

**PREDICTIVE LARGE REACTION NETWORK MODELING VIA  
DATA-DRIVEN METHODS: APPLICATION TO BIOMASS CONVERSION**

by

Geun Ho Gu

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Spring 2018

© 2018 Geun Ho Gu  
All Rights Reserved

**PREDICTIVE LARGE REACTION NETWORK MODELING VIA  
DATA-DRIVEN METHODS: APPLICATION TO BIOMASS CONVERSION**

by

Geun Ho Gu

Approved: \_\_\_\_\_  
Eric M. Furst, Ph.D.  
Chair of the Department of Chemical and Biomolecular Engineering

Approved: \_\_\_\_\_  
Babatunde A. Ogunnaike, Ph.D.  
Dean of the College of Engineering

Approved: \_\_\_\_\_  
Ann L. Ardis, Ph.D.  
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Dionisios G. Vlachos, Ph.D.  
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Markos A. Katsoulakis, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Michael T. Klein, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Bingjun Xu, Ph.D.  
Member of dissertation committee

## ACKNOWLEDGMENTS

I would like to convey my utmost gratitude to my advisor, Prof. Dion Vlachos. His emphasis on leadership, communication, and professional development is seldom seen in academia and has prepared me to confidently face the next challenge. Beyond his education, he has been a father figure to me; his exemplary passion and persist to achieve goals has inspired me to be ambitious. I will forever be grateful that he kept his faith and continued his guidance during difficult times. I am also grateful that he let me freely pursue topics that interested me towards the end of my Ph.D.

I would like to thank multiple colleagues, without whom this thesis would not have been possible. I am most indebted to Dr. Shengguang Wang and Dr. Matthew A. Christiansen whom I have persistently bothered in my early years to learn gruesome density functional theory and microkinetic modeling. Despite my tenacity, they were always sincere and caring. I am sincerely grateful to Dr. Sai Konda and Dr. Glen Jenness to help me with scientific writing, which has been the most difficult aspect of the research for me. In addition, Dr. Glen Jenness has been a mentor to me, lighting the way through tough times as well as introducing me to proper programming. I also would like to thank Alexander Mironenko, Marcel P. Nunez, Dr. Ryan E. Patet, and Vassili Vorotnikov not only for their many useful discussions but also for their unwavering friendship. Especially, Marcel, my 5 year-long former roommate, has always inspired me with his discipline and the will to stay ahead for his Ph.D. and career. He may not realize, but he had a tremendous impact on my professionalism. In addition, I would like to thank other former and current Vlachos group members as

well as Professor Petr Plechac, Professor Dominique Guillot, and Professor Markos Katsoulakis for their useful discussions regarding mathematics and machine learning. I would like to thank my undergraduate researcher, Miriam Lee. Lastly, I would like to thank my committee members, Professor Bingjun Xu and Professor Michael T. Klein for helpful feedbacks.

I would like to extend my gratitude to my friends and their support over the past years. Being far away from home, my friends have been the second family to go back to during holidays. A group of six close friends from my high school, self-named sleepy pandas, has helped me to stay rooted and provided entertainments whenever I needed. My three former roommates from undergraduate have listened to my problems like their own problems without any therapy session fees. I am also grateful to my percussion band members in First State Symphonic Band for their companionship and for providing entrainment every Thursday night practice.

Finally, I would like to thank my family back in the home. My parents have provided me the abundant finance and arranged the best possible educational environment during my high school, undergraduate, and Ph.D. Only a handful young Koreans get opportunities to come abroad to receive the best educations. I have only received such opportunity, because of my amazing parents. Without a doubt, I would not have been half the man I am today without their dedication. I am looking forward to going back, and catch up with building memories that have been missing for past 13 years.

## TABLE OF CONTENTS

LIST OF TABLES .....	xi
LIST OF FIGURES .....	xiii
ABSTRACT .....	xxiii

### Chapter

1	INTRODUCTION .....	1
1.1	Conversion of Biomass to Fuels and Chemicals .....	1
1.2	Predictive Modeling via Multiscale Modeling and Descriptors .....	3
1.2.1	Multiscale Modeling .....	3
1.2.2	Descriptor-based Modeling .....	5
1.2.3	Prediction of Materials' Activity and Selectivity .....	9
1.3	Gaps and Challenges of the Current Paradigm of Predictive Modeling for Biomass Conversion .....	10
1.4	Dissertation Scope .....	14
2	MECHANISM OF DEHYDRATION OF PHENOLS ON NOBLE METALS VIA FIRST-PRINCIPLES MICROKINETIC MODELING .....	16
2.1	Introduction .....	16
2.2	Methods .....	19
2.2.1	Reaction Network .....	19
2.2.2	Density Functional Theory (DFT) Calculations .....	21
2.2.3	Brønsted-Evans-Polanyi (BEP) and Transition State Scaling (TSS) Relations .....	22
2.2.4	Parameterization of the Microkinetic Model .....	22
2.3	Brønsted-Evans-Polanyi (BEP) and Transition State Scaling (TSS) Relations .....	24
2.4	Energetics of the <i>p</i> -cresol HDO Reaction Network .....	28
2.5	Conversion and Selectivity to Major Gas Products .....	32
2.6	Conclusions .....	39
2.7	Acknowledgements .....	40

3	GROUP ADDITIVITY FOR THERMOCHEMICAL PROPERTY ESTIMATION OF LIGNIN MONOMERS ON PT(111).....	41
3.1	Introduction .....	41
3.2	Methods .....	42
3.2.1	Density Functional Theory (DFT) Calculations .....	42
3.2.2	Group Additivity .....	43
3.2.3	Training Set Sampling .....	44
3.3	Result and Discussion.....	46
3.3.1	Adsorption Conformations of Bio-oil Lignin Monomers on Pt(111) .....	46
3.3.2	Conjugation-Character Analysis .....	47
3.3.3	Group-Additivity-Model Description.....	53
3.3.4	Model Assessment.....	55
3.3.5	Strain Description.....	57
3.4	Conclusions .....	59
3.5	Acknowledgements .....	59
4	GROUP ADDITIVITY FOR AQUEOUS PHASE THERMOCHEMICAL PROPERTIES OF ALCOHOLS ON PT(111).....	61
4.1	Introduction .....	61
4.2	Methods .....	65
4.2.1	Density Functional Theory and Polarizable Continuum Model..	65
4.2.2	Thermodynamic Property Calculations of Molecules.....	66
4.2.3	Group Additivity .....	70
4.2.4	Training Set .....	73
4.2.5	Cavitation Energy Corrections .....	73
4.3	Result and Discussion.....	74
4.3.1	Influence of the Solvation on the DFT Geometries.....	75
4.3.2	Influence of the Solvation on Reactions.....	76
4.3.3	Group Additivity Schemes .....	79
4.3.4	Model Validation.....	82
4.4	Conclusions .....	85
4.5	Acknowledgements .....	86

5	THERMOCHEMISTRY OF GAS-PHASE AND SURFACE SPECIES VIA LASSO-ASSISTED SUBGRAPH SELECTION .....	88
5.1	Introduction .....	88
5.2	Methods .....	91
5.2.1	Data Sets and Graph Mining .....	91
5.2.2	Least Absolute Shrinkage and Selection Operator (LASSO).....	95
5.2.3	Semi-Supervised Learning .....	97
5.2.4	Cross Validation .....	103
5.3	Result and Discussion.....	104
5.3.1	Comparison of LASSO and Group-Additivity Models.....	104
5.3.2	Subgraph Search Space Reduction and Performance of Semi-Supervised Learning .....	108
5.3.3	Comparison to Machine Learning Models Using the QM9 Data Set .....	110
5.3.4	Application to Surface Adsorbates .....	112
5.4	Conclusions .....	119
5.5	Acknowledgements .....	120
6	PREDICTING MULTI-DENTATE ADSORBATE CONFORMATION ON METAL SURFACES .....	122
6.1	Introduction .....	122
6.2	Methods .....	123
6.2.1	Representation of Adsorbate Conformation and Stability.....	123
6.2.2	Conformation and Stability Sampling <i>via</i> DFT.....	128
6.2.3	Automatic Surface Adsorbate Conformation Generation Algorithm .....	129
6.2.3.1	Binding Rules .....	130
6.2.3.2	Efficiency of Enumeration Algorithm.....	134
6.2.3.3	Canonicalization and Ambiguity Removal of Conformations .....	135
6.2.4	Determining Conformation Stability using Machine Learning.	138
6.3	Results and Discussion .....	139
6.3.1	Rule Mining.....	139

6.3.2	Model Assessment.....	141
6.3.3	An Example.....	143
6.3.4	Additional Post Processing.....	144
6.4	Conclusions.....	144
6.5	Acknowledgements.....	145
7	CONCLUSIONS AND OUTLOOK.....	146
7.1	Dissertation Summary.....	146
7.2	Future Work.....	149
	REFERENCES.....	151
Appendix		
A	MECHANISM OF DEHYDRATION OF PHENOLS ON NOBLE METALS VIA FIRST-PRINCIPLES MICROKINETIC MODELING – SUPPORTING INFORMATION.....	184
A.1	Nomenclature of the p-cresol Derivatives.....	184
A.2	Generation of the Reaction Network.....	184
A.3	Parameterization of the Microkinetic Model.....	186
A.3.1	Vibrational Frequency Calculations and Statistical Mechanical Approximation.....	190
A.3.2	Lateral Interactions.....	197
A.4	Adsorption and Transition State Structures.....	199
A.5	MKM Parameters Adjustments.....	203
A.6	Reaction Energies.....	204
A.7	MKM Results.....	205
B	GROUP ADDITIVITY FOR THERMOCHEMICAL PROPERTY ESTIMATION OF LIGNIN MONOMERS ON PT(111) – SUPPORTING INFORMATION.....	206
B.1	Adsorption Conformation of Lignin Monomer Ring.....	206
B.2	Schematics of Binding Type Couples.....	211
B.3	Charge Flow Analysis.....	213
B.4	Methoxy Strain Correction.....	213

C	GROUP ADDITIVITY FOR AQUEOUS PHASE THERMOCHEMICAL PROPERTIES OF ALCOHOLS ON Pt(111) – SUPPORTING INFORMATION .....	215
	C.1 Temperature Independence of Gibbs Free Energy of Solvation Assumptions .....	215
	C.2 PCM Method Validation .....	216
	C.3 Group Additivity Method Validation .....	218
	C.4 Cavitation Energy Correction .....	218
	C.5 Ethanol Reforming Reaction Network Detail .....	224
D	THERMOCHEMISTRY OF GAS-PHASE AND SURFACE SPECIES VIA LASSO-ASSISTED SUBGRAPH SELECTION – SUPPORTING INFORMATION .....	230
	D.1 Supporting Figures .....	230
E	MICROKINETIC MODELING OF AQUEOUS PHASE ETHYLENE GLYCOL REFORMING .....	235
	E.1 Introduction .....	235
	E.2 Method.....	235
	E.2.1 Reaction Network Enumeration .....	235
	E.2.2 Microkinetic Model Parameterization.....	236
	E.3 Model Validation and Mechanistic Insights.....	237
F	PERMISSION FOR REPRINT .....	242

## LIST OF TABLES

Table 2.1. Slopes and intercepts with 95% confidence intervals for BEP and TSS relations (uncorrected for zero-point energy).....	28
Table 2.2. Mean and max absolute errors (AE) of reaction barriers computed using each correlation. ....	28
Table 3.1. BCP analysis between two carbons in gas phase and on the surface. C2 and C3 molecules are used where valence is capped with hydrogen. Subscription of C indicates the binding-site type. Structures are shown in Figure B.6 in the appendix. ....	50
Table 3.2. C-C BCP dependence on the binding of the nearest neighboring carbon. (C) represents a -CH <sub>3</sub> group, and (C <sub>σ</sub> ) represents a -CH <sub>2</sub> group adsorbed on a σ site. Structures are shown in Figure B.7. ....	53
Table 3.3. Level of site description in prior and new group additivity schemes.....	54
Table 4.1. List of group centers and peripherals that constitute groups (See the body for group description), corrections and unique groups.....	82
Table 4.2. Mean and Maximum (Max) Absolute Errors (AEs) and Root Mean Square Error (RMSE) for 10-Fold Cross-Validation of Thermodynamic Properties in kcal/mol .....	84
Table 5.1. Comparison between the performance of the previous group additivity models and the graph selection method.[16, 21, 23].....	118
Table 6.1. Binding Rule Information. ....	140
Table 6.2. Error matrix of the logistic model. ....	142
Table A.1. Reaction rules and examples of input into RING.....	185
Table A.2. Number of reactions and intermediates in the HDO reaction network of <i>p</i> -cresol on Pt(111). The table on the left lists the intermediates in each functional form similar to Figure 2.1. ....	186

Table A.3. C-H and O-H scission reactions used for development of BEP/TSS relations. ....	188
Table A.4. C-OH scission reactions used for development of BEP/TSS relations. ...	190
Table A.5. Binding energies and vibrational frequencies of species computed via DFT. ....	191
Table A.6. Binding energies of surface species whose thermodynamics is approximated using the equations above.....	196
Table A.7. List of parameter adjustments for the MKM model. ....	203
Table C.1. Regression Coefficient.....	218
Table C.2. Cavitation energy computed using DFT and the model in kcal/mol. ....	219
Table C.3. The $\Delta G_{r,(vac)}$ and $\Lambda^{\circ}_{sol}$ of the 151 reactions examined in ethanol reforming reaction network are shown, computed using the DFT and group additivity. Here, the group additivity $\Delta G_{r,(vac)}$ and $\Lambda^{\circ}_{sol}$ are computed from cross-validation computed the Gibbs free energy of formation and the Gibbs free energy of solvation of each molecule. The molecule names are in-line with those provided in the XLSX file in ref.[216].....	225
Table E.1. Comparison between experimental[359] and model reaction orders of ethylene glycol and hydrogen. For ethylene glycol reaction order, conditions are 1, 4, and 10 wt% ethylene glycol feed at 225 °C, 29.3 bar ( $H_2$ ), 9.36 mol ethylene glycol $\cdot$ mol <sup>-1</sup> $\cdot$ site h <sup>-1</sup> . For hydrogen reaction order, conditions are 10 wt% ethylene glycol feed at 225 °C, 9.36 mol ethylene glycol $\cdot$ mol <sup>-1</sup> $\cdot$ site h <sup>-1</sup> with 30.3, 32.3, 39.3 bar of $H_2$ .....	238

## LIST OF FIGURES

- Figure 1.1. Schematic of multi-scale modeling showing different scales and models at each scale as well as couplings between scales. The computational fluid dynamics (CFD) illustration is adopted from Ref.[15]..... 4
- Figure 1.2. (a) Visualized reaction network of glycerol reforming and (b) exponentially scaling of the number of intermediates and reactions with a number of carbon in linear polyols. Adopted from Sutton et al.[18] The nodes and edges in (a) represent intermediates and reactions, respectively. .... 6
- Figure 1.3. Group additivity examples for (a) propane and (b) di-methylene ether on Pt(111). In this method, patterns in the molecule are identified (4 color-coded groups and their heat of formation values are shown), and their thermochemistry values are summed together.  $\Delta_f H^\circ_{GA}$ ,  $\Delta_f H^\circ_{DFT}$ ,  $\Delta_f H^\circ_{exp}$  are the heat of formation calculated from group additivity, DFT, and experiments, respectively..... 8
- Figure 1.4. Example of predictive catalyst mapping adopted from [33]. This map shows how activity and selectivity of ethylene glycol conversion changes with oxygen and carbon binding energy as key descriptors. The heat color indicates the turn-over-frequency, and the each divided region indicates favored selectivity shown in the white circle. Different catalysts are indicated with circles. Ni-Pt is superior to Pt for reforming and WC and Mo<sub>2</sub>C are selective but not as active catalysts for deoxygenation. Ethane is produced by in situ reforming to H<sub>2</sub>, followed by HDO. .... 10
- Figure 1.5. Overview of challenges and gaps in predictive modeling of biomass conversion on solid catalysts. .... 11
- Figure 2.1. Proposed HDO mechanisms of phenolic compounds in the literature. .... 18
- Figure 2.2. Summary of the overall p-cresol HDO reaction network. Ring carbon position is numbered as shown on p-cresol. Throughout the study, we will identify the original location of hydroxyl group as the  $\alpha$ -carbon. Horizontal arrows represent ring (de)hydrogenation and vertical arrows represent bond scission/formation indicated in color. .... 20

- Figure 2.3. (a) BEP relations for the C-OH scission reactions of three homologous series. Raw data is shown in Table A.4 and BEP relations in Table 2.1. (b) Corresponding TSS based on the FS energy. (c) Structural characteristics of each BEP relation with examples (bond lengths in Å). Identical color codes are maintained in the plots and the figures. The Table (panel a) summarizes the types of TS structures. Barriers and reaction energies are reported as electronic energies with no zero-point energy corrections for these fast screening methods..... 27
- Figure 2.4. Minimum energy path of ring hydrogenation of p-cresol to methyl cyclohexanol along with C-H, O-H, and C-OH reactions. For each reaction, upper number is reaction barrier and lower number is reaction energy in eV. Reaction energies (electronic energies without zero point correction) are calculated using DFT. Reaction barriers,  $E_a$ , are calculated using DFT (black) or BEP relations developed herein (cyan). Color convention of arrows as in Scheme 1..... 29
- Figure 2.5. C-OH scission barriers vs. degree of ring hydrogenation. Typical hydrogenation barriers fall within the green shaded region. Open (closed) symbols are BEP (DFT) computed barriers. Filled green triangles correspond to C-OH scission of methyl cyclohexanol ..... 31
- Figure 2.6. Alcohol intermediates along the minimum energy path demonstrating lifting as the ring carbons are hydrogenated (the number indicates the number of hydrogenated carbons). During hydrogenation the  $sp^3$  hybridized ring carbons give place to a chair shape structure with less strain. .... 32
- Figure 2.7. Comparison of model (lines) with (a) Nie et al.'s[62] and (b) Foster et al.'s experimental data[58] (solid symbols). Experimental conditions were (a)  $1.25 \times 10^{-2}$  atm *m*-cresol, 1.0 atm  $H_2$  at 573 K on 1.0 wt% Pt/SiO<sub>2</sub>, and (b)  $7.5 \times 10^{-3}$  atm *m*-cresol, 0.5 atm  $H_2$ , 0.5 atm  $N_2$ , at 533 K on 1.55 wt% Pt/SiO<sub>2</sub>..... 33
- Figure 2.8. Comparison of model (lines) and experimental data[62] (solid symbols). Experimental conditions are  $1.25 \times 10^{-2}$  atm (a) methyl cyclohexanone and (b) methyl cyclohexanol, 1.0 atm  $H_2$  at 573 K on 1.0 wt% Pt/SiO<sub>2</sub>. The model agrees well with the experimental data. .... 34
- Figure 2.9. Reaction network for HDO of *p*-cresol to toluene at short contact times of  $W/F=2$  gh/mol for Nie et al.'s[62] experimental condition. Double (single) arrows represent reversible (irreversible) reactions. Thickness of the arrows indicates the magnitude of the net flux. The network is based on fluxes at the exit of the reactor. .... 35

Figure 3.1. Local binding sites observed in lignin monomers on Pt(111). M represents metal atom. Top (bottom) row shows the side (top) view of the binding site. ....	47
Figure 3.2. Parity plots for group-additivity computed versus DFT-estimated $\Delta H_{f,298}$ for (a) regression and (b) validation. AE stands for absolute error. ....	56
Figure 3.3. Deviation of bond angles in each group from their average. The more narrow the distribution, the more alike the angles are for a group among all species in the training set. ....	58
Figure 4.1. Various considered states for the system, summarized in a thermodynamic cycle. $i$ is an arbitrary species, and $i'$ is the species $i$ adsorbed on the surface. Boxes represents a molecule in the gaseous state, in the bulk liquid, adsorbed on Pt(111), and adsorbed and solvated on Pt(111) going from left to right and from top to bottom. The three thermodynamic properties in each box represent the parameters predicted by our method. We assume that the solvation effect is temperature independent, and thus the entropy and heat capacity are the same in vacuum and in water. ....	70
Figure 4.2. DFT structures of the HCOH fragment adsorbed on a Pt(111) slab: (left) the H down configuration (right) the H up configuration. Associated solvation energies are also given. ....	75
Figure 4.3. Box plot for the change of (a) the Gibbs energy of a reaction upon solvation $A_{sol}$ (see text for definition) and (b) the solvation Gibbs energy $\Delta G_{sol}$ of each species using DFT in the ethanol reforming reaction network (without water-gas shift reactions and species). Black dots and white dots represent mean and outliers, respectively. The box represents the second and third quartiles while the whisker represents the first and fourth quartiles. Outliers are at least 1.5 times the box range (between the second and third quartile) from the edge of the box. ....	79
Figure 4.4. Parity plot between 10-fold cross-validation computed prediction and DFT data for (a) Gibbs free energy of formation, $\Delta G_{f,(gp)}$ , (b) Gibbs free energy of solvation, $\Delta G_{sol}$ , (c) Gibbs free energy of reaction, $\Delta G_{r,(vac)}$ , and (d) the change of reaction energies upon solvation, $A_{sol}$ , as defined in equation (4.14) and (4.15). 151 reactions of the ethanol reaction network are considered (see the Appendix for the details). ....	83

- Figure 5.1. Examples of graph descriptors used in estimating thermochemistry: (a) gaseous propane[162, 260] and (b) benzene on Pt(111).[21] The left/top legend indicates the element represented by each vertex.  $\Delta_f H^\circ_i$  and  $\Delta_f H^\circ_{GA,i}$  indicate standard enthalpy of formation, either experimental (a) or the DFT computed value (b) and group additivity computed standard enthalpy of formation for datum  $i$ , respectively.  $\beta_t$  is the contribution of graph  $t$  to the  $\Delta_f H^\circ_{GA}$ ,  $x_{it}$  is the number of appearances of graph descriptor  $t$  in the molecular graph of datum  $i$ , and  $\Delta_f H^\circ_{GA,i} = \sum_t x_{it} \beta_t$ ..... 90
- Figure 5.2. (a) Flow chart of the adsorbate graph enumeration algorithm and (b) example of the algorithm using benzene on Pt(111)[21] (only subgraphs with a number of vertices  $\leq 3$  are shown). “1” and “2” in arrows in (b) represent the steps in the flow chart of (a)..... 95
- Figure 5.3. Workflow of the graph selection method using LASSO. .... 97
- Figure 5.4. Number of unique subgraphs and descriptor vector memory vs. number of secondary carbons in alkanes. The solid and dashed lines indicate explicitly computed and extrapolated (estimated) data, respectively. The Spyder python software[302] is used for memory estimation. .... 98
- Figure 5.5. Illustration of the difference between LASSO and LAR-LASSO. Here,  $\times$  indicates the addition or removal of a descriptor. LAR-LASSO moves along the path of  $\lambda$  and tracks the additions or removals, whereas LASSO computes the regression coefficients at a single value of  $\lambda$ . The dashed line shows an example of the single LASSO solution obtained at  $\lambda = 10$ . .... 99
- Figure 5.6. Example of gSpan’s meta-graph enumeration (up to adding 3 edges). The algorithm employs mathematical techniques to avoid generating duplicate nodes. .... 101
- Figure 5.7. Parity plot between experimental  $\Delta_f H^\circ$  and cross-validation predicted  $\Delta_f H^\circ$  computed using (a) the Benson’s gas group additivity,[306] and various LASSO-trained models (b)-(d): (b)  $\lambda$  is chosen using the Morgan fingerprints[307] so the cross validation mean square error is minimal, (c)  $\lambda$  is chosen so the number of descriptors,  $P$ , is similar to that of group additivity, and (d)  $\lambda$  is chosen so the cross validation mean square error is minimal. The Morgan fingerprints and the exhaustively enumerated subgraph space consist of 835 and 16513 descriptors, respectively. CV RMSE is the cross validation root mean square error, and MeanAE and MaxAE are the mean and max absolute error, respectively..... 105

Figure 5.8. Graph descriptors (a) used in group additivity and (b) selected by LASSO for $P = 60$ . Only graphs related to alkanes are shown. R indicates any atom. The value below a graph indicates the graph's contribution to $\Delta_f H^\circ$ in kcal/mol. ....	108
Figure 5.9. (a) Computation time and (b) number of enumerated graph descriptors vs. maximum allowed edges in subgraphs for the exhaustive enumeration method and the semi-supervised learning using the dataset of 695 hydrocarbon molecules. The dashed line indicates the extrapolation of the exhaustive enumeration method. The semi-supervised learning is performed until $\lambda = 0.1$ . The exhaustive enumeration was performed using the Rdkit[292] C++ package (python API), and the semi-supervised learning performed in C++. ....	109
Figure 5.10. Parity plot between experimental data and prediction of (a) Benson's group additivity and (b) the semi-supervised model at optimal $\lambda$ for the 695 hydrocarbon data set. ....	110
Figure 5.11. Parity plot between DFT data and predicted data using the subgraph selection method for 10 % of the QM9 data set. ....	112
Figure 5.12. Parity plot between DFT-computed[21] and predicted $\Delta_f H^\circ$ for the lignin monomer dataset. ....	113
Figure 5.13. Molecular graphs and subgraphs of benzene and its hydrogenated derivatives. The arrows indicate the subgraphs encountered in each molecule. ....	114
Figure 5.14. Principal coordinates analysis plot projected to 2D using the distance measurement shown in equation (5.9) for (a) the lignin, (b) the furanics, and (c) the linear oxygenates adsorbate datasets. The color scheme represents the absolute error in $\Delta_f H^\circ$ (kcal/mol) where data points with an absolute error higher than 20 kcal/mol are the same color as 20 kcal/mol error. ....	119
Figure 5.15. Correlation between absolute errors and the sum of the distance of each data for (a) lignin, (b) furanics, and (c) linear oxygenates adsorbate datasets. The dashed line indicates the threshold values used to filter outliers. ....	119
Figure 6.1. Carbon in $\text{CH}_3$ , $\text{CH}_2$ and $\text{CH}$ with varying number of valence electrons on atop, bridge and hollow sites. The superscript indicates the binding atom's valency and the subscript the type of binding site. S (U) indicate a stable (unstable) conformation. ....	125

- Figure 6.2. Examples of the effect of first-nearest neighbors on stability. Atoms with orange arrows are the central carbons. S (U) indicates a stable (unstable) conformation. .... 127
- Figure 6.3. Overview of the conformation enumeration algorithm. The algorithm involves three steps: (1) exhaustive enumeration of conformations using mined transformation rules; (2) mining conformations' subgraphs and their frequencies; and (3) prediction of stability using machine learning. Black, red, and blue edges indicate bonds between two adsorbate atoms, between an adsorbate and a metal atom, and between two metal atoms, respectively. M indicates a surface metal atom. 1 and 0 in the stability column indicates stable and unstable conformation, respectively. .... 130
- Figure 6.4. An example of conformation generation of  $\text{CH}_2\text{CHCH}_2$  using the rules in Figure 6.5. FR is the first atom binding rule in Figure 6.5a, NR1 and NR2 are the neighbor atom binding rules in Figure 6.5c and e. The second and third column represent intermediate structures. Only one type of site (atop site) is shown here for simplicity. .... 131
- Figure 6.5. Graph theory representation of binding rules. (a) The first atom binding rule used in Figure 6.4 for generation of a conformation; (b) An example of the first atom binding rule in (a); (c) The neighbor atom binding rule; (d) An example of the neighbor atom binding rule in (c); (e) Another neighbor atom binding rule; (f) An example of the neighbor atom binding rule in (e). In  $\text{C}^\alpha_\beta$  notation,  $\alpha$  is the valency of a carbon atom, and  $\beta = \text{A, B, H or u}$  is atop, bridge, hollow or unassigned binding site of the carbon atom. The index in a circle for (a), (b), (c), (d), (e) and (f) shows how atoms are mapped to each other. .... 133
- Figure 6.6. Three lattice subgraphs possible for a metal-metal-metal lattice subgraph of a conformation. Algorithm 6.2 systematically and efficiently adds unoccupied metal atoms to represent a conformation canonically and unambiguously. .... 136
- Figure 6.7. Parity plot of the logistic model. Relative frequency is plotted for better visualization. .... 143
- Figure 6.8. Conformation enumeration of  $\text{CH}_2=\text{C}(\text{CH}_2)_2$ . The results agree with the DFT calculations on Pd(111). The numbers indicate the stability computed by the logistic regression. .... 144

Figure 7.1. Example of (a) unstable C-OH scission transition state structure on the terrace (DFT calculations in Chapter2), and (b) expected C-OH scission transition state structure on step (based on heuristics). .....	150
Figure A.1. Numbering of the ring carbon atoms. Structures from left to right: PC(OH), PC(O), PC, and TOL. An intermediate X hydrogenated at ring positions 2, 4, and 6 is denoted as H246X. For example, H23456PC(OH) refers to <i>p</i> -cresol with hydrogenated ring carbon at positions 2, 3, 4, 5, and 6.....	184
Figure A.2. Reaction energy profile (top) of hydrogenation of <i>p</i> -cresol (PC(OH)) following the minimum energy path (red arrows) and corresponding species (bottom) as hydrogens are being added. ....	187
Figure A.3. Structures of hydrogenated ring-carbons that lead to H-deficient derivatives with sp <sup>3</sup> hybridized $\alpha$ -carbon. ....	189
Figure A.4. BEP relations for C-H and O-H scission reactions. The training set is presented in Table A.3.....	189
Figure A.5. Lateral interaction for the effect of hydrogen coverage on the binding energy of the <i>p</i> -cresol derivatives vs. the number of Pt atoms involved in adsorption. ....	198
Figure A.6. Transition state structures of C-H bond scission reactions. ....	200
Figure A.7. Transition state structures of O-H bond scission reactions. ....	200
Figure A.8. Transition state structures of C-OH bond scission reactions. ....	201
Figure A.9. H-OH scission transition state structure.....	201
Figure A.11. Adsorption structures of the <i>p</i> -cresol derivatives with H deficient oxygen. ....	202
Figure A.12. Adsorption structures of the <i>p</i> -cresol derivatives with radical $\alpha$ -carbon. ....	202
Figure A.13. Adsorption structures for the toluene derivatives. ....	203
Figure A.14. (a) C-OH scission reaction energy vs. degree of ring hydrogenation for the alcohols in the entire reaction network. (b) Number of C-OH scission reactions encountered vs. degree of ring hydrogenation. The combination of these two factors leads to efficient C-OH scission upon sufficient ring hydrogenation. ....	204

Figure A.15. Surface coverages of <i>p</i> -cresol simulation for Nie et al. experimental data[62] shown in Figure 2.5 in the main text. Hydrogen is the most dominant species on the surface, followed by toluene derivatives with 3 hydrogenated ring carbons (TOL+3H), <i>p</i> -cresol with 3 hydrogenated ring carbons (PC(OH)+3H), and <i>p</i> -cresol with H deficient oxygen and 2 hydrogenated ring carbons (PC(O)+2H). .....	205
Figure A.16. Molar equilibrium composition with (a) and without (b) hydrocarbon products accounted for. Panel (b) is reminiscent of near the entrance region of the reactor. Conditions are $1.25 \times 10^{-2}$ atm <i>p</i> -cresol, 1.0 atm H <sub>2</sub> . .....	205
Figure B.1. Adsorption conformations of benzene ring and its hydrogenated derivatives. Legend at the top of the figure shows the binding type given the structure. The name convention follows the previous benzene adsorption study by Morin et al.[348] .....	207
Figure B.2. Adsorption conformations of benzene ring with one H-deficient ring carbon and its hydrogenated derivatives. Blue circles indicates Pt atoms. ....	208
Figure B.3. Adsorption conformations of benzene ring with one surface interacting oxygen substituent and its hydrogenated derivatives. ....	209
Figure B.4. Adsorption conformations of benzene ring with one H-deficient ring carbon and one surface interacting oxygen substituent and their hydrogenated derivatives. ....	209
Figure B.5. Adsorption conformations of benzene ring with two surface interacting oxygen substituents and its hydrogenated derivatives. ....	210
Figure B.6. Schematics of binding type couples listed in Table 1 along with expected <i>p</i> -orbital lobes in the same phase.[45, 151, 152, 349, 350] BCPs are shown in pink balls and Pt atoms in blue circles. ....	211
Figure B.7. Schematics of structures listed in Table 2 along with expected <i>p</i> -orbital lobes in the same phase. ....	212
Figure B.8. Electron density difference upon $\pi$ -mode adsorption of a C <sub>2</sub> H <sub>2</sub> fragment. Charge occupation of the conjugated $\pi$ electron system depends on binding of substituent groups. ....	213

Figure B.9. (a) An example of a strained molecule with a –OCH group. (b) An example of a strained molecule with a –OCH <sub>2</sub> group. (c) Schematics of strained fragments. ....	214
Figure C.1. Parity graph of vibrational energy between vacuum and PCM calculations. Total of 26 molecules are tested. Mean absolute deviation is 0.10 kcal/mol and max absolute deviation is 0.82 kcal/mol. ....	216
Figure C.2. Parity plot between the model computed solvation free energy and the experimental solvation free energy. Three models are shown: (1) PCM computed at 0 K without cavitation energy at dilute condition, (2) PCM computed at 0 K with cavitation energy at dilute condition, and (3) COSMO-SAC[355] computed at 298 K at 1 M. Experimental data[356] are obtained at 298 K at 1M condition.....	217
Figure C.3. Figure 4.3 is produced using group additivity method. Group additivity method reproduces the general trends seen by DFT well. ....	218
Figure D.1. Illustration of L1 norm penalty constraining least square loss function. The red surface indicates the quadratic loss where the outer ellipse indicates a higher loss. The blue surface indicates the L1 norm penalty where the outer diamond corresponds to a higher penalty. In this illustration, $\beta = \{\beta_1, \beta_2\}$ converges to the center of the red surface without the L1 norm penalty, whereas, $\beta$ converges to a solution with $\beta_1 = 0$ for the L1 norm penalty.....	230
Figure D.2. Illustration of LAR-LASSO tracking predicted response vector, $X\beta$ , along $\lambda$ . $X\beta$ is projected to 2 dimensions from the dimension of $N$ , the number of data points. $\mathbf{0} \in \mathbb{R}^P$ is the zero vector. LAR-LASSO starts at $X\beta = \mathbf{0}$ with initial descriptor selected, then $\beta$ is moved along the direction of $\mathbf{y}$ until the addition or removal of a descriptor occurs. The direction of $X\beta$ changes at each addition or removal event. Eventually, LAR-LASSO terminates when the number of parameter, $P$ , is equal to $N$ when the selected descriptor space has enough degrees of freedom to completely describe the data.....	231
Figure D.3. The learning curve of Benson’s group additivity model applied to a (a) 217 and (b) 695 gas species data set. Leave-one-out cross-validation was used with 100 training-set data shuffling for (a) and 10 for (b).....	231
Figure D.4. Parity plot between DFT-computed and predicted $\Delta_f H^\circ$ for (a) furanics[16] and (b) linear oxygenates[23]. ....	232

Figure D.5. Outliers in furanics ((a) and (c)) and linear oxygenates ((b) and (d)) data set. Adsorbates in (a) and (b) are mostly unsaturated molecules expected to be outliers in chemical space, and those in (b) and (d) are outliers in the sampled data, indicating insufficient sampling. ....	233
Figure D.6. Learning curve of the model selected by LASSO using the entire data set for (a) the furanics, (b) the linear oxygenates, and (c) the lignin data sets. The curves are the average of 100 training set shuffles for (a) and (b) and of 10 shuffles for (c). ....	234
Figure E.1. Comparison between experimental (filled bars) and model (empty bars) turn-over-frequency (TOF) for CO <sub>2</sub> and H <sub>2</sub> over Pt/Al <sub>2</sub> O <sub>3</sub> . Four experiments from [358] are considered. The conditions are: temperature (K), pressure (bar, H <sub>2</sub> ), and liquid hourly space velocity (h <sup>-1</sup> ):(1) 423, 4.8, 1.9 (2) 453, 10.0, 3.8 (3) 483, 19.6, 42.9 and (4) 498, 25.8, 128.6. Calculations done at 1-3% conversion (surface to volume ratio of ~5 cm <sup>-1</sup> ) and experiments at 1-10% conversion. ....	238
Figure E.2. Major reaction pathway leading to CO in (a) vapor phase (210 °C, 1 bar, 0.20 gas mol flow rate/Pt site [s <sup>-1</sup> ], 25% ethylene glycol and 75% steam; ethylene glycol steam reforming conditions are comparable to those reported in ref[127]) and (b) aqueous phase (150 °C, 4.8 bar, 1.9 liquid hourly space velocity [h <sup>-1</sup> ], 5 wt% ethylene glycol). CO is further converted to CO <sub>2</sub> and H <sub>2</sub> O via the water-gas-shift reaction. ....	240
Figure E.3. Gas composition (mol%) vs. temperature for (a) aqueous phase and (b) vapor phase. Conditions are the same as in Figure E.2. ....	241
Figure E.4. Surface coverage of major species vs. temperature for (a) aqueous phase and (b) vapor phase. Conditions are the same as in Figure E.2. .	241

## ABSTRACT

US energy demand is projected to continue increasing, and exploiting sustainable resources is critical to minimizing risks in economy, geopolitics, and environment. Lignocellulosic biomass is a key sustainable source of carbon for the production of renewable fuels and chemicals. Compared to petroleum processes, biomass utilization requires selective de-functionalization, calling for innovations in catalysis. Previously, metal surface catalysts have shown promising performance for hydrogen generation and hydrodeoxygenation (HDO) in biomass conversion processes. However, mechanistic understandings are lacking and the search for optimal catalysts continues. In this regard, density functional theory (DFT) studies can aid, but DFT is computationally too expensive to investigate large reaction networks of biomass monomers on metal surfaces. Previous research introduced a semi-empirical based framework to reduce the computational cost and rapidly build microkinetic models (MKMs), but several tasks remain to realize the framework. This thesis aims to narrow the gaps in this framework.

One gap is modeling capability for lignin monomer derivatives whose theoretical investigation has been minimal prior to this thesis. This thesis has two overarching goals. First, we build reliable semi-empirical methods mainly for thermochemistry and to an extent for kinetics. Second, we apply these methods to the HDO mechanism of a lignin model compound, cresol, on a metal surface for the first time.

Three chapters are dedicated to improving and expanding the capability of thermochemistry prediction for molecules on metal surfaces. With the reaction network identified, group additivity for lignin monomers is developed. The new group additivity introduces more sophisticated descriptors based on the theoretical basis of the group additivity and electron density analysis. The cross-validation shows 2.8 kcal/mol mean absolute error (MeanAE) with 591 data points, an improvement over the previous framework by 2.3 kcal/mol in the MeanAE. The model is capable of rapidly predicting ~14,000 molecules (accounting different binding geometries) in the reaction network of a lignin model compound, guaiacol. Next, group additivity for solvated molecules on a metal surface is developed that provides an excellent first approximation of solvation energy at MeanAE of 1.0 kcal/mol at no computational cost. To further improve the thermochemistry prediction framework, a machine learning method called LASSO was introduced. LASSO is used to automatically select the most informative patterns (descriptors) of the molecules in datasets. The application to the 591 lignin dataset reveals a MeanAE of 2.08 kcal/mol, achieving the sub 0.1 eV error for the first time for an adsorbate group additivity method. In order to simplify the user interface for these thermochemistry prediction methods as well as to automatically build datasets, a machine-learning algorithm for predicting adsorption geometries is presented. Methods are made available on Github.

Next application of these methods on MKM is carried out. Our DFT-refined MKM of cresol on Pt(111) demonstrates a novel mechanism consistent with multiple experiments, suggesting that Pt alone is capable deoxygenation. The MKM and DFT results suggest that cresol undergoes deoxygenation upon sufficient ring hydrogenation. Innovations in the semi-empirical model frameworks are made to build

the MKM. In order to rapidly estimate activation barriers of ~500 reactions, Brønsted-Evans-Polanyi (BEP) relations were developed. The previous approach for building BEP relations results in a significant MeanAE of 17.3 kcal/mol for the rate-limiting C-OH scission reactions. A new framework introduced herein categorizes transition states based on their structure and reduces the MeanAE to 1.6 kcal/mol. In addition, the computational cost in building a model for lateral interactions has been reduced by coarse-graining surface species (from 150 to 7 species).

## Chapter 1

### INTRODUCTION

#### 1.1 Conversion of Biomass to Fuels and Chemicals

As the US energy demand is projected to grow by 50 - 60 percent by 2030, relying on conventional fossil fuels brings economic, geopolitical, and environmental uncertainties.[1] These risks can be mitigated by diversifying resources for chemicals and energy production. Potential candidates include natural gas, solar, biomass, carbon dioxide, wind, geothermal and nuclear resources. Out of these, biomass is the only sustainable source of carbon. In this regard, the United States Department of Energy (DOE) has set a goal of displacing 30% of gasoline and diesel,[2, 3] and 25% of all petroleum-derived chemicals from biomass by 2022.[4] Biomass utilization has potential for production of many chemicals due to its diverse functional groups.[5]

Compared to petroleum, biomass is highly polymerized and over-functionalized.[6] These characteristics present new challenges. As the current petrochemical industry has focused on functionalization of hydrocarbons, innovations in selective deoxygenation are necessary to remove undesired functional groups.[7] In addition, while various compounds in petroleum can often be separated using mature distillation technology, lignocellulosic biomass is a solid composed of carbohydrate and aromatic oxygenate polymers. Inherently, the biomass utilization requires depolymerization to break down the polymer linkages, resulting in high cost.[8] Lastly, the high oxygen content of biomass causes depolymerized biomass to be of low energy density, corrosive, and prone to re-polymerization to undesired products.[6, 9]

Hence, innovations in deoxygenation are critical to render biomass useful. Catalysis can play a key role in realizing the economic conversion of biomass.[10]

A number of catalytic routes have been demonstrated for biomass conversion.[5] One of the most researched pathways is the production of 5-hydroxymethylfurfural (HMF) from cellulose and furfural from hemicellulose via a three-step acid-catalyzed hydrolysis, isomerization, and dehydration process.[7] HMF is readily accessible from biomass and one of the DOE's top 10 bio-based platform chemicals.[5]

Fast pyrolysis is an alternative process to thermally decomposed lignocellulose at temperatures between 400 – 650 °C at short retention times.[6] Fast pyrolysis of biomass forms a liquid mixture of various linear, branched, furanic, and aromatic oxygenates monomers called bio-oil. Even though the composition of bio-oil is complex, bio-oil has great potential as it contains naturally occurring C<sub>6</sub> aromatics to form key chemicals: benzene, toluene, and xylene (BTX). BTX represents ~5% of all the chemicals produced by volume and ~25% of the total revenue in the petrochemical industry.[11, 12]

The products either of acid-catalyzed hydrolysis or pyrolysis need to be further upgraded by removing oxygen. In this respect, hydrodeoxygenation (HDO) has widely been explored where oxygen is selectively removed as water using hydrogen gas over heterogeneous catalysts. Metal catalysts have been investigated for HDO as hydrogen gas readily adsorbs, dissociates and reacts on metal surfaces. Depolymerized monomers can also be utilized in aqueous phase reforming, popularized by Dumesic and co-workers.[13] In this process, hydrogen is produced along with CO<sub>2</sub> and alkanes in the aqueous phase using solid metal catalysts. This thesis focusses on computational

work on metal catalysts for these downstream biomass conversion processes. The next section elaborates on the framework of computational catalysis.

## **1.2 Predictive Modeling via Multiscale Modeling and Descriptors**

### **1.2.1 Multiscale Modeling**

Multiscale modeling simulates a system consisting of phenomena occurring at multiple time and length scales. Simulation of each scale involves a different model. The multiscale approach address this scale disparity by coupling adjacent scales via either coarse-graining of small-scale models (bottom-up approach) or linking models via feedback information among different scale models (top-down approach). In computational catalysis, the bottom-up approach is dominant and is the focus of this thesis.

There are at least four different time and length scales in reaction engineering as illustrated in Figure 1.1. The smallest scale (the bottom of Figure 1.1) involves quantum phenomena where the interactions between electrons and protons are simulated. In this simulation, energies of various states and atomic coordinates can be computed to obtain binding energies, reaction barriers, and vibrational frequencies. Typically, density functional theory (DFT) is adopted for this task. DFT solves an approximate representation of Schrödinger equation called Kohn-Sham equation. This equation is based on a proof by Hohenberg that the ground state energy of a system is a functional of electron density of the system.[14] This theorem effectively reduces the complexity of solving the many-electron problem of the Schrödinger equation to “one-electron” problem, making computation feasible. Analysis at this scale is crucial for

predictive modeling, as fundamental electronic characteristics of a material correlate to its catalytic activity.

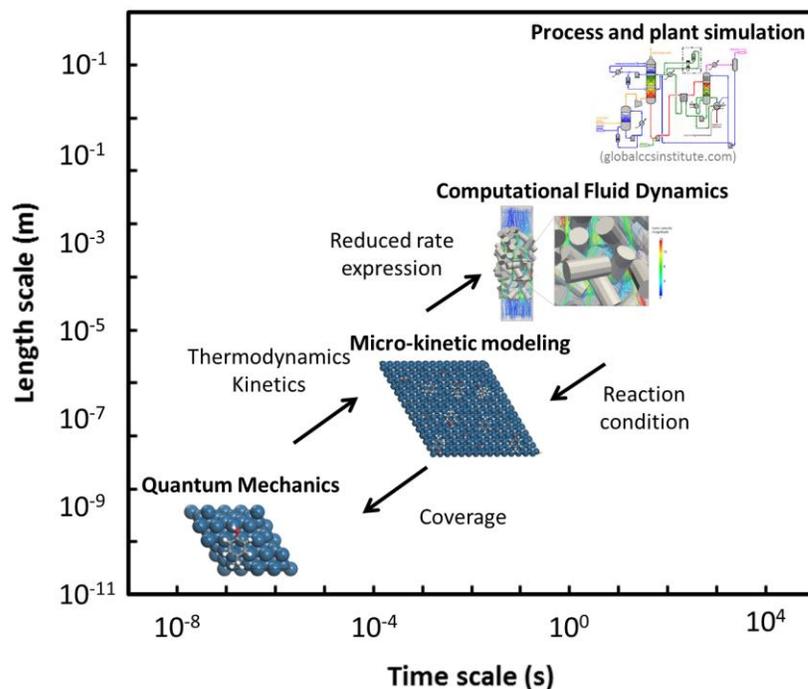


Figure 1.1. Schematic of multi-scale modeling showing different scales and models at each scale as well as couplings between scales. The computational fluid dynamics (CFD) illustration is adopted from Ref.[15]

At the mesoscopic scale (the second from the bottom of Figure 1.1), the collective behavior of individual reactions is averaged over time and space to obtain the observable reaction rate. Temperature-dependent parameters, such as the Gibbs free energy of formation of species and transition states, are rigorously coarse-grained from quantum-scale simulations using statistical thermodynamics (See supporting information of [16] for details). Two methods are widely applied for the mesoscopic

simulation to estimate reaction rates: lattice kinetic Monte Carlo (KMC) and mean-field micro-kinetic modeling (MKM). In lattice KMC, the configuration of adsorbates on the catalytic surface is tracked with time. Each configuration defines a state of the system and a rate of transition to different states based on microscopic rates, called propensities. The propensities are employed to stochastically simulate the evolution of the system. While the KMC offers atomically resolved spatial and temporal information, MKM simulation is faster and easier to parameterize due to the mean-field approximation. The mean-field approximation assumes that adsorbates on a catalytic surface are spatially uniform. As a result, the spatially average concentration of each adsorbate is tracked and is used to compute reaction rates. This approximation eliminates the spatial dependence and stochastic nature of the system. Typically, MKM is employed for simulation of large biomass molecules.

The next largest scale (second from the top in Figure 1.1) involves transport and flow phenomena. Typically, reaction networks are coarse-grained to produce simplified reaction networks and compute rates from reduced rate expressions. Finally, information from this scale is used for plant and process optimization. This thesis focuses on micro and mesoscopic scales, as fundamental catalytic reactivity descriptors are inherent in electronic interactions.[17]

### **1.2.2 Descriptor-based Modeling**

Traditional multiscale-modeling is computationally impractical for simulation of biomass conversion. Due to the large size of biomass molecules, the number of intermediates and reactions is very large, as shown in Figure 1.2. Parameterizing thermochemistry and kinetics of the glucose reaction network using DFT would take ~1 billion CPU hours[18], an intractable task for current supercomputers.

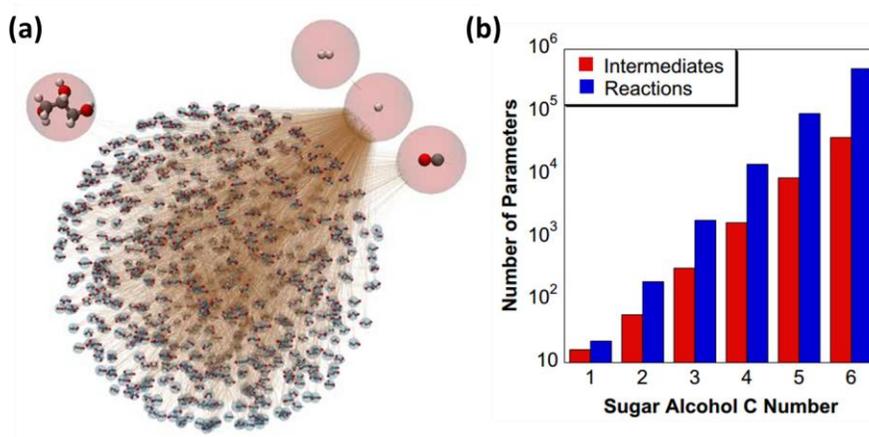


Figure 1.2. (a) Visualized reaction network of glycerol reforming and (b) exponentially scaling of the number of intermediates and reactions with a number of carbon in linear polyols. Adopted from Sutton et al.[18] The nodes and edges in (a) represent intermediates and reactions, respectively.

To overcome the computational cost, semi-empirical methods have been previously explored. While semi-empirical models are not as accurate as DFT, they rely on simple algebraic equations, eliminating the computational cost. In addition, the accuracy of the semi-empirical based MKM can be improved to DFT-level accuracy via hierarchical refinement of sensitive parameters.[19] Thus, development of semi-empirical methods is crucial for biomass modeling.

One of the most important semi-empirical relations is the linear scaling relation (LSR) between species energies. Specifically, Nørskov and co-workers [20] have demonstrated that the binding energy of a partially hydrogenated heteroatom AL<sub>x</sub> linearly correlates with the atomic binding energy of the heteroatom A

$$BE_{AL_x} = \gamma BE_A + \xi \quad (1.1)$$

Here,  $BE_{AL_x}$  is the binding energy of an atom, A, with  $x$  number of hydrogens,  $BE_A$  is the binding energy of an atom, A, and  $\xi$  is the intercept related to the correlation.  $\gamma$  is a function of  $AL_x$ 's valency,

$$\gamma(x) = \frac{x_{\max} - x}{x_{\max}} \quad (1.2)$$

where  $x_{\max}$  is the maximum number of valence electron of atom A (e.g., 4 for C, and 3 for N). For example, the binding energy of  $CH_3$  scales with the binding energy of C with a slope of 0.25.

In order to use the LSRs for large adsorbates, the thermochemistry of all adsorbates on one metal is still necessary. To accomplish this task, Benson's group additivity framework has been extended from gas-phase molecules to surface adsorbates.[16, 21-23] The group additivity predicts thermochemistry by summing up the properties of molecular patterns, called groups, in a molecule:

$$\Phi_i = \sum_{j=1}^{N_{\text{groups}}} n_{ij} \varphi_j \quad (1.3)$$

Here,  $\Phi_i$  is a thermodynamic property of molecule  $i$  (such as entropy or heat of formation),  $N_{\text{groups}}$  is the number of groups in the group additivity model,  $n_{ij}$  is the number of occurrences of group  $j$  in datum  $i$ , and  $\varphi_j$  is the thermodynamic property of group (pattern)  $j$ . A couple of examples of group additivity methods are provided in Figure 1.3.

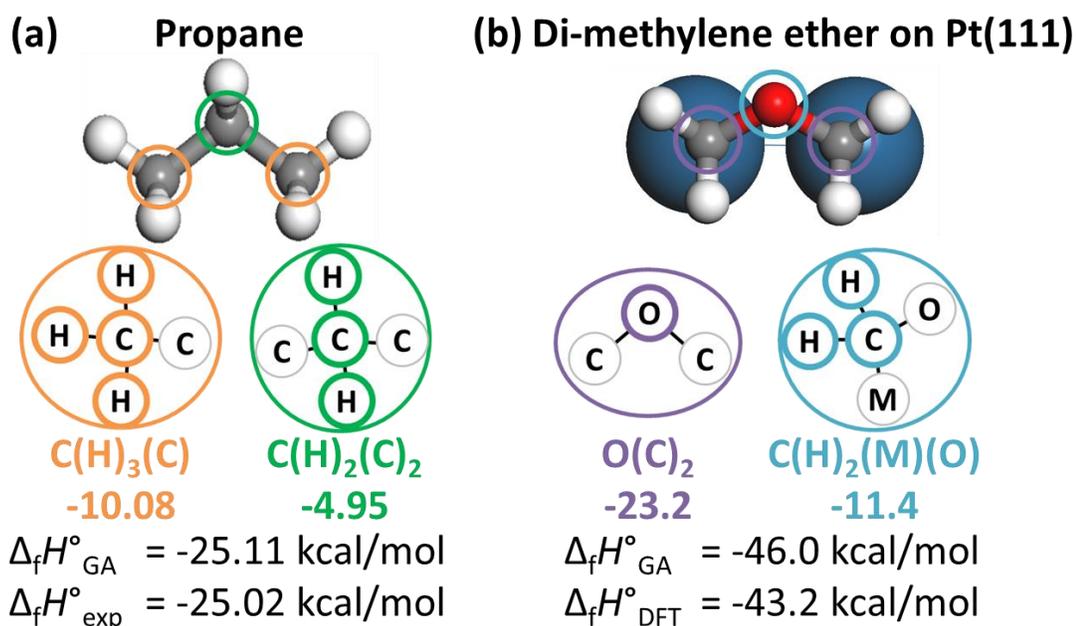


Figure 1.3. Group additivity examples for (a) propane and (b) di-methylene ether on Pt(111). In this method, patterns in the molecule are identified (4 color-coded groups and their heat of formation values are shown), and their thermochemistry values are summed together.  $\Delta_f H^\circ_{GA}$ ,  $\Delta_f H^\circ_{DFT}$ ,  $\Delta_f H^\circ_{exp}$  are the heat of formation calculated from group additivity, DFT, and experiments, respectively.

With the thermochemistry predicted, another linear correlation called Brønsted-Evans-Polanyi relation can be used to compute activation energy,  $E_A$ , of a homologous series, i.e., a family of reactions, from the reaction energy,  $E_{rxn}$ ,

$$E_A = \alpha E_{rxn} + \beta \quad (1.2)$$

where  $\alpha$  and  $\beta$  are the slope and intercept of the correlation. Traditionally, a reaction family is defined based on two atoms undergoing bond scission (e.g., C-C, C-O, C-H, O-H scission), and correlations are built for each family. This categorization has been found to roughly hold true over different catalytic surfaces.[24]

### 1.2.3 Prediction of Materials' Activity and Selectivity

Key to developing economic catalytic processes is the identification of active, selective, stable, and cheap catalysts.[10] Computational studies can aid identifying active materials using Sabatier's principle.[25] The idea behind Sabatier's principle is that catalyst cannot bind molecules too strong as molecules poison the catalytic surface or too weak as reactants are not activated. With the advancement of computational catalysis, descriptors based on species binding strengths can be computed and used to discover catalysts.[26-33]

For example, an activity and selectivity map of ethylene glycol conversion based on oxygen and carbon binding energies as descriptors was constructed [33] and adopted in Figure 1.4. Ethylene glycol is a model compound of sugar polyols. Previously, platinum was known to be the best catalyst for reforming chemistry, but through this mapping, nickel overlayer on top of the platinum was predicted to be more active. Experiments demonstrated that Ni on Pt to be more active in reforming chemistry. In addition, tungsten and molybdenum carbides were discovered to selectively form ethylene, one of the most used monomers. Such a framework provides rapid material screening capabilities and has tremendous potential for biomass utilization.

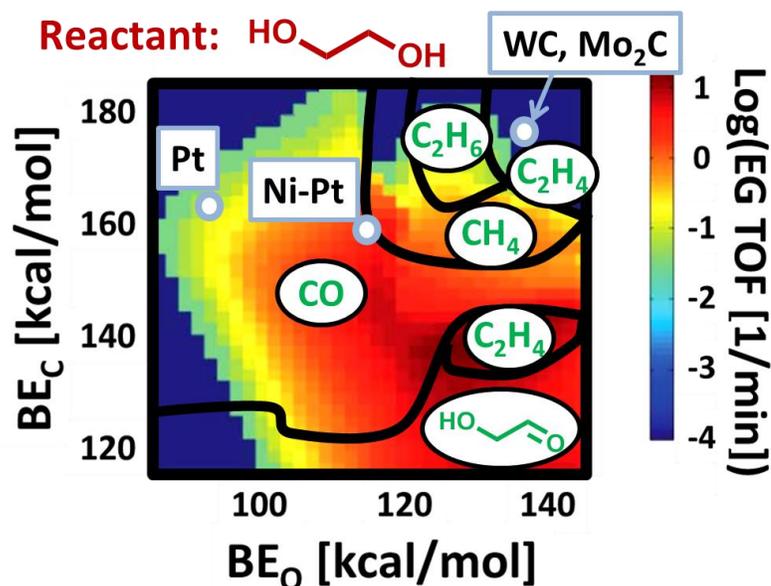


Figure 1.4. Example of predictive catalyst mapping adopted from [33]. This map shows how activity and selectivity of ethylene glycol conversion changes with oxygen and carbon binding energy as key descriptors. The heat color indicates the turn-over-frequency, and the each divided region indicates favored selectivity shown in the white circle. Different catalysts are indicated with circles. Ni-Pt is superior to Pt for reforming and WC and Mo<sub>2</sub>C are selective but not as active catalysts for deoxygenation. Ethane is produced by in situ reforming to H<sub>2</sub>, followed by HDO.

### 1.3 Gaps and Challenges of the Current Paradigm of Predictive Modeling for Biomass Conversion

The current framework requires a number of tasks for it to be comprehensive, accurate and easy-to-use. These tasks are summarized in Figure 1.5. Previously, the Vlachos research group modeled aliphatic oxygenates and furanics (for HMF), [19, 22, 23, 34, 35] but lignin monomers have not been studied (Figure 1.5a). As discussed above, lignin is one of the most abundant components of the biomass (18 - 35 wt%)[6] with the potential to be converted to BTX. However, modeling studies of lignin

monomers have been lacking. In addition, the roles of  $C_6$  aromaticity and ring strain for multi-dentate adsorption and kinetics remain elusive.[22, 35, 36]

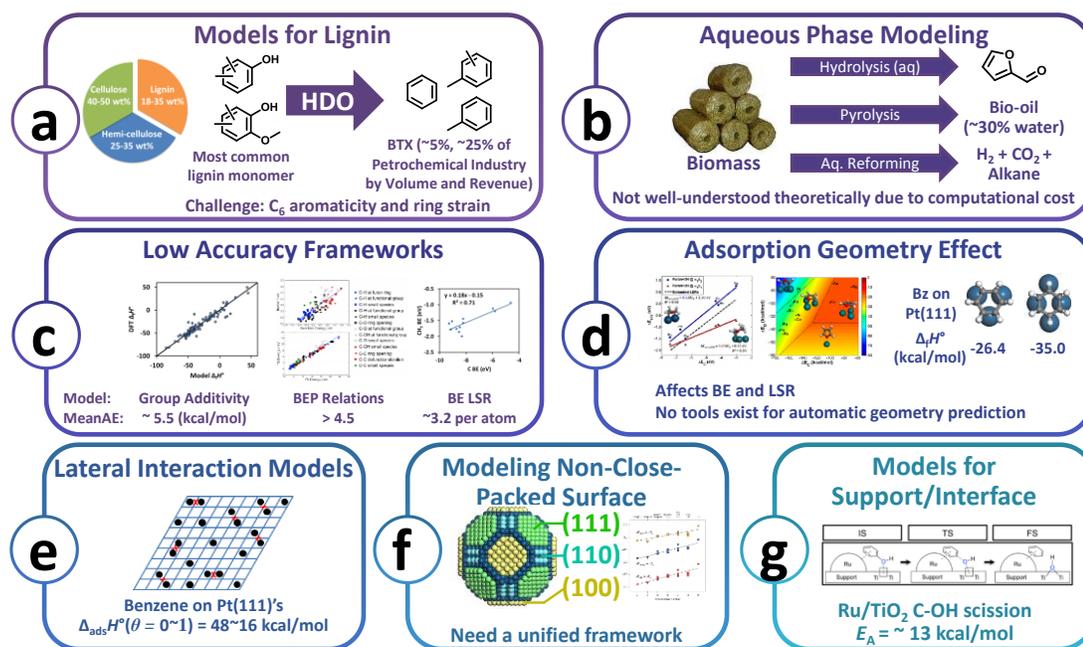


Figure 1.5. Overview of challenges and gaps in predictive modeling of biomass conversion on solid catalysts.

Another gap is our ability to model chemistry on metals in the aqueous phase (Figure 1.5b). The hydrolysis and pyrolysis involve aqueous phase (aqueous phase medium for hydrolysis, and aqueous product for pyrolysis). Developing efficient solid catalysts in aqueous phase could simplify the multi-step biomass conversion. In addition, fundamental understanding of the aqueous phase reforming mechanism is lacking. Recent works have investigated a subset of a reaction network  $C_3$  molecules via hybrid DFT and molecular mechanics approach,[37-41] but investigation of entire reaction networks remains too computationally intensive. Combined with hierarchical

refinement methodologies, a solvation energy semi-empirical model could provide significant insights into aqueous phase processing at reduced computational cost.[18]

Despite their efficiency, semi-empirical tools exhibit low accuracy (Figure 1.5c). Previous group additivity models have large cross-validation error ( $>5.5$  kcal/mol).[35, 36] This error is especially important as microkinetic models have been found to be sensitivity to thermochemistry.[42] Similarly, BEP relations have a high mean absolute error (MeanAE) ( $>4.5$  kcal/mol), [43] which combined with group additivity, result in as much as four orders of magnitude standard deviation in turnover-frequency (TOF).[42] When errors are too large, local sensitivity analysis may not correctly identify the rate-determining-steps (RDS) and lead to improper hierarchical model refinement. In this case, correlated global sensitivity analysis needs to be employed.[42] The model error is worse when the thermochemistry and kinetics parameters are transferred to other metals via LSRs.[44] The LSR's error is additive with the number of adsorbate atoms bonding to the surface, where the MeanAE for a single surface-bonded atom is about 3.2 kcal/mol.[20] As the biomass molecules are multi-dentate, errors from the LSRs are expected to be larger, resulting in poor model predictions.

Vorotnikov et al. has found that LSRs can be improved by accounting for the binding site of furanics molecules.[34] The preferred binding conformation changes based on the fundamental characteristics of the surface. Similarly, multi-dentate adsorbates, such as benzene, have a binding energy difference of 10 kcal/mol between two binding sites on Pt(111).[45] As benzene is the backbone of lignin monomers, binding sites play a critical in describing lignin monomers' thermochemistry. Clearly, the ability of automatically predicting adsorption conformation is essential (Figure

1.5d). While gas group additivity has been implemented via a web interface by the NIST Webbook and MIT ([rmg.mit.edu](http://rmg.mit.edu)), adsorbate group additivity requires prior knowledge of the adsorbates' binding geometry. Currently, this step is manual which is tedious and time-consuming for a reaction network involving hundreds of species. Development of an algorithm to automate the conformation prediction will aid computational studies.

Thermochemistry prediction requires also the inclusion of lateral interactions. Benzene's adsorption energy ranges from 48 to 16 kcal/mol as the coverage varies between 0 and 1 (Figure 1.5e).[46] As microkinetic models are typically parameterized with zero-coverage DFT calculations, thermochemistry parameters are expected to have at least ~30 kcal/mol error, which would then propagate errors in activation energies. For example, Sayes et al. has shown that the toluene hydrogenation energy profile dramatically changes with coverage effects.[47] Estimating lateral interactions for large biomass reaction networks is expected to take significant DFT calculations. Developing a semi-empirical framework for estimating lateral interactions will aid in correctly capturing the physics while reducing computation time for building a MKM.

The last two tasks relate to the extension of the framework. Previously, semi-empirical method capabilities have been focused on closely packed surfaces (111 for fcc, 0001 for hcp, and 110 for bcc metals). Though the optimal material prediction may be independent of the studied facet,[25, 28, 48, 49] developing tools for other facets can aid understanding structural effects. Another phenomenon that has not been studied much is the effect of support. The support may affect the solid catalyst via: (1) changing the electronic structure of the metal surface, (2) creating interfacial sites with

different reactivity than the metal, (3) carrying out reactions itself and participating in spillover and reverse effects, and (4) changing the physical structure of the catalyst.[50, 51]

#### 1.4 Dissertation Scope

This thesis addresses five challenges shown in Figure 1.5 (a, b, c, d, and e). Metal-catalyzed biomass conversion has been extended for the first time to lignin monomers as well as sugar alcohols in the aqueous phase. Major methodological developments have enabled these kinetic studies: (1) transition state structure-sensitive BEP relations; (2) a lateral interaction model based on the number of the surface bonding atoms; (3) a binding-site based group additivity framework; and (4) a machine-learning method for thermochemistry prediction. The organization of this thesis is outlined next.

While the HDO of phenolics has been widely studied experimentally, a consistent mechanism has been lacking. Chapter 2 builds a DFT-based microkinetic modeling for the conversion of *p*-cresol, a lignin model compound, on Pt(111). This model elucidates the mechanism and is consistent with all experimental results. In addition, an improved BEP-relation and a lateral interaction model are introduced. The new BEP relationship is capable of predicting DFT barriers near the chemical accuracy.

Chapter 3 develops a group additivity model for lignin components on Pt(111). We improve the prior group additivity framework via examining the theoretical background of group additivity and implementing electronic analysis. Bader et al.[52] has demonstrated that the regressed group energy is conserved across molecules as the electron distribution within a group is transferable to other molecules. In order to

select the best groups, electron density (Bader) analysis is employed. The method reduces the previous mean absolute error from 5.15 to 2.81 kcal/mol.

Chapter 4 extends group additivity to solvated alcohol adsorbates on Pt(111). Adsorbate group additivity is combined with solvation group additivity. The DFT solvation energy is predicted within chemical accuracy (MeanAE of  $\sim 1$  kcal/mol). The effect of solvation on the reaction network thermochemistry of ethanol steam reforming is studied. The solvation group additivity scheme, in conjunction with published BEPs, is applied to develop a MKM for aqueous phase reforming of glycerol for the first time in Appendix E.

Despite the advancement in the group additivity framework, the group additivity's cross-validation MeanAE remains high (2.81 kcal/mol). Chapter 5 exploits machine-learning techniques to further improve our model. A descriptor selection regression called Least Absolute Shrinkage and Selection Operator (LASSO) is implemented to automatically find the most informative graph patterns that describe thermochemistry accurately. The LASSO-trained model performs better than the previous group additivity (reduces MeanAE from 2.81 to 2.08 kcal/mol). The LASSO method is promising especially for large data sets.

To address the manual adsorption conformation, a machine learning framework for adsorbate conformation prediction is introduced in Chapter 6. This algorithm learns from a diverse adsorption conformation data computed using DFT, and rapidly predicts adsorption configurations given the gas structure of a molecule. The algorithm shows promising accuracy to eliminate manual construction of conformation for performing DFT calculations.

Finally, Chapter 7 summarizes this thesis. Future work is also suggested.

## Chapter 2

### MECHANISM OF DEHYDRATION OF PHENOLS ON NOBLE METALS VIA FIRST-PRINCIPLES MICROKINETIC MODELING

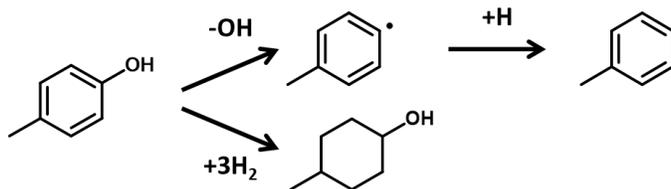
#### 2.1 Introduction

Conversion of biomass to petroleum-like fuels and chemicals is an active area of research.[53] Fast pyrolysis techniques provide one such viable option to convert biomass into a depolymerized liquid bio-oil.[6] The high oxygen content in the bio-oil makes it unsuitable for applications.[6] Consequently, current research has been centered around developing catalysts for hydrodeoxygenation (HDO),[54-62] a process involving high hydrogen pressure and temperature to remove oxygen as water.[6] More specifically, HDO of phenolic compounds has attracted significant attention because alkyl phenols constitute a major portion of the remaining oxygenates from catalytic pyrolysis over zeolites (17 – 35%).[63-68]

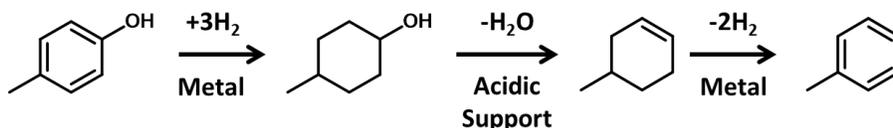
Unlike short-chain hydrogenated alcohols, such as ethanol, whose Brønsted- and Lewis-catalyzed dehydration mechanisms are well-established and involve OH and H removal from the  $\alpha$ -C and  $\beta$ -C, respectively,[69-73] phenols, such as phenol and cresol, cannot follow the same mechanism due to having an unsaturated ring. Several mechanisms have been proposed for HDO of these alcohols, as summarized in Figure 2.1 for the case of p-cresol. Direct dehydroxylation followed by hydrogenation (Figure 2.1a) has been proposed for oxophilic metals such as Fe, Ru, and CoMoS.[74-77] This mechanism cannot be operable on low oxophilicity metals like Pt. A bi-functional mechanism has been the most popular and entails full ring hydrogenation

on the metal to form methyl cyclohexanol, followed by dehydration on the acid support to form methyl cyclohexene and dehydrogenation on the metal to form toluene, as shown in Figure 2.1b.[58, 59, 61, 78-83] This mechanism may be applicable on metal/Brønsted acid catalysts. However, this mechanism falls short in explaining HDO on non-acid-based supported catalysts (Rh/C, Pd/C, Pt/C, Pt/SiO<sub>2</sub>, Pd-Fe/C, Ni/SiO<sub>2</sub>).[58, 62, 84-87] Lobo and co-workers confirmed that methyl cyclohexanol is not dehydrated on Pt/SiO<sub>2</sub>, and suggested that, instead of methyl cyclohexanol, reactive methyl cyclohexadienol or methyl cyclohexenol with hydrogenated  $\alpha$ -carbon undergoes dehydration.[58] Additional experiments feeding methyl cyclohexanol or methyl cyclohexanone did not lead to toluene, further indicating that mechanism B is not applicable on these catalysts.[62] In order to resolve this enigma, Resasco and co-workers suggested that a fast keto-enol tautomerization followed by carbonyl hydrogenation to form methyl cyclohexadienol precedes dehydration[62, 88, 89] (Figure 2.1c). Furthermore, in order to explain the formation of oxygenated products, methyl cyclohexanone and methyl cyclohexanol, they postulated a parallel ring hydrogenation path (lower path in Figure 2.1c). Given the lack of acidity of Pt/SiO<sub>2</sub> and the fact that keto-enol tautomerization is often acid catalyzed, this mechanism may not be plausible either. Despite many experimental advances, the mechanism by which phenols dehydrate remains unclear.

**(a) Direct Dehydroxylation**



**(b) Hydrogenation Followed by Dehydration and Dehydrogenation**



**(c) Keto-Enol, Partial Hydrogenation Followed by Dehydration**

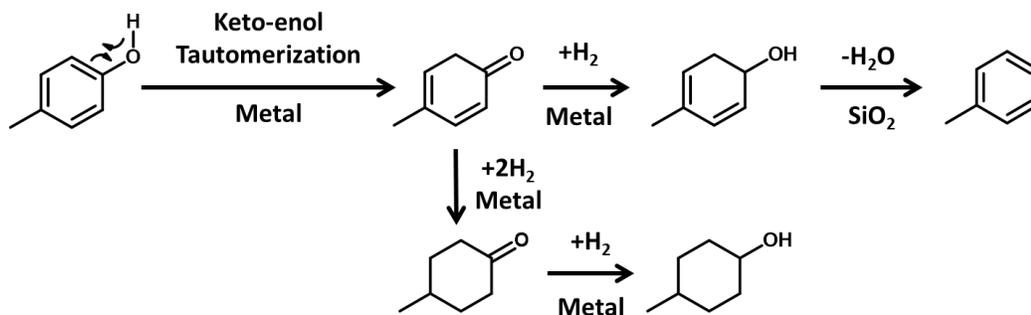


Figure 2.1. Proposed HDO mechanisms of phenolic compounds in the literature.

Density functional theory (DFT) studies have been limited to investigating adsorption and a select number of elementary reactions of phenols on metal surfaces, such as Pt(111),<sup>[90, 91]</sup> Rh(111),<sup>[91]</sup> and Ni(111).<sup>[92, 93]</sup> More recently, a few paths of the HDO mechanism of *m*-cresol were studied via DFT on Pt(111)<sup>[77]</sup>, Ru(111)<sup>[77]</sup>, and Ru/TiO<sub>2</sub><sup>[87]</sup>, but quantitative comparison to experimental data is still to be reported.

In this work, we provide insights into the HDO mechanism of *p*-cresol on the Pt(111) surface. We encompass all the relevant reactions in a comprehensive reaction

network and build a microkinetic model. We develop Brønsted-Evans-Polanyi (BEP) and Transition State Scaling (TSS) relations for fast parameterization and refine the microkinetic model using DFT. Our results reconcile various experimental data starting from various substrates.[58, 62] Reaction path analysis demonstrates a novel HDO mechanism on Pt whereby the ring is activated prior to dehydroxylation but stable species, such as methyl cyclohexanol and methyl cyclohexanone, are not intermediates toward toluene but rather, they are side products. We show that ring hydrogenation changes the conformation of the molecule and is eventually responsible for facile dehydroxylation. Interestingly, while the fraction of toluene increases at the expense of methyl cyclohexanol and methyl cyclohexanone with increasing space velocity, this is not due to reactions in series but rather due to sharing common intermediates.

## **2.2 Methods**

### **2.2.1 Reaction Network**

Previous experimental studies of HDO of cresol observed 4 products: methyl cyclohexanol, methyl cyclohexanone, methyl cyclohexane, and toluene.[58, 62] The saturated products imply hydrogenation of the ring carbon, which can occur at 6 different ring locations with different combinations. Consequently, the extensive reaction network of p-cresol (Figure 2.1) required an automatic mechanism generation procedure using the RING software[94] (see A.2 of the appendix regarding the enumeration rules) to account for 4 different types of elementary reactions (Figure 2.2): (1) hydrogenation of the ring to produce alcohol intermediates ending with methyl cyclohexanol, (2) dehydrogenation of the O-H group of alcohol derivatives to

form the corresponding ketones, leading to methyl cyclohexanone, (3) dehydroxylation of the alcohol derivatives to form hydrocarbon species that are H-deficient at the  $\alpha$ -carbon, and (4) hydrogenation of the resulting hydrocarbon at the dehydroxylated  $\alpha$ -carbon followed by dehydrogenation or hydrogenation of the carbon ring to produce stable species, such as toluene and methyl cyclohexane. The reaction network is comprised of 464 elementary reactions (500 when including the adsorption steps) between 144 surface species.

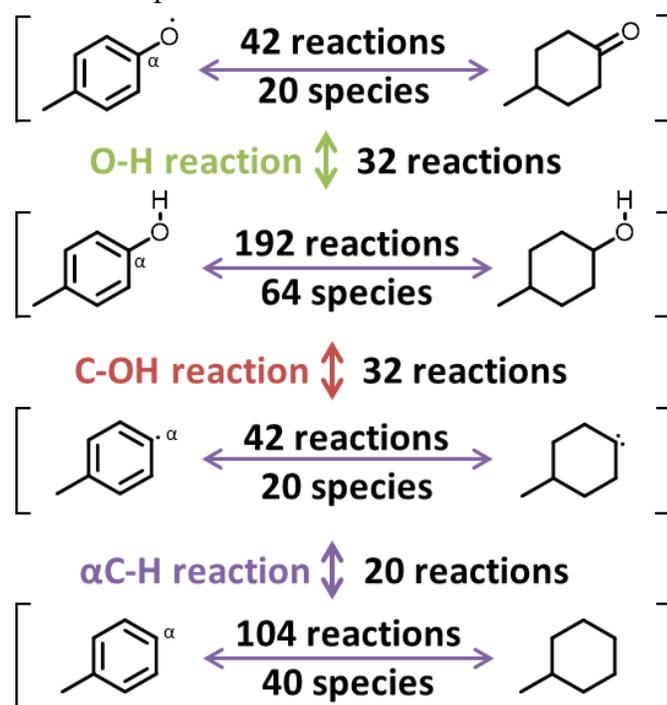


Figure 2.2. Summary of the overall p-cresol HDO reaction network. Ring carbon position is numbered as shown on p-cresol. Throughout the study, we will identify the original location of hydroxyl group as the  $\alpha$ -carbon. Horizontal arrows represent ring (de)hydrogenation and vertical arrows represent bond scission/formation indicated in color.

### 2.2.2 Density Functional Theory (DFT) Calculations

We performed plane-wave DFT calculations using the Vienna ab initio Simulation Package (VASP), version 5.3.2.[95] The electron-electron exchange and correlation energies are computed using the PBE-D3, Perdew, Burke, and Ernzerhof functional with the dispersion correction.[96, 97] The inclusion of the dispersion corrections was based on their important role revealed in benzene-adsorption studies.[98, 99] Furthermore, the core electrons were treated with the projector augmented-wave (PAW) pseudopotentials. [100, 101]

We optimized the lattice constant of Pt using the Birch-Murnaghan equation of state.[102, 103] The Pt bulk was simulated using the tetrahedron method with Blochl corrections and a  $15 \times 15 \times 15$  Monkhorst-Pack  $k$ -point mesh.[100, 104] The lattice constant of Pt was calculated to be  $3.917 \text{ \AA}$  using the PBE-D3 functional, in close agreement with the experimental value of  $3.92 \text{ \AA}$  .[105]

The Pt slab was simulated using a four layer deep  $4 \times 4$  unit cell. The bottom two layers of Pt were held fixed at the bulk position, whereas the top two layers were allowed to relax. The vacuum between the slabs was set at  $20 \text{ \AA}$  to minimize the interaction in the  $z$ -direction. The Brillouin zone was integrated using a  $5 \times 5 \times 1$   $k$ -mesh, with a Methfessel-Paxton smearing of  $0.1 \text{ eV}$  ( $0.01 \text{ eV}$  for gas-phase calculations).[106] For the plane-wave set, a cutoff energy of  $400 \text{ eV}$  was used. The adsorbate-slab system was relaxed until all the forces were smaller than  $0.05 \text{ eV/\AA}$  . The transition states (TSs) were located using the climbing image nudged elastic band (CI-NEB) as well as the dimer methods.[107, 108]

The adsorption energy was calculated as  $E_{\text{ads}} = E_{\text{slab+i}} - E_{\text{slab}} - E_{\text{i}}$ , where  $E_{\text{slab+i}}$ ,  $E_{\text{slab}}$ , and  $E_{\text{i}}$  are the total energies of the adsorbate-slab system, the clean slab and the

adsorbate in the gas-phase, respectively. The gas-phase molecules were calculated in a  $20 \times 20 \times 20 \text{ \AA}$  cell. The reference energy was taken to be infinitely separated species.

### **2.2.3 Brønsted-Evans-Polanyi (BEP) and Transition State Scaling (TSS) Relations**

As we show below, published BEP and TSS are inadequate for some of the lignin reactions. Therefore, we develop BEP and TSS relations for various homologous series using a select set of DFT calculations. The data sets are comprised of 29 reactions for the C-H scission, 8 reactions for the O-H scission, and 15 reactions for the C-OH scission. These reactions and the associated species primarily lie along the minimum energy hydrogenation pathway between *p*-cresol and methyl cyclohexanol. (Illustration and data are provided in Figure A.2 and Table A.3 and A.4, respectively) We discuss the statistics of BEP and TSS in the results section below.

### **2.2.4 Parameterization of the Microkinetic Model**

We performed the microkinetic modeling using our in-house reactor code built around CHEMKIN<sup>TM</sup>.<sup>[109]</sup> Given the large number of reactions and species (Figure 2.2), we employed and developed DFT-based fast screening methods for parameter estimation (see overview in Saliccioli et al.<sup>[19]</sup>). Specifically, we invoke a statistical mechanics approximation<sup>[110-112]</sup> for estimating the thermochemistry of surface species from the corresponding gaseous species (see A.3.1 in the appendix for detail); the thermochemistry of the latter is computed at the G4-level using Gaussian.<sup>[113]</sup> Reaction barriers were estimated using BEPs.

In order to improve accuracy, we adopt the refinement methodology described by Saliccioli et al.,<sup>[19]</sup> where species and reactions are refined with subsequent DFT calculations.<sup>[18]</sup> In this study, we perform two sets of MKM refinement: (1) Inclusion

of the lateral interactions in an iterative fashion, where we identified the most abundant surface intermediates by running the microkinetic model and compute their interactions via DFT. This procedure is repeated until convergence. As the reaction network involves a large number of species, we developed a lateral interaction model that accounts for the number of surface-adsorbed carbon atoms (see Figure A.5 and A.3.2 in the appendix). The model is used to assign lateral interactions for the alcohol, ketone and hydrocarbon derivatives with 0, 1, and 2 hydrogenated ring carbons. Due to a large fraction of the reaction network being partially equilibrated, the dominant surface species are sensitive to these interactions. However, the overall conclusions of our work are unaffected. (2) Following the inclusion of the lateral interactions, the barriers and thermochemistry of important reactions were refined. Of the 144 surface species, the thermodynamic properties of 26 species encountered in the minimum energy hydrogenation pathway and 26 kinetically relevant species (identified via sensitive analysis) were calculated using DFT calculations and zero-point energy (ZPE) corrections in these species were included. Parameter adjustments for this study are discussed in A.5 of the appendix.

To ensure thermodynamic consistency, we referenced the VASP-calculated surface reaction barriers and intermediates' thermochemistry to NIST experimental values of hydrogen, toluene, and water as discussed in Refs.[19, 114] Here, we preserve the gas-phase and surface reaction thermochemistry and adjust the adsorption thermochemistry. For the gas-phase thermochemistry of *p*-cresol, toluene, methyl cyclohexanol, methyl cyclohexanone and methyl cyclohexane, the G4 quantum chemical method[113] was used to calculate the electronic energies and vibrational frequencies. Then these quantities were referenced to the NIST experimental values to

compute thermodynamic properties (see supporting information of Vorotnikov et al.[35] for the referencing method).

### **2.3 Brønsted-Evans-Polanyi (BEP) and Transition State Scaling (TSS) Relations**

Previously, our group has developed BEP relations for ethanol on Pt(111)[43]. When these BEPs were employed for a select group of reactions (see Methods), the mean absolute error (AE) in the reaction barriers was 0.13 eV, 0.50 eV, and 0.75 eV, for the C-H, O-H and C-OH scission reactions, respectively, i.e., the classic BEP relations, where the homologous series is based on the two atoms that undergo bond breaking, are sufficient for the C-H but inadequate for O-H and C-OH scission. Such shortcomings have recently been reported for other molecules. For example, Zaffran et al. and Sutton et al. observed improved accuracy by accounting for the adjacency of the C-H scission to the –OH group for small aliphatic molecules (< 5 C atoms) on Pt(111)[43] and Rh(111)[115]. Wang et al. concluded that structural similarity in reactants and transition states of a homologous series is a prerequisite for improved accuracy and developed C-O and C-C scission BEPs by accounting for the neighboring atoms of carbon and oxygen of furan derivatives and small molecules on Pd(111).[116] Lee et al. also observed a similar behavior for the C-O scission of methoxy and hydroxyl groups in guaiacol.[117]

Here, we extend the definition of the homologous series and group the data based on the binding of the initial state (IS), transition state (TS), and final state (FS) structures, as shown in Figure 2.3. We identify two TS structures depending on the number of atoms involved. A 3-centered TS entails the  $\alpha$ -carbon and the –OH group bound to the same Pt atom, whereas a 4-centered TS entails the  $\alpha$ -carbon and –OH

group bound to two separate Pt atoms (Figure 2.3c). In addition, we find that the hybridization of the  $\alpha$ -carbon in the FS of 4-centered TS is important.

In IS1, the  $\alpha$ -carbon and an adjacent ring carbon share a Pt atom resulting in a  $\pi$ -mode binding. This IS leads to a 3-centered TS, because upon dehydroxylation, the  $\alpha$ -carbon forms a  $sp^2$  hybridized C-Pt bond (bond angles being  $120^\circ$ ) with a nearby Pt atom on which the -OH group is also bound to. The  $\pi$ -mode binding occurs frequently for unsaturated species due to the relatively large Pt lattice spacing compared to the bond distances between the ring carbons. In IS2, the center of the ring is over a bridge site, and the  $\alpha$ -carbon is  $\sigma$ -bonded and is off-centered with respect to the Pt atom, as shown in Figure 2.3c. During C-OH scission, the -OH group forms a bond with the Pt atom on which the  $\alpha$ -carbon is bound to, resulting in a 3-centered TS. IS2 is again observed frequently in unsaturated molecules as the off-centered  $\alpha$ -carbon is caused by strain (see below for strain discussion). Finally, IS3 has the  $\alpha$ -carbon  $\sigma$ -bonded directly to a Pt atom and the TS involves the -OH group bound to a neighboring Pt atom, leading to a 4-centered TS. The  $\alpha$ -carbon in the FS is typically  $sp^2$  hybridized owing to the strain restricting the  $\alpha$ -carbon from becoming  $sp^3$  hybridized (a few exceptions of  $sp^3$  hybridized structures are shown in Figure A.3). Strain on the  $\alpha$ -carbon needs to be minimized for direct binding to a Pt atom, and, as a result, IS3 is observed more frequently for partially hydrogenated species.

Figure 2.3a and b shows BEP and (FS-based) TSS relations. The blue, red, and green colored data correspond to a 3-centered TS and a 4-centered TS with  $sp^2$  and  $sp^3$  hybridized FS, respectively. Reactions with a 4-centered TS entail lower reaction barriers than those with a 3-centered TS by  $\sim 1$  eV for the same reaction energy. This trend is similar to observations made by Neurock and co-workers,[118] who proposed

lower stability of the 3-centered TS due to the destabilization from overcrowding.  $sp^3$  hybridization leads to a stabilization of the FS energy by  $\sim 0.25$  eV (green data) compared to  $sp^2$  hybridization (red data), decreasing the reaction energy by 0.25 eV.

In implementing the “structure-sensitive” BEP relations in the microkinetic model, we computed all the adsorption configurations of *p*-cresol and its hydrogenated derivatives following the adsorption study for benzene derivatives by Morin et al.[119] We used adsorption structures of alcohol derivatives that lead to a 4-centered TS (IS3), due to their lower dehydroxylation barrier, provided such structures were stable; otherwise structures with 3-centered TS (IS1 and IS2) were employed.

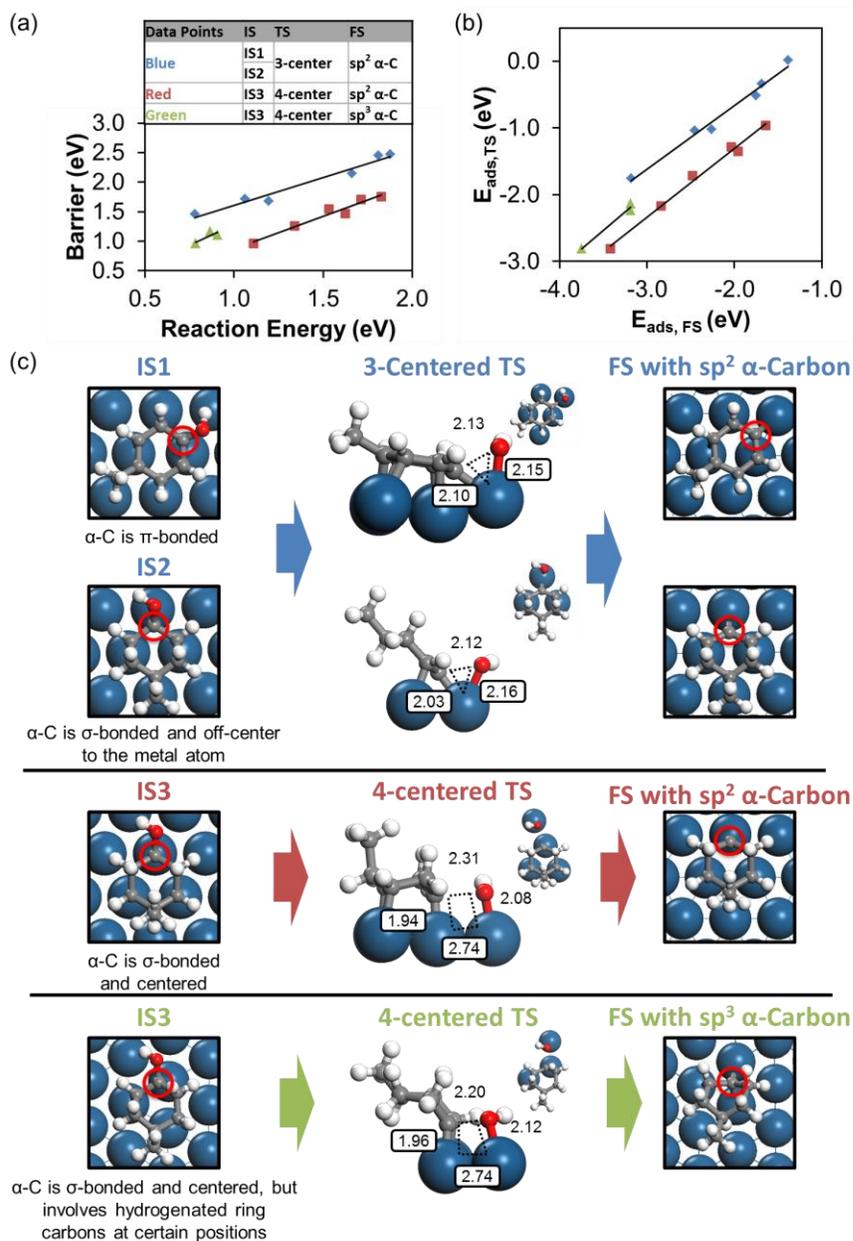


Figure 2.3. (a) BEP relations for the C-OH scission reactions of three homologous series. Raw data is shown in Table A.4 and BEP relations in Table 2.1. (b) Corresponding TSS based on the FS energy. (c) Structural characteristics of each BEP relation with examples (bond lengths in Å). Identical color codes are maintained in the plots and the figures. The Table (panel a) summarizes the types of TS structures. Barriers and reaction energies are reported as electronic energies with no zero-point energy corrections for these fast screening methods.

A synopsis of BEP and TSS relations is given in Table 2.1 and Table 2.2. The accompanying graphical representation is shown in Figure A.4. Significant reduction in the mean AE of O-H and C-OH scission was found from 0.50 and 0.75 eV to 0.04 and ~0.06 eV, respectively. For barrier estimation, the barriers obtained using the FS based TSS are slightly more accurate than those using the BEP relations, but owing to the ease of implementation, the latter were used for microkinetic modeling.

Table 2.1. Slopes and intercepts with 95% confidence intervals for BEP and TSS relations (uncorrected for zero-point energy).

Correlation	<i>n</i>	BEP Relations		TSS Relations -IS		TSS Relations -FS	
		$\alpha$	$\beta$ (eV)	$\alpha$	$\beta$ (eV)	$\alpha$	$\beta$ (eV)
C-H	29	0.65 ± 0.24	0.91 ± 0.09	0.94 ± 0.10	0.51 ± 0.36	1.14 ± 0.08	0.93 ± 0.26
O-H	8	1.04 ± 0.40	0.23 ± 0.18	0.99 ± 0.35	0.64 ± 1.04	1.06 ± 0.11	-0.16 ± 0.23
C-OH (3-centered TS)	6	0.94 ± 0.25	0.67 ± 0.37	2.12 ± 1.76	4.09 ± 4.88	0.95 ± 0.17	1.25 ± 0.38
C-OH (4-centered TS; sp <sup>2</sup> )	6	1.11 ± 0.33	-0.24 ± 0.51	1.66 ± 0.47	3.54 ± 1.51	1.03 ± 0.14	0.75 ± 0.34
C-OH (4-centered; sp <sup>3</sup> )	3	1.33 ± 12.21	-0.06 ± 10.44	1.38 ± 1.09	2.39 ± 3.79	1.12 ± 1.98	1.38 ± 6.69

*n* is the number of regressed data points, and  $\alpha$  and  $\beta$  are coefficients of the linear correlation,  $y = \alpha\Delta E_{\text{reaction}} + \beta$

Table 2.2. Mean and max absolute errors (AE) of reaction barriers computed using each correlation.

Correlation	BEP Relations		TSS Relations-IS		TSS Relations-FS	
	Mean AE (eV)	Max AE (eV)	Mean AE (eV)	Max AE (eV)	Mean AE (eV)	Max AE (eV)
C-H	0.11	0.30	0.16	0.50	0.09	0.45
O-H	0.04	0.12	0.15	0.20	0.04	0.10
C-OH (3-centered TS)	0.07	0.12	0.23	0.54	0.07	0.11
C-OH (4-centered TS; sp <sup>2</sup> )	0.07	0.12	0.10	0.20	0.05	0.09
C-OH (4-centered; sp <sup>3</sup> )	0.05	0.07	0.02	0.02	0.03	0.05

## 2.4 Energetics of the *p*-cresol HDO Reaction Network

Figure 2.4 shows the elementary reactions in converting *p*-cresol to methyl cyclohexanol along the minimum energy path, i.e., the path where hydrogen is

sequentially added to the most stable position (see Figure A.2 for more details). The reactions leading to ketone and hydrocarbon derivatives are also depicted. Reaction energies, computed via DFT, and barriers, computed via BEP or DFT, are also shown. These DFT values were the main data used to build the BEP and TSS discussed above.

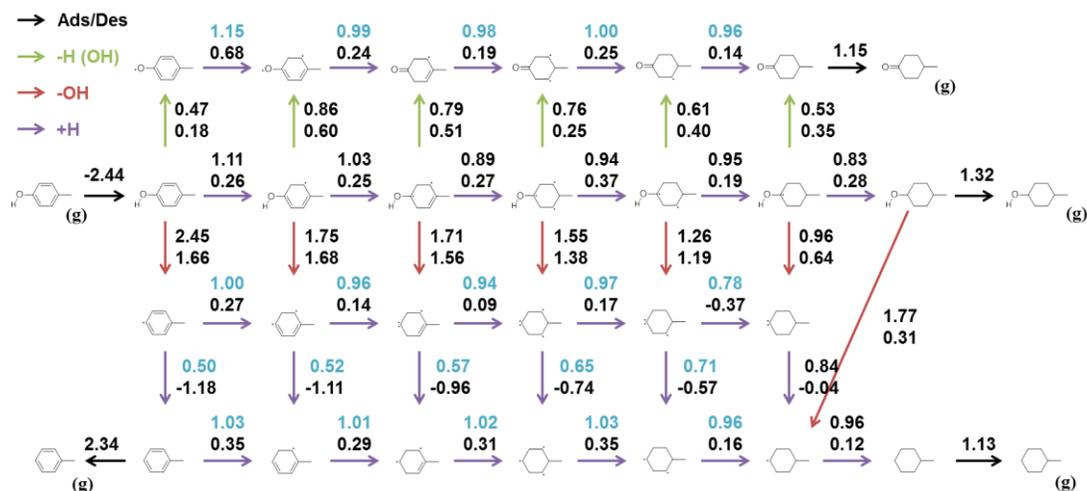


Figure 2.4. Minimum energy path of ring hydrogenation of p-cresol to methyl cyclohexanol along with C-H, O-H, and C-OH reactions. For each reaction, upper number is reaction barrier and lower number is reaction energy in eV. Reaction energies (electronic energies without zero point correction) are calculated using DFT. Reaction barriers,  $E_a$ , are calculated using DFT (black) or BEP relations developed herein (cyan). Color convention of arrows as in Scheme 1.

Examination of the data indicates that the O-H scission barriers ( $0.67 \pm 0.12$  eV) are lower than the hydrogenation reactions ( $1.02 \pm 0.05$  eV), and the reverse O-H formation barriers ( $0.29 \pm 0.08$  eV) are even lower. Thus, O-H reactions are expected to be fast and equilibrated. This finding is consistent with the general understanding that phenols are more acidic than alcohols. The desorption energy for all stable products is higher (1.1 – 1.3 eV for saturated vs. 2.3 – 2.4 eV for unsaturated) owing

to the dispersion interactions. This finding indicates that desorption may be slower than hydrogenation/dehydrogenation, potentially affecting product selectivity.

Figure 2.5 depicts the C-OH scission barriers for all the C-OH scission reactions in the reaction network. The C-OH scission barrier of the unsaturated *p*-cresol is 2.45 eV, strongly suggesting that direct dehydroxylation on Pt is unlikely, refuting mechanism A (Scheme 1). This finding is consistent with a recent DFT calculation of *m*-cresol on Pt.[77] In addition, the dehydroxylation barrier of the methyl cyclohexanol is significantly high, 1.77 eV, indicating that complete ring hydrogenation followed by dehydration on Pt is not effective. (mechanism B, Scheme 1) Similarly, the dehydroxylation barriers of stable alcohol species with hydrogenated  $\alpha$ -carbon are expected to be high as dehydroxylation from non-surface bound  $\alpha$ -carbon is not catalyzed by the Pt surface. This indicates that the suggested mechanism where stable alcohols with hydrogenated  $\alpha$  and  $\beta$  carbon undergo dehydration is not favorable on Pt.[58]

Pt can easily hydrogenate the ring, with typical hydrogenation barriers falling in the green shaded region in Figure 2.5. As the carbon atoms get hydrogenated, the ring lifts gradually off the surface (see Figure 2.6), resulting in the dehydroxylation reaction becoming less endothermic (see Figure A.14a; 1.68 eV at zero hydrogenation vs. 0.31 eV at full saturation) and C-OH scissions involving more frequently 4-centered rather than 3-centered TS with significantly lower barriers (see Figure A.14b). While the locations of hydrogens added to the ring affect the barrier (vertical scatter in Figure 2.6), species along the minimum energy path expose some of the lowest barriers. Interestingly, the minimum C-OH scission barrier decreases profoundly with increasing degree of hydrogenation. Upon 3-5 sequential

hydrogenations, some of the dehydroxylation barrier are  $\sim 1$  eV hinting to the possibility that dehydroxylation is likely upon sufficient hydrogenation of the ring. Importantly, complete ring hydrogenation does not kinetically favor dehydroxylation, pointing to the fact that methyl cyclohexanol and cyclohexanone are not main products from which hydrocarbons form (see also below). Instead, dehydroxylation is facile on Pt for sufficiently hydrogenated species but not necessarily for stable species.

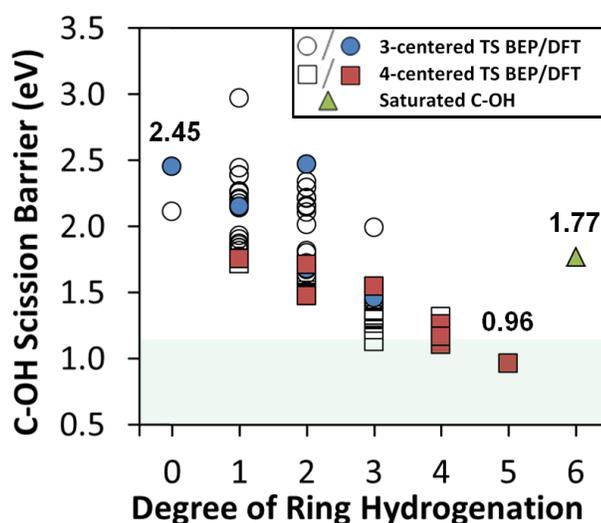


Figure 2.5. C-OH scission barriers vs. degree of ring hydrogenation. Typical hydrogenation barriers fall within the green shaded region. Open (closed) symbols are BEP (DFT) computed barriers. Filled green triangles correspond to C-OH scission of methyl cyclohexanol

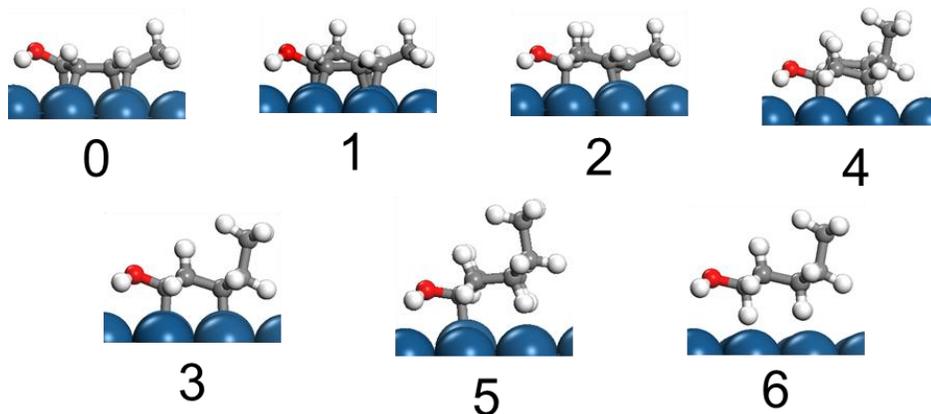


Figure 2.6. Alcohol intermediates along the minimum energy path demonstrating lifting as the ring carbons are hydrogenated (the number indicates the number of hydrogenated carbons). During hydrogenation the  $sp^3$  hybridized ring carbons give place to a chair shape structure with less strain.

## 2.5 Conversion and Selectivity to Major Gas Products

Here we compare to experimental data reported by Nie et al.[62] and Foster et al.[58] (Figure 2.7). Our model is for HDO of *p*-cresol while the experimental data is for *m*-cresol. *p*-cresol is investigated instead to lessen the computational burden as *p*-cresol's -OH and -CH<sub>3</sub> symmetry results in 64 less alcohol intermediates than those in *m*-cresol. Given the similarity of these isomers, no significant differences in reactivity are expected.

Overall, the model predictions are in very good agreement with the experimental results, indicating that the metal sites of Pt are sufficient to carry out the chemistry in the absence of an acid functionality. Interestingly, while methyl cyclohexanone is the dominant product in Nie's data, toluene dominates in Foster's data. This appears as a contradiction in experimental results. However, a major difference between these data is the space velocity. Careful examination of both the

data and model results indicates that methyl cyclohexanol and to a much lesser extent methyl cyclohexanone form early on in the reactor (short contact times) and they then convert to toluene downstream (longer contact times). The reduction of oxygenates downstream to toluene is typically taken in textbooks as a signature of reactions in series. If this were true, this will contradict experiments starting from either oxygenate, which do not give toluene under those conditions. We resolve this paradox below. Some differences in the distribution of cyclohexanone vs. cyclohexanol between data are seen but this may arise from sensitivity to particle size, catalyst precursors, and pretreatment conditions, as previous studies have shown.[120-124] These factors are not currently captured by our model. Good agreement is also observed when methyl cyclohexanol and methyl cyclohexanone are used as feeds (Figure 2.8). These data were obtained at short contact times relevant to those experimental data.

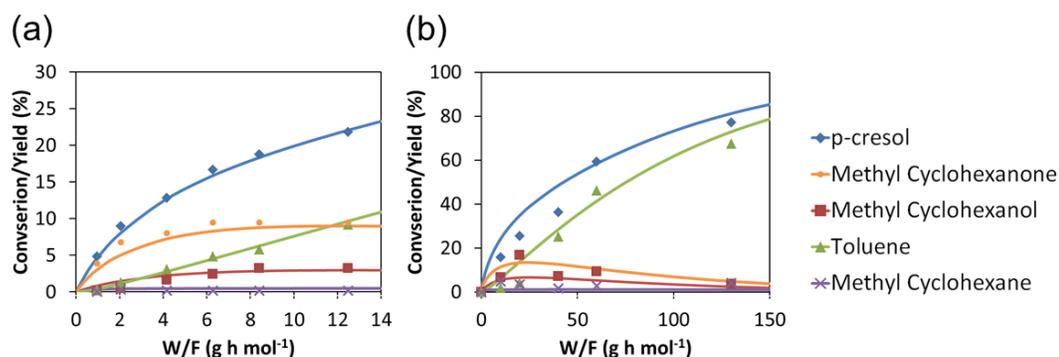


Figure 2.7. Comparison of model (lines) with (a) Nie et al.'s[62] and (b) Foster et al.'s experimental data[58] (solid symbols). Experimental conditions were (a)  $1.25 \times 10^{-2}$  atm *m*-cresol, 1.0 atm H<sub>2</sub> at 573 K on 1.0 wt% Pt/SiO<sub>2</sub>, and (b)  $7.5 \times 10^{-3}$  atm *m*-cresol, 0.5 atm H<sub>2</sub>, 0.5 atm N<sub>2</sub>, at 533 K on 1.55 wt% Pt/SiO<sub>2</sub>.

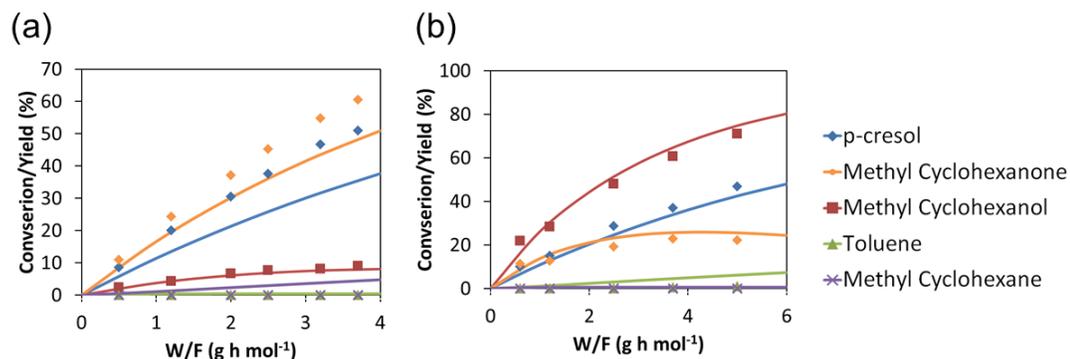


Figure 2.8. Comparison of model (lines) and experimental data[62] (solid symbols). Experimental conditions are  $1.25 \times 10^{-2}$  atm (a) methyl cyclohexanone and (b) methyl cyclohexanol, 1.0 atm  $H_2$  at 573 K on 1.0 wt% Pt/SiO<sub>2</sub>. The model agrees well with the experimental data.

Next we conduct reaction pathway analysis to understand the HDO mechanism (Figure 2.9). Here we compute forward, backward, and net rates and identify paths of high flux and paths that are equilibrated. The pathway leading to toluene starts with adsorption of *p*-cresol, and continues on with hydrogenation of 3-5 ring carbons (either in the alcohol and ketone hydrogenation network). Subsequently, C-OH bond scission occurs consistent with the energetics shown in Figure 2.5, followed by hydrogenation at the H-deficient  $\alpha$ -carbon, and dehydrogenations of the ring to produce adsorbed toluene, which subsequently desorbs to the gas phase. A critical aspect here is dehydroxylation does not necessarily happen from a stable compound but rather from an intermediate and thus building a full microkinetic model is essential. The adsorption of *p*-cresol is slightly irreversible whereas the alcohol and ketone network is partially-equilibrated. The derivatives, which are deficient in  $\alpha$ -C, hydrogenate irreversibly to the hydrocarbon network. Importantly, the oxygenates are

not stable intermediates from which toluene forms, i.e., the formation of toluene and oxygenates are not reactions in series. Rather, the oxygenates are products of parallel reactions that share common intermediates with the toluene network.

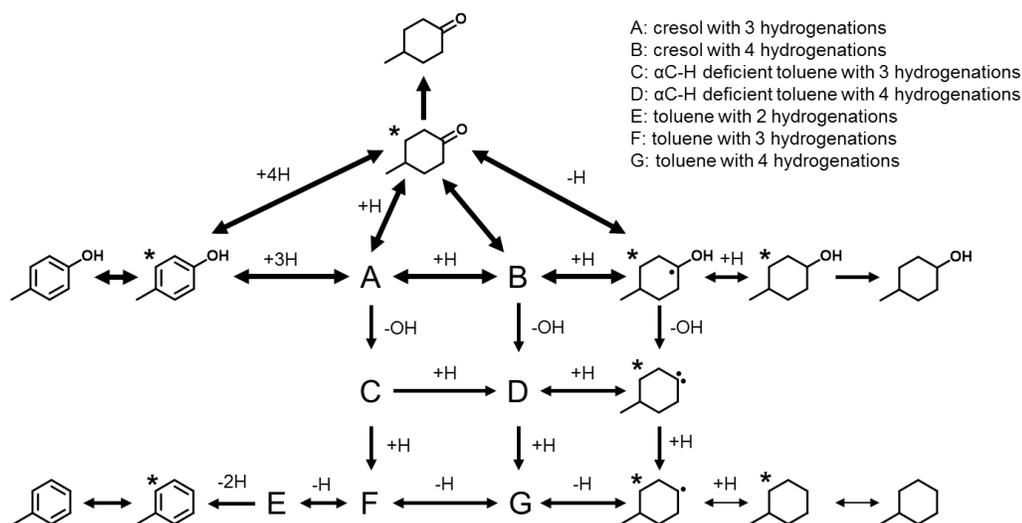


Figure 2.9. Reaction network for HDO of *p*-cresol to toluene at short contact times of  $W/F=2$  gh/mol for Nie et al.'s[62] experimental condition. Double (single) arrows represent reversible (irreversible) reactions. Thickness of the arrows indicates the magnitude of the net flux. The network is based on fluxes at the exit of the reactor.

Thermodynamics also plays an important role in driving the dehydroxylation reaction. The dehydroxylation reactions on Pt are very endothermic with an energy ranging from 1.6 to 0.6 eV, and with a low reverse reaction barrier due to the late transition state (Figure 2). Thus, it is natural to expect that the reverse reaction (oxygenation) would be fast. However, our reaction path analysis reveals that the dehydroxylation reactions are irreversible due to the low hydroxyl coverage on the

surface ( $10^{-13}$ ). This low coverage results from the thermodynamics highly favoring water formation.

The formation of oxygenated products, such as methyl cyclohexanol and cyclohexanone, is kinetically limited by desorption at short contact times and thermodynamically controlled at long contact times. As suggested by DFT results, desorption barriers (1.1-1.3 eV) are higher than the hydrogenation barriers ( $1.02 \pm 0.05$  eV). Combined with the high hydrogen coverage, hydrogenation is fast compared to desorption; the latter controls the formation of the saturated oxygenated products. Reaction path analysis at various locations of the reactor (not shown) indicates that at low space velocities, the formation of methyl cyclohexanol and methyl cyclohexanone is irreversible. In contrast, at longer contact times, the methyl cyclohexanol and methyl cyclohexanone reactions are in partial equilibrium and these compounds gradually convert through the partially hydrogenated intermediates to toluene because thermodynamics favors hydrocarbons over oxygenates under these conditions (see equilibrium composition in Figure A.16a). This explains the reduction in concentration of oxygenates at longer space velocities shown in Figure 2.7b. Our reaction network supports the concept of common intermediates being at play rather than reactions in series.

Finally, the hydrogenation and dehydrogenation reaction barriers of toluene derivatives determine the selectivity between toluene and methyl cyclohexane upon dehydroxylation. Equilibrium composition also favors toluene formation at higher temperatures, suppressing methyl cyclohexane formation at high space velocity. This indicates that cyclohexane formation is sensitive to the hydrogen pressure, and agrees

with experimental data for Pt/C[86, 125], where methyl cyclohexane is observed at high H<sub>2</sub> pressure (40 atm H<sub>2</sub>, and 82 atm H<sub>2</sub> respectively).

Compared to previously suggested mechanisms (Figure 2.1), our microkinetic results elucidate a new HDO mechanism for phenolic compounds. Direct dehydroxylation from any stable unsaturated with hydrogenated  $\alpha$ -carbon or saturated compound is not favored due to high barriers. Consistently, concerted dehydration of 3-methyl-3,5-cyclohexadienol on Pt(111) suggested by Resasco and co-workers[77] exhibits a high barrier of 1.58 eV. Our model indicates that methyl cyclohexanone and methyl cyclohexanol form in parallel with toluene formation, different from previous sequential hydrogenation mechanisms. In addition, the keto-enol tautomerization pathway, which was proposed before, and observed in aqueous phase to follow a concerted mechanism with a low barrier,[90] is not an important pathway for the HDO of *p*-cresol. Specifically, a concerted mechanism where H is transferred from OH to the nearby C is slow (we computed a barrier of 2.69 eV for hydrogen transfer). Instead, this intermediate can form in two sequential elementary steps: OH dehydrogenation and H addition to the nearby C. The barriers of the aforementioned reactions turned out to be 0.47 and 1.01 eV, in close agreement with results for *m*-cresol (0.40 and 0.80 eV).[77] Our microkinetic model accounts for these elementary steps early on in the network but indicates that these reactions are equilibrated and are not kinetically relevant.

The most abundant surface species are hydrogen, derivatives of *p*-cresol and toluene derivatives (A and F in Figure 2.9) with 3 hydrogenations (see Figure A.15 for detailed coverages). The hydrogen surface coverage, of up to 60%, facilitates hydrogenation of *p*-cresol derivatives toward saturated intermediates to promote ring

activation. The next most abundant species are intermediates with three hydrogenations with a total coverage up to 26%. Although hydrogenation is thermodynamically uphill, repulsive lateral interactions decrease with increasing level of hydrogenation as a result of the molecule being lifted off the surface (see Figure 2.6), resulting in increased stability of the hydrogenated intermediates. Recently, Sabbe et al.[126] have shown a downhill energy profile of benzene hydrogenation on Pt, as a consequence of co-adsorption with hydrogen. This result supports our observation that saturated molecules have higher coverage than the unsaturated molecules.

In summary, Pt alone is capable of performing HDO chemistry because of the following key factors: (1) Pt is an excellent hydrogenation catalyst. As hydrogenation proceeds, the ring lifts gradually off the surface, reducing the reaction energy and barrier for dehydroxylation; (2) frequent appearance of low barrier 4-centered TSS as the ring is being saturated; (3) high hydrogen coverage shifting the equilibrium towards saturated products on the surface; (4) reaction conditions favoring deoxygenated products. These factors lead to a drastically different mechanism in comparison with the existing pathways, where the dehydroxylation occurs after sufficient ring hydrogenation. Dehydroxylation may occur earlier (less ring hydrogenation) on more oxophilic metals, like Ni, Fe and Ru, which are expected to have lower dehydroxylation barrier to render it competitive with hydrogenation. However, oxygen removal is harder on these materials due to their lower hydrogenation ability.

## 2.6 Conclusions

We have analyzed the hydrodeoxygenation (HDO) reaction network of a prototypical phenolic compound, *p*-cresol on Pt catalyst. We enumerated all the possible reactions and parameterized the network consisting of four sequences of interconnected pathways. Our results are in excellent agreement with various experimental data on Pt/SiO<sub>2</sub>. In contrast to the general belief, and consistent with the experimental data, our results demonstrate that Pt alone can perform effective HDO upgrade of lignin-derived phenolic compounds. Interestingly, full hydrogenation of the ring is neither required nor is preferred. However, partial hydrogenation is necessary to activate the ring and lead to dehydroxylation. Similarly, dehydrogenation of the deoxygenated ring to toluene is thermodynamically favored over methyl cyclohexane. Our results indicate that the network prefers a much longer sequence of elementary reactions with an overall uphill and downhill energy landscape rather than the more intuitive paths through methyl cyclohexanol and/or methyl cyclohexane. Stable species such as methyl cyclohexanol and methyl cyclohexanone are not key intermediates, as suggested before.[127] Rather, they share common intermediates with the hydrocarbons. Due to thermodynamics, at longer reaction times, these species are consumed through the common pool of intermediates to hydrocarbons. Our results also provide insights into catalyst screening that can prove useful in future studies. For example, more oxophilic metals, such as Ni and Ru, may require less ring activation and thus lower hydrogen pressure while having a lower cost than noble metals, like Pt, due to the dehydroxylation being more exothermic and of lower reaction barrier.

## **2.7 Acknowledgements**

This research was supported by a USDA-NIFA-BRDI grant (No. 2012-10008-20271) via agreement with USDA-ARS (No. 59-1935-008). The authors acknowledge computational resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant OCI-1053575. In connection with XSEDE, the computations were performed on high performance computing clusters administered by the Texas Advanced Computing Center (TACC) and on the Kraken cluster administered by the National Institute for Computational Sciences (NICS). This research also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. G. G. is grateful to Dr. Matthew A. Christiansen, Dr. Jonathan Sutton, Vassili Vorotnikov, and Dr. Shengguang Wang for useful discussions.

## Chapter 3

### GROUP ADDITIVITY FOR THERMOCHEMICAL PROPERTY ESTIMATION OF LIGNIN MONOMERS ON Pt(111)

#### 3.1 Introduction

Biomass is receiving increasing attention as a promising replacement of fossil fuels for renewable energy and chemicals.[53] Typical biomass conversion involves depolymerization via hydrolysis or pyrolysis.[6] While the depolymerization is relatively facile, removal of extraneous oxygen remains a challenge in the production of petrochemical-like products. In this regard, the research community has focused on understanding hydrodeoxygenation (HDO),[128] a process involving high hydrogen pressures (1–100 bar) and temperatures (200–400 °C) to remove oxygen as water.[6] While there have been several theoretical studies on the HDO of depolymerized lignin monomers using first-principles,[77, 87, 90, 91, 93, 117, 129-133] it was only recently that the full reaction network was modeled through *ab-initio* and kinetic models[134] due to the complexity involved in the reaction network.

In order to rapidly parametrize microkinetic models, our research has focused on developing first-principles-based semiempirical or screening methods,[18, 33] wherein thermodynamic and kinetic parameters are estimated using adsorbate group additivity and Brønsted–Evans–Polanyi relationships, respectively. Combined with DFT refinement and uncertainty quantification, first-principles-based semiempirical methods are a powerful hierarchical approach to building and understanding complex reaction mechanisms.[19, 42, 44] Uncertainty quantification reveals that errors

propagated from the thermodynamic-property estimation have significantly more impact on the activity and selectivity prediction than those arising from the kinetic-parameter estimation.[18, 42, 44] Thus, the development of an accurate adsorbate group additivity scheme is key to building accurate kinetic models for large lignin derivatives.

Previously, we have developed BEP relationships for relevant elementary reactions,[117, 134] with the HDO mechanism of *p*-cresol being revealed through microkinetic modeling.[134] However, the full mechanism for the HDO of other derivatives, such as guaiacol, has yet to be elucidated due to the large reaction network involved. Here, we develop a group additivity scheme for lignin derivatives on Pt(111) surface, a material that exhibits a high activity and selectivity,[61, 135-137] and re-examine the framework of the previous adsorbate group additivity[22, 35, 36, 138] by means of the quantum theory of atoms in molecules (QTAIM). We reformulate the group definition in agreement with the theoretical background of the group additivity,[139] and demonstrate significant improvement over the previous method.

## 3.2 Methods

### 3.2.1 Density Functional Theory (DFT) Calculations

DFT calculations were performed using the Vienna ab initio Simulation Package (VASP), version 5.3.2.[140] Exchange and correlation energies are computed with the Perdew, Burke, and Ernzerhof (PBE) functional with the DFT-D3 dispersion correction of Grimme *et al.*[96, 97] The core electrons are treated with the projector augmented-wave function (PAW) method.[100, 101] We used a plane-wave basis set with a 400 eV kinetic energy cutoff for the valence electrons.

The lattice constant of Pt was optimized using the Birch–Murnaghan equation of state,[102, 103] with a  $15 \times 15 \times 15$  Monkhorst-Pack  $k$ -point mesh.[104, 141] The lattice constant of Pt was calculated to be  $3.917 \text{ \AA}$ , which is in close agreement with the experimental value of  $3.92 \text{ \AA}$ .[105] The Pt slab was modeled with a  $(4 \times 4)$  unit cell using four layers. The bottom two layers were held fixed at the bulk position, whereas the top two layers were allowed to relax. A  $20 \text{ \AA}$  vacuum was inserted to minimize the interaction between the supercells normal to the surface. The Brillouin zone was integrated using a  $5 \times 5 \times 1$  Monkhorst-Pack  $k$ -mesh,[104] with a Methfessel–Paxton smearing of  $0.1$ .[106] The adsorbate-slab system was relaxed using quasi-Newton and conjugate-gradient algorithm with a  $0.05 \text{ eV/\AA}$  convergence threshold.

The chemical bonds were subsequently analyzed using Bader’s QTAIM[142] using the *critic2* software package.[143] In order to ensure convergence of the bond critical point (BCP) properties, the integration grid was increased by a factor of 5 ( $0.03 \text{ \AA}/\text{grid point}$ ). The valence charge density is employed for BCP analysis as it is located outside of the pseudopotential core region. All BCPs are converged within  $|\nabla\rho| = 10^{-12} \text{ e}^-/\text{r}_{\text{bohr}}^4$ .

Thermodynamic properties ( $\Delta H_{f,298}^o$ ,  $\Delta S_{298}^o$  and  $C_{p,T}^o$ ) were computed with the DFT energetics and vibrational frequencies using statistical mechanics (see ref[35] for details), followed by referencing to the standard state using the method described in ref.[35]

### 3.2.2 Group Additivity

Utilization of group theory provides an estimate for the various thermodynamic properties by summing up contributions from “groups,”

$$\Phi = \sum_{i=1}^{N_{\text{groups}}} \nu_i \varphi_i \quad (3.1)$$

where  $\Phi$  is thermodynamic property of a molecule,  $\nu_i$  is number of occurrence of group  $i$ , and  $\varphi_i$  is group  $i$ 's contribution to the thermodynamic properties. Parameterization of the  $\varphi_i$  terms involves a linear regression,

$$\varphi_{(N \times 1)} = \mathbf{v}_{(M \times N)}^{-1} \Phi_{(M \times 1)} \quad (3.2)$$

where  $\varphi$  is the group thermodynamic property vector,  $\mathbf{v}$  is the group coefficient matrix,  $\Phi$  is the molecular thermodynamic property vector, and  $M$  and  $N$  are the total number of data points and groups in the model, respectively. Linear regression of the group contributions involves a rank-deficient  $\mathbf{v}$  matrix due to a linear dependency between the groups. Originally, the linear dependency was lifted by assigning the contribution of a dependent group with those of a similar group.[144] For convenience, we applied the Moore-Penrose pseudoinverse of  $\mathbf{v}$  to assign a “minimum norm solution” to all the dependent groups (see ref[145] for details).

### 3.2.3 Training Set Sampling

We define our population set as all the reaction intermediates in the HDO of lignin monomers. Previous studies of bio-oil have shown lignin monomers consist of a benzene ring with different combinations of three substituents: methoxy (-OCH<sub>3</sub>), hydroxyl (-OH), and alkyl (-CH<sub>2</sub>)<sub>x</sub>CH<sub>3</sub>.[146] Multiring dimers are not observed as the dimer linkages are readily broken under pyrolysis conditions;[147] as such we

focused on elucidating the C-OH[134] and C-OCH<sub>3</sub> scission[117] mechanisms. These studies demonstrate the C-OH scission requires partial ring hydrogenation, and C-OCH<sub>3</sub> scission requires dehydrogenation of the methoxy. On the basis of these observations, the reaction intermediates were enumerated with the following elementary reactions: (1) C<sub>ring</sub>-H formation, (2) C-OH scission, (3) C<sub>methoxy</sub>-H scission (dehydrogenation all the way down to 1 H), (4) C-OCH<sub>x</sub> scission, (5) O-CH<sub>x</sub> scission, and (6) O-H scission. Guaiacol and its 3,4,5,6-methyl derivatives are employed as lignin model molecules as their reaction network includes those of other lignin monomers observed in bio-oil. Reaction network generation was performed using the R.I.N.G. software package,[148, 149] resulting in a total of 6820 species (14036 intermediates when different adsorption conformations are distinguished) and 19153 elementary reactions.

We develop an improved group additivity method that accounts for the adsorption conformation of the adsorbates. The training set consists of a total of 591 data points with 59 intermediates obtained from previous studies.[117, 134] As a result, 201 calculations were performed to build the adsorption-conformation library for the lignin monomers. We also included an additional 331 randomly selected samples generated by R.I.N.G. Structures (CONTCAR and xyz format), energetics, and vibrational frequencies of the training set are provided in the supporting information of Gu et al.[21]

### 3.3 Result and Discussion

#### 3.3.1 Adsorption Conformations of Bio-oil Lignin Monomers on Pt(111)

For the lignin monomers, the adsorption conformations are complex due to the various hydrogenation combinations of the ring. To analyze the conformations systematically, we classify each conformation based on the type of binding of the binding (central) atom, as shown in Figure 3.1. The type of binding captures the energetics in part; strain induced from the nearest neighbors needs also to be accounted for (see below). Examination of the binding reveals a unique set of adsorption conformations given the location of the atoms available for binding to the surface as well as the number of valence electrons of atoms interacting with the surface. Adsorption conformations of the methoxy group ( $-\text{OCH}_x$ ,  $1 < x < 3$ ) can be decoupled from those of the rest of the adsorbate due to the bridging oxygen. Nonbinding substituent groups, such as  $=\text{O}$ ,  $-\text{OH}$ ,  $-\text{OCH}_3$ , and  $-\text{CH}_3$ , do not affect these conformations, indicating that a change in the hybridization and the steric hindrance of the substituent groups is not strong enough to affect the adsorption conformations. In this fashion, a total of 201 ring conformations were identified; the chemical drawings of all conformations are shown in Appendix B. Within all the ring adsorption conformations found, the face-centered cubic and hexagonal close-packed variation affect  $\Delta H_{f,298}$  by 0.68 kcal/mol on average, and 3.42 kcal/mol in maximum with higher deviation seen for the unsaturated molecules likely due to strain.

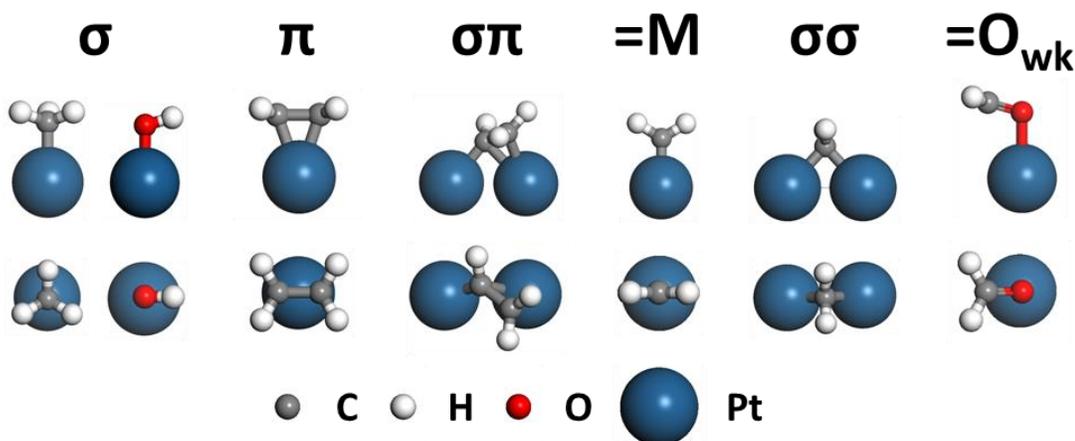


Figure 3.1. Local binding sites observed in lignin monomers on Pt(111). M represents metal atom. Top (bottom) row shows the side (top) view of the binding site.

### 3.3.2 Conjugation-Character Analysis

Our previous surface group additivity scheme[22, 35, 36] was built under the assumption that the aromaticity and the conjugation of the adsorbates are completely lost upon adsorption; thus the bonds of surface bound atoms were treated as a single bond in the gas phase. However, the molecular orbitals of an adsorbate split into bonding and antibonding orbitals upon interaction with the  $d$ -band of the metal. Therefore, we expect the C–C bond in the gas phase involving only  $s$ - and  $p$ -orbitals to be different from the C–C bond on the surface where each carbon is adsorbed in a  $\sigma$  configuration. Specifically, ethylene adsorption on Ni(111)[150] and Rh(111)[151] shows only a portion of the electrons are donated to the surface from the  $\pi$  orbital, and *vice versa* with following electron back-donation from the surface to the  $\pi^*$  orbital. Additionally, the molecular orbitals from benzene adsorption on Pt(111)[45] and ethylene adsorption on Ag(001),[152] Cu(110)[153] and Ni(11)[153] show the lobe of

$p_z$ -orbitals (specifically, HOMO/ $\pi$ ) pointing toward the surface is localized on the surface metal atom(s), while the other lobe pointing normal to the surface remains conjugated (when they are in-phase).[154] This indicates that a “single bond” between two adsorbed atoms is not the same as a single bond in the gas phase.

While previous DFT studies show that the chemical environment of adsorbates differs from that in the gas phase, it is not clear how it affects the bond classification in group additivity. The virial theorem allows for the spatial partitioning of the electronic energy that has previously been used to understand group additivity using DFT.[139] The energy of each atom (or group of atoms) in the system is computed by integrating the energy density within the Bader volume. Comparison of the partitioned atomic energy with the regressed group additivity values reveals the partitioned energy of a group changes depending on the connecting groups; however the regressed group energy is conserved as the partitioned energy of the connected group changes in an equal and opposite amount. Furthermore, the energy of a group is transferable to other molecules as the charge distribution within a Bader volume is identical between different molecules. Finally, the electron delocalization does not affect the energy of a group as delocalized electrons from other groups and the local electrons sum up to construct an identical charge distribution, regardless of where the delocalized electron(s) comes from.[139, 155] While more accurate energy partitioning methods have recently been introduced,[156, 157] it is clear that the electron distribution within the groups needs to be consistent. Therefore, we investigate bonds in adsorbates by means of the QTAIM bond critical point to quantitatively analyze the charge topology in order to correctly categorize homologous bonds for the application of group additivity.

In Table 3.1, the properties of the electronic density at the BCPs for the binding configurations in Figure 3.1 are compared to those of gas phase ethane, ethylene, acetylene, and benzene. The magnitude of electronic density ( $\rho$ ) and the Laplacian of the electronic density ( $\nabla^2\rho$ ) are related to the strength of the bond, and the sign of  $\nabla^2\rho$  indicates whether the bond is covalent (negative) or ionic/van der Waals (positive). Furthermore, we compute the ellipticity, which is defined as the anisotropy of the electron density orthogonal to the bond direction, that is, the  $\pi$  characteristics of the bond,

$$\varepsilon = \frac{\lambda_1}{\lambda_2} - 1 \quad (1)$$

where  $\lambda_1$  and  $\lambda_2$  are the lowest and second lowest Hessian eigenvalues of the electron density. Table 3.1 shows the  $\varepsilon$  of ethane is 0 while the  $\varepsilon$  of ethylene and benzene are 0.30 and 0.17, respectively. Thus, the conjugation is nonexistent for a single bond (ethane), is stronger for a double bond (ethylene), and is weaker for a delocalized bond (benzene). The  $\varepsilon$  of acetylene is 0 as the two  $\pi$  orbitals are orthogonal to each other giving rise to a circular anisotropic geometry. The bond order,  $n$ , can be computed as a function of  $\rho$ , [158, 159]

$$n = e^{A(\rho-B)} \quad (2)$$

where  $A$  and  $B$  are fitted using ethane, ethylene, and acetylene as our standards. While the fitted bond order is empirical, it provides a convenient measure to compare the electron density at the BCP with those of the standard gas-phase molecules.

Table 3.1. BCP analysis between two carbons in gas phase and on the surface. C2 and C3 molecules are used where valence is capped with hydrogen. Subscription of C indicates the binding-site type. Structures are shown in Figure B.6 in the appendix.

Phase	Bond Type	$\rho$ ( $e^-/r_{\text{bohr}}^3$ )	$\nabla^2\rho$ ( $e^-/r_{\text{bohr}}^5$ )	$\epsilon$	$n$
Gas	C-C (ethane)	0.242	-0.61	0.00 <sup>3</sup>	1.00
	C=C (ethylene)	0.354	-1.35	0.30 <sup>3</sup>	2.00
	C≡C (acetylene)	0.420	-1.58	0.00 <sup>3</sup>	3.00
	C-C (benzene)	0.316	-1.12	0.17 <sup>3</sup>	1.58
Adsorbed	C <sub>σ</sub> -C <sub>σ</sub>	0.267	-0.78	0.06	1.17
	C <sub>σ</sub> -C <sub>σσ</sub>	0.276	-0.82	0.03	1.23
	C <sub>σ</sub> -C <sub>π</sub>	0.278	-0.85	0.08	1.25
	C <sub>σ</sub> -C <sub>=M</sub>	0.287	-0.92	0.01	1.32
	C <sub>σ</sub> -C <sub>σπ</sub>	0.278	-0.83	0.05	1.25
	C <sub>σ</sub> -C <sub>πσ</sub>	0.263	-0.73	0.06	1.14
	C <sub>σ</sub> -C	0.252	-0.67	0.01	1.06
	C <sub>σσ</sub> -C <sub>σσ</sub>	0.297	-0.96	0.07	1.41
	C <sub>σσ</sub> -C <sub>π</sub>	0.278	-0.85	0.05	1.25
	C <sub>σσ</sub> -C <sub>=M</sub>	0.330	-1.17	0.13	1.72
	C <sub>σσ</sub> -C <sub>σπ</sub>	Unstable			
	C <sub>σσ</sub> -C <sub>πσ</sub>	Unstable			
	C <sub>σσ</sub> -C	0.258	-0.72	0.02	1.10
	C <sub>π</sub> -C <sub>π</sub> <sup>1</sup>	0.283	-0.88	0.06	1.28
	C <sub>π</sub> -C <sub>π</sub> <sup>2</sup>	0.305	-1.04	0.15	1.48
	C <sub>π</sub> -C <sub>=M</sub>	0.289	-0.94	0.05	1.34
	C <sub>π</sub> -C <sub>σπ</sub>	0.276	-0.82	0.03	1.24
	C <sub>π</sub> -C <sub>πσ</sub>	Unstable			
	C <sub>π</sub> -C	0.257	-0.72	0.03	1.10
	C <sub>=M</sub> -C <sub>=M</sub>	0.356	-1.33	0.25	2.02
	C <sub>=M</sub> -C <sub>σπ</sub>	0.317	-1.10	0.07	1.59
	C <sub>=M</sub> -C <sub>πσ</sub>	0.264	-0.70	0.09	1.14
	C <sub>=M</sub> -C	0.271	-0.82	0.03	1.19
	C <sub>σπ</sub> -C <sub>σπ</sub>	Unstable			
	C <sub>σπ</sub> -C <sub>πσ</sub> <sup>1</sup>	0.275	-0.81	0.03	1.22
	C <sub>σπ</sub> -C <sub>πσ</sub> <sup>2</sup>	0.299	-0.97	0.13	1.42
	C <sub>σπ</sub> -C	0.255	-0.70	0.02	1.08
	C <sub>πσ</sub> -C <sub>πσ</sub>	Unstable			
	C <sub>πσ</sub> -C	0.255	-0.70	0.02	1.08

<sup>1</sup> At separate sites. <sup>2</sup> At the same site. <sup>3</sup> These values are in agreement with the previous studies.[158-160]

Overall,  $\varepsilon > 0$  and  $n > 1$  suggest the bonds between the adsorbed carbons are partially conjugated, signifying the adsorbate C–C bonds are different from those of the gas phase. The majority of the adsorbate’s  $\varepsilon$  and  $n$  are between those of the gas-phase ethane and benzene, indicating the presence of a weak conjugation with the magnitude of the conjugation being dependent on the overlap of the  $p$ -orbital lobes. For example, the conjugation character of the  $C_{\sigma\sigma}$ – $C_{=M}$  bond is strong ( $\varepsilon = 0.13$ ,  $n = 1.72$ ) as the molecular orbital of the  $C_{\sigma\sigma}$  atom, which binds carbon to the two Pt, aligns with the  $p$ -orbital of the  $C_{=M}$  atom, which double-bonds to the Pt (see  $C_{\sigma\sigma}$ – $C_{=M}$  in Figure B.6) Alternatively, the  $C_{\sigma}$ – $C_{=M}$  bond shows minimal conjugation ( $\varepsilon = 0.01$ ,  $n = 1.32$ ) as the  $p$ -orbitals do not align (see  $C_{\sigma}$ – $C_{=M}$  in Figure B.6). The C–C bond in the  $\pi$  and  $\sigma\pi$  local binding sites (see Figure 3.1) shows a fairly strong conjugation comparable to the benzene C–C bond (see  $C_{\pi}$ – $C_{\pi}^2$  and  $C_{\sigma\pi}$ – $C_{\pi\sigma}^2$  in Table 3.1). This is expected as the  $\pi$ -orbital of the C–C bond is retained and interacts with the surface. By taking the average of  $\varepsilon$ , the conjugation strength in increasing order is:  $C < C_{\sigma} < C_{\sigma\pi} < C_{\sigma\sigma} < C_{\pi} < C_{\pi\sigma} < C_{=M}$ . This ordering is justified as the  $p$ -orbital character becomes more discernible going from left to right (from C being  $sp^3$  hybridized to  $C_{\pi}$  with a retained  $\pi$ -orbital to  $C_{=M}$  with a double bond). The conjugation of the  $C_{=M}$ – $C_{=M}$  bond closely resembles the C=C bond of ethylene, suggesting the C=M bond is likely the C–M bond, and the M–C=C–M configuration is more stable than M=C–C=M. Our analysis indicates that categorizing bonds based on the binding type is important for adsorbate group-additivity as the conjugation strength is site sensitive.

Next, we explore the effect of the binding of nearest neighbor atoms on the conjugation to account for possible strain effects mentioned above. Table 3.1 shows that the C–C bond in benzene exhibits less conjugation than in ethylene. For ethylene,

the  $p$ -orbital has another  $p$ -orbital with which to conjugate, whereas for benzene each  $p$ -orbital has two  $p$ -orbitals to conjugate with, allowing electron delocalization. This delocalization effect for adsorbates is investigated in the case of  $\sigma$  binding of nearest neighbor(s) in Table 3.2. Overall, the degree of the C–C bond conjugation is dependent on the binding of the nearest neighbors. Specifically, electrons may localize or delocalize depending on the binding type. For example, conjugation of the  $C_\sigma$ – $C_\sigma$  bond is weakened in the presence of  $C_\sigma$  neighbors ( $\varepsilon = 0.06, 0.05, 0.05$  for 0, 1, 2  $C_\sigma$  neighbors), which indicates delocalization. The opposite trend is observed for the  $C_\pi$ – $C_\pi$  bond ( $\varepsilon = 0.14, 0.15, 0.16$  for 0, 1, 2  $C_\sigma$  neighbors) with a similar trend for the  $C_{\sigma\pi}$ – $C_{\pi\sigma}$  bond. Charge-flow analysis (shown in Figure B.8 in the appendix) suggests that the  $\pi$  electrons from the neighboring  $C_\sigma$  are donated to the  $\pi$  system of the  $C_\pi$ – $C_\pi$ , which strengthens the conjugation. For the  $C_\sigma$ – $C_{\sigma\sigma}$  bond, the conjugation is strengthened with  $C_\sigma$  neighbor attached at the  $C_\sigma$  side, and is weakened when attached at the  $C_{\sigma\sigma}$  side. Therefore, electron (de)localization is dependent on the binding type and further supports the inclusion of type-specific terms into the group additivity framework. Additionally, the binding of the nearest neighbor changes the geometrical alignment of the  $p$ -orbital by inducing strain. For example, the  $C_\sigma$ – $C_{=M}$  bond shows a minimal conjugation without binding neighbor, but a strong conjugation with a  $C_\sigma$  neighbor attached at  $C_{=M}$ . This is attributed to the  $C_\sigma$  neighbor straining the  $C_{=M}$  carbon to lay flat, resulting in a better alignment of the  $p$ -orbital as shown in (C) $C_\sigma$ – $C_{=M}(C_\sigma)$  in Figure B.7. Thus, the nearest neighbors' effect is not only due to the binding type dependence of the electron (de)localization, but also to the geometric alignment of the  $p$ -orbital.

Table 3.2. C-C BCP dependence on the binding of the nearest neighboring carbon. (C) represents a  $-\text{CH}_3$  group, and  $(\text{C}_\sigma)$  represents a  $-\text{CH}_2$  group adsorbed on a  $\sigma$  site. Structures are shown in Figure B.7.

Bond	$\rho$ ( $e^-/\text{r}_{\text{bohr}}^3$ )	$\nabla^2\rho$ ( $e^-/\text{r}_{\text{bohr}}^5$ )	$\varepsilon$	$n$
(C) $\text{C}_\sigma$ - $\text{C}_\sigma$ (C)	0.261	-0.73	0.06	1.12
(C) $\text{C}_\sigma$ - $\text{C}_\sigma$ ( $\text{C}_\sigma$ )	0.270	-0.77	0.05	1.19
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_\sigma$ ( $\text{C}_\sigma$ )	0.271	-0.77	0.05	1.19
(C) $\text{C}_\sigma$ - $\text{C}_\pi$ ( $\text{C}_\pi$ )	0.274	-0.81	0.07	1.22
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_\pi$ ( $\text{C}_\pi$ )	0.276	-0.83	0.09	1.23
(C) $\text{C}_\sigma$ - $\text{C}_{\sigma\sigma}$ (C)	0.270	-0.77	0.02	1.19
(C) $\text{C}_\sigma$ - $\text{C}_{\sigma\sigma}$ ( $\text{C}_\sigma$ )	0.264	-0.73	0.01	1.15
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_{\sigma\sigma}$ (C)	0.267	-0.76	0.05	1.17
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_{\sigma\sigma}$ ( $\text{C}_\sigma$ )	0.261	-0.73	0.06	1.12
(C) $\text{C}_\sigma$ - $\text{C}_{=M}$ (C)	0.281	-0.87	0.02	1.27
(C) $\text{C}_\sigma$ - $\text{C}_{=M}$ ( $\text{C}_\sigma$ )	0.281	-0.85	0.08	1.27
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_{=M}$ (C)	0.285	-0.89	0.04	1.31
( $\text{C}_\sigma$ ) $\text{C}_\sigma$ - $\text{C}_{=M}$ ( $\text{C}_\sigma$ )	0.290	-0.91	0.10	1.34
(C) $\text{C}_{\sigma\pi}$ - $\text{C}_{\pi\sigma}$ (C)	0.286	-0.86	0.07	1.31
( $\text{C}_\sigma$ ) $\text{C}_{\sigma\pi}$ - $\text{C}_{\pi\sigma}$ (C)	0.287	-0.88	0.11	1.32
(C) $\text{C}_{\sigma\pi}$ - $\text{C}_{\pi\sigma}$ ( $\text{C}_\sigma$ )	0.296	-0.96	0.12	1.39
( $\text{C}_\sigma$ ) $\text{C}_{\sigma\pi}$ - $\text{C}_{\pi\sigma}$ ( $\text{C}_\sigma$ )	0.294	-0.94	0.14	1.38
(C) $\text{C}_\pi$ - $\text{C}_\pi$ (C)	0.304	-1.02	0.14	1.47
(C) $\text{C}_\pi$ - $\text{C}_\pi$ ( $\text{C}_\sigma$ )	0.301	-0.99	0.15	1.43
( $\text{C}_\sigma$ ) $\text{C}_\pi$ - $\text{C}_\pi$ ( $\text{C}_\sigma$ )	0.291	-0.92	0.16	1.35

### 3.3.3 Group-Additivity-Model Description

Previous group additivity models used a “valence-based scheme” where the number of valence electrons of the atom forming a bond with the surface atom(s) was specified in each group.[35] For example, the  $\text{C}(\text{C})(\text{H})_2(\text{M})$  group in prior work would represent carbon adsorbed in any of the three different binding modes:  $\sigma$ -site,  $\pi$ -site,

and  $\sigma\pi$ -site, as shown in Figure 3.1. This method reduces the number of fitted parameters, but the binding-energy difference arising from the type of binding mode is not accounted for.

On the basis of the discussion above, we introduce two new group additivity schemes to account for the dependence on the binding type, as summarized in. The major difference between the two models is the strategy employed in defining the group centers and their nearest neighbors. The BCP analysis has shown that the degree of bond conjugation depends on the binding type of the two atoms, as well as the conjugation of the nearest neighbors and their binding mode. Therefore, we derive a “site-based” model that resolves the type of bond of both the group center and the nearest neighbors. This model requires a high number of fitted parameters due to its higher spatial resolution. In order to reduce the number of fitted parameters, we also introduce a “conjugation-based” model which lumps some of the nearest neighbor terms. Specifically, while  $C_{=O}$  and  $C_{=M}$  both have full double bond orbital available for conjugation, only a portion of the  $p$ -orbital is available for the  $C_{\sigma}$ ,  $C_{\sigma\sigma}$ ,  $C_{\pi}$ , and  $C_{\sigma\pi}$  sites. Thus, those terms are lumped as a single type  $C_{pp}$ , which indicates that a portion of the  $p$ -orbital is available for conjugation.

Table 3.3. Level of site description in prior and new group additivity schemes.

	Valence-based <sup>1</sup>	Conjugation-based	Site-based
Group center <sup>2</sup>	Element and hybridization	+ Site resolved	+ Site resolved
Nearest Neighbor <sup>2</sup>	Element and hybridization	+ Lumped based on binding mode $C_{\sigma}, C_{\sigma\sigma}, C_{\pi}, C_{\sigma\pi} \rightarrow C_{pp}$	+ Site resolved

<sup>1</sup> Groups only describe number of valence electrons interacting with the surface.

<sup>2</sup> For a group, A(B)<sub>x</sub>(C)<sub>y</sub>(D)<sub>z</sub>, A represents the group center, and B, C, D represent nearest neighbors.

### 3.3.4 Model Assessment

Following parameter regression, we compare the group additivity estimates with the DFT computed values in Figure 3.2a. We focus on  $\Delta H_{f,298}$  as it is the largest source of error for the computation of the kinetically relevant  $\Delta G_{f,298}$  ( $\Delta S_{298}$  and  $C_p$  contribute minimally to the  $\Delta G_{f,298}$  error[35]). For model validation (Figure 3.2b), we adopt the leave-one-out cross-validation method.[161] In this method, the  $\Delta H_{f,298}$  regression is performed without one datum which is then estimated and compared with the real  $\Delta H_{f,298}$ . This omission–regression–estimation process is repeated for all the data points to provide a better assessment of the model than the regression alone. Additionally, this approach has an advantage over the conventional validation method where the data set is partitioned to 70% training set and 30% validation set: because the most amount of data possible is used to regress the model parameter, the error attributed to the lack of data points is minimized.

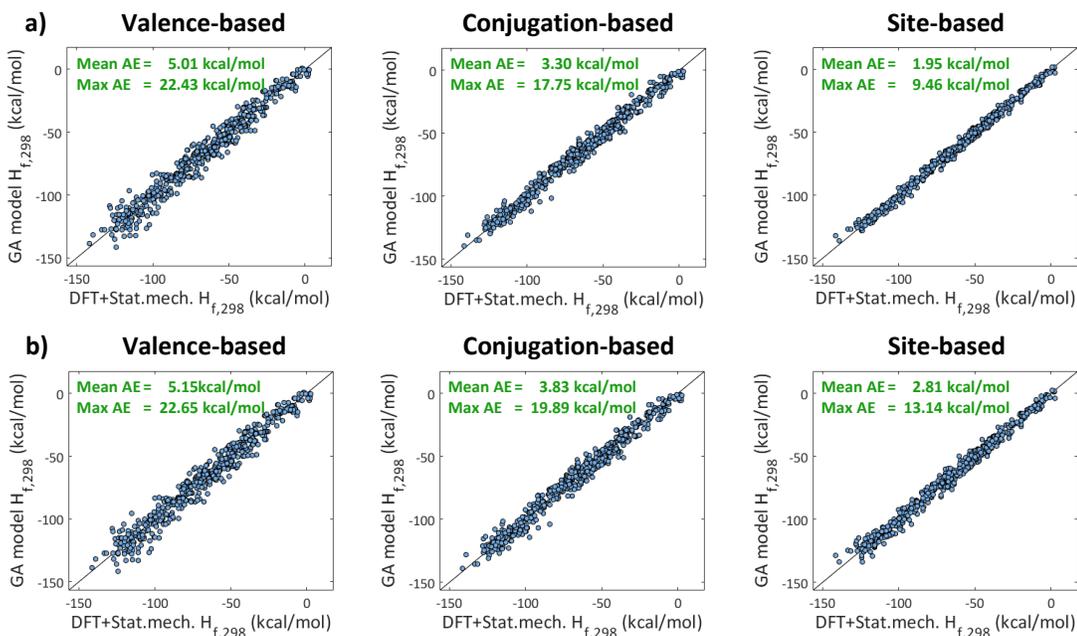


Figure 3.2. Parity plots for group-additivity computed versus DFT-estimated  $\Delta H_{f,298}$  for (a) regression and (b) validation. AE stands for absolute error.

The valence based model performs poorly, with a mean absolute error (AE) of 5.15 kcal/mol. The conjugation- and site-based models perform significantly better. With the description of binding type of both the group center and its nearest neighbor(s), conjugation is described better. The mean AE disparity between the regression and validation suggest overfitting of the model. The valence-based model shows the smallest disparity (0.21 kcal/mol) as only 17 parameters are regressed, followed by the conjugation-based model (0.52 kcal/mol) with 95 regressed parameters (rank = 83), and the site-based model (0.86 kcal/mol) with 202 regressed parameters (rank = 169). However, given the same amount of data, overfitting is expected to be the highest with the number of parameters is largest, and despite the

overfitting, the site-based model remains the best model as the mean and max AE from the cross-validation are the lowest. Prior surface models invoked various corrections, such as weak H–M interaction, furan ring deformation, adsorbed furan ring, and  $C(M)_2-C(M)-C(M)$  and ring strain[35, 36] to correlate the data well. These are no-longer necessary as information about the nearest neighbors accounts for the physical phenomena described by these corrections. We introduced only one correction for the excess strain observed when a methoxy group is bound to the surface, while the rest of the adsorbate remains in a certain adsorption conformation. This correction is discussed in detail in Appendix B.

### 3.3.5 Strain Description

The group additivity method often falls short for strained systems. For gas phase molecules, strain-corrections are assigned for the various ring sizes.[144, 162] For adsorbates, strain is complex as multiple rings are created with multiple binding sites. Our previous model empirically accounted for strain by considering a ring formed by two surface binding atoms (e.g.  $M-C-C-C-M$ , and  $M-C-C-M$ ).[36] This method is satisfactory for linear aliphatics where the adsorbate involves one strained ring,[22, 36] but, for the multidentate ring molecules, such as benzene, the strain of the six-carbon ring is often affected by other strained rings formed by the carbons and surface atoms; thus computation of ring-strain would require knowledge about multiple strained angles.

Here, we show that the site-based model captures the strain effect the best. Previous strain studies described the angle-strain as a function of the angle of the central atom and two neighbor atoms making up the angle.[144, 163-166] In terms of group additivity, this means the strain energy would be captured by each group better

if the angles between the group center and the nearest neighbors are closer to constant among the species in the training set as the group already captures element and hybridization. To assess this, we plot a histogram of the deviations between each angle in the training set and its average in each group, as shown in Figure 3.3. The site-based model gives the most narrow-distribution, that is, the most consistent angle description in each group. The deviation is due to the strain induced by other groups in the adsorbate. The largest deviations are observed for  $C_{\sigma\sigma}(C)_2$  and  $C_{=M}(C)_2$  groups. These groups' two nearest neighbors point in opposite direction of each other (see  $C_{\sigma\sigma}$  and  $C_{=M}$  in Figure 3.1) but the two nearest neighbors must curve inward in order to enclose the six-carbon ring, and, as a result, these group centers are sensitive to the location of the next nearest surface binding atoms. On the other hand,  $C_{\sigma}$ ,  $C_{\pi}$ , and  $C_{\sigma\pi}$  provide naturally inward curved substituent directions (see Figure 3.1) and are less sensitive to strain induced from other groups.

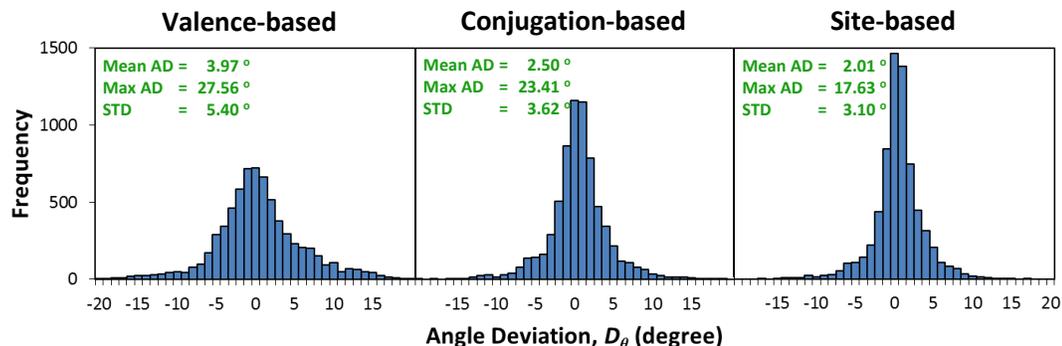


Figure 3.3. Deviation of bond angles in each group from their average. The more narrow the distribution, the more alike the angles are for a group among all species in the training set.

### **3.4 Conclusions**

We developed an adsorbate group additivity scheme for bio-oil lignin monomers and single-ring aromatic hydrocarbons on Pt(111). Using the reaction network of HDO of guaiacol as an example, we defined an ensemble of 591 adsorbates for DFT calculations and subsequent regression of group values. Using QTAIM, we were able to define bonds in the adsorbates in accordance with the theoretical backbone of the group additivity. We have found that the binding type of adsorbates with surface sites not only affects the binding strength, but also the conjugation character of the bonds in the adsorbates. We have introduced a model (site-based) where the groups are defined based on the binding type of the group centers and of their first nearest neighbors with surface sites. In a simpler version of this model (conjugation-based model), some of the nearest neighbors are lumped to a single group (partial conjugation). Both models show significant improvement over previous models in accurately describing the thermochemistry of adsorbates. Other corrections invoked in prior work are no-longer necessary. The group additivity can be used to build large microkinetic models at significantly lower cost followed by the DFT refinement and to estimate thermodynamics on different mono-/bi-metallic surfaces using linear scaling relationships.

### **3.5 Acknowledgements**

This research was supported by the NSF under grant number CHE-1415828. The authors acknowledge computational resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant OCI-1053575. In connection with XSEDE, the computations were performed on high performance computing clusters administered

by the Texas Advanced Computing Center (TACC) and on the Kraken cluster administered by the National Institute for Computational Sciences (NICS). G. G. is grateful to Dr. Glen R. Jenness for the paper writing improvement.

## Chapter 4

### GROUP A DDITIVITY FOR AQUEOUS PHASE THERMOCHEMICAL PROPERTIES OF ALCOHOLS ON PT(111)

#### 4.1 Introduction

Biomass processing occurs often in water. Aqueous phase reforming is one of the processes and converts polyols and sugars in water to H<sub>2</sub> and CO<sub>2</sub> at relatively low temperatures under pressure, as popularized by Dumesic and co-workers.[167] An alternative large scale use of biomass conversion is bio-oil production via pyrolysis, which contains 15-30% water[6], and which needs to be further upgraded via hydrodeoxygenation. Here solvation effects are important due to the deactivation of well-studied NiMo sulfides in aqueous environment.[168] Another alternative is the selective transformation of sugar derivatives to largely used monomers such as acrylic or adipic acid.[169, 170] All of these processes involve hydrogenations, C–C and C–OH bond breakings in the presence of water, which impose different conditions than for the current petro-chemistry. As demonstrated by several experimental studies, the presence of water is not innocent, even beyond catalyst stability, driving the need for new catalytic systems.[13, 171-173]

The difference between chemistry in the gas- and solution-phase can be dissected into several distinct effects: (1) solubility influences the concentration of reactants, intermediates and products and hence impacts the kinetics even if the rate constants would not change compared to the gas-phase ; (2) the solvent can participate directly in the reaction, either as a reactant or as co-catalyst (e.g., in relay mechanisms

of proton transfers); (3) the averaged interaction between the solvent and the solute changes the energetics compared to the gas-phase, stabilizing large dipole moments and charge-separations (heterolytic bond dissociations); (4) entropic contributions arise as the available phase-space can be significantly different in solution compared to the gas-phase, both in terms of translation/rotations and accessible configurations; and (5) the occurrence of homogeneous reactions in solvent phase such as dehydration and aldol condensation.[174]

Despite the wide use of solvent in biomass conversion, only a limited number of theoretical studies have explicitly addressed solvation effects.[39-41, 90, 175-179] This shortcoming is due to the methodology: there are no computationally affordable, broadly validated and general methods to include solvent effects in heterogeneous catalysis. The most rigorous way to describe reactivity in condensed phases is to perform thermodynamic integration at the *ab initio* molecular dynamics (AIMD) level.[180, 181] However, these computations are prohibitively expensive, as equilibration of the solid-liquid interface and convergence of the free energy would require more than 100 ps of simulations,[182] but today's computing power only allows to obtain about 10 ps for metallic surfaces with a reasonable effort.[183] Therefore, different approximations have been proposed: inclusion of only a couple of solvent molecules (called micro-solvation),[184, 185] adsorption of ice-like water layers on metallic surfaces,[186, 187] the use of a combination of optimization and AIMD,[90, 178, 179] the classical treatment of solvent interactions[37, 176] or the application of implicit solvent models.[38, 188, 189] In terms of interpretable differences between gas-phase and solution-phase reactivity, the advantage of implicit solvents is that solvation energies are directly accessible, which is not the case for

AIMD simulations. On the other hand, the main limitation of implicit solvents is that no direct participation of the solvent can be described, in contrast to micro-solvation approaches. As a compromise between efficiency and general accuracy, the combination of an implicit solvent with micro-solvation has been explored.[185, 190-193]

The main conclusions of the various theoretical studies are that adsorbate bindings are weaker in solution than in gas-phase given that the surface needs to be desolvated and only roughly half of the adsorbate is exposed to the solvent.[40] Second, reaction barriers are minimally affected, except for proton transfers for which water can act as a relay or at least activates O–H bonds[185] and when hydroxyl groups are well-exposed to the solvent in the course of the elementary step.[37, 39] As a minor variant, elementary steps that drastically change the surface dipole moment are also influenced by the solvent.[39, 41] This phenomenon can, in the case of acidic protons, lead to heterolytic bond breakings at the water/metal interface.[184] Furthermore, solvent effects are critical when surfaces are partially charged, either in electrochemistry[194] or when assessing the influence of a base which promotes the reaction through co-adsorption.[195] Despite these progress, investigation of the effect of solvent on the entire reaction networks using density functional theory (DFT) remains challenging due to the tremendous computational burden, especially for large alcohols and polyols.[18]

Previously, group contribution methods were pioneered by Benson to predict thermodynamic properties of molecules at a greatly reduced computational cost.[162] Those methods use molecular fragments to correlate thermodynamic properties. The simplest form is to estimate the property of a given component as the sum of its

molecular fragments or groups and is simply called group additivity.[162] It is based on the observation that state properties of molecules can be expressed as simple sums over values assigned to the functional groups. As an example, elongating a linear alkane by one additional methylene ( $\text{CH}_2$ ) group decreases the heat of formation of the alkane by  $\sim 5$  kcal/mol. This simple picture can be refined by applying various corrections for neighboring groups. Nevertheless, the property evaluations are extremely simple compared to quantum mechanical computations. This method was initially developed for closed-shell gas molecules,[162, 196, 197] and it has since been extended to open-shell radicals,[198] transition states,[199, 200] and carbenium ions.[201] Additionally, group additivity has been applied to pure liquids and dilute solutes,[202] and it has been shown to accurately predict various state properties such as the volume, density, viscosity, boiling point, melting point, critical temperature, and pressure.[203-206]

The group additivity scheme accounts for intra-molecular interactions, whereas solution non-ideality requires inclusion of inter-molecular interactions, whose strength depends on the composition of the mixture. To address this issue, group contribution methods that include group interactions were developed. For instance, universal quasichemical functional-group activity coefficients (UNIFAC),[207, 208] computes the average interaction between groups using Flory-Huggins statistical-mechanics model. This method employs the surface area and volume of each group, in addition to the interactions between groups as descriptors, which are then correlated to activity coefficients. More recently, the group contribution method has been extended to equations of state, with the statistical associating fluid theory (SAFT) and conductor-like screening model-segment activity coefficient (COSMO-SAC) models

demonstrating promising performance.[209, 210] These methods apply to bulk fluid mixtures but do not consider catalyst surfaces. The Delaware group has been developing group contribution methods[138] to predict adsorbate properties for hydrocarbons, oxygenates, furanics and aromatics,[35, 36, 211] which, combined with linear scaling relationships, can predict thermochemistry on metal surfaces.[22, 34] However, such a group additivity method for the prediction of thermodynamic properties of adsorbates **in the presence of a solvent** has yet to be developed.

In this paper, we demonstrate that the introduction of an implicit solvent changes the reaction equilibrium constants of specific reaction types (e.g., C–H and O–H dissociations). Then, we build a group additivity method for Gibbs free energy of formation of linear alcohol adsorbates in vacuum and solvent using the polarizable continuum method (PCM) for implicit solvation. Cross-validation reveals that the mean absolute error (MeanAE) is 1.0 kcal/mol and 3.3 kcal/mol for the solvation free energy and the Gibbs free energy of formation, respectively. We further assess our model by comparing DFT computed reaction energies to those predicted via group additivity.

## 4.2 Methods

### 4.2.1 Density Functional Theory and Polarizable Continuum Model

The DFT computations are carried out with the Vienna Ab initio Simulation Package VASP 5.3.5.[212] The electronic structure is described within the generalized gradient approximation using the PBE[96] exchange correlation functional. The dispersion correction is included using the dDsC approach.[213, 214] The projector augmented wave method (PAW) is used to describe the ion-electron interactions.[100,

101] The plane wave basis set is cut off at 400 eV. The electronic energy is converged to  $10^{-6}$  eV and the forces to  $0.02 \text{ eV } \text{\AA}^{-1}$ . The solvation is described using an implicit solvation model as implemented in the package VASPsol.[188, 215] Dipole correction in the z direction is applied for adsorbates to avoid periodic summation of induced dipole and solvent model effect.[176]

The platinum surface is simulated as a  $p(4 \times 4)$  slab consisting of four layers. The two bottom layers are kept fixed at the optimal bulk position in which the Pt-Pt distance is  $2.80 \text{ \AA}$ . The vacuum distance is large enough to avoid spurious interactions (the norm of the vector perpendicular to the surface is  $22.86 \text{ \AA}$ ). The integration of the Brillouin zone is performed using a Monkhorst-Pack mesh of with  $3 \times 3 \times 1$  k-points for surfaces. The gas phase calculations are done at the gamma point in a cubic box of  $20 \text{ \AA}$  in length.

#### 4.2.2 Thermodynamic Property Calculations of Molecules

The Gibbs free energy of formation of a species  $i$  in gas-phase,  $\Delta G_{f,i(\text{gp})}^{\circ}$ , is

$$\Delta G_{f,i(\text{gp})} = \Delta H_{f,i(\text{gp})} - TS_{i(\text{gp})} \quad (4.1)$$

where  $\Delta H_{f,i(\text{gp})}$  is the enthalpy of formation of the molecule  $i$  in gas-phase,  $T$  is the temperature, and  $S_{i(\text{gp})}$  is the entropy of the species  $i$  in gas-phase computed using statistical mechanics, atoms' coordinates and vibrational frequencies (See ref[35] for details). We consider two type of species  $i$ : a gaseous molecule and an adsorbate.  $\Delta H_{f,i(\text{gp})}$  is estimated as

$$\Delta H_{f,i(\text{gp})} = \begin{cases} E_{i(\text{vac})} + H_{\text{rt},i(\text{gp})} + H_{\text{v},i(\text{gp})} + \Delta H_{\text{ref},i} & : i \text{ is gaseous molecule} \\ E_{i(\text{vac})}^* - E_{*(\text{vac})} + H_{\text{v},i(\text{gp})} + \Delta H_{\text{ref},i} & : i \text{ is adsorbate} \end{cases} \quad (4.2)$$

where  $E_{i(\text{vac})}$  is the DFT electronic energy of the gaseous species  $i$  in vacuum,  $E_{i^*(\text{vac})}$  is the DFT electronic energy of the adsorbate  $i$  and the slab,  $E_{^*(\text{vac})}$  is the DFT electronic energy of the empty slab,  $H_{\text{rt},i(\text{gp})}$  and  $H_{\text{v},i(\text{gp})}$  are the rotational (r), translational (t) and vibrational (v) contribution to the heat of formation for the species  $i$  computed in the gas-phase, respectively (See ref[35] for calculation details). For a strongly bounded adsorbate, the rotational and translational degrees of freedom of atomic motion become vibrational upon binding, thus  $H_{\text{rt},i(\text{gp})}$  does not exist for adsorbates. To change from DFT reference system (infinitely separated atoms and electrons) to conventional heat of formations, we introduce the adjustment  $\Delta H_{\text{ref},i}$  (see refs [35] and [34] for computational details and the Supporting Information of Gu et al.[216] for the structures and properties of reference molecules). For the computation of the Gibbs free energy of formation of a solute  $i$  in water or an adsorbate  $i$  at water/metal interface ( $\Delta G_{\text{f},i(\text{aq})}^\circ$ ), we assume that solvation does not depend on temperature (see the Appendix for further discussion) and that the entropy of a species is the same in gas-phase and water. In particular the later approximation is rather severe for adsorption from the liquid phase on the catalyst: Although for ideal solutions the entropy of a solute is the same as the one in the corresponding gas-phase when appropriately taking into account the available volume,[217, 218] the entropy in aqueous solution for real solutes differs by roughly a factor 2.[174, 219] Since an adsorbate competes with the solvent for adsorption sites, the entropy of the adsorbate is, in general, even more challenging to assess than the entropy in solution. Although the structure of the solid-liquid interface remains a matter of debate, it is well established that the solvent properties at the interface are different compared to the bulk liquid.[220-223] Determining the change in entropy upon adsorption requires large-scale molecular

dynamics simulations, which are difficult to converge, both in terms of size and time scale.[222, 224-226] While AIMD would be ideal for such studies, the corresponding computational effort is tremendous.[223, 227-229] On the other hand, force field methods, which make such investigations affordable,[230, 231] suffer from low availability for general interfaces and very limited accuracy, best exemplified for the seemingly trivial water-metal interface, for which many potentials exist,[181, 227, 232, 233] but only for the case of Cu/H<sub>2</sub>O does there exist a well validated force field.[225] Because of all these considerations, the simplistic approximation taken herein makes our approach transparent, and once established, the corresponding entropy corrections can be trivially applied to our model. Hence,  $\Delta G_{f,i(\text{aq})}$  is

$$\begin{aligned} \Delta G_{f,i(\text{aq})} &= \Delta H_{f,i(\text{aq})} - TS_{i(\text{aq})} \\ S_{i(\text{aq})} &\cong S_{i(\text{gp})} \end{aligned} \quad (4.3)$$

where  $\Delta H_{f,i(\text{aq})}$  and  $S_{i(\text{aq})}$  are the enthalpy of formation and the entropy of species  $i$  in aqueous solvent, respectively. The effect of the implicit solvation is included in  $\Delta H_{f,i(\text{aq})}$  as the change in electronic energy:

$$\Delta H_{f,i(\text{aq})} = \Delta H_{f,i(\text{gp})} + E_{\text{sol},i} + E_{\text{cav},i} \quad (4.4)$$

$$= \begin{cases} E_{i(\text{aq})} - E_{i(\text{vac})} & : i \text{ is gaseous molecule or solute} \\ \left( E_{i^*(\text{aq})} - E_{i^*(\text{vac})} \right) - \left( E_{*(\text{aq})} - E_{*(\text{vac})} \right) & : i \text{ is adsorbate} \end{cases} \quad (4.5)$$

Here  $E_{\text{sol},i}$  is the solvation electronic energy of a species  $i$ ,  $E_{i(\text{aq})}$  is the DFT electronic energy of the solute  $i$  in implicit water,  $E_{i^*(\text{aq})}$  is the electronic energy of the solvated adsorbate  $i$  and the slab,  $E_{*(\text{aq})}$  is the electronic energy of solvated empty slab (all without cavitation energy). The cavitation energy approximates the energetic cost to insert a solute in a solvent, such as solvent reorganization and (Pauli) repulsion. Its

energy is usually assumed to be proportional to the solvent accessible surface area. We introduce an empirical cavitation energy correction term,  $E_{cav,i}$ . The computation of cavitation energy using PCM is difficult for the adsorbates and the slab due to numerical instability. Instead, we compute cavitation energy for a small subset of our data, and regress the cavitation energy contribution for each element. The computation of  $E_{cav,i}$  is discussed below in detail. The Gibbs free energy of solvation of a species  $i$ ,  $\Delta G_{sol,i}^\circ$ , is the Gibbs free energy to transfer a species  $i$  from the gas-phase to the liquid bulk, which, using the equations above, is

$$\Delta G_{sol,i} = \Delta G_{f,i(aq)} - \Delta G_{f,i(vac)} = E_{sol,i} + E_{cav,i} \quad (4.6)$$

The accuracy of the PCM is further assessed in Figure C.1 in the Appendix, where the DFT-computed values of  $\Delta G_{sol,i}$  are in a good agreement with the experimental  $\Delta G_{sol,i}$  at 298 K, with a mean absolute deviation of 1.1 kcal/mol. Unfortunately, such data are not available for species at the interface. Hence, the accuracy of the herein applied PCM cannot be assessed. Furthermore, computationally feasible alternatives to implicit solvents to produce a consistent set of undoubtedly reliable solvation free energies at the interface are, to the best of our knowledge, not yet available. More reliable solvation free energies could be obtained with an accurate force field[176] or with long-time-scale ab initio molecular dynamics simulations; in the first case, the force field is currently missing, and the second option is not yet computationally affordable.

In addition to the heat of formation and entropy, we also compute the heat capacity at any temperature  $T$  for the species  $i$ ,  $C_{P,T,i(gp)}$  ( $C_{P,T,i(aq)} \cong C_{P,T,i(gp)}$ ; See ref[35] for calculation detail) for the group additivity regression in order to transfer properties to different temperatures. We surveyed  $C_{P,T,i(gp)}$  between 100 and 1500 at 100 K

intervals. Figure 4.1 summarizes the relationship between gaseous molecules, molecule in the bulk liquid, adsorbates, and solvated adsorbates in a thermodynamic cycle.

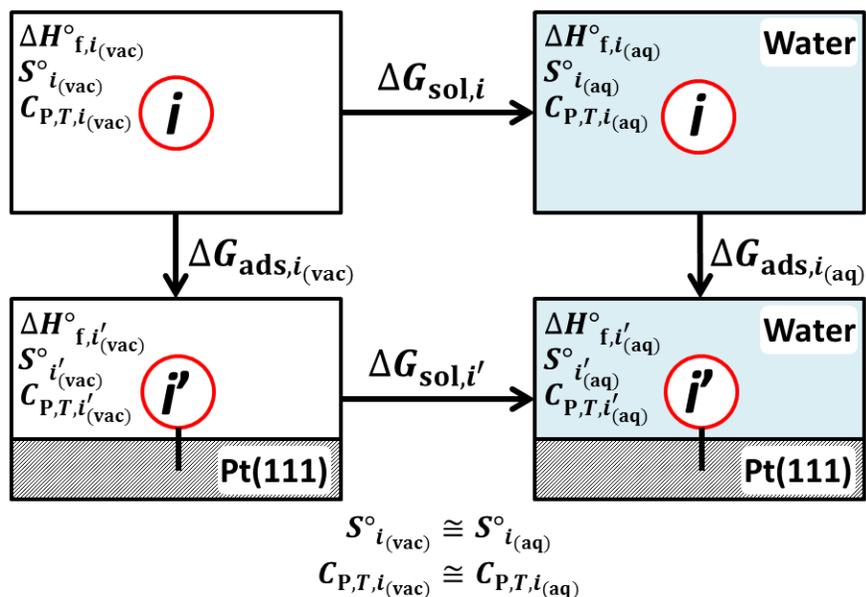


Figure 4.1. Various considered states for the system, summarized in a thermodynamic cycle.  $i$  is an arbitrary species, and  $i'$  is the species  $i$  adsorbed on the surface. Boxes represents a molecule in the gaseous state, in the bulk liquid, adsorbed on Pt(111), and adsorbed and solvated on Pt(111) going from left to right and from top to bottom. The three thermodynamic properties in each box represent the parameters predicted by our method. We assume that the solvation effect is temperature independent, and thus the entropy and heat capacity are the same in vacuum and in water.

### 4.2.3 Group Additivity

The group additivity method assumes that a thermodynamic property,  $\Phi$ , is a linear combination of the thermodynamic properties of groups found within a molecule

$$\Phi_i = \sum_{j=1}^{N_{\text{groups}}} n_{ij} \varphi_j \quad (4.7)$$

where  $n_{ij}$  is the number of times group  $j$  appears in a species  $i$  and  $\varphi_j$  is group  $j$ 's contribution to the thermodynamic property.  $\varphi_j$  values can be derived via linear regression of experimental data or quantum chemical calculations (*vide infra*).

Conventionally, the linear regression is performed for three thermodynamic properties: standard heat of formation, standard entropy, and heat capacity at temperature  $T$ . [162, 198] To account for the solvation effect, we perform the regression to one additional property, the standard heat of formation of solvated species. In order to perform regression to properties of both gaseous molecules and adsorbate molecules together, we found that subtracting rotational and translational contribution from the heat of formation of gaseous molecules improves the fitting. [22, 35, 36] Hence, the dependent variables for the regression are

$$\begin{aligned} \Delta H_{f,i(\text{gp}),\text{GA}} &= \begin{cases} \Delta H_{f,i(\text{gp})} - H_{\text{rt},i(\text{gp})} & : i \text{ is gaseous molecule} \\ \Delta H_{f,i(\text{gp})} & : i \text{ is adsorbate} \end{cases} \\ \Delta H_{f,i(\text{aq}),\text{GA}} &= \begin{cases} \Delta H_{f,i(\text{aq})} - H_{\text{rt},i(\text{vac})} & : i \text{ is gaseous molecule or solute} \\ \Delta H_{f,i(\text{aq})} & : i \text{ is adsorbate} \end{cases} \\ C_{P,T,i(\text{gp}),\text{GA}} &= \begin{cases} C_{P,T,i(\text{gp})} - C_{P,T,\text{rt},i(\text{gp})} & : i \text{ is gaseous molecule} \\ C_{P,T,i(\text{gp})} & : i \text{ is adsorbate} \end{cases} \end{aligned} \quad (4.8)$$

where the subscript “GA” represents the part of the quantity which is expressed in terms of a group additivity scheme.  $C_{P,T,\text{rt},i(\text{gp})}$  is rotational and translational contribution to the heat capacity. For the entropy, we do not correct for

rotational and translational contribution but correct for the symmetric number ( $\sigma$ ) contribution following Bensen's group additivity formulation:[162]

$$S_{i(\text{gp}),\text{GA}} = \begin{cases} S_{i(\text{gp})} + \ln(\sigma) & : i \text{ is gaseous molecule} \\ S_{i(\text{gp})} & : i \text{ is adsorbate} \end{cases} \quad (4.9)$$

We assume that there is no symmetry contribution for adsorbates. The rotational and translational corrections are only applied to the heat of formation and heat capacity as the rotational and translational contributions do not depend on the groups, but are a constant shift (e.g.  $H_{\text{rt},i(\text{gp})} = C_{\text{p,rt},i(\text{gp})} = 4RT$  for nonlinear molecules, or  $7/2RT$  for linear molecules). However, the rotational and translational corrections for entropy depend on the atoms' weight and coordinates, and hence group additivity approach can describe these terms. For the computation of gaseous molecule properties, these subtracted terms needed to be added back when using the group additivity.

The group additivity parameters,  $\boldsymbol{\varphi}_{(N \times 1)}$ , are computed by exploiting the Moore-Penrose pseudoinverse indicated by  $^+$ , in the linear regression:

$$\boldsymbol{\varphi} = (\mathbf{n}^T \mathbf{n})^+ \mathbf{n}^T \boldsymbol{\Phi} \quad (4.10)$$

Here  $N$  is the number of groups,  $\mathbf{n}_{(M \times N)}$  is the group contribution matrix,  $M$  is the number of data points, and  $\boldsymbol{\Phi}_{(M \times 1)}$  is the vector of thermodynamic parameters. The predictive accuracy of the model is assessed using  $k$ -fold cross validation.[234, 235] In this method, the data set is randomly partitioned into  $k$  subsamples of equal size. The model is trained using  $k-1$  subsamples, and the remaining subsample is used as the validation set to compute the model error. This process is repeated  $k$  times with each subsample used as a validation set once. Since the data partitioning involves

randomization, Monte Carlo sampling is employed to compute the average error of each molecule. We used 10 as the value of  $k$  and 1000 repetitions of random partitioning which results in 95% confidence interval of predicted values less than 0.1 kcal/mol for  $\Delta H_{f,i(\text{gp})\cdot\text{GA}}$  and  $\Delta H_{f,i(\text{aq})\cdot\text{GA}}$ , and 0.1 cal/(mol K) for  $S_{i(\text{gp})\cdot\text{GA}}$  and  $C_{P,T,i(\text{gp})\cdot\text{GA}}$ .

#### 4.2.4 Training Set

Our training set consists of 200 data points that are divided into three subcategories: (i) 27 gas molecules, (ii) 53 adsorbates found in the ethanol aqueous phase reforming reaction network, and (iii) 110 adsorbates and 10 gas molecules from the butanetraol reaction network.[236] The adsorbates in the ethanol reaction network include the dissociation products of ethanol as well as species from the water-gas-shift reaction and are used to examine the solvation effect on the reaction energies. The adsorbates from the reaction network of butanetraol supplement the polyol backbone of our group-additivity scheme. This reaction network, generated using the R.I.N.G. software package,[148, 149] contains a total of 5374 species, of which 120 species have been chosen randomly under the constraint that each nonunique group is found in at least seven points of the complete training set. The gas molecules are computed to test the effect of the PCM on the vibrational frequencies (see Figure C.2 in the Appendix). The computed thermodynamic properties and structures of the training set species are also provided in the Supporting Information of Gu et al.[216]

#### 4.2.5 Cavitation Energy Corrections

The cavitation energy correction introduced above,  $E_{\text{cav},i}$ , corrects the solvation energy computed using DFT calculations without the cavitation energy. To determine

the corrections, we compute cavitation energies,  $E_{cav,i}$ , of 20 solutes and nine solvated adsorbates using the equation below:

$$E_{cav,i} = E_{sol,i,+c} - E_{sol,i,-c} \quad (4.11)$$

Here  $E_{sol,i,+c}$  and  $E_{sol,i,-c}$  are  $E_{sol,i}$  using equation (4.5) with the cavitation energy accounted and unaccounted, respectively. Here,  $E_{sol,i,+c}$  is computed using energy cut off of 600 eV to improve the accuracy of this numerically delicate quantity. In a similar spirit to group additivity, we perform a regression of cavitation contribution for each element (C, H, O) to 29 data

$$E_{cav,i} = \sum_j^{C,H,O} n_{ij} e_{cav,j} \quad (4.12)$$

where  $n_{ij}$  is the number of occurrence of element  $j$  in species  $i$ , and  $e_{cav,j}$  is the cavitation energy of element  $j$  computed using the linear regression method introduced above. The MeanAE and MaxAE (training error, not cross-validation error) are 0.18 and 0.47 kcal/mol, respectively. The training set data and regression coefficients are provided in Supporting Information.

### 4.3 Result and Discussion

Having presented the methodological approach, we now turn to the results of our study. First, we discuss the influence of the implicit solvent on geometries and reaction free energies. Then, we present the performance of our group additivity scheme for gas- and solution-phase species and reactions.

### 4.3.1 Influence of the Solvation on the DFT Geometries

Upon solvation by water using a continuum model, isolated molecules showed limited structural changes with a root-mean-square deviation (RMSD) of 0.02 Å and a maximum RMSD of 0.09 Å. The RMSD have been evaluated using the Kabsch algorithm.[237-239] This algorithm optimally superposes the two molecules to minimize translational and rotational error and computes the RMSD. For the surface species, we disregarded the change in surface atom's position. These small positional changes are induced by the dipolar interactions with the continuum solvent mostly affecting the relative orientation of the C–O bonds. The structures of the adsorbed species are slightly more sensitive to the solvation with a mean RMSD of 0.03 Å and a maximum RMSD of 0.16 Å. This maximum corresponds to a change in the orientation of the OH bonds of COH–C–CH<sub>2</sub>–CHOH that rotates from pointing toward the surface to pointing toward the solvent. To better analyze this effect, we considered this reorientation for surface intermediates in the ethanol steam reforming reaction network. While the conformer with the OH bond pointing *down* is the most stable in vacuum, pointing the OH bond *up* becomes more stable in water. This effect, illustrated in the simple case of the CHOH fragment in Figure 4.2, is assigned to a better solvation of the OH by up 3 to 5 kcal/mol compared to OH *down*.

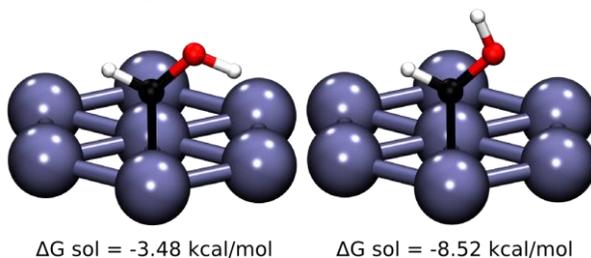


Figure 4.2. DFT structures of the HCOH fragment adsorbed on a Pt(111) slab: (left) the H down configuration (right) the H up configuration. Associated solvation energies are also given.

### 4.3.2 Influence of the Solvation on Reactions.

We assess the influence of PCM on ethanol aqueous reforming, including water-gas shift reactions, by considering the corresponding standard Gibbs free energy of reactions in gas-phase (gp) and solution (aq):

$$\begin{aligned}\Delta G_{r,(gp)} &= \sum_i v_i \Delta G_{f,i(gp)} \\ \Delta G_{r,(aq)} &= \sum_i v_i \Delta G_{f,i(aq)}\end{aligned}\tag{4.13}$$

where the reaction involves species  $i$  with the stoichiometric coefficient  $v_i$  with  $v_i > 0$  if  $i$  is a product. To assess the change in the standard Gibbs free energy of reaction upon solvation, we further define  $A_{sol}$  as the change of a reaction energy upon solvation:

$$A_{sol} = \Delta G_{r,(aq)} - \Delta G_{r,(gp)}\tag{4.14}$$

When  $A_{sol}$  is negative, the reaction is favored by solvation, the Gibbs energy of reaction at the solid-liquid interface being more exergonic than in the gas-phase. This change can also be seen as the sum of the solvation energies of each species  $i$  involved in the chemical reaction with a stoichiometric coefficient  $v_i$ :

$$A_{sol} = \sum_i^{N_{species}} v_i \Delta G_{sol,i}\tag{4.15}$$

$A_{sol}$  is reported as a box plot on Figure 4.3a for each type of reaction under consideration: adsorption, C–H, O–H, C–C and C–O bond scissions.

In general, the adsorption of close-shell molecules is only slightly disfavored by the solvation with a mean of 0.2 kcal/mol for  $A_{sol}$ . The most affected species is acetaldehyde with  $A_{sol} = 2.4$  kcal/mol: its adsorption at the Pt(111) surface is

disfavored by the stronger solvation in the solution ( $\Delta G_{\text{sol}} = -5.6$  kcal/mol) than at the surface ( $\Delta G_{\text{sol}} = -3.3$  kcal/mol).

Since the solvation energy does not change significantly upon adsorption, the analysis of the solvation energy of different species is done jointly for the molecules in bulk water and the ones adsorbed at the interface. The species have been split into three groups to represent the solvation Gibbs energy  $\Delta G_{\text{sol}}$  with box plots in Figure 4.3b: (1) *Hydroxyl* contains at least one hydroxyl group, including water; (2) *aprotic Oxygenate* includes all the species that contain an oxygen atom but not a hydroxyl group, for instance acetaldehyde; (3) *hydrocarbon* selects species that do not contain any oxygen atom. As expected, *hydrocarbon* members are weakly solvated by water (on average,  $\Delta G_{\text{sol}} = -1.7$  kcal/mol, standard deviation of  $\sigma = 1.0$  kcal/mol). *Aprotic oxygenate* are more solvated (on average,  $\Delta G_{\text{sol}} = -2.9$  kcal/mol,  $\sigma = 0.9$  kcal/mol) with a small standard deviation ( $\sigma$ ). The *hydroxyl* species show a stronger Gibbs solvation energy (on average,  $\Delta G_{\text{sol}}^{\circ} = -5.6$  kcal/mol,  $\sigma = 1.9$  kcal/mol) and, more importantly, a large standard deviation, with a maximum and minimum of  $-2.5$  and  $-9.8$  kcal/mol, respectively. For the surface species, the strong variations can be directly related to the two possible orientations of the OH bond, *up* or *down*. In most cases, the O–H *down* is the most stable configuration and is poorly solvated (for instance,  $\Delta G_{\text{sol}}^{\circ} = -3.5$  kcal/mol in CHOH). In some cases, the O–H *up* is the most stable configuration and is strongly solvated such as in CCHOH ( $\Delta G_{\text{sol}}^{\circ} = -9.3$  kcal/mol, the *up* configuration being more stable by 0.3 kcal/mol).

While most bond dissociations are made more exergonic by the solvation ( $A_{\text{sol}} < 0$ ), the OH scission is generally disfavored ( $A_{\text{sol}} > 0$ ). This  $A_{\text{sol}}$  dependency on the reaction type is in agreement with that computed using the report by Behtash *et al.*[39-

41] With the loss of the most solvated functional group, the OH scission leads to a net loss in solvation, explaining why such a reaction is less likely in water than at the metal/gas-phase interface. This thermodynamic observation contrasts with our previous kinetic study using microsolvation, where the O–H breaking was found to be slightly accelerated by the presence of a water molecule.[240] On the other hand, it agrees with the study of Lopez and co-workers, who have employed implicit solvation in combination with some explicit water molecules.[241] In the absence of a detailed study at an explicit water/Pt(111) interface, it is not possible to assert that one or the other study captures the essential effect. The outliers identified in Figure 4.3a correspond to reactions where the solvation effect between reactants and products is the most unbalanced. For instance, the three outliers of the C–O scission all correspond to reactions that generate the OH fragment. Two reactions are strongly disfavored at the water-metal interface, due to a strong solvation of the reactants (CCOH,  $\Delta G_{\text{sol}} = -9.8$  kcal/mol and COH,  $\Delta G_{\text{sol}} = -7.9$  kcal/mol) that exhibit an OH bond that is pointing toward the solvent. Conversely, the CH<sub>3</sub>COH fragmentation is favored at the water/Pt interface, since this species is only weakly solvated when adsorbed on Pt(111) ( $\Delta G_{\text{sol}} = -2.5$  kcal/mol), with a OH bond that points downward.

In, summary, most of the steps along the ethanol reforming process are favored by the presence of the water solvent except O–H scission (disfavored) and adsorption (neutral). The solvation Gibbs energy of an adsorbate can be related to the presence of a hydroxyl group and more precisely to its orientation: when an OH bond points toward the bulk water, the corresponding adsorbate is particularly well solvated. This aspect allows rationalizing and predicting the effect of water on most elementary steps.

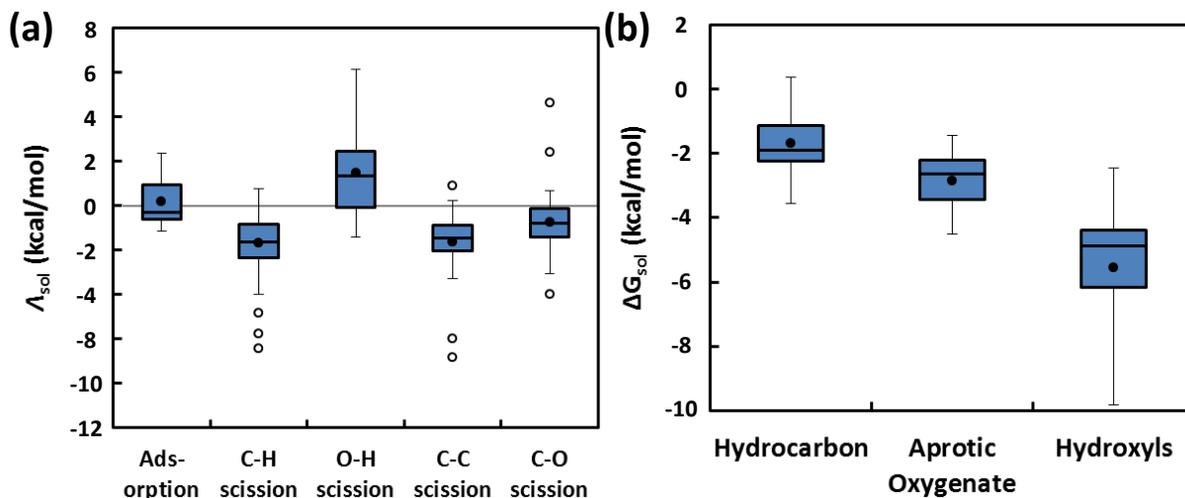


Figure 4.3. Box plot for the change of (a) the Gibbs energy of a reaction upon solvation  $\Delta G_{sol}$  (see text for definition) and (b) the solvation Gibbs energy  $\Delta G_{sol}$  of each species using DFT in the ethanol reforming reaction network (without water-gas shift reactions and species). Black dots and white dots represent mean and outliers, respectively. The box represents the second and third quartiles while the whisker represents the first and fourth quartiles. Outliers are at least 1.5 times the box range (between the second and third quartile) from the edge of the box.

### 4.3.3 Group Additivity Schemes

In this section, we present a group additivity scheme to predict  $\Delta G_{f,i(\text{gp})}$  and  $\Delta G_{f,i(\text{aq})}$ . These two quantities are computed using the equations described in the Methods. The group additivity scheme for these quantities follows the strategy of Benson's original group additivity. Groups are given as  $A(B)_b(C)_c\dots$  where A is a group center, and B and C are group peripherals, and b and c are the numbers of connectivity to the B and C group peripherals. A, B, and C are typically an atom and sometimes a group of atoms. A, B, and C are further categorized by their hybridization. For example,  $C_=\text{}$  indicates a carbon with a double bonded neighbor. Also, if the bond order between group center and group peripherals are higher than

single bond, the bond order is described (e.g., A(=B) for A double bonded to B). Every atom must be accounted once as a group center with the exception of hydrogen.

Previously, we have extended this formulation to the adsorbate on surface.[19, 35, 36] In these schemes, the valence electrons covalently interacting with the surface electrons are denoted as  $(\text{Pt})_x$  group peripheral. For example, the group  $\text{C}(\text{H})(\text{Pt})_2(\text{O})$  contains a carbon central atom single-bonded to peripheral groups, a hydrogen and an oxygen, and has 2 valence electrons interacting with the surface denoted as  $(\text{Pt})_2$ . This scheme does not distinguish whether the central carbon sits on an atop, bridge, or hollow site; thus, the scheme only requires adsorbates' connectivity to compute their energy. We also adopt ring-strain correction as implemented in a previous study.[36] These corrections account for the strain of the ring formed by surface atoms and adsorbate atoms. For example, a ring is formed when two connected carbons are bonded to two surface atoms, forming  $\text{Pt}-\text{C}-\text{C}-\text{Pt}$  ring where two Pt atoms are connected. Additionally, we introduce 16 pair corrections that describe the valence electrons of pairs of central atoms, e.g.,  $\text{C}(\text{Pt})_3\text{C}(\text{Pt})_3$ ,  $\text{C}(\text{Pt})_3\text{C}(\text{Pt})_2$ ,  $\text{C}(\text{Pt})_3\text{C}(\text{Pt})$ ,  $\text{C}(\text{Pt})_2\text{C}(\text{Pt})_2$ , and so on. These corrections accounts for the partial  $\pi$ -mode interaction between the two surface-bound organic atoms. Previous studies[139, 211] have shown that when two connected atoms are bonded to the surface, the bond between two connected atoms remains partially conjugated whereas we previously assumed that the bond between two atoms are single bonds. These inadequately defined bond order causes an error as the group additivity relies on transferable electron spatial density profile of the homologous groups between molecules. Thus, groups accounting for conjugation are essential to the group additivity scheme.[211] These corrections improve the model significantly (reduce the MeanAE in  $\Delta G_{f,(\text{gp})}$  from 4.6 to 3.3

kcal/mol using 10-fold cross-validation). Additionally, we add a group specific to an outlier, CCOH. CCOH has a unique structure that leads to high deviation for  $\Delta G_{\text{sol}}$ . This correction improves the MaxAE of  $\Delta G_{\text{sol}}$  from 6.4 to 5.3 kcal/mol.

Following the formulation described above, the model results in total of 82 groups/corrections. The effective rank of this model is 75; i.e., it is rank-deficient. Five of seven linearly dependent descriptors are  $\text{C(Pt)}_3\text{C(Pt)}_3$ , OC,  $\text{OC(Pt)}_1$ ,  $\text{OC(Pt)}_2$ , and  $\text{OC(Pt)}_3$  which are part of the new corrections introduced in this work. Since these corrections are already implicitly accounted by the scheme, we remove these corrections. The last two rank deficiency comes from  $\text{C}_= (\text{CO})(\text{C}_=)(\text{H})$  group's dependence to  $\text{CO}(\text{C}_=)(\text{H})$ , and  $\text{O}(\text{CO})(\text{H})$  group's dependence to  $\text{CO}(\text{O})(\text{Pt})$  (determined using reduced row echelon form). These groups are found in gas/solute molecules, which are not the focus of this study, and the related methods have been previously developed extensively, thus we combine these groups together as a single multigroup. The list of group centers and peripherals as well as corrections and unique groups are summarized in Table 4.1, and their values are listed in the Appendix.

Table 4.1. List of group centers and peripherals that constitute groups (See the body for group description), corrections and unique groups

Group Centers	Group Peripherals	Corrections/Outlier Correction
C	C	Corrections:
CO <sup>a</sup>	CO <sup>a</sup>	C(Pt) <sub>x</sub> C(Pt) <sub>x</sub>
O	O	C(Pt) <sub>x</sub> O(Pt) <sub>x</sub>
C=	C=	(5 linearly dependent corrections removed)
C≡	C≡	Surface ring strain[36]
	H	
	Pt	Outlier Correction: CCOH (Chemisorbed)

<sup>a</sup> CO indicates C=O carbonyl group.

#### 4.3.4 Model Validation

The predictive accuracy of the group additivity model is assessed using *k*-fold cross validation as described in the computational details.[234, 235] Parts a and b of Figure 4.4 show the average of predicted values using the group additivity models vs. the DFT computed  $\Delta G_{f,(gp)}$  and  $\Delta G_{sol}$ . Cross validation reveals a reasonable MeanAE in  $\Delta G_{f,(gp)}$  of 3.3 kcal/mol, and the MaxAE of 16.5 kcal/mol as shown in Table 4.2. The large errors are attributed to unaccounted binding modes (e.g., adsorption conformation of polyols). Our previous group additivity model has shown that including binding modes' information to the group descriptors results in large improvement of the cross-validation error.[211] However, including these unaccounted binding modes requires a significantly larger data set. Since the main utility of the group additivity is to rapidly screen large reaction networks before computing the most promising intermediates at the DFT level, a huge computational effort is not justified. The MeanAE and MaxAE for  $\Delta G_{sol}$  are as small as 1.0 and 5.4 kcal/mol, respectively. Importantly, the MeanAE is within chemical accuracy;

therefore, the solvation free energy scheme from group additivity provides a good first approximation to estimate the thermochemistry.

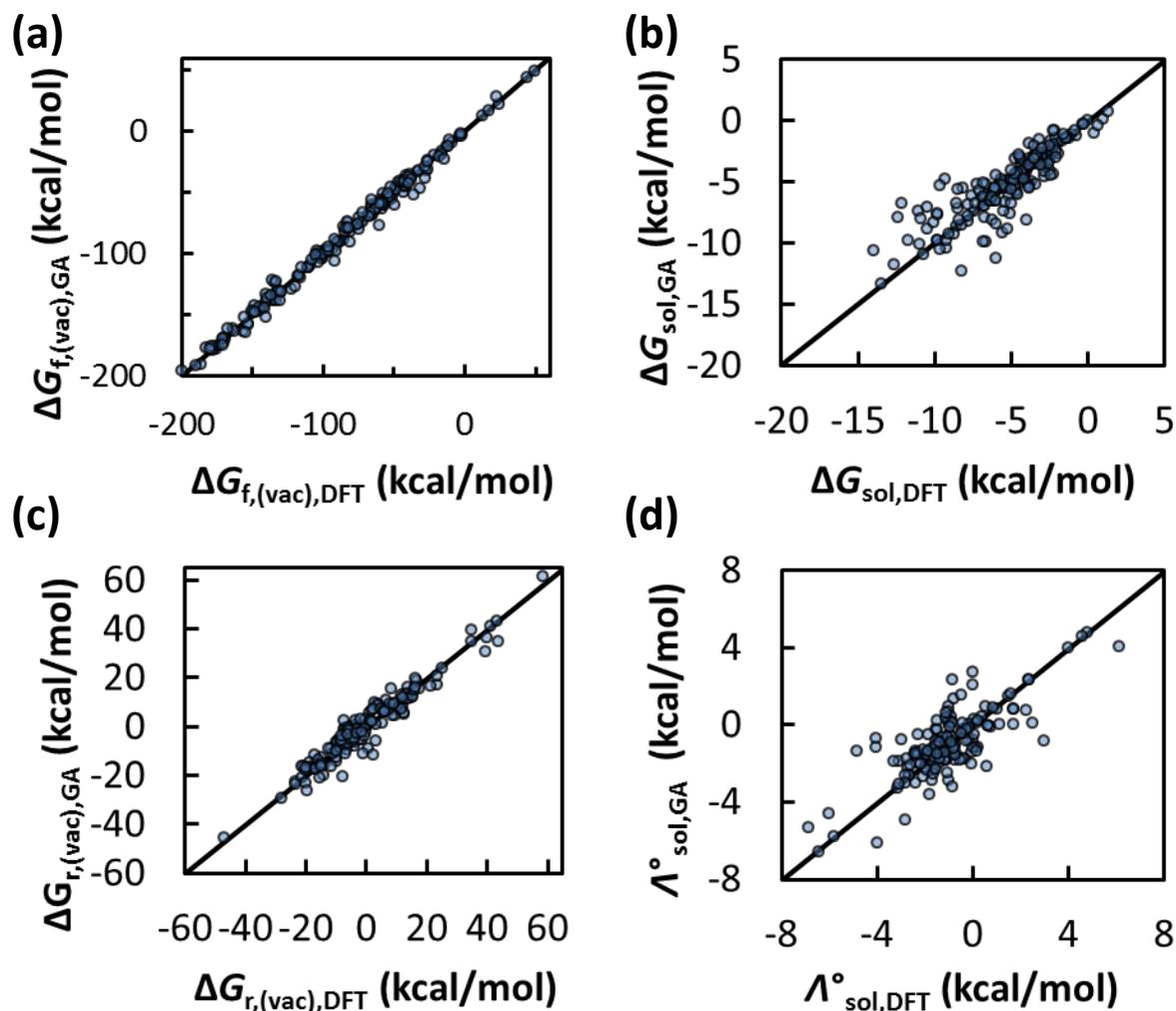


Figure 4.4. Parity plot between 10-fold cross-validation computed prediction and DFT data for (a) Gibbs free energy of formation,  $\Delta G_{f,(gp)}$ , (b) Gibbs free energy of solvation,  $\Delta G_{sol}$ , (c) Gibbs free energy of reaction,  $\Delta G_{r,(vac)}$ , and (d) the change of reaction energies upon solvation,  $\Delta_{sol}$ , as defined in equation (4.14) and (4.15). 151 reactions of the ethanol reaction network are considered (see the Appendix for the details).

Table 4.2. Mean and Maximum (Max) Absolute Errors (AEs) and Root Mean Square Error (RMSE) for 10-Fold Cross-Validation of Thermodynamic Properties in kcal/mol

Properties	MeanAE	MaxAE	RMSE
$\Delta G_{f,(gp)}$	3.3	16.5	4.8
$\Delta G_{f,(aq)}$	3.3	17.5	4.8
$\Delta G_{sol}$	1.0	5.4	1.6
$\Delta G_{r,(gp)}$	2.9	13.8	4.0
$\Delta G_{r,(aq)}$	2.8	15.1	3.9
$A_{sol}$	0.9	3.8	1.2

We also assess the model’s predictive ability for the (151) reaction free energies in the ethanol steam reforming reaction network.[236] Specifically, we compare the Gibbs free energy of reaction in gas-phase,  $\Delta G_{r,(gp)}$ , and the change in Gibbs free energy of reaction upon solvation,  $A_{sol}$

Parts c and d of Figure 4.4 show the parity plot of the group additivity computed values vs. the DFT-derived  $\Delta G_{r,(gp)}$  and  $A_{sol}$ . The MeanAE and MaxAE for  $\Delta G_{r,(gp)}$  are 2.9 and 13.8 kcal/mol, which are smaller than those for  $\Delta G_{r,(gp)}$ . Considering that the number of groups constituting a molecule is typically larger than the number of groups changing in a reaction, the variance is likely smaller for the reaction energy, explaining the smaller error. The MeanAE and MaxAE for  $A_{sol}$  are reasonable at 0.9 and 3.8 kcal/mol, respectively. However, as can be seen in Figure 4.4d, the parity plot is not perfect, i.e., the predicted change of solvation energies along a reaction is not very well correlated to the DFT data. One of the main issues is the evidenced reorientation of OH-groups discussed above: in the GA, the orientation of the OH group is not accounted for, implying that these subtle effects cannot be investigated by the present group additivity scheme. However, since the Gibbs

reaction energy prediction is rather accurate, the main goal of the GA is still achieved: rapid prediction of various reaction energies at a solvated metal surface.

For a further assessment of the quality of the solvation free energy group additivity, we compare the trends we obtained in Figure 4.3 with the DFT calculations with the trends we obtain using our group additivity scheme (Figure C.3 in the Appendix). Overall, the solvation free energy group additivity captures the trend very well. The mean and median between the DFT and the group additivity are close, signifying that our model captures the general dependency on the reaction type and functional groups. However, the boxes and whiskers are tighter for the model than the DFT (i.e., variance is not captured by the model). Especially, the model computed  $\Delta_{\text{sol}}$  for O–H scission and  $\Delta G_{\text{sol}}$  for *Hydroxyls* have significantly lower variance than the DFT-computed values. Future work will be based on information from explicit water simulations and include heterolytic bond breakings involving protons, which may be important in biomass conversion. These improvements would, ultimately, allow to investigate the pH dependence of reaction schemes at the group additivity level.

#### 4.4 Conclusions

The impact of an implicit solvent on adsorption of alcohols on the Pt(111) surface has first been investigated by DFT computations. We evidenced only minor changes in the geometries upon solvation, with the reorientation of an O–H bond pointing toward the surface in vacuum and being more stable pointing into the solvent being the rare exception. Examination of the ethanol reforming reaction network reveals, however, that equilibrium constants of surface reactions change significantly in solvent depending on the reaction type. In aqueous environment, C–H scission is slightly favored as the oxygen group of the oxygenate is raised toward bulk water,

whereas O–H scission is thermodynamically disfavored as the oxygen binds to the surface which decreases the oxygen-water interaction. In order to cut down the computational expense of such studies, we have developed a simple group additivity scheme for the Gibbs free energy of linear alcohols at the Pt(111) water interface by computing the thermodynamic properties of 200 adsorbates and gaseous species using DFT and an implicit solvent. The 10-fold cross-validation yields a MeanAE for the Gibbs free energy of formation and the Gibbs free energy of solvation of 3.3 and 1.0 kcal/mol, respectively, and smaller error for reaction energies (MeanAE = 2.8 kcal/mol). Despite severe approximations regarding the entropy of adsorption, this development allows addressing large and complex reaction networks, e.g., hydrodeoxygenations of polyols such as glycerol, erythritol, or even sorbitol, in order to narrow down the most likely reaction routes, which then can be studied by DFT computations. Furthermore, the corresponding solvent effects can be improved in order to refine the subtle enthalpy-entropy balance at the solid-liquid interface.

#### **4.5 Acknowledgements**

G. G. is grateful to Dr. Glen R. Jenness for the paper writing improvement. We acknowledge Angela Norton for performing COSMO-SAC calculations (Figure C.2 in the Appendix). G. G. developed the group additivity model, B. S. performed DFT calculations, D. G. V., P. S., C. M., and S. N. S. conceptualized the research, and advised the research direction. This research was supported by the NSF under grant number CHE-1415828. This work benefited from the support of the project MuSiC ANR-14-CE06-0030 of the French National Research Agency (ANR). This work was granted access to the HPC resources of CINES and IDRIS under the allocation A0010810119 and x2015080609 made by GENCI. Computational resources from the

*Pôle Scientifique de Modélisation Numérique (PSMN)* at ENS Lyon are greatly acknowledged as well.

## Chapter 5

### THERMOCHEMISTRY OF GAS-PHASE AND SURFACE SPECIES VIA LASSO-ASSISTED SUBGRAPH SELECTION

#### 5.1 Introduction

Modeling of chemical kinetics plays a key role in a wide variety of fields such as catalysis, biology, and combustion. While first-principles modeling can provide insights into reaction mechanisms and better catalysts for reactions of small molecules,[242-247] it is prohibitive for large molecules and reaction networks due to the computational burden of electronic structure calculations, especially for reactions occurring on catalysts.[18, 21, 134, 248-250] Development of semi-empirical methods, which can quickly estimate thermochemistry and activation energies, is thus essential. In addition, semi-empirical methods have been combined with microkinetic modeling and uncertainty quantification and demonstrated a powerful hierarchical approach to building and understanding complex reaction mechanisms.[19, 42, 44] Given that uncertainty quantification reveals higher model sensitivity to thermochemical parameters, the development of a more accurate thermochemistry estimation tool as well as model uncertainty quantification is crucial.[19, 42]

Graph theory is one such widely employed approach for estimating thermochemistry[52, 202-206, 251-259] that represents species with graphs. A vertex represents an atom or a group of atoms and bonds among them, and an edge represents a bond between two atoms. A graph made of such vertices and edges representing a chemical compound is called a molecular graph. A graph, whose vertices and edges

are subsets of a molecular graph, is called subgraph, and, typically, the subgraph frequency — the number of appearances of a subgraph in a molecular graph — correlates with thermodynamic properties.[52, 202-206, 251-259] The most prominent thermochemistry estimation method is the group additivity (GA) pioneered by Benson,[162, 260-262] which estimates the standard enthalpy and entropy of formation of species. GA has been proven effective for simple hydrocarbons and oxygenates with mean absolute error (MAE) for standard enthalpy of formation, on average, comparable to experimental precision (within 1.0 kcal/mol of experimental values).[162, 260-262] Since its original application to gaseous molecules, GA has been extended to solvated species, ions, radicals, and surface adsorbates.[21-23, 34, 216, 249, 252, 263-267]

GA involves two graph descriptors, namely groups and corrections, as shown in Figure 5.1. A group graph consists of a single atom, its nearest neighbor atoms, and the covalent bonds between them (coarse-grained or pseudo atoms (e.g., C=O) are occasionally used). Such graph descriptors have physical basis.[52, 259] A graph of corrections usually involves a larger number of vertices and edges than a group and describes non-nearest-neighbor electronic effects, such as steric hindrance, resonance structures, and ring strain. Corrections are introduced when large errors in prediction using only groups occur.[162, 249, 250, 260, 268] In most cases, both group and correction graphs are heuristically selected.[162, 248-250, 260, 264, 266, 268]

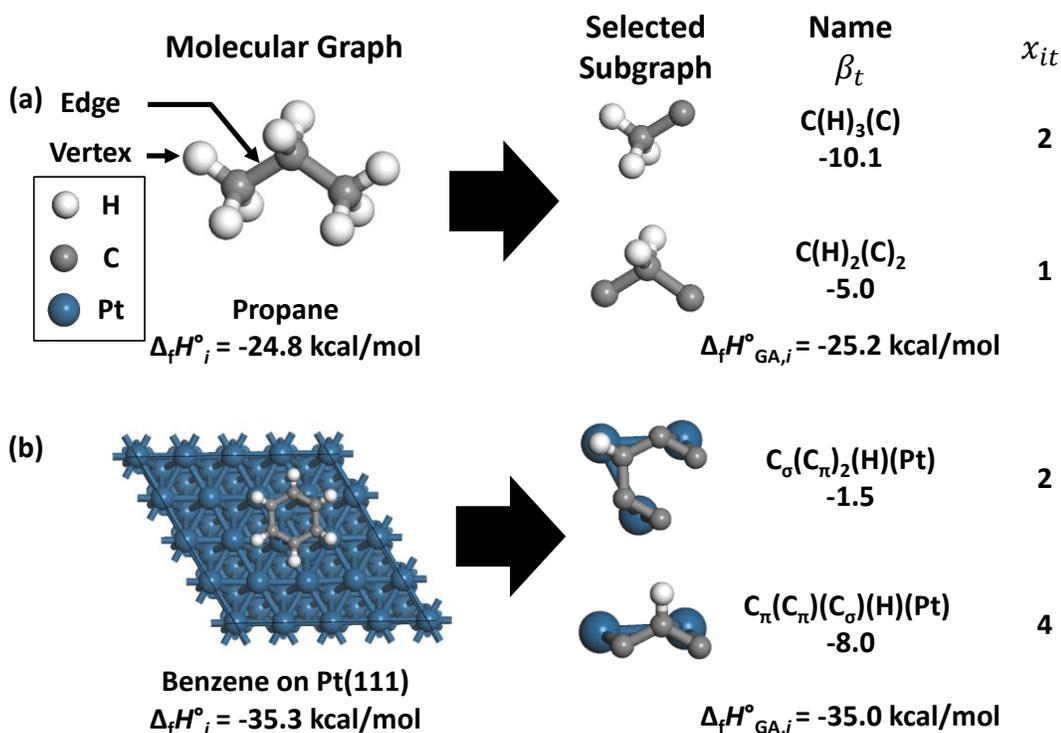


Figure 5.1. Examples of graph descriptors used in estimating thermochemistry: (a) gaseous propane[162, 260] and (b) benzene on Pt(111).[21] The left/top legend indicates the element represented by each vertex.  $\Delta_f H^\circ_i$  and  $\Delta_f H^\circ_{\text{GA},i}$  indicate standard enthalpy of formation, either experimental (a) or the DFT computed value (b) and group additivity computed standard enthalpy of formation for datum  $i$ , respectively.  $\beta_t$  is the contribution of graph  $t$  to the  $\Delta_f H^\circ_{\text{GA}}$ ,  $x_{it}$  is the number of appearances of graph descriptor  $t$  in the molecular graph of datum  $i$ , and  $\Delta_f H^\circ_{\text{GA},i} = \sum_t x_{it} \beta_t$ .

Heuristics-based graph selection requires manual identification of corrections, which is often difficult (e.g., for alkyl peroxides,[269, 270] proteins and enzymes,[271, 272] or ring strain corrections for complex adsorbates[16, 21-23]). For example, strain energies of rings are significantly affected by the electronegativity of rings' substituents[165, 273-276] but ring strain corrections in previous models do not account for substituents.[162, 250, 260] Importantly, selection of graphs is not

optimal, leading to potentially lower accuracy and ad hoc introduction of new rules.[16, 21-23, 277-285] Development of a systematic approach for graph selection could lead to models with the improved predictive ability and enable subgraph-based modeling of more complex systems.

In this work, we exploit a feature selection method called LASSO to systematically and optimally choose subgraphs for estimating thermochemistry. We gather all hydrocarbon gas molecules from the NIST Chemistry WebBook database[286] and the Burcat's thermodynamic database[287] and exploit LASSO with complete graph enumeration for small molecules. A new semi-supervised learning framework, the Least Angle Regression (LAR)-LASSO method with branch pruning, is introduced for selecting graphs for large molecules to keep computational cost and memory tractable. Improved accuracy of the new methods compared to the traditional group additivity is demonstrated. We also compare this framework with the state-of-the-art machine learning models by applying it to QM9 database.[288] In addition, we introduce an adsorbate subgraph enumeration algorithm and apply LASSO to previously published surface science data.[16, 21, 23] The framework shows a good predictive ability for lignin monomer adsorbates[21] but is not effective for other adsorbate data sets[16, 23] due to poor data sets. We suggest a simple approach to identify such poorly sampled data.

## **5.2 Methods**

### **5.2.1 Data Sets and Graph Mining**

We obtain thermodynamic properties from 5 different data sets: (1) 738 gaseous hydrocarbons from the NIST Chemistry WebBook database[286] and the

Burcat's thermodynamic database[287] obtained as of February 2017; (2) ~134,000 gaseous molecules in the QM9 data set;[288] (3) 150 linear oxygenate adsorbates from Saliccioli et al.[23]; (4) 87 furanic adsorbates from Vorotnikov et al.;[16] and (5) 591 lignin monomer adsorbates from Gu et al.[21] For (1), (3), (4) and (5), we exclude data points that contain groups encountered only once in previous GA schemes. These data points are outliers as their electron density profile of the described atom is non-transferrable.[52, 289] This removal results in 695, 137, 76, and 591 data in the order of sets listed above. QM9 is a popular data set for testing machine learning models. This data set contains density functional theory computed properties of ~134,000 molecules with up to 9 heavy atoms (C, O, F, and N) found in the GDB-17 chemical database.[288, 290] This data set is used in its entirety and differs from data set (1) in that (1) is a smaller data set and contains molecules with as many as 42 heavy atoms as well as radical species, which are known to be difficult to build models for. We use the QM9 data set to directly compare the proposed LASSO method to published machine-learning models. We focus on the standard enthalpy of formation,  $\Delta_f H^\circ$ , as the electronic energy accounts for the majority of the error in the Gibbs free energy of formation.[16]

We obtain available molecular graphs for gas molecules, and when these are missing, we convert their molecule names to SMILES strings, using OPSIN,[291] and then to molecular graphs. The molecular graphs are then converted to subgraphs and frequencies of appearance. All graph operations are performed using the RDKit software.[292] We lump together hydrogen atoms connected to the same atom for simplicity. For example, instead of assigning a vertex to each hydrogen on a methyl group, a single vertex that represents three hydrogens is used.

For adsorbates, we construct molecular graphs from atomic coordinates. We determine connectivity between atoms,  $i$  and  $j$ , using the following inequality:

$$d_{ij} < tol(cr_i + cr_j) \quad (5.1)$$

Here,  $d_{ij}$  is the distance between the atom  $i$  and  $j$ ,  $tol$  is a tolerance factor, and  $cr_i$  and  $cr_j$  are the covalent radii of the atoms  $i$  and  $j$ , respectively. The covalent radius, fitted using 228,000 experimentally computed bond lengths,[293, 294] is the average size of a covalently bonded atom. The sum of two atoms' covalent radii is a good estimate of the bond length, and by multiplying with a tolerance factor, connectivity is estimated well. Here, we use a value of  $tol$  of 1.35. This factor is chosen by inspecting electron density, and performing Bader analysis and charge flow analysis for the bond's existence given interatomic distances in our previous work.[21] Motivated by Bader analysis,[21] all bonds between adsorbate atoms (C, H, and O) are taken to be single bonds, giving occasionally leftover electrons denoted in the atom's label. For example, a carbon in benzene is labeled as C1 to indicate that it has one leftover electron after the aromatics bonds are set to single bonds. This modification improves fitting as the number of surface interacting-electrons is correlated to binding.[295] Hydrogen atoms are lumped together as discussed above.

As shown in Figure 5.1, the molecular graphs of adsorbates on a lattice can be complex if one includes all surface atoms. To cope with this, we introduce an algorithm that enumerates all possible subgraphs using atoms in adsorbates (i.e., C, H, and O) only, and adds surface atoms bonded to adsorbate-atoms in a subgraph but ignores bare surface atoms. The adsorbate atoms' connectivity to surface atoms captures binding site information which is crucial for the description of electronic energy.[21] This enumeration algorithm is summarized in Figure 5.2 and is available

at <https://github.com/vlachosgroup/GraphLearning>. This approach does not currently differentiate molecules on other miller index surfaces. Thus, a separate model is needed for each surface. It may be possible to build a unified model that accounts for the local site-environment at which binding happens, based on surface atom coordination numbers, as the latter has recently been shown to scale with binding energy.[295] In order to use models based on molecular graphs for adsorbates, an algorithm that can predict adsorption geometry from the gas molecular graph is necessary. This algorithm, based on data-based machine learning methods, is currently work in progress.[296]

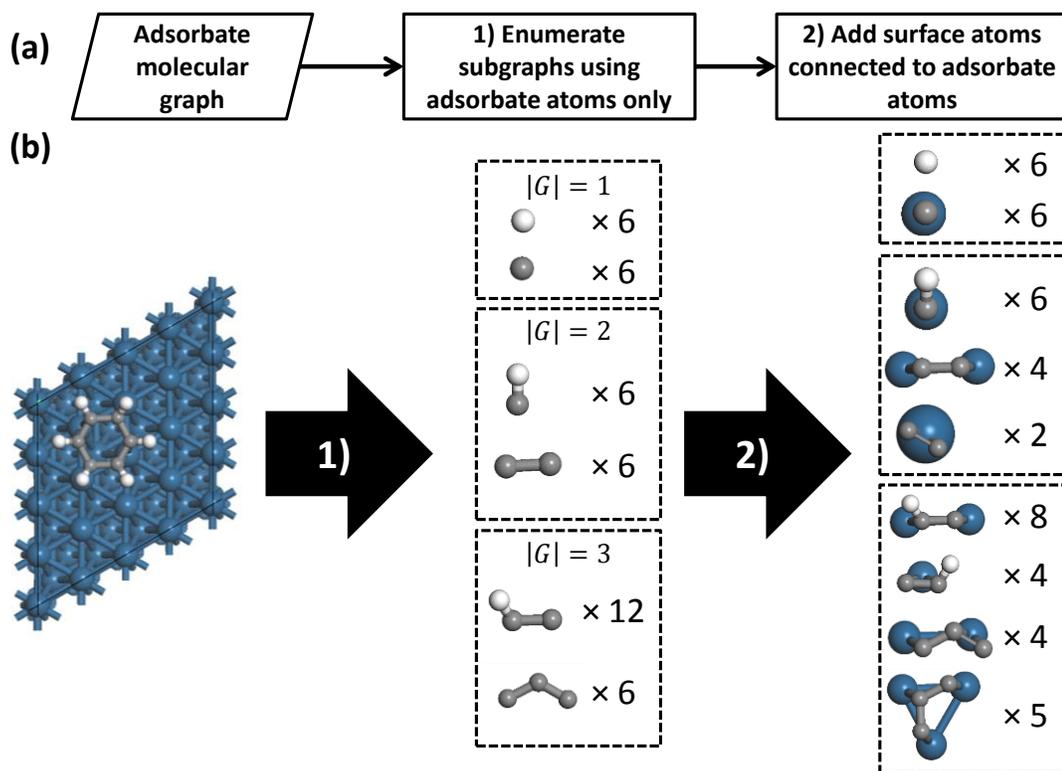


Figure 5.2. (a) Flow chart of the adsorbate graph enumeration algorithm and (b) example of the algorithm using benzene on Pt(111)[21] (only subgraphs with a number of vertices  $\leq 3$  are shown). “1)” and “2)” in arrows in (b) represent the steps in the flow chart of (a).

### 5.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Here we introduce a feature selection algorithm, LASSO, which is a constrained ordinary least square regression for automatic and optimal selection of graphs.[297] Based on previous studies, we correlate graph descriptors to thermochemistry with a linear model:[52, 202-206, 251-258]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r} \quad (5.2)$$

$\mathbf{Y} \in \mathfrak{R}^N$  is the response variable (here the standard enthalpy of formation),  $N$  is the number of data,  $\mathbf{X} \in \mathfrak{R}^{N \times P}$  is the descriptor matrix where the element  $x_{ij}$  indicates

the frequency of graph descriptor  $j$  in the molecular graph of datum  $i$ ,  $P$  is the number of descriptors,  $\boldsymbol{\beta} \in \mathfrak{R}^P$  is the regression coefficient vector, and  $\mathbf{r} \in \mathfrak{R}^N$  is the model residual or error. For the ordinary least square method, the following loss function is minimized to compute  $\boldsymbol{\beta}$ :

$$L(\boldsymbol{\beta}) = \frac{1}{2N} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.3)$$

In LASSO, the  $L_1$  norm of  $\boldsymbol{\beta}$  is minimized together with the loss function:

$$\boldsymbol{\beta}(\lambda) = \min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \quad (5.4)$$

Here,  $\lambda$  is a tunable parameter called the regularization parameter. The surface of the  $L_1$ -norm penalty in  $\boldsymbol{\beta}$  space is pointed on each axis, thus favoring sparse solutions (see Figure D.1 for illustration).  $\lambda$  controls the relative importance of the  $L_1$  norm of  $\boldsymbol{\beta}$ . A larger  $\lambda$  selects less descriptors and *vice versa*. The  $L_1$  norm penalizes a large number of regression coefficients and sets coefficients of insignificant descriptors to zero. The parameter  $\lambda$  is typically chosen such that the cross-validation error is minimized (see below). Subgraphs with non-zero coefficients are the LASSO-selected graph descriptors. We perform LASSO optimization using the implementation in MATLAB<sup>TM</sup>. For the QM9 set, we perform LASSO using parallel block coordinate descent method in C++ for efficient regression,[298] as QM9 set regression involves a highly sparse  $120,496 \times 110,498,191$  data matrix (data available at <http://files.ccei.udel.edu/p/QM9SubgraphData/>). The workflow of the graph selection using LASSO is presented in Figure 5.3.

LASSO can be seen as a feature selection machine learning type of algorithm. The  $L_1$  regularization is chosen to encourage sparsity in the parametric fit and the extension to the graph mining is novel. As a result, the LASSO algorithm used here

differs from the textbook-like LASSO. As compared to neural networks, the method does not involve intuition to tune parameters. Below, we compare its accuracy to other methods.

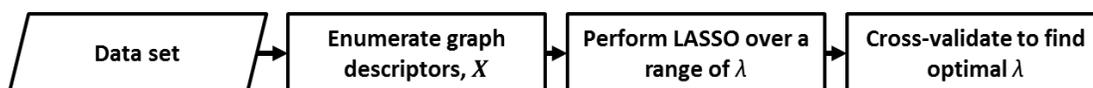


Figure 5.3. Workflow of the graph selection method using LASSO.

### 5.2.3 Semi-Supervised Learning

Selecting optimal graphs for LASSO regression requires complete enumeration of all subgraphs. However, exhaustive enumeration scales exponentially with the size of the molecule, as shown in Figure 5.4. For example, for icosane ( $C_{20}H_{42}$ ,  $x = 18$  in Figure 5.4), there are  $10^{12}$  unique subgraphs, and storing the subgraph frequencies requires 10 PB of memory. To overcome the complete enumeration challenge, the Least Angle Regression (LAR)-LASSO method,[299] a method analogous to LASSO, is combined with a frequent subgraph mining algorithm called gSpan.[300, 301] Here, the set of subgraphs is pruned during regression to prevent enumerating all subgraphs. The pruning method involves computing an upper bound score of not-yet-enumerated subgraphs and suppressing the enumeration of subgraphs with a low score. This simultaneous dimension reduction of descriptor space while conducting regression is called *semi-supervised learning*. The pruning method in previous semi-supervised learning methods has only been applicable to logical graph descriptors (the descriptor value is 1 if the subgraph appears at least once or 0 if not). Modification of the pruning method is required for graph frequency descriptors.

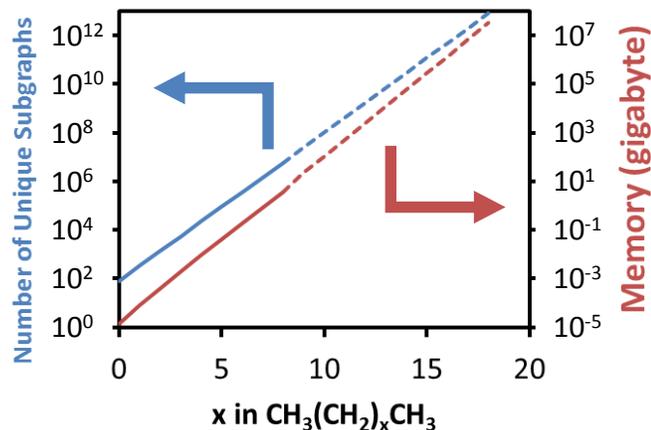


Figure 5.4. Number of unique subgraphs and descriptor vector memory vs. number of secondary carbons in alkanes. The solid and dashed lines indicate explicitly computed and extrapolated (estimated) data, respectively. The Spyder python software[302] is used for memory estimation.

LAR-LASSO starts at the highest possible  $\lambda$  that chooses the first descriptor (Figure 5.5), and,  $\boldsymbol{\gamma}$ , the direction of  $\boldsymbol{\beta}$  along  $\lambda$ , is set as the equiangular direction of  $\mathbf{X}$ , that is  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}$ , [299] where  $\mathbf{s} \in \mathfrak{R}^P$  is the vector of 1 and -1, computed by applying the sign function to  $\mathbf{r}^T \mathbf{x}_t$ . Then,  $\lambda$  decreases and  $\boldsymbol{\beta}$  is tracked. While  $\lambda$  decreases, a descriptor is either added or removed. The former occurs when a descriptor vector is as correlated to the model residual as  $\boldsymbol{\gamma}$  is. This descriptor is added to the active set,  $\mathbf{A}$ , of non-zero coefficients. If one of the active descriptor coefficients becomes zero while moving in  $\boldsymbol{\gamma}$ , then that descriptor is removed. Upon change in the number of descriptors,  $\boldsymbol{\gamma}$  is updated as  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}$ . [299] This process is repeated until the number of coefficients is equal to the number of data and the algorithm is terminated. This algorithm is illustrated in Figure D.2.

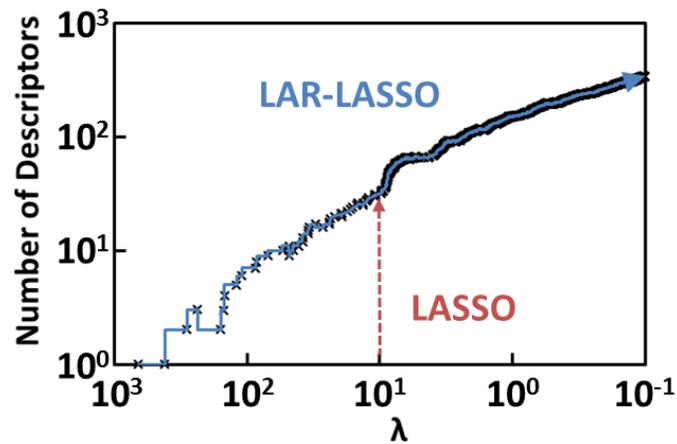


Figure 5.5. Illustration of the difference between LASSO and LAR-LASSO. Here,  $\times$  indicates the addition or removal of a descriptor. LAR-LASSO moves along the path of  $\lambda$  and tracks the additions or removals, whereas LASSO computes the regression coefficients at a single value of  $\lambda$ . The dashed line shows an example of the single LASSO solution obtained at  $\lambda = 10$ .

Taking infinitesimal steps to traverse along  $\lambda$  is impractical. Instead, the step sizes for all potential moves are computed and the smallest step size is selected. The practical implementation of LAR-LASSO is shown in the pseudo-code (Algorithm 5.1). Here,  $\mathbf{x}_t \in \mathfrak{R}^N$  is the frequency vector of the descriptor  $t$ . In line 1, the algorithm searches through the descriptor space and finds the first descriptor, and in the next line, it adds it to the active set, and  $\boldsymbol{\gamma}$  is set. Lines 4 and 5 find the minimal step sizes for addition and removal of a descriptor, respectively. The step in  $\boldsymbol{\beta}$  is taken in line 7 (technically, the step size,  $\Delta\lambda$ , is equal to  $d \times N$ ). The direction is updated as equiangular direction (line 10) and this process is repeated until the number of selected descriptors is equal to the number of data.

---

**Algorithm 5.1.** Semi-Supervised LAR-LASSO Algorithm.

---

```
1: Find descriptor  $t$  with  $\max |Y^T x_t|$  (Algorithm 5.2)
2:  $A = \{t\}, \beta = 0, s = \text{sgn}(Y^T x_t)$ , add  $x_t$  to  $X$ ,  $\gamma = (X^T X)^{-1} s$ 
3: while  $P \neq n$  do
4:   Find descriptor  $t$  with minimal  $d_t$  and set to  $d_1$  (Algorithm 5.2)
5:    $d_2 = \min\{d > 0: (\beta + d\gamma)_i = 0, i \in A\}$ 
6:   Find step size:  $d = \min(d_1, d_2)$ 
7:   Take step:  $\beta \leftarrow \beta + d\gamma$ 
8:   If  $d = d_1$ , add  $t, x_t, 0, \text{sgn}(r^T x_t)$  to  $A, X, \beta, s$ 
9:   If  $d = d_2$ , remove  $i^{\text{th}}$  entry from  $A, X, \beta, s$ 
10:   $\gamma = (X^T X)^{-1} s$ 
11:end while
```

---

The determination of the step size for adding a descriptor  $t$ ,  $d_t$ , is derived through the Karush-Kuhn-Tucker conditions (see ref[299] for derivation), which result in

$$d_t = \min_+ \left\{ \frac{x_i^T r - x_t^T r}{x_i^T v - x_t^T v}, \frac{x_i^T r + x_t^T r}{x_i^T v + x_t^T v} \right\}, t \notin A, i \in A, \quad (5.5)$$

where  $v = X\gamma$

and  $\min_+$  stands for the operation of taking minimum among positive elements for forward traversing in  $\lambda$ .

The key to the semi-supervised learning is in finding the descriptor with minimal  $d_t$  (lines 1 and 4 in Algorithm 5.1) where the subgraph search space reduction occurs. The search through the subgraph search space is performed via gSpan[300] that organizes graph descriptors in a tree (a meta-graph) where each node represents a unique graph descriptor and a child node is a supergraph built from its parent node by adding one more edge. The algorithm visits each graph descriptor node from a root (single vertex graphs) to branches using a recursive algorithm where each recursion represents an addition of an edge. An example of gSpan meta-graph is shown in Figure 5.6.

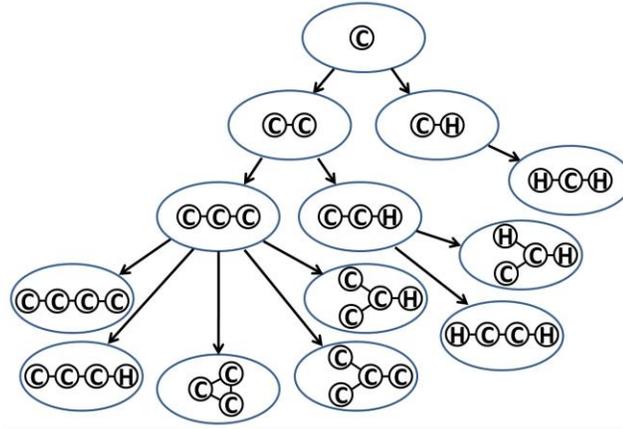


Figure 5.6. Example of gSpan’s meta-graph enumeration (up to adding 3 edges). The algorithm employs mathematical techniques to avoid generating duplicate nodes.

In a previous gSpan-based LAR-LASSO algorithm,[301] each node was visited recursively and  $d_1$  was updated when  $d_t$  with a lower value was found. To reduce the subgraph search space, the enumeration of supergraphs is suppressed by computing the lower bound of  $d_t$  for the supergraphs. This method is called branch pruning as the generation of the child nodes is suppressed. This pruning method has been previously applied to logical graph descriptors.[301, 303-305] For a logical descriptor, if graph  $t'$  is a supergraph of  $t$  (denoted by  $t' \supseteq t$ ), then  $x_{it'} \leq x_{it}$ , and this provides a pruning condition[301]

$$b_r + d_1 b_v < |\mathbf{x}_i^T \mathbf{r}| - d_1 |\mathbf{x}_i^T \mathbf{v}|, i \in \mathbf{A}, \text{ where}$$

$$b_r = \max \left\{ \sum_{r_i < 0} |r_i| x_{it}, \sum_{r_i > 0} |r_i| x_{it} \right\}, \quad (5.6)$$

$$b_v = \max \left\{ \sum_{v_i < 0} |v_i| x_{it}, \sum_{v_i > 0} |v_i| x_{it} \right\}$$

Here,  $r_i$  and  $v_i$  are the elements of the vector  $\mathbf{r}$  and  $\mathbf{v}$  for the datum  $i$ , respectively. If equation (5.6) is satisfied, then,  $d_{t'} < d_1$  for all  $t' \supseteq t$ . [301] In describing thermochemistry, graph frequencies are used as  $x_{it}$ . [52, 202-206, 251-259] In this case,  $x_{it'} \not\leq x_{it}$ . For example, the C-C graph is observed six times in benzene in Figure 5.2, but its supergraph, H-C-C, occurs 12 times. Thus, the assumptions of equation (5.6) do not hold. We modify equation (5.6) to provide an approximate pruning condition. Here, we keep track of variable  $x'_{it}$  that contains the maximum frequency among all the supergraphs of  $t$  enumerated so far for the datum  $i$ . Then, we replace  $x_{it}$  with  $x'_{it}$  in equation (5.6). Algorithm 5.2 shows the pseudo-code for this modified search method. Here,  $d_1$  is initially set to infinity (line 2), and the algorithm recursively visits each node;  $x'_{it''}$  where  $t'' \subseteq t$  is updated in lines 10-14;  $d_1$  is set to  $d_t$  if a lower  $d_t$  is found. Then, the pruning condition is checked (line 21) and if it holds, the recursive function is returned, which prevents enumeration of supergraphs of  $t$ .

---

**Algorithm 5.2.** Modified gSpan graph search

---

```

1: procedure Subgraph Search
2:    $d_1 = \infty$ 
3:   for  $t \in$  Single vertex graphs do
4:     gSpan search( $t$ )
5:   end for
6:   return  $d_1, t$ 
7: end procedure
8: function gSpan search ( $t$ )
9:    $x'_{it} = x_{it}$ 
10:  for  $t' \subseteq t$  do
11:    if  $x'_{it'} < x'_{it}$  then
12:       $x'_{it'} = x'_{it}$ 
13:    end if
14:  end for
15:  if  $t \notin A$ ,

```

---

---

```

16:   Compute  $d_t$  using equation (5.5)
17:   if  $d_t < d_1$  then
18:      $d_1 = d_t$ 
19:   end if
20: end if
21: if pruning condition equation (5.6) is true then
22:   return
23: end if
24: Enumerate  $t''$  using the gSpan algorithm
25: for  $t'' \supseteq t$  do
26:   gSpan search ( $t''$ )
27: end for
28:end function

```

---

For the initial descriptor search, the following equation is used instead of equation (5.5) in line 10:[301]

$$d_t = \left| \sum_i y_i x'_{it} \right| \quad (5.7)$$

The pruning condition, equation (5.6), is also replaced with[301]

$$d_1 > \max \left\{ \sum_{y_i < 0} |y_i| x'_{it}, \sum_{y_i > 0} |y_i| x'_{it} \right\} \quad (5.8)$$

We have implemented this algorithm in C++ and made it available at <https://github.com/vlachosgroup/LARLASSO-gSpan>.

#### 5.2.4 Cross Validation

We assess the predictive ability of LASSO models using leave-one-out cross validation.[234, 235] Here, the LASSO regression is performed without one datum and the response value of the leftover datum is predicted using the trained model. This process is repeated for the entire data set to obtain the predicted value for each datum. For LAR-LASSO and the QM9 data set, we perform hold-out cross-validation where 90% of the data is used for regression and the rest is used for testing. Despite the

pruning, the LAR-LASSO regression is computationally expensive for the entire NIST's and BURCAT's data set. Thus, hold-out cross-validation is performed to reduce computational time.

### **5.3 Result and Discussion**

#### **5.3.1 Comparison of LASSO and Group-Additivity Models**

In order to understand the potential of LASSO, we obtain cross validation errors for various LASSO models trained using the exhaustively enumerated subgraph space focusing on the gas-phase dataset. To reduce computational time, we select data with less than 12 atoms. This results in 217 data points out of 695 data points. Figure 5.7 shows the cross validation results for two values of  $\lambda$ , namely when the number of parameters,  $P$ , is the same to that of the Benson's GA model[306] (for a fair comparison) and for the optimal  $\lambda$  at which the cross validation error is minimal. Figure 5.7 depicts also the cross-validation error of the graph selection method using Morgan fingerprints at optimal  $\lambda$ . [307] Morgan fingerprinting generates a tree graph rooted at each atom at a chosen depth – the maximum number of edges from the root vertex considered.

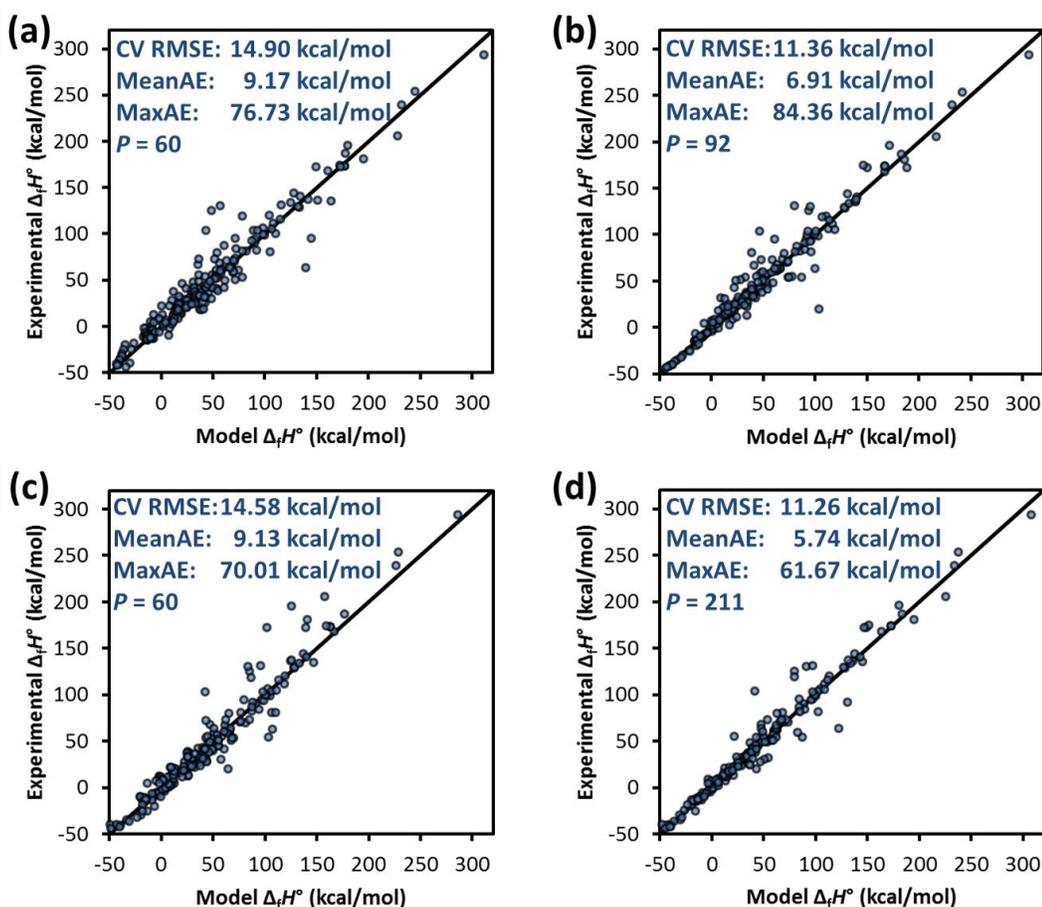


Figure 5.7. Parity plot between experimental  $\Delta_f H^\circ$  and cross-validation predicted  $\Delta_f H^\circ$  computed using (a) the Benson's gas group additivity,[306] and various LASSO-trained models (b)-(d): (b)  $\lambda$  is chosen using the Morgan fingerprints[307] so the cross validation mean square error is minimal, (c)  $\lambda$  is chosen so the number of descriptors,  $P$ , is similar to that of group additivity, and (d)  $\lambda$  is chosen so the cross validation mean square error is minimal. The Morgan fingerprints and the exhaustively enumerated subgraph space consist of 835 and 16513 descriptors, respectively. CV RMSE is the cross validation root mean square error, and MeanAE and MaxAE are the mean and max absolute error, respectively.

Figure 5.7 demonstrates that the LASSO-trained models (panels b, c, and d) possess lower CV RMSE than the Benson's GA (panel a). As expected, the model

trained using all possible graph descriptors at optimal  $\lambda$  (Figure 5.7d) performs the best. The LASSO-trained model with the same number of parameters (non-optimal  $\lambda$ ; Figure 5.7c) performs slightly better than the Benson's GA model indicating that LASSO selects comparable graph descriptors. The model trained using the Morgan fingerprint space performs better than the Benson's GA model but worse than the model trained with the exhaustively enumerated subgraph space, signifying that Morgan fingerprint space misses important subgraphs. However, the model trained using the Morgan fingerprint space performs better than the model trained using exhaustive graph space with the same number of parameters (CV RMSE 11.36 vs. 11.75 kcal/mol). This indicates that performing regression by exhaustively enumerating the graph space results in the selection of non-true descriptors (overfitting), and the Morgan fingerprint more selectively enumerates descriptive subgraphs. Therefore, developing a graph space enumeration method that selectively enumerates descriptive graphs has a potential to improve this framework.

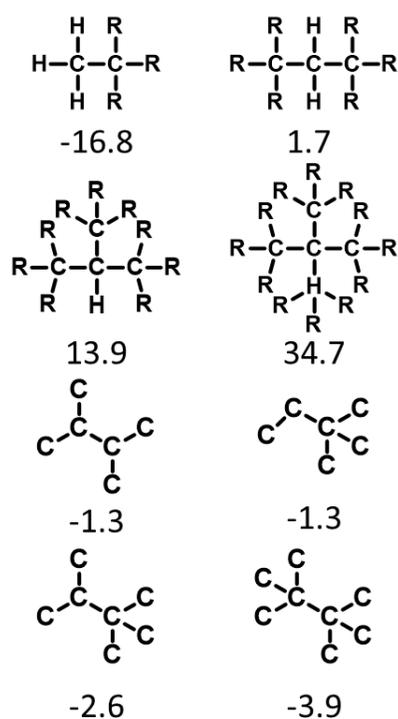
All models have significant errors in multi-cyclo and radical molecules such as benzvalene, HCCC, cyclopropenyl, cyclopentadienyl, Bicyclo[1.1.0]but-1(3)-ene, and cyclobutadiene. These molecules contain ring and electronic subgraph patterns not observed elsewhere in the data set (outliers in descriptor space[289]), making the development of a predictive LASSO-trained model impossible unless similar molecules are introduced into the dataset. Part of the errors is attributed to the lack of sufficient data points as the learning curve in Figure D.3a shows that the test and training set RMSEs have not fully converged.

Figure 5.7d shows that the LASSO method can outperform the conventional, intuitively-based GA scheme. The suboptimality of the Benson's GA descriptors is

shown by the unexpected sign of the corrections' contribution. The graph descriptors in third and fourth rows of Figure 5.8a represent the corrections for gauche interactions. The contributions for these corrections are positive (destabilizing) in the Benson's GA, but negative (stabilizing) in the present work. We suspect that these gauche interaction corrections are compensating for missing radical and ring corrections. Due to the missing corrections,  $\Delta_f H^\circ$  of radical and ring molecules are under-estimated, which, in return results in over-estimation of other molecules such as closed-shell, non-ring molecules.

The LASSO-selected graphs (Figure 5.8b) differ from those of GA (Figure 5.8a). In LASSO, single atom graphs likely account for the formation energy from each elements' standard state to a reference state. Graphs with 2 and 3 vertices resemble C-H, C-H<sub>2</sub>, and C-H<sub>3</sub> bonds and likely account for bond energy between atoms. Graph descriptors that focus on bond energies rather than an atom's environment resemble those of the so-called bond-centered GA.[281] Finally, larger graphs are conceptually similar to the GA's corrections and have a small contribution.

### (a) Group Additivity



### (b) LASSO Selected

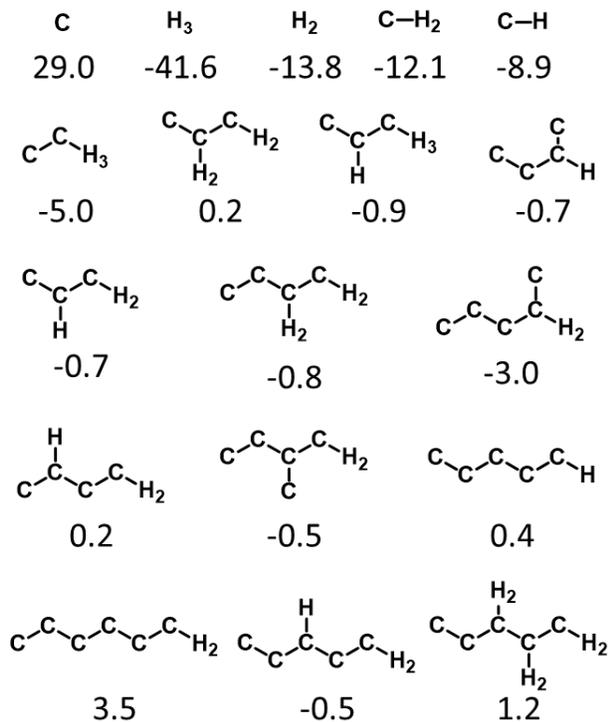


Figure 5.8. Graph descriptors (a) used in group additivity and (b) selected by LASSO for  $P = 60$ . Only graphs related to alkanes are shown. R indicates any atom. The value below a graph indicates the graph's contribution to  $\Delta_f H^\circ$  in kcal/mol.

### 5.3.2 Subgraph Search Space Reduction and Performance of Semi-Supervised Learning

To assess the effectiveness of the semi-supervised learning method, we compare the computational time and the number of enumerated graph descriptors of the exhaustive enumeration method and the semi-supervised learning method using the entire hydrocarbon dataset. We control the subgraph space size by limiting the maximum number of edges in subgraphs during the enumeration as shown in Figure 5.9. Unlike the exponential explosion in subgraphs when complete enumeration is

done, the computational cost for the semi-supervised method plateaus, enabling the extension of the method to larger molecules.

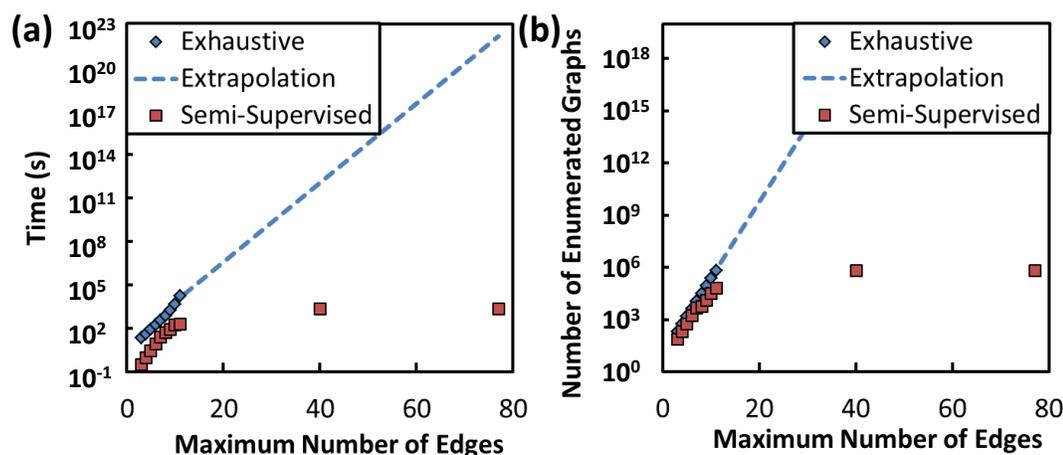


Figure 5.9. (a) Computation time and (b) number of enumerated graph descriptors vs. maximum allowed edges in subgraphs for the exhaustive enumeration method and the semi-supervised learning using the dataset of 695 hydrocarbon molecules. The dashed line indicates the extrapolation of the exhaustive enumeration method. The semi-supervised learning is performed until  $\lambda = 0.1$ . The exhaustive enumeration was performed using the Rdkit[292] C++ package (python API), and the semi-supervised learning performed in C++.

The semi-supervised learning model and the Benson's GA scheme are compared for the entire 695 hydrocarbon data set in Figure 5.10. Overall, the former performs better than the latter. Both models exhibit significant errors for outliers in descriptor space, such as [18]-Annulene, 9,10-dihydro-Phenanthrene, Trispiro[2.0.2.0.2.0]nonane, and Tetraspiro[2.0.0.0.2.1.1.1]undecane as well as a number of radicals. In order to assess the performance of the approximate pruning condition, we apply the semi-supervised learning to the 217 hydrocarbon data set and

compare to complete enumeration. The RMSE, MeanAE and MaxAE at optimal  $\lambda$ , (13.94, 6.38, and 72.06 kcal/mol, respectively) vs. (11.26, 5.74, and 61.67 kcal/mol), are comparable.

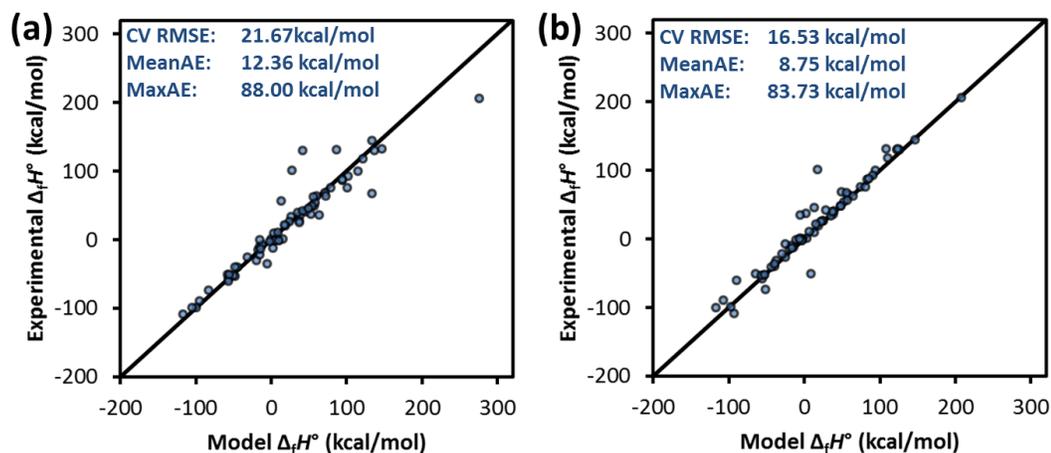


Figure 5.10. Parity plot between experimental data and prediction of (a) Benson's group additivity and (b) the semi-supervised model at optimal  $\lambda$  for the 695 hydrocarbon data set.

### 5.3.3 Comparison to Machine Learning Models Using the QM9 Data Set

In predicting chemical properties, several machine learning models have been introduced using the data set QM9. Most widely used regression models are kernel ridge regression,[308-316] neural network[309, 317-319] and convolutional network[320-323] with molecular representations such as bag of bonds,[311] Coulomb matrix,[309, 312, 313, 315, 318] BAML,[311] SOAP,[316] and many others.[308, 310, 317, 319, 320, 323, 324] Similar to this work, a couple of models are based solely on molecular graphs. Li et al. presented gated graph neural networks whose inputs are node and edge features and an adjacency matrix. Each atom has

hidden representation which is updated using its feature as well as the neighboring edge and atom features until convergence is achieved. The parameters of the update functions are learned to correlate with the data.[319] Similarly, Kearnes et al. and Duvenaud et al. presented molecular graph convolution neural networks where atom and bond features undergo transformation to an abstract representation of atoms and bonds. The abstract atoms are, then, correlated to data using a function of one's choice.[321, 322] Recently, Faber et al. implemented both methods to build models for 0 K atomization energy, achieving 3.46 kcal/mol and 0.97 kcal/mol MeanAE, respectively.[323]

Compared to published state-of-the-art methods involving abstraction of atoms and bonds in molecular graphs, our approach lies on finding optimal descriptors that possess more theoretical basis. Bader et al. discusses that the groups in group additivity describe structures transferable among molecules.[52] Using DFT, transferable structures are found to have similar electron density profiles, resulting in the transferable energy of groups according to the virial theorem. In a similar spirit, this framework operates under the assumption that, if a subgraph implicitly describes a structure consistent over all molecules in the data set, their number of occurrence in a molecule linearly scales with the contribution of this subgraph's structure to the energy of the molecule.

To compare the performance directly with state-of-the-art machine learning models, we fully enumerated subgraphs and performed LASSO on the QM9 data set. Validation results are shown in Figure 5.11. Here, the MeanAE is 1.39 kcal/mol, comparable to some state-of-the-art and better than several published machine learning models. In addition, our model offers high interpretability as well as

accessible computation of uncertainty (a topic beyond the scope of this work), as the literature of statistical inference for LASSO is ample.[325-331] On the other hand, model inference of kernel ridge regression has been only studied in the case of heteroscedastic regression,[332] and leading efforts to quantify uncertainty for deep learning frameworks, which involve generalization of deep learning and Bayesian statistics, are currently in embryonic stage.[333-335]

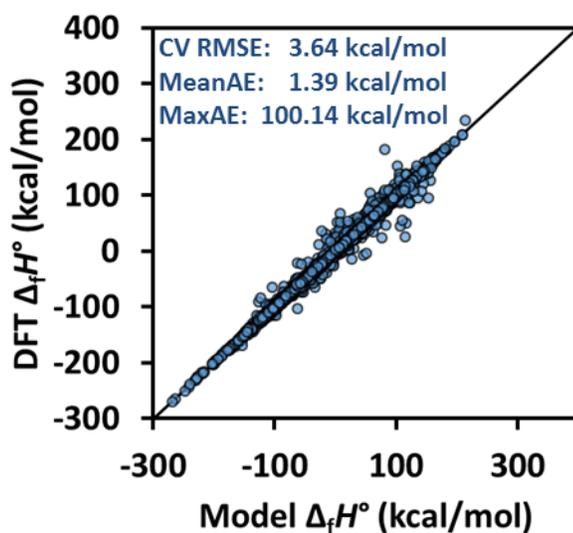


Figure 5.11. Parity plot between DFT data and predicted data using the subgraph selection method for 10 % of the QM9 data set.

#### 5.3.4 Application to Surface Adsorbates

We use the adsorbate graph mining algorithm to enumerate all possible adsorbate subgraphs and perform LASSO to three data sets taken from our prior work (lignin monomers[21], furanics[16], and linear oxygenates[23]). The parity plot for the lignin monomer dataset is shown in Figure 5.12 (for furanics and linear oxygenates,

see Figure D.4). Selected subgraphs of benzene and its hydrogenated derivatives are shown for illustration in Figure 5.13.

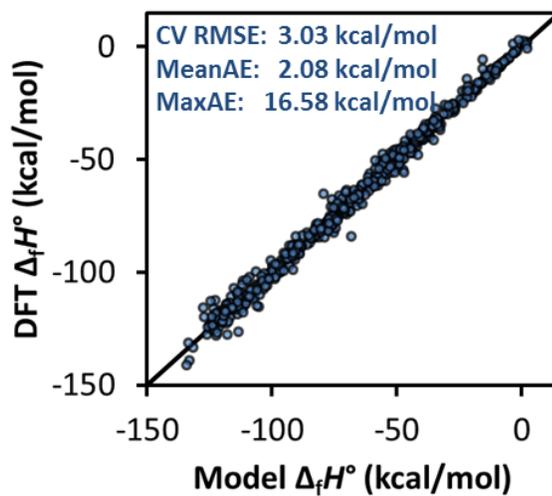


Figure 5.12. Parity plot between DFT-computed[21] and predicted  $\Delta_f H^\circ$  for the lignin monomer dataset.

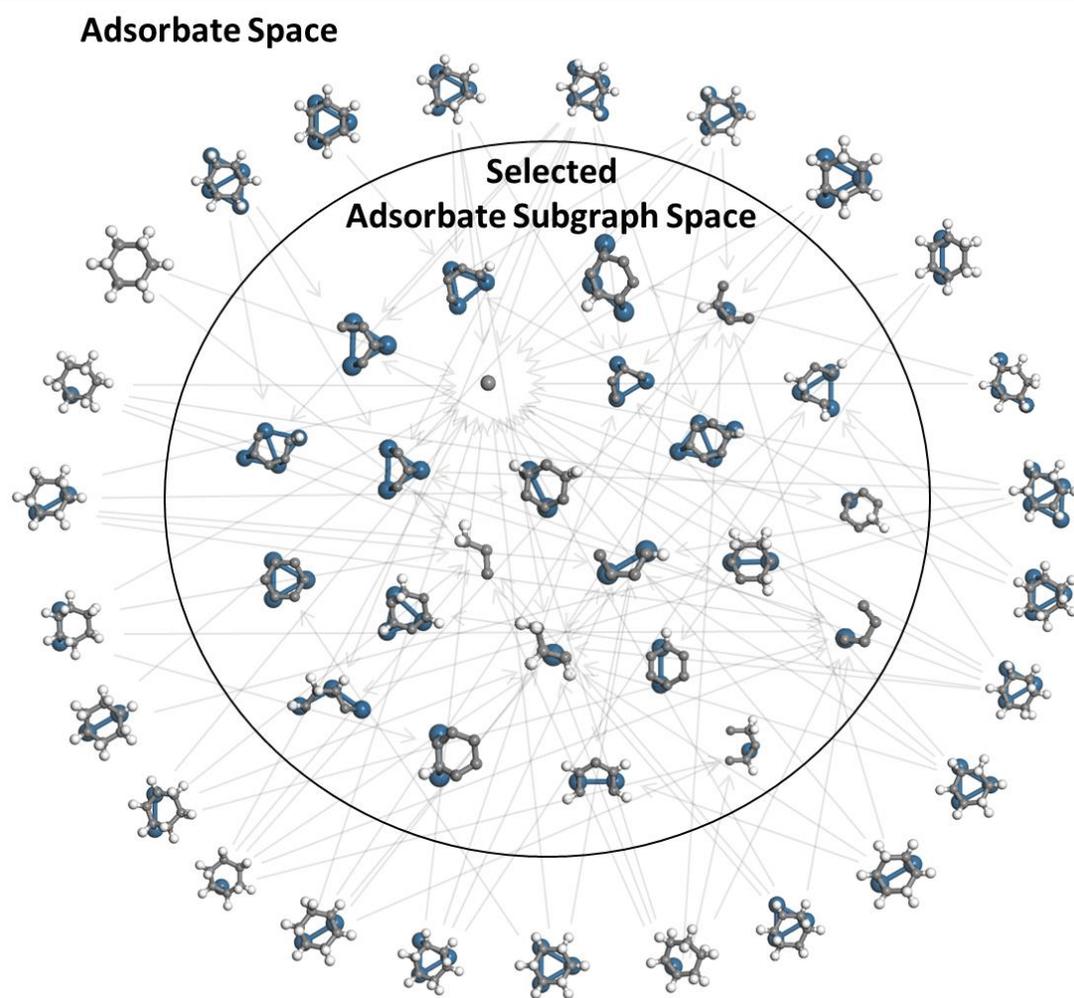


Figure 5.13. Molecular graphs and subgraphs of benzene and its hydrogenated derivatives. The arrows indicate the subgraphs encountered in each molecule.

Figure 5.12 demonstrates the great performance of the adsorbate fingerprint combined with LASSO (MeanAE and MaxAE of 2.08 and 16.58 kcal/mol, respectively). The cross-validation statistics for the other data sets are shown in Table 5.1. Despite the great performance for lignin adsorbate dataset, the graph enumeration

combined with LASSO performs poorly for the furanics and linear oxygenate datasets, as shown in Table 5.1. We attribute this to the large number of outliers in the descriptor space due to poor diversity in the data and the low number of data. To illustrate this fact, we have performed classical multi-dimensional scaling or principal coordinates analysis (Figure 5.14).[336] In this method, coordinates  $\in \mathfrak{R}^N$  for each datum are initially randomly set and are optimized to match the distance between data points specified by a distance matrix, where each element contains a numerical measurement of the “dissimilarity” between data. Then, similar to the principal component analysis, the eigenvectors in the coordinate space are found, and the two dimensions that have the maximal variance are plotted for visualization. In this work, the dissimilarity measure between data  $i$  and  $j$ ,  $D_{ij}$ , is computed by

$$D_{ij} = \left( \sum_{t \in A} \left| \frac{O_{it} - O_{jt}}{n_t^\alpha} \right|^\varepsilon \right)^{\frac{1}{\varepsilon}} \quad (5.9)$$

$$\text{where } O_{it} = \begin{cases} 1, & x_{it} > 0 \\ 0, & x_{it} = 0 \end{cases} \quad n_t = \sum_{i=1}^N O_{it}$$

Here,  $O_{it}$  is the logical graph descriptor,  $n_t$  is the number data containing graph  $t$ ,  $\alpha$  is the tunable parameter for the effect of division by  $n_t$ , and  $\varepsilon$  is the order of Minkowski distance. This dissimilarity is a variation of the classical Euclidean distance or  $\left( \sum_{t \in A} |O_{it} - O_{jt}|^2 \right)^{1/2}$ . The modifications to Euclidean distance are made to exaggerate the distance of data containing descriptors with small  $n_t$  as these data tend to have high cross-validation error due to the biased descriptor coefficient. The division of  $O_{it} - O_{jt}$  by  $n_t^\alpha$  enlarges the distance if the given descriptor is not frequently observed in the data set. A high  $\varepsilon$  value amplifies large  $(O_{it} - O_{jt})/n_t^\alpha$ . Here, we use  $\alpha$  and  $\varepsilon$  of 0.5 and 4, respectively, and plot the optimized coordinates in

two eigenvectors with the highest variance, as shown in Figure 5.14. A highly dense data set is desired to avoid outliers and have high  $n_t$  for each descriptor. Such is the case for the lignin adsorbate dataset (Figure 5.14a). However, for the furanics and linear oxygenates, several data points are far away from other data points (i.e., outliers in descriptor space) and have high cross-validation errors. These datasets are not well suited for LASSO training.

The furanics and linear oxygenates contain a large number of outliers. The previous group additivity model for the lignin data set uses binding-site resolved descriptors.[21] As the graph mining method considers the binding site via the adsorbate atoms' connectivity to surface atoms, the lignin data set is well-sampled in the descriptor space. On the other hand, descriptors in the group additivity models for the furanics and linear oxygenate data sets consider the number of electrons interacting with the surface instead of the binding site. Thus, these data sets are not well sampled in the descriptor space.

In sampling the descriptor space, two types of outliers are observed: (1) molecules exhibiting unique characteristics and patterns in the entire descriptor space, such as  $H_2$ , CO and  $CO_2$ ; and (2) molecules with patterns scarce in the data set (poor sampling). Building a model for the former molecules is impossible as these data appear only once. Examination of outliers in the furanics and linear oxygenates sets reveals that the data set contains outliers of both types. Figure D.5a and b shows outliers in the furanics and linear oxygenate data sets, respectively, which are highly unsaturated species and possess rare structures due to complex connectivity patterns. The cluster at the bottom right corner of Figure 5.14 represents unsaturated hydrocarbons shown in Figure D.5b. Cross-validation errors for these hydrocarbons

can be improved by sampling *larger* unsaturated molecules of similar adsorption conformation. On the other hand, Figure D.5c and d shows relatively saturated adsorbates. These adsorbates are unstrained and are expected to contain structures transferable to many adsorbates in the descriptor space even for the same size. The issue here is poor data sampling. Identification of outliers in the entire descriptor space is beneficial to improve sampling.

The furanics and linear oxygenate data set also lack a sufficient number of data points, resulting in high errors as demonstrated by the learning curve (Figure D.6). A large gap between the test and training set RMSEs indicates overfitting of a model and a low training set RMSE indicates the high complexity of a model. Figure D.6a shows that the selected model is simple as the training set RMSE is high, but a large gap is still present between the training and test set RMSEs. This indicates overfitting of a relatively simple model. Since the LASSO framework balances over- and under-fitting, the graph signifies lack of data points to support a more complex model. On the other hand, Figure D.6c shows that the selected model for lignin data set is complex enough to keep the training set RMSE low while having a lower test set RMSE due to the ample data. The relatively small number of data in these sets is due in part to the large computational cost of DFT for surface species as compared to gas-phase species.

In order to remove outliers in descriptor space, data points are filtered using a threshold of the sum of distance. The correlation between the sum of distance and the model absolute error as well as the threshold is shown in Figure 5.15. The LASSO performance without outliers for linear oxygenate and furanics data sets are shown in the last two rows of Table 5.1. The performance of LASSO significantly improves for

both data sets. Especially, the performance for linear oxygenates without outliers exceeds that of group additivity. Thus, the development of a reliable outlier detection method poses great potential for LASSO fitting improvement as well as data sampling.

Table 5.1. Comparison between the performance of the previous group additivity models and the graph selection method.[16, 21, 23]

Model	MeanAE (kcal/mol)	MaxAE (kcal/mol)	RMSE (kcal/mol)
Lignin; GA	2.72	15.09	3.61
Furanics; GA	5.76	31.61	8.51
Linear oxygenates; GA	5.75	89.10	10.86
Lignin; LASSO	2.08	16.58	3.03
Furanics; LASSO	9.06	60.40	14.87
Linear oxygenates; LASSO	6.51	46.55	9.67
Furanics; LASSO without outliers	7.73	30.07	10.66
Linear oxygenates; LASSO without outliers	3.79	30.22	7.56

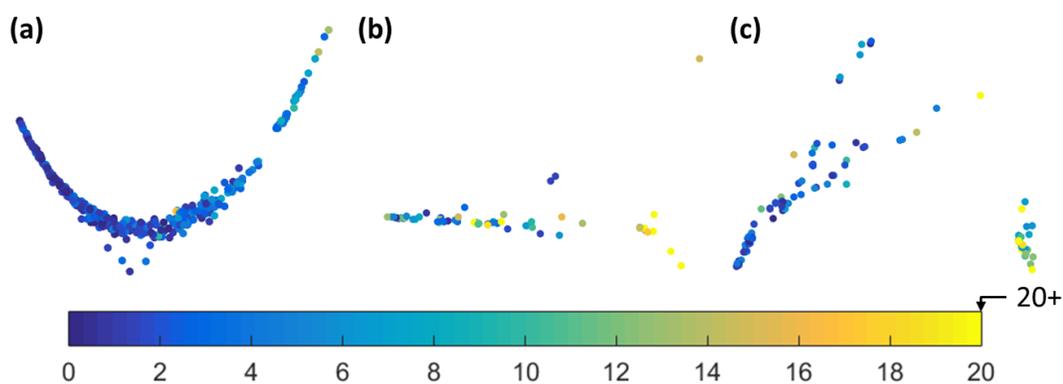


Figure 5.14. Principal coordinates analysis plot projected to 2D using the distance measurement shown in equation (5.9) for (a) the lignin, (b) the furanics, and (c) the linear oxygenates adsorbate datasets. The color scheme represents the absolute error in  $\Delta_f H^\circ$  (kcal/mol) where data points with an absolute error higher than 20 kcal/mol are the same color as 20 kcal/mol error.

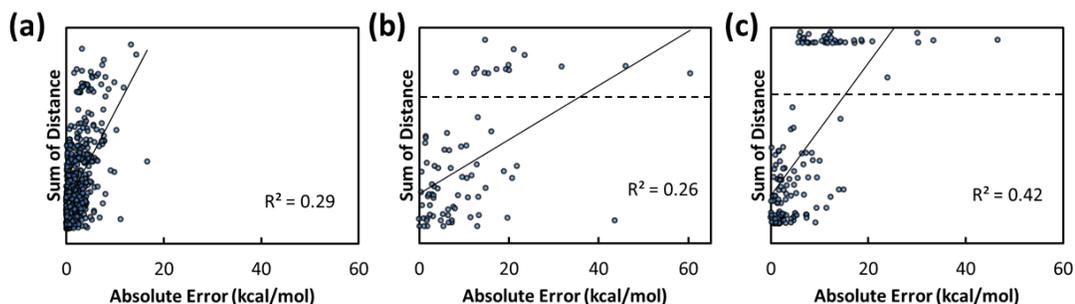


Figure 5.15. Correlation between absolute errors and the sum of the distance of each data for (a) lignin, (b) furanics, and (c) linear oxygenates adsorbate datasets. The dashed line indicates the threshold values used to filter outliers.

## 5.4 Conclusions

In this work, we have exploited feature selection concepts to improve accuracy and automate descriptor selection as an alternative to the traditional group additivity scheme for predicting thermochemistry of gas phase species and surface adsorbates.

We considered four different datasets: (1) gas-phase hydrocarbons from the NIST webbook and BURCAT’s thermodynamic databases, (2) gas molecules in the QM9 dataset, and surface adsorbates consisting of (3) lignin monomers, (4) furanics, and (5) linear oxygenates. The LASSO trained model using all possible subgraphs for a small subset of the gas dataset demonstrates the high potential of this framework in automatically selecting subgraphs and improving accuracy. We find that the graphs selected by LASSO resemble the bond centered group additivity model.[281] The computational time and the memory of exhaustive graph enumeration scale exponentially with the number of vertices in molecular graphs. In order to reduce the cost of graph enumeration, we have presented a semi-supervised learning framework based on LAR-LASSO, which performs similar to the rigorous exhaustive enumeration LASSO method but its cost no-longer scales with the molecular graph size. This framework has been also compared to state-of-the-art machine learning models for the popular QM9 data set. The framework shows comparable accuracy to several published machine learning models without the need to intuitively tune parameters as often happens in neural networks. We have also introduced an effective adsorbate graph mining algorithm that, given a healthy data set, possesses improved accuracy. We also presented a simple distance measurement to identify outliers in descriptor space, which may be removed to improve fit, or similar data to be additionally sampled.

## 5.5 Acknowledgements

G. G. is grateful to Professor Dominique Guillot, Professor Markos Katsoulakis, Marcel P. Nunez and Jonathan Lym for useful discussions. This research was supported by the Defense Advanced Research Project Agency under grant

W911NF-15-2-0122. The semi-supervised learning code as well as the python package that converts atomic coordinates to molecular graph and enumerate adsorbate subgraphs are available at <https://github.com/vlachosgroup>.

## Chapter 6

### PREDICTING MULTI-DENTATE ADSORBATE CONFORMATION ON METAL SURFACES

#### 6.1 Introduction

Biomass conversion provides a renewable pathway to produce chemicals and fuels. While density functional theory (DFT) has provided insights into reaction mechanisms and catalyst design for conversion of small molecules on metal surfaces,[242-247] the computational cost of DFT is prohibitive for investigating large reaction networks, typically encountered when converting biomass molecules.[18, 21, 134, 248-250] In this respect, semi-empirical methods have been developed to estimate thermochemistry of adsorbates.[21-23, 34, 35, 216, 337] These studies have revealed that accounting for the adsorption mode ( $\sigma$  vs.  $\eta$ ) as well as the binding sites (one-, two-, three-fold sites) is crucial for estimating thermochemistry accurately.[21, 34, 337] While manually manipulating atomic coordinates to generate potentially stable adsorption conformations for small species is rather easy, it becomes increasingly tedious and slow to do so for large multi-dentate adsorbates that bind to a catalyst via many heteroatoms. In this regard, adsorption conformation sampling methods have been introduced.[338, 339] However, these methods use DFT as ‘a calculator’ for sampling minima on potential energy surfaces, and thus, they are unsuitable for large mechanisms.[18] Developing an adsorption conformation prediction tool can enable large reaction-network modeling.

Recently, molecular graphs of every possible closed-shell organic molecule with up to 17 heavy atoms (among C, O, F, and N) have been enumerated.[290] Such large chemical space has resulted in large DFT data sets, with the most popular one being the QM9.[288] This dataset involves 134,000 molecules and has resulted in the development of several machine-learning models whose accuracy has been found to be better than that of hybrid DFT methods.[310, 320, 323] Development of such accurate models for surface species is lacking owing to two reasons. First, molecular graphs cannot be automatically created given that adsorbates can take multiple conformations on a surface and these can change with catalyst surface and catalyst itself. Second, the computational cost is high, leading to using lower accuracy DFT methods and smaller molecules consisting typically of up to 3 heteroatoms. Building a molecular graph enumeration tool, which accounts for the conformation of adsorbates, can overcome the first problem and facilitate advances in heterogeneous catalysis.

This Chapter introduces a machine-learning algorithm that learns from a dataset without human intervention, and automatically generates adsorption conformations for adsorbates on close-packed metal surfaces (111 for fcc, 0001 for hcp, and 110 for bcc metals). The dataset is derived from 808 DFT calculations performed herein, involving various adsorption conformations for carbon-based molecules on Pd(111). We find that the algorithm possesses promising accuracy.

## **6.2 Methods**

### **6.2.1 Representation of Adsorbate Conformation and Stability**

Adsorbate's conformations on a surface are represented using molecular graph theory concepts, where a vertex and an edge are abstract objects describing an atom

and the connectivity between a pair of atoms, respectively. A conformation may be stable or unstable. For example, gas phase cyclohexane's boat and half-chair conformations are unstable while the twisted boat and chair conformations are stable. The unstable boat and half-chair conformation relax to the stable twisted boat and chair conformations, respectively. Similarly, there are stable and unstable conformations for adsorbates. Describing conformations for adsorbates typically involves the binding site and the valency of a heteroatom. Heteroatoms of an adsorbate can potentially bind to one-, two-, or three-fold sites (called atop, bridge, or hollow sites, respectively). The valency is equal to  $x_{\max} - x$ , where  $x_{\max}$  represents the maximum number of ligands of an atom, and  $x$  is the actual number of ligands. For example,  $\text{CH}_3$  has a valency of  $4 - 3 = 1$ . The valency represents the number of electrons interacting with surface atoms. Heteroatoms with valence of zero do not bind to a binding site, due to the lack of electrons interacting with the surface. For example, carbon of  $\text{RCH}_3$  does not bind and sticks out of the surface. Hereafter, we focus on carbons atoms for illustration.

To describe adsorption of a heteroatom, we introduce the notation  $\text{C}_{\beta}^{\alpha}$ , where  $\alpha$  is the valency of a carbon and  $\beta$  denotes the type of binding site ( $\beta = \text{A, B, H, u}$  for atop, bridge, hollow or unoccupied site). For multiple heteroatoms, we employ a notation similar to the group additivity scheme. For example,  $\text{C}_{\text{A}}^1(\text{C}_{\text{B}}^1)_2(\text{C}_{\text{H}}^1)$  indicates that the central atom  $\text{C}_{\text{A}}^1$  (leftmost) is attached to two  $\text{C}_{\text{B}}^1$  atoms (middle) and one  $\text{C}_{\text{H}}^1$  atom (rightmost in the formula).

The stability of a single heteroatom is strongly affected by its valency and its binding site, as demonstrated in Figure 6.1 for a simple case of  $\text{CH}_x$  adsorption on

three binding sites. Heuristically, a single heteroatom prefers a site to create a number of bonds with metal atoms equal to its valency.

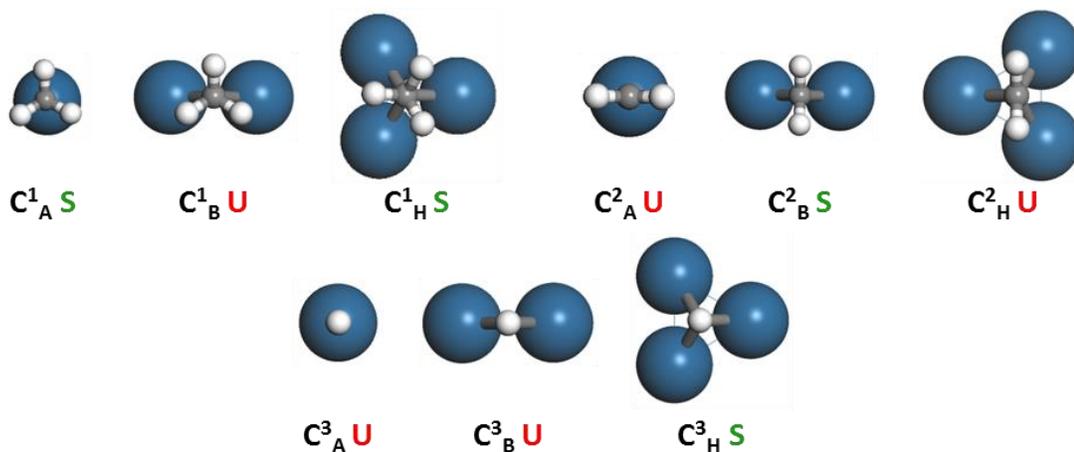


Figure 6.1. Carbon in  $CH_3$ ,  $CH_2$  and  $CH$  with varying number of valence electrons on atop, bridge and hollow sites. The superscript indicates the binding atom's valency and the subscript the type of binding site. S (U) indicate a stable (unstable) conformation.

The stability of a conformation is strongly affected also by the neighbor atoms' binding site and valency, as illustrated in Figure 6.2. Due to the mismatch between C-C and metal-metal bond distances and preferred bond angles, heteroatoms sometimes share surface atoms, as shown in the second conformation of  $C^1_A(C^1_A)$  in Figure 6.2. For the same reason, some conformations are strained compared to a single atom binding, as shown in the first conformation of  $C^1_A(C^1_A)$  row in Figure 6.2. As an example, the two  $C^1_A(C^1_A)_2$  conformations in Figure 6.2 are unstable due to the elongated C-C bond and large C-C-C angle (these instabilities are determined using DFT; see below). Another key observation is that  $C^2_A$  (Figure 6.1) is unstable on the

atop site as it sits on the saddle point of the potential surface, but it is stable when two neighboring atoms are binding to the surface, as shown in  $C^2_A(C^1_A)_2$  (Figure 6.2). The two binding neighbors change the potential energy surface, making the  $C^2_A$  conformation stable. What these examples demonstrate is that the stability of a single heteroatom (see Figure 6.1) is insufficient to predict the stable conformations of an adsorbate consisting of multiple heteroatoms. For this reason, we define the *local conformation* subgraph (a fragment of a molecular graph) that describes the binding structure around an adsorbed atom. It is centered on a heteroatom and contains its nearest neighbor heteroatoms and the surface atoms connected to all of these. The role of local conformation is discussed below.

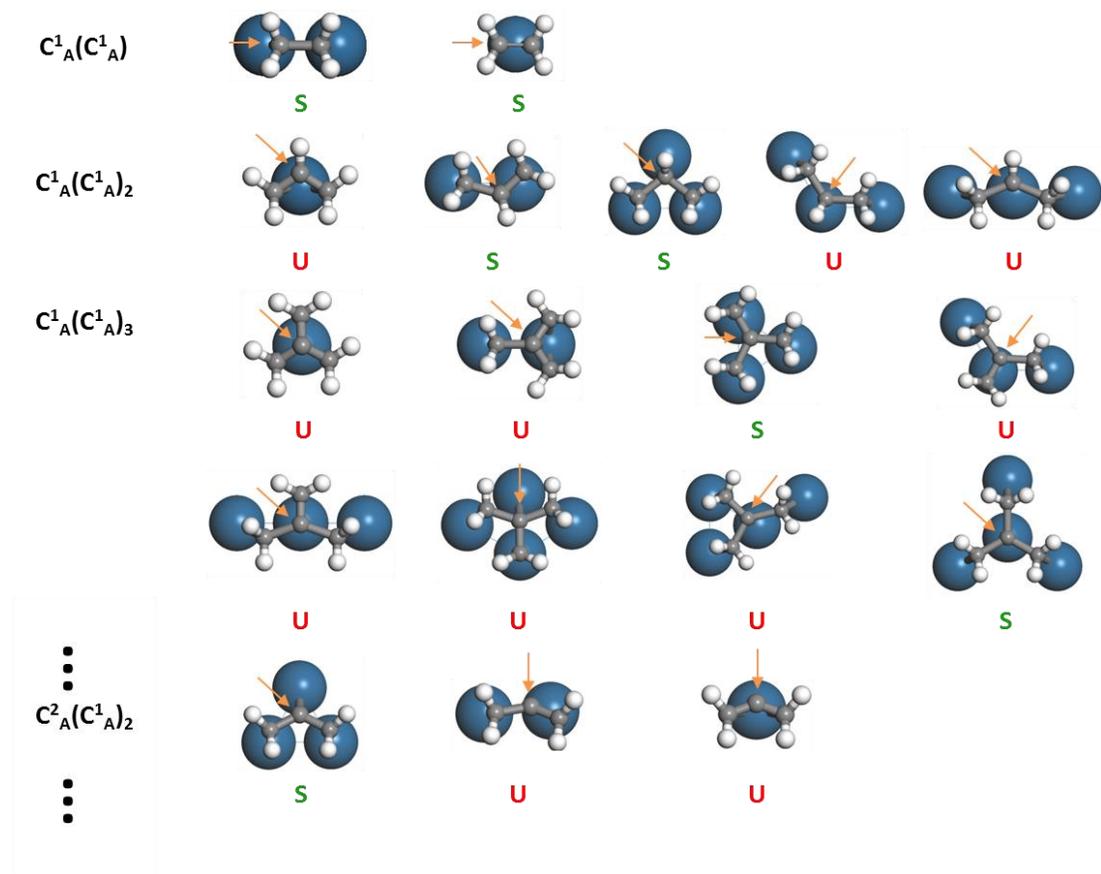


Figure 6.2. Examples of the effect of first-nearest neighbors on stability. Atoms with orange arrows are the central carbons. S (U) indicates a stable (unstable) conformation.

From our experience, these binding conformations are typically transferable between strong-binding metals. The metal-metal distance in 12 metals (Ag, Au, Co, Cu, Fe, Ir, Ni, Pd, Pt, Re, Rh, Ru) is fairly constant ( $2.67 \pm 0.10 \text{ \AA}$ ), thus the strain in the same conformation across metals is expected to be similar. However, for weakly carbon binding metals, such as Ag, Au, and Cu, the preferred binding sites are significantly different. For example, these surface atoms disfavor binding with multiple adsorbate atoms (e.g., the second conformation of  $C^1_A(C^1_A)$  in Figure 6.2).

On these metals, molecules with multiple non-zero valency heteroatoms (such as benzene and furan) prefer to be physisorbed. Here we focus on strong binding metals.

Our overarching objective is to collect a data set of conformations and their stability, and extract their connectivity to enable generation of conformations and prediction of their stability. The sections below describe the data collection process and algorithm building using those data.

### **6.2.2 Conformation and Stability Sampling *via* DFT**

The stability of a conformation is determined using DFT by examining whether this conformation relaxes to another one during the DFT ionic relaxation. In detail, atomic coordinates of an adsorbate and the slab are manipulated so that the coordinates describe the conformation's connectivity (confirmed by using the atomic coordinate to molecular graph conversion method introduced in Chapter 5). These atomic coordinates are inputted into DFT for ionic relaxation calculations. After the calculation converges, the atomic coordinates are redeemed and converted back to a molecular graph. The converged final conformation is added to the data set as a stable conformation. If the resulting conformation is not isomorphic to the initial conformation, the initial conformation is deemed unstable. Intermediate structures generated during ionic relaxation undergo the same stability classification.

To build our data set, we have constructed 808 conformations on Pd(111) involving an atom adsorbed on a binding site with first and second-nearest neighbor atoms on various binding sites occasionally shared by heteroatoms. These conformations are heuristically chosen. The resulting data set of 808 initial structures is supplied to the DFT solver. In a DFT calculation, conformations in each ionic relaxation step are extracted and those that are non-isomorphic with the ground state

conformation are termed unstable. This approach results in 2,404 conformations with assigned stability.

### **6.2.3 Automatic Surface Adsorbate Conformation Generation Algorithm**

The adsorption conformation prediction algorithm involves largely three steps, as illustrated in Figure 6.3. The first step enumerates the conformations of a species given its gas-phase structure. The conformation enumeration algorithm is inspired by the reaction network generation algorithm.[94, 149] While the reaction network generation algorithm enumerates species in a reaction network by exhaustively applying reaction rules to reactants and the generated species, the algorithm in this work uses “binding rules” to enumerate conformations. Here a rule describes a graph transformation/graph rewriting procedure. Thus, the two algorithms mainly differ by the rules. Conformations are built starting from a gas-phase structure by assigning a binding site to one heteroatom at a time. The sections below describe the binding rules considering  $\text{CH}_2\text{CHCH}_2$  as an example along with the methodology for harnessing these rules from the collected conformation and stability data discussed above.

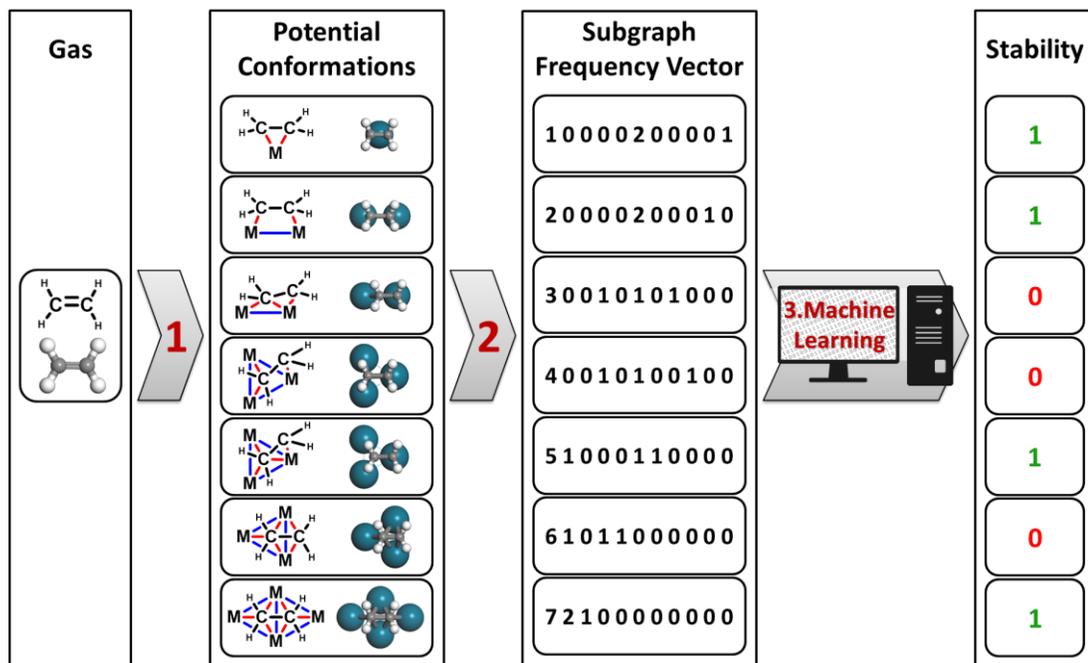


Figure 6.3. Overview of the conformation enumeration algorithm. The algorithm involves three steps: (1) exhaustive enumeration of conformations using mined transformation rules; (2) mining conformations' subgraphs and their frequencies; and (3) prediction of stability using machine learning. Black, red, and blue edges indicate bonds between two adsorbate atoms, between an adsorbate and a metal atom, and between two metal atoms, respectively. M indicates a surface metal atom. 1 and 0 in the stability column indicates stable and unstable conformation, respectively.

### 6.2.3.1 Binding Rules

The enumeration is initiated by binding a single heteroatom (with valency greater than zero) randomly chosen on a metal site (denoted as the first atom binding rule). Typically 3 conformations are created to account for adsorption on an atop, bridge or hollow site. Then, a nearest neighbor site of the occupied metal site is assigned to a nearest neighbor heteroatom of the molecule (denoted as the neighbor atom binding rule). By doing so, the resulting molecular graph does not contain a disconnected lattice graph. This procedure is repeated until metal sites are assigned to

all heteroatoms atoms (with valency greater than zero), resulting in a potential conformation. The sequence of atoms chosen is irrelevant in the algorithm since complete enumeration is performed. For species that can form loops, e.g.,  $\text{CH}_2=\text{CH}-\text{CH}_2-\text{CH}_2-\text{CH}=\text{CH}_2$ , similar to polymers, a bridge-like rule can be applied to allow multiple, disconnected bindings.

An example of a conformation generation for  $\text{CH}_2\text{CHCH}_2$  is shown in Figure 6.4. Here, the first atom binding rule is applied to the primary or secondary carbon. Since the two primary carbons are equivalent (isomorphic conformations), only one of them is selected. Once the C atom is bound, two neighbor atom binding rules are next applied to each structure. The three conformations generated are shown in the last column in Figure 6.4.

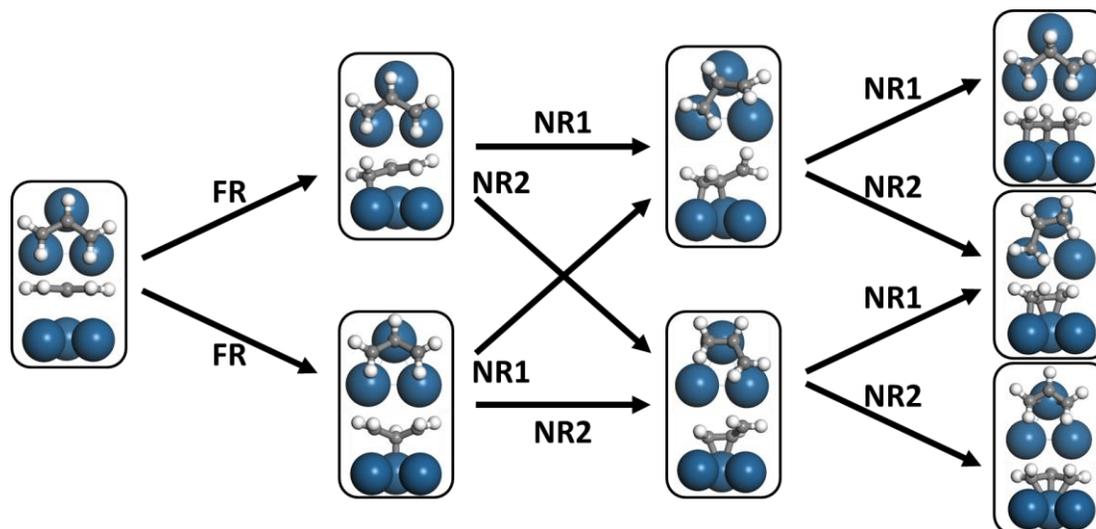


Figure 6.4. An example of conformation generation of  $\text{CH}_2\text{CHCH}_2$  using the rules in Figure 6.5. FR is the first atom binding rule in Figure 6.5a, NR1 and NR2 are the neighbor atom binding rules in Figure 6.5c and e. The second and third column represent intermediate structures. Only one type of site (atop site) is shown here for simplicity.

Next we describe the binding rules using graph theory concepts. The illustrated first atom binding rule (Figure 6.5) shows adsorption of a carbon with valency 1 to an atop site. This operation is achieved via graph transformation. The rule contains a “pattern graph” which represents an unbound carbon with valency of 1 and an empty metal atom. This pattern is searched in the initial conformation containing the gaseous molecule and the lattice. If a match is found, the matched subgraph is replaced with the “replacement graph,” which, in this rule, represents a carbon with valency 1 adsorbed at an atop site. A more intuitive interpretation is that an edge is added between the said carbon atom and a surface atom. An example of finding the occurrence of pattern graph and replacement is shown in Figure 6.5b. This example involves the adsorption of the primary carbon. The secondary carbon is another candidate for applying this first atom binding rule, as shown in Figure 6.4. The numbers in the circle show how atoms in the pattern graph, replacement graph, and the matches are mapped to each other. Next, a neighbor atom binding rule is illustrated (Figure 6.5c). The pattern graph of this rule identifies an unbound carbon with valency of one ( $C^1_u$ ) connected to another carbon with valency of one already bound to an atop site ( $C^1_A$ ) and its connected metal atom  $M_1$ , and an unoccupied metal atom  $M_2$  connected to  $M_1$ . In the replacement graph,  $C^1_u$  is adsorbed to  $M_2$ . Another neighbor atom binding rule is illustrated in (Figure 6.5e). This rule is similar to Figure 6.5c where an unbound carbon becomes bound to an occupied metal atom instead of an unoccupied metal atom. These rules are applied twice to generate the final conformations.

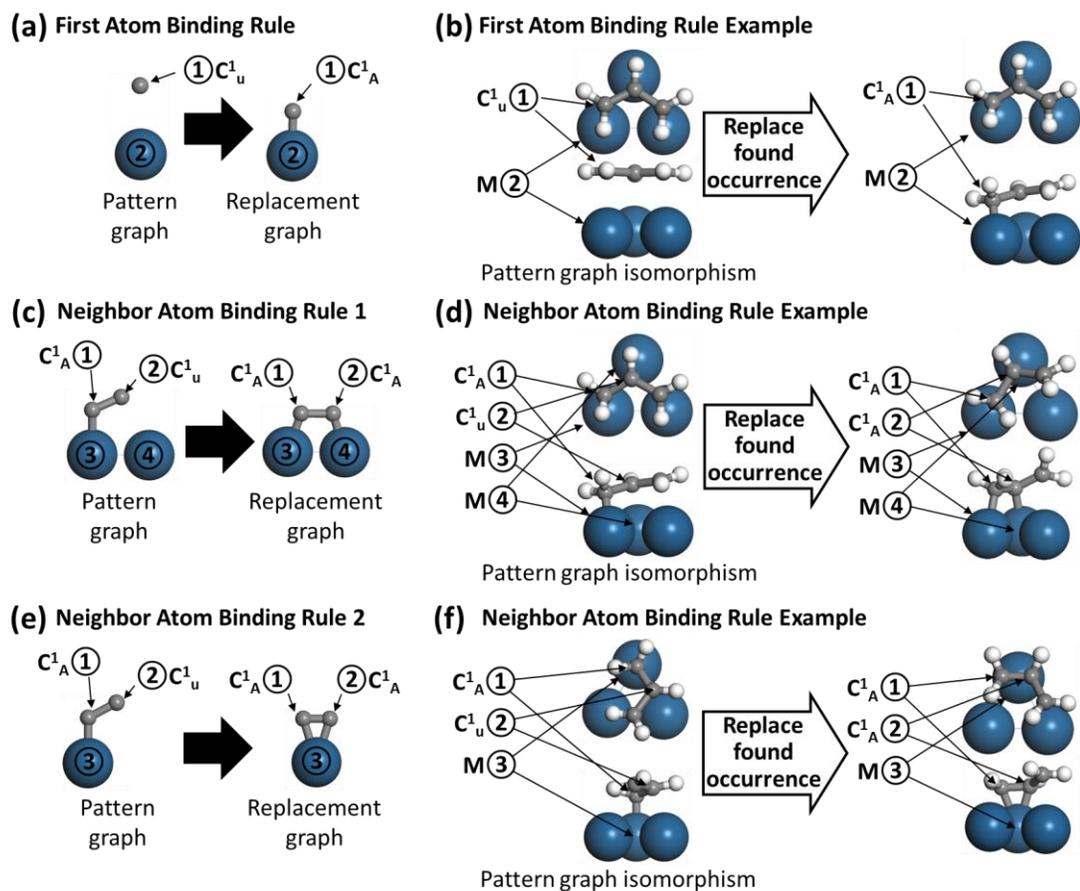


Figure 6.5. Graph theory representation of binding rules. (a) The first atom binding rule used in Figure 6.4 for generation of a conformation; (b) An example of the first atom binding rule in (a); (c) The neighbor atom binding rule; (d) An example of the neighbor atom binding rule in (c); (e) Another neighbor atom binding rule; (f) An example of the neighbor atom binding rule in (e). In  $C_{\beta}^{\alpha}$  notation,  $\alpha$  is the valency of a carbon atom, and  $\beta = A, B, H$  or  $u$  is atop, bridge, hollow or unassigned binding site of the carbon atom. The index in a circle for (a), (b), (c), (d), (e) and (f) shows how atoms are mapped to each other.

The first and neighbor atom binding rules are mined from stable conformations from the 2,404 dataset built above. For the first atom binding rules, we extract every heteroatom bound to the surface and its binding surface sites among the stable conformations in the dataset. The resulting graphs are the replacement graphs. Pattern graphs are generated by removing the edges between the heteroatoms and the surface sites in the replacement graphs. For the neighbor atom binding rules, every two connected adsorbate atoms and their surface sites from stable conformations of the dataset are extracted. The resulting graphs are the replacement graphs. Pattern graphs are formed in a similar way by removing the related edges.

#### **6.2.3.2 Efficiency of Enumeration Algorithm**

The computational time for the resulting algorithm is high due to the large number of intermediate conformations. To overcome this problem, a set of “stable local conformations” is created from the dataset to prune conformations during generation. Specifically, upon application of a binding rule, if a local conformation in the structure is not in the stable set, the structure is removed. This approach is summarized in Algorithm 6.1 together with the enumeration algorithm. Lines 1-4 represent the initial adsorption of one atom to the surface using the first atom binding rule. The remaining of the algorithm exhaustively applies the neighbor atom binding rule to the conformation pool. This algorithm has been implemented using Python and the Rdkit package.[292]

---

**Algorithm 6.1.** Conformation Enumeration.

---

- 1: **for** each heteroatom **do**
- 2:   apply first atom binding rule
- 3:   add resulting structure to new species list
- 4: **end for**
- 5: **while** the new species list is not empty **do**
- 6:   pop a species from the new species list
- 7:   apply all binding rules
- 8:   remove lattice ambiguity in each species using Algorithm 6.2
- 9:   remove previously enumerated species
- 10:   remove species with unobserved local conformations
- 11:   add remaining species to the new species list
- 12: **end while**

---

### 6.2.3.3 Canonicalization and Ambiguity Removal of Conformations

A naïve way to represent an adsorbate conformation is to simply include all occupied metal atoms (M) in a graph. However, this approach sometimes results in an ambiguous lattice structure. In particular, when a conformation contains a metal site M with only one or two non-connected neighboring metal sites, the lattice structure become ambiguous. Figure 6.6 shows such a case where the conformation contains a metal-metal-metal line-like lattice subgraph. As the angle of the metal-metal-metal is not described, the M-M-M structure is not unique.

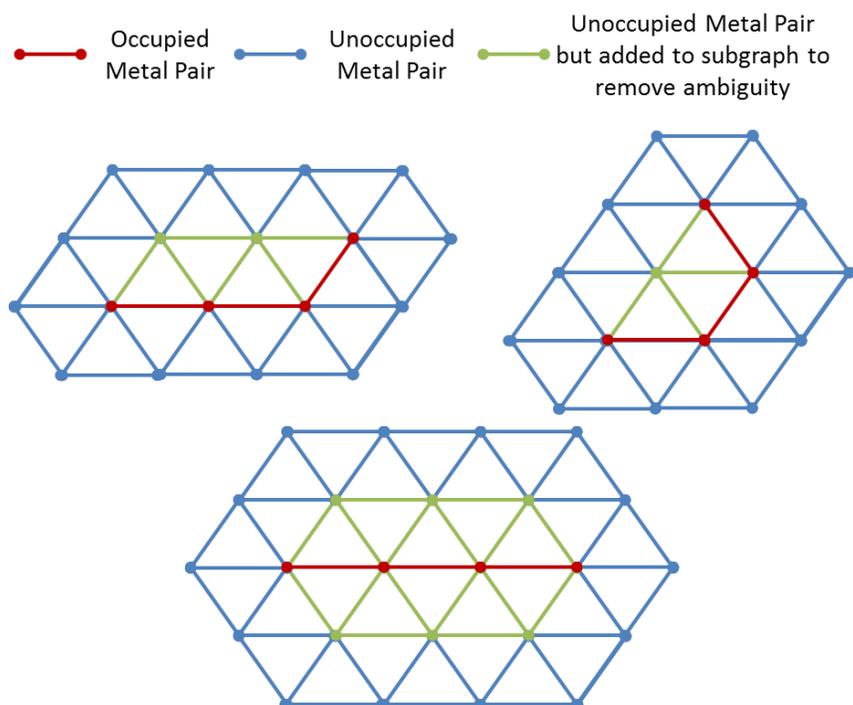


Figure 6.6. Three lattice subgraphs possible for a metal-metal-metal lattice subgraph of a conformation. Algorithm 6.2 systematically and efficiently adds unoccupied metal atoms to represent a conformation canonically and unambiguously.

In order to remove ambiguity, each metal atom in a conformation needs to be connected to at least two metal atoms connected to each other (see Figure 6.6). We introduce a simple and efficient algorithm to remove the ambiguity in the lattice of the conformation by adding appropriate unoccupied atoms to the conformation, as shown in Algorithm 6.2. There exists more than one way to add unoccupied atoms. The presented algorithm is designed to always result in the same lattice subgraph given an initial lattice subgraph in the conformation. This canonicalization feature, where a unique representation is outputted, when multiple representations are possible to achieve a goal, is critical to removing duplicate conformations. This algorithm is

applied whenever new conformations are produced during the enumeration. In detail, unoccupied metal atoms with more than 2 metal atoms in the conformation graph are iteratively added to the conformation. At the end of the iteration, the lattice subgraph of the conformation becomes planar without a concave corner, as demonstrated in Figure 6.6. One exception is if the lattice subgraph is a line. In this case, all the unoccupied metal sites with two occupied metal sites are added. Lines 1-4 in Algorithm 6.2 initialize the map  $O: i \rightarrow n_i$  where  $i$  is the index of unoccupied metal atom  $i$ , and  $n_i$  is the number of metal atom neighbors of  $i$  that are in the conformation molecular graph. Lines 5-7 check whether the lattice in the conformation is a line, and, if it is, the algorithm adds unoccupied sites appropriately as discussed above and returns a new conformation without the lattice ambiguity (see Algorithm 6.1). Lines 9-17 iteratively add metal atoms with more than two metal atoms to the conformation until no more metal atoms are added to the conformation. The canonical conformations are then converted to SMILES using Rdkit,[292] which are algorithmically guaranteed to canonically describe a molecular graph, i.e., to compare isomorphism between different conformations and ensure uniqueness of structures.[340]

---

**Algorithm 6.2.** Conformation Canonicalization.

---

```
1:  $O = \{ \}$ 
2: for each unoccupied metal atom  $i$  do
3:    $O(i) \leftarrow$  Number of occupied metal neighbors of atom  $i$ 
4: end for
5: if all elements in  $O$  are 2 do
6:   add all atoms in  $O$  to conformation and return
7: end if
8: add atoms in  $O$  with more than two to AtomToProcess list
9: while AtomToProcess is not empty do
10:  pop an atom  $i$  from AtomToProcess and add  $i$  to conformation
11:  for atom  $j$  in  $O$  do
12:    if  $i$  is a neighbor of  $j$  do
13:       $O(j) \leftarrow O(j) + 1$ 
14:    end if
15:    if  $O(j)$  is three do
16:      add atom  $j$  to AtomToProcess list
17:    end if
16:  end for
17:end while
```

---

#### 6.2.4 Determining Conformation Stability using Machine Learning

The next two steps of the algorithm involve determining the stability of a generated conformation (Steps 2 and 3 in Figure 6.3). In step 2, conformations are decomposed to descriptors, namely, subgraphs and their frequency – number of occurrences. We hypothesize that subgraphs may represent an unphysical structure, and their frequency correlates to the stability of a conformation. Subgraphs are mined selectively using the adsorbate graph mining method of previous work (this differs from local conformations).[337] In this method, all possible subgraphs of the adsorbate without the surface atoms are enumerated, and, then, surface atoms connected to heteroatoms are added for each subgraph. These selectively mined subgraphs have been shown to correlate well with thermodynamic properties, and we

expect them to correlate well with the conformation stability. Finally, descriptor vectors are inputted into a logistic regression model as implemented in scikit-learn.[341] Logistic regression is a classifier, which outputs the probability of having a feature, i.e., the conformation stability in this work. The input involves the linear function of the descriptor variables.

Training the logistic regression model involves decomposition of all the collected 2,404 conformations to subgraphs followed by training using the stability data. An  $L_1$  penalty of the descriptor coefficients is applied to encourage sparsity of the descriptor coefficients. This reduces the number of isomorphisms required for building descriptor variables. The magnitude of  $L_1$  penalty is optimized by leave-one-out cross-validation.

## 6.3 Results and Discussion

### 6.3.1 Rule Mining

For the first atom binding rule, nine rules are mined which are  $C^1_A$ ,  $C^1_B$ ,  $C^1_H$ ,  $C^2_A$ ,  $C^2_B$ ,  $C^2_H$ ,  $C^3_A$ ,  $C^3_B$ , and  $C^3_H$ . Therefore, carbons can adsorb at atop, bridge, or hollow site irrespective of their valency. This result is counter intuitive, as atoms are typically expected to adsorb at a site that result in a number of C-M bonds equal to their valency. However, when multiple heteroatoms bind to the surface, each heteroatom can adsorb at unintuitive binding sites due to strain, signifying the complexity of adsorption of multi-dentate species. A total of 47 pairs of connected heteroatoms are mined, as shown in Table 6.1. Depending on the graph symmetry, one or two neighbor atom binding rules are mined from these pairs, resulting in 86 neighbor atom binding rules. As mentioned above, two atoms can share surface atoms

upon binding. For example, the  $C^1_A-C^1_A$  pair in the first row of Table 6.1 may share a surface atom or interact with two separate surface atoms. Interestingly, for each pair of binding sites, a minimum and maximum number of sharable surface atoms exist. For example,  $C^x_A-C^y_{A \text{ or } B \text{ or } H}$  and  $C^x_B-C^y_B$  pairs could share one surface atom or none. However, the carbons of a  $C^x_B-C^y_H$  pair must share one surface atom because the C-C bond cannot stretch far enough without dissociation. For  $C^x_H-C^y_H$  pairs, the only observed structure is when two surface atoms are shared. We have extracted a total of 250 local conformations from stable conformations.

Table 6.1. Binding Rule Information.

Atom1		Atom2		Number of Shared Surface Atom
Valency	Site	Valency	Site	
1	A	1	A	0, 1
1	A	1	B	0, 1
1	A	1	H	1
1	B	1	B	1
1	B	1	H	2
1	H	1	H	2
1	A	2	A	0
1	B	2	A	1
1	B	2	H	2
1	A	3	H	0, 1
1	B	3	H	2
1	H	3	H	2
2	A	1	A	1
2	A	1	B	1
2	A	1	H	1
2	B	1	A	0, 1
2	B	1	B	0
2	B	1	H	1, 2
2	H	1	A	1

Table 6.1. continued.

2	H	1	H	2
2	A	2	A	0, 1
2	A	2	B	0, 1
2	A	2	H	1
2	B	2	B	0, 1
2	B	2	H	1, 2
2	H	2	H	2
2	A	3	H	1
2	B	3	B	0
2	B	3	H	1, 2
2	H	3	H	2
3	B	1	A	0
3	H	1	B	1
3	B	2	B	1
3	H	3	H	2

### 6.3.2 Model Assessment

To thoroughly test the logistic regression model, cross-validation is performed. At the optimal  $L_1$  penalty, the accuracy of the model is 85%. The detailed statistics and parity plot are shown in Table 6.2 and Figure 6.7. In our model, 12% of the unstable conformations are falsely labeled. However, the cross-validation accuracy is likely over-predicted, as each datum is highly dissimilar to the others. The dataset is very sparse, with local conformations typically sampled only once. The dataset needs more redundancy to improve the cross-validation score. In addition, the data contains fundamental outliers where no similar datum exists in chemical space (all possible molecular structures) for extrapolation. For example,  $C^2_1$ , as discussed above, is unstable by itself but becomes stable with two anchoring neighbors. Thus, the instability of  $C^2_1$  cannot be predicted without the datum itself. The actual accuracy is likely somewhere between the cross-validation and the training set accuracy. The

training set accuracy is 98% with only 2% false negatives, thus the prediction is expected to be reliable.

Table 6.2. Error matrix of the logistic model.

	<b>Cross validation error</b>		<b>Training set error</b>	
	True Condition Positive	True Condition Negative	True Condition Positive	True Condition Negative
<b>Predicted Condition Positive</b>	283	114	462	13
<b>Predicted Condition Negative</b>	219	1624	40	1725

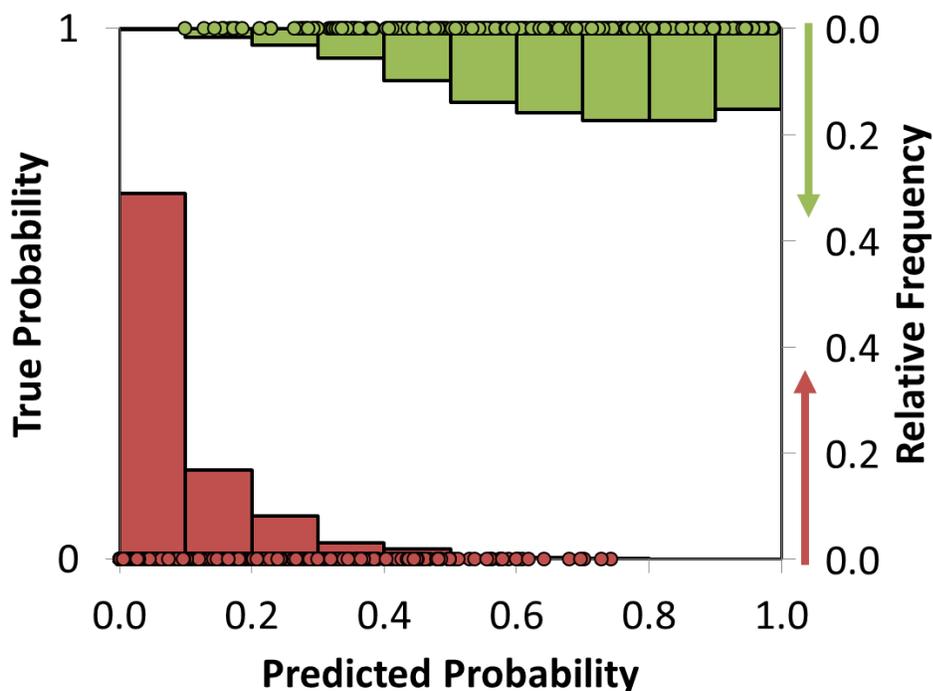


Figure 6.7. Parity plot of the logistic model. Relative frequency is plotted for better visualization.

### 6.3.3 An Example

Here, we present results for a complex molecule, namely,  $\text{CH}_2=\text{C}(\text{CH}_2)_2$ . The script only takes the SMILES of the gas structure as input, which is `[CH2][C]([CH2])[CH2]`. The script automatically generates 8 conformations and suggests that 5 of them are stable, as demonstrated in Figure 6.8.

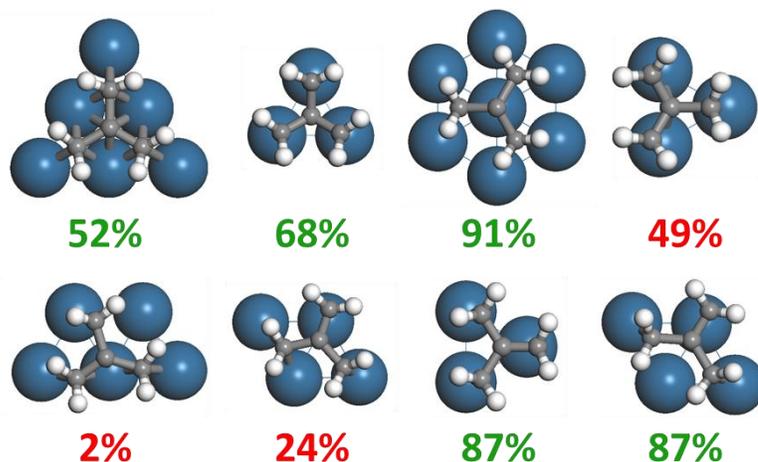


Figure 6.8. Conformation enumeration of  $\text{CH}_2=\text{C}(\text{CH}_2)_2$ . The results agree with the DFT calculations on Pd(111). The numbers indicate the stability computed by the logistic regression.

### 6.3.4 Additional Post Processing

The atomic coordinates for the dataset in this work was constructed manually. To avoid building data set manually in the future, we have modified the universal force-field to convert conformations to roughly optimized coordinates (using the implementation in the Rdkit software package).[164, 292] These coordinates can be used for visualizing molecular graphs or as a good initial guess for DFT calculations.

## 6.4 Conclusions

This chapter introduced a new algorithm to rapidly predict adsorption conformations of large adsorbates on a catalyst surface. 808 conformations were generated and passed onto DFT, and this gave 2,404 data by removing unstable conformations during the ionic relaxation of DFT calculations. In order to enumerate potential conformations, the algorithm introduces fundamental binding rules that are mined from stable conformations in a training dataset. The binding rules are designed

to avoid disconnected lattice structures. In order to prune the dimensionality of generated conformations, structures with unobserved stable local conformations are omitted. For instability caused by a structure that extends beyond the first nearest neighbor, the adsorbate graph mining method from Chapter 5 is employed, and logistic regression is applied to predict the stability of generated conformations. The cross-validation accuracy is 85%, but the accuracy should be between 85% and the training set accuracy 98% due to the non-redundancy and outliers in the dataset. Future work should involve accounting for binding sites of oxygen heteroatoms to address biomass chemical transformations.

## **6.5 Acknowledgements**

G. H. G. developed the algorithm and Miriam Lee, an undergraduate from University of Massachusetts Amherst, performed the DFT calculations, as an REU student in CCEI. This research was supported by the NSF under grant number CHE-1415828. This work benefited from the support of the project MuSiC ANR-14-CE06-0030 of the French National Research Agency (ANR).

## Chapter 7

### CONCLUSIONS AND OUTLOOK

#### 7.1 Dissertation Summary

In this thesis, several computational gaps for biomass conversion on metal surfaces have been addressed. Group additivity schemes for lignin monomers as well as solvated sugar alcohols have been developed for the first time. The previous semi-empirical model framework has been improved via developing better descriptors and by implementing machine-learning. These new tools enable modeling of large biomass conversion reaction networks. In addition, a simple lateral interaction model has been introduced that is applied to diverse species in a reaction network. The frameworks introduced herein are extendable to other biomass monomers as well as to other surface facets.

Chapter 2 carried out the first microkinetic study of a lignin model compound, cresol, on Pt to provide mechanistic understanding unattainable from experiments. The simulation results are consistent with multiple experiments, supporting that Pt(111) is capable of deoxygenation chemistry without an acidic support. This work elucidated a new mechanism where dehydroxylation becomes facile upon sufficient hydrogenation of the ring. In addition, our work implied that oxophilic metal such as Fe and Ni would likely dehydrate cresols via direct deoxygenation mechanism. In order to parameterize a large number of reactions, BEP relationships were significantly improved using transition state structure sensitivity. In addition, a model for lateral

interactions was introduced that can be extended to the modeling of other lignin monomers.

Besides the deoxygenation of the hydroxyl group explored in Chapter 2, the methoxy group is another common oxygenate functional group in lignin monomers. To understand the methoxy groups' deoxygenation mechanism in the guaiacol HDO was investigated in a co-authored work, by Lee et al.[117] via DFT. However, MKM has never been applied to guaiacol HDO. Thus, Chapter 3 discussed a group additivity scheme for the guaiacol reaction network. 591 molecules were sampled using DFT. The previous group additivity framework was found to be inadequate for describing conjugation and ring strain imposed by the C<sub>6</sub> ring. To identify groups that are transferable, Bader analysis was employed. The analysis revealed that groups must include the binding site of the central and the nearest neighbor atoms. The new framework reduced the MeanAE from 5.15 to 2.81 kcal/mol.

Despite the presence of water in biomass conversion, the effect of water is poorly understood. Chapter 4 introduces a group additivity for solvated adsorbates on the metal surface. Specifically, sugar alcohols in aqueous solution on Pt(111) were investigated. The Gibbs free energy of solvation is predicted within MeanAE of ~ 1.0 kcal/mol. The model provides an accurate estimation of solvation effect, at a dramatically reduced computational cost. Analysis revealed that O-H scission is thermodynamically disfavored and C-H, C-C, and C-O scissions are thermodynamically favored in aqueous solution.

Chapter 5 discusses a feature selection machine learning method, LASSO, that automates selection of groups (optimal descriptors) especially for adsorbates that possess complex graphic structures for group additivity. For each datum, all possible

patterns were enumerated. Then, the patterns' occurrence in the datum was correlated to the heat of formation and LASSO automatically selected the most informative set of patterns for describing the datum's heat of formation. This framework is significantly more accurate than the traditional group additivity approach (9.17  $\rightarrow$  5.74 kcal/mol improvement in MeanAE). In addition, a graph mining method was introduced to extract useful patterns from adsorbate DFT calculations. This method, combined with LASSO, improved the lignin group additivity MeanAE from 2.81 kcal/mol to 2.08 kcal/mol, achieving  $< 0.1$  eV (2.31 kcal/mol) error for the first time for an adsorbate group additivity scheme. Statistical analysis showed that a bigger data set is needed to improve the fitting further. As the computational time for pattern enumeration scales exponentially with molecular size, a semi-supervised method was developed where patterns are enumerated during regression, and uninformative patterns are pruned from the descriptor space to reduce computational burden. Using this method, the computational cost no-longer scales with the graph size. Lastly, a simple method to identify outliers was introduced in order to remove datums whose energy cannot be predicted, as well as to improve data sampling techniques.

Chapter 6 introduced a data-based algorithm to rapidly predict adsorption conformations of large adsorbates on a close-packed metal surface. 2,404 conformation and stability data were computed using DFT calculations. To enumerate conformations, this chapter introduced fundamental binding rules mined from stable adsorption conformations. The algorithm implements canonicalization features to remove ambiguity for the lattice in the conformation. To compute the stability, the conformation is decomposed into subgraphs using the graph mining method introduced in Chapter 5 and logistic regression is applied to predict the stability. The

accuracy of the stability classification is expected to be above 85%. This work introduces a streamlined algorithm to generate inputs for building a big data set as well as to compliment the developed group additivity method. Future work involves accounting for binding sites of oxygen heteroatoms to address biomass chemical transformations. Lastly, Appendix E shows results for the mechanism of the aqueous phase reforming for ethylene glycol on Pt(111).

## 7.2 Future Work

Figure 1.5 discussed computational gaps, some of which were addressed in this thesis. Critical future tasks include the development of a comprehensive lateral interaction framework (Figure 1.5e) and the improvement of LSRs (Figure 1.5c). Despite the improvement of the group additivity and the development of the machine-learning model, the learning curve for the lignin dataset in Chapter 5 shows more data is necessary. With the development of the automatic adsorption conformation prediction algorithm, building data points could be fully automated without human intervention. Building a big data and applying a single machine learning framework is promising to minimize model errors and would simplify the user interface. In addition, these frameworks can be employed as web applications to boost other computational research in the field. Also, microkinetic modeling could be coupled with these semi-empirical methods to generate reaction network on the fly, so unfavorable reactions and species are unaccounted to minimize computation time of large microkinetic models.

The BEP relations in Chapter 2 show that a three center-atom transition state is unfavorable due to the increased coordination of the Pt atom in the transition state (Figure 7.1a). This unfavorable three center-atom transition state could be bypassed by

designing the catalytic site. A simple example is presented in Figure 7.1b. On step sites, the Pt atom on the step is significantly closer to the OH group than those on the terrace slab and a four center atom transition state is expected. A similar observation has been reported for olefin C-C and C-H scission where the cracked components involve smaller reaction energy and barriers due to the nearby step site surface atom.[342] Facile olefin cracking has been experimentally observed on smaller catalyst particles.[342] Interestingly, deoxygenation of phenol has been shown to become more facile with decreasing particle size for Ni/SiO<sub>2</sub>, reinforcing this hypothesis.[343] Our results indicate that controlling particle size and shape provide simple strategies to tune reactivity.

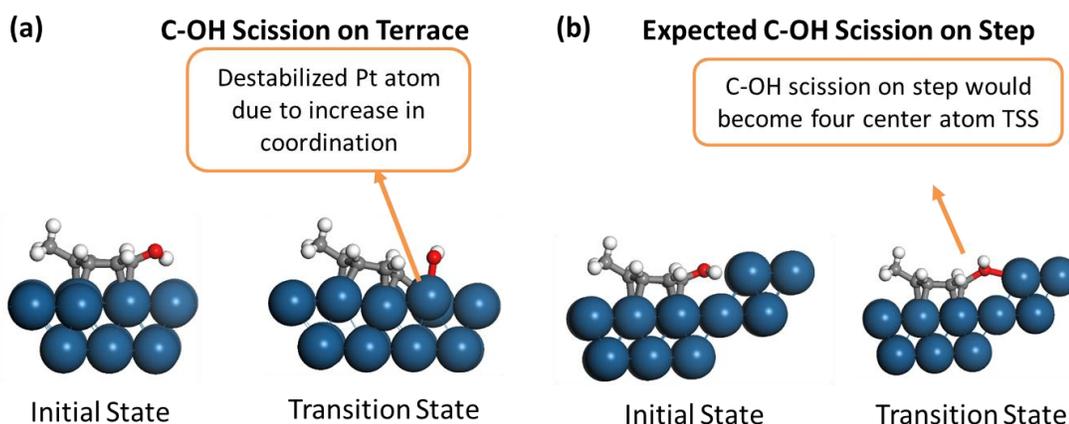


Figure 7.1. Example of (a) unstable C-OH scission transition state structure on the terrace (DFT calculations in Chapter2), and (b) expected C-OH scission transition state structure on step (based on heuristics).

## REFERENCES

1. *Facing the Hard Truths about Energy*. US National Petroleum Council: Washington DC, 2007.
2. Renewable Fuel Standard. U.S. Department of Energy: <https://www.afdc.energy.gov/laws/RFS> (accessed 10/23).
3. *U.S. Billion-ton update*. U.S. Department of Energy: Washington D.C., 2011.
4. Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., Frederick, W. J., Hallett, J. P., Leak, D. J., Liotta, C. L., Mielenz, J. R., Murphy, R., Templer, R., and Tschaplinski, T. The Path Forward for Biofuels and Biomaterials. *Science* **2006**, *311*, 484-489.
5. Corma, A., Iborra, S., and Velty, A. Chemical Routes for the Transformation of Biomass into Chemicals. *Chem. Rev.* **2007**, *107*, 2411-2502.
6. Dickerson, T., and Soria, J. Catalytic Fast Pyrolysis: A Review. *Energies* **2013**, *6*, 514.
7. Vlachos, D. G., Chen, J. G., Gorte, R. J., Huber, G. W., and Tsapatsis, M. Catalysis Center for Energy Innovation for Biomass Processing: Research Strategies and Goals. *Catal. Lett.* **2010**, *140*, 77-84.
8. Gallezot, P. Conversion of biomass to selected chemical products. *Chem. Soc. Rev.* **2012**, *41*, 1538-1558.
9. Heltzel, J., Patil, S. K. R., and Lund, C. R. F. Humin Formation Pathways. In *Reaction Pathways and Mechanisms in Thermocatalytic Biomass Conversion II: Homogeneously Catalyzed Transformations, Acrylics from Biomass, Theoretical Aspects, Lignin Valorization and Pyrolysis Pathways*, Schlaf, M.; Zhang, Z. C., Eds. Springer Singapore: Singapore, 2016; pp 105-118.
10. Vlachos, D. G., and Caratzoulas, S. The roles of catalysis and reaction engineering in overcoming the energy and the environment crisis. *Chem. Eng. Sci.* **2010**, *65*, 18-29.

11. Pallegriano, J. L. Energy and Environmental Profile of the U.S. Chemical Industry. Office of Industrial Technologies, U. S. D. o. E., Ed. Columbia, Maryland, 2000.
12. *Petrochemicals Market Analysis By Product By Region And Segment Forecasts, 2014 - 2025*; Grand View Research: 2016.
13. Cortright, R. D., Davda, R. R., and Dumesic, J. A. Hydrogen from Catalytic Reforming of Biomass-Derived Hydrocarbons in Liquid Water. *Nature* **2002**, *418*, 964-967.
14. Hohenberg, P., and Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864-B871.
15. Teixeira, C. A. d. M. Direct numerical simulation of fixed-bed reactors: Effect of random packing Universidade Do Porto, Porto, Portugal, 2013.
16. Vorotnikov, V., Wang, S., and Vlachos, D. G. Group additivity for estimating thermochemical properties of furanic compounds on Pd (111). *Ind. Eng. Chem. Res.* **2014**, *53*, 11929-11938.
17. Hammer, B., and Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* **1995**, *343*, 211-220.
18. Sutton, J. E., and Vlachos, D. G. Building large microkinetic models with first-principles' accuracy at reduced computational cost. *Chem. Eng. Sci.* **2015**, *121*, 190-199.
19. Saliccioli, M., Stamatakis, M., Caratzoulas, S., and Vlachos, D. G. A review of multiscale modeling of metal-catalyzed reactions: Mechanism development for complexity and emergent behavior. *Chem. Eng. Sci.* **2011**, *66*, 4319-4355.
20. Abild-Pedersen, F., Greeley, J., Studt, F., Rossmeisl, J., Munter, T. R., Moses, P. G., Skúlason, E., Bligaard, T., and Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99*, 016105.
21. Gu, G. H., and Vlachos, D. G. Group additivity for thermochemical property estimation of lignin monomers on Pt (111). *J. Phys. Chem. C* **2016**, *120*, 19234-19241.

22. Saliccioli, M., Chen, Y., and Vlachos, D. G. Density Functional Theory-Derived Group Additivity and Linear Scaling Methods for Prediction of Oxygenate Stability on Metal Catalysts: Adsorption of Open-Ring Alcohol and Polyol Dehydrogenation Intermediates on Pt-Based Metals. *J. Chem. Phys. C* **2010**, *114*, 20155-20166.
23. Saliccioli, M., Edie, S., and Vlachos, D. Adsorption of acid, ester, and ether functional groups on Pt: fast prediction of thermochemical properties of adsorbed oxygenates via DFT-based group additivity methods. *J. Phys. Chem. C* **2012**, *116*, 1873-1886.
24. Wang, S., Temel, B., Shen, J., Jones, G., Grabow, L. C., Studt, F., Bligaard, T., Abild-Pedersen, F., Christensen, C. H., and Nørskov, J. K. Universal Brønsted-Evans-Polanyi Relations for C–C, C–O, C–N, N–O, N–N, and O–O Dissociation Reactions. *Catal. Lett.* **2011**, *141*, 370-373.
25. Medford, A. J., Vojvodic, A., Hummelshøj, J. S., Voss, J., Abild-Pedersen, F., Studt, F., Bligaard, T., Nilsson, A., and Nørskov, J. K. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* **2015**, *328*, 36-42.
26. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I., and Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. **2006**, *5*, 909.
27. Greeley, J., and Mavrikakis, M. Alloy catalysts designed from first principles. **2004**, *3*, 810.
28. Greeley, J., and Nørskov, J. K. Combinatorial Density Functional Theory-Based Screening of Surface Alloys for the Oxygen Reduction Reaction. *J. Phys. Chem. C* **2009**, *113*, 4932-4939.
29. Bligaard, T., Andersson, M. P., Jacobsen, K. W., Skriver, H. L., Christensen, C. H., and Nørskov, J. K. Electronic-Structure-Based Design of Ordered Alloys. *MRS Bull.* **2006**, *31*, 986-990.
30. Nørskov, J. K., Abild-Pedersen, F., Studt, F., and Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 937-943.
31. Nørskov, J. K., Bligaard, T., Rossmeisl, J., and Christensen, C. H. Towards the computational design of solid catalysts. **2009**, *1*, 37.

32. Hansgen, D. A., Vlachos, D. G., and Chen, J. G. Using first principles to predict bimetallic catalysts for the ammonia decomposition reaction. **2010**, *2*, 484.
33. Guo, N., Caratzoulas, S., Doren, D. J., Sandler, S. I., and Vlachos, D. G. A perspective on the modeling of biomass processing. *Energy Environ. Sci.* **2012**, *5*, 6703-6716.
34. Vorotnikov, V., and Vlachos, D. G. Group Additivity and Modified Linear Scaling Relations for Estimating Surface Thermochemistry on Transition Metal Surfaces: Application to Furanics. *J. Phys. Chem. C* **2015**, *119*, 10417-10426.
35. Vorotnikov, V., Wang, S., and Vlachos, D. G. Group Additivity for Estimating Thermochemical Properties of Furanic Compounds on Pd(111). *Ind. Eng. Chem. Res.* **2014**, *53*, 11929-11938.
36. Saliccioli, M., Edie, S. M., and Vlachos, D. G. Adsorption of Acid, Ester, and Ether Functional Groups on Pt: Fast Prediction of Thermochemical Properties of Adsorbed Oxygenates via DFT-Based Group Additivity Methods. *J. Phys. Chem. C* **2012**, *116*, 1873-1886.
37. Faheem, M., and Heyden, A. Hybrid Quantum Mechanics/Molecular Mechanics Solvation Scheme for Computing Free Energies of Reactions at Metal–Water Interfaces. *J. Chem. Theory Comput.* **2014**, *10*, 3354-3368.
38. Faheem, M., Suthirakun, S., and Heyden, A. New Implicit Solvation Scheme for Solid Surfaces. *J. Phys. Chem. C* **2012**, *116*, 22458-22462.
39. Behtash, S., Lu, J., Mamun, O., Williams, C. T., Monnier, J. R., and Heyden, A. Solvation Effects in the Hydrodeoxygenation of Propanoic Acid over a Model Pd(211) Catalyst. *J. Phys. Chem. C* **2016**, *120*, 2724-2736.
40. Behtash, S., Lu, J., Walker, E., Mamun, O., and Heyden, A. Solvent Effects in the Liquid Phase Hydrodeoxygenation of Methyl Propionate over a Pd(1 1 1) Catalyst Model. *J. Catal.* **2016**, *333*, 171-183.
41. Behtash, S., Lu, J., Faheem, M., and Heyden, A. Solvent Effects on the Hydrodeoxygenation of Propanoic Acid over Pd(111) Model Surfaces. *Green Chem.* **2014**, *16*, 605-616.
42. Sutton, J. E., Guo, W., Katsoulakis, M. A., and Vlachos, D. G. Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic modelling. *Nat. Chem.* **2016**, *8*, 331-337.

43. Sutton, J. E., and Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2*, 1624-1634.
44. Sutton, J. E., and Vlachos, D. G. Effect of errors in linear scaling relations and Brønsted–Evans–Polanyi relations on activity and selectivity maps. *J. Catal.* **2016**, *338*, 273-283.
45. Saeys, M., Reyniers, M.-F., Marin, G. B., and Neurock, M. Density Functional Study of Benzene Adsorption on Pt(111). *J. Phys. Chem. B* **2002**, *106*, 7489-7498.
46. Ihm, H., Ajo, H. M., Gottfried, J. M., Bera, P., and Campbell, C. T. Calorimetric Measurement of the Heat of Adsorption of Benzene on Pt(111). *J. Phys. Chem. B* **2004**, *108*, 14627-14633.
47. Saeys, M., Reyniers, M.-F., Thybaut, J. W., Neurock, M., and Marin, G. B. First-principles based kinetic model for the hydrogenation of toluene. *J. Catal.* **2005**, *236*, 129-138.
48. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I., and Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **2006**, *5*, 909.
49. Sehested, J., Larsen, K. E., Kustov, A. L., Frey, A. M., Johannessen, T., Bligaard, T., Andersson, M. P., Nørskov, J. K., and Christensen, C. H. Discovery of technical methanation catalysts based on computational screening. *Top. Catal.* **2007**, *45*, 9-13.
50. Stakheev, A. Y., and Kustov, L. M. Effects of the support on the morphology and electronic properties of supported metal clusters: modern concepts and progress in 1990s. *Appl. Catal., A* **1999**, *188*, 3-35.
51. Cuenya, B. R. Synthesis and catalytic properties of metal nanoparticles: Size, shape, support, composition, and oxidation state effects. *Thin Solid Films* **2010**, *518*, 3127-3150.
52. Bader, R. F., and Bayles, D. Properties of atoms in molecules: group additivity. *J. Phys. Chem. A* **2000**, *104*, 5579-5589.

53. Perlack, R. D., Wright, L. L., Turhollow, A. F., Graham, R. L., Stokes, B. J., and Erbach, D. C. *Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-ton Annual Supply*. United States. Department of Energy; United States. Department of Agriculture; Oak Ridge National Laboratory: 2005.
54. Ardiyanti, A. R., Khromova, S. A., Venderbosch, R. H., Yakovlev, V. A., and Heeres, H. J. Catalytic hydrotreatment of fast-pyrolysis oil using non-sulfided bimetallic Ni-Cu catalysts on a  $\delta$ -Al<sub>2</sub>O<sub>3</sub> support. *Appl. Catal., B* **2012**, 117-118, 105-117.
55. Ausavasukhi, A., Huang, Y., To, A. T., Sooknoi, T., and Resasco, D. E. Hydrodeoxygenation of m-cresol over gallium-modified beta zeolite catalysts. *J. Catal.* **2012**, 290, 90-100.
56. Bu, Q., Lei, H. W., Zacher, A. H., Wang, L., Ren, S. J., Liang, J., Wei, Y., Liu, Y. P., Tang, J. M., Zhang, Q., and Ruan, R. A review of catalytic hydrodeoxygenation of lignin-derived phenols from biomass pyrolysis. *Bioresour. Technol.* **2012**, 124, 470-477.
57. de Miguel Mercader, F., Groeneveld, M. J., Kersten, S. R. A., Geantet, C., Toussaint, G., Way, N. W. J., Schaverien, C. J., and Hogendoorn, K. J. A. Hydrodeoxygenation of pyrolysis oil fractions: process understanding and quality assessment through co-processing in refinery units. *Energy Environ. Sci.* **2011**, 4, 985-997.
58. Foster, A., Do, P. M., and Lobo, R. The Synergy of the Support Acid Function and the Metal Function in the Catalytic Hydrodeoxygenation of m-Cresol. *Top. Catal.* **2012**, 55, 118-128.
59. Hong, D.-Y., Miller, S. J., Agrawal, P. K., and Jones, C. W. Hydrodeoxygenation and coupling of aqueous phenolics over bifunctional zeolite-supported metal catalysts. *Chem. Commun.* **2010**, 46, 1038-1040.
60. Horáček, J., Šťávková, G., Kelbichová, V., and Kubička, D. Zeolite-Beta-supported platinum catalysts for hydrogenation/hydrodeoxygenation of pyrolysis oil model compounds. *Catal. Today* **2013**, 204, 38-45.
61. Mortensen, P. M., Grunwaldt, J.-D., Jensen, P. A., and Jensen, A. D. Screening of Catalysts for Hydrodeoxygenation of Phenol as a Model Compound for Bio-oil. *ACS Catal.* **2013**, 3, 1774-1785.
62. Nie, L., and Resasco, D. E. Kinetics and mechanism of m-cresol hydrodeoxygenation on a Pt/SiO<sub>2</sub> catalyst. *J. Catal.* **2014**, 317, 22-29.

63. Mohan, D., Pittman, C. U., and Steele, P. H. Pyrolysis of Wood/Biomass for Bio-oil: A Critical Review. *Energy Fuels* **2006**, *20*, 848-889.
64. Mullen, C. A., and Boateng, A. A. Chemical Composition of Bio-oils Produced by Fast Pyrolysis of Two Energy Crops†. *Energy Fuels* **2008**, *22*, 2104-2109.
65. Wang, H., Male, J., and Wang, Y. Recent Advances in Hydrotreating of Pyrolysis Bio-Oil and Its Oxygen-Containing Model Compounds. *ACS Catal.* **2013**, *3*, 1047-1070.
66. Mullen, C. A., and Boateng, A. A. Accumulation of Inorganic Impurities on HZSM-5 Zeolites during Catalytic Fast Pyrolysis of Switchgrass. *Ind. Eng. Chem. Res.* **2013**, *52*, 17156-17161.
67. Mullen, C. A., Boateng, A. A., Mihalcik, D. J., and Goldberg, N. M. Catalytic Fast Pyrolysis of White Oak Wood in a Bubbling Fluidized Bed. *Energy Fuels* **2011**, *25*, 5444-5451.
68. Jae, J., Coolman, R., Mountziaris, T. J., and Huber, G. W. Catalytic fast pyrolysis of lignocellulosic biomass in a process development unit with continual catalyst addition and removal. *Chem. Eng. Sci.* **2014**, *108*, 33-46.
69. Christiansen, M. A., Mpourmpakis, G., and Vlachos, D. G. Density Functional Theory-Computed Mechanisms of Ethylene and Diethyl Ether Formation from Ethanol on  $\gamma$ -Al<sub>2</sub>O<sub>3</sub>(100). *ACS Catal.* **2013**, *3*, 1965-1975.
70. Jenness, G. R., Christiansen, M. A., Caratzoulas, S., Vlachos, D. G., and Gorte, R. J. Site-Dependent Lewis Acidity of  $\gamma$ -Al<sub>2</sub>O<sub>3</sub> and Its Impact on Ethanol Dehydration and Etherification. *J. Phys. Chem. C* **2014**, *118*, 12899-12907.
71. Roy, S., Mpourmpakis, G., Hong, D.-Y., Vlachos, D. G., Bhan, A., and Gorte, R. J. Mechanistic Study of Alcohol Dehydration on  $\gamma$ -Al<sub>2</sub>O<sub>3</sub>. *ACS Catal.* **2012**, *2*, 1846-1853.
72. Janik, M. J., Macht, J., Iglesia, E., and Neurock, M. Correlating Acid Properties and Catalytic Function: A First-Principles Analysis of Alcohol Dehydration Pathways on Polyoxometalates. *J. Phys. Chem. C* **2009**, *113*, 1872-1885.
73. Carr, R. T., Neurock, M., and Iglesia, E. Catalytic consequences of acid strength in the conversion of methanol to dimethyl ether. *J. Catal.* **2011**, *278*, 78-93.

74. Furimsky, E. Catalytic hydrodeoxygenation. *Appl. Catal., A* **2000**, *199*, 147-190.
75. Odebunmi, E. O., and Ollis, D. F. Catalytic hydrodeoxygenation: I. Conversions of o-, p-, and m-cresols. *J. Catal.* **1983**, *80*, 56-64.
76. Hensley, A. J. R., Wang, Y., and McEwen, J.-S. Phenol Deoxygenation Mechanisms on Fe(110) and Pd(111). *ACS Catal.* **2015**, *5*, 523-536.
77. Tan, Q., Wang, G., Nie, L., Dinse, A., Buda, C., Shabaker, J., and Resasco, D. E. Different Product Distributions and Mechanistic Aspects of the Hydrodeoxygenation of m-Cresol over Platinum and Ruthenium Catalysts. *ACS Catal.* **2015**, *5*, 6271-6283.
78. Song, W., Liu, Y., Barath, E., Zhao, C., and Lercher, J. A. Synergistic effects of Ni and acid sites for hydrogenation and C-O bond cleavage of substituted phenols. *Green Chem.* **2015**, *17*, 1204-1218.
79. He, J., Zhao, C., and Lercher, J. A. Impact of solvent for individual steps of phenol hydrodeoxygenation with Pd/C and HZSM-5 as catalysts. *J. Catal.* **2014**, *309*, 362-375.
80. Echeandia, S., Pawelec, B., Barrio, V. L., Arias, P. L., Cambra, J. F., Loricera, C. V., and Fierro, J. L. G. Enhancement of phenol hydrodeoxygenation over Pd catalysts supported on mixed HY zeolite and Al<sub>2</sub>O<sub>3</sub>. An approach to O-removal from bio-oils. *Fuel* **2014**, *117, Part B*, 1061-1073.
81. Zhao, C., Kasakov, S., He, J., and Lercher, J. A. Comparison of kinetics, activity and stability of Ni/HZSM-5 and Ni/Al<sub>2</sub>O<sub>3</sub>-HZSM-5 for phenol hydrodeoxygenation. *J. Catal.* **2012**, *296*, 12-23.
82. Güvenatam, B., Kurşun, O., Heeres, E. H. J., Pidko, E. A., and Hensen, E. J. M. Hydrodeoxygenation of mono- and dimeric lignin model compounds on noble metal catalysts. *Catal. Today* **2014**, *233*, 83-91.
83. Feng, B., Kobayashi, H., Ohta, H., and Fukuoka, A. Aqueous-phase hydrodeoxygenation of 4-propylphenol as a lignin model to n-propylbenzene over Re-Ni/ZrO<sub>2</sub> catalysts. *J. Mol. Catal. A: Chem.* **2014**, *388-389*, 41-46.
84. Sun, J., Karim, A. M., Zhang, H., Kovarik, L., Li, X. S., Hensley, A. J., McEwen, J.-S., and Wang, Y. Carbon-supported bimetallic Pd-Fe catalysts for vapor-phase hydrodeoxygenation of guaiacol. *J. Catal.* **2013**, *306*, 47-57.

85. Shin, E.-J., and Keane, M. A. Gas-Phase Hydrogenation/Hydrogenolysis of Phenol over Supported Nickel Catalysts. *Ind. Eng. Chem. Res.* **2000**, *39*, 883-892.
86. Ohta, H., Kobayashi, H., Hara, K., and Fukuoka, A. Hydrodeoxygenation of phenols as lignin models under acid-free conditions with carbon-supported platinum catalysts. *Chem. Commun.* **2011**, *47*, 12209-12211.
87. Nelson, R. C., Baek, B., Ruiz, P., Goundie, B., Brooks, A., Wheeler, M. C., Frederick, B. G., Grabow, L. C., and Austin, R. N. Experimental and Theoretical Insights into the Hydrogen-Efficient Direct Hydrodeoxygenation Mechanism of Phenol over Ru/TiO<sub>2</sub>. *ACS Catal.* **2015**, *5*, 6509-6523.
88. de Souza, P. M., Rabelo-Neto, R. C., Borges, L. E. P., Jacobs, G., Davis, B. H., Sooknoi, T., Resasco, D. E., and Noronha, F. B. Role of Keto Intermediates in the Hydrodeoxygenation of Phenol over Pd on Oxophilic Supports. *ACS Catal.* **2015**, *5*, 1318-1329.
89. Nie, L., de Souza, P. M., Noronha, F. B., An, W., Sooknoi, T., and Resasco, D. E. Selective conversion of m-cresol to toluene over bimetallic Ni-Fe catalysts. *J. Mol. Catal. A: Chem.* **2014**, *388-389*, 47-55.
90. Yoon, Y., Rousseau, R., Weber, R. S., Mei, D., and Lercher, J. A. First-Principles Study of Phenol Hydrogenation on Pt and Ni Catalysts in Aqueous Phase. *J. Am. Chem. Soc.* **2014**, *136*, 10287-10298.
91. Honkela, M. L., Bjork, J., and Persson, M. Computational study of the adsorption and dissociation of phenol on Pt and Rh surfaces. *Phys. Chem. Chem. Phys.* **2012**, *14*, 5849-5854.
92. Delle Site, L., Alavi, A., and Abrams, C. F. Adsorption energies and geometries of phenol on the (111) surface of nickel: An *ab initio* study. *Phys. Rev. B* **2003**, *67*, 193406.
93. Ghiringhelli, L. M., Caputo, R., and Site, L. D. Phenol near Ni(111), Ni(110), and Ni(221) surfaces in a vertical ring geometry: A density functional study of the oxygen-surface bonding and O-H cleavage. *Phys. Rev. B* **2007**, *75*, 113403.
94. Rangarajan, S., Bhan, A., and Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Chem. Eng. Sci.* **2012**, *46*, 141-152.

95. Kresse, G., and Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169-11186.
96. Perdew, J. P., Burke, K., and Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865-3868.
97. Grimme, S., Antony, J., Ehrlich, S., and 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*.
98. Liu, W., Carrasco, J., Santra, B., Michaelides, A., Scheffler, M., and Tkatchenko, A. Benzene adsorbed on metals: Concerted effect of covalency and van der Waals bonding. *Phys. Rev. B* **2012**, *86*, 245405.
99. Wei, L., Victor, G. R., Guo-Xu, Z., Biswajit, S., Xinguo, R., Matthias, S., and Alexandre, T. Structure and energetics of benzene adsorbed on transition-metal surfaces: density-functional theory with van der Waals interactions including collective substrate response. *New J. Phys.* **2013**, *15*, 053046.
100. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953-17979.
101. Kresse, G., and Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758-1775.
102. Murnaghan, F. D. The Compressibility of Media under Extreme Pressures. *Proc. Natl. Acad. Sci. U.S.A.* **1944**, *30*, 244-247.
103. Birch, F. Finite Elastic Strain of Cubic Crystals. *Phys. Rev.* **1947**, *71*, 809-824.
104. Monkhorst, H. J., and Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188-5192.
105. Ashcroft, N., and Mermin, D. *Solid state physics*. Thomson Learning: 1976.
106. Methfessel, M., and Paxton, A. T. High-precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B* **1989**, *40*, 3616-3621.
107. Henkelman, G., and Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010-7022.

108. Henkelman, G., Uberuaga, B. P., and Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901-9904.
109. Coltrin, M. E., Kee, R. J., and Rupley, F. M. *Surface CHEMKIN (Version 4.0): A Fortran package for analyzing heterogeneous chemical kinetics at a solid-surface---gas-phase interface*; SAND-90-8003B; Other: ON: DE92003583 United States10.2172/6128661Other: ON: DE92003583Thu Feb 07 17:18:19 EST 2008OSTI; NTIS; GPO Dep.SNL; EDB-92-009854English; 1991; p Medium: ED; Size: Pages: (91 p).
110. Kandoi, S., Greeley, J., Sanchez-Castillo, M., Evans, S., Gokhale, A., Dumesic, J., and Mavrikakis, M. Prediction of Experimental Methanol Decomposition Rates on Platinum from First Principles. *Top. Catal.* **2006**, *37*, 17-28.
111. Mhadeshwar, A. B., and Vlachos, D. G. Hierarchical, multiscale surface reaction mechanism development: CO and H<sub>2</sub> oxidation, water–gas shift, and preferential oxidation of CO on Rh. *J. Catal.* **2005**, *234*, 48-63.
112. Saliccioli, M. S., Chen, Y., and Vlachos, D. G. Microkinetic Modeling and Reduced Rate Expressions of Ethylene Hydrogenation and Ethane Hydrogenolysis on Platinum. *Ind. Eng. Chem. Res.* **2011**, *50*, 28-40.
113. Curtiss, L. A., Redfern, P. C., and Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.
114. Blaylock, D. W., Ogura, T., Green, W. H., and Beran, G. J. O. Computational Investigation of Thermochemistry and Kinetics of Steam Methane Reforming on Ni(111) under Realistic Conditions. *J. Phys. Chem. C* **2009**, *113*, 4898-4908.
115. Zaffran, J., Michel, C., Auneau, F., Delbecq, F., and Sautet, P. Linear Energy Relations As Predictive Tools for Polyalcohol Catalytic Reactivity. *ACS Catal.* **2014**, *4*, 464-468.
116. Wang, S., Vorotnikov, V., Sutton, J. E., and Vlachos, D. G. Brønsted–Evans–Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules. *ACS Catal.* **2014**, *4*, 604-612.

117. Lee, K., Gu, G. H., Mullen, C. A., Boateng, A. A., and Vlachos, D. G. Guaiacol Hydrodeoxygenation Mechanism on Pt(111): Insights from Density Functional Theory and Linear Free Energy Relations. *ChemSusChem* **2015**, *8*, 315-322.
118. Neurock, M., Pallassana, V., and van Santen, R. A. The Importance of Transient States at Higher Coverages in Catalytic Reactions. *J. Am. Chem. Soc.* **2000**, *122*, 1150-1153.
119. Morin, C., Simon, D., and Sautet, P. Intermediates in the hydrogenation of benzene to cyclohexene on Pt(111) and Pd(111): A comparison from DFT calculations. *Surf. Sci.* **2006**, *600*, 1339-1350.
120. Mahata, N., Raghavan, K. V., Vishwanathan, V., Park, C., and Keane, M. A. Phenol hydrogenation over palladium supported on magnesia: Relationship between catalyst structure and performance. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2712-2719.
121. Matos, J., and Corma, A. Selective phenol hydrogenation in aqueous phase on Pd-based catalysts supported on hybrid TiO<sub>2</sub>-carbon materials. *Appl. Catal., A* **2011**, *404*, 103-112.
122. Park, C., and Keane, M. A. Catalyst support effects: gas-phase hydrogenation of phenol over palladium. *J. Colloid Interface Sci.* **2003**, *266*, 183-194.
123. Watanabe, S., and Arunajatesan, V. Influence of Acid Modification on Selective Phenol Hydrogenation Over Pd/Activated Carbon Catalysts. *Top. Catal.* **2010**, *53*, 1150-1152.
124. Xiang, Y., Kong, L., Xie, P., Xu, T., Wang, J., and Li, X. Carbon Nanotubes and Activated Carbons Supported Catalysts for Phenol in Situ Hydrogenation: Hydrophobic/Hydrophilic Effect. *Ind. Eng. Chem. Res.* **2014**, *53*, 2197-2203.
125. Wan, H., Chaudhari, R., and Subramaniam, B. Catalytic Hydroprocessing of p-Cresol: Metal, Solvent and Mass-Transfer Effects. *Top. Catal.* **2012**, *55*, 129-139.
126. Sabbe, M. K., Canduela-Rodriguez, G., Reyniers, M.-F., and Marin, G. B. DFT-based modeling of benzene hydrogenation on Pt at industrially relevant coverage. *J. Catal.* **2015**, *330*, 406-422.
127. Sutton, J. E., Panagiotopouou, P., Veryldos, X. E., and Vlachos, D. G. Combined DFT, Microkinetic, and Experimental Study of Ethanol Steam Reforming on Pt. *J. Phys. Chem. C* **2013**, *117*, 4691-4706.

128. Mortensen, P. M., Grunwaldt, J. D., Jensen, P. A., Knudsen, K. G., and Jensen, A. D. A review of catalytic upgrading of bio-oil to engine fuels. *Appl. Catal., A* **2011**, *407*, 1-19.
129. Delle Site, L., Alavi, A., and Abrams, C. F. Adsorption energies and geometries of phenol on the (111) surface of nickel: An *ab initio* study. *Phys. Rev. B* **2003**, *67*, 193406.
130. Griffin, M. B., Ferguson, G. A., Ruddy, D. A., Bidy, M. J., Beckham, G. T., and Schaidle, J. A. Role of the Support and Reaction Conditions on the Vapor-Phase Deoxygenation of m-Cresol over Pt/C and Pt/TiO<sub>2</sub> Catalysts. *ACS Catal.* **2016**, *6*, 2715-2727.
131. Chiu, C.-c., Genest, A., Borgna, A., and Rösch, N. Hydrodeoxygenation of Guaiacol over Ru(0001): A DFT Study. *ACS Catal.* **2014**, *4*, 4178-4188.
132. Lu, J., Behtash, S., Mamun, O., and Heyden, A. Theoretical Investigation of the Reaction Mechanism of the Guaiacol Hydrogenation over a Pt(111) Catalyst. *ACS Catal.* **2015**, *5*, 2423-2435.
133. Robinson, A., Ferguson, G. A., Gallagher, J. R., Cheah, S., Beckham, G. T., Schaidle, J. A., Hensley, J. E., and Medlin, J. W. Enhanced Hydrodeoxygenation of m-Cresol over Bimetallic Pt–Mo Catalysts through an Oxophilic Metal-Induced Tautomerization Pathway. *ACS Catal.* **2016**, 4356-4368.
134. Gu, G. H., Mullen, C. A., Boateng, A. A., and Vlachos, D. G. Mechanism of Dehydration of Phenols on Noble Metals via First-Principles Microkinetic Modeling. *ACS Catal.* **2016**, *6*, 3047-3055.
135. Wan, H., Chaudhari, R. V., and Subramaniam, B. Catalytic Hydroprocessing of p-Cresol: Metal, Solvent and Mass-Transfer Effects. *Top. Catal.* **2012**, *55*, 129-139.
136. Foster, A. J., Do, P. T. M., and Lobo, R. F. The Synergy of the Support Acid Function and the Metal Function in the Catalytic Hydrodeoxygenation of m-Cresol. *Top. Catal.* **2012**, *55*, 118-128.
137. Runnebaum, R. C., Nimmanwudipong, T., Block, D. E., and Gates, B. C. Catalytic conversion of compounds representative of lignin-derived bio-oils: a reaction network for guaiacol, anisole, 4-methylanisole, and cyclohexanone conversion catalysed by Pt/[gamma]-Al<sub>2</sub>O<sub>3</sub>. *Catal. Sci. Technol.* **2012**, *2*, 113-118.

138. Kua, J., Faglioni, F., and Goddard, W. A. Thermochemistry for Hydrocarbon Intermediates Chemisorbed on Metal Surfaces:  $\text{CH}_n\text{-m}(\text{CH}_3)_m$  with  $n = 1, 2, 3$  and  $m \leq n$  on Pt, Ir, Os, Pd, Rh, and Ru. *J. Am. Chem. Soc.* **2000**, *122*, 2309-2321.
139. Bader, R. F. W., and Bayles, D. Properties of Atoms in Molecules: Group Additivity. *J. Phys. Chem. A* **2000**, *104*, 5579-5589.
140. Kresse, G., and Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169-11186.
141. Blöchl, P. E., Jepsen, O., and Andersen, O. K. Improved tetrahedron method for Brillouin-zone integrations. *Phys. Rev. B* **1994**, *49*, 16223-16233.
142. Bader, R. F. W. *Atoms in Molecules*. Oxford University Press: Oxford, U.K., 1994.
143. Otero-de-la-Roza, A., Johnson, E. R., and Luaña, V. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Comput. Phys. Commun.* **2014**, *185*, 1007-1018.
144. Benson, S. W., Cruickshank, F. R., Golden, D. M., Haugen, G. R., O'Neal, H. E., Rodgers, A. S., Shaw, R., and Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279-324.
145. Noor, E., Haraldsdóttir, H. S., Milo, R., and Fleming, R. M. T. Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.* **2013**, *9*, e1003098.
146. Mullen, C. A., and Boateng, A. A. Chemical Composition of Bio-oils Produced by Fast Pyrolysis of Two Energy Crops. *Energy Fuels* **2008**, *22*, 2104-2109.
147. Mar, B. D., Qi, H. W., Liu, F., and Kulik, H. J. Ab Initio Screening Approach for the Discovery of Lignin Polymer Breaking Pathways. *J. Phys. Chem. A* **2015**, *119*, 6551-6562.
148. Rangarajan, S., Bhan, A., and Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Chem. Eng. Sci.* **2012**, *46*, 141-152.

149. Rangarajan, S., Bhan, A., and Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Description of RING. *Chem. Eng. Sci.* **2012**, *45*, 114-123.
150. Fahmi, A., and van Santen, R. A. Density functional study of acetylene and ethylene adsorption on Ni(111). *Surf. Sci.* **1997**, *371*, 53-62.
151. Calhorda, M. J., Lopes, P. E. M., and Friend, C. M. Theoretical studies of ethylene adsorption and oxidation on clean and oxygen covered rhodium (111). *J. Mol. Catal. A: Chem.* **1995**, *97*, 157-171.
152. Kokalj, A., Dal Corso, A., de Gironcoli, S., and Baroni, S. Adsorption of ethylene on the Ag(0 0 1) surface. *Surf. Sci.* **2002**, *507-510*, 62-68.
153. Öström, H., Föhlisch, A., Nyberg, M., Weinelt, M., Heske, C., Pettersson, L. G. M., and Nilsson, A. Ethylene on Cu(1 1 0) and Ni(1 1 0): electronic structure and bonding derived from X-ray spectroscopy and theory. *Surf. Sci.* **2004**, *559*, 85-99.
154. Yamagishi, S., Jenkins, S. J., and King, D. A. Symmetry and site selectivity in molecular chemisorption: Benzene on Ni{111}. *J. Chem. Phys.* **2001**, *114*, 5765-5773.
155. Cyrański, M. K. Energetic Aspects of Cyclic Pi-Electron Delocalization: Evaluation of the Methods of Estimating Aromatic Stabilization Energies. *Chem. Rev.* **2005**, *105*, 3773-3811.
156. Rodríguez, J. I., Ayers, P. W., Götz, A. W., and Castillo-Alvarado, F. L. Virial theorem in the Kohn-Sham density-functional theory formalism: Accurate calculation of the atomic quantum theory of atoms in molecules energies. *J. Chem. Phys.* **2009**, *131*, 021101.
157. Yu, M., Trinkle, D. R., and Martin, R. M. Energy density in density functional theory: Application to crystalline defects and surfaces. *Phys. Rev. B* **2011**, *83*, 115113.
158. Bader, R. F. W., Slee, T. S., Cremer, D., and Kraka, E. Description of conjugation and hyperconjugation in terms of electron distributions. *J. Am. Chem. Soc.* **1983**, *105*, 5061-5068.
159. Wiberg, K. B., Bader, R. F. W., and Lau, C. D. H. Theoretical analysis of hydrocarbon properties. 1. Bonds, structures, charge concentrations, and charge relaxations. *J. Am. Chem. Soc.* **1987**, *109*, 985-1001.

160. López, C. S., Faza, O. N., Cossío, F. P., York, D. M., and de Lera, A. R. Ellipticity: A Convenient Tool To Characterize Electrocyclic Reactions. *Chem. Eur. J* **2005**, *11*, 1734-1738.
161. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis, Third Edition*. CRC Press: 2013.
162. Benson, S. W., and Buss, J. H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *J. Chem. Phys.* **1958**, *29*, 546-572.
163. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.
164. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., and Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024-10035.
165. Gilbert, T. M. Application of a Semi-homodesmotic Approach in Estimating Ring Strain Energies (RSEs) of Highly Substituted Cyclobutanes: RSEs for c-C<sub>4</sub>R<sub>8</sub> That Make Sense. *J. Phys. Chem. A* **2014**, *118*, 6060-6067.
166. Wiberg, K. B., Bader, R. F. W., and Lau, C. D. H. Theoretical analysis of hydrocarbon properties. 2. Additivity of group properties and the origin of strain energy. *J. Am. Chem. Soc.* **1987**, *109*, 1001-1012.
167. Huber, G. W., Shabaker, J. W., and Dumesic, J. A. Raney Ni-Sn Catalyst for H<sub>2</sub> Production from Biomass-Derived Hydrocarbons. *Science* **2003**, *300*, 2075-2077.
168. Laurent, E., and Delmon, B. Influence of Water in the Deactivation of a Sulfided NiMo $\gamma$ -Al<sub>2</sub>O<sub>3</sub> Catalyst During Hydrodeoxygenation. *J. Catal.* **1994**, *146*, 281-291.
169. Beerthuis, R., Rothenberg, G., and Shiju, N. R. Catalytic Routes Towards Acrylic Acid, Adipic Acid and  $\epsilon$ -Caprolactam Starting from Biorenewables. *Green Chem.* **2015**, *17*, 1341-1361.
170. Yan, N., Zhao, C., Luo, C., Dyson, P. J., Liu, H., and Kou, Y. One-Step Conversion of Cellobiose to C<sub>6</sub>-Alcohols Using a Ruthenium Nanocluster Catalyst. *J. Am. Chem. Soc.* **2006**, *128*, 8714-8715.

171. Xiao, C.-x., Cai, Z.-p., Wang, T., Kou, Y., and Yan, N. Aqueous-Phase Fischer–Tropsch Synthesis with a Ruthenium Nanocluster Catalyst. *Angew. Chem. Int. Ed.* **2008**, *47*, 746-749.
172. Lee, J., Xu, Y., and Huber, G. W. High-Throughput Screening of Monometallic Catalysts for Aqueous-Phase Hydrogenation of Biomass-Derived Oxygenates. *Appl. Catal., B* **2013**, *140–141*, 98-107.
173. Wang, Y., Yao, J., Li, H., Su, D., and Antonietti, M. Highly Selective Hydrogenation of Phenol and Derivatives over a Pd@Carbon Nitride Catalyst in Aqueous Media. *J. Am. Chem. Soc.* **2011**, *133*, 2362-2365.
174. Wertz, D. H. Relationship between the Gas-Phase Entropies of Molecules and Their Entropies of Solvation in Water and 1-Octanol. *J. Am. Chem. Soc.* **1980**, *102*, 5316-5322.
175. Desai, S. K., and Neurock, M. First-Principles Study of the Role of Solvent in the Dissociation of Water over a Pt-Ru Alloy. *Phys. Rev. B* **2003**, *68*, 075420.
176. Steinmann, S. N., Sautet, P., and Michel, C. Solvation Free Energies for Periodic Surfaces: Comparison of Implicit and Explicit Solvation Models. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31850-31861.
177. Taylor, C. D., and Neurock, M. Theoretical Insights into the Structure and Reactivity of the Aqueous/Metal Interface. *Curr. Opin. Solid State Mater. Sci.* **2005**, *9*, 49-65.
178. Akpa, B. S., D’Agostino, C., Gladden, L. F., Hindle, K., Manyar, H., McGregor, J., Li, R., Neurock, M., Sinha, N., Stitt, E. H., Weber, D., Zeitler, J. A., and Rooney, D. W. Solvent Effects in the Hydrogenation of 2-Butanone. *J. Catal.* **2012**, *289*, 30-41.
179. Liu, J., Cao, X.-M., and Hu, P. Density Functional Theory Study on the Activation of Molecular Oxygen on a Stepped Gold Surface in an Aqueous Environment: a New Approach for Simulating Reactions in Solution. *Phys. Chem. Chem. Phys.* **2014**, *16*, 4176-4185.
180. Carter, E. A., Ciccotti, G., Hynes, J. T., and Kapral, R. Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.* **1989**, *156*, 472-477.
181. Sprik, M., and Ciccotti, G. Free Energy from Constrained Molecular Dynamics. *J. Chem. Phys.* **1998**, *109*, 7737-7744.

182. Jindal, G., and Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **2016**, *120*, 9913-9921.
183. Bellarosa, L., García-Muelas, R., Revilla-López, G., and López, N. Diversity at the Water–Metal Interface: Metal, Water Thickness, and Confinement Effects. *ACS Cent. Sci.* **2016**, *2*, 109-116.
184. Desai, S. K., Pallassana, V., and Neurock, M. A Periodic Density Functional Theory Analysis of the Effect of Water Molecules on Deprotonation of Acetic Acid over Pd(111). *J. Phys. Chem. B* **2001**, *105*, 9171-9182.
185. Michel, C., Zaffran, J., Ruppert, A. M., Matras-Michalska, J., Jedrzejczyk, M., Grams, J., and Sautet, P. Role of Water in Metal Catalyst Performance for Ketone Hydrogenation: a Joint Experimental and Theoretical Study on Levulinic Acid Conversion into Gamma-Valerolactone. *Chem. Commun.* **2014**, *50*, 12450-12453.
186. Cao, D., Lu, G. Q., Wieckowski, A., Wasileski, S. A., and Neurock, M. Mechanisms of Methanol Decomposition on Platinum: A Combined Experimental and *ab Initio* Approach. *J. Phys. Chem. B* **2005**, *109*, 11622-11633.
187. Zope, B. N., Hibbitts, D. D., Neurock, M., and Davis, R. J. Reactivity of the Gold/Water Interface During Selective Oxidation Catalysis. *Science* **2010**, *330*, 74-78.
188. Mathew, K., Sundararaman, R., Letchworth-Weaver, K., Arias, T. A., and Hennig, R. G. Implicit Solvation Model for Density-Functional Study of Nanocrystal Surfaces and Reaction Pathways. *J. Chem. Phys.* **2014**, *140*, 084106.
189. Garcia-Ratés, M., and López, N. Multigrid-Based Methodology for Implicit Solvation Models in Periodic DFT. *J. Chem. Theory Comput.* **2016**, *12*, 1331-1341.
190. Pliego, J. R., and Riveros, J. M. The Cluster–Continuum Model for the Calculation of the Solvation Free Energy of Ionic Species. *J. Phys. Chem. A* **2001**, *105*, 7241-7247.
191. Asthagiri, D., Pratt, L. R., and Ashbaugh, H. S. Absolute Hydration Free Energies of Ions, Ion–Water Clusters, and Quasichemical Theory. *J. Chem. Phys.* **2003**, *119*, 2702-2708.

192. Bryantsev, V. S., Diallo, M. S., and Goddard Iii, W. A. Calculation of Solvation Free Energies of Charged Solutes Using Mixed Cluster/Continuum Models. *J. Phys. Chem. B* **2008**, *112*, 9709-9719.
193. Wang, H.-F., and Liu, Z.-P. Formic Acid Oxidation at Pt/H<sub>2</sub>O Interface from Periodic DFT Calculations Integrated with a Continuum Solvation Model. *J. Phys. Chem. C* **2009**, *113*, 17502-17508.
194. Steinmann, S. N., Michel, C., Schwiedernoch, R., and Sautet, P. Impacts of Electrode Potentials and Solvents on the Electroreduction of CO<sub>2</sub>: a Comparison of Theoretical Approaches. *Phys. Chem. Chem. Phys.* **2015**, *17*, 13949-13963.
195. Wang, P., Steinmann, S. N., Fu, G., Michel, C., and Sautet, P. Key Role of Anionic Doping for H<sub>2</sub> Production from Formic Acid on Pd(111). *ACS Catal.* **2017**, *7*, 1955-1959.
196. Eigenmann, H. K., Golden, D. M., and Benson, S. W. Revised Group Additivity Parameters for the Enthalpies of Formation of Oxygen-Containing Organic Compounds. *J. Phys. Chem.* **1973**, *77*, 1687-1691.
197. Stein, S. E., Golden, D. M., and Benson, S. W. Predictive Scheme for Thermochemical Properties of Polycyclic Aromatic Hydrocarbons. *J. Phys. Chem.* **1977**, *81*, 314-317.
198. Benson, S. W. Thermochemistry and Kinetics of Sulfur-Containing Molecules and Radicals. *Chem. Rev.* **1978**, *78*, 23-35.
199. Willems, P. A., and Froment, G. F. Kinetic Modeling of the Thermal Cracking of Hydrocarbons. 1. Calculation of Frequency Factors. *Ind. Eng. Chem. Res.* **1988**, *27*, 1959-1966.
200. Willems, P. A., and Froment, G. F. Kinetic Modeling of the Thermal Cracking of Hydrocarbons. 2. Calculation of Activation Energies. *Ind. Eng. Chem. Res.* **1988**, *27*, 1966-1971.
201. Martinis, J. M., and Froment, G. F. Alkylation on Solid Acids. Part 2. Single-Event Kinetic Modeling. *Ind. Eng. Chem. Res.* **2006**, *45*, 954-967.
202. Luria, M., and Benson, S. W. Heat Capacities of Liquid Hydrocarbons. Estimation of Heat Capacities at Constant Pressure as a Temperature Function, Using Additivity Rules. *J. Chem. Eng. Data* **1977**, *22*, 90-100.

203. Constantinou, L., Gani, R., and O'Connell, J. P. Estimation of the Acentric Factor and the Liquid Molar Volume at 298 K Using a New Group Contribution Method. *FFE* **1995**, *103*, 11-22.
204. Constantinou, L., and Gani, R. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697-1710.
205. Wu, D. T. Proceedings of the Fourth International Conference: Prediction of Viscosities of Liquid Mixtures by a Group Contribution Method. *FFE* **1986**, *30*, 149-156.
206. Ihmels, E. C., and Gmehling, J. Extension and Revision of the Group Contribution Method GCVOL for the Prediction of Pure Compound Liquid Densities. *Ind. Eng. Chem. Res.* **2003**, *42*, 408-412.
207. Abrams, D. S., and Prausnitz, J. M. Statistical Thermodynamics of Liquid Mixtures: A New Expression for the Excess Gibbs Energy of Partly or Completely Miscible Systems. *AIChE J.* **1975**, *21*, 116-128.
208. Fredenslund, A., Jones, R. L., and Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086-1099.
209. Chapman, W. G., Gubbins, K. E., Jackson, G., and Radosz, M. New Reference Equation of State for Associating Liquids. *Ind. Eng. Chem. Res.* **1990**, *29*, 1709-1721.
210. Lymperiadis, A., Adjiman, C. S., Galindo, A., and Jackson, G. A Group Contribution Method for Associating Chain Molecules Based on the Statistical Associating Fluid Theory (SAFT- $\gamma$ ). *J. Chem. Phys.* **2007**, *127*, 234903.
211. Gu, G. H., and Vlachos, D. G. Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111). *J. Chem. Phys. C* **2016**, *120*, 19234-19241.
212. Kresse, G., and Hafner, J. *Ab initio* Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47*, 558-561.
213. Steinmann, S. N., and Corminboeuf, C. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *J. Chem. Theory Comput.* **2011**, *7*, 3567-3577.

214. Steinmann, S. N., and Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134*, 044117.
215. Fishman, M., Zhuang, H. L., Mathew, K., Dirschka, W., and Hennig, R. G. Accuracy of Exchange-Correlation Functionals and Effect of Solvation on the Surface Energy of Copper. *Phys. Rev. B* **2013**, *87*, 245402.
216. Gu, G. H., Schweitzer, B., Michel, C., Steinmann, S. N., Sautet, P., and Vlachos, D. G. Group Additivity for Aqueous Phase Thermochemical Properties of Alcohols on Pt(111). *J. Phys. Chem. C* **2017**, *121*, 21510-21519.
217. Ben-Naim, A. *Solvation Thermodynamics*. Springer US: New York, 1987.
218. Ben-Naim, A. *Statistical Thermodynamics for Chemists and Biochemists*. Springer US: New York, 1992.
219. Cooper, J., and Ziegler, T. A Density Functional Study of SN2 Substitution at Square-Planar Platinum(II) Complexes. *Inorg. Chem.* **2002**, *41*, 6614-6622.
220. Golze, D., Iannuzzi, M., Nguyen, M.-T., Passerone, D., and Hutter, J. Simulation of Adsorption Processes at Metallic Interfaces: An Image Charge Augmented QM/MM Approach. *J. Chem. Theory Comput.* **2013**, *9*, 5086-5097.
221. Schiros, T., Andersson, K. J., Pettersson, L. G. M., Nilsson, A., and Ogasawara, H. Chemical Bonding of Water to Metal Surfaces Studied with Core-Level Spectroscopies. *J. Electron. Spectrosc. Relat. Phenom.* **2010**, *177*, 85-98.
222. Limmer, D. T., Willard, A. P., Madden, P., and Chandler, D. Hydration of Metal Surfaces Can be Dynamically Heterogeneous and Hydrophobic. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 4200-4205.
223. Sakong, S., Forster-Tonigold, K., and Groß, A. The Structure of Water at a Pt(111) Electrode and the Potential of Zero Charge Studied from First Principles. *J. Chem. Phys.* **2016**, *144*, 194701.
224. Carlsson, J., and Åqvist, J. Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding. *J. Phys. Chem. B* **2005**, *109*, 6448-6456.

225. Natarajan, S. K., and Behler, J. Neural Network Molecular Dynamics Simulations of Solid-Liquid Interfaces: Water at Low-Index Copper Surfaces. *Phys. Chem. Chem. Phys.* **2016**, *18*, 28704-28725.
226. Kelly, E., Seth, M., and Ziegler, T. Calculation of Free Energy Profiles for Elementary Bimolecular Reactions by ab Initio Molecular Dynamics: Sampling Methods and Thermostat Considerations. *J. Phys. Chem. A* **2004**, *108*, 2167-2180.
227. Price, D. L., and Halley, J. W. Molecular Dynamics, Density Functional Theory of the Metal–Electrolyte Interface. *J. Chem. Phys.* **1995**, *102*, 6603-6612.
228. Lee, M.-S., Peter McGrail, B., Rousseau, R., and Glezakou, V.-A. Structure, Dynamics and Stability of Water/scCO<sub>2</sub>/Mineral Interfaces from ab initio Molecular Dynamics Simulations. *Sci. Rep.* **2015**, *5*, 14857.
229. Cheng, T., Xiao, H., and Goddard, W. A. Full Atomistic Reaction Mechanism with Kinetics for CO Reduction on Cu(100) from ab initio Molecular Dynamics Free-Energy Calculations at 298 K. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 1795-1800.
230. Cantu, D. C., Wang, Y.-G., Yoon, Y., Glezakou, V.-A., Rousseau, R., and Weber, R. S. Heterogeneous Catalysis in Complex, Condensed Reaction Media. *Catal. Today* **2017**, *289*, 231-236.
231. Schravendijk, P., van der Vegt, N., Delle Site, L., and Kremer, K. Dual-Scale Modeling of Benzene Adsorption onto Ni(111) and Au(111) Surfaces in Explicit Water. *ChemPhysChem* **2005**, *6*, 1866-1871.
232. Li, X., and Ågren, H. Molecular Dynamics Simulations Using a Capacitance–Polarizability Force Field. *J. Phys. Chem. C* **2015**, *119*, 19430-19437.
233. Spohr, E. Computer simulation of the water/platinum interface. *J. Phys. Chem.* **1989**, *93*, 6171-6180.
234. Bengio, Y., and Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* **2004**, *5*, 1089-1105.
235. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada, 1995; pp 1137-1143.

236. Sutton, J. E., Panagiotopoulou, P., Verykios, X. E., and Vlachos, D. G. Combined DFT, Microkinetic, and Experimental Study of Ethanol Steam Reforming on Pt. *J. Phys. Chem. C* **2013**, *117*, 4691-4706.
237. Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Gound. Crystallogr.* **1976**, *32*, 922-923.
238. Walker, M. W., Shao, L., and Volz, R. A. Estimating 3-D Location Parameters Using Dual Number Quaternions. *CVGIP: Image Underst.* **1991**, *54*, 358-367.
239. Kromann, J. C., Steinmann, C., larsbratholm, aandi, and Lauritzen, K. P. RMSD: Small Release Update. 10.5281/zenodo.46697.
240. Zaffran, J., Michel, C., Delbecq, F., and Sautet, P. Towards More Accurate Prediction of Activation Energies for Polyalcohol Dehydrogenation on Transition Metal Catalysts in Water. *Catal. Sci. Technol.* **2016**, *6*, 6615-6624.
241. Garcia-Ratés, M., García-Muelas, R., and López, N. Solvation Effects on Methanol Decomposition on Pd(111), Pt(111), and Ru(0001). *J. Phys. Chem. C* **2017**, *121*, 13803-13809.
242. Sabbe, M. K., Van Geem, K. M., Reyniers, M.-F., and Marin, G. B. First principle-based simulation of ethane steam cracking. *AIChE J.* **2011**, *57*, 482-496.
243. Ranzi, E. A Wide-Range Kinetic Modeling Study of Oxidation and Combustion of Transportation Fuels and Surrogate Mixtures. *Energy Fuels* **2006**, *20*, 1024-1032.
244. Tran, L. S., Sirjean, B., Glaude, P.-A., Fournet, R., and Battin-Leclerc, F. Progress in detailed kinetic modeling of the combustion of oxygenated components of biofuels. *Energy* **2012**, *43*, 4-18.
245. Saliccioli, M., and Vlachos, D. G. Kinetic Modeling of Pt Catalyzed and Computation-Driven Catalyst Discovery for Ethylene Glycol Decomposition. *ACS Catal.* **2011**, *1*, 1246-1256.
246. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I., and Norskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **2006**, *5*, 909-913.
247. Norskov, J. K., Bligaard, T., Rossmeisl, J., and Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **2009**, *1*, 37-46.

248. Ince, A., Carstensen, H.-H., Reyniers, M.-F., and Marin, G. B. First-principles based group additivity values for thermochemical properties of substituted aromatic compounds. *AIChE J.* **2015**, *61*, 3858-3870.
249. Ince, A., Carstensen, H.-H., Sabbe, M., Reyniers, M.-F., and Marin, G. B. Group additive modeling of substituent effects in monocyclic aromatic hydrocarbon radicals. *AIChE J.* **2017**, *63*, 2089-2106.
250. Sabbe, M. K., Saeys, M., Reyniers, M.-F., Marin, G. B., Van Speybroeck, V., and Waroquier, M. Group Additive Values for the Gas Phase Standard Enthalpy of Formation of Hydrocarbons and Hydrocarbon Radicals. *J. Phys. Chem. A* **2005**, *109*, 7466-7480.
251. Cabani, S., Gianni, P., Mollica, V., and Lepori, L. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.* **1981**, *10*, 563-595.
252. Chickos, J. S., Hesse, D. G., and Liebman, J. F. A Group Additivity Approach for the Estimation of Heat-Capacities of Organic Liquids and Solids at 298 K. *Struct. Chem.* **1993**, *4*, 261-269.
253. Elbro, H. S., Fredenslund, A., and Rasmussen, P. Group contribution method for the prediction of liquid densities as a function of temperature for solvents, oligomers, and polymers. *Ind. Eng. Chem. Res.* **1991**, *30*, 2576-2582.
254. Mavrovouniotis, M. L. Estimation of standard Gibbs energy changes of biotransformations. *J. Bio. Chem.* **1991**, *266*, 14440-14445.
255. Mavrovouniotis, M. L. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **1990**, *36*, 1070-1082.
256. Perron, G., and Desnoyers, J. E. Volumes and heat capacities of benzene derivatives in water at 25°C: Group additivity of the standard partial molal quantities. *FFE* **1979**, *2*, 239-262.
257. Růžička, V., and Domalski, E. S. Estimation of the Heat Capacities of Organic Liquids as a Function of Temperature using Group Additivity. II. Compounds of Carbon, Hydrogen, Halogens, Nitrogen, Oxygen, and Sulfur. *J. Phys. Chem. Ref. Data* **1993**, *22*, 619-657.
258. Shaw, R. Heat capacities of liquids. Estimation of heat capacity at constant pressure and 25.deg., using additivity rules. *J. Chem. Eng. Data* **1969**, *14*, 461-465.

259. Cyrański, M. K. Energetic aspects of cyclic pi-electron delocalization: evaluation of the methods of estimating aromatic stabilization energies. *Chem. Rev.* **2005**, *105*, 3773-3811.
260. Benson, S. W., Cruickshank, F., Golden, D., Haugen, G. R., O'neal, H., Rodgers, A., Shaw, R., and Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279-324.
261. Cohen, N., and Benson, S. Estimation of heats of formation of organic compounds by additivity methods. *Chem. Rev.* **1993**, *93*, 2419-2438.
262. Eigenmann, H., Golden, D., and Benson, S. Revised group additivity parameters for the enthalpies of formation of oxygen-containing organic compounds. *J. Phys. Chem.* **1973**, *77*, 1687-1691.
263. Bhattacharya, A., and Shivalkar, S. Re-tooling Benson's Group Additivity Method for Estimation of the Enthalpy of Formation of Free Radicals: C/H and C/H/O Groups. *J. Chem. Eng. Data* **2006**, *51*, 1169-1181.
264. Bjorkman, K. R., Sung, C.-Y., Mondor, E., Cheng, J. C., Jan, D.-Y., and Broadbelt, L. J. Group Additivity Determination for Enthalpies of Formation of Carbenium Ions. *Ind. Eng. Chem. Res.* **2014**, *53*, 19446-19452.
265. Cohen, N. Revised group additivity values for enthalpies of formation (at 298 K) of carbon-hydrogen and carbon-hydrogen-oxygen compounds. *J. Phys. Chem. Ref. Data* **1996**, *25*, 1411-1481.
266. Dellon, L. D., Sung, C.-Y., Robichaud, D. J., and Broadbelt, L. J. Group Additivity Determination for Oxygenates, Oxonium Ions, and Oxygen-Containing Carbenium Ions. *Ind. Eng. Chem. Res.* **2017**, *56*, 10259-10270.
267. Holmes, J. L., and Aubry, C. Group Additivity Values for Estimating the Enthalpy of Formation of Organic Compounds: An Update and Reappraisal. 2. C, H, N, O, S, and Halogens. *J. Phys. Chem. A* **2012**, *116*, 7196-7209.
268. Marsi, I., Viskolcz, B., and Seres, L. Application of the Group Additivity Method to Alkyl Radicals: An ab Initio Study. *J. Phys. Chem. A* **2000**, *104*, 4497-4504.
269. Sebbar, N., Bockhorn, H., and Bozzelli, J. W. Structures, thermochemical properties (enthalpy, entropy and heat capacity), rotation barriers, and peroxide bond energies of vinyl, allyl, ethynyl and phenyl hydroperoxides. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3691-3703.

270. Sebbar, N., Bozzelli, J. W., and Bockhorn, H. Thermochemical Properties, Rotation Barriers, Bond Energies, and Group Additivity for Vinyl, Phenyl, Ethynyl, and Allyl Peroxides. *J. Phys. Chem. A* **2004**, *108*, 8353-8366.
271. Hedwiga, G. R., and Hinz, H.-J. Group additivity schemes for the calculation of the partial molar heat capacities and volumes of unfolded proteins in aqueous solution. *Biophys. Chem.* **2002**, *100*, 239-260.
272. Avbelj, F., and Baldwin, R. L. Limited validity of group additivity for the folding energetics of the peptide group. *Proteins* **2006**, *63*, 283-289.
273. Bauzá, A., Quiñonero, D., Deyà, P. M., and Frontera, A. Estimating ring strain energies in small carbocycles by means of the Bader's theory of 'atoms-in-molecules'. *Chem. Phys. Lett.* **2012**, *536*, 165-169.
274. Bauzá, A., Streubel, R., and Frontera, A. On the relationship between ring strain energies and 'atoms-in-molecules' properties in N2P2 rings. *Chem. Phys. Lett.* **2014**, *597*, 40-44.
275. Cortés-Guzmán, F., and Bader, R. F. W. Role of functional groups in linear regression analysis of molecular properties. *J. Phys. Org. Chem.* **2004**, *17*, 95-99.
276. Gao, H., Bader, R. F. W., and Cortés-Guzmán, F. Energy additivity in branched and cyclic hydrocarbons. *Can. J. Chem.* **2009**, *87*, 1583-1592.
277. Khachatryan, A. A., Shamsutdinova, Z. I., and Varfolomeev, M. A. Group additivity approach for determination of solvation enthalpies of aromatic compounds in 1-butyl-3-methylimidazolium tetrafluoroborate based on solution calorimetry data. *J. Mol. Liq.* **2017**, *236*, 278-282.
278. Khachatryan, A. A., Shamsutdinova, Z. I., and Varfolomeev, M. A. Enthalpies of solution and enthalpies of solvation of chloro- and nitro-substituted benzenes in 1-butyl-3-methyl imidazolium based ionic liquids at 298.15 K: Additivity of group contributions. *Thermochim. Acta* **2016**, *645*, 1-6.
279. Sumathi, R., Carstensen, H. H., and Green, W. H. Reaction Rate Prediction via Group Additivity Part 1: H Abstraction from Alkanes by H and CH3. *J. Phys. Chem. A* **2001**, *105*, 6910-6925.
280. Ammon, H. L., and Mitchell, S. A New Atom/Functional Group Volume Additivity Data Base for the Calculation of the Crystal Densities of C, H, N, O and F-Containing Compounds. *PEP* **1998**, *23*, 260-265.

281. Yu, J., Sumathi, R., and Green, W. H. Accurate and Efficient Method for Predicting Thermochemistry of Polycyclic Aromatic Hydrocarbons – Bond-Centered Group Additivity. *J. Am. Chem. Soc.* **2004**, *126*, 12685-12700.
282. Sumathi, R., Carstensen, H. H., and Green, W. H. Reaction Rate Prediction via Group Additivity, Part 2: H-Abstraction from Alkenes, Alkynes, Alcohols, Aldehydes, and Acids by H Atoms. *J. Phys. Chem. A* **2001**, *105*, 8969-8984.
283. Sumathi, R., Carstensen, H.-H., and Green, W. H. Reaction Rate Predictions Via Group Additivity. Part 3: Effect of Substituents with CH<sub>2</sub> as the Mediator. *J. Phys. Chem. A* **2002**, *106*, 5474-5489.
284. Kolesov, V. P., Dorofeevab, O. V., Iorishb, V. S., Papina, T. S., Lukyanova, V. A., and Melkhanova, S. V. Experimental measurements and a group additivity approach for estimating the standard molar enthalpies of formation of dioxins. *Mendeleev Commun.* **1999**, *9*, 143-144.
285. Adamczyk, A. J., Reyniers, M.-F., Marin, G. B., and Broadbelt, L. J. Exploring 1,2-Hydrogen Shift in Silicon Nanoparticles: Reaction Kinetics from Quantum Chemical Calculations and Derivation of Transition State Group Additivity Database. *J. Phys. Chem. A* **2009**, *113*, 10933-10946.
286. Linstrom, P. J., and Millards, W. G. NIST Chemistry WebBook, NIST Standard Reference Database Number 69. National Institute of Standards and Technology: Gaithersburg MD, 20899, (accessed February 10, 2017).
287. Burcat, A. Extended Third Millennium Thermodynamic Database for Combustion and Air-Pollution Use with updates from Active Thermochemical Tables. <http://burcat.technion.ac.il/dir/BURCAT.BAK> (accessed February 10, 2017).
288. Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
289. Casalegno, M., Sello, G., and Benfenati, E. Definition and Detection of Outliers in Chemical Space. *J. Chem. Inf. Model.* **2008**, *48*, 1592-1601.
290. Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864-2875.

291. Lowe, D. M., Corbett, P. T., Murray-Rust, P., and Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51*, 739-753.
292. Glandrum, G. version 2016\_3\_1; RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
293. Cordero, B., Gomez, V., Platero-Prats, A. E., Reves, M., Echeverria, J., Cremades, E., Barragan, F., and Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832-2838.
294. Sanderson, R. T. Electronegativity and bond energy. *J. Am. Chem. Soc.* **1983**, *105*, 2259-2261.
295. Calle-Vallejo, F., Loffreda, D., KoperMarc, T. M., and Sautet, P. Introducing structural sensitivity into adsorption–energy scaling relations by means of coordination numbers. *Nat. Chem.* **2015**, *7*, 403-410.
296. Gu, G. H., and Vlachos, D. G. Adsorption Geometry Prediction via Machine Learning. *In preparation*.
297. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, 267-288.
298. Richtárik, P., and Takáč, M. Parallel Coordinate Descent Methods for Big Data Optimization. *arXiv:1212.0873* **2012**.
299. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407-499.
300. Yan, X., and Han, J. In *gspan: Graph-based substructure pattern mining*, Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, IEEE: 2002; pp 721-724.
301. Tsuda, K. Entire regularization paths for graph data. In *Proceedings of the 24th international conference on Machine learning*, ACM: Corvallis, Oregon, USA, 2007; pp 919-926.
302. *Spyder*, version 3.1.4 64 bit on Windows 10; 2017.
303. Pan, S., Wu, J., Zhu, X., Long, G., and Zhang, C. Finding the best not the most: regularized loss minimization subgraph selection for graph classification. *Pattern Recognit.* **2015**, *48*, 3783-3796.

304. Chiappa, S., Saigo, H., and Tsuda, K. A Bayesian Approach to Graph Regression with Relevant Subgraph Selection. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics: 2009; pp 295-304.
305. Saigo, H., Nowozin, S., Kadowaki, T., Kudo, T., and Tsuda, K. gBoost: a mathematical programming approach to graph classification and regression. *Mach. Learn.* **2009**, *75*, 69-89.
306. Benson, S. W. *Thermochemical kinetics: methods for the estimation of thermochemical data and rate parameters*. Wiley: 1968.
307. Rogers, D., and Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
308. Collins, C. R., Gordon, G. J., Lilienfeld, O. A. v., and Yaron, D. J. Constant Size Molecular Descriptors For Use With Machine Learning. *arXiv:1701.06649* **2017**.
309. Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O. A., Tkatchenko, A., and Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404-3419.
310. Huang, B., and Lilienfeld, O. A. v. The "DNA" of chemistry: Scalable quantum machine learning with "amons". *arXiv:1707.04146* **2017**.
311. Huang, B., and von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
312. Ramakrishnan, R., and von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *Chimia* **2015**, *69*, 182-186.
313. Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
314. Janet, J. P., and Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939-8954.

315. Montavon, G., Hansen, K., Fazli, S., Rupp, M., Biegler, F., Ziehe, A., Tkatchenko, A., Lilienfeld, A. V., and Müller, K.-R. Learning Invariant Representations of Molecules for Atomization Energy Prediction. *Adv. Neural Inf. Process. Syst.* **2012**, 25, 440.
316. Bartók, A. P., Kondor, R., and Csányi, G. On representing chemical environments. *arXiv:1209.3140* **2012**.
317. Yao, K., Herr, J. E., Brown, S. N., and Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, 8, 2689-2694.
318. Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R., and Lilienfeld, O. A. v. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, 15, 095003.
319. Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. Gated Graph Sequence Neural Networks. *arXiv:1511.05493* **2015**.
320. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. SchNet - a deep learning architecture for molecules and materials. *arXiv:1712.06113* **2017**.
321. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv:1509.09292* **2015**.
322. Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **2016**, 30, 595-608.
323. Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., Vinyals, O., Kearnes, S., Riley, P. F., and von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, 13, 5255-5264.
324. Yao, K., Herr, J., Brown, S., and Parkhill, J. Bond Energies from a Diatomics-in-Molecules Neural Network. *arXiv:1703.08640* **2017**.
325. Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. Exact post-selection inference, with application to the lasso. *Ann. Statist.* **2016**, 44, 907-927.

326. Liu, H., Xu, X., and Li, J. J. A Bootstrap Lasso + Partial Ridge Method to Construct Confidence Intervals for Parameters in High-dimensional Sparse Linear Models. *arXiv:1706.02150* **2017**.
327. Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. A significance test for the lasso. *Ann. Statist.* **2014**, *42*, 413-468.
328. Ratkovic, M., and Tingley, D. Sparse Estimation and Uncertainty with Application to Subgroup Analysis. *Polit. Anal.* **2017**, *25*, 1-40.
329. Reid, S., Tibshirani, R., and Friedman, J. A Study of Error Variance Estimation in Lasso Regression. *arXiv:1311.5274* **2013**.
330. Taylor, J., and Tibshirani, R. J. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 7629.
331. Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact Post-Selection Inference for Sequential Regression Procedures. *arXiv:1401.3889* **2014**.
332. Cawley, G. C., Talbot, N. L. C., and Chapelle, O. Estimating Predictive Variances with Kernel Ridge Regression. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, Quiñero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp 56-77.
333. Gal, Y. Uncertainty in Deep Learning. University of Cambridge, Cambridge, UK, 2016.
334. Shi, J., Chen, J., Zhu, J., Sun, S., Luo, Y., Gu, Y., and Zhou, Y. ZhuSuan: A Library for Bayesian Deep Learning. *arXiv:1709.05870* **2017**.
335. Wang, H., and Yeung, D.-Y. Towards Bayesian Deep Learning: A Survey. *arXiv:1604.01662* **2016**.
336. Borg, I., and Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*. Springer: 1997.
337. Gu, G. H., Plechac, P., and Vlachos, D. G. Thermochemistry of Gas-phase and Surface Species via LASSO-assisted Subgraph Selection. *Reaction Chemistry & Engineering* **2018**, *Just Accepted*.

338. Carr, S. F., Garnett, R., and Lo, C. S. Accelerating the search for global minima on potential energy surfaces using machine learning. *J. Chem. Phys.* **2016**, *145*, 154106.
339. Peterson, A. A. Global Optimization of Adsorbate–Surface Structures While Preserving Molecular Identity. *Top. Catal.* **2014**, *57*, 40-53.
340. Weininger, D., Weininger, A., and Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101.
341. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825-2830.
342. Van Santen, R. A. Complementary Structure Sensitive and Insensitive Catalytic Relationships. *Acc. Chem. Res.* **2009**, *42*, 57-66.
343. Mortensen, P. M., Grunwaldt, J.-D., Jensen, P. A., and Jensen, A. D. Influence on nickel particle size on the hydrodeoxygenation of phenol over Ni/SiO<sub>2</sub>. *Catal. Today* **2016**, *259*, 277-284.
344. Gao, W., Zheng, W. T., and Jiang, Q. Dehydrogenation of benzene on Pt(111) surface. *J. Chem. Phys.* **2008**, *129*, 164705.
345. Kandoi, S., Greeley, J., Sanchez-Castillo, M. A., Evans, S. T., Gokhale, A. A., Dumesic, J. A., and Mavrikakis, M. Prediction of Experimental Methanol Decomposition Rates on Platinum from First Principles. *Top. Catal.* **2006**, *37*, 17-28.
346. Saliccioli, M., Chen, Y., and Vlachos, D. G. Microkinetic Modeling and Reduced Rate Expressions of Ethylene Hydrogenation and Ethane Hydrogenolysis on Platinum. *Ind. Eng. Chem. Res.* **2011**, *50*, 28-40.
347. Sutton, J. E., Panagiotopoulou, P., Verykios, X. E., and Vlachos, D. G. Combined DFT, Microkinetic, and Experimental Study of Ethanol Steam Reforming on Pt. *J. Phys. Chem. C* **2013**, *117*, 4691-4706.
348. Morin, C., Simon, D., and Sautet, P. Intermediates in the hydrogenation of benzene to cyclohexene on Pt(1 1 1) and Pd(1 1 1): A comparison from DFT calculations. *Surf. Sci.* **2006**, *600*, 1339-1350.

349. Kobayashi, H., Kato, H., Tarama, K., and Fukui, K. A molecular orbital calculation of adsorption of ethylene and acetylene on nickel cluster. *J. Catal.* **1977**, *49*, 294-304.
350. Miller, G. J. *Solids and Surfaces: A Chemist's View of Bonding in Extended Structures*. Von R. Hoffmann. VCH Publishers, New York/VCH Verlagsgesellschaft, Weinheim 1988. X, 142 S., geb., \$ 24.95. – ISBN 0-89573-709-4/3-527-26905-3. *Angew. Chem.* **1989**, *101*, 1570-1571.
351. Ribeiro, R. F., Marenich, A. V., Cramer, C. J., and Truhlar, D. G. Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation. *J. Phys. Chem. B* **2011**, *115*, 14556-14562.
352. Giesen, D. J., Cramer, C. J., and Truhlar, D. G. Entropic Contributions to Free Energies of Solvation. *J. Phys. Chem.* **1994**, *98*, 4141-4147.
353. Wilson, S., and Diercksen, G. H. F. *Problem Solving in Computational Molecular Science: Molecules in Different Environments*. 1997.
354. Kreevoy, M. M., and Truhlar, D. G. *Investigations of Rates and Mechanisms of Reactions Part I*. John Wiley and Sons: New York, 1986.
355. Pye, C. C., Ziegler, T., van Lenthe, E., and Louwen, J. N. An Implementation of the Conductor-Like Screening Model of Solvation within the Amsterdam Density Functional Package. Part II. COSMO for Real Solvents. *Can. J. Chem.* **2009**, *87*, 790.
356. Cabani, S., Gianni, P., Mollica, V., and Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous-Solution. *J. Solution Chem.* **1981**, *10*, 563-595.
357. Faheem, M., Saleheen, M., Lu, J., and Heyden, A. Ethylene glycol reforming on Pt(111): first-principles microkinetic modeling in vapor and aqueous phases. *Catal. Sci. Technol.* **2016**, *6*, 8242-8256.
358. Huber, G. W., Shabaker, J. W., Evans, S. T., and Dumesic, J. A. Aqueous-phase reforming of ethylene glycol over supported Pt and Pd bimetallic catalysts. *Appl. Catal., B* **2006**, *62*, 226-235.
359. Shabaker, J. W., Davda, R. R., Huber, G. W., Cortright, R. D., and Dumesic, J. A. Aqueous-phase reforming of methanol and ethylene glycol over alumina-supported platinum catalysts. *J. Catal.* **2003**, *215*, 344-352.

## Appendix A

### MECHANISM OF DEHYDRATION OF PHENOLS ON NOBLE METALS VIA FIRST-PRINCIPLES MICROKINETIC MODELING – SUPPORTING INFORMATION

#### A.1 Nomenclature of the *p*-cresol Derivatives

For convenience, we refer to *p*-cresol as “PC(OH)”, where PC stands for the methyl ring without OH. The *p*-cresol derivatives undergoing O-H scission are referred to as “PC(O)”. Similarly, PC refers to derivatives after C-OH scission. Following the  $\alpha$ -carbon hydrogenation of PC, the molecules are referred to as TOL. We number the ring carbons as shown in Figure A.1. Fully hydrogenated molecules are: MCH(OH), methyl cyclohexanol; MCH(O), methyl cyclohexanone; and MCH, methyl cyclohexane.

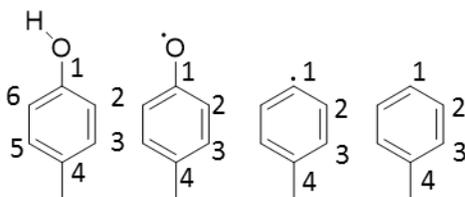


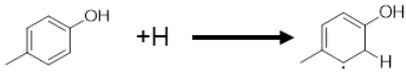
Figure A.1. Numbering of the ring carbon atoms. Structures from left to right: PC(OH), PC(O), PC, and TOL. An intermediate X hydrogenated at ring positions 2, 4, and 6 is denoted as H246X. For example, H23456PC(OH) refers to *p*-cresol with hydrogenated ring carbon at positions 2, 3, 4, 5, and 6.

#### A.2 Generation of the Reaction Network

We generate the network using the RING[94, 149] software. The classifications of the elementary reactions are listed in Table A.1. The number of

associated reactions and intermediates are listed in Table A.2. C-C cracking is not considered, as light hydrocarbons are not observed experimentally.[58, 62] The C-H and O-H reactions are included, as a previous study on Pt(111) using BEP relations demonstrated low barriers,[43] and saturated ketone and alcohol were observed experimentally[58, 62]. C-OH bond scission is a main elementary reaction type for deoxygenation of *p*-cresol.

Table A.1. Reaction rules and examples of input into RING.

Reaction rules	Example of reaction
Hydrogenation of ring carbon, C-H <sup>a</sup>	
C-OH bond scission	
O-H bond scission	

<sup>a</sup>C-H scission for the carbon with (C)<sub>2</sub>(H)(Pt) connectivity is not considered as a previous benzene dehydrogenation study shows it disfavored.[344] Additionally, the high hydrogen to *p*-cresol ratio used in the experiment results in high hydrogen coverage, disfavoring surface intermediates with H-deficient ring carbons.

Table A.2. Number of reactions and intermediates in the HDO reaction network of *p*-cresol on Pt(111). The table on the left lists the intermediates in each functional form similar to Figure 2.1.

Functional form of $\alpha$ -carbon	Surface species	Stable gas species	Reaction	Count
Alcohol (OH)	64	17	Adsorption	36
Ketone (O)	20	5	C-H scission/formation	400
H-deficient/ Dehydroxylated	20	0	C-OH scission/formation	32
Hydrocarbon/ Toluene	40	10	O-H scission/formation	32
Total	144	32	Total	500

### A.3 Parameterization of the Microkinetic Model

Figure A.2 shows the minimum energy path as the ring is being hydrogenated. At each degree of hydrogenation, there are multiple positions where H can be added. The energy level of the most stable species is shown in red and subsequent hydrogenation starts from this species. The reaction energies and barriers studied by DFT are summarized in Table A.3 and Table A.4.

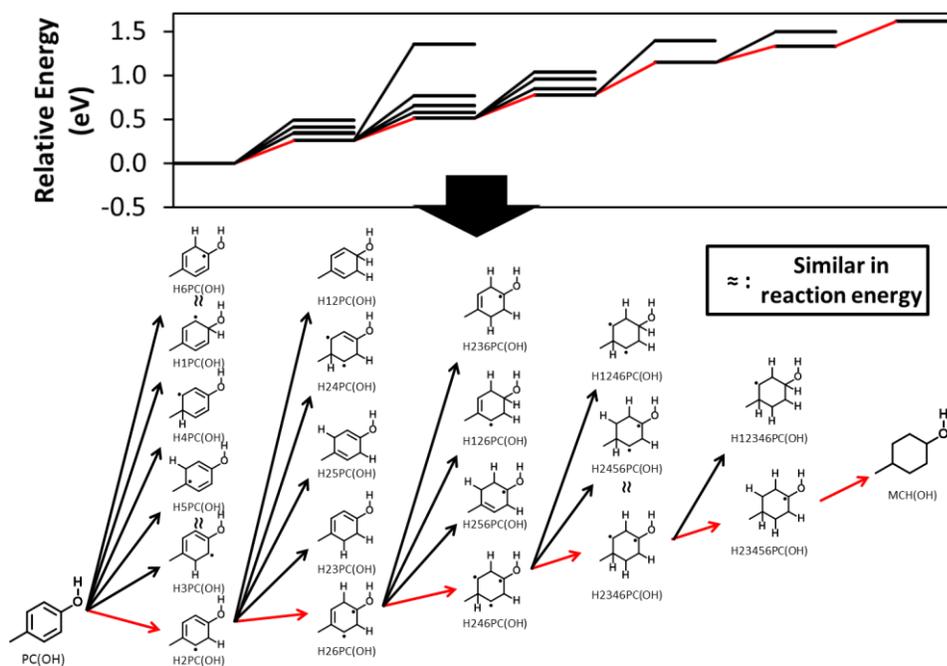


Figure A.2. Reaction energy profile (top) of hydrogenation of *p*-cresol (PC(OH)) following the minimum energy path (red arrows) and corresponding species (bottom) as hydrogens are being added.

Table A.3. C-H and O-H scission reactions used for development of BEP/TSS relations.

Type	Reactant	Product	Reaction energy (eV)	Barrier (eV)
C-H	H2PC(OH)(S)	PC(OH)(S)	-0.26	0.85
	H26PC(OH)(S)	H2PC(OH)(S)	-0.25	0.78
	H246PC(OH)(S)	H26PC(OH)(S)	-0.27	0.63
	H2346PC(OH)(S)	H246PC(OH)(S)	-0.37	0.57
	H23456PC(OH)(S)	H2346PC(OH)(S)	-0.19	0.76
	MCHOH(S)	H23456PC(OH)(S)	-0.28	0.55
	H1PC(OH)(S)	PC(OH)(S)	-0.49	0.78
	H3PC(OH)(S)	PC(OH)(S)	-0.34	0.73
	H4PC(OH)(S)	PC(OH)(S)	-0.41	0.78
	H5PC(OH)(S)	PC(OH)(S)	-0.35	0.84
	H6PC(OH)(S)	PC(OH)(S)	-0.49	0.63
	H12PC(OH)(S)	H2PC(OH)(S)	-1.10	0.14
	H23PC(OH)(S)	H2PC(OH)(S)	-0.32	0.73
	H24PC(OH)(S)	H2PC(OH)(S)	-0.51	0.64
	H25PC(OH)(S)	H2PC(OH)(S)	-0.40	0.84
	H23456PC(S)	H2345PC(S)	-0.24	0.63
	H2345TOL(S)	H2345PC(S)	-0.05	1.15
	H23456TOL(S)	H23456PC(S)	0.04	1.12
	H345TOL(S)	H34TOL(S)	-0.10	0.88
	H2345TOL(S)	H234TOL(S)	-0.36	0.66
	H2345TOL(S)	H245TOL(S)	-0.35	0.57
	H2345TOL(S)	H345TOL(S)	-0.48	0.45
	H2345TOL(S)	H235TOL(S)	-0.40	0.55
	H12345TOL(S)	H2345TOL(S)	0.07	0.93
	H23456TOL(S)	H2345TOL(S)	-0.15	0.52
	H23456TOL(S)	H2356TOL(S)	-0.06	0.86
	H23456TOL(S)	H2346TOL(S)	-0.16	0.81
	MCH(S)	H23456TOL(S)	-0.12	0.84
	MCH(S)	H12345TOL(S)	-0.32	0.64
	O-H	PC(OH)(S)	PC(O)(S) <sup>a</sup>	0.18
PC(OH)(S)		PC(O)(S) <sup>b</sup>	0.26	0.46
H2PC(OH)(S)		H2PC(O)(S) <sup>a</sup>	0.60	0.86
H2PC(OH)(S)		H2PC(O)(S) <sup>b</sup>	0.48	0.85
H26PC(OH)(S)		H26PC(O)(S) <sup>a</sup>	0.51	0.79
H246PC(OH)(S)		H246PC(O)(S) <sup>a</sup>	0.25	0.76
H2346PC(OH)(S)		H2346PC(O)(S) <sup>a</sup>	0.40	0.61
H23456PC(OH)(S)		H23456PC(O)(S) <sup>a</sup>	0.35	0.53

<sup>a</sup> C=O lifted off the surface. <sup>b</sup> C=O chemisorbed to the surface.

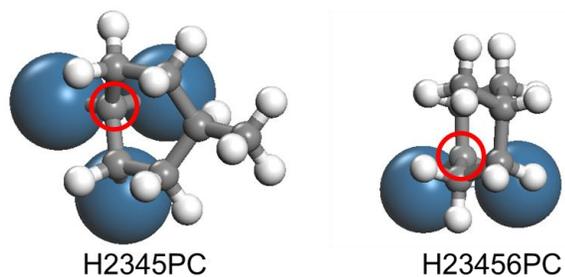


Figure A.3. Structures of hydrogenated ring-carbons that lead to H-deficient derivatives with  $sp^3$  hybridized  $\alpha$ -carbon.

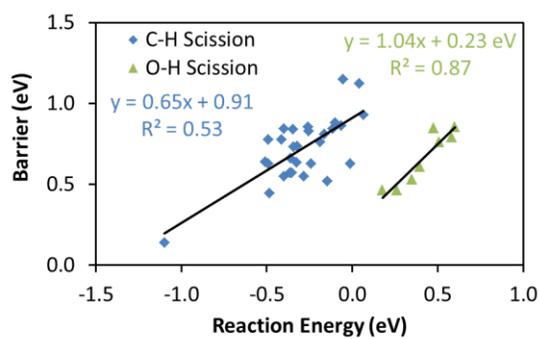


Figure A.4. BEP relations for C-H and O-H scission reactions. The training set is presented in Table A.3.

Table A.4. C-OH scission reactions used for development of BEP/TSS relations.

IS	TS	FS	Reactant	Product	Reaction energy (eV)	Barrier (eV)
1	3-centered	sp <sup>2</sup> α-C	PC(OH)(S)	PC(S)	1.81	2.45
1	3-centered	sp <sup>2</sup> α-C	H25PC(OH)(S)	H25PC(S)	1.88	2.47
1	3-centered	sp <sup>2</sup> α-C	H34PC(OH)(S)-1 <sup>a</sup>	H34PC(S)-1	1.20	1.68
1	3-centered	sp <sup>2</sup> α-C	H34PC(OH)(S)-2 <sup>a</sup>	H34PC(S)-2	1.06	1.72
2	3-centered	sp <sup>2</sup> α-C	H2PC(OH)(S)-1 <sup>a</sup>	H2PC(S)	1.66	2.15
2	3-centered	sp <sup>2</sup> α-C	H345PC(OH)(S)	H345PC(S)	0.78	1.46
3	4-centered	sp <sup>2</sup> α-C	H2PC(OH)(S)-2 <sup>a</sup>	H2PC(S)	1.83 <sup>b</sup>	1.75
3	4-centered	sp <sup>2</sup> α-C	H23PC(OH)(S)	H23PC(S)	1.62 <sup>b</sup>	1.47
3	4-centered	sp <sup>2</sup> α-C	H26PC(OH)(S)	H26PC(S)	1.71 <sup>b</sup>	1.71
3	4-centered	sp <sup>2</sup> α-C	H246PC(OH)(S)	H246PC(S)	1.53 <sup>b</sup>	1.55
3	4-centered	sp <sup>2</sup> α-C	H2346PC(OH)(S)	H2346PC(S)	1.34 <sup>b</sup>	1.26
3	4-centered	sp <sup>2</sup> α-C	H23456PC(OH)(S)	H23456PC(S)	1.11 <sup>b</sup>	0.96
3	4-centered	sp <sup>3</sup> α-C	H2345PC(OH)(S)	H2345PC(S)	0.91	1.10
3	4-centered	sp <sup>3</sup> α-C	H3456PC(OH)(S)	H2345PC(S)	0.87	1.16
3	4-centered	sp <sup>3</sup> α-C	H23456PC(OH)(S)	H23456PC(S)	0.79	0.96

<sup>a</sup>Same intermediates but adsorbed in different conformation. <sup>b</sup>These reaction energies are higher than the corresponding barriers at infinite separation but lower at co-adsorbed state, implying an attractive interaction between cresol derivatives and OH. For C-OH scission reaction energy in Figure 2.4 is computed at co-adsorbed state.

### A.3.1 Vibrational Frequency Calculations and Statistical Mechanical Approximation

We compute vibrational frequencies of 26 species along the minimum energy hydrogenation path, along with those of H<sub>2</sub>O, OH, and H. The frequencies of 26 additional species are calculated for MKM refinement. Vibrational frequencies and binding energies are presented in Table A.5.

Table A.5. Binding energies and vibrational frequencies of species computed via DFT.

Species	B.E. (eV)	$\nu$ (cm <sup>-1</sup> )
PC(S)	3.44	3079, 3073, 3059, 3057, 3053, 3023, 2937, 1436, 1429, 1405, 1359, 1346, 1300, 1273, 1240, 1177, 1136, 1064, 1024, 1004, 931, 926, 902, 882, 871, 826, 796, 719, 626, 568, 542, 522, 480, 417, 372, 328, 298, 226, 210, 185, 154, 111
H2PC(S)	4.68	3077, 3043, 3037, 3014, 3010, 3002, 2935, 2920, 1421, 1420, 1415, 1373, 1350, 1293, 1273, 1215, 1170, 1136, 1131, 1082, 1042, 1012, 996, 948, 935, 926, 899, 859, 834, 778, 692, 598, 576, 517, 493, 485, 416, 348, 335, 328, 241, 203, 192, 140, 119
H26PC(S)	5.90	3067, 3050, 3020, 2987, 2985, 2951, 2927, 2912, 2894, 1428, 1420, 1413, 1405, 1383, 1348, 1299, 1266, 1206, 1198, 1168, 1130, 1109, 1052, 1038, 1000, 994, 958, 945, 919, 899, 870, 831, 778, 680, 593, 579, 506, 483, 405, 375, 343, 321, 313, 251, 198, 152, 130, 94
H234PC(S)	4.77	3056, 3044, 3036, 3009, 2979, 2963, 2958, 2957, 2955, 2821, 1452, 1447, 1428, 1404, 1358, 1314, 1310, 1273, 1251, 1230, 1204, 1166, 1123, 1110, 1085, 1077, 1016, 999, 961, 955, 930, 897, 884, 868, 831, 752, 696, 608, 594, 527, 496, 433, 404, 303, 265, 223, 206, 191, 114, 74, 59
H235PC(S)	5.05	3064, 3031, 2988, 2986, 2971, 2970, 2950, 2915, 2908, 2875, 1434, 1423, 1416, 1398, 1385, 1341, 1302, 1278, 1225, 1204, 1182, 1164, 1146, 1119, 1096, 1074, 1029, 1007, 970, 967, 930, 901, 878, 850, 823, 759, 675, 583, 506, 476, 445, 429, 343, 318, 276, 270, 237, 219, 157, 99, 63
H236PC(S)	4.85	3029, 3026, 2999, 2987, 2969, 2947, 2933, 2908, 2899, 2868, 1429, 1417, 1408, 1400, 1393, 1348, 1302, 1273, 1240, 1202, 1184, 1168, 1160, 1095, 1078, 1065, 1026, 1000, 965, 946, 932, 912, 892, 853, 798, 756, 676, 576, 558, 486, 448, 402, 362, 347, 304, 296, 248, 219, 147, 114, 96
H245PC(S)	5.16	3052, 3030, 3026, 3010, 3003, 2985, 2952, 2919, 2911, 2785, 1447, 1442, 1422, 1402, 1355, 1310, 1277, 1269, 1248, 1239, 1196, 1151, 1146, 1121, 1085, 1042, 1031, 1022, 990, 987, 927, 903, 873, 870, 811, 768, 684, 591, 579, 502, 483, 393, 379, 320, 269, 245, 213, 159, 133, 85, 58
H246PC(S)	7.72	3054, 3044, 3001, 2997, 2980, 2968, 2966, 2915, 2892, 2814, 1449, 1447, 1425, 1408, 1351, 1293, 1283, 1266, 1234, 1188, 1182, 1170, 1105, 1095, 1079, 1025, 1015, 998, 991, 982, 924, 915, 882, 851, 782, 762, 726, 612, 576, 574, 558, 418, 403, 361, 317, 260, 209, 194, 164, 117, 108
H2345PC(S)	2.16	3041, 3030, 3014, 2993, 2966, 2959, 2954, 2948, 2935, 2907, 2791, 1449, 1447, 1436, 1407, 1394, 1362, 1324, 1310, 1274, 1267, 1239, 1232, 1211, 1183, 1150, 1123, 1096, 1092, 1059, 1015, 987, 959, 909, 901, 869, 845, 831, 763, 688, 525, 512, 481, 377, 357, 327, 248, 232, 216, 153, 128, 95, 69, 39
H2346PC(S)	5.96	3052, 3043, 3012, 3010, 2982, 2966, 2957, 2938, 2905, 2864, 2728, 1450, 1447, 1431, 1407, 1391, 1358, 1314, 1298, 1270, 1247, 1212, 1199, 1184, 1164, 1126, 1101, 1091, 1028, 1018, 1006, 985, 944, 925, 903, 888, 829, 792, 742, 685, 602, 574, 537, 419, 403, 379, 302, 264, 217, 185, 158, 113, 78, 65
H2356PC(S)	6.07	3045, 3030, 3009, 3006, 2986, 2971, 2950, 2932, 2928, 2921, 2916, 1438, 1430, 1421, 1412, 1404, 1390, 1345, 1316, 1288, 1236, 1225, 1208, 1187, 1181, 1154, 1124, 1099, 1060, 1048, 998, 973, 960, 935, 884, 865, 839, 794, 754, 692, 530, 485, 440, 431, 343, 305, 286, 269, 261, 233, 181, 92, 60, 51
H23456PC(S)	4.75	3046, 3038, 3003, 3001, 2969, 2962, 2960, 2949, 2934, 2928, 2715, 2708, 1456, 1453, 1431, 1422, 1394, 1372, 1366, 1325, 1319, 1312, 1281, 1260, 1249, 1205, 1184, 1152, 1138, 1093, 1081, 1046, 1036, 975, 969, 936, 931, 916, 864, 850, 757, 753, 644, 567, 477, 439, 385, 379, 329, 323, 218, 171, 156, 146, 105, 54, 45
PC(O)(S)	3.26	3068, 3062, 3049, 3043, 3016, 3001, 2929, 1655, 1425, 1420, 1402, 1353, 1332, 1288, 1267, 1242, 1152, 1138, 1069, 1034, 1011, 987, 956, 935, 897, 882, 839, 733, 724, 681, 582, 574, 519, 467, 427, 408, 404, 324, 310, 239, 191, 138, 117, 104, 94
H2PC(O)(S)	2.78	3070, 3040, 3018, 3013, 3007, 2989, 2920, 2887, 1513, 1425, 1413, 1403, 1382, 1350, 1313, 1277, 1273, 1229, 1156, 1128, 1112, 1052, 1012, 1004, 975, 935, 914, 895, 861, 784, 704, 695, 591, 560, 546, 445, 429, 422, 376, 325, 312, 255, 253, 208, 124, 119, 91, 55

Table A.5 continued.

H26PC(O)(S)	3.16	3052, 3046, 3020, 2996, 2984, 2932, 2918, 2896, 2891, 1639, 1433, 1415, 1406, 1375, 1361, 1348, 1328, 1295, 1274, 1227, 1175, 1161, 1134, 1093, 1045, 1022, 1002, 975, 949, 923, 893, 832, 747, 718, 696, 575, 544, 467, 450, 425, 371, 333, 308, 262, 233, 219, 192, 123, 92, 76, 69
H246PC(O)(S)	4.65	3057, 3050, 2995, 2987, 2977, 2970, 2968, 2905, 2899, 2759, 1652, 1448, 1444, 1375, 1362, 1346, 1311, 1299, 1284, 1269, 1251, 1199, 1167, 1155, 1109, 1071, 1060, 1038, 1004, 1002, 971, 964, 927, 866, 811, 776, 753, 699, 588, 546, 511, 509, 437, 405, 322, 288, 250, 241, 207, 174, 111, 109, 63, 62
H2346PC(O)(S)	2.83	3049, 3043, 3025, 3004, 2966, 2962, 2940, 2927, 2885, 2883, 2637, 1659, 1449, 1448, 1394, 1371, 1364, 1352, 1321, 1317, 1292, 1273, 1258, 1223, 1206, 1179, 1135, 1128, 1085, 1058, 1038, 1022, 979, 974, 921, 901, 866, 780, 769, 714, 680, 569, 518, 485, 447, 403, 379, 279, 246, 214, 203, 151, 114, 90, 67, 51, 29
MCH(O)(S)	1.35	3056, 3050, 3008, 3005, 2983, 2978, 2972, 2957, 2937, 2930, 2660, 2650, 1645, 1455, 1450, 1390, 1375, 1373, 1370, 1356, 1329, 1323, 1313, 1299, 1267, 1244, 1229, 1203, 1191, 1150, 1108, 1081, 1061, 1044, 992, 985, 932, 904, 882, 872, 770, 719, 682, 664, 504, 478, 432, 405, 374, 344, 221, 216, 172, 109, 96, 91, 58, 38, 25, 18
PC(OH)(S)	2.44	3410, 3095, 3082, 3053, 3043, 3036, 3007, 2933, 1452, 1425, 1421, 1401, 1364, 1353, 1326, 1283, 1262, 1201, 1151, 1136, 1079, 1038, 1001, 951, 939, 922, 909, 868, 842, 741, 703, 632, 580, 548, 507, 445, 427, 412, 392, 367, 319, 310, 238, 203, 188, 137, 124, 102
H2PC(OH)(S)	3.68	3300, 3085, 3048, 3028, 3024, 3022, 3000, 2928, 2889, 1423, 1417, 1409, 1401, 1349, 1335, 1305, 1299, 1267, 1235, 1172, 1139, 1102, 1085, 1051, 1010, 1009, 981, 937, 921, 918, 876, 805, 730, 690, 610, 579, 558, 493, 454, 437, 421, 415, 344, 319, 294, 247, 242, 203, 132, 121, 114
H26PC(OH)(S)	5.00	3479, 3062, 3050, 3033, 2988, 2987, 2917, 2916, 2877, 2871, 1431, 1418, 1411, 1406, 1389, 1350, 1345, 1312, 1300, 1271, 1235, 1182, 1145, 1129, 1092, 1063, 1038, 1011, 1008, 981, 960, 937, 896, 865, 842, 748, 681, 594, 561, 510, 473, 441, 423, 407, 346, 319, 314, 277, 259, 238, 196, 109, 99, 88
H234PC(OH)(S)	3.16	3308, 3050, 3043, 3007, 3004, 2992, 2965, 2959, 2932, 2931, 2666, 1452, 1445, 1414, 1385, 1357, 1335, 1330, 1314, 1301, 1277, 1264, 1248, 1183, 1157, 1137, 1117, 1087, 1054, 1048, 1011, 987, 981, 914, 900, 880, 850, 771, 716, 676, 608, 536, 505, 469, 460, 426, 412, 392, 268, 247, 236, 223, 198, 140, 99, 81, 70
H235PC(OH)(S)	3.74	3285, 3033, 3029, 3009, 2978, 2975, 2945, 2926, 2909, 2862, 2841, 1424, 1420, 1416, 1398, 1396, 1340, 1328, 1311, 1303, 1285, 1240, 1218, 1190, 1176, 1129, 1113, 1092, 1084, 1025, 1021, 990, 970, 952, 906, 870, 826, 757, 732, 663, 575, 541, 514, 463, 425, 407, 405, 327, 310, 283, 249, 245, 220, 191, 108, 101, 76
H236PC(OH)(S)	3.7	3495, 3022, 3020, 3001, 2972, 2965, 2960, 2913, 2894, 2890, 2873, 1427, 1414, 1406, 1400, 1395, 1350, 1341, 1314, 1301, 1289, 1243, 1227, 1182, 1167, 1127, 1110, 1082, 1078, 1022, 1019, 1004, 974, 957, 885, 877, 837, 761, 733, 657, 568, 543, 483, 471, 416, 408, 403, 333, 310, 297, 253, 236, 231, 189, 108, 104, 70
H245PC(OH)(S)	3.86	3525, 3054, 3032, 3015, 3006, 2978, 2975, 2955, 2896, 2888, 2744, 1448, 1443, 1419, 1399, 1357, 1338, 1312, 1286, 1278, 1262, 1239, 1215, 1173, 1143, 1134, 1105, 1062, 1057, 1041, 1017, 1007, 981, 977, 936, 890, 860, 790, 753, 700, 616, 583, 557, 456, 438, 431, 388, 368, 303, 265, 246, 230, 216, 133, 108, 91, 68
H246PC(OH)(S)	6.63	3508, 3056, 3052, 3011, 2992, 2971, 2941, 2901, 2890, 2857, 2783, 1447, 1443, 1402, 1386, 1351, 1335, 1312, 1289, 1280, 1276, 1266, 1189, 1172, 1127, 1114, 1093, 1073, 1037, 1036, 1001, 998, 973, 966, 941, 871, 837, 834, 775, 708, 606, 583, 511, 506, 458, 422, 390, 369, 332, 302, 251, 231, 206, 168, 113, 104, 88
H256PC(OH)(S)	3.72	3506, 3025, 3013, 2999, 2987, 2971, 2947, 2902, 2870, 2855, 2808, 1426, 1410, 1401, 1393, 1391, 1345, 1337, 1322, 1288, 1279, 1245, 1222, 1183, 1167, 1143, 1092, 1080, 1058, 1022, 1018, 998, 974, 955, 882, 880, 843, 762, 736, 653, 569, 543, 494, 458, 416, 410, 399, 333, 309, 299, 249, 240, 227, 210, 107, 101, 79

Table A.5. continued.

H346PC(OH)(S)	3.88	3598, 3050, 3035, 3018, 3005, 3001, 2959, 2957, 2898, 2888, 2750, 1446, 1443, 1420, 1397, 1356, 1331, 1307, 1284, 1275, 1267, 1245, 1214, 1182, 1145, 1131, 1098, 1081, 1046, 1042, 1010, 1005, 981, 974, 939, 892, 861, 789, 755, 698, 618, 587, 559, 453, 440, 408, 382, 364, 305, 259, 244, 234, 218, 134, 106, 92, 62
H356PC(OH)(S)	3.81	3486, 3019, 3016, 3014, 2972, 2965, 2959, 2921, 2902, 2844, 2781, 1422, 1416, 1414, 1396, 1388, 1342, 1332, 1307, 1295, 1265, 1250, 1207, 1185, 1170, 1123, 1107, 1097, 1051, 1025, 1020, 994, 972, 944, 913, 870, 829, 757, 737, 663, 579, 544, 470, 454, 422, 407, 396, 330, 307, 298, 252, 244, 228, 217, 110, 103, 78
H456PC(OH)(S)	3.23	3550, 3052, 3046, 3008, 2999, 2997, 2979, 2964, 2953, 2934, 2702, 1452, 1444, 1414, 1387, 1358, 1349, 1319, 1318, 1297, 1277, 1262, 1234, 1174, 1156, 1132, 1089, 1061, 1048, 1040, 1010, 983, 974, 916, 894, 880, 855, 775, 722, 677, 606, 553, 496, 473, 452, 436, 413, 387, 276, 252, 234, 214, 199, 151, 98, 85, 68
H2345PC(OH)(S)	2.16	3328, 3043, 3023, 3010, 3006, 2994, 2970, 2945, 2936, 2893, 2884, 2849, 1454, 1446, 1424, 1417, 1404, 1365, 1340, 1321, 1316, 1301, 1291, 1286, 1247, 1217, 1195, 1158, 1149, 1101, 1085, 1072, 1054, 1018, 992, 983, 940, 902, 879, 832, 744, 724, 653, 563, 534, 503, 450, 415, 390, 364, 313, 268, 243, 234, 201, 147, 102, 75, 55, 27
H2346PC(OH)(S)	4.69	3421, 3047, 3041, 2994, 2993, 2989, 2962, 2935, 2905, 2897, 2840, 2608, 1450, 1448, 1413, 1386, 1381, 1358, 1335, 1322, 1301, 1286, 1280, 1273, 1230, 1195, 1148, 1130, 1118, 1099, 1071, 1048, 1037, 1012, 981, 969, 929, 899, 875, 822, 762, 746, 660, 596, 500, 485, 461, 455, 414, 389, 354, 269, 247, 239, 225, 200, 143, 88, 81, 72
H2356PC(OH)(S)	4.56	3506, 3024, 3002, 2988, 2984, 2960, 2949, 2932, 2915, 2909, 2905, 2898, 1434, 1425, 1424, 1415, 1405, 1400, 1346, 1340, 1321, 1298, 1294, 1256, 1241, 1228, 1211, 1178, 1136, 1103, 1091, 1070, 1027, 1018, 999, 971, 893, 871, 866, 817, 734, 730, 646, 519, 489, 463, 452, 409, 404, 320, 282, 260, 252, 237, 228, 222, 159, 79, 71, 29
H2456PC(OH)(S)	4.68	3520, 3051, 3045, 2996, 2994, 2970, 2962, 2921, 2894, 2894, 2829, 2607, 1447, 1445, 1409, 1387, 1376, 1354, 1337, 1316, 1304, 1285, 1277, 1264, 1232, 1184, 1143, 1135, 1123, 1069, 1066, 1043, 1031, 1014, 979, 969, 929, 896, 873, 829, 762, 739, 662, 595, 500, 486, 460, 432, 415, 385, 352, 268, 251, 238, 220, 194, 139, 94, 73, 67
H3456PC(OH)(S)	1.93	3615, 3046, 3029, 3012, 2996, 2992, 2984, 2950, 2931, 2895, 2876, 2686, 1451, 1447, 1418, 1403, 1391, 1361, 1338, 1320, 1310, 1297, 1283, 1278, 1251, 1215, 1178, 1153, 1144, 1098, 1077, 1069, 1042, 1013, 991, 982, 935, 908, 872, 828, 741, 731, 655, 581, 528, 458, 426, 390, 375, 344, 318, 257, 238, 231, 193, 152, 96, 73, 57, 31
H23456PC(OH)(S)	2.92	3401, 3049, 3046, 3008, 3005, 2966, 2963, 2962, 2954, 2896, 2893, 2625, 2614, 1455, 1450, 1405, 1391, 1372, 1364, 1358, 1339, 1326, 1313, 1305, 1290, 1281, 1259, 1229, 1189, 1150, 1142, 1090, 1076, 1070, 1050, 1046, 987, 976, 921, 917, 910, 874, 770, 741, 726, 632, 512, 478, 461, 458, 393, 374, 372, 310, 278, 241, 213, 161, 149, 97, 75, 60, 48
MCH(OH)(S)	1.31	3635, 3048, 3042, 3015, 3012, 3008, 2991, 2964, 2949, 2944, 2943, 2598, 2581, 2124, 1456, 1451, 1427, 1410, 1373, 1371, 1354, 1351, 1330, 1318, 1309, 1295, 1286, 1253, 1236, 1198, 1180, 1150, 1139, 1130, 1099, 1070, 1057, 1046, 1037, 976, 972, 943, 930, 900, 866, 796, 757, 748, 636, 498, 462, 444, 389, 372, 340, 323, 259, 224, 211, 135, 102, 100, 78, 55, 39, 29
TOL(S)	2.33	3098, 3086, 3085, 3053, 3045, 3039, 3019, 2941, 1436, 1423, 1415, 1387, 1368, 1352, 1325, 1307, 1261, 1144, 1140, 1087, 1015, 1004, 966, 928, 924, 909, 891, 866, 837, 829, 721, 631, 566, 529, 503, 474, 394, 370, 338, 333, 231, 193, 161, 140, 118
H2TOL(S)	3.66	3073, 3039, 3030, 3027, 3010, 3008, 2994, 2923, 2864, 1426, 1417, 1407, 1405, 1349, 1327, 1320, 1306, 1273, 1210, 1155, 1138, 1089, 1051, 1028, 1000, 992, 972, 946, 925, 892, 872, 819, 795, 712, 613, 562, 546, 491, 444, 411, 346, 333, 315, 249, 202, 156, 132, 112
H26TOL(S)	4.91	3063, 3046, 3028, 3021, 2979, 2958, 2916, 2904, 2856, 2853, 1432, 1415, 1408, 1403, 1385, 1349, 1331, 1312, 1307, 1277, 1204, 1178, 1142, 1118, 1059, 1052, 1040, 995, 993, 976, 952, 937, 875, 865, 834, 800, 717, 592, 552, 519, 452, 430, 357, 328, 322, 299, 266, 200, 125, 122, 91

Table A.5. continued.

H34TOL(S)	2.81	3068, 3061, 3049, 3044, 3026, 3018, 3005, 2965, 2901, 2565, 1446, 1444, 1415, 1399, 1366, 1349, 1321, 1289, 1268, 1245, 1235, 1196, 1148, 1108, 1080, 1071, 1019, 1009, 990, 962, 919, 900, 886, 844, 826, 799, 734, 605, 558, 536, 481, 443, 376, 332, 260, 231, 205, 145, 131, 86, 71
H234TOL(S)	3.12	3058, 3046, 3039, 3023, 3015, 3005, 3000, 2960, 2913, 2885, 2624, 1450, 1448, 1419, 1405, 1381, 1364, 1350, 1330, 1308, 1296, 1263, 1259, 1199, 1155, 1139, 1097, 1090, 1045, 1023, 991, 980, 956, 906, 891, 879, 852, 823, 754, 682, 584, 527, 511, 453, 409, 361, 265, 248, 216, 189, 124, 97, 73, 68
H235TOL(S)	3.69	3024, 3017, 3009, 2987, 2972, 2970, 2934, 2905, 2900, 2875, 2855, 1423, 1417, 1417, 1396, 1391, 1340, 1323, 1315, 1293, 1282, 1233, 1207, 1181, 1172, 1115, 1105, 1094, 1055, 1018, 988, 966, 959, 946, 884, 859, 828, 814, 741, 683, 554, 531, 463, 454, 405, 342, 322, 280, 246, 242, 206, 130, 107, 80
H245TOL(S)	3.81	3051, 3033, 3011, 3005, 3001, 2987, 2955, 2953, 2891, 2865, 2781, 1447, 1444, 1416, 1397, 1355, 1323, 1307, 1288, 1277, 1270, 1243, 1214, 1173, 1146, 1114, 1075, 1061, 1037, 1026, 1008, 986, 973, 960, 915, 866, 851, 822, 771, 736, 597, 561, 534, 479, 431, 373, 328, 271, 250, 217, 141, 137, 93, 68
H246TOL(S)	6.53	3050, 3045, 3008, 3001, 2999, 2962, 2909, 2908, 2855, 2853, 2746, 1446, 1443, 1397, 1382, 1350, 1321, 1321, 1287, 1286, 1272, 1266, 1177, 1169, 1114, 1094, 1091, 1039, 1031, 1025, 1001, 994, 979, 953, 930, 845, 840, 828, 787, 766, 601, 586, 548, 493, 467, 363, 335, 306, 259, 221, 185, 140, 111, 99
H345TOL(S)	3.19	3051, 3031, 3004, 3001, 2994, 2988, 2955, 2950, 2890, 2886, 2596, 1448, 1445, 1413, 1399, 1364, 1348, 1318, 1305, 1281, 1266, 1232, 1224, 1220, 1140, 1129, 1110, 1089, 1062, 1011, 1009, 1000, 969, 937, 874, 870, 844, 812, 777, 702, 563, 535, 463, 460, 417, 389, 306, 240, 236, 199, 149, 132, 80, 63
H2345TOL(S)	1.84	3048, 3028, 3007, 3002, 2989, 2985, 2959, 2949, 2912, 2899, 2891, 2739, 1451, 1446, 1414, 1401, 1397, 1361, 1335, 1327, 1305, 1303, 1289, 1281, 1249, 1216, 1173, 1156, 1137, 1090, 1077, 1057, 1021, 992, 974, 961, 924, 903, 857, 823, 809, 732, 657, 549, 496, 473, 411, 369, 337, 268, 244, 189, 143, 118, 74, 60, 22
H2346TOL(S)	4.60	3044, 3038, 3001, 2999, 2989, 2979, 2959, 2910, 2909, 2877, 2790, 2581, 1450, 1447, 1406, 1383, 1374, 1355, 1333, 1322, 1295, 1287, 1281, 1265, 1224, 1179, 1137, 1126, 1097, 1091, 1044, 1037, 1027, 1007, 975, 956, 930, 891, 844, 837, 797, 756, 676, 595, 534, 474, 463, 424, 359, 274, 261, 232, 200, 153, 111, 81, 70
H2356TOL(S)	4.50	3021, 2990, 2989, 2982, 2980, 2964, 2946, 2921, 2917, 2912, 2901, 2898, 1430, 1425, 1421, 1409, 1402, 1395, 1345, 1324, 1311, 1305, 1292, 1235, 1227, 1218, 1202, 1170, 1115, 1098, 1092, 1078, 1026, 992, 976, 968, 866, 864, 857, 818, 767, 722, 698, 503, 480, 448, 432, 351, 280, 259, 251, 245, 227, 161, 82, 76, 24
H12345TOL(S)	3.01	3042, 3025, 3012, 2994, 2968, 2946, 2922, 2921, 2915, 2880, 2877, 2600, 2453, 1442, 1440, 1424, 1392, 1383, 1367, 1360, 1336, 1312, 1305, 1291, 1260, 1251, 1228, 1202, 1193, 1172, 1108, 1105, 1086, 1084, 1048, 1025, 1010, 962, 945, 906, 870, 834, 817, 809, 746, 618, 471, 462, 440, 408, 318, 310, 252, 221, 215, 141, 117, 82, 62, 39
H23456TOL(S)	2.88	3047, 3042, 3003, 2999, 2973, 2965, 2949, 2929, 2925, 2889, 2885, 2604, 2586, 1455, 1449, 1399, 1385, 1368, 1360, 1356, 1335, 1333, 1315, 1296, 1292, 1281, 1243, 1226, 1184, 1144, 1105, 1098, 1097, 1053, 1042, 1039, 972, 963, 936, 913, 889, 847, 824, 754, 731, 639, 547, 449, 445, 387, 380, 305, 294, 215, 164, 153, 102, 86, 65, 49
MCH(S)	1.12	3048, 3047, 3013, 3008, 3006, 2995, 2990, 2966, 2945, 2944, 2931, 2561, 2535, 2471, 1456, 1449, 1424, 1411, 1365, 1354, 1350, 1340, 1337, 1330, 1304, 1302, 1296, 1292, 1256, 1236, 1186, 1170, 1145, 1112, 1110, 1066, 1050, 1028, 1014, 962, 945, 936, 878, 850, 833, 789, 747, 745, 591, 468, 447, 390, 375, 304, 222, 208, 184, 113, 98, 80, 47, 40, 29
H2O(S)	0.46	3690, 3583, 1546, 525, 507, 160, 92, 92, 54
OH(S)	2.99	3638, 924, 510, 119, 96, 52
H(S)	0.54	1097, 627, 597

A statistical mechanical approximation method[111, 345, 346] is used to estimate the thermodynamics of surface species. The enthalpy of each surface species ( $\Delta H_{i,surf}$ ) is computed by subtracting the DFT binding energy from the enthalpy of formation of the corresponding gas-phase species as shown below:

$$\Delta H_{i,surf} = \Delta H_{i,gas,G4} - Q(T=0)_{i,VASP} - Q_i(T) \quad (\text{A.1})$$

Here,  $Q_i(T=0)_{i,VASP}$  is the 0 K binding energy (without zero-point energy correction) obtained from periodic DFT calculations using the VASP software, and  $Q_i(T)$  is a temperature dependent correction approximated using statistical mechanics to account for the change in degrees of freedom. 0 K binding energies are listed in Table A.6. For the chemisorbed *p*-cresol derivatives,  $Q_i(T)$  is given by the loss of 3 translational ( $3 \times 0.5RT$ ) and rotational ( $3 \times 0.5RT$ ) degrees of freedom, and a gain of 6 vibrational degrees of freedom ( $6 \times RT$ ):

$$Q_i(T) = -1.5RT - 1.5RT + 6RT = 3RT \quad (\text{A.2})$$

The entropy of each surface species is calculated from the following approximation reviewed in Saliccioli et al.[346]

$$\Delta S_{i,surf} = F_{loc}(\Delta S_{i,gas,G4} - S_{i,trans,3D}) \quad (3)$$

$$S_{i,trans,3D} = R \left[ \ln \left( \frac{(2\pi m k_B T)^{3/2}}{h^3} \right) + \ln \left( \frac{k_B T}{P} \right) + \frac{5}{2} \right] \quad (\text{A.3})$$

Using the above equation, the entropy of surface species, *i*, ( $\Delta S_{i,surf}$ ) is computed by subtracting the translational entropy ( $\Delta S_{i,trans,3D}$ ) from the gas-phase entropy ( $\Delta S_{i,gas,G4}$ ). The multiplying factor  $F_{loc}$  is fitted using known values (here is taken as 0.99).

Table A.6. Binding energies of surface species whose thermodynamics is approximated using the equations above.

Species	B.E. (eV)	Species	B.E. (eV)	Species	B.E. (eV)
H3PC(S)	5.04	H16PC(OH)(S)	2.89	H12345PC(OH)(S)	3.06
H4PC(S)	4.57	H23PC(OH)(S)	2.79	H12346PC(OH)(S)	3.06
H23PC(S)	4.21	H24PC(OH)(S)	4.99	H12356PC(OH)(S)	2.72
H24PC(S)	4.22	H25PC(OH)(S)	2.78	H12456PC(OH)(S)	3.00
H25PC(S)	3.80	H34PC(OH)(S)	2.96	H13456PC(OH)(S)	3.07
H34PC(S)	4.46	H35PC(OH)(S)	5.04	H1TOL(S)	3.64
H35PC(S)	6.52	H36PC(OH)(S)	2.83	H3TOL(S)	3.58
H345PC(S)	4.50	H45PC(OH)(S)	2.86	H4TOL(S)	3.61
H3PC(O)(S)	4.51	H46PC(OH)(S)	4.98	H12TOL(S)	2.84
H4PC(O)(S)	2.66	H56PC(OH)(S)	2.84	H13TOL(S)	4.93
H23PC(O)(S)	2.81	H123PC(OH)(S)	3.08	H14TOL(S)	2.65
H24PC(O)(S)	3.71	H124PC(OH)(S)	3.75	H23TOL(S)	2.75
H25PC(O)(S)	3.49	H125PC(OH)(S)	3.97	H24TOL(S)	4.93
H34PC(O)(S)	3.44	H126PC(OH)(S)	3.29	H25TOL(S)	2.76
H35PC(O)(S)	6.08	H135PC(OH)(S)	6.64	H35TOL(S)	4.86
H234PC(O)(S)	2.01	H136PC(OH)(S)	3.82	H123TOL(S)	3.05
H235PC(O)(S)	4.26	H145PC(OH)(S)	3.90	H124TOL(S)	3.70
H236PC(O)(S)	1.86	H146PC(OH)(S)	3.72	H125TOL(S)	3.87
H345PC(O)(S)	4.41	H156PC(OH)(S)	3.10	H126TOL(S)	3.23
H245PC(O)(S)	4.36	H134PC(OH)(S)	3.92	H134TOL(S)	3.64
H2345PC(O)(S)	2.76	H345PC(OH)(S)	3.31	H135TOL(S)	6.47
H2356PC(O)(S)	2.70	H1234PC(OH)(S)	1.92	H236TOL(S)	3.62
H1PC(OH)(S)	3.70	H1235PC(OH)(S)	4.54	H1234TOL(S)	1.80
H3PC(OH)(S)	3.67	H1236PC(OH)(S)	2.02	H1235TOL(S)	4.52
H4PC(OH)(S)	3.71	H1245PC(OH)(S)	4.63	H1236TOL(S)	1.78
H5PC(OH)(S)	3.61	H1246PC(OH)(S)	4.72	H1245TOL(S)	4.44
H6PC(OH)(S)	3.69	H1256PC(OH)(S)	1.93	H1246TOL(S)	4.63
H12PC(OH)(S)	2.92	H1345PC(OH)(S)	4.87	H1345TOL(S)	4.68
H13PC(OH)(S)	5.02	H1346PC(OH)(S)	4.85	H12346TOL(S)	2.91
H14PC(OH)(S)	2.75	H1356PC(OH)(S)	4.57	H12356TOL(S)	2.75
H15PC(OH)(S)	5.10	H1456PC(OH)(S)	1.85		

### A.3.2 Lateral Interactions

We develop a lateral interaction model for self and binary interactions as shown below:

$$Q_i(\theta_1, \dots, \theta_n) = Q_{i,0} + 2w_{ii}\theta_i + \sum_{j=1, i \neq j}^n w_{ij} \theta_j \quad (\text{A.4})$$

where  $Q_i$  is the binding energy of species  $i$ ,  $Q_{i,0}$  is the binding energy of species  $i$  in the zero coverage limit,  $w_{ii}$  is the lateral interaction of species  $i$  with itself,  $\theta_i$  is the coverage of species  $i$ ,  $w_{ij}$  is the lateral interaction of species  $i$  with species  $j$ . We included H-H lateral interactions adopted from previous work on Pt(111) (-3 kcal/mol/monolayer).[347] Here, we introduce a simple lateral interaction model shown in Figure A.5. We calculated lateral interactions for various species with hydrogen by adding hydrogen to the unit cell and calculate the slopes. These species are along the minimum energy hydrogenation path. (PC(OH), H2PC(OH), H26PC(OH), H246PC(OH), H2346PC(OH), H23456PC(OH), and MCHOH) We find that the lateral interactions correlate linearly with number of Pt atoms involved in adsorption. For microkinetic modeling, we include lateral interactions for *p*-cresol derivatives with 0, 1, or 2 hydrogenations (using the correlation in Figure A.5). Our results are overall consistent with those of Sabbe et al. who found that the hydrogenation energy profile for benzene is downhill when the benzene is co-adsorbed with hydrogen as opposed to uphill at the zero-coverage limit.[126]

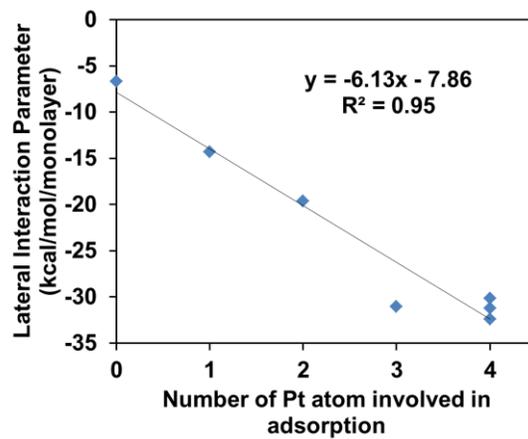


Figure A.5. Lateral interaction for the effect of hydrogen coverage on the binding energy of the *p*-cresol derivatives vs. the number of Pt atoms involved in adsorption.

#### A.4 Adsorption and Transition State Structures

In this section, we present the adsorption and transition state structures computed via DFT. The most stable adsorption structure of *p*-cresol is bridge site, consistent with benzene.[119] For the adsorption structures along the minimum energy hydrogenation path, we simply add hydrogen to the desired ring carbon position and relax the structure. We observe that the molecule remains at the same bridge site. Further addition of hydrogen at C6 creates a structure that is similar to both bridge and hollow site adsorption. For the next 4 added hydrogen atoms, the molecule relaxes to the hollow site. For toluene adsorption, we used the same adsorption configurations as *p*-cresol. For the dehydroxylated species, the adsorption configuration does not change except for the dehydroxylated *p*-cresol and H23456PC. The  $\alpha$ -carbon of dehydroxylated *p*-cresol moves away from the original metal atom and binds to the adjacent Pt atom with  $sp^2$  hybridization. The  $\alpha$ -carbon of dehydroxylated H23456PC relaxes into the bridge site with two Pt atoms via  $\sigma$  bonds.

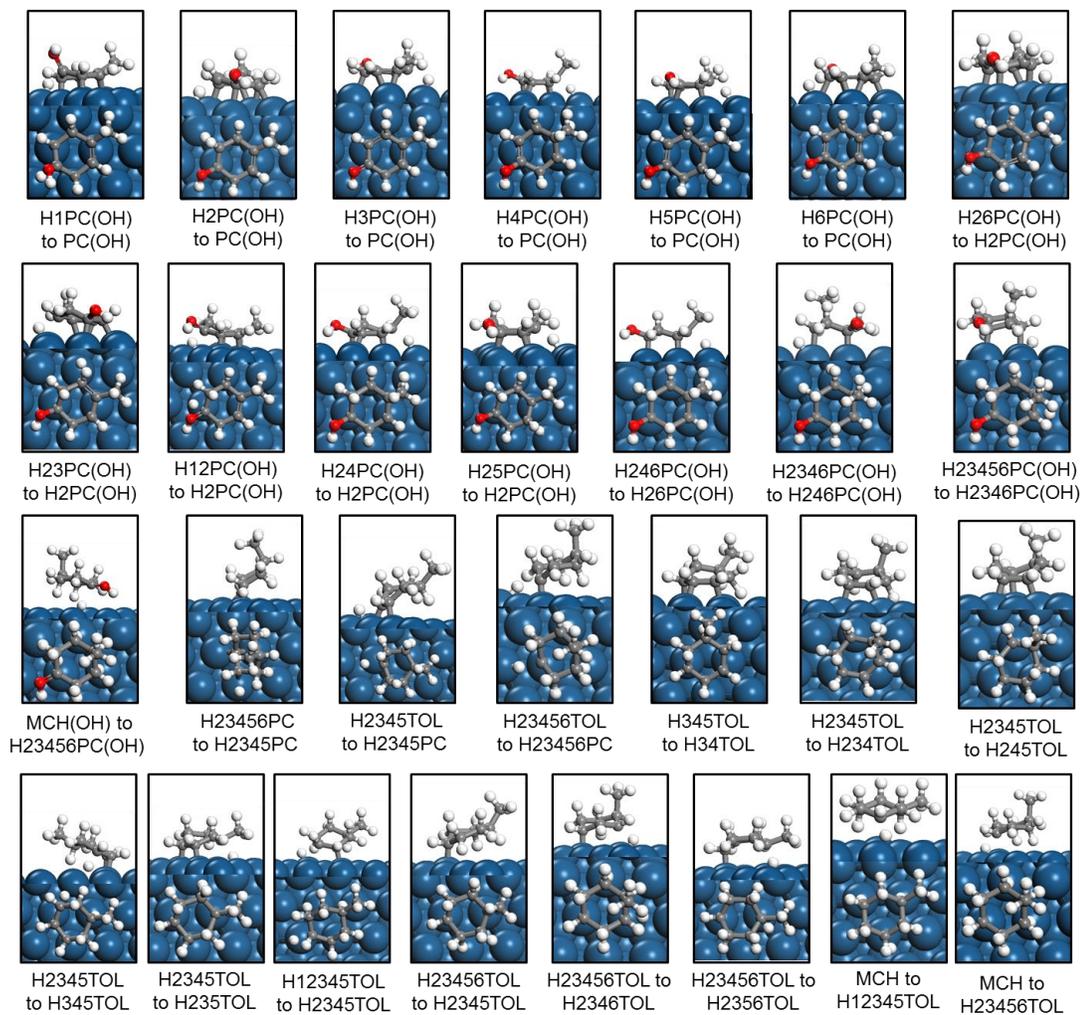


Figure A.6. Transition state structures of C-H bond scission reactions.

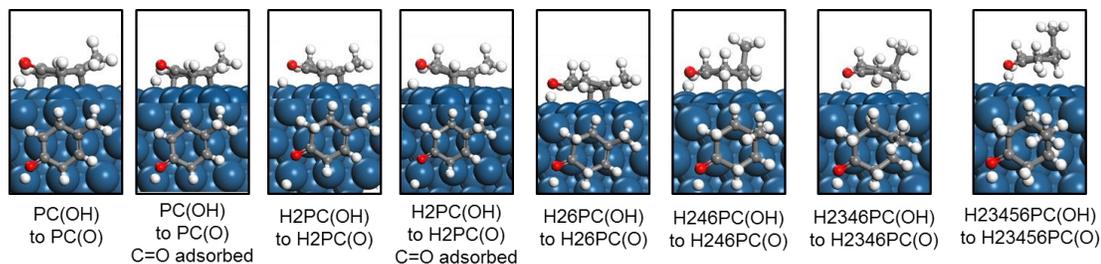


Figure A.7. Transition state structures of O-H bond scission reactions.

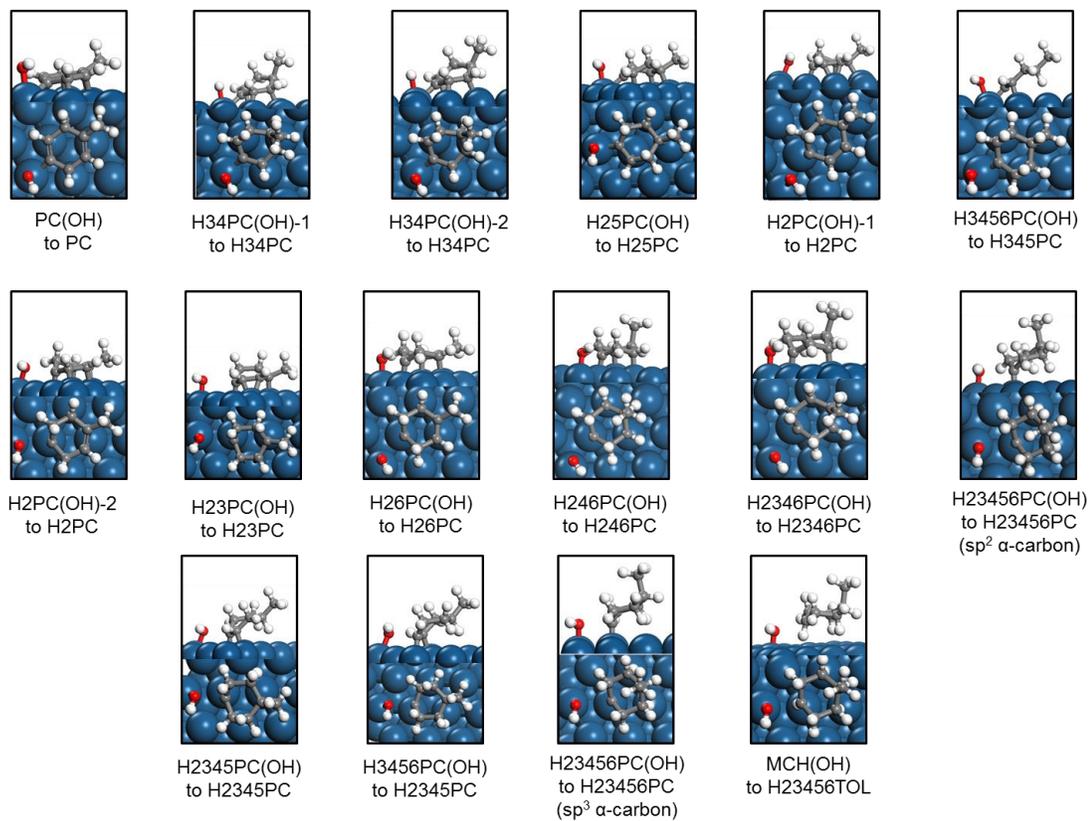
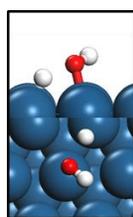


Figure A.8. Transition state structures of C-OH bond scission reactions.



Water  
Formation

Figure A.9. H-OH scission transition state structure.

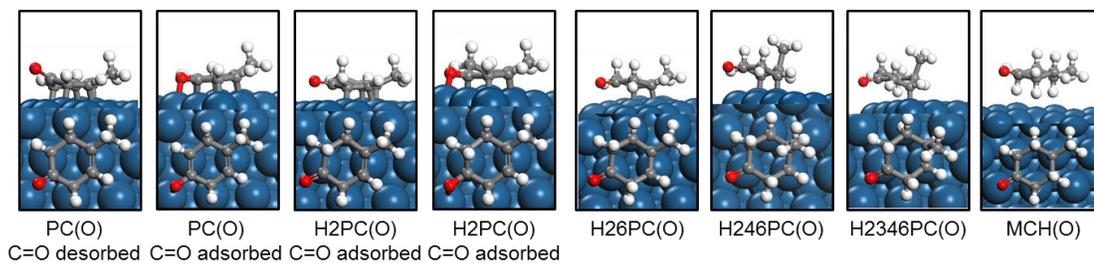


Figure A.11. Adsorption structures of the *p*-cresol derivatives with H deficient oxygen.

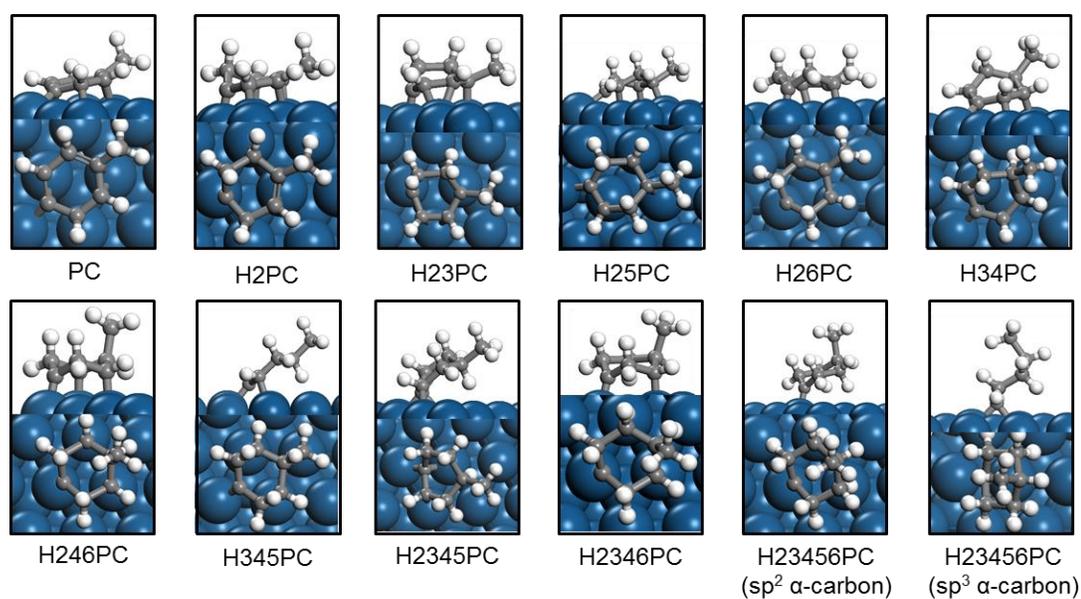


Figure A.12. Adsorption structures of the *p*-cresol derivatives with radical  $\alpha$ -carbon.

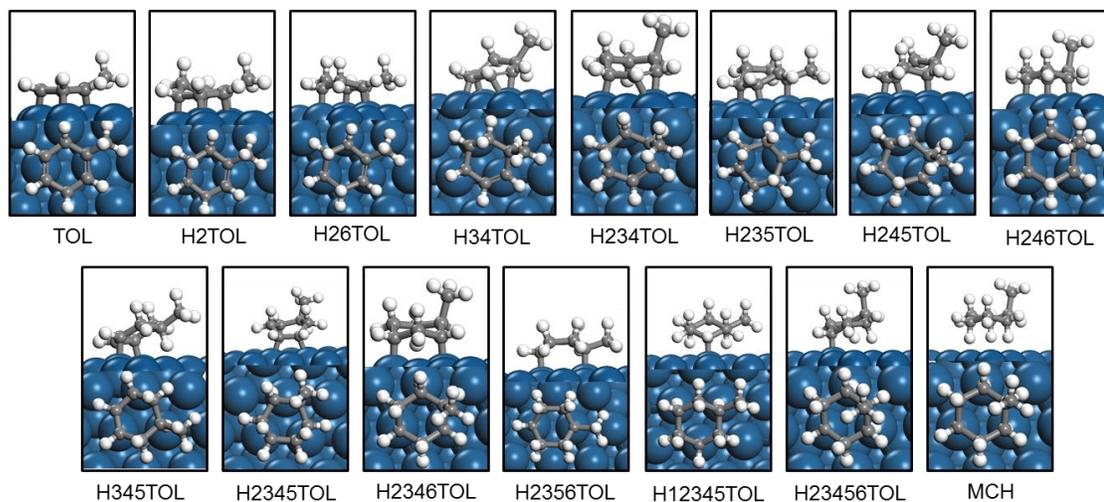


Figure A.13. Adsorption structures for the toluene derivatives.

## A.5 MKM Parameters Adjustments

Table A.7. List of parameter adjustments for the MKM model.

Adjusted parameter	Amount of adjustments
$\Delta H_f$ of methyl cyclohexanone	-0.14 eV
$\Delta H_f$ of methyl cyclohexanol	-0.12 eV
Pre-exp. of methyl cyclohexanol desorption	40 fold (equivalent to 0.18 eV decrease in barrier)
Pre-exp. of methyl cyclohexanone desorption	5 fold (equivalent to 0.08 eV decrease in barrier)
Pre-exp. of C-OH scission reactions	0.5 fold (equivalent to 0.03 eV increase in barrier)
Pre-exp. of C-H scission reactions for toluene derivatives with 1 or 2 hydrogenated ring carbons	10 fold (equivalent to 0.11 eV decrease in barrier)

## A.6 Reaction Energies

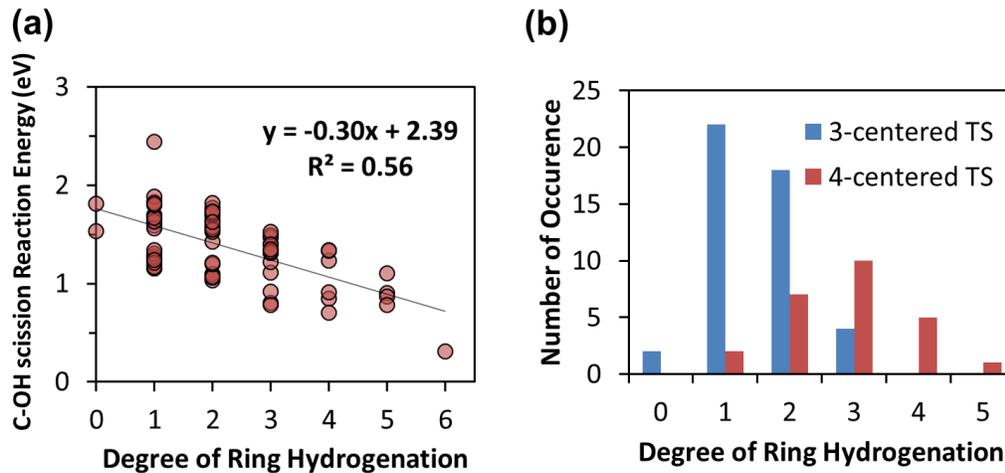


Figure A.14. (a) C-OH scission reaction energy vs. degree of ring hydrogenation for the alcohols in the entire reaction network. (b) Number of C-OH scission reactions encountered vs. degree of ring hydrogenation. The combination of these two factors leads to efficient C-OH scission upon sufficient ring hydrogenation.

### A.7 MKM Results

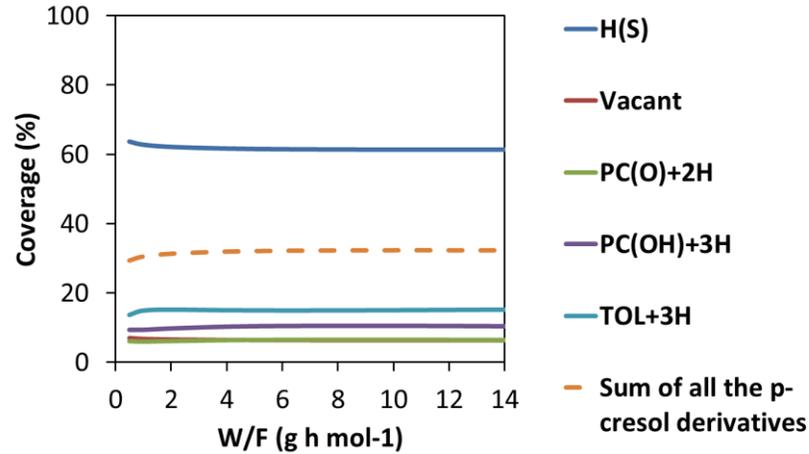


Figure A.15. Surface coverages of *p*-cresol simulation for Nie et al. experimental data[62] shown in Figure 2.5 in the main text. Hydrogen is the most dominant species on the surface, followed by toluene derivatives with 3 hydrogenated ring carbons (TOL+3H), *p*-cresol with 3 hydrogenated ring carbons (PC(OH)+3H), and *p*-cresol with H deficient oxygen and 2 hydrogenated ring carbons (PC(O)+2H).

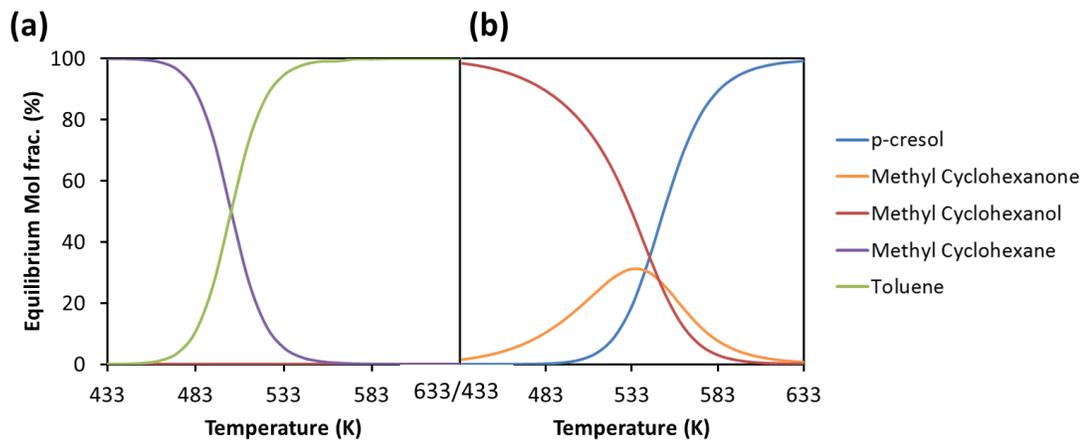


Figure A.16. Molar equilibrium composition with (a) and without (b) hydrocarbon products accounted for. Panel (b) is reminiscent of near the entrance region of the reactor. Conditions are  $1.25 \times 10^{-2}$  atm *p*-cresol, 1.0 atm H<sub>2</sub>.

## Appendix B

### GROUP ADDITIVITY FOR THERMOCHEMICAL PROPERTY ESTIMATION OF LIGNIN MONOMERS ON PT(111) – SUPPORTING INFORMATION

#### B.1 Adsorption Conformation of Lignin Monomer Ring

Using the binding type approach, we identified a total of 107 ring conformations (201 including *fcc/hcp* variation). This section shows chemical drawings of conformations including hydrogenated derivatives of: (1) benzene ring (Figure B.1), (2) benzene ring with one H-deficient ring carbon (caused by C<sub>ring</sub>-O scission) (Figure B.2), (3) benzene ring with one surface interacting oxygen substituent (whether  $\sigma$  binding or =O<sub>wk</sub> binding) (Figure B.3), (4) benzene ring with one H-deficient ring carbon and one surface interacting oxygen substituent (Figure B.4), and (5) benzene ring with two surface interacting oxygen substituents (Figure B.5).

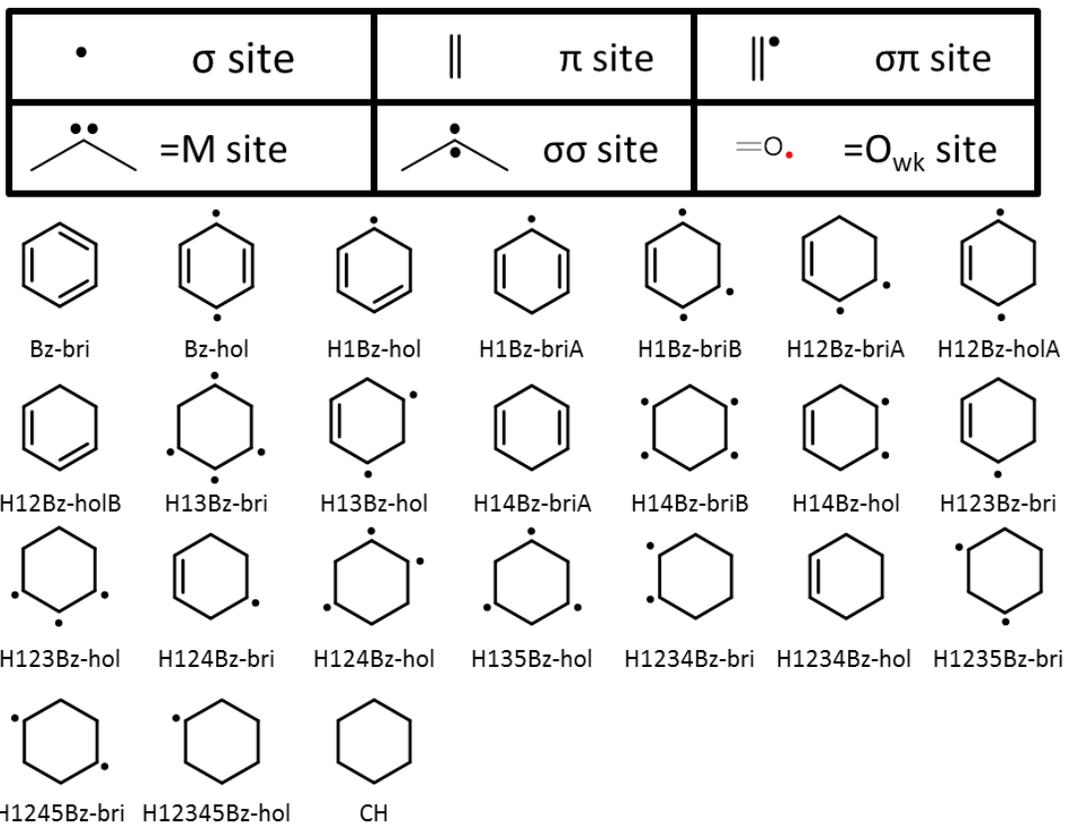


Figure B.1. Adsorption conformations of benzene ring and its hydrogenated derivatives. Legend at the top of the figure shows the binding type given the structure. The name convention follows the previous benzene adsorption study by Morin et al.[348]

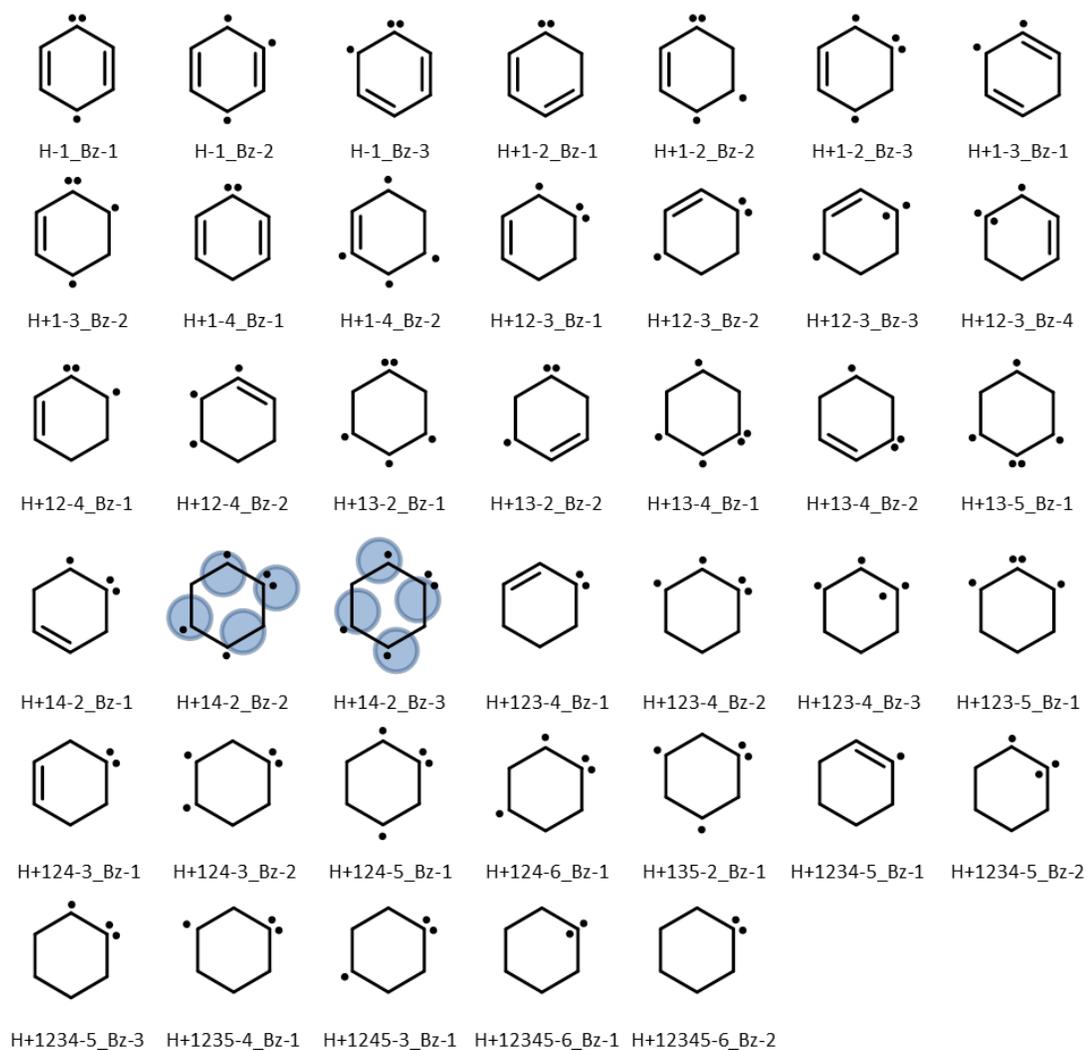


Figure B.2. Adsorption conformations of benzene ring with one H-deficient ring carbon and its hydrogenated derivatives. Blue circles indicates Pt atoms.

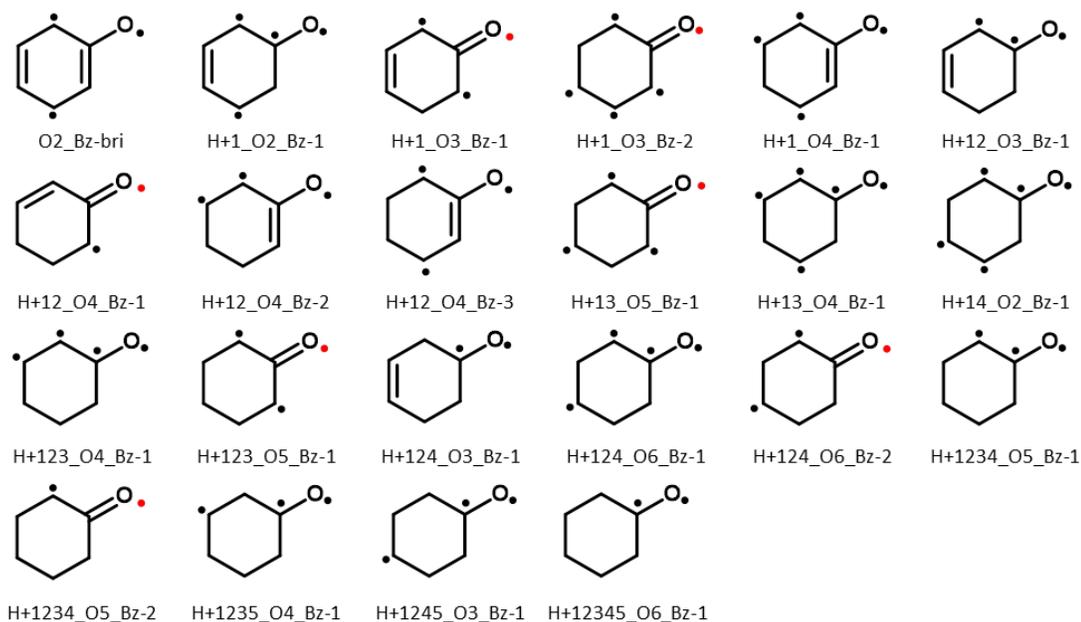


Figure B.3. Adsorption conformations of benzene ring with one surface interacting oxygen substituent and its hydrogenated derivatives.

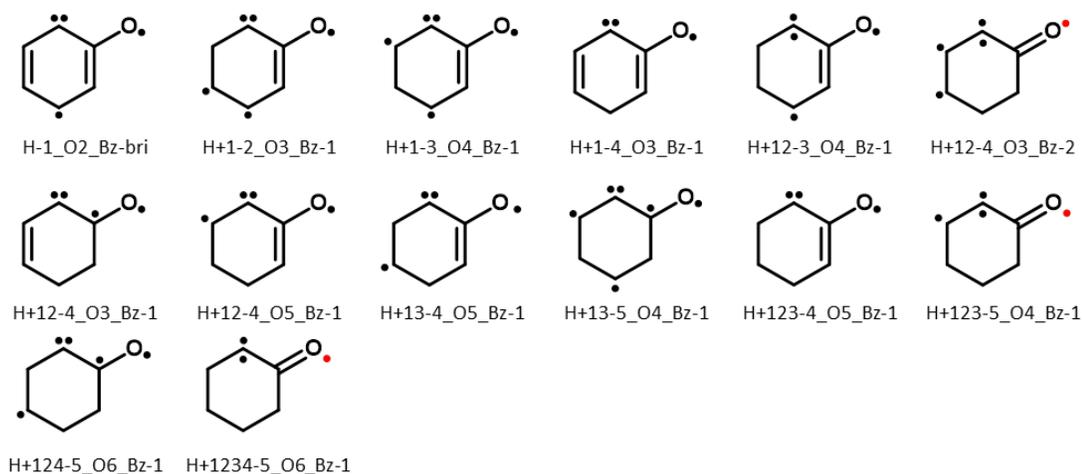


Figure B.4. Adsorption conformations of benzene ring with one H-deficient ring carbon and one surface interacting oxygen substituent and their hydrogenated derivatives.

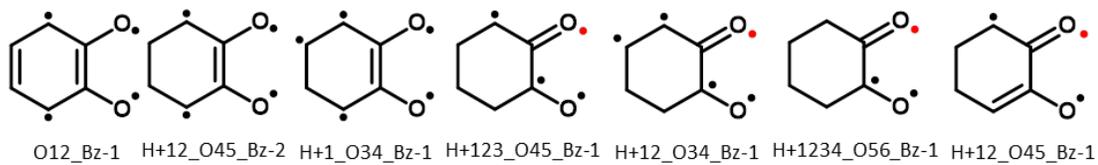


Figure B.5. Adsorption conformations of benzene ring with two surface interacting oxygen substituents and its hydrogenated derivatives.

## B.2 Schematics of Binding Type Couples

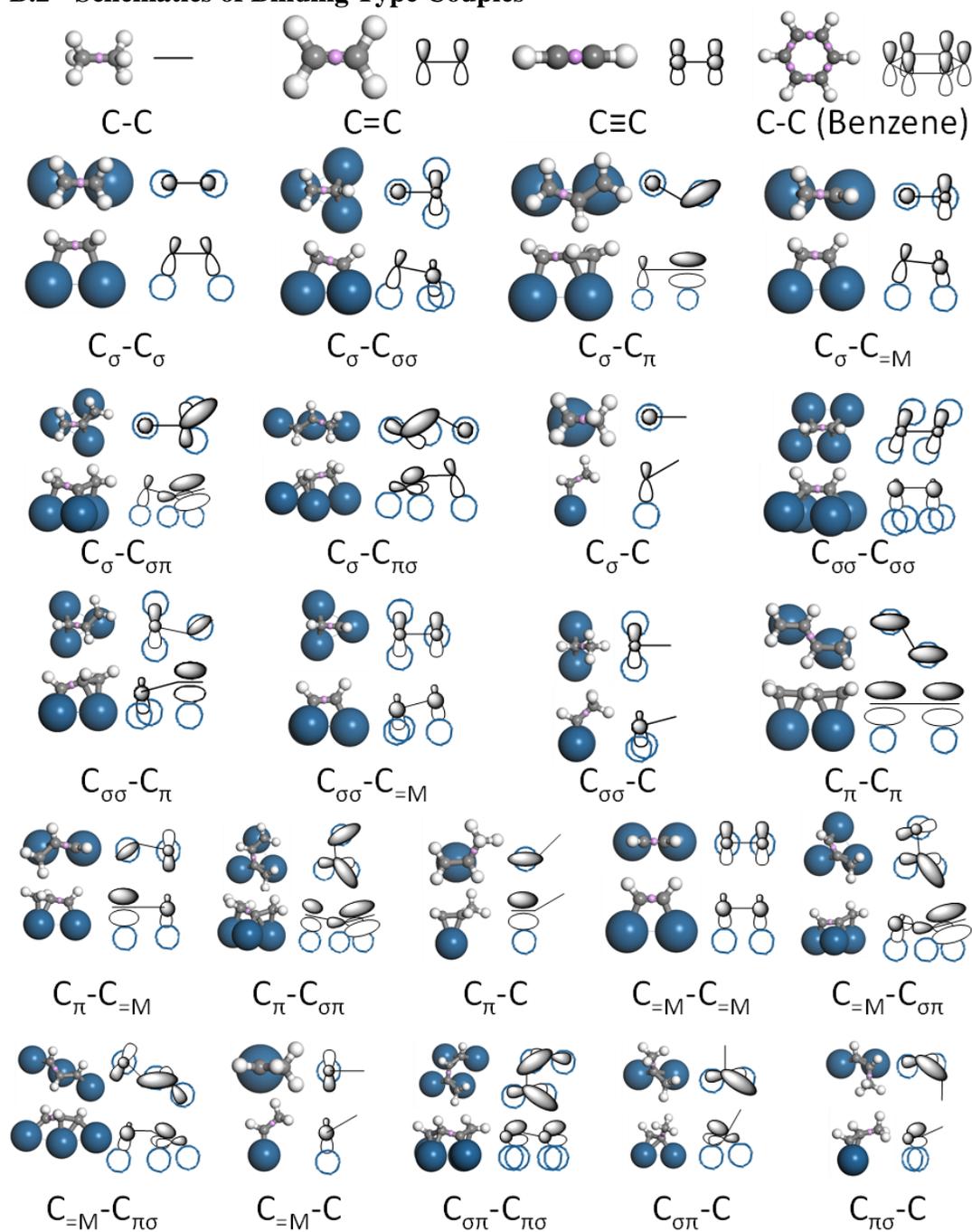


Figure B.6. Schematics of binding type couples listed in Table 1 along with expected *p*-orbital lobes in the same phase.[45, 151, 152, 349, 350] BCPs are shown in pink balls and Pt atoms in blue circles.

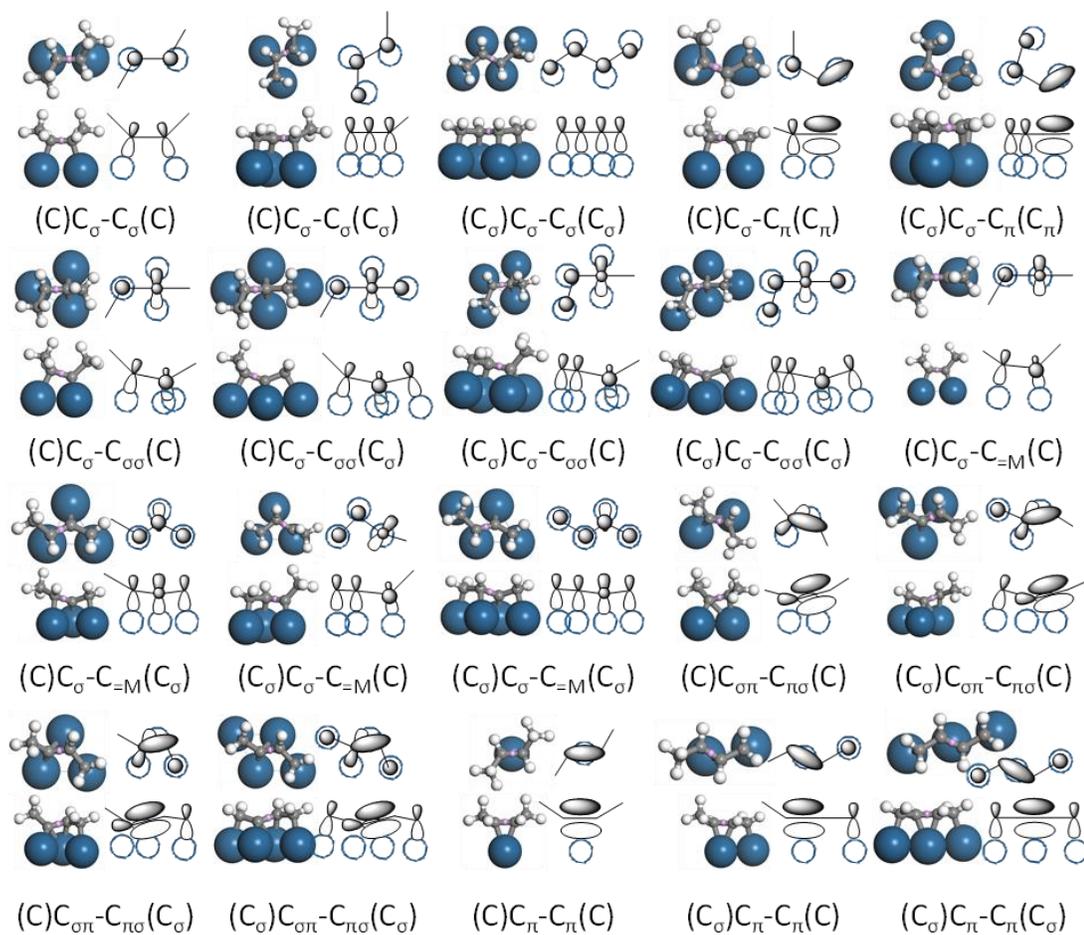


Figure B.7. Schematics of structures listed in Table 2 along with expected  $p$ -orbital lobes in the same phase.

### B.3 Charge Flow Analysis

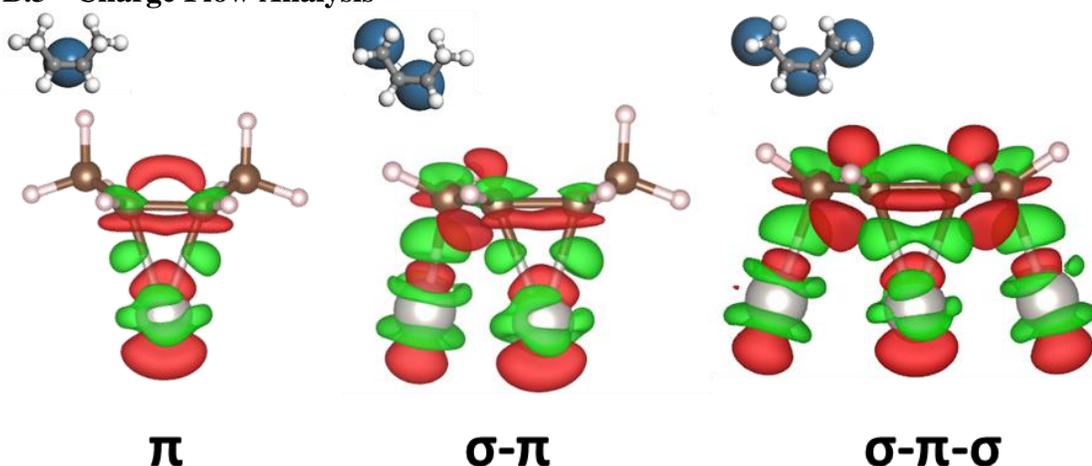


Figure B.8. Electron density difference upon  $\pi$ -mode adsorption of a  $C_2H_2$  fragment. Charge occupation of the conjugated  $\pi$  electron system depends on binding of substituent groups.

### B.4 Methoxy Strain Correction

In the training set, we observed strong strain related to  $-OCH$  and  $-OCH_2$  groups. The strain is visually observable where the methoxy and the ring are pulled towards the bridging oxygen, and the Pt atoms are pulled out of the surface (Figure B.9a, b). These strained fragments involve methoxy group bound to the surface, and the  $\alpha$ -carbon and the two  $\beta$ -carbons bound to three separate Pt atoms (Figure B.9c). We assigned two different corrections for the fragment involving  $-OCH$  and  $-OCH_2$  groups.

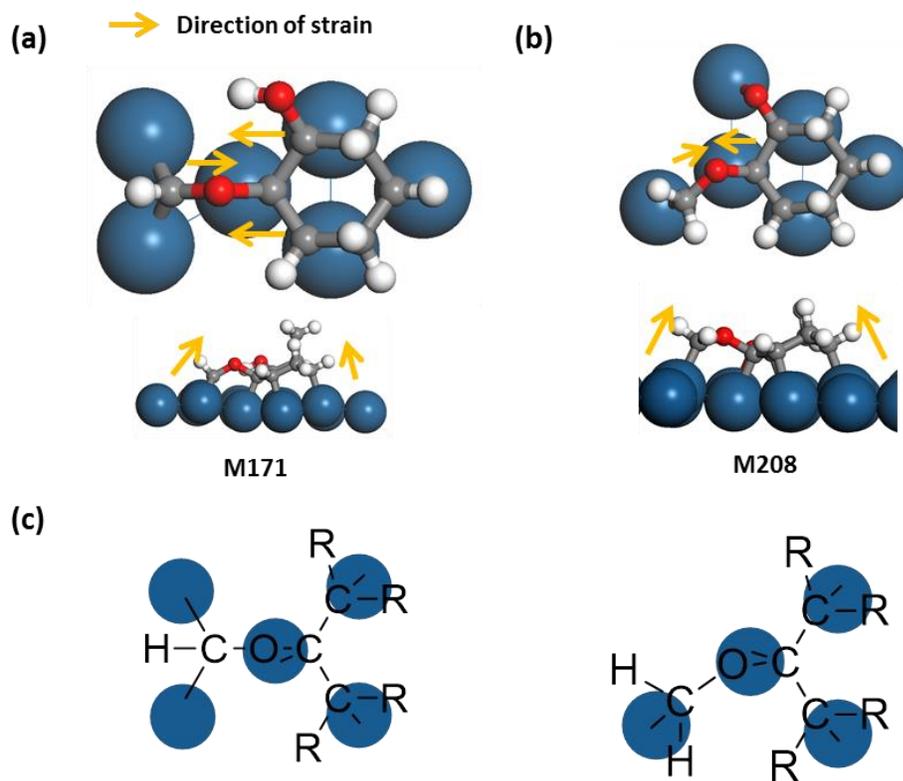


Figure B.9. (a) An example of a strained molecule with a  $\text{-OCH}$  group. (b) An example of a strained molecule with a  $\text{-OCH}_2$  group. (c) Schematics of strained fragments.

## Appendix C

### GROUP A ADDITIVITY FOR AQUEOUS PHASE THERMOCHEMICAL PROPERTIES OF ALCOHOLS ON Pt(111) – SUPPORTING INFORMATION

#### C.1 Temperature Independence of Gibbs Free Energy of Solvation Assumptions

Our test calculations (see Figure C.1) and previous study[351] also suggest that vibrational contribution is minimally affected by solvation. The solvation effect to translational contribution is often called liberation free energy and can be computed using the translational partition function of the gas phase,[351-353] as long as the reference state remains the same as those of the gas phase translational contribution calculation. However, the solvation effect to rotational contribution is a rather debated area, but the computation of it is difficult as the molecular rotation becomes hindered as solvational cage forms around the solute.[351, 354] Particularly, the PCM method cannot simulate the solvation hindered rotation as it requires explicit water molecules for their structural perturbation. Nonetheless, PCM method validation section below shows that this assumption holds up well.

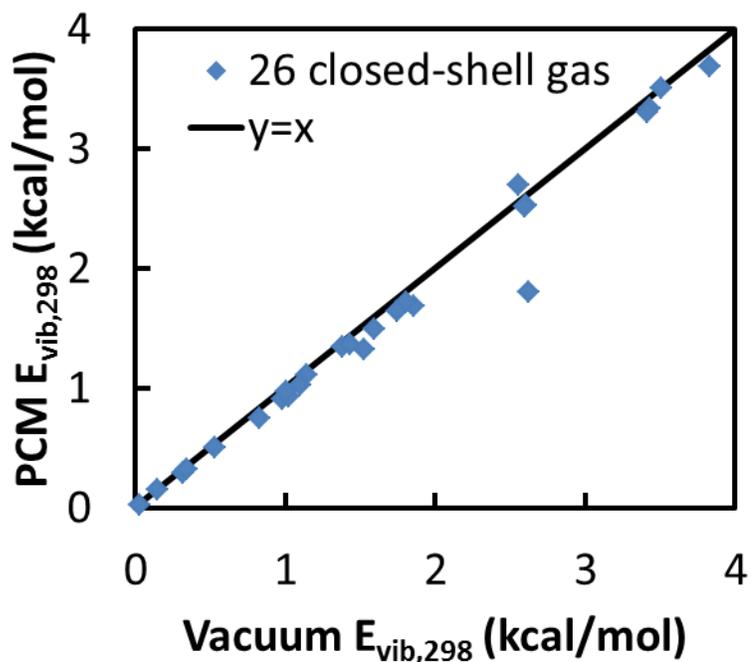


Figure C.1. Parity graph of vibrational energy between vacuum and PCM calculations. Total of 26 molecules are tested. Mean absolute deviation is 0.10 kcal/mol and max absolute deviation is 0.82 kcal/mol.

## C.2 PCM Method Validation

Figure C.2 shows the comparison between the PCM and the experimental data. We also included COSMO-SAC calculation results for the methodology comparison. Overall, the COSMO-SAC performs the best at the mean absolute error of 0.68 kcal/mol. For the PCM, the mean absolute error is 1.17 kcal/mol and 1.90 kcal/mol for when cavitation energy is included and not included, respectively. The PCM is expected to have errors as it computes solvation energy at the dilute limit instead of 1M of the experimental condition.

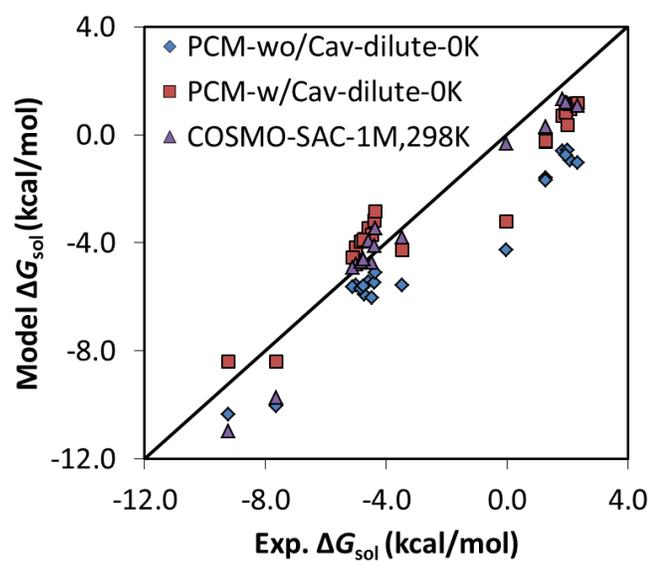


Figure C.2. Parity plot between the model computed solvation free energy and the experimental solvation free energy. Three models are shown: (1) PCM computed at 0 K without cavitation energy at dilute condition, (2) PCM computed at 0 K with cavitation energy at dilute condition, and (3) COSMO-SAC[355] computed at 298 K at 1 M. Experimental data[356] are obtained at 298 K at 1M condition.

### C.3 Group Additivity Method Validation

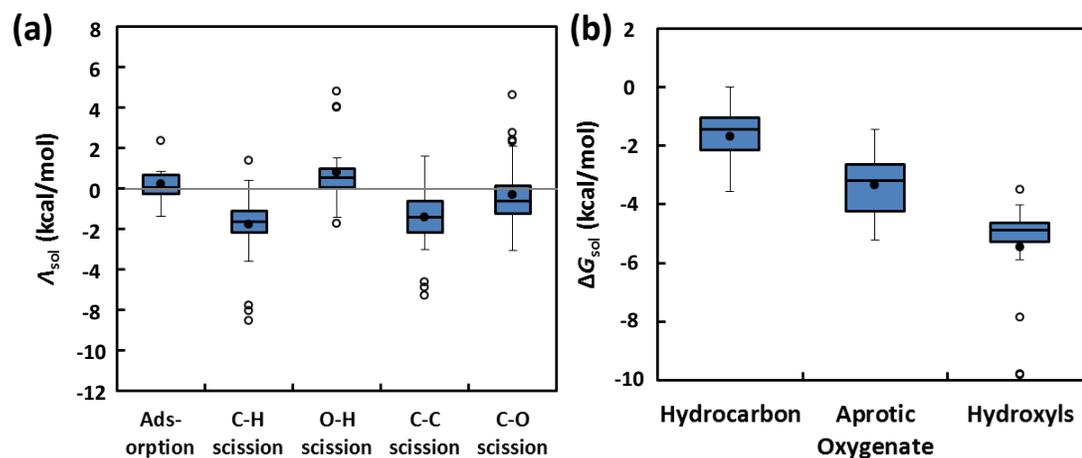


Figure C.3. Figure 4.3 is produced using group additivity method. Group additivity method reproduces the general trends seen by DFT well.

### C.4 Cavitation Energy Correction

Table C.1 shows the regression coefficient,  $e_{cav}$  from equation (4.12) and Table C.2 shows the DFT computed cavitation energy using equation (4.11) as well as the predicted value using the regression method introduced in equation (4.12). The MeanAE and MaxAE (training error, not the cross-validation error discussed in Chapter 4 Methods) of this model is insignificant at 0.18 and 0.47 kcal/mol.

Table C.1. Regression Coefficient

Name	$e_{cav}$ (kcal/mol)
C	0.33
H	0.06
O	0.23

Table C.2. Cavitation energy computed using DFT and the model in kcal/mol.

Name	C	H	O	$E_{\text{cav,DFT}}$	$E_{\text{cav,model}}$
acetaldehyde	2	4	1	1.35	1.12
acetylene	2	2	0	1.1	0.77
butan1ol	4	10	1	2.06	2.12
butan2ol	4	10	1	2.01	2.12
butane	4	10	0	1.88	1.89
ethan12diol	2	6	2	1.64	1.47
ethane	2	6	0	1.34	1.01
ethanol	2	6	1	1.45	1.24
ethene	2	4	0	1.36	0.89
methane	1	4	0	0.94	0.56
methanol	1	4	1	1.13	0.79
pentan1ol	5	12	1	2.34	2.57
pentan2ol	5	12	1	2.33	2.57
pentan3ol	5	12	1	2.26	2.57
pentane	5	12	0	2.19	2.34
prop1ene	3	6	0	1.49	1.33
propan123triol	3	8	3	1.97	2.14
propan1ol	3	8	1	1.74	1.68
propan2ol	3	8	1	1.71	1.68
propane	3	8	0	1.6	1.45
formaldehyde	1	2	1		0.67
ethenol	2	4	1		1.12
acrolein	3	4	1		1.45
allylalcohol	3	6	1		1.56
propenol	3	6	1		1.56
propionaldehyde	3	6	1		1.56
G01	3	4	2		1.68
G02	3	2	3		1.79
G03	4	4	3		2.24
G04	6	4	5		3.35
G05	4	4	3		2.24
G06	5	6	3		2.68
G07	6	8	3		3.12
G08	7	10	3		3.57
G09	7	8	4		3.68

Table C.2. continued.

G10	3	4	3		1.91
CO <sub>2g</sub>	1	0	2		0.79
C	1	0	0		0.33
CC	2	0	0		0.66
CCH <sub>2</sub> O	2	2	1		1
CCH <sub>2</sub> OH	2	3	1		1.06
CCHO	2	1	1		0.94
CCHOH	2	2	1		1
CCO	2	0	1		0.89
CCOH	2	1	1		0.94
CH	1	1	0		0.39
CH <sub>2</sub>	1	2	0		0.44
CH <sub>2</sub> C	2	2	0		0.77
CH <sub>2</sub> CH	2	3	0		0.83
CH <sub>2</sub> CH <sub>2</sub>	2	4	0		0.89
CH <sub>2</sub> CH <sub>2</sub> O	2	4	1		1.12
CH <sub>2</sub> CH <sub>2</sub> OH	2	5	1		1.18
CH <sub>2</sub> CHO	2	3	1		1.06
CH <sub>2</sub> CHOH	2	4	1		1.12
CH <sub>2</sub> CO	2	2	1		1
CH <sub>2</sub> COH	2	3	1		1.06
CH <sub>2</sub> O	1	2	1		0.67
CH <sub>2</sub> OH	1	3	1	0.82	0.73
CH <sub>3</sub>	1	3	0	0.78	0.5
CH <sub>3</sub> C	2	3	0		0.83
CH <sub>3</sub> CH	2	4	0		0.89
CH <sub>3</sub> CH <sub>2</sub>	2	5	0		0.95
CH <sub>3</sub> CH <sub>2</sub> O	2	5	1	1.33	1.18
CH <sub>3</sub> CH <sub>2</sub> OH	2	6	1	1	1.24
CH <sub>3</sub> CHO	2	4	1		1.12
CH <sub>3</sub> CHOH	2	5	1		1.18
CH <sub>3</sub> CO	2	3	1		1.06
CH <sub>3</sub> COH	2	4	1		1.12
CH <sub>3</sub> O	1	3	1	0.74	0.73
CH <sub>3</sub> OH	1	4	1	0.9	0.79
CHC	2	1	0		0.71
CHCH	2	2	0		0.77

Table C.2. continued.

CHCH2O	2	3	1		1.06
CHCH2OH	2	4	1		1.12
CHCHO	2	2	1		1
CHCHOH	2	3	1		1.06
CHCO	2	1	1		0.94
CHCOH	2	2	1		1
CHO	1	1	1		0.62
CHOH	1	2	1		0.67
CO	1	0	1	0.74	0.56
COH	1	1	1	0.66	0.62
H	0	1	0		0.06
H2O	0	2	1		0.35
O	0	0	1		0.23
OH	0	1	1		0.29
CHOO	1	1	2		0.85
CO2	1	0	2		0.79
COOH	1	1	2	0.8	0.85
CH4	1	4	0		0.56
M001	4	7	4		2.64
M002	4	6	4		2.58
M003	4	6	4		2.58
M004	4	8	4		2.7
M005	4	6	4		2.58
M006	4	6	4		2.58
M007	4	4	4		2.47
M008	4	4	3		2.24
M009	4	4	4		2.47
M010	4	5	2		2.06
M011	4	2	4		2.35
M012	4	2	4		2.35
M013	4	2	4		2.35
M014	4	3	4		2.41
M015	4	3	4		2.41
M016	4	6	3		2.35
M017	4	2	1		1.66
M018	4	5	3		2.29
M019	4	3	2		1.95

Table C.2. continued.

M020	4	6	3	2.35
M021	4	3	3	2.18
M022	4	6	1	1.89
M023	4	3	1	1.72
M024	4	0	1	1.54
M025	4	4	2	2.01
M026	4	3	3	2.18
M027	4	3	3	2.18
M028	4	0	3	2
M029	4	6	2	2.12
M030	4	4	4	2.47
M031	4	9	3	2.53
M032	4	7	1	1.95
M033	4	4	1	1.77
M034	4	6	3	2.35
M035	4	5	3	2.29
M036	4	0	3	2
M037	4	3	3	2.18
M038	4	0	3	2
M039	4	3	2	1.95
M040	4	3	2	1.95
M041	4	5	3	2.29
M042	4	5	1	1.83
M043	4	6	3	2.35
M044	4	5	2	2.06
M045	4	3	3	2.18
M046	4	5	2	2.06
M047	4	4	4	2.47
M048	4	3	2	1.95
M049	4	6	2	2.12
M050	4	2	2	1.89
M051	4	2	3	2.12
M052	4	4	3	2.24
M053	4	5	1	1.83
M054	4	4	3	2.24
M055	4	5	3	2.29
M056	4	3	2	1.95

Table C.2. continued.

M057	4	5	2	2.06
M058	4	4	3	2.24
M059	4	5	1	1.83
M060	4	4	2	2.01
M061	5	10	2	2.68
M062	4	4	2	2.01
M063	4	4	2	2.01
M064	4	3	3	2.18
M065	4	3	3	2.18
M066	4	6	2	2.12
M067	4	7	1	1.95
M068	4	5	2	2.06
M069	4	5	3	2.29
M070	3	1	2	1.5
M071	4	3	1	1.72
M072	3	1	1	1.27
M073	3	2	1	1.33
M074	5	8	2	2.57
M075	5	8	3	2.8
M076	5	8	3	2.8
M077	6	8	3	3.12
M078	4	7	1	1.95
M079	5	6	1	2.22
M080	5	7	1	2.28
M081	4	4	2	2.01
M082	4	3	1	1.72
M083	4	3	0	1.49
M084	4	2	2	1.89
M085	4	3	2	1.95
M086	4	3	3	2.18
M087	4	4	2	2.01
M088	4	4	1	1.77
M089	4	2	2	1.89
M090	4	2	3	2.12
M091	4	1	1	1.6
M092	4	3	2	1.95
M093	4	3	1	1.72

Table C.2. continued.

M094	4	6	2	2.12
M095	3	4	3	1.91
M096	4	5	3	2.29
M097	4	7	1	1.95
M098	4	4	2	2.01
M099	4	3	2	1.95
M100	4	3	2	1.95
M101	4	5	2	2.06
M102	4	5	2	2.06
M103	4	3	4	2.41
M104	4	2	3	2.12
M105	4	2	3	2.12
M106	4	4	4	2.47
M107	4	5	3	2.29
M108	4	2	3	2.12
M109	4	3	4	2.41
M110	4	4	4	2.47

### C.5 Ethanol Reforming Reaction Network Detail

The 151 reactions examined in the Chapter 4 are extracted from the previous study.[236] The reaction network contains dissociation products of ethanol as well as water gas shift reactions. The reaction reactants and products as well as  $\Delta G_{r,(vac)}$  and  $\Lambda^{\circ}_{sol}$  (see Chapter 4 Methods for definition) are listed in the Table C.3.

Table C.3. The  $\Delta G_{r,(vac)}$  and  $\Lambda^{\circ}_{sol}$  of the 151 reactions examined in ethanol reforming reaction network are shown, computed using the DFT and group additivity. Here, the group additivity  $\Delta G_{r,(vac)}$  and  $\Lambda^{\circ}_{sol}$  are computed from cross-validation computed the Gibbs free energy of formation and the Gibbs free energy of solvation of each molecule. The molecule names are in-line with those provided in the XLSX file in ref.[216]

Reaction Type	Reactant 1	Product 1	Product 2	$\Delta G_{r,(vac)}$	$\Lambda^{\circ}_{sol}$	$\Delta G_{r,(vac)}$	$\Lambda^{\circ}_{sol}$
$\alpha$ H abstraction	CH3CH2OH	CH3CHOH	H	-9.79	-0.72	-14.34	-1.04
$\alpha$ H abstraction	CH3CHOH	CH3COH	H	-9.02	0.2	-6.07	-1.14
$\alpha$ H abstraction	CH2CH2OH	CH2CHOH	H	-15.02	-1.05	-12.6	-1.7
$\alpha$ H abstraction	CH2CHOH	CH2COH	H	-9.01	0.09	-5.24	-1.17
$\alpha$ H abstraction	CHCH2OH	CHCHOH	H	-16.94	0.58	-11.49	-2.15
$\alpha$ H abstraction	CHCHOH	CHCOH	H	-9.8	-4.02	-12.76	-0.7
$\alpha$ H abstraction	CCH2OH	CCHOH	H	-7.28	-0.95	-1.23	-1.24
$\alpha$ H abstraction	CCHOH	CCOH	H	12.29	-6.46	5.35	-6.55
$\alpha$ H abstraction	CH3CH2O	CH3CHO	H	-13.91	-1.24	-15.09	-0.9
$\alpha$ H abstraction	CH3CHO	CH3CO	H	-18.59	-1.36	-14.51	-1.98
$\alpha$ H abstraction	CH2CH2O	CH2CHO	H	-14.17	-2.72	-19.92	-2.19
$\alpha$ H abstraction	CH2CHO	CH2CO	H	-20.72	-0.97	-16.18	-0.49
$\alpha$ H abstraction	CHCH2O	CHCHO	H	-19.54	-1.82	-18.13	-3.62
$\alpha$ H abstraction	CHCHO	CHCO	H	-16.95	-0.84	-15.7	-0.8
$\alpha$ H abstraction	CCH2O	CCHO	H	-28.14	-2.41	-29.48	-3
$\alpha$ H abstraction	CCHO	CCO	H	-10.11	-0.01	-6.36	-0.78
$\alpha$ H abstraction	CH3OH	CH2OH	H	-9.32	-1.07	-8.49	-1.72
$\alpha$ H abstraction	CH2OH	CHOH	H	-8.4	0.73	-8.99	-0.13
$\alpha$ H abstraction	CHOH	COH	H	-17.26	-5.79	-17.26	-5.79
$\alpha$ H abstraction	CH3O	CH2O	H	-11.46	-1.86	-9.26	-1.82
$\alpha$ H abstraction	CH2O	CHO	H	-23.26	-2.43	-22.68	-1.55

Table C.3. continued.

$\alpha$ H abstraction	CHO	CO	H	-21.15	0.4	-21.15	0.4
$\alpha$ H abstraction	CH <sub>4</sub>	CH <sub>3</sub>	H	0.67	-2.64	0.67	-2.64
$\alpha$ H abstraction	CH <sub>3</sub>	CH <sub>2</sub>	H	0.34	-1.96	0.34	-1.96
$\alpha$ H abstraction	CH <sub>2</sub>	CH	H	-16.74	-1.65	-16.74	-1.65
$\alpha$ H abstraction	CH	C	H	10.05	-0.33	10.05	-0.33
$\alpha$ H abstraction	ethane <sup>a</sup>	CH <sub>3</sub> CH <sub>2</sub>	H	3.18	-3.06	-5.95	-1.87
$\alpha$ H abstraction	CH <sub>3</sub> CH <sub>2</sub>	CH <sub>2</sub> CH <sub>2</sub>	H	-6.81	-2.08	-0.44	-2.65
$\alpha$ H abstraction	CH <sub>3</sub> CH	CH <sub>2</sub> CH	H	-5.72	-1.69	0.03	-2.86
$\alpha$ H abstraction	CH <sub>3</sub> C	CH <sub>2</sub> C	H	11.91	-1.71	8.22	-2.26
$\alpha$ H abstraction	CH <sub>2</sub> CH <sub>2</sub>	CH <sub>2</sub> CH	H	-0.76	-1.9	0.44	-1.6
$\alpha$ H abstraction	CH <sub>2</sub> CH	CHCH	H	-3.89	-1.97	-6.25	-2.2
$\alpha$ H abstraction	CH <sub>2</sub> C	CHC	H	14.93	-1.27	13.67	-0.63
$\alpha$ H abstraction	CHCH	CHC	H	15	-0.77	11.74	1.35
$\alpha$ H abstraction	CHC	CC	H	13.7	-0.03	15.88	-1.45
$\beta$ H abstraction	CH <sub>3</sub> CH <sub>2</sub> OH	CH <sub>2</sub> CH <sub>2</sub> OH	H	0.01	-2.36	-2.47	-1.95
$\beta$ H abstraction	CH <sub>3</sub> CHOH	CH <sub>2</sub> CHOH	H	-5.22	-2.68	-0.73	-2.61
$\beta$ H abstraction	CH <sub>3</sub> COH	CH <sub>2</sub> COH	H	-5.21	-2.8	0.1	-2.63
$\beta$ H abstraction	CH <sub>2</sub> CH <sub>2</sub> OH	CHCH <sub>2</sub> OH	H	-1.1	-2.39	0.03	-1.39
$\beta$ H abstraction	CH <sub>2</sub> CHOH	CHCHOH	H	-3.02	-0.75	1.14	-1.84
$\beta$ H abstraction	CH <sub>2</sub> COH	CHCOH	H	-3.81	-4.86	-6.38	-1.37
$\beta$ H abstraction	CHCH <sub>2</sub> OH	CCH <sub>2</sub> OH	H	-19.06	0	-16.91	-1.13
$\beta$ H abstraction	CHCHOH	CCHOH	H	-9.4	-1.54	-6.65	-0.22
$\beta$ H abstraction	CHCOH	CCOH	H	12.69	-3.98	11.46	-6.08
$\beta$ H abstraction	CH <sub>3</sub> CH <sub>2</sub> O	CH <sub>2</sub> CH <sub>2</sub> O	H	-6.15	-1.67	-1.73	-1.97
$\beta$ H abstraction	CH <sub>3</sub> CHO	CH <sub>2</sub> CHO	H	-6.41	-3.15	-6.57	-3.27
$\beta$ H abstraction	CH <sub>3</sub> CO	CH <sub>2</sub> CO	H	-8.53	-2.77	-8.24	-1.77
$\beta$ H abstraction	CH <sub>2</sub> CH <sub>2</sub> O	CHCH <sub>2</sub> O	H	-2	-1.89	0.45	-1.45
$\beta$ H abstraction	CH <sub>2</sub> CHO	CHCHO	H	-7.37	-0.99	2.24	-2.88
$\beta$ H abstraction	CH <sub>2</sub> CO	CHCO	H	-3.6	-0.86	2.71	-3.19
$\beta$ H abstraction	CHCH <sub>2</sub> O	CCH <sub>2</sub> O	H	0.61	-1.11	-9.08	-1.83
$\beta$ H abstraction	CHCHO	CCHO	H	-7.99	-1.7	-20.43	-1.21
$\beta$ H abstraction	CHCO	CCO	H	-1.15	-0.86	-11.09	-1.19
$\beta$ H abstraction	CH <sub>3</sub> CH <sub>2</sub>	CH <sub>3</sub> CH	H	-1.85	-2.29	-0.04	-1.39
$\beta$ H abstraction	CH <sub>3</sub> CH	CH <sub>3</sub> C	H	-21.45	-1.45	-16.38	-0.82
$\beta$ H abstraction	CH <sub>2</sub> CH	CH <sub>2</sub> C	H	-3.82	-1.47	-8.19	-0.22
O-H abstraction	CH <sub>3</sub> CH <sub>2</sub> OH	CH <sub>3</sub> CH <sub>2</sub> O	H	10.64	1.04	7.91	0.82
O-H abstraction	CH <sub>3</sub> CHOH	CH <sub>3</sub> CHO	H	6.52	0.53	7.16	0.96
O-H abstraction	CH <sub>3</sub> COH	CH <sub>3</sub> CO	H	-3.05	-1.03	-1.28	0.12

Table C.3. continued.

O-H abstraction	CH <sub>2</sub> CH <sub>2</sub> OH	CH <sub>2</sub> CH <sub>2</sub> O	H	4.48	1.73	8.65	0.79
O-H abstraction	CH <sub>2</sub> CHOH	CH <sub>2</sub> CHO	H	5.33	0.06	1.32	0.3
O-H abstraction	CH <sub>2</sub> COH	CH <sub>2</sub> CO	H	-6.38	-1	-9.61	0.98
O-H abstraction	CHCH <sub>2</sub> OH	CHCH <sub>2</sub> O	H	3.58	2.23	9.07	0.73
O-H abstraction	CHCHOH	CHCHO	H	0.98	-0.18	2.42	-0.74
O-H abstraction	CHCOH	CHCO	H	-6.17	3	-0.53	-0.84
O-H abstraction	CCH <sub>2</sub> OH	CCH <sub>2</sub> O	H	23.26	1.12	16.9	0.04
O-H abstraction	CCHOH	CCHO	H	2.39	-0.34	-11.36	-1.73
O-H abstraction	CCOH	CCO	H	-20.01	6.12	-23.07	4.05
O-H abstraction	CH <sub>3</sub> OH	CH <sub>3</sub> O	H	9.82	2.52	7.26	0.1
O-H abstraction	CH <sub>2</sub> OH	CH <sub>2</sub> O	H	7.67	1.73	6.5	0.01
O-H abstraction	CHOH	CHO	H	-7.19	-1.42	-7.19	-1.42
O-H abstraction	COH	CO	H	-11.07	4.77	-11.07	4.77
O-H abstraction	H <sub>2</sub> O	OH	H	12.83	4	12.83	4
O-H abstraction	OH	O	H	-2.47	1.51	-2.47	1.51
C-C dissociation	CH <sub>3</sub> CH <sub>2</sub> OH	CH <sub>3</sub>	CH <sub>2</sub> OH	-0.31	-2.3	0.19	-1.55
C-C dissociation	CH <sub>3</sub> CHOH	CH <sub>3</sub>	CHOH	1.08	-0.86	5.53	-0.64
C-C dissociation	CH <sub>3</sub> COH	CH <sub>3</sub>	COH	-7.16	-6.85	-5.65	-5.29
C-C dissociation	CH <sub>2</sub> CH <sub>2</sub> OH	CH <sub>2</sub>	CH <sub>2</sub> OH	0.01	-1.91	3	-1.56
C-C dissociation	CH <sub>2</sub> CHOH	CH <sub>2</sub>	CHOH	6.63	-0.13	6.61	0
C-C dissociation	CH <sub>2</sub> COH	CH <sub>2</sub>	COH	-1.62	-6.01	-5.41	-4.62
C-C dissociation	CHCH <sub>2</sub> OH	CH	CH <sub>2</sub> OH	-15.62	-1.17	-13.76	-1.83
C-C dissociation	CHCHOH	CH	CHOH	-7.08	-1.03	-11.27	0.2
C-C dissociation	CHCOH	CH	COH	-14.54	-2.8	-15.78	-4.9
C-C dissociation	CCH <sub>2</sub> OH	C	CH <sub>2</sub> OH	13.49	-1.49	13.19	-1.02
C-C dissociation	CCHOH	C	CHOH	12.36	0.19	5.42	0.09
C-C dissociation	CCOH	C	COH	-17.19	0.86	-17.19	0.86
C-C dissociation	CH <sub>3</sub> CH <sub>2</sub> O	CH <sub>3</sub>	CH <sub>2</sub> O	-3.28	-1.61	-1.22	-2.36
C-C dissociation	CH <sub>3</sub> CHO	CH <sub>3</sub>	CHO	-12.63	-2.8	-8.81	-3.02
C-C dissociation	CH <sub>3</sub> CO	CH <sub>3</sub>	CO	-15.19	-1.05	-15.45	-0.64
C-C dissociation	CH <sub>2</sub> CH <sub>2</sub> O	CH <sub>2</sub>	CH <sub>2</sub> O	3.21	-1.91	0.85	-2.35
C-C dissociation	CH <sub>2</sub> CHO	CH <sub>2</sub>	CHO	-5.88	-1.61	-1.9	-1.72
C-C dissociation	CH <sub>2</sub> CO	CH <sub>2</sub>	CO	-6.31	-0.24	-6.87	-0.83
C-C dissociation	CHCH <sub>2</sub> O	CH	CH <sub>2</sub> O	-11.53	-1.67	-16.34	-2.55
C-C dissociation	CHCHO	CH	CHO	-15.25	-2.27	-20.88	-0.48
C-C dissociation	CHCO	CH	CO	-19.45	-1.03	-26.32	0.71
C-C dissociation	CCH <sub>2</sub> O	C	CH <sub>2</sub> O	-2.1	-0.88	2.79	-1.05
C-C dissociation	CCHO	C	CHO	2.79	-0.9	9.6	0.41

Table C.3. continued.

C-C dissociation	CCO	C	CO	-8.25	-0.49	-5.19	1.58
C-C dissociation	ethane <sup>a</sup>	CH3	CH3	9.55	-3.31	4.55	-1.92
C-C dissociation	CH3CH2	CH3	CH2	6.71	-2.2	10.84	-2
C-C dissociation	CH3CH	CH3	CH	-8.18	-1.56	-5.87	-2.26
C-C dissociation	CH3C	CH3	C	23.32	-0.44	20.55	-1.77
C-C dissociation	CH2CH2	CH2	CH2	13.86	-2.09	11.62	-1.32
C-C dissociation	CH2CH	CH2	CH	-2.12	-1.84	-5.56	-1.36
C-C dissociation	CH2C	CH2	C	11.75	-0.69	12.67	-1.47
C-C dissociation	CHCH	CH	CH	-14.97	-1.52	-16.05	-0.81
C-C dissociation	CHC	CH	C	-19.92	-1.07	-17.74	-2.49
C-C dissociation	CC	C	C	-23.57	-1.36	-23.57	-1.36
C-O dissociation	CH3CH2OH	CH3CH2	OH	8.71	-0.82	4.5	-1.12
C-O dissociation	CH3CHOH	CH3CH	OH	16.65	-2.39	18.8	-1.48
C-O dissociation	CH3COH	CH3C	OH	4.22	-4.04	8.5	-1.15
C-O dissociation	CH2CH2OH	CH2CH2	OH	1.89	-0.54	6.53	-1.82
C-O dissociation	CH2CHOH	CH2CH	OH	16.15	-1.39	19.56	-1.73
C-O dissociation	CH2COH	CH2C	OH	21.34	-2.95	16.62	-0.78
C-O dissociation	CHCH2OH	CH2CH	OH	2.23	-0.05	6.94	-2.04
C-O dissociation	CHCHOH	CHCH	OH	15.28	-2.6	12.18	-2.09
C-O dissociation	CHCOH	CHC	OH	40.08	0.64	36.67	-0.04
C-O dissociation	CCH2OH	CH2C	OH	17.48	-1.52	15.66	-1.12
C-O dissociation	CCHOH	CHC	OH	39.68	-1.84	30.56	-0.51
C-O dissociation	CCOH	CC	OH	41.09	4.59	41.09	4.59
C-O dissociation	CH3OH	CH3	OH	6.08	0.17	6.31	-1.33
C-O dissociation	CH2OH	CH2	OH	15.74	-0.72	15.15	-1.57
C-O dissociation	CHOH	CH	OH	7.4	-3.09	7.4	-3.09
C-O dissociation	COH	C	OH	34.71	2.37	34.71	2.37
C-O dissociation	CH3CH2O	CH3CH2	O	-4.4	-0.35	-5.88	-0.43
C-O dissociation	CH3CHO	CH3CH	O	7.66	-1.4	9.17	-0.92
C-O dissociation	CH3CO	CH3C	O	4.8	-1.5	7.3	0.24
C-O dissociation	CH2CH2O	CH2CH2	O	-5.06	-0.76	-4.6	-1.11
C-O dissociation	CH2CHO	CH2CH	O	8.35	0.06	15.77	-0.52
C-O dissociation	CH2CO	CH2C	O	25.24	-0.44	23.76	-0.25
C-O dissociation	CHCH2O	CH2CH	O	-3.83	-0.77	-4.6	-1.26
C-O dissociation	CHCHO	CHCH	O	11.83	-0.92	7.28	0.16
C-O dissociation	CHCO	CHC	O	43.77	-0.85	34.72	2.31
C-O dissociation	CCH2O	CH2C	O	-8.25	-1.13	-3.71	0.35
C-O dissociation	CCHO	CHC	O	34.82	0.01	39.45	2.73

Table C.3. continued.

C-O dissociation	CCO	CC	O	58.63	-0.01	61.69	2.06
C-O dissociation	CH3O	CH3	O	-6.21	-0.84	-3.42	0.08
C-O dissociation	CH2O	CH2	O	5.59	-0.94	6.17	-0.07
C-O dissociation	CHO	CH	O	12.11	-0.17	12.11	-0.17
C-O dissociation	CO	C	O	43.3	-0.89	43.3	-0.89
Adsorption	ethanol	CH3CH2OH		-3.03	-0.43	-2.52	0.08
Adsorption	acetaldehyde	CH3CHO		1.59	2.36	2.27	2.34
Adsorption	ethene	CH2CH2		-20.07	-1.16	-16.97	-1.4
Adsorption	acetylene	CHCH		-46.87	0.64	-45.79	-0.07
Adsorption	methanol	CH3OH		-0.64	-1.13	-3.52	0.61
Adsorption	formaldehyde	CH2O		-1.46	1.71	-2.04	0.84
Adsorption	methane	CH4		9.04	-0.24	9.04	-0.24
Adsorption	CO2 <sup>g</sup> <sup>b</sup>	CO2 <sup>b</sup>		2.02	-0.46	2.02	-0.46
water-gas-shift	COOH	CO	OH	6.28	0.22	6.28	0.22
water-gas-shift	CHOO	CHO	O	16.16	-1.47	16.16	-1.47
water-gas-shift	COOH	CO2	H	-10.1	1.61	-10.1	1.61

<sup>a</sup> Dissociative adsorptions are used for ethane reactions. <sup>b</sup> CO2<sup>g</sup> is the gaseous carbon dioxide, and CO2 is the Physisorbed carbon dioxide.

## Appendix D

### THERMOCHEMISTRY OF GAS-PHASE AND SURFACE SPECIES VIA LASSO-ASSISTED SUBGRAPH SELECTION – SUPPORTING INFORMATION

#### D.1 Supporting Figures

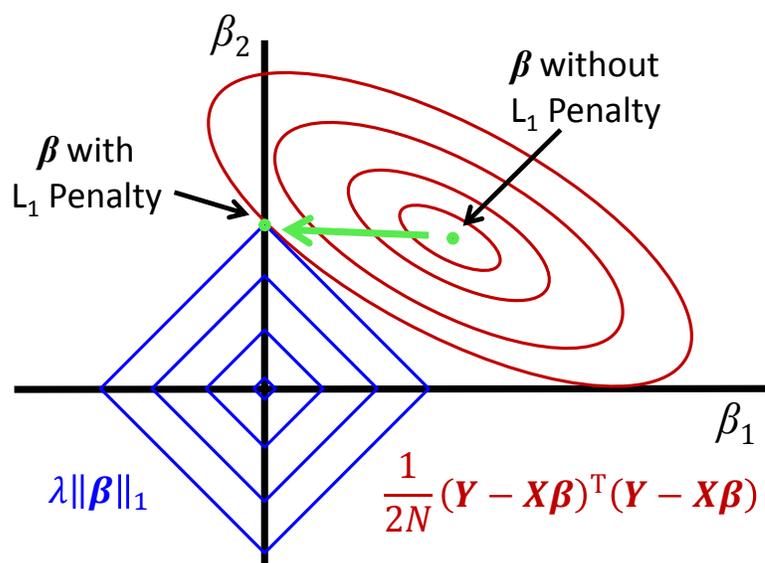


Figure D.1. Illustration of L1 norm penalty constraining least square loss function. The red surface indicates the quadratic loss where the outer ellipse indicates a higher loss. The blue surface indicates the L1 norm penalty where the outer diamond corresponds to a higher penalty. In this illustration,  $\boldsymbol{\beta} = \{\beta_1, \beta_2\}$  converges to the center of the red surface without the L1 norm penalty, whereas,  $\boldsymbol{\beta}$  converges to a solution with  $\beta_1 = 0$  for the L1 norm penalty.

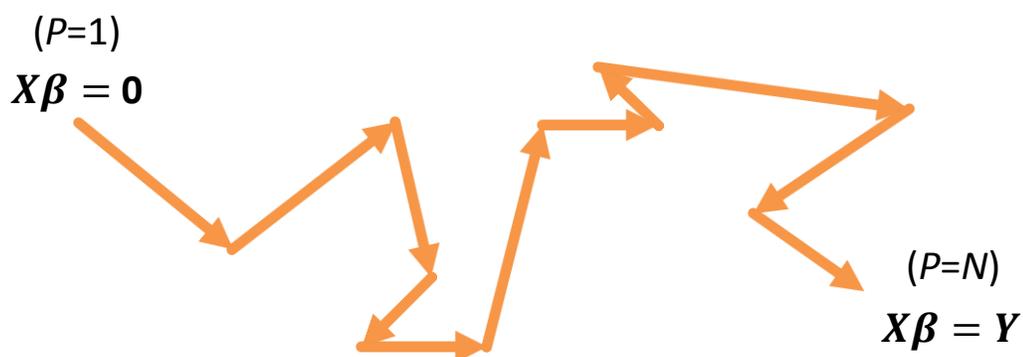


Figure D.2. Illustration of LAR-LASSO tracking predicted response vector,  $X\beta$ , along  $\lambda$ .  $X\beta$  is projected to 2 dimensions from the dimension of  $N$ , the number of data points.  $\mathbf{0} \in \mathbb{R}^P$  is the zero vector. LAR-LASSO starts at  $X\beta = \mathbf{0}$  with initial descriptor selected, then  $\beta$  is moved along the direction of  $\gamma$  until the addition or removal of a descriptor occurs. The direction of  $X\beta$  changes at each addition or removal event. Eventually, LAR-LASSO terminates when the number of parameter,  $P$ , is equal to  $N$  when the selected descriptor space has enough degrees of freedom to completely describe the data.

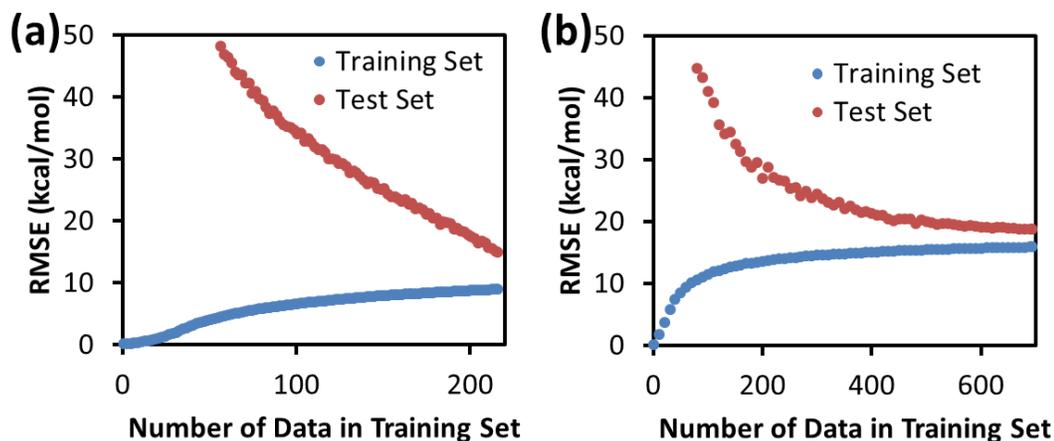


Figure D.3. The learning curve of Benson's group additivity model applied to a (a) 217 and (b) 695 gas species data set. Leave-one-out cross-validation was used with 100 training-set data shuffling for (a) and 10 for (b).

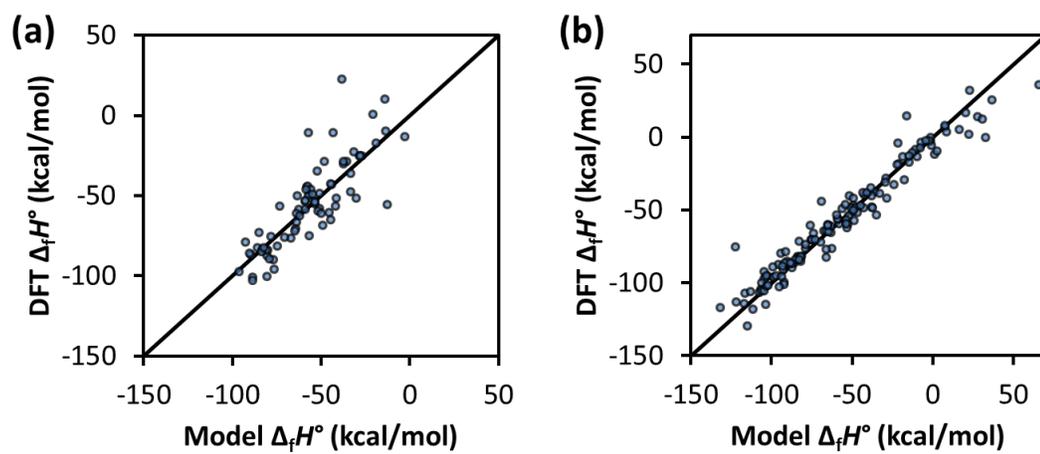


Figure D.4. Parity plot between DFT-computed and predicted  $\Delta_f H^\circ$  for (a) furanics[16] and (b) linear oxygenates[23].

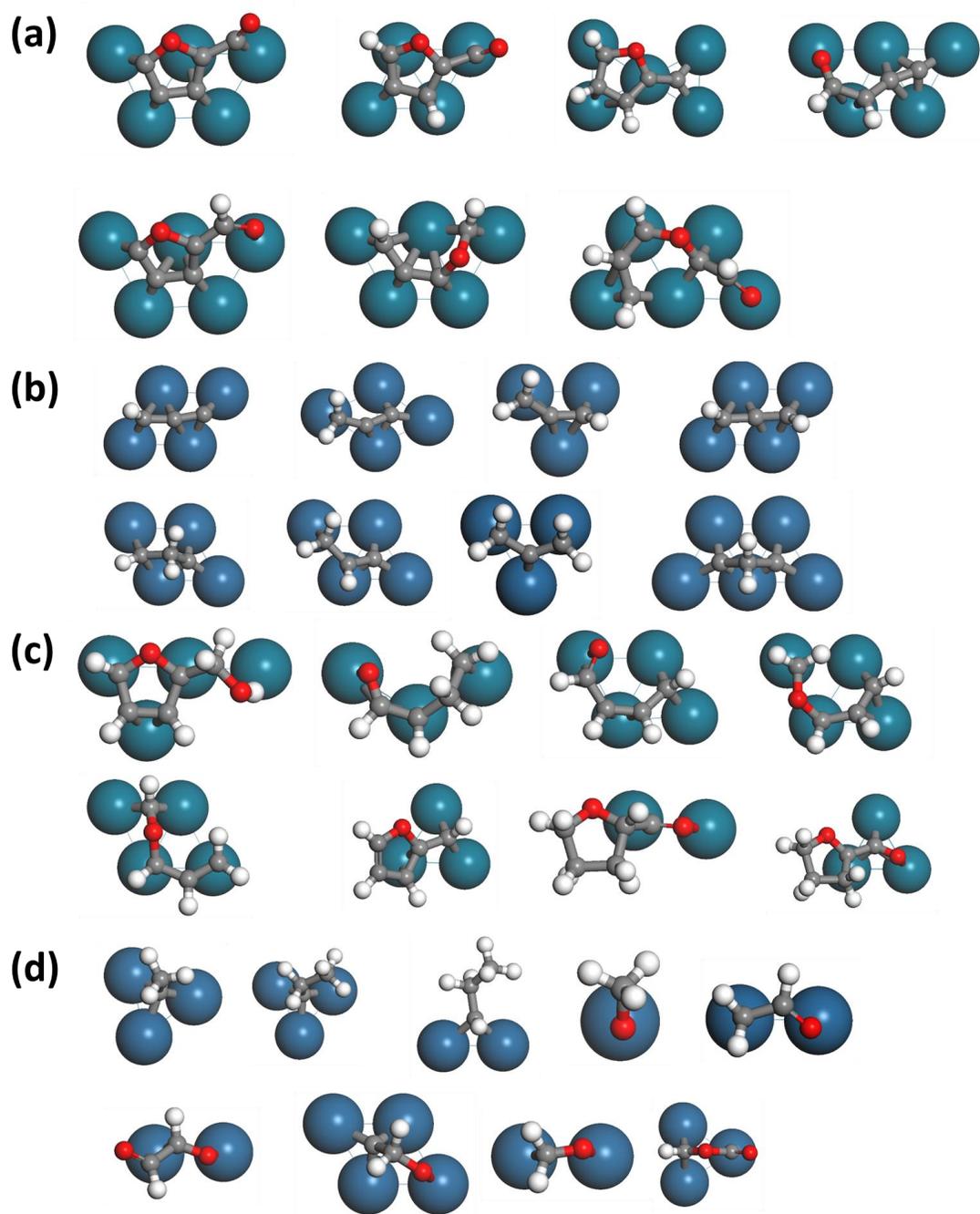


Figure D.5. Outliers in furanics ((a) and (c)) and linear oxygenates ((b) and (d)) data set. Adsorbates in (a) and (b) are mostly unsaturated molecules expected to be outliers in chemical space, and those in (b) and (d) are outliers in the sampled data, indicating insufficient sampling.

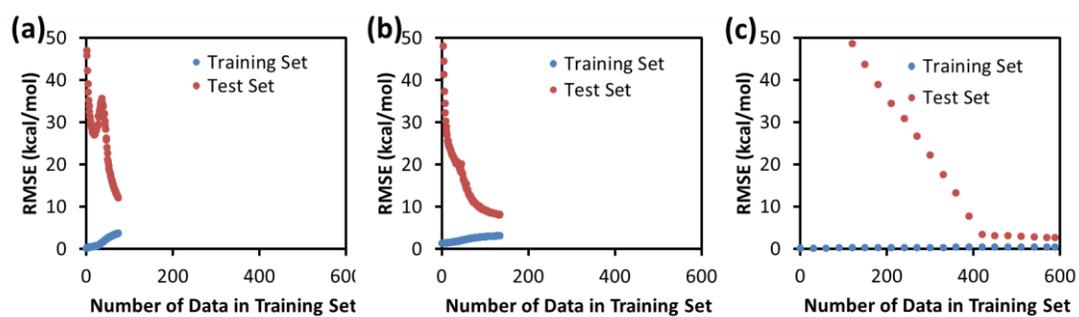


Figure D.6. Learning curve of the model selected by LASSO using the entire data set for (a) the furanics, (b) the linear oxygenates, and (c) the lignin data sets. The curves are the average of 100 training set shuffles for (a) and (b) and of 10 shuffles for (c).

## Appendix E

### MICROKINETIC MODELING OF AQUEOUS PHASE ETHYLENE GLYCOL REFORMING

#### E.1 Introduction

Hydrogen production from biomass by aqueous-phase reforming (APR) has been popularized by Dumesic and co-workers.[13] Since then, many studies have focused on understanding the reforming mechanism in vapor and aqueous phases.[13, 18, 114, 127, 357-359] While computations can provide valuable mechanistic insights and potentially help with catalyst design, such effort has been hindered from the high computational intensity of (1) computing kinetics and thermochemistry for biomass molecules involving more than three heteroatoms and (2) handling aqueous-phase conditions. To address these issues, we have recently developed an aqueous-phase adsorbate group additivity scheme on Pt.[216] Here we use the group additivity and BEP relations,[43] and provide preliminary insights into APR of ethylene glycol on Pt(111).

#### E.2 Method

##### E.2.1 Reaction Network Enumeration

The reaction network of ethylene glycol is enumerated via the automatic mechanism generation method RING.[94] We applied C-H, O-H, C-C, and C-O scission rules, resulting in 193 reactions among 86 species. In addition, the water-gas-

shift reaction (WGSR) is important.[13] Thus we added 3 more species and 12 reactions relevant to the WGSR.[236]

### E.2.2 Microkinetic Model Parameterization

Out of 89 species, the thermochemistry of 53 species is available in the dataset used to develop the solvation group additivity scheme. The thermochemistry for the rest of the species is computed using the group additivity scheme mentioned above. The aqueous and vapor phase barriers of most reactions (161 reactions) are available in the literature.[357] For those missing (32 C-O scission reactions), we used the vapor-phase barriers from our previous ethanol study on Pt[236] for both vapor and aqueous phase MKMs in this study. A tacit assumption is that thermochemistry is affected by the solvent and the difference in barriers between the vacuum and the aqueous phase is small. Indeed calculations by Heyden and co-workers in aqueous phase indicate that about 80% of all barriers fall within 0.05 eV of the vapor phase counterparts.[41] For any other missing barriers (12 C-OH scission reactions), we used the BEP relations from our previous work.[43] We have implemented lateral interactions using the following equation:

$$E_{\text{BE},i} = E_{\text{BE},i,0} + \sum_j w_{ij}\theta_j \quad (\text{E.1})$$

where  $E_{\text{BE},i}$  is the binding energy of species  $i$ ,  $E_{\text{BE},i,0}$  is the zero-coverage binding energy of species  $i$ ,  $w_{ij}$  is the lateral interaction of species  $j$  on species  $i$ , and  $\theta_j$  is the coverage of species  $j$ . We assign 15 kcal mol<sup>-1</sup> coverage<sup>-1</sup> lateral interaction to each species from itself as suggested by Faheem et al.[357] For CH<sub>3</sub>C and CH, 30 kcal/mol<sup>-1</sup> coverage<sup>-1</sup> lateral interaction was assigned from itself as well as from CO and CH<sub>3</sub>C by 15 kcal mol<sup>-1</sup> coverage<sup>-1</sup> as suggested by Sutton et al.[127]

We performed microkinetic modeling using our in-house reactor code built around CHEMKIN<sup>TM</sup>.<sup>[109]</sup> We assume that the gaseous, closed-shell ethylene glycol derivatives (hydroxyacetaldehyde, ethanol, acetaldehyde, ethane, ethylene, acetylene, methanol, acetone, methane) are in equilibrium with the respective surface species, and thus bulk solvation is bypassed. These compounds are in very low concentrations in this system. We compute the thermochemistry ( $\Delta_f H^\circ$ ,  $S^\circ$ ,  $C_{P,T}$ ) of liquid water and aqueous-phase ethylene glycol (high concentration in water) using UNIFAC, and we assume that the liquid water and aqueous-phase ethylene glycol are in equilibrium with physisorbed water and ethylene glycol on Pt(111), respectively.

### **E.3 Model Validation and Mechanistic Insights**

Here, we simulate experimental data reported by Dumesic and co-workers<sup>[358, 359]</sup> of ethylene glycol APR over Pt/Al<sub>2</sub>O<sub>3</sub> as shown in Figure E.1. Overall, turn-over-frequency (TOF) predictions are all within one order of magnitude (Figure E.1). In addition, the reaction orders with respect to ethylene glycol and hydrogen agree qualitatively with experimental data (Table E.1), indicating that the species involved in the rate-determining step and site blocking are likely captured well.

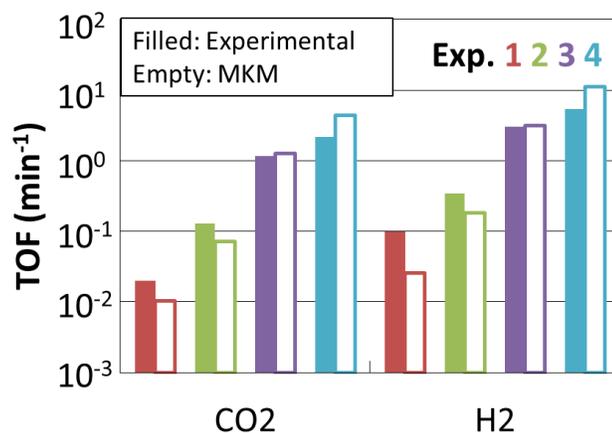


Figure E.1. Comparison between experimental (filled bars) and model (empty bars) turn-over-frequency (TOF) for CO<sub>2</sub> and H<sub>2</sub> over Pt/Al<sub>2</sub>O<sub>3</sub>. Four experiments from [358] are considered. The conditions are: temperature (K), pressure (bar, H<sub>2</sub>), and liquid hourly space velocity (h<sup>-1</sup>):(1) 423, 4.8, 1.9 (2) 453, 10.0, 3.8 (3) 483, 19.6, 42.9 and (4) 498, 25.8, 128.6. Calculations done at 1-3% conversion (surface to volume ratio of ~5 cm<sup>-1</sup>) and experiments at 1-10% conversion.

Table E.1. Comparison between experimental[359] and model reaction orders of ethylene glycol and hydrogen. For ethylene glycol reaction order, conditions are 1, 4, and 10 wt% ethylene glycol feed at 225 °C, 29.3 bar (H<sub>2</sub>), 9.36 mol ethylene glycol · mol<sup>-1</sup> · site h<sup>-1</sup>. For hydrogen reaction order, conditions are 10 wt% ethylene glycol feed at 225 °C, 9.36 mol ethylene glycol · mol<sup>-1</sup> · site h<sup>-1</sup> with 30.3, 32.3, 39.3 bar of H<sub>2</sub>.

Species	Exp.	Model
Ethylene Glycol	0.4	0.54
Hydrogen	-0.5	-0.48

The reaction pathway of ethylene glycol reforming in vapor and aqueous phases is shown in Figure E.2. Our simulation suggests two major pathways for vapor phase initiated by (1) C-H scission or (2) O-H scission. These two initial reactions are the rate-determining step for each respective pathway. Subsequently, the molecule undergoes a number of dehydrogenations, followed by C-CO cracking. This finding is

in agreement with previous microkinetic models for ethylene glycol vapor-phase reforming.[245, 357]. In aqueous phase, the pathway initiated by O-H scission becomes less favorable due to the lower barrier of the competitive C-H scission pathway (19.1 vs. 17.3 kcal/mol for vapor and aqueous phase, respectively) compared to the vapor phase as well as an increase in the C-H scission barrier of the OCHCH<sub>2</sub>OH for formation of the OCCH<sub>2</sub>OH (6.2 vs. 7.1 kcal/mol respectively for vapor and aqueous phase). This is in contrast to a previous aqueous phase microkinetic model study, where O-H scission pathway remains active.[357] The transition state for the initial C-H scission is likely stabilized as the underside hydrogen of the ethylene glycol becomes exposed to the bulk water. On the other hand, the transition state of the C-H scission of OCHCH<sub>2</sub>OH is stabilized less in the aqueous phase relative to that in the vapor phase, resulting in a higher barrier in aqueous phase. Overall, CO\* (adsorbed CO) formation pathways are similar. Specifically, CO is formed from C-H scission of CHO at 35% and 28%, from O-H scission of COH at 16% and 23%, and from C-C scission CHOHCO at 49% and 45% in aqueous and vapor phase, respectively. The small changes (within a few kcal/mol) of the barriers mentioned above are within the error of DFT and thus the changes in pathways may be sensitive to the parameterization of the model. Interestingly, a key difference is the high concentration of water that solvates the catalyst surface and promotes the WGSR.

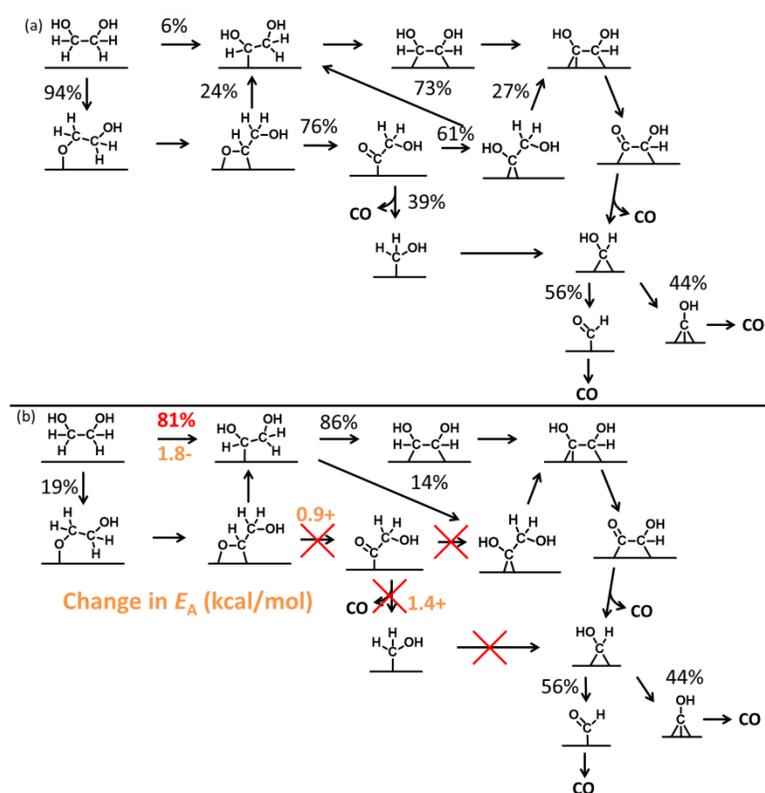


Figure E.2. Major reaction pathway leading to CO in (a) vapor phase (210 °C, 1 bar, 0.20 gas mol flow rate/Pt site [s<sup>-1</sup>], 25% ethylene glycol and 75% steam; ethylene glycol steam reforming conditions are comparable to those reported in ref[127]) and (b) aqueous phase (150 °C, 4.8 bar, 1.9 liquid hourly space velocity [h<sup>-1</sup>], 5 wt% ethylene glycol). CO is further converted to CO<sub>2</sub> and H<sub>2</sub>O via the water-gas-shift reaction.

A dramatic difference is observed in the outlet gas-composition, as shown in Figure E.3. The selectivity in the aqueous phase is 100% towards the reforming (CO<sub>2</sub> and H<sub>2</sub>) whereas CO is observed in the vapor phase. This is due to the high water coverage on the surface in the aqueous phase, as shown in Figure E.4. The high water coverage drives the WGSR in the aqueous phase, increasing the hydrogen selectivity.

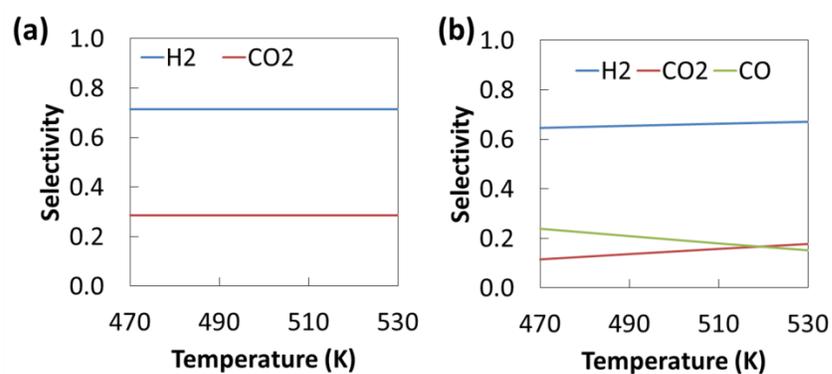


Figure E.3. Gas composition (mol%) vs. temperature for (a) aqueous phase and (b) vapor phase. Conditions are the same as in Figure E.2.

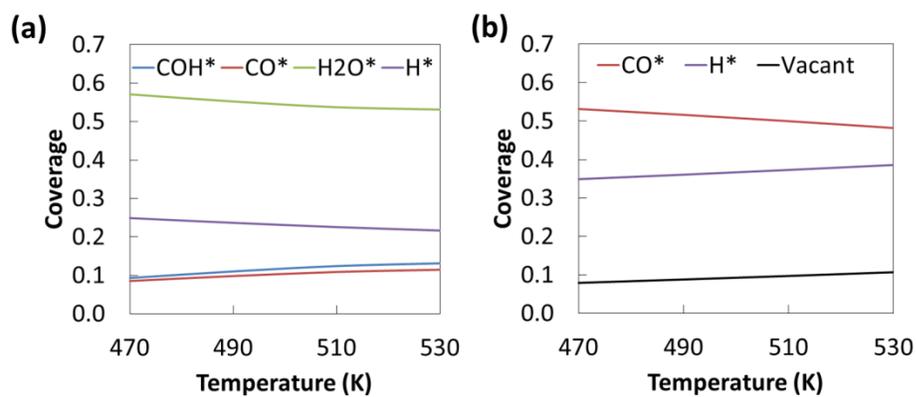


Figure E.4. Surface coverage of major species vs. temperature for (a) aqueous phase and (b) vapor phase. Conditions are the same as in Figure E.2.

## Appendix F

### PERMISSION FOR REPRINT



**Title:** Mechanism of Dehydration of Phenols on Noble Metals via First-Principles Microkinetic Modeling

**Author:** Geun Ho Gu, Charles A. Mullen, Akwasi A. Boateng, et al

**Publication:** ACS Catalysis

**Publisher:** American Chemical Society

**Date:** May 1, 2016

Copyright © 2016, American Chemical Society

**LOGIN**

If you're a [copyright.com user](#), you can login to RightsLink using your [copyright.com credentials](#). Already a [RightsLink user](#) or want to [learn more?](#)

#### PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK

CLOSE WINDOW

Copyright © 2018 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement](#). [Terms and Conditions](#). Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)



**Title:** Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111)  
**Author:** Geun Ho Gu, Dionisios G. Vlachos  
**Publication:** The Journal of Physical Chemistry C  
**Publisher:** American Chemical Society  
**Date:** Sep 1, 2016  
Copyright © 2016, American Chemical Society

[LOGIN](#)  
If you're a [copyright.com](#) user, you can login to RightsLink using your [copyright.com](#) credentials. Already a [RightsLink](#) user or want to [learn more?](#)

### PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)



**Title:** Group Additivity for Aqueous Phase Thermochemical Properties of Alcohols on Pt(111)

**Author:** Geun Ho Gu, Benjamin Schweitzer, Carine Michel, et al

**Publication:** The Journal of Physical Chemistry C

**Publisher:** American Chemical Society

**Date:** Oct 1, 2017

Copyright © 2017, American Chemical Society

**LOGIN**

If you're a [copyright.com](#) user, you can login to RightsLink using your [copyright.com](#) credentials. Already a [RightsLink](#) user or want to [learn more?](#)

#### PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

[BACK](#)[CLOSE WINDOW](#)

## Thermochemistry of Gas-phase and Surface Species via LASSO-assisted Subgraph Selection

G. H. Gu, P. Plechac and D. Vlachos, *React. Chem. Eng.*, 2018, Accepted Manuscript, **DOI:** 10.1039/C7RE00210F

If you are not the author of this article and you wish to reproduce material from it in a third party non-RSC publication you must [formally request permission](#) using RightsLink. Go to our [Instructions for using RightsLink page](#) for details.

Authors contributing to RSC publications (journal articles, books or book chapters) do not need to formally request permission to reproduce material contained in this article provided that the correct acknowledgement is given with the reproduced material.

Reproduced material should be attributed as follows:

- For reproduction of material from NJC:  
Reproduced from Ref. XX with permission from the Centre National de la Recherche Scientifique (CNRS) and The Royal Society of Chemistry.
- For reproduction of material from PCCP:  
Reproduced from Ref. XX with permission from the PCCP Owner Societies.
- For reproduction of material from PPS:  
Reproduced from Ref. XX with permission from the European Society for Photobiology, the European Photochemistry Association, and The Royal Society of Chemistry.
- For reproduction of material from all other RSC journals and books:  
Reproduced from Ref. XX with permission from The Royal Society of Chemistry.

If the material has been adapted instead of reproduced from the original RSC publication "Reproduced from" can be substituted with "Adapted from".

In all cases the Ref. XX is the XXth reference in the list of references.

If you are the author of this article you do not need to formally request permission to reproduce figures, diagrams etc. contained in this article in third party publications or in a thesis or dissertation provided that the correct acknowledgement is given with the reproduced material.

Reproduced material should be attributed as follows:

- For reproduction of material from NJC:  
[Original citation] - Reproduced by permission of The Royal Society of Chemistry

(RSC) on behalf of the Centre National de la Recherche Scientifique (CNRS) and the RSC

- For reproduction of material from PCCP:  
[Original citation] - Reproduced by permission of the PCCP Owner Societies
- For reproduction of material from PPS:  
[Original citation] - Reproduced by permission of The Royal Society of Chemistry (RSC) on behalf of the European Society for Photobiology, the European Photochemistry Association, and RSC
- For reproduction of material from all other RSC journals:  
[Original citation] - Reproduced by permission of The Royal Society of Chemistry

If you are the author of this article you still need to obtain permission to reproduce the whole article in a third party publication with the exception of reproduction of the whole article in a thesis or dissertation.

Information about reproducing material from RSC articles with different licences is available on our [Permission Requests page](#).