

Generative Pretrained Transformer for Heterogeneous Catalysts

Dong Hyeon Mok and Seoin Back*



Cite This: <https://doi.org/10.1021/jacs.4c11504>



Read Online

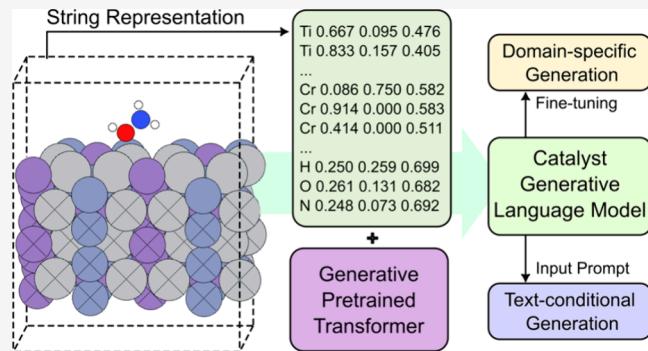
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Discovery of novel and promising materials is a critical challenge in the field of chemistry and material science, traditionally approached through methodologies ranging from trial-and-error to machine-learning-driven inverse design. Recent studies suggest that transformer-based language models can be utilized as material generative models to expand the chemical space and explore materials with desired properties. In this work, we introduce the catalyst generative pretrained transformer (CatGPT), trained to generate string representations of inorganic catalyst structures from a vast chemical space. CatGPT not only demonstrates high performance in generating valid and accurate catalyst structures but also serves as a foundation model for generating the desired types of catalysts by text-conditioning and fine-tuning. As an example, we fine-tuned the pretrained CatGPT using a binary alloy catalyst data set designed for screening two-electron oxygen reduction reaction (2e-ORR) catalyst and generated catalyst structures specialized for 2e-ORR. Our work demonstrates the potential of generative language models as generative tools for catalyst discovery.



Downloaded via XI'AN JIAOTONG UNIV on November 25, 2024 at 11:10:15 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

The discovery of novel and advanced materials with superior performance surpassing that of existing materials is essential across various fields.^{1–4} As new materials are developed, performance benchmarks rise, making it more challenging to identify better materials. This increases the difficulty and cost of traditional methods, such as experimental trial-and-error approach or computationally intensive simulations such as density functional theory (DFT) calculations.⁵

In this context, machine learning (ML) has become a major technology for materials discovery, enabling the more efficient exploration of vast chemical spaces. There are two primary ML-driven approaches to materials searches: high-throughput virtual screening (HTVS) and inverse design.⁶ HTVS starts with a data set within the screening scope and gradually narrows down the search space to focus on materials with desired properties. Initial candidates are filtered by multiple criteria, leaving only those that satisfy all of the requisite property thresholds as promising candidates. ML models enhance the screening process's efficiency by predicting which candidates are likely to fall outside the desired criteria, allowing for their early exclusion.^{7–10} However, HTVS suffers from the drawback that the search scope is limited to the user-selected database. Expanding the chemical space through elemental substitution is a simple and common solution, yet this approach is unable to introduce structural diversity.^{9,11}

Inverse design, in contrast, reverses the traditional material design process by starting with the desired set of properties. Unlike HTVS, which relies on a predefined set of candidates,

inverse design leverages a broader search space and proposes new materials by refining structures until the desired properties are optimally achieved (global optimization) or by generating materials from a learned distribution of materials with desired properties (generative model).^{12,13} In recent years, the rapid advancement of generative models, commonly used in creating images, audio, and video, has prompted significant interest in their applications within the materials domain.^{14–16} Various generative techniques have been investigated, ranging from variational autoencoders (VAE)¹⁷ and generative adversarial network (GAN)¹⁸ to diffusion models.¹⁹ More recently, autoregressive language models based on transformer architectures,^{20,21} known for utilizing the self-attention mechanism and serving as the backbone of large language models (LLM), have shown promising results in generating both organic and inorganic materials.^{22,23}

Since language models were originally designed to handle discrete sequences of variable lengths, they are particularly suited for generating crystal structures, which consist of discrete entities (atomic symbols) and can vary in length (number of atoms). This inherent flexibility makes language models more adept at this task compared to other generative

Received: August 21, 2024

Revised: November 17, 2024

Accepted: November 19, 2024

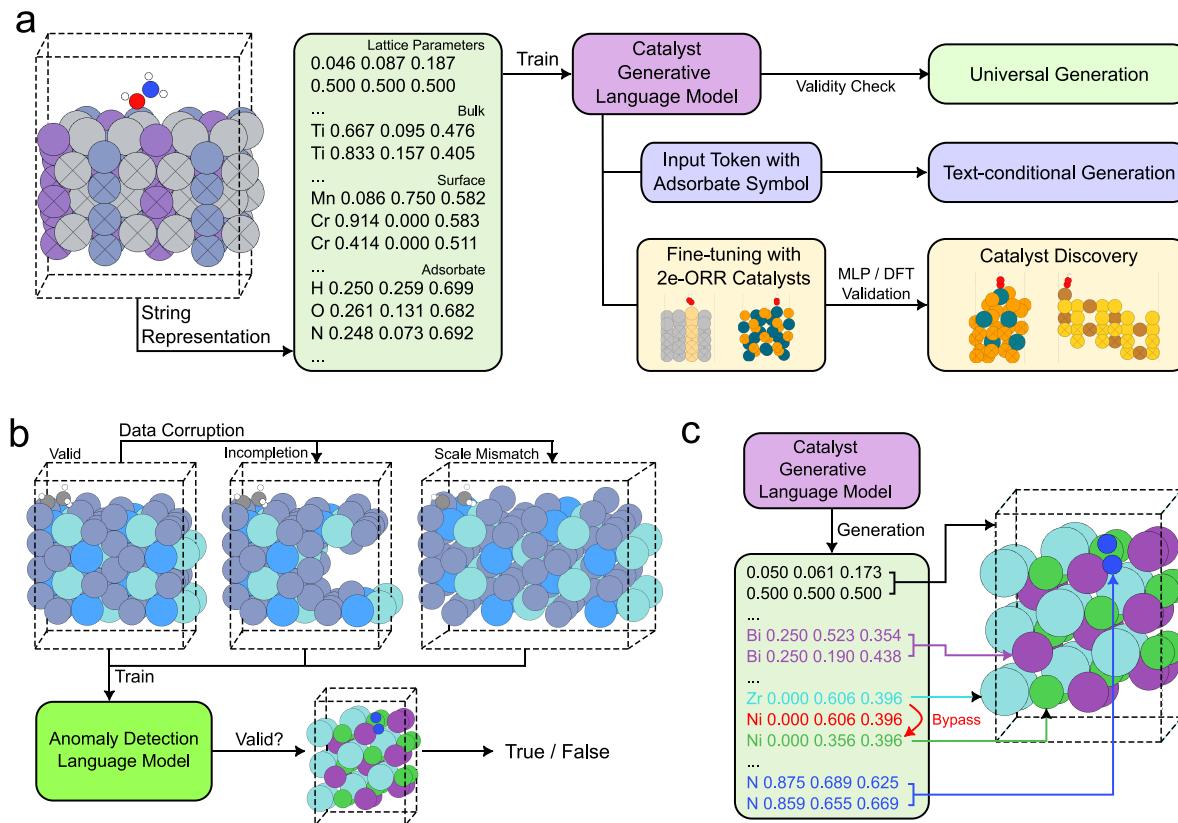


Figure 1. (a) Schematic of the workflow for training a catalyst generative model. The trained model was further applied for adsorbate-specific and domain-specific generation by using text-conditioning and fine-tuning, respectively. After evaluating various validity metrics, candidate catalysts were validated with MLP predictions and DFT calculations. (b) Schematic of the anomaly detection model to check the catalyst validity of generated structures. The detection model was trained on intentionally corrupted data set to classify the validity of string representations of structures. (c) Catalyst generation using a bypass method to address atomic overlapping. The generation of new atoms located close to already existing atoms was ignored in the process that converts string representations to 3D structures.

models, such as diffusion models, which require additional considerations to handle the discrete and variable nature of crystal structures.²⁴ Flam-Shepherd and Aspuru-Guzik trained a GPT-based language model from scratch on sequences of tokenized atomistic data, including molecules, protein binding pockets and crystals.²² They reported that their models performed comparably to or even surpassed other generative models. Antunes et al. demonstrated that CIF-format representation of crystals is also valid for training language models for generative purposes.²⁵

In this study, we developed a generative pretrained transformer (GPT)-based approach for catalyst discovery and demonstrated the potential of the model as a foundation model for catalyst generation. We encoded catalyst structures into textual sequence data and trained the model from scratch using a massive catalyst data set. Since there is no standardized metric for evaluating catalyst generation performance, we developed an anomaly detection model to assess the validity of generated catalyst structures. To demonstrate the practical utility of our approach for catalyst discovery, we employed text-conditional generation and a fine-tuning strategy. The model could generate catalysts with specific adsorbate atoms from an input prompt with adsorbate symbols. Fine-tuning the pretrained model on a small and biased data set of binary alloy catalysts, specifically designed for two-electron oxygen reduction reaction (2e-ORR) catalysts,^{26,27} successfully modified model's generation space. Finally, we evaluated the generated catalysts using machine learning potentials (MLP)

and subsequent density functional theory (DFT) calculations, demonstrating that our strategy can generate new catalyst structures that have not been previously collected in a database. This work highlights the potential of GPT models as useful tools for catalyst discovery.

2. RESULTS AND DISCUSSION

We implemented the generative pretrained transformer 2 (GPT-2) architecture²⁸ to generate catalyst structures, hereafter referred to as “CatGPT”. The CatGPT was trained autoregressively on two million catalyst structures containing both surface and adsorbate atoms, collected from the Open Catalyst 2020 Structure to Energy and Forces task data set (OC20-S2EF 2M data set).²⁹ These structures were represented as a corpus consisting of tokenized strings of lattice parameters, atomic symbols, and 3D coordinates (see Methods section). For validation, we used the Open Catalyst 2020 in-domain validation set (OC20-S2EF Val-ID data set).

Since there are no systematically defined metrics for assessing the validity and quality of generated catalyst structures, we developed a model to detect invalidly generated structures based on bidirectional encoder representations from transformers (BERT)³⁰ and a validity metric for generated catalyst structures, hereafter referred to as the “anomaly detection model” and “catalyst validity”, respectively.

To manipulate the generation space of the CatGPT model for different purposes, we employed a text-conditioning and

Table 1. Catalyst Structure Generation Performance of CatGPT and CatGPT-BP^a

| method | validity | | coverage | | property distribution | |
|------------------|----------------------|--------------------|----------|------------|-----------------------|-------------------|
| | structural validity↑ | catalyst validity↑ | recall↑ | precision↑ | EMD (ρ)↓ | EMD (N_{el})↓ |
| (validation set) | 1.000 | 0.997 | 0.999 | 1.000 | 0.014 | 0.019 |
| CatGPT | 0.920 | 0.913 | 0.999 | 0.999 | 0.217 | 0.034 |
| CatGPT-BP | 1.000 | 0.899 | 0.999 | 0.999 | 0.182 | 0.028 |

^aTo calculate coverage and property distributions of the generated structures, 10 000 randomly sampled structures from the validation set (OC20 S2EF Val-ID data set) were used as the ground truth. The coverage and property distribution of validation set were calculated with respect to 10 000 randomly sampled structures from the training set.

fine-tuning strategy. Text-conditional generation using input prompts including adsorbate symbols allowed the model to generate catalysts with specific adsorbate atoms without requiring additional training or data generation. Fine-tuning the pretrained CatGPT model constrained the composition and geometry of catalyst structures to align with the fine-tuning data set while preserving generation performance of the pretrained model. This approach has the potential to discover the desired catalysts by integrating the model with MLP predictions and DFT calculations. The scheme of this work is summarized in Figure 1a.

2.1. Evaluation Metrics of Generated Structures.

Generative models for crystal structures have used two metrics to evaluate the validity of generated structures.³¹ First, compositional validity assesses the charge neutrality of the system through semiconducting materials from analogy and chemical theory (SMACT).³² Second, structural validity checks for overlapping atoms based on atomic distance criteria. In this work, atoms were considered to overlap if the distance between them was less than 0.5 Å.

However, it should be noted that catalyst structures consisting of a catalyst surface and adsorbate differ from bulk structures used in previous works. Generally, surface structures are prepared by cleaving bulk structures often resulting in nonstoichiometric surfaces; thus, the charge neutrality of the system is not always guaranteed. In addition, as the CatGPT generates catalyst surfaces with adsorbates simultaneously, the compositional validity is not applicable. Although structural validity is partially useful, the following two invalid examples, identified in generated catalyst structures that passed the structural validity check, indicate that structural validity is necessary but not sufficient: (1) incomplete structures and (2) scale mismatches.

Incompletion occurs when the model terminates generation prematurely, resulting in structures with unusually empty regions or isolated atoms (Figure S1a). While incomplete generated structures could be interpreted as the generation of porous or defected catalyst structures, there are no such structures in the training data, so this outcome is unintentional. Scale mismatch refers to invalid generation when the lattice parameters are disproportionately large relative to the number of atoms in the unit cell, leading to disconnected atomic bonds (Figure S1b). Despite these anomalies being obviously recognized as unsuitable for catalyst structures by experts, there are no systematic criteria for identifying them.

Thus, we developed an anomaly detection model using the BERT architecture to evaluate the catalyst validity of the generated structures (Figure 1b). This model can detect incompleteness and scale mismatch from string representation of catalyst structures. Because the model receives the string representation as input, it can directly evaluate the string representation of generated structures from the CatGPT

without any conversion step. The model was trained to perform binary classification on a half-corrupted OC20 S2EF 2M data set, where a random half of the original data set was intentionally modified to represent anomalies of incompleteness and scale mismatch. To create the incompleteness feature, we randomly removed between 20% and 80% of atoms from the structures. For the scale mismatch feature, we expanded the unit cell dimensions by 150–200% while maintaining the fractional coordinates. We note that reducing the unit cell dimensions was not considered, as it could be filtered out by increasing the distance cutoff for structural validity (Figure S2). The detection model demonstrated a high accuracy of 96% in distinguishing valid and invalid catalyst structures for the half-corrupted validation set (Figure S3a). We employed this model to evaluate the “catalyst validity” of the generated structures. Overall, we evaluated the validity of generated catalyst structures in two aspects: (1) structural validity and (2) catalyst validity.

Coverage and property distribution are key metrics for evaluating how well the generated data replicate the validation data, which serves as the ground-truth, meaning real catalyst structures. We used the method described by Xie et al.³¹ to evaluate these metrics. Coverage assesses the structural and compositional similarity between the ensemble of generated and the ground-truth structures, while property distribution evaluates the similarity in property distributions between generated and ground-truth structures. The structures and compositions are converted into vector forms using CrystalNN and Magpie fingerprints, respectively, as implemented in Matminer.³³ This allows us to calculate the similarity of the catalysts arithmetically, enabling us to determine the coverage. Coverage is assessed by computing the pairwise distances between the vectorized structure and composition of the ground-truth structures and those of the generated structures.³⁴ Coverage Recall (Precision) quantifies the proportion of ground-truth (generated) structures whose feature vector distances to the nearest generated (ground-truth) structures fall within a predefined cutoff criterion (0.2 for structure and 8.5 for composition, respectively). Property distribution is evaluated by computing the Earth Mover’s Distance (EMD)³⁵ between the property distributions of the generated and ground-truth structures sets. Density (ρ , g/cm³) and number of unique elements (N_{el}) of the structures are used as properties.

To establish the upper and lower bounds of the metrics and to rationalize the anomaly detection model as a sanity check for catalyst validity, we computed these metrics for the validation set. The catalyst validity score of 0.997 observed for the validation set supports the suitability of this metric for evaluating the quality of generated catalyst structures (Table 1). Notably, the low compositional validity score for validation structures regardless of the presence of adsorbates (0.133 and

Table 2. Text-Conditional and Universal Generation Performance of CatGPT Trained on the String Representation with Adsorbate Symbols^a

| adsorbate | match rate↑ | validity | | coverage | | property distribution | |
|----------------------------------|-------------|----------------------|--------------------|----------|------------|-----------------------|-------------------|
| | | structural validity↑ | catalyst validity↑ | recall↑ | precision↑ | EMD (ρ)↓ | EMD (N_{el})↓ |
| *O | 0.992 | 0.924 | 0.887 | 0.913 | 0.998 | 0.244 | 1.492 |
| *OH | 0.980 | 0.902 | 0.892 | 0.990 | 1.000 | 0.237 | 0.569 |
| *C | 0.989 | 0.906 | 0.850 | 0.984 | 0.996 | 0.275 | 1.590 |
| *COCH ₂ O | 0.986 | 0.914 | 0.912 | 0.978 | 0.997 | 0.251 | 0.429 |
| *CH ₂ OH | 0.983 | 0.897 | 0.888 | 0.980 | 0.999 | 0.196 | 0.461 |
| *CHOCH ₂ OH | 0.981 | 0.903 | 0.933 | 0.977 | 0.999 | 0.223 | 0.435 |
| *NHNH | 0.967 | 0.912 | 0.898 | 0.985 | 1.000 | 0.205 | 0.588 |
| *NO ₂ NO ₂ | 0.952 | 0.923 | 0.907 | 0.991 | 0.994 | 0.253 | 0.522 |
| *N | 0.975 | 0.904 | 0.876 | 0.989 | 0.994 | 0.150 | 1.524 |
| *NO ₃ | 0.977 | 0.921 | 0.905 | 0.988 | 0.997 | 0.177 | 0.567 |
| average | 0.978 | 0.911 | 0.891 | 0.988 | 0.997 | 0.221 | 0.667 |
| random | | 0.883 | 0.904 | 0.998 | 0.996 | 0.232 | 0.296 |

^a1000 structures were conditionally generated per adsorbate, while an additional 1000 structures were universally generated without conditions (random). The match rate represents the ratio of catalyst structures in which the adsorbate matched the conditioned adsorbate symbol. Coverage and property distributions of the generated structures were calculated with respect to 10 000 randomly sampled structures from the validation set (OC20 S2EF Val-ID data set), which served as the ground truth.

0.557 with and without adsorbate atoms, respectively) confirms that charge neutrality is not a reliable measure for assessing the valid generation of catalyst structures.

2.2. Performance of Generations by CatGPT. We generated 10 000 catalyst structures using the “⟨bos⟩” token, which indicates the beginning of catalyst sequences for random catalyst generation, and evaluated their validity, coverage and property distributions. The generation was performed using a sampling strategy with the default generation parameters provided in the HuggingFace library, as detailed in the Methods section “Model Structures and Training”. Validity and coverage were evaluated for all generated structures, while property distributions were assessed for 5000 valid structures randomly sampled from those that satisfied all validity criteria.

The results demonstrated that CatGPT performs exceptionally well in generating high-quality catalyst structures and accurately retrieving a distribution similar to the validation set, as evidenced by nearly perfect precision and recall scores, respectively. However, we observed that it occasionally generates catalyst structures with overlapping atoms, indicating some atoms are within 0.5 Å of each other, resulting in a slightly low structural validity score (Table 1). To address the generation of overlapping atoms without modifying the model architecture, we introduced a new feature to bypass overlapping atoms when converting string representations of catalysts into 3D structures (Figure 1c). This feature takes advantage of the autoregressive nature of the language model, which predicts tokens sequentially, unlike other generation models that create entire structures at once by decoding latent vectors or denoising Gaussian noise. The generative model with this feature, named CatGPT-BP, shows perfect structural validity with only a minimal decrease in the catalyst validity score. The changes in the EMD of the property distributions is also minor, especially considering the comparatively large standard deviations (1.11 and 0.86 for ρ and N_{el} , respectively) observed in the property distributions (Figure S4a). Additionally, the slightly lower catalyst validity value compared to the validation set, despite nearly maximum structural validity and coverage, highlights the difficulty of detecting anomalies using other metrics and the necessity of the catalyst validity metric.

We also note that filtering with catalyst validity does not alter the data distribution (Figure S5).

Since an ideal generative model should be invariant to translation (or rotation) and permutation, we prepared two augmented data sets to introduce translational and permutational features (details in the Methods section). As shown in the training loss curves (Figure S6a), while the translation and rotation of structures barely affected performance, permutational augmentation, which involves shuffling the order of atoms, clearly decreased the performance. The difficulty in accounting for permutation invariance in CatGPT implies that the sequence in the string representation of catalyst structures is itself informative.²⁴ This is more clearly demonstrated in the generation probability distribution of the end token (Figure S6b).

In the model trained without the augmented permutation data set, the tokens of adsorbate atoms increased the probability of end token generation, effectively acting as indicators for the termination of generation. However, in the model trained with the augmented permutation data set, the probability increased gradually with sequence length, making it difficult to identify the end of generation. This led to repeated generations of atoms in the same position. As a result, the model augmented with the permutation data set demonstrated relatively poor generative performance (Table S1).

3. DISCUSSION

In the field of materials inverse design via generative models, it is generally considered more practical to focus on generating catalysts for specific purposes rather than covering the entire chemical space. With this perspective, we aimed to manipulate the model’s generation space through text-conditioning, fine-tuning, and adjustments of generation parameters.

3.1. Adsorbate Conditional Generation. Since most electrochemical reactions involve specific intermediates, it is important to generate catalysts with particular adsorbates conditionally for practical applications. We tested text-conditional generation for adsorbates by adding adsorbate symbols, which represent the type of adsorbates, in front of the current catalyst representation C. In the new representation,

Table 3. Evaluated Metrics of Fine-Tuned CatGPT Models Using the 2e-ORR Data Set^a

| method | validity | | coverage | | property distribution | | 2e-ORR validity | |
|----------------------|-------------|-----------|----------|------------|-----------------------|----------------|-----------------|------------|
| | structural↑ | catalyst↑ | recall↑ | precision↑ | EMD (ρ)↓ | EMD (N_e)↓ | composition | adsorption |
| (validation set) | 1.000 | 0.988 | 1.000 | 1.000 | 0.111 | 0.000 | 1.000 | 1.000 |
| fine-tuned CatGPT | 0.965 | 0.971 | 0.985 | 0.992 | 0.123 | 0.000 | 0.994 | 0.950 |
| fine-tuned CatGPT-BP | 1.000 | 0.969 | 0.985 | 0.994 | 0.139 | 0.000 | 0.992 | 0.950 |

^aAs in Table 1, the structural, catalyst, and 2e-ORR validity metrics and coverages were evaluated for all generated structures, while property distribution were assessed for 500 valid structures randomly sampled from those satisfying structural and catalyst validity. The coverage and property distribution of the validation set were calculated with respect to 1000 randomly sampled structures from the training set.

Table 4. Validity, Coverage, and Diversity of the Fine-Tuned CatGPT Models at Different Temperatures and Input Queries^a

| temperature | validity | | 2e-ORR validity | | coverage | | diversity | |
|---------------|------------|--------------|-----------------|---------|------------|-------------|-----------|--|
| | catalytic↑ | composition↑ | adsorption↑ | recall↑ | precision↑ | uniqueness↑ | novelty↑ | |
| 1.0 | 0.969 | 0.992 | 0.950 | 0.985 | 0.994 | 0.312 | 0.131 | |
| 1.0 (lattice) | 0.468 | 0.812 | 0.619 | 0.833 | 0.894 | 0.988 | 0.951 | |
| 0.5 | 0.991 | 0.999 | 0.985 | 0.945 | 0.999 | 0.115 | 0.117 | |
| 0.7 | 0.985 | 0.998 | 0.980 | 0.976 | 0.997 | 0.204 | 0.084 | |
| 1.5 | 0.878 | 0.947 | 0.851 | 0.979 | 0.984 | 0.647 | 0.412 | |
| 2.0 | 0.750 | 0.807 | 0.637 | 0.917 | 0.960 | 0.945 | 0.763 | |

^aBecause the metrics were assessed using the bypassing method, the structural validities are 1.00; thus they are omitted from the table. The metrics assessed without the bypassing method can be found in Supporting Information Table S2. The diversity metrics were assessed for the samples that satisfy all validity criteria. “Lattice” indicates the result of generated structures sampled from lattice parameters, while the others were sampled from “⟨bos⟩” tokens.

denoted as C_{ads} , an adsorbate symbols S corresponding to the adsorbate elements is included as follows:

$$C_{\text{ads}} = (S, l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, x_1, y_1, z_1, \dots, e_N, x_N, y_N, z_N)$$

Text-conditional generation was performed by using input prompts that include adsorbate symbols rather than starting with the “⟨bos⟩” token for universal generation. For example, generation initiated with “*O” ideally results in catalysts adsorbed with *O.

Using the model trained on the OC20-S2EF-2M data set represented as C_{ads} , we generated 1000 catalysts for 10 selected adsorbates (*O, *OH, *C, *COCH₂O, *CH₂OH, *CHO-CH₂OH, *NHNH, *NO₂NO₂, *N, and *NO₃), which vary in complexity and elements. We evaluated the model’s capability to generate the intended adsorbate, as well as the validity and diversity of the generated structures (Table 2). The results showed that the model generated surfaces with the correct adsorbate nearly 100% of the time, thus enabling effective text-conditional generation for adsorbates.

3.2. Fine-Tuning for 2e-ORR Catalysts Discovery. Fine-tuning, which involves retraining a pretrained model and its parameters on a specific data set, is widely used across various fields and model types, including GPT, to adapt the pretrained model for a specific domain or purpose. To evaluate whether fine-tuning could effectively tune the generation space of CatGPT while maintaining reasonable validity and diversity, we fine-tuned the model on a small and biased data set. In our previous works, it was proposed that two-electron oxygen reduction reaction (2e-ORR), which electrochemically produces H₂O₂ from O₂, is facilitated on single-site binary alloys consisting of oxophilic and oxophobic elements.^{26,27} The ensemble effect of isolated oxophilic sites weakens the Gibbs free energy of O^{*} (ΔG_{O^*}), a selectivity descriptor of 2e-ORR, by enforcing the on-top site adsorption of O^{*}. When ΔG_{O^*} exceeds $\Delta G_{\text{H}_2\text{O}_2^*}$ (3.52 eV), the maximum selectivity is

considered achieved. The ligand effect of oxophobic elements surrounding the active site controls the Gibbs free energy of OOH* (ΔG_{OOH^*}), an activity descriptor of 2e-ORR, which exhibits maximum activity when its value is 4.22 eV.

To investigate these effects and discover promising 2e-ORR catalysts, Back et al.²⁶ previously generated a binary alloy data set based on two rules: (1) the binary alloys should consist of oxophobic and oxophilic elements (composition rule) and (2) the active sites should be the ontop site of oxophilic atoms (adsorption rule). To ensure that the learned distribution of the fine-tuned model follows the fine-tuning data set distribution, we defined composition and adsorption validity as additional metrics to confirm whether the generated catalyst structures satisfy these rules. We note that the catalyst anomaly detection model to evaluate the catalyst validity was also fine-tuned with a half-corrupted 2e-ORR data set and achieved high accuracy (Figure S3b).

The results from 1000 sampled structures demonstrated that fine-tuning the pretrained model can modify the generation data distribution toward the fine-tuning while maintaining or outperforming the generation performance of the pretrained model (Table 3). Particularly, a composition and adsorption validity score of over 0.95 indicates that the model successfully learns the rules of the fine-tuning data set and generates appropriate catalyst structures for 2e-ORR with only about 1700 fine-tuning data points, which is an amount insufficient to train a model from scratch to generate any valid structures. This is further supported by the zero EMD of the N_e , indicating that all generated structures are binary (Figure S7). Additionally, it was confirmed that the bypassing method remains a practical approach for the fine-tuned model.

3.3. Adjustments of Generation Parameters. Considering that the purpose of generative models is to discover novel catalysts, the diversity, uniqueness, and novelty of generated catalysts are as important as valid generation. Since the probability distribution of autoregressive models can be modified by generation parameters, such as temperature and

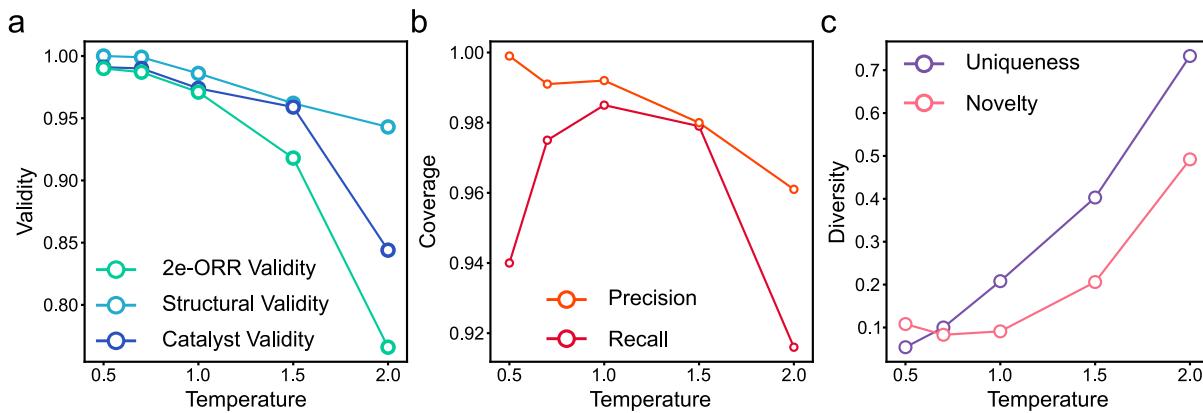


Figure 2. (a) Validity, (b) coverage, and (c) diversity as a function of sampling temperatures. 2e-ORR validity is the ratio of generated structures satisfying both adsorption and composition validity.

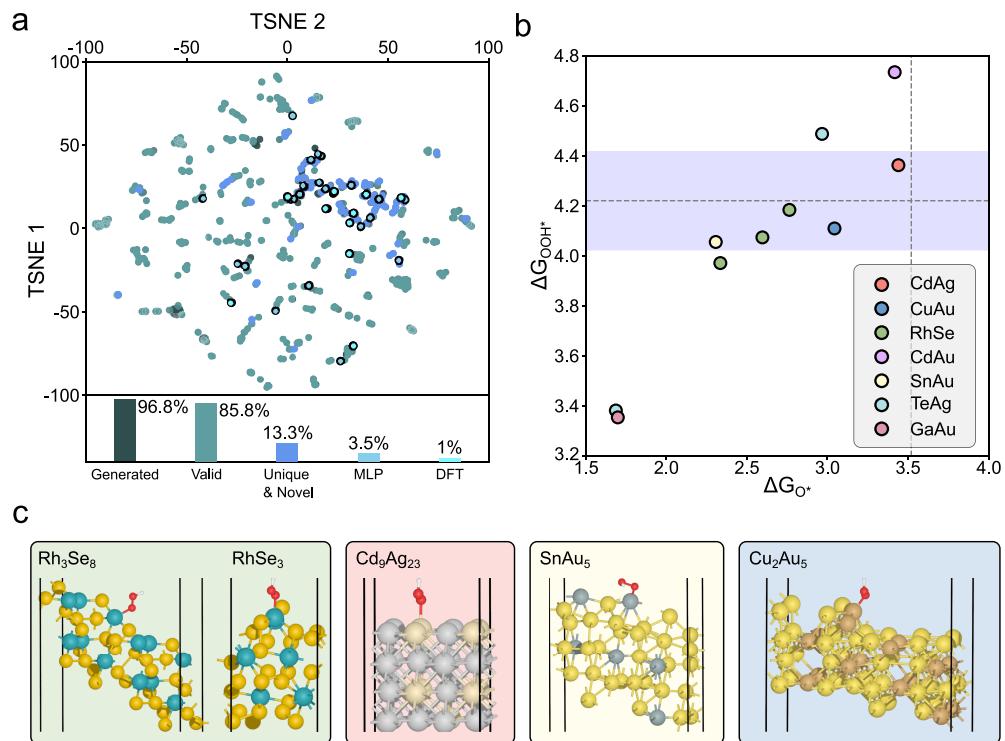


Figure 3. (a) Two-dimensional visualization of the screening process using structure featurization and dimension reduction with t-distributed stochastic neighbor embedding (t-SNE).³⁹ The bar graph shows the percentage of the remaining structures in each screening scheme. (b) DFT calculated ΔG_{O^*} and ΔG_{OOH^*} of 10 catalysts. The horizontal dashed line indicates ΔG_{OOH^*} to achieve the maximum 2e-ORR activity, while the vertical dashed line indicates ΔG_{O^*} to achieve the maximum 2e-ORR selectivity. (c) The atomic structures of active catalysts whose activities are close to the maximum (4.22 ± 0.2 eV, visualized as the blue area in part b).

input queries, we compared the validity and diversity of generated structures from the 2e-ORR fine-tuned model under different temperatures and queries. To evaluate diversity, we used metrics of uniqueness and novelty.³⁶ Uniqueness is the proportion of validly generated structures that are unique ($N_{\text{unique}}/N_{\text{valid}}$), while novelty is the proportion of validly generated structures that are not contained in the training set ($N_{\text{novel}}/N_{\text{valid}}$). The detection of duplicated structures in both cases was performed by “StructureMatcher” module implemented from Pymatgen.³⁷

To observe the effect of input queries, we collected the lattice parameters of 1000 structures randomly selected from the OC20 S2EF Val-ID set and used them as queries for generating new structures. Note that these lattice parameters

were not included in the fine-tuning data set. Because the conditional probabilities of generation vary with the given lattice parameters, the uniqueness and novelty of the model greatly increased. However, the proportion of invalid structures also increased, with around 60% of structures being classified as anomalous based on catalytic validity metric (Table 4). This result indicates that the fine-tuned model is weak in extrapolating validly generated out-of-domain catalyst structures whose lattice parameters are not covered in the fine-tuning data set distribution.

The effect of generation parameters is measured by changing the temperatures (τ), which controls the confidence of the model by adjusting the probability distribution of the next token prediction within the softmax function. At low

temperatures, the probability distribution of the next token prediction becomes sharper, resulting in more valid generations but with less diversity. In contrast, high temperatures flatten the distribution, leading to a higher diversity of predictions but less valid generations. We confirmed this trade-off between diversity and validity from the results of the fine-tuned model, both with and without the bypassing method (Figure 2 and Table S2). Additionally, because the temperature modifies the probability distribution of the trained model, the change in temperature affects the coverage. Coverage recall decreased at both low and high temperatures, because the generated structure distribution is either too localized or too broad, respectively, to recover the validation data. Coverage precision, which indicates the ratio of high-quality structures, follows the same trend as the validity.

3.4. Application for 2e-ORR Catalysts Discovery. We validated the generated catalyst structures by performing geometry optimizations and calculating ΔG_{OOH^*} and ΔG_{O^*} . Instead of directly performing computationally expensive density functional theory (DFT) calculations, we initially employed machine learning potential (MLP), specifically the pretrained EquiformerV2 model.³⁸ Based on the MLP-optimized structures and the predicted binding Gibbs free energies (ΔG_{OOH^*} and ΔG_{O^*}), we collected catalysts satisfying the activity condition with a considerable margin ($3.22 \text{ eV} < \Delta G_{\text{OOH}^*}, \text{MLP} < 5.22 \text{ eV}$) and subsequently performed DFT calculations to obtain both ΔG_{OOH^*} and ΔG_{O^*} .

From the 1000 generated catalyst structures with a temperature parameter of 1.5, 968 structures were successfully converted into atomic structures in 3D space, and 858 structures satisfied all validity metrics. Among these valid structures, we identified 133 unique and novel structures. MLP optimization and binding energy prediction resulted in 35 structures satisfying the activity criteria. Subsequently, clean catalyst surfaces and ${}^*\text{O}$ adsorbed surfaces were constructed from the generated surfaces by removing ${}^*\text{OOH}$ adsorbates and placing ${}^*\text{O}$ in the positions where ${}^*\text{OOH}$ had been adsorbed. Then, DFT optimizations were performed to calculate the ΔG_{OOH^*} and ΔG_{O^*} . Among them, 10 catalyst structures were successfully optimized within a time constraint (24 h or 999 relaxation steps) (Figure 3a,b). The displacement of atomic positions during optimization was comparable to that of the ground-truth structures (Figure S8), providing quantitative evidence that the quality of the generated structures is not compromised compared to that of ground-truth structures. DFT results confirmed that all these structures met the activity and selectivity criteria, with five exhibiting significant 2e-ORR activity in the range $4.22 \pm 0.2 \text{ eV}$ (Figure 3c). We note that none of these structures exist in the training data, and the suggested binary alloy combinations have not been reported as 2e-ORR catalysts in literature. This result highlights that the GPT-based catalyst generative model is a potential platform for the discovery of novel catalysts via fine-tuning with a relatively small data set.

4. CHALLENGES AND FUTURE DIRECTIONS

While catalysts can be easily modeled using generated crystal structures from crystal generative models, it is very challenging to recover bulk crystals from generated catalyst structures. This makes it infeasible to obtain bulk information, which is crucial for proposing novel catalysts. Stability factors, such as formation energy and surface energy, need to be considered to identify stable candidates, and these require DFT

calculations of the bulk structures. Additionally, obtaining crystallographic information such as the space group is challenging without bulk structures. These limitations imposed by the absence of bulk highlight the need to develop strategies to recover bulk structures from surfaces for catalyst discovery. Furthermore, although the CatGPT generates new structures, it struggles to suggest new combinations outside of the training data, restricting the model's role in expanding elemental diversity.

The catalyst validity metric introduced in this work has been effective in anomaly detection but has certain limitations. The definition of catalyst anomalies is subjective, depending on individual perspectives, making the catalyst validity less universally applicable compared to more objective metrics, such as "structure validity", which is less influenced by subjectivity. For example, the current anomaly detection model identifies most porous materials as anomalies (Figure S9). Additionally, because the anomaly detection model is a deep neural network that requires GPU usage for optimal operation and fine-tuning to handle out-of-domain data, its accessibility for additional applications using CatGPT is limited. These challenges are partially addressed by using a density-based cutoff. Specifically, classifying structures with atomic densities below the 25th percentile of the ground-truth data (1.421 g/cm^3) achieved over 80% accuracy for both in-domain and out-of-domain data. This approach serves as a suitable rule of thumb for rough anomaly detection, as it does not require machine learning (Table S3 and Figures S10 and S11).

Since CatGPT is based on the GPT architecture, several optimization techniques applicable to GPT can also be utilized in CatGPT. For examples, Antunes et al.²⁵ reported that a crystal generative model based on GPT architecture could be guided to generate stable materials using Monte Carlo tree search and formation energy prediction models. Similarly, Mazuz et al.⁴⁰ optimized a molecule-generative model based on a transformer decoder to generate SMILES strings with desired properties using policy gradient reinforcement learning. We expect that these optimization strategies, combined with machine learning potentials to predict binding energies, will guide CatGPT to conditionally generate catalysts with the desired binding energies, enabling more efficient catalyst discovery.

Additionally, while CatGPT is not a model utilizing natural language, we observed that LLMs finetuned on prompts related to catalyst generation will provide functional flexibility through prompt engineering.⁴¹ For example, Gruver et al. showed that elemental substitution via the infilling task of finetuned LLMs resulted in a higher percentage of materials with desired properties compared to random substitution, indicating the potential for conditional generation with structural and compositional chemical space expansion.²⁴ Other strategies to leverage LLMs for inorganic materials can also be applied to catalyst systems.^{42,43}

5. CONCLUSIONS

While several works have progressed in developing generative language models for materials discovery, none have applied these models to catalyst surface systems. This may be due to a lack of validation methods for complex systems, which have more atoms and greater structural diversity than simple crystals, and substantial data requirements (at least 10 000

samples) for training and validating generative models on their systems.

In this work, we introduce autoregressive language models for generating catalyst structures. The results demonstrated that the catalyst generative model based on GPT-2 could generate catalysts close to real structures but struggled to generate valid structures without atomic overlapping. We addressed the overlapping issue with the bypassing method, which leverages the sequential nature of the autoregressive model without degrading its overall generative performance. Furthermore, we demonstrated adsorbate-specific generation through text-conditioning as a potential application of the CatGPT. In addition, the model can be utilized as a universal catalyst generative pretrained model for a fine-tuning strategy. With only about 1700 fine-tuning data points of 2e-ORR catalyst structures, the fine-tuned model successfully generated structures within the fine-tuning data distribution while maintaining the generative performance, suggesting novel and promising catalysts for 2e-ORR. These results suggest the potential of autoregressive language models not only as generative model for catalyst structures but also as foundation model for catalyst discovery.

6. METHODS

6.1. Representations of Catalyst Structures. Bulk crystal structures can be represented as three-dimensionally repeating systems.³¹ Similar to crystals, catalyst surface structures can be fully represented by the lattice parameters of the unit cell, atomic symbols, and respective fractional (or absolute) coordinates of atoms. Thus, the tuple C , for representing catalyst surfaces consisting of N atoms, is configured as follows:

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, y_1, z_1, \dots, e_N, x_N, y_N, z_N)$$

This includes the length (l_1, l_2, l_3) and angle ($\theta_1, \theta_2, \theta_3$) information of the unit cell lattice parameters, atomic symbols (e_i) and their fractional coordinates set (x_i, y_i, z_i) in the unit cell.

During the tokenization process, the components in tuple C are represented as string tokens within the vocabulary space, which comprises textual tokens for all atomic symbols in the periodic table, numerical tokens for the lattice parameters, and coordinates of the crystal. Among the various tokenization strategies for numbers, we employed coordinate-level tokenization, which directly represents each fraction as a distinct token, for example, “0.000”, “0.001”, ..., “1.000”. To apply these tokenization strategies to lattice parameters (lengths and angles), we divide them by 180 so that they share the same numerical space as the fractional coordinates. Since the ordering of atoms itself can hold structural information in the catalyst structures, the pairs of atomic symbols and coordinates in tuple C are arranged in the order of bulk, surface, and adsorbate. We note that atoms in the first layer, excluding adsorbates, were categorized as surface atoms, while the rest were treated as bulk atoms, following the method of ref 29.

6.2. Model Structures and Training. We used the GPT-2 architecture²⁸ from HuggingFace to generate 3D catalyst structures tokenized as components of tuple C . The tokens t_i in the sequence of tokenized structures are predicted by a categorical probability distribution $p(t_i|t_{0:i-1})$, where $t_{0:i-1}$ is the sequence of input tokens up to $i - 1$. To generate a sequence of tokenized structure x , the joint probability is modeled by the transformer as follows:

$$p(x) = \prod_{i=1}^n p(t_i|t_{0:i-1}) \quad (1)$$

Task conditioning of GPT-2 for multitask learning was not utilized in this work, so the only difference from GPT-1²⁰ model is the architecture, such as the position of layer normalization in each transformer decoder block.

For the anomaly detection model, we utilized BERT. This architecture is designed to predict masked tokens and thereby understand the inherent features of data, making it suitable for classification tasks rather than generation. To detect anomalies of incompleteness and scale mismatch, we added a regression head to the embedding of the initial token, enabling the detection of anomalies from the generated string representations of catalyst structures.

We trained both the GPT-based generative model and BERT-based detection model with 12 self-attention layers, 8 attention heads, an embedding size of 512, and a batch size of 144 on GeForce RTX 3090 GPU. Training epochs are set to 10. The other parameters for model and generation, such as temperature, top_k, and top_p, were set to their default values in HuggingFace unless otherwise specified.

6.3. Data Set Details. To train the scratch GPT-2 model to generate string representations of catalyst structures, we utilized OC20-S2EF data set that contains single-point calculated heterogeneous catalyst structures generated during the DFT relaxation.²⁹ Although there are about 250 million data points in the OC20 database, we used 2 million catalyst structures to train the CatGPT considering the training cost (1.48 h/1K batch steps on one RTX 3090 GPU). For validation, the OC20-S2EF Val-ID data set was used.

For training the BERT as an anomaly detection model, 50% of the OC20-S2EF 2M data set was randomly sampled and corrupted in two ways (incomplete structures and scale mismatches) in equal proportions. These corrupted samples were labeled as 0, and the other 50% was left unchanged and labeled as 1. The same scheme was applied to the OC20-S2EF Val-ID data set for validation of the detection model.

To fine-tune the GPT-2 model pretrained with OC20-S2EF 2M data set, we used a data set of 1,721 catalyst structures generated for screening two-electron oxygen reduction reaction catalysts in our previous studies (2e-ORR data set).²⁶ The 2e-ORR data set consists of binary alloy surface structures composed of oxophilic and oxophobic elements, featuring oxophilic active sites surrounded by oxophobic atoms (Figure S12). We randomly selected 80% of the 2e-ORR data set as the training set and 10% as the validation set to finetune the pretrained model. The remaining 10% was used as the test set to evaluate the performance of the fine-tuned model. The same corruption scheme described above was applied to the 2e-ORR data set for fine-tuning the anomaly detection model.

6.4. Metrics Details. Structural validity checks the overlapping of atoms based on atomic distances and cell volume. The cutoff values for atomic distances and cell volume to determine atomic overlap are set to 0.5 Å and 0.1 Å³, respectively. Catalyst validity is determined by the anomaly detection model, which classifies catalyst structures to be anomalous. Similar to a logistic regression model, the detection model predicts continuous values between 0 and 1 from the string representation of structures, where values closer to 1 indicate a more valid catalyst. In this work, we set the cutoff for classifying anomalies at 0.5, based on the observation that any cutoff value between 0.2 and 0.8 achieved an F_1 -score over 0.995 detecting anomalies (Figure S13). Coverage measures the similarity between generated and ground-truth structures using the pairwise distance of vectorized structures and compositions by CrystalNN and Magpie fingerprints, respectively. The CrystalNN fingerprint converts periodic structures into vectors that represent the local coordination information of sites within the structures.⁴⁴ The Magpie fingerprint converts the element stoichiometry of structures into weighted elemental properties based on stoichiometric attributions.⁴⁵ The distance cutoffs were set to 0.2 for structure and 8.5 for composition, which correspond to about 2/3 of the average pairwise distances of structure and composition vectors (0.380 and 13.356, respectively) of 10 000 randomly sampled OC20-S2EF 2M data set.

6.5. Data Augmentations. To ensure that generative models for catalyst structures maintain invariance to both unit cell rotation and spatial translation of atomic coordinates within lattice boundaries (rotation and translation invariance, respectively), we constructed an augmented training and validation data set. Specifically, 33% of the augmented data set undergoes a uniform translation of atoms in randomly selected vector directions, while another 33% undergoes

rotations within the $x-y$ plane of the unit cell at random scales. The remaining data set was left unchanged.

Ideally, the order of the atoms should not affect the generative performance of the model (permutation invariance). However, due to the nature of autoregressive language models, which generate atoms sequentially, we discovered that atom ordering significantly influences the generation outcomes. To investigate the impact of atom sequencing on generative performance, we prepared a data set wherein the atomic orderings were randomly shuffled throughout the entire data set.

6.6. Computational Details. To evaluate the 2e-ORR activity and selectivity of screened structures, spin-polarized DFT calculations were performed using the Vienna ab initio simulation package (VASP, version 5.4.4)^{46,47} with the projected-augmented wave (PAW) pseudopotential method⁴⁸ and the generalized gradient approximation revised Perdew–Burke–Ernzerhof (GGA-RPBE) exchange–correlation functional.⁴⁹ Geometry optimizations were conducted with convergence criteria of 10^{-4} eV for energy and 0.05 eV/Å for force, and a kinetic energy cutoff of 400 eV was applied. Monkhorst–Pack k -point mesh⁵⁰ was configured ($k_1 \times k_2 \times 1$) to ensure 25 Å $< a_n \times k_n$ ($n = 1, 2$) < 30 Å, where a_1 and a_2 represent the unit vector sizes in the x and y directions, respectively.

To calculate the Gibbs free energy of adsorbates, the computational hydrogen electrode (CHE) method⁵¹ was used, which equates the chemical potential of $\frac{1}{2}\text{H}_2$ (g) with $(\text{H}^+ + \text{e}^-)$ at 0 V_{RHE} under standard conditions. The Gibbs free energy of O^{*} and OOH^{*} adsorption was calculated as follows:

$$\Delta G_{\text{O}^*} = E_{\text{O}^*} - E_{\text{slab}} - E_{\text{H}_2\text{O}} - E_{\text{H}_2} + \Delta(\text{ZPE} + \int C_p \, dT - TS)$$

$$\Delta G_{\text{OOH}^*} = E_{\text{OOH}^*} - E_{\text{slab}} + 2E_{\text{H}_2\text{O}} - \frac{3}{2}E_{\text{H}_2} + \Delta(\text{ZPE} + \int C_p \, dT - TS)$$

where E_{O^*} and E_{OOH^*} are DFT energies of O^{*} and OOH^{*} adsorbed surfaces, and E_{H_2} and $E_{\text{H}_2\text{O}}$ are DFT energies of H₂ and H₂O molecules, respectively. Zero-point energies (ZPE), enthalpic contribution ($\int C_p \, dT$), and entropic contribution (TS) were calculated using the harmonic oscillator and ideal gas approximations for adsorbates (O^{*}, OOH^{*}) and gas molecules (H₂, H₂O), respectively, as implemented in atomic simulation environment (ASE).⁵² The correction values are given in Table S4. In addition, since the MLP model used in this work predicts the DFT binding energy ($\Delta E_{\text{OOH}^*,\text{ML}}$) directly, the ML predicted Gibbs free energy of the adsorption of the OOH^{*} was calculated as follows:

$$\Delta G_{\text{OOH}^*,\text{ML}} = \Delta E_{\text{OOH}^*,\text{ML}} + \Delta(\text{ZPE} + \int C_p \, dT - TS)$$

A solvation effect for OOH^{*} (-0.25 eV) was also included.⁵³ Due to the poor description of the DFT energy of O₂ by GGA functionals,⁵⁴ an experimental reaction Gibbs free energy of ORR was used instead, with ΔG_{rxn} values of -4.92 eV for O₂ + 2H₂ \rightarrow 2H₂O and -1.40 eV for O₂ + H₂ \rightarrow H₂O₂.

ASSOCIATED CONTENT

Data Availability Statement

OC20 data set can be found in <https://fair-chem.github.io/core/datasets/oc20.html>. 2e-ORR data set can be found in <https://github.com/SeoinBack/CatGPT>. The code developed in this work and relevant information can be found in Github (<https://github.com/SeoinBack/CatGPT>).

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacs.4c11504>.

Example images of invalid structures, property distributions of ground-truth and generated structures, training

loss for augmented data, atomic displacements for generated structures, and additional plots and figures for validity metrics ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Seoin Back – Department of Chemical and Biomolecular Engineering, Institute of Emergent Materials, Sogang University, Seoul 04107, Republic of Korea; orcid.org/0000-0003-4682-0621; Email: sback@sogang.ac.kr

Author

Dong Hyeon Mok – Department of Chemical and Biomolecular Engineering, Institute of Emergent Materials, Sogang University, Seoul 04107, Republic of Korea; orcid.org/0000-0001-5319-7052

Complete contact information is available at:

<https://pubs.acs.org/10.1021/jacs.4c11504>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Nano & Material Technology Development Program through the National Research Foundation of Korea (NRF), funded by Ministry of Science and ICT (Grants RS-2024-00406517 and RS-2024-00448287). D.H.M. acknowledges support from Sogang University Research Grant of 2024 (202412020.01).

REFERENCES

- Hautier, G.; Jain, A.; Ong, S. P. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.* **2012**, *47*, 7317–7340.
- Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; et al. Accelerated discovery of CO₂ electrocatalysts using active machine learning. *Nature* **2020**, *581* (7807), 178–183.
- Alberi, K.; Nardelli, M. B.; Zakutayev, A.; Mitas, L.; Curtarolo, S.; Jain, A.; Fornari, M.; Marzari, N.; Takeuchi, I.; Green, M. L.; et al. The 2019 materials by design roadmap. *J. Phys. D Appl. Phys.* **2019**, *52* (1), 013001.
- Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2018**, *2* (4), 0121.
- Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533* (7601), 73–76.
- Noh, J.; Gu, G. H.; Kim, S.; Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **2020**, *11* (19), 4871–4881.
- Li, Z.; Achenie, L. E.; Xin, H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catal.* **2020**, *10* (7), 4377–4384.
- Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO₂ reduction and H₂ evolution. *Nat. Catal.* **2018**, *1* (9), 696–703.
- Kim, J.; Mok, D. H.; Kim, H.; Back, S. Accelerating the Search for New Solid Electrolytes: Exploring Vast Chemical Space with Machine Learning-Enabled Computational Calculations. *ACS Appl. Mater. Interfaces* **2023**, *15* (45), 52427–52435.
- Yohannes, A. G.; Lee, C.; Talebi, P.; Mok, D. H.; Karamad, M.; Back, S.; Siahrostami, S. Combined high-throughput DFT and ML screening of transition metal nitrides for electrochemical CO₂ reduction. *ACS Catal.* **2023**, *13* (13), 9007–9017.

- (11) Mok, D. H.; Li, H.; Zhang, G.; Lee, C.; Jiang, K.; Back, S. Data-driven discovery of electrocatalysts for CO₂ reduction using active motifs-based machine learning. *Nat. Commun.* **2023**, *14* (1), 7303.
- (12) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361* (6400), 360–365.
- (13) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4* (4), 828–849.
- (14) Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, 2204.06125.
- (15) Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. *arXiv* **2023**, 2306.15687.
- (16) Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv* **2024**, 2402.17177.
- (17) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, 1312.6114.
- (18) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63* (11), 139–144.
- (19) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**; NeurIPS, 2020; pp 6840–6851.
- (20) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *Preprint* **2018**, 1–12.
- (21) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, 1706.03762 (accessed Nov 12, 2024).
- (22) Flam-Shepherd, D.; Aspuru-Guzik, A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv* **2023**, 2305.05708.
- (23) Born, J.; Manica, M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mach. Intell.* **2023**, *5* (4), 432–444.
- (24) Gruber, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. Fine-tuned language models generate stable inorganic materials as text. *arXiv* **2024**, 2402.04379.
- (25) Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *arXiv* **2023**, 2307.04340.
- (26) Back, S.; Na, J.; Ulissi, Z. W. Efficient discovery of active, selective, and stable catalysts for electrochemical H₂O₂ synthesis through active motif screening. *ACS Catal.* **2021**, *11* (5), 2483–2491.
- (27) Mok, D. H.; Back, S.; Siahrostami, S. Validating ΔΔG Selectivity Descriptor for Electrosynthesis of H₂O₂ from Oxygen Reduction Reaction. *Angew. Chem.* **2024**, *136*, No. e202404677.
- (28) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1* (8), 9.
- (29) 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* (10), 6059–6072.
- (30) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, 1810.04805.
- (31) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv* **2021**, 2110.06197.
- (32) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Skelton, J. M.; Morita, K.; Walsh, A. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software* **2019**, *4* (38), 1361.
- (33) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.
- (34) Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; Jaakkola, T. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems* **34**; NeurIPS, 2021; pp 13757–13769.
- (35) Rubner, Y.; Tomasi, C.; Guibas, L. J. A metric for distributions with applications to image databases. *Sixth International Conference on Computer Vision*; IEEE, 1998; pp 59–66 (IEEE Catalog No. 98CH36271).
- (36) Simonovsky, M.; Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. *Proceedings, Artificial Neural Networks and Machine Learning-ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018*; Springer, 2018; Part I 27, pp 412–422.
- (37) 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.
- (38) Liao, Y.-L.; Wood, B.; Das, A.; Smidt, T. Equiformer2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv* **2023**, 2306.12059.
- (39) Van der Maaten, L.; Hinton, G.; et al. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9* (11), 2579–2605.
- (40) Mazuz, E.; Shtar, G.; Shapira, B.; Rokach, L. Molecule generation using transformers and policy gradient reinforcement learning. *Sci. Rep.* **2023**, *13* (1), 8799.
- (41) Kang, Y.; Kim, J. ChatMof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **2024**, *15* (1), 4705.
- (42) Choudhary, K. AtomGPT: Atomistic Generative Pretrained Transformer for Forward and Inverse Materials Design. *J. Phys. Chem. Lett.* **2024**, *15*, 6909–6917.
- (43) Jia, S.; Zhang, C.; Fung, V. LLMatDesign: Autonomous Materials Discovery with Large Language Models. *arXiv* **2024**, 2406.13163.
- (44) Zimmermann, N. E.; Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **2020**, *10* (10), 6063–6081.
- (45) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comp. Mater.* **2016**, *2* (1), 16028.
- (46) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6* (1), 15–50.
- (47) Kresse, G.; Hafner, J. Ab initio molecular dynamics for open-shell transition metals. *Phys. Rev. B* **1993**, *48* (17), 13115.
- (48) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59* (3), 1758.
- (49) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B* **1999**, *59* (11), 7413.
- (50) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13* (12), 5188.
- (51) Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L.; Kitchin, J. R.; Bligaard, T.; Jonsson, H. Origin of the overpotential for oxygen reduction at a fuel-cell cathode. *J. Phys. Chem. B* **2004**, *108* (46), 17886–17892.
- (52) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dufak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29* (27), 273002.
- (53) Zhang, Q.; Asthagiri, A. Solvation effects on DFT predictions of ORR activity on metal surfaces. *Catal. Today* **2019**, *323*, 35–43.

(54) Klüpfel, S.; Klüpfel, P.; Jónsson, H. The effect of the Perdew-Zunger self-interaction correction to density functionals on the energetics of small molecules. *J. Chem. Phys.* **2012**, *137* (12), 124102.