyst systems into graphs, effectively capturing the inherent structural intricacies of the atomic arrangements. This enables the models to discern intricate connections between structures and properties.[12−17] However, the conversion of a 3D system structure into a graph mandates precise spatial coordinates for each atom, involving the meticulous identification of the nearest neighbors within predefined proximity thresholds for individual atoms. The need for precise spatial comprehension, which is not easily attainable, can potentially introduce constraints during the initial phases of the screening process, especially when researchers aim to employ easily observable features for the screening procedure. Furthermore, GNNs function as a black box, making it difficult to assess the specific influence of

Received: October 16, 2023
Revised: November 15, 2023
Accepted: November 16, 2023
Published: November 30, 2023



https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

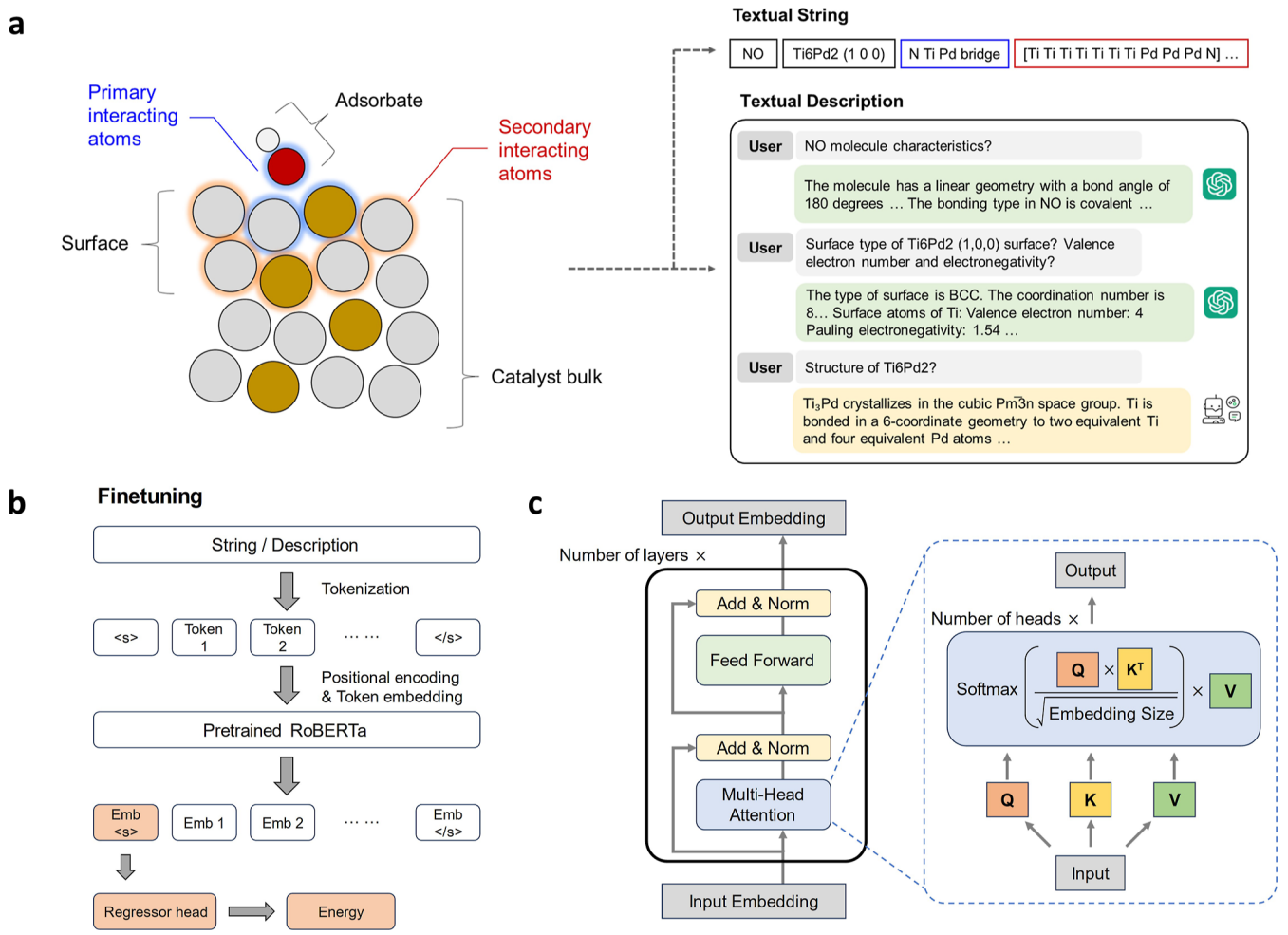


Figure 1. Overview of CatBERTa: (a) transformation of structural data into a textual format. The structural data undergoes conversion into two types of textual inputs: strings and descriptions. (b) Visualization of the fine-tuning process. The embedding from the special token "<s>" is input to the regression head, comprising a linear layer and an activation layer. (c) Illustration of the Transformer encoder and a multihead attention mechanism.

distinct physical attributes within the system on the property prediction.

The utilization of textual representations for describing adsorbate−catalyst systems presents an intriguing alternative to graph-based approaches. Unlike graph representations, textual descriptions offer a natural way to incorporate observable features in a human-interpretable manner. When modeling materials, we often possess metadata that captures specific observable attributes. For catalyst systems, this typically includes aspects such as bulk composition, the type of bulk structure, and surface orientation. Given the availability of such metadata, it makes sense to leverage it rather than rely solely on the actual atomic structure through simulation tools. Several established textual string-based representations exist for molecular and crystal structures, such as SMILES,[18] SELFIES,[19] InChI,[20] and MOFid,[21] each encoding the atomic system's structure in specific formats. Moreover, the emergence of generative language models has spurred ongoing research into generating textual descriptions for molecules, enhancing human interpretability.[22,23] Once an atomic system is represented as textual data, it can be processed by deep neural networks. However, the segmentation of text into tokens presents a significant challenge, as placing essential tokens throughout the text's potentially extensive span can introduce complexities for various neural network architectures.[24]

In recent years, an array of Transformer-based models, such as BERT,[25] RoBERTa,[26] GPT,[27] ELMo,[28] and LLAMA,[29] have showcased exceptional proficiency across diverse natural language processing tasks. The Transformer's distinctive strength lies in its adept application of an attention mechanism,[30] enabling the model to identify meaningful relationships among sentence tokens without relying solely on previous hidden states. This advancement has sparked a keen interest in exploring the Transformer's potential within chemistry and materials science.[31−37] For instance, TransPolymer employs polymer sequence representations for predicting polymer properties, leveraging pretraining through masked language modeling on unlabeled data.[34] Similarly, MOFormer, a structure-agnostic Transformer model, utilizes text string representations of metal−organic frameworks to predict their properties.[35]

Drawing on the Transformer-based large language model's (LLM) advanced abilities in comprehending textual inputs, we propose the CatBERTa model, designed with the aim of predicting catalyst properties through textual representations. This distinctive trait opens a new avenue for property prediction in catalysis research, potentially bypassing the

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

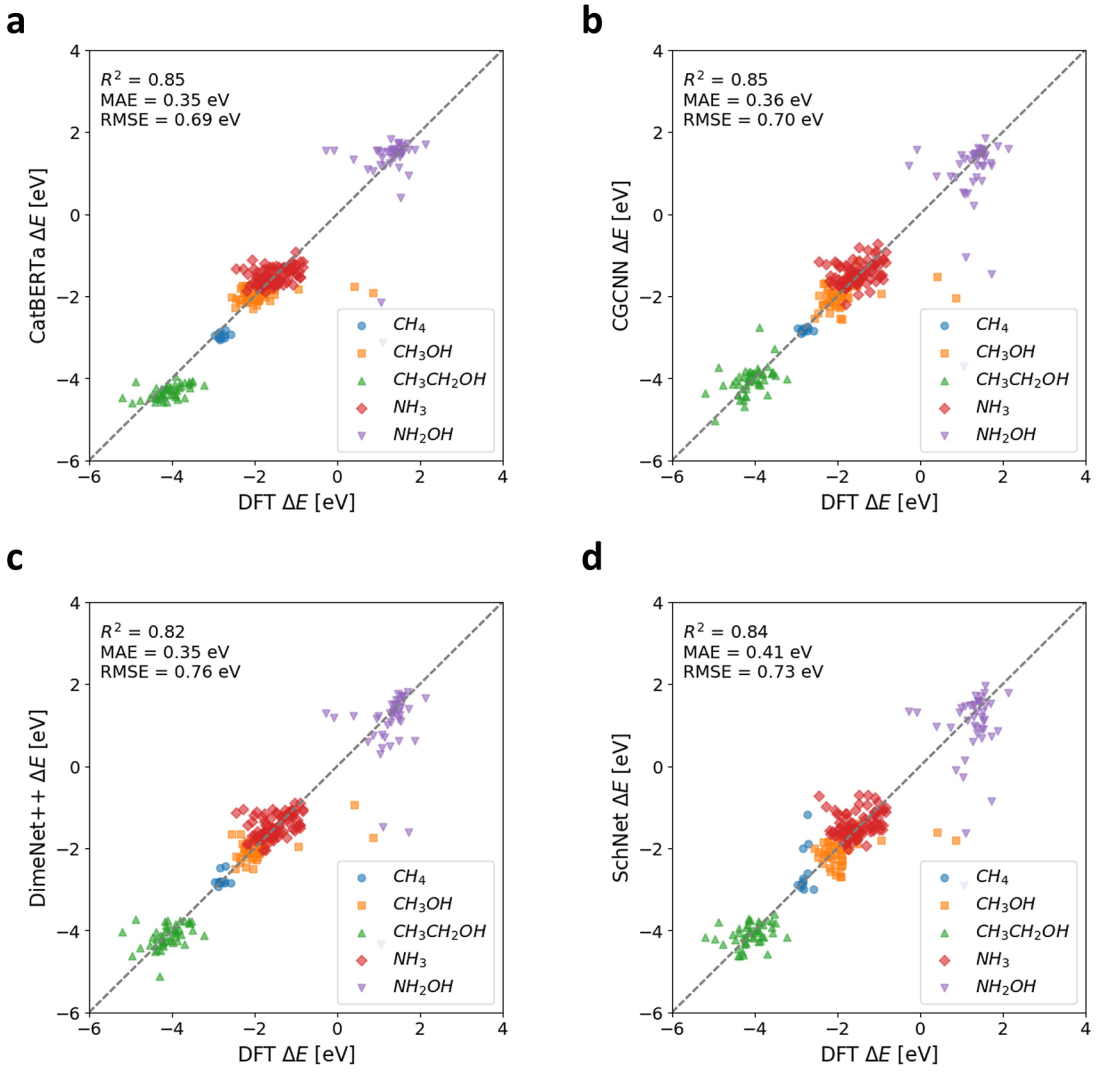


Figure 2. Comparison between DFT-calculated energy values (x-axis) and model-predicted energy values (y-axis) for systems featuring specific adsorbates: CH_4, CH_3OH, CH_3CH_2OH, NH_3, and NH_2OH. The models used for predictions are (a) CatBERTa, (b) CGCNN, (c) DimeNet++, and (d) SchNet.

need for precise 3D atomic coordinates. Moreover, this approach enables us to exercise control over the inclusion of features in the input data in a way that can be easily interpreted by humans. Scrutinizing self-attention scores provides insight into the varying importance levels assigned to different input features. Manipulating input data, coupled with interpretive analysis of self-attention scores, provides insight regarding the identification of the most effective features for predicting the desired properties. Furthermore, CatBERTa serves as a litmus test for Transformer-based LLMs, assessing their capacity to grasp the scientific significance of specific features through data-driven methods.

The prediction of the adsorption energy (ΔE_{ads}) stands out as a primary focus in ML modeling for catalysis research. It serves as a pivotal descriptor directly linked to catalyst reactivity.[38−42] The Open Catalyst 2020 (OC20) data set encompasses an extensive collection of over 1.2 million DFT relaxations of adsorbate−catalyst systems.[10] This data set serves as the foundation for training ML models. Over successive development of GNNs for molecular property prediction has exhibited enhanced predictive accuracy for adsorption energies in the OC20 data set. Illustrated by examples such as GemNet-OC[16] and SCN,[17] leading-edge GNNs have achieved remarkable results in energy prediction. Notably, they have almost approached the DFT level accuracy by achieving mean absolute error (MAE) values as low as 0.28 eV.[43] Given the importance of adsorption energy prediction in catalysis research, our proposed CatBERTa model showcases its capability in property prediction for this task.

This paper presents CatBERTa, a Transformer-based model designed to predict the adsorption energy of adsorbate−catalyst systems solely through textual inputs. CatBERTa improves interpretability by enabling researchers to integrate human-interpretable features into input data and identify the tokens in the textual representation that capture more model attention. This understanding aids in pinpointing the crucial catalyst attributes. Furthermore, the CatBERTa model demonstrates the efficacy and potential of Transformer-based LLMs for predicting properties of catalytic systems. By integrating the capabilities of extensive language models with the demands of catalyst discovery, we aim to streamline the process of effective catalyst screening.

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

RESULTS AND DISCUSSION

CatBERTa Framework. We utilize a pretrained Transformer model, trained on a large corpus of natural language text, to predict the energy of the adsorbate−catalyst system, allowing us to thoroughly explore the feature space. To generate training data, we transform the initial 3D structures of DFT trajectories sourced from the OC20 data set into easily understandable textual representations. Within this context, we use the publicly available 10k and 100k initial-structure-to-relaxed energy (IS2RE) data sets as textual inputs for training our model. Our focus is on predicting the relaxed energy from the initial structure using initial structures from these 10k and 100k DFT trajectories. As illustrated in Figure 1a, our conversion process involves two distinct approaches: first, transforming the structure into a textual string containing exclusively the desired input features; second, generating a natural language description that elucidates the material's structure and characteristics.

At the core of our framework, we employ the encoder of RoBERTa, a robustly optimized variant of the Bidirectional Encoder Representations from Transformers (BERT) model. This serves as the engine for our framework. The generated textual inputs undergo tokenization and are then passed through the pretrained RoBERTa encoder (Figure 1b). Subsequently, the first token embedded from the encoder is used in the regression head to predict the scalar energy value. The model undergoes fine-tuning using both the 10k and 100k textual data sets, with detailed data set information provided in the Methods section.

The foundation of Transformer encoders comprises a series of stacked self-attention and pointwise fully connected layers, as depicted in Figure 1c. In contrast to the architectures of recurrent neural networks (RNNs), the Transformer model relies on the self-attention mechanism to establish meaningful connections between tokens positioned at different points within a sequence. This process of self-attention involves scaled dot-product attention, centered around the query, key, and value matrices. In our specific setup, the Transformer encoder consists of 12 hidden layers, each of which accommodates 12 attention heads. Additional details concerning the hyperparameters are listed in Table S4.

Input Feature Exploration. The adsorbate−catalyst system involves the adsorption of an adsorbate molecule onto a catalytic surface. The amount of energy required for adsorption hinges upon numerous factors such as the adsorbate and catalyst composition, surface orientation, and particular adsorption configuration. Our research endeavors to unravel the underlying factors that drive energy predictions for adsorbate−catalyst systems, leveraging the Transformer which is predrilled for large language modeling purposes. In pursuit of our goal, we curate a diverse array of input text strings and descriptions sourced from the structural data in the OC20 data set. Each of these inputs illuminates distinct facets of the adsorbate−catalyst system, contributing to a comprehensive understanding.

We evaluate the prediction accuracy of each textual format across both the full validation data set and a smaller subset, as shown in Figure 2. We select systems with adsorbates, such as CH_4, CH_3OH, CH_3CH_2OH, NH_3, and NH_2OH, which exhibit a strong prediction accuracy (below 0.4 eV) when using DimeNet++. Evaluating accuracy across the full data set offers insights into how the input features generalize across diverse adsorbate−catalyst systems. Furthermore, if a subset achieves high accuracy with GNNs, it suggests that their important features are adeptly represented by the structure. This becomes a valuable test for gauging how well our text strings, emphasizing structural attributes, can elucidate the link between structure and energy.

Our subsequent efforts involve refining the input data by incorporating adsorption configurations (Table 1). We begin by introducing information about the primary interacting atoms within both the adsorbate and the catalyst bulk, as illustrated in string type 2. We set out on a quest to enrich our feature set by introducing additional atomic properties in string type 3, spanning atomic mass, periodicity, dipole polarizability, electronegativity, and electron affinity. For string type 4, we extend the adsorption configuration feature landscape by incorporating secondary interacting atoms within the catalyst bulk. The identification of atoms situated within a covalent radius of the primary interacting atoms on the surface is facilitated through the Pymatgen package.[44]

Table 1. Different Compositions of Features in Input Strings and Their Corresponding MAE for the Validation Set[a]

no	textual string format	MAE [eV]	
		subset	total
string 1	<s> adsorbate symbol </s> bulk symbol (Miller index) </s>	0.44	0.85
string 2	string 1 + [primary interacting atoms, site type]	0.37	0.79
string 3	string 1 + [primary interacting atoms, site type] </s>[atomic properties]	0.47	0.79
string 4	string 1 + [primary interacting atoms, site type] [secondary interacting atoms]	0.35	**0.75**
string 5	string 1 + [primary interacting atoms, site type, bond distance] [secondary interacting atoms]	**0.34**	0.77

[a]The outcomes result from training the model on a data set of 100k and its subset featuring adsorbates such as CH_4, CH_3OH, CH_3CH_2OH, NH_3, and NH_2OH. The best performance cases are highlighted in bold text. Actual example strings are provided in Table S1.

In crafting string-type inputs, we compose the textual string only with the essential target features in a condensed manner. This approach minimizes the inclusion of semantically vacant vocabulary, which could introduce unwanted noise. Our approach involves experimenting with five types of textual input strings. Initially, we concentrate on information relevant to the adsorbate, bulk, and surface, deliberately excluding explicit adsorption configuration details. This initial approach results in an MAE of 0.85 eV across the entire validation set and 0.44 eV for the subset. While this level of performance does not meet the stringent requirements of practical applications, it is crucial to acknowledge that even these preliminary features contribute to capturing the fundamental correlation (refer to Figure S2).

Our subsequent efforts involve refining the input data by incorporating adsorption configurations. We begin by introducing information about the primary interacting atoms within both the adsorbate and the catalyst bulk. This leads to a notable 16% reduction in the subset MAE to 0.37 eV compared with string type 1. Furthermore, it results in a 7% decrease in the overall MAE, settling at 0.79 eV. This aligns with our hypothesis: subsets demonstrating high accuracy with

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

Table 2. Different Compositions of Features in the Input Description and Their Corresponding MAE for the Validation Set[a]

no	entity	input feature	tool	MAE [eV]	
				subset	total
description 1	system	adsorbate symbol, bulk symbol, Miller index, primary interacting atoms, adsorption site type	Pymatgen	0.50	0.84
	adsorbate	boding type, angle, length, molecule size, dipole moment, orbital characteristics	ChatGPT		
	catalyst	composition, space group, bonding geometry, length, atom arrangement	RoboCrystallographer		
description 2	system	adsorbate symbol, bulk symbol, Miller index, primary interacting atoms, adsorption site type	Pymatgen	0.43	0.79
	adsorbate	central atom, coordination number	ChatGPT		
	catalyst	surface type, coordination number, valence electron number, electronegativity	ChatGPT		

[a]The outcomes arise from training the model using a data set of 100k and a subset with specific adsorbates. If the model is exclusively trained using system descriptions, the resulting MAE is 0.79 eV.

GNNs should also perform well when the textual string emphasizes structural attributes.

Encouraged by this advancement, we aim to expand our feature set, incorporating various atomic properties, spanning atomic mass, periodicity, dipole polarizability, electronegativity, and electron affinity. Despite these enrichments, the resulting enhancements in overall accuracy are marginal, and intriguingly, the subset's accuracy saw a reduction. This suggests that for subsets already performing commendably with mere structural descriptions in the text format, the supplementary atomic properties might introduce unnecessary noise. When compared with results from string type 2, it appears that these additional properties could benefit certain other subsets, especially given that the overall MAE remains consistent across the two types.

We extend the adsorption configuration feature landscape by incorporating secondary interacting atoms within the catalyst bulk. The identification of atoms situated within a covalent radius of the primary interacting atoms on the surface is facilitated through the Pymatgen package.[44] This refinement contributes to an improvement, reducing the overall MAE to 0.75 eV and the subset MAE to 0.35 eV from previous values of 0.79 and 0.37 eV, respectively. This supports the great performance of GNN since the graph topology naturally embeds the primary and secondary interacting atoms by capturing the whole geometry.

However, our attempt to further enhance accuracy by incorporating bond lengths of primary interacting atoms is counterproductive for the entire data set. The overall MAE increases to 0.77 eV, which suggests model overfitting in the presence of the distance data. Yet, for the subset, the inclusion of distance either slightly improves or maintains performance when compared to string type 4. This indicates that the distance feature does not consistently serve as a dependable predictor across data sets. These insights enable a more informed exploration of the feature landscape, facilitating assessments of the effects of specific target feature additions.

Unlike the previous string-based approach that requires manual feature integration, which can be a time-consuming task, we strive to streamline the process by automating the generation of input text. To achieve this, we leverage the capabilities of generative language models, specifically ChatGPT[45] and RoboCrystallographer,[46] to generate textual descriptions in response to our queries. Using this method, we construct input text with targeted features in three main sections: the first section provides an overview of the system, covering the adsorbate, catalyst bulk composition, and surface details; the second section focuses on the adsorbate; and the final section addresses the catalyst, as depicted in Table 2. The query prompts employed to generate these descriptions can be found in Table S2. Although ChatGPT's outputs might have imperfections, we aim to showcase an automated method for generating feature-rich textual descriptions. A detailed assessment of the accuracy of these descriptions is available in Supporting Information S4.

Initially, we constructed the text descriptions with general chemical and structural attributes, operating independently from prior knowledge of adsorption energy modeling. For an in-depth exploration of adsorbate characteristics, we obtain information on bonding types, molecular sizes, bond angles, lengths, and dipole moments using ChatGPT 3.5 version.[45] For catalyst bulk descriptions, RoboCrystallographer[46] provides textual representations of the catalyst bulk in the OC20 data. This approach produces detailed textual representations of both the adsorbate and the catalyst. When the model is trained solely on the system section, containing features equivalent to those of string type 2, the resulting overall MAE stands at 0.79 eV. This underscores the capacity of the language model to discern critical features through its inherent self-attention mechanism. However, the augmentation with adsorbate and catalyst bulk descriptions leads to a marginal increase in the overall MAE, reaching 0.84 eV. This increase indicates that the augmented features introduce some level of noise into the model without substantively enhancing data fitting.

In the subsequent refined phase, the description is generated with scientifically grounded information, derived from the modeling study of adsorption energy.[47] Recognizing the profound influence of factors like the coordination number of the central adsorbate atom and its interactions with counterparts on the surface, coupled with the electronegativity of these surface atoms,[47] we engage ChatGPT-3.5 in generating textual portrayals of these features. This approach yields an overall MAE of 0.79 eV. It is noteworthy that this MAE remains lower than the outcome derived from previous text descriptions encompassing general chemical characteristics. This discovery underscores that features within science-based descriptions do not compromise prediction accuracy, unlike the outcomes obtained from general chemical characteristics. Consequently, this revelation implies that the coordination number and electronegativity hold a greater magnitude of significance compared to structural and characteristic descriptions of the adsorbate and the bulk crystal. Our data-driven approach provides support to the modeling study conducted by Gao et al. (2020),[47] offering a cross-check methodology for

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

Figure 3. Visualization of the attention scores from CatBERTa. The left column displays attention scores from the initial hidden layer, while the right column presents visualizations from the final hidden layer. The top row uses string type 4, the middle row has description type 1, and the bottom row displays description type 2 for the same system.

the physics-based approach. This method enables the assessment of the specific features' influence on property predictions.

For subset accuracy, description type 2 has a higher MAE compared to string type 2 even though both contain equivalent system information and they exhibit similar overall MAEs. This suggests that for subsets already performing well simply with structural information, the added depiction of coordination number and electronegativity does not enhance accuracy. However, since the overall MAE remains consistent, other subsets likely benefit from this additional information.

Self-Attention Visualization. The attention score, which acts as an indicator of the relationship between two tokens, offers insights into CatBERTa's acquisition of chemical knowledge and the individual contributions of each token to prediction outcomes. To illustrate, we choose an example adsorbate−catalyst system with NH_3 as the adsorbate and VCr_3 with a Miller index of (2, 1, 0) as the catalyst. This selection is based on systems displaying an MAE below 0.05 eV for all instances of string type 4 and description types 1 and 2. Employing the Attention Visualizer package,[48] we calculate the average attention scores from the 12 attention heads for each token and merge them for each vocabulary. In Figure 3, we present a visualization of these scores for both the initial and final hidden layers, where a stronger color represents a higher attention score.

In the attention score of the first hidden layer, strong relationships between nearby tokens are apparent, as evidenced by dispersed attention scores across the text. As the tokens pass through a sequence of the hidden layers, they become concentrated on specific parts of the text. This transition leads to fewer vocabulary elements receiving high attention in the final hidden layer, in comparison to the attention observed in the initial hidden layer.

When presented with input in the string format, the model inherently places greater emphasis on the interactions between surface and adsorbate atoms. Notably, in the context of secondary interacting atoms such as [Cr Cr Cr Cr V V V N] [Cr Cr Cr Cr V V V N], where the initial Cr atoms within both brackets pertain to primary interacting atoms, the model exhibits heightened attention toward the N atom—an

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

Table 3. Attention Scores between the "<s>" Token and Other Tokens within Individual String Sections across the Full Validation Data Set[a]

	before fine-tuning				after fine-tuning			
no	<s>	adsorbate	catalyst	configuration	<s>	adsorbate	catalyst	configuration
string 1	0.36	0.09	0.55		0.05	0.44	0.52	
string 2	0.23	0.06	0.34	0.37	0.03	0.20	0.38	0.40
string 4	0.21	0.06	0.32	0.41	0.04	0.24	0.13	0.59
string 5	0.21	0.05	0.29	0.45	0.03	0.23	0.22	0.52

[a]The attention scores are derived from the last hidden layer and averaged across all 12 attention heads. The column labeled "<s>" indicates the self-attention score between the "<s>" token and itself. Refer to Tables 1 and S1 for detailed strings.

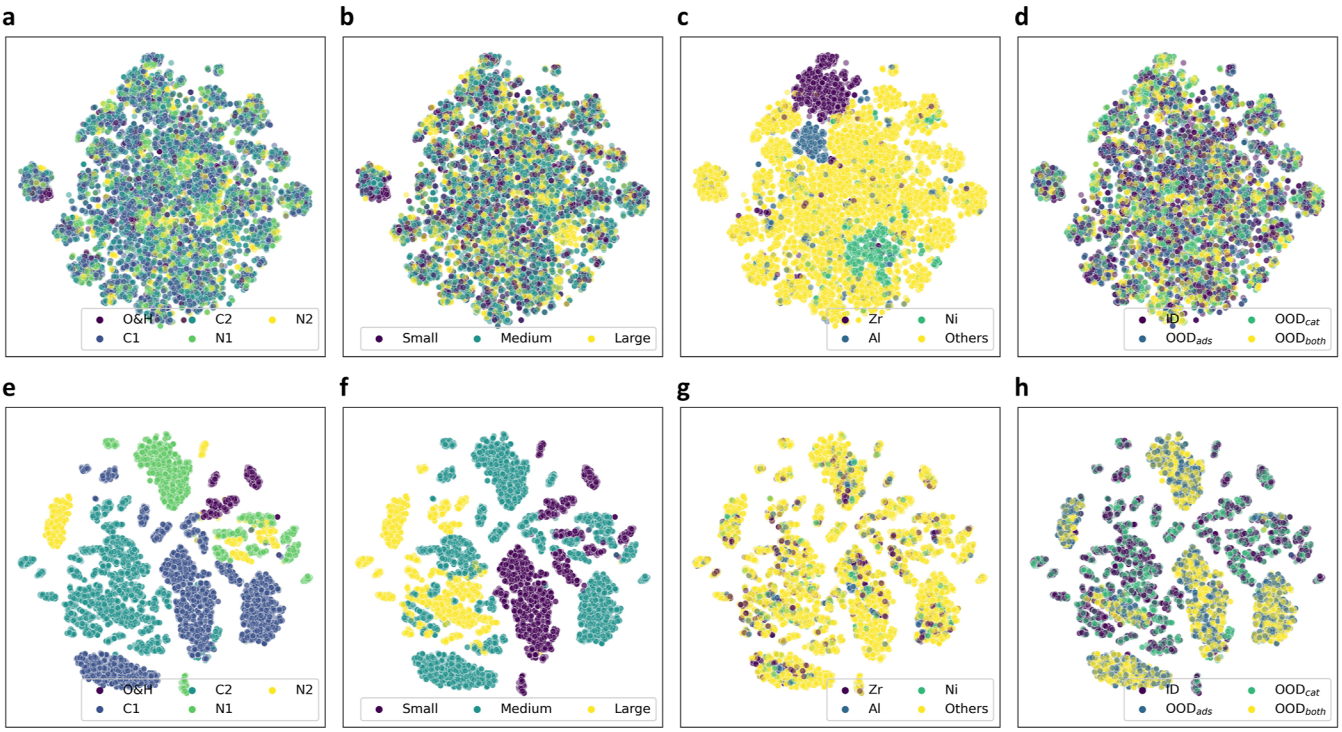


Figure 4. *t*-SNE visualization of first "<s>" token embeddings before and after fine-tuning. The upper row plots (a–d) display *t*-SNE plots before fine-tuning, while the lower row plots (e–h) represent after fine-tuning. (a,e) illustrate the adsorbate type. (b,f) indicate the size of the adsorbate molecule. For molecules with fewer than 3 atoms, it is small; for more than 5 atoms, it is large. (c,g) portray the bulk material which consists of certain atoms, specifically the top three common elements in the OC20 data set: Zr, Al, and Ni. (d,h) depict the validation set splits. The in-domain split comprises adsorbates and catalysts seen during training, whereas the out-of-domain splits include adsorbates or bulk materials not encountered during training.

adsorbate atom engaged in interactions, despite the absence of explicit emphasis on the significance of interacting atoms in both entities.

Upon contrasting the attention scores from the final layer for text descriptions 1 and 2, a noticeable pattern emerges: the attention score in description 1 displays greater dispersion compared with that in description 2. The adsorbate and the catalyst are the main focuses of the model in the setting of description 2, with elements like bond count, electron amount, and electronegativity receiving secondary consideration. Conversely, in description 1, numerous vocabulary terms with limited semantic significance attract substantial attention. This could potentially contribute to the reduced accuracy observed in the case of description 1.

The attention score corresponding to each section in the final layer provides insights into the level of attention apportioned to each section during the energy prediction process. Specifically, we calculate the attention scores between the "<s>" token and other tokens, given that the embedding of the "<s>" token is processed through the regression head to produce the scalar energy value. Furthermore, the attention score of the "<s>" token in the final layer is averaged across all 12 attention heads. Subsequent calculations yield the average attention scores for various sections: the adsorbate, catalyst, and configuration, which are listed in Table 3.

A comparison of attention scores before and after fine-tuning reveals that fine-tuning with the energy label noticeably redirects attention toward the adsorbate section. This transition also can be seen in Figure 3, where the adsorbate symbol gains heightened attention as tokens pass through the hidden layers. Moreover, the inclusion of primary and secondary interacting atoms within string types 2 and 4 prompts a substantial attention shift toward the configuration section, correlating with a tangible enhancement in accuracy. Intriguingly, a converse trend emerges with string type 5: extending the configuration section to incorporate distances for primary interacting atoms diminishes the attention score to 0.52, compared to 0.59 observed in the case of string type 4.

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032–16044

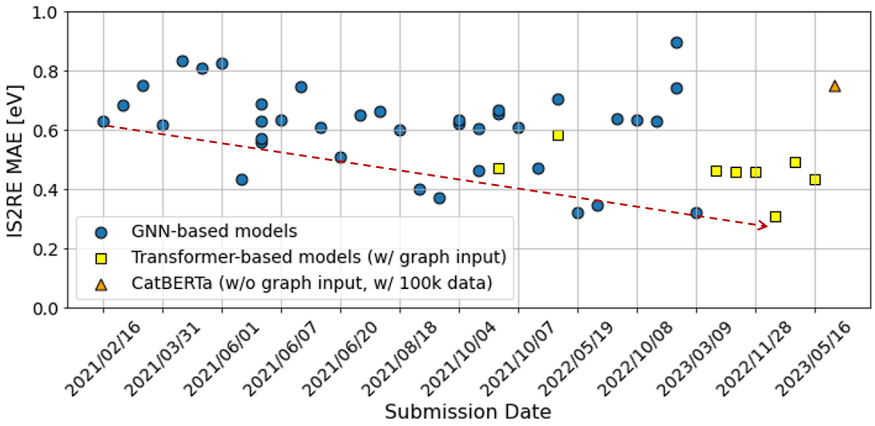


Figure 5. Open Catalyst Project leaderboard results. Model performance is evaluated using the IS2RE task with the OC20 data set. The MAE values are obtained from the public leaderboard of the Open Catalyst Project.[43]

Table 4. MAE Values for the Energy Prediction across Various Models

		MAE [eV] (lower is better)					
data size	model	subset	total	ID	OOD_{ads}	OOD_{cat}	OOD_{both}
10k	CGCNN	0.66	0.83	0.88	0.85	0.80	0.79
	SchNet	0.73	0.90	0.89	0.93	0.87	0.90
	DimeNet++	0.51	0.76	0.77	0.80	0.72	0.75
	CatBERTa	0.46 ± 0.02	0.82 ± 0.02	0.75 ± 0.02	0.95 ± 0.02	0.71 ± 0.01	0.88 ± 0.02
100k	CGCNN	0.36	0.63	0.58	0.72	0.55	0.66
	SchNet	0.41	0.61	0.61	0.61	0.64	0.58
	DimeNet++	0.35	0.57	0.54	0.63	0.52	0.60
	CatBERTa	0.35 ± 0.01	0.75 ± 0.01	0.65 ± 0.01	0.90 ± 0.01	0.61 ± 0.01	0.86 ± 0.02

We can gain an initial understanding of interpretability with respect to the model's emphasis on specific features. However, the attention scores might lack a comprehensive representation of the interaction between tokens and prediction outcomes.[49] This disparity arises from the omission of value matrix considerations in their evaluative approach. We acknowledge that a comprehensive resolution of the challenge in interpreting attention demands thorough additional investigation and concurrent research efforts.

Features Captured by Latent Space. The impact of the training process can be illustrated by examining how the representations are captured in the latent space. This phenomenon is effectively visualized through the *t*-SNE plot shown in Figure 4. The *t*-SNE plots specifically correspond to results from string type 4. The embeddings of the first "<s>" token, which traverse the regression head, are subjected to *t*-SNE analysis for visualization. The upper row of plots showcases *t*-SNE visualizations derived from a pretrained RoBERTa model that has not undergone our fine-tuning procedure with energy prediction. Conversely, the bottom row of plots corresponds to the *t*-SNE representations generated after our fine-tuning process.

Before undergoing fine-tuning, the embeddings exhibit a constraint in effectively capturing the type and size of the adsorbate. Embeddings associated with the same type and size of adsorbate do not cluster together; instead, they appear intermixed, as showcased in Figure 4a,b. Conversely, the model demonstrates adeptness in distinguishing the bulk that contains specific atoms. The embeddings stemming from systems containing Zr, Ni, and Al form distinct clusters, as illustrated in Figure 4c. Notably, given that a substantial proportion of atoms listed in both the primary and secondary interacting lists belong to the catalyst bulk, it is unsurprising that the model allocates more attention toward the bulk section in contrast to the adsorbate section.

In contrast, post-fine-tuning, the embeddings effectively capture distinctions in the type and size of the adsorbate. This is evident as embeddings featuring the same adsorbate type and size cluster together in the latent space, as depicted in Figure 4e,f. The embeddings extracted from systems specifically containing certain atoms in the catalyst bulk no longer form distinct clusters but instead become intermixed, as shown in Figure 4g. The fine-tuning process, conducted without explicit specifications of adsorbate types and sizes, directs the model's focus more keenly toward the adsorbate section. This observation is consistent with the increased attention score for the adsorbate section post-fine-tuning, as shown in Table 3.

Furthermore, notable is the model's competence in discerning the size of the adsorbate after fine-tuning, as this attribute can be deduced from the number of tokens allocated to the adsorbate section. Unlike the adsorbate type, which is inherently reflected in the tokens, the size relies on the quantity of tokens allocated to this section. Hence, this implies that the integration of positional embedding enhances the extraction of more meaningful information from the provided tokens.

Energy Prediction Results. Since the project's inception, numerous models have been submitted, leading to a consistent reduction in the overall MAE over time, as shown in Figure 5. GNNs have prominently driven this progress. Furthermore, the recent strides in submissions have been made by the emergence of Transformer-based or Transformer-assisted models like Equiformer[50] and Molformer.[51] These models

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

have demonstrated noteworthy enhancements in performance, which closely parallel the ongoing surge in the prominence of the Transformer architecture. Both these GNNs and cutting-edge Transformer models rely on node and edge embeddings derived from graph representations, which necessitate precise atomic coordinates to establish a connectivity network. Specifically, for the IS2RE task, the atomic coordinates in the initial structure must be provided in their entirety to construct the graph representation. In contrast, our CatBERTa model, presented in this research, stands apart as the first Transformer-based approach that operates without the need for graph embeddings. Instead, it leverages human-interpretable textual input to grasp a physical understanding of the system.

The prediction results of CatBERTa fall within the range of 0.3−0.9 eV, the results seen in the outcomes of the other submitted models. To evaluate the performance of CatBERTa, we conduct a comparative analysis against earlier versions of GNNs, and the results are presented in Table 4. Specifically, we employ CatBERTa trained and validated using string type 4 and compare its performance with publicly available checkpoints of CGCNN, SchNet, and DimeNet++, all of which were trained on the same data set size as CatBERTa. Our evaluation encompasses distinct data splits, including in-domain (ID) samples drawn from the training distribution, as well as out-of-domain (OOD) samples for adsorbates (OOD_{ads}), catalysts (OOD_{cat}), and both entities (OOD_{both}), which encompasses previously unseen adsorbate and catalyst compositions.[10]

When trained with the 10k data set, CatBERTa demonstrates an overall MAE of 0.82 eV, outperforming both CGCNN and SchNet. Its heightened performance stands out, especially in predictions for the selected subset, a group of systems recognized for their strong performance in GNNs, where CatBERTa outshines other GNN models. Moreover, for both ID and OOD catalyst splits, CatBERTa displays greater accuracy than its GNN counterparts. These results highlight the effectiveness of the CatBERTa model when trained with limited data. However, its predictive capacity shows a relatively diminished performance when dealing with unfamiliar adsorbate molecules, signifying a limitation of the LLM-based CatBERTa model.

Despite CatBERTa being trained on an expanded data set of 100k instances, resulting in an appreciable 8.5% reduction in MAE, this advancement remains comparatively modest when compared to the substantial improvements observed in other GNNs. These models exhibit MAE reductions ranging from 24 to 32%. For the 100k training case, CatBERTa falls short of surpassing its predecessors among GNNs in terms of performance. However, when the selected subset is focused, CatBERTa maintains competitive accuracy, even if its overall MAE does not surpass other GNN benchmarks. This suggests that textual strings, highlighting structural features, can effectively discern the relationship between the structure and the energy. As depicted in Figure 5, the performance of CatBERTa aligns with the range of other submitted results. It is noteworthy that this level of performance is achieved using solely a 100k data set, in contrast to the results achieved using the complete 460k training data set.

Error Cancellation for Chemically Similar Systems. Every model shown in Figure 5, including CatBERTa, is designed to predict individual energy values. Nevertheless, the prediction of relative energy differences holds paramount practical implications compared to the individual energy prediction itself, notably in tasks such as screening for optimal catalysts employing volcano plots,[52−54] computing reaction energies,[39,54,55] and finding the most stable configuration.[7,42] These applications hinge on discerning energy disparities between chemically similar systems. For instance, the energy difference between the two adsorbed states can be derived by subtracting their respective adsorption energies. Similarly, a volcano plot can be constructed by contrasting the adsorption energies of adsorbate−catalyst systems featuring the same catalyst but varying adsorbates. Thus, an evaluation of the energy difference prediction performance is imperative.

The difference in adsorption energy ($\Delta\Delta E_{ij}$) between two systems, denoted as i and j, is computed by subtracting two ML-predicted energies (ΔE_i, ΔE_j), as expressed in the equation below

$$(\Delta E_i + \varepsilon_i) - (\Delta E_j + \varepsilon_j) = \Delta\Delta E_{ij} + \varepsilon_{ij} \quad (1)$$

When energy differences are calculated by subtracting individual energy predictions, systematic errors are canceled out. In cases where the signs of errors in these two energy predictions (ε_i, ε_j) align, there is the potential for a partial reduction in the error of the energy difference (ε_{ij}). This phenomenon is grounded in the principle of error propagation, elucidated in Supporting Information S7. Consequently, the precision of energy difference computation can be significantly enhanced if the errors associated with individual energy predictions are effectively canceled out through subtraction. Hence, the extent of error cancellation emerges as a significant metric, alongside the precision of individual energy prediction, capable of reflecting the potential to enhance the accuracy of energy difference calculations.[56]

Moreover, owing to the approximations inherent in DFT functionals that often give rise to systematic errors with correlations to the atomic structure of a system, these errors see considerable cancellation when comparing energies of akin systems.[57−59] Consequently, ML models intended to surrogate DFT calculations should effectively replicate the error cancellation phenomena observed in DFT calculations.

The magnitude of error cancellation in ML models can be quantified with subgroup error cancellation ratio (SECR) as follows[56]

$$\text{SECR}\ [\%] = 1 - \frac{\text{RMSE in subgroup}}{\text{RMSE in total pairs}} \quad (2)$$

In this equation, the root-mean-square error denotes the standard deviation of the error distribution. The equation illustrates how much the error distribution narrows within a specific subgroup of energy differences in comparison to the entire set of energy difference pairs. As a result, a higher SECR value indicates strong error cancellation, a factor that can enhance the precision of the energy difference prediction. In the denominator, the set of total pairs represents all pairs in each split, including in-domain, out-of-domain-adsorbate, out-of-domain-catalyst, and out-of-domain-both. On the other hand, the numerator's subgroup is composed of pairs of systems within each split that possess at least one shared element, whether it is an adsorbate molecule or a catalyst. For instance, NH−$Al_{20}Rh_8$ and NH−N_2Ti_4 systems share the same adsorbate molecule as NH, while OCH_3−Sc_3Al and $COCH_2O$−Sc_3Al share the common catalyst.

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

Table 5. MAE and SECR Values across the Various Models[a]

	MAE [eV] (lower is better)				SECR [%] (higher is better)			
model	ID	OOD_{ads}	OOD_{cat}	OOD_{both}	ID	OOD_{ads}	OOD_{cat}	OOD_{both}
CGCNN	0.86	0.94	0.81	0.87	1.7	9.1	4.0	8.1
SchNet	0.90	0.88	0.85	0.86	2.5	4.0	1.9	3.8
DimeNet++	0.81	0.86	0.78	0.85	2.0	5.2	2.2	3.4
CatBERTa	0.95	1.01	0.87	0.95	2.7	18.2	3.8	19.3

[a]These results are obtained from the model trained using a 100k data set.

An interesting observation is that CatBERTa exhibits more pronounced error cancellation in the out-of-domain-adsorbate and out-of-domain-both splits compared to those of other GNNs. This is evident through its higher SECR values, reaching up to 19.3%, as illustrated in Table 5. Notably, even leading GNNs like GemNet-OC[16] do not achieve such a level of robust cancellation.[56] Previous study reveals that a high embedding similarity leads to robust error cancellation.[56] This observation finds further support in the *t*-SNE plots shown in Figure 4d,h, where the embeddings of out-of-domain-adsorbate and out-of-domain-both are tightly clustered after the fine-tuning process. Hence, CatBERTa shows strong error cancellation in both splits even though its accuracy in predicting individual energy values for these divisions falls behind other models. Particularly for the out-of-domain-both, CatBERTa exhibits a higher MAE of 0.2−0.3 eV when compared to other models in individual energy prediction. However, the discrepancy reduces considerably to 0.08−0.10 eV when it comes to energy differences primarily due to the significantly stronger error cancellation. This distinctly highlights CatBERTa's potential to enhance error cancellation capabilities, thereby contributing to the accurate prediction of energy differences, as compared to GNNs.

CONCLUSIONS

The introduction of the CatBERTa model in this study presents a Transformer-based approach for predicting energy using textual input data. By harnessing the power of a Transformer encoder pretrained on extensive natural language data, this model seamlessly incorporates features into input data in a format that is easily interpretable by humans. CatBERTa serves as a compelling example for LLMs, demonstrating its capacity to comprehend the textual representations of the adsorbate−catalyst system.

The ablation study involving distinct feature-laden input texts illuminates the significance of interacting atoms as descriptors for comprehending adsorption configurations. Conversely, the intentional inclusion of atomic properties and bond distances does not contribute to overall accuracy enhancement. Considering that the interacting atoms within the adsorbate attract increased attention, we can infer that the model effectively captures their crucial role in the adsorption energy prediction. As such, attention score analysis enables the evaluation of emphasized features within the input text, distinguishing CatBERTa from GNNs. Moreover, through the fine-tuning process, the model becomes adept at prioritizing the adsorbate-related aspects as opposed to those associated with the catalyst.

CatBERTa delivers predictive accuracy comparable to that achieved by earlier versions of GNNs. Notably, it demonstrates enhanced efficacy, particularly when trained on limited-size data sets. Furthermore, subsets that exhibit high accuracy with GNNs also yield consistent results through CatBERTa despite its overall energy prediction accuracy slightly trailing that of GNNs across the full data set. This highlights the potential utility of the CatBERTa model in property prediction tasks. The primary limitation in accuracy mainly arises from unseen adsorbate molecules, thereby paving a pathway for future exploration. Additionally, CatBERTa demonstrates a notable ability for error cancellation when subtraction of energies of similar systems, surpassing the magnitude observed in GNNs. This robust error cancellation underscores CatBERTa's potential to enhance the accuracy of the energy difference prediction.

In summary, we establish a foundational framework for predicting catalyst properties based on text. While the focus here is on adsorption energy, the approach can be extended to other properties, such as the HOMO−LUMO gap and stability, related to adsorbate−catalyst systems, given an apt data set. This framework helps researchers estimate the properties of adsorbate−catalyst systems using easily discernible attributes such as adsorbate composition, bulk composition, surface orientation, and adsorption site. This approach enables preliminary assessments despite inherent accuracy limitations. Furthermore, our framework aids in the identification of effective descriptors for predicting catalyst properties through a data-driven methodology. By seamlessly integrating target features into the input text and subsequently training the model, we assess their contributions to improved accuracy, indicative of their capacity to encapsulate essential system characteristics. This methodological approach facilitates a comprehensive evaluation of how effectively these features capture relationships pertinent to property prediction.

METHODS

OC20 Data Set. In this study, we utilized the OC20 data set, which is the most comprehensive and diverse data set available for heterogeneous catalysts.[10] The OC20 data set comprises 1.2 million DFT relaxations, employing the revised Perdew−Burke−Ernzerhof (RPBE) functional.[60] Each relaxation involved approximately 200 single-point calculations, resulting in a data set containing approximately 250 million structures along with their corresponding energies.

The OC20 data set contains training data sets of varying sizes for the IS2RE task: 10k, 100k, and 460k. We focus on utilizing the 10k and 100k data sets for model training. Since the DFT relaxation trajectories for the test data set in the OC20 are not publicly accessible, our evaluation and comparison of model performance are centered on the validation data set. As for the validation set, we randomly selected 10k data points from the openly accessible OC20 validation data set, distributing them evenly across four distinct splits: in-domain, out-of-domain-adsorbate, out-of-domain-catalyst, and out-of-domain-both. Through deliberate consid-

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

eration of these distinct partitions, we intend to establish a balanced and representative validation set that effectively assesses the performance of the CatBERTa model.

Text Generation. The textual string is composed of three components: adsorbate, catalyst, and adsorption configuration. The adsorbate segment solely includes the symbol of the adsorbate, whereas the catalyst component combines the bulk composition of the catalyst and its Miller index. For these segments, we draw from the pre-existing metadata of the OC20 data set. To elucidate the adsorption configuration, we utilized the Pymatgen package to pinpoint both primary and secondary interacting atoms. This process determines atomic connectivities within the adsorbate by relying on a predefined cutoff radius that corresponds to the covalent radius in this study. Additionally, the Pymatgen package aids in determining bond lengths within this connectivity network, pertinent information for string type 5. For string type 3, we integrate various atomic attributes, including atomic weight, polarizability, electron affinity, and electronegativity. These specifics are retrieved by inputting the atomic symbol into the element object within the mendeleev package. Given that adsorption energy hinges on interatomic interactions, we have incorporated electron-related properties alongside fundamental information like atomic number, period, and atomic weight.

Generative language models, capable of producing text based on input queries, can generate feature-rich textual descriptions. For the catalyst bulk structure in description type 1, we employ RoboCrystallographer—a tool specifically designed for generating text-based representations of crystal structures.[46] This process involves feeding the material ID from the Materials Project[61] into the RoboCrystallographer package. For generating text related to the adsorbate in description type 1, as well as both the adsorbate and the catalyst bulk in description type 2, we employ ChatGPT-3.5. The specifications of the API settings are outlined in Table S3.

For description type 1, our goal is to capture the general attributes of the adsorbate molecule. To achieve this, we prompt ChatGPT to elaborate on the chemical attributes of a specific molecule or an atom, highlighting aspects such as bonding type, molecular dimensions, bond angles and lengths, orbital features, and dipole moments. The bonding and molecular size information aids in grasping the adsorbate molecule's structure, while the orbital attributes and dipole moment assist in understanding potential substrate bonding.

Conversely, description type 2 is crafted with the understanding that the adsorption energy is influenced by the coordination number and electronegativity of both the central adsorbate atom and the surface atoms it interacts with. Thus, for the adsorbate segment, ChatGPT is specifically prompted to identify the adsorbate's central atom and its coordination number given the chemical symbol of the adsorbate. For the catalyst component, we instruct ChatGPT to discern the surface type based on its bulk composition and Miller index. Additionally, we inquire about the number of valence electrons and electronegativity.

Transformer Encoder. The Transformer encoder consists of multiple stacked layers, each containing two primary components: a multihead self-attention mechanism and a position-wise feed-forward network. An essential feature of Transformers is their ability to process input tokens in parallel, unlike the sequential processing of RNNs.

The self-attention mechanism assigns varying attention scores to different tokens in the input based on their relevance to the current processing token. It operates using three vector representations: query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$). The attention scores are derived by calculating the dot products of the $\boldsymbol{Q}$ and $\boldsymbol{K}$ vectors, which are then passed through a softmax function. The result is used as a weight for the $\boldsymbol{V}$ vectors, aggregating information from different parts of the sequence. The "multihead" aspect signifies the use of multiple parallel attention layers, allowing the model to capture diverse types of relationships within the data.

Given that Transformers lack an inherent sense of order or position, positional encodings are added to the embeddings at the input layer to provide information about the position of each token in the sequence. These encodings ensure that the model can account for the sequence's order, an essential feature for tasks such as translation or sequence prediction.

Each attention output is passed through a feed-forward network, which is identical across different positions but with different parameters for each layer of the encoder. It acts to transform the attention-derived features and is followed by layer normalization and residual connections, enhancing the model's training stability and convergence speed.

Pretraining and Fine-Tuning. Pretraining, in the context of language models, involves training a model on an extensive corpus before fine-tuning it for a specific downstream task. In this study, we utilize the RoBERTa model,[26] which underwent pretraining on an extensive collection of textual data sourced from the BookCorpus and English Wikipedia data sets, amounting to over 160 GB. Unlike BERT,[25] which masks 15% of tokens in each sequence per epoch, RoBERTa employs dynamic masking, allowing masked tokens to change across epochs. This phase equips the model to predict masked words within sequences, discern syntactic and semantic structures, and absorb vast amounts of general knowledge from the training corpus. Given its established proficiency in interpreting English text, including atomic symbols, we opt to use the pretrained RoBERTa model without introducing any custom pretraining.

For the fine-tuning stage, the pretrained model's weights undergo slight adjustments using a more compact, task-specific data set—specifically, 10k and 100k textual data drawn from OC20 data set in this study. This step refines the model weights for the energy prediction task. In alignment with our regression objective, we substitute RoBERTa's classification head with a custom linear layer designed to produce a singular scalar value as output. The embedding of the first "<s>" token after passing through the encoder is fed into the regression head to generate an energy prediction.

We adopt MAE as our chosen loss function, which is coupled with the AdamW optimizer. We employ a strategy called grouped layer-wise learning rate decay,[62] where the learning rate gradually decreases across groups of layers using a multiplicative decay rate. Group assignment is based on layer depth. Specifically, the four lower layers constitute group 1, with an initial learning rate set to 1×10^{-6}. The subsequent four middle layers form group 2, with their initial learning rate scaled by a factor of 1.75. The uppermost quartet of layers, located near the output, has their initial learning rate increased by a factor of 3.5. This strategic approach acknowledges that different layers assimilate distinct insights from sequences. The layers closer to the output focus on capturing local and specific information, necessitating higher learning rates. In contrast, the lower layers near the input excel in understanding broader and more generalized knowledge. Moreover, we introduce early

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044

stopping mechanisms, which halt training when validation performance shows no improvement over a designated number of epochs, mitigating the risk of overfitting to the training data set.

■ ASSOCIATED CONTENT

Data Availability Statement

Both the Python code and the data employed in this study are available on GitHub at the following link: https://github.com/hoon-ock/CatBERTa.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.3c04956.

Textual data examples, interactions with ChatGPT, OpenAI API settings, reliability of descriptions generated by ChatGPT, hyperparameters for CatBERTa fine-tuning, prediction results for each input case, principle of error propagation, data composition in energy difference calculation, and energy difference prediction results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Amir Barati Farimani − *Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0002-2952-8576; Email: barati@cmu.edu

Authors

Janghoon Ock − *Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0009-0000-0370-4212

Chakradhar Guntuboina − *Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.3c04956

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the Bradford and Diane Smith Fellowship for their pivotal funding support. Additionally, the authors express gratitude to Rishikesh Magar and Yayati Jadhav for enlightening discussions on Transformer-based models and commend Brook Wander for her valuable guidance in analyzing the Open Catalyst 2020 data set.

■ REFERENCES

(1) Sperger, T.; Sanhueza, I. A.; Schoenebeck, F. Computation and Experiment: A Powerful Combination to Understand and Predict Reactivities. *Acc. Chem. Res.* **2016**, *49*, 1311−1319.

(2) Chen, B. W. J.; Xu, L.; Mavrikakis, M. Computational Methods in Heterogeneous Catalysis. *Chem. Rev.* **2021**, *121*, 1007−1048.

(3) Wander, B.; Broderick, K.; Ulissi, Z. W. Catlas: an automated framework for catalyst discovery demonstrated for direct syngas conversion. *Catal. Sci. Technol.* **2022**, *12*, 6256−6267.

(4) Tran, R.; Wang, D.; Kingsbury, R.; Palizhati, A.; Persson, K. A.; Jain, A.; Ulissi, Z. W. Screening of bimetallic electrocatalysts for water purification with machine learning. *J. Chem. Phys.* **2022**, *157*, 074102.

(5) Nguyen, L.; Tao, F. F.; Tang, Y.; Dou, J.; Bao, X.-J. Understanding Catalyst Surfaces during Catalysis through Near Ambient Pressure X-ray Photoelectron Spectroscopy. *Chem. Rev.* **2019**, *119*, 6822−6905.

(6) Pielsticker, L.; Zegkinoglou, I.; Han, Z.-K.; Navarro, J. J.; Kunze, S.; Karslıoğlu, O.; Levchenko, S. V.; Roldan Cuenya, B. Crystallographic Orientation Dependence of Surface Segregation and Alloying on PdCu Catalysts for CO2 Hydrogenation. *J. Phys. Chem. Lett.* **2021**, *12*, 2570−2575.

(7) Lan, J.; Palizhati, A.; Shuaibi, M.; Wood, B. M.; Wander, B.; Das, A.; Uyttendaele, M.; Zitnick, C. L.; Ulissi, Z. W. AdsorbML: Accelerating Adsorption Energy Calculations with Machine Learning. *arXiv* **2022**, arXiv:2211.16486.

(8) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64*, 2311−2323.

(9) Zitnick, C. L.; Chanussot, L.; Das, A.; Goyal, S.; Heras-Domingo, J.; Ho, C.; Hu, W.; Lavril, T.; Palizhati, A.; Riviere, M.; et al. An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage. *arXiv* **2020**, arXiv:2010.09435.

(10) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059−6072.

(11) Musielewicz, J.; Wang, X.; Tian, T.; Ulissi, Z. FINETUNA: fine-tuning accelerated molecular simulations. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 03LT01.

(12) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(13) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv* **2017**, arXiv:1706.08566.

(14) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *arXiv* **2020**, arXiv:2003.03123.

(15) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. *arXiv* **2020**, arXiv:2011.14115.

(16) Gasteiger, J.; Shuaibi, M.; Sriram, A.; Günnemann, S.; Ulissi, Z.; Zitnick, C. L.; Das, A. GemNet-OC: Developing Graph Neural Networks for Large and Diverse Molecular Simulation Datasets. *arXiv* **2022**, arXiv:2204.02782.

(17) Zitnick, C. L.; Das, A.; Kolluru, A.; Lan, J.; Shuaibi, M.; Sriram, A.; Ulissi, Z.; Wood, B. Spherical Channels for Modeling Atomic Interactions. *arXiv* **2022**, arXiv:2206.14331.

(18) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(19) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(20) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminf.* **2015**, *7*, 23.

(21) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification schemes for metal−organic frameworks to enable rapid search and cheminformatics analysis. *Cryst. Growth Des.* **2019**, *19*, 6682−6697.

(22) Christofidellis, D.; Giannone, G.; Born, J.; Winther, O.; Laino, T.; Manica, M. Unifying molecular and textual representations via multi-task language modelling. *arXiv* **2023**, arXiv:2301.12586.

(23) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between molecules and natural language. *arXiv* **2022**, arXiv:2204.11817. arXiv preprint

(24) Rajan, K.; Steinbeck, C.; Zielesny, A. Performance of chemical structure string representations for chemical image recognition using transformers. *Digital Discovery* **2022**, *1*, 84−90.

(25) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
(26) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
(27) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. *Improving Language Understanding by Generative Pre-training*, 2018. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.
(28) Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
(29) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
(30) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I. Attention is All you Need. *Advances in Neural Information Processing Systems*; MIT Press, 2017.
(31) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, arXiv:2010.09885.
(32) Wang, S., Guo, Y., Wang, Y., Sun, H., Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019; pp 429−436.
(33) Jadhav, Y.; Berthel, J.; Hu, C.; Panat, R.; Beuth, J.; Barati Farimani, A. StressD: 2D Stress estimation using denoising diffusion model. *Comput. Methods Appl. Mech. Eng.* **2023**, *416*, 116343.
(34) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. *npj Comput. Mater.* **2023**, *9*, 64.
(35) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal−Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145*, 2958−2967.
(36) Huang, H.; Magar, R.; Xu, C.; Farimani, A. B. Materials Informatics Transformer: A Language Model for Interpretable Materials Properties Prediction. *arXiv* **2023**, arXiv:2308.16259.
(37) Guntuboina, C.; Das, A.; Mollaei, P.; Kim, S.; Farimani, A. B. PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction. *J. Phys. Chem. Lett.* **2023**, *14*, 10427.
(38) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nature Chem.* **2009**, *1*, 37−46.
(39) Yang, B.; Burch, R.; Hardacre, C.; Headdock, G.; Hu, P. Understanding the Optimal Adsorption Energies for Catalyst Screening in Heterogeneous Catalysis. *ACS Catal.* **2014**, *4*, 182−186.
(40) Wang, S.; Temel, B.; Shen, J.; Jones, G.; Grabow, L. C.; Studt, F.; Bligaard, T.; Abild-Pedersen, F.; Christensen, C. H.; Nørskov, J. K. Universal Brønsted-Evans-Polanyi Relations for C−C, C−O, C−N, N−O, N−N, and O−O Dissociation Reactions. *Catal. Lett.* **2011**, *141*, 370−373.
(41) Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2*, 1624−1634.
(42) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.
(43) Open Catalyst Project Leaderboard. https://opencatalystproject.org/leaderboard.html (accessed Oct 16, 2023).
(44) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(45) OpenAI. *ChatGPT 3.5*, Version 3.5, 2022. https://openai.com/chatgpt
(46) Ganose, A. M.; Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Commun.* **2019**, *9*, 874−881.
(47) Gao, W.; Chen, Y.; Li, B.; Liu, S.-P.; Liu, X.; Jiang, Q. Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates. *Nat. Commun.* **2020**, *11*, 1196.
(48) Ala Alam Falaki, R. G. Attention Visualizer Package: Revealing Word Importance for Deeper Insight into Encoder-Only Transformer Models. *arXiv* **2023**, arXiv:2308.14850.
(49) Hao, Y.; Dong, L.; Wei, F.; Xu, K. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 12963−12971.
(50) Liao, Y.-L.; Smidt, T. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. *arXiv* **2023**, arXiv:2206.11990.
(51) Yuan, Z.; Zhang, Y.; Tan, C.; Wang, W.; Huang, F.; Huang, S. Molecular Geometry-aware Transformer for accurate 3D Atomic System modeling. *arXiv* **2023**, arXiv:2302.00855.
(52) Li, Y.; Li, Q.; Wang, H.; Zhang, L.; Wilkinson, D.; Zhang, J. Recent Progresses in Oxygen Reduction Reaction Electrocatalysts for Electrochemical Energy Applications. *Electrochem. Energy Rev.* **2019**, *2*, 518−538.
(53) Ooka, H.; Huang, J.; Exner, K. S. The sabatier principle in electrocatalysis: Basics, limitations, and extensions. *Front. Energy Res.* **2021**, *9*, 654460.
(54) Cheng, J.; Hu, P.; Ellis, P.; French, S.; Kelly, G.; Lok, C. M. Brønsted−Evans−Polanyi Relation of Multistep Reactions and Volcano Curve in Heterogeneous Catalysis. *J. Phys. Chem. C* **2008**, *112*, 1308−1311.
(55) Liang, S.; Huang, L.; Gao, Y.; Wang, Q.; Liu, B. Electrochemical Reduction of CO2 to CO over Transition Metal/N-Doped Carbon Catalysts: The Active Sites and Reaction Mechanism. *Advanced Science* **2021**, *8*, 2102886.
(56) Ock, J.; Tian, T.; Kitchin, J.; Ulissi, Z. Beyond independent error assumptions in large GNN atomistic models. *J. Chem. Phys.* **2023**, *158*, 214702.
(57) Collins, E. M.; Raghavachari, K. Effective Molecular Descriptors for Chemical Accuracy at DFT Cost: Fragmentation, Error-Cancellation, and Machine Learning. *J. Chem. Theory Comput.* **2020**, *16*, 4938−4950.
(58) Plessow, P. N.; Studt, F. How Accurately Do Approximate Density Functionals Predict Trends in Acidic Zeolite Catalysis? *J. Phys. Chem. Lett.* **2020**, *11*, 4305−4310.
(59) Hautier, G.; Ong, S. P.; Jain, A.; Moore, C. J.; Ceder, G. Accuracy of Density Functional Theory in Predicting Formation Energies of Ternary Oxides from Binary Oxides and its Implication on Phase Stability. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2012**, *85*, 155208.
(60) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved Adsorption Energetics within Density-Functional Theory using Revised Perdew-Burke-Ernzerhof Functionals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 7413−7421.
(61) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(62) Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K. Q.; Artzi, Y. Revisiting Few-sample BERT Fine-tuning. *arXiv* **2021**, arXiv:2006.05987.

https://doi.org/10.1021/acscatal.3c04956
ACS Catal. 2023, 13, 16032−16044