nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00930-7

Multimodal language and graph learning of adsorption configuration in catalysis

Received: 14 January 2024

Accepted: 21 October 2024

Published online: 27 November 2024

Check for updates

Janghoon Ock[1], Srivathsan Badrinarayanan[1], Rishikesh Magar[2], Akshay Antony[2] & Amir Barati Farimani[2]

Adsorption energy is a reactivity descriptor that must be accurately predicted for effective machine learning application in catalyst screening. This process involves finding the lowest energy among different adsorption configurations on a catalytic surface, which often have very similar energies. Although graph neural networks have shown great success in computing the energy of catalyst systems, they rely heavily on atomic spatial coordinates. By contrast, transformer-based language models can directly use human-readable text inputs, potentially bypassing the need for detailed atomic positions or topology; however, these language models often struggle with accurately predicting the energy of adsorption configurations. Our study improves the predictive language model by aligning its latent space with well-established graph neural networks through a self-supervised process called graph-assisted pretraining. This method reduces the mean absolute error of energy prediction for adsorption configurations by 7.4–9.8%, redirecting the model's attention towards adsorption configuration. Building on this, we propose using generative large language models to create text inputs for the predictive model without relying on exact atomic positions. This demonstrates a potential use case of language models in energy prediction without detailed geometric information.

Machine learning approaches, particularly graph neural networks (GNNs), have emerged as efficient surrogates to computationally expensive density functional theory (DFT) simulations[1–4]. These advancements can accelerate energy and force predictions for high-throughput materials screening[5–9]. The successful application of machine-learning-based DFT surrogate modelling in catalysis can enable the identification of optimal catalyst materials for specific reactions, which is crucial for advancing energy storage technologies and sustainable chemical processes. The importance of such techniques has drawn attention beyond the chemical engineering and chemistry communities, extending into the AI for science field[10].

Despite the considerable success of GNNs in machine learning applications in the catalysis domain, obtaining their input data can be challenging since they require atomic positions or topology. Constructing graph representations of structures relies on identifying nearest neighbours within specific proximity thresholds for each atom[11–14]. However, achieving such precise coordinates can be difficult, limiting the applicability of GNNs primarily to theoretical studies. For instance, even with experimentally validated adsorption energy data from the literature, using this information in modelling remains difficult because replicating the exact atomic positions of the adsorbate–catalyst systems from experiments is problematic[15,16].

Recent advancements in language model applications offer a promising alternative to relying on exact atomic coordinates as input data[17–20]. Language models can process human-readable text descriptions of atomic systems instead of building an input with atomic coordinates. For example, the MOFormer model encodes metal–organic frameworks (MOFs) as text string representations, called MOFid, which, unlike graph representations[17], include chemical information on building blocks and topology codes. The TransPolymer model encodes

[1]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. e-mail: barati@cmu.edu

polymers using the SMILES strings of their repeating units along with attributes such as the degree of polymerization, polydispersity and chain conformation[19]. Furthermore, the ability to process textual input allows us to incorporate experimentally obtainable attributes into the input data. We aim to extend these successes from the materials science domain to the catalysis domain. For instance, the CatBERTa model takes textual input for adsorbate–catalyst systems to predict the energy of the system[20].

Identification of adsorption energy is an important task in catalysis because it is a key reactivity descriptor in catalyst screening[3,21,22]. A single adsorbate–catalyst pair can have numerous adsorption configurations, which vary by adsorption site and molecule orientation on the catalytic surface[8,23]. The minimum energy among these configurations is considered the adsorption energy. Due to the subtle differences between these configurations, their energies can be very similar. Therefore, to accurately identify the adsorption energy, the model must be capable of distinguishing these subtle energy differences, which can range from 0.1 to 0.3 eV around the minimum energy[8]. Although the language models offers the potential to bypass the need for exact atomic positions, which are critical for building graph representations in many models used for adsorbate–catalyst systems, its accuracy remains a concern. Improving the model's accuracy is essential to effectively apply this text-based approach to adsorption configuration energy prediction tasks.

To address this challenge, our study introduces graph-assisted pretraining—a multimodal learning method that leverages graph modality to improve the prediction accuracy of the language model for adsorption configurations. Multimodal learning has already been successfully applied to the materials and chemical sciences to boost model performance[17,24,25]. We aim to extend this success to the catalysis domain, particularly to enhance the predictive capability of transformer-based language models in adsorption energy prediction. Graph-assisted pretraining transfers the structural knowledge captured in graph embeddings to text embeddings in a self-supervised manner. This transfer of knowledge from a learned embedding space to the language model will help in our application of adsorption configuration energy predictions.

We also aim to show a potential use case of the large language model (LLM) in making predictions without relying on precise atomic positions. For this, we use the generative capabilities of LLMs to generate desired textual input data for our predictive language model for energy prediction. Language models' generative capabilities have recently shown success in structure generation for inorganic crystals[26,27]. In this study we specifically fine-tune CrystaLLM to generate crystallographic information files (CIFs) for relevant adsorbate–catalyst systems instead of inorganic crystals. Here, the fine-tuned CrystaLLM takes textual information about the chemical composition of the system, along with its surface orientation. We then use the generated CIFs to derive the input string for our predictive language model. This method allows us to make energy predictions without knowing the full structure of the adsorbate–catalyst configurations.

Results and discussion

Framework

The language-model-based approach for catalyst energy prediction leverages textual data for both training and inference. We have developed a multimodal pretraining framework, called graph-assisted pretraining, to bridge the established graph-based approach with a text-based approach within a shared latent space (Fig. 1). This method is introduced to enhance the accuracy of adsorption configuration energy predictions. This framework uses the CatBERTa model, which uses the RoBERTa encoder for text processing and a linear regression header to predict catalyst system energies (Fig. 1b)[20]. Furthermore, the EquiformerV2 model is employed as a graph encoder for its capability to encode precise atomic structures (Fig. 1c)[10,28]. In this framework, both text and graph embeddings are aligned in a self-supervised manner during pretraining. The model then undergoes a fine-tuning stage, in which it is trained in a supervised manner using energy labels derived from DFT calculations. Importantly, the fine-tuning step relies exclusively on text input data, without the need for graph representations (refer to the Methods for more details).

We conduct two types of downstream inference: one to assess the effect of graph-assisted pretraining and the other to demonstrate the model's capability to predict energy without precise knowledge of the adsorbate–catalyst structures. Both are depicted in Fig. 2. First, we made predictions on the test set strings derived from the machine-learning-relaxed structure to evaluate the impact of graph-assisted pretraining on prediction accuracy. The CatBERTa model, which takes textual strings as input, is trained using textual data derived from machine-learning-relaxed structures to predict the energy of a relaxed configuration. Second, we generate indicative structures in CIF format using a generative language model to illustrate the model's potential in predicting energies without relying on exact structures. These were attained by providing the chemical composition and surface orientation of the adsorbate and catalyst as input. The generated CIFs are converted into textual strings that are compatible with CatBERTa model inputs.

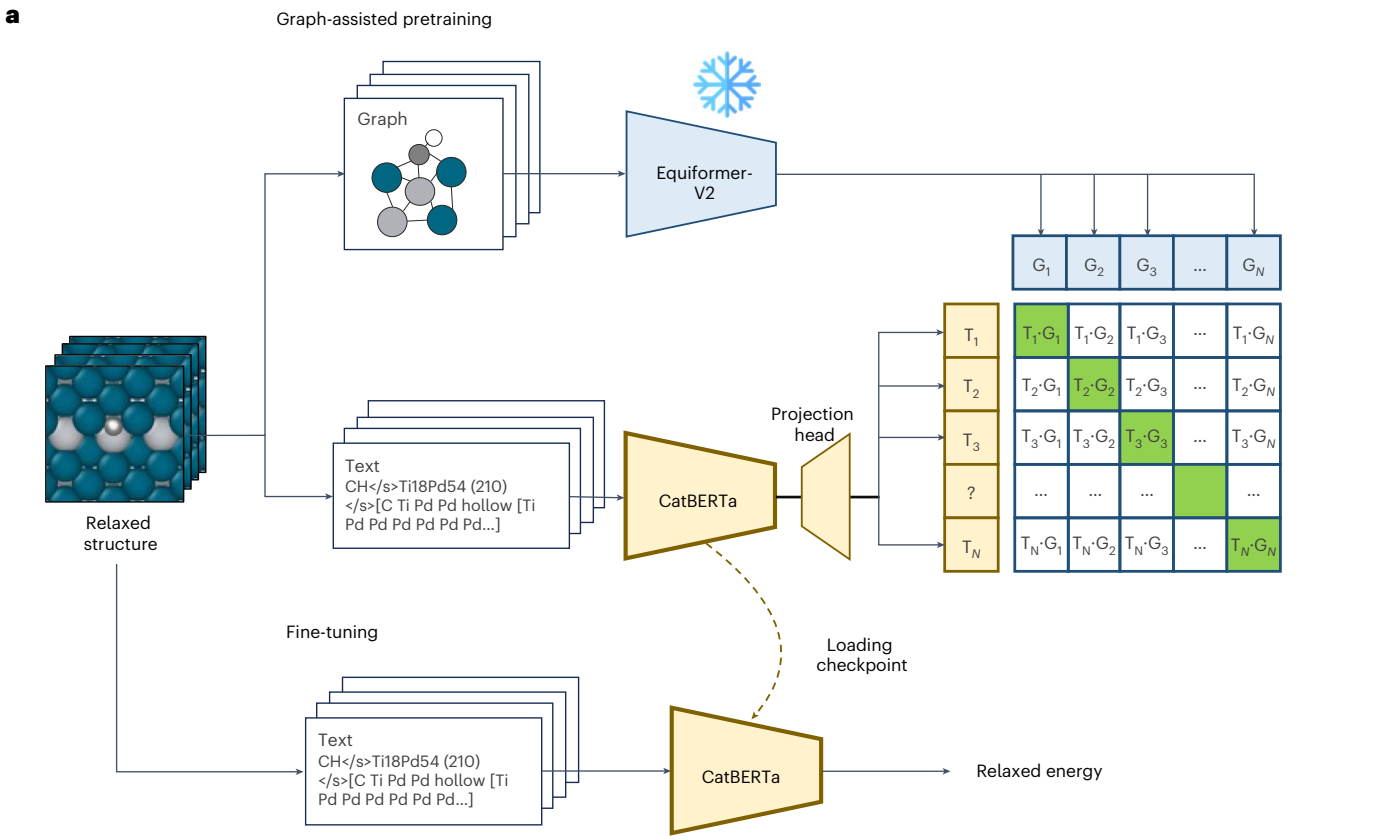
The textual strings are generated by converting structural information into a specific format containing three sections, as illustrated in the dashed box at the bottom right of Fig. 2. The first section represents the adsorbate's chemical symbol; the second section includes the catalyst's chemical symbols and Miller index, indicating its chemical composition and surface orientation, respectively; the final section describes the adsorption configuration, capturing the primary and secondary interacting atoms in the adsorbate and the top layers of the catalyst surface, identified using the Pymatgen library[29]. Refer to the Methods for further details on this structure-to-text conversion process.

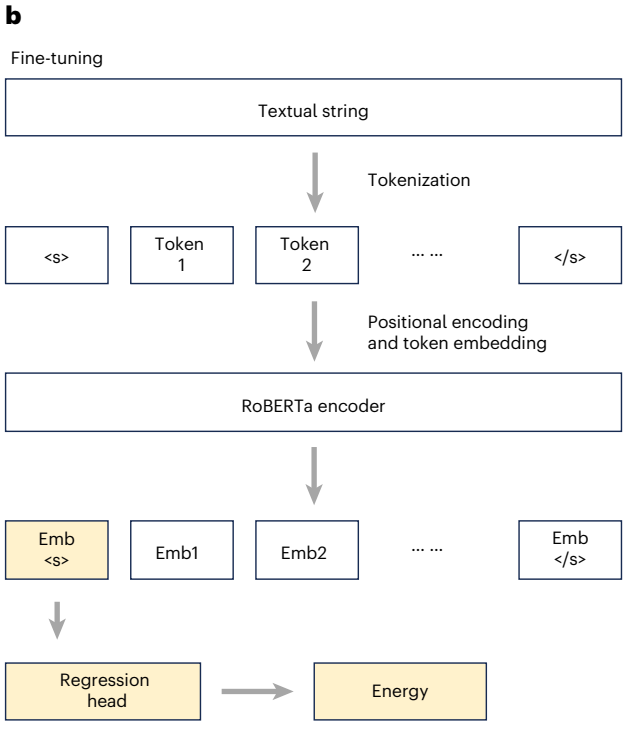
Data pipeline

The textual string input for CatBERTa training is derived from the relaxed structures in the Open Catalyst 2020 (OC20) and Open Catalyst 2020 Dense (OC20-Dense) datasets. For both CatBERTa-involved training and CrystaLLM fine-tuning, training and validation are conducted using texts sourced from DFT-relaxed structures. Specifically, for the first case, we convert the relaxed structures to string representations and use them for the training and validation process. For the latter case, we create CIFs for the relaxed structures and then use them to fine-tune CrystaLLM.

In the first case of graph-assisted pretraining evaluation, predictions are made on strings generated from machine-learning-relaxed structures. These machine-learning-relaxed structures, along with their DFT-calculated energy labels, are provided by the Open Catalyst Project Challenge 2023[10]. GNNs, such as GemNet-OC, SCN and eSCN, are used for the machine learning relaxation process. The machine learning relaxations are conducted on out-of-domain splits in the OC20-Dense dataset, yielding 11,508, 11,630 and 11,755 relaxed structures from the respective models. DFT single-point calculations are performed on machine-learning-relaxed structures to obtain valid DFT energies. These were prepared by the dataset provider and included in the publicly available dataset. Our model's accuracy is then evaluated using approximately 920 of these machine-learning-relaxed structures with valid DFT energies. We quantify the uncertainty of our model's predictions by calculating the standard deviation across predictions for structures relaxed using GemNet-OC, SCN and eSCN. Their individual results are listed in Supplementary Table 5. For embedding and attention score analysis, we use the entire set of machine-learning-relaxed structures—ranging from 11,508 to 11,755—regardless of whether these structures have verified DFT energies. See Supplementary Section 2 for details on the data split.

https://doi.org/10.1038/s42256-024-00930-7





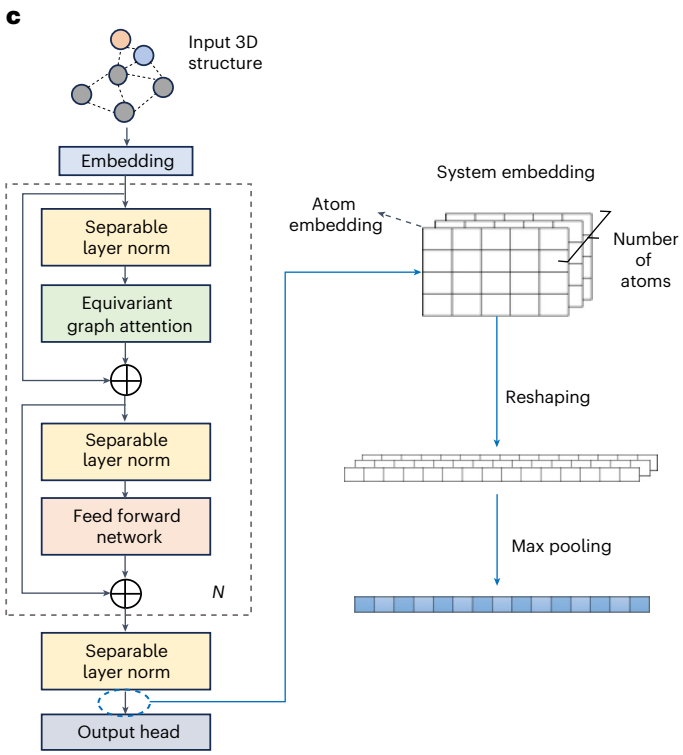


Fig. 1 | Overview of the model training framework. a, The training process consists of two steps: graph-assisted pretraining and energy prediction fine-tuning. **b**, The CatBERTa model is used as the text encoder. **c**, The EquiformerV2 model serves as the graph encoder, and the graph embedding from the final layer is converted to a 1D format by reshaping and max pooling the collection of atom embeddings. Panels adapted from: **b**, ref. 20 under a Creative Commons licence CC BY 4.0; **c**, ref. 28 under a Creative Commons licence CC BY 4.0.

https://doi.org/10.1038/s42256-024-00930-7

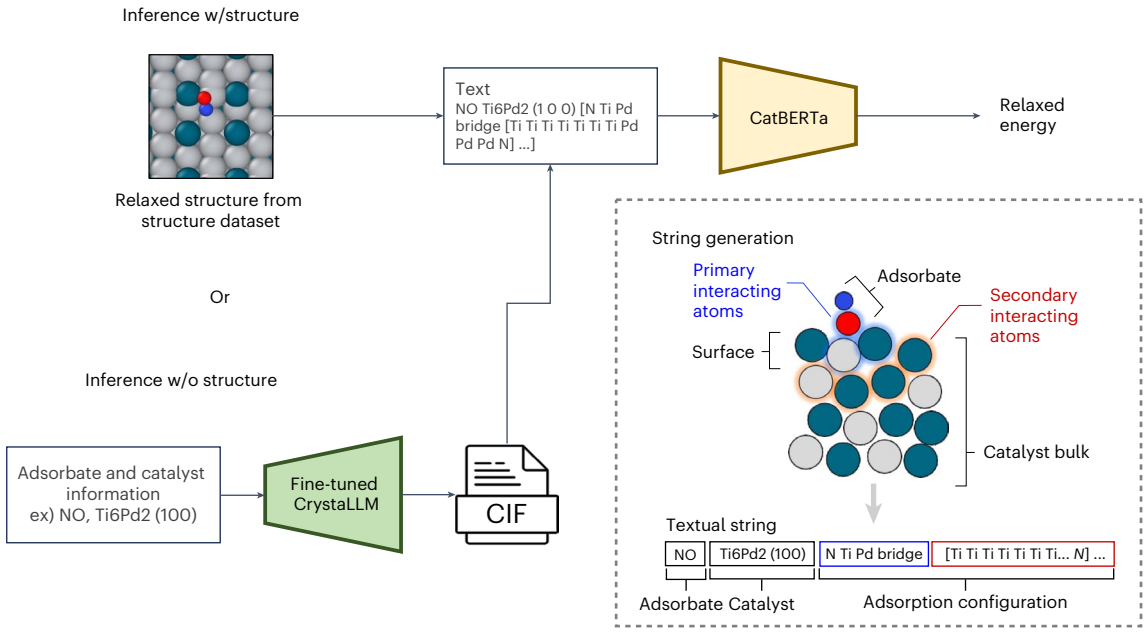


Fig. 2 | Model inference framework. Both structure data from the Open Catalyst datasets and CIFs generated by fine-tuned CrystaLLM can be converted into textual strings compatible with CatBERTa input, following the string conversion logic shown in the dashed box. Generated CIFs provide structure information, including atomic positions, types and unit cell details.

Table 1 | Performance comparison of CatBERTa with and without graph-assisted pretraining

	GAP data (size)	Fine-tuning data (size)	Prediction results		Improvement from GAP	
			MAE (eV) (↓)	R^2 (↑)	MAE (%) (↓)	R^2 (%) (↑)
CatBERTa	–	OC20 (460,000)	0.713±0.014	0.584±0.014	–	–
	–	OC20-Dense (16,000)	0.542±0.011	0.712±0.008	–	–
	–	Combined (476,000)	0.378±0.005	0.863±0.005	–	–
GAP-CatBERTa	OC20 (460,000)	OC20 (460,000)	0.643±0.020	0.691±0.015	−9.82	+18.32
	OC20 (460,000)	OC20-Dense (16,000)	0.502±0.010	0.764±0.008	−7.38	+7.30
	Combined (476,000)	Combined (476,000)	0.346±0.005	0.882±0.002	−8.47	+2.20

'Combined' refers to a combination of OC20 and OC20-Dense datasets. GAP, graph-assisted pretraining.

For inference on the LLM-derived strings, predictions are made on the basis of strings derived from the adsorbate and catalyst, as well as the surface orientation. The aim is to demonstrate the potential of generating plausible textual string representations using the LLM framework. A subset of adsorbate and catalyst pair information is chosen from the original OC20-Dense training set, which contains 235 unique adsorbate–catalyst pairs. The downselection process is detailed in the Methods. We extract only the adsorbate, catalyst and Miller index information from these pairs, and use them as initial prompts for the fine-tuned CrystaLLM framework.

Graph-assisted pretraining

Graph-assisted pretraining—a core component of our framework—is designed to transfer knowledge from graph embeddings to text embeddings. This approach bridges the gap between GNNs, which show great performance in energy and force predictions, and language models, which process human-interpretable text but do not take the entire structure as input. We select EquiformerV2 as the graph encoder for its excellent performance with the OC20 dataset[28,30]. The CatBERTa model serves as the text encoder, producing text embeddings that are then projected to match the dimensions of the graph embeddings. To align these embeddings, we apply a contrastive loss to increase the similarity between embeddings from the same adsorbate–catalyst configurations, with the graph encoder remaining frozen.

Applying this method to the embedding space offers utility and flexibility. It operates solely on embeddings without downstream-task-specific labels, such as regression labels or classification categories. This means that we do not need to obtain labels for this pretraining stage. The downstream fine-tuning process can remain text-only, like the standard CatBERTa method. Once a properly pretrained checkpoint that bridges graph and text modalities is established, it can be applied to multiple downstream tasks. Using the embedding space also enhances generalizability, allowing the method to be applied to various encoders, provided their embedding sizes match. This graph-assisted pretraining method considerably improves prediction accuracy and adaptability across different datasets and tasks.

The graph-assisted pretraining method results in a substantial reduction in mean absolute error (MAE), as shown in Table 1, with reductions ranging from ~7.4% to ~9.8%. To evaluate the enhancement from graph-assisted pretraining, we compare the prediction results of CatBERTa with and without this pretraining method. In all cases, graph-assisted pretraining improves downstream prediction accuracy.

Notably, pretraining with OC20 also benefits fine-tuning only with OC20-Dense, despite there being no overlap between these datasets. This indicates that graph-assisted pretraining can serve as a transferable pretraining strategy, bridging the gap between high-performing

https://doi.org/10.1038/s42256-024-00930-7

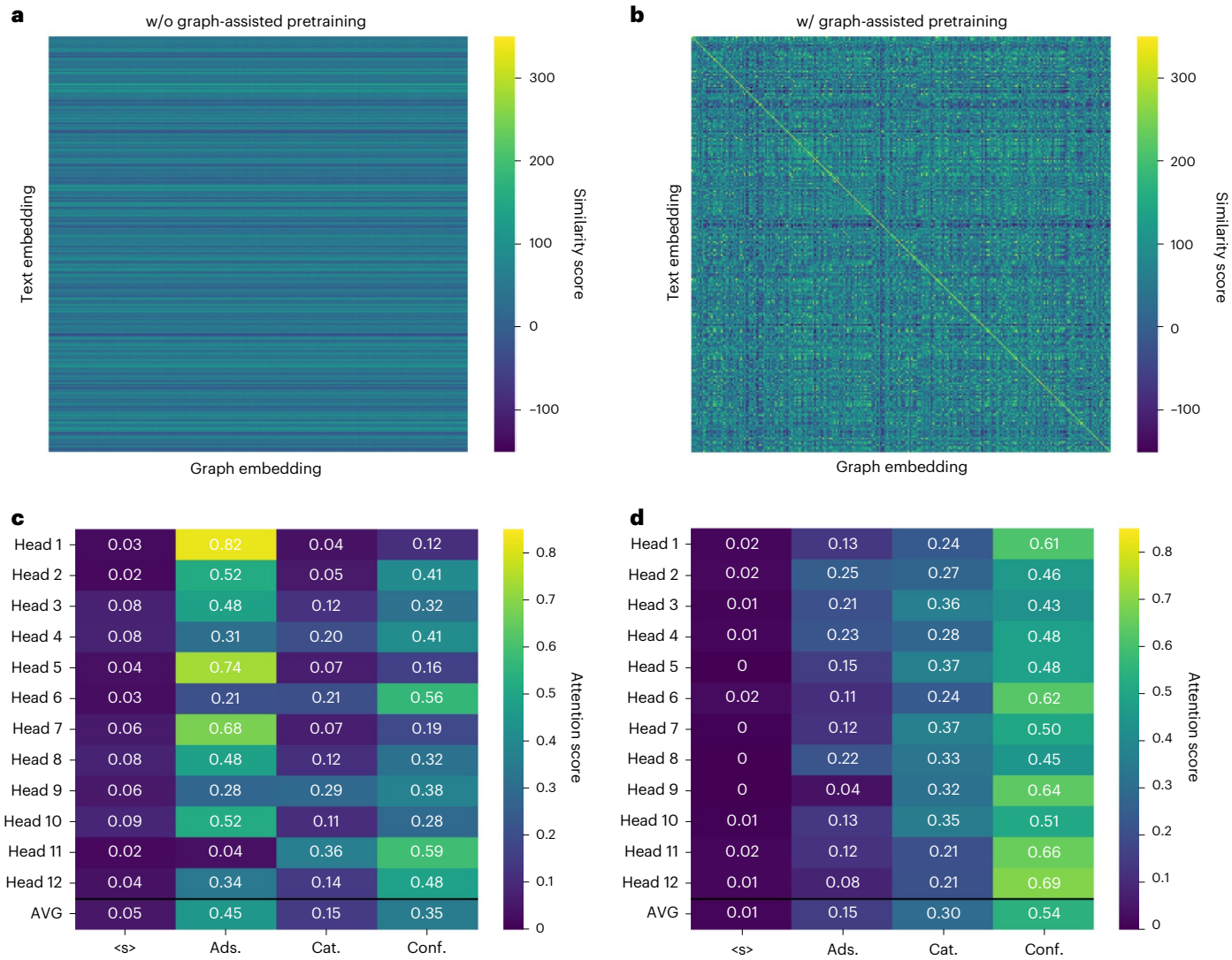


Fig. 3 | Analysis of similarity scores and sectional attention with and without graph-assisted pretraining. a,**b**, Similarity score analysis without (**a**) and with (**b**) graph-assisted pretraining. **c**,**d**, Sectional attention score comparisons without (**c**) and with (**d**) graph-assisted pretraining. These results are derived from model predictions, which were trained on the OC20 dataset and evaluated using text strings from the GemNet-OC-relaxed structures.

GNNs and emerging transformer-based language model approaches. It demonstrates the potential of self-supervised pretraining on one dataset to enhance performance on downstream tasks involving a different dataset. Prediction visualization and further analysis are provided in Supplementary Figs. 2 and 3. Although the substantial accuracy improvement after incorporating OC20-Dense is attributed to the expansion of in-domain systems, what stands out is that a comparable level of improvement is still observed after applying graph-assisted pretraining.

Enhancement in the latent space and attention score

As graph-assisted pretraining is applied to the embeddings, it is essential to examine the latent space to observe its effects. Graph-assisted pretraining can align the graph and text embeddings from the same adsorbate–catalyst configurations. Figure 3a,b shows the similarity matrix of graph and text embeddings. After applying graph-assisted pretraining, a clear diagonal line appears in the similarity matrix, indicating the alignment between embeddings in the latent space.

Figure 3a,b shows the similarity matrices of graph and text embeddings. A clear diagonal line appears after applying graph-assisted pretraining, indicating the alignment between embeddings in the latent space. By comparing the left and right panels, we can clearly observe that the similarity score of actual pairs becomes higher than that of random pairs, which are not supposed to have correlations. The horizontal stripes in the similarity matrix before applying graph-assisted pretraining are due to the weight initialization in the final layer of EquiformerV2. Clustering in the latent space is shown in Supplementary Fig. 1.

The analysis of attention scores in the final layer provides insights into how the model allocates attention. Our input string consists of three sections, as discussed earlier, with the '<s>' token at the beginning as a default setting of the tokenizer. The section-wise attention scores reveal the model's focus on each section. We extract and average the attention scores for the '<s>' token, which is fed to the regression head, across three distinct sections: the adsorbate, the catalyst, and the adsorption configuration. We also compute the attention score of the '<s>' token with respect to itself. The section-wise-averaged attention scores are presented in Fig. 3c,d.

Graph-assisted pretraining redirects the model's attention towards the adsorption configuration section. This attention redirection occurs across all twelve attention heads. Specifically, although the vanilla CatBERTa model primarily concentrates on the adsorbate section, the graph-assisted pretraining reallocates the model's focus towards both the catalyst and configuration sections. This shift in attention aligns with the general physical understanding that the interaction of

https://doi.org/10.1038/s42256-024-00930-7

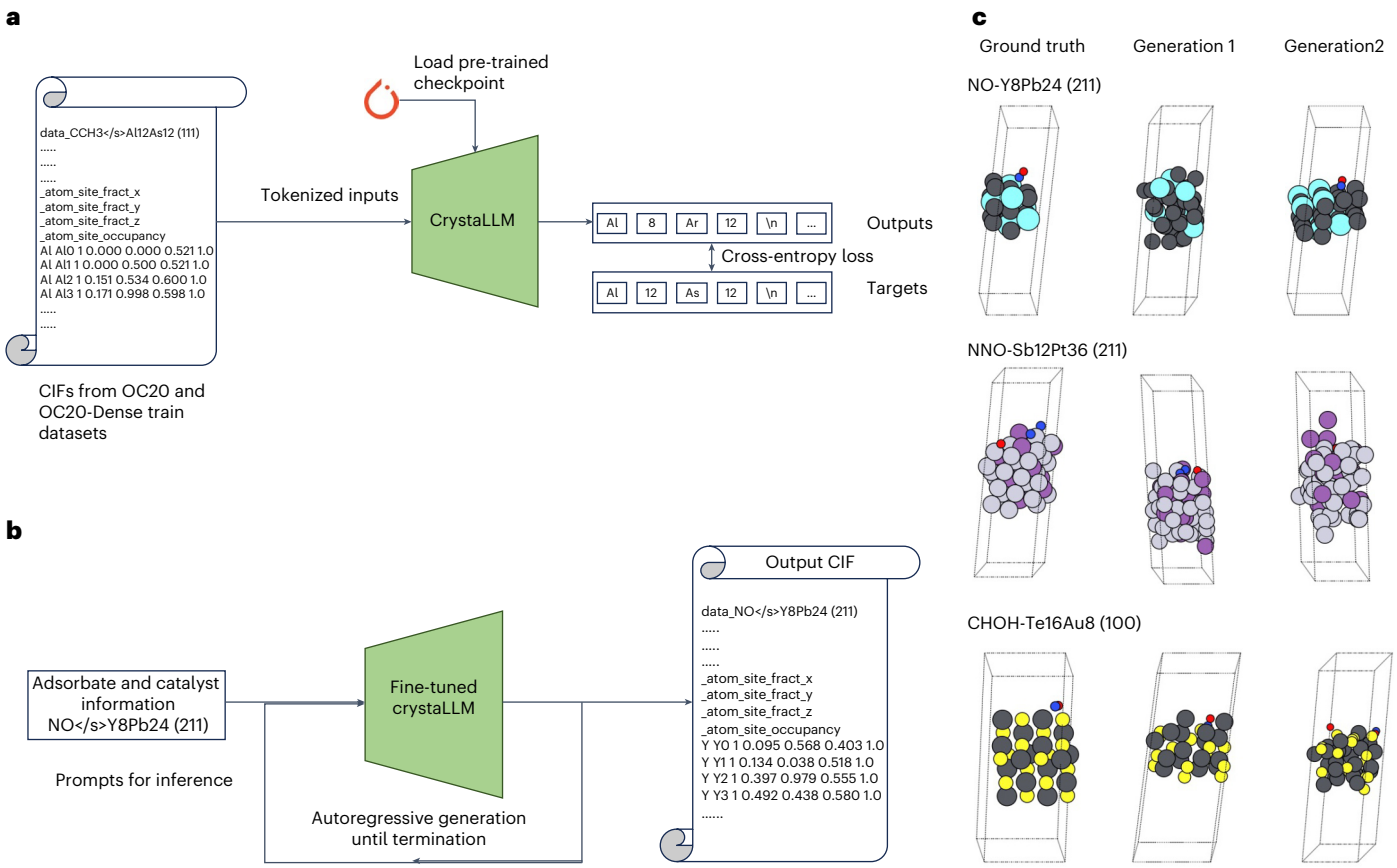


Fig. 4 | CrystaLLM framework. a, The fine-tuning step using the CIFs from the relaxed structures in the OC20 and OC20-Dense training datasets. **b**, The inference process using the provided adsorbate and catalyst pair information. **c**, Visualization examples. These ground-truth systems are sourced from the OC20 validation set, matching composition and surface orientation.

the adsorbate with the catalytic surface as a whole is more important than focusing on the adsorbate and surface individually[31,32].

Energy prediction without coordinate dependence

One benefit of using language models and language representations is to bypass the need for explicit atomic coordinates. The CatBERTa model is capable of processing text descriptions with or without atomic position information; however, for optimal performance, the model still relies on incorporating neighbour atom information within the textual input, as demonstrated in previous predictions. To address this, we explore using an additional LLM to generate the necessary input data solely on the basis of adsorbate and catalyst information. The CatBERTa input string contains three sections, as illustrated in Fig. 2. The main idea is to derive the last configuration section from the first two sections, which pertain to the adsorbate and catalyst details.

For this purpose, we use CrystaLLM, which was originally trained to generate CIFs of inorganic crystals[26]. We fine-tune the pretrained CrystaLLM using CIFs from relaxed structures in the Open Catalyst training datasets (Fig. 4). Please refer to the Methods and Supplementary Table 2 for details on the process and hyperparameters. CrystaLLM autoregressively predicts the next tokens in the CIF until it encounters two consecutive '\n' tokens. The initial prompt to the CrystaLLM is set as the first two parts of the CatBERTa input string, which includes adsorbate and catalyst information, along with the Miller index. These first two parts of the CatBERTa input string are derived from the metadata of the Open Catalyst dataset[3,8], encompassing the adsorbate chemical symbol, catalyst chemical symbol and Miller index—information which relies on atomic geometry and is experimentally obtainable.

As the model autoregressively generates the next token on the basis of the given tokens, the model completes the rest of the CIF from the given starting prompt. For example, the input for the fine-tuned CrystaLLM might be 'data_CCH3 </s> Al12As12(111)', and the output from the model would be the corresponding CIF, which contains indicative structural information. 'Indicative' means that, although the generated CIFs do not necessarily guarantee atomistic validity, they still provide some information on neighbouring atoms, which can assist in improving CatBERTa's predictions (see Fig. 4c). The generated CIFs cannot be used as input for GNNs because they lack precise atomic coordinates.

For unknown structure systems, we can use the proposed generative language model approach along with the predictive CatBERTa model to obtain energy predictions. From the OC20-Dense training dataset, which contains 235 unique adsorbate–catalyst pairs, we downsampled 66 pairs on the basis of the type and number of elements in the adsorbate and catalyst bulk. These selected pairs of adsorbate and catalyst information are used as starting prompts, which are fed into the fine-tuned CrystaLLM to generate CIFs (see Fig. 4b). We iterated through three generations and selected the CIFs whose composition of generated atoms matched the given adsorbate and catalyst chemical symbols within a certain threshold (see Methods for details on the process). These CIFs are then converted into textual string inputs for CatBERTa prediction. Each adsorbate–catalyst pair can have numerous adsorption configurations in the OC20-Dense dataset. The total number of adsorption configurations is 5,141 for the 66 downsampled adsorbate–catalyst pairs.

We benchmark the energy predictions using these adsorbate–catalyst pairs to determine whether the LLM-derived configuration strings can assist with the energy prediction. For this, we compared

https://doi.org/10.1038/s42256-024-00930-7

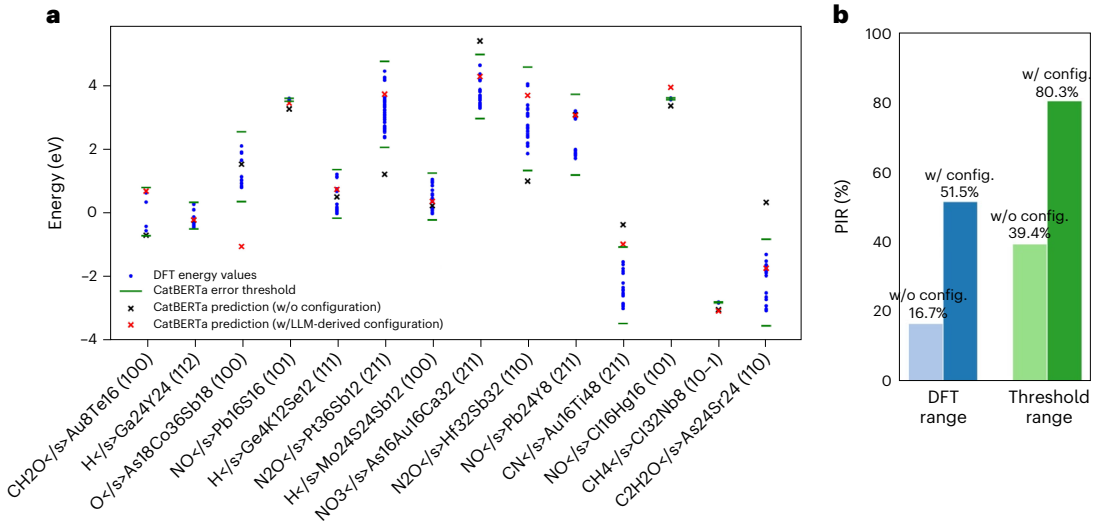


Fig. 5 | Enhancement from LLM-derived strings as input for the CatBERTa model. a, Fourteen example adsorbate–catalyst pairs are sampled from the 66 pairs. Blue dots represent the DFT energy of different adsorption configurations for each adsorbate–catalyst pair. The number of adsorption configurations for the fourteen example pairs ranges from 4 to 130, with a mean value of 62.5. **b**, Prediction inclusion ratio (PIR) for each case quantifies the improvement in prediction accuracy across 66 pairs. The term 'config.' refers to the LLM-derived configurations strings.

two types of predictions: one is made only with the adsorbate and catalyst section, whereas the other is made on the strings including LLM-derived configuration strings. We then compare those predicted values with the actual DFT energy values of possible adsorption configurations for those chosen adsorbate–catalyst pairs in the OC20-Dense dataset (Fig. 5a). For this prediction, we used CatBERTa—pretrained and fine-tuned with the OC20 dataset—making these predictions entirely out-of-domain because the adsorbate–catalyst pairs are from the OC20-Dense dataset.

This comparison allows us to evaluate whether the LLM-derived strings can help predictions by bringing them within the plausible energy range across possible configurations. To quantify this, we define the prediction inclusion ratio as:

$$\mathrm{PIR}\,(\%) = \frac{N_{\text{in-range}}}{N_{\text{total}}} \tag{1}$$

where $N_{\text{in-range}}$ represents the number of predictions falling within the desired range, and N_{total} is the total number of predictions.

In Fig. 5a the blue points represent the DFT-calculated energy values for various adsorption configurations of each adsorbate–catalyst pair. The red crosses represent CatBERTa predictions with LLM-derived configurations, whereas the black crosses represent predictions without them. The range of blue points indicates the potential variations in DFT energy of adsorption configurations for each pair. The green lines above and below the blue points indicate CatBERTa's intrinsic error threshold, with an average value of 0.24 eV for the exemplary fourteen systems. This threshold for each adsorbate–catalyst combination is derived from the standard deviation of CatBERTa's predictions on the actual strings of those adsorption configurations, corresponding to the blue points. As the CatBERTa prediction itself has intrinsic error from the DFT values, we add this value to the minimum and subtract it from the maximum values of the blue points, respectively, to establish the error threshold.

Figure 5b demonstrates that incorporating LLM-derived configuration strings considerably increases the likelihood of CatBERTa's predictions falling within the DFT energy and threshold ranges. After incorporating the LLM-derived configuration strings, the likelihood of CatBERTa's predictions being within the target range more than doubles for both cases. This suggests that, despite the generated structures not being highly accurate, the information extracted from them can still benefit downstream prediction tasks.

Conclusion

We introduced a multimodal pretraining approach that integrates graph and text embeddings within the latent space. This approach facilitates connections between different model set-ups, enhancing the application of language models in prediction tasks. In the field of catalysis, the predictive language model CatBERTa can handle textual data with various features. Our graph-assisted pretraining method improves the accuracy of the language model by guiding the text modality using the graph modality. We also leveraged the autoregressive generative capabilities of the language model to predict energies without requiring precise atomic structures. By using a transformer-based language model to generate input strings, we create textual representations for the CatBERTa model based solely on chemical composition and surface orientation. This allows us to make energy predictions for adsorbate–catalyst systems using experimentally obtainable information. Although the current framework has limitations in prediction accuracy and generation validity, it provides starting points for more detailed simulations or experimental validation.

Recent advancements in LLMs have hugely impacted the chemical and materials sciences, especially in areas such as autoregressive generation, data retrieval and autonomous scientific discovery[33–36]. Our study explores the potential of LLMs for both generative and predictive modelling. Moving forward, we aim to develop a more comprehensive, language-based platform for catalyst design by improving both predictive and generative capabilities, integrating them into a single LLM, incorporating additional functional tools, and equipping the platform with reasoning and planning capabilities in an agent-like framework.

Methods

Open Catalyst dataset

The OC20 dataset stands as the most extensive and varied dataset for heterogeneous catalysts—it encompasses over 1.2 million DFT relaxations, all of which use the revised Perdew–Burke–Emzerhof functional[3,37]. This dataset features various tasks, including Initial-Structure-to-Energy (IS2RE), Initial-Structure-to-Relaxed-Structure (IS2RS) and Structure-to-Energy-and-Force (S2EF). In this work we focus on the data for the IS2RE and IS2RS tasks, which comprise 460,328 DFT relaxations. Our

objective is to predict the relaxed energy of each adsorbate–catalyst configuration on the basis of its final relaxed structure, leading us to specifically select the last frame of these relaxation trajectories.

The OC20-Dense dataset was developed[8] to investigate the global minimum energy (also known as the adsorption energy) of adsorbate–catalyst pairs. The dataset—although extensive in types of adsorbates and catalytic surfaces—lacks variation in adsorption configurations. It addresses this by densely enumerating these configurations. The initial configurations of adsorbates on surfaces are produced using both heuristic and random approaches[8,29,38]. These configurations then undergo relaxations using both machine learning and DFT methods. The OC20-Dense dataset contains 995 distinct adsorbate–catalyst pairs (evenly selected from the in-domain and three out-of-domain splits) from the OC20 validation set (see Supplementary Section 2). As a result, the entire OC20-Dense set has no overlap with the OC20 training set. The OC20-Dense training set, drawn from the entire OC20-Dense dataset, comprises 15,450 data entries and serves as an optional addition. The validation set for the OC20-Dense is created by randomly selecting 9,001 data entries from the three out-of-domain splits of the OC20 validation set. For the test set, we used machine-learning-relaxed structures from the OC20-Dense dataset, specifically from the same three out-of-domain splits in the OC20 validation set, provided by the Open Catalyst Challenge 2023[10].

GNN-relaxed structures

As part of the Open Catalyst Challenge 2023, the Open Catalyst Project has provided a set of machine-learning-relaxed structures along with their energies calculated using DFT. These structures, originating from the OC20-Dense validation set, were relaxed using models such as GemNet-OC, SCN and eSCN. Any relaxed structures that are invalid or lack valid DFT energy values were excluded by the dataset creator[8,10]. This includes cases in which the adsorbate fails to bind to the surface, decomposes into different atoms or molecules, or causes substantial alterations to the surface from its original state. After filtering out these invalid configurations, the remaining counts for the machine-learning-relaxed test sets using the GemNet-OC, SCN and eSCN models are 11,508, 11,630 and 11,755 structures, respectively. Within these datasets, only a subset of structures—919, 922 and 922, respectively—have valid DFT-verified energy values. Our accuracy analysis concentrates on these approximately 920 machine-learning-relaxed structures, each supported by a reliable DFT energy assessment. Meanwhile, the embedding and attention score analyses fully utilize predictions on all the valid machine-learning-relaxed structures.

Structure-string conversion

The input data is entirely text based, adhering to the string-type input format outlined in the original CatBERTa paper[20]. We generate textual strings by converting the relaxed structures in the OC20 and OC20-Dense datasets (Fig. 2). Our textual input format is structured into three segments: the adsorbate, the catalytic surface, and the depiction of the adsorption configuration. Specifically, the adsorbate segment simply contains its elemental symbol. For the catalytic surface part, we integrate information about the catalyst's overall composition, along with its Miller index. For these two segments, the information is sourced from the pre-existing metadata of the OC20 dataset. The depiction of the adsorption configuration is achieved by pinpointing both the primary and secondary atoms involved in the interaction. This method was selected due to its proven effectiveness in predicting energy outcomes in past research[20]. In this process we identify these interacting elements using the Pymatgen library. First, we establish atomic connectivity on the basis of a predefined cutoff radius, which is a covalent radius of the atom. We then pinpoint the atoms connected to those in the adsorbate and surface. The connected atoms of the adsorbate atoms are classified as primary interacting atoms, whereas the neighbouring atoms of the primary interacting atoms on the surface are grouped as secondary interacting atoms.

To convert structures from LLM-generated CIFs, we employ a more lenient and simplified approach. Initially, we identify and specify only the adsorbate atom closest to the surface. Next, we gather the primary neighbour atoms surrounding this adsorbate atom. Following this, we collect the secondary neighbour atoms from the primary neighbours. In this process, we use a multiplier of four for the cutoff radius, which means that the neighbour atoms are those within four times the covalent radius. This approach is based on the understanding that the structures in the generated CIF are indicative and not exact.

CatBERTa

In this work we employ the CatBERTa model as a predictive language model. This text-based model is specifically designed and trained for predicting relaxed energy in adsorbate–catalyst systems. The model incorporates the RoBERTa encoder, originally pretrained on an extensive natural language corpus that includes resources such as BookCorpus and English Wikipedia, cumulatively exceeding 160 GB (ref. 39). The RoBERTa model—diverging from the conventional BERT model[40], which masks a fixed 15% of tokens in each epoch during training—adopts a dynamic masking approach. This method alters the masked tokens variably across different training epochs, thereby improving the model's proficiency in predicting masked words and grasping syntactic and semantic nuances.

The CatBERTa model is fine-tuned for an energy-prediction task. The original RoBERTa's classification head is replaced with a regression head, comprising a linear and activation layer. This modification allows CatBERTa to generate a singular scalar value of energy predictions. For this prediction, the embedding of the special '<s>' token, after encoder processing, serves as input for this regression head. The training hyperparameters and the architecture details of the CatBERTa model are provided in Supplementary Table 1, whereas pretraining and fine-tuning strategies are listed in Supplementary Table 4.

EquiformerV2

Equiformer is a GNN that is SE(3)/E(3)-equivariant, adeptly fusing the inductive biases of equivariance with the dynamic strengths of transformers[30]. Equiformer stands out by demonstrating that transformers can be effectively adapted to 3D atomistic graphs. This is achieved by two main factors. First, Equiformer modifies the traditional transformer by substituting SE(3)/E(3)-equivariant operations for the original operations. Second, Equiformer introduces equivariant graph attention as the attention mechanism.

In the pretraining stage, we use the EquiformerV2 embedding for graph representation purposes due to its excellent performance in the OC20 dataset. EquiformerV2 (ref. 28), a refined version of the original Equiformer, brings to the table a host of enhancements. These improvements encompass: the replacement of SO(3) convolutions with eSCN convolutions, the introduction of attention renormalization, the incorporation of separable S2 activation, and the application of separable layer normalization[28]. Such advancements have elevated EquiformerV2, especially in the energy and force predictions task on the OC20 dataset. The model demonstrates state-of-the-art accuracy in its performance, achieving an MAE of 0.22 eV for the S2EF task and 0.31 eV for the IS2RE task, outperforming other benchmarked models.

The graph embeddings, in our case, are extracted after the final layer normalization in the EquiformerV2 model, which precedes the energy and force prediction stage (Fig. 1c). Within this model, each atom is represented as a node, and each atom node is characterized by a two-dimensional embedding tensor, collectively forming a three-dimensional tensor for the entire system. The size of the system tensor is defined by the number of atoms, the count of spherical channels, and the maximum degree of spherical harmonics involved[28]. The extraction of graph embeddings begins with reshaping the two-dimensional atom embedding tensor into a one-dimensional

tensor. Max-pooling is then applied across all these one-dimensional atom embeddings in the system, yielding a single, comprehensive embedding for each system. In our study, the final embedding of each system is represented by a tensor with a size of 3,200. Consequently, during the graph-assisted pretraining phase, text embeddings, initially sized at 768, undergo a linear projection head to match the 3,200-size tensor.

Contrastive loss

Graph-assisted pretraining synergistically aligns text and graph embeddings through a self-supervised framework. Utilizing the graph encoder EquiformerV2 in a static (frozen) state, this method is specifically designed to transfer the insights from the graph to the text modality.

Inspired by the methodology used in Contrastive Language-Image Pretraining[41], the graph-assisted pretraining strategy incorporates both text and graph encoders. The pretraining mechanism is centred around the optimization of a symmetric cross-entropy loss[42]. This optimization process aims to increase the similarity between embeddings from matching text-graph pairs while decreasing the similarity between those from unmatching pairs. By using a contrastive loss function, the primary objective is to establish a meaningful correlation between text and graph embeddings. The overarching goal is to precisely align embeddings from corresponding text-graph pairs while effectively differentiating between non-corresponding pairs.

The mathematical formulation of the loss function is defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(I_i,T_i)/\tau}}{\sum_{j=1}^{N}e^{\mathrm{sim}(I_i,T_j)/\tau}\mathbb{I}_{\{i\neq j\}}} - \frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(T_i,I_i)/\tau}}{\sum_{j=1}^{N}e^{\mathrm{sim}(T_i,I_j)/\tau}\mathbb{I}_{\{i\neq j\}}} \quad (2)$$

In this expression, T_i and G_i represent the embeddings of the ith text and graph, respectively. The function $\mathrm{sim}(G_i, T_j)$ calculates the cosine similarity between the embeddings of the ith graph and jth text. We also introduce τ as a temperature parameter, which serves to appropriately scale the similarity scores within the model.

Fine-tuning CrystaLLM

CrystaLLM is a GPT-2-based large language model designed to generate crystal structures in CIF format for a given composition and, optionally, a specified space group[26]. Adapted from the nanoGPT[43] model implementation, the model is trained from scratch with a vocabulary size of 371, specifically for inorganic crystal systems. The training data were sourced from various databases, including the Materials Project[44], the Open Quantum Materials Database[45] and NOMAD[46]. The tokenizer for the CrystaLLM operates on character bytes. The CrystaLLM (large) model, which we use in our framework, consists of 16 layers with 16 heads each, and individual block sizes of 2,048. This model has been trained for 48,000 epochs, enabling it to predict text-based structural CIF representations of various crystal systems by leveraging the generative capabilities of language models.

By fine-tuning the pretrained model over the combined training data from the OC20 and OC20-Dense datasets for 6,000 epochs, we leverage the learned embedding space of the existing model. This allows us to transfer the corresponding knowledge to our specific task. Compared with the pretraining learning rate of 1×10^{-3}, the fine-tuning learning rate is set to 6×10^{-4} to further enhance the model's generalization capabilities. Consequently, our fine-tuned CrystaLLM generates CIFs for target adsorbate–catalyst systems using only the adsorbate and catalyst bulk chemical symbols, along with surface orientation.

CrystaLLM subsampling

In this study, we applied targeted selection criteria to extract adsorbate–catalyst pairs from the OC20-Dense training dataset. Two filters were used to derive the final 66 downsampled adsorbate–catalyst pairs: the first ensures that the number of atoms conforms to the token limits of the original CrystaLLM model, whereas the second focuses on composition matching to maintain a minimum level of accuracy in the generated configurations.

As we fine-tuned the original CrystaLLM model, we adhered to the token length constraints of the tokenizer, which is set at 3,000. We selected systems from the OC20-Dense dataset where the adsorbates contain no more than five atoms each, across 51 unique adsorbates. Notably, each adsorbate is composed of a maximum of three different types of elements, resulting in a total of 27 unique, filtered adsorbates. We also imposed a constraint on the number of atoms in the catalyst, selecting catalyst systems with no more than 72 atoms. Applying these filters to the OC20-Dense training dataset yielded 108 unique adsorbate–catalyst pairs out of the initial 235 pairs.

We filtered out nonsensical compositions in the generated CIFs to ensure a minimum level of generation accuracy. Specifically, the adsorbate had to exactly match the given composition, while we allowed a tolerance of up to 12 atoms of deviation in the catalyst. By applying this composition-matching filter, the 108 adsorbate–catalyst pairs were reduced to 66. Due to the stochastic nature of the generative model, the number of valid systems can vary across iterations.

Following this process, we selected one CIF for each pair that passed the criteria and converted it into the desired string configurations. If more than one generation out of three passed the criteria for a given pair, we randomly selected one CIF. We randomly selected 14 pairs from the in-domain split for exemplary visualization in Fig. 5. It is noteworthy that the 66 adsorbate–catalyst systems correspond to 5,141 overall configurations in the original training dataset, each with a DFT-calculated energy value.

Data availability

Access to the Open Catalyst 2020 dataset is available via GitHub at https://github.com/FAIR-Chem/fairchem (ref. 47). The Open Catalyst 2020 Dense dataset and relevant data about the Open Catalyst Challenge 2023 are available via GitHub at: https://github.com/Open-Catalyst-Project/AdsorbML (ref. 48). The preprocessed data, formatted for compatibility with the training framework, is available via figshare at https://doi.org/10.6084/m9.figshare.27208356.v2 (ref. 49).

Code availability

The Python code used in this study is available on Zenodo at https://doi.org/10.5281/zenodo.13917199 (ref. 50) and on GitHub at https://github.com/hoon-ock/multi-view.

References

1. Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
2. Zitnick, C. L. et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. Preprint at https://arxiv.org/abs/2010.09435 (2020).
3. Chanussot, L. et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
4. Reiser, P. et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **3**, 93 (2022).
5. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
6. Wander, B., Broderick, K. & Ulissi, Z. W. Catlas: an automated framework for catalyst discovery demonstrated for direct syngas conversion. *Catal. Sci. Technol.* **12**, 6256–6267 (2022).
7. Tran, R. et al. Screening of bimetallic electrocatalysts for water purification with machine learning. *J. Chem. Phys.* **157**, 074102 (2022).
8. Lan, J. et al. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Comput. Mater.* **9**, 172 (2023).

9. Cao, Z., Barati Farimani, O., Ock, J. & Barati Farimani, A. Machine learning in membrane design: from property prediction to AI-guided optimization. *Nano Lett.* **24**, 2953–2960 (2024).
10. Open Catalyst Challenge. *Open Catalyst Project* https://opencatalystproject.org/challenge.html (2023).
11. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
12. Schütt, K. T. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds Guyon, I. et al.) 991–1001 (Curran Associates, 2017).
13. Gasteiger, J. et al. GemNet-OC: developing graph neural networks for large and diverse molecular simulation datasets. *Trans. Mach. Learn. Res.* u8tvSxm4Bs (2022).
14. Pablo-García, S. et al. Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks. *Nat. Comput. Sci* **3**, 433–442 (2023).
15. Studt, F. Grand challenges in computational catalysis. *Front. Catal.* **1**, 658965 (2021).
16. Giulimondi, V., Mitchell, S. & Pérez-Ramírez, J. Challenges and opportunities in engineering the electronic structure of single-atom catalysts. *ACS Catal.* **13**, 2981–2997 (2023).
17. Cao, Z., Magar, R., Wang, Y. & Barati Farimani, A. Moformer: self-supervised transformer model for metal-organic framework property prediction. *J. Am. Chem. Soc.* **145**, 2958–2967 (2023).
18. Balaji, S. & Magar, R. Gpt-molberta: Gpt molecular features language model for molecular property prediction. Preprint at https://arxiv.org/abs/2310.03030 (2023).
19. Xu, C., Wang, Y. & Barati Farimani, A. Transpolymer: a transformer-based language model for polymer property predictions. *npj Comput. Mater.* **9**, 64 (2023).
20. Ock, J., Guntuboina, C. & Barati Farimani, A. Catalyst energy prediction with catberta: unveiling feature exploration strategies through large language models. *ACS Catal.* **13**, 16032–16044 (2023).
21. Wang, S. et al. Universal Brønsted–Evans–Polanyi relations for C–C, C–O, C–N, N–O, N–N, and O–O dissociation reactions. *Catal. Lett.* **141**, 370–373 (2011).
22. Sutton, J. E. & Vlachos, D. G. A theoretical and computational analysis of linear free energy relations for the estimation of activation energies. *ACS Catal.* **2**, 1624–1634 (2012).
23. Ock, J., Tian, T., Kitchin, J. & Ulissi, Z. Beyond independent error assumptions in large GNN atomistic models. *J. Chem. Phys.* **158**, 214702 (2023).
24. Huang, H. & Barati Farimani, A. Multimodal learning of heat capacity based on transformers and crystallography pretraining. *J. Appl. Phys.* **135**, 165104 (2024).
25. Badrinarayanan, S., Guntuboina, C., Mollaei, P. & Farimani, A. B. Multi-peptide: multimodality leveraged language-graph learning of peptide properties. Preprint at https://arxiv.org/abs/2407.03380 (2024).
26. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. Preprint at https://arxiv.org/abs/2307.04340 (2024).
27. Gruver, N. et al. Fine-tuned language models generate stable inorganic materials as text. In *Proc. Twelfth International Conference on Learning Representations* https://openreview.net/forum?id=vN9fpfqoP1 (OpenReview, 2024).
28. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. Equiformerv2: improved equivariant transformer for scaling to higher-degree representations. In *Proc. Twelfth International Conference on Learning Representations* https://openreview.net/forum?id=mCOBKZmrzD (OpenReview, 2024).
29. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
30. Liao, Y.-L. & Smidt, T. Equiformer: equivariant graph attention transformer for 3D atomistic graphs. In *Proc. Eleventh International Conference on Learning Representations* https://openreview.net/forum?id=KwmPfARgOTD (OpenReview, 2023).
31. Gao, W. et al. Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates. *Nat. Commun.* **11**, 1196 (2020).
32. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Theory-guided machine learning finds geometric structure–property relationships for chemisorption on subsurface alloys. *Chem* **6**, 3100–3117 (2020).
33. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
34. M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
35. Jadhav, Y., Pak, P. & Farimani, A. B. Llm-3D print: large language models to monitor and control 3D printing. Preprint at https://arxiv.org/abs/2408.14307 (2024).
36. Jadhav, Y. & Farimani, A. B. Large language model agent as a mechanical designer. Preprint at https://arxiv.org/abs/2404.17525 (2024).
37. Hammer, B., Hansen, L. B. & Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew–Burke–Ernzerhof functionals. *Phys. Rev. B* **59**, 7413–7421 (1999).
38. Boes, J. R., Mamun, O., Winther, K. & Bligaard, T. Graph theory approach to high-throughput surface adsorption structure generation. *J. Phys. Chem. A* **123**, 2281–2285 (2019).
39. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
40. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2019).
41. Radford, A. et al. Learning transferable visual models from natural language supervision. Preprint at https://arxiv.org/abs/2103.00020 (2021).
42. Van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2019).
43. Karpathy, A. NanoGPT. *GitHub* https://github.com/karpathy/nanoGPT (2024).
44. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
45. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
46. Draxl, C. et al. NOMAD: a distributed web-based platform for managing materials science research data. *J. Open Source Softw.* **8**, 5388 (2023).
47. FAIR-Chem/fairchem. *GitHub* https://github.com/FAIR-Chem/fairchem (2024).
48. Open-Catalyst-Project/AdsorbML. *GitHub* https://github.com/Open-Catalyst-Project/AdsorbML (2024).
49. Ock, J., Badrinarayanan, S., Magar, R., Antony, A. & Barati Farimani, A. *Language and Graph Multimodal Data for Heterogeneous Catalyst* (FigShare, 2024); https://doi.org/10.6084/m9.figshare.27208356.v2
50. Ock, J. hoon-ock/multi-view: release. *Zenodo* https://doi.org/10.5281/zenodo.13922448 (2024).

Acknowledgements
We thank Meta Fundamental AI Research (FAIR) for providing the publicly available Open Catalyst Project dataset and organizing the Open Catalyst Challenge. We note the assistance of ChatGPT for grammar and typo corrections in preparing the manuscript.

Author contributions
J.O., S.B. and A.B.F. designed the research study. J.O., S.B., R.M. and A.A. developed the method, wrote the code and performed the analysis. All authors wrote and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00930-7.

Correspondence and requests for materials should be addressed to Amir Barati Farimani.

Peer review information *Nature Machine Intelligence* thanks Sergio Pablo-García and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.