Supporting Information

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

Jinwoong Nam[1], Charanyadevi Ramasamy[1], Daniel E. Raser[1], Gustavo L. Barbosa Couto[1], Lydia Thies[2], David Hibbitts[2], and Fuat E. Celik[1]*

[1]*Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, 98 Brett Road, Piscataway, NJ 08854, USA*

[2]*Department of Chemical Engineering, University of Florida, 1030 Center Drive Gainesville, FL 32611, USA*

* Corresponding author:

E-mail address: fuat.celik@rutgers.edu (Fuat. E. Celik)

S1. Coordinates of the Relaxed Structures from DFT Calculation for All 384 Species Used in This Study

All relaxed structures of 384 species obtained from DFT calculation are saved in the attached file below.

CONTCAR files

S2. Referencing the Enthalpy of Formation[41]

The referencing was performed based on the NIST heat of formation with the equation below where E_C and E_H are calculated to adjust the energy of each element from DFT calculation.

$$\boldsymbol{H}_{\mathrm{f,298K}}^{ref,NIST} = \boldsymbol{H}_{\mathrm{298K}}^{ref,DFT} + \boldsymbol{X}_{\mathrm{comp}}^{ref} \begin{pmatrix} E_\mathrm{C} \\ E_\mathrm{H} \end{pmatrix}$$

Ethane, ethylene, methane, and hydrogen in the gas phase were used as reference molecules. $\boldsymbol{H}_{\mathrm{f,298K}}^{ref,NIST}$ and $\boldsymbol{H}_{\mathrm{298K}}^{ref,DFT}$ are the 4×1 matrices where the entries are the enthalpies of formation of each reference molecule, from NIST and DFT calculation, respectively. $\boldsymbol{X}_{\mathrm{comp}}^{ref}$ is the matrix describing the composition of carbon and hydrogen for each reference molecule in which each row represents the reference species, and each column corresponds to each element (C for the first, H for the second column). By solving the equation with a Least Squares approach, E_C and E_H are obtained.

S3. Molecular Fingerprints Components

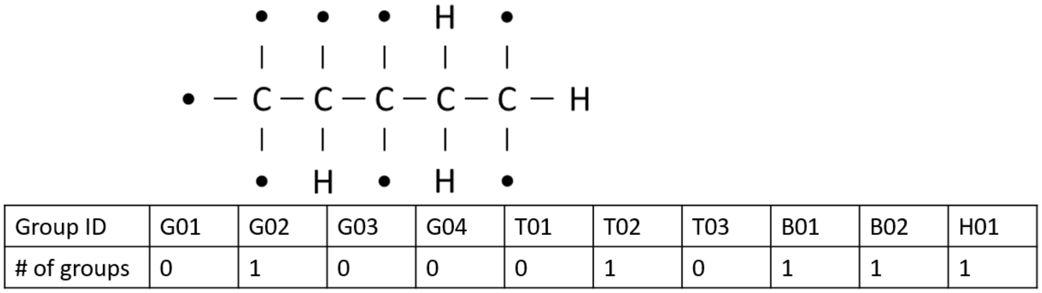
S3.1 Group Additivity Fingerprint (GA)[41, 43-46]

The group additivity components that describe the species used in this study are displayed in Table S1. An example of a species represented by the GA is depicted in Figure S1.

Table S1. List of the group components for hydrocarbon adsorbates used in the GA

Group ID	Group Components
G01	C(C)(H)3
G02	C(C)2(H)2
G03	C(C)3(H)
G04	C(C)4
T01	C(C)(H)2(•)
T02	C(C)2(H)(•)
T03	C(C)3(•)
B01	C(C)(H)(•)2
B02	C(C)2(•)2
H01	C(C)(•)3

(•) denotes a free valency



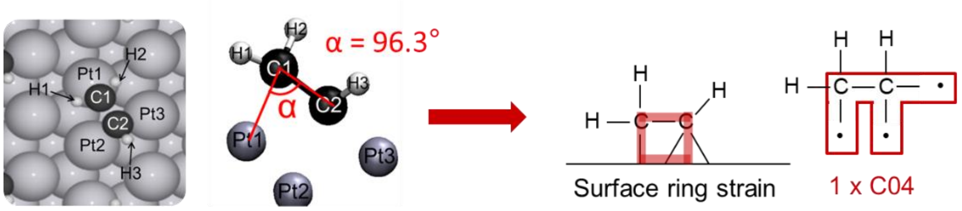
Group ID	G01	G02	G03	G04	T01	T02	T03	B01	B02	H01
# of groups	0	1	0	0	0	1	0	1	1	1

(•) denotes a free valency

Figure S1. Example of the GA assigned to the adsorbate

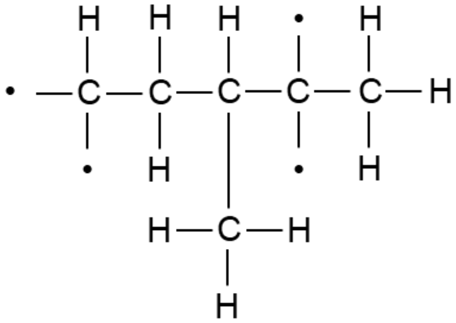
S3.2 Group Additivity with Surface Structure (GASS)[41, 43-46]

An example of considering the surface strain effect for adsorbed species along with five additional group components used is highlighted in Figure S2. An example of a species represented by the GASS is depicted in Figure S3.



Group ID	Group
C01	C(•)3C(•)2
C02	C(•)3C(•)
C03	C(•)2C(•)2
C04	C(•)2C(•)
C05	C(•)C(•)

Figure S2. Example of surface ring strain with vinyl on Pt(111) and the table of correction groups used in the GASS



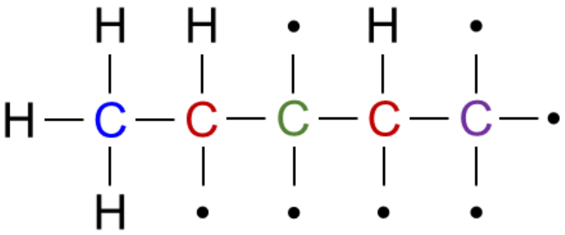
G01	G02	G03	G04	T01	T02	T03	B01	B02	H01	C01	C02	C03	C04	C05
1	0	1	0	0	1	0	2	1	0	0	0	1	1	0

(•) denotes a free valency

Figure S3. Example of the GASS assigned to the adsorbate

S3.3 Flat Molecular Fingerprint (FMF)[47]

15 FMF components used in this study and their application are shown in Figure S4.



C0	C1	C2	C3	C-H	C0-C0	C0-C1
1	2	1	1	5	0	1

C0-C2	C0-C3	C1-C1	C1-C2	C1-C3	C2-C2	C2-C3	C3-C3
0	0	0	2	1	0	0	0

C0: saturated carbon (no free valence)

C1-C3: carbon with one, two, and three free valencies

Figure S4. Example of the FMF assigned to the adsorbate

S3.4 Sequential Valency-Connectivity Fingerprint (SVCF)

Sequential Valency-Connectivity Fingerprint (SVCF) was developed to incorporate sequential information on the type of carbons and bonds within adsorbates. To keep the consistency of the carbon numbering for any studied species, the rules were defined based on IUPAC nomenclature guidelines[48] as explained in S3.4.1. All SVCF components along with an example of their application to the adsorbate are shown in Figure S5.

S3.4.1 Carbon numbering rules for the SVCF

1. Selection of the main chain (longest chain)
    - Always pick the longest chain as the main chain, then other parts as branches
    - If multiple chains have the same longest chain length, pick up the chain that has the highest total free valency as the main

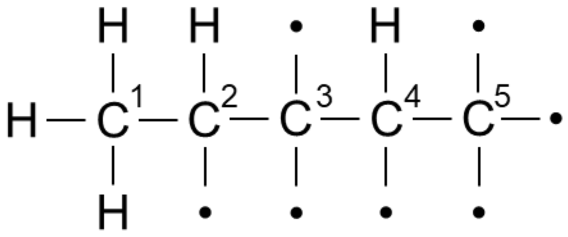
2. Main counting rules for acyclic species
    - For a single chain without any branches, start numbering with an end carbon that has a higher free valency number. If both end carbons have the same free valency, see the second carbons from the end, and the third, and so on
    - For a chain having functional groups or substitutional groups (e.g. methyl) as branches, count the carbons in the main chain first, then the branch carbons
    - For branches, always count the carbons directly attached to the main chain first

3. Specific counting rules for acyclic species
    - When a single branch is attached to a symmetrical main chain, make the carbon number that has a branch attached the lowest
    - When multiple branches exist at the same location of the main chain, the carbons in the branch that have a higher total free valency are numbered first, then the carbons in other branches
    - When multiple branches exist at different locations of the main chain,
        - i. For a non-symmetrical main chain, count first the branched carbon attached to a lower carbon number in the main chain
        - ii. For a symmetrical main chain, count first the branch that has a higher total free valency

4. Cyclic species
    - Start numbering with a carbon that has a higher free valency number, then count the carbons in a clockwise direction. If multiple carbons have the same free

valency, see the second carbons in the clockwise direction, and, the third, and so on

- If the ring has linear branches,
  - i. Count the carbons in the (main) ring first, then branch carbons
  - ii. Follow the same rules as the linear species



1st C	2nd C	3rd C	4th C	5th C	6th C
0	1	2	1	3	0

$C_{1st} - C_{2nd}$	$C_{1st} - C_{3rd}$	$C_{1st} - C_{4th}$	$C_{1st} - C_{5th}$	$C_{1st} - C_{6th}$	$C_{2nd} - C_{3rd}$	$C_{2nd} - C_{4th}$	$C_{2nd} - C_{5th}$
1	0	0	0	0	1	0	0

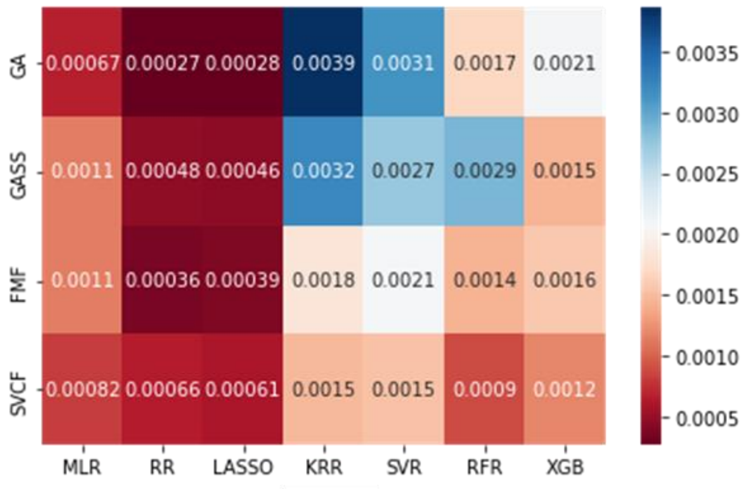
$C_{2nd} - C_{6th}$	$C_{3rd} - C_{4th}$	$C_{3rd} - C_{5th}$	$C_{3rd} - C_{6th}$	$C_{4th} - C_{5th}$	$C_{4th} - C_{6th}$	$C_{5th} - C_{6th}$
0	1	0	0	1	0	0

# of C	# of H
5	5

Figure S5. Example of the SVCF assigned to the adsorbate

S4. Heat Maps of the Second (variance) and Third (skewness) Statistical Moments for the Distributions of the Mean and Maximum Absolute Errors Across the Iterations

(a) Variance of mean absolute errors in enthalpy of formation per carbon (in $(\text{kcal/mol}_C)^2$)



(b) Variance of maximum absolute errors in enthalpy of formation per carbon (in $(\text{kcal/mol}_C)^2$)

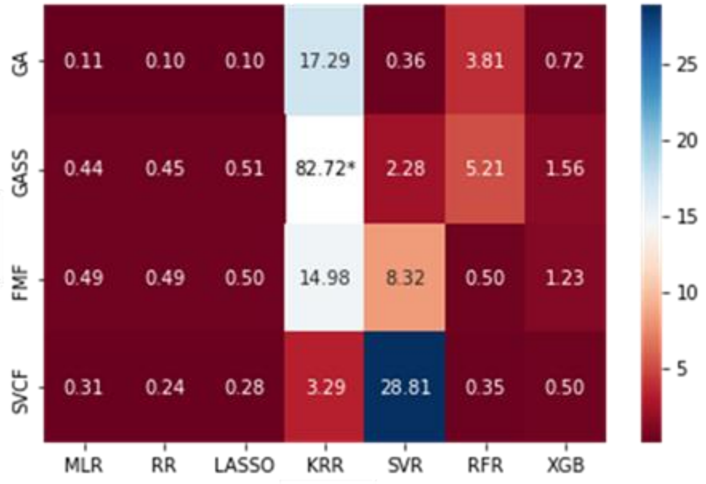
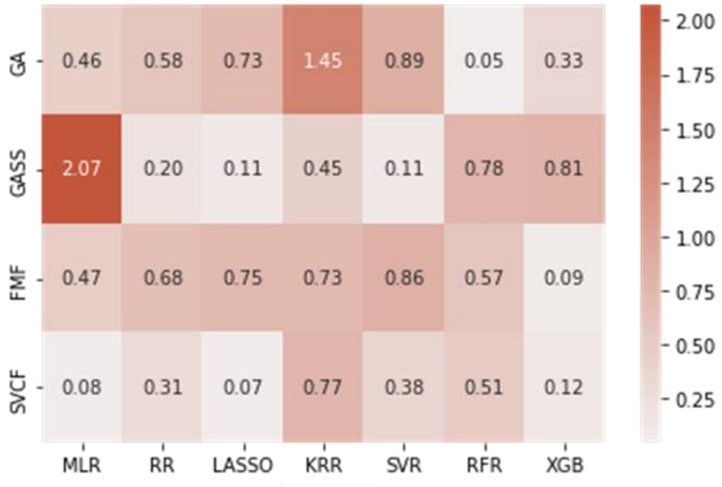


Figure S6. Heat maps for the variance of (a) mean and (b) maximum absolute errors of each model in enthalpy of formation per carbon across the 30 iterations. All numbers in both heat maps are in (kcal/molC)2. *In b), the variance of the KRR+GASS model is excluded in coloring (shown in white) for effective comparison of other models as the number is much higher than others.

(a) Absolute skewness of mean absolute errors in enthalpy of formation per carbon



(b) Absolute skewness of maximum absolute errors in enthalpy of formation per carbon

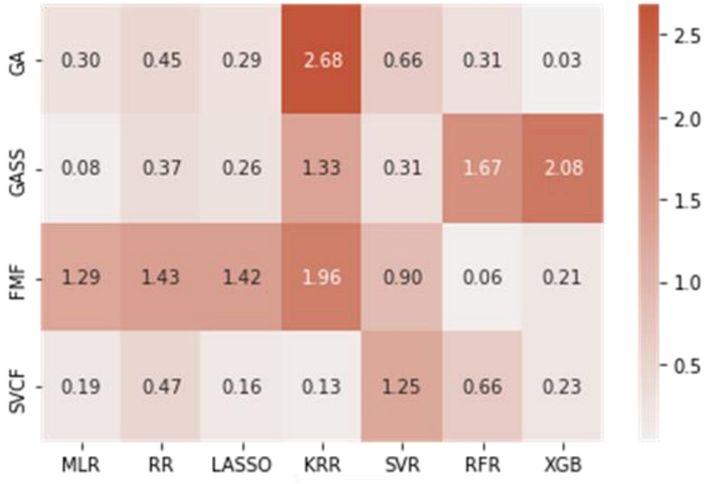


Figure S7. Heat maps for the absolute skewness of (a) mean and (b) maximum absolute errors of each model in enthalpy of formation per carbon across the 30 iterations. The absolute values of skewness for each distribution are shown.

S5. Statistical Hypothesis Testing with Individual Machine Learning Models

The independent samples t-test was performed for the distributions of the error metric with all models using KRR and XGB to statistically identify the difference in performance. The results of the calculated p-values are presented in Table S2.

Table S2. Calculated *p*-values for the independent samples t-test between KRR and XGB models

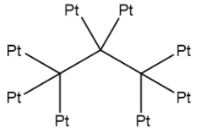
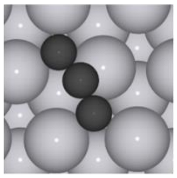
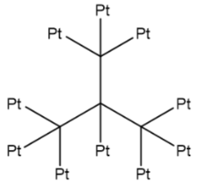
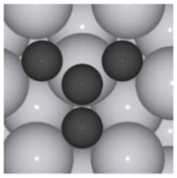
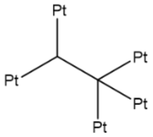
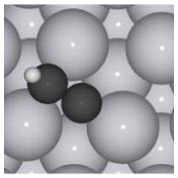
	KRR+ GA	KRR+ GASS	KRR+ FMF	KRR+ SVCF	XGB+ GA	XGB+ GASS	XGB+ FMF	XGB+ SVCF
KRR+GA	1	**0.0894**	0.0010	**0.6687**	0.0001	0.0	0.0	0.0
KRR+GASS	**0.0894**	1	0.0	0.0125	0.0248	0.0008	0.0078	0.0
KRR+FMF	0.0010	0.0	1	0.0002	0.0	0.0	0.0	0.0
KRR+SVCF	**0.6687**	0.0125	0.0002	1	0.0	0.0	0.0	0.0
XGB+GA	0.0001	0.0248	0.0	0.0	1	**0.2254**	**0.7074**	0.0
XGB+GASS	0.0	0.0008	0.0	0.0	**0.2254**	1	**0.3660**	0.0
XGB+FMF	0.0	0.0078	0.0	0.0	**0.7074**	**0.3660**	1	0.0
XGB+SVCF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1

*Numbers in bold indicate the p-values larger than the significance level of 0.05

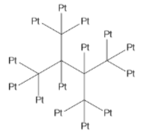
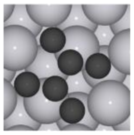
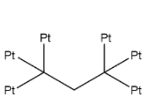
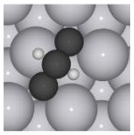
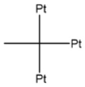
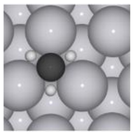
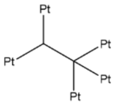
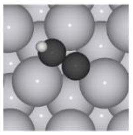
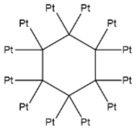
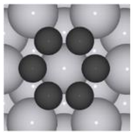
S6. Analysis of the Species with the Maximum Absolute Errors

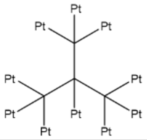
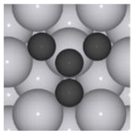
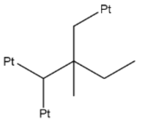
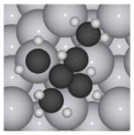
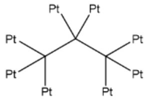
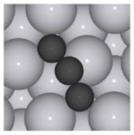
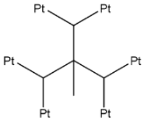
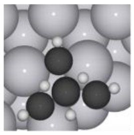
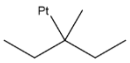
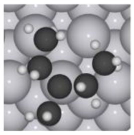
Figure S8. List of the species with the maximum absolute errors for (a) all twelve linear models of MLR, RR, and LASSO, and (b) the selected four non-linear models (KRR+GASS, SVR+SVCF, KRR+GA, KRR+FMF). Please note in this representation that "Pt" indicates a free valency that "could" binds to a Pt surface atom. In many species, these remain unsaturated (i.e. pi-bonded) instead.

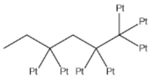
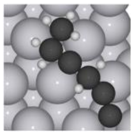
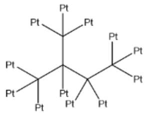
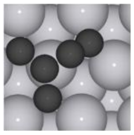
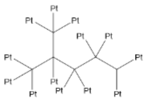
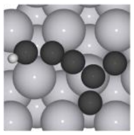
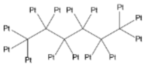
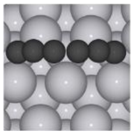
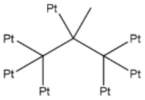
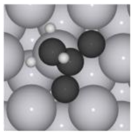
(a) Species with the maximum absolute errors in 360 iterations for 12 linear models using MLR, RR, and LASSO

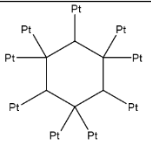
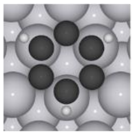
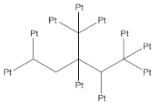
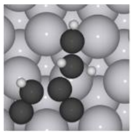
Species	Relaxed structure (Top view)	# of cases
		294
		63
		3

(b) Species with the maximum absolute errors in 120 iterations for the selected non-linear models

Species	Relaxed structure (Top view)	# of cases
		34
		19
		18
		11
		10

Species	Relaxed structure (Top view)	# of cases
		6
		5
		3
		3
		3

Species	Relaxed structure (Top view)	# of cases
		2
		1
		1
		1
		1

Species	Relaxed structure (Top view)	# of cases
		1
		1

S7. Comparison of the Errors Between Normal and Randomly Disturbed Datasets

The mean absolute errors of the ML models trained with the normal and randomly disturbed datasets are presented in Tables S3 and S4.

Table S3. Mean absolute errors in enthalpy of formation per carbon for the single iteration using the normal datasets (in kcal/mol$_C$)

	MLR	**RR**	**LASSO**	**KRR**	**SVR**	**RFR**	**XGB**
GA	1.96	1.95	1.95	1.24	1.50	1.60	1.32
GASS	1.79	1.79	1.77	1.23	1.65	1.72	1.37
FMF	1.93	1.86	1.86	1.11	1.43	1.48	1.30
SVCF	2.14	2.07	2.08	1.22	1.76	1.37	1.07

Table S4. Mean absolute errors in enthalpy of formation per carbon for the single iteration using the randomly disturbed (RD) datasets (in kcal/mol$_C$)

	MLR	**RR**	**LASSO**	**KRR**	**SVR**	**RFR**	**XGB**
GA	1.96	1.95	1.95	1.32	1.65	1.69	1.50
GASS	1.79	1.80	1.78	1.26	1.74	1.77	1.56
FMF	1.93	1.88	1.87	1.25	1.65	1.53	1.36
SVCF	2.12	2.07	2.11	1.26	1.73	1.39	1.18

S8. Effect of the Error Cancellation for Individual Species

By comparing mean errors with mean absolute errors across the iterations, possible error cancellation effects that can misleadingly improve the prediction of some species were investigated for KRR+FMF and XGB+SVCF models.

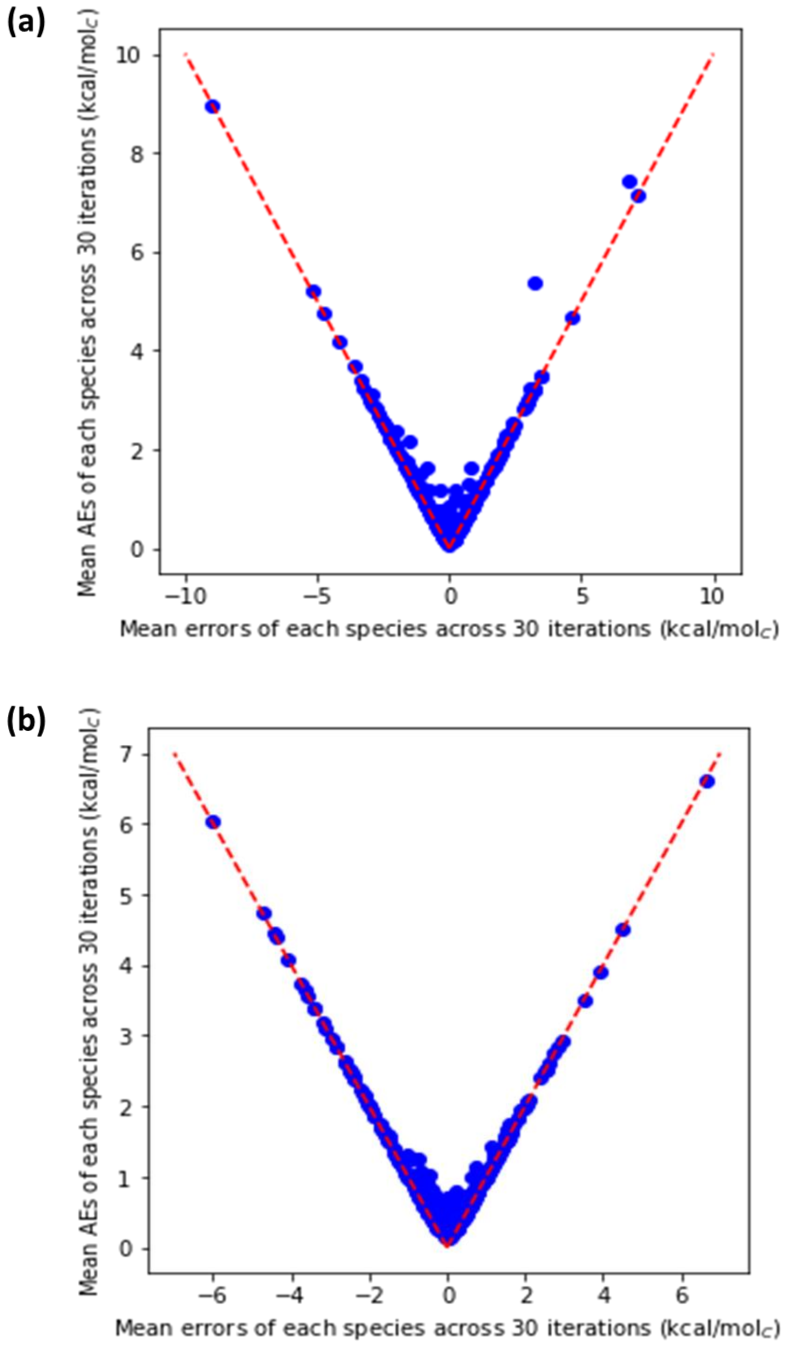


Figure S9. Plots of mean errors and mean absolute errors of individual species for (a) KRR+FMF and (b) XGB+SVCF. The red dotted lines denote $y = |x|$.

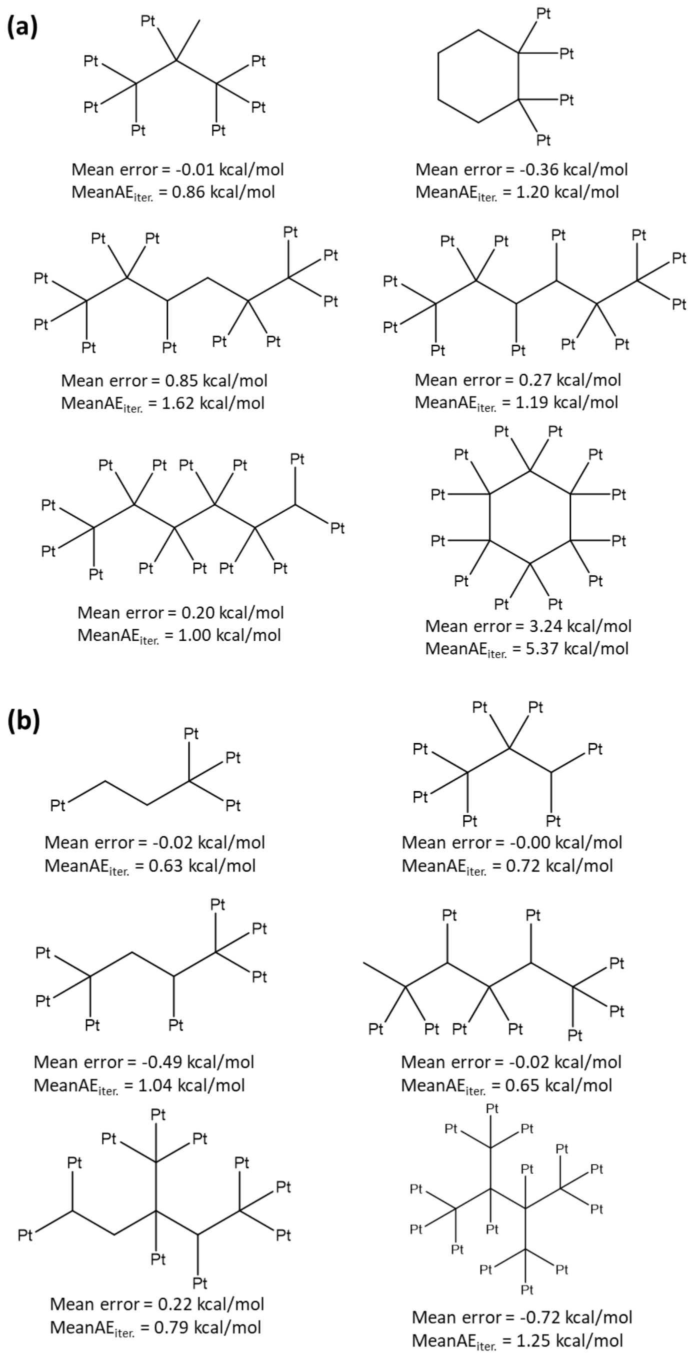

(a)
Mean error = -0.01 kcal/mol
MeanAE$_{iter.}$ = 0.86 kcal/mol
Mean error = -0.36 kcal/mol
MeanAE$_{iter.}$ = 1.20 kcal/mol
Mean error = 0.85 kcal/mol
MeanAE$_{iter.}$ = 1.62 kcal/mol
Mean error = 0.27 kcal/mol
MeanAE$_{iter.}$ = 1.19 kcal/mol
Mean error = 0.20 kcal/mol
MeanAE$_{iter.}$ = 1.00 kcal/mol
Mean error = 3.24 kcal/mol
MeanAE$_{iter.}$ = 5.37 kcal/mol
(b)
Mean error = -0.02 kcal/mol
MeanAE$_{iter.}$ = 0.63 kcal/mol
Mean error = -0.00 kcal/mol
MeanAE$_{iter.}$ = 0.72 kcal/mol
Mean error = -0.49 kcal/mol
MeanAE$_{iter.}$ = 1.04 kcal/mol
Mean error = -0.02 kcal/mol
MeanAE$_{iter.}$ = 0.65 kcal/mol
Mean error = 0.22 kcal/mol
MeanAE$_{iter.}$ = 0.79 kcal/mol
Mean error = -0.72 kcal/mol
MeanAE$_{iter.}$ = 1.25 kcal/mol
Pt

Figure S10. List of the species with a deviation of (a) larger than 0.75 kcal/mol$_C$ from $y = |x|$ in Figure S9(a) and (b) larger than 0.5 kcal/mol$_C$ from $y = |x|$ in Figure S9(b). For each species, the mean error of all iterations and Mean AEiter. are specified. Please note in this representation that "Pt" indicates a free valency that "could" bind to a Pt surface atom. In many species, these remain unsaturated (i.e. pi-bonded) instead.

S9. Ensemble Average Voting Model of KRR+GA and XGB+GA

The ensemble model of KRR+GA and XGB+GA was used for massive enthalpy prediction of 3115 C_2 to C_6 acyclic hydrocarbon adsorbates. The parity plot of this model is shown in Figure S11. ML predicted enthalpies for each species were the mean values of 30 iterations.

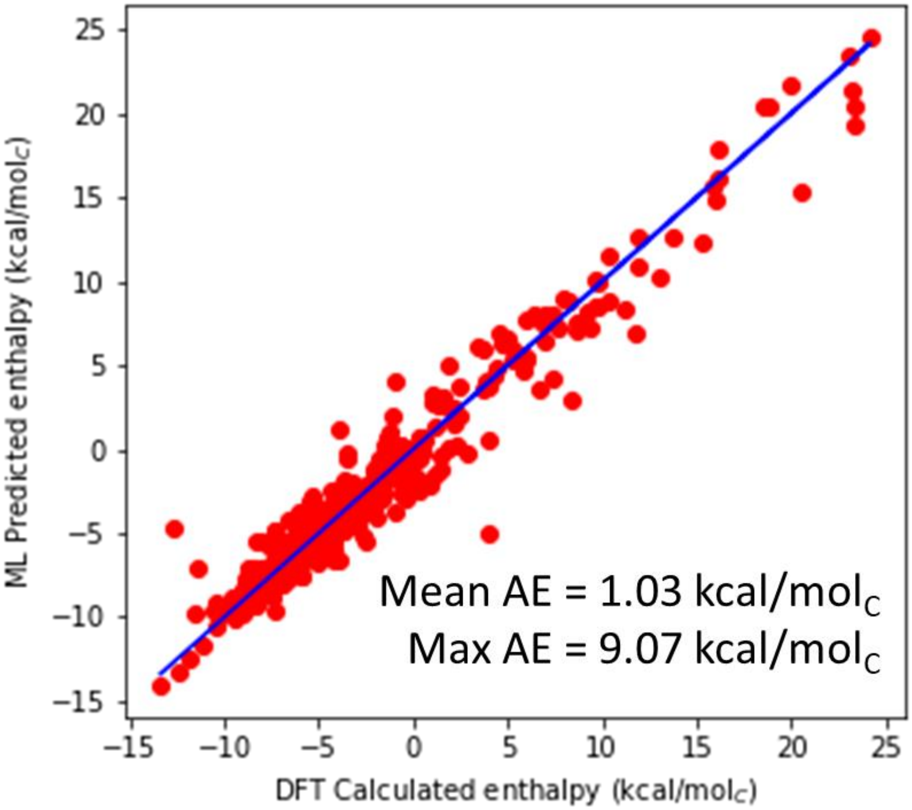


Figure S11. Parity plot of the ML predicted and DFT calculated enthalpy of formation per carbon for the ensemble average voting model of KRR+GA and XGB+GA. Mean AE: Mean absolute error, Max AE: Maximum absolute error

S10. Hyperparameters Used for Two Best Individual Models (KRR+FMF, XGB+SVCF)

Table S5. Best hyperparameters used for KRR+FMF in 150 different folds

Alpha	Degree	Gamma	Kernel	The number of times used
0.005	2	0.01	rbf	43
0.001	2	0.01	rbf	26
0.01	4	0.01	polynomial	22
0.05	5	0.01	polynomial	20
0.005	3	0.01	polynomial	9
0.01	3	0.01	polynomial	6
0.005	4	0.01	polynomial	4
0.01	2	0.01	polynomial	3
0.3	2	0.1	polynomial	3
0.005	2	0.01	polynomial	3
0.2	2	0.1	polynomial	2
1	3	0.1	polynomial	2
0.05	4	0.01	polynomial	2
0.1	5	0.01	polynomial	1
0.05	2	0.01	rbf	1
0.01	2	0.01	rbf	1
0.5	2	0.1	polynomial	1
0.01	5	0.01	polynomial	1

Table S6. Best hyperparameters used for XGB+SVCF in 150 different folds

colsample_bytree	learning_rate	max_depth	min_child_weight	n_estimators	subsample	The number of times used
1	0.1	3	3	200	0.5	75
1	0.1	5	3	200	0.5	54
0.5	0.1	3	3	200	0.5	5
1	0.1	3	3	200	1	5
1	0.1	7	3	200	0.5	5
0.5	0.1	5	3	200	1	3
1	0.1	5	3	200	1	2
1	0.1	3	5	200	0.5	1

S11. Prediction results of the ensemble average voting model of KRR+FMF and XGB+SVCF with respect to carbon number and cyclicity

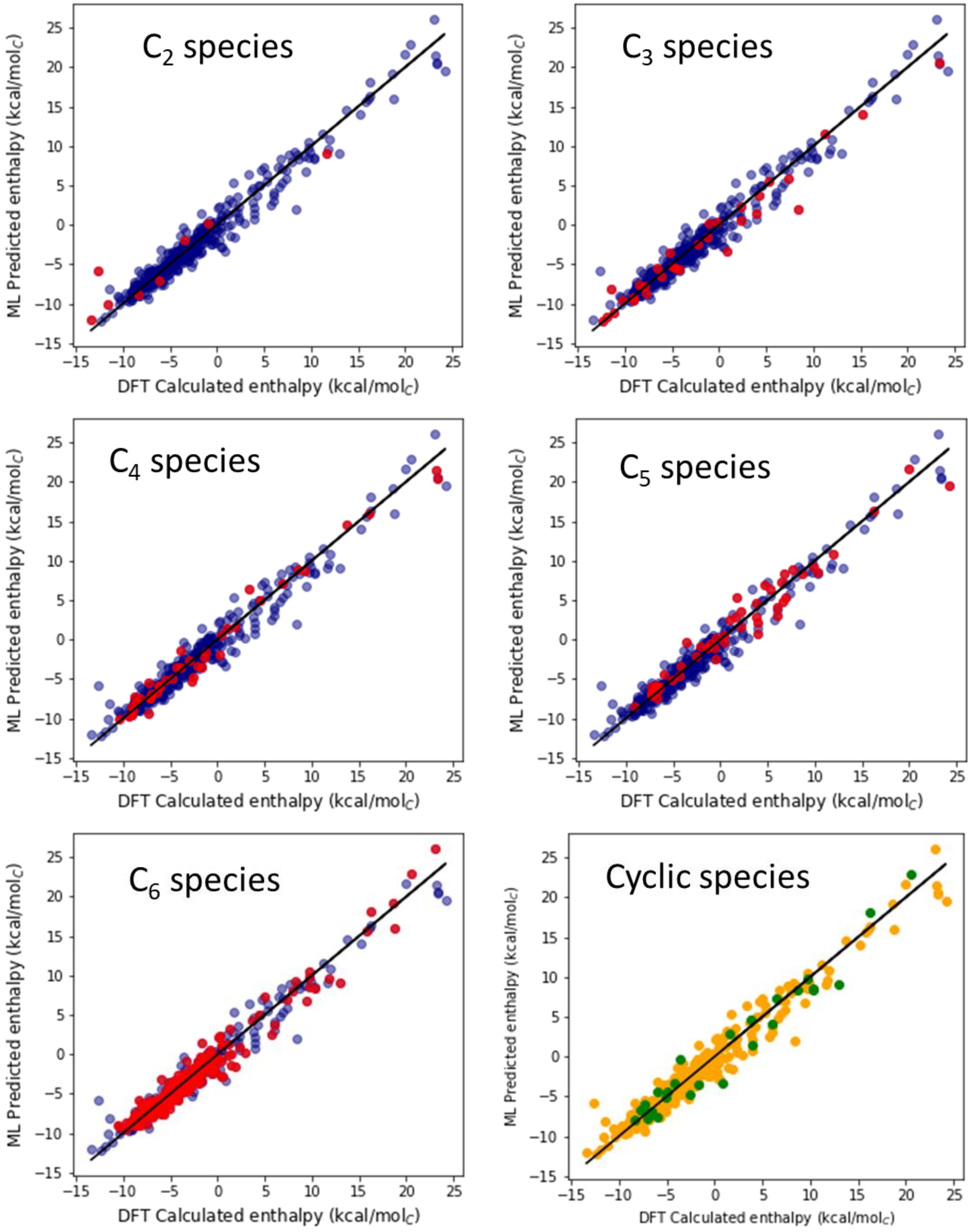


Figure S12. Parity plots of the ML predicted and DFT calculated enthalpy of formation per carbon for the ensemble average voting model of KRR+FMF and XGB+SVCF. In each of the C_n plots, C_n (n=2,3,4,5,6) species are displayed in red while all remainders, C_m $(m \neq n)$ species, are in navy. In the final plot, cyclic species are in green and acyclic species are in gold.

Reference

(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. https://doi.org/10.1021/ie502049a.

(43) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546–572. https://doi.org/10.1021/cr60259a002.

(44) Benson, S. W.; Golden, D. M.; Haugen, G. R.; Shaw, R.; Cruickshank, F. R.; Rodgers, A. S.; O'neal, H. E.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69*, 279–324. https://doi.org/10.1021/cr60259a002.

(45) Kua, J.; Faglioni, F.; Goddard, W. A. Thermochemistry for Hydrocarbon Intermediates Chemisorbed on Metal Surfaces: CH(n-m)(CH3)(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. https://doi.org/10.1021/ja993336l.

(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. https://doi.org/10.1021/acs.jpcc.6b06430.

(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. https://doi.org/10.1021/acs.jctc.9b00986.

(48) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry*; The Royal Society of Chemistry, 2013. https://doi.org/10.1039/9781849733069.