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

pubs.acs.org/JPCC

Article

Comparative Study on the Machine Learning-Based Prediction of Adsorption Energies for Ring and Chain Species on Metal Catalyst Surfaces

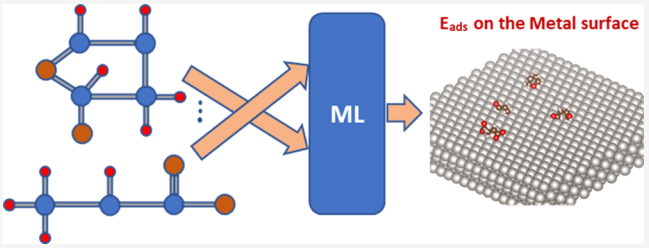
Asif J. Chowdhury,[‖] Wenqiang Yang,[‖] Andreas Heyden,* and Gabriel A. Terejanu*

Cite This: *J. Phys. Chem. C* 2021, 125, 17742−17748

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computation of adsorption and transition-state energies for a large number of surface intermediates for numerous active site models poses significant computational overhead in computational screening of catalysts. Machine learning (ML) techniques can be used to predict part of these energies. To predict the energies, ML models need to be fed appropriate metal and species descriptors. For complex surface chemistries, the structures of the intermediate species can vary greatly. In this paper, working with the hydrodeoxygenation of succinic acid on six different metal surfaces, we have studied the effect of linear and non-linear ML models used along with pen-and-paper-based species descriptors and two categories of metal descriptors on two different categories of intermediate species: chain and ring. More specifically, our computations include the prediction of chain species when trained on only chain species and also when trained on both chain and ring species. Similar computations were performed for predictions of ring species. In each case, the results of linear ML models were compared with kernel-based non-linear models. Our results indicate that ring species data do not improve the prediction of chain species. Similarly, chain species data do not improve the prediction of ring species. The use of non-linear ML models, however, did help to minimize the prediction errors compared to the linear models. The study also shows that electronic or adsorption energy-based metal descriptors along with bond count-based species fingerprints can achieve a mean absolute error (MAE) of less than 0.2 eV for complex chain molecules when used with an appropriate machine learning model.



1. INTRODUCTION

For discovery of heterogeneous catalysts through computational catalyst screening, microkinetic reaction models are usually developed, which are based on parameters computed from density functional theory (DFT) and transition-state theory (TST).[1−3] There is a large computational cost associated with the computation of adsorption and transition-state energies for each elementary reaction in the reaction network on different metal surfaces.[4] When energy data for each intermediate and transition-state species are available for a number of surfaces, the energies on other surfaces have been predicted using typically linear scaling relations based on metal descriptors.[5−7] However, when all species data are not available for any metal surface, the predictive model must incorporate species descriptors. Previous studies[8−10] have shown that non-linear ML models outperform linear models in this case. It has also been shown[11,12] that flat molecular fingerprints based on SMILES notation[13] give good predictive results when training and testing sets contain similar-sized molecules.

In this paper, we have studied and compared the predictive results between chain-structured intermediate species and ring-structured species on different metal surfaces for different splittings of train and test sets. Specifically, we worked with the data of adsorption energies for surface species for the hydrodeoxygenation of succinic acid on six different metal surfaces (Pd, Pt, Rh, Ru, Cu, and Ni). Using two different sets of electronic metal descriptors[14] and adsorption energy-based metal descriptors, a flat constant-sized SMILES-based species descriptor, and both linear and kernel-based non-linear ML models, we have run predictions for four different splittings of train and test sets: first, when training on a random subset of the chain species and predicting on the rest of them for each metal surface; second, similar to the first case but with training and testing performed on ring structures; third, training on all the ring structures and a random subset of chain structures and testing on the rest of the chain molecules; and fourth, training on the full chain data and a random subset of ring data and predicting on the rest of the ring data. The key questions that the current study seeks to answer are as follows: what is the predictive accuracy of pen-and-paper-based descriptors for

Received: June 21, 2021
Revised: July 26, 2021
Published: August 10, 2021

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748

complex ring-structured intermediates that absorb at multiple surface sites and thus whose adsorption geometry is difficult to describe without the use of coordinates that are unknown in a prediction model? Does inclusion of chain data help with the prediction of ring structures and vice versa? In terms of accuracy, is there any advantage of using electronic metal descriptors over adsorption energy based ones? How linear and non-linear ML models compare for predictions of both chain- and ring-structured surface intermediates? Thus, the goal of the paper was not to develop any novel descriptor or machine learning model. Instead, this is a comparative study on the efficiency of established predictive models for chain and ring species on metal surfaces.

2. METHODOLOGY

In this section, we begin with a description on data collection and preparation. Then, we discuss the choice of metal descriptors, species descriptors, and ML models. Finally, the process of splitting the combined chain and ring data into the training and testing sets is explained.

2.1. Data Collection and Data Preparation. Since adsorption energies vary widely based on the structure of the metal surface,[1] we have only used data of the hydrodeoxygenation of succinic acid for similar, closed-packed metal surface structures in the current work: Pd(111), Pt(111), Rh(111), Ru(0001), Cu(111), and Ni(111); all were obtained from VASP[15] calculations with the PBE-D3 functional. Data consist of 186 intermediate species for both ring and chain structures for each of the six metal surfaces.

The energy data were prepared to have the same reference values. For example, the adsorption energy for an intermediate surface species $C_xH_yO_z$ was calculated as

$$E_{C_xH_yO_z} = E^{DFT}_{C_xH_yO_z} - E^{DFT}_{*} - xE_C - yE_H - zE_O$$

where

$$E_C = E^{DFT}_{CH_4(g)} - 2E^{DFT}_{H_2(g)}$$

$$E_H = \frac{1}{2}E^{DFT}_{H_2(g)}$$

$$E_O = E^{DFT}_{H_2O(g)} - E^{DFT}_{H_2(g)}$$

Here, E^{DFT}_{*} is the energy of the free site (clean slab) and E^{DFT}_{X} denotes the adsorption energy of species X from the DFT calculations. The species energies are summarized in Table S1 of the Supporting Information, and coordinate files of all optimized species structures on all metal surfaces are also available in the Supporting Information.

2.2. Computational Methods. All calculations were carried out using the Vienna ab initio simulation package (VASP)[16−19] based on density functional theory (DFT) with the projector augmented wave (PAW) method .[20,21] The generalized gradient approximation (GGA)[22] with the Perdew−Burke−Ernzerhof (PBE) functional[23,24] was used to treat the electron exchange and correlation effects. An energy cutoff of 420 eV is used for all calculations, and the energy convergence criterion was set to 10^{-7} eV. All structures were relaxed until the Hellmann−Feynman force on each atom was smaller than 0.03 eV $Å^{-1}$. Considering that dispersion interactions have a significant effect on the adsorption and desorption processes of long-chain hydrocarbon molecules on surfaces[25,26] and that the PBE functional is unable to describe these van der Waals interactions, we included them into the calculations based on Grimme's DFT-D3 methodology.[26] To conduct the calculations, a 4 × 4 metal (111) surface slab with 4 metal atom layers (64 metal atoms in total) was constructed to simulate the metal surface. A 15 Å vacuum gap was set to the direction perpendicular to the surface plane to avoid the interactions between the periodic slabs. For all surface calculations, the bottom two layers were fixed to their bulk positions, while the top two layers were fully relaxed in all directions. The Brillouin zone integration was sampled by the 3 × 3 × 1 *k*-point mesh using the Monkhorst−Pack scheme.[27] Dipole corrections were applied to the direction perpendicular to the surface, and all the calculations on the Ni(111) surface are spin-polarized.

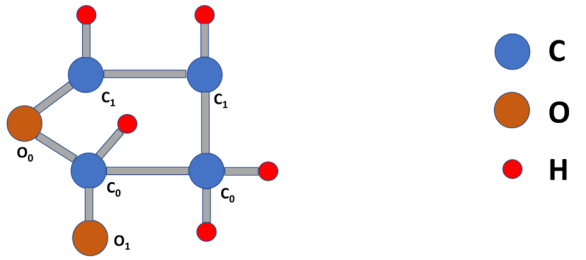
2.3. Metal Descriptors. Two different categories of metals descriptors were used in our work. First, we used the published values of electronic metal descriptors[14] for the six metal surfaces used in our study. The eight metal descriptors used are as follows: Pauling electronegativity, ionization potential, radius of *d*-orbitals, surface work function, *d*-band center, *d*-bandwidth, *d*-band filling, and density of sp-states at the Fermi level. Second, our predictive trials also included species adsorption energy-based metal descriptors (adsorption energies of CHCHCO, OH, and C) used in previous works.[8] We note that electronic descriptors have the advantage over adsorption energy descriptors that they can be obtained using look-up tables rather than performing expensive adsorption energy calculations for all surfaces. Also, there is hope that these descriptors are less surface chemistry-dependent given their rather general nature. On the other hand, the catalysis community has a lot of experience with the use of adsorption energies as metal descriptors that can be very reliable for some chemistries.[5−8,28,29]

2.4. Species Descriptor. Species descriptors can also be divided into two broad categories: coordinate-based and non-coordinate-based. The coordinate-based descriptors are computed using distance measures between each pair or triplet of atoms inside the species. Some of the commonly used methods in this category are Coulomb matrix[30] and bag-of-bonds.[31] The disadvantage of using atomic coordinates is that they have to be obtained by DFT or other semi-empirical methods which defeat the purpose that we do not want to run these expensive computations for all species. The other category is based on the count of different bond types around each atom[11] or in the molecule as a whole.[32]

In the current study, we have used the constant-sized flat molecular fingerprint based on previously published work[11] which is similar to extended connectivity fingerprint (ECFP)[12] and other constant-sized molecular descriptors.[32] The fingerprint, obtained from the molecule's SMILES notation, consists of the number of different types of bonds in the molecule. Here, the fingerprint vector contains more information than what is available with a basic bond count scheme which just calculates the number of C−C bonds or the number of C−O bonds and so on. Instead, each carbon or oxygen atom is denoted with the number of free valencies, and the fingerprint is made up of these more granular-level bond counts such as how many oxygen atoms with one free valence are connected to saturated carbon atoms and so forth. The fingerprint is described in Figure 1 with an example.

We note that particularly for ring species (but also for longer-chain species), describing the specific adsorption site is not easily done with pen-and-paper descriptors that do not use

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748



C_0	C_1	C_2	C_3	O_0	O_1	C-H	C=O	O-H	C-O_0	C-O_1
2	2	0	0	1	1	5	1	0	2	1

C_0-C_0	C_0-C_1	C_0-C_2	C_0-C_3	C_1-C_1	C_1-C_2	C_1-C_3	C_2-C_2	C_2-C_3	C_3-C_3
1	1	0	0	1	0	0	0	0	0

Figure 1. Molecular fingerprint for a ring-structured surface species. Here, C_0 denotes a saturated carbon (no free valence). C_1, C_2, and C_3 denote carbon atoms with one, two, and three free valencies, respectively. Similarly, O_0 is a saturated oxygen, whereas O_1 is an oxygen atom with one free valence. The fingerprint vector (shown at the bottom of the image) contains the number of different saturated or unsaturated atoms and the number of bonds between them.

coordinates. Under such conditions, we are most interested in the most stable adsorption configuration and an additional objective of this study is investigating what accuracy can be achieved for databases that only contain the most stable adsorption configuration of various adsorbed species with descriptors (fingerprints) that do not explicitly distinguish adsorption configurations [indirectly, the species descriptors can probably describe the adsorption configurations in that adsorbate species with, for example, carbon and oxygen atoms with one or two unpaired electrons (described in our fingerprint as C_1, C_2, and O_1) along with the corresponding C_1−C_2 bond and C_1−O_1 bond indices etc. describing sites within the adsorbate that prefer forming bonds with specific metal surface sites).

2.5. Machine Learning Models. The ML models in the current work can be broadly divided into two categories—linear models and kernel-based non-linear models. The goal was to compare the effects of different settings of species descriptors and metal descriptors on these categories of ML models.

The linear models used were linear ridge regression and lasso.[33] Both methods are linear regressions but with different types of regularizers: the L2 regularizer for ridge and L1 for lasso. The kernel-based models used were kernel ridge regression (KRR),[34] support vector regression (SVR),[35] and the Gaussian process (GP).[36] The GP model has an additional benefit over the other models in that besides the predictions, it also supplies the uncertainty measures around the predictions that can be useful in later stages during the calculation of a catalyst's turnover frequency or other macroscopic quantities of interest.[37,38] More advanced and complex ML models based on different structures of neural networks[12,39−45] or molecular graphs[46,47] have also been proposed but have not been used here. As will be shown later, our predictive trials indicate that all kernel-based models with tuned hyperparameters outperform linear models. The results of kernel-based models, however, have no statistically significant difference among themselves. This finding is consistent with previous works[8] and suggests that there is no basis to prefer one kernel-based model over another in terms of prediction accuracy. However, it should be pointed out that the GP provides the extra information about the uncertainty of the predictions compared to the other models and hence can be of importance for some scenarios where one has to study the effects of uncertainty propagation on the macroscopic quantities of interest.

A GP treats each data point as a random variable where any subset of these variables[48] forms a multivariate normal distribution. The relation between any pair of data points is defined by the kernel.[49] GP predicts the values for test points given the values of the training points. The uncertainty of the prediction is higher in the region of data space where the concentration of training points is low. Also, the opposite happens in the regions where there is a high number of training points. Support vector regression (SVR) predicts based on a subset of the training data which are called support vectors. Kernel ridge regression (KRR) provides closed form estimates while using a different loss function compared to SVR. The hyperparameters of the models such as which kernel to use and the parameters of the kernels such as length scale were tuned using 5-fold cross-validation.[50] We found Gaussian kernels to perform better than the Laplacian kernel.[51] The hyperparameters were tuned using cross-validation on the training set. The tuned hyperparameters were then used to predict on the testing sets to yield the performance measure of a trial run. The performance measures were averaged over all the trial runs to get the final prediction results.

2.6. Splitting Data into Training and Testing Sets. Using both linear and non-linear ML models, our predictive trials ran for four different types of splittings of the combined data set of chain- and ring-structured intermediate species.

First, for each metal surface, we trained on a random subset of the chain species and predicted on the rest. Out of the 186 species for a metal surface, 160 were randomly chosen and added to the training set, and the remaining 26 species were added to the testing set—with the process repeated for each metal. The training set thus obtained was used to train each of the ML models, and in each case, the predictions were done on the testing set.

Second, similar to the first case but working with ring structures instead of chain structures, we split the 186 ring species for a metal surface randomly into 160 to be added to the training set and the remaining 26 to be added to the testing set and finally perform the usual training of the ML model on the training set and then predict on the testing set.

Third, the chain data were split into 160−26 as in the first case and appended to the training and testing sets. This time, however, all the ring structure data are appended to the training set. This is the case where we have both ring and chain structures in training but only chain structures in the testing set. The prediction results compared with the first case would ascertain whether there is any statistically significant benefit obtained by including the ring structures in the training.

Fourth, similar to the third case but predicting on ring structures instead of chain structures, we performed the 160−26 split of ring structure data for each metal and then at the end appending all of the chain data to the training set and testing on the rest of the ring structures. Again, this will help us to see if inclusion of chain data increases the prediction accuracy on ring structures or not.

3. RESULTS AND DISCUSSION

The chain structure data for the hydrodeoxygenation of succinic acid contains information on 186 intermediate species

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748

across 6 metal surfaces, making the total size of the chain data set 1116. Similarly, there is a 1116-sized data set for the ring-structured species. The results of the predictions on the chain species are shown in Figure 2 and Table 1, and those of ring species are shown in Figure 3 and Table 2. For both the tables, the first 10 rows show predictions when using electronic metal descriptors, whereas the bottom 10 rows are for energy-based metal descriptors. There is no apparent advantage of using adsorption energies as metal descriptors relative to the tabulated electronic metal descriptors. For each set of the 10 rows, the first 5 show predictions for linear and non-linear ML models when training on either chain or ring species and predicting on the same category of species, whereas the second 5 rows are for predictions of linear and non-linear ML models when training on both ring and chain species and predicting on either chain (for the first table) or ring species (for Table 2).

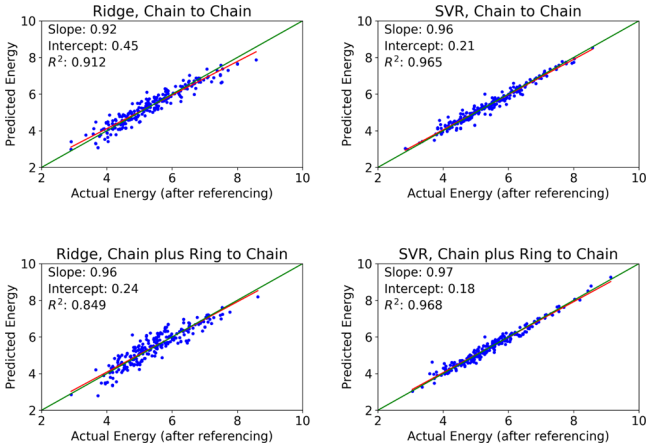


Figure 2. Predicted adsorption energy (after referencing) vs actual adsorption energy (after referencing) for predictions of chain-structured species. The top row shows prediction results for linear (on the left) and non-linear SVR (on the right) models when predicting chain data while training on chain data. The bottom row shows prediction results for linear (on the left) and non-linear SVR (on the right) models when predicting chain data while training on the rest of the chain plus all the ring structure data.

For each case, the whole process is repeated 100 times each time selecting a different random subset. The absolute error (AE) for each run is obtained by taking the absolute difference between the predicted and real energies. For both the tables, the mean and standard deviation of these absolute errors are shown in the fourth and fifth columns, respectively. Repeating the experiments 100 times also gives us 100 different MAEs. The standard deviations of these MAEs are shown in the sixth column and are generally quite small. For both Figures 2 and 3, the scatter plots show a random subset of prediction results of 100 different runs on the randomly selected test sets for all metal surfaces.

The key takeaway from these results are as follows: first, the prediction MAE for ring species using kernel-based ML models such as SVR is around 0.2 eV; second, inclusion of ring data when predicting chain and inclusion of chain data when predicting ring do not help to improve the predictions (to ensure that this was not because of different data distributions between the ring data and the chain data, we performed data normalization, but it did not have a statistically significant difference on the prediction accuracy); third, non-linear models outperform linear models; and fourth, the difference

Table 1. Results of Prediction of Chain-Structured Species[a]

	case	metal desc	model	MAE (eV)	SD of AEs (eV)	SD of MAEs (eV)
1	chain to chain	electronic	Ridge	0.247	0.200	0.014
2	chain to chain	electronic	Lasso	0.251	0.243	0.017
3	chain to chain	electronic	KRR	0.130	0.121	0.009
4	chain to chain	electronic	GP	0.133	0.119	0.009
5	chain to chain	electronic	SVR	0.126	0.123	0.008
6	chain plus ring to chain	electronic	Ridge	0.315	0.250	0.019
7	chain plus ring to chain	electronic	Lasso	0.321	0.249	0.017
8	chain plus ring to chain	electronic	KRR	0.154	0.131	0.008
9	chain plus ring to chain	electronic	GP	0.137	0.129	0.009
10	chain plus ring to chain	electronic	SVR	0.139	0.129	0.010
11	chain to chain	energy	Ridge	0.245	0.204	0.015
12	chain to chain	energy	Lasso	0.245	0.208	0.016
13	chain to chain	energy	KRR	0.130	0.122	0.010
14	chain to chain	energy	GP	0.129	0.124	0.011
15	chain to chain	energy	SVR	0.127	0.124	0.010
16	chain plus ring to chain	energy	Ridge	0.318	0.245	0.020
17	chain plus ring to chain	energy	Lasso	0.315	0.240	0.018
18	chain plus ring to chain	energy	KRR	0.158	0.134	0.009
19	chain plus ring to chain	energy	GP	0.137	0.139	0.011
20	chain plus ring to chain	energy	SVR	0.136	0.141	0.009

[a]The first five rows and rows 11 to 15 show the results for different ML models when training on a randomly selected subset of chain data and predicting on the rest of the chain species for all metal surfaces. Rows 6 to 10 and 16 to 20 show the results for linear and non-linear models when training on a randomly selected subset of chain data and all ring species and predicting on the rest of the chain species for all metal surfaces.

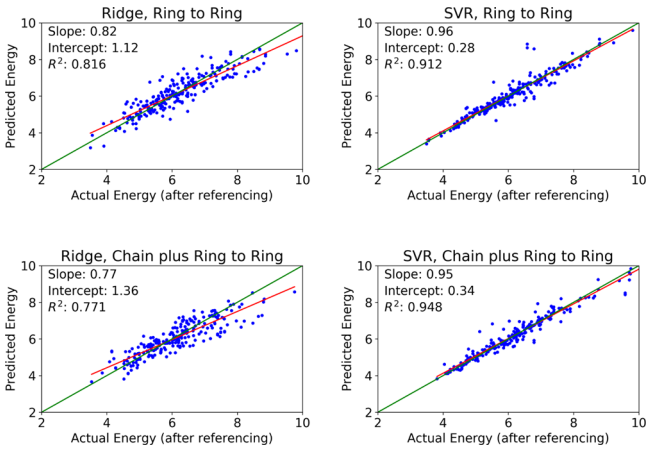


Figure 3. Predicted adsorption energy (after referencing) vs actual adsorption energy (after referencing) for predictions of ring-structured species. The top row shows prediction results for linear (on the left) and non-linear SVR (on the right) models when predicting ring data while training on ring data. The bottom row shows prediction results for linear (on the left) and non-linear SVR (on the right) models when predicting ring data while training on the rest of the ring plus all the chain structure data.

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748

Table 2. Results of Prediction of Ring-Structured Species[a]

	case	metal desc	model	MAE (eV)	SD of AEs (eV)	SD of MAEs (eV)
1	ring to ring	electronic	Ridge	0.380	0.336	0.025
2	ring to ring	electronic	Lasso	0.389	0.342	0.027
3	ring to ring	electronic	KRR	0.201	0.241	0.019
4	ring to ring	electronic	GP	0.203	0.237	0.019
5	ring to ring	electronic	SVR	0.192	0.245	0.017
6	chain plus ring to ring	electronic	Ridge	0.440	0.401	0.028
7	chain plus ring to ring	electronic	Lasso	0.428	0.395	0.027
8	chain plus ring to ring	electronic	KRR	0.212	0.242	0.019
9	chain plus ring to ring	electronic	GP	0.199	0.254	0.018
10	chain plus ring to ring	electronic	SVR	0.202	0.248	0.018
11	ring to ring	energy	Ridge	0.385	0.331	0.026
12	ring to ring	energy	Lasso	0.392	0.329	0.027
13	ring to ring	energy	KRR	0.203	0.229	0.018
14	ring to ring	energy	GP	0.204	0.226	0.020
15	ring to ring	energy	SVR	0.190	0.248	0.018
16	chain plus ring to ring	energy	Ridge	0.435	0.401	0.030
17	chain plus ring to ring	energy	Lasso	0.434	0.401	0.029
18	chain plus ring to ring	energy	KRR	0.209	0.243	0.019
19	chain plus ring to ring	energy	GP	0.200	0.252	0.017
20	chain plus ring to ring	energy	SVR	0.204	0.251	0.018

[a]The first five rows and rows 11 to 15 show the results for different ML models when training on a randomly selected subset of chain data and predicting on the rest of the chain species for all metal surfaces. Rows 6 to 10 and 16 to 20 show the results for linear and non-linear models when training on a randomly selected subset of ring data and all chain species and predicting on the rest of the ring species for all metal surfaces.

between prediction accuracy of electronic metal descriptors and adsorption energy-based metal descriptors is not statistically significant.

Another observation on the results is that the prediction errors on ring species are significantly higher than those on chain species. One possible explanation for this is that dehydrogenated ring species prefer to form strong bonds with specific sites of the surface metal atoms (atop vs bridge vs three-fold hollow); however, this leads to significant strain in the ring structure, and the optimized structures, for significantly dehydrogenated ring species such as adsorbed C_4O on Pt(111), often possess both elongated and/or compressed bonds within the ring atoms and the metal sites. Figure 4 illustrates that the spread of the energy values is bigger for ring species compared to the chain species in our data set. Also, the ring structure data contain more outliers, and we found that these high-energy outliers are consistently significantly dehydrogenated species with elongated and/or compressed bonds. Fingerprints that only consider nearest neighbor atoms are limited in capturing the properties of such species, and more complex fingerprints are needed if it is desired to describe the properties of deeply dehydrogenated surface ring species. However, given that these species are usually high-energy species that are likely not kinetically

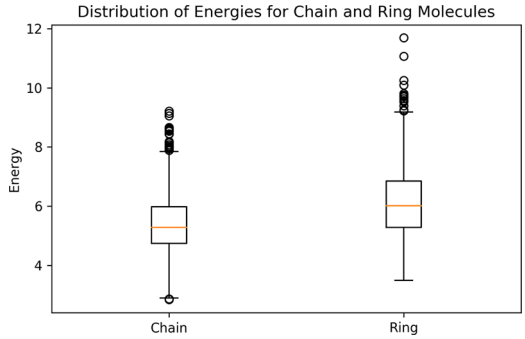


Figure 4. Box plots comparing the energy distributions in chain and ring data sets. The ring data set has not only a bigger spread but also more outliers.

relevant and that more training data are required for models with more complex fingerprints, it might be acceptable to use the fingerprints of this study and have a higher prediction error for deeply dehydrogenated ring species.

4. CONCLUSIONS

Working with two data sets on the most stable chain and ring structures on six different metals surfaces, our comparative study on predicting adsorption energies of these two different structures has revealed some key insights. We have seen that although ring structures had a higher predictive error, it was still below 0.2 eV when working with simple SMILES-based flat fingerprints and electronic or adsorption energy-based metal descriptors along with regular non-linear ML models. Our results also indicate that the non-linear models perform better than the linear models when the predictive model requires species descriptors as well as metal descriptors. Another key outcome from the current study is that information on chain species does not help in predicting ring species and vice versa for current species descriptors. We highlight that these results have been obtained with species descriptors that do not explicitly describe specific adsorption sites but that we used a database containing only the most stable adsorption configuration and energy as it is currently typical in many databases.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.1c05470.

Energy data and SMILES representation for all species; species fingerprints; electronic metal descriptors; adsorption energy metal descriptors; and coordinate files for species (PDF)

Cu(111) adsorption coordinates, Ni(111) adsorption coordinates, Pd(111) adsorption coordinates, Pt(111) adsorption coordinates, Rh(111) adsorption coordinates, and Ru(0001) adsorption coordinates (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Andreas Heyden − *Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States;* orcid.org/0000-0002-4939-7489; Email: heyden@cec.sc.edu

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748

Gabriel A. Terejanu − *Department of Computer Science, University of North Carolina at Charlotte, Charlotte, North Carolina 28262, United States*; Email: gterejan@uncc.edu

Authors

Asif J. Chowdhury − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Wenqiang Yang − *Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.1c05470

Author Contributions

‖A.J.C. and W.Y. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grant no. DMREF-1534260 (the bulk of the machine learning research and the DFT data on the HDO of succinic acid) and the U.S. Department of Energy Office of Science, Office of Basic Energy Sciences, Catalysis Science Program, under Award DE-SC0007167 (most of the DFT data on the HDO of propanoic acid). K.E.A. acknowledges financial support from the National Science Foundation under grant no. OIA-1632824. Finally, this work was partially supported by the South Carolina SmartState Center for Strategic Approaches to the Generation of Electricity (SAGE). Computational resources provided by XSEDE facilities located at San Diego Supercomputer Center (SDSC) and Texas advanced Computing Center (TACC) under grant number TG-CTS090100, U.S. Department of Energy facilities located at the National Energy Research Scientific Computing Center (NERSC) under contract no. DE-AC02-05CH11231 and Pacific Northwest National Laboratory (Ringgold ID 130367, grant proposals 51163 and 51711) and the High-Performance Computing clusters located at the University of South Carolina are gratefully acknowledged.

■ REFERENCES

(1) Nørskov, J. K.; Studt, F.; Abild-Pedersen, F.; Bligaard, T. *Fundamental Concepts in Heterogeneous Catalysis*; John Wiley and Sons: Hoboken, New Jersey, 2014; Chapter 2, pp 17−19.

(2) de Carvalho, T. P.; Catapan, R. C.; Oliveira, A. A. M.; Vlachos, D. G. Microkinetic Modeling and Reduced Rate Expression of the Water-Gas Shift Reaction on Nickel. *Ind. Eng. Chem. Res.* **2018**, *57*, 10269−10280.

(3) Wittreich, G. R.; Alexopoulos, K.; Vlachos, D. G.. In *Handbook of Materials Modeling: Applications: Current and Emerging Materials*; Andreoni, W., Yip, S., Eds.; Springer International Publishing, 2020; pp 1377−1404.

(4) Bo, C.; Maseras, F.; López, N. The role of computational results databases in accelerating the discovery of catalysts. *Nat. Catal.* **2018**, *1*, 809−810.

(5) 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.

(6) Busch, M.; Wodrich, M. D.; Corminboeuf, C. Linear Scaling Relationships and Volcano Plots in Homogeneous Catalysis - Revisiting the Suzuki Reaction. *Chem. Sci.* **2015**, *6*, 6754−6761.

(7) 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.

(8) 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.

(9) Li, X.; Chiong, R.; Hu, Z.; Cornforth, D.; Page, A. J. Improved Representations of Heterogeneous Carbon Reforming Catalysis Using Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6882−6894.

(10) Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Machine learning in chemical reaction space. *Nat. Commun.* **2020**, *11*, 5505.

(11) 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 PMID: 31962041.

(12) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754 PMID: 20426451.

(13) 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.

(14) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752−2759.

(15) Hafner, J. Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *J. Comput. Chem.* **2008**, *29*, 2044−2078.

(16) 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.

(17) Kresse, G.; Hafner, J. Ab initio molecular dynamics for open-shell transition metals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *48*, 13115−13118.

(18) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47*, 558−561.

(19) 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.

(20) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953−17979.

(21) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758−1775.

(22) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(23) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244−13249.

(24) Perdew, J. P.; Yue, W. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B* **1986**, *33*, 8800−8802.

(25) Karp, E. M.; Silbaugh, T. L.; Crowe, M. C.; Campbell, C. T. Energetics of Adsorbed Methanol and Methoxy on Pt(111) by Microcalorimetry. *J. Am. Chem. Soc.* **2012**, *134*, 20388−20395 PMID: 23181692.

(26) 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.

(27) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.

(28) Wang, Y.; Qiu, W.; Song, E.; Gu, F.; Zheng, Z.; Zhao, X.; Zhao, Y.; Liu, J.; Zhang, W. Adsorption-energy-based activity descriptors for electrocatalysts in energy storage applications. *Natl. Sci. Rev.* **2017**, *5*, 327−341.

(29) Huang, H.-C.; Li, J.; Zhao, Y.; Chen, J.; Bu, Y.-X.; Cheng, S.-B. Adsorption energy as a promising single-parameter descriptor for

single atom catalysis in the oxygen evolution reaction. *J. Mater. Chem. A* **2021**, *9*, 6442−6450.
(30) Rupp, M.; Tkatchenko, A.; Müller, K. R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(31) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(32) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.
(33) Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. *IEEE International Conference on Advances in Computer Applications (ICACA)*; 2016; pp 18−20.
(34) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.
(35) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. *Advances in Neural Information Processing Systems*; 1997.
(36) Seeger, M. Gaussian Processes for Machine Learning. *Int. J. Neural Syst.* **2004**, *14*, 69−106.
(37) Walker, E. A.; Mitchell, D.; Terejanu, G. A.; Heyden, A. Identifying Active Sites of the Water-Gas Shift Reaction over Titania Supported Platinum Catalysts under Uncertainty. *ACS Catal.* **2018**, *8*, 3990−3998.
(38) Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A. Uncertainty Quantification Framework Applied to the Water-Gas Shift Reaction over Pt-Based Catalysts. *J. Phys. Chem. C* **2016**, *120*, 10328−10339.
(39) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(40) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(41) Morawietz, T.; Behler, J. A density-functional theory-based neural network potential for water clusters including van der Waals corrections. *J. Phys. Chem. A* **2013**, *117*, 7356−7366 PMID: 23557541.
(42) Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
(43) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO2 reduction. *ACS Catal.* **2017**, *7*, 6600−6608.
(44) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(45) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(46) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(47) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P..Convolutional networks on graphs for learning molecular fingerprints. *Proceedings of the 28th International Conference on Neural Information Processing Systems*; Cambridge, MA, USA, 2015; Vol. 2, pp 2224−2232.
(48) Rasmussen, C. E. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2−14, 2003, Tübingen, Germany, August 4−16, 2003, Revised Lectures*; Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer: Berlin, Heidelberg, 2004; pp 63−71.
(49) Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Ann. Stat.* **2008**, *36*, 1171−1220.
(50) Kohavi, R. A. Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*; San Francisco, CA, USA, 1995; Vol. 2, pp 1137−1143.
(51) Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.

https://doi.org/10.1021/acs.jpcc.1c05470
J. Phys. Chem. C 2021, 125, 17742−17748