

Screening Cu-Zeolites for Methane Activation Using Curriculum-Based Training

Jiawei Guo, Tyler Sours, Sam Holton, Chenghan Sun, and Ambarish R. Kulkarni*



Cite This: *ACS Catal.* 2024, 14, 1232–1242



Read Online

ACCESS |



Metrics & More



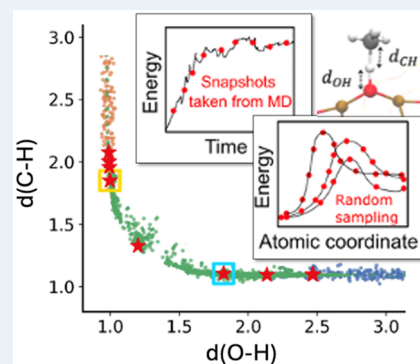
Article Recommendations



Supporting Information

ABSTRACT: Machine learning (ML), when used synergistically with atomistic simulations, has recently emerged as a powerful tool for accelerated catalyst discovery. However, the application of these techniques has been limited by the lack of interpretable and transferable ML models. In this work, we propose a curriculum-based training (CBT) philosophy to systematically develop reactive machine learning potentials (rMLPs) for high-throughput screening of zeolite catalysts. Our CBT approach combines several different types of calculations to gradually teach the ML model about the relevant regions of the reactive potential energy surface. The resulting rMLPs are accurate, transferable, and interpretable. We further demonstrate the effectiveness of this approach by exhaustively screening thousands of $[\text{CuOCu}]^{2+}$ sites across hundreds of Cu-zeolites for the industrially relevant methane activation reaction. Specifically, this large-scale analysis of the entire International Zeolite Association (IZA) database identifies a set of previously unexplored zeolites (i.e., MEL, ATN, EWO, and CAS) that show the highest ensemble-averaged rates for $[\text{CuOCu}]^{2+}$ -catalyzed methane activation. We believe that this CBT philosophy can be generally applied to other zeolite-catalyzed reactions and, subsequently, to other types of heterogeneous catalysts. Thus, this represents an important step toward overcoming the long-standing barriers within the computational heterogeneous catalysis community.

KEYWORDS: curriculum-based training (CBT), Cu-exchanged zeolites, methane activation, reactive machine learning potential (MLP), heterogeneous catalysis, high throughput screening, density functional theory (DFT)



Developing catalysts for the selective and efficient conversion of methane to methanol continues to be a grand challenge for the heterogeneous catalysis community.^{1,2} Among the several types of catalysts that have been explored, industrially relevant materials such as Cu- and Fe-exchanged zeolites have received significant attention.^{2–5} These studies have been motivated by the hope of finding an elusive active site motif stabilized within a “Goldilocks” zeolite topology that shows high methanol selectivity at reasonable methane conversions.^{6–12} Recent studies have shown consistent improvements in the reported methanol yields as high as 200 $\mu\text{mol/g}$ for MAZ.⁵ Despite these advances, however, only a small fraction of the possible zeolite material space remains experimentally explored. As the synthesis and testing of hundreds of zeolite topologies and their possible variations (e.g., Si/Al ratio, Al distributions, etc.) is cumbersome and prohibitively expensive, computational screening can serve as an efficient tool to accelerate this search.^{13–16}

Indeed, descriptor-based analyses using density functional theory (DFT) have become an integral part of the heterogeneous catalysis research community.^{17–20} These approaches, often accelerated by machine learning (ML) models, have been successfully used to screen large catalyst libraries for several industrially relevant reactions.^{21–24} Within the context of zeolite-catalyzed methane activation, DFT has

also provided valuable mechanistic insights.^{6,8–10,25} But these studies are typically limited to comparing a few zeolites or a handful of active site motifs. Thus, to the best of our knowledge, the large-scale screening of zeolite-based catalysts has not been reported for any reaction.

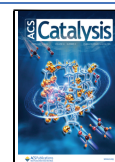
The key bottleneck is that the current state-of-the-art computational approaches, which rely on DFT-based barrier calculations or approximate descriptor-based analyses, are not well suited for the large-scale screening of zeolites. For example, identifying the “best” $[\text{CuOCu}]^{2+}$ motif—a known active site for methane activation—requires detailed analysis of 15,255 unique $[\text{CuOCu}]^{2+}$ sites across 231 fully connected IZA topologies (obtained through the MAZE package,²⁶ details in SI Section 1 and Figure S3). These sites span a range of Cu–Cu distances, Cu–O–Cu angles, Al–Al distances and vary in their local confining environment (Figures S3, S5, S6, Table S1). This suggests that descriptor-based approaches

Received: November 2, 2023

Revised: December 15, 2023

Accepted: December 18, 2023

Published: January 10, 2024



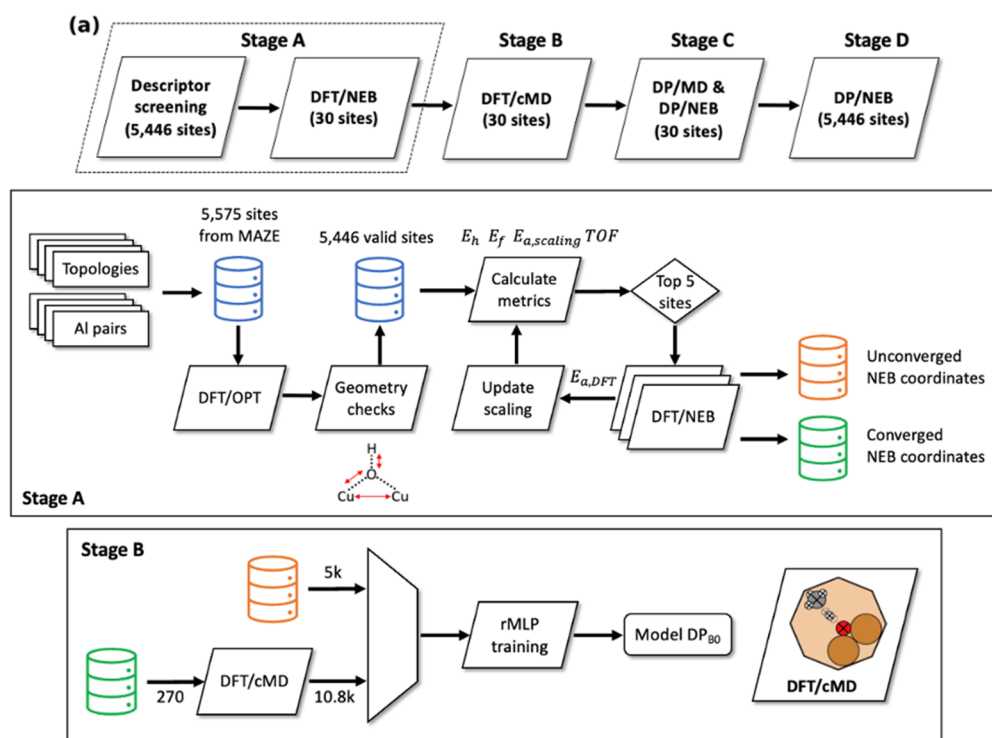


Figure 1. (a). Overview of the four-stage multistage active learning algorithm. Flowchart for training **Stage A**: Descriptor-based initial site sampling (30 sites from 27 topologies) and DFT/NEB calculations. Detailed structure generation and sampling criteria can be found in the SI, [Sections 1 and 2](#). Flowchart for training **Stage B**: Initial rMLP model training using DFT/NEB and DFT/cMD sampled configurations. Color scheme: Cu (brown), O (red), H (white), C (gray).

are likely to be less reliable and transferable (vide infra), and detailed DFT-based nudged elastic band calculations (denoted as DFT/NEB) will be necessary. Thus, large scale screening of zeolite-based catalysts for any reaction remains inaccessible to the heterogeneous catalysis community. We note that the potential limitations of descriptor-based strategies have been discussed previously.^{39,40}

The central challenge is related to the computational cost of using DFT to identify transition states and calculate the reaction barriers in zeolites. For example, while force field-based methods have been widely used to screen large libraries of metal–organic frameworks and zeolites for adsorption or diffusion-based separations,^{13,19,27,28} such interatomic potentials are generally unsuitable for describing reactive phenomena. On the other hand, using DFT to develop reactive potentials is also challenging. Although considerable strides have been made in using machine learning based potentials (MLPs) to calculate adsorption energies or to accelerate molecular dynamics simulations,^{29–32} an analogous approach for describing reactions requires an efficient algorithm for sampling of the “reactive” region of the potential energy surface (PES).^{33,34} But as reactions are rare event phenomena, the relevant portion of PES is not known *a priori*. Although this information can be obtained using a DFT/NEB calculation, this approach defeats the purpose of developing a surrogate ML model. Thus, in our opinion, overcoming the above “chicken-or-egg” issue remains the central bottleneck in developing cheap and accurate ML-based alternatives to traditional DFT-based barrier calculations.

As a step toward addressing this broad challenge, we now report a novel curriculum-based training (CBT) strategy for developing reactive machine learning-based potentials

(rMLPs) capable of describing reactions in zeolites. Specifically, we present a multistage active learning algorithm that combines several different types of calculations to gradually but efficiently “teach” the ML model about the relevant regions of the PES across different zeolite topologies. Importantly, we show that such a progressive training approach has significant implications for quantifying the interpretability, confidence, and transferability of nominally “black box” ML models. We demonstrate the latter aspect by applying our model to the entire IZA database³⁵ and a few hypothetical zeolite topologies³⁶ not included in the initial training process. We emphasize that active learning is now widely used within the molecular modeling community to accelerate the development of highly accurate machine learning potentials, providing opportunities for future chemical reaction kinetics studies.^{41–44}

Second, from a practical catalyst screening perspective, our rMLP replaces expensive DFT/NEBs without any appreciable loss in accuracy. We illustrate this by using the previously reported $[\text{CuOCu}]^{2+}$ motif as a prototypical active site for methane activation.^{7,12} Specifically, we calculated the transition state geometries and the C–H bond activation energies for thousands of distinct $[\text{CuOCu}]^{2+}$ active sites across 52 zeolite topologies at DFT accuracy. Our predictions are within 0.07 eV of the DFT computed energy barriers and are completed within minutes.

Finally, the ability to explicitly calculate the reaction barriers for zeolite catalyzed reactions has important implications for the larger nanoporous materials community. Whereas current theoretical studies rely on comparing one or a few active sites or topologies, this approach allows us to compare the performance of *ensembles of all possible active sites* across 192 different zeolites. By rescaling our computationally predicted

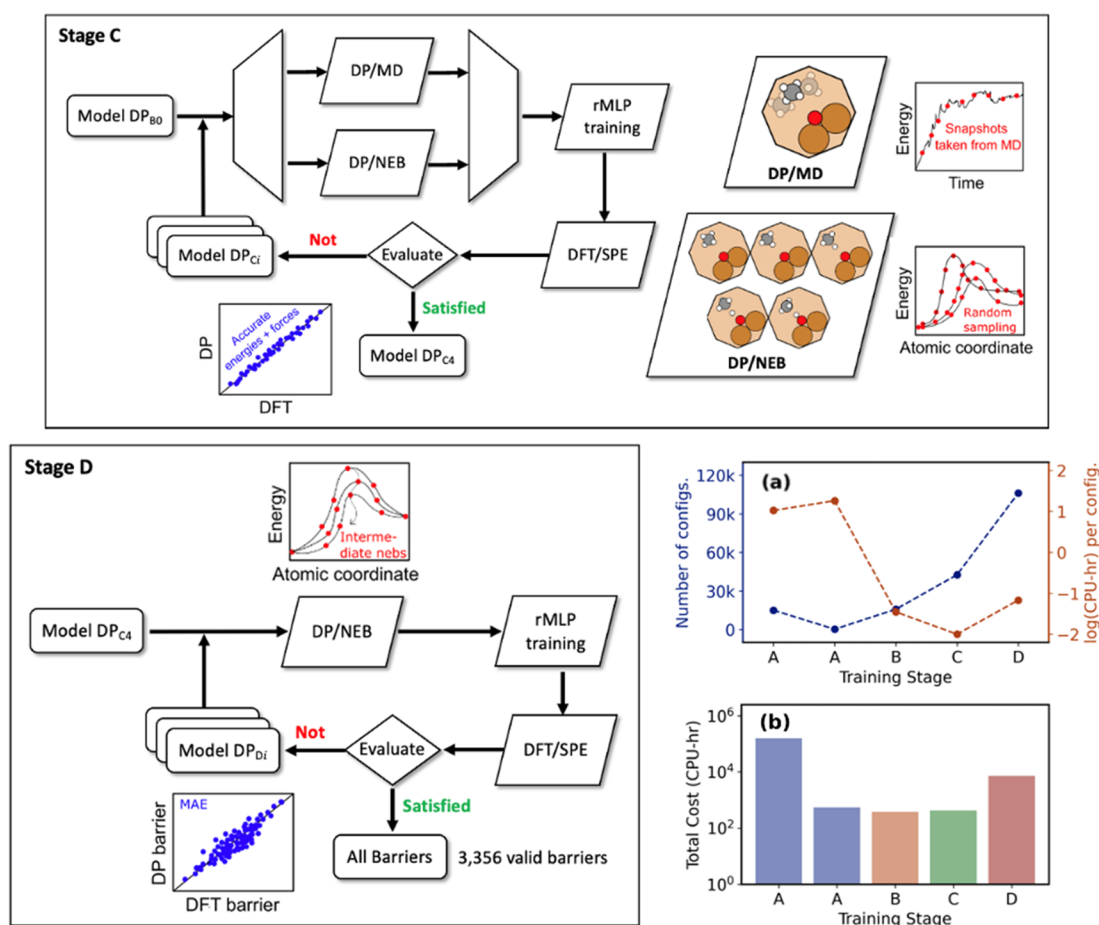


Figure 2. Flowchart for training **Stage C**: rMLP model training for the 30 initial sites using configurations sampled from DP only (DP/MD and DP/NEB). Flowchart for training **Stage D**: Extend the rMLP model training to all possible $[\text{CuOCu}]^{2+}$ sites using DP/NEB. (a). The entire workflow balances the computational cost of the calculations (initially expensive, few structures) with the configurational diversity (cheap calculations but many more structures) to efficiently sample the potential energy surface (PES). (b). Overall computational cost at each training stage.

rates with known experimental measurements, we have identified a set of previously unexplored zeolite structures (i.e., MEL, ATN, EWO, and CAS) that show the highest ensemble-averaged performance for methane activation. Although this work has focused on the methane activation reaction, we anticipate that similar strategies can be applied to other zeolite-catalyzed reactions and, subsequently, to other types of heterogeneous catalysts.

CURRICULUM-BASED TRAINING

The overall goal of this work is to accelerate the discovery of promising Cu-exchanged zeolites for methane activation by explicitly calculating the C–H activation barriers for all possible $[\text{CuOCu}]^{2+}$ active sites that can be formed within each topology (Figures S1–S3, S8). We chose to explore 52 zeolite topologies obtained from the IZA database (detailed in Table S1). Our training data set includes several well-studied materials (e.g., CHA, MOR, MFI, etc.; Figure S4) as well as other topologies that offer a high density and diversity of possible $[\text{CuOCu}]^{2+}$ sites within the smallest simulation volumes (Figure S5). Taken together, the resulting CuZEO23 database consists of 3,912 1Al sites (i.e., Cu^+ , $[\text{CuOH}]^+$) and 5,575 2Al sites (i.e., $[\text{CuOCu}]^{2+}$, $[\text{CuOH-Cu}]^{2+}$) across 52 topologies. Further details are presented in the SI, Figures S7–S10.

We used an active learning approach to efficiently identify the best $[\text{CuOCu}]^{2+}$ motif for C–H bond activation within the 52 topologies considered here. As shown in Figure 1a, our workflow is divided into four distinct stages, which aims to maximize the configurational diversity of our data set while reducing the computational cost at each stage. For example, in **Stage A**, we combine geometry-based screening, DFT-calculated quantities, known descriptors for C–H bond activation energy (e.g., hydrogen binding energy), and detailed DFT/NEB calculations (i.e., 5 per iteration) to recursively update the scaling relations (Figure 1). After 6 iterations, Stage A provides a set of 30 promising $[\text{CuOCu}]^{2+}$ sites across 27 topologies. These DFT-based transition state (TS) geometries and reaction coordinates serve as the initial seeds for developing an rMLP that can be generalized to any zeolite. The first step of the training curriculum is discussed next.

In **Stage B**, the above 30 $[\text{CuOCu}]^{2+}$ sites are used to generate the necessary DFT data for training the first iteration of our rMLP (Figure 1). Specifically, we use DFT to perform constrained MD simulations (denoted as DFT/cMD). Here, the positions of CH_4 (or $\text{CH}_3 - \text{H}$) and the oxygen of the $[\text{CuOCu}]^{2+}$ site are frozen. The resulting DFT/cMD configurations are combined with a subset of unconverged DFT/NEB snapshots (from **Stage A**) to train an initial rMLP. We use a deep neural network potential as implemented in

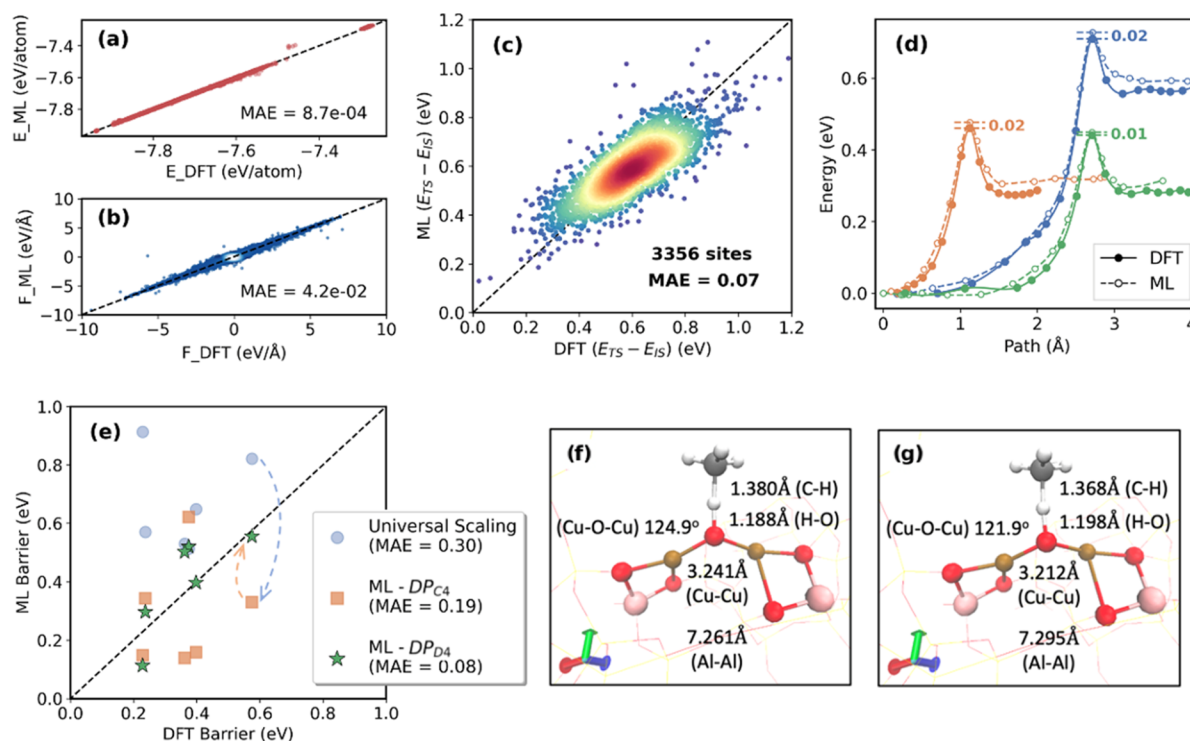


Figure 3. (a–c). Parity plots comparing ML-predicted (a) energies per atom, (b) forces per dimensions, and (c) $(E_{TS} - E_{IS})$, with DFT-calculated values for data sets not included in the training pool. (d). ML(DP_{D4})-predicted minimum energy pathways (MEPs) vs DFT-calculated MEPs (blue: HEU, green: MOR, orange: RUT) using 17 NEB images. (e). Parity plot comparing the performance of universal scaling with the DP models in predicting reaction barriers for six randomly selected validation sites from CHA, HEU, JSW, MFI, MOR, and RUT, using 5 images for DFT/NEBs and 17 images for DP/NEBs. (f–g). Transition state geometries predicted by ML(DP_{D4}) (f) vs DFT (g), taking CHA as an example. Coloring: Si in the background (yellow), Al (pink), Cu (brown), O (red), H (white), C (gray).

DeePMD-kit.^{30,31} We denote the resulting rMLP as DP_{B0}, where DP indicates the ML model used in this study (i.e., DeePMD-kit) and the subscript indicates the zeroth iteration of the model obtained at **Stage B** of the overall protocol. The rMLP training details are presented in the SI, [Section 4](#), [Figure S11–S17](#), [Tables S2–S4](#).

Following **Stage B**, our algorithm utilizes only DFT-based single point energy calculations (denoted as DFT/SPE); no further DFT/MDs or DFT/NEBs are necessary. This part of the training curriculum exposes the rMLP to different regions of the PES. For example, **Stage C** uses an iterative loop that combines rMLP-based molecular dynamics (denoted as DP/MD) with rMLP-based nudged elastic band calculations (denoted as DP/NEB) to expand the training data set while simultaneously improving the accuracy of the model ([Figure 2](#)). Here, the DP/MD simulations improve the model performance close to the initial and final state regions of the PES, while the DP/NEB targets the reactive region of the process. **Stage C** is terminated after 4 iterations when the resulting rMLP (i.e., DP_{C4}) achieves DFT accuracy for the 30 sites identified previously. The above sampling procedure aims to minimize the number of DFT/NEB calculations (due to their high computational costs) necessary to train a potential. Although we recognize that other sampling strategies or hyperparameter optimizations can improve the efficacy of the algorithm described above, such detailed studies are beyond the scope of this work.

Stage D, the final active learning step shown in [Figure 2](#), is central to this study. While the training curriculum in **Stages B** and **C** is limited to the initial 30 sites, we now expand the sampled configuration space to the entire data set of 5,446

[CuOCu]²⁺ sites. Specifically, owing to the rigorous training curriculum used above (i.e., several iterations at each stage), DP_{C4} and later models (DP_{D1–D4}) are sufficiently reliable for executing the entire NEB workflow, including initial and final state optimizations (i.e., DP/OPT), for previously unseen sites and topologies. Thus, the intermediate configurations obtained from DP/NEBs are “good enough” to be directly used for DFT/SPE calculations for any [CuOCu]²⁺ site in any zeolite. (In contrast, using **Stage B** models for unseen topologies will result in unphysical geometries.)

Note that the above workflow relies on rapidly and reliably exploring the configuration space using cheap rMLP runs (i.e., DP/NEB, DP/MD, etc.) This is followed by expensive but easily parallelizable DFT/SPE calculations to “label” a smaller subset of sampled configurations. To ensure that labeled configurations are both physically relevant and sufficiently dissimilar from the data already seen by the model, we use built-in uncertainty metrics available within DeePMD-Kit. Specifically, the configurations chosen for DFT/SPE lie within the ϵ_t range of 0.05 and 0.4 eV/Å. Here, ϵ_t measures the uncertainty of the rMLP-predicted forces across 4 distinct DP models trained on the same data set.

Taken together, the entire model development procedure includes 180,700 DFT-calculated configurations obtained using scaling relationships, SPEs, MDs, CMDs, and NEBs. At first glance, the overall training protocol may seem overly complicated. However, the actual implementation is relatively straightforward, as the individual steps are automated and can be parallelized. Specifically, [Figures 2a](#) and [2b](#) show the number of configurations sampled at each stage, the cost of each calculation, and the overall computational cost of the

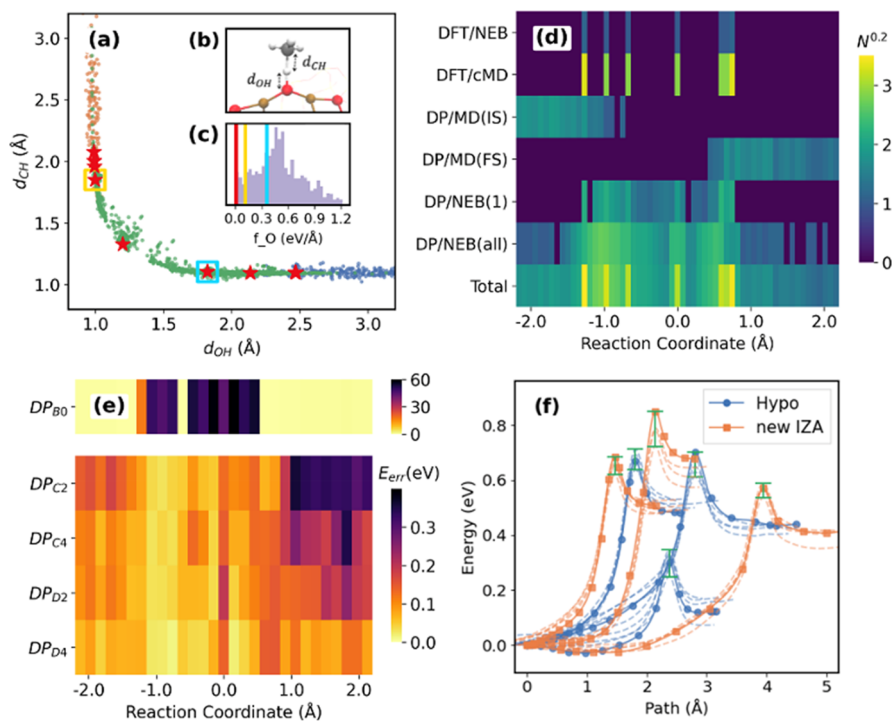


Figure 4. (a). Reaction coordinate described by distances: d_{OH} and d_{CH} (labeled in (b)), including all MOR configurations sampled throughout the entire training protocol shown in Figures 1 and 2. (c). Distribution of forces on the oxygen atom of $[CuOCu]^{2+}$ sites (denoted as f_O) from all MOR configurations sampled using DFT/cMD. Vertical lines highlight f_O sampled from DFT/NEB at training Stage A. (d). Heatmap showing the number of MOR configurations sampled at each stage of the training. (e). Heatmap showing the energy error across the DP model at different training stages. (f). Model performance on previously unseen IZA topologies and hypothetical zeolites.

training process. We now demonstrate the performance of the DP_{D4} model in predicting DFT-calculated properties at a fraction of the computational cost.

MODEL PERFORMANCE AND THE LIMITATIONS OF DESCRIPTOR-BASED APPROACHES FOR ZEOLITES

Figure 3 summarizes the performance of the DP_{D4} model for our test data set. The parity plots in Figures 3a and 3b show that DP_{D4} -predicted per-atom energies and forces (x -, y -, and z -components) are in excellent agreement with the DFT-calculated values. We observe impressive mean absolute errors (MAEs) of 0.87 meV/atom and 42 meV/Å for energies and forces, respectively. Here, the test set is obtained using 10% of the overall configurations sampled from both the local minima and the reactive regions; these configurations have not been used during model development.

Building on our ability to reproduce DFT energies and forces, we now show that DP_{D4} can be used to accurately predict C–H activation energies across thousands of $[CuOCu]^{2+}$ sites. Figure 3c compares DP_{D4} predicted barriers with DFT calculated values for 3,356 sites across 52 topologies with an overall MAE of 0.07 eV (detailed in the SI, Figures S12–S13). Due to the high cost of performing full DFT/NEBs for thousands of sites, the DFT barriers shown in Figure 3c are obtained using DFT/SPE on DP-optimized initial, final, and transition states geometries.

A more stringent test of the model is to compare the entire reaction coordinate obtained by using independent DFT and DP_{D4} calculations. To ensure an unbiased comparison, we randomly selected three $[CuOCu]^{2+}$ sites from three different topologies (Figures S14–S16) and use identical stopping

criteria (SI Section 3) for both DFT/and DP/NEB. As shown in Figure 3d for one $[CuOCu]^{2+}$ site in MOR, almost identical minimum energy paths (MEPs) are obtained for DFT and DP_{D4} . Note that a sufficiently high density of intermediate images (Figure 3d uses 17 images) is necessary to obtain a good resolution for a DP/NEB calculation (Figure S17).

Encouraged by the accuracy (0.07 eV MAE, Figure 3c) and computational efficiency (100× faster, Figure 2b, Table S2) of the DP_{D4} model, we now evaluate the efficacy of the traditional descriptor-based philosophy for zeolites. This is illustrated in Figure 3e using six randomly chosen $[CuOCu]^{2+}$ sites across six topologies that were not included in Stage A. Specifically, using hydrogen affinity as a descriptor to predict activation energies, we observe a significantly higher MAE (0.3 eV) for the descriptor approach compared to that of the DP_{D4} predictions (0.08 eV MAE for DP_{D4} /NEB). Interestingly, the intermediate DP_{C4} model, which is trained only on 30 sites (from Stage A), shows an MAE that lies between the two approaches (0.19 eV).

Second, unlike descriptor-based strategies that only provide energy estimates, rMLPs also provide the entire reaction coordinate and transition state geometries that can be used for further analysis. For example, using CHA as an example, we show that the transition state geometries predicted by the DP_{D4} model (Figure 3f) are nearly identical to those obtained using DFT (Figure 3g). The entire data set of DP_{D4} calculated transition state geometries is provided in the SI.

Before proceeding further, we note that more than 2,000 sites resulted in physically unrealistic TS geometries during DP_{D4} /NEBs. This is because either the confining environment of the zeolite was too small or the $[CuOCu]^{2+}$ motif was inaccessible to methane for those sites. We emphasize that this

important aspect is not captured when H binding energy is used as a descriptor.

Thus, in addition to demonstrating the capabilities of the DP_{D4} model, the above analysis suggests that caution must be used while adsorption energy-based descriptors are used to predict reaction barriers in zeolites. We now discuss how the model learns, how confident is it in its predictions, and how it can be generalized to unseen configurations.

MODEL INTERPRETABILITY, CONFIDENCE, AND TRANSFERABILITY

The three questions posed above are all closely related to the question of the interpretability of rMLPs. Here, we define “model interpretability” as being able to quantitatively answer three specific questions: (1) what has the model learned from the training data that it has seen, (2) how reliable are its predictions, and (3) can the model be applied to a system that it has not seen before. As discussed in this section, the multistage active learning protocol can be used to address all three questions.

To answer these questions, we first examine what the model has “learned” from the training data supplied during the various stages in the CBT protocol. As an illustrative example, we used a single [CuOCu]²⁺ site that was part of the initial 30 DFT/NEBs. Analogous to the idea of collective variables used in rare event sampling methods, we use d_{OH} and d_{CH} to describe the reaction coordinate. Specifically, the red stars in Figure 4a show the d_{OH} and d_{CH} distances (Figure 4b) corresponding to each image of the converged DFT/NEB for the chosen [CuOCu]²⁺ site.

Using these configurations as a starting point, Stage B uses DFT/cMD runs to explore the “nearby” PES. Although the relative positions of the CH₄ (or CH₃–HO) remain unchanged (as the d_{OH} and d_{CH} are constrained), DFT/cMD is a key step for stabilizing the model. Specifically, as shown by the force histogram in Figure 4c, this iterative protocol samples a wider range of forces on the oxygen of the [CuOCu]²⁺ site than what was available during DFT/NEB. Thus, Stage B of the training curriculum “teaches” the model about the PES corresponding to nonequilibrium values of d_{OH} and d_{CH} . Additionally, this step also samples the dynamics of the zeolite backbone, as shown previously by Sours et al.³⁷

The impact of the subsequent steps is also shown in Figure 4a. Specifically, Stage C samples the initial states (IS, blue) and final states (FS, orange) using DP/MD, as well as the reactive region, using DP/NEB (orange). Subsequently, this approach is extended to all possible [CuOCu]²⁺ sites within the chosen topology (MOR) in Stage D. Here, the uncertainty-based sampling procedure discussed above “fills-in” the remaining sections of the reaction coordinate (green). In summary, Figure 4a shows how the chosen curriculum progressively exposes the model to the relevant PES for C–H activation while avoiding unphysical configurations that are far away from the reaction coordinate (e.g., the top right region of Figure 4a).

We now show that each step of our training approach samples a different but important region of the PES. The heat map in Figure 4d shows the number of configurations (denoted as N) sampled at each stage of the training algorithm. The reaction coordinate (i.e., the x-axis of Figure 4d) is defined as the distance (measured by d_{OH} and d_{CH}) of each configuration in Figure 4a from the corresponding values for the transition state. Thus, from left to right, Figure 4d shows the IS region, TS region, and FS region. Similar to the

discussion above, Figure 4d quantitatively shows that starting from the DFT/NEB images for a single [CuOCu]²⁺ site within the MOR topology, the model is “taught” to fill the gaps across the entire reaction coordinate for all [CuOCu]²⁺ sites across the 52 topologies (final row of Figure 4d).

An analogous approach is used to plot the data in Figure 4e. Here, the heat map shows the mean absolute errors (compared with DFT) from models obtained during the different stages of the training process. As model DP_{B0} (row #1) was trained using DFT/cMD of the 30 initial NEBs, large errors are observed in the TS region (shown by purple shades). On the other hand, as several initial DFT/NEB images (red stars in Figure 4a) lie closer to the IS or the FS, smaller errors are observed in the IS and FS regions.

Next, compared to the short but expensive DFT/cMD (0.1 ps) used above, Stage C uses longer DP/MD runs (50 ps) and several DP/NEB iterations to efficiently sample the PES. This improved training and sampling protocol, which is enabled by a reasonably DP_{B0} model, rapidly increases the accuracy of the models. By the end of the training protocol, DP_{D4} model shows errors < 0.2 eV for all MOR configurations across the entire reaction coordinate. Interestingly, this MAE value is higher than the MAE of activation energies for MOR (0.05 eV). This is because DP_{D4} predictions are more accurate (using DFT as the reference) for configurations which are located closer to the IS, FS, or TS (*vide infra*).

This type of progressive error analysis also provides important insights about the subtle learning process. For example, by comparing the bottom left and top right regions of Figure 4e, we generally observe lower errors for IS configurations (lighter colors, bottom left) than the FS (darker colors, top right). Deeper analysis suggests that this phenomenon arises due to the higher rotational degrees of freedom of the methane molecule compared to the methyl radical. Simply stated, the methane molecule explores the configuration space of relevant configurations more quickly, which results in the superior performance of the ML model in the IS region (*vide infra*).

We now turn to addressing the other two questions, which are related to the transferability and reliability of our model. To explore these aspects, we use the DP_{D4} model to calculate methane activation barriers for [CuOCu]²⁺ sites using three IZA topologies (SZR, EMT, and LTL) and three hypothetical zeolites. As shown in Figure 4f (solid lines), we obtain physically realistic reaction coordinates and transition state geometries (detailed in SI, Figures S33–S38). Furthermore, to illustrate the model confidence, we use the ensemble of penultimate models (DP_{D3,i}, dashed lines) and compare the discrepancy in their predictions with the final DP_{D4} model. These ensembles show a high degree of self-consistency (<0.1 eV), especially for configurations closer to the transition state.

Thus, the above analysis emphasizes the interpretability, confidence, and transferability of the DP_{D4} model in reproducing DFT-quality PES. We now revisit our original motivation—the challenge of identifying the best Cu-zeolite for methane activation.

STABILITY AND ACTIVITY OF ENSEMBLE OF ACTIVE SITES

Following Latimer and others,⁹ the overall rate of methane activation depends on the interplay between the stability (denoted as E_{β} , eq 1) and reactivity of the active sites. Specifically, an inverse correlation between these two

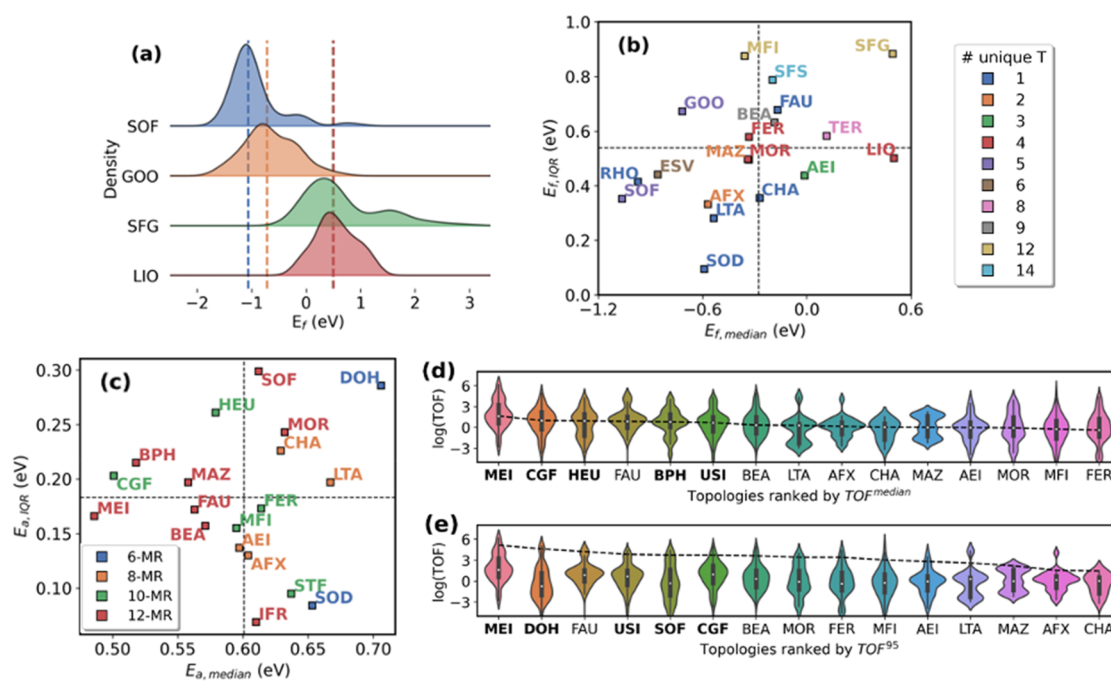


Figure 5. (a) E_f histograms of four zeolites with distinct shapes of distributions, including all valid $[\text{CuOCu}]^{2+}$ sites within each topology. Quadrant plots showing median and interquartile range (denoted as IQR) of the (b) E_f distributions for $E_{f,median}$ below 0.6 eV (color-coded by the number of unique T-sites) and (c) ML-predicted barrier (E_a) distributions (color-coded by the maximum window size). Only the top 10 most studied and the few outlier topologies are shown. The complete data set can be found in the SI, Figures S18–26, Tables S5–S6. Vertical dashed lines highlight the median of the distributions (SFG median overlaps with LIO). (d–e) Violin plots ranking zeolite performance by (d) median TOF (TOF^{median}) and (e) TOF at the 95th percentile (TOF^{95}). As a comparison, we focus on the stepwise methane activation process involving activation at 450 °C with 1 atm O_2 , reaction at 80 °C, and extraction with water. Results for the higher reaction temperature at 200 °C (also seen in the literature) are summarized in SI Figure S32. Only the top 10 most studied topologies and the top 5 others (highlighted in bold fonts) are shown; the complete data set is available in Figures S29–S31.

quantities results in a volcano relationship for the turnover frequency (eq 4). In the context of zeolite catalyzed methane activation, the framework topology controls the properties of the $[\text{CuOCu}]^{2+}$ sites that can be stabilized within the pores.

$$E_f = E_{\text{CuOCu}} + E_{\text{silica}} - \left(E_{Z_1\text{Cu}} + E_{Z_2\text{Cu}} + \frac{1}{2}E_{\text{O}_2} \right) \quad (1)$$

$$E_h = E_{\text{CuOHCu}} - E_{\text{CuOCu}} - \frac{1}{2}E_{\text{H}_2\text{O}} + \frac{1}{4}E_{\text{O}_2} \quad (2)$$

$$E_{a,scaling} = 0.75E_h + 1.09 \quad (3)$$

$$TOF = \frac{\sqrt{P_{\text{O}_2}} e^{-G_f/k_B T}}{1 + \sqrt{P_{\text{O}_2}} e^{-G_f/k_B T}} \frac{k_B T}{h} e^{-G_a/k_B T} \quad (4)$$

As an illustrative example, Figure 5a shows the distributions of formation energies of all possible $[\text{CuOCu}]^{2+}$ sites across four different zeolite topologies obtained using DFT. We observe that the SOF topology is more successful at stabilizing $[\text{CuOCu}]^{2+}$ sites than LIO. More generally, the formation energies of most $[\text{CuOCu}]^{2+}$ sites lie within -2.4 to 2.4 eV. However, very few topologies show similar shapes for the formation energy histograms.

These trends are further quantified by plotting the median values of the distribution ($E_{f,median}$) versus the interquartile range (IQR, $E_{f,IQR}$) in Figure 5b. Thus, even though similar median stabilities are observed for both SFG and LIO, the higher IQR for SFG (0.88 eV) indicates that a greater diversity of $[\text{CuOCu}]^{2+}$ sites can be formed compared to LIO (IQR =

0.50 eV). We also observe a weak trend where zeolites with a fewer unique T-sites (e.g., SOD, LTA) populate the lower left region indicating that smaller T-site diversity results in more stable and more uniform sites. On the other hand, a greater diversity of unique T-sites (e.g., MFI, SFS, SFG) results in more diverse $[\text{CuOCu}]^{2+}$ sites. Analogous analysis for the hydrogen binding energy (E_h , eq 2) and the scaling-predicted C–H activation energy ($E_{a,scaling}$, eq 3) can be found in the SI, Sections 5.1 and 5.2, respectively. Although we find that the previously reported scaling relation between E_f and E_h holds true, significant deviations are observed (Figure S28).

Next, we examine how the framework topology impacts the distribution of activation energies for all possible $[\text{CuOCu}]^{2+}$ molecules that can be formed within a given zeolite. Using a similar approach as above, Figure 5c shows a quadrant plot for the DP_{D4} calculated methane activation energies ($E_{a,DPD4}$) for 3,356 sites across 52 zeolite topologies. Although the correlation is weaker, it is interesting that 12 MR zeolites (e.g., MEI, FAU, MAZ, BPH, and BEA) show the lowest median activation energies. This suggests that on average 12 MR zeolites can form more active $[\text{CuOCu}]^{2+}$ sites than other smaller or larger ring zeolites. As mentioned previously, ca. 2,000 $[\text{CuOCu}]^{2+}$ sites that result in unphysical TS geometries have been excluded.

ENSEMBLE-AVERAGE RATES OF METHANE ACTIVATION

We now turn to our original goal of comparing the rates of methane activation across different zeolite frameworks. As the ensembles of possible $[\text{CuOCu}]^{2+}$ sites yield a distribution of

stabilities and activation energies, the application of eq 4 for each site also results in a distribution of C–H activation rates for each zeolite. Thus, we use two metrics: (1) the median TOF (TOF^{median}) and (2) the TOF at the 95th percentile (TOF^{95}) for comparing the different framework. These metrics are valuable as precise identification of the AI distributions remains experimentally challenging.

The two TOF metrics are plotted as violin plots in Figures 5d and 5e for the top 15 zeolites considered here. This representation also captures the distribution of the TOFs for all viable $[CuOCu]^{2+}$ sites. Both metrics suggest that MEI is the top performer. This analysis also highlights considerable differences in the distributions of the TOFs across the various topologies. More importantly, since we now have a complete database of DFT-quality TOFs for all valid $[CuOCu]^{2+}$ sites across 52 zeolites, we can extend this analysis to make experimentally relevant predictions. Specifically, as discussed in Section 5.2 of the SI, we now scale our theory-predicted TOF^{median} for CHA using the experimentally measured rates for Cu-Chabazite (0.0057 s^{-1} at $80 \text{ }^\circ\text{C}$).¹²

The experimentally rescaled rates are summarized in Figure 6 using the above two TOF metrics. Among the initial data set

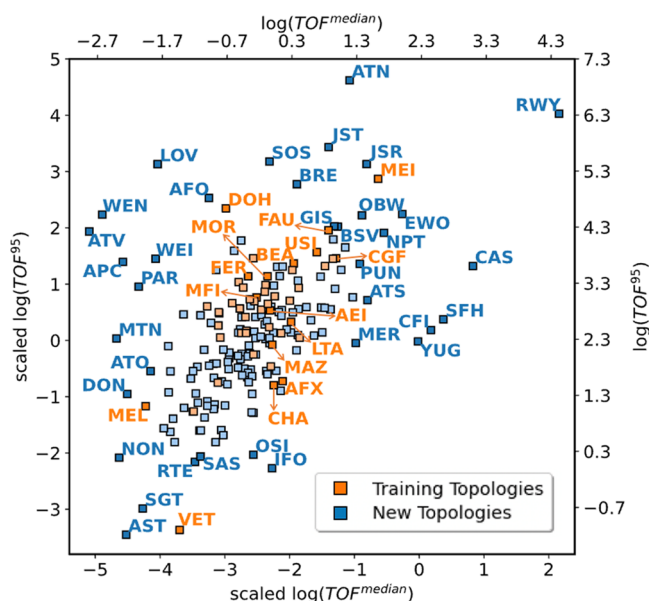


Figure 6. Scatter plot of TOF^{median} and TOF^{95} on the top and right two axes, evaluated across 192 different IZA topologies (52 included in the training set in orange, 150 not included in the rMLP training in blue). Topologies with small primitive cells of less than 50 atoms are excluded from the analysis due to lower model accuracy. The bottom and left axes are corresponding TOF values scaled from the experimentally obtained rate constant. The most well-studied zeolites and outlier topologies are highlighted (darker color) and labeled in the figure. A constant E_f of -0.7 eV is used for the TOF evaluations of all new IZA sites.

of 52 topologies, our analysis shows that known aluminosilicates such as MEI and FAU zeolites are among the top 20 candidates. In contrast, smaller pore materials such as CHA and MAZ do not form stable $[CuOCu]^{2+}$ sites. These trends, which assume that only the $[CuOCu]^{2+}$ site is responsible for the catalytic activity, are broadly consistent with previous reports.⁸ The underlying caveats, key assumptions, and potential showstoppers associated with this theory-guided recommendation are discussed in SI, Section 5.3.

Finally, we extended this approach to screen all possible $[CuOCu]^{2+}$ sites across 192 experimentally reported zeolite topologies contained within the IZA database. As summarized in Figure 6, we identify several other structures (e.g., ATN, RWY, JST, JSR, EWO, and CAS) that appear promising, a subset of which has been reported in their all-silica, aluminosilicate, aluminophosphate, or germanosilicate analogs. Thus, in addition to MEI, we suggest these materials (i.e., ATN, EWO, and CAS) as potential candidates for further experimental testing. Although initially motivated by the direct methane to methanol conversion reaction, we emphasize that our catalytic activity results also serve as a guide to develop nonplatinum group metal catalysts the abatement of low-level methane emissions.³⁸ We emphasize that this work provides a useful tool to screen the catalytic activity of a single type of Cu site within aluminosilicate zeolites. Current work in our team is focused on developing related approaches that can address other important factors, such as product selectivity and site stability.

CONCLUSION

To the best of our knowledge, this work represents the first example of using machine learning-based potentials to screen a large library of zeolite catalysts for any reaction. Specifically, we have proposed a new philosophy called curriculum-based training (CBT) to systematically develop reactive machine learning potentials (rMLPs) for describing reactions. Our multistage active learning approach combines several different types of calculations to gradually teach the ML model about the relevant regions of the potential energy surface. Despite the black box nature of the underlying ML model, the CBT philosophy is advantageous, as it can be used to quantify the interpretability, confidence, and transferability of the rMLP. From a practical catalyst screening perspective, the resulting rMLP replaces expensive DFT calculations without sacrificing accuracy. This was demonstrated by exhaustive screening of a large library of Cu-zeolites for the industrially relevant methane activation reaction. Importantly, our analysis identified a set of previously unexplored zeolite structures that show the highest ensemble-averaged rates for methane activation. Although this study has focused on a specific system, we believe that the CBT philosophy can be generally applied to other zeolite-catalyzed reactions and, subsequently, to other types of heterogeneous catalysts. Thus, this work represents an important step toward overcoming the long-standing barriers within the computational heterogeneous catalysis community.

METHODS

Stage A: Descriptor-Based Initial Site Sampling and DFT/NEB Calculations. In Stage A, we use DFT-calculated energies of the various Cu species (obtained by brute force DFT calculations) to identify the initial set of promising sites for C–H activation. A ranking system that uses four metrics (i.e., E_f , E_h , $E_{a,scaling}$, and TOF) is used to select the top 5 sites from an initial set of 5,446 configurations. Detailed descriptions of these metrics and underlying assumptions can be found in the SI, Section 2.4. Then we perform traditional DFT/NEB calculations to obtain the “correct” methane activation energy ($E_{a,DFT}$). The true DFT activation energy is used to develop a new scaling relationship, which is then used to re-evaluate the $E_{a,scaling}$ of all 5,446 sites. The new site rankings (since $E_{a,scaling}$ is updated) are used to identify the next

set of top 5 sites for DFT/NEB calculations. This process is repeated 6 times until a set of 30 DFT-calculated reaction coordinates, transition state geometries, and activation energies is identified. A fraction of unconverged reaction coordinates (~150 configurations per site), obtained during the NEB convergence steps, is reused for the model development in Stage B.

Stage B: Initial rMLP Model Training. At Stage B, the set of 30 promising $[\text{CuOCu}]^{2+}$ sites identified at Stage A (across 27 different topologies) is used to generate the necessary training data for rMLP development. Specifically, a set of 270 configurations obtained from the 30 converged NEB calculations is used to run constrained MD simulations (0.5 fs time step, 200 fs total simulation time steps, 298 K). Here, we freeze the positions of CH_4 (or $\text{CH}_3 - \text{H}$) and the oxygen of the $[\text{CuOCu}]^{2+}$ site; this ensures that the algorithm samples configurations in the neighborhood of the NEB images. A total of 10,800 configurations are obtained from these constrained MD. These are combined with the 5,000 configurations from unconverged DFT NEBs to train an initial rMLP using a deep neural network potential as implemented in DeepMD-kit. Details of the rMLP training are presented in the SI, Section 4. Beyond this stage, our algorithm requires only DFT-based single point energy calculations (denoted as DFT/SPE). No further DFT/MDs or DFT/NEBs are used.

Stage C: rMLP Model Training on the 30 Initial Sites. While Stage B used constrained DFT/MD and DFT/NEBs to generate the necessary training data, Stage C uses the rMLP developed above (i.e., DP_{B0}) to obtain additional relevant configurations. Specifically, to ensure that the entire “length” of the reaction coordinate is well sampled, we use DP/MD and DP/NEB. These two approaches sample different regions of the PES. While DP/MD mainly samples the local minima of the PES (e.g., IS: CH_4 and FS: $\text{CH}_3 - \text{H}$), the DP/NEBs sample configurations at various stages of the C–H bond breaking (and O–H bond formation) process.

The DP/MD is performed using LAMMPS using an approach similar to that for DP-GEN. Briefly, an ensemble of four models is trained. Each model has an identical architecture but uses different random seeds to initialize the neural net parameters. One of these four models is used to run a 50 ps NVT-MD (0.5 fs time step), and the temperature ramped from 298 to 500 K. This provides a diverse set of configurations close to the initial state (IS: $\text{CH}_4 + [\text{CuOCu}]^{2+}$) and final state (FS: $\text{CH}_3 + [\text{CuOHCu}]^{2+}$) of the system. The DP/NEB is performed using the Atomic Simulation Environment (ASE). Several strategies are used to induce randomness and avoid biased sampling. For example, our DP/NEB iterations use different numbers of intermediate NEB images (ranging from 11 to 21) and the total number of NEB iterations. Also, the initial and final states of the 30 DFT-derived NEB trajectories are “rattled” and then interpolated; this approach adds diversity to the “initial guesses” (i.e., the starting points) of the reaction coordinates for DP/NEBs.

New training configurations sampled from both DP/MD and DP/NEB are selected by evaluating the uncertainty of the force predictions from the ensemble of models (denoted as ϵ_f , detailed in the SI, Section 4.1, Figure S11). We performed DFT/SPE for configurations where the ϵ_f parameter lies between 0.05 and 0.4 eV/Å; this ensures that sampled configurations are physically relevant and also sufficiently dissimilar from the existing data. As ϵ_f measures the uncertainty of predictions between the four ensembles of

models, this metric has been shown to be correlated to the errors between the DFT forces and the DP prediction.

The above two approaches are used iteratively to expand the training data set while also improving the accuracy of the model. Across 4 iterations within Stage C, we added a total of 42,600 configurations – 36,400 near the initial and the final states (from DP/MD) and 6,100 sampled from the DP-NEBs (Table S2). The model refinement loop is stopped when the model-predicted energy and forces show near-DFT accuracy for the 30 sites identified in Stage A, with energy MSE of 2.36×10^{-1} eV and force MSE of 4.42×10^{-3} eV/Å. The MSEs of the sequentially obtained models, which show systematic improvements, are presented in Table S3. Although the training protocol may seem overly complicated, the actual implementation is relatively straightforward (as the individual steps are automated) and computationally efficient (as the DFT/SPEs can be parallelized).

Stage D: Extend the rMLP Model Training to All Possible $[\text{CuOCu}]^{2+}$ Sites. The final active learning step, Stage D, is central to this study. In this training block, we expand the active site configuration space from the initial 30 promising sites to all possible 5,446 sites across 52 different topologies. Additionally, due to the training data obtained from DP/MD (Stage C), the model can now perform geometric optimization of the initial and final states for the previously unseen $[\text{CuOCu}]^{2+}$ sites. More importantly, the DP-optimized initial and final state geometries are sufficiently realistic to initiate an NEB calculation that relies entirely on the DP_{C4} PES. To address the increased diversity of the PES and achieve more efficient sampling considering the scope of the active site space (from 30 sites to 5,446), we adopt a new sampling scheme for the DP/NEBs. We extract unconverged reaction coordinates from different stages of the convergence (20%, 60%, and the converged coordinate). This methodology, used for each active learning iteration in Stage D, requires one NEB run per site instead of multiple NEB runs per site (as in Stage C). In addition, we also include reaction coordinates at 40% and 80% from the final convergence into the training set for sites with more complicated PESs or where the initial guess transition state is further from the “true” transition state geometry as these sites tend to require more NEB steps to converge. Similar to the strategy in Stage C, the new configurations are evaluated with DFT/SPE and selectively add into the training pool based on ϵ_f . The above approach is necessary since multiple DP/NEBs and DFT/SPEs (used for 30 sites in Stage C) are not feasible for 5,446 sites. Restated simply, while Stage C repeatedly samples the DP/NEB converged reaction coordinate for 30 sites to improve the model (i.e., DP_{C1} to DP_{C4}), Stage D also includes the configurations encountered while approaching the converged reaction coordinate for 5,446 sites.

In addition to comparing energies and forces obtained from the DP model to DFT, we also calculate the activation energy barrier of each $[\text{CuOCu}]^{2+}$ site using DFT (Figure S12). This is calculated using DFT/SPE of the initial and the transition states (i.e., $E_{\text{TS}} - E_{\text{IS}}$), where the geometries are obtained from the current DP model. The E_a parity plots in Figure S12 for models DP_{C4} to DP_{D3} (The final DP_{D4} model is shown in the main text, Figure 2c.) review systematic but gradually slowing improvements in the MAE, which is an indication of no additional gain from the training. A similar trend is also observed in ϵ_f uncertainty, where negligible improvement was observed from DP_{D2} to DP_{D3} (Figure S11). In this stage,

106,000 new configurations are included in the training (Table S2).

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acscatal.3c05275>.

Detailed discussion of the computational methods and assumptions, Figures S1–S38, Tables S1–S6. The resulting ML model, demo scripts, and final databases are available online at https://github.com/JiaweiGuo-0313/DPNEB_CuZeo_demo, <https://drive.google.com/drive/folders/1Cr8BXbAMtf-N2vHxpcRjKEd75zWDeGpp>, and <https://colab.research.google.com/drive/16gibtUmkB37Uu54nn3-7AMZIQnMkGL3S> (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ambarish R. Kulkarni – Department of Chemical Engineering, University of California, Davis, California 95616, United States; orcid.org/0000-0001-9834-8264; Email: arkulkarni@ucdavis.edu

Authors

Jiawei Guo – Department of Chemical Engineering, University of California, Davis, California 95616, United States;

orcid.org/0000-0001-9373-4297

Tyler Sours – Department of Chemical Engineering, University of California, Davis, California 95616, United States

Sam Holton – Department of Chemical Engineering, University of California, Davis, California 95616, United States

Chenghan Sun – Department of Chemical Engineering, University of California, Davis, California 95616, United States

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acscatal.3c05275>

Author Contributions

J.G. performed the DFT calculations, analyzed the results, and wrote the original draft. J.G., T.S., and A.R.K. designed the active learning algorithm. J.G. and A.R.K. prepared the final manuscript. All authors contributed to the discussion, provided suggestions, and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The development of the CBT algorithm was supported by NSF #2048260. J.G. further acknowledges partial support from a Chevron Fellowship. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award 2022ERCAP0021791.

■ REFERENCES

- (1) Ravi, M.; Ranocchiari, M.; van Bokhoven, J. A. The direct catalytic oxidation of methane to methanol—a critical assessment. *Angew. Chem., Int. Ed.* **2017**, *56* (52), 16464–16483.
- (2) Ravi, M.; Sushkevich, V. L.; Knorpp, A. J.; Newton, M. A.; Palagin, D.; Pinar, A. B.; Ranocchiari, M.; van Bokhoven, J. A. Misconceptions and challenges in methane-to-methanol over transition-metal-exchanged zeolites. *Nature Catalysis* **2019**, *2* (6), 485–494.
- (3) Park, M. B.; Ahn, S. H.; Mansouri, A.; Ranocchiari, M.; van Bokhoven, J. A. Comparative Study of Diverse Copper Zeolites for the Conversion of Methane into Methanol. *ChemCatChem* **2017**, *9* (19), 3705–3713.
- (4) Liu, L.; Corma, A. Confining isolated atoms and clusters in crystalline porous materials for catalysis. *Nature Reviews Materials* **2021**, *6* (3), 244–263.
- (5) Newton, M. A.; Knorpp, A. J.; Sushkevich, V. L.; Palagin, D.; van Bokhoven, J. A. Active sites and mechanisms in the direct conversion of methane to methanol using Cu in zeolitic hosts: A critical examination. *Chemical Society Reviews* **2020**, *49* (5), 1449–1486.
- (6) Alayon, E. M. C.; Nachttegaal, M.; Bodi, A.; Ranocchiari, M.; van Bokhoven, J. A. Bis(μ -oxo) versus mono(μ -oxo)dicopper cores in a zeolite for converting methane to methanol: An in situ XAS and DFT investigation. *Phys. Chem. Chem. Phys.* **2015**, *17* (12), 7681–7693.
- (7) Vanelderden, P.; Snyder, B. E. R.; Tsai, M. L.; Hadt, R. G.; Vancauwenbergh, J.; Coussens, O.; Schoonheydt, R. A.; Sels, B. F.; Solomon, E. I. Spectroscopic definition of the copper active sites in mordenite: Selective methane oxidation. *J. Am. Chem. Soc.* **2015**, *137* (19), 6383–6392.
- (8) Kulkarni, A. R.; Zhao, Z. J.; Siahrostami, S.; Nørskov, J. K.; Studt, F. Monocopper Active Site for Partial Methane Oxidation in Cu-Exchanged 8MR Zeolites. *ACS Catal.* **2016**, *6* (10), 6531–6536.
- (9) Latimer, A. A.; Kulkarni, A. R.; Aljama, H.; Montoya, J. H.; Yoo, J. S.; Tsai, C.; Abild-Pedersen, F.; Studt, F.; Nørskov, J. K. Understanding trends in C-H bond activation in heterogeneous catalysis. *Nat. Mater.* **2017**, *16* (2), 225–229.
- (10) Zhao, Z. J.; Kulkarni, A.; Vilella, L.; Nørskov, J. K.; Studt, F. Theoretical Insights into the Selective Oxidation of Methane to Methanol in Copper-Exchanged Mordenite. *ACS Catal.* **2016**, *6* (6), 3760–3766.
- (11) Snyder, B. E. R.; Vanelderden, P.; Schoonheydt, R. A.; Sels, B. F.; Solomon, E. I. Second-Sphere Effects on Methane Hydroxylation in Cu-Zeolites. *J. Am. Chem. Soc.* **2018**, *140* (29), 9236–9243.
- (12) Rhoda, H. M.; Plessers, D.; Heyer, A. J.; Bols, M. L.; Schoonheydt, R. A.; Sels, B. F.; Solomon, E. I. Spectroscopic Definition of a Highly Reactive Site in Cu-CHA for Selective Methane Oxidation: Tuning a Mono- μ -Oxo Dicopper(II) Active Site for Reactivity. *J. Am. Chem. Soc.* **2021**, *143* (19), 7531–7540.
- (13) Lin, L.-C.; Berger, A. H.; Martin, R. L.; Kim, J.; Swisher, J. A.; Jariwala, K.; Rycroft, C. H.; Bhowan, A. S.; Deem, M. W.; Haranczyk, M.; Smit, B. In silico screening of carbon-capture materials. *Nat. Mater.* **2012**, *11* (7), 633–641.
- (14) Pahls, D. R.; Ortuño, M. A.; Winegar, P. H.; Cramer, C. J.; Gagliardi, L. Computational Screening of Bimetal-Functionalized Zr6O8MOF Nodes for Methane C-H Bond Activation. *Inorg. Chem.* **2017**, *56* (15), 8739–8743.
- (15) Xie, P.; Pu, T.; Aranovich, G.; Guo, J.; Donohue, M.; Kulkarni, A.; Wang, C. Bridging adsorption analytics and catalytic kinetics for metal-exchanged zeolites. *Nature Catalysis* **2021**, *4* (2), 144–156.
- (16) Felvey, N.; Guo, J.; Rana, R.; Xu, L.; Bare, S. R.; Gates, B. C.; Katz, A.; Kulkarni, A. R.; Runnebaum, R. C.; Kronawitter, C. X. Interconversion of atomically dispersed platinum cations and platinum clusters in zeolite ZSM-5 and formation of platinum gem-dicarbonyls. *J. Am. Chem. Soc.* **2022**, *144* (30), 13874–13887.
- (17) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **2009**, *1* (1), 37–46.

- (18) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Structure-Activity Relationships That Identify Metal-Organic Framework Catalysts for Methane Activation. *ACS Catal.* **2019**, *9* (4), 3576–3587.
- (19) Sun, Y.; Rogge, S. M.; Lemaire, A.; Vandenbrande, S.; Wieme, J.; Siviour, C. R.; Van Speybroeck, V.; Tan, J.-C. High-rate nanofluidic energy absorption in porous zeolitic frameworks. *Nat. Mater.* **2021**, *20* (7), 1015–1023.
- (20) Xu, L.; Papanikolaou, K. G.; Lechner, B. A.; Je, L.; Somorjai, G. A.; Salmeron, M.; Mavrikakis, M. Formation of active sites on transition metals through reaction-driven migration of surface atoms. *Science* **2023**, *380* (6640), 70–76.
- (21) Kitchin, J. R. Machine learning in catalysis. *Nature Catalysis* **2018**, *1* (4), 230–232.
- (22) Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; Sun, S.; Voznyy, O.; Tan, C.-S.; Askerka, M.; Che, F.; Liu, M.; Seifitokaldani, A.; Pang, Y.; Lo, S.-C.; Sargent, E. H.; et al. Accelerated discovery of CO₂ electrocatalysts using active machine learning. *Nature* **2020**, *581* (7807), 178–183.
- (23) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **2021**, *11* (10), 6059–6072.
- (24) Tran, R.; Lan, J.; Shuaibi, M.; Wood, B. M.; Goyal, S.; Das, A.; Heras-Domingo, J.; Kolluru, A.; Rizvi, A.; Shoghi, N.; Sriram, A.; Therrien, F.; Abed, J.; Voznyy, O.; Sargent, E. H.; Ulissi, Z.; Zitnick, C. L. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **2023**, *13* (5), 3066–3084.
- (25) Li, C.; Ferri, P.; Paris, C.; Moliner, M.; Boronat, M.; Corma, A. Design and synthesis of the active site environment in zeolite catalysts for selectively manipulating mechanistic pathways. *J. Am. Chem. Soc.* **2021**, *143* (28), 10718–10726.
- (26) Antonio, D. D.; Guo, J.; Holton, S. J.; Kulkarni, A. R. Simplifying computational workflows with the Multiscale Atomic Zeolite Simulation Environment (MAZE). *SoftwareX* **2021**, *16*, 100797.
- (27) Verploegh, R. J.; Kulkarni, A.; Boulfelfel, S. E.; Haydak, J. C.; Tang, D.; Sholl, D. S. Screening diffusion of small molecules in flexible zeolitic imidazolate frameworks using a DFT-parameterized force field. *J. Phys. Chem. C* **2019**, *123* (14), 9153–9167.
- (28) Goeminne, R.; Vanduyfhuys, L.; Van Speybroeck, V.; Verstraelen, T. DFT-quality adsorption simulations in metal-organic frameworks enabled by machine learning potentials. *J. Chem. Theory Comput.* **2023**, *19* (18), 6313–6325.
- (29) Behler, J. Perspective: Machine learning potentials for Atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
- (30) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.
- (31) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; E, W. DP-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **2020**, *253*, 107206.
- (32) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **2021**, *20* (6), 750–761.
- (33) Polino, D.; Parrinello, M. Kinetics of aqueous media reactions via ab initio enhanced molecular dynamics: The case of urea decomposition. *J. Phys. Chem. B* **2019**, *123* (31), 6851–6856.
- (34) Schreiner, M.; Bhowmik, A.; Vegge, T.; Busk, J.; Winther, O. Transition1x - A dataset for building generalizable reactive machine learning potentials. *Scientific Data* **2022**, *9* (1), 779.
- (35) Baerlocher, Ch.; McCusker, L. B. Database of Zeolite Structures. <http://www.iza-structure.org/databases/>. Accessed 13 December 2023.
- (36) Treacy, M.; Foster, M. Atlas of Prospective Zeolite Structures. <http://www.hypotheticalzeolites.net>. Accessed 13 December 2023.
- (37) Sours, T. G.; Kulkarni, A. R. Predicting structural properties of pure silica zeolites using deep neural network potentials. *J. Phys. Chem. C* **2023**, *127* (3), 1455–1463.
- (38) Brenneis, R. J.; Johnson, E. P.; Shi, W.; Plata, D. L. Atmospheric- and low-level methane abatement via an Earth-abundant catalyst. *ACS Environmental Au* **2022**, *2* (3), 223–231.
- (39) Rebollar, L.; Intikhab, S.; Oliveira, N. J.; Yan, Y.; Xu, B.; McCrum, I. T.; Snyder, J. D.; Tang, M. H. Beyond adsorption descriptors in hydrogen electrocatalysis. *ACS Catal.* **2020**, *10* (24), 14747–14762.
- (40) Andersen, M.; Levchenko, S. v.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9* (4), 2752–2759.
- (41) Schreiner, M.; Bhowmik, A.; Vegge, T.; Jørgensen, P. B.; Winther, O. NeuralNEB—neural networks can find reaction paths fast. *Machine Learning: Science and Technology* **2022**, *3*, 045022.
- (42) Kulichenko, M.; Barros, K.; Lubbers, N.; Li, Y. W.; Messerly, R.; Tretiak, S.; Smith, J. S.; Nebgen, B. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science* **2023**, *3* (3), 230–239.
- (43) Brezina, K.; Beck, H.; Marsalek, O. Reducing the Cost of Neural Network Potential Generation for Reactive Molecular Systems. *J. Chem. Theory Comput.* **2023**, *19* (19), 6589–6604.
- (44) Yoo, P.; Sakano, M.; Desai, S.; Islam, M. M.; Liao, P.; Strachan, A. Neural network reactive force field for C, H, N, and O systems. *Npj Computational Materials* **2021**, *7* (1), 9.