THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

Predicting the Enthalpy of Hydrocarbon Radicals Adsorbed on Pt(111) Using Molecular Fingerprints and Machine Learning

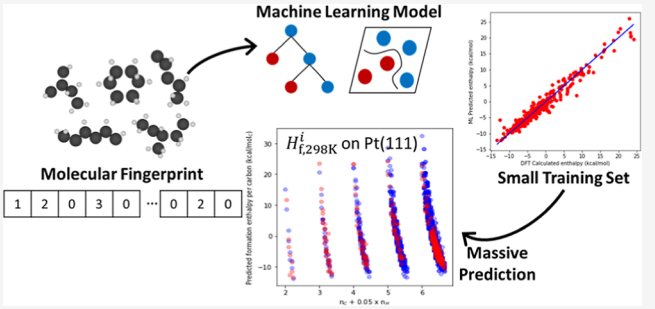
Jinwoong Nam, Charanyadevi Ramasamy, Daniel E. Raser, Gustavo L. Barbosa Couto, Lydia Thies, David Hibbitts, and Fuat E. Celik*

Cite This: J. Phys. Chem. C 2024, 128, 5030−5043

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The reliable prediction of properties for the adsorbates, including their enthalpy, has been a long-standing challenge as a first key step in studying surface reactions. It is especially difficult when large adsorbates are involved as the interactions between the adsorbates and surface atoms are complex. Here, we developed machine learning (ML) models for the prediction of the formation enthalpy of various C_2 to C_6 hydrocarbon adsorbates on the Pt(111) surface based on 384 density functional theory calculations. Focusing on larger and more intricate adsorbates, two-thirds of the total species were C_6 species. Four molecular descriptors that represent the valency and bonding of individual carbons within the adsorbates were generated without intensive computation. They were subsequently used as the features of the ML models with three linear and four nonlinear algorithms. The models were developed with 30 different samplings of train/test sets, and their results were statistically analyzed to ensure the performance of the models. Nonlinear models, especially kernel ridge regression and extreme gradient boosting, outperformed linear models with lower absolute errors. The top two accurate models, based on these algorithms, also displayed remarkable robustness in predicting various species. Employing ensemble average voting with these two models, we achieved the lowest mean absolute error of 0.94 kcal/mol$_C$. Finally, ML was used to estimate the formation enthalpy of 3115 hydrocarbon adsorbates on Pt(111), highlighting the promise of these methods to study more complicated reaction networks.



1. INTRODUCTION

One of the biggest challenges in computational catalysis arises from the intricate nature of catalytic surface reactions, where the investigation of numerous surface intermediates and active sites may be required. As a first step to understanding their surface chemistry, it is critical to reliably obtain the thermochemical properties of the adsorbates, including enthalpy, since they are crucial inputs to microkinetic models for the reactions of interest. While many studies have relied on quantum-mechanical computation such as density functional theory (DFT) calculations to derive their thermochemical properties,[1] the sole use of computationally expensive DFT calculations would be implausible when a great number of adsorbates are involved. Alternatively, approaches combining quantum-chemical calculations with machine learning (ML) have recently emerged and received much attention, showing the promise for reasonable estimation of adsorption properties with affordable computational efforts.[2−21] For example, a vast range of catalytic materials or active sites, including different facets of metals or metal alloys, have been examined by leveraging estimation of the adsorption energy of small adsorbates with ML models.[2−9] Ideally, we would anticipate significantly accelerated and more exhaustive screening of the active sites, surpassing the limitations imposed by traditional linear scaling relationships.[22−24] In addition, other studies have sought to develop generalized ML potentials that seek to replicate the quantum mechanical information obtained from DFT without the computational cost involved.[3,10,11]

Despite extensive prior studies, challenges still exist in predicting adsorption properties using ML. This encompasses the limited examination of adsorbates to relatively simple species, such as carbon, oxygen, nitrogen, and carbon monoxide, among other examples, as opposed to exploring the larger and more complex species involved in many industrially relevant reactions. Such reactions include Fischer−Tropsch synthesis of long-chain hydrocarbons[25,26] and coke formation during light alkane dehydrogenation,[23,27,28] to name a few, for which hydrocarbons ranging from C_1 to C_6+ would be surface intermediates. In these reactions, the number

Received: December 17, 2023
Revised: February 25, 2024
Accepted: February 27, 2024
Published: March 18, 2024

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

of alkane-derived surface radicals increases rapidly with the number of carbons. For example, considering only C−H and C−C bond formation and cleavage of C_1 to C_6 hydrocarbons, we obtain an estimate of more than 3000 surface intermediates and 38,000 elementary reactions by using the Rule Input Network Generator (RING).[29] As such, considerable computational costs are expected to study them, even with a single-facet model. However, it is unclear whether the ML models developed by small species could reliably be scaled to larger adsorbates.

Furthermore, the ML models in previous studies have often employed electronic features such as information from the site-projected density of states,[12−14] e.g., d-band center,[15] Coulomb matrix,[16] or the geometric features like bond lengths in the relaxed structure.[17] However, given that the acquisition of these descriptors requires the use of computationally intensive techniques such as DFT or other semiempirical methods, their inclusion may be undesirable for predictive analysis with ML. This can be especially problematic when we intend to predict the adsorption energies of large adsorbates. As the size of the adsorbates increases, the number of combinatorial cases for DFT calculations to find their minimum energy and structure will multiply significantly. Moreover, relaxation of large multidentate adsorbate structures is subject to intricate intra-adsorbate and surface−adsorbate interactions, presenting a formidable challenge in reducing the number of calculation cases by proposing selected candidate structures relying on simple adsorption rules. For instance, the preferred sp^3-hybridized carbon structure of C_1 and C_2 hydrocarbons on Pt(111) found in our previous studies[30,31] may no longer apply for larger adsorbates where some carbons are unable to interact with the surface. Throughout this study, we encountered cases in which the best binding geometries displayed more complicated structures that are not found in smaller species.

Additionally, when evaluating the performance of ML models, randomly splitting the available datasets into training sets and test sets could introduce data sampling bias, especially in cases where the available datasets are small and/or not all numerical values of descriptors are evenly distributed throughout the datasets. Random sampling can inadvertently over- or under-represent some descriptor values in training sets, leading to larger than expected errors in the test sets. A more systematic approach would be to perform multiple trials with differently randomized training/test sets.[9,15,16,32] While some previous studies implemented a single randomized train/test split,[17,19,20] we have implemented 30 randomized trials followed by statistical analyses.

In this work, our primary goal is to predict the formation enthalpy of hydrocarbon radicals adsorbed on Pt(111) with ML models while seeking to tackle the aforementioned challenges. We developed ML models with C_2 to C_6 hydrocarbon species on Pt(111) using four molecular fingerprints describing the local valency and bonding around each carbon within the adsorbates. Here, the descriptors used for ML models are readily generated without the need for DFT calculations. C_6 species account for two-thirds of the total dataset with which ML models were developed, aiming for reasonable enthalpy prediction for larger hydrocarbon adsorbates. In addition to linear models serving as benchmarks, kernel-trick and tree-based ML models were assessed. Statistical hypothesis testing of the distribution of error metrics across multiple trials of randomized samplings was employed to properly assess the models' performance, sensitivity, and overall robustness. Then, an ensemble voting method was implemented for two independent individual models to further reduce the bias and obtain the best predictive performance. Lastly, as a case study, the enthalpies were estimated for a comprehensive list of 3115 C_2−C_6 acyclic hydrocarbons on Pt(111) using our ML models, from which the low-energy species of interest in reaction pathway analysis can be suggested. Overall, we discuss the sensitivity of the descriptors, the performance, validity, and robustness of the ML models based on chemistry and ML algorithms, and the promise of employing ML models to navigate potential energy surfaces.

2. METHODS

2.1. DFT and Enthalpy Calculations. Periodic DFT calculations were performed to calculate the energy of the adsorbates using the Vienna Ab initio Simulation Package (VASP).[33,34] The Perdew−Burke−Ernzerhof functional with the dispersion correction (PBE-D3) was used to compute electron−electron exchange and correlation.[35,36] The interactions between ion cores and valence electrons were considered using the projector augmented wave,[37,38] and the plane wave energy cutoff was 400 eV. The metal slab was based on the Pt(111) surface of the fcc structure and modeled by a (4×4) surface unit cell with four atomic layers for a total of 64 Pt atoms. A vacuum layer of 12 Å was applied to separate any two successive slabs in the z-direction (normal to the surface). The Brillouin zone was sampled using a $(4 \times 4 \times 1)$ Gamma-centered Monkhorst−Pack k-point mesh,[39] following a convergence test for adsorption energies for sampling mesh size. The bottom two layers of each metal slab were fixed in their bulk positions, while the top two layers were allowed to relax in all calculations.

Based on the energies calculated from DFT, enthalpies of formation ($H_{f,298\ K}$) for 384 hydrocarbon adsorbates were computed with the equation below using pMuTT[40]

$$H^i_{f,298K} = E^i_{ads,0K} - E_{slab,0K} - n^i_C E_C - n^i_H E_H + E^i_{vib,298K}$$

$H^i_{f,298\ K}$ is the enthalpy of the formation of adsorbate i. $E^i_{ads,0K}$ and $E_{slab,0K}$ are the total energies for the slab with adsorbate i on it and the clean slab at 0 K, respectively, both from the DFT calculation. The referencing terms, $n^i_C E_C$ and $n^H_i E_H$, consist of the number of each element ($n^i_C E_C$, n^H_i) in adsorbate i multiplied by its energetic adjustment (E_C, E_H). The detailed information on referencing, including the calculations of E_C and E_H, can be found in Section S2 of the Supporting Information.[41] $H^i_{Vib,298\ K}$ is the temperature contribution from the vibrational frequencies of adsorbate i, including the zero-point energy correction. Finally, the calculated $H^i_{f,298\ K}$ was normalized by the number of carbon atoms, yielding the enthalpy per carbon. This was done to facilitate comparison across species of different molecular weights, but it should be noted that the enthalpy per $mol_{species}$ varies, e.g., 1 kcal/mol_C of enthalpy corresponds to 2 kcal/mol for a C_2 species and 6 kcal/mol for a C_6 species. Throughout this study, the formation enthalpy per carbon was used as a target property for the ML models to properly compare the enthalpies on the same scale between the species with different carbon numbers.

2.2. Molecular Fingerprints. The adsorbates used in this study were described with four molecular fingerprints, and they subsequently served as feature sets for the ML models. It was reported in the previous ML study with small adsorbates that

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

the features related to "free" adsorbates affected the most in the prediction of binding energies.[20] It includes the valency, molecular weight, and number of bonds of the main element of the adsorbates, the contribution of which exceeds any other features derived from adsorption sites. This finding was transferred and expanded to our chemistry, where the information on the adsorption site of larger species is limited. Here, we introduce four molecular fingerprints in which local valency and chemical bond information on carbons in the adsorbates are represented in different ways. A binary acyclic/cyclic descriptor was incorporated in all four fingerprints described in the following sections since the dataset has both acyclic and cyclic species. It should be highlighted that the molecular fingerprints used can easily be generated from chemical notations of the adsorbates, such as SMILES (simplified molecular input line entry system)[42] because only information about the free adsorbates is needed.

2.2.1. Group Additivity Fingerprint. Group additivity fingerprint (GA) schemes[41,43−46] have been used since 1958 to parametrize molecular structure by decomposing molecules into their constituent groups. For our hydrocarbon adsorbates, each group consists of carbon and its nearest bonded atoms, including other carbons, hydrogens, and free valencies. Ten distinct groups are necessary to describe the whole set of adsorbates, as shown in Table S1. These groups were systematically assigned for each species using RING, which implements the group assignments to the entire list of species and outputs the GA fingerprints. An example is depicted in Figure S1.

2.2.2. Group Additivity with Surface Structure. We obtained the next molecular fingerprint, GA with surface structure (GASS), by adding additional terms to account for surface distortion.[41,46] When applying GA for surface species, it is often reasonable to add some correction groups to consider the distorted structure of adsorbates caused by a confined surface environment or surface strain. As shown in Figure S2, vinyl is adsorbed on Pt(111), forming 3 C−Pt bonds with surface Pt atoms. While an ideal sp^3-hybridized carbon structure has a bond angle of 109.5° in the gas phase, one of the bond angles measured in the vinyl adsorbate on Pt(111) is 96.3° due to deformation of the bond angle caused by surface ring strain. To account for this surface strain effect the adsorbates experience, five correction groups (group ID: C01 to C05) were added, as shown in Figure S2. These groups consist of two carbons sharing a C−C bond, both with at least one free valency. The groups vary only in the number of C−H bonds on each carbon in the group. It should be noted that the correction groups are taken from the free adsorbates without any surface binding geometry information. An example of the GASS is illustrated in Figure S3.

GA is flexible enough to incorporate even more correction groups beyond GASS, but a priori selection of the most important groups to include is difficult without observing the surface ring structures and strains that the relaxed adsorbates exhibit in DFT calculations.

2.2.3. Flat Molecular Fingerprint. Flat molecular fingerprint (FMF) is another way of describing adsorbates by counting the types of carbons based on free valency and the number of bonds between them.[47] In this method, carbons are classified into four types; C_0 to C_3, where C_0 is a saturated carbon with no free valency, and C_1, C_2, and C_3 are carbons with one, two, and three free valencies, respectively. Then, the number of each type of carbon is counted, followed by the number of C−H bonds and C_x−C_y (x = 0 to 3 and y = 0 to 3) bonds within the adsorbates. An example of FMF is depicted in Figure S4.

2.2.4. Sequential Valency-Connectivity Fingerprint. While the FMF displays a total count of different types of carbons and bonds, it lacks information about their sequential ordering. Alternatively, we have built the sequential valency-connectivity fingerprint (SVCF). In the SVCF, the carbon numbers are initially designated based on the rules primarily derived from the IUPAC nomenclature guidelines[48] (see Section 3.4.1 of Supporting Information for the detailed rules). Then, the number of free valencies on each carbon, binary information on bonding between each C−C pair in the molecule, and the number of total carbons and hydrogens are specified in the SVCF. In this way, the local environment of each carbon is described in sequence. An example is shown in Figure S5.

2.3. ML Methods. With the advent of the era of data and artificial intelligence, a variety of ML algorithms have been used for predictive analytics in many applications, such as disease diagnoses in healthcare,[49−52] fraud detection and prevention in finance,[53−55] and object recognition and classification in image and video analysis,[56−58] among many others. Different algorithms showed distinct advantages and disadvantages depending on the types of datasets, the size of the datasets, and the problems to solve.[59] Likewise, for the prediction of adsorption properties in surface chemistry, multiple ML models from different algorithmic classes have been tried and evaluated, covering both linear and nonlinear regression, with the inclusion of techniques like regularization and the kernel trick, alongside a selection of nonlinear models.[6,8,47,60−62] Herein, we employed both linear and nonlinear ML models with feature sets derived from four molecular fingerprints. The linear algorithms used in the study include multiple linear regression (MLR), ridge regression (RR),[63] and LASSO.[64] Then, the kernel trick was used to introduce the nonlinearity, from which kernel ridge regression (KRR)[65] and support vector regression (SVR)[66] were utilized. The tree-based nonlinear models, random forest regression (RFR),[67] and XGBoost regression (XGB),[68] were also used. Four feature sets were obtained by standardizing each molecular fingerprint, which resulted in a mean of zero and a unit variance for each set. The combination of these seven algorithms and four feature sets resulted in the creation of 28 distinct ML models.

To train, validate, and test the ML models, the Python-based library "scikit-learn" was used.[69] The complete dataset, consisting of 384 species calculated using DFT, was randomly divided into five subsets, with each subset containing 20% of the total dataset. Four subsets (80% of the data) were used to train and validate, and the rest of one subset (20% of the data) was used to test the models. This 80/20% train/test split was repeated five times, ensuring that one subset used for testing was different each time. In this manner, prediction results for the whole species were obtained. The hyperparameters of the ML algorithms used were optimized with a grid search using k-fold cross-validation (k = 5).

Training and testing ML models with the five subsets described in the previous paragraph are considered one iteration. The formation enthalpy of every species was predicted, and errors were calculated after each iteration. To ensure the precision and robustness of prediction results by the ML models, we performed 30 independent iterations with a random sampling of five subsets for each iteration, generating the distribution of the enthalpy prediction for each model.

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

When performing each iteration, it is noted that specific fingerprint values that are rare or absent in the training set could cause divergence in the prediction of test species in MLR. Regularization would ensure convergence, but the results would still not be reliable, which makes these cases undesirable. For selected pairs of models, the independent sample *t*-test was conducted using the mean absolute errors, aiming to ascertain the presence of statistically significant differences in predictive performance.

Finally, we implemented an ensemble of average voting, selecting two individual models from the best-performing ones. Better prediction of the enthalpies could be achieved by ensemble methods with individual base estimators as it would enhance generalizability and robustness over a single regressor.[5,55] In this method, the mean of the enthalpy of each species was calculated from two individual predictions obtained by two independent algorithms to obtain a final estimate.

2.4. Adsorbates Used in This Study. The adsorbates investigated in this study range from aliphatic and olefinic C_2 to C_6 hydrocarbons and their derived radicals, including both acyclic and ringed species.

While an ideal comprehensive study would include all hydrocarbon radicals in this range, the computational expenses of DFT calculations necessitate the selection of smaller subset species. Rather than deliberately choosing highly representative and exemplary species to yield a better predictive model, the species selected in this study came from all previously calculated species and structures by the authors.[23,30,31,70,71] This more closely resembles the development of practical ML approaches by using the best available data to generate useful predictions with the minimum additional computational effort.

C_6 species account for two-thirds of the total species, as we mainly seek the prediction of the enthalpies of relatively larger species. The information about the number of species used in the study is summarized in Table 1. The dataset of the adsorbates excludes any CH_x (x = 0 to 3) and dicarbon as their structures are unique in molecular fingerprints and binding geometries but not useful for the prediction of whole other species. For GA of CH_x species, different groups from the ones in Table S1 are needed to describe their structures since the adsorbates have only one carbon. For example, a group component of C(H)2(•)2 is required to represent CH_2, whereas this group is not used to explain any other species in the dataset. This is because all groups used for C_2+ species have at least one carbon as the nearest atom bonded to the center carbon. Dicarbon also represents a unique structure where two carbons with three-free valency are directly bonded to each other. As a result, dicarbons may have distinctive adsorbate descriptors. For example, in the FMF, the "C3−C3" component is 1 for dicarbon, whereas it is 0 for all other species. This uniqueness makes this species a structural outlier, which is neither beneficial for predicting others nor accurate in predicting the enthalpy of this species by ML models trained on other species.

Table 1. Number of Adsorbates Used in This Study

number of carbons	number of species	
	acyclic	cyclic
C2	8	0
C3	29	2
C4	39	2
C5	36	12
C6	244	12
total	356	28

3. RESULTS AND DISCUSSION

An overview of the development of the ML models and their use for the massive prediction of the enthalpies of new hydrocarbon adsorbates is illustrated in Figure 1. DFT calculations were used to obtain surface enthalpies per carbon for all 384 species in this study. Four molecular fingerprints were generated for each adsorbate and were input to the ML models as features, while the DFT-calculated enthalpies were set as target properties. Each ML model was developed 30 times with different samplings of train-test splits (30 iterations), resulting in a distribution of error metrics (e.g., mean and maximum deviation or error) for each model. In postanalysis, the performance of the models was evaluated for precision and accuracy (i.e., model robustness), and the best individual models were ensembled to create the best ML predictions. Finally, the model was used to estimate the enthalpy of a significant number of new hydrocarbon adsorbates on Pt(111).

3.1. Analysis of Prediction Errors Arising from Stochastic Split-Fold Train/Test Sets. For each of the nonoverlapping 5-fold train/test set splits, the species assigned to train/test sets are arbitrary. However, each ML algorithm's parameter estimation depends on the specific data input, i.e., the species included in that particular split. The randomness of the species selected for a split introduces stochasticity into the ML results. This is examined by performing 30 iterations of the 5-fold splits, creating 30 independent trials of the complete ML approach, allowing for statistical analysis of the algorithm and fingerprint robustness across 384 species.

3.1.1. Analysis of Mean Absolute Errors. For a single implementation of the ML models, 30 iterations of 5-fold splits were performed. For each of these iterations, the mean of absolute error across all 384 species ($MeanAE_{species}$) was calculated, yielding 30 values of $MeanAE_{species}$ for each model. The distributions of these $MeanAE_{species}$ are displayed in Figure 2a. In similar ways, the maximum value among the absolute errors of 384 species ($MaxAE_{species}$) is taken in every single iteration, creating the distributions of 30 $MaxAE_{species}$, as presented in Figure 2b. In addition, for each boxplot in Figure 2, the first (mean), second (variance), and third (skewness) moments of the distribution are given in the corresponding heat maps in Figures 3,S6, and S7, respectively.

In Figures 2 and 3, the $MeanAE_{species}$ is lower in nonlinear models (KRR, SVR, RFR, and XGB) than in linear models (MLR, RR, and LASSO), regardless of the feature sets used. The poorer performance of linear models was unsurprising given that the enthalpy of adsorption is a highly complex function of the number of carbons in adsorbates. As shown in Figure 3a, the KRR (1.14−1.22 kcal/mol_C) and XGB (1.08−1.26) models have lower $MeanAE_{species}$ than any other models. In KRR, good predictive performance can be expected using an appropriate kernel function and its parameters,[72,73] which were obtained by allowing the ML algorithm to attempt many different combinations and select the one with the lowest loss function. A sampling of the kernel function and parameter choices explored is included in Table S5. For XGB, the decision tree branch based on rank or threshold rather than actual values. This algorithmic characteristic may be

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

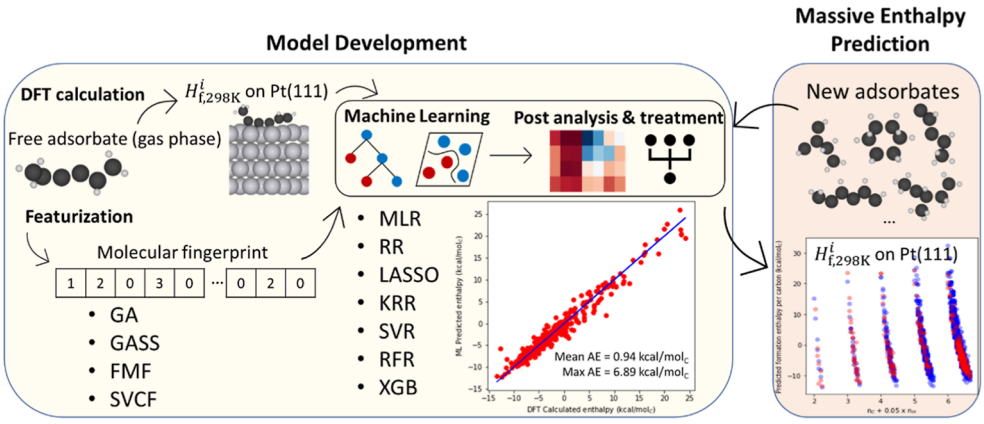


Figure 1. Schematic of the development of ML models and their use of enthalpy prediction for a large set of new species.

particularly beneficial for regression with discrete data, which is the case for the four feature sets used here.[68] Meanwhile, in Figure S6a, while all models have reasonably low variances, the variance of $MeanAE_{species}$ across the iterations is generally lower in linear models (0.00027−0.0011 $kcal^2/mol_C^2$) than in nonlinear models (0.0009−0.0039). This can also be seen in Figure 2, where MLR, RR, and LASSO have smaller box lengths compared to those of the other algorithms. It implies that linear algorithms produce consistent or stable predictions with different samplings of the dataset. Consequently, the linear models have higher precision but lower accuracy, whereas the nonlinear models have lower (but sometimes reasonable) precision and higher accuracy.

A statistical hypothesis test, the independent samples *t*-test, was conducted to statistically confirm the difference in performance between KRR and XGB with respect to their $MeanAE_{species}$. The calculated *p*-values from the *t*-test between various combinations of KRR and XGB models are listed in detail in Table S2. While the differences between several model pairs lacked statistical significance (*p*-value >0.05), the *p*-values between XGB + SVCF and all other models were lower than a significance level of 0.05, as were those for KRR + FMF. In fact, KRR + FMF showed only minor overlap with two other models (KRR + GA, $p = 0.0010$; KRR + SVCF, $p = 0.0002$) and XGB + SVCF showed $p = 0$ for all other models, making these two models both statistically the most accurate and most independent of all ML algorithm + molecular fingerprint combinations studied here, and their independence from one another indicates that this accuracy is not coincidental but inherent.

When comparing the four fingerprints, Figure 3a shows that the best fingerprint choice depends on the class of the ML algorithm used. GASS was the best when employed with the linear models MLR, RR, and LASSO. FMF was the most preferred, with kernel-based KRR and SVR algorithms. The two tree-based algorithms, RFR and XGB, had the best results with SVCF. This demonstrates that no single fingerprint simply outperforms others, but the performance of each fingerprint depends on its compatibility with the ML algorithm used. However, fingerprint choice causes only small variations in the $MeanAE_{species}$ within one ML algorithm, with the best and worst fingerprint choice being within ~0.2 $kcal/mol_C$ with two major exceptions. SVR showed a range of 0.34 $kcal/mol_C$ between the best (FMF) and worst (SVCF) fingerprints, and SVCF showed much poorer compatibility with the linear models than any other fingerprint choice.

3.1.2. Analysis of Maximum Absolute Errors. For $MaxAE_{species}$, as with $MeanAE_{species}$, no single fingerprint or algorithm on its own surpasses the others. In Figure 2b, MLR, RR, and LASSO showed distributions of $MaxAE_{species}$ similar to each other for any given fingerprint, while the remaining algorithms varied. Notably, high variances in $MaxAE_{species}$ were observed for the kernel models, including KRR + GASS (82.72 $kcal^2/mol_C^2$, see Figure S6b), SVR + SVCF (28.81), KRR + GA (17.29), and KRR + FMF (14.98), whereas all linear models had much smaller variances (0.10−0.51).

This difference in variance may be related to which species has $MaxAE_{species}$ for each iteration. The whole list of species of $MaxAE_{species}$ for the models specified above (all 12 linear models, KRR + GASS, SVR + SVCF, KRR + GA, and KRR + FMF) can be found in Figure S8. It should be noted that most species in Figure S8 are highly unsaturated hydrocarbons, which suggests an increased complexity of adsorption and binding modes as unsaturation increases. The relaxed 3D structures of these species and other highly unsaturated species can be obtained from the output files in Section S1.

For 12 models using MLR, RR, and LASSO, three species were responsible for the $MaxAE_{species}$ (see Figure S8a), and tricarbon was the dominant species, accounting for more than 80% of the total cases, 294 out of 360 total iterations (=3 algorithms ×4 features ×30 iterations). Its enthalpy was always underpredicted in any iteration by −17 to −10 $kcal/mol_C$. Thus, the same species consistently underpredicted by a large amount but within a small range resulted in a lower variance of the distribution of $MaxAE_{species}$ in linear models. Conversely, in four nonlinear models (KRR + GASS, SVR + SVCF, KRR + GA, and KRR + FMF), 17 species are attributed to $MaxAE_{species}$, which is a much higher number than the three species found in 12 linear models. No single species comprised more than 30% of the cases (Figure S8b). In addition, it was found for some species that the errors fluctuated much across the iterations within the same model. For example, one of the isomers of C_6H_0 is responsible for the $MaxAE_{species}$ of the KRR + GASS model in 17 out of 30 iterations, and its errors with this model had a wide range of −6.90 to 40.18 $kcal/mol_C$. This shows that higher variances of $MaxAE_{species}$ for some kernel models are ascribed to not only the variety of species but also the dependency of accuracy on data samplings for the same species. That is, unlike the linear models, the prediction of enthalpy for specific species with nonlinear models can be sensitive to what species are sampled in training sets. This trend is also observed in the tree-based nonlinear models, but

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

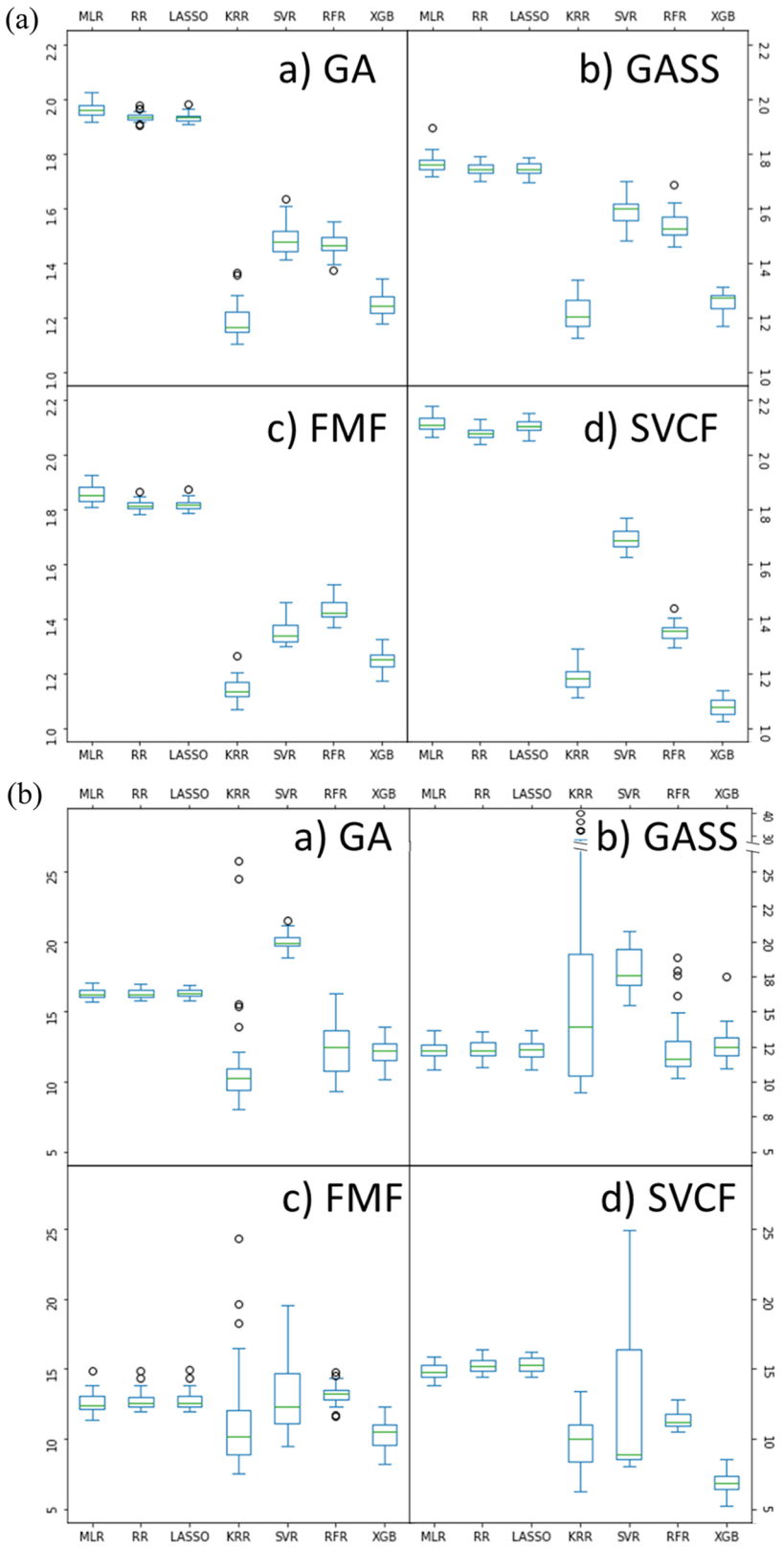


Figure 2. Distribution of (a) mean and (b) maximum absolute errors of each model in enthalpy of formation per carbon (in kcal/mol$_C$). Each single data point in the distributions indicates the mean/maximum absolute errors of 384 species in one iteration. Distribution of each model includes 30 data points obtained from 30 iterations. Box represents the data between the first (Q1) and third (Q3) quartiles, and the green line within the box denotes the median (Q2). Distance between Q1 and Q3 is called the interquartile range (IQR). Upper and lower whiskers extend to the furthest data points that are within Q3 + 1.5 × IQR and Q1 − 1.5 × IQR, respectively. Data points outside the whisker limits are considered outliers and are marked as white dots.

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

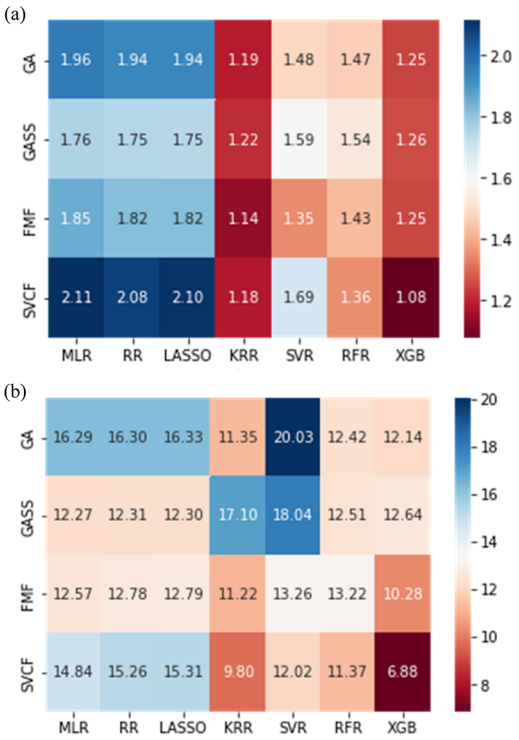


Figure 3. Heat maps for the (a) mean and (b) maximum absolute errors of each model in enthalpy of formation per carbon across the 30 iterations. All numbers in both heat maps are in kcal/mol$_C$.

far more moderately, as shown in the heatmap in Figure S6b. It is possibly due to the algorithmic benefits of decision trees and ensemble methods, which impart robustness to outliers.[67,68] When the decision trees are built based on the feature values, they can create splits that minimize the impact of individual extreme data points. In addition, the ensemble methods used in RFR (bagging) and XGB (boosting) with multiple decision trees help mitigate the impact of the outliers present in individual trees. To ensure the robustness of the prediction, it would eventually be important to try to include many chemically similar species in the training set. In terms of ML models, this would mean sampling the species with similar compositions of feature components in the training set. Therefore, a diversity of components in terms of richness and evenness, as well as the elimination of data sparsity, would be important for the features of the training data.

3.2. Examination of the Validity of ML Models. To examine the credibility of the ML models, we performed computational tests by intentionally adding two-dimensional vectors of random integers (0 to 4) to each feature, followed by the development of ML regression models using the raw and random vector-added features for one iteration. Then, the errors of the randomly disturbed (RD) dataset were compared to the original estimators. The mean absolute errors of a single iteration for the models trained on normal and RD datasets are presented in Tables S3 and S4. The mean absolute errors of MLR, RR, and LASSO are comparable between the normal and RD datasets, with a difference (=RD − normal) of −0.02 to 0.03 kcal/mol$_C$. For more accurate nonlinear models, the mean absolute errors were increased by 0.02−0.22 kcal/mol$_C$ with the RD datasets, except for SVR + SVCF (decreased by 0.03). It demonstrates that the random noise added to the descriptors cannot improve the estimation of the enthalpies in most models but rather worsens it, especially for more rigorous models, revealing the reliability of molecular-fingerprint-based ML models. That is, the reasonable accuracy of the estimation should be attributed to properly describing the local valency and bond types of carbons in the adsorbates with the molecular fingerprints and not random values inserted into the descriptors. The effect of a random disturbance either worsens the prediction or has a negligible impact.

3.3. Evaluation of Individual ML Models and Ensemble of Average Voting. *3.3.1. Overall Performance of Individual ML Models.* The mean and variance from the distribution of MeanAE$_{species}$ for each model were investigated to assess both the accuracy and the robustness of the individual ML models, with two plots shown in Figure 4. Two variances are employed in the analysis. First, the variance plotted in Figure 4a is calculated from the distribution comprising 30 MeanAE$_{species}$ obtained from each iteration, i.e., Var$_{iter}$(MeanAE$_{species}$), also given in Figure S6a. Larger values of Var$_{iter}$(MeanAE$_{species}$) indicate greater variance in MeanAE$_{species}$ values when different samplings are applied, and conversely, smaller values imply greater consistency. Comparing these metrics in Figure 4a, XGB + SVCF is the best model, having both the smallest error and among the lowest variances, with KRR + FMF and KRR + SVCF also showing good performance. The high precision (low variance) but low accuracy (high error) of linear algorithms (MLR, RR, and LASSO) is also evident, with all of them clustered in the lower right quadrant of the plot.

One can also compare the variance and error across each individual species, averaged over the 30 independent estimations obtained from the iterations. Figure 4b plots the mean absolute errors of each iteration across 384 species (MeanAE$_{iter}$) against the variance of the errors across the 384 species, Var$_{species}$(MeanAE$_{iter}$). It should be noted that the MeanAE$_{species}$ = MeanAE$_{iter}$ as either represents an average over species and iterations, so the order of the averaging does not matter. Here, Var$_{species}$(MeanAE$_{iter}$) represents how consistent the predictions of the enthalpy for a single species can be across 30 trials with a given model. That is, larger values indicate that the errors for some species are significantly larger than the errors for others, while smaller values indicate that the model is able to predict enthalpies for all species more consistently, giving greater confidence in the chemical enthalpy prediction for any given species in the dataset. Here, the limitations of the three linear algorithms (MLR, RR, and LASSO) in the upper right quadrant of the plot indicate both high mean error but also a high extent of variation in the quality or confidence in those predictions depending on the species (e.g., behavior of highly unsaturated species captured poorly as discussed above), even though those poor predictions are predicted consistently across all 30 trials. Conversely, the models with KRR and XGB are in the bottom left quadrant, representing their consistent predictions of the enthalpies of specific species regardless of iteration. Hence, it is anticipated that KRR and XGB would exhibit relatively uniform errors centered around the mean for any predicted species. Figure 4b reaffirms XGB + SVCF, KRR + FMF, and KRR + SVCF as the three most accurate and robust models for these 384 hydrocarbon adsorbates.

In combination, Figure 4a,b reveals the robustness of the models to statistical sampling and chemical representation challenges inherent in ML from limited datasets. Figure 4a captures the statistical robustness of the models by indicating

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

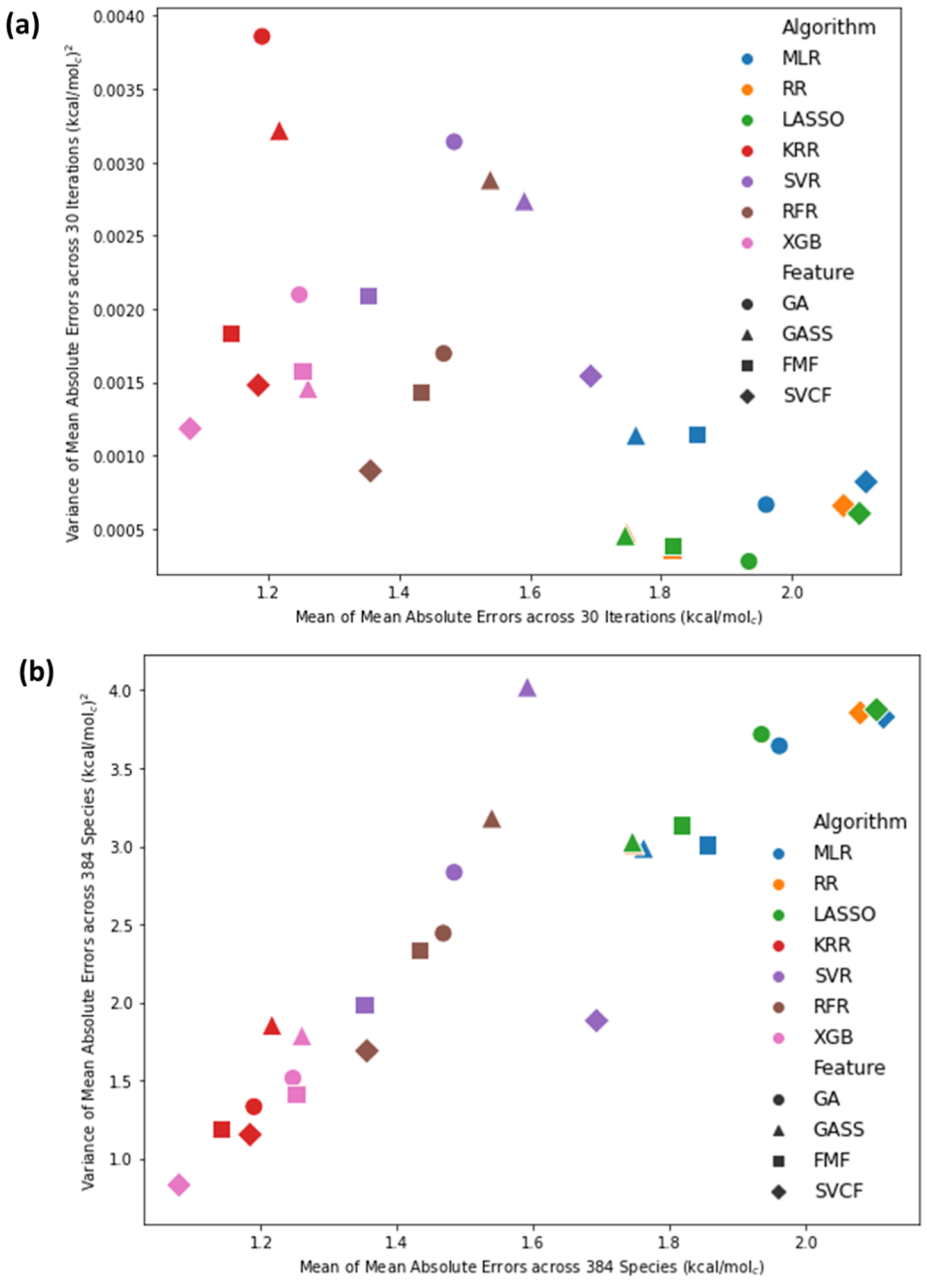


Figure 4. Mean and variance plots for the distribution of mean absolute errors for each model in enthalpy of formation per carbon. (a) Mean and variance for each model were calculated across 30 iterations. Variance quantifies the amount of dispersion in errors for a set of randomly sampled iterations. (b) Mean and variance for each model were computed across 384 species. Variance measures the amount of spread in errors across 384 species.

how consistent the results are for a given sampling or iteration of each model. Figure 4b shows the chemical robustness of the models, revealing which models are expected to give similar confidence in their predictions for any of the species.

3.3.2. Prediction of Individual Species across Iterations. Predicting enthalpies for individual species is further explored in the parity plots of Figure 5, where the enthalpy prediction with ML for 384 species, averaged over 30 iterations, is compared to the DFT-calculated enthalpies. Some trends emerged between different ML approaches applied at the level of individual chemical species. Notably, MLR, RR, and LASSO have not only quantitatively similar mean and variance to each other in distributions but also similar predictions for individual species, as seen in the parity plots. Significant errors are observed with the species of high per-carbon formation enthalpies in these models. Given most of them are highly unsaturated species, large deviations would mainly be due to the enthalpy's nonlinear trend with local valency and bond types of adsorbates. These limitations were generally improved upon when nonlinear models were applied. Among these, XGB + SVCF, KRR + FMF, and KRR + SVCF were the most accurate and robust models in Figure 4. It is also found that these models present consistently reliable predictions for most of the species in Figure 5. The parity plots also show that the deviations from the parity line for these models did not depend on the enthalpy, supporting, once again, that these models gave consistently reliable predictions for all species. RFR- and SVR-based models, which have higher $\mathrm{Varl}_{\mathrm{species}}(\mathrm{MeanAE}_{\mathrm{iter}})$ values,

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

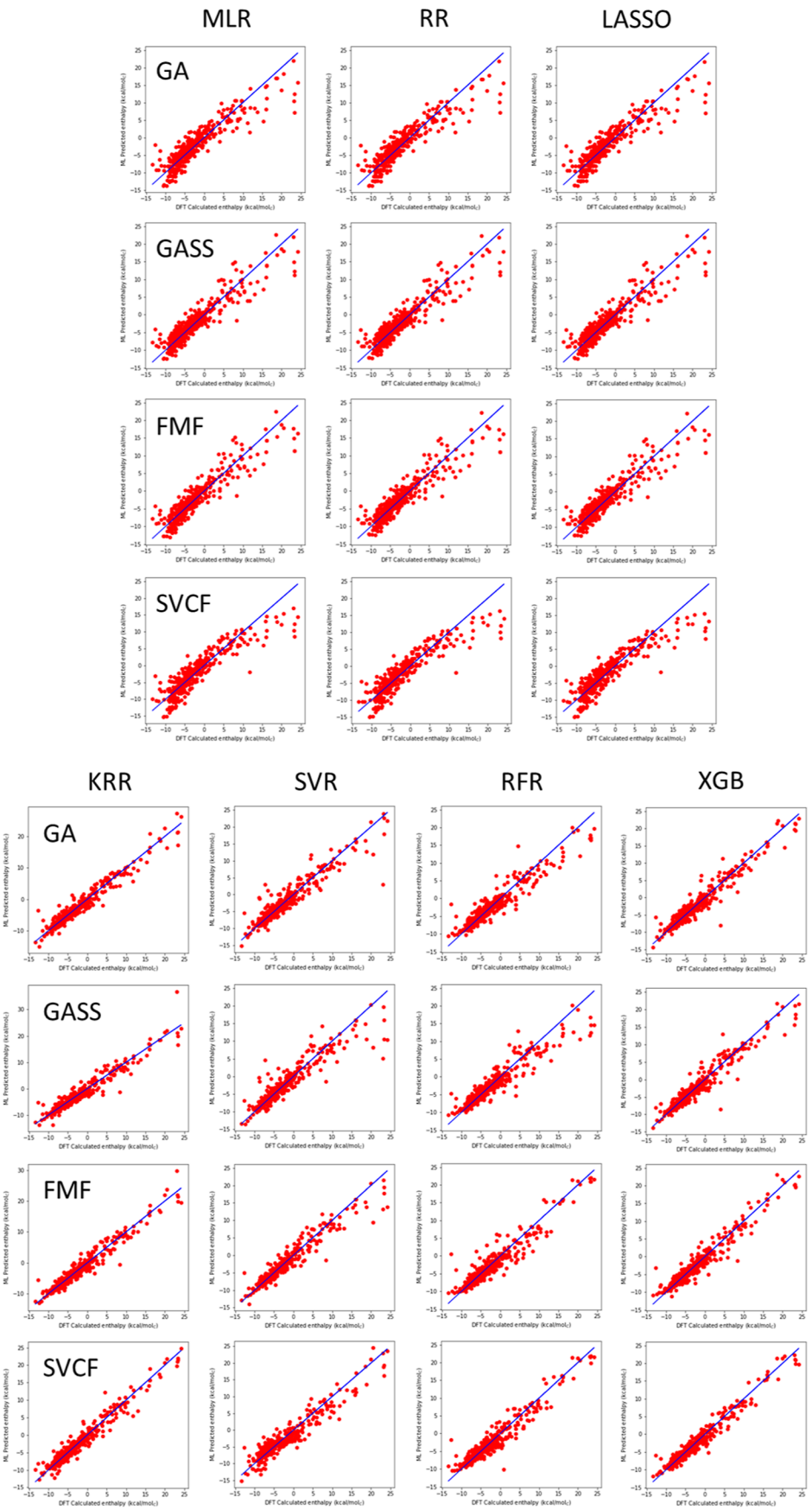


Figure 5. Parity plots for DFT-calculated and ML-predicted enthalpies of formation per carbon. Each row represents the molecular fingerprints (or features) and each column represents algorithms.

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

show poorer performances at high and very low enthalpies in the larger spreads seen in the parity plots.

When the mean enthalpy prediction is calculated by averaging 30 iterations for use in the parity plots, some information about the deviations between individual iteration errors relative to the mean value is lost. Namely, the mean error for a set of 30 predictions need not be equal to the mean absolute error if the predictions are *both* overpredicted and underpredicted in different iterations. While the significance of the mean absolute error does not change, the predicted mean enthalpy may arise from error cancellation between these disparate iterations. In Figure S9, the mean errors of each species through 30 iterations are on the x-axis and the $\mathrm{MeanAE_{iter}}$ is on the y-axis for KRR + FMF and XGB + SVCF. If error cancellation reduces mean error relative to mean absolute error, the data points will have a large deviation from $y = |x|$, specifically the points that appear closer to the vertical line $x = 0$. On the other hand, the species whose predictions are consistent will be on or near the $y = |x|$ line, i.e., the mean error and mean absolute error are equivalent. It is noted in Figure S9 that most species are on or very close to the parity lines in both models, indicating a few effects of error cancellation. However, several species still exhibit deviations. For example, six species calculated by KRR + FMF have a difference larger than 0.75 $\mathrm{kcal/mol_C}$ from $y = |x|$ and are listed in Figure S10a. The low precision of the enthalpy prediction across the iterations, leading to error cancellation, could be misleading when interpreting the overall predictive reliability of the models. For XGB + SVCF, the species are more aligned with the parity lines in Figure S9b, which implies there are fewer species benefiting from error cancellation when calculating the enthalpy. No species has a deviation larger than 0.75 $\mathrm{kcal/mol_C}$ from $y = |x|$, and there are six species with deviations between 0.75 and 0.5 $\mathrm{kcal/mol_C}$, as shown in Figure S10b.

3.3.3. Ensemble Average Voting of the Individual ML Models. Finally, to maximize the general accuracy of the ML model, an ensemble of average voting was implemented with two models, KRR + FMF and XGB + SVCF, the best two individual estimators in terms of accuracy and robustness. The parity plot of the ensemble model of KRR + FMF and XGB + SVCF is displayed in Figure 6. As expected, the accuracy of the ensemble model was improved with a mean absolute error of 0.94 $\mathrm{kcal/mol_C}$, which is lower than any other individual regressor. It exemplifies that more reliable estimation can be attained through the voting ensemble method for most species by complementing independent individual models with each other.[5,74]

The performance of the ensemble model for subsets of the data can be found in Figure S12, where C_n species (n = 2, 3, 4, 5, and 6) are highlighted against the whole dataset as well as cyclic/acyclic species. No significant differences are observed in prediction accuracy as a function of carbon number, with predictions within each group roughly consistent with the dataset as a whole. The same can be said when comparing cyclic and acyclic species. It is rather the species with very low or very high formation enthalpies that show the largest deviations from the parity line.

3.4. Prediction of the Enthalpies of Exhaustive $\mathrm{C_2}$ to $\mathrm{C_6}$ Acyclic Hydrocarbon Adsorbates. The development of our ML models should contribute to significantly reducing the computational efforts for the investigation of reaction networks involving surface intermediates of large hydrocarbons. One way to accomplish this would be by estimating enthalpies for an extensive list of surface species using the ML models and suggesting species of low energies and significance on a potential energy diagram. This would essentially create priorities for species requiring more DFT calculations and/or ideally lead to generating preferred reaction pathways.

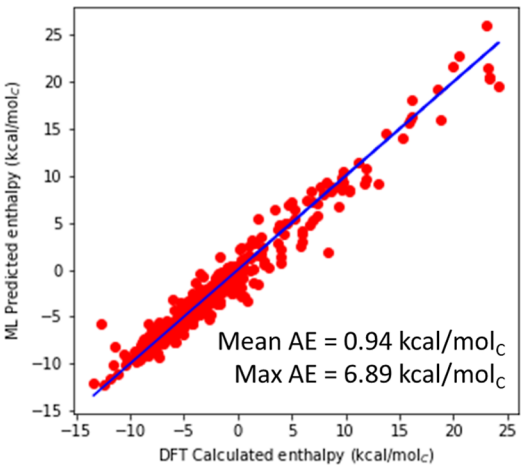


Figure 6. Parity plot of the ML-predicted and DFT-calculated enthalpy of formation per carbon for the ensemble average voting model of KRR + FMF and XGB + SVCF. Mean AE: mean absolute error and Max AE: maximum absolute error.

To showcase it, we have first enumerated 3115 $\mathrm{C_2}$ to $\mathrm{C_6}$ acyclic hydrocarbon adsorbates in SMILES that are potentially involved in carbon chain growth reactions on metal surfaces. Their enthalpies on Pt(111) were estimated using the ensemble models of KRR + GA and XGB + GA, which are two of the superior individual models, as shown in Figure 4. GA was selected simply because algorithms for generating the feature sets from SMILES are available in RING and readily available for other researchers. To ensure reliability, 30 iterations were carried out, and the mean of predictions throughout the iterations was used as a final output. The predicted formation enthalpy per carbon and total formation enthalpy are presented in blue for every $\mathrm{C}_x\mathrm{H}_y$ (x = 2 to 6, y = 0 to 13) in Figure 7. The 384 species for which DFT-derived values were available in this study are shown in red. It is first noted that while the total formation enthalpy has a large difference in its range from $\mathrm{C_2}$ (−30 to 35 kcal/mol) to $\mathrm{C_6}$ species (−80 to 200 kcal/mol), the range of formation enthalpy per carbon is relatively small, for instance, −15 to 20 $\mathrm{kcal/mol_C}$ for $\mathrm{C_2}$ and −15 to 35 $\mathrm{kcal/mol_C}$ for $\mathrm{C_6}$ species. This reinforces the usage of enthalpy per carbon as a target variable for the ML models instead of total enthalpy to avoid the overestimation of errors in larger adsorbates. In addition, it is highlighted how effectively the ML model was developed by using only a subset of the hydrocarbon adsorbates and was leveraged to predict the enthalpies of a massive list of species.

In Figure 7, the formation enthalpy of the adsorbates with the same carbon numbers is negatively correlated with the number of hydrogens. That is, as the hydrocarbons become more saturated, their formation enthalpies on Pt(111) generally decrease. To examine the energetically favorable species among the structural isomers, one can find species of lower enthalpies at the same x-values in Figure 7. For example, only 9 species have total formation enthalpies lower than −27 kcal/mol among the 390 isomers studied for $\mathrm{C_6H_7}$, which can be taken at x = 6.35. In similar ways, all thermodynamically

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

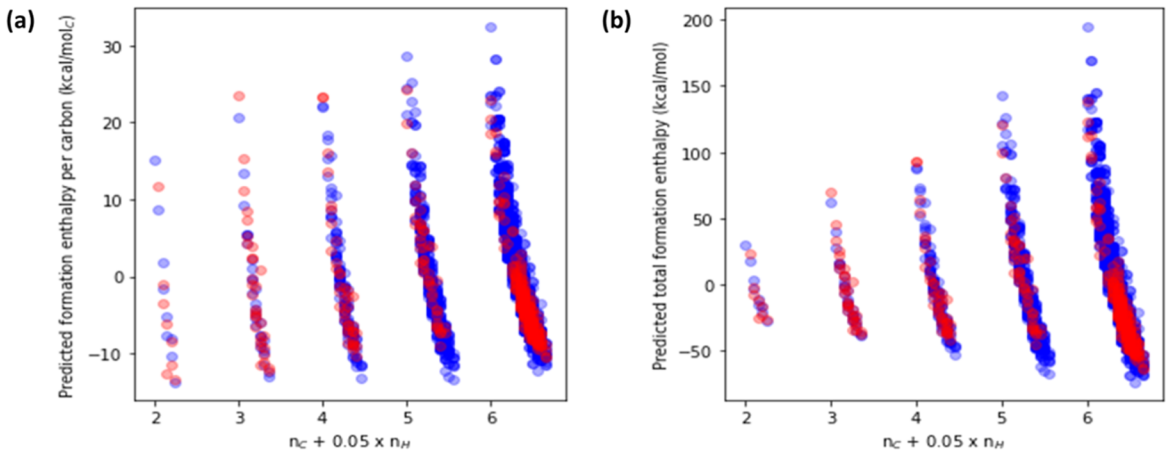


Figure 7. Prediction of (a) the enthalpy of formation per carbon and (b) the total enthalpy of formation for 3115 acyclic C_2 to C_6 hydrocarbon adsorbates on Pt(111) with the ensemble average voting model of KRR + GA and XGB + GA. All ML model-predicted enthalpies are in blue. For comparison, DFT-derived enthalpies used to develop the ML model are displayed in red. Number of carbons (n_C) and hydrogens (n_H) is denoted in the x-axis. n_C is an integer (2, 3, 4, 5, and 6) and n_H is multiplied by 0.05 first, then added to n_C. For example, C_2H_4 is 2 + 0.05 × 4 = 2.20 and C_5H_9 is 5 + 0.05 × 9 = 5.45 in x-axis.

preferred adsorbates can be obtained at each saturation level of hydrocarbons, giving some insights into possible reaction pathways.

4. CONCLUSIONS

In this study, we explored ML models predicting the formation enthalpies of various C_2 to C_6 hydrocarbon radicals adsorbed on Pt(111). The predicted enthalpies per carbon were compared with the values obtained from the DFT calculation for 384 species. By including C_6 species in two-thirds of the total datasets, our models mainly targeted relatively large hydrocarbons, expanding the space of adsorbates in the investigation of adsorption properties with ML. To capture the nonlinear behavior of the enthalpy resulting from the complex interaction, two algorithms with the kernel trick and two tree-based ML algorithms were applied, and three linear models were also compared as a benchmark. Four molecular fingerprints were developed and served as features for the ML models with which the valency and bond of individual carbons within the adsorbates are effectively described. Notably, no DFT calculation or any other intensive computation was required to generate these fingerprints. To ensure the reliable evaluation of individual ML models, each model was trained, validated, and tested with 30 different train/test samplings.

It was found that nonlinear models (KRR, SVR, RFR, and XGB) have lower mean absolute errors than the linear models (MLR, RR, and LASSO), outperforming the accuracy in general because of their better explanatory power for the nonlinear complexity of adsorption. In particular, KRR and XGB have the lowest mean absolute errors due to their algorithmic adequacy to our feature sets. By the independent sample t-test performed for the combination of KRR and XGB used models, XGB + SVCF and KRR + FMF were the two best statistically distinct models in terms of accuracy. For molecular fingerprints, note that no single fingerprint dominates the predictive performance. Rather, the best fingerprint with the lowest mean absolute error was dependent on the combination of features and algorithms. The accuracy of the prediction for specific species can be sensitive to the train-test samplings, especially for largely unsaturated hydrocarbon radicals or species with rare fingerprint compositions. Also, the validity of the ML models was proved by testing them with the introduction of random disturbances to each descriptor, where we found an increase of errors in most of the cases.

The robustness of the models was assessed with the mean and variance of the distribution of mean absolute errors, from which XGB + SVCF and KRR + FMF had not only the best accuracy but also the best robust models. The prediction of individual species displayed these results as well in the parity plots with the minimized effect of error cancellation. To improve the accuracy beyond the individual models, an ensemble of average votes was introduced with the two best individual models. The ensemble model predicted the adsorbate formation enthalpy per carbon most accurately, with a mean absolute error of 0.94 kcal/mol$_C$.

Finally, a case study was carried out to exhibit the use of our ML models when investigating the preferred pathways on surface reactions. The formation enthalpy per carbon for a total of 3115 C_2 to C_6 acyclic hydrocarbon adsorbates on Pt(111) was predicted using the ensemble models of KRR + GA and XGB + GA. It was plotted for the number of carbons and hydrogens, and the species with lower enthalpies at the same C_xH_y (x = 2 to 6 and y = 0 to 13) level were easily identified. Collectively, screening these species at each level of hydrocarbons would reduce computational efforts and provide potential reaction pathways.

Our study exemplified the promise to extend the ML study for the properties of adsorbates to more complex species. We believe this study helps shed light on tackling the challenges for the prediction of adsorbates in order to study more complicated reaction networks.

■ ASSOCIATED CONTENT

Data Availability Statement

The code used to perform this work is available at https://github.com/jinwoong-nam/predicting_enthalpy_machine_learning.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.3c08227.

Coordinates of the relaxed structures from DFT calculation for all 384 species used in this study;

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

referencing the enthalpy of formation; molecular fingerprint components; heat maps of the second (variance) and third (skewness) statistical moments for the distributions of the mean and maximum absolute errors across the iterations; statistical hypothesis testing with individual ML models; analysis of the species with the maximum absolute errors; comparison of the errors between normal and randomly disturbed datasets; effect of the error cancellation for individual species; ensemble average voting model of KRR + GA and XGB + GA S10; hyperparameters used for two best individual models (KRR + FMF and XGB + SVCF); prediction results of the ensemble average voting model of KRR + FMF and XGB + SVCF with respect to carbon number and cyclicity (PDF)

AUTHOR INFORMATION

Corresponding Author

Fuat E. Celik − *Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-5891-6375; Email: fuat.celik@rutgers.edu

Authors

Jinwoong Nam − *Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-3711-4023

Charanyadevi Ramasamy − *Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, United States;* orcid.org/0009-0001-8219-1730

Daniel E. Raser − *Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, United States*

Gustavo L. Barbosa Couto − *Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, United States;* orcid.org/0009-0008-8161-9281

Lydia Thies − *Department of Chemical Engineering, University of Florida, Gainesville, Florida 32611, United States*

David Hibbitts − *Department of Chemical Engineering, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0001-8606-7000

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.3c08227

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (CBET-1705746). This work used the Extreme Science and Engineering Discovery Environment (XSEDE) located at the Pittsburgh Supercomputing Center (PSC), which is supported by the National Science Foundation grant no. ACI-1548562 (allocation no. CTS200045), the Rutgers Office of Advanced Research Computing, which is supported by Rutgers and the State of New Jersey, and the Rutgers School of Engineering High-Performance Computing for computational resources. J.N. would like to thank Dr. Lyndsay M. Schaeffer, Dr. Keith Carroll, and Dr. Mark Vandeven for their professional mentoring as well as valuable discussions on ML.

REFERENCES

(1) Nørskov, J. K.; Abild-Pedersen, F.; Studt, F.; Bligaard, T. Density Functional Theory in Surface Chemistry and Catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 937−943.

(2) 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* **2017**, *5*, 24131−24138.

(3) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059−6072.

(4) Andersen, M.; Reuter, K. Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Acc. Chem. Res.* **2021**, *54*, 2741−2749.

(5) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *J. Phys. Chem. Lett.* **2020**, *11* (9), 3185−3191.

(6) Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. Convolutional Neural Network of Atomic Surface Structures to Predict Binding Energies for High-Throughput Screening of Catalysts. *J. Phys. Chem. Lett.* **2019**, *10* (15), 4401−4408.

(7) Mamun, O.; Winther, K. T.; Boes, J. R.; Bligaard, T. A Bayesian Framework for Adsorption Energy Prediction on Bimetallic Alloy Catalysts. *npj Comput. Mater.* **2020**, *6*, 177.

(8) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem* **2020**, *6*, 3100−3117.

(9) Noh, J.; Back, S.; Kim, J.; Jung, Y. Active Learning with Non-Ab Initio Input Features toward Efficient CO2 Reduction Catalysts. *Chem. Sci.* **2018**, *9*, 5152−5159.

(10) von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem* **2020**, *4*, 347−358.

(11) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(12) Xu, W.; Andersen, M.; Reuter, K. Data-Driven Descriptor Engineering and Refined Scaling Relations for Predicting Transition Metal Oxide Reactivity. *ACS Catal.* **2021**, *11*, 734−742.

(13) Nanba, Y.; Koyama, M. NO Adsorption on 4d and 5d Transition-Metal (Rh, Pd, Ag, Ir, and Pt) Nanoparticles: Density Functional Theory Study and Supervised Learning. *J. Phys. Chem. C* **2019**, *123*, 28114−28122.

(14) Fung, V.; Hu, G.; Ganesh, P.; Sumpter, B. G. Machine Learned Features from Density of States for Accurate Adsorption Energy Prediction. *Nat. Commun.* **2021**, *12*, 88.

(15) Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528−3533.

(16) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J. Phys. Chem. C* **2018**, *122*, 28142−28150.

(17) Wexler, R. B.; Martirez, J. M. P.; Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (13), 4678−4683.

(18) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.

(19) Praveen, C. S.; Comas-Vives, A. Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces. *ChemCatChem* **2020**, *12*, 4611−4617.

(20) Zong, X.; Vlachos, D. G. Exploring Structure-Sensitive Relations for Small Species Adsorption Using Machine Learning. *J. Chem. Inf. Model.* **2022**, *62*, 4361−4368.
(21) Kulik, H. J.; Hammerschmidt, T.; Schmidt, J.; Botti, S.; Marques, M. A. L.; Boley, M.; Scheffler, M.; Todorović, M.; Rinke, P.; Oses, C.; et al. Roadmap on Machine Learning in Electronic Structure. *Electron. Struct.* **2022**, *4*, 023004.
(22) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99*, 016105.
(23) Hook, A.; Celik, F. E. Predicting Selectivity for Ethane Dehydrogenation and Coke Formation Pathways over Model Pt-M Surface Alloys with Ab Initio and Scaling Methods. *J. Phys. Chem. C* **2017**, *121*, 17882−17892.
(24) Xiao, L.; Ma, F.; Zhu, Y.-A.; Sui, Z.-J.; Zhou, J.-H.; Zhou, X.-G.; Chen, D.; Yuan, W.-K. Improved Selectivity and Coke Resistance of Core-Shell Alloy Catalysts for Propane Dehydrogenation from First Principles and Microkinetic Analysis. *Chem. Eng. J.* **2019**, *377*, 120049.
(25) Savost'yanov, A. P.; Yakovenko, R. E.; Narochnyi, G. B.; Bakun, V. G.; Sulima, S. I.; Yakuba, E. S.; Mitchenko, S. A. Industrial Catalyst for the Selective Fischer−Tropsch Synthesis of Long-Chain Hydrocarbons. *Kinet. Catal.* **2017**, *58* (1), 81−91.
(26) Kibby, C.; Jothimurugesan, K.; Das, T.; Lacheen, H. S.; Rea, T.; Saxton, R. J. Chevron's Gas Conversion Catalysis-Hybrid Catalysts for Wax-Free Fischer−Tropsch Synthesis. *Catal. Today* **2013**, *215*, 131−141.
(27) Lian, Z.; Si, C.; Jan, F.; Zhi, S.; Li, B. Coke Deposition on Pt-Based Catalysts in Propane Direct Dehydrogenation: Kinetics, Suppression, and Elimination. *ACS Catal.* **2021**, *11*, 9279−9292.
(28) Wang, H.-Z.; Sun, L.-L.; Sui, Z.-J.; Zhu, Y.-A.; Ye, G.-H.; Chen, D.; Zhou, X.-G.; Yuan, W.-K. Coke Formation on Pt-Sn/Al2O3 Catalyst for Propane Dehydrogenation. *Ind. Eng. Chem. Res.* **2018**, *57*, 8647−8654.
(29) Rangarajan, S.; Bhan, A.; Daoutidis, P. Rule-Based Generation of Thermochemical Routes to Biomass Conversion. *Ind. Eng. Chem. Res.* **2010**, *49*, 10459−10470.
(30) Nam, J.; Celik, F. E. Effect of Tin in the Bulk of Platinum-Tin Alloys for Ethane Dehydrogenation. *Top. Catal.* **2020**, *63*, 700−713.
(31) Hook, A.; Massa, J. D.; Celik, F. E. Effect of Tin Coverage on Selectivity for Ethane Dehydrogenation over Platinum-Tin Alloys. *J. Phys. Chem. C* **2016**, *120*, 27307−27318.
(32) Joseph, V. R. Optimal Ratio for Data Splitting. *Stat. Anal. Data Min.* **2022**, *15*, 531−538.
(33) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(34) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169−11186.
(35) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(36) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(37) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953−17979.
(38) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758−1775.
(39) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B: Solid State* **1976**, *13*, 5188−5192.
(40) Lym, J.; Wittreich, G. R.; Vlachos, D. G. A Python Multiscale Thermochemistry Toolbox (PMuTT) for Thermochemical and Kinetic Parameter Estimation. *Comput. Phys. Commun.* **2020**, *247*, 106864.
(41) Vorotnikov, V.; Wang, S.; Vlachos, D. G. Group Additivity for Estimating Thermochemical Properties of Furanic Compounds on Pd(111). *Ind. Eng. Chem. Res.* **2014**, *53*, 11929−11938.
(42) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(43) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546−572.
(44) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69*, 279−324.
(45) Kua, J.; Faglioni, F.; Goddard, W. A. Thermochemistry for Hydrocarbon Intermediates Chemisorbed on Metal Surfaces: $CH_{n-m}(CH_3)_m$ with n = 1, 2, 3 and m ≤ n on Pt, Ir, Os, Pd, Rh, and Ru. *J. Am. Chem. Soc.* **2000**, *122*, 2309−2321.
(46) Gu, G. H.; Vlachos, D. G. Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111). *J. Phys. Chem. C* **2016**, *120* (34), 19234−19241.
(47) Chowdhury, A. J.; Yang, W.; Abdelfatah, K. E.; Zare, M.; Heyden, A.; Terejanu, G. A. A Multiple Filter Based Neural Network Approach to the Extrapolation of Adsorption Energies on Metal Surfaces for Catalysis Applications. *J. Chem. Theory Comput.* **2020**, *16*, 1105−1114.
(48) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry*; The Royal Society of Chemistry, 2013.
(49) Shafiha, R.; Bahcivanci, B.; Gkoutos, G. V.; Acharjee, A. Machine Learning-Based Identification of Potentially Novel Non-Alcoholic Fatty Liver Disease Biomarkers. *Biomedicines* **2021**, *9*, 1636−1716.
(50) Mi, X.; Zou, B.; Zou, F.; Hu, J. Permutation-Based Identification of Important Biomarkers for Complex Diseases via Machine Learning Models. *Nat. Commun.* **2021**, *12*, 3008−3012.
(51) Zhu, J.; Ren, Z.; Lee, C. Toward Healthcare Diagnoses by Machine-Learning-Enabled Volatile Organic Compound Identification. *ACS Nano* **2021**, *15*, 894−903.
(52) Sinha, P.; Churpek, M. M.; Calfee, C. S. Machine Learning Classifier Models Can Identify Acute Respiratory Distress Syndrome Phenotypes Using Readily Available Clinical Data. *Am. J. Respir. Crit. Care Med.* **2020**, *202*, 996−1004.
(53) Ileberi, E.; Sun, Y.; Wang, Z. A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection. *J. Big Data* **2022**, *9*, 24.
(54) Alarfaj, F. K.; Malik, I.; Khan, H. U.; Almusallam, N.; Ramzan, M.; Ahmed, M. Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms. *IEEE Access* **2022**, *10*, 39700−39715.
(55) Forough, J.; Momtazi, S. Ensemble of Deep Sequential Models for Credit Card Fraud Detection. *Appl. Soft Comput.* **2021**, *99*, 106883.
(56) Wu, X.; Sahoo, D.; Hoi, S. C. H. Recent Advances in Deep Learning for Object Detection. *Neurocomputing* **2020**, *396*, 39−64.
(57) Wang, C.; Liu, B.; Liu, L.; Zhu, Y.; Hou, J.; Liu, P.; Li, X. *A Review of Deep Learning Used in the Hyperspectral Image Analysis for Agriculture*; Springer Netherlands, 2021; Vol. *54*.
(58) Zhou, S. K.; Le, H. N.; Luu, K. V.; V Nguyen, H.; Ayache, N. Deep Reinforcement Learning in Medical Imaging: A Literature Review. *Med. Image Anal.* **2021**, *73*, 102193.
(59) Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160−221.
(60) Chang, C.; Medford, A. J. Classification of Biomass Reactions and Predictions of Reaction Energies through Machine Learning. *J. Chem. Phys.* **2020**, *153*, 044126.
(61) Nayak, S.; Bhattacharjee, S.; Choi, J.; Lee, S. C. Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy. *J. Phys. Chem. A* **2020**, *124*, 247−254.

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043

(62) 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.* **2021**, *12*, 5288.
(63) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55−67.
(64) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267−288.
(65) Saunders, C.; Gammerman, A.; Vovk, V. Ridge Regression Learning Algorithm in Dual Variables. *Proceedings of the Fifteenth International Conference on Machine Learning* 1998; pp 515−521.
(66) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(67) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(68) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*; 2016; pp 785−794.
(69) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(70) Flaherty, D. W.; Hibbitts, D. D.; Iglesia, E. Metal-Catalyzed C-C Bond Cleavage in Alkanes: Effects of Methyl Substitution on Transition-State Structures and Stability. *J. Am. Chem. Soc.* **2014**, *136*, 9664−9676.
(71) Hibbitts, D. D.; Flaherty, D. W.; Iglesia, E. Effects of Chain Length on the Mechanism and Rates of Metal-Catalyzed Hydrogenolysis of n-Alkanes. *J. Phys. Chem. C* **2016**, *120*, 8125−8138.
(72) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical Diversity in Molecular Orbital Energy Predictions with Kernel Ridge Regression. *J. Chem. Phys.* **2019**, *150*, 204121.
(73) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K. R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(74) Wang, B.; Gu, T.; Lu, Y.; Yang, B. Prediction of Energies for Reaction Intermediates and Transition States on Catalyst Surfaces Using Graph-Based Machine Learning Models. *Mol. Catal.* **2020**, *498*, 111266.

https://doi.org/10.1021/acs.jpcc.3c08227
J. Phys. Chem. C 2024, 128, 5030−5043