DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Understanding Catalysts With Density Functional Theory and Machine Learning

Citation

Hoyt, Robert. 2018. Understanding Catalysts With Density Functional Theory and Machine Learning. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:41129179

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Understanding Catalysts with Density Functional Theory and Machine Learning

A DISSERTATION PRESENTED

BY

ROBERT HOYT

TO

THE DEPARTMENT OF PHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

PHYSICS

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

APRIL 2018

©2018 – ROBERT HOYT

ALL RIGHTS RESERVED.

Dissertation Advisor: Professor Efthimios Kaxiras Robert Hoyt

Understanding Catalysts with Density Functional Theory and Machine Learning

ABSTRACT

Catalysts are the cornerstone of the chemicals industry, whose products are used in nearly all human endeavors. At the core of catalysis lies the intricate relationship between their atoms and electrons, where quantum mechanics dictates interactions with reactants, products, and electromagnetic fields. This is the "electronic structure" of catalysts, and studying this structure provides deep insight into the understanding and design of novel catalytic materials. This thesis focuses on understanding a small subset of promising heterogeneous catalytic systems using density functional theory (DFT), from oxygen evolution over polyiodide-doped graphene to the nonadiabatic dissociation of hydrogen over Cu nanoclusters. Some of these studies emphasize the importance of nonadiabatic behavior, especially magnetization transitions in Cu nanoclusters upon hydrogen dissociation. Further insights into catalytic properties can be obtained by comparing DFT calculations to corresponding machine learning predictions. For example, differences between DFT and empirical data-driven kernels highlight important discontinuous quantum mechanical effects in H adsorption on dilute Ag alloys. The studies presented here are examples of how detailed electronic structure calculations can be used to develop a deeper understanding of catalysts and how they might be improved.

Contents

1 ELECTRONIC STRUCTURE THEORY FOR CATALYSIS **1**
- 1.1 The Schrödinger Equation 1
  - 1.1.1 Dimensional Analysis of the Schrödinger Equation 2
  - 1.1.2 Quantum-Classical Dynamics 3
  - 1.1.3 Adiabatic Dynamics 4
  - 1.1.4 Born-Oppenheimer Approximation 4
  - 1.1.5 Potential Energy Surfaces 6
- 1.2 Density Functional Theory 7
  - 1.2.1 Exchange-Correlation Functionals 11
  - 1.2.2 Basis Sets and Superposition Error 12
  - 1.2.3 Electron Spin and Spin-Orbit Coupling 15
- 1.3 Zero-Point Vibrational Energy 18
- 1.4 Classical Transition State Theory 20
  - 1.4.1 Canonical Ensemble Derivation 20
  - 1.4.2 Gambling with Numerical Error 23
  - 1.4.3 Finding Transition States 24
- 1.5 Time-Dependent DFT and Ehrenfest Dynamics 28
  - 1.5.1 Molecular Photodissociation 31
  - 1.5.2 H Atom Spin Transitions 33
- 1.6 Conclusion 34

2 POLYIODIDE-DOPED GRAPHENE **36**
- 2.1 Introduction 36
- 2.2 Computational Details 39
- 2.3 Results and Discussion 41
  - 2.3.1 Formation Energies 41
  - 2.3.2 Electronic Structure 44
  - 2.3.3 Charge Transfer 47
  - 2.3.4 Polyiodide Stability 48
  - 2.3.5 Mechanism of Catalysis 50
- 2.4 Conclusion 51

3 ANHYDROUS ALCOHOL DEHYDROGENATION ON CU(111) **53**
- 3.1 Introduction 54
- 3.2 Computational Details 57
  - 3.2.1 DFT and MD Parameters 57
  - 3.2.2 Structures and Energetics 58
- 3.3 Dehydrogenation Intermediates 63
  - 3.3.1 Alcohols 65
  - 3.3.2 Alkoxys 65
  - 3.3.3 Aldehydes 66
  - 3.3.4 Hydrogen 67
  - 3.3.5 Coverage 67
  - 3.3.6 Ethanol Hydrogen Bonding 68
- 3.4 Surface Reactivity 72
  - 3.4.1 Alcohols to Alkoxys 75
  - 3.4.2 Alkoxys to Aldehydes 76

3.4.3 Aldehyde Desorption . . . 77
3.4.4 Temperature Comparison with Experiments . . . 78
3.5 Chemical Defects: Isolated Pt Atoms . . . 80
3.6 Catalytic Implications . . . 83
3.6.1 Branching Energies for Longer Alcohols . . . 83
3.6.2 Competitive Binding . . . 84
3.6.3 Hydrogen Recombination . . . 85
3.7 Conclusions . . . 85

4 NONADIABATIC H_2 DISSOCIATION ON CU_{13} NANOCLUSTERS **88**
4.1 Introduction . . . 89
4.2 Results and Discussion . . . 90
4.2.1 Electronic Structure . . . 90
4.2.2 Magnetization-Dependent Reactivity . . . 92
4.2.3 Magnetization Transition Probabilities . . . 94
4.3 Possible Catalytic Improvements . . . 95
4.3.1 Magnetic Excitations . . . 95
4.3.2 Substitution and Charge States . . . 96
4.3.3 Co-Catalysts . . . 98
4.4 Conclusion . . . 98

5 MACHINE LEARNING **100**
5.1 Introduction . . . 100
5.2 Machine Learning of Regression Models . . . 101
5.2.1 Kernels . . . 102
5.2.2 Kernel Trick . . . 103
5.2.3 Kernel Matrix as a Covariance Matrix . . . 104
5.2.4 Kernel Regression . . . 106
5.2.5 Principal Component Analysis and Clustering . . . 133
5.3 Model Training . . . 139
5.3.1 Cost Functions: LASSO and Ridge . . . 139
5.3.2 LASSO versus Ridge Regression . . . 140
5.3.3 Maximum Likelihood and Bayesian Optimization . . . 141
5.4 H Adsorption on Ag Alloys . . . 146
5.4.1 Introduction . . . 147
5.4.2 Data and Models . . . 148
5.4.3 Results and Discussion . . . 149
5.4.4 Data Mining . . . 149
5.4.5 Initial Features and Models . . . 151
5.4.6 Physical Insights . . . 154
5.4.7 Final Model . . . 159
5.4.8 Conclusion . . . 163

6 CONCLUSIONS **165**

A MAGNETIZATION SWITCHES IN CU_{13} **181**
A.1 Computational Settings . . . 181
A.2 Basis Set Optimization . . . 182
A.3 Landau-Zener Approximation . . . 183
A.4 DFT Estimate of Spin-Orbit Coupling . . . 184

B DETAILS OF MACHINE LEARNING ON AG ALLOYS **187**
- B.1 Density Functional Theory Calculations . . . 187
- B.2 Data Mining . . . 188
- B.3 Outliers . . . 191
- B.4 Densities of States . . . 192
- B.5 List of Features . . . 194
- B.6 Top Features . . . 196

Acknowledgments

My first encounter with Harvard was driving past it with my dad, seemingly the one campus we *didn't* visit during our college "grand tours" a decade ago. After he pointed it out, I thought "yeah right, who actually goes here?" Despite the winters, it turns out that it's full of terrific people.

Some journeys end much as they begin: in this case with hydrogen. From my first group, Xi Wang and I had quite the adventure studying H_2 production from *Chlamydomonas reinhardtii* and the related mysteries of photosynthetic electron transfer. Her enthusiasm and attention to detail was an inspiration. Tony Zhou, Jesse Crossno, Evan Walsh, and Daniel Wintz introduced me to graduate student life here and made lab feel like a second home. Like my early career in experimental laser optics, it was fun while it lasted.

From my second group, I will start by thanking Efthimios Kaxiras for being a tremendous advisor and building a great group. His guidance and generosity convinced me to stay in graduate school. Grigory Kolesov and Oscar Grånäs put in serious work guiding me through the finer points of nonadiabatic dynamics and helping me settle in to the group as I switched into computational physics. Horseradish knows how Grigory's Perl scripts work, but they sure saved me a lot of time. In these early years Dmitry Vinichenko was a true comrade as well, assisting me in the serious business of ensuring that the chemistry department's pool table and Hemenway's free weights were well used. As my research transitioned to surface catalysis, Matt Montemore's expertise, help, and more than a little patience was crucial as I developed most of this thesis' content. Many of these projects would not have been possible without additional help from Wei Chen and Ioanna Fampiou. I'd also like to thank Trevor Rhone for our discussions of machine learning as well as my other officemates Daniel Larson and Cedric Flamant. Best of luck to Cedric throughout his own graduate school adventures.

Outside Cruft there were innumerable other sources of help and motivation. Cynthia Friend, her group members, and the IMASC collaboration have given me a great opportunity to learn about catalysis and how these DFT calculations fit in with the rest of the field, and my collaboration with Zhitao Wang and Charlie Sykes deepened my understanding of alcohol dehydrogenation in particular. I also appreciate Cynthia Friend and Philip Kim serving on my thesis committee. Outside of lab, living with Louis Ryan, Kevin Thompson, Levi Dudte, and Lee Weinstein was a whole other adventure. Finally, and most importantly, I'd like to thank my wonderful parents Richard and Sylvia Hoyt for a lifetime of encouragement and opportunities.

Further Acknowledgements by Chapter

Chapter 1 contains content that was previously published in the following article:

- Grigory Kolesov, Oscar Grånäs, **Robert Hoyt**, Dmitry Vinichenko, and Efthimios Kaxiras. "Real-Time TD-DFT with Classical Ion Dynamics: Methodology and Applications." *J. Chem. Theory Comput.* **12**(2) 466 (2016)

Chapter 2 contains content that was previously published in the following article:

- **Robert A. Hoyt**, E. Marielle Remillard, Ekin D. Cubuk, Chad D. Vecitis, and Efthimios Kaxiras. "Polyiodide-Doped Graphene." *J. Phys. Chem. C* **121**(1) 609 (2017)

Chapter 3 contains content from the following publications:

- **Robert A. Hoyt**, Matthew M. Montemore, Robert J. Madix, E. Charles H. Sykes, and Efthimios Kaxiras. " Anhydrous Methanol and Ethanol Dehydrogenation on Cu(111) Step Defects." *in preparation*
- Zhi-Tao Wang, **Robert A. Hoyt**, Mostafa El-Soda, Robert J. Madix, Efthimios Kaxiras, and E. Charles H. Sykes. "Dry Dehydrogenation of Ethanol on Pt–Cu Single Atom Alloys." *Top Catal.* **1** (2017)

Chapter 4 and Appendix A contain content being prepared for the following publication:

- **Robert A. Hoyt**, Matthew M. Montemore, and Efthimios Kaxiras. "Nonadiabatic Hydrogen Dissociation on Copper Nanoclusters." *in preparation*

Chapter 5 and Appendix B contain content being prepared for the following publication:

- **Robert A. Hoyt**, Matthew M. Montemore, Ioanna Fampiou, Wei Chen, and Efthimios Kaxiras. "H Adsorption Energies on Dilute Ag Alloys: A Data Science Approach." *in preparation*

1

Electronic Structure Theory for Catalysis

Section 1.5 is adapted from the following publication:

> Grigory Kolesov, Oscar Grånäs, **Robert Hoyt**, Dmitry Vinichenko, and Efthimios Kaxiras. "Real-Time TD-DFT with Classical Ion Dynamics: Methodology and Applications." *J. Chem. Theory Comput.* **12** 2016, 466–476.

1.1 THE SCHRÖDINGER EQUATION

The Schrödinger Equation is the core of practical electronic structure theory. The following groups related terms together for further discussion:

$$i\hbar\frac{\partial\psi(\mathbf{R},\mathbf{r},t)}{\partial t} = \left(\mathcal{H}_K + \mathcal{H}_{\text{ion}} + \mathcal{H}_{SOC} + V_{\text{rest}}(\mathbf{R},\mathbf{r})\right)\psi(\mathbf{R},\mathbf{r},t) \tag{1.1a}$$

$$\mathcal{H}_K \equiv \frac{\hbar^2}{2}\left(\sum_I \frac{\nabla_I^2}{M_I} + \sum_i \frac{\nabla_i^2}{m}\right) \tag{1.1b}$$

$$\mathcal{H}_{\text{ion}} \equiv \sum_{I<J}\frac{Z_I Z_J e^2}{r_{IJ}} + \sum_{i<j}\frac{e^2}{r_{ij}} - \sum_{Ij}\frac{Z_I e^2}{r_{Ij}} \tag{1.1c}$$

$$\mathcal{H}_{\text{SOC}} \equiv -\frac{1}{2m^2c^2}\sum_i \mathbf{s}_i\cdot\left(\nabla_i V_i(\mathbf{r}_i)\times\vec{p}_i\right) \tag{1.1d}$$

These equations use the following notation:

$\mathbf{R}$	Coordinates of all nuclei
$\mathbf{r}$	Coordinates of all electrons
I,J	A nucleus; $\mathbf{R}_I$ is the position of atom I
i,j	An electron; $\mathbf{r}_i$ is the position of electron i
Z_I	The atomic number of atom I
$r_{\alpha\beta}$	Distance between particles α and β

$\mathbf{s}$	Electron spin operator
$V_i(\mathbf{r})$	The total potential acting on electron i, including nuclei and other electrons
V_{rest}	Catch-all term for other interactions including electric fields, relativistic effects beyond $\mathcal{H}_{\text{SOC}}$, etc.

The expression for $\mathcal{H}_{\text{SOC}}$ is approximate, but generally contains the most important relativistic effects in catalytic studies.

Solving the Schrödinger Equation for the wavefunction ψ gives the probability density for observing the particles at any given set of positions $\mathbf{R}$ and $\mathbf{r}$. The operators in Equation (1.1a) are straightforward, but explicit solutions are only available for very simple systems including the hydrogen atom and the harmonic oscillator. The main obstacle for larger systems is the *curse of dimensionality*. Most applications of electronic structure theory in this thesis are catalyst surfaces, with approximately 50-100 atoms and several hundred valence electrons. Each of these particles has three spatial coordinates, yielding roughly 1,000 coordinates in total. Even if the equation was solved on a numerical grid with only 2 points per dimension, one would need 2^{1000} total grid points. Computers with such vast memory do not exist.

1.1.1 Dimensional Analysis of the Schrödinger Equation

The simplifications described in the following subsections are best motivated by applying dimensional analysis to a simplified Schrödinger Equation for the hydrogen atom:

$$i\hbar\frac{\partial\psi(\mathbf{r},\mathbf{R},t)}{\partial t} = \left(-\frac{\hbar^2}{2m}\nabla_e^2 - \frac{\hbar^2}{2M_H} - \nabla_H^2\frac{e^2}{r}\right)\psi(\mathbf{r},\mathbf{R},t) \tag{1.2a}$$

$$i\frac{\partial\psi(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{\tau})}{\partial\tau} = \left(-\frac{1}{2}\nabla_e^2 - \frac{1}{2}\frac{m}{M_H}\nabla_H^2 - \frac{1}{\rho}\right)\psi(\boldsymbol{\rho},\boldsymbol{P},\boldsymbol{\tau}) \tag{1.2b}$$

$$L \equiv \frac{\hbar^2}{m e^2} \approx 0.529\ \text{Å} \tag{1.2c}$$

$$T \equiv \frac{\hbar^3}{m e^4} \approx 0.024\ \text{fs} \tag{1.2d}$$

where the hydrogen atom is located near $\mathbf{R} = 0$. Making the variable replacements $t \to T\boldsymbol{\tau}$, $\mathbf{r} \to L\boldsymbol{\rho}$, and $\mathbf{R} \to L\mathbf{P}$, then solving for L and T in terms of physical constants, yields the dimensionless Schrödinger Equation in Equation (1.2b). These length scales give tremendous insight into the equations. First, we see that electronic wavefunctions vary significantly on the scales of 1 Bohr and 24 attoseconds. Variations over smaller distance and time scales are generally minor. Second, the length scale for the hydrogen atom's wavefunction is $\frac{M_H}{m} \approx 14,000$ times smaller than that of the electron. This reveals a companion to the curse of dimensionality: the "curse of scale." Handling both nuclear and electronic quantum effects simultaneously, especially with finite-precision arithmetic, is challenging because extremely fine grids are required to accurately track the nuclear degrees of freedom.

1.1.2 QUANTUM-CLASSICAL DYNAMICS

The "curse of scale" can be eliminated by treating the nuclei as point particles. This strategy is usually motivated though perturbation theory. Since nuclei are much more massive than electrons, $\frac{1}{M_I} \ll \frac{1}{m}$, the nuclear contribution to the kinetic energy in Equation (1.1b) is small and can be considered a perturbing Hamiltonian. In most cases simply neglecting the nuclear contribution entirely (i.e. the zeroth order solution) is sufficiently accurate. In other cases where nuclear quantum effects are important, e.g. electron-phonon coupling, they can be included using first order or even higher order perturbation theory. An even simpler motivation for neglecting the nuclear kinetic energy term is to consider dimensional analysis. As discussed above and shown in Equation (1.2b), the length scale for nuclear wavefunctions is vastly smaller than the scale for electronic wavefunctions. Solutions to the Schrödinger Equation are therefore approximately separable by inspection. In particular, relative to electronic length scales, the nuclear solutions are well approximated as point particles. This assertion is more rigorously justified in Section 1.1.4 by deriving the Born-Oppenheimer approximation.

In any case, this "classical ion" approximation demotes **R** to a parameter. Solutions to the Schrödinger Equation are then found at fixed **R** values, neglecting the quantum details of the nuclei entirely. Catalyst design almost always invokes this simplification so most of this thesis considers nuclei to be point particles. However, the classical ion approximation ignores several effects that can be catalytically relevant:

1. Zero-point vibrations. Treating nuclei as classical particles neglects the zero-point energy (ZPE) of vibrational modes. ZPE is particularly significant for bonds involving H atoms.
2. Quantum tunneling. Demoting **R** to a parameter neglects the spatial delocalization of nuclear wavefunctions. In the limit of low temperatures and atomic mass, delocalizaed atoms can tunnel past the classical activation energy barrier.
3. Electron-phonon coupling. Rapid atomic motion, such as in molecular beams and bond activation, perturbs electronic states and thus couples them. Raman scattering, electron-hole pair excitations,[39,129] and "electronic friction"[53] are notable examples.

Corrections for most of these effects can be included in the classical ion approach when necessary. For example, zero-point vibrational energy corrections are applied in my work on alcohol reactivity presented in

Chapter 3, and electron-phonon coupling is implicitly included in Ehrenfest dynamics as discussed below in Section 1.5.

1.1.3 ADIABATIC DYNAMICS

Another major simplification, often combined with the classical ion approximation, is to only consider the electronic ground states. This is also motivated by dimensional analysis. Time scales for atomic motion are set by the system's highest frequency bond vibration. The highest frequencies are reached in H-containing compounds, H–F in particular with a vibrational period of approximately 7.5 fs. The standard "safe" molecular dynamics time step is then about 1 fs. In comparison, the electronic time scale of $T \approx 24$ as is more than 100 times shorter. Electrons can react almost instantaneously to most changes in atomic positions and remain in the ground state for the given arrangement of of atoms $\mathbf{R}$. This is the *adiabatic approximation*.

Within the adiabatic approximation we solve the time-independent Schrödinger Equation at the atomic positions $\mathbf{R}$. Then, rather than propagate the equations with short time steps of order 24 as, we simply assume the electrons remain in the ground state and propagate the atomic positions with the much longer molecular dynamics time step of order 1 fs. The adiabatic approximation works well for most thermal catalysis studies. Nevertheless, similar to the classical ion approximation, there are some exceptions:

1. Photocatalysis. Optical excitations involve electronic transitions so they must be considered explicitly in the Schrödinger Equation. These excitations can yield very different effective potential energy surfaces than the electronic ground state and facilitate chemical reactions. For example, as discussed in Section 1.5.1, some excited-state compounds spontaneously dissociate.

2. Rapid wavefunction changes. In some cases the ground state electronic wavefunctions can be particularly sensitive to changes in atomic positions. Ground state wavefunctions can then change quickly on the times scales electronic motion, requiring explicit consideration of the electrons' real-time response to atomic motion. Section 1.5.2 discusses the simple case of H adsorption on metal surfaces, and Chapter 4 considers similar effects in H_2 dissociation.

1.1.4 BORN-OPPENHEIMER APPROXIMATION

The Born-Oppenheimer (BO) approximation is closely related to the classical ion approximation, and relies on the time independent Schrödinger Equation. Rather than consider the full wavefunction $\psi(\mathbf{r}, \mathbf{R}, t)$, the

BO approximation assumes that ψ can be separated into electronic and nuclear components:

$$\psi(\mathbf{r},\mathbf{R}) = \psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R}) \tag{1.3}$$

Similar to the discussion of the classical ion approximation, we note that the length scale for nuclear wavefunctions is much smaller than for electronic wavefunctions. Since both ψ_e and ψ_N are normalized, it follows that the nuclear *gradient* is much *larger* for the nuclear wavefunction than for the electronic wavefunction. In other words,

$$\nabla_I \psi_N(\mathbf{R}) \gg \nabla_I \psi_e(\mathbf{r};\mathbf{R}) \tag{1.4}$$

The nuclear kinetic energy terms are proportional to ∇_I^2 for each nucleus I, and are the only term to contain factors of ∇_I. Therefore we start by evaluating $\nabla_I^2\psi$, using the chain rule, then consider Equation (1.4):

$$\begin{aligned}
\nabla_I^2 \psi(\mathbf{r},\mathbf{R}) &\approx \nabla_I^2 \left(\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R})\right) && \text{(1.5a)}\\
&= \left(\nabla_I^2 \psi_e(\mathbf{r};\mathbf{R})\right) \psi_N(\mathbf{R}) + \psi_e(\mathbf{r};\mathbf{R}) \left(\nabla_I^2 \psi_N(\mathbf{R})\right) \\
&\quad + 2\nabla_I \psi_e(\mathbf{r};\mathbf{R}) \cdot \nabla_I \psi_N(\mathbf{R}) && \text{(1.5b)}\\
&\approx \psi_e(\mathbf{r};\mathbf{R}) \left(\nabla_I^2 \psi_N(\mathbf{R})\right) && \text{(1.5c)}
\end{aligned}$$

The last line, Equation (1.5c), is the core of the BO approximation. It follows from Equation (1.4) since this is the largest of the three terms from the chain rule.

Finally, consider the full Hamiltonian $\mathcal{H}$ from the Schrödinger Equation in Equation (1.1a). We rewrite $\mathcal{H}_K$ as $T_e + T_N$ to explicitly keep track of the nuclear kinetic energy. Moreover, we also rewrite the full Hamiltonian as $\mathcal{H} = T_e + T_N + \mathcal{H}_{\text{other}}$ to simplify our consideration of the BO approximation. Note that neither T_e nor $\mathcal{H}_{\text{other}}$ have terms containing ∇_I. We see how the BO approximation in Equation (1.5c) simplifies the Schrödinger Equation:

$$\begin{aligned}
E\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R}) &= \left(T_e + T_N + \mathcal{H}_{\text{rest}}\right)\left(\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R})\right) && \text{(1.6a)}\\
&= \left(T_e + \mathcal{H}_{\text{rest}}\right)\left(\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R})\right) + T_N\left(\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R})\right) && \text{(1.6b)}\\
&\approx \psi_N(\mathbf{R})\left(T_e + \mathcal{H}_{\text{rest}}\right)\psi_e(\mathbf{r};\mathbf{R}) + \psi_e(\mathbf{r};\mathbf{R})T_N\psi_N(\mathbf{R}) && \text{(1.6c)}
\end{aligned}$$

In Equation (1.6c), the only approximation used is the one discussed from Equation (1.5c). We also use the fact that $T_e + \mathcal{H}_{\text{rest}}$ has no ∇_I operators to rearrange the first term on the right-hand side. Finally, we divide both sides by $\psi_e(\mathbf{r};\mathbf{R})\psi_N(\mathbf{R})$:

$$E = \frac{(T_e + \mathcal{H}_{\text{rest}})\,\psi_e(\mathbf{r};\mathbf{R})}{\psi_e(\mathbf{r};\mathbf{R})} + \frac{T_N\,\psi_N(\mathbf{R})}{\psi_N(\mathbf{R})} \tag{1.6d}$$

The final expression, Equation (1.6d), is then explicitly separable. Rather than solving for a combined electronic and nuclear wavefunction, with the BO approximation we can instead find separate solutions for ψ_e and ψ_N.

Inspecting the electronic contribution to Equation (1.6d) recovers the classical ion approximation. Since the first term on the right-hand side is the only one depending on $\mathbf{r}$, it must be a constant with respect to $\mathbf{R}$. For nefarious reasons we can let this constant be $E_g(\mathbf{R})$. This choice of notation yields the same time-independent Schrödinger Equation resulting from the classical ion approximation:

$$(T_e + \mathcal{H}_{\text{rest}})\,\psi_e(\mathbf{r};\mathbf{R}) = E_g(\mathbf{R})\,\psi_e(\mathbf{r};\mathbf{R}) \tag{1.7}$$

In addition, once $E_g(\mathbf{R})$ is known, then the nuclear wavefunction can be found using Equation (1.6d):

$$T_N\,\psi_N(\mathbf{R}) = (E - E_g(\mathbf{R}))\,\psi_N(\mathbf{R}) \tag{1.8}$$

In principle, the BO approximation is an improvement over the classical ion approximation since it also provides ψ_N. However, in practice the BO approximation is almost always only invoked to justify the classical ion approximation. In these cases the BO approximation is equivalent to making both the adiabatic and classical ion approximations.

1.1.5 POTENTIAL ENERGY SURFACES

Solutions for $E_g(\mathbf{R})$ are known as *potential energy surfaces*. This is because E_g only varies with the atomic positions, i.e. a scalar function of $3N$ coordinates for N atoms. Thermodynamics and transition state theory, including saddle points and local minima, involve only the total energies and forces. As discussed previously, these energies are well approximated by solving the electron-only Schrödinger Equation from Equation (1.7). Moreover, the BO approximation is widely used in molecular dynamics since atomic forces are easily obtained through Newton's equations as $F_I = -\nabla_I E_g(\mathbf{R})$. Rather than explicitly consider the complex electronic wavefunction, we only need the energy E_g and its gradients $\nabla_I E_g$ with respect to the atomic positions. Knowing E_g alone is therefore sufficient for calculating catalytic rates, activation energies, and vibrational properties.

1.2 DENSITY FUNCTIONAL THEORY

Equation (1.7) highlights a simple yet surprisingly important fact about systems in the adiabatic and classical ion limits: the ground state wavefunction directly corresponds to the ground state energy. This equivalence suggests that a dramatic reduction in dimensionality is possible: for a given set of atomic positions $\mathbf{R}$ and n electrons, the $3n$-dimensional $\psi_g(\mathbf{r})$ directly corresponds to *scalar* E_g. In light of Section 1.1.5, we see that knowing E_g alone is still sufficient to obtain catalytic insight. Does a low-dimensional shortcut to E_g actually exist?

Fortunately the answer is "yes." First proven by Hohenberg and Kohn in their eponymous Hohenberg-Kohn (HK) theorems,[62] it can be shown using the variational principle that the ground state density is a *functional* of the total electron density $n(\mathbf{r})$:

$$E_g(\mathbf{R}) = E_g(\mathbf{R})\,[n_g(\mathbf{r})] \tag{1.9}$$

Rather than solve for the $3n$-dimensional wavefunction ψ_g, the HK theorems prove that we only really need the 3-dimensional total electron density. Kohn was awarded half of the 1998 Nobel Prize in Chemistry for this work.

Later work by Levy[97] provides a cleaner and more general proof for density functional theory as it is used today. Instead of repeating the proofs, here I provide a "derivation" of DFT in the same spirit of Levy's work. Neglecting the nuclear kinetic energy and spin-orbit coupling terms, the time-independent Schrödinger Equation can be rewritten as follows:

$$E\,\psi(\mathbf{r};\mathbf{R}) = \left(T_e + \mathcal{H}_{\text{ion}} + V_{\text{rest}}(\mathbf{r},\mathbf{R})\right)\psi(\mathbf{r};\mathbf{R}) \tag{1.10a}$$

$\mathcal{H}_{\text{ion}}$ contains electron-electron repulsion, electron-ion attraction, and ion-ion repulsion. Since the latter depends only on the atomic positions $\mathbf{R}$, it is a constant $C(\mathbf{R})$ that can be added to E after solving the electronic Schrödinger Equation. This leaves the following quantities:

$$= \left(T_e + \sum_{i<j} \frac{e^2}{r_{ij}} - \sum_{Ij} \frac{Z_I e^2}{r_{Ij}} + V_{\text{rest}}(\mathbf{r}, \mathbf{R}) \right) \psi(\mathbf{r}; \mathbf{R}) \tag{1.10b}$$

$$= \left(T_e + \sum_{i<j} \frac{e^2}{r_{ij}} - \sum_{j} \left(\sum_{I} \frac{Z_I e^2}{r_{Ij}} \right) + V_{\text{rest}}(\mathbf{r}, \mathbf{R}) \right) \psi(\mathbf{r}; \mathbf{R}) \tag{1.10c}$$

This rearrangement demonstrates that the Schrödinger Equation (aside from nuclear kinetic energy and spin-orbit interactions) contains only one and two-body operators. Similar to neglecting the repulsion between ions, we also set aside the terms in $V_{\text{rest}}(\mathbf{r}, \mathbf{R})$ that only affect the ions since they can be calculated after solving the electronic Schrödinger Equation as well. We are then left with $V_{\text{rest}}(\mathbf{r}, \mathbf{R}) \rightarrow V_{\text{rest}}(\mathbf{r})$. Electrons are indistinguishable particles so they must all experience the same potential operator $V_{\text{rest},i}$. Therefore $V_{\text{rest}} = \sum_i V_{\text{rest},i}$ is a single-body operator along with T_e and electron-ion interaction. For simplicity we also assume that these are *local* operators, such that $V_{\text{rest},i} \leftrightarrow V_{\text{rest}}(\mathbf{r}_i)$. Most operators of catalytic interest are local or at least approximately local (e.g. van der Waals interactions). For convenience we collect the one-body operators into $\mathcal{H}_1$, one-body Hamiltonian:

$$\mathcal{H}_1 \equiv T_e - \sum_i \sum_J \frac{Z_J e^2}{r_{iJ}} + \sum_i V_{\text{rest},i} \tag{1.11}$$

The last remaining term in the full Hamiltonian, electron-electron interaction, is a two-body operator. Defining this operator as V_{ee}, we obtain the following form for the electronic part of the Schrödinger Equation:

$$E \psi(\mathbf{r}; \mathbf{R}) \psi(\mathbf{r}; \mathbf{R}) = (\mathcal{H}_1 + V_{\text{ee}})\, \psi(\mathbf{r}; \mathbf{R}) \tag{1.12}$$

With the definition of Equation (1.12) the "derivation" of DFT is surprisingly straightforward. As discussed previously, we seek the electronic total ground state energy E_g instead of the wavefunction. In the spirit of the HK theorems and Levy's work, we use the variational principle:

$$E_g = \min_{\psi} \left\langle \psi \middle| \mathcal{H}_1 + V_{\text{ee}} \middle| \psi \right\rangle \tag{1.13a}$$

$$= \min_{\psi} \left(\int \mathrm{d}\mathbf{r}_1 \cdots d\mathbf{r}_n\, \psi^* \mathcal{H}_1\, \psi + \left\langle \psi \middle| V_{\text{ee}} \middle| \psi \right\rangle \right) \tag{1.13b}$$

$$= \min_{\psi} \left(\int \mathrm{d}\mathbf{r} \lim_{\mathbf{r}' \to \mathbf{r}} d\mathbf{r}_2 \cdots d\mathbf{r}_n \mathcal{H}_1 \psi(\mathbf{r}', \mathbf{r}_2 \cdots \mathbf{r}_n)^* \psi(\mathbf{r}, \mathbf{r}_2 \cdots \mathbf{r}_n) + \left\langle \psi \middle| V_{\text{ee}} \middle| \psi \right\rangle \right) \tag{1.13c}$$

Introducing the limit lets us replace the integral over the $n-1$ degrees of freedom *not* acted on by the single-body operators with the first-order density matrix γ.

$$\equiv \min_{\psi} \left(\int d\mathbf{r} \lim_{\mathbf{r}' \to \mathbf{r}} \mathcal{H}_1 \gamma(\mathbf{r}'|\mathbf{r}) + \langle \psi | V_{\text{ee}} | \psi \rangle \right) \tag{1.13d}$$

Rather than minimize over all wavefunctions directly, we can instead minimize over all wavefunctions ψ^γ that correspond to a given first-order density matrix γ, then minimize over all first-order density matrices. This is mathematically equivalent, but conceptually important.

$$\equiv \min_{\gamma} \min_{\psi^\gamma \to \gamma} \left(\int d\mathbf{r} \lim_{\mathbf{r}' \to \mathbf{r}} \mathcal{H}_1 \gamma(\mathbf{r}'|\mathbf{r}) + \langle \psi^\gamma | V_{\text{ee}} | \psi^\gamma \rangle \right) \tag{1.13e}$$

$$\equiv \min_{\gamma} \left(\int d\mathbf{r} \lim_{\mathbf{r}' \to \mathbf{r}} \mathcal{H}_1 \gamma(\mathbf{r}'|\mathbf{r}) + \min_{\psi^\gamma \to \gamma} \langle \psi^\gamma | V_{\text{ee}} | \psi^\gamma \rangle \right) \tag{1.13f}$$

We now define $W[\gamma] \equiv \min_{\psi^\gamma \to \gamma} \langle \psi^\gamma | V_{\text{ee}} | \psi^\gamma \rangle$. $W[\gamma]$ is a functional that represents the minimum possible Coulomb repulsion for a pure state corresponding to γ. Finally, in general any single-particle density matrix can be written as a sum over single-particle orbitals ψ_i as $\gamma(\mathbf{r}'|\mathbf{r}) = \sum_i n_i \psi_i(\mathbf{r}')^* \psi_i(\mathbf{r})$. Minimizing over γ is therefore equivalent to minimizing over all n_i and ψ_i. With this final replacement, we obtain reduced density matrix functional theory (RDMFT):

$$E_g = \min_{n_i, \psi_i} \left(\sum_i n_i \left\langle \psi_i \left| \frac{\hbar^2}{2m} \nabla^2 + \sum_I \frac{Z_I e^2}{|\mathbf{r} - \mathbf{R}_I|} + V_{\text{rest}}(\mathbf{r}) \right| \psi_i \right\rangle + W[n_i, \psi_i] \right) \tag{1.13g}$$

The corresponding density $\rho(\mathbf{r})$ is easily written in terms of n_i and ψ_i:

$$\rho(\mathbf{r}) = \sum_i n_i |\psi_i(\mathbf{r})|^2 \tag{1.14}$$

Symbolically this completes RDMFT: the total electronic energy E_g is a functional of the first order density matrix. Instead of solving for the $3n$-dimensional wavefunction, we can obtain E_g by minimizing over 3-dimensional single-particle orbitals $\psi_i(\mathbf{r})$ and occupation numbers n_i instead.

Although Equation (1.13g) is exact in principle, in practice the functional W is unknown. This is because the coulomb repulsion term V_{ee} is a two-body operator. While its expectation value can be written exactly in terms of $\gamma^{(2)}(\mathbf{r}, \mathbf{r}'|\mathbf{r}, \mathbf{r}')$, the second-order density matrix, the space of all $\gamma^{(2)}$ matrices corresponding to a system n electrons is unknown.[125] Moreover, even if it were known, this domain is likely much larger than the space of all first-order density matrices. The lack of $\gamma^{(2)}$ is a problem because V_{ee} is a significant contribution to E_g, so even small relative errors could be large on the scale of k_BT at typical catalytic temperatures.

Fortunately, the largest contribution to W in Equation (1.13g) is the Hartree term $V_H(\mathbf{r})$ from Hartree Fock theory, and this term can be written exactly in terms of $\gamma^{(1)}$. We can then rewrite Equation (1.13g) in terms of $V_H(\mathbf{r})$ and the remaining part of the functional W'.

$$W[\gamma] \equiv \int \mathrm{d}\mathbf{r}\, d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + W'[\gamma] \tag{1.15a}$$

$$V_H(\mathbf{r}) \equiv \int \mathrm{d}\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \tag{1.15b}$$

$$E_g = \min_{n_i,\psi_i} \left(\sum_i n_i \left\langle \psi_i \left| \frac{\hbar^2}{2m}\nabla^2 + \sum_I \frac{Z_I e^2}{|\mathbf{r}-\mathbf{R}_I|} + V_H(\mathbf{r}) + V_{\text{rest}}(\mathbf{r}) \right| \psi_i \right\rangle + W'[n_i,\psi_i] \right) \tag{1.15c}$$

Introducing $V_H(\mathbf{r})$ highlights that the Hartree term is single-body operator that can be included with the others introduced before.

Although RDMFT is already a major simplification, $\gamma(\mathbf{r}|\mathbf{r}')$ is a *six*-dimensional quantity while the density $\rho(\mathbf{r})$ is only a *three*-dimensional quantity. Moving to density functional theory, i.e. using ρ as the fundamental quantity instead of γ, yields a significant reduction in computational effort so that larger systems can be studied. To accomplish this, we consider a system of *noninteracting* particles instead, whose corresponding density $\rho(\mathbf{r})$ minimizes the total energy $E_g[\rho]$. The wavefunction of a noninteracting system is a single Slater determinant, whose occupation numbers are either 0 or 1 (Slater determinants either do or don't contain a particular single-particle wavefunction). The sacrifice we make is that the single-particle states ψ_i no longer have any connection to actual electronic states in principle.* Instead of minimizing over n_i and ψ_i quantities in DFT, we minimize over $\rho(\mathbf{r})$:

$$E_g = \min_{\rho} \left(\sum_i \left\langle \psi_i \left| \frac{\hbar^2}{2m}\nabla^2 + \sum_I \frac{Z_I e^2}{|\mathbf{r}-\mathbf{R}_I|} + V_{\text{rest}}(\mathbf{r}) \right| \psi_i \right\rangle + E_{xc}[\rho] \right) \tag{1.16a}$$

$$= \min_{\rho} \left(\sum_i \left\langle \psi_i \left| \frac{\hbar^2}{2m}\nabla^2 \right| \psi_i \right\rangle + \int \mathrm{d}\mathbf{r}\rho(\mathbf{r}) \left(\sum_I \frac{Z_I e^2}{|\mathbf{r}-\mathbf{R}_I|} + V_{\text{rest}}(\mathbf{r}) \right) + E_{xc}[\rho] \right) \tag{1.16b}$$

$$\rho(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2 \tag{1.16c}$$

DFT is similar to RDMFT, and is still exact in principle. The first two terms in Equation (1.16b) are especially similar since only the n_i factors have been removed. The subtle difference is that eliminating the

---

*In practice the ψ_i can usually be interpreted as single-particle orbitals. They appear to be meaningful whenever molecular orbital theory applies.

n_i factors breaks the connection between the KS states ψ_i and γ, so their corresponding kinetic energy values are only approximate. This approximation is generally excellent. The remaining contribution to kinetic energy, as well as $V_{ee} - V_H$, is now handled by the *exchange-correlation functional* $E_{\text{xc}}[\rho]$. E_{xc} contributes only a few percent of the total energy but it describes most of chemical bonding.

Computational implementations of DFT re-cast Equation (1.16b) as a linear system of equations. To do so, we define the *exchange-correlation potential* V_{xc} as the functional derivative of $E_{\text{xc}}[\rho(\mathbf{r})]$ with respect to $\rho(\mathbf{r})$:

$$V_{\text{xc}}(\mathbf{r}) \equiv \left.\frac{\delta E_{\text{xc}}[\rho]}{\delta \rho}\right|_{\mathbf{r}} \tag{1.17}$$

The method of Lagrange multipliers, to enforce the normality condition $\langle \psi_i | \psi_i \rangle = 1$, then produces the following linear system of equations:

$$\mathcal{H}_{\text{DFT}}[\rho]\,|\psi_i\rangle = \varepsilon_i |\psi_i\rangle \tag{1.18}$$

$$\mathcal{H}_{\text{DFT}}[\rho] \equiv \left(-\frac{\hbar^2}{2m}\nabla^2 + \sum_I \frac{Z_I e^2}{|\mathbf{r} - \mathbf{R}_I|} + V_{\text{rest}}(\mathbf{r}) + V_{\text{xc}}[\rho](\mathbf{r}) \right) \tag{1.19}$$

Equations (1.18) (one for each ψ_i) are known as the Kohn-Sham (KS) equations,[84] introduced by Kohn and Sham shortly after publication of the Hohenberg-Kohn theorems. Although it appears that these are mean-field equations at first glance, an exact expression for V_{xc} retains all of the original system's interactions. The KS equations are solved self-consistently using standard matrix diagonalization routines to obtain the KS wavefunctions ψ_i in some convenient basis set (see Section 1.2.2 for details).

1.2.1 EXCHANGE-CORRELATION FUNCTIONALS

DFT's accuracy for catalytic studies rests entirely on E_{xc}. Its exact functional form is unknown, but there are significant ongoing efforts to create and improve approximations to it. The earliest functional is the *local density approximation* (LDA), which uses the exact expression for the homogeneous electron gas.[62] It is surprisingly effective given its simplicity, but LDA significantly overestimates bond energies.[24] The next generation of functionals include gradients of the density, yielding the *generalized gradient approximation* (GGA). One of the most popular functionals in solid state physics is the GGA functional of Perdew, Burke, and Ernzerhof,[138] published in 1996. It is simply known as "the PBE functional," and it is used in all work featured in this thesis despite being developed over 20 years ago. PBE has remained at the forefront of physics since it is among the most accurate *local functionals*, where $V_{\text{xc}}[\rho](\mathbf{r})$ only depends on $\mathbf{r}$ and the density at that point $V_{\mathrm{xc}}(\mathbf{r}, \rho(\mathbf{r}), \nabla\rho(\mathbf{r}), \cdots)$. This form is approximate because the real $V_{\mathrm{xc}}[\rho](\mathbf{r})$ is a functional of ρ *everywhere*, not just at a particular $\mathbf{r}$ of interest, but considering this extended behavior requires computationally expensive integrations. PBE is an excellent tradeoff between accuracy and speed. As local functionals, LDA and GGA are comparatively inexpensive but fail to correctly treat long-range interactions and suffer from self-interaction error.[24]

Most significant improvements to PBE in general require making V_{xc} *nonlocal*, requiring nested integrations in Equations (1.16b) and (1.18). For example, *hybrid functionals* include a fraction of the Hartree-Fock exchange energy, generally referred to as "exact exchange" despite using KS states as if they were true single-electron states. Evaluating the exchange interaction requires pair-wise integrals over all pairs of KS states ψ_i and ψ_j, significantly increasing computational costs. The most popular hybrid functional is B3LYP,[4,95,157,178] a weighted combination of Hartree-Fock exchange with exchange and correlation functionals from both LDA and GGA.[157] Another route to significant improvements in some cases is to use *empirical functionals*, which are specifically fitted to a dataset of known quantities. Fitting is a controversial process because parameters that minimize error for one type of system or set of molecules can yield increased errors for other systems.[†] Another recent approach is to include the local kinetic energy density along with the local density and its gradients, resulting in *meta-GGA* functionals. A recent and popular example is the SCAN functional,[161] which appears to offer significant and systematic improvements to PBE while remaining computationally inexpensive. Part of meta-GGAs' successes are their effective nonlocality; the kinetic energy density is a function of the KS eigenstates which vary across the whole system. Meta-GGAs have been shown to include some of the long-range interactions GGA functionals lack, in this sense they have been described as having "ultranonlocality."[127]

1.2.2 BASIS SETS AND SUPERPOSITION ERROR

Computational implementations of DFT must express the continuous functions $\psi_i(\mathbf{r})$ with a finite collection of numbers. In practice this is done by defining a *basis set*, a collection of primitive functions such as plane waves or Gaussians, and then representing $\psi_i(\mathbf{r})$ as a linear combination of these functions. The mathematical approach follows:

$$\psi_i(\mathbf{r}) = \sum_{\mu} c_{i\mu} \phi_\mu(\mathbf{r}) \tag{1.20a}$$

$$\psi_i^T = (c_{i1}, c_{i2}, \cdots) \tag{1.20b}$$

[†] It is likely that a version of the "No Free Lunch" theorems in Chapter 5 applies to empirical potentials.

For finite basis sets it is generally simpler to consider the vector $\boldsymbol{\psi}_i$ of coefficients rather than the actual function $\psi_i(\mathbf{r})$.

After applying the "Fourier trick" with respect to each of the basis orbitals, the KS states described by Equations (1.18) can be obtained by solving a generalized eigenvalue equation:

$$H_{\text{DFT}}\left[\rho\right]\boldsymbol{\psi}_i = \varepsilon_i S \boldsymbol{\psi}_i \tag{1.21a}$$

$$\left(H_{\text{DFT}}\left[\rho\right]\right)_{\mu\nu} = \int \mathrm{d}\mathbf{r}\, \phi_\mu^*(\mathbf{r}) \mathcal{H}_{\text{DFT}}\, \phi_\nu(\mathbf{r}) \tag{1.21b}$$

$$S_{\mu\nu} = \int \mathrm{d}\mathbf{r}\, \phi_\mu^*(\mathbf{r}) \phi_\nu(\mathbf{r}) \tag{1.21c}$$

$$\rho = \sum_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^\dagger \tag{1.21d}$$

where ε_i is the Lagrange multiplier associated with KS state $\psi_i(\mathbf{r})$. To distinguish matrices from operators I use script letters for operators and capital letters for matrices. Equation (1.21a) is a generalized eigenvalue problem. At each iteration of a DFT calculation, Equation (1.21a) is solved for the new vectors $\boldsymbol{\psi}_i$ and eigenvalues ε_i. The new vectors produce a new density matrix ρ. This process is repeated until *self-consistent*, such that ρ does not change from one iteration to the next. The ε_i values have no inherent physical meaning, but similar to the KS states themselves, in practice they can interpreted as the energies of single electron states. In fact, DFT band structures are simply plots of ε_i for all KS states along high-symmetry directions in the Brillouin zone.

An important consideration in practical DFT calculations is highlighted by Equations (1.20a) and (1.21d): the total density $\rho(\mathbf{r})$ is constrained by the choice of basis functions $\phi_\mu(\mathbf{r})$. For example, $\rho(\mathbf{r}')$ is zero if $\phi_\mu(\mathbf{r}') = 0$ for all orbital ϕ_μ. In general we can describe a particular basis set by a collection of coefficients $\boldsymbol{\beta}$, perhaps describing a real-space grid or the number and shape of orbitals. Minimizing with respect to density in Equation (1.16b) now involves optimizing both the coefficients $c_{i\mu}$ *and* the basis set itself through $\boldsymbol{\beta}$:

$$E_g = \min_{\boldsymbol{\beta}} \min_{c_{i\mu}|\boldsymbol{\beta}} \sum_i \left\langle \psi_i \left| \frac{\hbar^2}{2m}\nabla^2 \right| \psi_i \right\rangle + \int \mathrm{d}\mathbf{r} \rho(\mathbf{r}) \left(\sum_I \frac{Z_I e^2}{|\mathbf{r} - \mathbf{R}_I|} + V_{\text{rest}}(\mathbf{r}) \right) + E_{xc}\left[\rho\right] \tag{1.22}$$

where we now explicitly minimize over basis set coefficients $c_{i\mu}$ and the basis set itself through $\boldsymbol{\beta}$.

While all DFT codes readily perform the "inner" optimization over the coefficients $c_{i\mu}$, almost none of them automatically perform the "outer" optimization with respect to the basis set itself. Poor basis sets constrain the density and systematically overestimate E_g. This systematic error leads to *basis set superposition error* (BSSE), where this "basis constraint" is mitigated by orbitals associated with another part of the system. For example, a poor basis for a molecule might be too inflexible to correctly charge density near a bond. When this molecule is brought close to a surface, orbitals located at the surface can provide additional flexibility and thus stabilize the bond following the variational principle, reducing the total energy. This decrease in total energy is indistinguishable from adsorption energy so BSSE systematically overestimates adsorption energy. BSSE effects are mitigated by optimizing the basis set itself; a fictitious enthalpy method[37] for this optimization is discussed in Appendix A.

Choosing an appropriate basis set is critical. There are three popular choices:

1. Plane waves: completely delocalized states which prevent BSSE. Energy and ρ are systematically improved by using more plane waves, although many high-frequency plane waves are needed to describe localized bonds. Plane waves are equivalent to uniform sampling by Fourier transforms so vacuum regions are treated with high precision, wasting computational effort.

2. Pseudoatomic orbitals (PAOs): DFT solutions for isolated atoms. Few PAOs are needed to describe localized bonds since they are perturbations of atomic orbitals. Atom-centered basis sets also lead to minimal wasted effort in vacuum regions. However, poorly-chosen orbitals can lead to significant BSSE, and there is no obvious systematic way to add additional basis functions.

3. Gaussians: Gaussian functions scaled by polynomials. They are not solutions for isolated atoms and are therefore more susceptible to BSSE than PAOs. Their mathematical form provides analytic overlap and exchange integrals, significantly reducing computational effort for exact exchange and post-Hartree-Fock methods.

Plane waves eliminate BSSE and are easy to converge so they are the best choice when computational resources are abundant. In this thesis plane wave calculations are used to understand the physical and catalytic properties of iodine-doped graphene in Chapter 2, the mechanism of anhydrous alcohol dehydrogenation on Cu(111) surfaces in Chapter 3, and the modeling of H adsorption energies on bimetallic Ag alloys in Chapter 5.

However, plane waves can become prohibitively expensive when studying large surfaces or nanoclusters. Part of the reason is that most practical DFT calculations employ periodic boundary conditions to facilitate computing the Hartree potential $V_H(\mathbf{r})$ from $\rho(\mathbf{r})$. This is because fast Fourier transforms are usually the fastest way to compute $V_H(\mathbf{r})$:

$$\nabla^2 V_H(\mathbf{r}) = 4\pi\rho(\mathbf{r}) \tag{1.23a}$$

$$-k^2 V_H(\mathbf{k}) = 4\pi\rho(\mathbf{k}) \tag{1.23b}$$

where the differential equation in Cartesian coordinates is algebraic in Fourier space $\mathbf{k}$. Equation (1.23b) is only valid if V_H and ρ are periodic, but surfaces, molecules, and other low-dimensionality structures are aperiodic along one or more dimensions. The common solution is to pad the non-periodic dimension(s) with *vacuum space* to minimize spurious interactions among the periodic images implied by Equation (1.23b). Typical vacuum distances for surfaces with small dipole moments are 10 to 20 Å, while the surface itself is typically less than 15 Å thick. Using plane waves is equivalent to a uniform mesh over the whole unit cell through Fourier transforms, so plane waves expend significant computational effort on vacuum regions that contribute little to the total energy.

Another major issue with plane waves is solving Equation (1.21a) for the KS states. Catalytic studies generally involve covalent bonds that must be described using many high-frequency plane waves. Plane wave basis sets therefore tend to be large, with many thousands of Fourier mode basis functions. Solving for the KS states involves diagonalizing H_{DFT}, and the runtime scales as $O\left(m^3\right)$ for m basis functions. Plane waves can be exceedingly expensive for more than about 100 atoms with substantial vacuum space.

Both the vacuum space and large matrix problems are mitigated using PAOs and Gaussians. Interatomic bonding (even in metals) is well described through perturbation theory, where neighboring atoms perturbs the spherically-symmetric atomic potential $\frac{Z_I e}{|\mathbf{r}-\mathbf{R}_I|}$. Reasonable basis functions $\phi_\mu(\mathbf{r})$ can then be constructed as $R_\ell(r)Y_{\ell m}(\hat{\mathbf{r}})$, i.e. hydrogen-like orbitals. PAO basis sets use $R_\ell(r)$ obtained from performing DFT calculations on isolated atoms, while Gaussian basis sets use a linear combination of several Gaussian functions scaled by polynomials. Hydrogen-like wavefunctions asymptotically decay as $r^\ell \exp(-\alpha r)$. A single PAO suffices for an isolated atom because it is the numerical solution, and only 23 total PAOs per atom are needed to obtain quality band structures for the bulk phases of most transition metals.[37,156] In contrast, superpositions of several Gaussians must be used for even an isolated atom because they have the wrong short and long-range asymptotic behaviors. Gaussian basis sets must be larger for the same accuracy at a given level of theory. Therefore PAOs are better than Gaussians for typical DFT calculations. On the other hand, exact exchange and (post-)Hartree-Fock calculations require extensive evaluation of overlap and exchange integrals. Gaussian orbitals yield analytic expressions for these integrals and are therefore faster for quantum chemistry calculations.

1.2.3 Electron Spin and Spin-Orbit Coupling

In the nonrelativistic Schrödinger Equation there is no explicit dependence on electron spin so spin commutes with the Hamiltonian. The same true for the DFT Hamiltonian as well. KS states are therefore joint eigenstates of the spin operator along some axis and $\mathcal{H}_{\text{DFT}}$, resulting in *spin orbitals* $\psi_{i\sigma}(\mathbf{r})|\sigma\rangle$ where σ indicates the spin state. The noninteracting KS "particles" are fermions just like real electrons so σ represents one of two spin-$\frac{1}{2}$ eigenstates usually referred to as "up" and "down." In addition, the spatial dependence $\psi_{i\sigma}$ is separate from the spin state $|\sigma\rangle$. In light of these considerations, such solutions are referred to as "colinear spin" since all spin directions are either "up" or "down" along some chosen spin axis. All KS states have either up or down character, so the total density can also be written as the sum of spin densities:

$$\rho(\mathbf{r}) = \rho_{\uparrow}(\mathbf{r}) + \rho_{\downarrow}(\mathbf{r}) = \sum_i \left|\psi_{i\uparrow}(\mathbf{r})\right|^2 + \sum_i \left|\psi_{i\downarrow}(\mathbf{r})\right|^2 \tag{1.24}$$

Since $\mathcal{H}_{\text{DFT}}$ and S_a commute, despite the up and down labels, it should be noted that the axis a is arbitrary–it has no relationship with the coordinate system chosen for $\mathbf{r}$.

Despite not having explicit spin dependence, $\mathcal{H}_{\text{DFT}}$ has *implicit* spin dependence. This is because $V_{\text{xc}}[\rho]$ includes spin-dependent exchange interactions, which vanish between any two KS states with opposite spin:

$$\left\langle \psi_{i\sigma}\psi_{j\sigma'} \left| \frac{1}{|\mathbf{r}-\mathbf{r}'|} \right| \psi_{j\sigma'}\psi_{i\sigma} \right\rangle = \int \mathrm{d}\mathbf{r}\, d\mathbf{r}'\, \frac{\psi^*_{i\sigma}(\mathbf{r})\psi^*_{j\sigma'}(\mathbf{r}')\psi_{j\sigma'}(\mathbf{r})\psi_{i\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \langle\sigma|\sigma'\rangle\langle\sigma'|\sigma\rangle \tag{1.25a}$$

$$= \int \mathrm{d}\mathbf{r}\, d\mathbf{r}'\, \frac{\psi^*_{i\sigma}(\mathbf{r})\psi^*_{j\sigma'}(\mathbf{r}')\psi_{j\sigma'}(\mathbf{r})\psi_{i\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \delta_{\sigma'\sigma}\delta_{\sigma\sigma'} \tag{1.25b}$$

If σ and σ' differ, then the states' exchange interaction vanishes. Functionals are therefore parameterized on the spin densities, i.e. $V_{\text{xc}}[\rho] = V_{\text{xc}}[\rho_{\uparrow}, \rho_{\downarrow}]$, to account for this.

For the single-particle equations and basis functions, spin is included by "splitting" each of the m spatial basis functions $\phi_\mu(\mathbf{r})$ into two spin orbitals $\phi_\mu(\mathbf{r})|\uparrow\rangle$ and $\phi_\mu(\mathbf{r})|\downarrow\rangle$. If these $2m$ total orbitals are grouped by spin, then $\boldsymbol{\psi}_{i\sigma}$ and H_{DFT} can be written as a collection of "spin blocks:"

$$\boldsymbol{\psi}_{i\sigma} = \left(c_{i1\uparrow}, \cdots, c_{im\uparrow}, c_{i1\downarrow}, \cdots, c_{im\downarrow}\right)^T \tag{1.26a}$$

$$H_{\text{DFT}} = \begin{pmatrix} H^{\uparrow\uparrow} & H^{\uparrow\downarrow} \\ H^{\downarrow\uparrow} & H^{\downarrow\downarrow} \end{pmatrix} \tag{1.26b}$$

$$H^{\sigma\sigma'}_{\mu\nu} \equiv \left\langle \phi_{\mu\sigma} \left| \mathcal{H}_{\text{DFT}} \right| \phi_{\nu\sigma'} \right\rangle \tag{1.26c}$$

The nonrelativistic Hamiltonian commutes with spin so the off-diagonal spin blocks $H^{\uparrow\downarrow}_{\text{DFT}}$ and $H^{\downarrow\uparrow}_{\text{DFT}}$ vanish:

$$H^{\sigma\sigma'} = H^{\sigma\sigma}\delta_{\sigma\sigma'} \tag{1.27}$$

Off-diagonal spin blocks in the overlap matrix, e.g. $S^{\uparrow\downarrow}$, vanish as well. The eigenvalues and eigenvectors within each block of a block-diagonal matrix are entirely decoupled. KS states for each spin direction can then be obtained independently. Magnetism and spin polarization in colinear spin systems result from the spin-dependent exchange iterations, leading to $H^{\uparrow\uparrow} \neq H^{\downarrow\downarrow}$ and thus different KS states.

Matters become more complicated with the introduction of relativistic effects. In solid-state and catalytic studies, the most important relativistic effect is spin-orbit coupling where electrons' intrinsic spin interacts with orbital angular momentum. The general expression for spin-orbit coupling is given in Equation (1.1d) and repeated here for the case of a single particle:

$$\mathcal{H}_{\mathrm{SOC}} = -\frac{1}{2m^2c^2}\,\mathbf{s}\cdot(\nabla V(\mathbf{r}) \times \vec{p}) \tag{1.28}$$

where $\mathbf{s}$ is the vector of Pauli matrices $\mathbf{s} = (\sigma_x, \sigma_y, \sigma_z)$, and $V(\mathbf{r})$ is the total potential operating on the electron. This includes the potential from ions and other electrons.

Although evaluating the gradient $\nabla V(\mathbf{r})$ for all contributions to $V(\mathbf{r})$ is complicated, most contributions result in mild gradients and therefore very small corrections. For example, the electron-electron repulsion contribution is small because the length scale of electronic wavefunctions, about 0.5 Å, produces small gradients. Similarly, the Rashba effect at surfaces contribute only small gradients as well. An order of magnitude estimate of the gradient is the work function of a surface divided by an angstrom, yielding electric fields on the order of 1 eV/Å. This leads to a typical contribution on the order of only 1 μeV. By far the largest contribution in catalytic materials is the ionic potential. For each ion I, its contribution to the potential $\frac{Z_I e^2}{|\mathbf{r}-\mathbf{R}_I|}$ diverges as $\mathbf{r}$ approaches $\mathbf{R}_I$, and its gradient diverges as well. In these cases the spin-orbit contribution to $\mathcal{H}$ can rise to the order of several electron volts for sufficiently heavy elements like Au.

The total potential $V(\mathbf{r})$ in Equation (1.28) can therefore be approximated simply as a $\frac{1}{r}$ potential if we define $\mathbf{R}_I$ as the origin. The gradient of this potential is directed along $\hat{\mathbf{r}}$ in spherical coordinates. With this approximation, we obtain the most common form of spin-orbit coupling:

$$\begin{aligned}
\mathcal{H}_{\mathrm{SOC}} &\propto \mathbf{s}\cdot(\nabla V(\mathbf{r}) \times \mathbf{p}) && \text{(1.29a)}\\
&\approx \mathbf{s}\cdot\left(\nabla\frac{Z_I e^2}{r} \times \mathbf{p}\right) && \text{(1.29b)}\\
&\propto -\mathbf{s}\cdot\left(\frac{1}{r^2}\hat{\mathbf{r}} \times \mathbf{p}\right) && \text{(1.29c)}\\
&\propto -\frac{1}{r^2}\mathbf{s}\cdot\mathbf{L} && \text{(1.29d)}
\end{aligned}$$

where $\mathbf{L}$ is the electron angular momentum operator. The last equation, with $\mathbf{L}\cdot\mathbf{s}$, reveals why this approximation is known as *LS coupling*. DFT codes include spin-orbit coupling by including the matrix elements of the LS operator in H_{DFT}. Localized basis sets almost always use products of radial and angular wavefunctions, i.e. $\phi_\mu(\mathbf{r}) = R_\mu(r)Y_\mu(\hat{\mathbf{r}})$, leading to simple expressions for the spin-orbit matrix elements $(H_{\mathrm{SOC}})_{\mu\nu}$.[33]

Spin-orbit coupling can have important implications for solid-state physics and surface catalysis. The first is spin splitting in surface states, where LS corrections of order 1 eV can significantly alter the energies and characters of states involved in adsorbate-surface bonding. Another is introducing more complex magnetic order. This arises because the $\mathbf{L}\cdot\mathbf{s}$ operator now explicitly couples electron spin to the atoms' valence orbitals, and the off-diagonal spin blocks $H^{\uparrow\downarrow}$ and $H^{\downarrow\uparrow}$ in Equation (1.26b) no longer vanish. The KS eigenvalue problem (Equation (1.21a)) now has complex-valued eigenvectors with nonzero coefficients for both $\phi_{\mu\uparrow}(\mathbf{r})$ and $\phi_{\mu\downarrow}(\mathbf{r})$. Contributions from each $\phi_{\mu\sigma}$ basis function vary in space, so each KS state's spin magnitude and direction can vary throughout the system.[49,66] The single-particle KS states cannot as spin orbitals in this case since spin direction and magnitude become coupled to the spatial wavefunction. Moreover, coupling between spins and orbitals can lead to anisotropic magnetism, where crystal structures can support preferred magnetic orders. In this thesis, spin-orbit coupling is primarily used to predict the effects of magnetism in Cu_{13} nanoclusters on the catalytic rate of H_2 dissociation in Chapter 4.

1.3 ZERO-POINT VIBRATIONAL ENERGY

Given the existence of the potential energy surface, $E_g(\mathbf{R})$, we consider its applications in catalytic studies. As discussed in Section 1.1.4, the Born-Oppenheimer approximation leading to $E_g(\mathbf{R})$ neglects nuclear quantum effects. For the most part these effects are minor at typical catalytic temperatures of 300 K and higher where quantum tunneling is impeded. For example, quantum tunneling effectively reduces the effective activation energy for H_2 dissociation over isolated Pt atoms in the top layer of the Cu(111) surface for temperatures up to 190 K, but the activation energy approaches its classical value by 300 K. An important factor that cannot be neglected, however, is *zero-point vibrational energy* (ZPE). Compounds are never fully localized at their minimum-energy geometry since the ionic positions $\mathbf{R}$ are also quantum degrees of freedom, so they have nonzero vibrational ground state energy similar to the simple harmonic oscillator.

Fortunately, it is relatively straightforward to include ZPE corrections in DFT calculations. We typically seek corrections for local minima where the forces on all atoms are zero. With $\mathbf{R}$ defined as the displacement from the minimum-energy geometry, the energy can be expanded in a Taylor series:

$$E(\mathbf{R}) = E_{\text{min}} + \frac{1}{2}\mathbf{R}^T H \mathbf{R} + O(R^3) \tag{1.30}$$

where H is the Hessian. Since the Hessian is symmetric, we can rewrite it as a sum over its eigenvalues λ_i and eigenvectors $\mathbf{u}_i$:

$$(H - \lambda_i I)\,\mathbf{u}_i = 0 \tag{1.31}$$

$$H = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T \tag{1.32}$$

A local minimum has all positive eigenvalues by definition.[‡] Equation (1.32) shows that energies increase quadratically and independently along each displacement direction $\mathbf{u}_i$ for small displacements. Similarly to eigenmodes, normal modes χ_i and and their frequencies ω_i are determined by analogy to the simple harmonic oscillator, with the "harmonic oscillation" ansatz:

$$\mathbf{R}(t) = \sum_i c_i \hat{\chi}_i \sin(\omega_i t - \theta_i) \tag{1.33}$$

where c_i are amplitudes of sinusoidal oscillations along each of the normal mode directions $\hat{\chi}_i$ with some phase θ_i. Application of Newton's equations of motion results in a similar eigenvalue equation for the normal modes and their frequencies:

$$\left(H - \omega_i^2 M\right)\chi_i = 0 \tag{1.34}$$

where M is the *mass matrix*, whose entries on the diagonal are M_I for coordinates belonging to atom I, and whose off-diagonal elements are zero.[§] Once Equation (1.34) is solved, we obtain the ZPE correction by summing over the zero-point energies of each of these independent harmonic oscillator modes:

$$E_{\text{ZPE}} = \frac{\hbar}{2}\sum_i \omega_i \tag{1.35}$$

It should also be noted that the normal mode directions χ_i are *not* the same as the Hessian eigenmodes $\mathbf{u}_i$: they are solutions to two different eigenvalue equations. In particular, note that that the Hessian eigenvalues in Equation (1.31) are multiplied by the identity matrix I, resulting in a standard eigenvalue problem, while the normal mode frequencies in Equation (1.34) are multiplied by the mass matrix M, resulting in a *generalized* eigenvalue problem. Although M is diagonal, diagonal matrices do generally commute with other matrices[¶] (including H) so Hessian eigenmodes and normal modes are different. This is a subtle but important distinction. For example, isotopic substitution changes M, normal modes, and their frequencies, but not H, its eigenvectors $\mathbf{u}_i$, or their curvatures.

[‡] The sum in Equation (1.32) implicitly does not include rigid translations or rotations which correspond to $\lambda_i = 0$.

[§] Technically M is defined as the matrix such that the kinetic energy can be written as $\frac{1}{2}\dot{\mathbf{R}}^T M \dot{\mathbf{R}}$. M is diagonal in Cartesian and mass-weighted Cartesian coordinates, which are most commonly used.

ZPE is especially important in activation energies for bonds involving H. For example, O–H bond vibrations have frequencies of order $3,000$ cm^{-1} so the corresponding ZPE correction is approximately 0.2 eV. In the alcohol dehydrogenation work presented in Chapter 3, this O–H vibrational frequency and other bond frequencies are negligibly affected by adsorption to Cu, so ZPE corrections to adsorption energies are minor. However, since the O–H bond no longer exists in the transition state leading to alcohol dehydrogenation, the ZPE correction to the activation energy for O–H bond breaking is roughly -0.2 eV! Vibrational frequencies for C–H bonds are similar in magnitude and thus have similar ZPE corrections as well for activation energies, so ZPE corrections should always be included for bonds involving H.

1.4 CLASSICAL TRANSITION STATE THEORY

1.4.1 CANONICAL ENSEMBLE DERIVATION

Classical transition state theory (TST) is the application of statistical mechanics to the prediction of reaction rates. Suppose we are interested in the elementary reaction step from compound A to compound B. These compounds can be defined as local minima of the potential energy surface. Molecular vibrations and phonon modes can then be described as the small deviations from these minima, while reactions correspond to more significant changes in molecular geometries. In general the reaction from A to B crosses a reaction surface, labeled S,[183] such that geometries displaced to one side of S will move toward A's local minimum, and microstates displaced on the other side will move toward B's local minimum. The relative probability of any two microstate geometries $\mathbf{R}$ and $\mathbf{R}'$ is given by their relative Boltzmann factors:

$$\frac{P(\mathbf{R}')}{P(\mathbf{R})} = \exp\left[-\beta\left(E\left(\mathbf{R}'\right) - E\left(\mathbf{R}\right)\right)\right] \tag{1.36}$$

Each microstate in the canonical ensemble has a location $\mathbf{R}$ and momentum $\mathbf{P}$ in phase space. We let $\mathbf{P}_{||}$ be the momentum tangent to the plane S at $\mathbf{R}$, and let $p_\perp$ be the momentum perpendicular to the plane. The flux across S at some point $\mathbf{R}$ in the surface is its speed $\frac{p_\perp}{M_\perp}$, where $M_\perp$ is the effective mass along $\hat{\mathbf{p}}_\perp$.[32,177]

[¶] They only commute when all entries on the diagonal are identical, i.e. if they are proportional to the identity matrix.

The relative flux per particle can be calculated as the flux across S, weighted by both the relative probability of each point $\left(\mathbf{R}, \mathbf{P}_{||}\right)$ in phase space as well as having a particular momentum $p_\perp$ perpendicular to S:

$$F = \frac{1}{Z}\int_S \mathrm{d}\mathbf{R}\, d\mathbf{P}_{||}\, dp_\perp\, e^{-\beta E(\mathbf{R})} \exp\left[-\frac{\beta}{2}\left(\frac{P_{||}^2}{M_{||}} + \frac{p_\perp^2}{M_\perp}\right)\right]\frac{p_\perp}{M_\perp} \tag{1.37a}$$

This expression is very general but difficult to evaluate because of the high dimesionality. However, we can see that flux contributions decay exponentially with $E\left(\mathbf{R}\right)$ due to the Boltzmann factors. Therefore most of the F is contributed by the lowest energy region(s) of S. If there is a single, small low-energy region on S, then S can be approximated as a plane tangent to this region whose normal vector $\hat{\mathbf{p}}_\perp$ is constant.[32,177] In this case the integrals over $\hat{\mathbf{p}}_\perp$ and $\mathbf{R}$ are separable:

$$\approx \frac{1}{Z}\left[\int_S \mathrm{d}\mathbf{R}\, d\mathbf{P}_{||}\, e^{-\beta E(\mathbf{R})}\exp\left(-\frac{\beta P_{||}^2}{2M_{||}}\right)\right]\int_0^\infty \mathrm{d}p_\perp \exp\left(-\frac{\beta p_\perp^2}{2M_\perp}\right)\frac{p_\perp}{M_\perp} \tag{1.37b}$$

$$= \frac{1}{Z}\left[\int_S \mathrm{d}\mathbf{R}\, \mathrm{d}\mathbf{P}_{||}\, e^{-\beta E(\mathbf{R})}\exp\left(-\frac{\beta P_{||}^2}{2M_{||}}\right)\right] kT \tag{1.37c}$$

The term in square brackets is the partition function integral over S. Defining this quantity Z^*, we obtain a general expression for the flux per particle:

$$F = kT\,\frac{Z^*}{Z} \tag{1.37d}$$

Rather than use the full partition function integral, we can redefine the partition functions as being relative to the minimum energy encountered in each region. Letting ΔE_A be the difference between the minimum energy along S and the minimum energy of A, we obtain a more familiar form for the flux:

$$F = kT\,\frac{Z^*}{Z}e^{-\beta(E_{\mathrm{TS}} - E_A)} \equiv kT\,\frac{Z^*}{Z}e^{-\beta\Delta E_A} \tag{1.38}$$

where Z and Z^* are now integrals of energies *relative* to their respective minima. Z and Z^* account for configurational entropy since low-energy distortions lead to Boltzmann factors near unity and thus larger partition function integrals. This is the statistical mechanics derivation of the Arrhenius rate law. [||] Finally, we note that this flux is defined per particle so the overall rate, in particles per second, is F multiplied by the

---

[||] Note that no shortcuts based on quantum mechanics, "attempt frequencies," or a fictitous "unstable frequency" are necessary to obtain the Arrhenius expression. These tricks deliver the right results for the wrong reasons.

concentration or coverage of A molecules. In some cases energies and the integrals change with concentrations, e.g. interadsorbate interactions and excluded-volume effects; these cases can be considered in the grand canonical ensemble.

Deriving Equation (1.38) exposes several quantities and assumptions of central importance to TST and calculating reaction rates. First, the minimum-energy geometry $\mathbf{R}_{\text{TS}}$ along S is the *transition state*. It is a first-order saddle point on the potential energy surface because it has negative curvature perpendicular to S by definition, and it is a local minimum with respect to all other degrees of freedom. Second, the quantity $\frac{Z^*}{Z}$ is commonly referred to as the "entropy factor." Transitions where S contains many accessible configurations, or where A contains few accessible configurations, lead to high reaction rates. The number of of accessible microstates in S and A can change differently with temperature as well. Empirically speaking the overall rate factor for bond breaking in molecules with several atoms is roughly 10^{13} Hz when the reaction leaves other parts of the potential energy surface along other degrees of freedom are unaffected. This is a common situation so 10^{13} Hz is commonly used when more detailed calculation or experimental results are unavailable. For example, breaking an alcohol O–H bond does not significantly change the properties of the alcohol's C–H bonds. However, some reactions lead to significant entropy changes and therefore have dramatically different rates. In molecular desorption, for example, the desorbed molecule has significantly higher configurational entropy than the adsorbed state so a typical desorption rate prefactors rise to $10^{15.5}$ Hz.[14]

The Hessian and the harmonic approximation are typically employed to further simplify Equation (1.38). Energies corresponding to small displacements are quadratic as shown in Equation (1.30), so for small displacements we can approximate $E(\mathbf{R})$ near the minimum-energy geometry $\mathbf{R}_A$ and the transition state as quadratic. Each mode, with vibrational frequency $\omega_i = 2\pi\nu_i$ contributes a factor of $\frac{1}{\nu_i}$.[177] High-frequency modes involve rapid increases in the energy with displacement and therefore small partition functions. With this approximation we obtain the "Vineyard prefactor" for the rate:

$$F = \frac{\prod_i^{3N} \nu_i}{\prod_i^{3N-1} \nu_i^{\text{TS}}} \tag{1.39}$$

This expression was popularized by Vineyard.[177] The vibrational frequencies are the same as those used for harmonic ZPE corrections, so finding the normal modes for adsorbates yields both the ZPE correction as well as the Vineyard prefactor with minimal additional effort.

Although Equation (1.39) is straightforward and nearly comes for free along with ZPE corrections, its actual utility is highly debatable for surface science. Among other approximations employed in its derivation, e.g. that the dividing surface S is a plane, it is subject to a number of questionable assumptions:

1. There is a single transition state, leading to a single effective plane that approximates S.
2. 100% of the flux across S results in a reaction, the "no-recrossing" assumption.
3. $E(\mathbf{R})$ is quadratic near the minima of A and S for all distortions that change energies by up to several kT. Otherwise the harmonic approximation fails. Most of the configurational entropy, and hence the prefactor, comes from very low-energy distortions. Unfortunately, for weakly adsorbed molecules these energies are so low that $E(\mathbf{R})$ is not quadratic. In fact, weakly-bound adsorbates can even have diffusion barriers lower than kT. In these cases the harmonic approximation significantly underestimates entropy for adsorbed molecules so desorption rates are overestimated by orders of magnitude.[14]
4. Vibrations along each normal mode satisfy the classical limit $\hbar\omega_i \ll kT$ where the classical partition function is valid. For high frequency modes, such as O–H bond vibrations as discussed previously, $\hbar\omega_i$ significantly exceeds kT. In these cases, alternate approaches to transition state theory should be used to handle vibrational quantization.[172]
5. The mode frequencies ν_i must be computed precisely enough that their product in Equation (1.39) has low numerical error.

Most of these assumptions have been thoroughly considered in the references provided above and elsewhere.

1.4.2 GAMBLING WITH NUMERICAL ERROR

However, an issue that has not been discussed in the literature is assumption 5: calculations may be insufficiently precise. Precision is particularly challenging to obtain since small Hessian errors can yield large relative errors for low frequencies. Moreover, the product of these frequencies in the Vineyard prefactor *accumulates* all these relative errors. Poor precision for one mode yields poor precision for the whole product. Adsorbates generally have several low-frequency modes, corresponding to unhindered translations and rotations in the gas phase, so low frequencies are ubiquitous.

Relative uncertainties in the harmonic frequencies can be quantified using Monte Carlo simulations. Finite differences are used to calculate the Hessian we can sample noise values from errors observed in the calculation. Defining the "error matrix" as $\mathcal{E} = \left(H - H^T\right)/2$, we can obtain new error matrices $\mathcal{E}'$ by sampling with replacement from the actual errors matrix $\mathcal{E}$. This resampling process, using a data set as an approximate histogram, is known as *bootstrapping*. Relative errors in frequencies are obtained by performing many frequency calculations with different bootstrapped errors:

1. Symmetrize the Hessian $H_{\text{sym}} \equiv (H + H^T)/2$
2. Obtain a bootstrapped error matrix $\mathcal{E}'$ by sampling with replacement from $\mathcal{E}$
3. Let $H_{\text{err}} = H_{\text{sym}} + \mathcal{E}'$ be the corresponding noisy Hessian
4. Symmetrize H_{err}, then calculate its normal mode frequencies and their product

Each trial produces a set of vibrational frequencies given its bootstrapped errors. Distributions of each mode's frequencies and their products from these Monte Carlo trials provides confidence intervals. For example, I applied this bootstrapping approach to the Hessian obtained from VASP for the minimum-energy ethanol geometry on the (111) step of the Cu(111) surface (see Chapter 3 for details). This test reveals that the 68% confidence interval for the product of frequencies is 300% of its "actual" value from H_{sym}. These ethanol Hessians were calculated with even higher precision than that needed to fully converge adsorption energies and geometries, so these large uncertainties should be considered typical for weakly-bound adsorbates. Calculating ratios of *two* products for the Vineyard prefactor would lead to even worse numerical error. Considerable care and great attention to detail is therefore needed when calculating the Vineyard prefactor in the presence of numerical error.

1.4.3 FINDING TRANSITION STATES

We now turn to the process of obtaining the transition state geometry and its energy. From the flux derivation leading to Equation (1.37d) we identified that the transition state is a *first-order saddle point*, the local maximum along the reaction path $\mathbf{R}(\chi)$ and a local minimum along all other directions. In the Vineyard prefactor this used with the harmonic approximation to simplify the partition function integrals. For an elementary reaction there is only a single local maximum along $\mathbf{R}(\chi)$ connecting the reactant and product geometries. Moreover, as a local maximum along the reaction path, the lowest-curvature mode will have the most negative curvature. Saddle points, by definition, also have zero gradients so forces should vanish. Consequently there are three objectives when searching for transition states:

1. zero forces
2. highest energy along the reaction path
3. most negative possible curvature

While many transition state search algorithms exist, two of them are particularly notable for their popularity and success. The most popular of these is the *nudged elastic band method* (NEB).[57,58,115] NEB discretizes the reaction path between two user-selected geometries into m steps, known as *images*. Along the path direction, forces along the discretized path are projected out and replaced with springs between adjacent images to keep them uniformly spaced along the path. The path direction, or *tangent* $\mathbf{t}_i$, for each image $\mathbf{R}_1, \cdots, \mathbf{R}_m$ is the unit vector toward its higher-energy neighbor. This single-sided derivative is less accurate than a centered difference calculation, but centered differences tend to yield kinks and discontinuities in the path.[57] With these definitions the "NEB force" is reasonably simple:

$$\mathbf{F}_{\text{NEB},i} = \left(1 - \mathbf{t}_i \mathbf{t}_i^T\right) \mathbf{F}_{\text{DFT},i} + k\, \mathbf{t}_i \left(d_{i,i+1} - d_{i,i-1}\right) \tag{1.40}$$

where $\mathbf{F}_{\text{NEB,i}}$ and $\mathbf{F}_{\text{DFT},i}$ are the NEB and actual forces at geometry $\mathbf{R}_i$, k is some spring constant, and $d_{i,j}$ is the distance $\left|\mathbf{R}_i - \mathbf{R}_j\right|$. Minimizing this force yields equally spaced images since the spring term only vanishes when distances to neighboring images is equal. The forces perpendicular to the path, $\left(1 - \mathbf{t}_i \mathbf{t}_i^T\right) \mathbf{F}_{\text{DFT},i}$, are minimized as usual. In the limit that many images are used, the highest-energy image will be close to the transition state.

In surface catalysis studies, however, using a large number of images is computationally impractical due to the large number of atoms and vacuum space. Reasonable approximations to the reaction path can be obtained by interpolating among only a few images, but in general none of the NEB images will lie at the transition state itself. A common improvement that addresses this issue is the *climbing-image* NEB method (cNEB).[58] At each cNEB step, a special "climbing force" is applied to the highest-energy image at $\mathbf{R}_*$. The transition state is a local maximum along the path so the climbing force *reverses* the force along the tangent direction:

$$\mathbf{F}_{\text{NEB},*} = \left(1 - 2\mathbf{t}_* \mathbf{t}_*^T\right) \mathbf{F}_{\text{DFT},*} \tag{1.41}$$

where the tangent $\mathbf{t}_*$ for this highest-energy image uses centered differences rather than the one-sided differences used for the other geometries. Optimized image geometries from cNEB now have their highest-energy geometry close to the transition state by construction. The climbing image lacks the spring force, however, so the image spacing on the reactant side of the path can differ from that of the product side.

NEB's relative simplicity and explicit path discretization makes it the most popular choice for accurate transition state searches. Its most significant strength is that the transition state is guaranteed to lie between the user-selected reactant and product geometries. Furthermore, information about the reaction *mechanism* can be obtained by observing the changes from one image to the next along the pathway. On the other hand, these advantages are closely related to two of the NEB method's major weaknesses: the user must specify both the reactant and product geometries, and the NEB method minimizes the collective force among m geometries. The first is frustrating because often times a key piece of the mechanism is known, e.g. breaking an alcohol O–H bond, but the initial and final geometries for this elementary step are unknown. Using the NEB method then requires running a separate NEB calculation for each possible pair of reactant and product geometries. The second consideration is that m geometries must be optimized instead of just one. Moreover, these force minimizations are now coupled by the distant-dependent spring forces and the path tangents $\mathbf{t}_i$. These further increase the computational costs of NEB calculations beyond the m-fold increase from minimizing m images instead of one.

Besides these more obvious issues, there were two other prominent problems encountered throughout the work described in this thesis. First, the cNEB method typically fails when the there are multiple transition states. Such cases are relatively common, such as reactants diffusing to a metastable geometry to facilitate breaking a bond of interest. Second, discontinuous paths prevent convergence because the one-sided tangent approximations $\mathbf{t}_i$ fail at each discontinuity. In the nanocluster work in Chapter 4, for example, the minimum-energy reaction path has sharp turns that lead to poor finite-difference approximations for the path tangents $\mathbf{t}_i$. In that work, between 15 and 23 images were needed to converge the path. The associated computational costs for the small nanocluster were acceptable, but full surface calculations with this many images would have been prohibitively expensive.

A complementary method that resolves most of these challenges is the *dimer method*.[56,59] Rather than discretize the reaction path, the dimer method is a local search that directly seeks a nearby transition state. This is done by minimizing the lowest eigenmode of the Hessian and minimizing energy along all other directions. This process naturally converges at first-order saddle points.

Similar to the NEB method, the dimer method uses modified forces. Two closely-spaced images, which put the "dimer" in "dimer method," are slightly displaced from each other along $\hat{\mathbf{n}}$, the *dimer axis*. The direction of $\hat{\mathbf{n}}$ is the dimer method's estimate for the Hessian eigenmode with the lowest curvature. The dimer is iteratively translated and rotated throughout the search as the following effective force is minimized:

$$\mathbf{F}_{\text{dimer}} = (1 - 2\hat{\mathbf{n}}\hat{\mathbf{n}}^T)\,\mathbf{F}_{\text{DFT}} \tag{1.42}$$

This minimizes force perpendicular to the dimer and *reverses* the force along the dimer, climbing to the nearest saddle point where the lowest curvature mode is most negative. This effective force is very similar to the cNEB method's climbing force in Equation (1.41).

Since computing the full Hessian at each iteration is prohibitively expensive, the dimer method uses only a few finite-difference calculations at each iteration to obtain just the lowest-frequency mode. Curvature is minimized by *rotating* the dimer (i.e. $\hat{\mathbf{n}}$) to the lowest-curvature mode. This is done in two steps. First, the current curvature c_1 and "torque" $\boldsymbol{\tau}$ are computed from the forces at each endpoint of the dimer:

$$c_0 = \hat{\mathbf{n}} \cdot \left(\mathbf{F}_{\mathrm{DFT},1} - \mathbf{F}_{\mathrm{DFT},0}\right) \tag{1.43a}$$

$$\boldsymbol{\tau} = \left(1 - \hat{\mathbf{n}}\hat{\mathbf{n}}^T\right)\left(\mathbf{F}_{\mathrm{DFT},1} - \mathbf{F}_{\mathrm{DFT},2}\right) \tag{1.43b}$$

where $\mathbf{F}_{\mathrm{DFT},0}$ and $\mathbf{F}_{\mathrm{DFT},1}$ are the forces.[**] The torque vector $\mathbf{t}$ is the steepest-descent direction for curvature, so the dimer method minimizes curvature by rotating $\hat{\mathbf{n}}$ by some angle θ toward $\hat{\boldsymbol{\tau}}$, producing a new mode direction $\hat{\mathbf{n}}'(\theta) = \cos(\theta)\hat{\mathbf{n}} + \sin(\theta)\hat{\boldsymbol{\tau}}$. Curvature is π-periodic with respect to θ so curvature along the rotated direction can be written as a linear combination of sines and cosines:

$$c(\theta) = A + B\cos(2\theta) + C\sin(2\theta) \tag{1.44}$$

where A, B, and C are constants. The curvature c_0 and its derivative $c'(0) = -\tau$ are given by the finite difference expressions above. To solve for all three variables we require a third equation, obtained by computing the curvature $c(\theta)$ along a rotated vector $\hat{\mathbf{n}}'(\theta)$. Choosing $\theta = \pi/4$ keeps the dimer method numerically stable and produces a simple equation for the rotation angle θ^* that minimizes the total curvature:[59]

$$\theta^* = \frac{1}{2}\tan^{-1}\left(\frac{\tau}{2c(\pi/4) - 2c(0) + \tau}\right) \tag{1.45}$$

This rotation process requires three DFT force calculations: one at each endpoint of the dimer, and a third for displacement along $\hat{\mathbf{n}}'(\theta = \pi/4)$.

The dimer method's chief advantage is that it searches *locally* for transition states rather than trying to converge the entire reaction pathway. This path gives information regarding the catalytic mechanism for the reaction, but as discussed above, only the transition state itself and its neighborhood on the dividing surface S are truly important. Including an extra gradient for minimizing the effective dimer force, typically using either the conjugate gradient or BFGS methods, the dimer method requires a total of four gradients per iteration while the NEB method requires $2m$ gradients. NEB calculations generally use at least three to five images so only four gradients is already a significant advantage. Once converged, the dimer geometry is located at the transition state and its orientation $\hat{\mathbf{n}}$ is aligned with the unstable mode. In addition, the dimer method only requires an initial guess for the transition state and the direction $\hat{\mathbf{n}}$ for the unstable mode. Since catalytic studies often focus on bond breaking, a reasonable guess for the transition state and unstable mode is to stretch the bond of interest, and use the corresponding displacement vector as the guess for $\hat{\mathbf{n}}$. The dimer method's inputs are therefore relatively easy to obtain from chemical intuition. Dimer method calculations are extensively used for the work on anhydrous alcohol dehydrogenation in Chapter 3 and on H_2 dissociation over Cu_{13} nanoclusters in Chapter 4.

[**] Technically these expressions should be normalized by the length of the dimer, but only ratios are important for the dimer method.

Dimer method calculations also have two significant disadvantages. The first, and most important, is that the dimer can get "distracted" as it searches for a transition state since it orients itself along the lowest-curvature mode *of the current geometry*. This issue primarily arises when searching for low-curvature transition states such as diffusion modes for weakly bound adsorbates. Instead of converging the low-curvature transition state for diffusion, distortions introduced by the initial guess or during force minimization lead the dimer to find corresponding bond activation or bending modes with much lower curvature instead. This issue is empirically rare for bond breaking, but relatively common for diffusion. The second disadvantage is that the dimer method is a local search and only returns the transition state geometry itself. Minimum-energy pathways require subsequent steepest-descent or conjugate-gradient optimizations toward the reactant and product states.

1.5 TIME-DEPENDENT DFT AND EHRENFEST DYNAMICS

This section is adapted from the following publication:

- Grigory Kolesov, Oscar Grånäs, **Robert Hoyt**, Dmitry Vinichenko, and Efthimios Kaxiras. "Real-Time TD-DFT with Classical Ion Dynamics: Methodology and Applications." *J. Chem. Theory Comput.* **12** 2016, 466–476.

The variational approach used to obtain DFT from a time-independent Hamiltonian can be made more general to handle time-dependent systems as well. This gives rise to time-dependent DFT (TDDFT), describing the time-dependent density and total energy corresponding to a time-dependent external potential. Accounting for atomic motion in this process, including some or all the nonadiabatic interactions discussed in Sections 1.1.3 and 1.1.4, is known as *nonadiabatic molecular dynamics* (NAMD). A wide variety of methods with different levels of sophistication and computational expense have been developed over the years. Most of these, however, are too expensive to implement for systems with hundreds of atoms over the long time scales of order 10-100 fs for chemical reactions. Among the less expensive approaches, some popular choices have drawbacks in terms of surface catalysis. For example, the fewest-switches surface hopping method[173] is prohibitively expensive for metallic surfaces with high densities of states, and even its modern successors require model Hamiltonians to be computationally feasible.[154] Another example is molecular dynamics with first-order "electronic frictions"[53] where strong nonadiabatic coupling can lead to diverging friction.[53] Therefore we will consider a specific approach to NAMD that scales to hundreds of atoms and is generally appropriate for metallic surfaces: Ehrenfest dynamics.

Ehrenfest dynamics can be derived in a number of ways, but the best and most elegant approach is presented by Kunert and Schmidt.[90] Whereas DFT is obtained by variationally minimizing total energy with respect to density, Kunert and Schmidt derive Ehrenfest dynamics by variationally minimizing a *mixed quantum-classical action*. The total action A is split into a classical part A_c, composed on N classical point masses describing the atoms, and a quantum part A_q describing n electrons. In this scheme the nuclei are asserted to be point particles from the beginning in the spirit of the classical ion approximation (Section 1.1.2) and the BO approximation (Section 1.1.4). Kunert and Schmidt then derive a collection of core equations that dictate dynamics:

$$A = A_c + A_q \tag{1.46a}$$

$$A_c = \int_{t_0}^{t_1} \mathrm{d}t \left(\sum_I \frac{M_I}{2} \dot{\mathbf{R}}_I^2(t) - U(\mathbf{R}(t), t) \right) \tag{1.46b}$$

$$A_q = \int_{t_0}^{t_1} \mathrm{d}t \left\langle \Psi(t) \middle| i\hbar \frac{\partial}{\partial t} - \mathcal{H}(\mathbf{R}(t), t) \middle| \Psi(t) \right\rangle \tag{1.46c}$$

$$\mathbf{F}_I = -\nabla_I U(\mathbf{R}(t), t) - \nabla_I E(\mathbf{R}(t), t) \qquad \forall I \in 1..N \tag{1.46d}$$

$$i\hbar \dot{\psi}_i(t) = \mathcal{H}_{\mathrm{DFT}} \left[\rho(\mathbf{r}', t')\right](t)\, \psi_i(t) \qquad \forall i \in 1..n \tag{1.46e}$$

where $U(\mathbf{R}(t), t)$ includes the ion-ion repulsion term and any other interactions only affecting the ions, and $E(\mathbf{R}(t), t)$ is the TDDFT total energy. Equations (1.46d) and (1.46e) are the result of minimizing the classical and quantum actions in the TDDFT formalism. In principle $\mathcal{H}_{\mathrm{DFT}}[\rho(\mathbf{r}, t')](t)$ is nonlocal in both space and time, but as an approximation we apply the "adiabatic" approximation[††] and use the local PBE functional as the time-dependent exchange-correlation functional. Equations (1.46e) are the time-dependent KS (TDKS) equations. Propagating the TDKS equations in time implicitly includes the effects of ionic motion through $\mathcal{H}_{\mathrm{DFT}}$.

[††] Adiabatic here refers to approximating the functional as being strictly local in time t, not the adiabatic approximation in Section 1.1.3.

As with time-independent DFT, a carefully chosen basis set is necessary to facilitate the calculation of $\mathcal{H}_{\text{DFT}}$ and to propagate the time-dependent KS states $\psi_i(t)$. For the work presented in this thesis, a pseudoatomic orbital (PAO) basis, with orbitals $\phi_\mu(\mathbf{r})$, are used to reduce computational effort. As discussed previously in Section 1.2.2, PAOs require the fewest orbitals to obtain a good description of bonding. Letting $\mathbf{c}_i$ be the time-dependent vector of coefficients for KS state i, such that $(\mathbf{c}_i)_\mu$ corresponds to basis function $\phi_\mu(\mathbf{r})$, the TDKS equations can be written as

$$i\hbar\dot{\mathbf{c}}_i = S^{-1}(H - iP)\,\mathbf{c}_i, \tag{1.47}$$

where S is the overlap matrix and H is the DFT Hamiltonian in the chosen basis set. The new addition compared to the time-independent KS equations is the matrix P, defined as $P_{\mu\nu} = \left\langle \phi_\mu \,\middle|\, \frac{d}{dt} \,\middle|\, \phi_\nu \right\rangle$. P arises from the chain rule; PAOs are centered at the ionic positions $\mathbf{R}_I(t)$ and therefore vary in time along with the coefficients $\mathbf{c}(t)$. Equation (1.47) defines the time dependence of each of the KS states. We then integrate the TDKS equations with respect to time using the time evolution operator $\hat{\mathcal{U}}(t+\Delta t, t)$:

$$\mathbf{c}_i(t+\Delta t) = \mathcal{U}(t+\Delta t, t)\mathbf{c}_i(t) \tag{1.48a}$$

$$\mathcal{U}(t+\Delta t, t) = \exp\left\{-i\Delta t\left[S^{-1}_{1/2}\left(H_{1/2} - iP_{1/2}\right)\right]\right\} + O\left(\Delta t^3\right) \tag{1.48b}$$

where the subscripts of $\frac{1}{2}$ indicate that S, H, and P are evaluated at the *midpoint* between time steps t and Δt, i.e. $H_{1/2} = \frac{1}{2}\left(H(t) + H(t+\Delta t)\right)$. Using the midpoint rule increases the order of accuracy relative to simply using the the matrices evaluated at t alone (Euler integration), making Equation (1.48b) accurate to second order in the time step Δt. Typical time steps are $\Delta t \lesssim 24$ as, the characteristic time scale for electrons. Choosing the midpoint rule also improves time-reversibility and thus energy conservation, but also increases computational costs since $H(t+\Delta t)$ and $\rho(\mathbf{r}, t+\Delta t)$ must be known. In practice this is accomplished by iteratively propagating until self-consistency is reached,[16] similar to the iterations performed for time-independent DFT. The matrix exponential in Equation (1.48b) is therefore calculated several times for each of the small time steps Δt. We find that computational effort is reduced as much as possible by using Padé approximants of order 3/3, augmented with a scaling-and-squaring strategy,[16] while still keeping the calculations stable and sufficiently accurate.

Solving for the time-dependent atomic positions $\mathbf{R}_I(t)$ is more straightforward since the nuclei are considered point particles. As in BO molecular dynamics the nuclei are propagated according to Newton's equations with standard Verlet integration. Using an atom-centered PAO basis complicates the force calculation in Equation (1.46d) since ∇_I affects the PAO basis functions as well as the Hamiltonian itself. The full expression for the forces is given in Equation (1.49):

$$\mathbf{F}_I = -\nabla_I \sum_{J \neq I} \frac{Z_I Z_J e^2}{R_{IJ}} - \mathrm{Tr}\left\{\rho \left[\nabla_I H - \mathbf{G}_I - \left(H S^{-1} \mathbf{d}_I + \mathbf{d}_I^{\dagger} S^{-1} H\right)\right]\right\} \tag{1.49}$$

$$(\mathbf{G}_I)_{\mu\nu} \equiv \left\langle \phi_\mu \left| \nabla_I \left(V_{\mathrm{H}} + V_{\mathrm{xc}}\right) \right| \phi_\nu \right\rangle \tag{1.50}$$

$$(\mathbf{d}_I)_{\mu\nu} \equiv \left\langle \phi_\mu \left| \nabla_I \right| \phi_\nu \right\rangle \tag{1.51}$$

G and **d** are vector matrices that account for how the orbitals are affected by the nuclear gradients ∇_I. These forces are similar to those in used in the ordinary Pulay force term.[141,156] Equation (1.49) explicitly shows that the nuclear forces F_I correspond to the total electron density ρ via the trace, regardless of whether ρ corresponds to a single electronic state or a superposition of many electronic states reached by nonadiabatic transitions. In this sense the Ehrenfest forces are considered "mean field" forces since they correspond to the mean potential energy surface. Ehrenfest forces are valid when the electronic density corresponds to a single electronic state, or to a superposition of several states with similar forces. On the other hand, when different electronic transitions result in significantly different forces, Ehrenfest dynamics will follow the a weighted mean of these forces and be misleading. The following subsections discuss two different applications of Ehrenfest dynamics to systems of catalytic interest.

1.5.1 MOLECULAR PHOTODISSOCIATION

Azobisisobutyronitrile (AIBN) is a common source of radicals in photopolymerization. AIBN dissociates upon exposure to UV light to form a N_2 molecule and a pair of isobutyronitrile radicals that facilitate polymerization through radical chemistry. Linear response calculations show that the HOMO to LUMO+1 excitation closely resembles this UV transition. The initial condition for the TDDFT simulation of this transition was therefore the HOMO to LUMO+1 state at the minimum energy geometry using the Δ-SCF method, where the highest energy occupied KS state was left unoccupied and the second-lowest energy KS state was occupied until self-consistency was reached. The TDDFT simulations the proceed by alternating between Equations (1.48b) and (1.49) with a time step of $\Delta t = 25$ as.

The dissociation proceeds as shown in Figure 1.1, producing the pair of isobutyronitrile radicals and a nitrogen molecule as anticipated. During the dissociation, the net spin of each radical oscillates substantially before arriving at an equilibrium value. These oscillations are the result of the initial excited state, with no net magnetization on each isobutyronitrile group, gradually coupling to magnetized radical states through

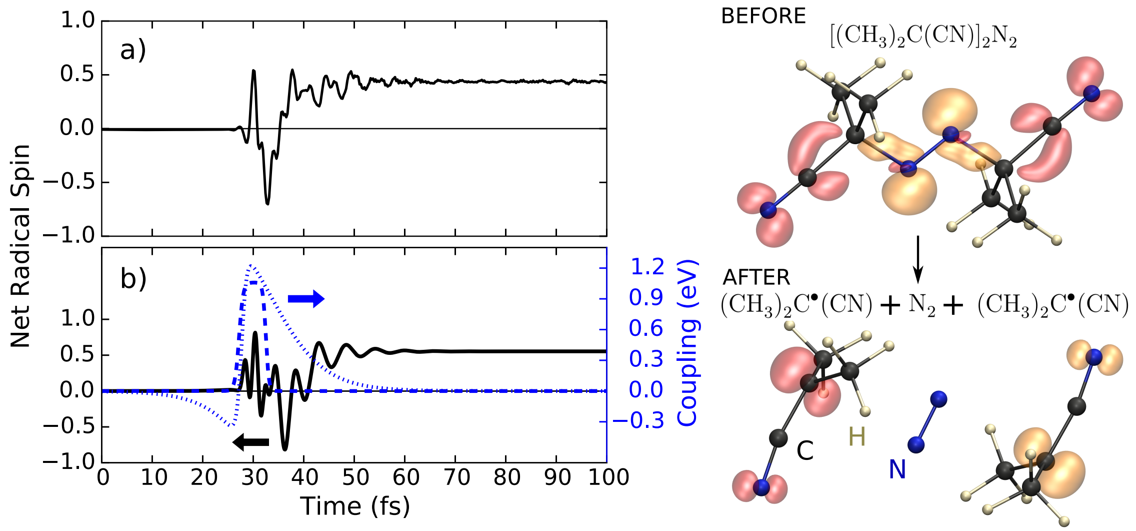


Figure 1.1: Net radical spin on one of the isobutyronitrile radicals calculated from: (a) the nonadiabatic propagation and (b) the model Hamiltonian. In (b), the short-dash line shows the term $a(t)$ in the effective Hamiltonian, the long-dash line shows $b(t)$. The geometry of the AIBN molecule before and after dissociation is shown on the right. The red and orange lobes are magnetization density isosurfaces of opposite sign.

the central N_2 molecule. These magnetized states interact nonadiabatically via the DFT Hamiltonian and therefore yield rapid oscillations in the magnetization density.

To qualitatively understand this behavior we consider a simple three-state model Hamiltonian $H_{\text{eff}}(t)$ involving the two magnetized radical states, $|\uparrow\downarrow\rangle$ and $|\downarrow\uparrow\rangle$, and the unmagnetized initial state $|\Delta\text{SCF}\rangle$. For simplicity, we model the matrix elements with sigmoidal time-dependence:

$$H_{\text{eff}} = \begin{pmatrix} E & a(t) & -a(t-t_0) \\ a(t) & 0 & b(t) \\ -a(t-t_0) & b(t) & 0 \end{pmatrix} \tag{1.52}$$

where the matrix elements represent the following couplings:

$$\langle \Delta\text{SCF}|H_{\text{eff}}(t)|\Delta\text{SCF}\rangle = E = 0.71 \text{ eV} \tag{1.53}$$

$$\langle \Delta\text{SCF}|H_{\text{eff}}(t)|\uparrow\downarrow\rangle = a(t) = -\langle \Delta\text{SCF}|H_{\text{eff}}(t)|\downarrow\uparrow\rangle \tag{1.54}$$

$$\langle \uparrow\downarrow|H_{\text{eff}}(t)|\downarrow\uparrow\rangle = b(t) \tag{1.55}$$

The coupling to the $|\downarrow\uparrow\rangle$ state is shifted by $t_0 = 1$ fs to model the slight initial asymmetries that lead to the $|\uparrow\downarrow\rangle$ state first dominating the overall magnetization. The numerical solution for the net radical spin is

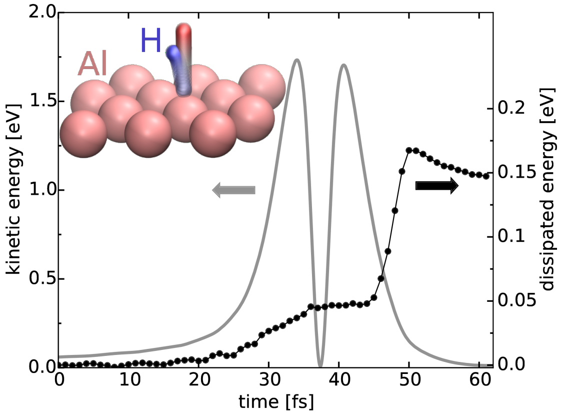


Figure 1.2: H atom kinetic energy and the total energy dissipated into the electronic degrees of freedom for H adsorption over the top site of the Al(111) surface. Insert shows the motion trail over the first 60 fs of the H atom adsorbing at the Al(111) top site. The H atom's initial position is shown in red, and its position at 60 fs is shown in blue. Only the top layer of Al atoms is shown for clarity.

calculated as the probability of being in the $|\uparrow\downarrow\rangle$ state minus that of being in $|\downarrow\uparrow\rangle$ state. A comparison between the DFT and model net radical spin, as well as the time-dependence of the $a(t)$ and $b(t)$ matrix elements, is shown in Figure 1.1. The initial onset of large oscillations, their amplitude, and the trailing fluctuations around 50 fs are qualitatively well-reproduced despite the significant simplifications implied by the model.

1.5.2 H ATOM SPIN TRANSITIONS

H adsorption on metallic surfaces exhibits an electronic phase transition where the H atom's net spin discontinuously vanishes once it is sufficiently close to the surface.[7,102,103,128,168] This distance is known as the spin transition point. This spin transition is a fundamentally nonadiabatic effect due to the square root discontinuity in the ground state spin polarization as it crosses the spin transition point, and previous studies calculated that this transition dissipates roughly 200 meV into electronic degrees of freedom over the first period of H atom motion as it collides with a metallic FCC (111) top site.[7,102]

We model the system representing H adsorption by a $2\sqrt{3}\times 2\sqrt{3}$ supercell of the Al(111) surface. To reduce computational effort as much as possible we use a single-zeta polarized (SZP) basis set, a real-space cutoff of 35 Ry, and only sample the Brillouin zone at the Γ-point. The Al slab consists of 6 layers in the (111) direction, with the bottom two fixed at their bulk positions. These settings are chosen to closely resemble those of Lindenblatt and Pehlke.[102] The amount of nonadiabatic energy dissipation is determined as the the difference between the electronic energy in the nonadiabatic calculation and the corresponding ground-state electronic energy with identical atomic positions. The difference between these two calculations is exactly the degree to which the electrons have been excited in the metal surface. The H atom starts 4 Å above the Al(111) top site and directly approaches the surface with an initial kinetic energy of 60 meV. The motion trail of the H atom colliding with the top site is shown in the inset of Figure 1.2, and our results for the H atom's kinetic energy and the nonadiabatic energy dissipation in the first period of the H atom's motion are presented in Figure 1.2. The two step-like jumps in the electronically-dissipated energy correspond to the H atom crossing the spin transition point, once on approach and again right after it recoils from its collision with the Al(111) top site. We calculate a maximum energy of 0.17 eV dissipated to the electronic degrees of freedom, close to the maximum of 0.18 eV found by Lindenblatt and Pehlke[102] despite the low computational parameters chosen. This dissipated energy is twice the amount transferred to the Al atoms' total kinetic energy, highlighting the importance that nonadiabatic effects can have on adsorption at metallic surfaces.

1.6 CONCLUSION

This chapter reviews the major considerations and approximations made in studying the electronic structure and catalytic properties of "large" systems on the order of 100 atoms. Directly solving the Schrödinger Equation is computationally impractical due to its high dimensionality, and conceptually unnecessary since only one and two-body operators are necessary to describe electron dynamics (see Section 1.2 and work by Levy[97]). These considerations, along with the Born-Oppenheimer approximation (see Section 1.1.4), lead to density functional theory. DFT reduces all of the many-body complexity of the Schrödinger Equation to a functional of the total electron density $\rho(\mathbf{r})$. With appropriate functionals, particularly local GGA functionals, large systems can be studied with relatively modest computational resources. The work presented in this thesis makes extensive use of DFT with the PBE functional to understand adsorption and heterogeneous catalysis at the atomic scale. Some discussion of relativistic effects is also provided for spin-orbit coupling in Section 1.2.3.

Moreover, application of the Born-Oppenheimer approximation leads to potential energy surfaces. The latter are a key ingredient in statistical mechanics in the canonical ensemble and classical transition state theory. Section 1.4 reviews the canonical ensemble description of transition state theory and the origins of the Arrhenius rate law, without questionable appeals to "attempt frequencies" or other conceptual shortcuts. This review provides a clear picture of transition state theory's limitations, and in particular the limitations of harmonic transition state theory where rates are calculated using the Vineyard prefactor.[177] The popular

NEB and dimer methods for locating transition states are introduced and compared with one another. Finally, this chapter also discusses ways to treat catalytically-relevant effects beyond the Born-Oppenheimer approximation, specifically zero-point vibrational motion in Section 1.3 and nonadiabatic dynamics in Section 1.5.

2

Polyiodide-Doped Graphene

Apart from minor modifications, this chapter originally appeared in the following publication:

> **Robert A. Hoyt**, E. Marielle Remillard, Ekin D. Cubuk, Chad D. Vecitis, and Efthimios Kaxiras. "Polyiodide-Doped Graphene." *J. Phys. Chem. C* **121**(1) 609 (2017)

ABSTRACT

Iodine-doped graphene has recently attracted significant interest as a result of its enhanced conductivity and improved catalytic activity. Using density functional theory calculations, we obtain the formation energy, desorption rate, and electronic properties for graphene systems doped with polyiodide chains consisting of one to six iodine atoms in the low-concentration limit. We find that I_3 and I_5 act as p-type surface dopants that shift the Fermi level 0.46 and 0.57 eV below the Dirac point, respectively. For these two molecules, molecular orbital theory and analysis of the charge density show that doping transfers electronic charge to iodine π^* molecular orbitals oriented perpendicular to the graphene sheet. For even-length polyiodides, we find that I_6 decomposes to I_2 and I_4, both of which readily desorb at 300 K. Adsorption energy calculations further show that I_3 acts as an effective catalyst for the oxygen reduction reaction on graphene by stabilizing the rate-limiting OOH intermediate.

2.1 INTRODUCTION

Despite a wide range of extraordinary features, practical applications of graphene in modern electronics remain limited due to the difficulties of controlling its electronic properties. For example, the lack of an electronic band gap restricts application in semiconductor electronics; the presence of dangling bonds on the exposed carbon atoms at the edges of graphene make samples highly reactive and prone to reactions with impurities;[68] and the low density of states near the Fermi level contributes to high sheet resistance such that even high quality graphene samples have a sheet resistance in excess of 1 kΩ/sq.[79] In order to use graphene in advanced electronics, it is necessary to develop better control of its structural[21,99,180] and electronic[17] properties. Doping provides one possible path in this direction: depending on the type of dopant, the concentration, and the synthesis method, it is possible to create band gaps ranging from 0.1 to 6.4 eV,[48,83,142] to increase conductivity,[197] to form ultra-thin insulators,[78] and to improve catalytic performance.[189]

Conventional ion implantation techniques introduce lattice defects that disrupt graphene's electronic band structure and ultimately degrade desirable properties, such as ballistic electron transport, that arise from the Dirac-like fermions[17] near the Fermi level. One way to introduce dopants without creating structural defects is to use adsorbed atoms or molecules. Iodine-doped graphene is of particular interest as a reversible p-type surface dopant: recent experiments have shown that iodine enhances conductivity[140,187] and catalytic response[3,189] making it ideal for use in optoelectronic devices as a transparent conducting film[48,73,78,137] and as a low-cost alternative to platinum catalysts in fuel cells.[189] Moreover, reversible iodine doping allows graphene to transition from a semi-metal to a metal and back without disrupting the sp^2 hybridization in pristine graphene.[187].

To date, theoretical studies suggest that iodine is adsorbed to the surface of graphene by van der Waals (vdW) interactions.[109,170,171] However, structures that contain a 1:1 ratio of C, H, and halogen X (X=F, Cl, Br, I) atoms, like CX and CHX, or a 2:1 ratio like C_2HX, are not likely to be stable at room temperature.[78] The main difference between experimental and theoretical studies has been the treatment of iodine. Thus far, all but one theoretical study considered isolated atoms of iodine or I_2 molecules interacting with the graphene sheets, whereas Raman spectroscopy and X-ray photoelectron spectroscopy data show that iodine dopants form I_3 and I_5 chains.[74,76,140,187] More recent work by Tristant and coworkers[171] focused on the formation of I_3 and I_5 on monolayer and bilayer graphene from high concentrations of I_2, as well as their geometric and vibrational properties. These authors, using molecular dynamics simulations, found vibrational spectra in good agreement with Raman spectroscopy, and from these studies concluded that I_3 and I_5 are stable on graphene near 300 K. What has not yet been been systematically explored for I_3 and larger polyiodides is the formation energy, thermodynamic stability, and charge transfer in the low-concentration limit of order 1 at. % or less found in experimental samples.[80,96,146,155,192] In particular, the doping behavior of I_4 and I_6 is unknown. These species could be formed from precursor I_2 subunits during the thermal annealing stage but are absent in experimental samples. Determining their structure and thermodynamic stability would complement the current understanding of the polyiodide-doped graphene system.

In a related, important application it is also unknown how polyiodides can serve as active sites for the oxygen reduction reaction (ORR). Recent work[69,188,193,194] demonstrated that oxygen reduction proceeds primarily through a four-electron associative pathway involving OOH adsorption on undoped and substitutionally-doped graphene. Jiao at al.[69] showed that the rate-limiting step is the weak adsorption of this OOH intermediate, which prefers regions of high spin and positive charge density on graphene clusters.[193,194] This preference is related to the fact that dopants at edge sites yield higher ORR activity than those in the bulk of a cluster. ORR activity in N-doped graphene can be improved by co-doping with adsorbed molecules such as polyiodides,[192], benzoate,[116], and 1-pyrenebutyrate[116] by a similar increase of positive charge induced by N atoms. In contrast to these covalently-bonded substitutional dopants, polyiodides are weakly-bound surface adsorbates. The mechanism for stabilizing the OOH intermediate by polyiodides is ambiguous since OOH could bind directly to adsorbed polyiodides rather than to positively-charged carbon atoms in graphene.

2.2 COMPUTATIONAL DETAILS

To explore the issues raised above we studied the electronic properties of graphene doped with adsorbed polyiodide chains and OOH using first-principles electronic structure calculations based on density functional theory (DFT). We obtain band structures and formation energies within DFT with the generalized gradient approximation exchange-correlation functional of Perdew, Burke, and Ernzerhof[138,139] and the projector-augmented wave (PAW) formalism with frozen core states, using the standard pseudopotentials in the VASP computational package.[49,85–88] We employ a 7×7 graphene supercell (containing 98 C atoms) in which polyiodides and OOH adsorbates are placed, with 24.7 Å vacuum spacing between images of the graphene sheet to minimize interactions. We determine the equilibrium geometries with the conjugate gradient relaxation method and a maximum force tolerance of 0.02 eV/Å. Convergence of total energies to within 1 meV per atom requires a 3×3×1 Monkhorst-Pack k-point sampling and a 500 eV plane-wave cutoff. We include vdW interactions using the methods of Grimme[46] (D2) and Tkatchenko-Scheffler[167] (TS). Gas phase calculations for polyiodide and OOH molecules were performed using cubic cells with side length 36 Å. For visualization of the results we use the VESTA[117] and VMD[64] graphics packages.

We calculate the formation energies, $E_f(N)$, and desorption energies, ΔE_N, for each iodine chain of size N from the expressions

$$\begin{aligned} E_f(N) &= E_{\text{tot}}(N) - E_{\text{G}} - \frac{N}{2}E_{\text{vac}}(2) \\ \Delta E_N &= E_{\text{vac}}(N) + E_{\text{G}} - E_{\text{tot}}(N) \end{aligned}$$

where $E_{\text{tot}}(N)$ is the total energy for the polyiodide-graphene system, E_{G} is the total energy of a pristine graphene sheet, and $E_{\text{vac}}(N)$ is the energy of an isolated I_N chain. Similarly, for OOH calculations we obtain the adsorption energy E_{ads} of OOH bound to system X relative to the independent systems using the expression

$$E_{\text{ads}} = E_{\text{tot}}(\text{OOH@}X) - E_{\text{tot}}(\text{OOH}) - E_{\text{tot}}(X) \tag{2.1}$$

To determine the minimum energy configurations for polyiodides, we place single iodine atoms at several high-symmetry points above the graphene lattice. To test the effects of rotation, we rotate the linear I_3 chains originating above the hollow site to form angles of 0, $\frac{\pi}{12}$, and $\frac{\pi}{6}$ radians relative to a C-C bond. To test the effects of translation, we consider a linear I_3 chain originating at the top site and running parallel to a C-C bond. We allow all these initial geometries to relax to their ground-state. The ground-state energies for all tested rotations and displacements differ by less than 2 meV, showing that there is no strong preference for the orientation of the iodine chains, and consequently we focus on a single orientation for the rest of the structures considered. For OOH adsorption, the minimum-energy geometries were found with a similar approach by testing multiple rotations and translations relative to undoped and I_3-doped graphene.

In addition to possible energy differences arising from molecular orientation, polyiodides can exhibit a wide range of structural conformations, especially when combined with salts and metallic structures[162]. We therefore consider a range of possible geometries to determine the preferred doping structure, some of which are depicted in Figure 2.1. These include: 1) linear geometries (denoted as I_N^L) for all polyiodide dopants; 2) a triangular geometry for I_3; 3) various shapes, like V, T, and dice-like for I_5, shown in Figure 2.1; 4) for I_6, two I_3^L units were placed 3.3 Å apart both collinearly, denoted as I_6^{LL}, and perpendicular to each other, denoted as I_6^T. In most cases, the geometry obtained after relaxation was nearly identical to the initial geometry. The only exceptions are the geometries for I_3, which always relaxed to I_3^L, and the geometry for I_5^L, which

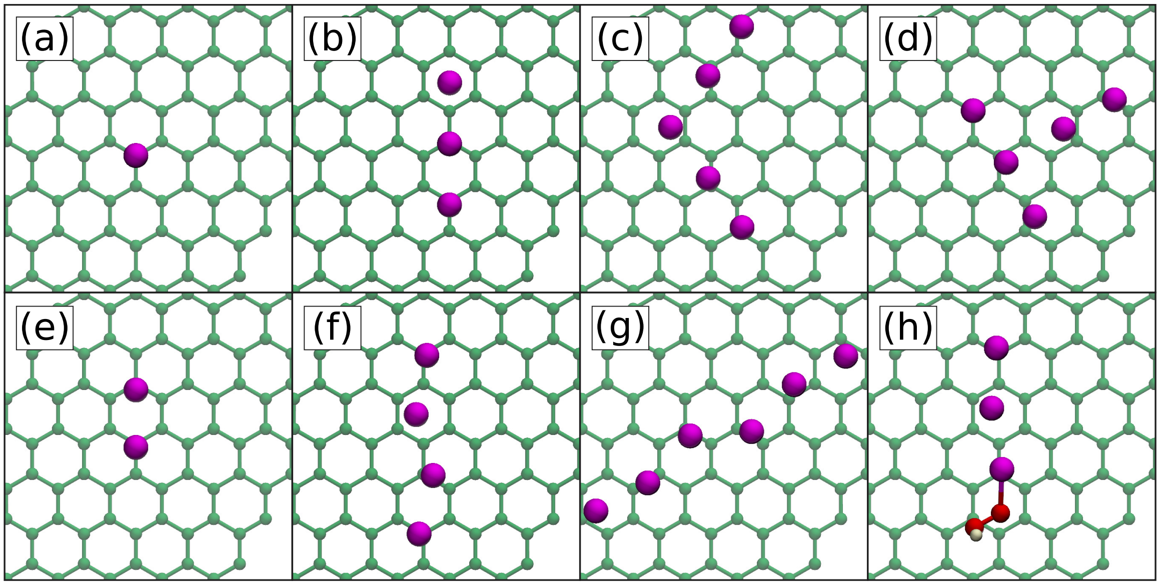


Figure 2.1: (a-c) Minimum-energy configurations for the odd-length polyiodides. (d) The next-most stable pentaiodide I_5^T. (e-g) Minimum-energy configurations for each even-length polyiodide. (h) Most stable adsorption geometry for OOH on I_3-doped graphene.

relaxed to I_5^V (see Figure 2.1(c)).

2.3 RESULTS AND DISCUSSION

2.3.1 FORMATION ENERGIES

The most stable geometries (lowest formation energy per iodine atom) are I_N^L for $N \neq 5$, and I_5^V; we denote these as simply I_N in the following. All relaxed structures are coplanar and parallel to the graphene sheet, and the formation energy generally decreases with increasing chain length: I_5 has the lowest formation energy (-0.376 eV/I) suggesting that it is the most stable polyiodide. The energy for adsorbed atomic I is positive since the formation energy is defined with reference to molecular iodine (I_2), and therefore includes the full bond dissociation energy for converting I_2 to atomic I. In all cases, the vdW corrections of either kind (D2 or TS) substantially decrease the formation energy for all chain lengths as seen in Figure 2.2(a). To investigate the iodine concentration's influence on the binding energy, we normalize by including multiple copies of the

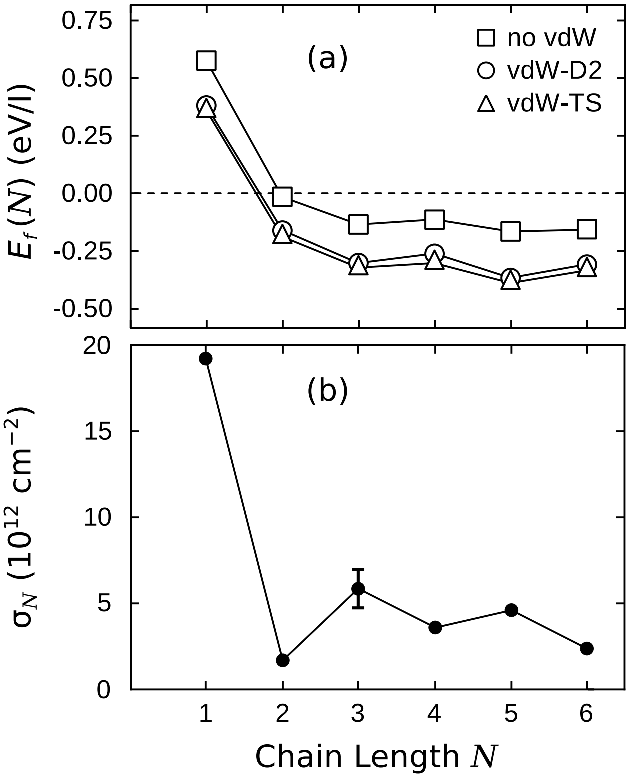


Figure 2.2: (a) Formation energy, $E_f(N)$ per iodine atom for chain length $N = 1-6$ with and without vdW corrections. (b) Induced hole density σ_N versus polyiodide chain length N.

shorter chains ($N = 1$ through 4) in the same unit cell to achieve approximately 6% and 8% doping. Higher iodine concentration increases the formation energies per iodine for I_1, I_3 and I_4, but does not affect the formation energy for I_2. The formation energy change is greater for the odd-length chains, with $E_f(1)$ increasing by 0.39 eV and $E_f(3)$ increasing by 0.06 eV, while $E_f(4)$ increases by only 0.02 eV.

The values of $E_f(N)$ and ΔE_N, along with the structural information for each case, are summarized in Table 2.1. The high total formation energies, all above -2 eV, and long iodine-graphene distances indicate only weak interactions between polyiodides and graphene. From comparing formation energies and final geometries, a key result is the tendency to form weaker bonds (about 3 Å) between I_2 subunits, a common feature among polyiodides.[162] These appear in the I_4 and I_6 geometries as well as in the I_2 subunit of I_5^T, implying weaker bonds than

Table 2.1: Formation energies per iodine atom (E_f), desorption energy (ΔE_N), desorption lifetime (τ_N), mean polyiodide-graphene distance ($d_{\text{I-G}}$), and nearest-neighbor distances ($d_{\text{I-I}}$) for the most stable length-N polyiodide. Numbers in parentheses are from Tristant et al.[170]

N	E_f (eV/I)	ΔE_N (eV)	τ_N (s)	$d_{\text{I-G}}$ (Å)	$d_{\text{I-I}}$ (Å)
1	0.370, (0.43)	0.884	1.1×10^{2}	3.60, (3.35)	17.3, (12.30)
2	−0.177, (−0.21)	0.354	1.4×10^{-7}	3.74, (3.62)	15.0, (12.30)
3	−0.312	1.150	3.3×10^{3}	3.70	12.6
4	−0.289	0.869	6.3×10^{1}	3.73	10.5
5	−0.376	1.595	1.0×10^{13}	3.68	10.2
6	−0.323	1.352	8.3×10^{9}	3.76	8.3

those found within I_3 and I_5. For the even-length chains, comparing formation energies reveals that each of these weak I_2-I_2 bonds yields approximately 0.4 eV, and is responsible for stabilizing the longer, even-length polyiodides. In contrast, the local minima for I_3 and I_5 are the result of favorable orbital hybridization, yielding shorter (~ 2.9 Å) bonds, and larger charge transfer from the graphene sheet as discussed below. Similar results for the geometries of the odd-length polyiodides were reported by Tristant et al.;[171] the lower binding energies and higher mean iodine-graphene distances reported here are consistent with the tendency of the GGA exchange-correlation functional to underbind structures relative to optB86b-vdW.[82]

The low formation energy per elongated I-I bond in I_4 and I_6 suggests low reaction barriers toward decomposing into smaller polyiodide chains. To explore this in more detail, we calculated the bond dissociation curves for I_4 and I_6 as a function of I-I bond length by abstracting an I_2 subunit from the rest of the chain. This was done by constraining the second and third iodine atoms along the chain transverse to the graphene plane while allowing all other degrees of freedom to relax, see Figure 2.3 for I_6. This is an application of the distinguished (or driven) coordinate method,[185] which provides a reasonable reaction path and transition state structure when the reaction path is dominated by the selected coordinate, in this case the length of the already-elongated 3.01 Å I-I bond. The resulting dissociation energies, 0.446 eV for I_6 and 0.418 eV for I_4, agree well with the total energy difference between the I_N precursor and the

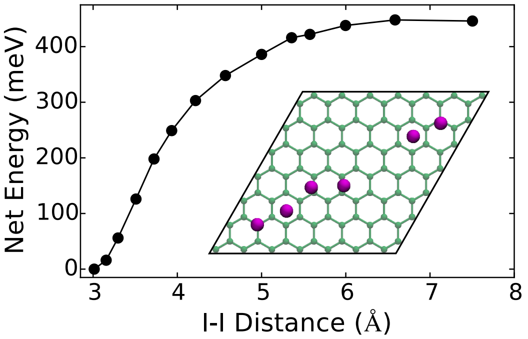


Figure 2.3: Bond dissociation curve for I_6 as a function of I_4-I_2 distance, with energies relative to the geometric ground state. The inset shows the dissociated geometry a distance of 7.50 Å.

resulting I_{N-2} and I_2 chains in separate supercells, 0.427 eV for I_6 and 0.445 eV for I_4.

2.3.2 Electronic Structure

To investigate the extent of orbital hybridization and the effect of iodine doping on graphene's electronic structure, we show results for I_2 and I_3 in Figure 2.4. The band structure and density of states (DOS) vary minimally with spin polarization and choice of vdW correction (we present only spin unpolarized results here). The presence of the Dirac cone, flat iodine bands, and the remarkable similarity between pristine graphene and the I_N-graphene band structure for all N values considered confirms that graphene's p_z-hybridization is not disrupted by the presence of iodine atoms (the differences are minimal on the scale of Figure 2.4). The Fermi level is always lower than the value of the Dirac point indicating that all polyiodide configurations considered are p-type dopants. As illustrated in Figure 2.4 for I_2 and I_3, the energy difference between the Dirac point and the Fermi level is generally higher for odd-length chains than even-length chains. The only exception is I_6, where the energy difference (0.48 eV) is substantially higher than that of I_1 (0.11 eV). This predicts greater charge transfer for odd-length polyiodides than even-length ones, as suggested in prior work for I_1 and I_2 by Tristant et al.[170]

To reveal the source of this difference, we project the DOS onto iodine-centered spherical harmonics, and find that the molecular orbitals of iodine are almost exclusively 5s and 5p hybrids.

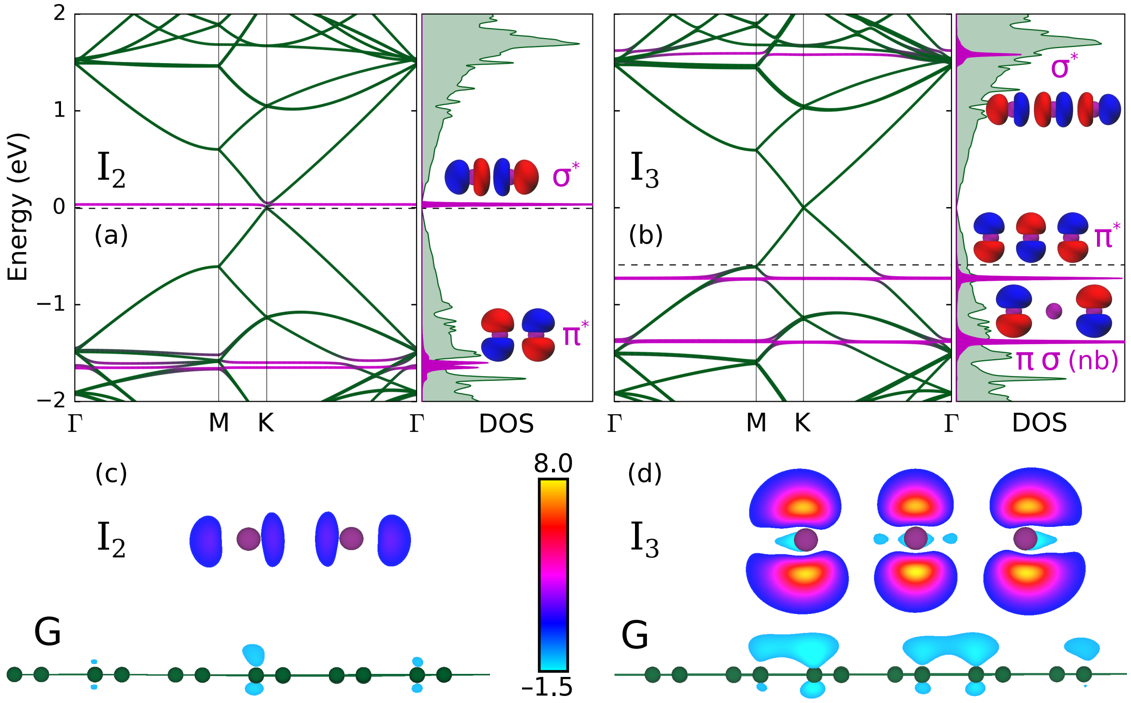


Figure 2.4: Band structures and density of states (DOS) for the minimum-energy configurations of: (a) I_2, (b) I_3. Dashed lines indicate the Fermi level, and violet bands highlight electronic contributions from iodine. The orbital diagrams show the spatial character of the corresponding iodine molecular orbitals. Charge transfer (units of 10^{-3} e Bohr^{-3}) along the chain for: (c) I_2, (d) I_3, shown on a plane perpendicular to the graphene sheet (G).

With a large 10 eV energy difference between the 5s and 5p orbitals, there is minimal s-p hybridization; for a chain of length N this yields N 5s molecular orbitals with energies centered around -13 eV, and $3N$ 5p hybrid orbitals centered around -2 eV relative to the Dirac point. The shapes and characters of the 5p-hybridized orbitals for I_2 and I_3 within the energy range of Figure 2.4(a)-(b) are shown alongside the DOS. Since all I_N in their minimum-energy configurations are coplanar and approximately linear, molecular symmetry further divides the 5p hybrids into N orbitals of σ character and $2N$ orbitals of π character. The strong interaction between 5p orbitals oriented along the chain yields larger energy gaps between σ molecular orbitals than between π molecular orbitals, as seen in Figure 2.4(a)-(b), thus explaining the difference between even and odd-length polyiodides. The odd-length neutral I_3 and I_5 chains have a lowest unoccupied molecular orbital (LUMO) with π^* character, and the small π-orbital energy gaps place these states at least 0.72 eV below the Dirac point, in good agreement with the maximum 0.8 eV Fermi level shift observed in experimental transport measurements by Wu and coworkers[187] at high iodine exposure. In contrast, the LUMO for even-length chains has σ^* character, and the larger energy gaps tend to place them much closer to the Dirac point. In the case of I_2 and I_4 this lowest unoccupied σ^* is very close to the Dirac point, while for I_6 the increasingly-dense spacing of σ^* orbitals places the LUMO somewhat lower, 0.48 eV below the Dirac point. The lone anomaly is I_1, where the lack of I-I bonding destabilizes the 5p orbitals by removing exchange and correlation effects present for $N \geq 2$. As a result, the nearly-degenerate 5p orbitals for I_1 lie only -0.11 eV below the Dirac point.

The spatial character of the transferred charge gives additional support to this molecular orbital interpretation. Figure 2.4(c) and (d) show isosurfaces of the transferred charge density for I_2 and I_3, defined as the difference between the charge density of the doped system and the sum of densities of the isolated, neutral polyiodide and undoped graphene sheet with the same atomic positions. The I_3 and I_5 chains primarily gain charge in π^* orbitals, while I_1 and the even-length chains gain charge in σ^* orbitals parallel to the graphene sheet. The spatial distribution of excess charge on the iodine chains agrees well with the density of the corresponding LUMO: note, for example, the similarity for I_3 between Figure 2.4(d) and the corresponding π^* LUMO that accepts the charge in Figure 2.4(b).

2.3.3 CHARGE TRANSFER

The charge transfer between iodine and graphene offers additional insight into the role of iodine dopants in catalysis and electronic devices. We employ Bader charge analysis[55,151,164] to determine how much charge the polyiodide dopants acquire in their optimal configurations. We obtain the net charge transfer Q_N by finding the total charge Q^B_{tot} contained in Bader volumes attributed to the iodine atoms, then subtracting the seven valence electrons per iodine atom to obtain the net charge. For a more direct comparison to experiment, where the iodine concentration is generally of order 1 at. %, the hole density is extrapolated to 1.0 at. % iodine doping by dividing the net charge Q_N by the corresponding area of graphene A_G, giving a hole density σ_N:

$$\sigma_N = (Q^B_{\text{tot}} - 7N)/A_G = Q_N/A_G$$

We show the results of this calculation in Figure 2.2(b). All three odd-length polyiodides draw more charge than the even-length polyiodides, generally consistent with the lower energy of the unoccupied π^* orbitals relative to the lowest unoccupied σ^* orbitals, and yield the highest hole densities induced on graphene. I_1 still draws more charge than I_6, even though the the dopant band is farther below the Dirac point for I_6 as mentioned above. Estimating charge transfer and hole density from the band structure alone, by integrating the DOS resulting from graphene's linear dispersion near the Dirac point,[17,170] would therefore predict the opposite order between I_6 and I_1.

The Hirshfeld partial charges,[60] calculated as part of the TS vdW correction, are qualitatively consistent with the Bader charges so the latter are expected to be more reliable. We also assess the magnitude of finite size effects by calculating the hole density for a pair of I_3 dopants. The error bar in Figure 2.2 indicates the difference: the $2\times I_3$ structure yields a lower hole density than a single I_3. Charge transfer for larger N may be underestimated compared to calculations performed with actual 1 at. % doping.

The good correspondence between charge transfer and the character of molecular orbitals explains the conductivity-enhancing effects of I_3 and I_5. With the weak π hybridization in these species, their π^* LUMO is lower than the Dirac point, and can therefore yield significant hole densities in graphene with both I_3 and I_5 yielding hole densities of approximately 5×10^{12} cm^{-2}. This value is in reasonable agreement with a range experimental results, which demonstrate 5×10^{12} cm^{-2} at 4 at. % I,[146] 2.6×10^{13} cm^{-2} at an estimated 1 to 1.7 at. % I,[73] and 4.75×10^{13} cm^{-2}.[187] In contrast, the even-length polyiodides have σ^* LUMOs that are much closer to the Dirac point due to the stronger 5p hybridization along the chain. The net charge transfer to these less-stable LUMOs is therefore reduced. The DC conductivity of graphene is proportional to its net charge density,[17] whether this density is created by an applied gate potential, impurities and adsorbates,[71,187] or even by photo-generated carriers.[81] For I_3 and I_5, the significant induced hole density in graphene agrees well with the observation that polyiodide doping significantly increases the conductivity.[187] These results are also consistent with the observation that polyiodide doping improves the conductivity of carbon nanotubes.[197]

2.3.4 POLYIODIDE STABILITY

An open question is the reason for the abundance of I_3 and I_5 reported in the experimental literature. Although we find I_5 to be the most stable per I atom, the formation energy of I_3 is only $0.9\,k_\mathrm{B}T$ lower per I atom than I_4, and is $0.4\,k_\mathrm{B}T$ higher per I atom than I_6 at 300 K. Formation energies alone cannot explain why I_3 is the majority polyiodide dopant reported by experiments. An observation that may help explain this is related to the hole densities: I_3 and I_5 show the largest charge transfer per adsorbed molecule, suggesting that electrostatics play a role in keeping these polyiodides bound to graphene. To explore this idea further, we obtain the desorption energy ΔE_N for each chain length to determine the stability of each I_N species. To estimate the corresponding desorption rates k_N and lifetimes k_N^{-1} we use the Eyring equation:[32] $k_N = (k_B T/h)\, e^{-\Delta E_N/k_B T} e^{\Delta S/k_B}$. The ΔS_N term in the full Eyring equation is assumed to be negligible since the polyiodides are only weakly bound to graphene, so that both adsorbed and free polyiodides have nearly identical vibrational modes; the only major difference is a hindered rotation into the graphene sheet for adsorbed chains. As a result, ΔS_N should be negligible and is unlikely to vary with N. Table 2.1 shows the resulting desorption energies and lifetimes. The desorption barriers are highest for I_3, I_5 and I_6, resulting in orders of magnitude longer desorption lifetimes that are in line with their larger charge transfer and dispersion interactions with graphene. Although these calculations neglect solvation and dielectric screening effects that would be present in an experimental setting, the relative rates presented here provide strong evidence that the high desorption barriers of I_3 and I_5 play a crucial role in their observed abundance.

Combined with the small bond dissociation energies for I_4 and I_6, the weak adsorption of I_2 explains the absence of the even-length polyiodides studied here, and likely all longer-length polyiodides as well. Any I_2 molecules remaining after the preparation of polyiodide-doped graphene should quickly desorb from the surface at 300 K given its low desorption barrier of 0.354 eV. For I_4 and I_6, although the desorption barriers are significantly higher, the corresponding dissociation barriers for I_2 abstraction are not. For example, the 0.445 eV dissociation barrier for I_4 to form a pair of I_2 molecules corresponds to a lifetime of only 4.8×10^{-6} s from the Eyring equation. Therefore I_4 and I_6 are expected to rapidly decompose toward I_2 at 300 K. For even larger polyiodides I_N with $N \geq 7$, these generally feature similar elongated I-I bonds between some combination of I_2, I_3, and I_5 subunits.[162] Assuming the corresponding bond dissociation energies to be similar to those for I_4 and I_6, the observed lack of large polyiodides can be explained by repeated thermal decomposition toward the shorter polyiodides considered here.

2.3.5 MECHANISM OF CATALYSIS

In addition to reversibly doping and increasing the conductivity of graphene, polyiodides have attracted recent interest for their role in making graphene an effective catalyst for oxygen reduction. The work of Jiao et al.[69] shows that the weak binding of the OOH intermediate presents a large free energy barrier of 1.1 eV, yielding slow ORR kinetics in undoped graphene. The presence of dopants can significantly increase the binding energy of OOH and therefore reduce the corresponding barrier. Of the two observed polyiodide dopants, we focus on I_3 since it has been observed to be the dominant active species in polyiodide-doped graphene.[189] To determine how strongly I_3 stabilizes the OOH intermediate, we find the minimum-energy adsorption geometry for OOH for undoped graphene (G), I_3-doped graphene (I_3-G), and for both neutral I_3 and charged I_3^- molecules in the gas phase. The adsorption energies, distances, and Bader charge values are shown in Table 2.2. On undoped graphene, OOH is only weakly adsorbed with a minimum O-graphene distance of 2.95 Å, and the molecular plane is nearly parallel to graphene. Similar to polyiodides, OOH adsorption energies vary negligibly with translations and rotations. The minimum-energy adsorption geometry for I_3-doped graphene is shown in Figure 2.1(h), and has a significantly stronger adsorption energy as a result of forming an O-I bond. The minimum O-graphene distance actually increases slightly in the presence of I_3, which shows that the lower adsorption energy comes from bonding to I_3 rather than a carbon atom in graphene. Compared to substitutional B-doped graphene[69] and N-doped graphene,[69,194] I_3 binds OOH by about -0.13 eV more and should therefore correspond to a smaller free energy barrier of 0.57 eV. This significant lowering

Table 2.2: Adsorption energy (E_{ads}), minimum O-graphene distance ($d_{\text{O-G}}$), O-I bond length ($d_{\text{O-I}}$), and net OOH Bader charge (Q_{OOH}) for each system interacting with OOH.

system	E_{ads} (eV)	$d_{\text{O-G}}$ Å	$d_{\text{O-I}}$ Å	Q_{OOH} (e)
G	−0.441	2.95		0.26
I_3-G	−0.968	3.06	2.23	0.35
I_3 gas	−1.440		2.12	0.33
I_3^- gas	−0.569		2.33	0.44

of the free energy barrier for OOH adsorption, due to direct bonding to I_3 dopants, explains how I_3 can serve as the active site for OOH adsorption and hence promote the early stages of oxygen reduction.

There is also an interesting relationship between OOH adsorption and the corresponding charge transfer, where the stronger binding of OOH on I_3-doped graphene is accompanied by a substantial increase in its partial charge. For reference, we compare the OOH adsorption strength to neutral I_3 and the corresponding I_3^- anion in the gas phase. I_3 yields significantly stronger OOH adsorption than I_3^-, see Table 2.2. I_3 on graphene is roughly halfway between neutral and anionic with a net Bader charge of 0.55 e, and its corresponding OOH adsorption energy and O-I bond length are between those of neutral I_3 and anionic I_3^- in the gas phase. This close agreement reflects the weak I_3-graphene interaction, and suggests that the reactivity of polyiodide-doped graphene could be improved by decreasing the net charge on the polyiodide dopants, for example by co-doping graphene with an electronegative element. This may explain the recent observation[192] that co-doping graphene with both polyiodides and N yields faster ORR kinetics than either dopant alone, despite lower I and N loading than the individual samples.

2.4 CONCLUSION

In summary, we studied the electronic properties of iodine-doped graphene to understand the possibility of using polyiodides as dopants (I_1 through I_6). To foster comparisons between these different polyiodides, we calculated the binding energies, electronic band structures, densities of states, charge transfer, and desorption rates for each polyiodide on graphene. We find that, in general, longer chain lengths and lower concentrations have more favorable formation energies, with little dependence on the relative orientation of the polyiodide molecule and the graphene lattice. Band structure calculations and molecular orbital theory reveal that polyiodide molecules act as p-type dopants and do not noticeably disrupt graphene's band structure. There are also clear distinctions in doping behavior among the different chain lengths as a result of orbital hybridization: the I_3 and I_5 chains acquire charge in π^* orbitals that are oriented with lobes perpendicular to graphene, while I_1 and even-length chains acquire charge in orbitals with lobes parallel to graphene. As a result of weak π hybridization and thus lower-energy LUMOs, the odd-length chains exhibit 25% greater charge transfer per iodine atom. This increased charge helps keep the experimentally-observed I_3 and I_5 species stably bound to graphene, while I_2 rapidly desorbs at 300 K. Furthermore, although I_4 and I_6 are reasonably stable with respect to desorption, they are unstable with respect to dissociation into separate I_2 molecules as a result of their elongated I-I bonds and low bond dissociation energies. This explains the observed lack of longer polyiodides I_N ($N \geq 7$), which generally feature longer bonds between shorter polyiodide subunits. We also find that the rate-limiting OOH intermediate is effectively stabilized by adsorbed I_3 chains, resulting in a lower free energy barrier that agrees with the fast ORR kinetics observed for polyiodide-doped graphene.

3

Anhydrous Alcohol Dehydrogenation on Cu(111)

Apart from minor modifications and Section 3.5, this chapter is being prepared for the following publication:

> **Robert A. Hoyt**, Matthew M. Montemore, Robert J. Madix, E. Charles H. Sykes, and Efthimios Kaxiras. “Anhydrous Methanol and Ethanol Dehydrogenation at Cu(111) Step Defects.” *in preparation*

Apart from minor modifications, content in Section 3.5 appeared in the following publication:

> Zhi-Tao Wang, **Robert A. Hoyt**, Mostafa El-Soda, Robert J. Madix, Efthimios Kaxiras, and E. Charles H. Sykes. “Dry Dehydrogenation of Ethanol on Pt-Cu Single Atom Alloys.” *Top. Catal.* **1** (2017)

ABSTRACT

Oxidative methanol dehydrogenation is a major industrial reaction with global formaldehyde production exceeding 30 million tonnes per year. Unfortunately, oxidative dehydrogenation produces water-aldehyde mixtures that require subsequent distillation. Anhydrous alcohol dehydrogenation is a promising alternative that produces H_2 instead of water. Pursuant to recent experimental work showing that highly stepped Cu(111) surfaces exhibit anhydrous dehydrogenation activity, we present first-principles density functional theory calculations for methanol and ethanol dehydrogenation at Cu(111) step edges to provide an atomistic understanding of the catalytic mechanism; these sites stabilize all intermediates while reducing activation energies. We find that van der Waals contributions to the energy account for more than 50% of adsorption energies, and their inclusion is essential in achieving good agreement with

experimental desorption temperatures. Furthermore, vibrational zero-point energy corrections significantly reduce the activation energy for all reaction steps considered here. Hydrogen bonding among ethanol intermediates at step edges is weakened by geometric frustration. These insights lead us to propose several ways for further improving undercoordinated Cu sites as anhydrous alcohol dehydrogenation catalysts.

3.1 INTRODUCTION

Improving the performance of industrial catalysts requires increasing their activity and selectivity while maintaining long catalyst lifetime. Alcohol dehydrogenation on heterogeneous catalysts is the most significant industrial pathway for formaldehyde production,[35,160] with global production exceeding 30 million tonnes per year. Current catalysts, either based on silver or oxides of iron, molybdenum and vanadium[38] require oxygen to facilitate alcohol oxidation and therefore produce water as a byproduct. Separating formaldehyde and unreacted methanol from water requires energy-intensive absorber and vacuum distillation stages.[35,199] If catalysts were able to function in the absence of oxygen, industrially valuable H_2 would be produced instead of water. Such catalysts are not commercially available yet, but further improvements to promising candidates may lead to better prospects. Cu is particularly promising due to its low cost, high selectivity,[160] and long catalyst lifetimes.

Substantial experimental work has addressed alcohol dehydrogenation on Cu surfaces. Much of this research has focused on oxidative dehydrogenation over single-crystal surfaces, starting with Cu(110)[12] and Cu(111).[150] The consensus is that O–H bond breaking occurs at low temperatures to produce stable alkoxy intermediates, followed by reaction rate-limited aldehyde desorption at higher temperatures. Hydrogen adatoms generated from breaking methoxy C–H bonds can also lead to methoxy hydrogenation and reaction rate-limited methanol desorption.[12] Detailed adsorbate-surface distance measurements for methanol and methoxy intermediates at low temperature have been made with X-ray standing wave adsorption experiments.[72] More detailed results for adsorbate geometry and induced surface relaxation are also available for methoxy on the Cu(111) surface.[61] Under ultra-high vacuum (UHV) conditions, defect-free Cu single-crystal surfaces (except Cu(210)[18]) are inert toward alkoxy dehydrogenation[11,12] such that a co-catalyst is required to facilitate C–H bond breaking. Early studies employed adsorbed oxygen atoms as the co-catalyst and produced water as a byproduct.[11,12,150] Later it was demonstrated that water itself serves as a co-catalyst for methoxy dehydrogenation to formaldehyde and H_2[10,153] through water-alcohol and water-alkoxy hydrogen bonding. In an effort to remove water from the reaction cycle entirely, more recent work on Cu(111) surfaces has demonstrated that surface defects, either isolated dopant metal atoms[153,181] or step edges,[182] can supplant water as an effective co-catalyst to facilitate anhydrous alcohol dehydrogenation. These defect studies also observed O–H bond breaking at low temperatures followed by C–H bond breaking at higher temperatures.

Several previous theoretical studies based on density functional theory (DFT) examined a variety of methanol dehydrogenation intermediates at flat and stepped Cu(111) surfaces. Greeley and Mavrikakis[45] found that methanol adsorbs to Cu top sites through its O atom, while formaldehyde adsorbs flat on Cu(111). They also find the same reaction mechanism as observed in experiments. A related study on 22-atom Cu(111)-terminated clusters[42] reported qualitatively similar energetics, but found that methanol adsorbs at top sites through the alcohol H atom, and that formaldehyde adsorbs perpendicular to the surface though weak O–Cu and H–Cu bonds. Methanol dehydrogenation intermediates have also been studied in the context of methanol synthesis with DFT and van der Waals (vdW) corrections on the stepped Cu(211) surface.[6,159] Step edges stabilized all adsorbates and improved reactivity, while vdW interactions were significant for higher molecular weight adsorbates, particularly those derived from CO_2 rather than CO. While

Cu(211) surfaces expose (100) steps, there also exist (111) steps, as discussed below. The (111) step has been studied with DFT for formaldehyde steam reforming using the Cu(221) surface,[100] where formaldehyde was found to adsorb above and parallel to the step edge.

The corresponding ethanol dehydrogenation pathway and its intermediates have been studied as well. A study of ethyl acetate synthesis on Cu(111) surfaces[98] found that ethanol adsorbs to the terrace through its alcohol H atom, while another study of ethanol adsorption found that it adsorbs to Cu(111) surfaces through its O atom instead.[165] Further work on CO hydrogenation to ethanol on Cu(211) surfaces found that all intermediates besides H are more stable at step edges,[196] and that H is most stable at three-fold hollow sites immediately below the step edge, while an earlier study[100] found that H is most stable at three-fold hollow sites immediately above the step edge. Although the predicted reaction mechanism for alcohol dehydrogenation agrees with experimental results, there is significant disagreement regarding adsorption geometries for alcohols and aldehydes.

To address the current lack of direct comparison between methanol and ethanol and between both Cu(111) step edges, as well as the lingering discrepancies among existing computational studies, we report here the adsorption geometries and energies for all key methanol and ethanol dehydrogenation intermediates on the Cu(111) terrace and its step edges. We take into account vdW interactions for all adsorbates and transition states as they have been shown to qualitatively alter adsorption energies and geometries for dehydrogenation intermediates on Cu(111) surfaces[159,165] and other organic molecules on Au(111) and (100) surfaces.[77] Furthermore, while alcohol-alcohol hydrogen bond energies are significant in the gas phase[34] and on transition metal surfaces,[10,93,112,153] hydrogen bonding behavior has not been studied at Cu(111) step edges. We also include vibrational zero point energy (ZPE) corrections for all intermediates to provide a more realistic description.

Our theoretical comparison between methanol and ethanol can help identify trends that extend toward larger alcohols like n-propanol, and the difference in behavior at step edges versus on the Cu(111) terrace can reveal how local surface structure contributes to adsorbate stability and affects reaction rates. These detailed calculations elucidate the importance of step edges and contribute to a better understanding of catalyst design principles.

3.2 COMPUTATIONAL DETAILS

3.2.1 DFT AND MD PARAMETERS

Density functional theory (DFT) calculations were performed using VASP version 5.4.2[85–88] with the exchange-correlation functional of Perdew, Burke, and Ernzerhoff (PBE).[138,139] To achieve precisions of 0.01 eV for minimum-energy geometries we use plane wave cutoffs of 400 eV and relax geometries until the maximum force component is less than 0.01 eV/Å in magnitude using the FIRE method[8] from the VASP Transition State Tools (VTST) package. The Cu(111) surface is modeled with a 3×3 supercell, while the (111) and (100) step edges are modeled using three-atom-wide supercells of the (553) and (533) surfaces, respectively. All unit cells contain six layers in the (111) direction, with the bottom two layers fixed in their bulk positions and with a bulk-optimized lattice constant of 3.547 Å. To minimize interactions between periodic slab images we ensure that at least 10 Å of vacuum separates atoms in adjacent slabs, including any adsorbates, and we apply dipole energy corrections perpendicular to the slab. Monkhorst-Pack k-point meshes[134] are selected to achieve similar sampling densities across all three surfaces. With the convention $\vec{A}_i \cdot \vec{b}_j \equiv \delta_{ij}$ for lattice vector $\vec{A}_i$ and reciprocal lattice vector $\vec{b}_j$, convergence is obtained using $m_i = |\vec{b}_i| \cdot 50$ Å points along lattice vector $\vec{A}_i$, rounding m_i to the nearest integer. This yields an 8×8 grid for the (111) surface, a 7×5 grid for the (111) step, and a 7×6 grid for the (100) step. For inter-adsorbate interactions, 4×4 Cu(111) supercells and four-atom-wide step surfaces were used with the same k-point density and otherwise-identical settings. Gas phase calculations use cubic unit cells with a side length of 30 Å. Molecular dynamics simulations use four total layers, plane wave cutoffs of 300 eV and $m_i = |\vec{b}_i| \cdot 35$ Å. These reduced values preserve most adsorption energies within 0.05 eV and yield very similar adsorbate geometries relative to the converged parameters. All energies include zero-point vibrational energy (ZPE) corrections, using 0.005 Å centered differences, and vdW corrections calculated with the Tkatchenko-Scheffler (TS) method.[167] All transition state geometries and energies are obtained using the dimer method,[56,59] as implemented in VTST, which improves the standard VASP version by rotating the dimer regardless of curvature to avoid pathological trajectories. Searches are initiated from several rotations and translations of the lowest-energy hydrogenated precursor with the cleaved O–H or C–H bond pre-stretched to approximately 1.3 Å to reduce its curvature. Dimer geometries are optimized using conjugate gradient minimization until the largest force is less than 0.02 eV/Å in magnitude. Reactant and product geometries are determined by displacing each transition state by ±0.25 Å along the unstable mode, then relaxing to the nearest local minimum using the conjugate gradient method. Finally, molecular dynamics (MD) simulations used a time step of 1 fs and a Nose-Hoover thermostat set to 150 K.

3.2.2 STRUCTURES AND ENERGETICS

The Cu(111) surface has two low index step facets as illustrated in Figure 3.1, with (111) or (100) orientations. We refer to undercoordinated atoms at step edges as “step atoms,” and to the bridge sites between them “step bridges.” Step atoms at both step edges are 7-fold coordinated and have identical radial distribution functions. The (111) and (100) step edges meet at 60angles to form kink sites with either 6-fold coordination (lower left in Figure 3.1) or 8-fold coordination (upper right in Figure 3.1). Although these kinks most likely have different reactivities, we did not consider them in detail since scanning tunneling microscopy (STM) images show that adsorbates

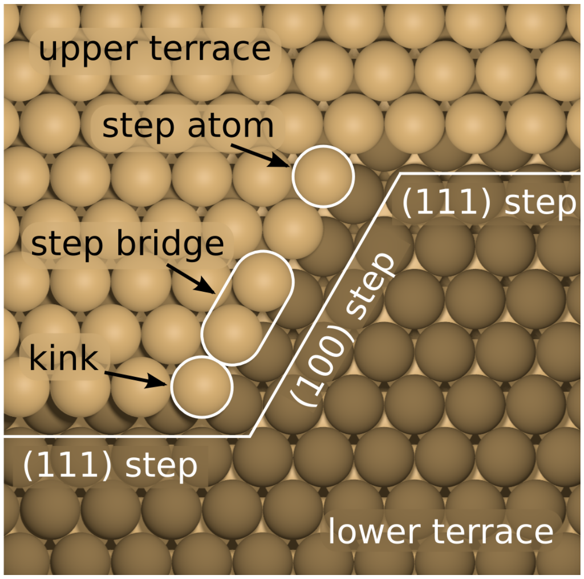


Figure 3.1: Primitive step features on the Cu(111) terrace. Cu atoms in the upper terrace are shaded light tan, while Cu atoms in the lower terrace are shaded dark tan.

are found along the entirety of step edges. Preliminary calculations also show that adsorption energies at 6-fold-coordinated kink sites are stronger, but their O–H transition states have higher energies so they are predicted to be less reactive. Both step types are experimentally stable and similarly prevalent on the Cu(111) surface throughout the alcohol dehydrogenation reactions and above 400 K. Furthermore, while adatom and vacancy defects at step edges are also possible, we calculate that the formation energy of an adjacent adatom-vacancy pair is significantly endothermic relative to clean step edges: 0.55 eV for the (111) step, and 0.65 eV for the (100) step. Adatoms and vacancies increase alcohol adsorption energies by about 0.1 eV compared to clean steps so adsorbates cannot stabilize these adatom-vacancy pairs. O–H transition states also have higher energies on adatoms and defects than on clean step edges. As a result we did not consider such defects in detail.

Adsorption energies are significantly increased by vdW interactions, especially for the larger ethanol dehydrogenation intermediates where the extra CH_2 subunit typically contributes an additional 0.18 eV to adsorption energies. Table 3.1 shows the vdW contribution to alcohol and

Table 3.1: Percent Contribution f_{vdW} of TS vdW Corrections to PBE+vdW Desorption Energies

Surface	MeOH	EtOH	CH_2O	CH_3CHO
(111) step	54	62	60	58
(100) step	61	66	53	64
terrace	79	87	111	111

aldehyde desorption energies, defined as the fraction f_{vdW} of the total adsorption energy due to the TS vdW correction

$$f_{\mathrm{vdW}}(A) = \left(E_{\mathrm{TS}}(A/X) - E_{\mathrm{TS}}(A) - E_{\mathrm{TS}}(X)\right)/E_{\mathrm{des}} \tag{3.1}$$

where E_{TS} is the TS vdW contribution to the PBE+vdW energy for intermediate A adsorbed to surface X or in the gas phase, and E_{des} is the PBE+vdW desorption energy. All values are taken from the same PBE+vdW calculation. TS vdW corrections account for at least 54% of desorption energies at step edges, and at least 79% on the flat Cu(111) terrace. This is largely due to variations in the PBE contributions since the TS vdW corrections for each adsorbate are similar across all three sites. f_{vdW} exceeds 100% for aldehydes on the terrace; aldehyde-surface vdW interactions are so dominant they result in geometries with a negative PBE contribution. Undercoordinated step atom and bridge sites provide stronger adsorbate-Cu interactions and hence reduce the relative contributions of vdW interactions at step edges. Our binding energy and f_{vdW} values for ethanol on the Cu(111) surface are in reasonable agreement with previous results[165] using the similar D3 vdW correction scheme.[47] For methanol and formaldehyde we obtain significantly larger adsorption energies for the same adsorption geometries reported by Greeley and Mavrikakis[45] without vdW corrections. Similar increases in stability due to vdW corrections are also seen in previous work on Cu(110),[20] though with a smaller typical trend of 0.14 eV/CH_2 due to the Cu(110) surface's lower atomic density at the surface.

In addition to increasing adsorbate stability, vdW contributions also affect adsorption geometries. The TS vdW corrections are exclusively attractive and therefore favor smaller adsorbate-surface distances. Previous comparisons of the PBE and PBE+vdW geometries of ethanol on Cu(111)[165] using D3 vdW corrections found that the C–C bond is approximately perpendicular to the surface with PBE alone, but is nearly parallel to the surface with PBE+vdW. Similarly, our alcohol adsorption geometries on the Cu(111) terrace feature slightly shorter adsorbate-Cu distances than previous methanol geometries with O–Cu bonding,[45] and much shorter than previous methanol[42] and ethanol[98] geometries with H–Cu bonding. The latter H–Cu adsorption geometries are unstable with PBE+vdW and relax to the O–Cu geometry. Force components were generally less than 0.1 eV/Å during the relaxation from H–Cu bonding to O–Cu bonding, demonstrating the importance of using stringent force criteria when working with weakly bound adsorbates. Adsorption geometries for aldehydes are similarly affected. Several previous studies without vdW corrections found that formaldehyde[42,43] and acetaldehyde[195] adsorb perpendicular to the surface with relatively large adsorbate-Cu distances, although flat formaldehyde adsorption geometries were only 0.02 eV less stable.[43] In contrast, perpendicular aldehyde geometries are unstable in PBE+vdW calculations and relax to lie flat on the terrace. Adsorption at step edges is similarly affected. Previous ethanol and acetaldehyde geometries at (100) steps[196] exhibited large angles between the C–C bond axis and the the upper terrace without vdW corrections, while the PBE+vdW geometries have C–C bonds parallel to the lower terrace. Similarly, vdW corrections reduced adsorbate-surface distances for alcohol dehydrogenation intermediates on Cu(110) surfaces as well.[20]

Another effect of vdW corrections is a reduction in the bulk-optimized lattice constant. Our PBE+vdW calculations yield a lattice constant of 3.55 Å, 1.9% lower than the experimental value of 3.615 Å. Conversely, GGA-only calculations tend to overestimate lattice constants.[26] The PW91 functional predicts 3.66 Å for Cu,[45] and PBE predicts 3.67 Å.[126] Compared to the latter PBE-only value, the PBE+vdW value is equivalent to a 3.3% isotropic compression of the Cu lattice and

Type	Site	Alcohol	Aldehyde	O–H	C–H
MeOH	(111) step	0.03	0.04	−0.21	−0.20
	(100) step	0.04	0.04	−0.21	−0.20
	terrace	0.04	0.01	−0.22	−0.21
EtOH	(111) step	−0.06	0.03	−0.22	−0.18
	(100) step	−0.06	0.03	−0.21	−0.22
	terrace	−0.06	0.03	−0.22	−0.23

Table 3.2: Contribution of Zero-Point Vibrational Energy to Desorption and Activation Energies in eV

should therefore affect its electronic and catalytic properties. For example, Greeley and Mavrikakis previously investigated the impact of a 4% lateral expansion on methanol dehydrogenation[45] on Cu(111) and found that expansion reduced the C–H activation energy by 0.06 eV and increased the stability of the methoxy intermediate by 0.18 eV. Similar effects are anticipated in the present study.

Zero-point vibrational energy can significantly alter potential energy landscapes, especially for activation energies of bonds with high vibrational frequencies. O–H and C–H bonds have typical frequencies around 3000 cm^{-1} and therefore contribute $\hbar\omega/2 \sim 0.2$ eV. Thus, ZPE corrections are expected to be significant in these cases since cleaved bonds do not contribute to transition state ZPE corrections. For adsorbates, ZPE corrections can still be significant since molecular normal modes may change as a result of adsorbate-surface interactions. We therefore calculated ZPE corrections for all reaction intermediates and gas phase species. Table 3.2 shows the resulting ZPE corrections to the PBE+TS desorption and activation energies for alcohols and aldehydes. ZPE corrections are small for desorption energies, where adsorbate-surface interactions are too weak to significantly alter molecular normal modes. The total correction is at most 10% across all 12 adsorbate-site combinations. On the other hand, ZPE corrections reduce activation energies by about 0.21 eV for all dehydrogenation reactions. This value is similar to the one predicted by the vibrational frequency of bonds which is approximately 3000 cm^{-1}. Comparing the stable normal modes of reactant and transition state geometries reveals no other significant differences for each reaction. Neglecting the approximately 0.21 eV ZPE corrections for bonds would therefore increase activation energies by up to 40% at step edges.

3.3 DEHYDROGENATION INTERMEDIATES

Steps increase the stability of all intermediates on the Cu(111) surface. Minimum-energy adsorption geometries at each site are shown in Figure 3.2 and their O–Cu bond lengths are shown in Table 3.3. Although both steps involve similar geometries for each intermediate, the reduced 7-fold coordination of step atoms leads to qualitatively different adsorption geometries compared to the flat Cu(111) terrace. All adsorbate-Cu bonds involve the O atom on all three surfaces, although the O–Cu coordination changes throughout the dehydrogenation process. A major finding is the similarity in adsorption and reactivity between methanol and ethanol intermediates. Both alcohols and their corresponding alkoxys have the same O–Cu coordination and very similar O–Cu bond lengths, as well as similar orientations relative to the surface. The corresponding aldehydes have weak preferences for different O–Cu coordinations, where formaldehyde prefers step bridges and acetaldehyde prefers step atoms.

Adsorbate	terrace	(111) step	(100) step
MeOH	2.23	2.17	2.22
EtOH	2.32	2.14	2.17
MeO	2.03	1.95	1.95
EtO	2.02	1.95	1.95
CH_2O	$2.99^{\perp}$	2.08	2.12
CH_3CHO	$3.06^{\perp}$	2.07	2.07

All distances are in Å.

$^{\perp}$ Distance from O atom to the Cu(111) surface plane

Table 3.3: O–Cu Bond Distances for Minimum-Energy Adsorbate Geometries

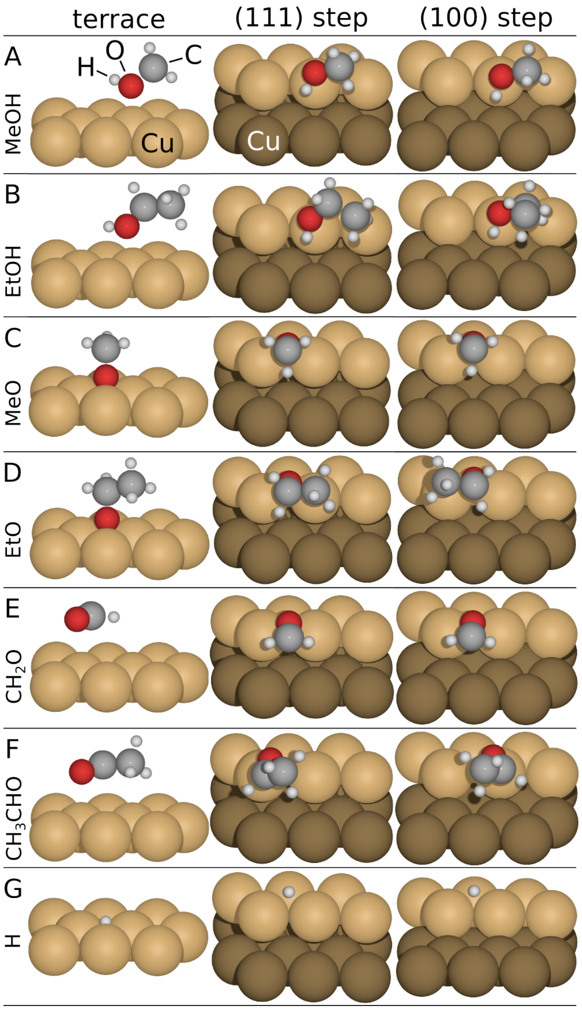

terrace
(111) step
(100) step
A
MeOH
H
O
C
Cu
Cu
B
EtOH
C
MeO
D
EtO
E
CH_2O
F
CH_3CHO
G
H

Figure 3.2: Minimum-energy adsorption geometries for isolated methanol and ethanol dehydrogenation intermediates. Cu atom colors indicate their position relative to the step: light tan for the upper terrace, and dark tan for the lower terrace.

3.3.1 ALCOHOLS

Methanol and ethanol prefer dative O–Cu bonding through one of the oxygen lone pairs to top sites on every surface facet. As shown in rows A and B of Figure 3.2, methanol and ethanol have similar geometries and adsorb through binding to the O atom. Ethanol is more strongly bound on average by -0.18 eV, due to vdW interactions between its additional CH_2 subunit and the Cu surface. At steps, adsorption is significantly stronger at the undercoordinated step atoms. This yields shorter O–Cu distances at step sites, 2.14 to 2.24 Å, versus the terrace values of 2.30 Å for methanol and 2.32 Å for ethanol. Since the O-bonded Cu atom relaxes somewhat toward the alcohol, the average vertical distance between the O atom and the Cu(111) surface layer is slightly larger, 2.39 Å, for both alcohols, in fair agreement with the experimental value of 2.69 ± 0.19 Å for methanol.[72] Alcohol adsorption energies vary by at most 0.04 eV between the (111) and (100) step edges for both alcohols, indicating negligible preference for either step edge. Alcohols on the terrace feature O–H bonds approximately parallel to the surface and O–C bonds oriented at angles 36-37from the surface. Alcohol geometries at steps have O–H bonds approximately perpendicular to the step edge and all carbon atoms are suspended over the lower terrace. The internal coordinates of adsorbed molecules are nearly identical to their gas phase values for all alcohol and step combinations. The only exception is ethanol in specific conformers. In the gas phase, there is negligible preference for trans over gauche ($<$ 0.01 eV difference in energy). On the terrace, the preference is more significant and the trans conformer is more stable than the gauche conformer by 0.06 eV. In contrast, trans geometries are unstable at step edges and relax to the gauche conformer.

3.3.2 ALKOXYS

Alkoxys have significantly different adsorption geometries on the terrace and steps as shown in rows C and D of Figure 3.2. Alkoxys on the terrace prefer fcc hollow sites, forming three O–Cu bonds of length 2.03 Å, and feature an O–C bond nearly perpendicular to the surface. We find that the O atom in methoxy and ethoxy both have average vertical distances from the Cu(111) surface layer of 1.39 Å, in good agreement with a previous experimental value of 1.31 ± 0.06 Å,[72] and this previous study determined that methoxy adsorbs to both fcc and hcp hollow sites with negligible preference between the two. Earlier photoelectron diffraction experiments[61] found a similar average vertical distance of 1.32 ± 0.05 Å from the methoxy O atom to its three nearest Cu neighbors, which agrees well with our value of 1.35 Å for both methoxy and ethoxy. Moreover, the latter study found that the methoxy C–O bond is perpendicular to the surface with a length of $1.42^{+0.10}_{-0.03}$ Å, in excellent agreement with the minimum-energy orientation and C–O bond length of 1.42 Å. In contrast to the terrace, alkoxys at step sites occupy undercoordinated bridge sites, with the O–C bond perpendicular to the step direction and nearly coplanar with the lower terrace. Alkoxys at steps exhibit an additional preference for a methyl C–H bond directed toward the lower terrace with an energy gain of 0.03 eV. In contrast to the alcohols, the (100) step binds alkoxys more strongly than the (111) step, by 0.06 eV for methoxy and 0.02 eV for ethoxy.

3.3.3 ALDEHYDES

Minimum-energy aldehyde adsorption geometries are shown in rows E and F of Figure 3.2. On the terrace there is negligible aldehyde-Cu chemical bonding, such that adsorption is dominated by vdW interactions, so distances to the Cu(111) surface plane are reported in Table 3.3 instead of the closest Cu atom. Aldehydes on the terrace also show negligible preference for any specific orientation or location as long as the molecular plane is nearly parallel to the terrace. On steps, the more reactive step atoms provide more stable and specific adsorption geometries consistent with dative O–Cu bonding. The lowest-energy geometries involve aldehyde O atoms bonded to either a step bridge or to a single step atom, with the rest of the aldehyde suspended over the lower terrace. Formaldehyde is most stable at step atoms and acetaldehyde is most stable at step bridges, but these preferences are weak since the alternate geometry is less stable by at most 0.04 eV. Formaldehyde has an additional metastable adsorption geometry above a step bridge with its O–C bond directed approximately parallel to the step direction, less stable by only 0.04 eV, suggesting that formaldehyde's potential energy surface is shallower than acetaldehyde's.

3.3.4 HYDROGEN

H atoms are most stable at three-fold hollow sites as shown in row G of Figure 3.2. On the Cu(111) terrace there is a negligible (< 0.01 eV) preference for hcp hollows. The H atom at the hollow site forms three H–Cu bonds of length 1.73 Å. At steps, hydrogen prefers the 3-fold hollow site closest to the step bridge on the upper terrace. This is an fcc hollow site for the (111) step, and an hcp hollow site for the (100) step. The H adsorption geometries have reduced symmetry, forming two H–Cu bonds of length 1.73 Å with the step bridge, and a longer bond of length 1.80 Å with the third Cu atom. Furthermore, while other hollow sites on the upper and lower terraces are metastable, hollow sites at steps are unstable. Top and bridge site geometries are also unstable. Using a clean slab and half the energy of a gas phase H_2 molecule as a reference, hydrogen adsorption energies are -0.28 eV on the terrace, -0.30 eV at the (111) step, and -0.32 eV on the (100) step. H atoms therefore have no significant preference for step edges relative to the Cu(111) terrace.

3.3.5 COVERAGE

To understand the effect of coverage on stability, we calculated the adsorption energies of ethanol and ethoxy in 4×4 Cu(111) unit cells and unit cells that include 4-atom-wide step edges. For the steps, stability is insensitive to coverage since the change in adsorption energy is negligible (≤ 0.02 eV). For the Cu(111) terrace, however, decreasing coverage increases stability. Ethanol is more stable by 0.06 eV and ethoxy is more stable by 0.04 eV in the 4×4 unit cell than in the 3×3

Table 3.4: Hydrogen Bond Energy E_{Hbond}, Donor-Acceptor H$\cdots$O distance $d_{\mathrm{H\cdots O}}$, Donor-Acceptor O–O Distance $d_{\mathrm{O–O}}$, Donor O–Cu Distance $d_{\mathrm{O–Cu}}$, and Acceptor O–Cu Distance $d'_{\mathrm{O–Cu}}$ for Hydrogen Bonding Geometries

Acceptor[a]	Location	E_{Hbond}(eV)	$d_{\mathrm{H\cdots O}}$(Å)	$d_{\mathrm{O–O}}$(Å)	$d_{\mathrm{O–Cu}}$(Å)	$d'_{\mathrm{O–Cu}}$(Å)
EtOH	gas phase	−0.25	1.87	2.84		
	terrace	−0.28	1.64	2.65	2.13	3.13
	(111) step	−0.09	1.63	2.64	2.03	3.56
	(100) step	−0.12	1.74	2.68	2.07	2.43
EtO	gas phase	−0.48	1.45	2.49		
	terrace	−0.06	1.65	2.65	2.17	2.05, 2.10
	(111) step	0.02	1.34	2.43	1.96	1.91
	(100) step	0.05	1.39	2.44	1.97	1.91

[a] The donor is ethanol in all cases.

unit cell. Although hydrogen bonding interactions lead to clustering between adjacent alcohols, intermediate-range alcohol-alcohol interactions are weakly repulsive.[93]

3.3.6 ETHANOL HYDROGEN BONDING

Alcohols in the gas phase exhibit significant hydrogen bonds with binding energies of up to 0.25 eV.[34] Hydrogen bonding to other alcohols and water can change their adsorption and reactivity on transition metal surfaces such as Cu[10,153] and Rh.[112] In particular, alcohol dehydrogenation on Cu surfaces can be co-catalyzed by water-alcohol hydrogen bonding. To investigate the potential for hydrogen bonding to influence reactivity, we calculate the energy of the hydrogen-bonded dimers relative to the isolated adsorbates:

$$E_{\mathrm{Hbond}} = (E\,[\mathrm{dimer}/X] + E\,[X]) - (E\,[\mathrm{donor}/X] + E\,[\mathrm{acceptor}/X]) \tag{3.2}$$

where $E\,[A/X]$ is the energy of adsorbate or dimer A on surface X. Negative values indicate that the hydrogen-bonded dimer is more stable than the isolated adsorbates.

Although EtOH–EtOH hydrogen bond energies are similar between the gas phase and Cu(111) terrace, hydrogen bonding is much weaker at step sites (see Table 3.4). This is a consequence of

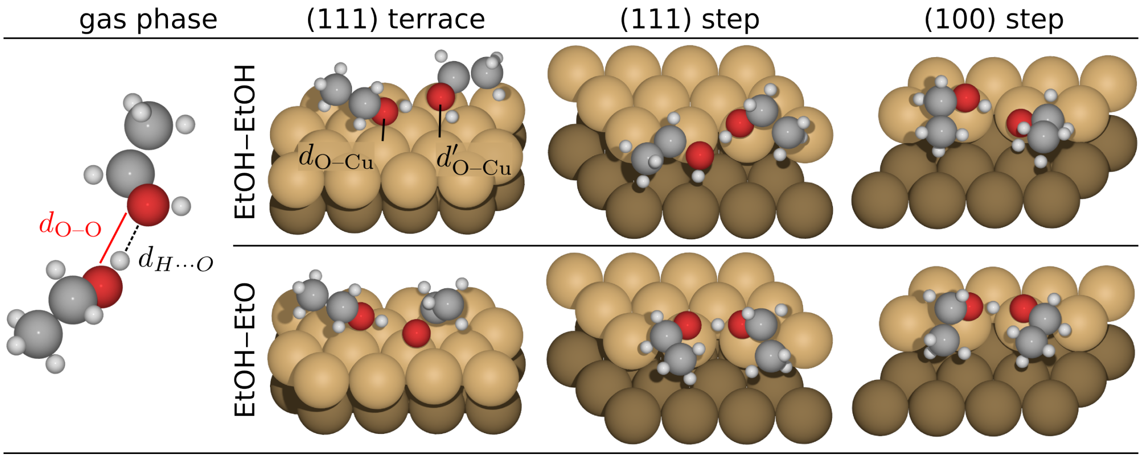


Figure 3.3: Minimum energy hydrogen-bonded geometries for EtOH–EtOH and EtOH–EtO dimers on each surface. Corresponding energies and geometry values are presented in Table 3.4.

two types of geometric frustration. First, the distance between adjacent step atoms, 2.51 Å is incompatible with the gas phase O$\cdots$O distance of 2.84 Å, as shown in Figure 3.3. Step sites' stronger bonding to ethanol increases the degree of frustration, weakening hydrogen bonds relative to the terrace and gas phase. Second, alcohol monomers feature O–H bonds nearly perpendicular to the step direction, at angles of 82for the (111) step and 74for the (100) step, but ethanol molecules identically adsorbed at adjacent step atoms would require a donor O–H bond parallel to the step direction instead. This mismatch in distance and O–H orientation destabilizes the EtOH–EtOH dimer relative to isolated ethanol monomers. The resulting step geometries favor hydrogen bonding: the acceptor ethanol's O–Cu bond is significantly lengthened on the (100) step, and broken altogether on the (111) step and terrace.

As in the EtOH–EtOH case, EtOH–EtO hydrogen bonding is stronger in the gas phase and on the terrace than on steps (see Table 3.4). In the gas phase, EtOH–EtO dimers are strongly bound due to the undercoordinated O atom on the acceptor ethoxy radical, and the hydrogen bond angle is nearly 180 On the terrace, ethanol prefers top sites while ethoxy prefers fcc hollow sites, but top-hollow site distances are incompatible with the gas phase O$\cdots$O distance of 2.49 Å. Hence, geometric frustration again reduces the binding energy of the dimer. Furthermore, since ethoxy is more strongly bound to the terrace than ethanol, hydrogen bonding is no longer strong enough to cleave all of the ethoxy acceptor O–Cu bonds. The minimum-energy geometry on the terrace partially resolves this geometric frustration by moving the acceptor ethoxy to a bridge site. Step dimers are similarly frustrated. Alcohols are most stable at step atoms while alkoxys are most stable at step bridges, but no atom-bridge distance matches the gas phase O···O distance of 2.49 Å. The resulting geometry is qualitatively different than on the terrace (see Figure 3.3). Both steps feature EtOH–EtO dimers where the EtOH and EtO fragments form a single O–Cu bond at adjacent top sites. EtOH–EtO dimers at steps are less stable than isolated EtOH and EtO due to significant geometry changes, but the hydrogen bond itself appears to be strong since the hydrogen bond angle and O···O distances are very similar to the gas phase values. The donor O–H bond is also longer than in the gas phase, 1.09 Å versus 1.04 Å, while the acceptor H···O distance is significantly shorter, 1.38 Å versus 1.65 Å.

Molecular dynamics (MD) simulations of isolated ethanol dehydrogenation intermediates and hydrogen-bonded EtOH–EtOH and EtOH–EtO dimers provide additional insight into hydrogen bonding. Isolated ethanol molecules on the terrace rotate almost freely about the O–Cu bond so O–H bond orientations change approximately every 100 fs. Such frequent rotations suggest a shallow potential energy surface. In contrast, ethanol trajectories on both steps show that the O–H bond is almost always directed toward the lower terrace which suggests a deeper potential energy well. This confirms the assertion made earlier, namely that geometric frustration for the donor ethanol O–H orientation is more significant at step sites. We also confirm that the O–Cu bonds for the accepting ethanol are broken on the terrace and (111) step. On the terrace, the donor ethanol remains adsorbed to the same Cu atom for MD simulation times up to 1 ps while the whole dimer rotates almost freely about the donor O–Cu bond. For the (111) step, the situation is similar where the donor O–H bond shows little energetic preference for a specific angle relative to the step direction over a range of 45 The behavior at the (100) step is significantly different, as the donor O–H bond does not rotate much and the acceptor ethanol remains bound to a step atom.

For EtOH–EtO dimers we gain similar insights into the bonding from MD simulation trajectories of 1 ps. On all surfaces the EtOH–EtO dimers are metastable and vibrate in place, without any of the nearly free rotations seen for EtOH–EtOH dimers. At both steps the H atom position relative to the O atoms varies significantly throughout the trajectories. Figure 3.4(a) shows the superimposed projections of O and H positions onto u, the O–O axis, and v, the perpendicular direction in the O-O-H plane, relative to the red O atom for each time step. These large variations are the result of the H atom "hopping" between the two O atoms approximately every 150 fs as shown in Figure 3.4(b), precluding clear donor and acceptor assignments. The dynamics are therefore more accurately described as two ethoxy fragments sharing a H atom. This shared H atom remains at least 1.10 Å from either O atom for 90% of the trajectory, indicating strong hydrogen bonding. This direct H exchange between alcohols and alkoxys could facilitate alcohol-alkoxy displacement reactions reported previously.[77]

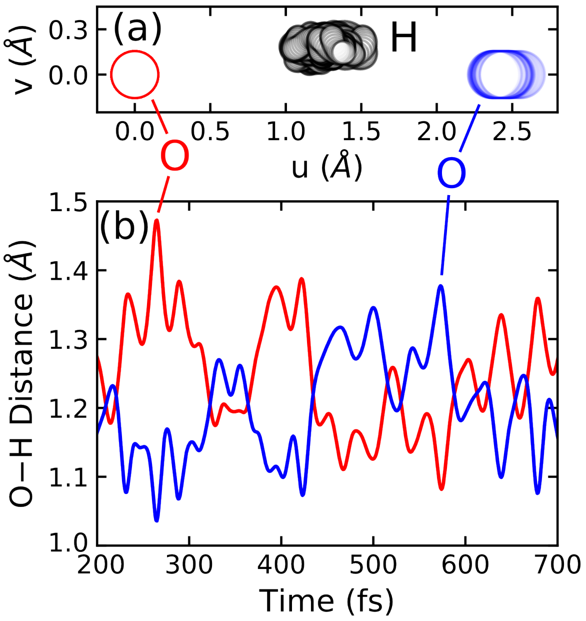


Figure 3.4: Molecular dynamics results for an EtOH-EtO dimer at the (111) step. (a) Projection of O and H atom positions along u, the O-O direction, and v, the perpendicular direction in the O-O-H plane. (b) Distance from the H atom to the O atom of the same color in panel (a).

3.4 SURFACE REACTIVITY

To compare reaction energies and stabilities among different alcohol and surface combinations, we use the DFT total energy of the gas phase alcohol and bare surfaces as a reference. All dehydrogenation intermediates, including H atoms, are considered adsorbed to the same kind of surface. Hydrogen diffusion away from product states is determined using the minimum energies of the isolated hydrogen and adsorbate in separate DFT unit cells

$$\Delta E = E\left[\mathrm{H}/X\right] + E\left[A/X\right] - E_{\text{product}}\left[\left(A+\mathrm{H}\right)/X\right] - E\left[X\right] \tag{3.3}$$

where A is the intermediate, X is the active site, and E_{product} is the DFT total energy of the nearest local minimum past the transition state. The resulting interaction energies are shown in Figure 3.5. Since peak aldehyde desorption temperatures are lower than or similar to peak H_2 recombination temperatures in temperature-programmed reaction spectroscopy (TPRS) experiments,[182] we do not include H_2 recombination in Figure 3.5; energy differences between step types include their relative H atom stabilities. The complete dehydrogenation reaction from gas phase alcohol to gas phase aldehyde and H_2 is significantly endothermic: 0.95 eV for methanol to formaldehyde, and 0.71 eV for ethanol to acetaldehyde. These values are within 0.06 eV of their experimental enthalpies of reaction.[1] Geometries for each of the 12 transition states are shown in Figure 3.6. Activated C–H bond lengths are significantly longer than activated O–H bond lengths, particularly at step edges, and similar to those found for the Cu(110) surface.[20] These lengths show that transition states for C–H activation are particularly late along the reaction pathway, in agreement with the low (≤ 0.25 eV) activation energies for aldehyde hydrogenation back to alkoxys at steps.

Reaction energy profiles for methanol and ethanol on all three active sites are shown in Figure 3.5. Steps decrease all activation energies and strengthen all adsorption energies relative to the Cu(111) terrace. The same is true compared to the Cu(110) surface as well,[20] with the exception of C–H bond activation energies which are similar. An important quantity for dehydrogenation activity is the relative probability of desorption versus O–H bond activation, which is exponential in the "branching energy" $\Delta E_A^{\text{O–H}} - \Delta E_{\text{desorb}}$. The branching energy is 0.39 eV for methanol and 0.27 eV for ethanol on the Cu(111) terrace, making desorption much more likely than O–H activation. In contrast, step edges provide significantly higher desorption energies and lower activation energies, yielding a negative branching energy and thus increasing catalytic activity. Steps are therefore predicted to contribute most of the surface reactivity under UHV conditions, in agreement with experimental results.[182]

Another major trend is the difference in stability and reactivity between methanol and ethanol. Ethanol and ethoxy are more stable than methanol and methoxy by about 0.18 eV on all three surfaces, respectively. The difference in adsorption energies between acetaldehyde and formaldehyde is more variable, ranging from 0.11 eV on the terrace to 0.28 eV on the (111) step.

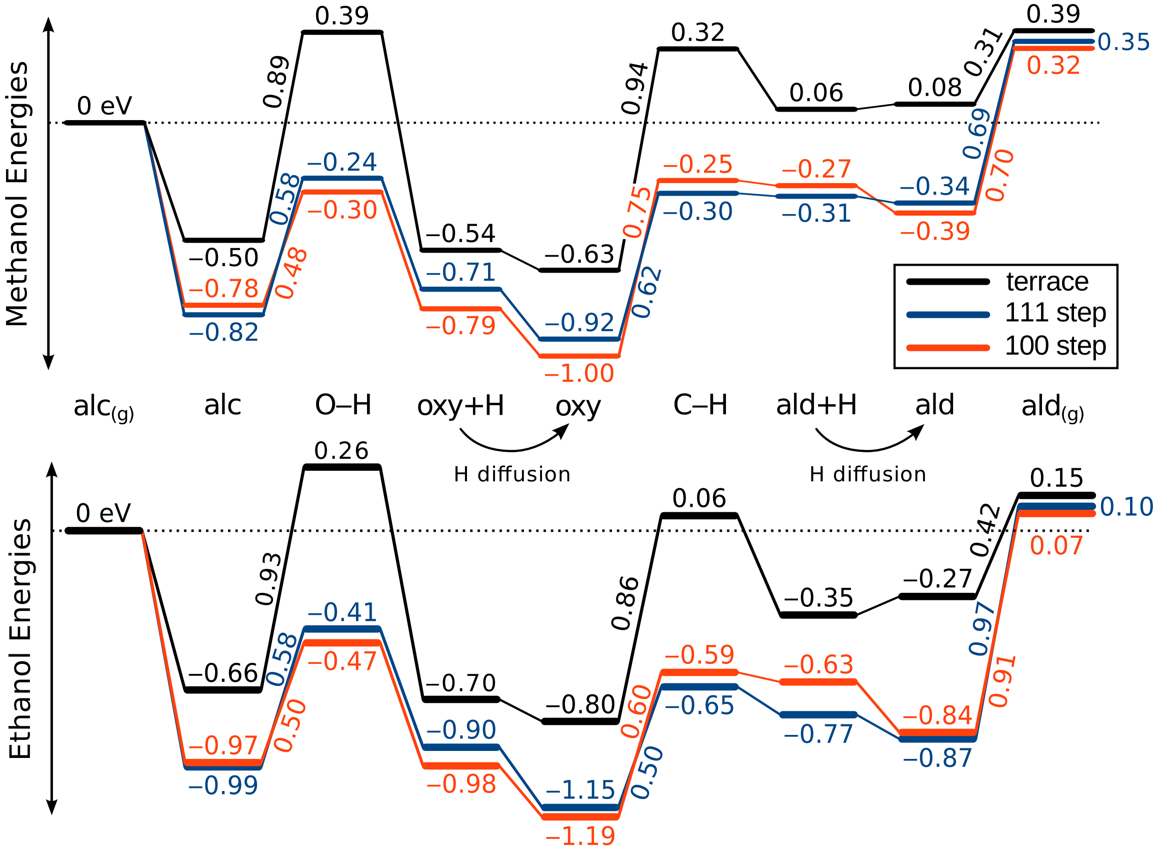


Figure 3.5: Energies, in eV, for methanol and ethanol dehydrogenation intermediates relative to gas phase alcohols and clean surfaces for the Cu(111) terrace (black), (111) step (blue), and (100) step (red). Steps are labeled as "alc" for alcohols, "oxy" for alkoxys, and "ald" for aldehydes. X+H indicates X coadsorbed with H, and X–H indicates the corresponding transition state. Energies for H_2 desorption are omitted for clarity; values include atomic H adsorption at each surface.

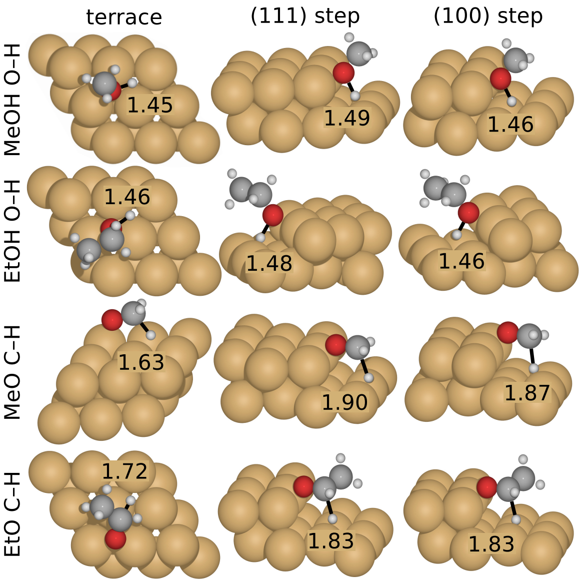


Figure 3.6: Transition states for O–H and C–H bond activation for the Cu(111) terrace and the (111) and (100) step edges. Each number indicates the activated bond's length in Angstroms.

There is little difference in O–H activation energies; the methanol and ethanol values are within 5% at each site. In contrast, ethoxy C–H activation energies are all lower than their methoxy counterparts, by about 10% for the terrace and 25% on both steps.

3.4.1 ALCOHOLS TO ALKOXYS

Steps significantly reduce the activation energy of O–H bond cleavage for both alcohols. Their O–H transition states have the alkoxy fragment adsorbed to a step bridge and a H atom near a three-fold hollow site on the lower terrace. The metastable product state consists of the alkoxy at the step bridge and H at the same 3-fold hollow site on the lower terrace. This state is less stable than having H and the alkoxy on separated slabs, primarily due to the significantly lower stability of H below the step relative to the terrace above the step (by about 0.2 eV).

On the terrace, the O–H bond is cleaved over a bridge site, with the alkoxy and H fragments adsorbed at adjacent hcp and fcc hollow sites, respectively. Isolated alkoxys are more stable by about 0.04 eV at fcc hollow sites but transition states with alkoxys at hcp sites are more stable by

≤ 0.03 eV. Following O–H cleavage the alkoxy diffuses across a Cu–Cu bridge to an adjacent fcc site. Alkoxy-H interactions are repulsive on the terrace, possibly due to Coulomb repulsion. Using methanol dehydrogenation as an example, the net Hirshfeld charges on the alcohol O and H atoms start at $q_{\mathrm{O}} = -0.16$ e and $q_{\mathrm{H}} = +0.14$ e, and decrease throughout the reaction to $q_{\mathrm{O}} = -0.22$ e and $q_{\mathrm{H}} = -0.12$ e in the coadsorbed product geometry. Compared to steps, higher Cu atom coordination and weaker alkoxy adsorption destabilize the transition state on the Cu(111) terrace.

3.4.2 ALKOXYS TO ALDEHYDES

Like O–H bond activation, C–H bond activation proceeds differently on step sites than on the Cu(111) terrace. To cleave the C–H bond on the terrace, the alkoxy O–C bond first tilts toward the surface to reduce the distance between the H atom and the terrace, followed by H abstraction to a fcc hollow site. The C–H activation energy is lower for ethoxy than for methoxy. In addition, C–H cleavage is significantly endothermic for both alkoxys so the activation energies for hydrogenation back to the alkoxy are 0.26 eV for formaldehyde and 0.41 eV for acetaldehyde.

Reactions at steps have significantly lower activation energies than on the terrace, a result of favorable alkoxy adsorption geometries. First, alkoxys already have O–C bonds approximately parallel to the terrace so the abstracted H atom starts approximately 0.5 Å closer to the Cu surface than on the flat Cu(111) terrace. Second, minimum-energy aldehyde geometries have similar O–C and C–C orientations relative to the step edge, reducing the amount of molecular distortion and hence the activation energy required to reach the transition state.

Alkoxys and aldehydes are both more stable at steps so the net reaction energies are approximately as endothermic as on the terrace. Combined with the lower transition state energy, at steps C–H bond formation has very low activation energy. The largest activation energy for aldehyde hydrogenation is only 0.25 eV so alkoxy dehydrogenation should occur readily at the 300-400 K aldehyde desorption temperatures reported in experiments.[182]

3.4.3 ALDEHYDE DESORPTION

We find that aldehyde desorption from Cu(111) step sites is strongly affected by kinetic competition. Aldehyde desorption temperatures are significantly higher in temperature-programmed reaction spectroscopy (TPRS) experiments, in which the alcohol is dosed, than in temperature-programmed desorption (TPD) experiments, in which the aldehyde itself is dosed. However, Figure 3.5 shows that alkoxy C–H bond activation is not the rate-limiting step. Using simple kinetic arguments similar to previous work,[113] we show that kinetic competition between reversible hydrogenation and irreversible desorption significantly increases the effective aldehyde desorption energy in TPRS experiments compared to TPD experiments.

Reversible aldehyde hydrogenation and irreversible desorption can be described using simple first and second-order rate expressions. We consider the rates of alkoxy dehydrogenation k_f, aldehyde hydrogenation k_r, and aldehyde desorption k_d per lattice site:

$$k_f = A_f\, \theta_{\mathrm{alk}}\, (1 - \Theta_{\mathrm{T}})\, e^{-\beta \Delta E_A^f} \tag{3.4}$$

$$k_r = A_r\, \theta_{\mathrm{ald}}\, \theta_{\mathrm{H}}\, e^{-\beta \Delta E_A^r} \tag{3.5}$$

$$k_d = A_d\, \theta_{\mathrm{ald}}\, e^{-\beta E_{\mathrm{ads}}} \tag{3.6}$$

$$\theta_o = \theta_{\mathrm{alk}} + \theta_{\mathrm{alk}} \tag{3.7}$$

where θ_X values are coverages, $\theta_o \sim 10\%$ is the coverage of active sites,[182] Θ_{T} is the total coverage on the terrace, and the prefactors $A_f = 10^{13}$ Hz, $A_r = 2 \times 10^{12}$ Hz, and $A_d = 10^{15.5}$ Hz are typical values for first-order reactions, second-order reactions, and first-order desorption,[14] respectively. Hydrogen recombines at higher temperatures than aldehyde desorption[182] so H atoms are kinetically trapped and thus $\theta_{\mathrm{H}} = \theta_{\mathrm{alk}} + 2\theta_{\mathrm{ald}} \geq \theta_o$. Moreover, all adsorbates except H are significantly more stable at steps so $\Theta_{\mathrm{T}} \approx \theta_{\mathrm{H}}$. Aldehyde hydrogenation barriers are much lower than their desorption energies at steps, so even the lower bound $\theta_{\mathrm{H}} = \theta_o$ gives $\frac{k_r}{k_d} \geq 2.4 \times 10^5$ at 300 K. The vastly higher hydrogenation rate yields an equilibrium between aldehyde hydrogenation and alkoxy dehydrogenation.

Equilibrium gives a fully-determined system of equations. Defining $B \equiv e^{-\beta\left(\Delta E_A^f - \Delta E_A^r\right)}$, where $B \leq 2.0 \times 10^{-5}$ at 300 K, we obtain the following important results to first order in B:

$$\frac{\theta_{\mathrm{ald}}}{\theta_{\mathrm{alk}}} \approx \frac{1-\theta_o}{\theta_o}\frac{A_f}{A_r}B \tag{3.8}$$

$$k_d \approx (1-\theta_o)\,A_d\,\frac{A_f}{A_r}\,e^{-\beta\left(E_{\mathrm{ald,(g)}} - E_{\mathrm{alk}}\right)} \tag{3.9}$$

Equation (3.8) predicts a formaldehyde to methoxy ratio of approximately $1{:}\left(2 \times 10^7\right)$, and an acetaldehyde to ethoxy ratio of approximately 1:500 at their peak TPRS desorption temperatures of 370 and 320, respectively. Alkoxys are predicted to be the majority species in all cases. Expressing k_d in terms of the dominant alkoxy coverage, $\theta_{\mathrm{ald}} \approx \theta_o$, results in Equation (3.9), showing the effective aldehyde desorption is the total energy difference between the gas phase aldehyde and the adsorbed alkoxy. Alkoxys are significantly more stable than their corresponding aldehydes so kinetic competition increases the effective aldehyde desorption temperature. These predictions are in excellent agreement with the observed TPRS temperatures as discussed below. The critical factor leading to these results is $\frac{k_r}{k_d} \gg 1$, a combination of high aldehyde desorption energy and low activation energy for C–H bond formation at step edges.

3.4.4 TEMPERATURE COMPARISON WITH EXPERIMENTS

To check the correspondence between our DFT calculations and recent experimental measurements, we use the Redhead Equation for first-order kinetics to calculate peak TPD temperatures $T_{\mathrm{max}}^{\mathrm{DFT}}$ from desorption energies

$$\frac{E_{\mathrm{A}}}{k\,T_{\mathrm{max}}^{2}} = \frac{v}{r}\exp\left(\frac{E_a}{k\,T_{\mathrm{max}}}\right) \tag{3.10}$$

where $r = 1.0$ K/s is the heating rate, k is the Boltzmann constant, and $v = 10^{15.5}$ Hz is a typical desorption prefactor for adsorbates like CO and methanol.[14] We compare to the experimental data of Wang et al.,[182] who report desorption temperatures for flat and stepped Cu(111) surfaces dosed with methanol, ethanol and acetaldehyde, as well as the reaction rate limited desorption temperatures of formaldehyde and acetaldehyde. The same experimental temperature is assigned to both step edges, and the activation energy for reaction rate-limited aldehyde desorption is assumed to be the total energy difference between adsorbed alkoxys and their gas phase aldehydes as discussed previously. As shown in Figure 3.7 with dark circles, the DFT calculations and chosen prefactor typically overestimate desorption temperatures by about 41 K. The linear correlation suggests that DFT calculations are reliable for obtaining catalytic trends. Furthermore, we extrapolate our ethanol results with the trend of 0.18 eV/CH_2 to predict experimental TPD temperatures for isopropanol and acetone[114] desorption, as well as 1-butanol, butyraldehyde, crotyl alcohol, and crotanaldehyde.[123] For these predictions we use the Redhead equation as above with the experimental heating rates. These predictions, shown in Figure 3.7 as open squares, follow a similar trend to the fitted line which suggests that the 0.18 eV/CH_2 PBE+vdW trend generalizes well to longer alcohols.

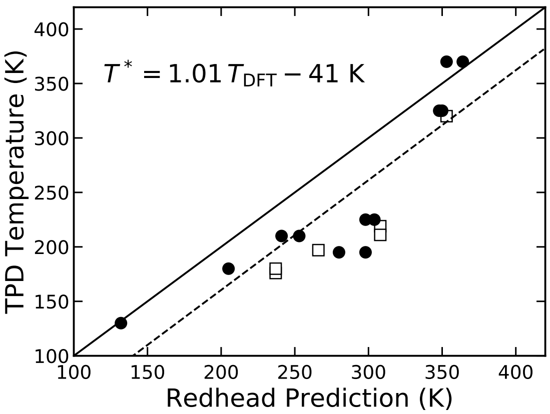


Figure 3.7: Parity plot of experimental TPD temperatures versus DFT predictions via the Redhead Equation (see text). Solid line: 1:1 parity. Dashed line: fit to methanol and ethanol results (dark circles). Open squares compare extrapolated desorption temperatures for isopropanol, 1-butanol, crotyl alcohol, and their aldehydes with experimental TPD values.

3.5 CHEMICAL DEFECTS: ISOLATED PT ATOMS

Whereas previous sections discuss dehydrogenation at Cu(111) step defects, isolated Pt atoms embedded in the Cu(111) surface have also been shown to catalyze ethanol dehydrogenation.[181] We refer to this Pt-substituted surface as the Pt–Cu single-atom alloy (SAA). Analogous calculations to those above, with the same computational settings, are used to map out the energetic landscape for the formation, spillover and reaction of ethoxy intermediates. Comparing these calculations to those on step defects highlights the most important considerations for anhydrous alcohol dehydrogenation on Cu(111) surfaces.

for ethanol are very similar between the Pt–Cu SAA and the Cu(111) surfaces, where ethanol adsorbs most strongly at Cu top sites in the trans conformer (Figure 3.8). Ethanol adsorption is stronger on the Pt–Cu SAA surface than on Cu(111) by 0.07 eV and the O–H bond is oriented toward the Pt atom. Attractive H–Pt interactions, consistent with hydrogen bonding, strengthen ethanol adsorption by 0.07 eV relative to the Cu(111) surface. These H–Pt interactions also facilitate O–H bond activation and stabilize the removed H atom after the dehydrogenation, significantly decreasing the activation barrier (see Figure 3.8). This reduction, combined with experimental observations of alcohol O–H bond activation, confirms the formation of the ethoxy intermediate. Inserting the O–H bond activation barrier (0.63 eV) on Pt–Cu SAAs into the Arrhenius equation with a prefactor of 10^{12} Hz and a reaction rate of 1 mHz predicts an annealing temperature of $\sim$ 210 K needed to induce the reaction.[15,145] This temperature is $\sim$ 50 K higher than the temperature at which the sample must be annealed in the STM experiment in order to observe the formation of the ethoxy intermediate. However, considering the accuracy of DFT calculations and the precision of STM temperature measurements ($\pm$20 K),[182] theory and experiment are in reasonable agreement with each other for the O–H bond activation energy on Pt–Cu SAAs.

The corresponding product states after O–H cleavage on Cu(111) terraces and Pt–Cu SAA surfaces are significantly different. On the Cu(111) surface the ethoxy group and H atom are adsorbed at adjacent threefold □Cu (Cu_3) hollow sites with ethoxy at the fcc hollow and H at the hcp hollow site, with an adsorption energy of -0.70 eV relative to gas phase ethanol. On the Pt–Cu SAA surface, the ethoxy–Pt interaction is repulsive so the ethoxy and H fragments are separated by two hollow sites (Figure 3.8). Ethoxy is coordinated to a Cu_3 fcc site and the H fragment is adsorbed at a Pt_1Cu_2 hcp site, with an adsorption energy -0.73 eV relative to gas phase ethanol. These energetics agree with the STM results showing that while ethoxy is formed at Pt sites, it spills over to pure Cu(111) sites at which it is 0.05 eV more stable than at Pt sites (Figure 3.8).

As shown in Figure 3.8 the calculated energy for C–H bond activation (0.63 eV) on the Pt–Cu SAA surface is also lower than on the Cu(111) surface (0.96 eV). Experimental results also suggest lower activation energies since acetaldehyde desorbs at lower temperatures in TPRS experiments on Pt–Cu SAAs compared to the Cu(111) surface. The final dehydrogenation product, acetaldehyde, has very similar nearly-flat adsorption geometries on both surfaces. Compared to ethoxy ($\sim$ 0.7 eV), coadsorbed acetaldehyde and H have a much smaller adsorption energy of

-0.35 eV on Cu(111) and ~ -0.55 eV eV on Pt–Cu SAAs relative to gas phase ethanol. The results also reveal that the acetaldehyde desorption barriers are 0.42 eV on Cu(111) and 0.53 eV on Pt–Cu SAAs. On both surfaces the desorption energy of acetaldehyde is comparable to the activation energy for C–H bond formation so kinetic competition is expected to be limited.

In addition, isolated Pt atoms reduce the TPRS temperature for recombination of molecular hydrogen, 260 K from Pt–Cu SAAs[181] compared to 330-370 K from the Cu(111) terrace,[182] despite stronger H adsorption to Pt atoms than Cu atoms. The reduced temperature is caused by isolated Pt atoms catalyzing hydrogen recombination, decreasing the activation energy significantly compared to the Cu(111) terrace.[36,105] Hydrogen coverages should then be lower on Pt–Cu SAAs than on flat or stepped Cu(111) surfaces at temperatures exceeding 260 K. Lower hydrogen coverages should then further reduce kinetic competition between acetaldehyde desorption and hydrogenation to ethoxy. These factors help explain why acetaldehyde desorbs at lower temperatures in TPRS experiments from Pt–Cu SAAs (310 K) compared to stepped Cu(111) surfaces (320 K) despite the higher C–H bond activation energy.

An important aspect of any catalytic cycle is the fate of the active site after the reaction, e.g. poisoning with byproducts. Our TPRS experiments show that the acetaldehyde yield remains constant after five reaction cycles on the same Pt–Cu SAA surface. This confirms that no deactivation occurs during ethanol dehydrogenation under UHV conditions. Furthermore, STM images after the deposition of ethanol and annealing to 450 K to complete the reaction cycle revealed a completely clean surface on which no residual surface-bound species are present, further confirming that no poisoning occurs. Importantly, the isolated Pt sites in the surface are stable up to this temperature and can catalyze repeated reactions. In terms of reaction selectivity, we did not detect any species other than the m/z fragments for ethanol, H_2 and acetaldehyde, indicating 100% selectivity for ethanol dehydrogenation.

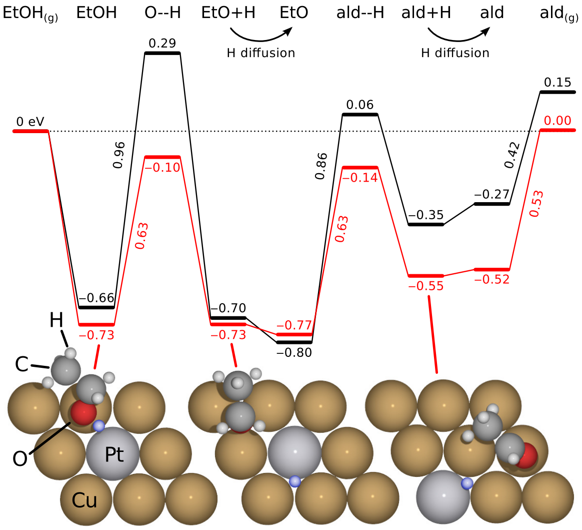


Figure 3.8: Selected stable structures and energies (eV) for all reaction steps for ethanol on Cu(111) (black) and the Pt–Cu SAA surface (red) relative to gas-phase ethanol and the corresponding bare surface. EtOH+H and ald+H are the nearest local minima past the transition state for O–H and C–H cleavage, respectively. EtOH, EtO and ald correspond to ethanol, ethoxy and acetaldehyde, respective. H_2 recombination is omitted for clarity; energy differences between the Cu(111) and Pt–Cu SAA surfaces include their relative H adsorption energies.

Taken in combination with experimental results,[181] these calculations show that ethanol undergoes the same two-step dehydrogenation on Pt–Cu(111) SAAs as on flat and stepped Cu(111) surfaces. Furthermore, while both Cu(111) step edges and surface Pt atoms can convert ethanol to acetaldehyde and H_2, surface Pt atoms have lower energetic barriers and higher conversion due to the spillover of the ethoxy intermediates to Cu. Thus, this work demonstrates the crucial role of dilute active metal sites on important reaction steps using model Pt–Cu single crystal alloy surfaces.

3.6 CATALYTIC IMPLICATIONS

3.6.1 BRANCHING ENERGIES FOR LONGER ALCOHOLS

Since O–H activation energies are very similar between methanol and ethanol, we expect them to be similar for longer alcohols as well. The limiting factor for UHV methanol and ethanol dehydrogenation on Cu(111) single crystals is their high branching energies. If each additional CH_2 subunit contributes a similar adsorption energy change of 0.18 eV without changing the O–H activation energy, then longer alcohols will increasingly favor reaction over desorption. A similar increase in reactivity is predicted for longer alcohols on the Cu(110) surface.[20] The ethanol branching energy, the difference between its O–H activation energy and desorption energy, is 0.27 eV on the terrace so desorption is strongly favored over reaction. However, with the trend of 0.18 eV/CH_2, longer alcohols will have increasingly favorable branching energies. For example, n-propanol should have a lower branching energy of approximately 0.09 eV and n-butanol should have a branching energy of approximately -0.09 eV. We therefore hypothesize that Cu(111) terraces can selectively dehydrogenate longer alcohols starting with n-butanol.

3.6.2 COMPETITIVE BINDING

Despite a higher calculated O–H activation energy, our previous investigation of isolated Pt atoms in Cu(111) single crystal surfaces[182] (single-atom alloys) found greater catalytic activity than steps for alcohol dehydrogenation. We believe this is the result of “self-poisoning” at steps where alkoxy products favorably compete with alcohol reactants for adsorption. On the Pt-Cu(111) surface, ethoxy-Pt interactions are repulsive and ethoxy groups adjacent the Pt active site are less stable than at longer distances. In contrast, in this study steps are the most stable binding sites for alkoxys. To achieve catalytic turnover at steps, alkoxy products must diffuse away from step edges to make way for additional alcohol adsorption. This likely slows down dehydrogenation kinetics at high alkoxy coverages. Furthermore, alkoxy-alcohol replacement is slightly endothermic (by < 0.1 eV) for ethanol at both steps and for methanol at the (100) step, effectively increasing the activation energy for subsequent O–H bond breaking. Making alkoxy adsorption weaker or altogether unstable could mitigate these kinetic and thermodynamic limitations. We therefore hypothesize that catalytic turnover for O–H activation could be improved if step edges were chemically modified to destabilize alkoxy adsorption. At high hydrogen coverage such modifications would also reduce aldehyde desorption temperatures since they are dictated by alkoxy stability as well.

3.6.3 HYDROGEN RECOMBINATION

In addition to destabilizing alkoxy adsorption, another way to reduce aldehyde desorption temperatures is to reduce hydrogen coverage. Alkoxys adsorb more strongly than their corresponding aldehydes, so reversible aldehyde hydrogenation to the alkoxy increases aldehyde desorption temperatures. Hydrogen desorbs from the Cu(111) terrace near 370 K at low coverages, a temperature similar to or higher than aldehyde desorption temperatures, so H coverage is expected to be high under reaction conditions. The activation energy for hydrogen recombination can be reduced by introducing other active sites on the surface. For example, isolated Pt atoms in the Cu(111) surface layer reduce the H_2 recombination temperature to approximately 255 K in TPD experiments for anhydrous ethanol dehydrogenation.[182] Reduced hydrogen coverage would limit aldehyde hydrogenation back to alkoxys and therefore reduce the peak TPD temperature for aldehyde desorption.

3.7 CONCLUSIONS

In this work we investigate anhydrous alcohol dehydrogenation on the Cu(111) surface and its steps. The presence of steps increases adsorption energies for all intermediates. Including TS vdW corrections further increases adsorption energies by more than 50% in all cases and favors shorter adsorbate-Cu distances. ZPE corrections yield a minor ($< 10\%$) effect on adsorption, but a significant reduction in O–H and C–H bond activation energies by ~ 0.21 eV. Finally, while hydrogen bonds are significant in the gas phase and on the terrace, geometric frustration limits their stability at step edges. EtOH–EtO dimers at step edges are less stable than isolated EtOH and EtO, but still exhibit strong H-bonding characteristics and a “H hopping” mechanism with sigificant variations in O–H distances.

Steps also significantly improve reactivity by reducing O–H and C–H activation energies to values lower than alcohol desorption energies. With O–H bond activation energies lower than desorption energies, steps will be highly reactive, while the opposite is found for the Cu(111) terrace. Step edges on Cu(111) are also predicted to be somewhat more reactive overall for anhydrous alcohol dehydrogenation than clean Cu(110) surfaces. Furthermore, the low C–H bond activation energies makes alkoxy dehydrogenation highly reversible. Acetaldehyde desorption from steps is therefore predicted to be limited due to kinetic competition so desorption temperatures are dictated by the greater alkoxy stability. Peak TPD and TPRS temperatures calculated at step edges using DFT based on this prediction are in good agreement with experimental results for methanol, ethanol, and acetaldehyde desorption from Cu(111). In contrast, kinetic competition does not take place on the Cu(111) terrace due to the higher C–H activation energies and lower aldehyde desorption energies.

In light of these results we propose three directions for further work on alcohol dehydrogenation on Cu(111) surfaces. First, the additional CH_2 subunit of ethanol increases its desorption energy by 0.18 eV relative to methanol while preserving the O–H activation energy. If this trend continues to larger alcohols, desorption energies are predicted to exceed O–H activation energies on the Cu(111) terrace starting with n-butanol and thus lead to length-selective alcohol dehydrogenation. Second, alcohols and alkoxys are similarly stable at step edges so catalytic turnover is limited. Alkoxy binding could be destabilized by doping, facilitating their replacement with unreacted alcohols to improve catalytic turnover and UHV aldehyde yield. Third, aldehyde desorption temperatures are increased in the presence of coadsorbed H due to reversible aldehyde hydrogenation. If surfaces were chemically modified to facilitate H_2 desorption at lower temperatures, the resulting decrease in H coverage would limit C–H hydrogenation rates and lead to aldehyde desorption at lower temperatures.

ACKNOWLEDGMENTS

We thank Professor Robert J. Madix for his constructive feedback. This work was supported by Integrated Mesoscale Architectures for Sustainable Catalysis, an Energy Frontiers Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award DE-SC0012573. Calculations were performed using resources supported by the U.S. Department of Energy, Office of Science: the Oak Ridge National Leadership Computing Facility at Oak Ridge National Laboratory, under Contract No. DE-AC05-00OR22725, and the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under Contract No. DE-AC02-05CH11231. Additional calculations were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

4

Nonadiabatic H_2 Dissociation on Cu_{13} Nanoclusters

Apart from minor modifications, this chapter is being prepared for the following publication:

> **Robert A. Hoyt**, Matthew M. Montemore, and Efthimios Kaxiras. “Nonadiabatic Hydrogen Dissociation on Copper Nanoclusters” *in preparation*

ABSTRACT

Hydrogenation is a major industrial process, spurring significant interest in the development of selective and efficient catalysts. Copper surfaces exhibit high selectivity, but are poor catalysts for the initial dissociation of molecular hydrogen. Various properties of materials have been previously considered as key factors in catalyst design; in this work, we identify the spin state as an additional design parameter for nanocatalysts. We study hydrogen dissociation on copper nanoclusters to understand how structure might improve this metal’s catalytic prospects. Nanoclusters, which can have large ground-state magnetic moments depending on symmetry and quantum size effects, can exhibit magnetization-dependent catalytic behavior. The most favorable transition state for hydrogen dissociation has a significantly lower activation energy on the nanocluster compared to that of single-crystal copper surfaces, but requires a switch in magnetization from 5 μ_B to 3 μ_B. Without this switch, the activation energy on nanoclusters is *higher* than that for single-crystal surfaces. The weak spin-orbit coupling in Cu hinders this important change in magnetization state, decreasing the kinetic rate of hydrogen dissociation by a factor of 16. We consider strategies to facilitate the magnetization switch through optical excitation, substitution, change in charge state,

and the presence of a co-catalyst. These considerations demonstrate how control of magnetic properties could be used to improve catalytic performance.

4.1 INTRODUCTION

Heterogeneous transition metal catalysts are ubiquitous in industry. Their wide-ranging applications, from catalytic cracking to highly selective hydrogenation, are derived from the large variety of electronic properties across the periodic table. These intrinsic properties are further modified through structural and compositional changes induced by alloying, oxidation, surface structure modification, metal-support interactions, and promotors; all these changes provide enormous flexibility in catalytic design. In recent years, progress in nanotechnology has introduced the nanocluster size and geometry as new catalyst design parameters.[184] As the nanocluster diameter approaches 1 nm and below, finite size effects can become especially pronounced and produce surprising effects like a discontinuous change in catalytic activity upon the addition of a single atom.[25] Here, we show that finite size effects can give rise to a significant magnetic moment in a nanocluster of Cu and this produces magnetization-dependent reactivity; this dependence on magnetization could be a path to magnetic control of catalytic activity.

We focus on selective hydrogenation on Cu, which is non-magnetic in its bulk form. The reason for our choice is that Cu surfaces exhibit high selectivity for industrially important reactions, but they require elevated temperatures and pressures to compensate for their poor H_2 dissociation kinetics.[51,91] While some improvements have been proposed for bulk surfaces, such as the use of dilute alloys,[91] the possibility of using Cu nanoclusters has been largely unexplored. The reduced atomic coordination and surface strain inherent in nanoclusters generally increase adsorption energies and reduce activation energies. Thus, we expect these effects in Cu nanoclusters to lead to improved H_2 dissociation kinetics.

We investigated H_2 dissociation on an icosahedral nanocluster of Cu using first-principles density functional theory (DFT) calculations in the SIESTA[156] code with the PBE exchange-correlation functional[138,139] (see Appendix A for details). Early studies using empirical interatomic potentials found icosahedral geometries to be most stable, but more recent work using first-principles quantum chemistry methods predict disordered minimum-energy geometries for Cu_{13} and other 13-atom fcc metal clusters.[23,132] Nevertheless, icosahedral Cu_{13} can be synthesized in solution using synchrotron radiolysis.[133] Electrochemical Cu_{13} synthesis has also been reported,[176] with sizes consistent with an icosahedral geometry, but the structure could not be determined precisely. More importantly, the 3-fold hollow sites of icosahedral Cu_{13} are locally identical to hcp hollow sites on the well-studied bulk Cu(111) surface; this similarity in structural features can reveal finite size effects with greater clarity.

4.2 RESULTS AND DISCUSSION

4.2.1 ELECTRONIC STRUCTURE

Icosahedral Cu_{13} nanoclusters have a significantly different electronic structure than Cu(111) surfaces. Quantum size effects produce discrete energy levels near while the icosahedral symmetry yields 5-fold degeneracies in both spin directions to produce 10 total valence states. The five valence electrons of Cu_{13} then follow Hund's rule to fully occupy the five majority-spin states and leave the minority-spin states unoccupied. The result is a large ground-state magnetization of 5 μ_B, as noted previously,[104] and an exchange splitting of 0.47 eV, as shown in Figure 4.1. Other 13-atom metal nanoclusters, either of icosahedral shape or with different structure, are also predicted to have significant magnetization, including Ni_{13}, Ag_{13}, and Au_{13}.[23,104,107] This appears to be a general tendency of nanoclusters, where the discrete densities of states combine with symmetry-induced multiple degeneracy to produce net magnetization through exchange

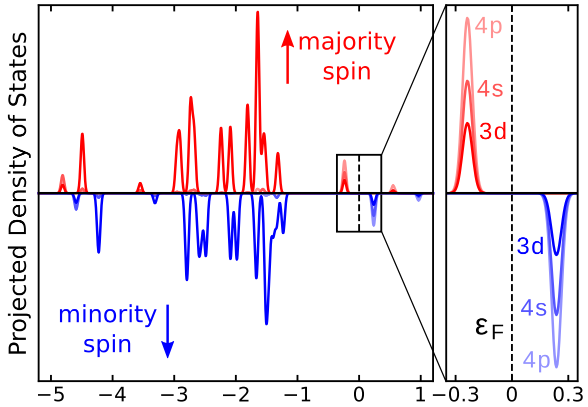


Figure 4.1: Spin-resolved projected density of states near the Fermi level icosahedral Cu_{13} for the majority (up, red) and the minority (down, blue) spin components. Contributions from different orbitals are identified for both spin directions.

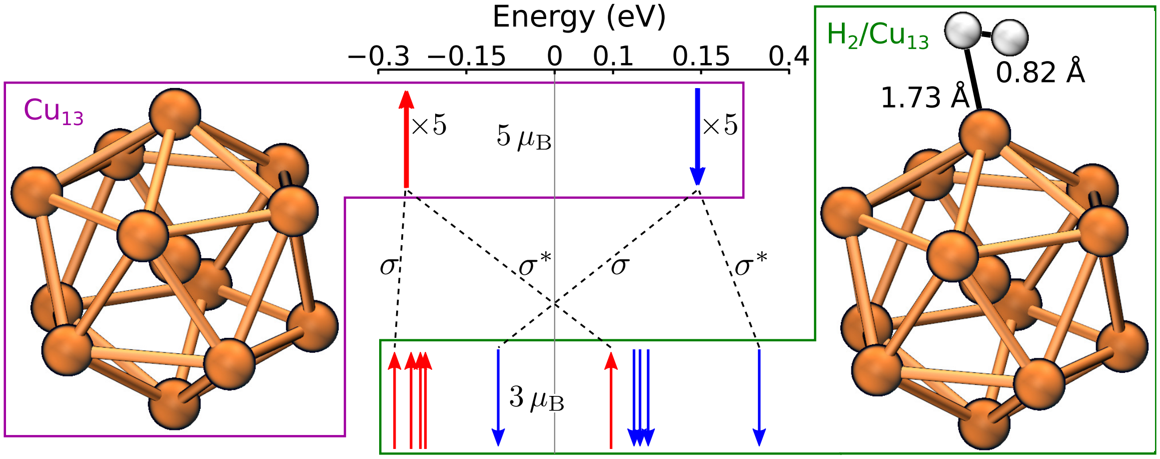


Figure 4.2: Illustration of how H-Cu hybridization drives the magnetization switch. Structural figures show the clean cluster (left, purple) and the one with H_2 adsorbed (right, green).

splitting.[107]

In contrast to bulk Cu and other transition metal surfaces, H_2 molecules can adsorb weakly on Cu_{13} intact (albeit weakly). There are two nearly isoenergetic geometries: one with a net magnetization of 5 μ_B and H–Cu bonds of length 1.90 Å, and another with 3 μ_B and H–Cu bonds of length 1.73 Å. The geometries are qualitatively similar, with the H–H bond axis centered above an apex Cu atom. The geometry with shorter bonds, shown in Figure 4.2, is marginally more stable with total energy lower by 0.04 eV. The reduced magnetization is caused by the shorter H–Cu bonds and increased hybridization with Cu orbitals, which pushes the energy of the antibonding majority-spin state above that of the bonding minority-spin state. This mechanism is illustrated in Figure 4.2, where the inversion of spin states produces the observed magnetization change by $-2\,\mu_\mathrm{B}$. Despite its greater stability, the lower magnetization geometry only has a weak adsorption energy of 0.09 eV, such that H_2 dissociation will most likely proceed from molecular collisions with Cu_{13} rather than thermal activation. The same is true of bulk Cu surfaces.[147]

4.2.2 MAGNETIZATION-DEPENDENT REACTIVITY

The initial magnetization of $5\,\mu_\mathrm{B}$ for isolated Cu_{13} differs from the magnetization of $3\,\mu_\mathrm{B}$ for the H_2/Cu_{13} complex, but the weak spin-orbit coupling of both H and Cu means that the magnetization switch required for adiabatic dissociation may not actually take place. We therefore obtain transition states and reaction pathways for a fixed magnetization state of $3\,\mu_\mathrm{B}$ and $5\,\mu_\mathrm{B}$ separately. As shown in Figure 4.3, the minimum-energy pathway for H_2 dissociation is more complicated than on the Cu(111) surface. H_2 first slides parallel to a Cu–Cu bridge (green to white in Figure 4.3), followed by bond cleavage and relaxation to the nearest local minimum (white to purple in Figure 4.3). The dissociated final state has one H atom at a hollow site and the other adsorbed at a Cu–Cu bridge site, the second-nearest bridge for the low magnetization structure and the third-nearest bridge for the high magnetization structure. In contrast to the smooth minimum-energy pathway on the Cu(111) surface, the minimum-energy pathways on Cu_{13} for both magnetization states have two major high-curvature regions as marked in Figure 4.3: one for the right-most H atom at the point of H_2 dissociation, and another for the left-most H atom during the subsequent descent to the final state.

The activation energy for H_2 dissociation is quite different in the two magnetization states: 0.20 eV in the low magnetization state and 0.67 eV in the high magnetization state. This dependence on magnetization is caused by increasingly large bonding-antibonding splitting along the minimum-energy pathway. In the adiabatic case, which involves a switch in magnetization

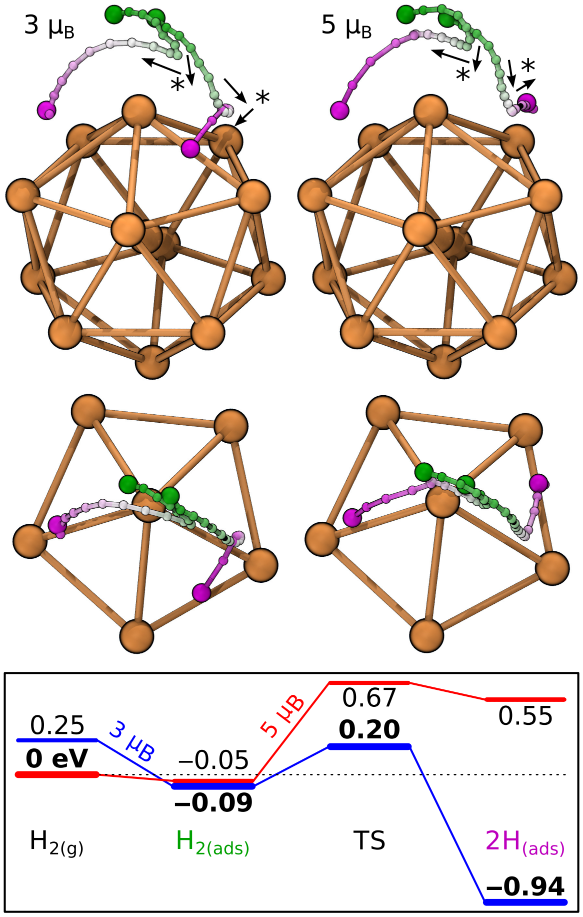


Figure 4.3: Above: minimum-energy pathways for H_2 dissociation at fixed magnetization of 3 μ_B and 5 μ_B. The green-white-purple gradient indicates reaction progress, with the initial and final states indicated by larger spheres. High curvature regions are marked with $*$. Below: energy of key intermediates, with the adiabatic ground-state energy in bold; the zero of energy is defined as the energy of the isolated H_2 and Cu_{13} at 5 μ_B.

from 5 μ_B to 3 μ_B, the high-energy antibonding state becomes unoccupied in favor of the lower-energy bonding state of opposite spin. The resulting activation energy is relatively small at 0.20 eV. In the limit of no spin-orbit coupling, however, the high-energy antibonding state remains occupied and the transition state is higher by 0.47 eV. In comparison, the Cu(111) and Cu(110) surfaces have reported activation energies around 0.4 eV to 0.5 eV[51,115,147] depending on the experimental or computational method. Since the low magnetization and high magnetization activation energies on Cu_{13} differ significantly, and straddle those of bulk Cu surfaces, H_2 dissociation rates depend critically on the probability of a magnetization switch.

4.2.3 MAGNETIZATION TRANSITION PROBABILITIES

To estimate the probability of a magnetization switch we ran 100 Born-Oppenheimer (BO) molecular dynamics simulations to obtain typical dissociation trajectories along the adiabatic potential energy surface. To ensure that each trajectory corresponds to H_2 dissociation, we start the simulation at the transition state geometry and run time backward to the associated state. Initial H_2 kinetic energies ranged from 0.05 to 0.40 eV, and velocity directions were chosen randomly. Since adiabatic dynamics are time-reversible, these initial conditions randomly sample H_2-Cu_{13} collisions which result in H_2 dissociation. The time-dependent magnetization for each trajectory exhibits the adiabatic magnetization switch from 3 μ_B to 5 μ_B as the H atoms recombine, involving the inversion of opposite-spin Kohn-Sham states shown in Figure 4.2. Analyzing each trajectory can then produce an estimate of the probability P_{LZ} of a magnetization switch using the Landau-Zener approximation (see Appendix A for details):

$$P_{LZ} = 1 - \exp(-2\pi\Gamma), \ \ \Gamma \equiv \frac{\delta^2/\hbar}{\left|\frac{\partial}{\partial t}\left(E_h(t) - E_l(t)\right)\right|} \tag{4.1}$$

where $E_h(t)$ and $E_l(t)$ are the potential energies at high magnetization (5 μ_B) low magnetization (3 μ_B), respectively, and $\delta = 8.15$ meV is the experimentally observed Rashba splitting for the Cu(111) surface.[163] The spin-orbit splitting can be estimated from first-principles using the DFT spin-orbit coupling Hamiltonian[33] ($\delta \approx 12$ meV, see Appendix A), but previous DFT calculations significantly overestimated the Cu(111) Rashba splitting[163] so we use the experimental value instead. Each randomized trajectory samples the denominator in the expression for Γ in Equation (4.1) (see Appendix A). P_{LZ} varies significantly with H_2 center of mass (CM) velocities away from Cu_{13} at the avoided crossing as shown in Figure 4.4. The fitting equation is derived by approximating the denominator of Γ in Equation (4.1), the crossing rate, as the CM velocity v_{CM}

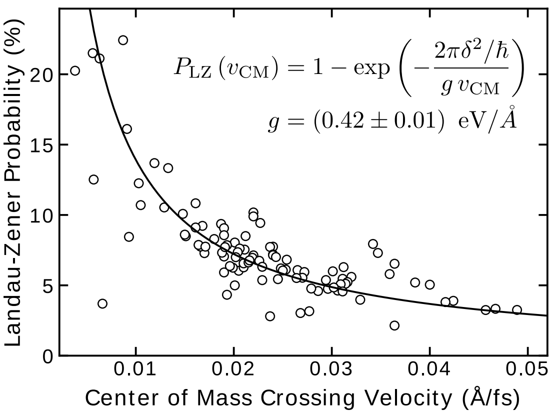


Figure 4.4: Landau-Zener probabilities P_{LZ} versus the H_2 center of mass velocity v_{CM} away from Cu_{13}. Open circles show results calculated from the 100 adiabatic molecular dynamics trajectories. The minimum mean absolute error fit of the Landau-Zener formula is indicated by the solid line.

away from the nanocluster multiplied by a scalar constant g, which is fitted to the data. Overall the median probability and interquartile range for P_{LZ} is $(6.2 \pm 1.4)\,\%$. Adiabatic transition state theory therefore overestimates reaction rates by a factor of $1/P_{LZ} \approx 16$.

4.3 POSSIBLE CATALYTIC IMPROVEMENTS

4.3.1 MAGNETIC EXCITATIONS

Since the weak spin-orbit coupling of Cu and H hinders the magnetization switch and hence reduces reaction rates, we now discuss several methods for improving Cu_{13} as a H_2 dissociation catalyst. One approach is to excite nanoclusters to the lower magnetization state in advance of a H_2-Cu_{13} collision. Such excitations would bypass the hindered magnetization switch and decrease activation energies by destabilizing the isolated nanocluster. Defining ΔE_{TS} to be the activation energy for H_2 dissociation relative to isolated H_2 and Cu_{13}, we use $\Delta E_{\mathrm{TS}}^{\mathrm{BO}}$ to refer to the Born-Oppenheimer limit where the system remains in its electronic ground state, and $\Delta E_{\mathrm{TS}}^{*}$ to refer to the excited initial state. Both values are shown for Cu_{13} and other clusters (discussed below) in Table 4.1. While $\Delta E_{\mathrm{TS}}^{\mathrm{BO}}$ is 0.20 eV, $\Delta E_{\mathrm{TS}}^{*}$ is negative so H_2 can dissociate over magnetically excited

Cu_{13} at arbitrarily slow collision velocities. These excitations could be induced using optical spin-flips and intense irradiation to compensate for the weak spin-orbit coupling of Cu.[66] Since the density of states is discrete near the Fermi level for both spin directions, the incident optical spectrum could be tuned to a narrow band around the exchange splitting of 0.47 eV ($\lambda \approx 2.7\ \mu$m) to specifically target the spin-flip excitation.

4.3.2 SUBSTITUTION AND CHARGE STATES

Another approach for improving catalytic rates is to modify the electronic structure of Cu_{13} so that a magnetization switch is no longer needed. The high activation energy for the high magnetization state is caused by occupying the highest antibonding state as shown in Figure 4.2. Depopulating this state using electron deficient clusters should result in activation energies that vary less with magnetization. To address this possibility, we consider electron deficient Cu clusters, Cu_{13}^{+}, or clusters formed by substitution of a surface Cu atom by an atom with fewer valence electrons, $Cu_{12}Ni$ and $Cu_{12}Pd$. Isolated Pd atoms in particular have been shown to facilitate H_2 dissociation on bulk Cu surfaces.[36,91] All three electron-deficient clusters we considered have magnetization of 4 μ_B in their ground state since one of the majority-spin antibonding states is no longer occupied. To investigate the effect of Ni and Pd substitution without changing the number of valence electrons, we also consider $Cu_{12}Ni^{-}$ and $Cu_{12}Pd^{-}$ anions with five valence electrons and ground-state magnetization of 5 μ_B. All clusters have nearly icosahedral geometries and similar densities of states near the Fermi level.

Table 4.1 summarizes the energetics of H_2 dissociation on all six nanoclusters we considered. H_2 adsorption only induces magnetization switches for clusters with five valence electrons (Cu_{13}, $Cu_{12}Ni^{-}$, and $Cu_{12}Pd^{-}$). However, as H–Cu hybridization increases along the minimum-energy pathway, clusters with ground state magnetization of 4 μ_B undergo a corresponding switch from 4 μ_B to 2 μ_B before reaching the transition state. Therefore the transition states for all six clusters

Table 4.1: Magnetization of the Isolated (m_I) and Adsorbed (m_{ads}) H_2/Nanocluster Systems, Adsorption Energy E_{ads}, and Activation Energies Relative to the Gas Phase (ΔE_{TS}) and Adsorbed H_2 (ΔE_{A})

Cluster	m_I	m_{ads}	E_{ads}	$\Delta E_{\text{TS}}^{\delta\to 0}$	$\Delta E_{\text{TS}}^{\text{BO}}$	ΔE_{TS}^{*}	$\Delta E_{\text{A}}^{\delta\to 0}$	$\Delta E_{\text{A}}^{\text{BO}}$
Cu_{13}	5	3	0.09	0.67	0.20	−0.05	0.72	0.29
Cu_{13}^{+}	4	4	0.34	0.29	0.10	−0.11	0.63	0.43
$Cu_{12}Ni$	4	4	0.77	−0.54	−0.58	−0.71	0.23	0.18
$Cu_{12}Ni^{-}$	5	3	0.69	−0.33	−0.60	−0.81	0.20	0.08
$Cu_{12}Pd$	4	4	0.53	−0.16	−0.19	−0.27	0.38	0.35
$Cu_{12}Pd^{-}$	5	3	0.47	−0.06	−0.30	−0.43	0.34	0.17

Energies are in eV, and magnetization values are in μ_B.

have lower magnetization than their corresponding ground states by 2 μ_B. Similar to bulk Cu surfaces, Ni and Pd dopants increase the H_2 adsorption energy while reducing the activation energy; all four clusters with substituted Cu atoms have lower activation energy than desorption energy so H_2 dissociation can proceed from both thermal activation and gas phase collisions. To distinguish between these possibilities, we use ΔE_{A} for activation energy relative to the adsorbed state, and ΔE_{TS} for the activation energy relative to isolated H_2 and the nanocluster. Superscripts indicate the relevant magnetic limits: $\delta \to 0$ for the zero spin-orbit coupling limit, and BO for the adiabatic limit following the Born-Oppenheimer approximation. $Cu_{12}Ni$ and $Cu_{12}Ni^{-}$ have the lowest activation energy regardless of the magnetization state.

As expected, electron-deficient clusters exhibit smaller changes in ΔE_{A} and ΔE_{TS} with respect to magnetization (see Table 4.1). For example, $\Delta E_{\text{TS}}^{\delta\to 0} - \Delta E_{\text{TS}}^{\text{BO}}$ is 0.47 eV for Cu_{13}, but only 0.19 eV for Cu_{13}^{+}. The adiabatic activation energy for all clusters is lower than the $\delta \to 0$ "fixed magnetization" limit, so adiabatic transition state theory overestimates H_2 dissociation rates for all six clusters. Magnetization differences in ΔE_{TS} are most important for Cu_{13} and Cu_{13}^{+} where H_2 dissociation mostly proceeds from collisions, while differences in ΔE_{A} are more important for clusters with substituted Cu atoms where thermally activated H_2 dissociation is likely.

Table 4.2: Incremental Adsorption Energies $E_{\text{ads}}^{(k)}$ and Activation Energies $\Delta E_{\text{TS}}^{(k)}$ for Multiple H_2 Adsorption on Cu_{13}

k	$E_{\text{ads}}^{(k)}$	$\Delta E_{\text{TS}}^{(k),\delta\to 0}$	$\Delta E_{\text{TS}}^{(k),\text{BO}}$
1	0.09	0.67	0.20
2	**0.30**	0.21	0.21
3	0.15	**0.23**	**0.01**

Energies are in eV, and the most stable arrangement and most common ΔE_{TS} values are shown in bold.

4.3.3 CO-CATALYSTS

Finally, we consider using co-catalysts to improve the performance of Cu_{13} in H_2 dissociation. One particularly simple strategy is to consider H_2 as its own co-catalyst: H_2 adsorption can induce the hindered magnetization switch for the dissociation of a subsequent H_2 molecule. This approach is feasible since H_2 grows more stable with increasing coverage. The adsorption and activation energy for increasing H_2 coverages is shown in Table 4.2, where $E_{\text{ads}}^{(k)}$ is the incremental adsorption energy of the kth H_2 molecule on $(k-1)H_2/Cu_{13}$. $2H_2/Cu_{13}$ has the highest incremental adsorption energy and should therefore be the most common complex at equilibrium. Adsorption and transition-state geometries still have ground-state magnetization of 3 μ_B except for H_2 dissociating over the $2H_2/Cu_{13}$ complex, where the magnetization is 1 μ_B due to further increases in H-Cu hybridization. We also find that the most common $2H_2/Cu_{13}$ arrangement also produces the lowest $\Delta E_{\text{TS}}^{\text{BO}}$ value of only 0.01 eV. Although a switch from 3 μ_B to 1 μ_B becomes necessary, the corresponding P_{LZ} is likely more than offset by the large 0.2 eV decrease in the activation energy.

4.4 CONCLUSION

In conclusion, we investigated the catalytic behavior of Cu_{13} clusters for H_2 dissociation and demonstrated that magnetization and spin-orbit coupling play an important role in this reaction. A combination of high symmetry in the cluster geometry and quantum size effects give Cu_{13} a large magnetic moment and yields magnetization-dependent catalytic behavior. The most favorable reaction pathway requires a magnetization switch, which is hindered by the weak spin-orbit coupling of H and Cu. The corresponding probability $P_{\mathrm{LZ}} \approx 6.2\%$ of a magnetization switch means that adiabatic transition state theory overestimates reaction rates for the H_2/Cu_{13} system by a factor of $1/P_{\mathrm{LZ}} \approx 16$. Therefore magnetization presents a new challenge for designing nanoscale catalysts for H_2 dissociation and hydrogenation. This magnetization dependence also offers new opportunities to control catalyst behavior by changing magnetic states. For example, ground-state Cu_{13} features a substantial collisional activation energy of 0.20 eV, but we predict that excited Cu_{13} which has lower magnetization will facilitate H_2 dissociation at arbitrarily slow collision velocities. The magnetization switch could be induced by optical spin-flip excitations. Appropriately designed catalysts with controllable magnetization, for example using optical or magnetic fields, could provide new ways to control catalyst activity without altering more conventional reaction parameters such as temperature, pressure, and reactant composition.

ACKNOWLEDGEMENTS

This work was supported by Integrated Mesoscale Architectures for Sustainable Catalysis, an Energy Frontier Research Center funded the U.S. Department of Energy, Office of Science, Basic Sciences under Award No. DESC0012573. Calculations were performed on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. We thank Professor Jaime Ferrer for providing his version of SIESTA with spin-orbit coupling.

5

Machine Learning

Apart from minor modifications, Section 5.4 is being prepared for the following publication:

> **Robert A. Hoyt**, Matthew M. Montemore, Ioanna Fampiou, Wei Chen, and Efthimios Kaxiras. "Understanding H Adsorption on Ag Alloys With Machine Learning." *in preparation*

5.1 INTRODUCTION

Before the advent of "fast computing machines,"[111] the only reliable data for materials and catalysts was to run experiments. Theoretical models certainly existed long before general-purpose computers, but their complexity was limited to those that can be solved or closely approximated by hand, and were generally applied as a way to rationalize experimental results *after* the fact. A famous example is the famous Newns-Anderson model[128] for adsorption on transition metals.[*] The model is based on empirical tight-binding Hamiltonians with approximate metal-hydrogen interactions. Theorists were not properly equipped to make reliable *first- principles* estimates of electronic and catalytic properties until the advent of Kohn-Sham DFT, and generalized gradient exchange-correlation potentials in particular. DFT has been highly successful but its speed and maximum practical system sizes remain limited, primarily since the cost of its underlying matrix operations scales cubically with the number of atoms. Machine learning algorithms represent a return to empiricism, but their dramatically lower computational costs could open the door to multiscale modeling and high-throughput screening with the accuracy of quantum chemistry methods if appropriately constructed.

---

[*] It should be noted that D. M. Newns hypothesized the "d-band model" decades before it was popularized by Hammer and Norskov in their famous "Why Gold is the Noblest of All the Metals" paper regarding H_2 dissociation. Moreover, silver is actually the noblest in this respect.[119]

Inspired by machine learning's meteoric growth over the past 20 years, and dramatic successes in advertising, finance, and computer vision, materials scientists have high hopes that machine learning will prove equally transformative in materials design. I remain cautiously optimistic. First, in the materials science community the phrase "machine learning" has an air of technical mysticism. Common misunderstandings, ambiguous jargon, and lack of good pedagogy complicate matters immensely. Second, the prevailing literature is primarily occupied with mimicking the results of DFT or other methods on familiar test sets. Major modeling innovations will be required to make machine learning in materials science suitable for the first-principles study of new systems. After all, DFT is widely used to gain first-principles insights into new systems, not to interpolate between the known results of more accurate quantum chemistry methods.

In this chapter I first discuss the theory behind machine learning for regression tasks. My hope is to substantially clarify machine learning, its terminology, and demonstrate the deep connections across popular models by introducing and discussing *kernel functions* in depth. Kernel functions are the very core of regression so understanding them gives deep insight into the models they produce. Next, I discuss the application of machine learning to predicting adsorption energies on dilute silver alloys.

5.2 MACHINE LEARNING OF REGRESSION MODELS

Although machine learning is a relatively late addition to my work, I consider this section to be the most important one in this thesis. It is broadly applicable to machine learning and data science in general. My hope is that it clearly explains the underpinnings of machine learning for single-variable regression. By discussing the *kernel* and the Mercer criterion from the very beginning, seemingly disparate models from linear regression to neural networks are immediately united under the same framework. These insights and connections are also related to principal component analysis and clustering methods. I hope this fresh take on regression is as useful to the reader as it was to me.

5.2.1 KERNELS

Our goal in regression is to construct a model $\hat{y}$ which estimates the value of an unknown, scalar function y of one or more variables $\mathbf{x}$. These vectors $\mathbf{x}$, called *feature vectors*, are points in *feature space*. Feature vectors $\mathbf{x}$ numerically describe some system of interest: surface alloys, customer reviews, stock market conditions, etc. In the context of statistics, vectors $\mathbf{x}$ describe sampling events and values y are the corresponding observed values. Common applications include interpolation, to estimate an unknown and generally noisy function between two known data points, and dimensionality reduction (or compression), to determine which subset of variables is most important. Both these applications are discussed in more detail below. A regression model is defined by the corresponding *kernel* that asserts how much the function's value $y(\mathbf{x})$ at $\mathbf{x}$ is related to its value $y(\mathbf{x}')$ at another point $\mathbf{x}'$. The kernel is a pair-wise scalar function $k(\mathbf{x},\mathbf{x}'|\beta)$ of any two points $\mathbf{x}$ and $\mathbf{x}'$, where β represents any parameters in the kernel. In the context of Gaussian processes, k asserts the covariance between the function's values at $\mathbf{x}$ and $\mathbf{x}'$. Kernels are are the core of regression methods since they dictate the relationship between the training values and new data points.

The key to a good learning model is therefore the selection of an appropriate kernel. By the same logic, there are poor choices as well. This raises an interesting question: is there a kernel with the best average performance over all possible functions y? The answer is *no*. David Wolpert's famous (albeit technical) "No Free Lunch" (NFL) theorems[186] prove that all models have identical mean performance in this case. This proves that that guessing zero is just as good as state of the art models. Fortunately, there are two important caveats to consider. First, while the NFL theorems prove identical mean performance over *all* functions, certain models can be much better than average for a particular class of functions. For example, most “real world” functions reliably exhibit at least some smoothness and continuity. Second, the NFL theorems only prove that the *mean* performance of all models is identical, so the *typical* (e.g. median) performance for some models can still be much better than average. The NFL theorems prove we must critically evaluate kernels to ensure they correctly reflect the target function’s properties. There are no silver bullets.

5.2.2 KERNEL TRICK

To interpret kernels as covariances, and imbue them with convenient mathematical properties, they are nearly always[†] chosen to satisfy what is popularly known as “Mercer’s condition:”[110]

$$k(\mathbf{x},\mathbf{x}'|\boldsymbol{\beta}) = \boldsymbol{\phi}(\mathbf{x}|\boldsymbol{\beta}) \cdot \boldsymbol{\phi}(\mathbf{x}'|\boldsymbol{\beta}) \tag{5.1}$$

where $\boldsymbol{\phi}$ is a vector-valued function that maps points in feature space $\mathbf{x}$ to points in *kernel space* $\boldsymbol{\phi}$. Mercer’s condition states that k can be nonlinear in feature space, but will be a simple inner product in its kernel space $\boldsymbol{\phi}$. This has major consequences. First, by inspection of the dot product, k is symmetric in $\mathbf{x}$ and $\mathbf{x}'$. Second, the dot product form shows that arbitrarily nonlinear behavior in feature space becomes linear in kernel space. In some cases, especially linear regression, both the kernel k and kernel space $\boldsymbol{\phi}$ are known and easy to compute. In most other cases, however, either k or $\boldsymbol{\phi}$ is unknown or computationally impractical.

A third property is that the dot product from ensures that the kernel is positive semidefinite.[110] To show this, consider any n samples at points $\mathbf{x}_i$ drawn from feature space. With the the *kernel*

[†] k-Nearest Neighbors (kNN) is the one prominent exception.

matrix[‡] $K_{ij} \equiv k(\mathbf{x}_i, \mathbf{x}_j|\beta) = \phi(\mathbf{x}_i|\beta) \cdot \phi(\mathbf{x}_j|\beta)$, then K is a symmetric and positive definite matrix. K is symmetric by inspection because k is symmetric. To prove K is positive definite, we calculate its inner product for an arbitrary test vector $\mathbf{u}$:

$$\begin{aligned} \mathbf{u}^T K\mathbf{u} &= \sum_{ij}^{n} u_i K_{ij} u_j && (5.2a)\\ &= \sum_{ij}^{n} u_i \phi(\mathbf{x}_i|\beta) \cdot \phi(\mathbf{x}_j|\beta) u_j && (5.2b)\\ &= \left(\sum_{i}^{n} u_i \phi(\mathbf{x}_i|\beta)\right) \left(\sum_{j} u_j \phi(\mathbf{x}_j|\beta)\right) && (5.2c)\\ &= \left(\sum_{i}^{n} u_i \phi(\mathbf{x}_i|\beta)\right)^2 && (5.2d)\\ \mathbf{u}^T K\mathbf{u} &\geq 0 && (5.2e) \end{aligned}$$

The combination of symmetry and being positive definite means that the kernel matrix K has positive semidefinite eigenvalues and its eigenvectors form an orthogonal basis. Furthermore, K can be rewritten using the spectral theorem, leading to the kernel-based definition of principle component analysis and clustering as discussed in Section 5.2.5.

5.2.3 KERNEL MATRIX AS A COVARIANCE MATRIX

Kernel matrices K have broad utility in machine learning. They explicitly appear in kernel regression models (see Section 5.2.4) and in kernel principal component analysis (see Section 5.2.5). However, given the importance of selecting appropriate kernels, it is important to be able to visualize and experiment with kernels to better understand their properties. This can be done using *Gaussian processes*.

A Gaussian process is a probability distribution over some domain, in our case feature space, where the value at every point is normally distributed. Knowing the covariance between two points $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and the mean $\boldsymbol{\mu}$, dictates the joint distribution for the observed values. In other words, the values are normally distributed, and are coupled to each other by known variances K_{ij}. The joint distribution is a multivariate normal:

[‡] More commonly known as the "Gramian," but Gramian matrices are more general than kernels so I find the term "kernel matrix" more concrete.

$$\hat{\boldsymbol{y}}\left(\mathbf{x}_1, \cdots, \mathbf{x}_n\right) \sim \mathcal{N}\left(\boldsymbol{\mu}, K\right) \tag{5.3}$$

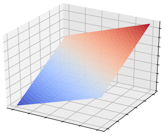
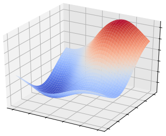
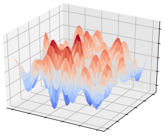
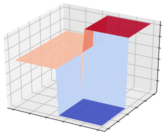
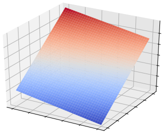
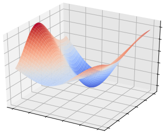
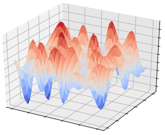
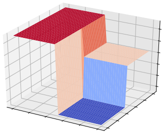
Equation (5.3) is also called the *Gaussian process prior*, and $\hat{\mathbf{y}}$ is the vector of model values $\hat{y}_i$. It is the probability distribution of observing any particular sequence of y_i values given sample points $\mathbf{x}_i$. Gaussian processes can be regression models in their own right, by *conditioning* this distribution given some already-observed training data. However, a thorough discussion of Gaussian processes and their applications as regression models is outside my expertise and could be a thesis in its own right.[30] Instead, I rewrite Equation (5.3) in their uncoupled form to provide a bit more clarity, then discuss a useful application for these equations:

$$K = U \Sigma U^T \equiv U \begin{pmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_n^2 \end{pmatrix} U^T \qquad \text{spectral theorem} \tag{5.4a}$$

$$\hat{\boldsymbol{y}} \sim \boldsymbol{\mu} + U \begin{pmatrix} \mathcal{N}\left(0, \sigma_1^2\right) \\ \vdots \\ \mathcal{N}\left(0, \sigma_n^2\right) \end{pmatrix} \tag{5.4b}$$

where $U \equiv (\hat{\mathbf{u}}_1 \cdots \hat{\mathbf{u}}_n)$ is the unitary matrix whose columns are the eigenvectors of K, and Σ is a diagonal matrix with K's eigenvalues along its diagonal. This whole process is a kind of Fourier analysis in disguise. In Fourier analysis, we transform the "unit vectors" $(\hat{\boldsymbol{y}})_i = \hat{y}_i$ into the basis of complex exponentials using a unitary transformation. Fourier transforms are broadly applicable

Table 5.1: Randomly Drawn Functions from Selected Kernels in Two Dimensions

Sample	Linear	RBF	Sinc	Tree
	$\mathbf{x}\cdot\mathbf{x}'$	$\exp(-\lvert\mathbf{x}-\mathbf{x}'\rvert^2)$	$\frac{\sin(\lvert\mathbf{x}-\mathbf{x}'\rvert)}{\lvert\mathbf{x}-\mathbf{x}'\rvert}$	$\frac{\delta_{\ell(\mathbf{x})\ell(\mathbf{x}')}}{\lvert\ell(\mathbf{x})\rvert}$
1				
2				

when operators are coupled in real space, but uncoupled in reciprocal space. In the same way, the unit vector values $\hat{y}_i$ are correlated according to K, but rotating to the basis described by U decorrelates them. Similar to Fourier modes, the basis U contains the "normal modes" of K that vary independently. This is what leads to the single-variable normal distributions in Equation (5.4b).

The most important insight from Gaussian processes is that Equation (5.3) allows us to sample, or *draw*, functions themselves from a kernel. After choosing some points $\mathbf{x}_i$, we construct K_{ij} and sample from the normal distribution to obtain typical $\hat{y}_i$ values. Multiple samples can be drawn quickly using the decoupled form of Equation (5.4b) once we have the eigenvectors and eigenvalues of K. In this case we can choose $\boldsymbol{\mu} = \mathbf{0}$ to focus entirely on variations of functions. Table 5.1 shows functions drawn from several kernels. Among others, linear, RBF, and tree kernels are discussed in more detail in the following sections.

5.2.4 KERNEL REGRESSION

Through some unfortunate series of events, many machine learning publications and presentations present "kernel learning" as being somehow distinct from linear regression, neural networks and random forests. In this section I demonstrate how all these common regression models are kernel methods. The question is not “is this kernel regression?”, but rather “which kernel is being used?”

For some kernel k satisfying the Mercer criterion, we can obtain the two general forms of regression used in machine learning. Consider a training set of n data points $\mathbf{x}_i$ and corresponding sampled values y_i. Kernel regression yields the following model $\hat{y}$ for new data points $\mathbf{x}$:

$$
\begin{aligned}
\hat{y}(\mathbf{x}) &= \sum_i^n \alpha_i k(\mathbf{x}_i, \mathbf{x}|\beta) && \text{(5.5a)}\\
&= \sum_i^n \alpha_i \phi(\mathbf{x}_i|\beta) \cdot \phi(\mathbf{x}|\beta) && \text{(5.5b)}\\
&= \left(\sum_i^n \alpha_i \phi(\mathbf{x}_i|\beta) \right) \cdot \phi(\mathbf{x}|\beta) && \text{(5.5c)}\\
&\equiv \mathbf{w} \cdot \phi(\mathbf{x}|\beta) && \text{(5.5d)}
\end{aligned}
$$

where the coefficients α_i are chosen to minimize some cost function C, usually including the total squared error. Equation (5.5a) is the form implied by mentions of “kernel learning” or “kernelized [method]” in the literature. In this form, the coefficients α_i, and possibly β as well (see sections 5.2.4.4 and 5.2.4.5), are optimized. These models are generally solved by forming the *kernel matrix*, $K_{ij} \equiv k(\mathbf{x}_i, \mathbf{x}_j|\beta)$ (see Section 5.2.2) to produce an equivalent matrix equation to be solved with standard linear algebra routines. While the solutions to these equations scale as $O(n^3)$, they can still be routinely used on training sets with up to tens of thousands of points.

On the other hand, when the kernel space ϕ is known, the dot product representation can be used to obtain the “linearized” form in Equation (5.5d). The dot product identity allows the sum over kernel vectors and their α_i coefficients to “collapse” to the constant vector of coefficients, or *support vector*, $\mathbf{w}$. Using Equation (5.5d) therefore yields a simple linear regression task. Taking advantage of this linear form is referred to as “the kernel trick.” Moreover, the linear form can be interpreted as plane in kernel space with $\mathbf{w}$ as its normal vector. In support vector methods, regression and classification are interpreted in terms of distances from this dividing surface.

Equation (5.5d) reveals that complicated and nonlinear boundaries in feature space correspond to simple hyperplanes in kernel space.

Overall, Equation (5.5d) is an extremely important result for machine learning. Nearly every common kernel satisfies Mercer's criterion so nonlinear models in feature space $\mathbf{x}$ are *necessarily* linear in feature space ϕ. One major consequence is that the linear form simplifies the statistical analysis of kernels so that their relative strengths and limitations can be better understood. Second, the linear form helps us choose appropriate kernels for regression tasks. For example, let $\Delta\hat{y}$ be the difference in model output between points $\mathbf{x}$ and $\mathbf{x}'$. Then we have

$$\Delta\hat{y} \equiv \hat{y}(\mathbf{x}) - \hat{y}(\mathbf{x}') \tag{5.6a}$$

$$= \mathbf{w}\cdot\phi(\mathbf{x}|\beta) - \mathbf{w}\cdot\phi(\mathbf{x}'|\beta) \tag{5.6b}$$

$$= \mathbf{w}\cdot\big(\phi(\mathbf{x}|\beta) - \phi(\mathbf{x}'|\beta)\big) \tag{5.6c}$$

This can be used to guide the selection of kernel functions. If we know that $\Delta\hat{y}$ should vanish, e.g. adsorbate stability at symmetrically equivalent adsorption sites, then we want $\phi(\mathbf{x}) \approx \phi(\mathbf{x}')$ to ensure that $\Delta\hat{y} \approx 0$ regardless of the value of $\mathbf{w}$. Equivalently, from Mercer's criterion, we would choose a kernel k that returns a large value (i.e. high covariance) between the two related points.

Some of the ambiguity in machine learning terminology is caused by the fact that, in most cases, either the kernel k or the kernel space ϕ is unknown. This either-or situation then leads to branching terminology even though both conditions are equivalent (at least in principle) due to Mercer's criterion. Methods where the kernel is known include linear regression and "kernel" support vector regression, while methods using explicit kernel spaces ϕ include neural networks and literature efforts aimed at "feature engineering."

5.2.4.1 LINEAR AND POLYNOMIAL REGRESSION

Linear regression corresponds to the "dot product" kernel

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}' \qquad \text{dot product kernel} \tag{5.7}$$

Comparison of Equations 5.1 and 5.7 shows that this is the special case where $\phi(\mathbf{x}) = \mathbf{x}$, i.e. feature space and kernel space are identical. When this ϕ is used in Equation (5.5d), we obtain the linear regression model:

$$\hat{y}(\mathbf{x}) = \mathbf{w} \cdot \phi = \mathbf{w} \cdot \mathbf{x} = w_1 x_1 + \cdots + w_m x_m \tag{5.8}$$

for m-dimensional feature vectors $\mathbf{x}$. Non-zero offsets in the model are generally avoided by first shifting $\mathbf{x}_i$ and y_i values to have zero mean. Another alternative is to add a constant "dummy" feature, $(x_1 \cdots x_m,) \rightarrow (x_1 \cdots x_m, 1)$, so that the new last element of $\mathbf{w}$ yields the linear offset. Shifting is generally the better strategy since an offset variable does not need to be optimized.

To interpret the dot product kernel in terms of covariance, consider the sketch of a one dimensional linear model in Figure 5.1. As discussed previously $k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$ is a good match for this data because the data clearly follow a linear trend in feature space $\mathbf{x}_i$. When both points are far from the origin along similar directions, such as points x_1 and x_2 in Figure 5.1, $k(x_1, x_2) \gg 0$. Large k values in this case reflect the fact that large values $y(x_1)$ strongly imply large $y(x_2)$ due to the slope, and vice-versa. Increasingly distant points have ever-increasing effects on the slope, leading to k's divergence as $|x_1|, |x_2| \rightarrow \infty$. Similarly, we see that positive $y(x_1)$ values imply negative $y(x_4)$ values through the slope as well. Finally, with the line centered at zero, the point x_3 near the origin is not expected to correlate with points elsewhere like x_1 and x_4. Points near the origin, like $x_3 \approx 0$, yield $k(x_3, x_1) = x_3 \cdot x_1 \approx 0$. This same logic applies to higher dimensions where

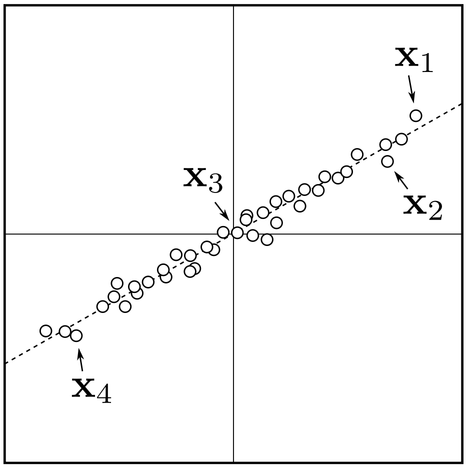


Figure 5.1: Sketch of a typical linear relationship. Values $x_1 \cdots x_4$ are labeled for reference for discussion in the main text.

inner products are used instead of scalar multiplication. Typical two-dimensional linear functions drawn from the Gaussian prior with the dot product kernel are shown in Table 5.1 for reference.

Polynomial regression is an extension of Equation (5.8), where an order p polynomial has up to p factors of $x_1 \cdots x_m$ in each term. For example, a quadratic model for a two-dimensional feature space $\mathbf{x}^T = (x_1, x_2)$ is

$$\hat{y}(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_{11} x_1^2 + w_{12} x_1 x_2 + w_{22} x_2^2 \tag{5.9}$$

Offsets can be handled just as discussed previously for linear regression. Comparison of Equations 5.5d and 5.9 shows that ϕ for quadratic regression is a vector of the polynomial terms:

$$\phi(\mathbf{x})^T = \left(x_1, x_2, x_1^2, x_1 x_2, x_2^2\right) \tag{5.10}$$

Vectors in kernel space ϕ can be constructed for cubic polynomials by appending 3-fold products, quartic polynomials by adding 4-fold products, etc. Two representative draws from Gaussian processes using the quadratic kernel are shown in Figure 5.2.

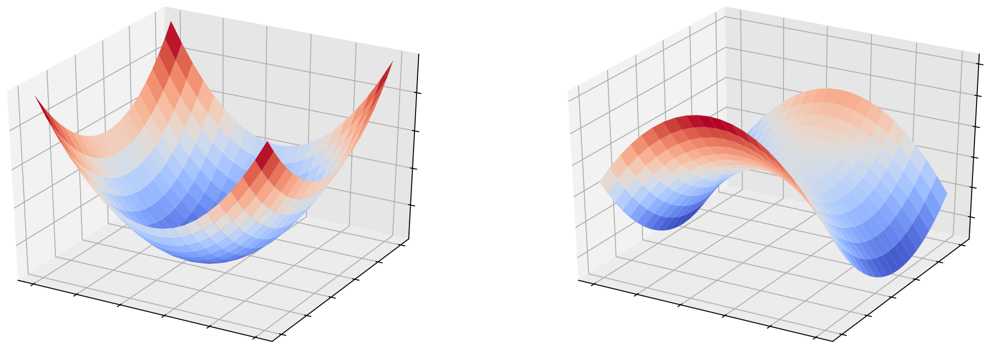


Figure 5.2: Two functions drawn randomly from a Gaussian process using the quadratic kernel as shown in Equations 5.3 and 5.4b. Each draw can yield positive or negative curvature along any axis.

5.2.4.2 SUPPORT VECTOR REGRESSION

Support vector regressors (SVRs) are another popular class of models where k is known instead of ϕ. Despite their deep connections to linear regression, machine learning terminology differs significantly. In the literature SVRs are interpreted as models consisting of a “decision boundary” $S(\mathbf{x}) = 0$, and $\hat{y}(\mathbf{x}')$ is the closest distance from $\mathbf{x}'$ to the surface S. The simplest dividing surface is a simple hyperplane. If we let the normal vector to the hyperplane be $\hat{\mathbf{w}}$, then the dividing surface is

$$S(\mathbf{x}) = \hat{\mathbf{w}} \cdot \mathbf{x} - c = 0 \tag{5.11}$$

Distances $\hat{y}(\mathbf{x})$ from this hyperplane can then be calculated. Letting $\mathbf{x}_p$ be any point on S, we obtain

$$\hat{y}(\mathbf{x}) = \hat{\mathbf{w}} \cdot \mathbf{x} - \hat{\mathbf{w}} \cdot \mathbf{x}_p \tag{5.12a}$$

$$= \hat{\mathbf{w}} \cdot \mathbf{x} - c \tag{5.12b}$$

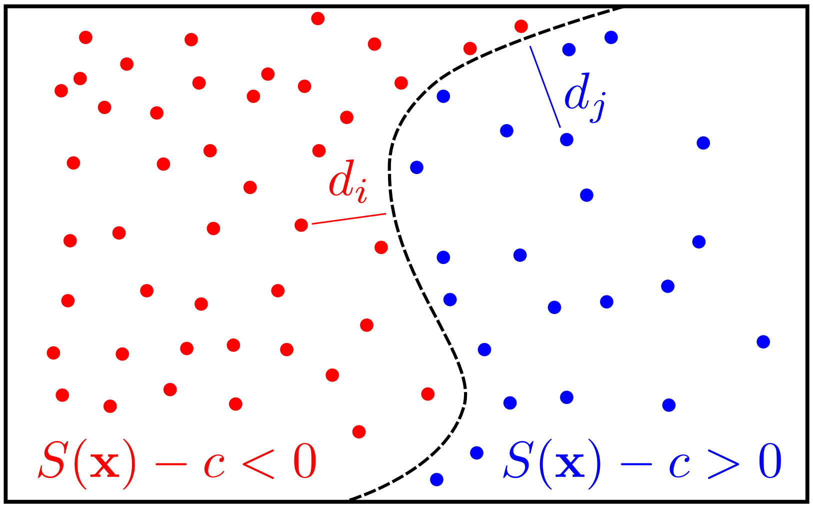


Figure 5.3: Sketch of support vector classification in two dimensions. The dashed line shows the dividing surface S. Points are classified according to which side of the surface they reside: red for $S(\mathbf{x}) - c < 0$, and blue for $S(\mathbf{x}) - c > 0$.

If we relax the requirement that $\hat{\mathbf{w}}$ must be a unit vector, then it becomes exactly the same as the the linear model in Equation (5.8) with an offset.

Matters are more interesting for nonlinear SVRs, where the surface S is nonlinear (i.e. curved) in feature space. They are implicitly used in support vector *classifiers* (SVCs), which seek a dividing surface S that separates data points in one subgroup from those in the other. An example of such a task is shown in Figure 5.3. SVCs then determine the class of any new point by determining which side of S they are on. S is usually chosen to minimize the total distance to all incorrectly classified points ("soft margin" training), so determining distances from S is an important task. However, framing a regression model in terms of distances to curved surfaces is rather unintuitive. For this reason so-called "kernel regression," described in the following section, is much more popular. SVR and kernel regression are equivalent by the Mercer criterion; only their interpretation is different.

5.2.4.3 KERNEL REGRESSION

As discussed previously, the name "kernel regression" is counterintuitive because nearly *all* regression is kernel regression. In an ideal world, then, "linear regression" would simply be renamed "linear kernel regression." Nevertheless I use kernel regression here in the same way it is used in the literature.

Kernel regression is the class of models where the kernel function k is known and used explicitly as in Equation (5.5a):

$$\hat{y}(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}|\beta) \tag{5.13}$$

for a set of training points $\mathbf{x}_i$ and kernel parameters β. Although ϕ is unknown in general, satisfying the Mercer criterion still guarantees that this model is equivalent to linear regression problem in kernel space as shown by Equations 5.5a and 5.5d. Kernel regression is also trivially linked to SVRs by noting that the constraint $\hat{y}(\mathbf{x}_p) = c$ defines a nonlinear surface, and that the distance of some point $\mathbf{x}$ from this surface can be written as $\hat{y}(\mathbf{x}) - \hat{y}(\mathbf{x}_p) = \hat{y}(\mathbf{x}) - c$. Moreover, invoking the Mercer condition shows that this surface is a linear hyperplane in kernel space:

$$\begin{aligned}
0 &= \hat{y}(\mathbf{x}) - c && (5.14a)\\
&= \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}|\beta) - c && (5.14b)\\
&= \left(\sum_i \alpha_i \phi(\mathbf{x}_i|\beta)\right) \cdot \phi(\mathbf{x}|\beta) - c && (5.14c)\\
&\equiv \mathbf{w} \cdot \phi(\mathbf{x}|\beta) - c && (5.14d)
\end{aligned}$$

Distances from this hyperplane therefore yield regression due to the same link between linear SVRs and linear regression.

Kernel regression is broadly applicable because a few of them, especially the RBF (or "squared exponential") kernel, tend to perform well for most real-world regression tasks. The RBF kernel is really a normal distribution in disguise:

$$k_{\text{RBF}}(\mathbf{x},\mathbf{x}'|\sigma) = \exp\left(-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2\sigma^2}\right) \tag{5.15}$$

The RBF kernel asserts that nearby values tend to have similar function values, i.e. are strongly correlated with $k_{\text{RBF}} \lesssim 1$, while distant points are uncorrelated with $k_{\text{RBF}} \gtrsim 0$. RBF kernel values monotonically approach zero as with increasing distance $|\mathbf{x}-\mathbf{x}'|$. For this reason RBF is often described as being a "localized" or "stationary" kernel. RBF kernels are infinitely differentiable and have analytic gradients. These locality and smoothness properties give the RBF kernel its broad applicability since most functions of interest are also smooth and continuous.

An advantage of kernel regression is that additional properties can be included in the kernel. For instance, values of a periodic functions are highly correlated at specific displacements along a particular direction. Correlation between points in these different fringes can be asserted using a cosine or other periodic function. If the periodic correlation only holds over limited length scales, the periodic kernel can be multiplied with a localized kernel such as RBF so long-range correlations vanish. Figure 5.4 shows several randomly drawn functions from such a kernel for several different length scales σ_{RBF}. Large σ_{RBF} values yield long-range correlations in the left-most panel of Figure 5.4, while shorter length scales yield only local cosine-like correlation. For distances much shorter than σ_{RBF}, $k_{\text{RBF}} \lesssim 1$ so most of the variation is due to the cosine kernel. For long distances, k_{RBF} decreases to 0 so distant points exhibit minimal correlation. Multiplying two kernels is similar to a logical `AND` operation in the sense that the product only yields significant correlation between two points if *both* kernels are large. This is just one example of how kernels can be combined to generate increasingly sophisticated models.[30]

With this more thorough discussion of kernels, we return to the No Free Lunch theorems. The

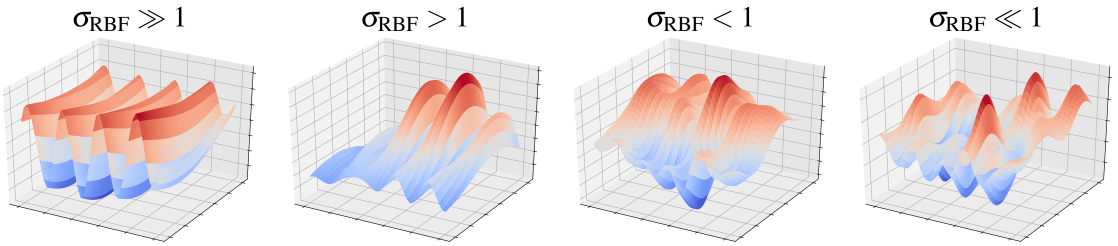


Figure 5.4: Random two-dimensional functions drawn from a Gaussian process with the composite kernel $k(\mathbf{x},\mathbf{x}') = k_{\cos}(\mathbf{x},\mathbf{x}') \cdot k_{\text{RBF}}(\mathbf{x},\mathbf{x}')$ and varying RBF length scales σ. The σ parameter is largest for the left-most panel, resulting in long-range periodicity, and smallest for the right-most panel, resulting in only short-range coherence.

RBF kernel (and other "general-purpose" kernels) perform reasonably well for most data sets, but only because it makes very mild assertions about smoothness and locality. Most functions have additional properties such as known asymptotic limits and symmetries that are not included by RBF. Kernels including these additional symmetries will outperform kernels lacking them.

The lack of specific function attributes in RBF and other general-purpose kernels leads to inefficient use of data. Starting from Equation (5.5a), we can write the regression model resulting from the RBF kernel:

$$\hat{y}_{\text{RBF}}(\mathbf{x}) = \sum_i \alpha_i \, k_{\text{RBF}}(\mathbf{x}_i, \mathbf{x} | \sigma) \tag{5.16a}$$

$$= \sum_i \alpha_i \exp\left(-\frac{|\mathbf{x}-\mathbf{x}_i|^2}{2\sigma^2}\right) \tag{5.16b}$$

Equation (5.16b) reveals that "RBF kernel regression" is another way of writing "a sum of Gaussians." With n training points, Equation (5.16b) shows there are n independent parameters. On the other hand, we generally know (or at least believe) that the sampled function is given by a small number p of physically important variables. The large excess of parameters $n \gg p$ suggests that the RBF kernel adds too much flexibility to the model. In many cases this excess flexibility can make model predictions *worse*, a phenomenon known as *overfitting*. Section 5.3.2 discusses methods designed to mitigate overfitting.

Returning to the specific example of periodic functions, the RBF kernel only gives the short-range correlation between data points so several Gaussians are required to describe just a single ridge in Figure 5.4. Modeling two fringes would require twice as many Gaussians, and three fringes would require three times as many. Moreover, the RBF kernel vanishes a long distances so $\hat{y}_{\mathrm{RBF}}$ will predict 0 for distant data points instead of following a periodic trend. Vanishing at long distances also leads to other unintended consequences such as misleading error estimates for distant points.[144] In comparison, a cosine kernel with the correct periodicity would yield a superposition-of-cosines model that provides better predictions and generalization with fewer samples, and does not vanish at long distances. While the RBF kernel is certainly reasonable for localized trends, we should keep its limitations and the NFL theorems in mind when selecting an appropriate model.

5.2.4.4 NEURAL NETWORKS

Neural networks (NNs) and deep learning are immensely popular. Computational implementations of neural networks were some of the first uses of general-purpose computers and a hot topic in artificial intelligence research in the late 1950s and 1960s.[131] The popularity of neural networks comes from a combination of empirical and mathematical evidence that they can "learn" arbitrarily complex functions[27] as well as their passing similarity to neurons. Anything that claims similarity to neurons and artificial intelligence is obviously cool. Unsurprisingly, the behavior of neural networks is therefore discussed with great ambiguity. In this section, however, I use the the previous discussion of kernels to offer a simple explanation for their behavior. In particular I demonstrate in Equations 5.19c and 5.19d that neural networks learn by approximating the kernel space function $\phi(\mathbf{x}|\boldsymbol{\beta})$.

The most common neural network design is the classic "multilayer perceptron." A schematic of the model is shown in Figure 5.5. The input nodes in layer $\ell = 0$, or perception neurons (e.g.

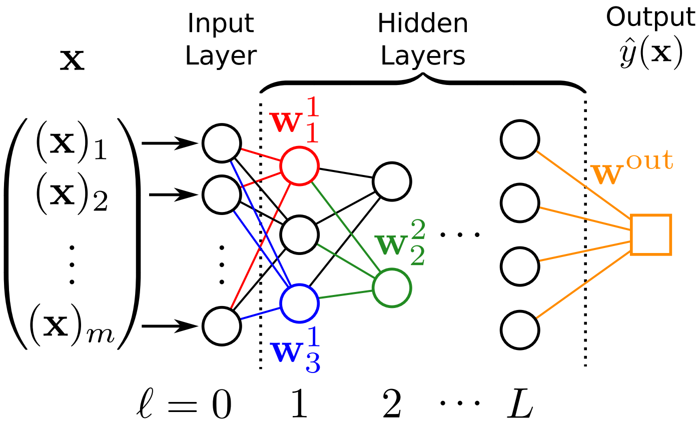


Figure 5.5: Sketch of a typical multilayer perceptron neural network. Input nodes are assigned the entries of $\mathbf{x}$. Subsequent nodes in hidden layer ℓ compute nonlinear functions of the weighted sum over node values from layer $\ell - 1$. Color coding highlights which edges correspond to several weight vectors $\mathbf{w}_j^\ell$. The output is drawn as a box to distinguish it from hidden nodes.

sensory information), are assigned the feature values $\mathbf{x}$. These nodes are then "connected" by edges (axons) to nodes in the next layer $\ell = 1$. Multiple layers, which may have different numbers of nodes, are then connected in a hierarchical way to obtain an L-layer network. Each node i in layer ℓ computes its own weighted sum s_i^ℓ based on the previous layer's values:

$$s_i^L = \mathbf{w}_i^\ell \cdot \boldsymbol{\sigma}^{\ell-1} + b_i^L \tag{5.17}$$

where $\boldsymbol{\sigma}^\ell$ is a vector of the previous layer's values, $\mathbf{w}_i^\ell$ is its weights vector, and b_i^ℓ is the node's own constant offset or "bias." The value s_i^ℓ can be roughly interpreted as an activation potential for the neuron. To give the network nonlinear behavior, the actual value the node is $\sigma_i^\ell = \sigma(s_i^\ell)$, where $\sigma(s)$ is a nonlinear function. Typical choices include simple sigmoids, hyperbolic tangents, or diode-like rectification functions. These serve as each neuron's response to its activation potential s_i^L. The final regression output of the L-layer network is a final weighted sum over the last layer's values $\boldsymbol{\sigma}^L$ with a final bias term b^{out}. In summary, neural networks are dictated by the following relationships:

$$\sigma_i^0 = x_i \tag{5.18a}$$

$$\sigma_i^\ell = \sigma\left(\mathbf{w}_i^\ell \cdot \boldsymbol{\sigma}^{\ell-1} + b_i^\ell\right) \tag{5.18b}$$

$$\hat{y}_{\mathrm{NN}}(\mathbf{x}) = \mathbf{w}^{\mathrm{out}} \cdot \boldsymbol{\sigma}^L\left(\mathbf{x}\middle|\left\{w_i^\ell, b_i^\ell\right\}\right) + b^{\mathrm{out}} \tag{5.18c}$$

The correspondence between $\mathbf{w}_j^\ell$ and inter-layer connections is shown for several nodes in Figure 5.5.

While the behavior of neural networks appears complicated at first glance, the resulting model is easily interpreted in terms of kernel regression. First, note that $\hat{y}_{\mathrm{NN}}$ in Equation (5.18c) is linear in the output layer values $\boldsymbol{\sigma}^L$. Comparing the "kernel trick" in Equation (5.5d) to Equation (5.18c) then reveals the deep (and almost trivial) connection between neural networks and kernel regression:

$$\hat{y}_{\mathrm{NN}}(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}|\boldsymbol{\beta}) \leftrightarrow \mathbf{w}^{\mathrm{out}} \cdot \boldsymbol{\sigma}^L(\mathbf{x}|\left\{w_i^\ell, b_i^\ell\right\}) + b^{\mathrm{out}} \tag{5.19a}$$

$$\Downarrow \tag{5.19b}$$

$$\mathbf{w} \leftrightarrow \mathbf{w}^{\mathrm{out}} \tag{5.19c}$$

$$\boldsymbol{\phi}(\mathbf{x}|\boldsymbol{\beta}) \leftrightarrow \boldsymbol{\sigma}^L\left(\mathbf{x}\middle|\left\{w_i^\ell, b_i^\ell\right\}\right) \tag{5.19d}$$

where the linear offset b^{out} can be handled as in linear regression by appending 1 to $\boldsymbol{\phi}$. Equation (5.19c) shows that the output weights are its support vector, and Equation (5.19d) shows that the neural network's last layer is its kernel space function $\boldsymbol{\phi}(\mathbf{x})$. For neural networks, the kernel parameters $\boldsymbol{\beta}$ are the collection of all its edge weights and bias terms. *Neural networks learn the kernel space* $\boldsymbol{\phi}(\mathbf{x})$ *leading to a linear model.* Increasingly nonlinear $\boldsymbol{\phi}$ functions can be produced by adding more layers to the network, while higher-dimensional kernel spaces are obtained by increasing the number of nodes per layer.

This deep connection clarifies the origins of neural networks' broad success in machine learning. Since the basic nonlinear σ functions are simple, they can be iteratively combined to form the more complicated functions required to map feature space to kernel space. The continuous and smooth properties of typical σ functions are also inherited by the resulting kernel space functions ϕ_{NN}. Neural networks are therefore a major example of regression models with known kernel space functions ϕ.

Compared to other kernel regression models, training a neural network involves optimizing both the support vector $\mathbf{w}$ and the kernel parameters $\beta \leftrightarrow \left\{w_i^\ell, b_i^\ell\right\}$ *at the same time*. In contrast to the linear kernel, with no parameters, and popular general-purpose kernels like RBF, with $O(1)$ parameters, neural networks often have at least several dozen parameters. In fact, with sufficient computational power and large data sets it is even possible to train and use neural networks with *billions* of parameters.[169] This flexibility lets neural networks handle wide ranges of regression tasks. Of course, this flexibility has associated costs: with large numbers of parameters, large training sets must be used to ensure all kernel parameters are well defined.

5.2.4.5 DEEP LEARNING

Deep learning refers to model fitting where both the coefficients and the kernel parameters β that enter the kernel are optimized along with the coefficients. This is often the case for *all* regression models, e.g. optimizing σ for the RBF kernel, but "deep" is generally reserved for very flexible kernels with many parameters. The most common form of deep learning is neural networks, where β is the collection of all neural network parameters in the "hidden layers" between the input layer and output value $\hat{y}$. As a result of the historically strong distinction between neural networks and kernel regression, the terminology has diverged. "Deep learning" in the literature refers to β optimization for neural networks and similar models, while optimizing β for kernel regression is

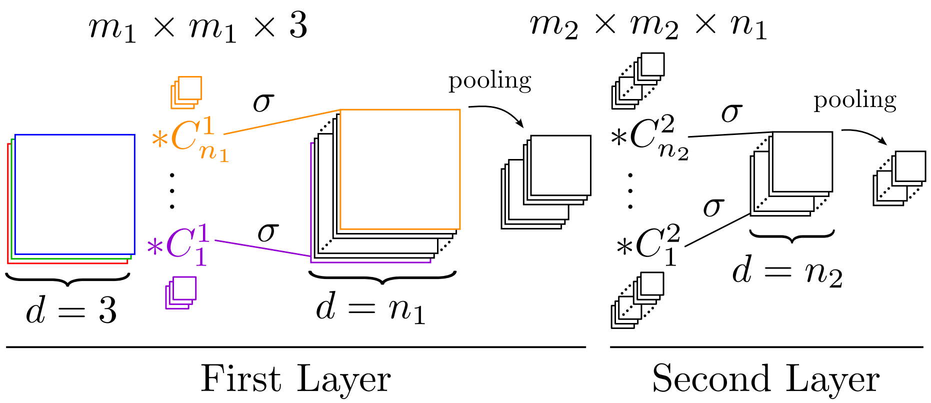


Figure 5.6: Sketch outlining the operations involved in the first two layers of a convolutional neural network. Large frame stacks represent images throughout the sequence of operations, while small stacks represent the convolutional filters. The correspondence between convolutional filters C and their resulting output frames is highlighted through color-coding for the first layer.

often called "hyperparameter optimization."

One of the major advances in deep learning, beyond neural networks themselves, arrived with the advent of *convolutional* neural networks. These networks were largely popularized by their excellent performance in optical character recognition without any hand-picked features or labeling.[94] Convolutional networks first perform several rounds of convolution operations, then pass the results on to a conventional neural network. As sketched in Figure 5.6, each layer ℓ in a convolutional network can be considered a stack of d^ℓ equally-sized images. The input (layer $\ell = 0$) might correspond to a color image, a stack of $d^0 = 3$ RGB images, or an audio waveform with one "image" per microphone. To obtain the next layer, $\ell = 1$, the input stack is convolved with n_1 smaller *filters* of size $m_1 \times m_1 \times d^0$, where m_1 is much less than the dimensions of the input image. As discussed previously, a nonlinear σ function is applied to each of these convolution values to obtain nonlinear behavior. An optional final step in the layer is a *pooling* operation which down-samples the image. A popular pooling operation is *max pooling*, where each $p \times p$ region (usually $p = 2$) in the image is replaced by its maximum value. These operations complete the first layer of the convolutional neural network. Subsequent layers proceed in exactly the same manner, creating d_ℓ new images from layer $\ell - 1$ using corresponding filters of size $m_\ell \times m_\ell \times d_{\ell-1}$.

Convolutional networks greatly outperform conventional neural networks for image recognition tasks. The main benefit they provide is *translational invariance*. Rather than training a conventional neural network to recognize the same object in several different locations, convolutions by definition scan each filter across the entire input image. A limited degree of noise tolerance is also introduced by having multiple pixels in each filter, as well as the pooling operations, which effectively act as a low-pass filter by coarse-graining the input image. Pooling operations' coarse-graining is also a lossy operation, introducing some spatial ambiguity and hence a limited degree of distortion and scale invariance.

Deep learning techniques are starting to make their way into materials science though graph convolutions. While image convolutions deal with pixels, graph convolutions deal with nodes and edges. Images can be considered to be a special kind of graph where each "node" (pixel) has the same local topology and number of edges (adjacent pixels). Graph convolutions can then "derived" by making the concept of convolutions more abstract. Graph convolutions are particularly relevant for materials science because nodes and edges can be considered abstract versions of atoms and bonds. Just as an image convolution combines information about a particular pixel with that of its neighbors, a graph convolution combines data about a node with that of adjacent nodes and their corresponding edges.[31] For molecular graph convolutions, atoms can described by values such as atomic number and valency while bonds can be described by values such as bond orders and aromaticity. At the first convolution layer, each filter function combine each atom's values with those of its neighbors and corresponding edges in the input molecule to generate a corresponding "output molecule." Similar to image convolutions, a set of filters then produces a stack of output molecules. Iterating this process several times then produces a graph convolutional neural network. Graph convolutions are a recent development but they have already demonstrated considerable success in predicting molecular properties including hydrophobicity and toxicity.[31] Theoretical understanding of graph convolutions and their properties is also under active development,[41] but more work remains to be done. For example, molecular graphs are generally irregular[§] so there is no obvious way to apply pooling operations.

5.2.4.6 DECISION TREES, RANDOM FORESTS, AND PARTITIONS

Decision trees perform regression using a series of if-else decisions, and hence strongly resemble binary tree data structures. Training a decision tree starts by placing its training data into the "root" node. The first if-else decision consists of selecting a *pivot*, p_f, a value corresponding to feature f such that all $(\mathbf{x}_i)_f \leq p$ points are placed in the "left child" node, and the remaining points with $(\mathbf{x}_i)_f > p$ are placed in the "right child" node. The feature f and pivot p_f are generally chosen to minimize the total variance of the child nodes. This process is then recursively applied to each of the children as much as possible until a limiting criterion is met. Typical limitations are a maximum tree depth, i.e. number of decisions, or number of data points per node. Nodes with no further decisions are *leaves*. Once trained, a prediction $\hat{y}_t(\mathbf{x})$ is made by evaluating all decisions on $\mathbf{x}$ to reach its leaf node, then returning the mean of all training values y_i in that leaf.

A sketch of a decision tree classifier is shown in Figure 5.7 to clarify our discussion. Starting with 100 points, the tree makes decisions to minimize the predicted error. The first decision is made using feature 3 with a pivot of -1, yielding mostly red points in the left branch. This node is relatively pure and no further partitions are made. The left branch, which now contains a nearly even mix of red and blue points, is then partitioned on feature 1 with a pivot of 2, resulting in a majority-red and majority-blue leaf. Both of these decisions reduce the overall classification error; the second decision in particular improves accuracy for the 89 points with $x_3 \geq -1$ from approximately 50% to 75%. Regression trees behave similarly, with each decision reducing the overall RMSE among the resulting children. Figure 5.7 also illustrates how all tree decisions are

---

[§] Atoms do not necessarily have the same number or type of bonds.

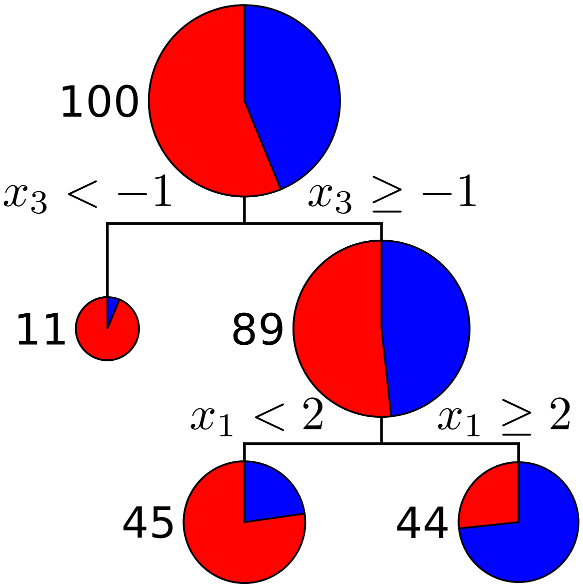


Figure 5.7: Sketch of a decision tree classifier trained on 100 data points belonging to either the "red" class or the "blue" class. Circle sizes indicate the number of points in each node, and the pie chart indicates how many points belong to each class. The first decision partitions data points using feature 3 with -1 as the pivot. A subsequent decision among points with $x_3 \geq -1$ is then made using feature 1 with 2 as the pivot.

mutually exclusive and made with respect to a single feature. The regions of feature space corresponding to each leaf node, or *leaf regions*, are therefore non-overlapping, axis-aligned and contiguous. The latter two properties also mean that all leaf regions are convex.

With these insights we can now derive the kernel corresponding to a decision tree. This derivation is inspired by previous work on forest and random partition kernels[28,152] but here we derive the tree kernel from first principles and obtain a more correct result. Let t be a decision tree trained on n data points and $\ell_t(\mathbf{x})$ be the leaf region containing $\mathbf{x}$. Each leaf regions ℓ_t contains $1 \leq |\ell_t(\mathbf{x})| \leq n$ training points. Finally, with $\delta_{\ell_t(\mathbf{x})\ell_t(\mathbf{x}')} = 1$ indicating when $\mathbf{x}$ and $\mathbf{x}'$ belong to the same leaf region, and 0 otherwise, we derive the decision tree kernel k_t:

$$\begin{aligned}
\hat{y}_t(\mathbf{x}) &= \langle y_i \rangle_{i \in \ell_t(\mathbf{x})} && (5.20a)\\
&= \frac{1}{|\ell_t(\mathbf{x})|} \sum_{i \in \ell_t(\mathbf{x})} y_i && (5.20b)\\
&= \frac{1}{|\ell_t(\mathbf{x})|} \sum_i y_i \delta_{\ell_t(\mathbf{x}_i)\ell_t(\mathbf{x})} && (5.20c)\\
&= \sum_i y_i \left(\frac{\delta_{\ell_t(\mathbf{x}_i)\ell_t(\mathbf{x})}}{|\ell_t(\mathbf{x})|} \right) && (5.20d)
\end{aligned}$$

Direct comparsion of Equations 5.20d and 5.5a yields the tree kernel and its coefficients:

$$\begin{aligned}
\alpha_i &= y_i && (5.20e)\\
k_t(\mathbf{x}, \mathbf{x}') &= \frac{\delta_{\ell_t(\mathbf{x})\ell_t(\mathbf{x}')}}{|\ell_t(\mathbf{x}')|} && (5.20f)
\end{aligned}$$

Before discussing k_t's properties we show that it satisfies the Mercer criterion. To demonstrate symmetry, there are two cases:

1. $\delta_{\ell_t(\mathbf{x})\ell_t(\mathbf{x}')} = 1$ so $\mathbf{x}$ and $\mathbf{x}'$ are contained in the same leaf region. In this case $|\ell_t(\mathbf{x})| = |\ell_t(\mathbf{x}')|$. This proves $k_t(\mathbf{x}, \mathbf{x}') = k_t(\mathbf{x}', \mathbf{x})$ as follows:

$$k_t(\mathbf{x}, \mathbf{x}') = \frac{\delta_{\ell_t(\mathbf{x})\ell_t(\mathbf{x}')}}{|\ell_t(\mathbf{x}')|} = \frac{\delta_{\ell_t(\mathbf{x}')\ell_t(\mathbf{x}')}}{|\ell_t(\mathbf{x})|} = k_t(\mathbf{x}', \mathbf{x}) \tag{5.21}$$

2. $\delta_{\ell_t(\mathbf{x})\ell_t(\mathbf{x}')} = 0$, i.e. $\mathbf{x}$ and $\mathbf{x}'$ are not in the same leaf. In this case the numerator is zero so symmetry is even easier to prove:

$$k_t(\mathbf{x}, \mathbf{x}') = \frac{0}{|\ell_t(\mathbf{x})|} = 0 = \frac{0}{|\ell_t(\mathbf{x}')|} = k_t(\mathbf{x}', \mathbf{x}) \tag{5.22}$$

Proving that k_t is positive semidefinite is more complicated. As usual, consider an arbitrary collection of m test points in feature space, with $K_{ij} = k_t(\mathbf{x}_i, \mathbf{x}_j)$. In general these m test points are independent of the tree's training points. Leaf regions $\ell_t(\mathbf{x})$ do not overlap so without loss of generality we can reorder the indices i so they are grouped by their leaf regions. We now label these regions as ℓ_r, where $r = 1..R$ enumerates the R regions containing at least one test point. In other words, the first $|\ell_1|$ indices $i = 1..|\ell_1|$ lie in leaf region ℓ_1, the next $|\ell_2|$ indices $i = |\ell_1| + 1..|\ell_1| + |\ell_2|$ lie in region ℓ_2, and so on. From the symmetry proof above, $K_{ij} = 0$ when $\mathbf{x}_i$ and $\mathbf{x}_j$ lie in different regions, so this ordering makes K block diagonal. We define these sub-blocks K_r for convenience. Moreover, for points i and j in the same leaf region ℓ_t^r, we have $K_{ij} = \frac{1}{|\ell_t^r|}$ for all i, j in the region. This makes all entries in each sub-block K_r identical. We can then re-write these constant blocks using the spectral theorem to obtain their eigenvalues:

$$K_r = \frac{1}{|\ell_r|} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1, \cdots, 1) + \sum_{j=2}^{|\ell_r|} 0 \cdot \mathbf{u}_j \mathbf{u}_j^T \tag{5.23}$$

where the $\mathbf{u}_j$ are vectors orthogonal to the first eigenvector of all ones. By inspection of Equation (5.23), each K_r block has one positive eigenvalue $(\lambda_1 = 1)$[¶] and the remaining $|\ell_r| - 1$ are zero. Each K_r block is therefore positive definite. Returning to K itself, the eigenvalues λ of a block diagonal matrix are the union of its sub-blocks' eigenvalues, i.e. $\{\lambda\}_K = \cup_r \{\lambda\}_{K_r}$. All sub-blocks K_r have eigenvalues that are positive or zero, so their union of all eigenvalues is also either positive or zero. K itself is therefore positive definite.

Now that we have proven k_t it is a valid kernel, we discuss some of its interesting properties. First, from the proofs above we see that $k_t(\mathbf{x}, \mathbf{x}')$ is piecewise constant: $\frac{1}{|\ell_t(\mathbf{x})|} = \frac{1}{|\ell_t(\mathbf{x}')|}$ if both points are in the same region, and zero otherwise. This reflects that tree values at $\mathbf{x}$ are $\mathbf{x}'$ are correlated via the mean of their leaf region, so all points in the same leaf region are identically correlated. Second, the correlation among points decreases with the number points in each leaf. If a leaf region has a single data point then $|\ell_t(\mathbf{x}')| = 1$ and $k_t = 1$. This is because the "mean" of a single y_i is identical to y_i itself, leading to perfect correlation. As the number of points in the leaf rises, though, correlation decreases as $\frac{1}{|\ell_t(\mathbf{x}')|}$ since each each y_i value only contributes that same fraction $\frac{1}{|\ell_t(\mathbf{x}')|}$ of the mean. It should be noted that the kernel k_t derived above is similar to that derived by Davies et al.,[28] but Equation (5.20f) contains the normalization factor of $\frac{1}{|\ell_t(\mathbf{x}')|}$ while that of Davies et al. does not. For "full" trees, i.e. those with a single data point per leaf, the normalization factor is 1 so our kernels are identical. However, in practice each leaf may have a different number of points. The kernel k_t derived above in Equation (5.20f) correctly accounts for the corresponding $\frac{1}{|\ell_t(\mathbf{x}')|}$ correlation.

[¶] The factor of $\frac{1}{|\ell_r|}$ in Equation (5.23) is the vector's normalization.

Additional properties can be gleaned from the tree construction process itself. First, training is fast compared to other nonlinear methods since predictions are simple means and the "cost functions" are just variances. Second, after training trees only need the sequence of decisions and the mean value y_i in each leaf region. Trees can then be stored in well-established binary tree structures for memory efficiency. Third, they are invariant to *monotone feature transformations*. This is because tree decisions simply partition on some value p_f, so for any transformation $f \rightarrow f'$ there exists a pivot value $p_f \rightarrow p'_{f'}$ that yields the same partitioning. For example, consider $x \rightarrow x^2$ for $x > 0$. If $p_x = 5$ best partitions the original data, then x^2 is identically partitioned at $p_{x^2} = 25$. Most other regression models are sensitive to monotone transformations. Monotone invariance allows forests to perform well without having to extensively test different data scaling and preconditioning strategies.

Despite these advantages, single decision trees are rarely used in practice. First, they have a strong tendency to *overfit* data, predicting $\hat{y}(\mathbf{x})$ with significant noise. This is because predictions are the average of only the small number of y_i values present in each leaf. Any noise in the sample values $\hat{y}$ is thus echoed back in the output–deep decision trees don't approximate the underlying function $y(\mathbf{x})$ so much as mimic its sample values $\hat{y}$. Second, individual decision trees have hard rectangular boundaries. For small perturbations $\boldsymbol{\epsilon}$, $\mathbf{x}$ and $\mathbf{x}+\boldsymbol{\epsilon}$ can straddle a decision boundary, leading to completely different decision paths and values.

Empirically it has been demonstrated that using an *ensemble* of randomized decision trees, or a *decision forest*,[13] strongly mitigates these drawbacks. Most decision forest regressors $\hat{y}_f$ return the mean of their trees' predictions $\hat{y}_t$. Randomized trees each have different leaf boundaries and decision pathways, and thus different predictions $\hat{y}_t(\mathbf{x})$. Averaging over a large number T of randomized trees mitigates noise and discontinuities in $\Delta\hat{y}_t$ at leaf boundaries because each tree only contributes a fraction $\frac{1}{T}$ to the forest's mean.

The challenge in such decision forests is to randomize each tree enough to change its boundaries and predictions, but not so much that it becomes useless for predictions. In other words, trees should be decorrelated from each other, but maintaining their correlation with the true function $y(\mathbf{x})$. Many randomization approaches exist, but the most popular by far are heavily based on the Breiman's random forest algorithm:[13]

1. *Bagging*: construct trees on bootstrapped subsets B_t drawn with replacement from the training data
2. *Random Inputs*: each decision considers a random subset of features F
3. *Optimal Pivots*: select the best pivot p_f among $f \in F$
4. *Prediction*: return the mean of all y_i where $i \in \ell_t(\mathbf{x})$ and $i \in B_t$

Various "global" choices must then be made regarding the number of features F to consider for each decision and maximum tree depths. Fortunately, the following choices are generally close to optimal for regression: consider $F \lesssim m$ of the m total features for each decision, and use unlimited depth.[13] Bagging gives each tree a different subset of the training data, leading to different pivot values and decision pathways. Similarly, using F random inputs adds additional variation to the decision pathways.

Subsequent work has suggested several improvements to the original random forest algorithm. The most popular of these is the *extra(-random) forest algorithm*,[40] which modifies step 3 by comparing a single *random* pivot value p_f for each of the F random inputs. Extra forests generally provide a small improvement in accuracy over random forests and are significantly faster to train because only a single pivot per feature is considered for each decision. Extra forests are used extensively in the work on predicting H adsorption energies for dilute Ag alloys presented in Section 5.4. Another possible improvement, *rotation forests*[148] further randomizes decision boundaries by applying a random unitary rotation U_t to the feature values of each tree's training data. Decisions are then made on the rotated axes described by U_t. Rotation forests are promising but most common machine learning libraries have not implemented it yet.

Now we can consider the *forest* kernel. A forest regressor returns the average of its trees, so the forest regressor $\hat{y}_f$ can be written in terms of its tree regressors:

$$\hat{y}_f(\mathbf{x}) = \frac{1}{T}\sum_t \hat{y}_t(\mathbf{x}) \tag{5.24a}$$

$$= \frac{1}{T}\sum_t \sum_{i \in B_t} y_i k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \tag{5.24b}$$

If we define δ_{iB_t} to be 1 if i is in B_t and 0 otherwise, we can further simplify this expression:

$$= \frac{1}{T}\sum_t \sum_i y_i \delta_{iB_t} k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \tag{5.24c}$$

$$= \frac{1}{T}\sum_i \sum_t y_i \delta_{iB_t} k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \tag{5.24d}$$

Interestingly this last expression *cannot* be written in terms of Equation (5.5a). This is because each tree's coefficients are different, with δ_{iB_t} changing from tree to tree. We can demonstrate this in general for two different models, with kernels $k_1(\mathbf{x},\mathbf{x}') \neq k_2(\mathbf{x},\mathbf{x}')$:

$$\begin{aligned}
\hat{y}_1(\mathbf{x}) + \hat{y}_2(\mathbf{x}) &= \sum_i \alpha_{i1}\, k_1(\mathbf{x}_i,\mathbf{x}) + \sum_i \alpha_{i2}\, k_2(\mathbf{x}_i,\mathbf{x}) && \text{(5.25a)}\\
&= \sum_i \alpha_{i1}\, k_1(\mathbf{x}_i,\mathbf{x}) + \sum_i \alpha_{i2}\, k_2(\mathbf{x}_i,\mathbf{x}) + \sum_i \alpha_{i1}\, k_2(\mathbf{x}_i,\mathbf{x}) - \sum_i \alpha_{i1}\, k_2(\mathbf{x}_i,\mathbf{x}) && \text{(5.25b)}\\
&= \sum_i \alpha_{i1}\, (k_1(\mathbf{x}_i) + k_2(\mathbf{x}_i)) + \sum_i (\alpha_{i2} - \alpha_{i1})\, k_2(\mathbf{x}_i,\mathbf{x}) && \text{(5.25c)}
\end{aligned}$$

Equation (5.25c) shows that the model sum corresponds to the sum of their kernels *only if the coefficients are the same*. Eliminating the explicit appearance of δ_{iB_t} would thus give us a forest kernel that is the sum of its trees' kernels. Returning to Equation (5.24b), we remove this bagging restriction to see what it implies:

$$\begin{aligned}
\frac{1}{T}\sum_t \sum_{i\in B_t} y_i\, k_t(\mathbf{x}_i,\mathbf{x}|\beta_t) \to \frac{1}{T}\sum_t \sum_i y_i\, k_t(\mathbf{x}_i,\mathbf{x}|\beta_t) && \text{(5.26a)}\\
= \frac{1}{T}\sum_t \sum_i y_i \frac{\delta_{\ell_t(\mathbf{x}_i|\beta_t)\ell_t(\mathbf{x}|\beta_t)}}{|\ell_t(\mathbf{x}|\beta_t)|} && \text{(5.26b)}
\end{aligned}$$

This last expression is the random forest kernel if trees predict the mean of *all* training point samples in a given leaf region, not just those in B_t. Bagging can still be used randomize partitions when generating trees (as part of the randomized β_t), but no samples should be left out when making predictions. In recognition of this change, and inspired by Davies et al.,[28] I call modified algorithm the *random partition forest*. We can now obtain the kernel k_f for the random partition forest:

$$\hat{y}_f(\mathbf{x}) = \cdots = \frac{1}{T}\sum_t \sum_i y_i k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \tag{5.27a}$$

$$= \sum_i y_i \left(\frac{1}{T} \sum_t k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \right) \tag{5.27b}$$

As with the regression tree, we finally compare to Equation (5.5a):

$$\alpha_i = y_i \tag{5.27c}$$

$$k_f(\mathbf{x}, \mathbf{x}'|\beta_f) = \frac{1}{T}\sum_t k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) = \langle k_t(\mathbf{x}_i, \mathbf{x}|\beta_t)\rangle_{t\in f} \tag{5.27d}$$

Again, it must be pointed out that k_f is the kernel corresponding to the random partition forest, not to the more common random forest introduced by Breiman. Nevertheless, tree construction is identical for both methods so the random partition forest kernel can be extracted from trees made for random and extra forests. Moreover, since bagging primarily serves to randomize partition boundaries rather than randomize predictions themselves, Equation (5.27d) should be a reasonable approximation to the kernel from Breiman's random forests. It should also be noted here that the culmination of bagging, random inputs, and random pivots can be considered to be parameters β for the forest kernel, drawn randomly from some distribution. From this perspective we can interpret extra forests as Monte Carlo samples drawn from the distributions of B_t, random inputs, and random partitions.

Although some properties of random [partition] forests have been discussed above, knowing the kernel gives us more insight. Tree kernels are rectangular so forest kernels are the mean of many such rectangles. A sketch of a forest kernel is shown in Figure 5.8 for different numbers of trees. By averaging over several trees, the forest kernel has softer boundaries. Each tree's kernel is chosen to be optimal for the target function, so similar to neural networks, k_f yields a data-driven

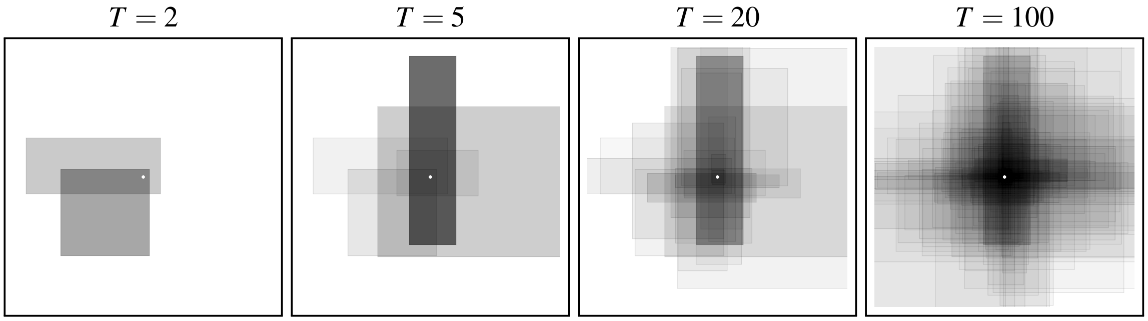


Figure 5.8: Sketch of a forest kernel $k_f(\mathbf{x}, \mathbf{0})$ in two dimensions using randomly-generated rectangular regions for varying numbers of trees T. The white circle indicates the origin, and k_f is plotted over the same region for varying values of $\mathbf{x} \in \mathbb{R}^2$. The forest kernel decreases monotonically with distance $|\mathbf{x}|$ from $\mathbf{x}' = 0$ along any fixed direction.

kernel. For a given point $\mathbf{x}'$, $k_f(\mathbf{x}, \mathbf{x}')$ is maximized at $k_f(\mathbf{x}', \mathbf{x}')$ where its trees' kernels are all nonzero ($\mathbf{x}'$ by definition is in its own leaf region). We can also see that that $k_f(\mathbf{x}, \mathbf{x}')$ decreases monotonically as $\mathbf{x}$ moves away from a fixed $\mathbf{x}'$ along some direction $\hat{\mathbf{u}}$. Letting t be the distance, then $\mathbf{x} = \mathbf{x}' + t\,\hat{\mathbf{u}}$. As t increases, each tree kernel in the forest is only nonzero in the convex region $\ell_t(\mathbf{x}')$, so sufficient displacement t may move $\mathbf{x}$ outside $\ell_t(\mathbf{x}')$ where the the tree's kernel is 0. Tree kernels are therefore monotonically decreasing functions along any direction $\hat{\mathbf{u}}$. Since the forest kernel is the mean of its trees' kernels, k_f must monotonically decreases along $\hat{\mathbf{u}}$ too. This proves that $k_f(\mathbf{x} + t_2\hat{\mathbf{u}}, \mathbf{x}') \leq k_f(\mathbf{x} + t_1\hat{\mathbf{u}}, \mathbf{x}')$ for $t_2 \geq t_1 \geq 0$, which I call *directionally monotonic* for lack of a better phrase. Forest kernels therefore exhibit a limited degree of localization, somewhat similar to the RBF kernel discussed previously in Section 5.2.4.3.

As proposed by Davies et al., forests can be generalized to obtain an entire class of partition-based models. We can obtain the generalization by expanding the forest kernel:

$$k_f(\mathbf{x}_i, \mathbf{x}|\beta_t) = \langle k_t(\mathbf{x}_i, \mathbf{x}|\beta_t) \rangle_{t \in f} \rightrightarrows \left\langle \frac{\delta_{\ell_p(\mathbf{x}|\beta_p)\ell_p(\mathbf{x}'|\beta_p)}}{|\ell_p(\mathbf{x})|} \right\rangle_{p \in e} \tag{5.28}$$

The right-hand side defines the *random partition kernel*, an ensemble e of *partitioners* p. Each partitioner subdivides feature space into one or more non-overlapping regions $\ell_p(\mathbf{x})$. Davies et al.

obtained the same expression, but without the $\frac{1}{|\ell_p(\mathbf{x})|}$ scaling discussed previously. Forests are a special case of random partition kernels: they are ensembles of partitioners that subdivide feature space into contiguous rectangular regions. However, there are no requirements about the shape of these regions in general so regions could be non-rectangular. For example, partitioners based on k-means clustering would not have rectangular boundaries. If each partitioner's leaf regions are convex, then the resulting ensemble would exhibit the same "directional monotonicity" that forests do. On the other hand, Equation (5.28) also permits non-convex and even disjoint regions as well. Partitioning models could be designed to handle periodicity and other nonlocal correlations. These partitions might be obtained using discrete Fourier transforms or wavelet analysis. Sophisticated partitioning schemes are beyond the scope of this thesis but could be promising directions for future work.

Finally, as ensemble methods, random forests can provide confidence intervals based on each tree's prediction. For any point $\mathbf{x}$, the random forest prediction $\hat{y}_f(\mathbf{x})$ is the mean of its trees' predictions $\{\hat{y}_t(\mathbf{x})\}$. Most implementations of random forests, such as the one in `scikit-learn` used in Section 5.4, report only the mean. However, the *distribution* of tree predictions can give insight into the forest's confidence in a prediction. If most trees predict similar values for $\hat{y}_t(\mathbf{x})$, then the forest has high confidence and can be interpreted as a narrow confidence interval. In contrast, widely varying predictions indicate poor confidence and a wide confidence interval. The simplest metric for the confidence interval is the standard deviation $\sigma_f(\mathbf{x})$ of forest predictions. Empirically, in the Ag alloys work in Section 5.4, these standard deviations are indeed correlated with the errors $\varepsilon_f(\mathbf{x})$. Defining the "forest z-score" as $z_f(\mathbf{x}) = \frac{\varepsilon_f(\mathbf{x})}{\sigma_f(\mathbf{x})}$, we found that z_f is approximately normally distributed such that z_f can be interpreted in the same way as z-scores in other contexts. Having quantifiable confidence intervals is a major benefit since it is always important to know the limitations and reliability of predictions, especially when accuracy can vary among different groups of points in dataset.

5.2.5 PRINCIPAL COMPONENT ANALYSIS AND CLUSTERING

Obtaining decorrelated features and sparsity is an important task when applying most regression models. In most practical data sets the feature values $\mathbf{x}$ themselves are not uniform and decorrelated. For example, consider a housing data set with variables including GPS location, population density, and distance to public transit. Population density and distance to public transit are correlated since public transit is generally not built in rural areas, and distance is a function of GPS coordinates so they are correlated as well. Models trained on all variables can then struggle to identify the most important features and their independent effects. Moreover, high dimensonal models with many features require many parameters, requiring more data to precisely determine the values of all coefficients.

These issues are mitigated if we map $\mathbf{x}_i$ to a new feature space, $\mathbf{z}_i$, where the new features $z_1, \cdots, z_m$ are *decorrelated*. By ranking the variance σ_i^2 along each of these directions from largest to smallest, we can also achieve sparsity by only keeping the components of $\mathbf{z}_i$ with high variance. These components describe most of the variation in the data set. A sketch of this process is shown in Figure 5.9. The data points are clearly distributed along a curve. If we could obtain the projection of each data point along the curve, then most of the two dimensional sample points are reasonably approximated using only the single coordinate z_1 along the curve. Similarly, the outliers indicated by filled circles can be detected by finding points with unusually large values along the low-variance dimension z_2. These kinds of projection become increasingly beneficial when we must consider dozens or even hundreds of features.

Both goals are achieved by *principal component analysis* (PCA). Relationships between correlated variables are nonlinear in general, such as those sketched in Figure 5.9. Following the kernel discussions above, however, we know that nonlinear correlations in feature space $\mathbf{x}_i$ will

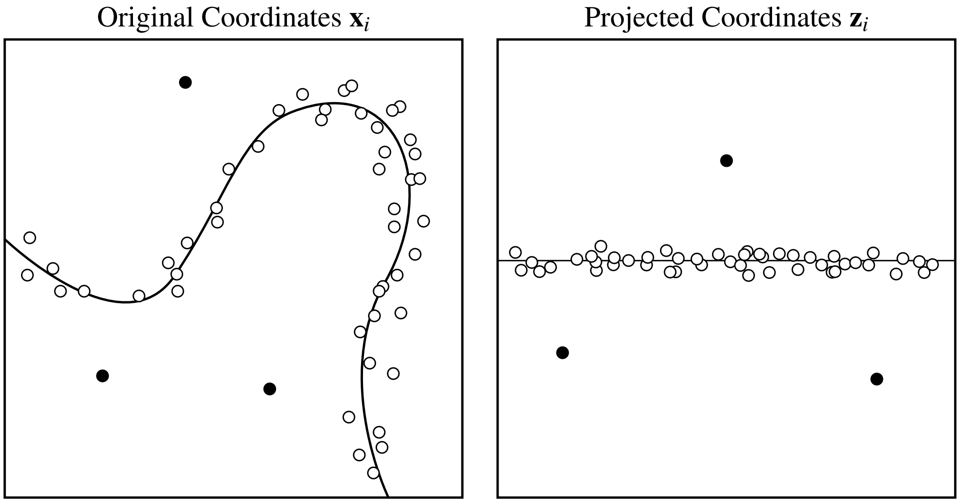

Original Coordinates $\mathbf{x}_i$
Projected Coordinates $\mathbf{z}_i$

Figure 5.9: A sketch demonstrating the utility of projecting features onto new axes. z_1 is distance along the curve, and z_2 is distance perpendicular to the curve. Outliers are indicated by filled circles.

become linear correlations in a well-chosen kernel space $\phi(\mathbf{x}_i)$. For simplicity, let $\phi_i \equiv \phi(\mathbf{x}_i)$. If we know the ϕ_i explicitly for all n sample points, e.g. linear and neural network regression, then we can obtain the correlation matrix directly:

$$C \equiv \frac{1}{n}\sum_{i=1}^{n} \phi_i \phi_i^T \tag{5.29a}$$

$$= \sum_i \lambda_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \tag{5.29b}$$

$$= U \Lambda U^T \tag{5.29c}$$

where the eigenvalues λ_i are the *principal values*, and the eigenvectors $\hat{\mathbf{u}}_i$ are the *principal components*. The matrix $U \equiv (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \cdots)$ is the matrix of principal components. Equation (5.29b) follows from the spectral theorem, and Equation (5.29c) shows that we can interpret U as a new basis: the rotation of kernel space vectors ϕ_i that decorrelates their entries.

Principal values λ_i give variances of the data along each dimension $\mathbf{u}_i$ in kernel space. Those directions with the largest variances then explain most of the variation among the corresponding $\mathbf{x}_i$.

By convention we sort the λ_i such that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots$ and so on. The rank of C depends on ϕ and n. Suppose ϕ is a d-dimensional vector, $\phi \in \mathbb{R}^d$. If $n < d$, then $\text{rank}(C) \leq n$ since the n kernel space points ϕ_i span at most n dimensions. If $n \geq d$, then $C \in \mathbb{R}^{d \times d}$ means that $\text{rank}(C) \leq d$. In both cases $\text{rank}(C) \leq n$ so the eigenvectors $\hat{\mathbf{u}}_i$ can be written as linear combinations of the kernel space points:

$$\hat{\mathbf{u}}_i = \sum_{\mu} \alpha^i_\mu \phi_\mu \tag{5.30}$$

As implied previously, seek the projection of $\mathbf{x}_j$ onto the ith largest principal component in kernel space:

$$(\mathbf{z}_j)_i \equiv \hat{\mathbf{u}}_i^T \phi_j \tag{5.31}$$

This notation is admittedly complicated, but it keeps our discussion entirely general so that we can apply PCA using any kernel of choice. If we have ϕ_i for all points, then Equation (5.31) completes our objective.

When only $k(\mathbf{x}_i, \mathbf{x}_j) = \phi_i \cdot \phi_j$ is known, however, we need to re-write these expressions to rely exclusively on inner products. Fortunately, the projections we want in Equation (5.31) are inner products. Using Equation (5.30) we can then rewrite them in terms of K:

$$(\mathbf{z}_j)_i \equiv \hat{\mathbf{u}}_i^T \phi_j = \left(\sum_{\mu} \alpha^i_\mu \phi^T_\mu \right) \phi_j = \sum_{\mu} \alpha^i_\mu K_{\mu j} = \left(\boldsymbol{\alpha}^{i^T} K \right)_j \tag{5.32}$$

The right-hand side of Equation (5.32) shows that the $\mathbf{x}_j$'s projection onto principal component $\hat{\mathbf{u}}_i$ is the the jth entry of $\boldsymbol{\alpha}^{i^T} K$. If we define $Z \equiv (\mathbf{z}_1 \cdots \mathbf{z}_n)$, and $A \equiv \left(\boldsymbol{\alpha}^1 \cdots \boldsymbol{\alpha}^n \right)$, then this equation can be rewritten as

$$Z = KA \tag{5.33}$$

so that the ith column of Z lists the projections of each point ϕ onto the ith principal component.

This matrix operation is easy to implement once K and A are known. We can obtain an eigenvalue equation for $\boldsymbol{\alpha}^i$ using the eigenvalue equation for $\hat{\mathbf{u}}_i$:

$$C\hat{\mathbf{u}}_i = \lambda_i \hat{\mathbf{u}}_i \qquad \text{eigenvector definition} \tag{5.34a}$$

$$\left(\frac{1}{n}\sum_j \phi_j \phi_j^T\right)\left(\sum_\mu \alpha_\mu^i \phi_\mu\right) = \lambda_i \sum_\mu \alpha_\mu^i \phi_\mu \qquad \text{Equations 5.29a, 5.30} \tag{5.34b}$$

$$\frac{1}{n}\sum_j \sum_\mu \phi_i K_{j\mu} \alpha_\mu^i = \lambda_i \sum_\mu \alpha_\mu^i \phi_\mu \tag{5.34c}$$

To eliminate the remaining vectors, we apply the Fourier trick by multiplying on the left with ϕ_ν, and seeing that this equation should hold for all ν

$$\phi_\nu^T \frac{1}{n}\sum_j \sum_\mu \phi_i K_{j\mu} \alpha_\mu^i = \phi_\nu^T \lambda_i \sum_\mu \alpha_\mu^i \phi_\mu \qquad \forall \nu \tag{5.34d}$$

$$\frac{1}{n}\sum_j \sum_\mu K_{\nu i} K_{j\mu} \alpha_\mu^i = \lambda_i \sum_\mu K_{\nu\mu} \alpha_\mu^i \qquad \forall \nu \tag{5.34e}$$

Equation (5.34e) is an eigenvalue equation in K:

$$K\boldsymbol{\alpha}^i = n\lambda_i \boldsymbol{\alpha}^i \tag{5.34f}$$

Equation (5.34f) shows we can get the projections $\boldsymbol{\alpha}^i$, the corresponding matrix of vectors A, and the principal values λ_i by finding the eigenvalues of K. Applying Equation (5.33) then yields the desired vectors $Z = (\mathbf{z}_1 \cdots \mathbf{z}_n)$. Before continuing, it should be noted that the matrix expression C is only the covariance matrix when the average kernel space values are centered at zero, i.e. $\sum_i \phi_i = 0$. Otherwise, the resulting correlations are biased. Therefore K is usually shifted to yield $\tilde{K}$, corresponding to kernel space vectors centered at the origin. This shifting can be accomplished as follows:

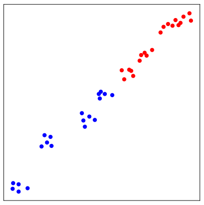
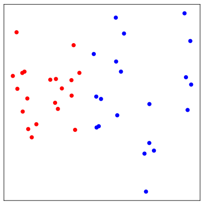
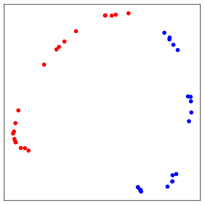
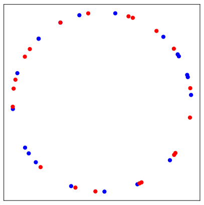
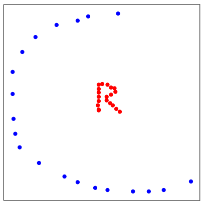
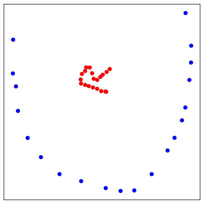
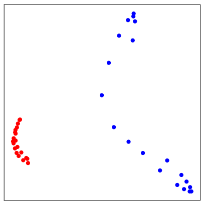
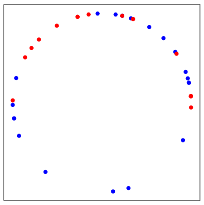
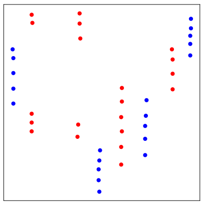
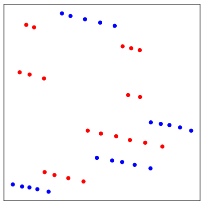
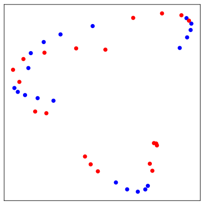
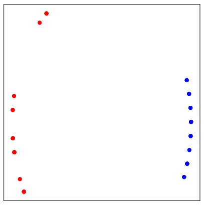
$$\tilde{K} = K - \mathbf{1}_{1/n}K - K\mathbf{1}_{1/n} + \mathbf{1}_{1/n}K\mathbf{1}_{1/n} \tag{5.35}$$

where $\mathbf{1}_{1/n}$ is the matrix where every entry $\left(\mathbf{1}_{1/n}\right)_{ij} = \frac{1}{n}$. This allows us to "center" K at the origin without directly knowing the values of ϕ_i. Equation (5.34f) is then applied to $\tilde{K}$.

Several examples of kernel PCA are shown in Table 5.2 for various combinations of data sets and kernels. As expected the linear kernel extracts linear relationships between input features. Linear PCA is the special case where $\mathbf{x}_i = \phi(\mathbf{x})$, so the eigenvectors U of the kernel matrix K can be interpreted as a rotation matrix for the original features. This explains why linear PCA rotates the input data. Interpreting the ϕ vectors for the RBF and cosine kernels is more difficult since they are infinite-dimensional: the power series expansions for $\exp(u)$ and $\cos(u)$ have an infinite number of terms. We can still interpret the results based on the behavior of k though. The RBF kernel asserts that nearby points are correlated with each other, so groups of points with small distances are mutually correlated. These points will then tend to grouped together with RBF PCA. The blue and red points in the "half moon" data set have this relationship, and RBF PCA clusters them together as expected. Finally, the cosine kernel asserts that points at periodically correlated along the horizontal direction. Points spaced at regular intervals are thus considered mutually correlated, and are grouped together by cosine PCA. This explains why cosine PCA performs poorly for data sets lacking periodic order, and performs well for the periodic "fringes" data set.

We can also see the performance of different kernels in reducing dimensions on real world data: the "red wine" data set.[29] These samples consist of 11 features describing the chemical properties of Portuguese red wines and their qualities on a 1-10 scale as determined by a panel of experts. The goal is to predict the quality of new wines using only the readily-available 11 chemical measurements. Figure 5.10 shows the results of the various approaches PCA and projection-based methods of obtaining a two dimensional representation of the original 11-dimensional data. The

Table 5.2: Kernel PCA Applied to Different Data Sets

Name	Feature Space	Linear Kernel $\mathbf{x}\cdot\mathbf{x}'$	RBF Kernel $\exp(-\lvert\mathbf{x}-\mathbf{x}'\rvert^2)$	Cosine Kernel $\cos(\lvert x_1 - x_1'\rvert)$
line				
half moon				
fringes				

most straightforward approach is to keep only the two most important features for linear regression. This is done using linear LASSO regression as described in section 5.3.2 and yields the “LASSO” projection in Figure 5.10. The other three panels show the results of PCA using the linear, RBF, and forest kernels. Generally speaking the LASSO and forest PCA projections yield the best distinction between wines of each quality; both methods show a clear separation from upper left to lower right. Linear and RBF PCA also perform reasonably, but moderately good and poor wines are mixed on the interior of the region.

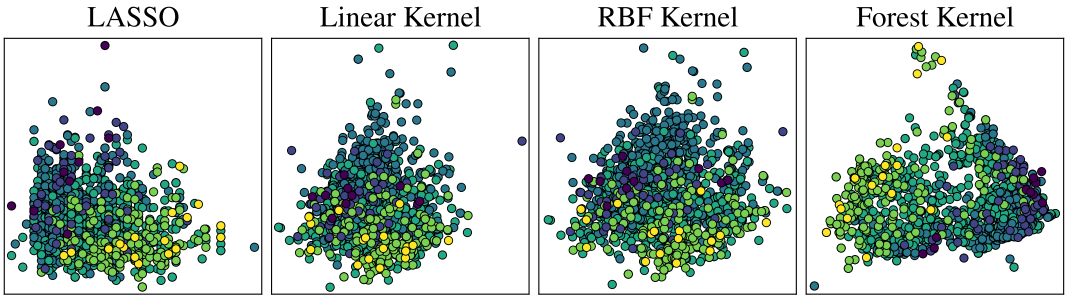


Figure 5.10: Several approaches for achieving two dimensional representations of the red wine data. Points are colored by quality; purple is worst and yellow is best. LASSO projects all points onto the two most important features as determined by linear LASSO regression (see Section 5.3.2). The other three are applications of kernel PCA. Forest and RBF kernels were optimized by maximizing their respective regression performances on the data set.

5.3 MODEL TRAINING

We now turn to the task of actually obtaining the coefficients α_i, and possibly β parameters as well, for a given regression model. There are two equivalent perspectives to model training:

1. Costs: optimize the coefficients α_i and parameter(s) β with respect to a cost function C

2. Likelihood: maximize the likelihood of observing training points y_i given the model

Both approaches are summarized below.

5.3.1 COST FUNCTIONS: LASSO AND RIDGE

Explicit cost functions pose finding the α_i in terms of an optimization problem. The most typical cost function is the [root mean] squared error:

$$C_{\mathrm{SE}} \equiv \sum_i \left(y_i - \hat{y}(\mathbf{x}_i)\right)^2 \tag{5.36}$$

Using squared error prioritizes models without large errors, but quadratic scaling leaves models susceptible to *outliers*, points with unusual y_i values resulting from noise or errors in data collection. Other cost functions, such as mean absolute error, can be used as well. Squared error is immensely popular because it leads to linear matrix equations involving the kernel matrix K and the vector of coefficients $\boldsymbol{\alpha}$. Other cost functions can provide greater resilience to noise but generally make optimization more computationally expensive.

For some models, especially those with many parameters and noisy data sets, minimizing squared error alone leads to the α_i coefficients becoming very large to replicate noise and errors in the data set instead of the underlying function. Augmenting the cost function a parameter-based penalty mitigates this behavior. Such augmentations are said to *regularize* the model. The most common regularization approaches are LASSO and ridge:

$$C_{\text{LASSO}} \equiv C_{\text{SE}} + \lambda_1 \left|\boldsymbol{\alpha}\right|_1 \tag{5.37a}$$

$$C_{\text{ridge}} \equiv C_{\text{SE}} + \lambda_2 \left|\boldsymbol{\alpha}\right|_2^2 \tag{5.37b}$$

where all λ parameters are positive. Both penalties are widely used.

5.3.2 LASSO VERSUS RIDGE REGRESSION

These cost functions have qualitatively different behavior. LASSO regularization is the L1 (or "Manhattan") norm. Empirically, increasing $\lambda_1 \to \infty$ almost always leads to a growing number of α_i coefficients becoming identically 0. LASSO is therefore useful for *dimensionality reduction* since the features corresponding to $\alpha_i = 0$ coefficients can be ignored. For example, the "LASSO projection" in Figure 5.10 as obtained by adjusting λ_1 until only two of the eleven features had nonzero coefficients in the linear model. LASSO penalties therefore yield sparse models with few terms, leading to its widespread use in compression. Unfortunately, the L1 norm generally leads to non-analytic solutions for $\boldsymbol{\alpha}$ as a function of λ_1, so C_{LASSO} must be iteratively minimized with numerical optimization methods.

On the other hand, the ridge penalty is the L2 norm which generally provides analytic matrix equations for $\boldsymbol{\alpha}$ for a given value of λ_2. Moreover, the L2 norm is invariant under unitary rotations. Unfortunately, L2 penalties do *not* tend to yield sparsity. As $\lambda_2 \to \infty$, $|\alpha_i|$ values do tend to decrease but almost never actually vanish. Sparsity can still be achieved with ridge regression, but often times redundant features will have approximately uniform coefficients. Determining which coefficients can be eliminated, e.g. with an empirical cutoff, must be done with great care. Ridge regression is therefore favored in cases where sparsity is less important than regularization.

Differences between LASSO and ridge regression can be explained by comparing frontiers of constant C_{SE} with those of constant L1 and L2 penalties. The squared error cost generally increases *toward* the origin ($\boldsymbol{\alpha} = 0$) since constraining the coefficients increases errors (the variational principle). Meanwhile, the regularization penalties increase *away* from the origin. Points where these frontiers *intersect* yield the lowest penalty for a given value of C_{SE} and thus the lowest overall cost. These intersections are sketched for both LASSO (L1) and ridge (L2) penalties in Figure 5.11. We see that LASSO tends to yield intersections at the axis-aligned "corners" of the L1 norm, corresponding to $\alpha_i = 0$. In contrast, the smooth L2 contours feature tangential intersections that generally do not correspond to $\alpha_i = 0$. This is why LASSO regression tends toward sparsity while ridge regression does not.

5.3.3 MAXIMUM LIKELIHOOD AND BAYESIAN OPTIMIZATION

The other approach to model parameters is *maximum likelihood estimation*. From this perspective, models and function values are considered to be probability distributions. We start with the error distribution, $p\left(y_i|\mathbf{x}_i, \boldsymbol{\alpha}\right)$, the probability p of observing y_i given its corresponding position $\mathbf{x}_i$ in feature space and model parameters $\boldsymbol{\alpha}$. This distribution is usually assumed to be Gaussian:

$$p\left(y_i|\mathbf{x}_i, \boldsymbol{\alpha}\right) \sim \mathcal{N}\left(\hat{y}_i\left(\mathbf{x}_i, \boldsymbol{\alpha}\right), \sigma^2\right) \tag{5.38}$$

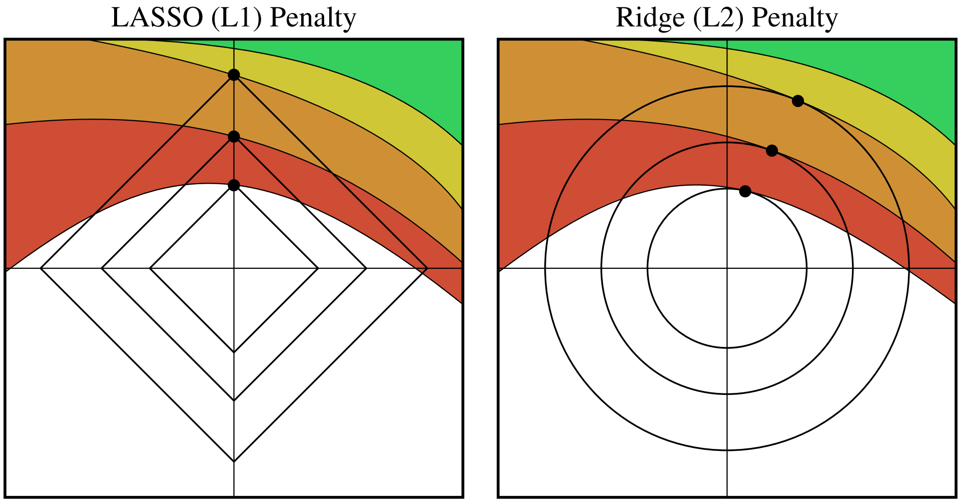


Figure 5.11: Sketches showing the intersections between contours of constant squared error cost and L1/L2 cost. Cel shading is applied according to squared error cost. Intersections between the L1/L2 contours and squared error contours are indicated by filled circles.

where $\hat{y}_i$ is a function of the features $\mathbf{x}_i$ and the model parameters $\boldsymbol{\alpha}$. Equivalently, we can define $\varepsilon_i \equiv y_i - \hat{y}_i$ such that $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

The most likely model maximizes the probability P of observing all the values y_i with respect to the parameters $\boldsymbol{\alpha}$. Assuming that errors are identically and independently distributed, we can re-write the optimization problem for as follows:

$$\max_{\boldsymbol{\alpha}} P\left(\{y_i\} \,|\, \{\mathbf{x}_i\}, \boldsymbol{\alpha}\right) = \prod_i P\left(y_i | \mathbf{x}_i, \boldsymbol{\alpha}\right) \tag{5.39a}$$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) \tag{5.39b}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_i \varepsilon_i^2\right) \tag{5.39c}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_i \left(y_i - \hat{y}_i\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)\right)^2\right) \tag{5.39d}$$

The only part of the right-hand side that depends on $\boldsymbol{\alpha}$ is the sum embedded in the exponential. The smaller this sum, the larger P becomes. Therefore we now have an equivalent *minimization* problem:

$$\max_{\boldsymbol{\alpha}} P\left(\{y_i\}\,|\,\{\mathbf{x}_i\}\,,\boldsymbol{\alpha}\right) \to \min_{\boldsymbol{\alpha}} \sum_i \left(y_i - \hat{y}_i\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)\right)^2 \equiv \min_{\boldsymbol{\alpha}} C_{\mathrm{SE}}\left(\boldsymbol{\alpha}\right) \tag{5.39e}$$

This last equation shows that maximizing the normal distribution's probability is exactly equivalent to minimizing the squared-error cost function C_{SE} (again assuming that errors are identically and independently distributed).

Based on the derivation above, we can generalize this connection for probabilities of the following form:

$$p\left(y_i|\mathbf{x}_i, \boldsymbol{\alpha}\right) = |a| \exp\left(-\frac{|\varepsilon_i|^p}{|b|}\right) \to \min_{\boldsymbol{\alpha}} \sum_i |\varepsilon_i|^p \tag{5.40}$$

Exponentials depending on some power p of the error lead to minimizing the p-norm of the errors. Equation (5.40) is the core link between minimizing the cost functions above and maximizing probability.

The approach above is immensely popular (i.e. ordinary least squares regression), but is a bit circuitous. In modeling we typically prefer $p\left(\boldsymbol{\alpha}|\,\{y_i\}\,;\{\mathbf{x}_i\}\right)$, the probability of observing the parameters $\boldsymbol{\alpha}$ given the observed values y_i and their feature vectors $\mathbf{x}_i$. Maximizing this probability gives the most likely weights, and p is a distribution yielding confidence intervals for the parameters.

Conditional probabilities can be obtained using Bayes theorem, resulting in *Bayesian optimization*. Bayesian approaches are particularly attractive because they provide easy ways to include regularization. We start with Bayes' theorem itself:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{5.41}$$

In the present context, A is the set of observed values $\mathbf{y}$, and B is a set of model parameters $\boldsymbol{\alpha}$. Rewriting Bayes' theorem with this notation yields the probability distribution for $\boldsymbol{\alpha}$:

$$P(\boldsymbol{\alpha}|\{y_i\};\{\mathbf{x}_i\}) = \frac{P(\{y_i\}|\boldsymbol{\alpha};\{\mathbf{x}_i\})P(\boldsymbol{\alpha})}{P(\{y_i\};\{\mathbf{x}_i\})} \tag{5.42}$$

The probability $P(\boldsymbol{\alpha})$ is the *Bayesian prior* on the weights. Similar to kernels, priors are arbitrary in the sense that they dictate the subsequent conditional distribution for $\boldsymbol{\alpha}$. Priors provide a statistically well-defined way to enforce constraints and biases on the parameters. For example, unusual or highly unlikely parameter combinations can be penalized by assigning them small probabilities in the prior. Bayesian approaches are therefore not first-principles methods, and statisticians have argued the relative merits and validity of Bayesian models (as opposed to frequentist models) for centuries. The main risk is that bad priors can lead to unreasonable bias into the results. For example, one could obtain any desired result $\boldsymbol{\alpha}_{\text{bias}}$ using $P(\boldsymbol{\alpha}) = \delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\text{bias}})$. On the other hand, well-chosen priors can improve resistance to outliers, leverage domain knowledge to compensate for small or incomplete data sets, and obtain uncertainty estimates via Equation (5.42).

The most popular Bayesian priors are selected to yield regularization. As discussed previously, keeping parameters small prevents models from fitting significant amounts of noise or being strongly biased by outliers. In terms of $P(\boldsymbol{\alpha})$ we therefore assert that coefficients are more likely to be small by using a prior that is maximized when $\boldsymbol{\alpha} = \mathbf{0}$. For example, we can set $P(\boldsymbol{\alpha})$ to be a normal distribution assuming that the parameters are independent and identically distributed with standard deviation σ_α:

$$P(\boldsymbol{\alpha}) = \prod_i \frac{1}{\sqrt{2\pi\alpha_w^2}} \exp\left(-\frac{\alpha_i^2}{2\sigma_\alpha^2}\right) \propto \exp\left(-\frac{|\boldsymbol{\alpha}|^2}{2\sigma_\alpha^2}\right) \tag{5.43}$$

To see how this prior affects the most likely value of $\boldsymbol{\alpha}$, we maximize the conditional probability $P(\boldsymbol{\alpha}|\{y_i\};\{\mathbf{x}_i\})$ given the training data:

$$\max_{\boldsymbol{\alpha}} P(\boldsymbol{\alpha}|\{y_i\};\{\mathbf{x}_i\}) = \max_{\boldsymbol{\alpha}} \frac{P(\{y_i\}|\boldsymbol{\alpha};\{\mathbf{x}_i\})\,P(\boldsymbol{\alpha})}{P(\{y_i\};\{\mathbf{x}_i\})} \tag{5.44a}$$

The denominator of the right-hand side is sometimes known as the *marginal distribution* of the training data itself and is constant with respect to $\boldsymbol{\alpha}$. Furthermore, we can replace $P(\{y_i\}|\boldsymbol{\alpha};\{\mathbf{x}_i\})$ with the expression we already found in Equation (5.39c):

$$= \max_{\boldsymbol{\alpha}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_i \varepsilon_i^2}{2\sigma^2}\right) P(\boldsymbol{\alpha}) \tag{5.44b}$$

$$= \max_{\boldsymbol{\alpha}} \exp\left(-\frac{\sum_i \varepsilon_i^2}{2\sigma^2}\right) P(\boldsymbol{\alpha}) \tag{5.44c}$$

We now replace $P(\boldsymbol{\alpha})$ with the Gaussian prior from Equation (5.43):

$$= \max_{\boldsymbol{\alpha}} \exp\left(-\frac{\sum_i \varepsilon_i^2}{2\sigma^2}\right) \exp\left(-\frac{|\boldsymbol{\alpha}|^2}{2\sigma_\alpha^2}\right) \tag{5.44d}$$

$$= \max_{\boldsymbol{\alpha}} \exp\left(-\left(\frac{\sum_i \varepsilon_i^2}{2\sigma^2} + \frac{|\boldsymbol{\alpha}|^2}{2\sigma_\alpha^2}\right)\right) \tag{5.44e}$$

This expression is maximized when the argument is minimized:

$$\rightarrow \min_{\boldsymbol{\alpha}} \frac{\sum_i \varepsilon_i^2}{2\sigma^2} + \frac{|\boldsymbol{\alpha}|^2}{2\sigma_\alpha^2} \tag{5.44f}$$

$$= \min_{\boldsymbol{\alpha}} \sum_i \varepsilon_i^2 + \frac{\sigma^2}{\sigma_\alpha^2}|\boldsymbol{\alpha}|^2 \tag{5.44g}$$

$$\max_{\boldsymbol{\alpha}} P\left(\boldsymbol{\alpha}|\left\{y_i\right\};\left\{\mathbf{x}_i\right\}\right) \equiv C_{\mathrm{SE}}\left(\boldsymbol{\alpha}\right) + \lambda_2|\boldsymbol{\alpha}|^2 \equiv C_{\mathrm{ridge}}\left(\boldsymbol{\alpha}\right) \tag{5.44h}$$

A Gaussian prior on the weights $\boldsymbol{\alpha}$ therefore corresponds to Ridge regression, the L2 penalty. Analogous to Equation (5.40), there is a general connection between exponential priors and cost functions:

$$P\left(\boldsymbol{\alpha}\right) \propto \exp\left(-\frac{|\boldsymbol{\alpha}|^p}{|b|}\right) \rightarrow C_p(\boldsymbol{\alpha}) = \lambda_p\,|\boldsymbol{\alpha}|^p \tag{5.45}$$

In other words, using an prior that is exponential in the p-norm, i.e. $P(\boldsymbol{\alpha}) \propto \exp(-|\boldsymbol{\alpha}|^p)$, corresponds to using the p-norm as a cost function.

Though lengthy, this derivation demonstrates an important result: *cost functions and Bayesian regularization are equivalent*. The advantage of explicit cost functions is that they directly lead to optimization problems, and tools for solving such optimization problems are widespread. On the other hand, the Bayesian approach yields distributions for the parameters. Moreover, when $P\left(\boldsymbol{\alpha}\right)$ has no analytic expression, Bayes' theorem makes Monte Carlo sampling straightforward since we can draw many samples from $P\left(\boldsymbol{\alpha}\right)$ and obtain empirical distributions for $\boldsymbol{\alpha}$ given the training data.

5.4 H ADSORPTION ON AG ALLOYS

ABSTRACT

Adsorption energies on surfaces are excellent descriptors of their chemical properties, including their catalytic performance. This has prompted a significant amount of work towards efficiently

predicting adsorption energies without performing expensive quantum chemical calculations, based on semi-empirical or data scientific models. Here, we develop machine learning models to predict H adsorption energies on stepped Ag surface alloys, based on a database of energies calculated with density functional theory. By starting with a simple model and iteratively refining the features extracted from the calculations, we determine the key characteristics of successful models and reduce the number of features required for accurate adsorption energy predictions. We find that nonlinear models are essential for accurate predictions. Moreover, detailed comparisons between model predictions and density functional theory calculations reveal a counterintuitive "self-passivation" exhibited by some dopants, as well as unexpected discontinuous electronic structure effects. Both these phenomena are explained using tight-binding Hamiltonians and analysis of the density of states. Finally, we show that we can estimate the reliability of a particular prediction, mitigating the effects of larger prediction errors in the small fraction of cases where they occur.

5.4.1 Introduction

Many important chemical processes involve hydrogenation, including biodiesel production, petrochemical conversion, fine chemical synthesis, and food processing.[19,63,191] For example, the selective hydrogenation of acetylene to ethylene is a necessary step to purify feedstocks for ethylene polymerization. Therefore, understanding and designing catalysts for hydrogenation reactions has been the subject of intensive research.

Adsorption energies of key intermediates have long proved useful in rationalizing trends in catalytic performance.[5] Hydrogen itself is an important intermediate in hydrogenation reactions, and its adsorption energy can correlate with that of other intermediates, such as hydrocarbons.[120] Therefore, to a first approximation, the hydrogen adsorption energy can be used to screen surfaces for their catalytic performance towards hydrogenation reactions.[5,22,106,158,198] After initial screening, more detailed studies of other important intermediates for a particular reaction can be performed prior to experimental synthesis, characterization, and testing. Ag has the weakest H adsorption energy of the transition metals,[121] and this inertness can lead to Ag alloys having high selectivity for hydrogenation.[191]

Several approaches have been tested for predicting adsorption energies without performing computationally intensive quantum chemical calculations. Some of these approaches are physically motivated, such as estimating electronic structure using tight-binding.[67] Other approaches are data driven,[2,143] culminating in the application of modern machine-learning techniques.[70,101,108,135,174] However, to apply more powerful machine-learning methods and ensure the generality of the approach, larger data sets are needed. The generated model and the process of creating it can in turn provide physical insights into complex problems.

In this work, we perform thousands of calculations of H adsorption energies on stepped Ag alloy surfaces, in order to apply powerful machine-learning methods and enable efficient prediction. We begin by mining the data for catalytic insights and constructing a very simple model. We examine the errors to tailor the machine-learning approach and iteratively develop more accurate models. Iterative improvements and detailed inspection of these models results in high accuracy (median absolute error of 0.3 kcal/mol) and provides a physical understanding of counterintuitive trends, including self-passivation. We find that the best models have low bias, allowing them to readily adapt to varying trends within the data, and that the best surface descriptions combine structural and chemical information.

5.4.2 DATA AND MODELS

Bimetallic alloys are formed from a Ag(211) host structure by replacing between one and eight atoms with a single transition metal element: Ti, Zr, Hf, V, Ta, Cr, Mo, W, Mn, Re, Fe, Ru, Co, Rh, Ir, Ni, Pd, Pt, Cu, Au, Zn, or Cd. The bottom layer is always composed of Ag to simulate the bulk, allowing for the remaining 32 atomic sites to serve as candidates for replacement. The H adsorption energy for each alloy surface is obtained using density functional theory (DFT) calculations with the VASP code[85,87] (see Appendix B for details). Geometry relaxations for each alloy start with the metal atoms placed in the corresponding minimum-energy positions of the pure Ag(211) structure, and H placed at the hcp hollow site adjacent to the step edge, which is its minimum-energy geometry on pure Ag(211).

All regression models are implemented in scikit-learn,[136] an open-source Python package supporting a wide variety of general-purpose machine learning methods. We focus on linear and quadratic polynomials, multilayer neural networks, kernel ridge regression, random forests,[13] and extra forests.[40] When applicable, we also consider ridge and LASSO regularization to adjust the balance between model accuracy and generality. All errors are calculated using randomized 4-fold cross-validation. Parameters are chosen to minimize total squared error on the training set (75%), and the test set (25%) is used to optimize model hyperparameters and evaluate model accuracy. For each model we also consider several preconditioning methods: standardization (zero mean, unit variance), principal component analysis (PCA), and quantile transformation.

5.4.3 RESULTS AND DISCUSSION

5.4.4 DATA MINING

We begin by examining the database to understand how various dopant positions affect H adsorption. For the structures with a single dopant atom, we calculated the standard deviation of adsorption energies for substitutions within the first, second, and third nearest-neighbor shells around the H atom, as well as more distant dopant locations. Standard deviations decline quickly with distance: 0.30 eV among first nearest-neighbors and 0.11 eV among second nearest-neighbors, but only 0.04 eV among third nearest-neighbors or more distant atoms (see

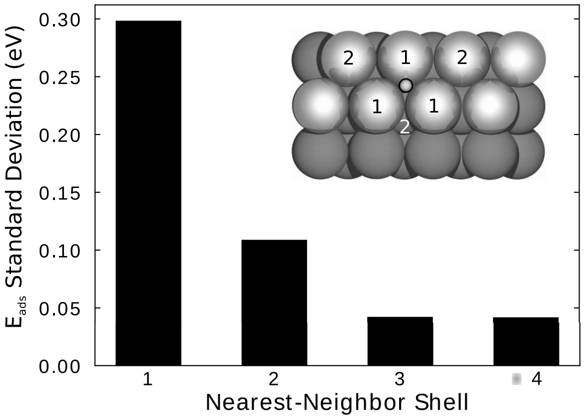


Figure 5.12: Standard deviation of adsorption energies for isolated dopant atoms located at varying distances from the H atom. Distances are labeled according to nearest-neighbor shells, e.g. ‘2’ refers to second nearest-neighbor dopants. Inset: geometry of the Ag(211) surface, with lighter atoms in the surface and darker atoms in the subsurface. Atoms in the first and second nearest-neighbor shells are labeled.

Appendix B, Figure 5.12). Therefore, a single dopant significantly affects adsorption energies only if it lies within the first and second nearest-neighbor shells, a total of four symmetrically inequivalent sites.

Doping in the first nearest-neighbor shell results in qualitatively different effects on adsorption than the second shell. Most dopants have higher d-orbital energies and radii than Ag and therefore tend to enhance H adsorption, as discussed in previous studies.[50,120,121] These dopants strengthen H adsorption energies when located adjacent to H in the surface, but have the opposite effect when located within the second nearest-neighbor shell (see Appendix B, Figure B.2). For elements with low d-orbital energies, specifically Cd, Zn, and Au, these effects are reversed. Therefore, elements that bind strongly to H also passivate nearby Ag atoms towards H adsorption.

We also examined the formation energies of the alloy surfaces, relative to the clean Ag(211) surface and the lowest-energy bulk phase of each dopant element. Our results indicate that alloy surfaces for most elements are thermodynamically unstable, with formation energies exceeding 0 eV, regardless of the number of dopants (see Appendix B, Figure B.3). These unstable elements prefer to lie as isolated atoms in the second or third layers. However, Zn, Cd, Au, Pd, and Pt form thermodynamically stable alloys, with dopant atoms in the surface layer. Under mild reaction conditions, metastable alloys may remain kinetically trapped in their configuration. Therefore, we also searched for arrangements with metastable formation energies between 0 and 1 eV. We find that Ti, Zr, Hf, Rh, Ni, and Cu have metastable arrangements for both bare surfaces and in the presence of a single H atom. Adsorption of a single H atom is insufficient to qualitatively change formation energies: no formation energy changes from > 0 eV to < 0 eV or from > 1 eV to < 1 eV, although higher H coverage or other reaction intermediates are likely to have larger effects.

5.4.5 INITIAL FEATURES AND MODELS

To apply a machine learning model, we must first extract features from each surface in our database to use as model inputs. Each atom in the Ag(211) structure is a symmetrically inequivalent site that can be occupied by either Ag or one of the 22 transition metal dopants. Therefore, it is natural to treat the system as a static lattice. Considering the 31 lattice sites within 8 Å of the H atom, we can then encode the entire structure using a binary matrix $\mathbf{B} \in \{0,1\}^{31\times 23}$, such that $B_{iJ} = 1$ if lattice site i is occupied by an atom of element J. With such a large number of features ($31 \times 23 = 713$) we apply LASSO-regularized linear regression. Regularization applies a penalty to large coefficients and therefore results in simpler models that are less likely to be overfitting. The root-mean-square error (RMSE) of this model is 150 meV (see Figure 5.13). Despite not supplying chemical descriptions of the elements, the linear model captures a significant amount of the variance in H adsorption, with an R^2 statistic of 0.81. However, there are outliers with errors exceeding 1 eV. Therefore, very simple models can provide a surprisingly reasonable estimate for most H adsorption energies, but improvements to the features and model are needed for precise estimates.

Errors from the simple model also provide insight into which data points may be unphysical or difficult to predict. The geometries of the outliers reveal two major sources of error. The first, and most severe, is related to reconstruction, where the surface geometry changes significantly from the input structure. In particular, "differential reconstruction," where the surface reconstructs differently for the bare slab compared to the H/slab systems, often results in high errors. A second class of outliers consists of cases where the H atom relaxes by moving to a different region of the slab, beyond the top and bridge sites adjacent to the initial configuration (see Appendix B for details). These outlier classes are color-coded in Figure 5.13. As indicated in the log-scale histogram in Figure 5.13(b), these two classes of points comprise a small fraction (2.5%) of the total data but dominate the total error. This is consistent with the lower prediction errors we obtained in the absence of relaxation, by predicting "adsorption energies" for the alloy surfaces in their initial bare slab and H/slab geometries.

To test whether these outliers and the high overall RMSE can be mitigated using different models, we compare LASSO- and ridge-regularized linear and polynomial regression, ridge-regularized multilayer neural networks and radial basis function (RBF) kernel regression, random forests,[13] and extra forests.[40] Although more sophisticated variants of these models have been considered previously,[31,94,166,200] these general-purpose models are widely supported in commonly used software packages.

To reduce overfitting and computational effort, we use a more parsimonious feature set for these comparisons, including both "arrangement" and "chemical" features that describe the locations and chemical identity of dopants, respectively. For structural features, we use a one-hot vector of length 31 indicating which sites are substituted: $\mathbf{b}_i = 1$ if lattice site i is occupied by a dopant. We then include the number of first, second, and third-nearest neighbors relative to the H atom, as well as the number of dopants farther away than third-nearest neighbor. This leads to 35 total arrangement features. These features are partially redundant, but linear combinations of features can yield more sparse and accurate models.[13]

Table 5.3: Lowest RMSE Values for Combinations of Models and Feature Scaling Strategies

Regularization	Model	Standardized	PCA	Quantile	Unscaled
LASSO	linear	212	212	**202**	213
LASSO	quadratic	134	135	**120**	137
Ridge	neural network	**106**	**106**	107	fail
Ridge	kernel ridge	111	111	**105**	117
n/a	random forest	**101**	120	102	**101**
n/a	extra forest	100	116	101	**99**

All values are in units of meV, with uncertainties of ±1 meV. The lowest RMSE value for each model is bolded.

For chemical features, we use commonly available atomic properties to account for the dopant element's chemical identity. Based on insights from previous studies,[50,67,124] we include each dopant's tight-binding d-orbital radius and energy[52] and its column in the d block. We also include the dopant's row in the d block, Pauling electronegativity, covalent radius, mass density, enthalpy of fusion, ionization energy, bulk modulus, Poisson ratio, and conductivity, similar to previous work.[101,108] Combining these with the structural features yields 47 total features. Our tests suggest that the accuracy of the nonlinear models is insensitive to the precise feature set, as long as there are enough features that the model can effectively distinguish between different elements.

The lowest RMSE results for each model using these features are shown in Table 5.3 for various feature scaling strategies. Standardization shifts each feature to have zero mean and unit variance, and PCA projects these standardized features onto the corresponding principal components. These are linear transformations. In contrast, quantile scaling is a nonlinear transformation that maps each feature value to the range $(0, 1)$ according to its percentile rank. The linear and quadratic models are generally out-performed by the other four models with lower bias for most scaling choices. Extra forests perform best with an RMSE of approximately 100 meV.

Forests are particularly successful due to their if-else decisions, which handle categorical features in a natural way. In particular, forests are ideally suited for the one-hot vectors we use to encode the structure. The benefits of if-else decisions also explain why PCA worsens forest performance, where the rotation in features tends to mix the one-hot bits with continuous features. Since extra forests yield the lowest RMSE values and are insensitive to the scale of individual features, we use their results in the subsequent discussion. Extra forests perform well across the entire range of adsorption energies, with an improved R^2 score of 0.92 and a median absolute error of only 17 meV.

We find that extra forests, along with all other models, still have high errors associated with differential reconstruction and H migration. In the following we exclude these points to focus on errors that are not due to artifacts of the relaxation. Differential reconstruction yields calculated adsorption energies that do not represent the chemical reactivity of the surface since two qualitatively different surfaces are included in the subtraction. Similarly, when the H atom relaxes to a completely different site, the model attempts to predict the adsorption energy far from the region of the intended hollow site. Because we have a large data set, we can ignore the 2.5% of points with significant atomic motion and still expect to not lose predictive capability for those structures. Retraining extra forests on the remaining 97.5% of the data set yields a significantly lower RMSE of 52 meV.

5.4.6 PHYSICAL INSIGHTS

We can gain insight into the model and into the underlying behavior of the system by examining cases where the model performs poorly. Prediction errors are highest for elements toward the left-hand side of the d block, where V and Mo have the highest element-wise RMSE values of 104 meV and 101 meV, respectively. These large errors are caused by similar configurations having significantly different adsorption energies. We quantified similarity by computing the forest's kernel k_f. Similar to other machine learning methods, the forest kernel $k_f(\mathbf{x}, \mathbf{x}')$ gives the correlation between any two data points described by $\mathbf{x}$ and $\mathbf{x}'$ according to the trained forest model. It is obtained by checking how often the two data points follow the same decision

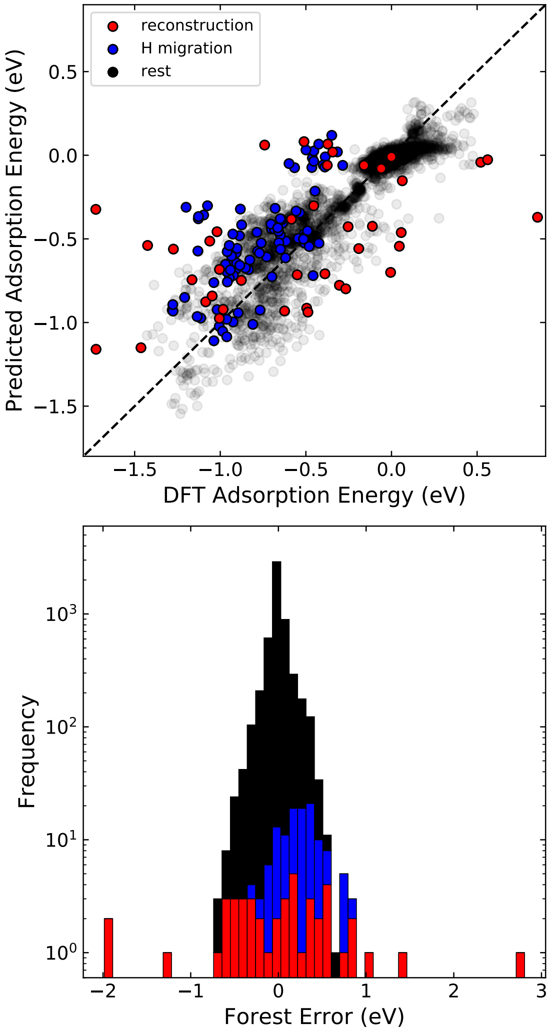


Figure 5.13: (a) Parity plot comparing simple linear model predictions to DFT calculations of H adsorption energies. The dashed line indicates 1:1 parity, and data points corresponding to surface reconstruction and H atom migration are color-coded red and blue, respectively. Four additional outliers due to reconstruction, with DFT adsorption energies of -2.83, -1.98, 1.07, and 2.03 eV are not shown. (b) Log-scale stacked error histogram with the same color coding as (a).

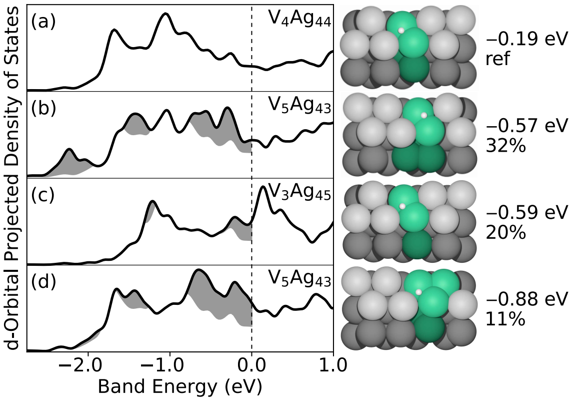


Figure 5.14: d-Orbital projected densities of states on V atoms for selected V/Ag(211) arrangements. Light-colored atoms are in the surface layer and dark-colored atoms are in the subsurface layers. (a) V tetramer with high prediction error. (b-d) The three most similar arrangements according to the forest kernel k_f (see text). Each plot is labeled with the corresponding H adsorption energy and k_f value. Grey regions indicate excess densities of states relative to the reference structure in (a).

pathways. Given k_f, the predicted adsorption energy for any data point $\mathbf{x}$ can be written as an average over all training values y_i, weighted by the corresponding kernel values $k_f(\mathbf{x}, \mathbf{x}_i)$ (see Section 5.2.4.6 for details). Data points $\mathbf{x}_i$ with the largest $k_f(\mathbf{x}, \mathbf{x}_i)$ values thus contribute the most to the predicted value at $\mathbf{x}$.

As an example, we use the forest kernel to better understand a large outlier, the V tetramer in Figure 5.14(a) whose adsorption energy is significantly overestimated. Inspection of the three most similar arrangements and their corresponding H adsorption energies, shown in Figure 5.14(b-d), reveals the source of the error: these most similar structures all have significantly stronger H adsorption. All four arrangements in Figure 5.14 have the same nearest neighbor shell for H, and similar second-nearest neighbor shells. Analyzing the density of states reveals that stronger adsorption is associated with higher d-orbital densities of states just below the Fermi level; these increases are highlighted by the gray regions in Figure 5.14(b-d). These additional states lead to stronger H adsorption energies, consistent with previous studies of O on Pt alloys.[65] For example, the last arrangement in Figure 5.14(d) yields the highest increase in the density of states near the Fermi level and has the strongest H adsorption energy among these four structures.

Similarly, further analysis of the errors reveals cases where adsorption energies tend to be counterintuitive. In particular, 2-dopant structures where both dopants lie in the H's first nearest-neighbor shell have an RMSE that is twice as high as the data set as a whole, and 30% higher than all cases with at least one dopant atom in this shell. Figure 5.15 compares adsorption energies for dimers and isolated dopant atoms along the step edge for all dopant elements considered here. One might expect that dimers of dopant atoms would yield roughly double the effect of a single dopant atom–in most cases, strengthening adsorption. Adsorption energies for dimers of Ti, Zr, and Hf at the step edge are indeed much stronger than the corresponding isolated atom. The same is true for most elements on the right-hand side of the d-block such as Co, Pd, Cu, and Zn. However, adsorption energies for dimers are significantly *weaker*, compared to an isolated atom, for most elements from the center of the d-block, especially Mo and Re. The trend in H adsorption energies for dimers and isolated atoms therefore changes dramatically across various elements. While previous work has generally indicated that adsorption energies change monotonically with the composition of the nearest neighbor shell,[44,122,190] we show that there are many exceptions to this.

Again, close inspection of the densities of states can rationalize this counterintuitive behavior. For the dimers, dopant-dopant hybridization is stronger than Ag-dopant hybridization, resulting in clear d-d bonding and antibonding peaks in most cases (see Figure B.4). Previous work has shown clearly that high densities of states at or just below the Fermi energy leads to strong adsorption.[130] Dimers of elements on the left side of the d-block have d-d antibonding peaks near the Fermi energy, allowing strong adsorption. For dimers of elements in the middle of the d-block, rising d-orbital energies push the antibonding peak above the Fermi energy where it no longer contributes significantly to adsorption, while the d-d bonding peak is still much lower than the Fermi energy. In these cases, the dopant-dopant hybridization results in weaker H adsorption. For elements on

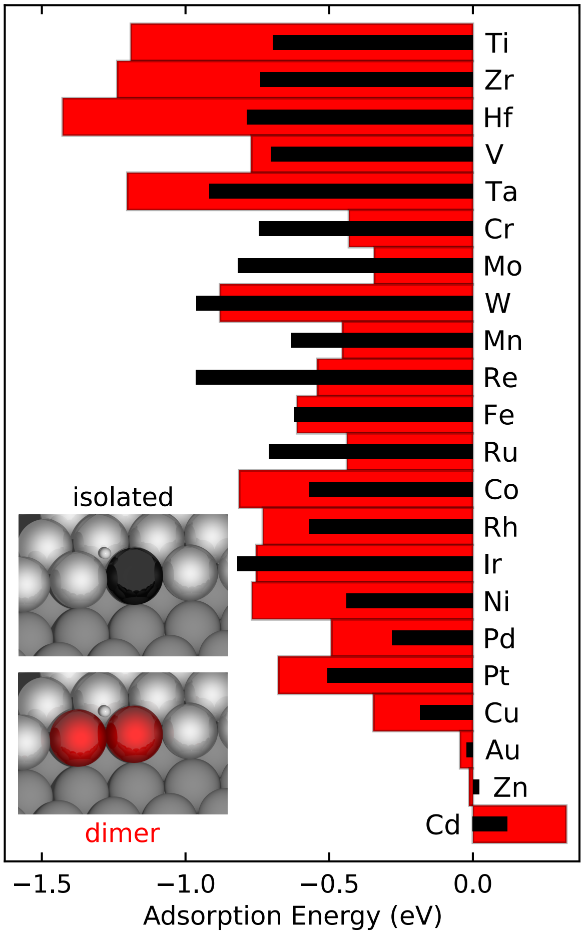


Figure 5.15: Comparison of H adsorption energies for dimers (wide red bars) and isolated atoms (narrow black bars) at the step edge for each dopant element. Inset: sketch of isolated atom and dimer arrangements, where atoms below the step edge are darker spheres.

the right side of the d-block, the d-d bonding peak lies near the Fermi energy, again allowing strong H adsorption.

Overall, the model performs worse in cases where the character of surface states varies discretely and discontinuously with the number and locations of dopant atoms. Small changes to the arrangement, such as the addition or removal of single dopant as shown in Figure 5.14(b-d), can significantly change surface state energies and occupations, and H adsorption energies are sensitive to these changes. This is analogous to finite size effects in small nanoclusters, where chemical reactivity is sensitive to the precise number of atoms and their configuration.[25] With weak dopant-Ag hybridization throughout the d-block, the dopant atoms' electronic structure can be understood in terms of a small nanocluster embedded in the Ag(211) lattice.

5.4.7 FINAL MODEL

In light of the observations described above, we considered a wide variety of surface features to improve forest predictions. We summarize these features here and leave additional details to Appendix B. Because of the sensitivity of the adsorption energy to particular arrangements of dopants, we added one-hot bits indicating the presence of dimer and trimer arrangements that we found to be important. We also added additional chemical features, including the formation energy of a single dopant element in the bulk-like third layer of the Ag(211) surface and H adsorption energies for isolated dopant atoms and dimers at the step edge. Since the combination of d-orbital properties and specific arrangements is important, we also included column-wise sums of the Coulomb matrix[149] using the dopant's column in the d block as its effective charge. Finally, to account for detailed changes in orbital hybridization, we included the eigenvalues of a simplified tight-binding Hamiltonian of the Ag(211) surface and the first five moments of the corresponding eigenvectors projected onto the three metal atoms closest to H. In total these contribute an additional 187 features to the original 47 features. We also tested an alternative encoding of the structure, where a radial distribution function centered at the H atom was calculated for each structure, and a weighted difference between these functions was used to define a kernel. Although this resulted in somewhat smaller errors than the one-hot vector for kernel regression, the difference was small (a few meV), and this scheme is very difficult to implement into a forest, which is a more accurate model for this dataset. Therefore, we continued to use the one-hot vector to encode the structure.

Predicting H adsorption energies using all 234 features leads to overfitting; therefore, we use recursive feature inclusion (RFI) to simplify the model as a form of regularization. RFI heuristically grows a selected subset F of the most important features. At each RFI step a new forest is trained on the union of F and each one of the remaining 234-$|F|$ features, then the feature yielding the lowest out-of-bag RMSE is added to F. Repeating RFI k times thus obtains $|F| = k$ of the most important features. RFI performs better than feature importances computed directly from the forest since the latter are more sensitive to correlations among features.

The lowest RMSE value of 49.7 meV is obtained by retaining only the top 25 features from RFI. The additional decision-making flexibility therefore leads to a slight improvement in model performance while significantly reducing the number of features. Figure 5.16 shows the corresponding parity plot and error histogram. Compared to the initial linear model with 713 one-hot bits, errors on the 97.5% of non-outlier points are significantly lower across the full range of H adsorption energies. In addition, fewer features can be included with only a small penalty to the RMSE. For example, retaining only the top 12 features results in an RMSE value of 52.1 meV, just below that of the original extra forest model's RMSE of 52.5 meV. This corresponds to a nearly four-fold reduction in the number of features while preserving the original forest's accuracy; hence, for future studies this feature set could reduce the amount of data needed for training. Considering many additional features is essential for this reduction since seven out of these top 12

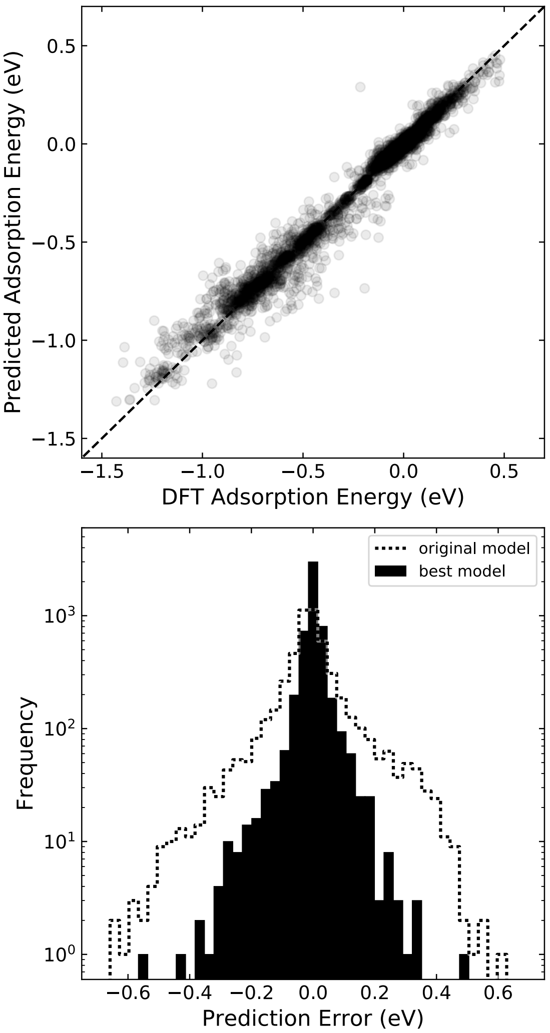


Figure 5.16: Parity plot and error histogram for the best extra forest model on the 97.5% of non-outlier data points. Solid bars in the histogram show the extra forest's error distribution, while the dashed outline shows the the corresponding error distribution for the original model of 713 one-hot bits on the same data.

features, and four of the top six features, are among those additional 187 features considered for the model. The top six features from RFI are the second and third moments of the tight-binding energies projected onto a step atom, the number of first-nearest neighbors, the element's column of the d block, and column sums of the Coulomb matrix corresponding to the nearest-neighbor atom on the terrace and the subsurface atom directly under the hcp hollow site. Using only these top six features yields an RMSE value of 61.0 meV, corresponding to a R^2 statistic of 0.97, showing that only a few well-chosen features can capture most of the variance in H adsorption energies.

Quantifying model uncertainty is important since the accuracy of predictions made by machine learning models, including extra forests, can vary across the data set. Previous work has considered sophisticated metrics for forest uncertainties,[179] but a simple and readily available metric is the "forest variance:" the variance of predictions made by trees in the forest. Forest predictions are means of its trees' predictions, so variances can be obtained easily. Letting ε be the prediction error of a particular point, and v_f be the forest variance, we define the forest z-score to be $z_f = \frac{\varepsilon}{\sqrt{v_f}}$. Z-scores here are calculated using 4-fold test-train splits, training the extra forest on 75% of the data and evaluating errors and prediction variances on the remaining 25% to avoid bias. These z-scores are close to normally distributed and forest variances are typically larger than model errors by approximately 50%. After correcting for this scaling, Figure 5.17 shows the resulting histogram of z-scores for the 97.5% of non-outlier data points. The good agreement between the z-score histogram and the normal distribution demonstrates that $0.68\sqrt{v_f}$ reliably estimates forest uncertainty for each data point across the full range of H adsorption energies. Hence, we have an indication of the reliability of a particular prediction, which allows us to put less trust in predictions with low reliability and mitigate the effects of prediction errors.

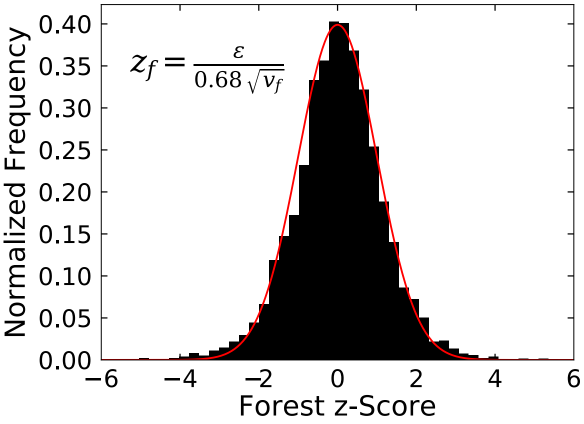


Figure 5.17: Histogram of forest z-scores, computed as the ratios of forest errors to the corresponding standard deviations among all trees' predictions. The red line shows the normal distribution with unit standard deviation for reference.

5.4.8 CONCLUSION

The process of constructing a machine learning model for our database of more than 5,000 H adsorption energies on stepped Ag alloy surfaces has given insight into both model construction and the physics underlying adsorption. First, we showed that a very simple linear model with no chemical or physical insight can give qualitatively useful predictions. However, nonlinear models significantly improve accuracy, and the proper featurization can significantly reduce the needed number of features and the amount of data needed for training. Extra forests were found to be the most accurate model, although other nonlinear models give fairly similar accuracy. The final model is quite accurate in the vast majority of cases, with a median absolute error of 0.3 kcal/mol. The model identifies structures with the same first nearest-neighbor shell and a similar second nearest-neighbor shell as being "similar" (or correlated) to each other, as we would intuitively expect, and the errors are high in cases where these similar structures have significantly different adsorption energies. These structures violate a widespread assumption in machine learning: that similar input values yield similar output values. The model also provides reliable confidence intervals for its predictions, crucial in screening studies to mitigate the possibility of missing promising candidates due to larger-than-expected errors.

We have also elucidated chemical trends and found unexpected behavior in adsorption on dilute alloys. Atoms that strengthen H adsorption when in H's first nearest neighbor shell tend to weaken adsorption when in H's second nearest neighbor shell. In some cases, H adsorption energies can be surprisingly sensitive to small changes in the composition of the H's second and even third nearest neighbor shells. We also found counterintuitive non-monotonic behavior as a function of the composition of the nearest neighbor shell. In particular, there are several cases where two nearby dopants lead to weaker H adsorption than a single nearby dopant. These counterintuitive results are driven by nanocluster-like electronic structure effects arising from weak dopant-Ag hybridization. The model tends to have higher errors in cases that are counterintuitive, suggesting that the model gains a similar "intuition" as human observers. Overall, we have found that the process of constructing a model and understanding its errors can bring physical insight beyond that given by the model itself.

ACKNOWLEDGEMENTS

We acknowledge useful discussions with Georgios Tritsaris. This work was performed as part of Integrated Mesoscale Architectures for Sustainable Catalysis, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award No. DE-SC0012573. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

6

Conclusions

This thesis discusses the details of computational electronic structure theory, starting from the Schrödinger Equation, and its applications to catalytic systems using transition state theory in the microcanonical ensemble. Several specific examples of these applications are presented with various levels of sophistication. Chapters 2 and 3 are traditional and straightforward applications of density functional theory to calculating adsorbate stability and activation energies. These studies demonstrate how atomistic insights and basic transition state theory can be used to characterize catalytic systems and resolve open experimental questions. More sophisticated applications of density functional theory and nonadiabatic dynamics are presented for photodissociation and atomic H adsorption in Sections 1.5.1 and 1.5.2, and H_2 dissociation over Cu_{13} nanoclusters in Chapter 4. These studies illustrate how electronic structure theory can still provide catalytic insight when classical transition state theory and the Born-Oppenheimer approximation fail. Finally, recent advances in machine learning and its applications to catalytic systems are discussed in detail. The strengths and weaknesses of machine learning models are exposed by comparison to density functional theory calculations of H adsorption energies on dilute Ag alloys. General-purpose machine learning models easily capture most of the variance in adsorption energy, but fail to capture the discrete, discontinuous effects of electronic quantization.

I hope the introductions to electronic structure (Chapter 1) and to machine learning (Chapter 5) are broadly useful. The various approximations made in moving from the full Schrödinger

Equation to time-independent density functional theory are particularly important. Violations of these approximations, such as the adiabatic and classical ion approximations, offer both challenges and opportunities for future work. For example, manipulating magnetic states could lead to the development of optically-switchable nanoscale catalysts. Similarly, the discussion of kernel learning and H adsorption on Ag alloys highlights the successes of machine learning in materials science and its ongoing challenges. The results of Section 5.4 demonstrates that even simple machine learning models can quickly and accurately predict adsorption energies on complex surfaces. On the other hand, detailed comparisons and analyzing outliers demonstrates that general-purpose models have serious limitations. For dilute Ag alloys, the electronic structure of the surface is sensitive to the precise number and locations of dopant atoms. Continuous models, such as random forests and neural networks, are poorly suited for predicting these effects. More specialized machine learning models for atomistic systems may fare better, such as convolutional neural networks, but work on their practical implementations and theoretical justifications is in its infancy. Using machine learning models for first-principles predictions is still a distant dream.

Bibliography

[1] H. Y. Afeefy, J. F. Liebman, and S. E. Stein. *Neutral Thermochemical Data.* National Institute of Standards and Technology, Gaithersburg, MD, 2018.

[2] A. N. Andriotis, G. Mpourmpakis, S. Broderick, K. Rajan, S. Datta, M. Sunkara, and M. Menon. Informatics guided discovery of surface structure-chemistry relationships in catalytic nanoparticles. *J. Chem. Phys.*, **140**, 094705 (2014).

[3] J. Y. Baek, I.-Y. Jeon, and J.-B. Baek. Edge-Iodine/Sulfonic Acid-Functionalized Graphene Nanoplatelets as Efficient Electrocatalysts for Oxygen Reduction Reaction. *J. Mater. Chem. A*, **2**, 8690 (2014).

[4] A. D. Becke. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.*, **98**, 5648 (1993).

[5] O. Y. Beeck. Hydrogenation Catalysts. *Proc. R. Soc. a*, **177**, 118 (1950).

[6] M. Behrens, F. Studt, I. Kasatkin, S. Kühl, M. Hävecker, F. Abild-Pedersen, S. Zander, F. Girgsdies, P. Kurr, B.-L. Kniep, M. Tovar, R. W. Fischer, J. K. Nørskov, and R. Schlögl. The Active Site of Methanol Synthesis over Cu/ZnO/Al2O3 Industrial Catalysts. *Science*, **336**, 893 (2012).

[7] D. Bird, M. Persson, J. Trail, and S. Holloway. Dynamics of the Spin Transition in the Adsorption of Hydrogen Atoms on Metals. *Surf. Sci.*, **566-568**, 761 (2004).

[8] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch. Structural Relaxation Made Simple. *Phys. Rev. Lett.*, **97**, 170201 (2006).

[9] P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, **50**, 17953 (1994).

[10] M. B. Boucher, M. D. Marcinkowski, M. L. Liriano, C. J. Murphy, E. A. Lewis, A. D. Jewell, M. F. G. Mattera, G. Kyriakou, M. Flytzani-Stephanopoulos, and E. C. H. Sykes. Molecular-Scale Perspective of Water-Catalyzed Methanol Dehydrogenation to Formaldehyde. *ACS Nano*, **7**, 6181 (2013).

[11] M. Bowker and R. Madix. XPS, UPS and Thermal Desorption Studies of Alcohol Adsorption on Cu(110): II. Higher Alcohols. *Surf. Sci.*, **116**, 549 (1982).

[12] M. Bowker and R. J. Madix. XPS, UPS and Thermal Desorption Studies of Alcohol Adsorption on Cu(110): I. Methanol. *Surf. Sci.*, **95**, 190 (1980).

[13] L. Breiman. Random Forests. *Mach. Learn.*, **45**, 5 (2001).

[14] C. T. Campbell, L. Árnadóttir, and J. R. V. Sellers. Kinetic Prefactors of Reactions on Solid Surfaces. *Zeitschrift für Phys. Chemie*, **227**, 1435 (2013).

[15] C. T. Campbell and J. R. V. Sellers. The Entropies of Adsorbed Molecules. *J. Am. Chem. Soc.*, **134**, 18109 (2012).

[16] A. Castro, M. A. L. Marques, and A. Rubio. Propagators for the Time-Dependent KohnâĂŞSham Equations. *J. Chem. Phys.*, **121**, 3425 (2004).

[17] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim. The Electronic Properties of Graphene. *Rev. Mod. Phys.*, **81**, 109 (2009).

[18] A. Chen and R. Masel. Direct Conversion of Methanol to Formaldehyde in the Absence of Oxygen on Cu(210). *Surf. Sci.*, **343**, 17 (1995).

[19] B. Chen, U. Dingerdissen, J. Krauter, H. Lansink Rotgerink, K. Möbus, D. Ostgard, P. Panster, T. Riermeier, S. Seebald, T. Tacke, and H. Trauthwein. New developments in hydrogenation catalysis particularly in synthesis of fine and intermediate chemicals. *Appl. Catal. A Gen.*, **280**, 17 (2005).

[20] W. Chen, E. D. Cubuk, M. M. Montemore, C. Reece, R. J. Madix, C. M. Friend, and E. Kaxiras. A Comparative Ab Initio Study of Anhydrous Dehydrogenation of Linear-Chain Alcohols on Cu(110). *J. Phys. Chem. C*, page acs.jpcc.8b01698 (2018).

[21] W. Chen, P. Cui, W. Zhu, E. Kaxiras, Y. Gao, and Z. Zhang. Atomistic Mechanisms for Bilayer Growth of Graphene on Metal Substrates. *Phys. Rev. B*, **91**, 045408 (2015).

[22] W. Chen, E. J. Santos, W. Zhu, E. Kaxiras, and Z. Zhang. Tuning the electronic and chemical properties of monolayer mos$_2$ adsorbed on transition metal substrates. *Nano Lett.*, **13**, 509 (2013).

[23] J. P. Chou, C. R. Hsing, C. M. Wei, C. Cheng, and C. M. Chang. Ab Initio Random Structure Search for 13-Atom Clusters of fcc Elements. *J. Phys.: Condens. Matter*, **25**, 125305 (2013).

[24] A. J. Cohen, P. Mori-Sánchez, and W. Yang. Challenges for Density Functional Theory. *Chem. Rev.*, **112**, 289 (2012).

[25] A. S. Crampton, M. D. Rötzer, C. J. Ridge, F. F. Schweinberger, U. Heiz, B. Yoon, and U. Landman. Structure Sensitivity in the Nonscalable Regime Explored via Catalysed Ethylene Hydrogenation on Supported Platinum Nanoclusters. *Nat. Commun.*, **7**, 10389 (2016).

[26] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov, and J. G. Ángyán. Assessing the Performance of Recent Density Functionals for Bulk Solids. *Phys. Rev. B*, **79**, 155107 (2009).

[27] G. Cybenko. Approximation by Superpositions of a Sigmoidal Function. *Math. Control. Signals, Syst.*, **2**, 303 (1989).

[28] A. Davies and Z. Ghahramani. The Random Forest Kernel and Other Kernels for Big Data from Random Partitions (2014).

[29] D. Dheeru and E. Karra Taniskidou. UCI Machine Learning Repository, 2017.

[30] D. K. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, jun 2014.

[31] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, F. Bombarell, T. Hirzel, and A. Aspuru-Guzik. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS 2015*, pages 2224–2232. Curran Associates, Inc., 2015.

[32] H. Eyring. The Activated Complex in Chemical Reactions. *J. Chem. Phys.*, **3**, 107 (1935).

[33] L. Fernández-Seivane, M. A. Oliveira, S. Sanvito, and J. Ferrer. On-Site Approximation for Spin-Orbit Coupling in Linear Combination of Atomic Orbitals Density Functional Methods. *J. Phys. Condens. Matter*, **18**, 7999 (2006).

[34] E. E. Fileti, P. Chaudhuri, and S. Canuto. Relative Strength of Hydrogen Bond Interaction in Alcohol-Water Complexes. *Chem. Phys. Lett.*, **400**, 494 (2004).

[35] A. W. Franz, H. Kronemayer, D. Pfeiffer, R. D. Pilz, G. Reuss, W. Disteldorf, A. O. Gamer, and A. Hilt. Formaldehyde, nov 2016.

[36] Q. Fu and Y. Luo. Catalytic Activity of Single Transition-Metal Atom Doped in Cu(111) Surface for Heterogeneous Hydrogenation. *J. Phys. Chem. C*, **117**, 14618 (2013).

[37] S. García-Gil, A. García, N. Lorente, and P. Ordejón. Optimal Strictly Localized Basis Sets for Noble Metal Surfaces. *Phys. Rev. B*, **79**, 075441 (2009).

[38] H. R. Gerberich and G. C. Seaman. *Formaldehyde*. John Wiley & Sons, Inc., Hoboken, NJ, USA, jan 2013.

[39] B. Gergen, H. Nienhaus, W. H. Weinberg, and E. W. McFarland. Chemically Induced Electronic Excitations at Metal Surfaces. *Science*, **294**, 2521 (2001).

[40] P. Geurts, D. Ernst, and L. Wehenkel. Extremely Randomized Trees. *Mach. Learn.*, **63**, 3 (2006).

[41] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural Message Passing for Quantum Chemistry (2017).

[42] J. R. B. Gomes and J. A. N. F. Gomes. A DFT Study of the Methanol Oxidation Catalyzed by a Copper Surface. *Surf. Sci.*, **471**, 59 (2001).

[43] J. R. B. Gomes, J. A. N. F. Gomes, and F. Illas. First-Principles Study of the Adsorption of Formaldehyde on the Clean and Atomic Oxygen Covered Cu(111) surface. *J. Mol. Catal. A Chem.*, **170**, 187 (2001).

[44] J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, and J. K. Nørskov. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.*, **5**, 909 (2006).

[45] J. Greeley and M. Mavrikakis. Methanol Decomposition on Cu(111): A DFT Study. *J. Catal.*, **208**, 291 (2002).

[46] S. Grimme. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.*, **27**, 1787 (2006).

[47] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.*, **132**, 154104 (2010).

[48] B. Guo, L. Fang, B. Zhang, and J. R. Gong. Graphene Doping: A Review. *Insciences J.*, **1**, 80 (2011).

[49] J. Hafner. Ab-Initio Simulations of Materials Using VASP: Density-Functional Theory and Beyond. *J. Comput. Chem.*, **29**, 2044 (2008).

[50] B. Hammer and J. Nørskov. Theoretical surface science and catalysisâĂŤcalculations and concepts. In *Adv. Catal.*, volume 45, pages 71–129. Elsevier, 2000.

[51] B. Hammer and J. K. Norskov. Why Gold is the Noblest of All the Metals. *Nature*, **376**, 238 (1995).

[52] W. A. Harrison. *Electronic Structure and the Properties of Solids : the Physics of the Chemical Bond*. Dover Publications, Inc., New York, New York, USA, 1989.

[53] M. Head-Gordon and J. C. Tully. Molecular Dynamics with Electronic Frictions. *J. Chem. Phys.*, **103**, 10137 (1995).

[54] G. Henkelman. *Methods for Calculating Rates of Transitions with Application to Catalysis and Crystal Growth*. PhD thesis, University of Washington, 2001.

[55] G. Henkelman, A. Arnaldsson, and H. Jónsson. A Fast and Robust Algorithm for Bader Decomposition of Charge Density. *Comput. Mater. Sci.*, **36**, 354 (2006).

[56] G. Henkelman and H. Jónsson. A Dimer Method for Finding Saddle Points on High Dimensional Potential Surfaces Using Only First Derivatives. *J. Chem. Phys.*, **111**, 7010 (1999).

[57] G. Henkelman and H. Jónsson. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.*, **113**, 9978 (2000).

[58] G. Henkelman, B. P. Uberuaga, and H. Jónsson. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.*, **113**, 9901 (2000).

[59] A. Heyden, A. T. Bell, and F. J. Keil. Efficient Methods for Finding Transition States in Chemical Reactions: Comparison of Improved Dimer Method and Partitioned Rational Function Optimization method. *J. Chem. Phys.*, **123**, 224101 (2005).

[60] F. L. Hirshfeld. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta*, **44**, 129 (1977).

[61] P. Hofmann, K.-M. Schindler, S. Bao, V. Fritzsche, D. Ricken, A. Bradshaw, and D. Woodruff. The Geometric Structure of the Surface Methoxy Species on Cu(111). *Surf. Sci.*, **304**, 74 (1994).

[62] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, **136**, B864 (1964).

[63] C. Hu, D. Creaser, S. Siahrostami, H. Grönbeck, H. Ojagh, and M. Skoglundh. Catalytic hydrogenation of C=C and C=O in unsaturated fatty acid methyl esters. *Catal. Sci. Technol.*, **4**, 2427 (2014).

[64] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graph.*, **14**, 33 (1996).

[65] M. P. Hyman and J. W. Medlin. Effects of Electronic Structure Modifications on the Adsorption of Oxygen Reduction Reaction Intermediates on Model Pt(111)-Alloy Surfaces. *J. Phys. Chem. C*, **111**, 17052 (2007).

[66] J. Ibañez-Azpiroz, A. Eiguren, E. Y. Sherman, and A. Bergara. Spin-Flip Transitions Induced by Time-Dependent Electric Fields in Surfaces with Strong Spin-Orbit Interaction. *Phys. Rev. Lett.*, **109**, 156401 (2012).

[67] N. İnoğlu and J. R. Kitchin. Identification of Sulfur-Tolerant Bimetallic Surfaces Using DFT Parametrized Models and Atomistic Thermodynamics. *ACS Catal.*, **1**, 399 (2011).

[68] D.-e. Jiang, B. G. Sumpter, and S. Dai. Unique Chemical Reactivity of a Graphene Nanoribbon's Zigzag Edge. *J. Chem. Phys.*, **126**, 134701 (2007).

[69] Y. Jiao, Y. Zheng, M. Jaroniec, and S. Z. Qiao. Origin of the Electrocatalytic Oxygen Reduction Activity of Graphene-Based Catalysts: A Roadmap to Achieve the Best Performance. *J. Am. Chem. Soc.*, **136**, 4394 (2014).

[70] R. Jinnouchi and R. Asahi. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.*, **8**, 4279 (2017).

[71] G. Jo, M. Choe, S. Lee, W. Park, Y. H. Kahng, and T. Lee. The Application of Graphene as Electrodes in Electrical and Optical Devices. *Nanotechnology*, **23**, 112001 (2012).

[72] S. M. Johnston, A. Mulligan, V. Dhanak, and M. Kadodwala. The Structure of Methanol and Methoxy on Cu(111). *Surf. Sci.*, **530**, 111 (2003).

[73] N. Jung, A. C. Crowther, N. Kim, P. Kim, and L. Brus. Raman Enhancement on Graphene: Adsorbed and Intercalated Molecular Species. *ACS Nano*, **4**, 7005 (2010).

[74] N. Jung, N. Kim, S. Jockusch, N. J. Turro, P. Kim, and L. Brus. Charge Transfer Chemical Doping of Few Layer Graphenes: Charge Distribution and Band Gap Formation. *Nano Lett.*, **9**, 4133 (2009).

[75] J. Junquera, Ó. Paz, D. Sánchez-Portal, and E. Artacho. Numerical Atomic Orbitals for Linear-Scaling Calculations. *Phys. Rev. B*, **64**, 235111 (2001).

[76] G. Kalita, K. Wakita, M. Takahashi, and M. Umeno. Iodine Doping in Solid Precursor-Based CVD Growth Graphene Film. *J. Mater. Chem.*, **21**, 15209 (2011).

[77] S. Karakalos, Y. Xu, F. Cheenicode Kabeer, W. Chen, J. C. F. Rodríguez-Reyes, A. Tkatchenko, E. Kaxiras, R. J. Madix, and C. M. Friend. Noncovalent Bonding Controls Selectivity in Heterogeneous Catalysis: Coupling Reactions on Gold. *J. Am. Chem. Soc.*, **138**, 15243 (2016).

[78] F. Karlický, K. Kumara Ramanatha Datta, M. Otyepka, and R. Zbořil. Halogenated Graphenes: Rapidly Growing Family of Graphene Derivatives. *ACS Nano*, **7**, 6434 (2013).

[79] I. N. Kholmanov, C. W. Magnuson, A. E. Aliev, H. Li, B. Zhang, J. W. Suk, L. L. Zhang, E. Peng, S. H. Mousavi, A. B. Khanikaev, R. Piner, G. Shvets, and R. S. Ruoff. Improved Electrical Conductivity of Graphene Films Integrated with Metal Nanowires. *Nano Lett.*, **12**, 5679 (2012).

[80] H. K. Kim, O. Renault, A. Tyurnina, J. F. Guillet, J. P. Simonato, D. Rouchon, D. Mariolle, N. Chevalier, and J. Dijon. Doping Characteristics of Iodine on As-Grown Chemical Vapor Deposited Graphene on Pt. *Ultramicroscopy*, **159**, 470 (2015).

[81] S. J. Kim, W. Song, S. Kim, M.-A. Kang, S. Myung, S. Sook Lee, J. Lim, and K.-S. An. Tunable Functionalization of Graphene Nanosheets for Graphene-Organic Hybrid Photodetectors. *Nanotechnology*, **27**, 075709 (2016).

[82] J. Klimeš, D. R. Bowler, and A. Michaelides. Van der Waals Density Functionals Applied to Solids. *Phys. Rev. B*, **83**, 195131 (2011).

[83] M. Klintenberg, S. Lebègue, M. I. Katsnelson, and O. Eriksson. Theoretical Analysis of the Chemical Bonding and Electronic Structure of Graphene Interacting with Group IA and Group VIIA Elements. *Phys. Rev. B*, **81**, 085433 (2010).

[84] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140 (1965).

[85] G. Kresse and J. Furthmüller. Efficiency of Ab Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.*, **6**, 15 (1996).

[86] G. Kresse and J. Furthmüller. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B*, **54**, 11169 (1996).

[87] G. Kresse and J. Hafner. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B*, **47**, 558 (1993).

[88] G. Kresse and J. Hafner. Ab Initio Molecular-Dynamics Simulation of the Liquid-Metal-Amorphous-Semiconductor Transition in Germanium. *Phys. Rev. B*, **49**, 14251 (1994).

[89] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, **59**, 1758 (1999).

[90] T. Kunert and R. Schmidt. Non-Adiabatic Quantum Molecular Dynamics: General Formalism and Case Study H2+ in Strong Laser Fields. *Eur. Phys. J. D - At. Mol. Opt. Phys.*, **25**, 15 (2003).

[91] G. Kyriakou, M. B. Boucher, A. D. Jewell, E. A. Lewis, T. J. Lawton, A. E. Baber, H. L. Tierney, M. Flytzani-Stephanopoulos, and E. C. H. Sykes. Isolated Metal Atom Geometries as a Strategy for Selective Heterogeneous Hydrogenations. *Science*, **335**, 1209 (2012).

[92] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. The Atomic Simulation Environment-a Python Library for Working with Atoms. *J. Phys.: Condens. Matter*, **29**, 273002 (2017).

[93] T. J. Lawton, J. Carrasco, A. E. Baber, A. Michaelides, and E. C. H. Sykes. Hydrogen-Bonded Assembly of Methanol on Cu(111). *Phys. Chem. Chem. Phys.*, **14**, 11846 (2012).

[94] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE*, **86**, 2278 (1998).

[95] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B*, **37**, 785 (1988).

[96] C.-H. Lee, G.-s. Kang, Y.-K. Lee, S. Lee, S. M. Jo, S. J. Yoo, J. H. Jang, D. C. Lee, and H.-I. Joh. Synthesis and Properties of Nitrogen and Iodine Co-Functionalized Graphene Oxide and Its Electrochemical Applications. *Sci. Adv. Mater.*, **8**, 28 (2016).

[97] M. Levy. Universal Variational Functionals of Electron Densities, First-Order Density Matrices, and Natural Spin-Orbitals and Solution of the v-Representability Problem. *Proc. Natl. Acad. Sci.*, **76**, 6062 (1979).

[98] R. Li, M. Zhang, and Y. Yu. A DFT Study on the Cu (111) Surface for Ethyl Acetate Synthesis from Ethanol Dehydrogenation. *Appl. Surf. Sci.*, **258**, 6777 (2012).

[99] X. Li, W. Cai, J. An, S. Kim, J. Nah, D. Yang, R. Piner, A. Velamakanni, I. Jung, E. Tutuc, S. K. Banerjee, L. Colombo, and R. S. Ruoff. Large-Area Synthesis of High-Quality and Uniform Graphene Films on Copper Foils. *Science*, **324**, 1312 (2009).

[100] X. Li and K. H. Lim. DFT Study of Steam Reforming of Formaldehyde on Cu, PdZn, and Ir. *ChemCatChem*, **4**, 1311 (2012).

[101] Z. Li, X. Ma, and H. Xin. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal. Today*, **280**, 232 (2017).

[102] M. Lindenblatt and E. Pehlke. Ab Initio Simulation of the Spin Transition During Chemisorption: H/Al(111). *Phys. Rev. Lett.*, **97**, 1 (2006).

[103] M. Lindenblatt, J. van Heys, and E. Pehlke. Molecular Dynamics of Nonadiabatic Processes at Surfaces: Chemisorption of H/Al(111). *Surf. Sci.*, **600**, 3624 (2006).

[104] R. C. Longo and L. J. Gallego. Structures of 13-Atom Clusters of fcc Transition Metals by Ab Initio and Semiempirical Calculations. *Phys. Rev. B*, **74**, 193409 (2006).

[105] F. R. Lucci, J. Liu, M. D. Marcinkowski, M. Yang, L. F. Allard, M. Flytzani-Stephanopoulos, and E. C. H. Sykes. Selective Hydrogenation of 1,3-Butadiene on Platinum–Copper Alloys at the Single-Atom Limit. *Nat. Commun.*, **6**, 8550 (2015).

[106] L. Luo, Z. Duan, H. Li, J. Kim, G. Henkelman, and R. M. Crooks. Tunability of the Adsorbate Binding on Bimetallic Alloy Nanoparticles for the Optimization of Catalytic Hydrogenation. *J. Am. Chem. Soc.*, **139**, 5538 (2017).

[107] W. Luo, S. J. Pennycook, and S. T. Pantelides. s-Electron Ferromagnetism in Gold and Silver Nanoclusters. *Nano Lett.*, **7**, 3134 (2007).

[108] X. Ma, Z. Li, L. E. K. Achenie, and H. Xin. Machine-Learning-Augmented Chemisorption Model for CO 2 Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.*, pages 3528–3533 (2015).

[109] P. V. C. Medeiros, A. J. S. Mascarenhas, F. de Brito Mota, and C. M. C. de Castilho. A DFT Study of Halogen Atoms Adsorbed on Graphene Layers. *Nanotechnology*, **21**, 485701 (2010).

[110] J. Mercer. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, **209**, 415 (1909).

[111] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**, 1087 (1953).

[112] C. Michel, F. Auneau, F. Delbecq, and P. Sautet. C-H versus O-H Bond Dissociation for Alcohols on a Rh(111) Surface: A Strong Assistance from Hydrogen Bonded Neighbors. *ACS Catal.*, **1**, 1430 (2011).

[113] C. Michel, J. Zaffran, A. M. Ruppert, J. Matras-Michalska, M. Jędrzejczyk, J. Grams, and P. Sautet. 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.*, **50**, 12450 (2014).

[114] B. P. Miller, O. J. Furlong, and W. T. Tysoe. Surface Chemistry of Isopropoxy Tetramethyl Dioxaborolane on Cu(111). *Langmuir*, **28**, 6322 (2012).

[115] G. Mills, H. Jónsson, and G. K. Schenter. Reversible Work Transition State Theory: Application to Dissociative Adsorption of Hydrogen. *Surf. Sci.*, **324**, 305 (1995).

[116] B. D. Mohapatra, S. P. Mantry, N. Behera, B. Behera, S. Rath, and K. S. K. Varadwaj. Stimulation of Electrocatalytic Oxygen Reduction Activity on Nitrogen Doped Graphene through Noncovalent Molecular Functionalisation. *Chem. Commun.*, **52**, 10385 (2016).

[117] K. Momma and F. Izumi. VESTA 3 for Three-Dimensional Visualization of Crystal, Volumetric and Morphology Data. *J. Appl. Crystallogr.*, **44**, 1272 (2011).

[118] H. J. Monkhorst and J. D. Pack. Special points for brillouin-zone integrations. *Phys. Rev. B*, **13**, 5188 (1976).

[119] M. M. Montemore. Personal Communication, 2017.

[120] M. M. Montemore and J. W. Medlin. A Unified Picture of Adsorption on Transition Metals through Different Atoms. *J. Am. Chem. Soc.*, **136**, 9272 (2014).

[121] M. M. Montemore and J. W. Medlin. Predicting and Comparing C–M and O–M Bond Strengths for Adsorption on Transition Metal Surfaces. *J. Phys. Chem. C*, **118**, 2666 (2014).

[122] M. M. M. Montemore, R. J. R. Madix, and E. Kaxiras. How Does Nanoporous Gold Dissociate Molecular Oxygen? *J. Phys. Chem. C*, **120**, 16636 (2016).

[123] A. Mulligan and M. Kadodwala. Destabilizing Effects of Thiols on Bonding to a Noble Metal: The Effects of Methanethiolate on the Bonding of Aldehydes and Alcohols on Cu(111). *J. Phys. Chem. C*, **114**, 21457 (2010).

[124] J. P. Muscat and D. M. Newns. Theoretical model of chemisorption on metals: I. Formalism. *Surf. Sci.* (1979).

[125] M. Nakata, B. J. Braams, K. Fujisawa, M. Fukuda, J. K. Percus, M. Yamashita, and Z. Zhao. Variational Calculation of Second-Order Reduced Density Matrices by Strong N-Representability Conditions and an Accurate Semidefinite Programming Solver. *J. Chem. Phys.*, **128**, 164113 (2008).

[126] S. Narasimhan and S. de Gironcoli. Ab Initio Calculation of the Thermal Properties of Cu: Performance of the LDA and GGA. *Phys. Rev. B*, **65**, 064302 (2002).

[127] V. U. Nazarov and G. Vignale. Optics of Semiconductors from Meta-Generalized-Gradient-Approximation-Based Time-Dependent Density-Functional Theory. *Phys. Rev. Lett.*, **107**, 216402 (2011).

[128] D. M. Newns. Self-Consistent Model of Hydrogen Chemisorption. *Phys. Rev.*, **178**, 1123 (1969).

[129] H. Nienhaus, H. S. Bergh, B. Gergen, A. Majumdar, W. H. Weinberg, and E. W. Mc Farland. Electron-Hole Pair Creation at Ag and Cu Surfaces by Adsorption of Atomic Hydrogen and Deuterium. *Phys. Rev. Lett.*, **82**, 446 (1999).

[130] J. K. Nørskov, F. Abild-Pedersen, F. Studt, and T. Bligaard. Density functional theory in surface chemistry and catalysis. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 937 (2011).

[131] M. Olazaran. A Sociological Study of the Official History of the Perceptrons Controversy. *Soc. Stud. Sci.*, **26**, 611 (1996).

[132] J. Oviedo and R. E. Palmer. Amorphous Structures of Cu, Ag, and Au Nanoclusters from First Principles Calculations. *J. Chem. Phys.*, **117**, 9548 (2002).

[133] H. Oyanagi, Y. Orimoto, K. Hayakawa, K. Hatada, Z. Sun, L. Zhang, K. Yamashita, H. Nakamura, M. Uehara, A. Fukano, and H. Maeda. Nanoclusters Synthesized by Synchrotron Radiolysis in Concert with Wet Chemistry. *Sci. Rep.*, **4**, 7199 (2015).

[134] J. D. Pack and H. J. Monkhorst. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B*, **13**, 5188 (1976).

[135] P. Pankajakshan, S. Sanyal, O. E. De Noord, I. Bhattacharya, A. Bhattacharyya, and U. Waghmare. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.*, **29**, 4190 (2017).

[136] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825 (2011).

[137] S. Pei, J. Zhao, J. Du, W. Ren, and H.-M. Cheng. Direct Reduction of Graphene Oxide Films into Highly Conductive and Flexible Graphene Films by Hydrohalic Acids. *Carbon N. Y.*, **48**, 4466 (2010).

[138] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, **77**, 3865 (1996).

[139] J. P. Perdew, K. Burke, and M. Ernzerhof. Erratum: Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, **78**, 1396 (1997).

[140] H. L. Poh, P. Šimek, Z. Sofer, and M. Pumera. Halogenation of Graphene with Chlorine, Bromine, or Iodine by Exfoliation in a Halogen Atmosphere. *Chem. - A Eur. J.*, **19**, 2655 (2013).

[141] P. Pulay. Ab Initio Calculation of Force Constants and Equilibrium Geometries in Polyatomic Molecules. *Mol. Phys.*, **17**, 197 (1969).

[142] P. Rani and V. K. Jindal. Designing Band Gap of Graphene by B and N Dopant Atoms. *RSC Adv.*, **3**, 802 (2013).

[143] E.-J. Ras, M. J. Louwerse, M. C. Mittelmeijer-Hazeleger, and G. Rothenberg. Predicting adsorption on metals: simple yet effective descriptors for surface catalysis. *Phys. Chem. Chem. Phys.*, **15**, 4436 (2013).

[144] C. E. Rasmussen and J. Quiñonero-Candela. Healing the Relevance Vector Machine Through Augmentation. In *Proc. 22nd Int. Conf. Mach. Learn. - ICML '05*, pages 689–696, New York, New York, USA, 2005. ACM Press.

[145] P. Redhead. Thermal Desorption of Gases. *Vacuum*, **12**, 203 (1962).

[146] O. Renault, A. Tyurnina, J.-P. Simonato, D. Rouchon, D. Mariolle, N. Chevalier, and J. Dijon. Doping Efficiency of Single and Randomly Stacked Bilayer Graphene by Iodine Adsorption. *Appl. Phys. Lett.*, **105**, 011605 (2014).

[147] K. Rendulic and A. Winkler. Adsorption and Desorption Dynamics as Seen Through Molecular Beam Techniques. *Surf. Sci.*, **299-300**, 261 (1994).

[148] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation Forest: A New Classifier Ensemble Method. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 1619 (2006).

[149] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.*, **108**, 058301 (2012).

[150] J. Russell, S. Gates, and J. Yates. Reaction of Methanol with Cu(111) and Cu(111) + O(ads). *Surf. Sci.*, **163**, 516 (1985).

[151] E. Sanville, S. D. Kenny, R. Smith, and G. Henkelman. Improved Grid-Based Algorithm for Bader Charge Allocation. *J. Comput. Chem.*, **28**, 899 (2007).

[152] E. Scornet. Random Forests and Kernel Methods. pages 1–32 (2015).

[153] J. Shan, F. R. Lucci, J. Liu, M. El-Soda, M. D. Marcinkowski, L. F. Allard, E. C. H. Sykes, and M. Flytzani-Stephanopoulos. Water Co-Catalyzed Selective Dehydrogenation of Methanol to Formaldehyde and Hydrogen. *Surf. Sci.*, **650**, 121 (2016).

[154] N. Shenvi, S. Roy, and J. C. Tully. Nonadiabatic Dynamics at Metal Surfaces: Independent-Electron Surface Hopping. *J. Chem. Phys.*, 130 (2009).

[155] P. Šimek, K. Klímová, D. Sedmidubský, O. Jankovský, M. Pumera, and Z. Sofer. Towards Graphene Iodide: Iodination of Graphite Oxide. *Nanoscale*, **7**, 261 (2015).

[156] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal. The SIESTA Method for Ab Initio Order-N Materials Simulation. *J. Phys.: Condens. Matter*, **14**, 2745 (2002).

[157] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, **98**, 11623 (1994).

[158] F. Studt, F. Abild-Pedersen, T. Bligaard, R. Z. Sørensen, C. H. Christensen, and J. K. Nørskov. Identification of Non-Precious Metal Alloy Catalysts for Selective Hydrogenation of Acetylene. *Science*, **320**, 1320 (2008).

[159] F. Studt, F. Abild-Pedersen, J. B. Varley, and J. K. Nørskov. CO and CO2 Hydrogenation to Methanol Calculated Using the BEEF-vdW Functional. *Catal. Letters*, **143**, 71 (2013).

[160] S. Su, P. Zaza, and A. Renken. Catalytic Dehydrogenation of Methanol to Water-Free Formaldehyde. *Chem. Eng. Technol.*, **17**, 34 (1994).

[161] J. Sun, A. Ruzsinszky, and J. P. Perdew. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.*, **115**, 036402 (2015).

[162] P. H. Svensson and L. Kloo. Synthesis, Structure, and Bonding in Polyiodide and Metal Iodide-Iodine Systems. *Chem. Rev.*, **103**, 1649 (2003).

[163] A. Tamai, W. Meevasana, P. D. C. King, C. W. Nicholson, A. de la Torre, E. Rozbicki, and F. Baumberger. Spin-Orbit Splitting of the Shockley Surface State on Cu(111). *Phys. Rev. B*, **87**, 075113 (2013).

[164] W. Tang, E. Sanville, and G. Henkelman. A Grid-Based Bader Analysis Algorithm Without Lattice Bias. *J. Phys.: Condens. Matter*, **21**, 084204 (2009).

[165] P. Tereshchuk and J. L. F. Da Silva. Ethanol and Water Adsorption on Close-Packed 3d, 4d, and 5d Transition-Metal Surfaces: A Density Functional Theory Investigation with van der Waals Correction. *J. Phys. Chem. C*, **116**, 24695 (2012).

[166] M. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *CrossRef List. Deleted DOIs*, **1**, 211 (2000).

[167] A. Tkatchenko and M. Scheffler. Accurate Molecular van der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.*, **102**, 073005 (2009).

[168] J. R. Trail, D. M. Bird, M. Persson, and S. Holloway. ElectronâĂŞHole Pair Creation by Atoms Incident on a Metal Surface. *J. Chem. Phys.*, **119**, 4539 (2003).

[169] A. Trask, D. Gilmore, and R. Matthew. Modeling Order in Neural Word Embeddings at Scale. In *Proc. 32nd Int. Conf. Mach. Learn.*, pages 2266–2275, Lille, France, 2015. Journal of Machine Learning Research.

[170] D. Tristant, P. Puech, and I. C. Gerber. Theoretical Study of Graphene Doping Mechanism by Iodine Molecules. *J. Phys. Chem. C*, **119**, 12071 (2015).

[171] D. Tristant, P. Puech, and I. C. Gerber. Theoretical Study of Polyiodide Formation and Stability on Monolayer and Bilayer Graphene. *Phys. Chem. Chem. Phys.*, **17**, 30045 (2015).

[172] D. G. Truhlar and B. C. Garrett. Variational Transition State Theory. *Annu. Rev. Phys. Chem.*, **35**, 159 (1984).

[173] J. C. Tully. Molecular Dynamics with Electronic Transitions. *J. Chem. Phys.*, **93**, 1061 (1990).

[174] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, and J. K. Nørskov. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO 2 Reduction. *ACS Catal.*, **7**, 6600 (2017).

[175] R. van Leeuwen. Mapping from Densities to Potentials in Time-Dependent Density-Functional Theory. *Phys. Rev. Lett.*, **82**, 3863 (1999).

[176] N. Vilar-Vidal, J. Rivas, and M. A. López-Quintela. Size Dependent Catalytic Activity of Reusable Subnanometer Copper(0) Clusters. *ACS Catal.*, **2**, 1693 (2012).

[177] G. H. Vineyard. Frequency Factors and Isotope Effects in Solid State Rate Processes. *J. Phys. Chem. Solids*, **3**, 121 (1957).

[178] S. H. Vosko, L. Wilk, and M. Nusair. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: a Critical Analysis. *Can. J. Phys.*, **58**, 1200 (1980).

[179] S. Wager, T. Hastie, and B. Efron. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *J. Mach. Learn. Res.*, **15**, 1625 (2014).

[180] W. L. Wang, E. J. G. Santos, B. Jiang, E. D. Cubuk, C. Ophus, A. Centeno, A. Pesquera, A. Zurutuza, J. Ciston, R. Westervelt, and E. Kaxiras. Direct Observation of a Long-Lived Single-Atom Catalyst Chiseling Atomic Structures in Graphene. *Nano Lett.*, **14**, 450 (2014).

[181] Z. T. Wang, R. A. Hoyt, M. El-Soda, R. J. Madix, E. Kaxiras, and E. C. H. Sykes. Dry Dehydrogenation of Ethanol on Pt-Cu Single Atom Alloys. *Top. Catal.*, pages 1–8 (2017).

[182] Z.-T. Wang, Y. Xu, M. El-Soda, F. R. Lucci, R. J. Madix, C. M. Friend, and E. C. H. Sykes. Surface Structure Dependence of the Dry Dehydrogenation of Alcohols on Cu(111) and Cu(110). *J. Phys. Chem. C*, **121**, 12800 (2017).

[183] E. Wigner. The Transition State Method. *Trans. Faraday Soc.*, **34**, 29 (1938).

[184] J. P. Wilcoxon and B. L. Abrams. Synthesis, Structure and Properties of Metal Nanoclusters. *Chem. Soc. Rev.*, **35**, 1162 (2006).

[185] I. Williams and G. Maggiora. Use and Abuse of the Distinguished-Coordinate Method for Transition-State Structure Searching. *J. Mol. Struct. THEOCHEM*, **89**, 365 (1982).

[186] D. H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Comput.*, **8**, 1341 (1996).

[187] Z. Wu, Y. Han, R. Huang, X. Chen, Y. Guo, Y. He, W. Li, Y. Cai, and N. Wang. Semimetallic-to-Metallic Transition and Mobility Enhancement Enabled by Reversible Iodine Doping of Graphene. *Nanoscale*, **6**, 13196 (2014).

[188] L. Yang, S. Jiang, Y. Zhao, L. Zhu, S. Chen, X. Wang, Q. Wu, J. Ma, Y. Ma, and Z. Hu. Boron-Doped Carbon Nanotubes as Metal-Free Electrocatalysts for the Oxygen Reduction Reaction. *Angew. Chemie Int. Ed.*, **50**, 7132 (2011).

[189] Z. Yao, H. Nie, Z. Yang, X. Zhou, Z. Liu, and S. Huang. Catalyst-Free Synthesis of Iodine-Doped Graphene via a Facile Thermal Annealing Process and its Use for Electrocatalytic Oxygen Reduction in an Alkaline Medium. *Chem. Commun.*, **48**, 1027 (2012).

[190] W.-Y. Yu, L. Zhang, G. M. Mullen, G. Henkelman, and C. B. Mullins. Oxygen Activation and Reaction on PdâĂŞAu Bimetallic Surfaces. *J. Phys. Chem. C*, **119**, 11754 (2015).

[191] F. Zaera. The Surface Chemistry of Metal-Based Hydrogenation Catalysis. *ACS Catal.*, **7**, 4947 (2017).

[192] Y. Zhan, J. Huang, Z. Lin, X. Yu, D. Zeng, X. Zhang, F. Xie, W. Zhang, J. Chen, and H. Meng. Iodine/Nitrogen Co-Doped Graphene as Metal Free Catalyst for Oxygen Reduction Reaction. *Carbon N. Y.*, **95**, 930 (2015).

[193] L. Zhang, J. Niu, M. Li, and Z. Xia. Catalytic Mechanisms of Sulfur-Doped Graphene as Efficient Oxygen Reduction Reaction Catalysts for Fuel Cells. *J. Phys. Chem. C*, **118**, 3545 (2014).

[194] L. Zhang and Z. Xia. Mechanisms of Oxygen Reduction Reaction on Nitrogen-Doped Graphene for Fuel Cells. *J. Phys. Chem. C*, **115**, 11170 (2011).

[195] M. Zhang, R. Yao, H. Jiang, G. Li, and Y. Chen. Catalytic Activity of Transition Metal Doped Cu(111) Surfaces for Ethanol Synthesis from Acetic Acid Hydrogenation: a DFT Study. *RSC Adv.*, **7**, 1443 (2017).

[196] R. Zhang, G. Wang, and B. Wang. Insights Into the Mechanism of Ethanol Formation from Syngas on Cu and an Expanded Prediction of Improved Cu-Based Catalyst. *J. Catal.*, **305**, 238 (2013).

[197] Y. Zhao, J. Wei, R. Vajtai, P. M. Ajayan, and E. V. Barrera. Iodine Doped Carbon Nanotube Cables Exceeding Specific Electrical Conductivity of Metals. *Sci. Rep.*, **1**, 83 (2011).

[198] Y. Zhou, W. Chen, P. Cui, J. Zeng, Z. Lin, E. Kaxiras, and Z. Zhang. Enhancing the hydrogen activation reactivity of nonprecious metal substrates via confined catalysis underneath graphene. *Nano Lett.*, **16**, 6058 (2016).

[199] L. F. Žilnik and J. Golob. Analysis of Separation of a Water-Methanol-Formaldehyde Mixture. *Acta Chim. Slov.*, **50**, 451 (2003).

[200] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **67**, 301 (2005).

A

Magnetization Switches in Cu_{13}

Apart from minor modifications, this appendix is part of a manuscript being prepared for publication:

> **Robert A. Hoyt**, Matthew M. Montemore, and Efthimios Kaxiras. “Nonadiabatic Hydrogen Dissociation on Copper Nanoclusters” *in preparation*

A.1 COMPUTATIONAL SETTINGS

The DFT calculations were carried out with the SIESTA[156] code using double-ζ plus polarization (DZP) basis sets and the GGA exchange-correlation functional of Perdew, Burke, and Ernzerhoff[138,139]. We used cubic unit cells with a side of length 14 Å to isolate nanoclusters from their periodic images, and a mesh cutoff of 290 Ry. The pseudoatomic orbitals were individually optimized using the `simplex` utility in SIESTA as described below to ensure accuracy in the electronic properties. Activation energy calculations were performed using the climbing-image nudged elastic band[54] (CI-NEB) method and the dimer method[56,59] as implemented in the Atomic Simulation Environment package[92] in Python. For the CI-NEB method, the discontinuities in the H_2 dissociation pathways (see Figure 4.3, main text) violate the first-order tangent approximation so more than a dozen images were needed to obtain convergence. In contrast, using the dimer method was more straightforward since it uses much smaller displacements. Finally, time steps of 1 fs were used for molecular dynamics simulations.

A.2 BASIS SET OPTIMIZATION

The finite pseudoatomic orbital (PAO) basis in SIESTA constrains the functional form of Kohn-Sham states and hence the total density, so in principle the PAO basis itself should be variationally optimized in addition to the wavefunction coefficients for all geometries. For example, basis set optimization improves the lattice constant of bulk metals as well as their vacuum charge densities.[37,75] However, in practice the lack of analytic gradients with respect to the basis set parameters makes this optimization computationally expensive. Optimized basis sets tend to be transferrable to chemically similar systems.[75] Without optimization the total density is artificially constrained by sub-optimal basis functions and this can lead to significant errors. In the present study, for example, the default basis set in SIESTA fails to produce molecular H_2 adsorption geometries on Cu_{13}. Therefore we optimized the basis set for each nanoparticle as described by García-Gil et al.[37] The overall strategy is to minimize a fictitious “basis enthalpy” H_{basis} with respect to the PAO parameters $\vec{\beta}$:

$$\min_{\vec{\beta}} H_{\text{basis}} = \min_{\vec{\beta}} \left[E_{\text{DFT}} \left(\vec{R} | \vec{\beta} \right) + p V_{\text{basis}} \left(\vec{\beta} \right) \right] \tag{A.1}$$

where E_{DFT} is the DFT total energy, $\vec{R}$ contains the atomic coordinates, V_{basis} is the total volume of the DFT basis set, and $p = 0.05$ GPa is the fictitious pressure. The pressure p dictates the tradeoff between accuracy (unconstrained orbitals) and speed (compact orbitals). In optimizing the H atom basis, we first optimized the valence shell of isolated H atoms, followed by optimizing all other parameters for the H_2 molecule to ensure good descriptions of both atomic and molecular hydrogen. For the metal atoms, we optimized all parameters for the ground-state nanocluster geometry as predicted by the default basis set in SIESTA.

A.3 LANDAU-ZENER APPROXIMATION

The Landau-Zener formula gives the transition probability from state A to state B in the presence of a constant coupling δ when their relative expected energies vary linearly in time. The probability P_{LZ} is given by

$$P_{\mathrm{LZ}} = 1 - \exp\left(-2\pi\Gamma\right), \;\; \Gamma \equiv \frac{\delta^2/\hbar}{\left|\frac{\partial}{\partial t}\left(E_A(t) - E_B(t)\right)\right|} \tag{A.2}$$

The key quantity of interest is the crossing rate, $r \equiv \left|\frac{\partial}{\partial t}\left(E_A(t) - E_B(t)\right)\right|$, in the denominator of Γ. In this case A is the state with high magnetization and B is the state with low magnetization. Most of the precession from high to low magnetization occurs when the relative energies are comparable to the coupling δ, that is, $|E_h - E_l| \lesssim \delta$, so the crossing rate is evaluated at the avoided crossing where $E_h = E_l$. We obtain the crossing rate from our molecular dynamics trajectories by approximating the time derivative using finite differences. Since the time-dependent magnetization $m(t)$ changes from $3\,\mu_{\mathrm{B}}$ to $5\,\mu_{\mathrm{B}}$ for each trajectory, the avoided crossing is located when $m(t) = 4\,\mu_{\mathrm{B}}$. The magnetization depends mostly on the relative occupation of the two opposite-spin states involved in the magnetization switch, so their relative energy $\Delta\varepsilon \equiv \varepsilon_{\uparrow} - \varepsilon_{\downarrow}$ can be extracted from $m(t)$. Defining the "strong coupling region" to be $|\Delta\varepsilon(t)| \leq 5\delta$, we then use the simulation time steps $T_{\pm}$ closest to $\Delta\varepsilon\left(T_{\pm}\right) = \pm 5\,\delta$ and re-calculate the fixed-magnetization DFT energies $E_h\left(T_{\pm}\right)$ and $E_l\left(T_{\pm}\right)$ at the geometries of those time steps. Letting $\Delta T = T_{+} - T_{-}$, these values yield the following finite difference approximation:

$$\begin{aligned} r &\equiv \left| \frac{\partial}{\partial t} \left(E_h(t) - E_l(t) \right) \right| \\ &= \left| \frac{\partial E_h(t)}{\partial t} - \frac{\partial E_l(t)}{\partial t} \right| \\ &\approx \left| \left(\frac{E_h(T_+) - E_h(T_-)}{\Delta T} \right) - \left(\frac{E_l(T_+) - E_l(T_-)}{\Delta T} \right) \right| \\ &= \frac{1}{\Delta T} \left| \Delta E_h - \Delta E_l \right| \end{aligned} \tag{A.3}$$

Finally, this approximation for r can be used to calculate Γ:

$$\Gamma \approx \frac{\delta^2 \hbar}{\frac{1}{\Delta T} \left| \Delta E_h - \Delta E_l \right|} = \frac{\delta^2 \Delta T / \hbar}{\left| \Delta E_h - \Delta E_l \right|} \tag{A.4}$$

We used Equation (A.4) to calculate P_{LZ} for all trajectories shown in Figure 4.4.

A.4 DFT ESTIMATE OF SPIN-ORBIT COUPLING

Kohn-Sham DFT provides exact ground-state densities and energies in principle, but the corresponding single Slater determinant “wavefunction” simply facilitates the calculation of kinetic energies and single-particle densities and therefore does not directly correspond to the many-body wavefunction. In particular, the spin state implied by the Slater determinant does not match that of the true many-body ground state in general. Spin-orbit couplings between many-body total spin states are therefore unavailable in Kohn-Sham DFT. However, the net magnetization is a property of the spin-resolved total density so magnetization changes are accessible in Kohn-Sham DFT. Furthermore, by the van Leeuwen theorem[175] (of which the Runge-Gross theorem can be considered a special case), real-time propagation of the Kohn-Sham states produces (in principle) the same time-dependent single-particle density as the actual many-body wavefunction. Therefore, when spin-flip excitations result in magnetization changes, we can approximate the spin-orbit coupling constant by appealing to magnetization changes within time-dependent DFT (TDDFT).

In TDDFT the Kohn-Sham equations are propagated according to the DFT Hamiltonian:

$$\frac{\partial}{\partial t}|\psi_i\rangle = -\frac{i}{\hbar}H_{\mathrm{DFT}}|\psi_i\rangle \tag{A.5}$$

The full DFT Hamiltonian can be considered to be a sum of the colinear spin Hamiltonian $H^{\sigma\sigma}$, which commutes with spin, and the spin-orbit coupling Hamiltonian $H_{\mathrm{soc}}^{\sigma'\sigma}$, which does not. Eigenstates of the colinear spin Hamiltonian can then be assigned spin quantum numbers along a chosen axis to yield pure-spin states in the “up” and “down” directions. In the absence of spin-orbit coupling, commutation between spin and the colinear Hamiltonian results in a constant net magnetization. The introduction of the spin-orbit coupling Hamiltonian, however, couples opposite-spin states and allows for spin precession, resulting in time-dependent net magnetization. The concept is most clear in the equivalent matrix form using a simplified basis of two opposite-spin states $|\uparrow\rangle$ and $|\downarrow\rangle$. These states represent the opposite-spin KS states in Figure 4.2 responsible for the magnetization switch:

$$\begin{aligned}
\frac{\partial \mathbf{p}}{\partial t} &= -\frac{i}{\hbar}\begin{pmatrix} \langle\uparrow|H|\uparrow\rangle & \langle\uparrow|H|\downarrow\rangle \\ \langle\downarrow|H|\uparrow\rangle & \langle\downarrow|H|\downarrow\rangle \end{pmatrix}\mathbf{p} \\
&= -\frac{i}{\hbar}\begin{pmatrix} \varepsilon_\uparrow & \langle\uparrow|H_{\mathrm{soc}}|\downarrow\rangle \\ \langle\downarrow|H_{\mathrm{soc}}|\uparrow\rangle & \varepsilon_\downarrow \end{pmatrix}\mathbf{p} \\
&\equiv -\frac{i}{\hbar}\begin{pmatrix} \varepsilon_\uparrow & \delta \\ \delta^\dagger & \varepsilon_\downarrow \end{pmatrix}\mathbf{p}
\end{aligned} \tag{A.6}$$

where $\varepsilon_\sigma \equiv \langle\sigma|H|\sigma\rangle$ and δ is the spin-orbit matrix element between the two opposite-spin states.

The maximum magnetization precession rate occurs at the avoided crossing, $\varepsilon_\uparrow - \varepsilon_\downarrow = 0$, where the eigenstates of H are equal spin superpositions $\frac{1}{2}\left(|\uparrow\rangle + |\downarrow\rangle\right)$. In this case an initially pure-spin state fully precesses between $|\uparrow\rangle$ and $|\downarrow\rangle$, and the net magnetization oscillates accordingly. Away from the avoided crossing, $\left|\varepsilon_\uparrow - \varepsilon_\downarrow\right| \gg |\delta|$ and the eigenstates are nearly identical to the pure-spin eigenstates of the colinear Hamiltonian. An initially pure-spin state then rotates about the true mixed-spin eigenstate and the net magnetization exhibits small oscillations. Including all of the Kohn-Sham states complicates the numerical solutions, but the precession of a one pure-spin state into another state of opposite spin is still maximized at avoided crossings, and the rate is still dictated by the corresponding spin-orbit coupling matrix element $\langle \psi_j^{\sigma'} | H_{soc} | \psi_i^{\sigma} \rangle$. Using $|\delta| \equiv |\langle\uparrow| H_{\text{soc}} |\downarrow\rangle|$ can therefore provide a reasonable DFT-based estimate of the spin-orbit splitting in the Landau-Zener approximation at the avoided crossing.

B

Details of Machine Learning on Ag Alloys

B.1 DENSITY FUNCTIONAL THEORY CALCULATIONS

Training data for machine learning was obtained by running high-throughput density functional theory calculations on Ag alloy surfaces. All calculations were performed using the projector augmented wave (PAW) method[9,89] within VASP.[85,87] The Perdew-Burke-Ernzerhof form of the generalized gradient approximation was employed to describe electron exchange and correlation.[138] Van der Waals interactions were taken into account using the Tkatchenko-Scheffler method.[167] Ag alloys were modeled using the stepped Ag(211) surface with 4 atoms along the step edge and 4 layers along the (111) direction. Atoms in the bottom layer along the (111) direction were fixed in their bulk positions. 28.0 Å of vacuum space was used along the z direction to prevent spurious image interactions. A $7 \times 7 \times 1$ Monkhorst-Pack[118] k-point grid was used to sample the Brillouin zone along with a plane-wave cutoff energy of 400 eV. Structural relaxations were performed until the maximum force component was less than 0.04 eV/Å. The optimized lattice constant of bulk Ag with these choices is 3.96 Å. Because of the large number of structures in the database, in the present study we do not consider the effects of spin polarization and zero-point vibrational energy which demand significantly more computational resources.

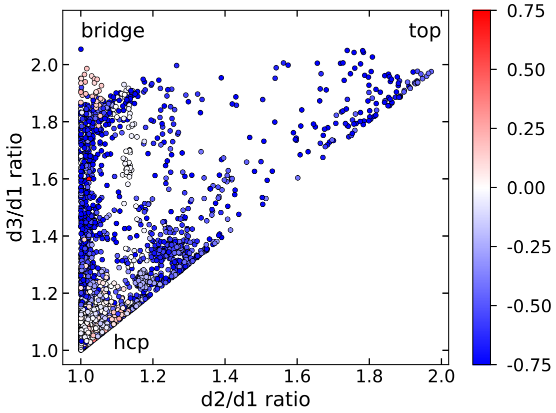


Figure B.1: Scatter plot of all 5457 energies against the ratio of the second and third-shortest H–metal bond to the first. Each point is colored by its adsorption energy (in eV) relative to the pure Ag(211) surface. A positive (negative) value means weaker (stronger) H adsorption.

B.2 DATA MINING

The final positions of the H atom are most often near the center of the hcp hollow, but a significant number of cases relax to off-center, to a bridge site, or to a top site (see Figure B.1).

For dopant positions in the first and second nearest-neighbor shells, Figure B.2 shows adsorption energies for selected elements. Adsorption energies are similar for nearest-neighbor dopants along the step edge and the terrace, with step dopants generally yielding slightly stronger H adsorption. There is a qualitative change in behavior between first and second nearest-neighbor dopants, where elements that strengthen H adsorption as first nearest-neighbors weaken H adsorption as second nearest-neighbors, and *vice versa*. Substitution of the subsurface atom immediately below the hcp hollow yields the least stable H adsorption energies.

We also ran many ($\approx$ 1000) symmetry-equivalent configurations, which allows us to assess the precision of our DFT method. Typically, the difference between adsorption energies for symmetry-equivalent configurations is around 1 meV (standard deviation of 1.5 meV), and only 4 cases have differences larger than 10 meV.

We compare formation energies of the most stable alloys across the d block in Figure B.3. Each

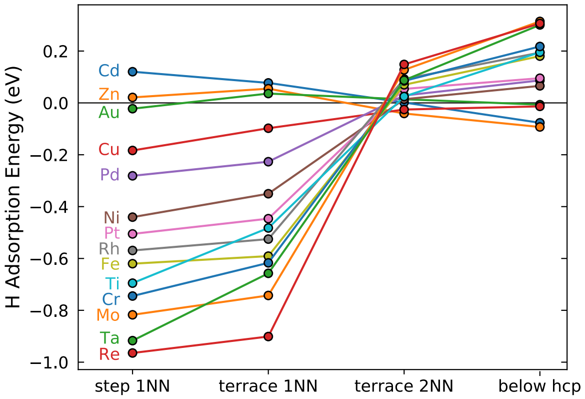
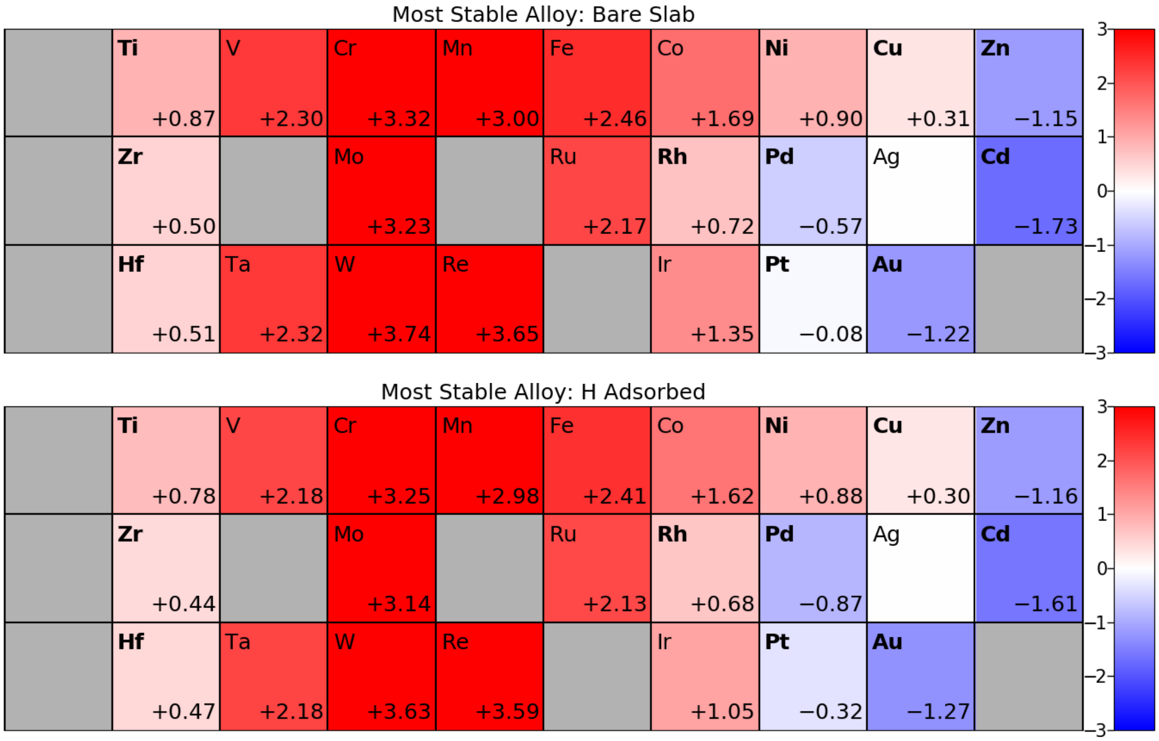


Figure B.2: H Adsorption energies for selected elements for each of the symmetrically inequivalent positions within the first nearest-neighbor (1NN) and second nearest-neighbor (2NN) shells.

plot shows the formation energy of each dopant's most stable arrangement for a bare slab and H-covered slab, highlighting trends across the d block. The most stable arrangements correspond to elements near Ag in the d block, particularly Zn and Cd, while the least stable arrangements are located near Mo. H adsorption does not significantly stabilize unstable dopants; the energetic cost to bring unstable elements to the surface is not sufficiently offset by stronger H adsorption.



Most Stable Alloy: Bare Slab

	Ti +0.87	V +2.30	Cr +3.32	Mn +3.00	Fe +2.46	Co +1.69	**Ni** +0.90	**Cu** +0.31	**Zn** −1.15
	Zr +0.50		Mo +3.23		Ru +2.17	**Rh** +0.72	**Pd** −0.57	Ag	**Cd** −1.73
	Hf +0.51	Ta +2.32	W +3.74	Re +3.65		Ir +1.35	**Pt** −0.08	**Au** −1.22	

Most Stable Alloy: H Adsorbed

	Ti +0.78	V +2.18	Cr +3.25	Mn +2.98	Fe +2.41	Co +1.62	**Ni** +0.88	**Cu** +0.30	**Zn** −1.16
	Zr +0.44		Mo +3.14		Ru +2.13	**Rh** +0.68	**Pd** −0.87	Ag	**Cd** −1.61
	Hf +0.47	Ta +2.18	W +3.63	Re +3.59		Ir +1.05	**Pt** −0.32	**Au** −1.27	

Figure B.3: Color-coded grid plots of thermodynamic stabilities, relative to Ag(211), for the most stable alloy surface for each element. Each tile is labeled with the formation energy in eV, and bolded symbols indicate elements with a minimum formation energy below 1 eV. Top: formation energies of bare slabs. Bottom: formation energies in the presence of a single H atom per unit cell, relative to H adsorbed on Ag(211).

B.3 OUTLIERS

Outliers due to reconstruction and changes in the H atom's adsorption site, the red and blue points in Figure 5.13, respectively, were identified by comparing the geometries of the relaxed H adsorption structure, relaxed slab, and initial positions before relaxation. "Differential reconstruction" points were identified using the simple criterion $|\mathbf{R}_{\text{ads}} - \mathbf{R}_{\text{slab}}| \geq 1.3$ Å, where $\mathbf{R}_{\text{ads}}$ and $\mathbf{R}_{\text{slab}}$ are vectors containing the relaxed metal atom positions for H adsorption and bare slab calculations, respectively. This is equivalent to the root mean square displacement between the two structures. Most adsorption energy calculations for structures exceeding the 1.3 Å threshold are unreliable since the relaxed H adsorption and bare slab structures have different coordinations. For example, in some cases asymmetric dopant atom arrangements lead to the formation of (111) step facets from the (100) facets originally present in the Ag(211) structure.

"H migration" points are identified in a similar way. The influence of H displacements $\Delta\mathbf{R}_H$ between the relaxed and initial structures on errors is anisotropic: the model is more sensitive to displacements along the step direction, $\Delta R_H^{||}$, than displacements perpendicular to the step edge, $\Delta R_H^{\perp}$. This is because the distance from the hcp hollow to an adjacent fcc hollow site along the step edge is approximately 1.4 Å, so displacements $\Delta R_H^{||}$ of order 1.0 Å correspond to H adsorption closer to fcc hollows. In contrast, distances to adjacent fcc hollow sites perpendicular to the step edge are longer so the hcp adsorption model remains valid for larger $\Delta R_H^{\perp}$. Displacement thresholds both along and perpendicular to the step edge were determined by inspecting the locations of the model's largest outliers, rejecting points meeting at least one of the following criteria:

1. $\Delta R_H^{\perp} > 2.3$ Å
2. $\left(\Delta R_H^{\perp} > 0.6 \text{ Å}\right)$ `and` $\left(\left|\Delta R_H^{||}\right| > 1.0 \text{ Å}\right)$

Approximately 2.5% of all data points meet the criteria for either differential reconstruction or H migration.

B.4 DENSITIES OF STATES

To gain insight into the self-passivation of step dimers, we examined the density of states of each case as noted in the main text (see Figure B.4).

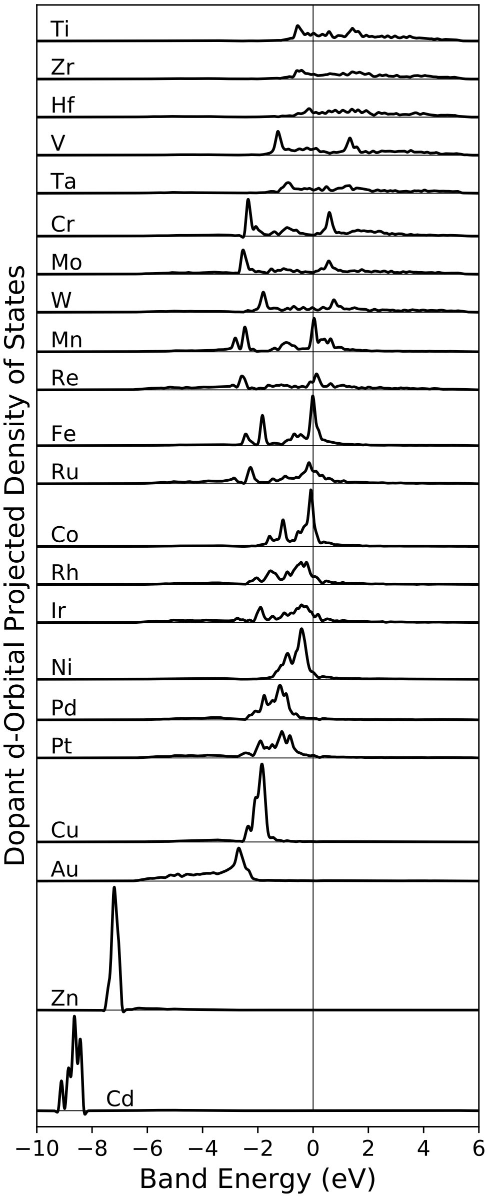


Figure B.4: Projected densities of states onto each dopant's d orbitals for the step dimers.

B.5 LIST OF FEATURES

As summarized in the main text, a wide variety of features were considered for the final forest model. These features can be generally categorized as arrangement features, chemical features, and hybrids of the two.

Arrangement features included the following:

- One-hot bits indicating if each of the 31 atoms within 8 Å of H is occupied by a dopant
- Total number of dopant atoms and the number of nearest-neighbor dopants at the step edge
- Shortest H–dopant and dopant–dopant bond length in the unrelaxed structure
- Standard deviation of dopant atom positions
- One hot bits for each of the 65 clusters present in a scatter plot of the shortest dopant-dopant bond length and standard deviation of metal atom positions
- One-hot bits indicating the existence of the following pairs of dopant atoms:
  - Step dimer with both atoms adjacent to the H atom
  - Any dimer where both dopants are in the surface layer
  - Any dimer where one atom is in the surface layer and the other is in the first subsurface layer
  - Dopants at opposite corners of the (111) step facet adjacent to the H atom

For each element, the following chemical features were included:

- Formation energies of an isolated atom and a dimer in the third layer of the slab
- Adsorption energy difference between an isolated step atom and a step dimer of that element

- Pauling electronegativity
- Covalent radius
- Mass density
- First ionization energy
- Bulk modulus
- Poisson ratio
- Conductivity
- Tight-binding d-orbital energy and radius[52]
- One-hot bits indicating if the bulk crystal structure is hexagonal close-packed or body-centered cubic
- Row and Column of the d block
- Formation energy per atom for the lowest-energy bulk structure relative to an isolated, spin-unpolarized atom

Finally, hybrid features including both arrangement and chemical information were also included:

- Column sums of the Coulomb matrix[149] constructed for atoms within 8 Å of the H atom, using each atom's column in the d block as the effective charge
- Eigenvalues of a simple tight-binding Hamiltonian for the slab geometry, using one effective orbital per atom. The d-orbital energies, radii, and coupling values V_{IJ} are those in Appendix E in Harrison[52].

- The first through fifth moments of the above tight-binding eigenvalues weighted by their projection onto each of the H atom's three nearest neighbors.
- Differences in the d-orbital energy and work function from Ag for each of the H atom's three nearest neighbors

B.6 TOP FEATURES

After applying RFI (see main text) to the remaining 97.5% of the data lacking differential reconstruction or H migration, we obtain the best extra forest RMSE values using the following 25 features:

1. third moment of tight binding eigenvalues projected onto the nearest-neighbor step atom
2. number of first nearest-neighbors
3. column sum of the Coulomb matrix corresponding to the nearest-neighbor terrace atom
4. second moment of tight binding eigenvalues projected onto the nearest-neighbor step atom
5. column sum of the Coulomb matrix corresponding to the atom directly underneath the H atom's hcp hollow site
6. shortest distance between dopant atoms using the relaxed Ag(211) positions
7. dopant element's column in the d block
8. number of second nearest-neighbors
9. formation energy of an isolated atom in the third layer of the Ag(211) structure
10. one-hot bit for the closest atom at the bottom of the step edge
11. number of third nearest-neighbors

12. shortest H–dopant distance using the relaxed Ag(211) positions
13. one-hot bit corresponding to one of the clusters in a scatter plot of shortest H–dopant bond versus standard deviation of dopant atom positions
14. work function of the nearest-neighbor step atom
15. one-hot bit for the second nearest-neighbor atom in the surface layer
16. the second-highest eigenvalue of the tight-binding Hamiltonian
17. column sum of the column matrix corresponding to the closest atom at the bottom of the step edge
18. one-hot bit for the atom directly under the H atom's hcp hollow site
19. dopant element's covalent radius
20. the 34th eigenvalue of the tight-binding Hamiltonian
21. the 25th eigenvalue of the tight-binding Hamiltonian
22. standard deviation of dopant atom positions
23. one-hot bit corresponding to another cluster of H–dopant bond lengths versus the standard deviation of metal atom positions
24. one-hot bit for the subsurface atom under the adjacent hcp hollow site along the step edge
25. dopant element's d-orbital energy