ORIGINAL ARTICLE

Examining Generalizability of AI Models for Catalysis

Shih-Han Wang[1] | **Hongliang Xin**[1] | **Luke E.K. Achenie**[1] | **Kamal Choudhary**[2*]

[1]Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

[2]Material Measurement Laboratory, National Institute of Standards and Technology, Maryland, 20899, USA.

Correspondence
Kamal Choudhary
Email: kamal.choudhary@nist.gov

In this work, we investigate the generalizability of problem-specific machine-learning models for catalysis across different datasets and adsorbates, and examine the potential of unified models as pre-screening tools for density functional theory calculations. We develop graph neural network models for 12 different datasets for catalysis and then cross-evaluate their performance. Unified models include ALIGNN-FF, MATGL, CHGNet, and MACE. Pearson correlation coefficient analysis indicates that generalizability improves when similar adsorbates are used for training and testing or when a larger database is employed for training. Results demonstrate that while the accuracy of the unified models has room for improvement, their excellent performance in predicting the trend of adsorption energies can be a valuable pre-screening tool for selecting potential candidates prior to resource-intensive DFT calculations in catalyst design, thereby reducing computational expenses. The tools used in this work will be made available at: `https://github.com/usnistgov/catalysismat`.

KEYWORDS
Machine learning, graph neural network, catalyst, adsorption energy, generalizability

1 | INTRODUCTION

Catalysts are widely used in various industries, including medicine and oil refining, as well as necessities such as catalytic converters for automobiles [1]. In catalytic reactions, the adsorption strength of the reactant or its fragments plays a critical role in determining the reaction rate [2]. If the adsorption is too strong, the catalyst surface will be poisoned and lose its activity. On the other hand, if the adsorption is too weak, the reaction will fail to initiate effectively. The relationship between adsorption energy and reaction rate can be plotted as a volcano-shape graph [3]. The top of the volcano is where the adsorption strength of an ideal catalyst should be, with just the right adsorption strength. In addition, Pedersen *et al.* [4] previously demonstrated that there is a linear relationship between the adsorption energies of two similar adsorbates (*e.g.* O vs. OH, N vs. NH_x, C vs. CH_x and S vs. SH) on the same adsorption site on different surfaces. The adsorption energy is also linearly related to the free formation energies of intermediates and transition states [5]. Based on the above characteristics, with descriptor-based microkinetic modeling, the adsorption energy can be used as a descriptor to predict the reactivity of the catalyst [5, 6]. The adsorption energy (E_{ad}) can be calculated using *ab initio* quantum chemical methods, such as density functional theory (DFT) [7]. However, these methods are computationally expensive, making searching for potential candidates in a vast design space a daunting task [8].

In the past decade, machine learning (ML) methods have emerged as an alternative tool for designing novel catalysts [9, 10]. Using adsorption or potential energies obtained through DFT calculations as targets, ML models are trained to learn the relationship between systems and their corresponding energies, enabling the creation of reliable pre-trained ML models for candidate screening. There are three common ways to utilize ML models for adsorption energy prediction, namely from initial structure to adsorption energy [11] (IS2AE), from initial structure to relaxed energy [12] (IS2RE), and machine learning potential [8] (MLP) approaches. The IS2AE approach trains an ML model to directly predict the adsorption energy of an adsorbate-catalyst system based on the corresponding initial adsorbate-catalyst structure, which has been proven effective [13, 5]. The IS2RE method is a variation of the IS2AE method, which predicts the potential energies of the adsorbate-catalyst system ($E_{ads-cat}$) and the clean surface (E_{cat}) separately. The adsorption energy is the energy change when an adsorbate is adsorbed from a distance to a clean surface, where each adsorbate has a reference value for the energy of its gas phase adsorbate (E_{gas}). Hence, with predicted $E_{ads-cat}$ and E_{cat}, and pre-tabulated E_{gas}, the adsorption energy can be calculated using Equation 1 [14]:

$$\Delta E_{ads} = E_{ads-cat} - E_{cat} - E_{gas} \tag{1}$$

One of the advantages of these two approaches is that model training only requires the initial structure and the corresponding adsorption energy or potential energy, eliminating the need to train the model with forces and trajectory structures, thereby reducing the computational resources needed for training. For prediction, only the initial structure is required, and there is no need to use the relaxed structure as the input graph, which requires the use of DFT calculations. However, the same adsorbate may have multiple different adsorption configurations or adsorption sites on the same surface, such as top, hollow, and bridge sites. An initial structure can only represent a certain adsorption site and configuration, but cannot represent all adsorption sites and configurations simultaneously. Hence, these two approaches cannot capture subtle configuration differences [8] and are limited to simple adsorbates on specific sites. It is worth noting that different databases may place adsorbate at different initial distances from the adsorption site, and differences in the initial distances of the adsorbate from the adsorption site will lead to differences in the graphs. Since IS2AE and IS2RE are distance-sensitive methods, different graphs will result in different input graph features,

further affecting the predicted adsorption energies. Therefore, graphs from different databases may not be used together for training.

The MLP approach is a more complex and computationally expensive but robust approach to address this issue, which requires extensive training datasets so that ML models can learn the energy and forces of different elements in various local environments. The training dataset size required to construct an MLP model containing dozens of elements for potential energy prediction is in the hundreds of thousands, millions, or even larger. This not only requires a significant number of DFT calculations but also demands substantial graphical processing unit (GPU) resources to train the model [15]. One of the advantages of MLP approach is that well-trained MLP models can be used as a force field to perform molecular dynamics (MD) calculations. Given an initial configuration, the structure is then relaxed until it reaches a local minimum of the potential energy surface (PES), which represents the potential energy under that configuration. With Equation 1, by considering different initial adsorption sites and various initial configurations of adsorbates, it becomes possible to calculate the adsorption energies for different sites and configurations on the same catalyst surface, eventually obtaining the global minimum, which represents the catalyst adsorption energy [8, 16].

In recent years, researchers have published numerous ML models, such as crystal graph convolutional neural network (CGCNN) [17], directional message passing neural network (DimeNet) [18, 19], SchNet [20], geometric message passing neural network (GemNet) [21, 22] and atomistic line graph neural network (ALIGNN) [23]. Although some ML models include both IS2AE, IS2RE, and MLP methods, most researchers adopt the IS2AE approach in their studies, which is the simplest method for problem-specific studies. These studies focus on specific adsorbates on alloys such as H, CO, O, OH, N, COOH, and CHO [13, 24, 11]. Such studies require the generation of problem-specific DFT datasets, which remains computationally expensive. In addition to ML models open to the public, numerous researchers have also developed a significant number of databases related to catalysts and adsorption energies, such as the theory-infused neural network (TinNet) [13], automatic graph representation algorithm (AGRA) [24], open catalyst project 2020 and 2022 (OCP20 and OCP22) [15, 12], JARVIS-DFT [25] and materials project (MP) [26]. Researchers can utilize these databases for model training, validation and testing. Generally, a model trained on a larger dataset that encompasses diverse systems is better equipped to capture various local environments, enhancing its robustness [12].

However, it is important to note that different databases employ varying DFT calculators, convergence conditions, parameter settings, and energy references, which makes it possible for even the same system to have different adsorption energies. Consequently, it becomes challenging to utilize all public databases simultaneously to train a universal model. Figure 1 illustrates the above three cases. Figure 1 (a) shows problem-specific case, which requires a relatively small training dataset (≈1,000 systems). However, since it is problem-specific, in most cases, researchers have to prepare the database by themselves, which can take several hundreds of CPU-core-hours for DFT calculations. Even with the help of supercomputers, this is still an expensive task. However, a relatively small training dataset makes training ML models very cheap. Even using a personal computer, training can be completed in a few GPU-hours. Furthermore, since the training and test systems are very similar to each other, the mean absolute error (MAE) of the pre-trained model on the test dataset is usually only about 0.1-0.2 eV [13].

Figure 1 (b,c) use OCP20 database [15], JARVIS-DFT and MP databases as examples for developing domain specific and unified MLP models. Although the OCP20 database contains hundreds of millions of images and took hundreds of billions of CPU-core-hours to construct, it is free and open, so researchers do not need to spend additional resources to prepare training datasets [8, 15]. But also because the training dataset is very large, the computational budget required to train a model ranges from hundreds to tens of thousands of GPU-hours, which is also the bottleneck of adopting these approaches.

Figure 1 (c) adopts JARVIS-DFT database [25, 27] and MP for MLP model training. While machine learning force

fields (MLFF)/machine learning potentials (MLP), have been developed for specific systems, recently a unified models were developed that can predict energies and forces of structurally and chemically diverse material systems across the periodic table for 89 elements. Such models include ALIGNN-FF [28], MATGL [29], CHGNet [30] and MACE[31]. We called the well-trained model on a large training dataset with periodic table elements as unified models. The MAE of potential energy and adsorption energy are on the order of 0.08 eV/atom and 0.8 eV, respectively.

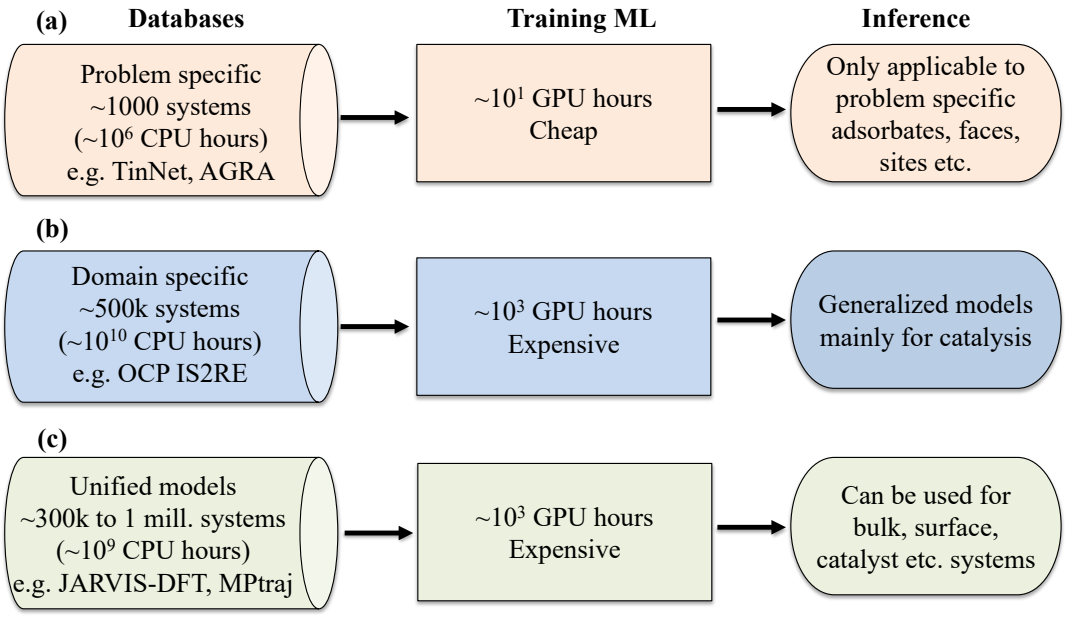

FIGURE 1 A schematic for process of preparing database and training: a) a problem-specific catalysis model, b) a large domain-specific model, c) unified force-field based models.

In this work, we aim to address the following questions: (1) How generalizable are these problem-specific models? Can pre-trained problem-specific models be used on different datasets or different adsorbates? (2) Can unified models trained on larger datasets accurately predict E_{ad} for smaller datasets? (3) Can unified models obviate the need for generating problem-specific catalyst datasets, allowing these large models to serve as screening tools for DFT calculations? In this work, we attempt to answer these important questions by systematically developing new models or utilizing the existing ones for catalyst application. This work also paves the way for carrying out such generalizability studies for other domain tasks beyond catalysis. The data and tools used in this work will be made publicly available at 1) GitHub (`https://github.com/usnistgov/catalysismat`), 2) JARVIS-Leaderboard [32] (`https://pages.nist.gov/jarvis_leaderboard/`), 3) JARVIS-web app[25, 27] for Catalysis (`https://jarvis.nist.gov/jcatalysis/`)

2 | METHODOLOGY

In this work, we utilize ALIGNN as the graph neural network (GNN) model, performing training and testing on the TinNet, AGRA, and OCP20 IS2RE datasets for problem-specific model development, while also evaluating ALIGNN-FF, MATGL, MACE, and CHGNet as unified models to assess the generalizability of machine learning models in predicting adsorption energies. Among the state-of-the-art ML models, ALIGNN has been shown to perform better than all other models in most cases [33]. ALIGNN models are trained with IS2AE method on TinNet and AGRA datasets. TinNet has N, O and OH based datasets with 329, 747 and 748 entries respectively. AGRA contains O, OH, CO, CHO and COOH based datasets with 1,000, 875, 193, 214, and 280 samples respectively. On the OCP20 IS2RE dataset, IS2RE approach and the equation 1 are used. For the pre-trained ALIGNN-FF wt01 dataset, an MLP model is trained with a weight of 0.1 for the force term [28, 25]. The ALIGNN-FF model was trained on 307,111 bulk material entries from the JARVIS-DFT dataset. However, for the test dataset, the ALIGNN-FF wt01 model was only used to predict the potential energy of the unrelaxed structure, and no MD calculations were performed to relax the structure. Similar unified force-field models such as MATGL [29], CHGNet [30] and MACE[31] were trained on materials project dataset with 1.5 million structural entries.

All training uses ALIGNN model architecture [23, 28]. In ALIGNN, a crystal structure is represented as a graph by mapping atoms at the lattice sites to nodes and bonds to the edges of the graph. Each node in the crystal graph is assigned 10 input node features based on its atomic species: electronegativity, group number, period number, covalent radius, number of s, p, and d valence electrons, first ionization energy, electron affinity, and the number of unpaired electrons. The inter-atomic bond distances are used as edge features with radial basis function up to 8 Å cut-off and 12 nearest-neighbors. This crystal graph is then used for constructing the corresponding line graph that encodes interatomic bond-distances as nodes and bond-angles as edge features. The ALIGNN model learns the relationship between a graph and the corresponding target property , *i.e.*, the adsorption energy of a specific adsorbate on a fully relaxed surface or the potential energy of a catalyst system. To evaluate generalizability, the ALIGNN model was trained with a specific dataset first and the pre-trained model was used to predict adsorption energies of unknown systems in other datasets.

3 | RESULTS AND DISCUSSION

While the ALIGNN (for direct property prediction) models have been shown to perform well for many tasks, we evaluate their performance for catalytic properties before evaluating the generalizability of the models for catalysis. In Table 1 , we benchmark the performance of the ALIGNN model on OCP20 and OCP22 datasets. Furthermore, Table 1 reports different model performances on different databases, where data come from Open Catalyst Project [34]. Among the models using the same training dataset in the OCP20 and OCP22 leaderboards, ALIGNN is by far the best performing model in most cases [34, 35, 32]. We find that ALIGNN outperforms many other GNN models such as CGCNN [17], DimeNet [18], DimeNet++ [19], SchNet [20], coGN [36], PaiNN [37] etc. for OCP20-10k, OCP20-100k, OCP20-all (510,214 entries) as well as the latest OCP22-direct only (45,890 entries) datasets. All of these datasets have well defined training-validation-test splits as described in the open catalyst project [35].

After establishing the model strength, now we examine the generalizability of such models across different training/test sets. Such analysis provides insight into the extrapolation strength of unified AI models [38]. For this study, Pearson correlation coefficients (PCC) is used as a criterion to evaluate the model generalizability. The definition of PCC is the covariance of the two variables divided by the product of their standard deviations, which is shown in

TABLE 1 Comparison of ML models on OCP20 and OCP22 datasets [34]

Database	Model	Energy MAE (eV)	Database	Model	Energy MAE (eV)
OCP20-10k	CGCNN	0.9881	OCP20-100k	CGCNN	0.682
	DimeNet	1.0117		DimeNet	0.6658
	SchNet	1.059		SchNet	0.7137
	DimeNet++	0.8837		DimeNet++	0.6388
	ALIGNN	**0.7623**		ALIGNN	**0.6289**
OCP20-All	CGCNN	0.6199	Direct OCP22-only	SchNet	3.424
	DimeNet	0.5999		DimeNet++	2.7393
	SchNet	0.6458		PaiNN	2.6997
	DimeNet++	**0.5639**		GemNet-dT	2.3804
	ALIGNN	0.5989		coGN	2.2121
	PaiNN	0.5728		ALIGNN	**2.1945**

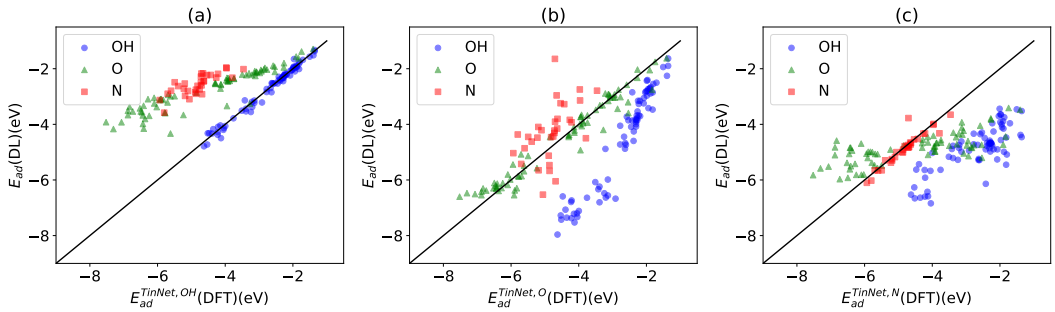

FIGURE 2 Parity plots are shown for (a) DFT-calculated adsorption energies on the held-out TinNet-OH test set versus predictions from various ALIGNN/deep learning (DL) models. These DL models were trained on TinNet-OH training data (blue), TinNet-O training data (green), and TinNet-N training data (red) and then tested on the TinNet-OH test set. Panels (b) and (c) display similar plots for DL models tested on the TinNet-O and TinNet-N test datasets, respectively. The corresponding Pearson correlation coefficients are provided in Table 2.

Equation 2. A PCC value of +1/-1 indicates perfect positive/negative correlation, while 0 signifies no correlation. When the pre-trained model exhibits perfect generalizability to the test dataset, the PCC approaches +1. Catalysis specific databases such as TiNet, AGRA, OCP20 have well defined training and test splits. For the generalizability study, we train a model on a specific train set (*e.g.*, TinNet N) and test the model on test sets of all the databases (such as TinNet N, AGRA OH, OCP20 etc.). Note that unified models such as ALIGNN-FF model was trained on bulk materials data and different DFT computational parameters, but we test that model here as well. The PCC values for the generalizability study is shown in Table 2. The bold number values represent the highest PCC values.

TABLE 2 Pearson Correlation Coefficient (PCC) values representing the generalizability of GNN models in catalysis-related datasets. The table compares problem-specific GNN models (first eight rows) with generalized or unified models. The values provide insights into the performance and transferability of GNN models across different adsorption environments, illustrating their potential to predict catalytic behavior in both targeted and broad catalytic systems. An average value for each row is included to highlight the overall accuracy and robustness of each model, along with a column average to summarize general trends across the datasets. A live version of this table will also be made available on the JARVIS-Leaderboard.

Database	TinNet			AGRA					Avg
Adsorbate	N	O	OH	O	OH	CO	CHO	COOH	
TinNet N (329), ALIGNN	**0.94**	0.73	0.80	0.52	0.54	0.28	-0.04	0.03	0.52
TinNet O (747), ALIGNN	0.51	**0.98**	0.94	0.75	0.96	0.10	0.82	0.43	0.69
TinNet OH (748), ALIGNN	0.78	0.93	**1.00**	0.79	0.96	-0.11	0.82	0.53	0.72
AGRA O (1000), ALIGNN	0.30	0.16	0.14	**1.00**	0.92	0.38	0.03	0.32	0.41
AGRA OH (875), ALIGNN	0.35	0.23	0.14	0.86	**1.00**	0.15	0.12	0.21	0.38
AGRA CO (193), ALIGNN	0.10	0.41	-0.12	0.10	0.04	**0.82**	0.86	-0.34	0.24
AGRA CHO (214), ALIGNN	0.45	0.21	0.05	0.12	-0.39	0.32	**0.97**	0.66	0.30
AGRA COOH (280), ALIGNN	0.40	0.40	0.34	0.14	-0.32	0.59	0.92	**0.90**	0.42
OCP20 IS2RE (510214), ALIGNN	0.50	**0.86**	0.64	0.82	0.70	0.10	**0.86**	0.03	0.56
JARVIS-DFT (307113), ALIGNN-FF	0.71	0.79	0.31	**0.80**	0.52	-0.38	0.56	0.21	0.44
MPtraj (1.5mill), MATGL	0.82	0.51	-0.41	**0.81**	0.74	0.37	-0.41	0.15	0.32
MPtraj (1.5mill), CHGNet	0.90	0.93	0.78	**0.95**	0.94	0.54	0.92	0.69	**0.83**
MPtraj (1.5mill), MACE	0.83	0.59	-0.39	**0.21**	0.71	0.06	-0.35	-0.18	0.19
Column Avg	0.61	**0.64**	0.39	**0.64**	0.61	0.26	0.53	0.32	-

Figure 2 (a) shows the parity plot of DFT-calculated adsorption energies of O versus OH at the top site of {111}-terminated alloy surfaces *i.e.* we train the model for TinNet-Oxygen-adsorption dataset and use that model to directly predict the adsorption energy for OH-catalyst systems to evaluate generalizability. We observe a linear relationship exists between the adsorption energies of two similar adsorbates on the same adsorption site though the values do not lie on the x = y line. Similarly, Figure 2 (b-d) show parity plots of DFT-calculated adsorption energies of N, O, and OH systems versus ALIGNN-predicted adsorption energies by pre-training models on N, O or OH systems. Taking O and OH systems as an example, there is only a single hydrogen difference between the O and OH systems. This difference has a minimal impact on the input features of the ML model, leading to a slight bias in the predicted adsorption energy in a specific direction. However, due to the similarity between the two systems and the linear relationship in the DFT data, a model trained on the O dataset exhibits a higher PCC value when used to predict the OH system, and vice versa. Even if the MAE is relatively large, the model can still capture the changing trend of adsorption energy accurately. The PCC values between different training datasets and test datasets are shown in Table 2. Key insights from Table 2 are as follows:

(1) Since systems in the same database use the same DFT calculation settings and these systems are relatively

similar, it is easier to obtain a high PCC value when we use the same database for training and testing (along th diagonal line of the top portion of table 2), hence, when the model is well-trained, the PCC value will be very close to +1. The off-diagonal values represent the out of domain generalizability PCC scores.

(2) TinNet datasets contain a diverse number of images than AGRA datasets. If the model is trained with the TinNet dataset, it will demonstrate good generalizability to the same adsorbate AGRA dataset, but the reverse is not necessarily true.

(3) Since both TinNet and AGRA datasets are adsorbate-specific, the model generalizability suffers when the adsorbate differs significantly, such as between N and COOH. Conversely, good generalizability is observed between similar adsorbates, such as O and OH, or CHO and COOH.

(4) The OCP20 dataset is an extensive collection of adsorption energy related data, encompassing tens of adsorbates and hundreds of thousands of structures. An ALIGNN model trained on the OCP20 IS2RE all dataset and then applied to predict adsorption energies for systems in the TinNet and AGRA datasets, which were not part of the training data, demonstrates good generalizability for most datasets. This supports the hypothesis that training on a large dataset enables the model to capture a wide range of local environments, enhancing its generalizability. However, this is not always the case, as seen with AGRA CO and COOH datasets.

(5) The ALIGNN-FF wt01 dataset is notably smaller than the OCP20 IS2RE dataset and based on bulk materials only. Despite this, it still exhibits some generalizability to unknown systems, even though ALIGNN-FF wt01 is a completely different dataset from the unknown systems.

(6) It is interesting to note that in (a) and (b) of Fig. 2, O and N predicted data points (red squares and green triangles) are closer to each other.

(7) The PCC values in some combinations are close to zero or even negative. This is because the predictions are based on the initial structure rather than the relaxed structure, so the predicted values will be affected by the different initial positions of the adsorbate in different datasets.

The row-wise average denotes overall PCC scores across different dataset. This average is highest for TinNet OH/ALIGNN model with a value of 0.72. The least value is observed for AGRA CO/ALIGNN model with 0.24. OCP20 IS2RE/ALIGNN model provides domain specific model that can predict energy of arbitrary composition and structures. This model performs consistently well for all the datasets with an average PCC value of 0.56.

While the above problem specific/domain specific models were developed using surface and adsorbate dataset, unified force-field models were trained on bulk materials only. Interestingly, we find that CHGNet model achieves highest average PCC score among all the models followed by ALIGNN-FF (0.44), MATGL (0.32) and MACE (0.19). The column wise average shown as the last row suggests that all these models perform comparatively well on oxygen catalyst datasets (TinNet o, AGRA O) and worst on AGRA CO dataset. This suggests the unified models can be indeed useful for general purpose catalyst design.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

Next, in order to democratize the trained model for pre-screening application, we developed a web-app [39] to quickly predict adsorption energy given the substrate and catalyst (*i.e.* substrate and adsorbate) atomic configurations. A user submits these two entries, which are then converted to respective atomic graphs. These two separate graphs are fed to pre-trained OCP20 IS2RE and ALIGNN-FF models which quickly predict their total energy. Both OCP20 IS2RE and ALIGNN-FF provide element specific chemical potentials, which are then used in the Equation 1 to predict

the adsorption energies. Currently, there is no geometric optimization performed with the app, but in the future, we plan to add that capability as well. It is worth mentioning that Ulissi *et al*. [8] are working on MLP models. The published model shows very credible adsorption energy prediction capabilities, but the CPU-core-hours and GPU-hours required for training models with 134M images are beyond the reach of our equipment, so this work only adopted relatively small datasets for training.

4 | CONCLUSION

In summary, we have answered some of the most important questions for evaluating the generalizability of AI models for catalysis. Our analysis reveals the following: (1) A pre-trained model exhibits good generalizability to unknown systems when two systems share high similarity. However, this hinges on the condition that the initial training dataset is sufficiently large and diverse. Small training datasets, as demonstrated by AGRA CO, may result in pre-trained models lacking fundamental predictive capabilities, (2) employing a substantial training dataset to train a unified model allows for reasonable prediction of adsorption energy trends for smaller, problem-specific datasets, as evidenced by the obtained PCC values, (3) in catalyst design, the difference and trends of adsorption energy predictions holds greater importance than the absolute values, as highlighted in [16]. Therefore, although the unified model primarily predicts the change trend in adsorption energy rather than accurately predicts the adsorption energy, it remains highly valuable in catalyst design, because the unified model has the ability to rapidly assess whether a specific system exhibits stronger or weaker adsorption strength than the reference system. Unified models are good to circumventing the need to generate problem-specific catalysts datasets as a screening tool for DFT calculations. Even when DFT calculations are ultimately required for precise predictions of candidate system adsorption energies, utilizing a unified model for pre-screening the design space substantially reduces the computational workload associated with DFT calculations. Therefore, this work underscores the value and utility of unified models as pre-screening tools for DFT calculations in catalyst design. A web application is available online that includes pre-trained ALIGNN models that allow the user to enter a user-specified catalyst system and use models to predict adsorption energies. Although this web application uses an MLP model, due to insufficient server computing resources, the web application can only calculate the corresponding adsorption energy based on the structure provided by the user, and can be further strengthened in the future.

Conflict of interest

The authors declare no conflict of interests.

Acknowledgements

S.H.W., L.E.K.A., H.X. and K.C. acknowledge the partial financial support from the NSF CAREER program. The computational resource used in this work is provided by the advanced research computing at Virginia Polytechnic Institute and State University, and Nisaba HPC cluster at National Institute of Standards and Technology.

references

[1] Chen BW, Xu L, Mavrikakis M. Computational methods in heterogeneous catalysis. Chemical Reviews 2020;121(2):1007–1048.

[2] Norsko J. Chemisorption on metal surfaces. Reports on Progress in Physics 1990;53(10):1253.

[3] Nørskov JK, Bligaard T, Rossmeisl J, Christensen CH. Towards the computational design of solid catalysts. Nature chemistry 2009;1(1):37–46.

[4] Abild-Pedersen F, Greeley J, Studt F, Rossmeisl J, Munter TR, Moses PG, et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. Physical review letters 2007;99(1):016105.

[5] Pillai HS, Li Y, Wang SH, Omidvar N, Mu Q, Achenie LE, et al. Interpretable design of Ir-free trimetallic electrocatalysts for ammonia oxidation with graph neural networks. Nature Communications 2023;14(1):792.

[6] Wang Y, Yang X, Xiao L, Qi Y, Yang J, Zhu YA, et al. Descriptor-based microkinetic modeling and catalyst screening for CO hydrogenation. ACS Catalysis 2021;11(23):14545–14560.

[7] Sholl DS, Steckel JA. Density functional theory: a practical introduction. John Wiley & Sons; 2022.

[8] Lan J, Palizhati A, Shuaibi M, Wood BM, Wander B, Das A, et al. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. npj Computational Materials 2023;9(1):172.

[9] Yang W, Fidelis TT, Sun WH. Machine learning in catalysis, from proposal to practicing. ACS omega 2019;5(1):83–88.

[10] Goldsmith BR, Esterhuizen J, Liu JX, Bartel CJ, Sutton C. Machine learning for heterogeneous catalyst design and discovery. John Wiley & Sons; 2018.

[11] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. Nature Catalysis 2018;1(9):696–703.

[12] Tran R, Lan J, Shuaibi M, Wood BM, Goyal S, Das A, et al. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. ACS Catalysis 2023;13(5):3066–3084.

[13] Wang SH, Pillai HS, Wang S, Achenie LE, Xin H. Infusing theory into deep learning for interpretable reactivity prediction. Nature communications 2021;12(1):5288.

[14] Ulissi ZW, Tang MT, Xiao J, Liu X, Torelli DA, Karamad M, et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO2 reduction. Acs Catalysis 2017;7(10):6600–6608.

[15] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, et al. Open catalyst 2020 (OC20) dataset and community challenges. Acs Catalysis 2021;11(10):6059–6072.

[16] Ock J, Tian T, Kitchin J, Ulissi Z. Beyond independent error assumptions in large GNN atomistic models. The Journal of Chemical Physics 2023;158(21).

[17] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters 2018;120(14):145301.

[18] Gasteiger J, Groß J, Gunnemann S. Directional Message Passing for Molecular Graphs. In: International Conference on Learning Representations (ICLR); 2020. .

[19] Gasteiger J, Giri S, Margraf JT, Gunnemann S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. In: Machine Learning for Molecules Workshop, NeurIPS; 2020. .

[20] Schutt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Muller KR. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems 2017;30.

[21] Gasteiger J, Becker F, Gunnemann S. Gemnet: Universal directional graph neural networks for molecules. Advances in Neural Information Processing Systems 2021;34:6790–6802.

[22] Gasteiger J, Shuaibi M, Sriram A, Gunnemann S, Ulissi Z, Zitnick CL, et al. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. arXiv preprint arXiv:220402782 2022;.

[23] Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Computational Materials 2021;7(1):185.

[24] Gariepy Z, Chen Z, Tamblyn I, Singh CV, Tetsassi Feugmo CG. Automatic graph representation algorithm for heterogeneous catalysis. APL Machine Learning 2023;1(3).

[25] Choudhary K, Garrity KF, Reid AC, DeCost B, Biacchi AJ, Hight Walker AR, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj computational materials 2020;6(1):173.

[26] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL materials 2013;1(1).

[27] Wines D, Gurunathan R, Garrity KF, DeCost B, Biacchi AJ, Tavazza F, et al. Recent progress in the JARVIS infrastructure for next-generation data-driven materials design. Applied Physics Reviews 2023 10;10(4):041302.

[28] Choudhary K, DeCost B, Major L, Butler K, Thiyagalingam J, Tavazza F. Unified graph neural network force-field for the periodic table: solid state applications. Digital Discovery 2023;2(2):346–355.

[29] Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science 2022;2(11):718–728.

[30] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel CJ, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence 2023;5(9):1031–1041.

[31] Batatia I, Kovacs DP, Simm G, Ortner C, Csanyi G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in Neural Information Processing Systems 2022;35:11423–11436.

[32] Choudhary K, Wines D, Li K, Garrity KF, Gupta V, Romero AH, et al. JARVIS-Leaderboard: a large scale benchmark of materials design methods. npj Computational Materials 2024;10(1):93.

[33] Du Y, Wang Y, Huang Y, Li JC, Zhu Y, Xie T, et al. M2Hub: Unlocking the Potential of Machine Learning for Materials Discovery. arXiv preprint arXiv:230705378 2023;.

[34] Open Catalyst Project;. Accessed: 2023-10-10. https://opencatalystproject.org/index.html.

[35] Open Catalyst Project GitHub;. Accessed: 2023-10-10. https://github.com/Open-Catalyst-Project.

[36] Ruff R, Reiser P, Stuhmer J, Friederich P. Connectivity Optimized Nested Graph Networks for Crystal Structures. arXiv preprint arXiv:230214102 2023;.

[37] Schutt K, Unke O, Gastegger M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In: International Conference on Machine Learning PMLR; 2021. p. 9377–9388.

[38] Choudhary K, Sumpter BG. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? AIP Advances 2023;13(9).

[39] JARVIS Catalysis;. Accessed: 2023-10-10. https://jarvis.nist.gov/jcatalysis/.