nature communications

Article

https://doi.org/10.1038/s41467-025-63860-x

Resolving chemical-motif similarity with enhanced atomic structure representations for accurately predicting descriptors at metallic interfaces

Received: 5 January 2025

Accepted: 25 August 2025

Published online: 01 October 2025

Check for updates

Cheng Cai [1,2,3] & Tao Wang [1,2,3]

Accurately predicting catalytic descriptors with machine learning (ML) methods is significant to achieving accelerated catalyst design, where a unique representation of the atomic structure of each system is the key to developing a universal, efficient, and accurate ML model that is capable of tackling diverse degrees of complexity in heterogeneous catalysis scenarios. Herein, we integrate equivariant message-passing-enhanced atomic structure representation to resolve chemical-motif similarity in highly complex catalytic systems. Our developed equivariant graph neural network (equivGNN) model achieves mean absolute errors <0.09 eV for different descriptors at metallic interfaces, including complex adsorbates with more diverse adsorption motifs on ordered catalyst surfaces, adsorption motifs on highly disordered surfaces of high-entropy alloys, and the complex structures of supported nanoparticles. The prediction accuracy and easy implementation attained by our model across various systems demonstrate its robustness and potentially broad applicability, laying a reasonable basis for achieving accelerated catalyst design.

Heterogeneous catalysis plays a crucial role in achieving energy sustainability and implementing chemical production processes, where the design of high-performance catalysts is an appealing research area for both experimentalists and theorists. In computational catalysis, descriptor-based high-throughput computational screenings have successfully identified promising catalysts for multiple essential reactions[1]. However, the high computational costs make whole chemical space screening challenging today. In this case, the surrogate machine learning (ML) model has been introduced to accelerate computational catalyst screening by replacing ab initio calculations with accurately predicted descriptor values at low costs[2]. Indeed, a reliable representation of the atomic structure, the numerical input used to describe chemical systems of interest, has been regarded as a crucial basis for developing good ML models[3,4]. Moreover, it is widely accepted that a robust representation of an atomic structure should possess the following merits[5]: (1) it should be applicable to the structures of the entire material domain of interest; (2) it should be more easily accessible and easier to compute than the target property; and (3) it should be capable of accurately reflecting and distinguishing the (dis)similarity between two similar structures.

In computational catalysis scenarios, the binding energies of the important intermediates on catalyst surfaces, which are derived from density functional theory (DFT) calculations, are widely used as descriptors to predict the activity and selectivity trends among

[1]Center of Artificial Photosynthesis for Solar Fuels and Department of Chemistry, School of Science and Research Center for Industries of the Future, Westlake University, 600 Dunyu Road, Hangzhou, Zhejiang Province, China. [2]Division of Solar Energy Conversion and Catalysis at Westlake University, Zhejiang Baima Lake Laboratory, Hangzhou, Zhejiang Province, China. [3]Institute of Natural Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou, Zhejiang Province, China. e-mail: twang@westlake.edu.cn

different catalysts[6]. However, the chemical motifs of diverse adsorbates on various material-based catalysts exhibit different levels of complexity, ranging from the prototype model with simple monodentate adsorbates on pure metal surfaces to complex catalytic systems, i.e., high entropy alloys (HEAs) and supported nanoparticles (Fig. 1). Apparently, atomic structure representations suffer from the challenge of completeness[7–10] as the complexity of chemical motifs increases, particularly in resolving chemical-motif similarity.

As illustrated in Fig. 1a, predicting the binding energies from the initial structures with simple adsorbates in monodentate adsorption motifs on ordered metal and alloy surfaces is a classical ML task in computational catalysis[11–28]. For example, a convolutional neural network (CNN) model using ab initio features derived from the electronic density of states (DOSnet)[20] demonstrated remarkable ML performance, with a mean absolute errors (MAE) of 0.10 eV across 11 diverse adsorbates. However, obtaining such ab initio features typically involves a high computational burden. Thus, ML models that use non-ab initio features are appealing. For example, models employing labeled site representations with non-ab initio features, along with crystal graph convolution neural networks[29] (CGCNN) and SchNet[30], achieved good prediction performances, with MAEs of 0.116 and 0.085 for CO* and H*, respectively[17]. Notably, an active ML model with a coordinate-based representation successfully accelerated the process of discovering active electrocatalysts for CO_2 reduction[14,19]. With respect to the complex adsorbates included in bi- and higher-dentate adsorption motifs on ordered metal surfaces[31–35], resolving the chemical-motif similarity is a challenging task if two distinct bidentate adsorption motifs have the same coordination environments. As illustrated in Fig. 1b, the bidentate adsorption motifs of the C*C*H and N*N*H intermediates on the hcp- and fcc-hollow adsorption sites of ordered metal surfaces are two specific examples.

Indeed, resolving chemical-motif similarity for chemically complex materials is an even more challenging task. The representative examples are high-entropy alloys (right side in Fig. 1c), which are typically composed of five or more principal elements and have unique active sites on highly disordered surfaces[36–42]. More specifically,

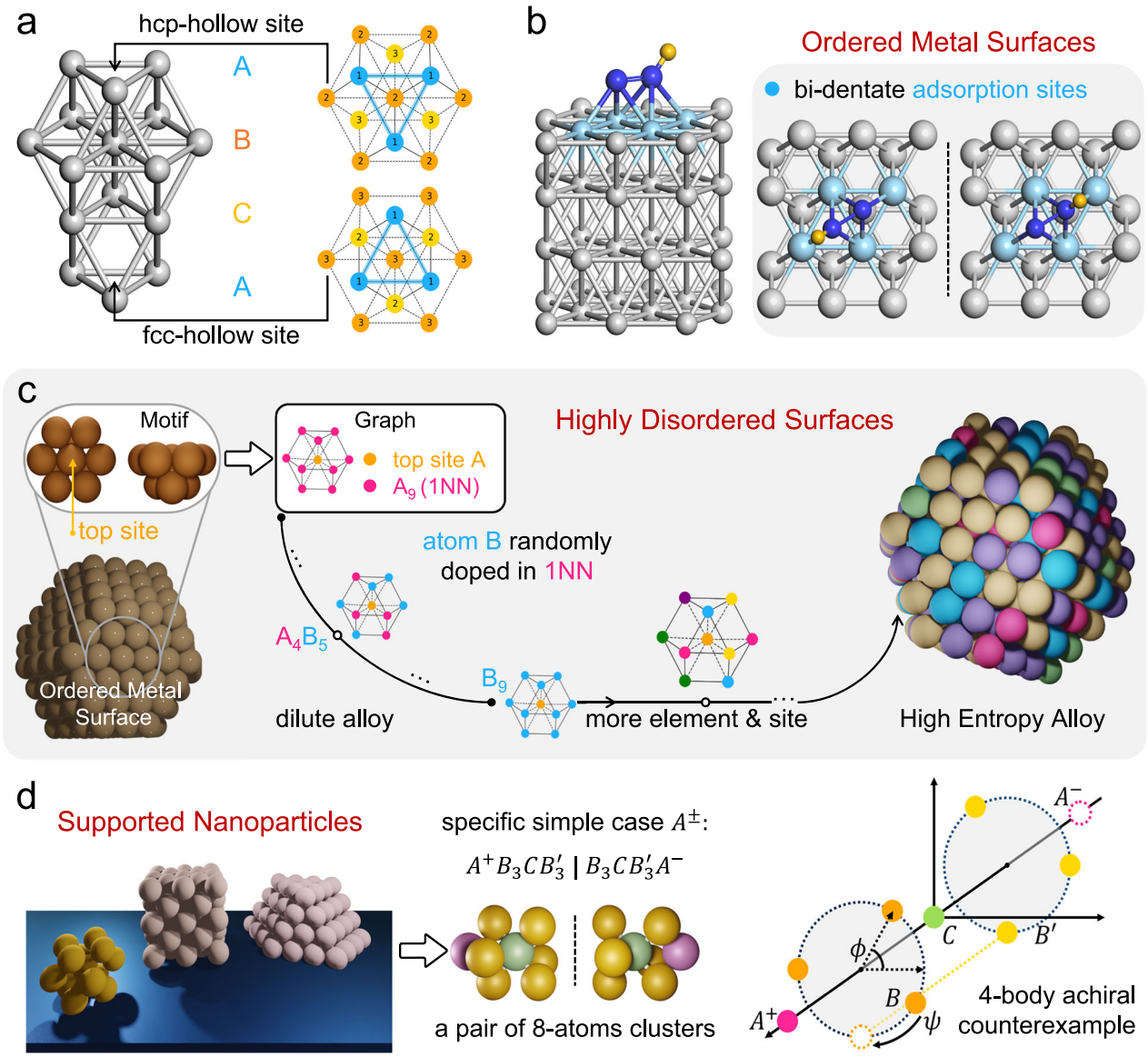

Fig. 1 | Chemical motifs at different metallic interfaces with various degrees of complexity. a Monodentate adsorption (the numbers in the cycles represent the layers) and (**b**) bidentate adsorption with hcp- and fcc-hollow sites on ordered catalyst surfaces. **c** Adsorption motifs on disordered catalyst surfaces. Left: the first nearest neighbors (1NN) around a top site on a dilute alloy as the simplest case; right: a high entropy alloy surface with much greater complexity. **d** Illustration of the supported cluster catalysis systems, and a specific case: a pair of 8-atom clusters as the 4-body achiral counterexample.

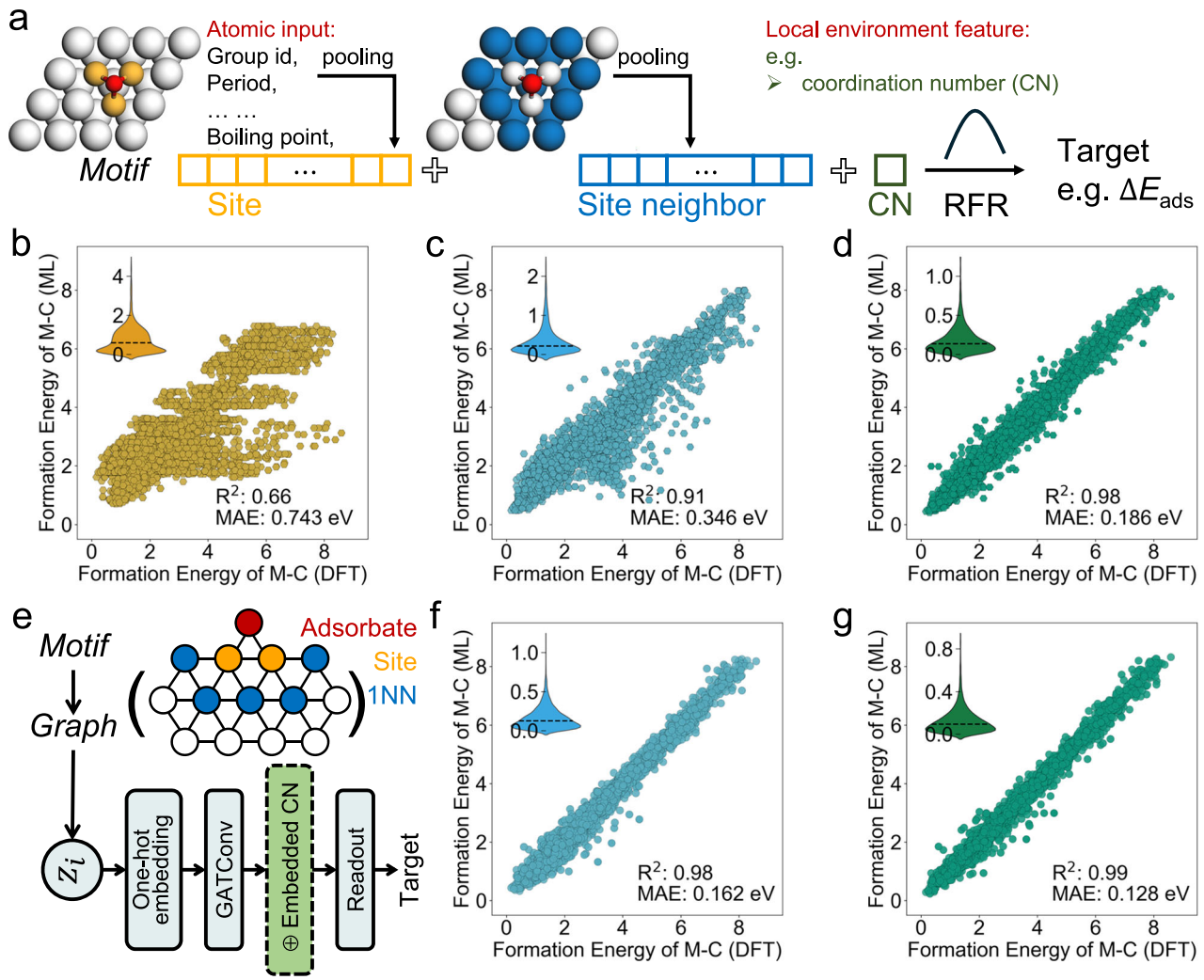

Fig. 2 | Schematic illustrations and prediction performances on the C_{ads} Database of the RFR and GAT models. a Schematic illustrations of site representation-based random forest regression (RFR) models. Parity plots of the density functional theory (DFT) calculated versus machine learning (ML) predicted formation energies of metal–carbon bonds (M–C) from the combined validation set in 5-fold cross-validation (CV) using the RFR models with different site representations through (**b**) sites, (**c**) sites and site neighbors, as well as (**d**) sites, site neighbors and coordination numbers (CN). **e** Schematic illustrations of the connectivity-based graph attention network (GAT) models without embedded CN (GAT-w/oCN) or with embedded CN (GAT-wCN) using GAT convolution (GATConv). Parity plots of the DFT-calculated versus ML-predicted formation energies of M–C with 5-fold CV using (**f**) GAT-w/oCN and (**g**) GAT-wCN models. The C_{ads} Database with 5096 entries is provided in the GitHub repository at Data Availability. Mean absolute error (MAE) and R^2 values are provided in parity plots from the RFR and GAT models; the violin plot in the inset shows the absolute error distributions; the inner dashed line represents the median (unit: eV).

considering a 13-atom group consisting of a central atom and its first coordination environment with 12 atoms in a five-element face-centered cubic (FCC) crystal, there are more than 100 million distinct chemical motifs in 9100 different chemical compositions[43]. This high chemical complexity extends far beyond that of mono-, bi-, and higher-dentate adsorption motifs on ordered catalyst surfaces. Moreover, dilute alloy (left side of Fig. 1c) and ternary alloy surfaces can be classified as simplified analogs of HEA surfaces[44–46].

Furthermore, for isolated clusters and the supported nanoparticles[47,48], a simple cluster exists as an achiral 4-body counterexample[7,10] with only eight atoms, as shown in Fig. 1d. It is very challenging to obtain a unique representation to resolve the similarity between the $A^{\pm}$ pairs. Thus, such highly complex catalytic systems necessitate the introduction of more information-rich and unique atomic structure representations to bypass the 4-body counterexample and resolve highly complicated adsorption motif similarity.

In this work, we developed a site representation-based approach to elucidate the significance and impact of atomic structure representations on the predictive performance of ML models, whereby graph fingerprints were manually constructed with varying degrees of atomic structure representations. Then, graph neural networks (GNNs) were employed to enhance the representations of atomic structures, where the edge features were constructed using a connectivity-based method. This allowed us to clarify the importance of resolving chemical-motif similarity. Two datasets composed of the catalytic systems presented in Figs. 1c, d were subsequently constructed to test the resolving power of equivariant message-passing for chemical-motif similarity with high complexity. Finally, we developed an equivariant GNN (equivGNN) model to provide robust representations of the adsorbate-metal motifs and accurately predict the associated energetic properties (e.g., binding energies). The model demonstrated superior prediction performance, with all MAEs <0.09 eV on several datasets spanning the diverse catalytic systems represented in Fig. 1. Notably, our model was easy to implement and outperformed the available prominent ML models that were specifically designed for individual systems, i.e., DOSnet[20], TinNet[23], WWL-GPR[33], GAME-Net[34], and augmented CGConv[37]. The universality and efficiency of our developed equivGNN model across different systems were proven.

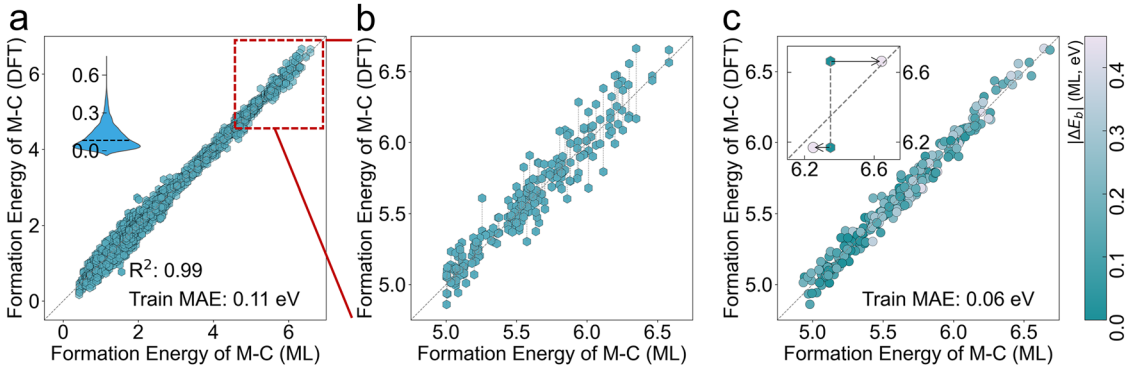

Fig. 3 | Enhancing prediction accuracy by distinguishing similar adsorption motifs. a Parity plots of the ML-predicted (GAT-w/oCN) versus DFT-calculated formation energy of M-C on the 3-fold-only C_{ads} database. Mean absolute error (MAE) and R^2 values are provided in the parity plot; the violin plot in the inset shows the absolute error distributions, and the inner dash line represents the median (unit: eV). The pairs of hcp-/fcc-hollow site adsorption motifs in the weak adsorption range (**b**) with identical ML predictions derived from the GAT-w/oCN joined with dashed lines. **c** Results colored with the differences in the ML predictions obtained from the GAT-wCN; the diagram in the inset shows the predictions from GAT-w/oCN and GAT-wCN plotted with the hexagons and circles, respectively. The 3-fold-only C_{ads} Database with 3262 entries is provided in GitHub repository at Data Availability. ML machine learning, DFT density functional theory, GAT graph attention network, CN coordination number.

Results

Prediction accuracy versus atomic structure representation levels

We start with a relatively simple case to illustrate the importance of atomic structure representations for developing surrogate ML models for predicting binding energies, i.e., monodentate adsorbates on the ordered catalyst surfaces shown in Fig. 1a. To decrease uncertainties caused by using complex data sources, we start with datasets of single adsorbates in the monodentate adsorption motifs on the close-packed surface of binary alloys, and more details are provided in the "**Methods**" section.

Owing to its good performance in systems with high similarity[24], the random forest regression (RFR) method is first chosen to evaluate the prediction performances achieved with different levels of site representations for the adsorption motifs in the atomic-carbon monodentate adsorption dataset (C_{ads} Dataset). Figure 2a depicts a schematic of RFR models with three different levels of site representations, and more details are shown in Supplementary Note 1. The resulting parity plots of the DFT-calculated versus ML-predicted formation energies of metal–carbon bonds (M–C) with fivefold cross-validation (5-fold CV) are shown in Figs. 2b–d. The plots reveal that the levels of site representations highly influence the performance of the RFR models. Notably, adding coordination numbers (CNs)[49,50] as a local environment feature greatly improves the model performance, and the MAE significantly decreases from 0.346 eV to 0.186 eV.

Notably, adsorbate-surface structures (adsorption motifs) can be naturally transformed into graph-structured data, where nodes represent atoms and edges represent the connections between pairs of atoms. By providing efficient frameworks to learn from graph-structured data, graph neural networks (GNNs) are easily applicable to predict the relationships between adsorption motifs and binding energies as a graph-level task. Thus, we further choose graph attention networks (GATs)[51] to enhance the atomic structure representation of adsorption motifs from unrelaxed structures and test them on the C_{ads} Dataset.

Figure 2e shows the schematics of connectivity-based GAT models, which employ atomic numbers as node inputs, and their edge weights are based on the connectivity of the adsorption motifs. The message-passing process in GAT updates the node features by aggregating messages from the node neighbors, and global pooling is used to extract graph characteristics from all node features for obtaining graph-level predictions. More details are shown in Supplementary Note 1. By mitigating the need for manual feature engineering, this approach simplifies the feature construction process. The parity plots of the DFT-calculated versus ML-predicted formation energy of M-C for GAT models without (GAT-w/oCN) and with CN (GAT-wCN) are shown in Fig. 2f, g, respectively. Overall, the connectivity-based GAT models perform better than the RFR models in predicting the formation energy of M-C on the C_{ads} Dataset. With the help of the CNs, the MAE decreases from 0.162 eV to 0.128 eV in the GAT models. However, this enormous enhancement in prediction performance, achieved by including the CNs, indicates that the original connectivity-based structure representation might be intrinsically deficient for the adsorption motifs in the C_{ads} Dataset.

False-positive prediction accuracy caused by the failure to distinguish the similarity of adsorption motifs

A close inspection reveals that the GAT-w/oCN model cannot produce unique structural representations for similar chemical motifs in systems at metallic interfaces (i.e., the monodentate adsorption motifs shown in Fig. 1a) because of the utilization of the connectivity among atoms as edge attributes. Here, we construct a dataset containing only the 3-fold adsorption structures included in the C_{ads} Dataset, termed the 3-fold-only C_{ads} dataset, to address the deficiency of connectivity-based GNNs in terms of distinguishing the chemical similarity of pairs of hcp-/fcc-hollow site adsorption motifs.

Figure 3a shows the resulting prediction performance, with a training MAE of 0.11 eV for the connectivity-based GAT-w/oCN model on the 3-fold-only C_{ads} dataset, where all the data points are set as training sets. However, the deficiency and misleading characterization exhibited by the GAT-w/oCN on the 3-fold-only C_{ads} dataset lie behind these seemingly reasonable prediction results. Figure 3b shows the parity plot of the ML-predicted versus DFT-calculated formation energy of M-C in the weak adsorption range. The pairs of hcp-/fcc-hollow site adsorption motifs are connected by short dashed lines, but the ML values predicted via the GAT-w/oCN are identical for the pairs. The results demonstrate that the GAT model with connectivity-based structure representations is deficient in its sensitivity for distinguishing between pairs of 3-fold adsorption motifs.

Figure 3c shows the parity plots of the ML-predicted (using GAT-wCN) versus DFT-calculated formation energy of M-C on the 3-fold-only C_{ads} database, and the pairs of hcp-/fcc-hollow site adsorption motifs are colored according to the differences between their ML predictions. The inset in Fig. 3c illustrates that the GAT-wCN, which contains the local environment features of the embedded CN, can resolve the discrepancy between the pairs of hcp-/fcc-hollow site

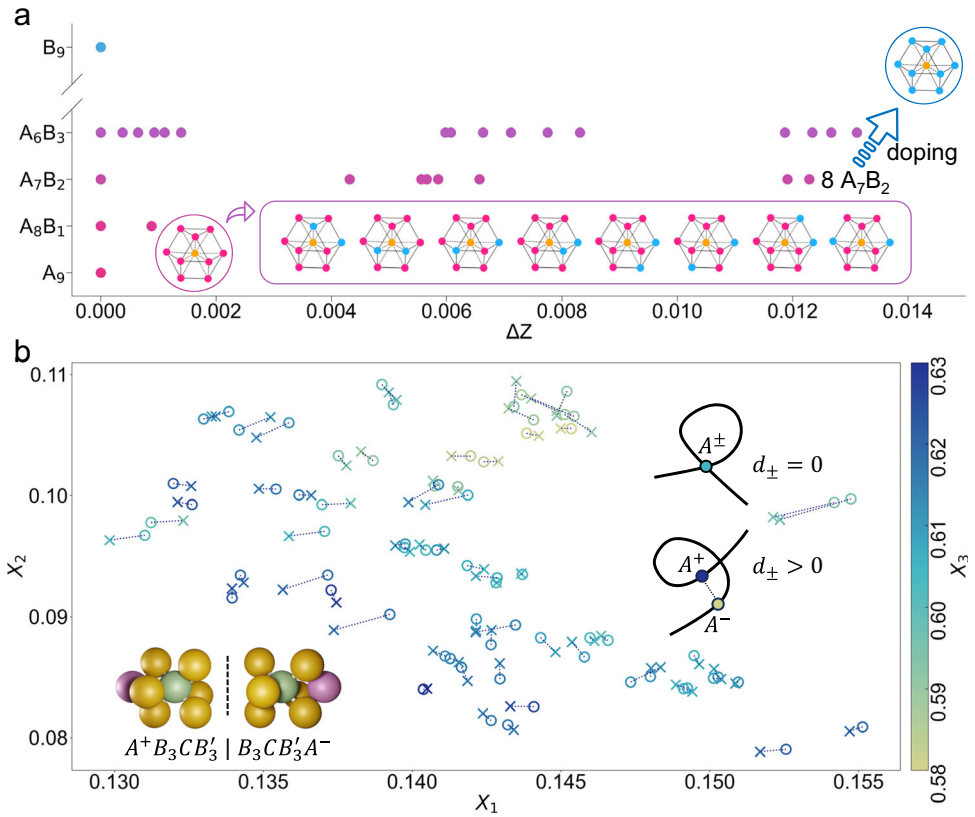

Fig. 4 | Resolving adsorption motif similarity via equivariant message-passing on highly complex catalytic systems. a Polya enumeration and distinguishing of the top site adsorption motifs on dilute alloy, which are plotted with dissimilarity ΔZ_i as the difference between graph fingerprint Z_i with the minimum value of the corresponding composition; the inset graphs represent the first nearest neighbors around top site with A_9, A_7B_2 and B_9 compositions; yellow points represent the top site, pink points are atom A and blue points are atom B. **b** Plots of the 70 $A^{\pm}$ pairs with their graph fingerprints, (X_1, X_2, X_3), plotted with pairs of circle and cross and joined with dashed lines; the inset image is a pair of $A^{\pm}$ as 4-body counterexample structure, and the inset diagrams show that the 4-body counterexample can be distinguished by performing equivariant message-passing that generates a dissimilarity measure $d_{\pm} > 0$.

adsorption motifs in the C_{ads} Database. As a result, the addition of CNs yields higher overall accuracy in terms of predicting the formation energy of M-C on the C_{ads} Dataset (Fig. 2g) than do the predictions produced by the GAT-w/oCN (Fig. 2f). Therefore, by improving the atomic structure representations, the importance of enhancing the power to resolve adsorption motif similarity for the development of surrogate ML models is evident.

Nevertheless, the efficacy of these subgraph-based site representation methods will diminish when attempting to resolve chemical motif similarity in systems with even greater complexity. For example, we tested the GAT-wCN on Complex Dataset[33], an external dataset in the literature that contains the adsorption motifs of complex adsorbates such as bi- and higher-dentate motifs adsorbed on metal and alloy surfaces (Fig. 1b). The GAT-wCN had inferior prediction performance, with a root-mean-square error (RMSE) of 0.30 eV (Supplementary Fig. 1) compared with 0.18 eV in the original work. Furthermore, the similarity problem becomes severe for chemically complex materials, such as the chemical motifs on highly disordered surfaces and supported nanoparticles (Figs. 1c, d). To improve the structure representations of adsorption motifs and guarantee accurate representations for any two distinct chemical motifs, high-order GNN models are necessary and highly desirable.

Enhanced atomic structure representations with the equivariant GNN model

To address the challenge mentioned above, GNNs with equivariant message-passing were applied in this work to resolve adsorption motif similarity between different catalytic systems with various degrees of complexity. To illustrate the power of equivariant message-passing for resolving chemical-motif similarity in complex systems, we start with the disordered surfaces of dilute alloys shown in Fig. 1c.

Considering a system with adsorbates forming monodentate adsorption on a close-packed surface of FCC metal A, the top site contains nine atoms in its first nearest neighbors (1NN). There are 512 (2^9) possible chemical motifs in the 10 different compositions that can be constructed when we randomly replace several of these nine atoms with another metal B. However, there are some identical motifs among the 512 possible chemical motifs, i.e., only two distinct motifs for A_8B composition among the nine possibilities. Here, we construct all 512 possible 1NN structures and convert them into graphs using radius graphs. The graph fingerprints Z_i are generated through a randomly initialized GNN model with equivariant message-passing but without training (implemented using the *e3nn* package[52]). Figure 4a shows the eight distinct adsorption motifs with A_7B_2 compositions, which are identified from 36 possibilities by kicking out all equivalent structures through equivariant message-passing. The 104 distinct motifs included in the 512 possibilities are shown in Supplementary Fig. 2. This result is confirmed by the Polya enumeration theorem[53], which is a general mathematical formalism based on group theory that can analytically count the number of distinct chemical motifs[43,54]. The details of the dataset construction and Polya enumeration processes are shown in Supplementary Note 2.

Next, we evaluate the resolving power of equivariant message-passing on more challenging catalytic systems in the heterogeneous catalysis field, specifically, the highly complex 4-body achiral counterexample, as illustrated in Fig. 1d. This counterexample consists of a

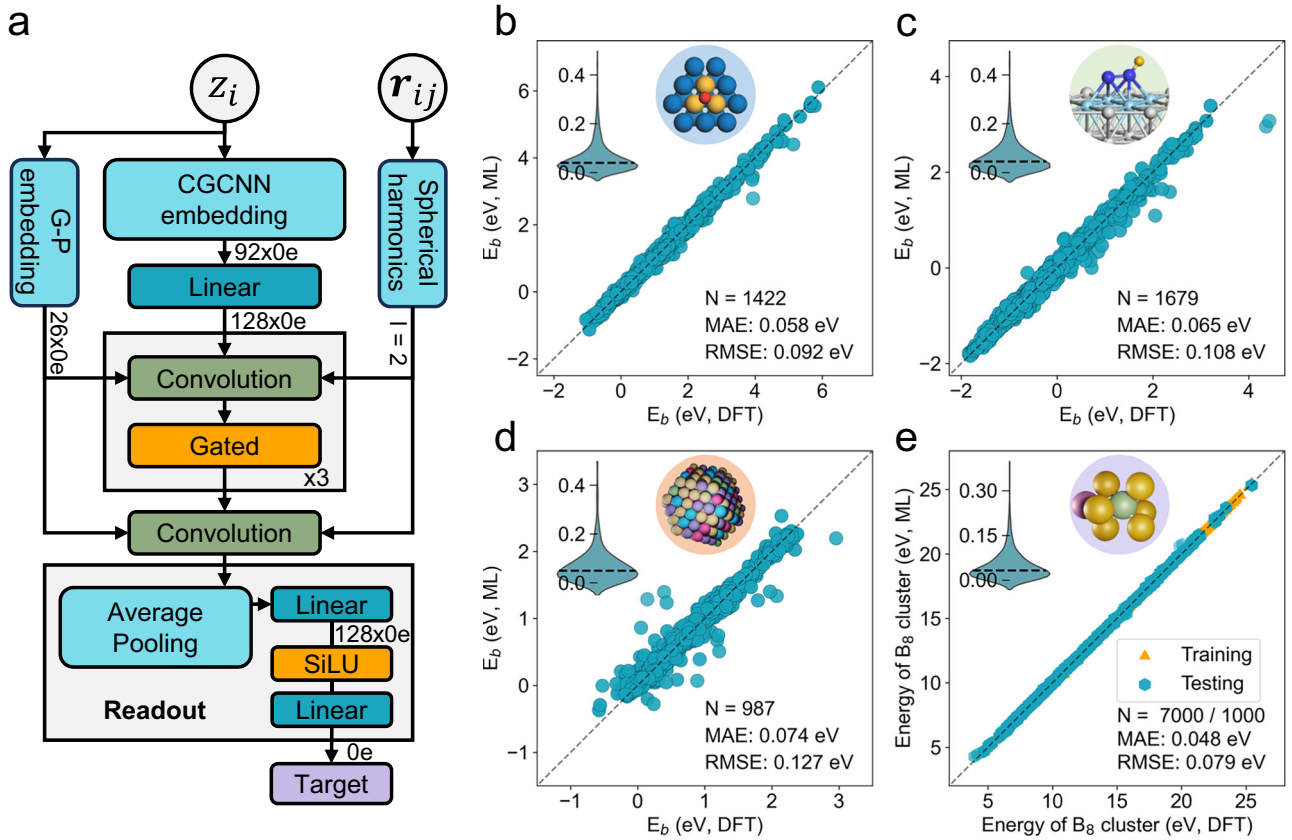

Fig. 5 | Schematic illustration and prediction performances of the equivGNN model. a Overview of the equivGNN architecture, which integrates atom embedding, equivariant message-passing, and readout blocks. Parity plots of the DFT-calculated versus ML-predicted binding energies derived from the combined validation set under 5-fold cross-validation (CV) via the equivGNN model on (**b**) Simple Dataset[33], **c** Complex Dataset[33], and (**d**) HEA Dataset[37]. **e** The prediction performance of the equivGNN on B_8-cluster Dataset[10] with a fixed training set. The samples, mean absolute errors (MAE) and root mean square errors (RMSE) are provided in parity plots from the equivGNN; the violin plots in the inset show the absolute error distributions, and the inner dash line represents the median (unit: eV); the inset images are simplified diagrams for catalytic systems with various degrees of complexity. DFT density functional theory, ML machine learning.

pair of inequivalent 8-atom clusters that share identical inversion-invariant 4-body correlation features. Here, we construct a dataset comprising 70 pairs of 8-atom clusters, denoted as $A^{\pm}$, with the construction details summarized in Supplementary Note 3. These clusters are subsequently converted to graphs and processed through an untrained and randomly initialized GNN model with equivariant message-passing, thereby generating graph fingerprints $\boldsymbol{X} \in R^3$. As shown in Fig. 4b, the 70 $A^{\pm}$ pairs are joined by short-dashed lines, plotted in a 2D space using coordinates (X_1, X_2) and colored with X_3 in their graph fingerprints. The distance between the graph fingerprints of any $A^{\pm}$ pair is greater than zero, which is denoted as $d_{\pm} > 0$, this quantifies the dissimilarity between the pairs. Unique atomic structure representations have been developed to address the 4-body counterexample. Consequently, the results in Fig. 4 demonstrate that equivariant message-passing can distinguish any possible chemical motif in the diverse heterogeneous systems in Fig. 1, regardless of their complexity. These capabilities identified in our study hold significant potential for the development of surrogate ML models that can provide highly accurate predictions of descriptor values.

Universally improved prediction accuracy with the equivGNN

The architecture of our developed equivariant GNN model (equivGNN) is illustrated in Fig. 5a, which integrates atom embedding, equivariant message-passing, and readout blocks. The details of the equivariant message-passing process are shown in the "**Methods**" section. For node inputs, we map the atomic numbers to 92-dimensional vectors using atomic embedding from CGCNN[29]. As shown in Supplementary Table 1, we explore the influences of self-connection layers with various node attributes on the prediction performance and training efficacy of the model. Compared with the z-type and atomic node attributes, the one-hot-encoded group and period (G-P) as atom attributions with 26-dimensional vectors have the highest model prediction performance with the lowest time consumption. Therefore, the G-P (26) embedding method for node attributes is employed in the self-connection layer. The development details of the equivGNN are shown in the "**Methods**" section.

Indeed, our developed equivGNN model exhibits superior prediction performance, with an MAE of only 0.076 eV on the C_{ads} Dataset under 5-fold CV, as shown in Supplementary Fig. 3a. Moreover, we apply the equivGNN to the 3-fold-only C_{ads} dataset, as shown in Supplementary Fig. 4, and the equivGNN can distinguish similar 3-fold adsorption sites on ordered metal or alloy surfaces. We turn to three additional datasets with simple adsorbates (H, O, and CH_3) in monodentate adsorption motifs on the ordered surfaces of 1998 binary alloys. The excellent prediction performances achieved by the equivGNN model on these datasets using unrelaxed structures are shown in Supplementary Figs. 3b–d, with MAEs of only 0.047, 0.087, and 0.062 eV for H, O, and CH_3, respectively. For the case involving a single monodentate adsorbate on ordered catalyst surfaces in a computational catalysis scenario, our model dramatically outperforms two prestigious models in this field, i.e., DOSnet[20], which utilizes the key features extracted from the electronic densities of states, and TinNet[23]. The details of the model training are shown in the "**Methods** section" and Supplementary Table 2.

Having demonstrated the efficacy of the equivGNN model in terms of predicting outcomes on datasets with a single adsorbate in

monodentate adsorption motifs, we proceed to evaluate the prediction performance of the equivGNN on datasets concerning different catalytic systems with various degrees of complexity that have been reported in the literature. These datasets include Simple Dataset[33] (Fig. 1a, simple adsorbates in monodentate adsorption motifs on ordered metal surfaces), Complex Dataset[33], and Organic Dataset[34] (Fig. 1b, complex adsorbates in bi- and higher-dentate adsorption motifs on ordered metal surfaces), HEA Dataset[37] (Fig. 1c, adsorption motifs on highly disordered surfaces), and B_8-clusters Dataset[10] (Fig. 1d, structures for supported nanoparticles). The early-stopping strategy is used during model training, and the other details of the training strategies applied to all the datasets are shown in Supplementary Note 4.

Monodentate adsorption motifs

First, for the heterogeneous catalysis systems of simple adsorbates on ordered surfaces with monodentate adsorption motifs (as illustrated in Fig. 1a), we evaluate our equivGNN model on an external dataset, i.e., Simple Dataset[33] with 1422 entries, which comprises 8 simple adsorbates (C, H, O, CO, OH, CH, CH_2 and CH_3) with monodentate adsorption motifs on ordered metal and alloy surfaces. Figure 5b shows the parity plots of the DFT-calculated versus ML-predicted binding energies derived from the combined validation set under 5-fold CV using our equivGNN model on the Simple Dataset. The prediction performance of the equivGNN is excellent, with an MAE of only 0.058 eV and an RMSE of 0.092 eV, which are much better than those of Gaussian process regression with a Wasserstein Weisfeiler–Lehman graph kernel (WWL-GPR), which attained an RMSE of 0.13 eV in the original work[33].

Higher-dentate adsorption motifs

Next, for complex adsorbates on ordered surfaces with bi- and higher-dentate adsorption motifs (as illustrated in Fig. 1b), we evaluate our equivGNN model on two external datasets, the Complex Dataset[33] and Organic Dataset[34]. These two datasets can be used to quickly evaluate the prediction performance of our ML model for systems with complex adsorbates on ordered metal surfaces.

The Complex Dataset with 1679 entries comprises 41 complex adsorbates (e.g., CHCO, CCHOH, CH_2CH_2O, and CH_3CH_2OH) with bi- and higher-dentate adsorption motifs on the fcc(111)/(211) facets of pure metals, which can be seen as a much denser sampling of diverse adsorption motifs than the OC20[31] dataset. On this more challenging complex dataset, our equivGNN model has superior performance, with an MAE of only 0.065 eV and an RMSE of 0.108 eV (as shown in Fig. 5c), outperforming the WWL-GPR model with an RMSE of 0.18 eV in the original work[33]. Another well-balanced and chemically diverse dataset, the Organic Dataset with 3108 adsorption structures of 207 closed-shell organic molecules adsorbed on 14 metals, is used to evaluate our equivGNN model. Notably, better prediction performance is achieved by the equivGNN on the Organic Dataset, with an MAE of only 0.089 eV (details in Supplementary Fig. 5). Graph-based adsorption on a metal-energy neural network (GAME-Net) yields an MAE of 0.18 eV[34]. Specifically, the MAE for aromatic compounds is 0.12 eV (as shown in Supplementary Fig. 6), and the accuracy improvement over that of GAME-Net (MAE of 0.34 eV) is greater than 60%.

Adsorption motifs on HEAs

For catalytic systems with adsorption motifs on highly ordered surfaces (as illustrated in Fig. 1c), our equivGNN model is subsequently evaluated on the HEA Dataset obtained from Chang et al. [37], which comprises 1984 entries for OH* and O* with unrelaxed monodentate adsorption motifs on the surfaces of quinary materials belonging to the FeCoNiIrRu HEAs family. Chang et al. trained an augmented CGCNN model[29] on the HEA Dataset with a digital annealer (DA)-driven similarity feature, which played a similar role as the CN embedding in the GAT models used in our study; their fixed training data had 1580 entries, and good prediction performance was achieved, with an MAE of 0.080 eV over 10000 training epochs. Compared with the above augmented CGCNN model, our equivGNN is trained for only 200 training epochs and achieves superior prediction performance on the HEA Dataset, with an MAE of 0.074 eV (shown in Fig. 5d) on the combined validation set under 5-fold CV. This result confirms the efficiency and robustness of the equivGNN in terms of predicting complex adsorption motifs on highly disordered surfaces.

Four-body achiral counterexample clusters

To further evaluate the performance of our equivGNN model in catalytic systems of supported nanoparticles (as illustrated in Fig. 1d), a prediction task is implemented utilizing the B_8-cluster Dataset[10], which contains 8000 entries with 4000 pairs of B_8 clusters of A^+ and A^- as the 4-body counterexample. Following the same training/test split ratio as that in the original study, our equivGNN model is trained on the first 3500 pairs of B_8 clusters, with the remaining 500 pairs allocated as the test set. On this test set, the model demonstrates minimal errors, with a mean energy error of 0.006 eV and an energy difference error of 0.002 eV for the 500 pairs. These results are significantly better than the original predictions, which yielded a mean error of 0.026 eV and a difference error of 0.035 eV. Furthermore, Fig. 5e shows the parity plot of the DFT-calculated versus ML-predicted energies of the B_8 clusters obtained using the equivGNN model, and the prediction performance is excellent, with an MAE of only 0.048 eV.

Indeed, the evaluation of our equivGNN model conducted on nine datasets constitutes a universal test domain, spanning different systems in the heterogeneous catalysis field with various degrees of complexity, as shown in Fig. 1. The prediction performances, with all MAEs below 0.09 eV, are also very noteworthy and have not yet been achieved by the previous ML models, i.e., DOSnet[20], TinNet[23], WWL-GPR[33], GAME-Net[34], and augmented CGConv[37]. To further understand catalytic reactions, we apply equivGNN on a coverage dataset[32] obtained from the work that published the ACE-GCN model. Supplementary Fig. 7 shows the parity plot of the DFT-calculated versus ML-predicted average NO* binding energies for an equivGNN model with NO* configurations consisting of 1–4 NO* molecules per unit cell. The prediction performance, with an MAE of 0.014 eV, is better than that of the original ML model (which attained an MAE of 0.02 eV), while the data split is the same as that employed in the original work. To explore the performance of our model on large databases, we apply the equivGNN to the formation energy dataset of the Materials Project with 132752 entries[55]. The equivGNN model achieves excellent performance, with an MAE of 17.4 meV $atom^{-1}$, and the accuracy improvement attained over M3GNET[56] is 10%, as shown in Supplementary Table 3. Furthermore, the implementation of our model is easy and has low computational and training costs. Thus, the universality and efficiency of our equivGNN model for different heterogeneous catalysis systems are evident, indicating its great potential for addressing the challenging complexities of heterogeneous catalysis systems.

Discussion

In this work, we specifically illustrated the importance and influence of atomic structure representations on ML models by implementing site representation-based RFR regression models on the datasets with simple adsorbate monodentate adsorption motifs. We subsequently implemented connectivity-based graph attention network (GAT) models to improve the representations of atomic structures, where edge features were constructed using a connectivity-based approach. We then identified and emphasized the importance of resolving chemical-motif similarity with enhanced atomic structure representations to avoid false-positive predictions when developing ML models. Eventually, to address the different degrees of complexity observed at metallic interfaces, we strived to enhance the representations of atomic structures using equivariant message-passing and developed the equivGNN model to

resolve adsorption motif similarity. Our developed ML model attained superior prediction accuracies on both simple monodentate datasets and several complex datasets in the literature, with all of its MAEs below 0.09 eV, which has not yet been achieved in previous studies. Notably, the complexities of these tested catalytic systems cover different scenarios in heterogeneous catalysis, i.e., simple or complex adsorbates with mono- or higher-dentate adsorption motifs on ordered metal surfaces; adsorption motifs on the highly disordered surfaces of high-entropy alloys; and complex structures for supported nanoparticles. Apart from the high prediction accuracy, the implementation of our model is facile, with low computational and training costs. The universally improved prediction accuracies for different datasets in the literature achieved by using our developed model indicate the importance of resolving chemical-motif similarity with equivariant message-passing-enhanced atomic structure representations, which have not yet attracted sufficient attention in the field of ML model development. Thus, our developed equivGNN-based surrogate ML model could be a promising alternative to computationally expensive DFT calculations, working as a helpful tool for screening large material spaces to accelerate the process of discovering new catalysts.

Methods

Computational details of the monodentate datasets

The dataset of the monodentate adsorption motifs on the close-packed surfaces of binary alloys was constructed. All DFT calculations are carried out with the QUANTUM ESPRESSO code[57], using the periodic plane-wave-based DFT method with pseudopotentials GBRV version 1.5[58], and the surface models are constructed based on the 1998 binary alloys from our previous work[59]. The contribution of the exchange-correlation to the electronic energy is described by the BEEF-vdW (Bayesian Error Estimation Functional - van der Waals) functional. The energy cutoffs for the plane wave and electron density are set to 500 and 5000 eV, respectively. The close-packed surfaces of the alloys are simulated using four-layer (2×2) supercells with two relaxed top layers and two constrained bottom layers. A vacuum layer with 12 Å is set between periodically repeated slabs. The $(4 \times 4 \times 1)$ Monkhorst–Pack k-point grids[60] are applied for sampling. Structure optimizations are done when the energy difference is lower than 10^{-5} eV and the forces are less than 0.05 eV/Å.

Equivariant message-passing

The process of enhancing atomic structure representation through equivariant message-passing has been demonstrated to possess a good ability to resolve the adsorption motif similarity encountered in heterogeneous catalysis systems, as shown in Fig. 4. The equivariant message-passing process is achieved through the equivariant *Tensor Product* in the *e3nn* 0.5.1 package[52], and the associated formula is as follows:

$$f_i' = \frac{1}{\sqrt{z}} \sum_{j \in \partial(i)} f_j \otimes h\left(|\boldsymbol{x}_{ij}|\right) Y\left(\frac{\boldsymbol{x}_{ij}}{|\boldsymbol{x}_{ij}|}\right) \tag{1}$$

where f_j and f_i' are the node input and output, respectively; $\boldsymbol{x}_{ij}$ is the relative vector; z is the average degree of the nodes; $\partial(i)$ is the set of neighbors of node i; h is a multi-layer perceptron; and Y represents the spherical harmonics.

Development of the equivGNN model

Our equivGNN model is developed on top of the open-source *e3nn* repository. First, we test an original GNN model with equivariant message-passing, with *e3nn*-v2103 as the baseline, on the C_{ads} Dataset. The graph nodes have the one-hot encoded atomic type (z) as their input with a 92-element-long array; this setting is labeled as z-type (92). The prediction performance of the model, with an MAE of 0.103 eV, as shown in Supplementary Table 1, is superior to that of the GAT-wCN (with an MAE of 0.128 eV, as shown in Fig. 2g) because of the enhanced atomic structure representations. However, the convolution block in *e3nn*-v2103 uses too many fully connected TensorProduct layers, which seem redundant and result in low training efficiency. To further improve the training efficiency of the model, the convolution block is replaced with a NequIP interaction block[61] by using linear layers as self-interactions. As shown in Supplementary Table 1, the training efficiency increases by more than 30% without a loss of accuracy.

Then, for node inputs, we map the atomic number to a 92-dimensional array using the atomic features embedding acquired from the CGCNN[29], this setting, labeled as atomic features (92), improves the accuracy by 11% without increasing the training cost. Next, we explore the influences of self-connection layers with various node attributes on the resulting prediction performance, training efficacy, and total number of parameters of the models, which are presented in Supplementary Table 1. When the 26-dimensional G-P embedding vectors are used as node attributes, this setting is labeled as G-P (26), the model parameters increase fivefold compared with those employed without node attributes, while the training speed decreases by 1 s/epoch, and the prediction error decreases significantly by 13%. Compared with using z-type and atomic features as node attributes, employing the one-hot encoded G-P (26) embedding as the node attributes yields the highest model prediction performance with the lowest time cost. Therefore, the G-P (26) embedding scheme for node attributes is employed in the self-connection layer. Notably, the accuracy and efficiency are greater than 20% (22% and 28%, respectively), more than those of the baseline. All the models are trained on the same randomly shuffled training (80%) and validation (20%) sets but without CV.

Graph construction

For adsorption systems, the distance criterion for the graph construction process is based on the covalent radius. Atoms i and j are connected by edges, where the distance between them is less than their total covalent radius multiplied by 1.35, as follows:

$$d_{ij}^{cut} = 1.35 \times (r_i^{cov} + r_j^{cov}) \tag{2}$$

where d_{ij}^{cut} is the cutoff distance between atom i and atom j; and r_i^{cov} and r_j^{cov} are the covalent radius of atom i and atom j, respectively.

For the Materials Project database, the graphs of the crystal structures are constructed using the k^{th} nearest distance, which is determined by the k^{th} nearest neighbors.

Model training

For the monodentate datasets, the training/test splits are set to 8:2, consistent with the data splits used for the Simple Dataset and Complex Dataset. For the Organic Dataset and HEA Dataset, the training/validation/test splits are set to 6:2:2 and 8:1:1, respectively. For B_8-cluster Dataset, the first 7000 entries (3500 pairs of B_8) are used for training, and the last 1000 entries (500 pairs) are reserved for testing, following the same split as described in the corresponding reference. For the coverage dataset, we use the same training/test splits in the original work, and it is 260:65. For all tasks implemented on the datasets, we use the weighted adaptive moment estimation (AdamW) optimizer with a weight decay rate of 1e-5 and a one-cycle learning rate scheduler. The batch size is set to 8. We use the mean squared error metric (MSELoss) as the loss function for training, except for the Organic and HEA datasets, for which the mean absolute error metric (L1Loss) is used. We only slightly adjust the learning rates and training epochs for different tasks conducted on the ten datasets, as shown in Supplementary Table 2. We use PyTorch 2.1.0 to implement our equivGNN models. For all the tasks implemented on the datasets, we use one NVIDIA RTX 4090 GPU (24 GB) to train our equivGNN models. For the formation energy prediction task performed on the Materials Project database, the batch size is set as 64, and two hidden layers with

448×0e + 64×1o + 64×2e hidden irreps are used in the equivGNN model, which is trained on one NVIDIA V100 GPU (32 GB) with the mean absolute error metric.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The DFT-calculated formation energies of metal–carbon bonds (M–C) and relaxed structures of C_{ads} Database and 3-fold-only C_{ads} Database have been deposited into a public repository: (https://github.com/woshicc/asp) and Zenodo[62]. All other data that support the findings of this study are available at (https://github.com/woshicc/asp). Source data are provided with this paper.

Code availability

The source code is available in a public repository: (https://github.com/woshicc/asp) and Zenodo[62].

References

1. Seh, Z. W. et al. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **355**, eaad4998 (2017).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
3. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
4. Hart, G. L. W., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
5. Damewood, J. et al. Representations of materials for machine learning. *Annu. Rev. Mater. Res.* **53**, 399–426 (2023).
6. Liu, X. & Peng, H.-J. Toward next-generation heterogeneous catalysts: empowering surface reactivity prediction with machine learning. *Engineering* **39**, 25–44 (2024).
7. Pozdnyakov, S. N. et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **125**, 166001 (2020).
8. Pozdnyakov, S. N., Zhang, L., Ortner, C., Csányi, G. & Ceriotti, M. Local invertibility and sensitivity of atomic structure feature mappings. *Open Res. Europe* **1**, 126 (2021).
9. Pozdnyakov, S. N. & Ceriotti, M. Incompleteness of graph neural networks for points clouds in three dimensions. *Mach. Learn.: Sci. Technol.* **3**, 045020 (2022).
10. Nigam, J., Pozdnyakov, S. N., Huguenin-Dumittan, K. K. & Ceriotti, M. Completeness of atomic structure representations. *APL Mach. Learn.* **2**, 016110 (2024).
11. Ma, X., Li, Z., Achenie, L. E. & Xin, H. Machine-learning-augmented chemisorption model for CO_2 Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
12. Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131–24138 (2017).
13. Noh, J., Back, S., Kim, J. & Jung, Y. Active learning with non-ab initio input features toward efficient CO_2 reduction catalysts. *Chem. Sci.* **9**, 5152–5159 (2018).
14. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **1**, 696–703 (2018).
15. Back, S. et al. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* **10**, 4401–4408 (2019).
16. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. *Chem* **6**, 3100–3117 (2020).
17. Gu, G. H. et al. Practical deep-learning representation for fast heterogeneous catalyst screening. *J. Phys. Chem. Lett.* **11**, 3185–3191 (2020).
18. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *npj Comput. Mater.* **6**, 177 (2020).
19. Zhong, M. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
20. Fung, V., Hu, G., Ganesh, P. & Sumpter, B. G. Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* **12**, 88 (2021).
21. Li, X., Chiong, R. & Page, A. J. Group and period-based representations for improved machine learning prediction of heterogeneous alloy catalysts. *J. Phys. Chem. Lett.* **12**, 5156–5162 (2021).
22. Roy, D., Mandal, S. C. & Pathak, B. Machine learning-driven high-throughput screening of alloy-based catalysts for selective CO_2 hydrogenation to methanol. *ACS Appl. Mater. Interfaces* **13**, 56151–56163 (2021).
23. Wang, S. H., Pillai, H. S., Wang, S., Achenie, L. E. K. & Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* **12**, 5288 (2021).
24. Liu, X., Cai, C., Zhao, W., Peng, H.-J. & Wang, T. Machine learning-assisted screening of stepped alloy surfaces for C_1 catalysis. *ACS Catal.* **12**, 4252–4260 (2022).
25. Li, X., Chiong, R., Hu, Z. & Page, A. J. A graph neural network model with local environment pooling for predicting adsorption energies. *Comput. Theor. Chem.* **1226**, 114161 (2023).
26. Malone, W. & Kara, A. Predicting adsorption energies and the physical properties of H, N, and O adsorbed on transition metal surfaces: A machine learning study. *Surf. Sci.* **731**, 122252 (2023).
27. Zhang, J. et al. Accurate and efficient machine learning models for predicting hydrogen evolution reaction catalysts based on structural and electronic feature engineering in alloys. *Nanoscale* **15**, 11072–11082 (2023).
28. Kirkvold, C., Collins, B. A. & Goodpaster, J. D. CatEmbed: a machine-learned representation obtained via categorical entity embedding for predicting adsorption and reaction energies on bimetallic alloy surfaces. *J. Phys. Chem. Lett.* **15**, 6791–6797 (2024).
29. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
30. Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A. & Muller, K. R. SchNet - a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
31. Chanussot, L. et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
32. Ghanekar, P. G., Deshpande, S. & Greeley, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nat. Commun.* **13**, 5788 (2022).
33. Xu, W., Reuter, K. & Andersen, M. Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation. *Nat. Comput. Sci.* **2**, 443–450 (2022).
34. Pablo-Garcia, S. et al. Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks. *Nat. Comput. Sci.* **3**, 433–442 (2023).
35. Tran, R. et al. The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **13**, 3066–3084 (2023).
36. Clausen, C. M., Nielsen, M. L. S., Pedersen, J. K. & Rossmeisl, J. Ab initio to activity: machine learning-assisted optimization of high-entropy alloy catalytic activity. *High. Entropy Alloy. Mater.* **1**, 120–133 (2022).
37. Chang, Y. et al. High-entropy alloy electrocatalysts screened using machine learning informed by quantum-inspired similarity analysis. *Matter* **7**, 1–15 (2024).

38. Clausen, C. M., Rossmeisl, J. & Ulissi, Z. W. Adapting OC20-trained equiformerV2 models for high-entropy materials. *J. Phys. Chem. C.* **128**, 11190–11195 (2024).
39. Nie, S. et al. Active learning guided discovery of high entropy oxides featuring High H_2-production. *J. Am. Chem. Soc.* **146**, 29325–29334 (2024).
40. Oliaei, H. & Aluru, N. R. Study of the adsorption sites of high entropy alloys for CO_2 reduction using graph convolutional network. *APL Mach. Learn.* **2**, 026103 (2024).
41. Xu, W., Diesen, E., He, T., Reuter, K. & Margraf, J. T. Discovering high entropy alloy electrocatalysts in vast composition spaces with multiobjective optimization. *J. Am. Chem. Soc.* **146**, 7698–7707 (2024).
42. Zhou, Q. et al. Analyzing the active site and predicting the overall activity of alloy catalysts. *J. Am. Chem. Soc.* **146**, 15167–15175 (2024).
43. Sheriff, K., Cao, Y. & Freitas, R. Chemical-motif characterization of short-range order with E(3)-equivariant graph neural networks. *npj Comput. Mater.* **10**, 215 (2024).
44. Jinnouchi, R. & Asahi, R. Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm. *J. Phys. Chem. Lett.* **8**, 4279–4283 (2017).
45. Deshmukh, G. et al. Active learning of ternary alloy structures and energies. *npj Comput. Mater.* **10**, 116 (2024).
46. Zhou, J. et al. Machine-learning-accelerated screening of Heusler alloys for nitrogen reduction reaction with graph neural network. *Appl. Surf. Sci.* **669**, 160519 (2024).
47. Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **4**, 37 (2018).
48. Bang, K. et al. Machine learning-enabled exploration of the electrochemical stability of real-scale metallic nanoparticles. *Nat. Commun.* **14**, 3004 (2023).
49. Calle-Vallejo, F., Martinez, J. I., Garcia-Lastra, J. M., Sautet, P. & Loffreda, D. Fast prediction of adsorption properties for platinum nanocatalysts with generalized coordination numbers. *Angew. Chem. Int. Ed.* **53**, 8316–8319 (2014).
50. Calle-Vallejo, F., Loffreda, D., Koper, M. T. & Sautet, P. Introducing structural sensitivity into adsorption-energy scaling relations by means of coordination numbers. *Nat. Chem.* **7**, 403–410 (2015).
51. Veličković, P. et al. Graph attention networks. In *International Conference on Learning Representations* (2018).
52. Geiger, M. & Smidt, T. e3nn: Euclidean neural networks. Preprint at https://arxiv.org/abs/2207.09453 (2022).
53. Pólya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Math.* **68**, 145–254 (1937).
54. Sheriff, K., Cao, Y., Smidt, T. & Freitas, R. Quantifying chemical short-range order in metallic alloys. *PNAS* **121**, e2322962121 (2024).
55. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).
56. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
57. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **21**, 395502 (2009).
58. Garrity, K. F., Bennett, J. W., Rabe, K. M. & Vanderbilt, D. Pseudopotentials for high-throughput DFT calculations. *Comput. Mater. Sci.* **81**, 446–452 (2014).
59. Wang, T., Li, G., Cui, X. & Abild-Pedersen, F. Identification of earth-abundant materials for selective dehydrogenation of light alkanes to olefins. *PNAS* **118**, e2024666118 (2021).
60. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
61. Batzner, S. et al. E(3)-Equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
62. Cai, C. & Wang, T. Resolving chemical-motif similarity with enhanced atomic structure representations for accurately predicting descriptors at metallic interfaces. https://doi.org/10.5281/zenodo.16713879 (Zenodo, 2025).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFA0911900), Zhejiang Provincial Natural Science Foundation of China (LR25B030001), Key R&D Program of Zhejiang (No. 2024SSYS0064) and the National Natural Science Foundation of China (22273076); T.W. thanks to the start-up packages from Westlake University, the Kunpeng research fund from Zhejiang Province, and the Research Center for Industries of the Future (RCIF) at Westlake University for supporting this work. We also thank Westlake University HPC Center for computational support.

Author contributions

T.W. conceived and supervised the project; C.C. implemented and developed the machine learning model and performed all the theoretical simulations; all authors discussed and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63860-x.

Correspondence and requests for materials should be addressed to Tao Wang.

Peer review information *Nature Communications* thanks Pengfei Ou and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.