UNIVERSITY OF CALIFORNIA, IRVINE

Data Directed Optimization of Catalysts

DISSERTATION

submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Chemical and Biomolecular Engineering

by

Hunter Noah Pauker

Dissertation Committee:

Assistant Professor Robert “Smith” Nielsen, Chair

Professor Plamen Atanassov

Professor Vy Dong

2024

Dedication

To

My parents, Tony and Kristee,
my brothers, Shane and Grant,
my girlfriend Rory,
my dogs, Kermit and Ferrous,
all my family,
the friends who I met here at UCI,
and the friends I already made.

You all helped me get to where I am today and made me who I am. Thank you.

A quote
It is not ok to just be smart. You must know everything.
(Murray Gell-Mann)

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF SCHEMES

Acknowledgements

I would first like to express my deepest thanks to my committee chair, Professor Robert "Smith" Nielsen. You have given me so much in the years I have been here. We had a rough start: beginning working together in the middle of the COVID-19 pandemic. We persisted through and even through a computer screen you taught me so much about science, and more importantly, how to be a scientist. Without your support, guidance, and encouragement I would not be where I am now. I could not be more thankful to have such a kind, patient, caring, and understanding man leading me in my pursuit for knowledge. It was truly a privilege to work with you, Smith.

I would like to thank my committee members, Professor Plamen Atanassov, and Professor Vy Dong. Your insight has helped me develop my writing to be more easily understood by a wider scientific audience. Thank you for the continual support and guidance you have given me in my time here at UCI. Your time and effort has been invaluable to me.

I would also like to thank Jasmine Gutierrez for the great amount of work she provided for these projects. Thank you for your patience in learning the tools of computational chemistry. I know you're going to be a great doctor someday soon.

This work utilized the infrastructure for high-performance and high-throughput computing, research data storage and analysis, and scientific software tool integration built, operated, and updated by the Research Cyberinfrastructure Center (RCIC) at the University of California, Irvine (UCI). The RCIC provides cluster-based systems, application software, and scalable storage to directly support the UCI research community. https://rcic.uci.edu

VITA

Hunter Noah Pauker

2017-18	Student at Prof. Michael Sailor's Summer School for Silicon Nanochemistry University of California, San Diego
2018-19	Research Assistant, for Prof. William Ristenpart in Pedagogy for Chemical Engineering, University of California, Davis
2019	B.S. in Chemical Engineering, University of California, Davis
2023	M.S. in Chemical Engineering, University of California, Irvine
2024	Ph.D. in Chemical and Biomolecular Engineering, University of California, Irvine

Field of Study

Catalysis and computational chemistry

Presentations

ACS Fall 2023 Presentation. Division: Computers in Chemistry. Session: Data Science for Catalysis: Automated-Synthesis, Process Optimization & Catalyst Discovery. *Data Directed Optimization of Catalysts.*

ABSTRACT OF THE DISSERTATION

Data Directed Optimization of Catalysts

by

Hunter Noah Pauker

Doctor of Philosophy in Chemical and Biomolecular Engineering

University of California, Irvine, 2024

Professor Robert “Smith” Nielsen, Chair

Optimizing matter via computational chemistry promises speed and chemical insight. However, many contemporary attempts oversimplify reaction mechanisms, use crude descriptions of materials, or utilize computational methods that lack rigor to exhaustively screen many compositions. Often these explorations of materials do not relate the properties of materials to one another or chemical concepts, and do not indicate whether an optimum composition has been discovered. To explore the possibility of quantitatively relating catalytic free energy surfaces (FES) to composition, we have computed (using DFT improved by select DLPNO-CCSD(T) calculations) FESs for two candidate reactions, and material descriptors for modular catalyst components. The electrocatalytic reduction of CO_2 to formate was studied using cyclopentadienyl complexes of first-row transition metals. A surprising variety of rate-limiting steps was observed in this family, and predicted H/D KIE values can be used as a rubric to interpret new experiments. pH- and potential-dependent microkinetic models were used to generate turnover frequencies for optimization, and degree-of-rate-control analysis underscores the necessity of including multiple intermediates and transition states in the FES. Principal component analysis was utilized to reveal three degrees of freedom for catalyst optimization and the maximum turnover frequency within the composition space. Methane oxidation was studied

using Ir^{III} complexes and ligands that varied in their σ, π, steric properties and charge. Various regression models related these chemical properties to DFT-derived FESs and predicted the FESs of catalysts not yet simulated. Compositions with beneficial traits were added to the training set, refining the model iteratively. The limitations of this approach were better understood and the combination of ligand properties that minimize the C-H cleavage barrier while preventing unwanted oxidation was identified.

Chapter 1: Formate Generation and Dissociation via CO2 Insertion into Metal Hydrides

Introduction

Global carbon dioxide emissions reached an all-time high in 2024[1], and these emissions are fueling the climate crisis. Non-renewable fossil fuels are driving these emissions up, and there is a greater need than ever to utilize more sustainable, renewable carbon resources. CO_2 has the potential to be one of the most abundant and affordable precursors of carbon based chemicals[2]. The first step to produce carbon-based energy or products from CO_2 is through reduction. Because CO_2 is so thermodynamically stable, finding methods to reduce CO_2 is a challenge that is as difficult as it is appealing. There has been progress made in the electrocatalytic reduction of CO_2 to carbon monoxide, formic acid, formaldehyde, and formate[3]. Formic acid is a useful chemical feed stock, and conversion of CO_2 to formate, and vice versa, has potential to store and release energy in a redox flow battery[4,5]. There is a great excess of CO_2 produced in all facets of industry, and this cheap by-product has the potential to be used as energy storage.

There has been much progress recently in reducing CO_2 through hydride transfer to the formate anion via organometallic homogeneous catalysis. Nitride carbide containing iron clusters in aqueous and mixed solvent have been shown by Berben et al. to reduce CO_2 to formate[6]. Furthermore, they found that this process was tunable to creating formate or hydrogen gas but it was difficult to balance these two processes. Using a $Ru(tpy)(bpy)PF_6$ complex, Hazari et al. have described two distinct forms of CO_2 insertion mechanisms to these organometallic catalysts[7,8]. Similar to our findings Hazari sees first a hydride donation from the metal to the CO_2 to create formate, then a rotation of that formate to be bound via its oxygen atom to the metal. Harazi states that “outersphere” mechanisms are rate limited by the first step, while

"innersphere" mechanisms are limited by the second step. Hazari et al. claim that the rate of this first step increases as the ligand becomes more electron rich or less sterically bulky. The two mechanisms are distinguished by kinetic isotope effects[7], and their rate is dependent on the presence of Lewis acids. After this, Yang and Ceballos utilized a $Pt(dmpe)_2$ complex for CO_2 insertion, designed with thermodynamic principles in mind[9]. This catalyst was found to be reversible, but surprisingly slow. Yang and Ceballos determined that CO_2 insertion or formate dissociation was found to be the rate limiting step.

Through the course of this paper, we will show using the tools of computational chemistry the wide variety of catalytic routes of formate production and dissociation for two families of organometallic catalysts. The first family of catalysts contain an Fe^{II} center, a Cp ligand, and either two monodentate or one bidentate ligand. The final catalyst studied is Ru^{II}(tpy)(bpy), a useful catalyst for benchmarking[10].

Methods

All calculations were performed in ORCA 4.0.2[11,12]. Free energies of molecules in solution were composed as

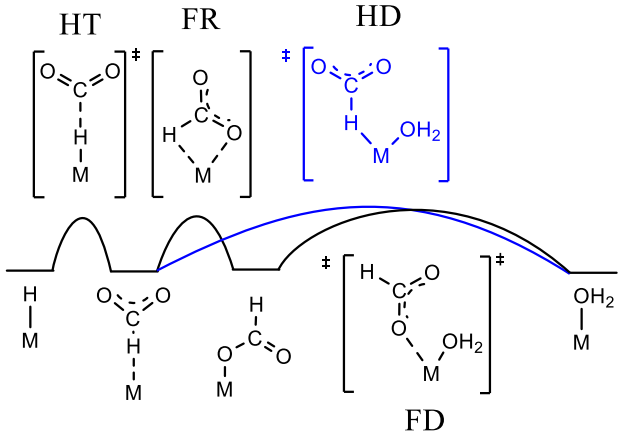
$$G(1M) = G_{solv} + kTln(24.5) + ZPE + PV + H_{vib} + H_{trans/rot} - T(S_{vib} + S_{trans} + S_{rot} + S_{elec}) \quad \textbf{Eqn1.1}$$

Two sets of DFT calculations were performed for each intermediate and transition state: a geometry optimization and hessian calculation in vacuum provided the thermal corrections, then a second geometry optimization was performed with the SMD[13] implicit solvation model with water as the solvent to acquire a more accurate solvated electronic energy. The vacuum calculation was performed with the BP86 GGA functional[14–16] with D3 dispersion corrections[17,18] and the SVP basis set[19]. The solvation calculation was performed with the PBE0-D3 hybrid functional[15–18,20,21], TZVP(-f) basis set, with diffuse functions[19] on oxygen. All

calculations used the Los Alamos pseudopotential and 3Z valence functions for metals[22–24] and Weigend J Auxiliary basis sets[25]. Nudged elastic band (NEB) calculations were performed with ORCAs built in NEB algorithms. The free energy of a proton in water at pH 7 was -279.8 kcal/mol[26], adopting –264.0 kcal/mol as the hydration energy of a proton at 1M[27]. The free energy of an electron at -0.536 V vs SHE was -86.3 kcal/mol, using a 4.28 V difference between SHE and vacuum consistent with the proton hydration energy [28].

Results and Discussion

Several different catalysts had their FES calculated via DFT. After analyzing the FES of these catalysts, a general scheme was developed for these catalysts. The scheme is as follows, first is formate generation, then formate must dissociate from the metal. This general scheme can be seen in Scheme 1.1. It should be noted that this is the broadest form of this scheme, and for some catalysts certain intermediates or transition states may not exist.



Scheme 1.1. Formate Generation and Dissociation Scheme

Previously Hazari et al.[7,8] showed that CO_2 insertion into a metal hydride often takes place via a certain chemical pathway: first the hydride is donated to CO_2 (we call this step hydride transfer), second the formate is bound via its hydrogen to the metal, third the formate rotates to

be bound to the metal via its oxygen (we call this step formate rotation). Hazari et al. coined reactions in which the hydride transfer transition state (HT) are rate limiting are "outersphere", and reactions in which the formate rotation transition state (FR) are rate limiting are "innersphere."[7,8] It was found for the family of catalysts we simulated with DFT that both transition states can be significant, and that there are formidable barriers to dissociate formate which cannot be ignored either.

Through NEB calculations, both HT and FR transition state barriers have been seen as stable for the $Cp(bpy)Fe^{II}$ and Ru(tpy)(bpy) catalyst. After either HT or FR barriers there are also barriers to dissociate the formate from the metal. Post HT transition state the formate may bind to the catalyst via the C-H bond, and post FR transition state the formate may bind via its oxygen to the metal atom. After the HT barrier the formate may then go through the FR transition state, but this would then require the formate to dissociate once bound via its oxygen to the metal atom. Instead, the formate can also dissociate once it is bound via the C-H bond to the metal. These dissociation barriers, HD for hydrogen bound formate dissociation, and FD for oxygen bound formate dissociation, are non-negligible and can be as large as the HT and FR barriers.

The catalyst $Cp(CO)_2Fe^{II}$ was found to not have a stable HT barrier via DFT simulation. This is what Hazari et al. would dub an "innersphere" reaction. These catalysts with a FR type kinetic scheme have a higher FR barrier than the FR barrier seen in catalysts displaying a HT scheme. This is believed to be because the reaction itself is incredibly uphill. Because the FR barrier is so high, it overshadows the HT barrier as the reaction as a whole is so uphill.

The catalyst $Cp(acac)Fe^{II}$ was not found to have a stable FR or HT transition state. This is because this catalyst is so hydridic, but that does not make it a good catalyst. The vacant metal +2 intermediate and the oxygen bound formate intermediate were found to be free energy sinks,

at nearly -27 kcal/mol relative to the starting hydride intermediate, which would make for a very slow TOF. This catalyst exemplifies why an overly hydridic catalyst is counterproductive. Although the rate of hydride donation is very fast, regeneration of the catalyst is incredibly slow. These findings underscore the wide variety of free energy surfaces that a catalyst going through this cycle may go through.

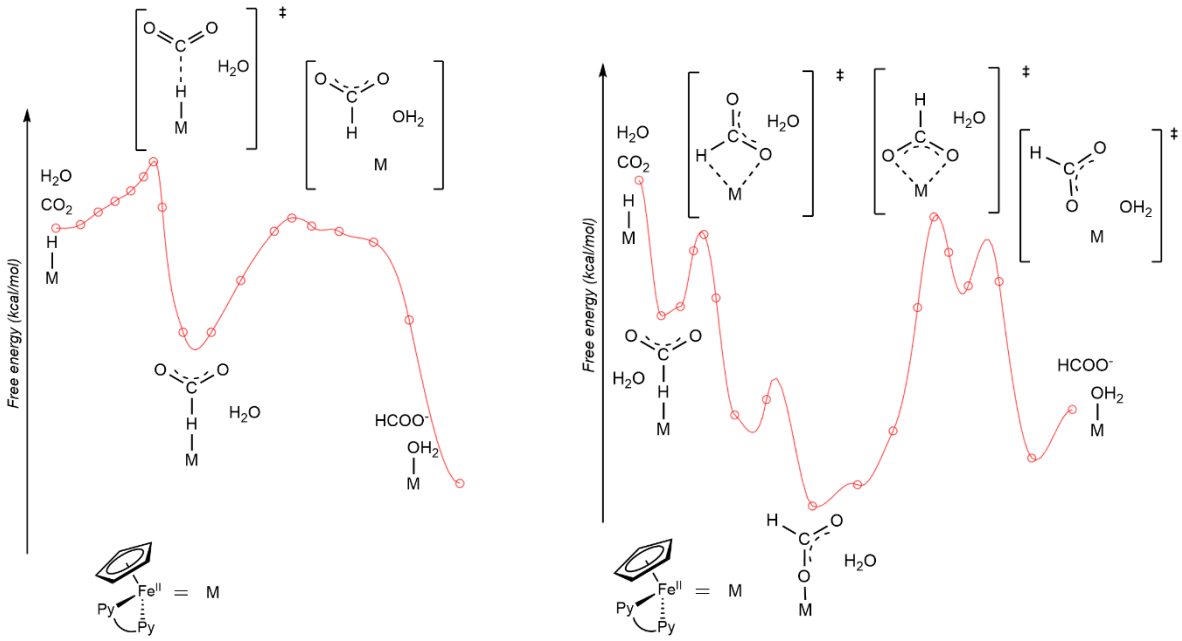


Figure 1.1. NEB Calculations of Cp(bpy)Fe^{II}. This figure illustrates the findings from NEB calculations for the Cp (bpy) Fe^{II} catalyst: a wide variety of transition states and intermediates may occur along the reaction path to free formate. Furthermore, it shows that some intermediates in one path may be present in the other, like the formate bound to the metal through the hydrogen bond.

The kinetic isotope effect (KIE) may be calculated via:

$$KIE=\exp((\Delta G_D-\Delta G_H)/RT), \qquad \textbf{Eqn 1.2}$$

Where D denotes the deuteride version of the catalyst, and H denotes the hydride version of the catalyst. In this calculation, only the thermal corrections were recalculated, and the electronic energy remained unchanged. Additionally in this equation, ΔG may be represented as:

$$\Delta G = G_{TS} - G_{reactant} \quad \textbf{Eqn 1.3}$$

Where G_{TS} is the Gibbs free energy of the transition state, and $G_{reactant}$ is the Gibbs free energy of the reactant immediately prior to the transition state. The specified atomic masses in these calculations were 1.008 AU for hydrogen and 2.014 AU for deuterium.

Each of these barriers discussed thus far also has a unique kinetic isotope effect (KIE) associated with it. The KIE may be used to differentiate between either barrier. Both HT and FD display inverse KIE while HD and FR display KIE. It should be noted that there is known error for DFT, on the order of several kcal/mol, which correlates to near 25% error in KIE. While this error is large, and non negligible, the importance of these DFT calculated KIE is to differentiate between KIE and inverse KIE. It should also be noted that the first line of FD barriers seen in Table 1.1 are relative to the oxygen bound formate intermediate, while all other barriers are relative to the $M^{n+2}H$ starting intermediate hydride. An additional line is added that shows the FD barrier relative to the starting intermediate denoted with an asterisk.

	DFT				Experimental	
	$Cp(CO)_2 Mn^I$	$Cp(CO)_2 Fe^{II}$	$Cp(bpy) Fe^{II}$	$Ru^{II}(tpy)(bpy)$	$[Ru(tpy)(bpy)D]PF_6$*	Ru(tpy)(bpy)
HT	1.05		1.14		CO_2 Insertion[7]:	Decarboxylation[29]:
HD	0.57	0.79	0.67	0.69	0.86±0.09	1.15±0.1
FR	0.58	0.74	0.59	0.75		
FD	1.07	1.16	1.02	0.97		
FD*	0.50	0.49	0.46	0.48		

Table 1.1. Kinetic Isotope Effect with Experimental Results from Hazari et al.[7] and Creutz et al[29].

In addition to these catalysts the FES was calculated via DFT for an additional catalyst, Ru(tpy)(bpy), the same as that utilized by Creutz et al[29], but distinct from the hexaflourophosphine catalyst utilized by Hazari et al.[7], this catalyst was found to have several transition state barriers during formate generation and dissociation. The first barrier, found via NEB calculation, was an unidentified transition state presumed to be hydride transfer. This barrier was not confirmed via normal DFT transition state geometry optimization. The barriers that were confirmed via DFT transition state geometry optimization were FR, HD, and FD barriers. Additionally the DFT calculated hydricity of this complex was found to be 38.3 and 17.8 kcal/mol for acetonitrile and water respectively[30]. This is comparable to the experimental values obtained by Creutz et al. of 39±3 and 22 kcal/mol[10].

The larger difference seen in DFT vs experimental between the hydricity in water vs acetonitrile, ~4 and ~1 kcal/mol respectively, is understandable. Changes in solvent quite drastically change hydricity[30]. The accuracy of the calculations compared to experimental work is practically the same if not better when the solvent is changed from water to other polar solvents.

Finally as seen in Table 1.1, although no KIE is calculated for a HT for the Ru complex, it would make sense that the transition state seen by Hazari et al[7]. is similar to our FR transition state. Furthermore, it would make sense from our results that the decarboxylation transition state seen by Creutz et al.[29] is similar to the FD transition state. It should be noted that although this FD transition state is formate leaving the oxygen bound formate intermediate, like that described

by Creutz et al, this barrier is not the same as in our FD transition state formate displaced by water. As mentioned before, there is inherent error in DFT especially with transition state metals, so several kcal/mol difference is to be expected. The trends of these transition states and intermediates are still quite clear despite these several kcal/mol differences. Because DFT calculations have a several kcal/mol error the KIE has a relatively larger error, however DFT calculations are still able to show the general trend of formate rotation displaying inverse KIE.

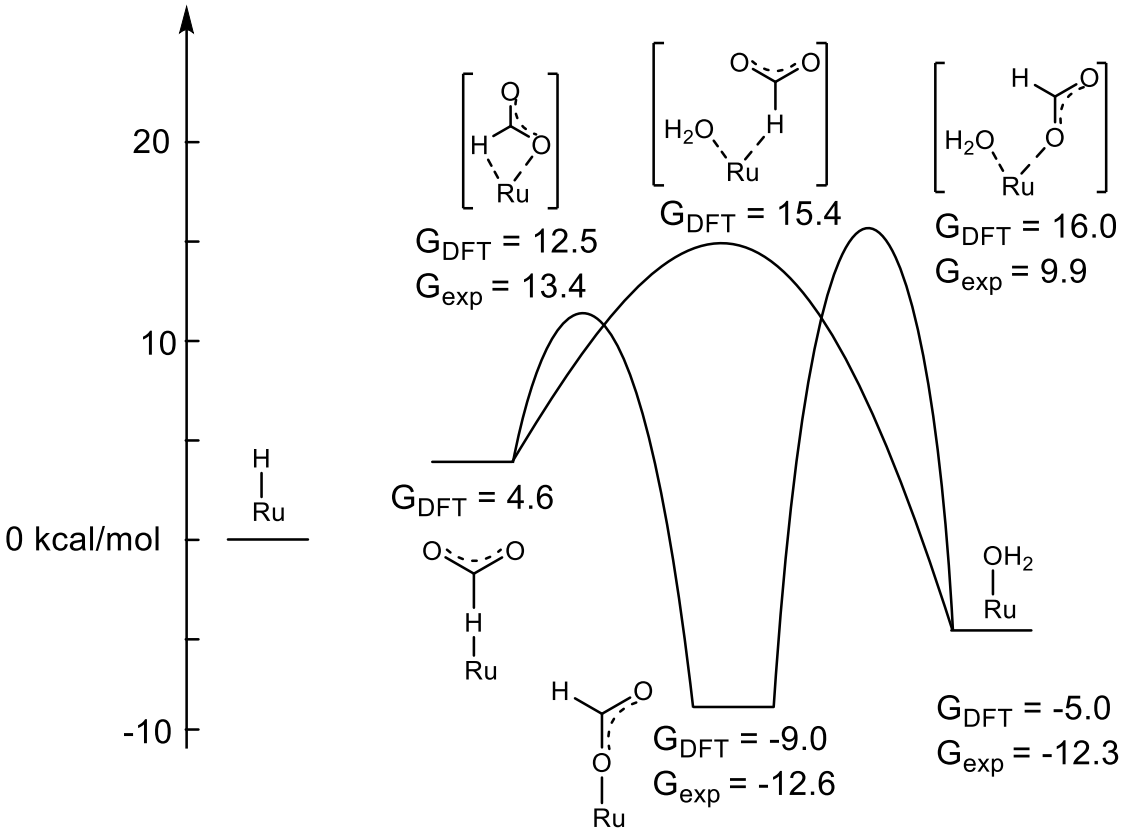


Figure 1.2. Ruthenium Free Energy Surfaces. Experimental data from Creutz et al[31].

The energies of transition states and intermediates were relatively similar to those reported by Hazari et al[7]. usually only having a several kcal/mol difference. Although the relative Gibbs free energies seen in Figure 1.2 are quite different between experimental and calculated, especially for the dissociation reaction, this is because they are compared to the $M^{n+2}H$ starting

intermediate. The overall TS energies compared are relatively similar between DFT and experimental.

Conclusion

Through calculating and analyzing the DFT generated FESs of CO_2 insertion and formate dissociation in this family of catalysts, several conclusions have been drawn. A wide variety of FES can appear on this family of catalysts, more than seen by Hazari et al. and their Ru(tpy)(bpy)PF_6 catalyst. This family of Cp ligand bearing catalysts may have a transition state barrier to donate the hydride to CO_2, dubbed the hydride transfer barrier (HT), and a barrier for the formate to rotate to become bound to metal, dubbed formate rotation (FR). After either barrier the formate may stick around or bind to the metal, and there is a dissociation barrier that must be overcome to free the formate from the metal through its hydrogen or its oxygen, dubbed hydride dissociation (HD) and formate dissociation (FD) respectively. All these barriers may be differentiated by kinetic isotope effect. Furthermore, it was found that catalysts may display both HT and FR barriers, or only a FR barrier.

Chapter 2: Linear Free Energy Relationships of Formate Generation Among First Row Transition Metal Hydride Catalysts

Introduction

Computational chemistry has been utilized for many years to better understand catalyst behavior. With computational costs decreasing and improvements in computational accuracy increasing, computational chemistry is a more appealing tool than ever to analyze catalyst behavior. Computational chemistry allows for large generation of data sets with relatively cheaper costs and time compared to performing experiments in a laboratory setting.

Catalyst families are known to have linear free energy relationships (LFERs)[30,32]. Among related compositions, the energy of two elementary steps is often found to correlate linearly or at least follow a predictable relationship. Understanding these relationships is instrumental in conceiving novel catalysts with desired properties[33]. The subject of this text is to develop mathematical tools for deliberately uncovering and quantifying LFERs among the steps of catalytic reactions. We present a case study by illustrating a process for discovering these relationships using the electrochemical reduction of CO_2 to formate. There has been precedent in the study of LFERs with hydride transfer. Dubois et al. mapped free energy relationships between hydride transfer and one electron reduction potentials of molecular catalysts through years of electrochemical experimental work[32,34,35]. Appel et al. described how the TOF correlated with the hydricity of Cobalt catalysts[36].

Do and Ngo studied the structure-activity relationship for a group of Cp^*Ir based catalysts performing ketone hydrogenation, finding that there is more than just hydride transfer that is influential to TOF in this family of catalysts[37]. Studies that elucidate LFERs in hydricity are

important but relationships among structural elements of catalyst composition and all steps of the reaction must be described to make use of these relationships.

When properties or components of the catalyst change so does the FES, and changing one intermediate or transition state energy will inherently change another. This means that within a family of catalysts modifying those catalysts to a "perfect" flat FES may not be possible. Like the Sabatier principle, a balance must be found in all barriers and intermediates in finding the smoothest, flattest FES.

A great challenge is how to discover and describe these LFERs quantitatively. The LFERs dictate constraints on the catalytic FES, and therefore limit catalytic performance. These relationships must be understood to optimize the FES for stability, activity, and selectivity. Furthermore, understanding these relationships may teach chemical lessons. Learning these chemical lessons can aid in tuning these catalysts and can also be useful in other applications.

Principal component analysis (PCA) is a method of analyzing multidimensional data in the form of an array[38]. New axes are constructed that are orthogonal linear combinations of the previous dimensions of the array. These PC dimensions lie in real space in the direction of the most variance of the original array. PCA eliminates linearly dependent dimensions and tells the user how many distinct linear dimensions can be used to describe the system. PCA is an incredibly useful tool in turning the FES in real space into a more legible form. By examining the domain of PC space one can realize the range of real space.

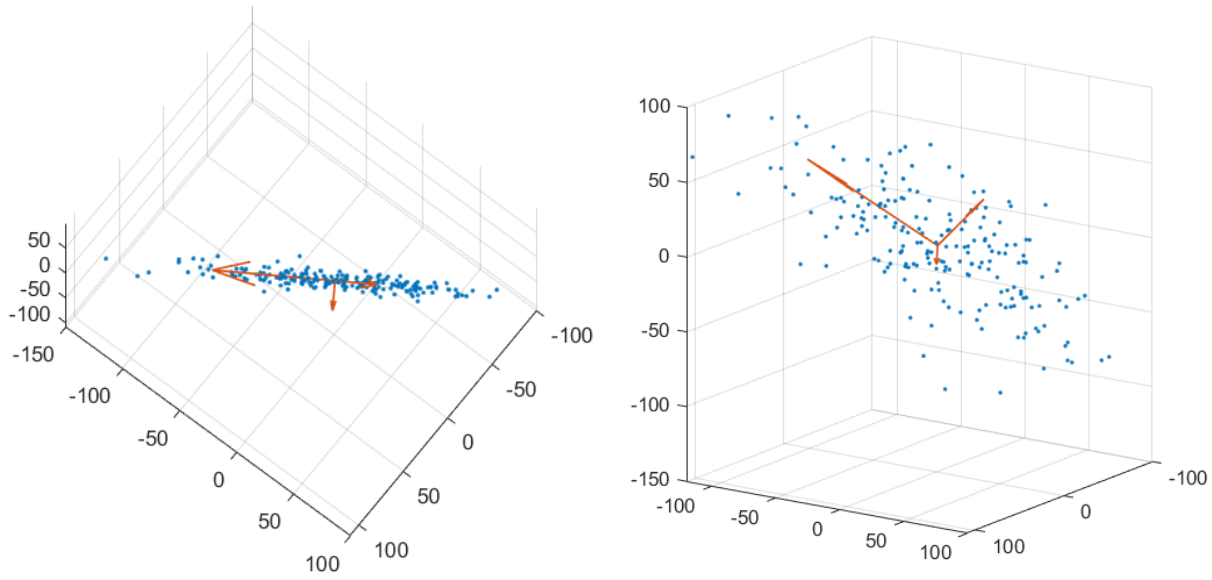


Figure 2.1. PCA Example Illustration. The PCA axes (red arrows) illustrate how principal component constructed axes lie in real space relative to the data points (blue circles).

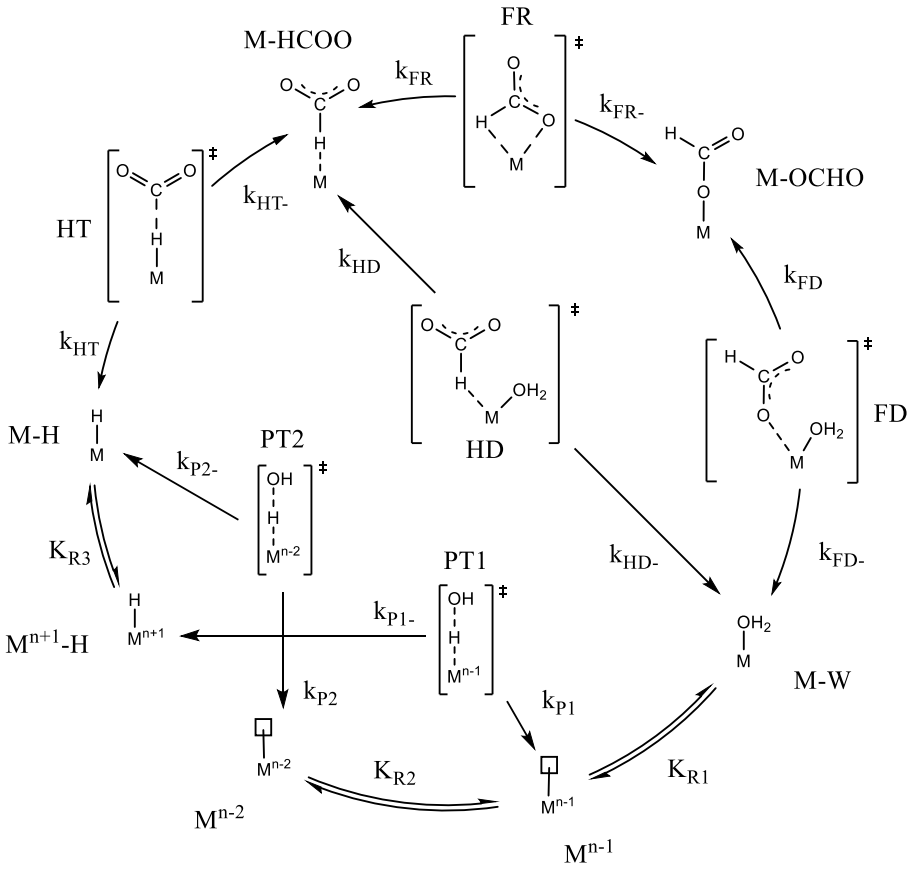
The 200 points in Figure 2.1 do not show actual chemical data and are just dummy values used to illustrate the power of PCA. Each data point is blue, and each principal component axis is a red arrow. Each arrow is proportional in size to the amount of variance said axis describes. Seeing these data points from two different perspectives makes it clear that although this data exists in three dimensions it essentially lies in a plane. The first two principal components describe this plane, and the third shows the relatively small spread in and out of this plane. The first principal component describes the majority of the variance or spread of this plane, and the second axis describes the second most amount of spread perpendicular to the first.

PCA will be used throughout the course of this paper to find quantitative measures of how far a catalyst can be changed along a certain dimension using the compositions in this study. A sort of principal component volcano pot is made that essentially allows one to navigate the reduced space to find a realistic optimum more easily. Furthermore, if dimensions can be interpreted physically, chemical lessons can be learned.

Results and Discussion

Great progress has been made recently in reducing CO_2 to formate via hydride transfer with organometallic catalysts, making it an ideal candidate to study LFERs. Berben et al. have reduced CO_2 to formate with nitride carbide containing iron clusters[6]. Hazari et al have described two distinct transition states and mechanisms during CO_2 insertion to Ru(tpy)(bpy)PF_6[7] and (^{tBu}PCP)NiH complexes[8]. The general catalytic cycle may be seen in Scheme 2.1. This is the most general catalytic cycle, as will be discussed in this text, other cycles exist which exclude certain transition states or intermediates.

The first transition state is the hydride being transferred from the metal to CO_2, dubbed hydride transfer (HT) transition state. The second transition state is the formate anion rotating from being bound via the hydrogen to being bound via the oxygen, dubbed formate rotation (FR) transition state. Utilizing DFT calculations with a different family of catalysts, with Fe^{II} center, two monodentate or one bidentate ligand, and a Cp ligand, we have confirmed the presence of these two transition states, and that these two transition states may happen one after the other (HT then FR). We have also determined that sometimes a catalyst does not have a stable HT barrier, and a FR barrier is only present. In cases like these, the FR barrier is often higher than catalysts where a HT barrier is stable. This is believed to be because the FR barrier overshadows the HT barrier. In addition to these barriers, DFT calculations have also confirmed the presence of nontrivial formate dissociation barriers. Once through the HT barrier, formate is bound via the hydrogen to the metal, and can dissociate via a hydrogen bound formate dissociation (HD) or can rotate through the FR barrier. Once bound via oxygen, formate may dissociate via the oxygen bound formate dissociation barrier (FD). These barriers are non-trivial and may be on the order of the HT and FR barriers.



Scheme 2.1. General Catalytic Cycle. This catalytic cycle is the most general. One other major catalytic cycle is seen for the family of catalysts discussed in this text. That cycle excludes the HT and HD transition states, and M-HCOO intermediate.

Post formate generation and dissociation, the catalyst must be regenerated. Utilizing DFT calculations we have also found that this region of chemical space can vary widely. The first step of regeneration was always found to be reduction. After this is either a protonation then a second reduction, or vice versa. The protonation barrier was assumed to be the same height for either protonation.

Turnover frequency (TOF) is an excellent way to quantify catalyst performance, and is a function of the conditions (potential, pH, and temperature), concentration of products and reactants, and the FES itself. With an established FES, as seen in Scheme 2.1, a microkinetic

model may be built to calculate TOF for each catalyst. Two models in fact can be built: one model assumes that a catalyst can go through the HT and HD barriers, and the other assumes the lack of HT and HD barriers. Utilizing two models, rather than one, more accurately quantifies catalyst performance. Utilizing the equilibria and rate limiting equations built off of the rates seen in Scheme 1, each microkinetic equilibria was built with the following assumptions: all intermediates are conserved, the concentration of CO_2 is constant, and both formate and formic acid are conserved in equilibrium. With these assumptions, balance equations around formate bound species (M-OCHO and M-HCOO for HT scheme and M-OCHO for FR scheme) and metal hydride species can be created which assumes these concentrations are constant in time. By substituting equilibria and rate constants in terms of Gibbs free energies, the concentrations may be solved in these balance and conservation equations. These concentrations, now as functions of Gibbs free energies and conditions, may be utilized to create the TOF as a function of the FES and conditions of the system.

Two distinct catalytic schemes were created: one with a HD and HT barriers, and one without these barriers, dubbed the HT and FR schemes respectively. Two distinct schemes means that two distinct TOFs, a TOF^{HT} and TOF^{FR}, can be utilized to quantify catalyst performance. Rather than using human intuition to decide which TOF should be used to quantify catalyst performance, a support vector machine (SVM) was trained and used to decide which TOF a catalyst should use. This SVM was trained on the partial FESs (FR, M-OCHO, FD, M^{n+2}, M-W, M^{n+1}, M^{n}, and $M^{n+3}H$ states) of the first seven catalysts as listed in Figure 2.4.

Some catalysts greatly prefer one regeneration pathway over the other, and for some catalysts a protonation before a second reduction or vice versa may be completely unstable. The variety in these regeneration steps means that the FES, and therefore the turnover frequency (TOF), is

greatly dependent on the conditions of the catalyst. The effect of pH and potential on TOF can be seen for the Cp(bpy)Fe^{II} catalyst in Figure 2.1. This figure displays absolute TOF, which is why there is a line with a negative slope down the figure, this line is when the reaction flips and runs in reverse. The bottom half of the figure is the forward reaction, and the top half is the backward reaction.

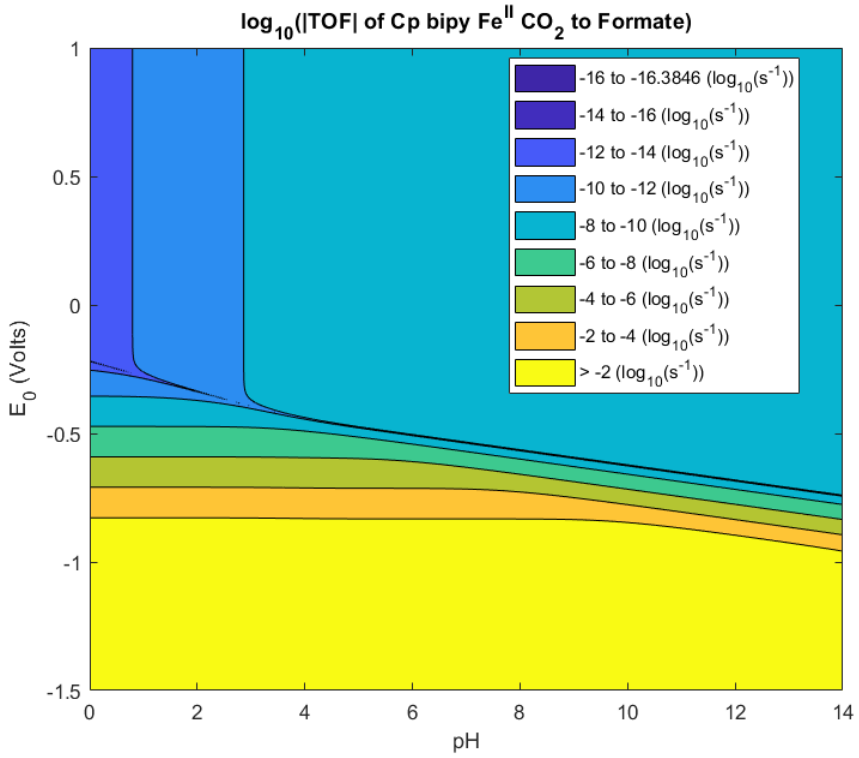


Figure 2.1. Absolute TOF for Cp(bpy)Fe^{II} catalyst. The downward sloping line through the middle divides the bottom half as the forward reaction (CO_2 to formate), and top half as the backwards reaction (formate to CO_2).

Imputation

Many more FESs were generated in this Cp family, now utilizing other metals in period 4, manganese and cobalt, and various other ligands. These ligands include but are not limited to those seen in Figure 2.2.

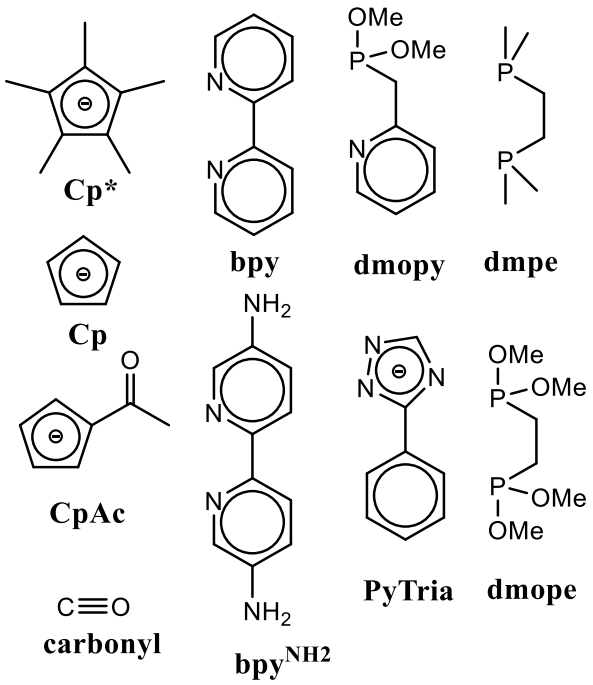


Figure 2.2. Ligand Family. The ligands seen in this figure are all ligands utilized in mapping the LFERs of this family of catalysts. All catalysts utilized one of the three Cp ligands, and one bidentate ligand or two carbonyl ligands.

With more FESs for catalysts in this family LFERs can be elucidated. Principal component analysis is an excellent way to analyze these LFERs. However, to perform PCA on these FESs, holes in these FESs need to be addressed. As stated previously, for some catalysts some intermediates or transition states are completely unstable, therefore the Gibbs free energy for that state may not be found for that catalyst. These holes may be filled in with PCA via imputation[39], a method of matrix completion which relies on PCA. The FES is arranged in an array, with the rows corresponding to the catalyst, and the columns corresponding to the intermediate or transition state. Each hole is filled with the average of the column, or intermediate free energy, then PCA is performed on the FES. From this value, a score and coefficient matrix is produced. The scores, or PC values, can be converted back to a FES by multiplying the score matrix by the

coefficient matrix, which produces a slightly different FES with some errors. The original FES is kept, any values that were originally holes are replaced with the score predicted values. This process is repeated until the holes converge on a value: PCA is performed on the new FES, a score is acquired, the score is turned back into a FES, and the original FES remains with only the holes replaced with the score predicted values.

$Cp(CO)_2Mn^I$	$Cp(bpy)Fe^{II}$	$Cp(dmpe)Fe^{II}$	$Cp^*(CO)_2Fe^{II}$	$Cp(CO)_2Fe^{II}$		Isodesmic Corr.
17.8	18.1	18.8	*23.5*	*24.0*	HT	4.2
14.9	8.1	13.5	36.0	36.2	FR	4.3
-2.6	-10.7	0.6	7.9	8.2	M-OCHO	3.6
19.0	12.0	26.7	36.6	29.8	FD	7.0
12.0	-7.5	2.0	34.1	37.7	M^{n+2}	2.6
-5.4	-11.7	-5.1	11.1	13.2	M-W	0.9
17.3	1.2	20.3	16.5	12.8	M^{n+1}	9.7
41.3	23.9	50.9	17.0	8.4	M^n	-5.8
0.2	-6.2	-5.8	28.1	32.0	$M^{n+3}H$	-0.9
23.4	15.2	*26.0*	*45.7*	44.6	HD	9.7
27.3	9.0	33.7	*34.4*	*30.7*	PT	9.0

$Cp(bpy^{NH2})Fe^{II}$	$CpAc(dmope)Mn^{I}$	$Cp(dmopy)Fe^{II}$	$Cp(pyTria)Co^{III}$	$Cp(dmope)Fe^{II}$	$Cp(bpy)Co^{III}$	$CpAc(dmpe)Fe^{II}$
16.4	*17.0*	20.6	26.0	25.8	*24.8*	20.8
8.4	*14.3*	20.0	34.6	31.6	36.2	18.0
-9.7	-2.8	-4.9	6.0	10.9	8.3	-0.9
10.9	20.0	19.2	*30.4*	38.3	38.9	27.8
-9.8	6.0	11.4	25.8	30.9	31.5	6.8
-13.0	-8.2	-4.9	9.3	9.1	13.2	-3.2
6.3	20.0	16.3	-2.8	21.0	-1.8	15.9
31.2	*49.7*	*37.8*	6.9	34.7	-1.7	43.5
-7.9	-17.5	-5.9	36.2	9.7	44.6	-1.4
14.7	22.5	27.0	40.0	41.4	42.9	29.0
13.1	*29.7*	*27.2*	*22.4*	37.8	*23.6*	30.8

PCA Predicted Best Catalyst	Cp*(dmpe)Fe^{II}	CpAc(dmope)Fe^{II}	CpAc(dmpe)Mn^{I}
19.0	14.9	20.7	15.5
14.5	13.1	36.7	-3.0
-8.7	0.8	9.7	-12.9
9.4	*19.9*	38.9	6.6
8.8	5.3	28.7	-9.1
-7.1	-6.5	8.7	-19.9
7.8	21.4	18.4	24.3
21.2	56.3	30.5	64.9
3.1	-15.5	12.4	-31.8
20.1	*23.4*	42.1	5.7
14.4	29.0	39.2	*23.7*

Table 2.1. Gibbs Free Energy Surface of Catalytic Cycle (kcal/mol). Additional values of all catalysts can be found in the SI. Imputed values are in italics. The best calculated FES was found through PCA composition space navigation.

The whole FES data may be seen in Table 2.1, with the imputed values in italics. It is important to note that when the holes are predicted from the scores, not all the possible PCA are utilized to predict these values. Most of the variance in data may be captured in only a few principal component axes. In this case only two principal components were utilized to predict the holes, as 93.2% of variance is captured within the first three principal component axes, and utilizing more principal components gave unrealistic values for certain intermediates. Utilizing

more than three principal components may lead to inaccurate imputation, skewing TOFs, as utilizing principal components that capture such small amounts of variance may just contribute to noise.

It should be noted that there is error with imputation, and that imputation can affect higher order (latter) PCA scores, and interpretation of those scores. As stated, DFT has inherent error, on the order of several kcal/mol, and imputation can have error on that lever or even larger. It is important to examine how imputation fills out data and use intuition to address poorly imputed data.

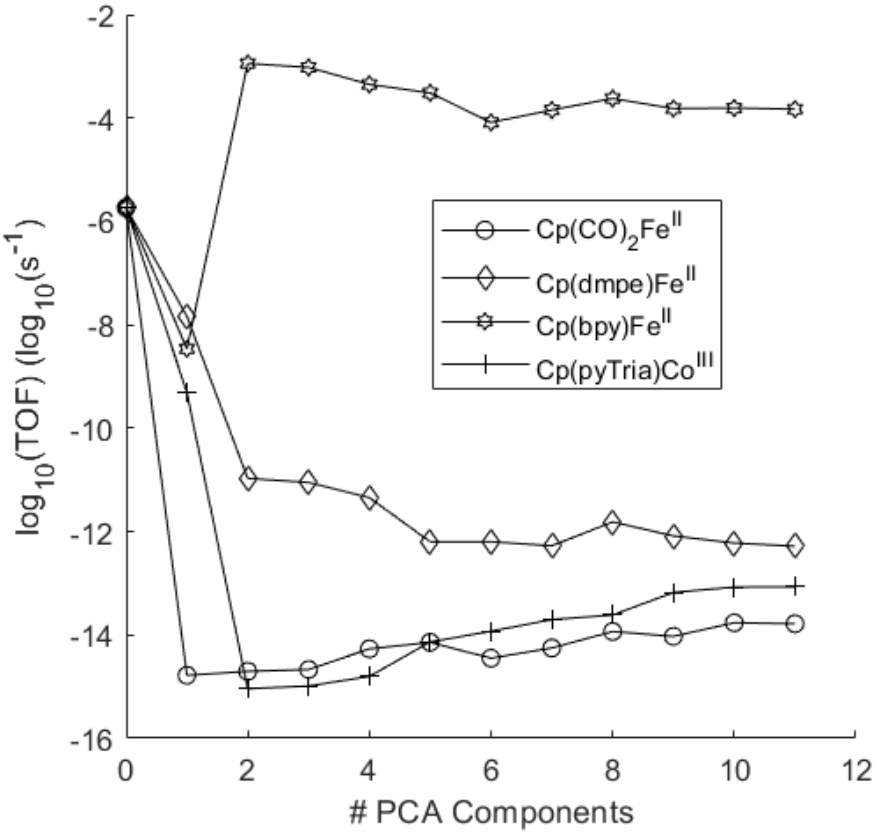


Figure 2.3. Number of Principal Components Effect on TOF. As the number of principal components increases, the accuracy of the predicted TOF increases, however with only 3 principal components the TOFs nearly converge.

As seen in Figure 2.3, the number of PC utilized to predict a FES greatly impacts the TOF. In this figure, the use of zero PCs corresponds to the FES being predicted for all catalysts as the average across all 15 catalysts. Some intermediates or transition states of the FES greatly impact the TOF by orders of magnitude when said energies fluctuate by only a few kcal/mol. Furthermore, it can be seen from this figure, that the TOFs quickly converge by 3 PCs. When too few PCs are utilized the chemical behavior of the LFERs are not accurately captured, however with more and more PCs utilized the behavior of these PC is captured but noise captured by latter PCs is introduced. This noise can lead to errors when imputing or predicting catalysts outside of the data set. A balance must be found in accurate predictions with minimal error.

As seen through imputation, principal component analysis can be incredibly useful because score space can easily be turned into real space, and vice versa. With scores imputed, PCA performed on all the FESs resulted in 98.7% of the variance being captured in the first three principal components. This means that for this family of catalysts there are essentially only three variables that can be changed to alter the FES of the catalyst. Finding the optimal catalyst along these three PC axes is much simpler than exploring a catalyst system with 11 dimensions.

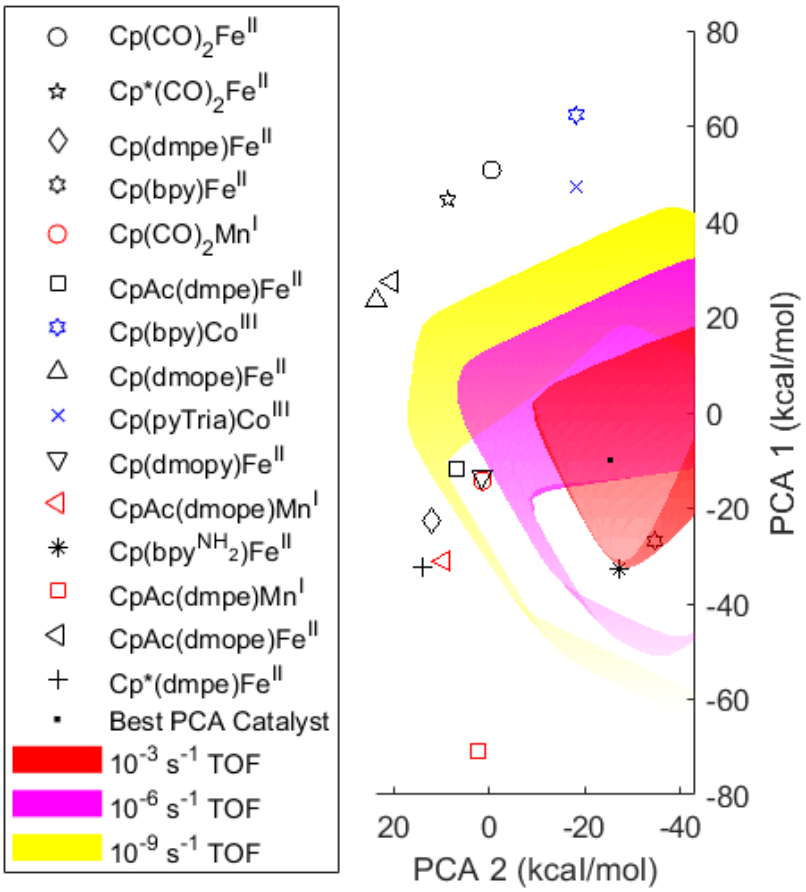


Figure 2.4. TOF Heat Map in PC Space. This heat map of TOF was created by creating a 50x50x50 grid spanning 2.5 times the standard deviation of scores of the 15 catalysts. Each point in PC space can be converted to a FES, and each FES can be converted to a TOF. Visualizing the LFERs in this way allows for ease of navigation in composition space, and realization of the maximum performance capabilities in this family of catalysts.

To navigate this space, a map first needs to be made, as seen in Figure 2.4. This heat map corresponds to the TOF of the PCA space, and where known catalysts lie in this space. A 50-by-50-by-50 evenly spaced grid was constructed that spanned roughly the maximum and minimum of the first three scores of this family of catalysts: two and half times the standard deviation of scores of each PC axis. It was assumed that all scores, besides the first three principal

components seen in this map, were zero. Of any catalyst in this family chosen at random zero is the most likely score, as it is the average across all principal component scores. Each point in this map was turned into a FES, and each FES was turned into a TOF. A transparent color heat map was then constructed of these catalysts corresponding to the $\log_{10}$(TOF).

Within this heat map the PCA predicted the best theoretical catalyst can also be found. This theoretical catalyst doesn't really exist, and the composition is unknown, it is simply the PCA point in space that corresponds to the fastest TOF. This TOF is 0.0360 s^{-1}, compared to DFT calculated fastest TOF, $Cp(bpy^{NH2})Fe^{II}$, ([2,2'-Bipyridine]-5,5'-diamine), of 0.00079 s^{-1}, is quite high, however, the change in PCA components to get from the best DFT calculated score to the best PCA predicted score is reasonable for this change in magnitude of TOFs. More than being the best predicted PCA point in space, this value of the best TOF is more indicative of the limits of the current family of catalysts.

Physical Meaning of Principal Components

Possibly the most useful aspect of PCA is interpreting what each principal component axis physically means. Making chemical sense of these axes can provide chemical lessons that can be applied to other scenarios, as well as better improve the performance of catalysts in the given family.

It is important to realize that when interpreting scores, the signal gets smaller with the later scores. This is because the first principal component captures the most variance across the data, and each axis thereafter captures less and less variance in the data. With a smaller magnitude the variance of these principal component values can be nearly the same order of magnitude as the error. All this is to say that the coefficient matrix, the operator that turns real world data (FES in this case) into scores, can be just as useful if not more useful than examining the scores. The

weights on the coefficient matrix are always nearly the same order of magnitude, regardless of the amount of variance in data. These weights can be telling of the largest influence on each principal component axis.

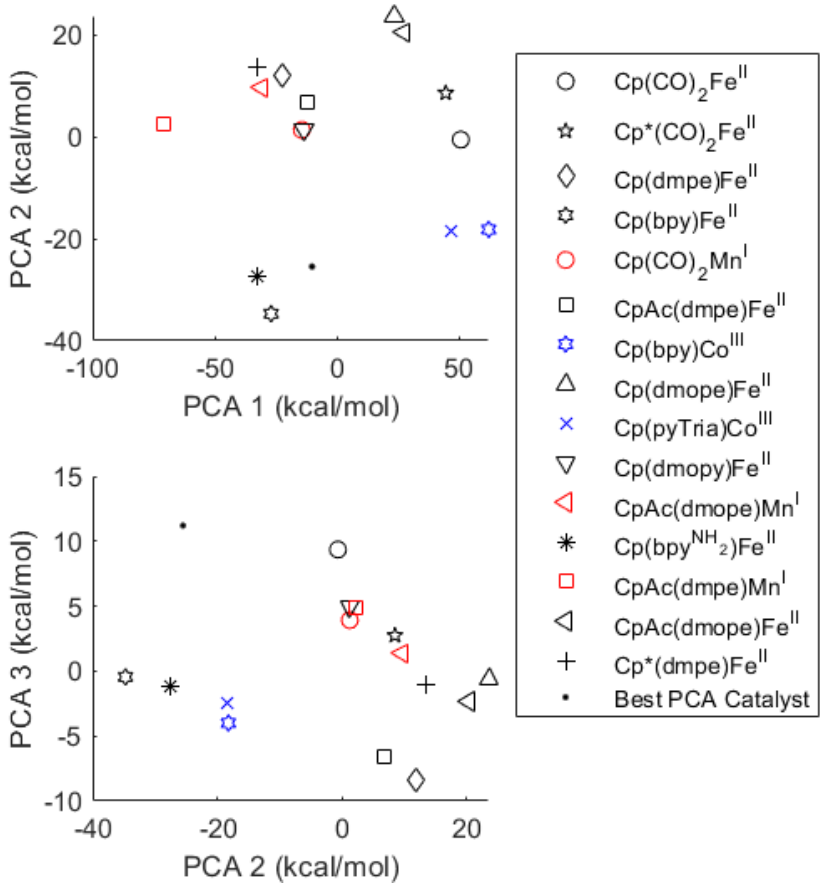


Figure 2.5. Catalysts in Principal Component Space. Viewing the catalysts in PC space is a powerful tool to qualitatively distinguish catalyst properties and discern the physical meaning of the PC axes.

Looking at both the scores as seen in Figure 2.5, and coefficient matrix, as seen in Table 2.2, one can interpret the physical meaning of the largest principal component axes. The first principal component, accounting for 82.0% of variance has to do with electrophilicity and is possibly influenced by charge. It is affected by all three components of composition of the catalyst, the Cp ring, the metal, and the L2 ligand. A large positive coefficient, see Table 2.2, is

seen in the coefficient for any intermediate where electron density is removed from either a pi or sigma orbital, like in the vacancy, water, and oxidized hydride intermediates.

The second principal component, accounting for 17.2% variance, appears to be influenced by the reduced intermediates, like the M^{n+1} and M^{n} states. This reducibility may be achieved by the metal or the ligand. This is evident through the bpy and pytria ligand containing catalysts having extreme values for the second score as seen in Figure 2.5. The cobalt compounds are extremely electrophilic which is how they achieve high reducibility. This score is helpful in interpreting why the (bpy)Fe complexes have the best performance of any catalyst. Although iron is hard to reduce it more easily allows for hydride transfer as it is not so electrophilic. Catalysts that contain bpy are easy to reduce because the ligand itself is reducible, which makes regeneration of the catalyst much easier.

The third score, accounting for 5.9% of the variance, is difficult to comprehend. It has the highest magnitude coefficients for the FD transition state and the vacancy M^{n+2} metal. When interpreting Figure 2.5, changes in composition are confusing. Changing one constituent on a catalyst may have changing the Cp to Cp* in $Cp(dmpe)Fe^{II}$ (black diamond to black plus) changes PC3 by +7.4 kcal/mol, while the same change with $Cp(CO)_2Fe^{II}$ changes PC3 by -6.6 kcal/mol. On the other hand changing from Fe to Mn with $Cp(CO)_2Fe^{II}$, $CpAc(dmope)Fe^{II}$, and $CpAc(dmpe)Fe^{II}$ changes PC3 by -5.4, +3.7, and +11.4 kcal/mol respectively. The latent value of PC3 is 4.6 kcal/mol as seen in Table 2.2. This is larger than we expect of nonsystematic DFT errors. This is evidence that changes in composition are not additive in PC3 so the relationship between composition and free energies in not linear. Because the FD barrier (negative contribution to PC3) is seven coordinate and the vacant intermediates (positive contribution to

PC3) are essentially five coordinate, PC3 appears to be related to steric crowing around the metal.

	PC1	PC2	PC3	PC4
HT	0.09	0.01	-0.03	0.00
FR	0.32	0.15	0.01	-0.45
M-OCHO	0.17	0.20	-0.21	-0.02
FD	0.23	0.25	-0.61	-0.16
M^{n+2}	0.39	0.31	0.61	0.06
M-W	0.27	0.13	-0.03	-0.04
M^{n+1}	-0.10	0.44	0.36	0.13
M^{n}	-0.43	0.53	-0.23	0.37
$M^{n+3}H$	0.54	-0.21	-0.15	0.74
HD	0.30	0.21	-0.09	-0.22
PT	0.07	0.45	-0.08	0.10
% Explained Variance	82.0	15.6	1.1	0.6
Latent (kcal/mol)	39.5	17.2	4.6	3.4

Table 2.2. Coefficient Matrix Principal Components 1-4.

Degree of rate control

The degree of rate control (DRC) initially described by Campbell et al.[40,41] is a method of computational catalyst screening. This technique is performed by first developing a microkinetic model of a catalyst reaction path. Once a catalysts FES is known the TOF can be calculated via the microkinetic model. The next step is to use simple numerical methods to approximate the change in TOF with respect to the change in each intermediate Gibbs free energy along the catalyst reaction path[40]. For a FES with n total intermediates and transition states, m, and a small

numerical perturbation, delta, the change in TOF with respect to changing intermediate m_i can be written as,

$$\frac{d\mathrm{TOF}}{dm_i} = -RT\,\frac{\log_{10}\mathrm{TOF}(m_1,\ldots,m_i+\delta,\ldots,m_n)-\log_{10}\mathrm{TOF}(m_1,\ldots,m_i-\delta,\ldots,m_n)}{2\delta} \qquad \textbf{Eqn 2.1}$$

The appeal of this method discussed by Campbell et al. is that it shows which steps in the FES do or do not have significant impact on the TOF[40]. Then if the FES of catalysts of similar enough composition are via computational chemistry, fewer steps in the FES need to be calculated for those catalysts. However, its use is beyond screening, and chemical insight can be gained from this method, especially for catalysts with highly varied compositions like the ones discussed in this text.

	HT	FR	M-OCHO	FD	M-W	M^{n+1}	M^{n}	$M^{n+3}H$	HD	PT	M-H
$Cp(CO)_2Fe^{II}$	0	1	0	0	0	0	0	0	0	0	-1
$Cp(dmpe)Fe^{II}$	0	0	0	0	-1	0	0	0	0	1	0
$Cp(bpy)Fe^{II}$	0.04	0	-0.96	0.96	-0.03	0	0	-0.01	0	0	0

Best PCA Catalyst	Cp*(dmpe)Fe^{II}	Cp(bpy^{NH2})Fe^{II}
0.67	0	0.11
0	0	0
-0.51	0	-0.16
0.12	0	0.16
-0.03	0	-0.81
0	0	0
0	0	0
0	-1	-0.03
0	0	0
0.22	1.00	0.74
-0.47	0	0

Table 2.3. Degree of Rate Control Effect on Catalyst TOF.

Examining table 2.3, the intermediates and transition states in which the TOF is most affected greatly vary as catalyst composition changes. For one of the best catalysts, Cp(bpy)Fe, the two most influential states are the metal-oxygen bound formate, and the formate dissociation barrier. Changing the bpy ligand to bpy^{NH2}, causes the most important states to be the water-metal complex, and the proton transfer barrier. Although these catalysts are nearly the same and have nearly the same TOF at ~10^{-3} s^{-1}, their most important states are quite different, due to the amine group on bpy changing the ease of reducibility of the catalyst. The best PCA predicted catalyst tells a different story than the rest, the biggest sinks are the oxygen bound formate and starting intermediate, while the barriers are hydride transfer, proton transfer, and formate dissociation. It should be noted that as the catalyst moves towards an optimum fastest TOF, the distribution of

what is the most influential rate determining step begins to become less clear, and different rates and sinks compete with each other. This can be seen with the change from $Cp(bpy)Fe^{II}$ to $Cp(bpy^{NH2})Fe^{II}$ then the PCA best predicted catalyst: at first it is clear there is one major sink and barrier, then as one barrier drops or sink rises another barrier or sink becomes more influential.

Other catalysts like $Cp(CO)_2Fe^{II}$, $Cp(dmpe)Fe^{II}$, and $Cp^*(dmpe)Fe^{II}$ are heavily influenced by FR and PT barriers, and $M^{n+3}H$ and M-W and $M^{n+2}H$ intermediates. All this is to say that at different points in composition space, the TOF is affected by different parts of the FES.

New Ligand Development

Utilizing the TOF heat map as seen in Figure 4, several novel compositions were created in an attempt to get closer to the maximum TOF. The dmope and dmopy ligands were created to try and make the dmpe ligand more electrophilic and a better pi acid. By having ligands with these properties, it was theorized the iron would be more electron deficient, therefore easier to reduce. Although the DFT generation of these FES was a great experiment, no catalysts containing either of these ligands had a better TOF than the Cp (bpy) Fe^{II} catalysts. It is clear from Figure 4 that while changing dmpe to dmope moved the catalyst farther away from peak performance, changing dmpe to dmopy was a move in the right direction towards the high TOF PCA space that bpy complexes lie in. However, this movement was not far enough in that direction.

Cobalt Example

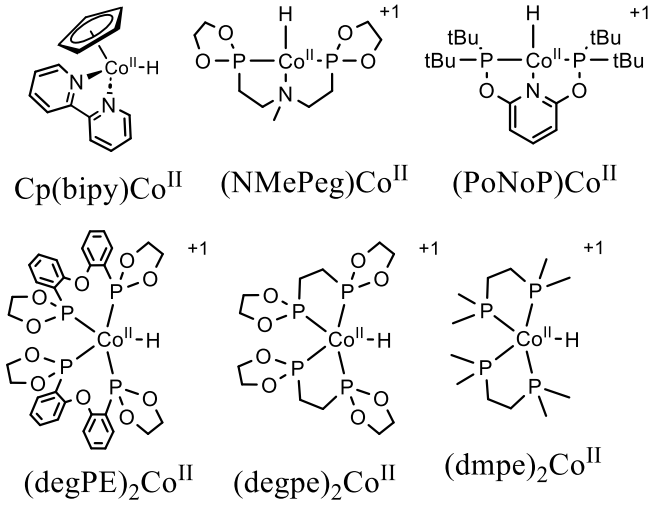


Figure 2.6. Cobalt catalyst compositions. In this analysis the ligand requirements were not quite as constrained as the previous analysis: here the ligands differ in functional groups, hapticity, and flexibility.

In previous the discussion we purposely chose modular ligands. In this analysis we wanted to see if the same method would still provide insight without as constrained or organized a composition space. In this analysis the catalysts utilized Co^{II} as the metal center, but differing types of ligands as seen in Figure 2.6. An abbreviated FES—formate formation transition state (TS) and M^{n-2}, M^{n-1}, $M^{n+3}H$, M-H, M-OCHO, and MW intermediates—was calculated for six distinct catalysts. With this abbreviated FES a simpler kinetic scheme was utilized to calculate the TOF to quantify catalyst performance. In addition to quantifying catalyst performance with TOF, PCA was also performed to assess the composition space. Although PCA was performed on six distinct points in free energy space, only two principal components described 97.2% of the variance. Using the same method as seen in Figure 2.4 these catalysts can be plotted on a two-dimensional TOF heat map as seen in Figure 2.7. These points lie on a line except for $Cp(bpy)Co^{II}$. What this means is that for all other five catalysts only one property is really changing as the ligands change. Looking at the coefficient matrix, in the first PC the largest

values are the oxidized states, while the lowest values are reduced states. For the second PC, the largest values by a wide margin are the two reduced states. For all the other catalysts these properties are a give and a take, as you change the catalyst's ability to be reduced you change its ability to be oxidized. To get off said line, and towards different chemical composition and behavior, the catalyst properties need to change drastically. The catalyst Cp(bpy)Co^{II} lies of the line because the bipyridine ligand is reducible.

Plotting catalysts in PCA space allows for ease of navigation in chemical composition. Without requiring any chemical intuition PCA space plots elucidate LFERs, and chemical lessons that might not be seen by looking at a FES table.

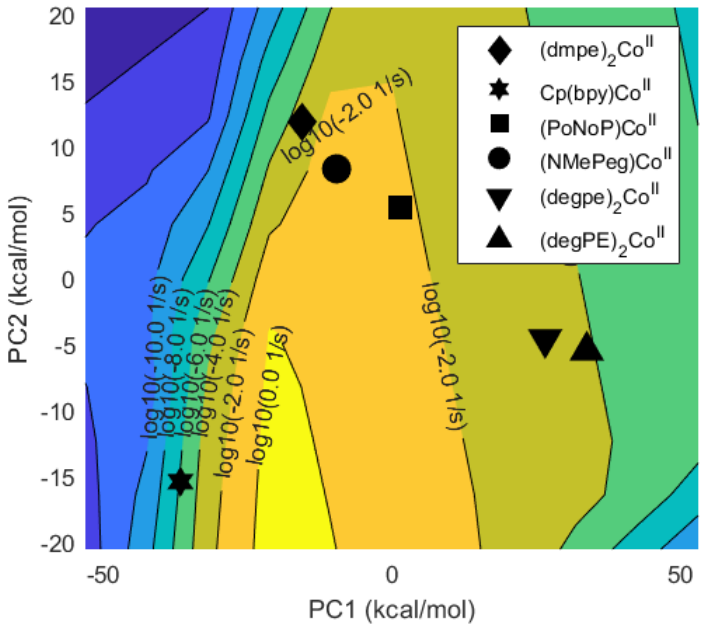


Figure 2.7. Co^{II} TOF Contour Plot in PCA Space. Like Figure 4, this figure correlates each point in PC to a FES, and each FES to a TOF. These figures allow for ease of navigation in chemical composition space, and paired with coefficients matrices can teach valuable chemical lessons.

Methods

All calculations were performed with ORCA 4.2.0 software[11,12]. For each Gibbs free energy calculation, two distinct DFT calculations were performed. The first was a simpler calculation in vacuum utilizing the bp86[15,16,42] functional and SVP basis set[19]. This calculation was utilized to calculate thermal corrections, and to more quickly get closer to an optimized geometry with a faster calculation, decreasing computational costs. The second DFT calculation performed utilized the PBE0 functional[15,16,20,21], TZVP(-f) basis set, and minimally augmented TZVP(-f) basis set on oxygen atoms[19], and the SMD implicit solvation model[13] with water as the solvent. This more rigorous calculation was utilized to more accurately calculate solvation effects and the single point energy. Both sets of DFT calculations utilized the Los Alamos pseudopotential and 3Z basis set and effective core potentials on the metal atoms[22–24], the D3BJ dispersion corrections[17,18], and the Weigend J auxiliary basis set[25].

For relative free energy calculations it was assumed the hydration energy of a proton was -264 kcal/mol[27], making the free energy of a proton in water at a pH of 7, -279.8 kcal/mol[26]. Similarly, the free energy of an electron at voltage of -0.536 V vs SHE, was assumed to be -86.3 kcal/mol[28].

There is known error with DFT, and to improve the accuracy of these calculations an isodesmic correction was made to all free energies. A more rigorous calculation, DLPNO-CCSD(T) with ZORA approximation, was performed on the full FES of two catalysts[43–47]. These calculations utilized the geometries from the converged solvated PBE0 calculations. A (3,4) infinite basis set extrapolation[48] utilizing the cc-pvNZ basis set where N is the cardinal number was performed for all values[49]. The difference between the CCSD(T) and DFT calculated Gibbs free energy was used as the isodesmic correction. These values are seen as the first line in Table 1.

Conclusion

Through the course of this text, we have examined the linear free energy relationships (LFER) of a family of catalyst containing a Cp ligand, two monodentate or one bidentate ligand, and a Period 4 metal. The free energy surfaces (FES) of these catalysts were generated via density functional theory (DFT) with an isodesmic correction calculated via CCSD(T). The free energy surfaces consist of generating formate via hydride insertion into CO_2, dissociation of said formate once bound to the metal, and finally regeneration of the catalyst via two reductions and a protonation. It was found that this family of catalysts displays wide variety in their FESs, and the path of least resistance may vary greatly from catalysts to catalyst. Catalyst performance was quantified via turnover frequency of each catalyst found via microkinetic model. It was found that the TOFs are heavily dependent on the conditions of the system (pH and applied voltage). Additionally, some catalysts may have intermediates or transition states that are unstable and do not exist for other catalysts. To better understand the LFERs of these catalysts, principal component analysis (PCA) was performed on this family of catalysts. To fill out unknown or nonexistent values of intermediate or transition state free energies, imputation, a form of matrix completion, was performed on a matrix consisting of the DFT generated FES of all catalysts in this family. Through examining the coefficient matrix generated by PCA, and examining the catalysts in PC space, two major properties were found to affect catalyst behavior. The first and most influential PC corresponded through electrophilicity, and the second PC corresponded to reducibility.

The PC space created by the known FES was examined to realize a maximum TOF: of this PC space itself of 0.0360 s^{-1}. This was several orders of magnitude different than the best DFT calculated TOFs of the $Cp(bpy)Fe^{II}$ catalysts at $\sim 10^{-3}$ s^{-1}.

A degree of rate control analysis (DRC) was performed, and it was found that the rate limiting steps and sinks of catalysts varied greatly in this family. As catalysts moved towards the optimum composition space, more steps in the kinetic cycle became more influential on the TOF as these steps began to compete with each other.

Finally, an additional study was performed on another family of Co^{II} catalysts with fewer composition constraints. Utilizing the same analysis described before, PCA found that there were two major principal component axes. The first PC corresponded to oxidizability of the catalyst, and the second corresponded to reducibility. By plotting these catalysts in PC space it further bolstered that PCA is clear way to visualize and navigate composition space of catalysts. This was evident when it was seen that the catalysts containing bipyridine (bpy) was far away from the other catalysts of this Co^{II} family, as the bpy ligand is reducible.

Chapter 3: Free Energy Surface Prediction via Quantified Catalyst Properties Case Study: C-H Activation

Introduction

Computation has long been used for discovering improved materials/molecules. When a computable figure of merit has been identified, discovery often happens by screening many candidates and choosing the materials which score highest. In catalysis, the figure of merit is often as simple as one or more bond energies[50] (to be optimized using the Sabatier Principle) or an activation barrier of a key step[51,52] (to be minimized) or features in a solids' band structure[53]. One of these descriptors are then used to predict relative catalyst activity[54]. While good candidates can be found via screening there are often downsides to this approach. First there is often no new understanding of structure/function relationships accumulates as the list of candidates is processed. Second, screening does not provide an indication of whether the best material has been found/whether the list of candidates contains the optimum composition (although the Sabatier Principle[55] allows this). Third, simple or 1D figures of merit do not capture competing qualities like stability, activity, selectivity. Finally, in catalysis the intermediate and elementary step which determine turnover frequency are not the same for all potential catalysts in all conditions, so knowledge of only part of the free energy surface (a reaction energy or barrier) is insufficient to accurately assign performance to candidates. As one strategy for overcoming these limitations, we have investigated the training of regression models to predict condensed catalytic free energy surfaces from the ligand properties of organometallic candidates. Successive generations of training data provided by DFT improve the accuracy and range of the models, leading to optimized compositions and revealing structure-activity relationships conserved in the best performers.

Multivariable linear regression techniques were utilized to relate quantified catalytic properties and performance as they were more readily interpretable to make sense of the relationship between these two phenomena. To get more accurate predictions from a regression trained on data with inherent error a relatively large amount of data is required, so simple high throughput calculations were utilized. As more DFT calculations were performed catalyst behavior was optimized and chemical lessons were learned along the way.

Methane oxidation, more specifically alkane functionalization, with an Ir^{III} catalyst was the case study utilized for this project. Goddard et al. have shown potential reaction pathways from methane to methanol utilizing DFT simulations of Au^{I} and Au^{III} in strong acid solvents[56]. There has been great precedence for utilizing bipyramidal Ir catalysts as a case study. Periana et al. showed viability of H/D exchange in methane with an Ir catalyst[57]. Later Cundari et al. investigated trades offs in designing ligands on Ir^{III} catalysts for C-H activation[58]. Finally Goldman et al. defined electronic factors in the oxidative addition of C-H bonds to Ir^{I} complexes[59]. With no desire to reinvent the wheel, this strong basis of the electronic factors influencing this chemical process were a great basis to investigate a way to optimize catalyst performance.

While this chemical process was chosen as the case study because of the strong foundation that electronic properties of ligands have on this process, there is great importance in alkane functionalization as well. Alkane functionalization allows for the conversion of hydrocarbon gases to liquids allowing for easier storage and transportation[60]. Alkane functionalization could also be utilized in fuel cells allowing for fuel to be stored and used easily and efficiently[60]. Finally, controlling alkane functionalization could change the approach to organic synthesis[61,62].

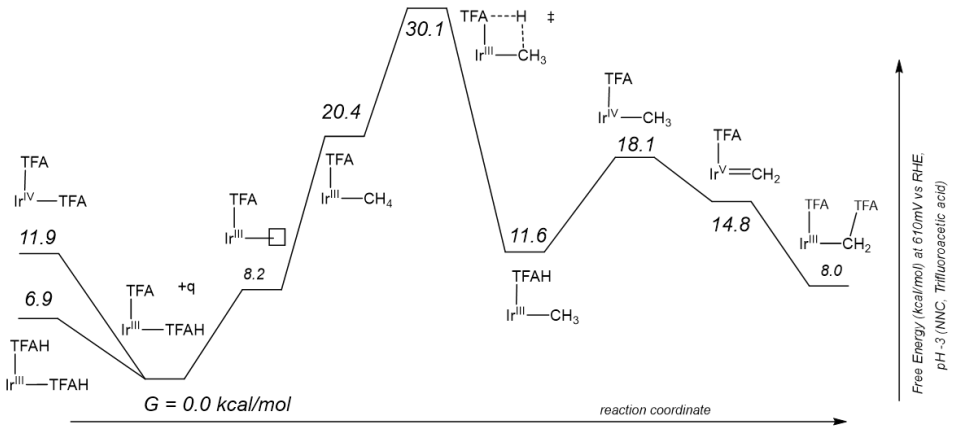


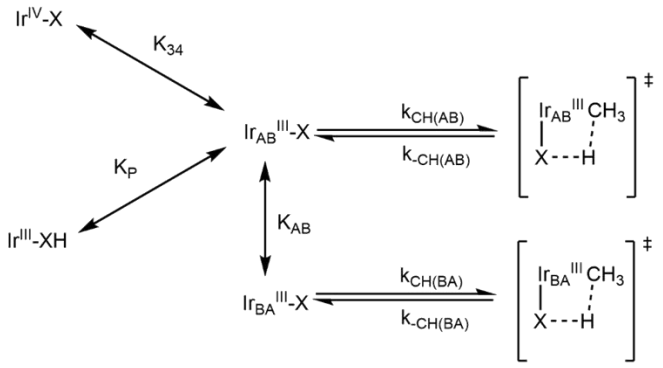
Figure 3.1. FES of Ir NNC oxidizing methane to methanol. The free energy surface (FES) of the full catalytic cycle of CH_4 oxidation for Ir^{III} NNC was calculated with DFT in TFA solvent. This scheme highlights the most important parts of the cycle are C-H activation, and the two off cycle intermediates, the protonated and oxidized intermediate.

Methods

Kinetics

The free energy surface (FES) of the reaction pathway of the Ir NNC complex from Periana et al.[57] was calculated in TFA solvent with DFT as seen in Figure 3.1. It was proposed to shorten the catalytic cycle to reduce the number of DFT calculations required to quantify catalytic performance per catalyst. In this scheme the largest barrier is the subject of interest, C-H activation. No intermediates that are created in this chemical process are as formidable of a barrier as this transition state, however two off-cycle intermediates can influence catalyst performance. These oxidized and protonated intermediates are not only indicative of the stability of the catalyst as they show ease of protonation and pKa of the catalyst, but they also influence how much starting intermediate is available. Concentration of starting intermediate influences turnover frequency, which is the main metric of catalytic performance used in this study. It was decided to reduce the kinetic scheme to just a few points on the FES: the starting intermediate,

the C-H activation barrier, and the protonated and oxidized intermediate. The final scheme, as seen in Scheme 3.1 decided on for this project allowed for the anion site and active site to switch places. This means there are potentially two ground states, and two transition states.



Scheme 3.1. Abbreviated catalytic cycle. The catalytic cycle can be shortened to the two off-cycle protonated and oxidized intermediates, the ground state, and the transition state. Since it was assumed that the active site and anion could switch places, there could be two possible ground states, and therefore two C-H activation transition states, assuming that ligands A and B are different.

This Ir^{III} catalyst as seen in Figure 3.2 has an active site, occupied by water in the starting intermediate, and an anion from the solvent, hydroxide from water. Positions A, B, C and C' are occupied by ligands with the A position trans to the active site, B trans to the anion, and C and C' trans to each other.

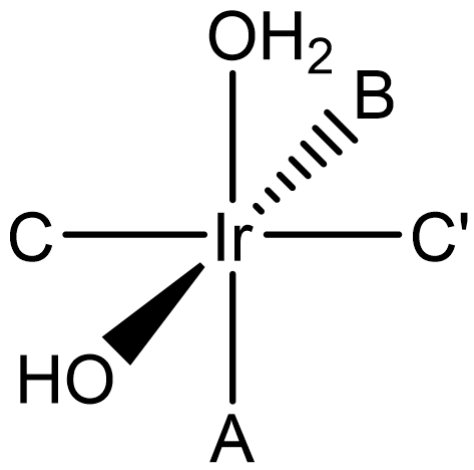


Figure 3.2. Our catalyst, ligand positions and anion and active site. This figure represents our catalyst, where A, B, C and C' are ligand positions. The water is the active site, and hydroxide is the anion (from the solvent water). Position A is trans to the active site, B is trans to the anion, and C and C' are trans to each other.

One of the goals of this project was to easily modulate the catalyst behavior with different kinds of ligands. However, the workflow of quantifying each multidentate ligand for a new calculation becomes tedious when utilizing a vast number of multidentate ligands, and changes to each multidentate ligand may be minimal. Rather than having a large database of multidentate ligands, smaller "toy" ligands, as seen in Figure 3.3, were developed to represent fragments of larger ligands. A database of relatively few "toy" ligands may then be used to represent a plethora of larger multidentate ligands. This not only allows for ease of modulation of electronic and steric properties with a relatively small ligand database, but also can reduce the number of atoms simulated via DFT, speeding up computational time.

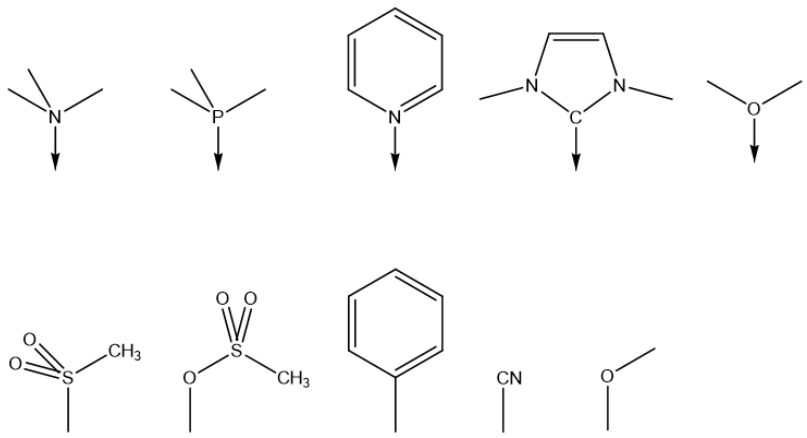


Figure 3.3. Toy Ligands. The monodentate ligands used in this report are coined "toy" ligands, essentially fragments of larger realistic ligands which. The first eight ligands utilized did not include the N-heterocyclic carbene and dimethyl ether ligands. Those two ligands were added at the end of analysis.

There were four properties of ligands that were believed to influence catalyst performance: sigma-donicity (how easily a ligand can donate as sigma bond), pi-accepticity (how easily a ligand can accept a pi bond), charge, and steric bulk[63]. To quantify these ligands their geometries were optimized via DFT on $Ru^{II}(NH_3)_4Cl$. From this calculation many methods were utilized to quantify these ligand properties. For example, sigma or pi orbital energies could quantify sigma-donicity and pi-accepticity, respectively. The sigma bond could also be quantified via Ru-Cl properties like bond distance and Mulliken charge on Cl. Steric bulk was quantified with $\%vol_{burried}$ via SambVca 2.1[64].

Computation

All computational chemistry simulations were run in ORCA 5.0.3[11,65]. Geometry optimizations were performed with density functional theory calculations. Vacuum calculations were performed with the B3Pω91 functional[15,16,42], D3BJ and def2/J corrections[17,18,25], def2-TZVP(-f) basis set[19], and Los Alamos pseudopotentials and 2Z valence functions for metals[22–24] For

calculations utilizing solvation, the SMD implicit solvation model was utilized[13]. The temperature was assumed to be 298K. The hydration energy of a proton at 1M was acquired by Tissandier et al. as -264.0 kcal/mol[27]. Correcting this energy to a pH of 7, the Gibbs free energy of a proton in solution was assumed to be -279.8 kcal/mol[26]. Like with the proton hydration energy it was assumed a 4.28 V difference between SHE and vacuum[28], the free energy of an electron was assumed to be -103.3 kcal/mol at 0.2 V vs SHE.

To speed up computational time thermal corrections were assumed to be constant across all catalysts. Only one set of thermal corrections was calculated, for the catalyst $Ir^{III}(OMe)(PMe_3)_3$, then utilized across all catalysts. Calculations were performed in vacuum except for the protonated intermediate. This is because the protonated intermediate has a +1 relative charge to everything else on the abbreviated FES, so its electronic energy is inflated.

The absolute Gibbs free energy was then equal to the sum of the electronic energy and thermal corrections,

$$G=E_{e^-}+E_{ZPE}+PV+H_{vib}+H_{trans}+H_{rot}-T*(S_{vib}+S_{e^-}+S_{trans}+S_{rot}) \quad \textbf{Eqn 3.1}$$

where the electronic energy in italics is unique to each calculation, and the thermal corrections are from the $Ir^{III}(OMe)(PMe_3)_3$ catalyst.

Results and Discussion

Ligand Data

There were numerous measurements made to describe each electronic property of these ligands, so it was believed that maybe all methods of describing each property may not be necessary to quantify the catalysts. As seen in figure 3.4, the pi orbital energies are linear dependent with each other. When examining the measurements to describe the electronic properties of each ligand, it was determined that some measurements were dependent on charge.

All orbital energies were found to depend on the charge of the ligand because the calculations were performed in vacuum, so the charged ligands had an inflated orbital energy. Furthermore, the Mulliken population on chlorine, meant to describe the sigma-donicity of the ligand, was found to be higher for anionic ligands than for the neutral ligands. More than all this, it was found that there was a high linear correlation between ways of measuring the same property.

To determine the linearity amongst the descriptor data principal component analysis (PCA) was performed on all the descriptor data. Principal component analysis is a method of analyzing multidimensional data in the form of an array. Principal component axes are constructed that are orthogonal linear combinations of the previous dimensions of the array, where the new dimensions in real space lay in the direction of most variance of the data. This analysis found that only four principal components (PCs) were needed to describe 99.1% of the variance amongst the descriptor data.

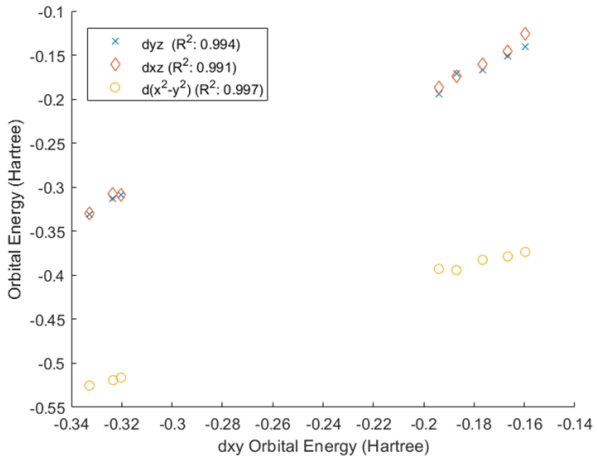


Figure 3.4. Ligand property correlation example. The pi orbital energies had strong linear correlation. The discrepancy between the top five values and bottom three is due to the bottom three being neutrally charged and the top five being anionic.

Many regression schemes are well suited to weeding out unnecessary predictor variables, so great apprehension is not necessary when attempting to describe each catalyst quantitatively. However, there is an inherent error in DFT, and without being able to pinpoint the sources of and consistencies this error turns into noise. As this was the first test case in predicting FESs from catalyst properties, it was not known which regression to use, or how to tailor them to balance bias and variance. Caution was taken to not introduce unnecessary noise into the descriptor variables by utilizing no more than one descriptor of each ligand.

To decrease redundancy in the descriptor data, measurements were selected to describe properties that were not dependent on charge. The sigma-donicity was described by the Ru-Cl bond distance. The pi-accepticity was described via the sum of the *dxz* and *dyz* pi orbital energy. The pi-accepticity was only quantified via orbital energies, so the effect of charge could not be separated from this measurement. However, the charge of the ligand was utilized as descriptor data, as the pi-accepticity alone would not be enough to describe the charge of the ligand. Finally, the steric bulk of each ligand was quantified with the $\%vol_{buried}$ via SambVca 2.1 utilizing the default settings[64].

As previously stated, the goal of this project is not necessarily to find the best catalyst, but to find the way to find the best catalyst in the most efficient way possible. It was believed that the best way to do this was to optimize towards certain catalyst behavior, in our case, highest turnover frequency. This would be done by first building a training set of FESs then using that training set to predict all possible 2304 catalysts. The best few catalysts predicted by the regression would have their FES generated via DFT, then added to the training set. This algorithm is visualized in Figure 3.5. This process would be repeated until the algorithm converged, or the best predicted catalysts already existed in the training set.

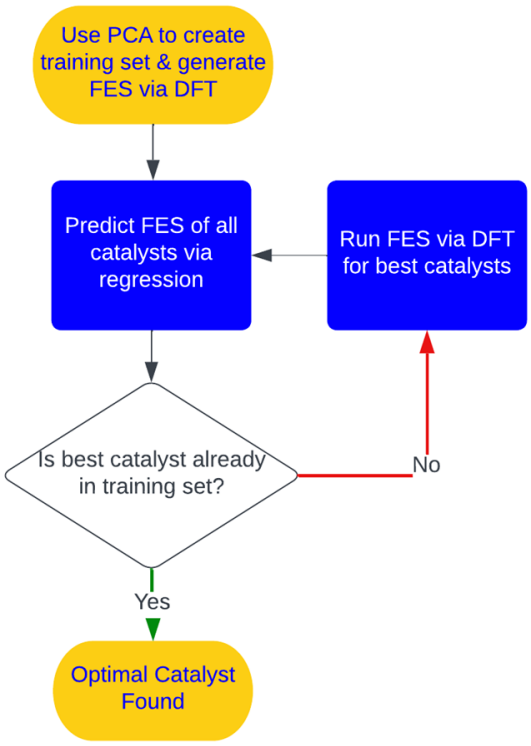


Figure 3.5. Catalyst optimization algorithm. This algorithm describes the general methodology to optimize towards a catalyst property with the "best catalyst" being that with the highest turnover frequency. In closed system the with a fixed number of catalysts, 2034, it is impossible for this algorithm to diverge.

With an astounding 2034 possible catalysts it is difficult to know where to start building a training set for predicting all of these catalysts and finding the best one. Chemical intuition and understanding of the goal at hand can be a good guide when taking a shot in the dark.

At the beginning of modeling, an educated guess was taken that using mostly neutral ligands on the catalyst would probably be preferable as they would provide the catalyst with a higher oxidation potential which would make it more difficult to change Ir^{III} into Ir^{IV}. Resistance to oxidation would allow for more relative starting intermediate, and because turnover frequency is dependent on starting intermediate concentration, create a higher turnover frequency. For this

reason the first round of catalysts whose free energy surfaces were mapped utilized trimethyl amine for most of their ligands. Trimethyl amine was utilized primarily in the C and C' locations on the catalyst, and the A and B ligands (mostly A) were varied. The A ligand was varied most because it is trans to the active site, so the ligand in position A should affect the catalyst behavior the most. The B ligand is trans to the solvent, so it was theorized that it would have the second largest effect on the catalyst. The C and C' ligands are trans to each other, so their effect was theorized to be the least influential on the catalyst.

The first 10 catalysts or so took an unexpectedly long time to converge via DFT due to the utilization of the trimethylamine ligand. Because this ligand is monodentate with many substituents, it can freely rotate which can make it take quite some time to find a stable conformation. The purpose of utilizing toy ligands was not only to be able to modulate catalyst behavior, but also reduce the time of each calculation so that many calculations could be performed. For this reason, the next round FESs calculated via DFT utilized catalysts with pyridine ligands in place of trimethylamine as they are neutral but do not have so many moving parts.

After calculating the FESs via DFT for around 20 catalysts of similar ligand compositions, partial least squares regression (PLSR) and elastic net regression were utilized to predict the FES of all possible catalysts (2304) with every ligand conformation of our ligand set. Elastic net and PLSR are two regression methods to relate multidimensional descriptor data to outcome data. Elastic net utilizes a penalty function to combine both techniques from LASSO and Ridge regression, both of which fit weights in multivariable regressions to relate descriptor data to outcome data. Partial least squares regression is slightly similar to PCA in that both descriptor and outcome data are first reduced in size through a linear transformation before being related to

each other through a regression equation. When utilizing the maximum number of PLS components in the regression, thousands of combinations of catalysts were predicted to have a negative transition state. Unphysical results like negative transition states means that the prediction quality is poor, and searching for a good catalyst amongst the thousands is impossible if the results are so inaccurate that potentially good candidates are considered poor ones.

Further work was done to look at the current training set of catalysts. When examining the in-sample predictions with PLSR and elastic net, the regression performed relatively well. Then a round robin style algorithm was performed to test prediction performance. In this algorithm a group of a few catalysts were set aside as validation data at random, then the remaining catalysts were utilized to train a regression and predict the validation data. This was performed again and again until all catalysts were predicted as validation data. The results of this style of algorithm showed that the error was much larger than when regressions were trained on the data. In fact it was found that sometimes with elastic net no correlation was found at all, meaning that elastic net predicted the same FES regardless of the predictor data. This result is indicative of too much bias in the regression.

It was clear that the model was doing a poor job at predicting catalysts that were highly different from those of the training set as seen in Figure 3.6. This comes as no surprise as PLSR is known for having poor prediction quality when predicting values that are very unlike the training set. Although at this point the training set wasn't small, the compositions in the training set were not well varied. The most logical solution at this point was to add more data to the training set, for adding more variance to the training set will give regressions a broader picture of the relationships between the descriptor variables of the catalysts and their free energy surfaces.

It was proposed to just start optimizing towards a desired property in the catalysts. This means to just add FESs via DFT of the best predicted catalysts. However, it was clear that prediction quality was quite low outside of the sample family. It was believed that a better, more thoughtful approach to creating an initial data set would provide for faster and more robust optimization. "Garbage in, garbage out."

One proposed idea was to add the FESs of the catalysts with the worst quality of predictions as the training set lacks catalysts in that region of chemical space. However, besides the catalysts that were predicted to have a negative TS the quality of prediction could not be quantified. Additionally, there were thousands of catalysts that had a predicted negative transition state. It was unfeasible to calculate the FES's of all those catalysts, and it would be taking a shot in the dark to choose any of those catalysts at random.

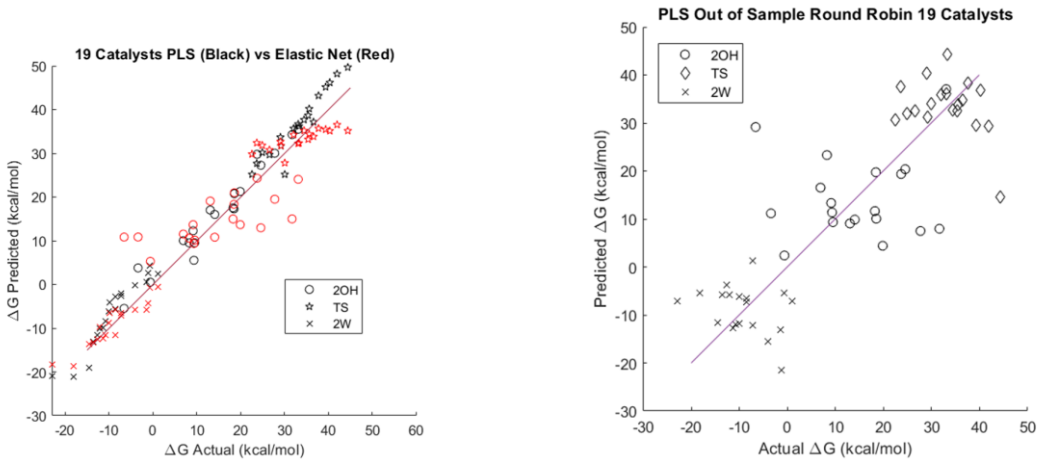


Figure 3.6. In and out of sample initial predictions. The in-sample catalyst predictions (left) were relatively accurate for both elastic net and PLS regressions. When a round robin style scheme to predict out of sample (right) was utilized neither regression techniques excelled.

It was clear what needed to be done, to add the most amount of variance possible to the training set with the fewest number of catalysts. To do so, catalysts need to be added to the

training data such that each catalyst is very different from each other, so that a broad picture of the relationships between catalyst properties and function may be captured by a regression. To generate this highly varied data set PCA was performed on the catalyst descriptor data for the thousands of catalysts, and the catalysts to be added to the training set were selected via PCA scores. Catalysts were chosen that lie on the outermost edges of principal component (PC) space, both negative and positive, for the first several PCs. This idea is illustrated in Figure 3.7, the most extreme PCA values are circled in red. Adding these catalysts greatly improved prediction quality and reduced the number of negative transition states.

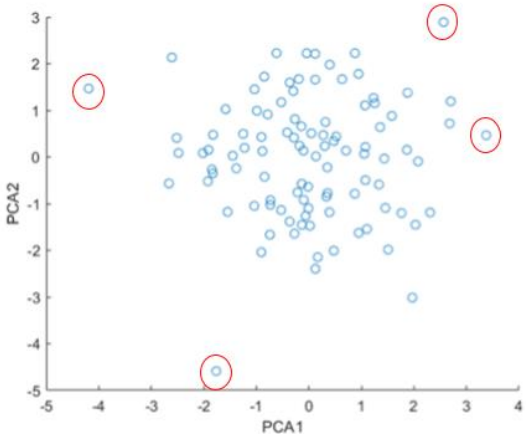


Figure 3.7. Data set building through PCA illustration. To add more variance to the data set PCA was performed on the descriptor data of all possible 2304 catalysts. This figure is an illustration of a PCA analysis performed on predictor data with the most extreme scores circled in red. These extremes would be chosen to add DFT generated FES data for.

Once the training set was updated and the predictions were improved, the search for the best catalyst began. The PLSR was used to predict the thousands of possible catalysts, and then the FES was found via DFT for the best catalyst amongst these predictions. It was at this point where it was determined a new kinetic scheme needed to be implemented as the active site and anion site were switching for many of the best catalysts. Prior to this point it was assumed that each

catalyst's active site would stay in place, and this assumption was only made to reduce the number of DFT calculations required for each catalyst. This realistically shouldn't be expected, and it makes more sense that the active site should be able to act as the anion site, and vice versa.

Once this change was made, this procedure reduced the number of total catalysts run as any catalysts where the A ligand was different from the B ligand could not be used in the training set unless the BA version of that catalyst was also there. To expand the training set, the same procedure was continued as before: 1) use PLSR to predict the thousands of possible catalysts (now slightly reduced to 1296 as AB catalysts were considered to be the same as BA catalysts), 2) find the "best" performing catalysts of the thousands—determined by highest TOF—3) calculate the FES of these "best" catalysts via DFT. This procedure was performed for several more catalysts until the algorithm converged, or the best catalyst predicted was the same after that best catalyst had been added to the training set. The out of sample results may be seen for elastic net and PLS in Figure 3.8. From this figure it is clear that error is much smaller than that seen in Figure 3.6, and elastic net and PLSR perform relatively similar.

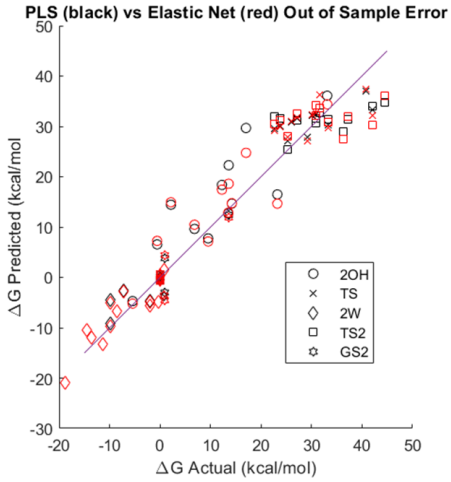


Figure 3.8. Final regression prediction results. After many catalysts FES had been generated and the optimization was farther along, out of sample predictions were made. These predictions were much lower in error, and elastic net and PLS perform relatively similarly.

The best catalyst was dependent on the number of PLS components utilized for predictions. However, it was seen that the best catalysts tended to have weaker sigma donors in the A and B position. For example, utilizing the maximum number of PLS components the best catalyst was found to be $Ir^{III}(OSO_2Me)(Ph)(Cyn)(Cyn)$. However, it wasn't clear whether this algorithm converged at a local maximum or an absolute maximum. It was proposed that maybe the current ligand set was the limiting factor in finding a global maximum. Better catalysts may exist that utilize ligands with just slightly different properties than the ligands utilized thus far. It was proposed to construct predictor data for theoretical ligands: ligands with unknown chemical structure that only exist as descriptor data. This allowed for catalyst properties to be varied on an even finer scale and with different combinations of data. Without knowing what these ligands would look like it was safer to not have wildly different properties than the current eight ligands utilized. Each of the theoretical ligand's properties (sigma-donicity, pi-accepticity, and steric bulk) could be one of three values, either the mean of said property of the eight ligands, or plus or minus two times the standard deviation of that property. This three-by-three-by-three grid (three possible values and three possible properties, sigma-donicity, pi-accepticity, and steric bulk) makes 27 ligands, and with the possibility of being anionic (-1), or neutral, makes 54 ligands. Four possible locations to put the ligands on the catalyst and A and B being the same position as well as C and C' being the same position then makes 4.3 million possible catalysts.

Utilizing the PLSR constructed with all the current catalysts with DFT acquired FESs, these 4.3 million catalysts with theoretical ligands were predicted and organized from best to worst.

With a relatively small training set compared to 4.3 million catalysts there were bound to be many with unphysical values. Obviously unphysical predictions were thrown out and the remaining catalysts were analyzed. Changing the number of PLS components greatly changed the quality and predictions of catalysts, and what was considered the best catalyst. Except when utilizing only one PLS component for predictions, one trend was clear: the catalysts with the highest TOFs had strong sigma donors in the C and C' position, weaker sigma donors in the A and B position, and mostly neutral ligands.

The reasoning behind this phenomenon is relatively intuitive for the A and B position. Weaker sigma donors trans to the anion allows the anion metal bond length to become shorter making the anion more basic. This allows the anion to more easily accept the proton from the methane. The methane is also stabilized by having a weak sigma donor trans to it. Neutral ligands impart stability against protonation and oxidation of the starting intermediate.

To attempt to improve the predictions for the 4.3M catalysts, the FES was calculated via DFT for two catalysts: $Ir^{III}(OMe)_2(PMe_3)_2$ and $Ir^{III}(OSO_2Me)_2(PMe_3)_2$. Both catalysts have strong sigma donors in the C and C' position, and weak sigma donors in the A and B positions. The FESs of these two catalysts did not have extraordinarily low TSs, however, once their data was used as part of the training data for the PLS regression it did improve the FES predictions by PLSR. The highest TOFs of the 4.3 million catalysts were no longer in the $10^{7\text{-}8}$ sec^{-1} range, but rather in the 10^4 sec^{-1}. These best FESs of the 4.3 million were still unphysical, with both the TSs hovering near 10 kcal/mol, however this was a significant improvement with the addition of only two catalysts to the training set.

Again, the FES of these best catalysts showed comparable properties, weak sigma donors for the A and B positions, strong sigma donors in the C and C' positions, and mostly neutral ligands.

Because of this, two new neutrally charged ligands were run on the $Ru(Cl)(NH_3)_3$ molecule to collect ligand property data. An N-heterocylcic carbene (NHC) was meant to be a slightly strong sigma donor. Dimethyl ether was the weakest sigma donor yet, according to the Cl-Ru bond distance. Four new catalysts with these ligands were added to the training set and there were minimal improvements to the predicted FESs of the 4.3 million catalysts. Looking back on all the data it was clear that there was error stemming from the use of monodentate ligands. These freely rotating monodentate ligands often formed strange configurations having hydrogen bonds that normally wouldn't occur in multidentate ligands. This error seen translates to noise which greatly decreases prediction quality by the regressions. This error needs to be considered when predicting FESs.

Tridentate Ligands

	NHC-Ir Dist (A)	H2O-I Dist (A)	Ir5s Orbital (Eh)	Relative Bond to Ir (kcal/mol)	d-pi antibond with OH (Hartree)	D-pi Antibond with NHC (Hartree)	D-pi antibond with nothing (Hartree)
Cp	2.022	2.231	-3.724	0	-0.2310	-0.2518	-0.2463
Tpm	2.018	2.116	-3.733	10.5	-0.2331	-0.2618	-0.2429
Tp	2.018	2.132	-3.715	32.3	-0.2216	-0.2466	-0.2274
Tpym	2.026	2.120	-3.736	13.9	-0.2373	-0.2787	-0.2479
Tpyr	2.027	2.131	-3.726	26.9	-0.2310	-0.2712	-0.2415
TpCF3	2.013	2.098	-3.740	-3.69	-0.2339	-0.2831	-0.2465
TpmCF3	2.017	2.091	-3.759	-24.2	-0.2460	-0.3022	-0.2981
C3POch3	1.983	2.073	-3.719	-34.3	-0.2218	-0.2259	-0.2530

TpymB	2.024	2.142	-3.719	37.2	-0.2252	-0.2858	-0.2622

Table 3.1. Tridentate Ligand Data

The final step of this process was to create realistic ligands with lessons learned thus far. Like the Ir NNC ligands, several tridentate ligands were theorized, constructed, and analyzed via DFT. While some well-known ligands were utilized, several novel ligands were constructed. Examples of these ligands can be seen in Figure 3.9. These tridentate ligands were meant to be like tris(pyrazolyl)borate (Tp) ligands that would rigidly occupy A, B, and C positions, with the C' position being occupied by the strong sigma donor N-heterocyclic carbene. Similar to the analysis with the toy ligands performed on the $Ru(Cl)(NH_3)_3$ molecule, these molecules were placed on the Ir atom along with an NHC ligand with a water and hydroxide bound (the starting intermediate) and their properties were recorded. These properties may be seen in Table 3.1. The NHC-iridium bond distance and water-iridium bond distance were used to quantify the sigma-donicity of the Tp ligand. The overall electron richness of the active site was quantified by the Ir 5s orbital energy. The complex and free ligand energies were also recorded to approximate the relative binding energy of the ligand to the metal. Finally, the pi-accepticity was quantified by visualizing the orbital energies and recording the various d-pi antibonds with OH, NHC, and nothing.

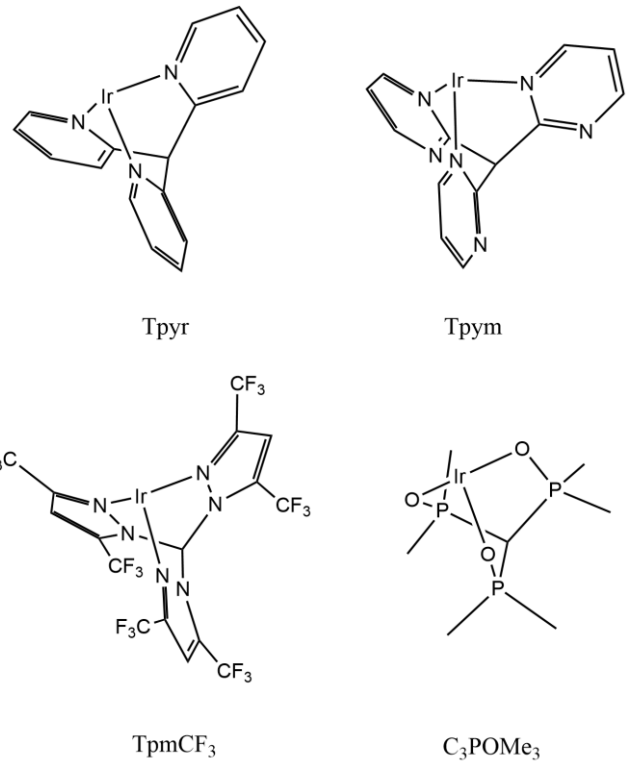


Figure 3.9. Tridentate ligands. These figures show the chemical structure of a few of the lesser-known tridentate ligands utilized in this analysis. Ligands like Tp and Tpm are relatively well known, however lesser-known ligands like Tpym and Tpyr are shown to distinguish their characteristics.

Many of these catalysts with Tp derivatives had their abbreviated FES simulated, but only two catalysts had their full FES calculated via DFT, tris(pyrazolyl)methane (Tpm) and tris(pyrimidyl)methane (Tpym). These free energies can be seen in Figure 3.10. Relative to $(NNC)Ir(TFA)_2$ the results were promising. Both Tpm and Tpym complexes had lower C-H activation transition state barriers and were more resistant to oxidation than the NNC complex, while the protonated intermediates remained endergonic.

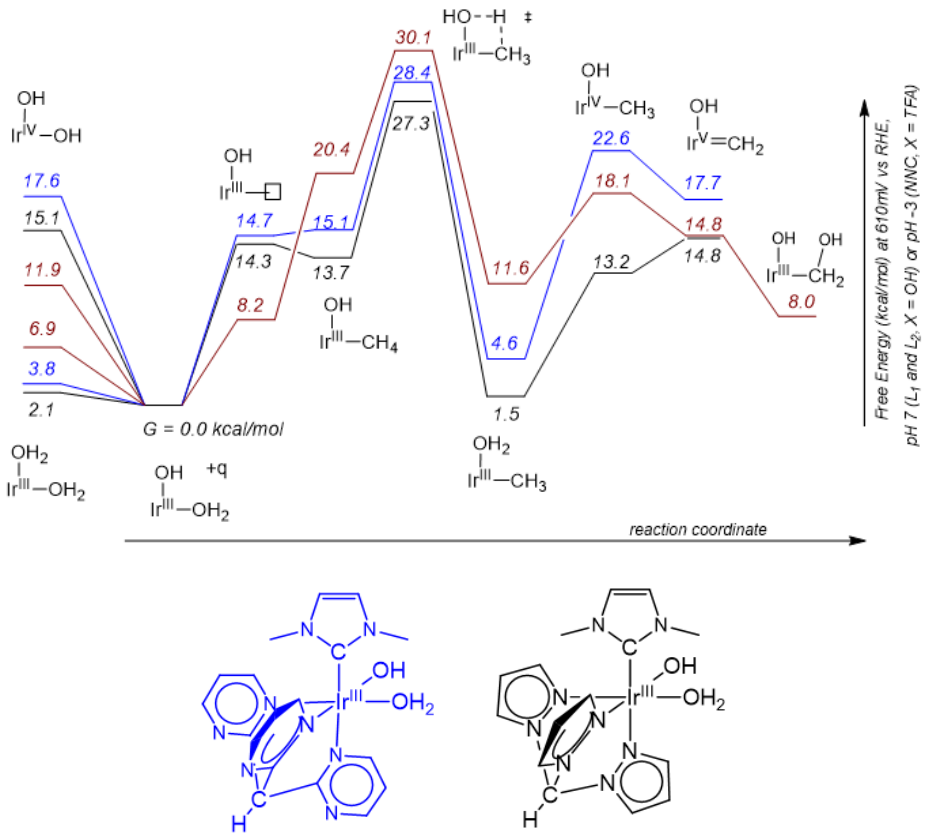


Figure 3.10. FES of tridentate ligand catalysts. The full FES was calculated with DFT in solvation for two catalysts, tris(pyrazolyl)methane (Tpm) and tris(pyrimidyl)methane (Tpym), with promising results. These catalysts had low C-H activation transition states, unfavorable oxidized intermediates, and endergonic protonated intermediates.

Categorical Regression

Post analysis another regression model was tested to see if it faired better: categorical regression. Categorical regression creates a binary style categorical model for predictor variables x, and then performs a standard linear regression for each outcome variable y. For example: if ligand position A can be ligands w, v and u, then the vector representing ligand A is three variables long and One Hot Encoded (if one value in the vector is one the others must be zero). In this way ΔG_{ij}, where i denotes the intermediate or transition state and j denotes a certain

catalyst, can be thought of as the average ΔG_i plus a sum of perturbations due to the presence of a particular ligand at one of the sites.

There are pros and cons of using categorical regression. The largest benefit is that it takes out inherent user bias. We believe that our measured properties, sigma-donicity, pi-accepticity, sterics, and charge, and the way we measure them are the most important predictors, but that could not be the case. Utilizing unbiased x variables may lead to better predictions that don't get skewed by these measured properties or our intuition.

The downside of categorical regression is that it is noncontinuous, which bounds the predictor data. Utilizing measured descriptors allows for more predictions outside of training data, and categorical descriptions limits predictions to strictly combinations of the training data. For example, if there are 10 ligands and the training data utilizes all 10 ligands on ligand positions A, C, and C', but only 7 ligands on position B, then FES predictions on catalysts that contain the other 3 ligands will be inaccurate and not properly described by predictor variables. Even with a relatively large and robust data set all possible ligands may not be utilized on all A, B, C, and C' positions. With a predictor data set like this, it would not be able to properly analyze all possible 2304 catalyst ligand combinations. More than this, the data is noncontinuous, so analyzing entirely different ligands is impossible. Some of the most insightful chemical lessons in this report came from analyzing the PLS predicted FESs of 4.3 million catalysts utilizing 54 "theoretical ligands." This was achieved because the these catalysts had ligand properties that were varied in a different manner than the known 10 ligands given in previous the analysis.

The other downside of categorical regression is interpretability in relation of structure to function. Categorical regression is able to inform the user what different ligand combinations may create the best performing catalyst within limitations of the training set, but doesn't provide

insight as to why. Because ligands are one hot encoded, categorical regression alone doesn't provide insight into what properties are useful in each ligand position. To perform this analysis, one would have to measure many different ligand properties like we did earlier in this text. Even once data on each ligand is collected, it must be determined that these properties are actually influential on the catalyst performance, and that the method of measuring these properties is actually relevant to the theorized properties themselves.

The results in a round robin style validation set resulted in Table 3.2. In this analysis only ligands A, B and C were utilized in the categorical regression. It was found that utilization of the C' ligand did not improve out of sample predictions. It is clear from this analysis that the PLS regression performed relatively better than the categorical except for the oxidized 2OH intermediate. Although the span in Gibbs free energy of the oxidized intermediate is larger than that of all other states, the larger error compared to the other states is most likely due to hydrogen bonds being created in this intermediate. The calculation for the oxidized intermediate were not performed in solvation, whereas the other states, ground state, and protonated intermediate were. This resulted in more hydrogen bonds being formed in these calculations. The simplest explanation as to why categorical regression predicts the oxidized intermediate free energies better than PLS is: hydrogen bonding effects are unrelated to our measured predictor properties. It would make sense that pi-accepticity, sigma-donicity, steric bulk, and charge have no predictive value of a hydrogen bond being present in the oxidized intermediate. However a certain ligand may be more likely to hydrogen bond than another, and categorical regression fits the free energies regardless of their chemical origin.

	2OH	TS	2W	TS2	GS2
PLS MSE	12.0	3.7	3.6	4.6	2.5

Categorical MSE	8.2	4.0	5.2	5.4	2.9

Table 3.2. Categorical vs PLS Mean Squared Error (MSE).

Conclusion

This body of work outlines the use of computational chemistry and regression techniques to relate quantified catalyst structure to function in an efficient manner to optimize catalyst behavior and learn chemical lessons along the way. Methane oxidation, more specifically C-H activation, performed by an Ir^{III} catalyst in water is used as a case study. Quantified ligand properties were used to predict FESs which could then be used to compute TOF to quantify catalyst performance.

Monodentate "toy" ligands were utilized in place of more realistic ligands to reduce computational costs, increasing the throughput, as well as to easily modulate electronic and steric properties of catalysts. Computational time and errors in FESs often increase because of the many conformations possible with freely rotating monodentate ligands. These toy ligands were quantified with DFT by optimizing their geometry on $Ru^{II}(NH_3)4Cl$. Properties of interest for the ligands include: sigma-donicity, pi-accepticity, charge, and steric bulk characterized by Ru-Cl bond distance, the sum of dxz+dyz orbital energy, charge, and %vol , respectively. Several methods for each property were initially measured for characterization. However, PCA revealed strong correlation between different ways of measuring any given property. Due to known errors of DFT it was believed there was no reason to introduce unnecessary noise into regressions.

During the process of building the data set to relate catalyst structure to function, both chemical and algorithmic lessons were learned. Chemical intuition can be a good start to building a data set, however having a broader data set allows for more accurate predictions in a broader catalytic space. Principal component analysis was found to be a useful tool to increase

variance of a data set allowing for better predictions across all chemical space. Optimizing towards a given property (adding data to the training set of the predicted best values), allows for chemical lessons to emerge in that chemical space as prediction quality increases even while errors can still be large. Finally, in this case study, it was found that even with relatively large errors, chemical trends were seen in the best catalysts. Neutral ligands impart stability against protonation and oxidation of the catalyst. While it is still unsure of the influence of having strong sigma donors on the C and C' ligand position, it is believed that weak sigma donors in the A and B positions most likely decrease the C-H transition state barrier height, and neutral ligands stabilize the catalyst against oxidation and protonation.

Finally, another similar analysis was performed on the DFT generated FES, utilizing categorical regression with one hot encoding rather than PLSR or elastic net. It was found that categorical regression was found to predict the oxidized intermediate better than PLSR. This is theorized to be likely due to the hydrogen bonds formed during the DFT geometry optimization of the oxidized intermediate, and the fact that these hydrogen bond effects aren't describable by or related to the measured predictor properties.

References

(1) IEA. *CO2 Emissions in 2023*; IEA: Paris, 2024. https://www.iea.org/reports/co2-emissions-in-2023.

(2) Liu, Q.; Wu, L.; Jackstell, R.; Beller, M. Using Carbon Dioxide as a Building Block in Organic Synthesis. *Nat. Commun.* **2015**, *6* (1), 5933. https://doi.org/10.1038/ncomms6933.

(3) Zhang, S.; Fan, Q.; Xia, R.; Meyer, T. J. CO2 Reduction: From Homogeneous to Heterogeneous Electrocatalysis. *Acc. Chem. Res.* **2020**, *53* (1), 255–264. https://doi.org/10.1021/acs.accounts.9b00496.

(4) Hosseini-Benhangi, P.; Gyenge, C. C.; Gyenge, E. L. The Carbon Dioxide Redox Flow Battery: Bifunctional CO2 Reduction/Formate Oxidation Electrocatalysis on Binary and Ternary Catalysts. *J. Power Sources* **2021**, *495*, 229752. https://doi.org/10.1016/j.jpowsour.2021.229752.

(5) Kanega, R.; Ishida, E.; Sakai, T.; Onishi, N.; Yamamoto, A.; Yasumura, H.; Yoshida, H.; Kawanami, H.; Himeda, Y.; Sato, Y.; Ohira, A. An Aqueous Redox Flow Battery Using CO2 as an Active Material with a Homogeneous Ir Catalyst**. *Angew. Chem. Int. Ed.* **2023**, *62* (47), e202310976. https://doi.org/10.1002/anie.202310976.

(6) Taheri, A.; Thompson, E. J.; Fettinger, J. C.; Berben, L. A. An Iron Electrocatalyst for Selective Reduction of CO2 to Formate in Water: Including Thermochemical Insights. *ACS Catal.* **2015**, *5* (12), 7140–7151. https://doi.org/10.1021/acscatal.5b01708.

(7) Heimann, J. E.; Bernskoetter, W. H.; Hazari, N. Understanding the Individual and Combined Effects of Solvent and Lewis Acid on CO2 Insertion into a Metal Hydride. *J. Am. Chem. Soc.* **2019**, *141* (26), 10520–10529. https://doi.org/10.1021/jacs.9b05192.

(8) Heimann, J. E.; Bernskoetter, W. H.; Hazari, N.; Mayer, J. M. Acceleration of CO2 Insertion into Metal Hydrides: Ligand, Lewis Acid, and Solvent Effects on Reaction Kinetics. *Chem. Sci.* **2018**, *9* (32), 6629–6638. https://doi.org/10.1039/C8SC02535E.

(9) Ceballos, B. M.; Yang, J. Y. Highly Selective Electrocatalytic CO2 Reduction by [Pt(Dmpe)2]2+ through Kinetic and Thermodynamic Control. *Organometallics* **2020**, *39* (9), 1491–1496. https://doi.org/10.1021/acs.organomet.9b00720.

(10) Matsubara, Y.; Fujita, E.; Doherty, M. D.; Muckerman, J. T.; Creutz, C. Thermodynamic and Kinetic Hydricity of Ruthenium(II) Hydride Complexes. *J. Am. Chem. Soc.* **2012**, *134* (38), 15743–15757. https://doi.org/10.1021/ja302937q.

(11) Neese, F. The ORCA Program System. *WIREs Comput. Mol. Sci.* **2012**, *2* (1), 73–78. https://doi.org/10.1002/wcms.81.

(12) Neese, F. Software Update: The ORCA Program System, Version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8* (1), e1327. https://doi.org/10.1002/wcms.1327.

(13) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. https://doi.org/10.1021/jp810292n.

(14) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(15) E. F. Valeev. Libint: A Library for the Evaluation of Molecular Integrals of Many-Body Operators over Gaussian Functions, 2024. http://libint.valeyev.net/.

(16) Lehtola, S.; Steigemann, C.; Oliveira, M. J. T.; Marques, M. A. L. Recent Developments in Libxc — A Comprehensive Library of Functionals for Density Functional Theory. *SoftwareX* **2018**, *7*, 1–5. https://doi.org/10.1016/j.softx.2017.11.002.
(17) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.
(18) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.
(19) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297–3305. https://doi.org/10.1039/B508541A.
(20) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110* (13), 6158–6170. https://doi.org/10.1063/1.478522.
(21) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105* (22), 9982–9985. https://doi.org/10.1063/1.472933.
(22) Hay, P. J.; Wadt, W. R. Ab Initio Effective Core Potentials for Molecular Calculations - Potentials for K to Au Including the Outermost Core Orbitals. *J. Chem. Phys.* **1985**, *82* (1), 299–310. https://doi.org/10.1063/1.448975.
(23) Feller, D. The Role of Databases in Support of Computational Chemistry Calculations. *J. Comput. Chem.* **1996**, *17* (13), 1571–1586. https://doi.org/10.1002/(SICI)1096-987X(199610)17:13<1571::AID-JCC9>3.0.CO;2-P.
(24) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. Basis Set Exchange: A Community Database for Computational Sciences. *J. Chem. Inf. Model.* **2007**, *47* (3), 1045–1052. https://doi.org/10.1021/ci600510j.
(25) Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8* (9), 1057–1065. https://doi.org/10.1039/B515623H.
(26) Chase, M. *NIST-JANAF Thermochemical Tables*, 4th ed.; Journal of physical and chemical reference data; American chemical society: Washington, D.C, 1998.
(27) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. The Proton's Absolute Aqueous Enthalpy and Gibbs Free Energy of Solvation from Cluster-Ion Solvation Data. *J. Phys. Chem. A* **1998**, *102* (40), 7787–7794. https://doi.org/10.1021/jp982638r.
(28) Truhlar, D. G.; Cramer, C. J.; Lewis, A.; Bumpus, J. A. Molecular Modeling of Environmentally Important Processes: Reduction Potentials. *J. Chem. Educ.* **2004**, *81* (4), 596. https://doi.org/10.1021/ed081p596.
(29) Creutz, C.; Chou, M. H.; Hou, H.; Muckerman, J. T. Hydride Ion Transfer from Ruthenium(II) Complexes in Water: Kinetics and Mechanism. *Inorg. Chem.* **2010**, *49* (21), 9809–9822. https://doi.org/10.1021/ic101124q.
(30) Wiedner, E. S.; Chambers, M. B.; Pitman, C. L.; Bullock, R. M.; Miller, A. J. M.; Appel, A. M. Thermodynamic Hydricity of Transition Metal Hydrides. *Chem. Rev.* **2016**, *116* (15), 8655–8692. https://doi.org/10.1021/acs.chemrev.6b00168.
(31) Creutz, C.; Chou, M. H. Hydricities of D6 Metal Hydride Complexes in Water. *J. Am. Chem. Soc.* **2009**, *131* (8), 2794–2795. https://doi.org/10.1021/ja809724s.

(32) Chen, S.; Rousseau, R.; Raugei, S.; Dupuis, M.; DuBois, D. L.; Bullock, R. M. Comprehensive Thermodynamics of Nickel Hydride Bis(Diphosphine) Complexes: A Predictive Model through Computations. *Organometallics* **2011**, *30* (22), 6108–6118. https://doi.org/10.1021/om200645x.
(33) Brereton, K. R.; Smith, N. E.; Hazari, N.; Miller, A. J. M. Thermodynamic and Kinetic Hydricity of Transition Metal Hydrides. *Chem. Soc. Rev.* **2020**. https://doi.org/10.1039/D0CS00405G.
(34) Liu, T.; Liao, Q.; O'Hagan, M.; Hulley, E. B.; DuBois, D. L.; Bullock, R. M. Iron Complexes Bearing Diphosphine Ligands with Positioned Pendant Amines as Electrocatalysts for the Oxidation of H2. *Organometallics* **2015**, *34* (12), 2747–2764. https://doi.org/10.1021/om501289f.
(35) Berning, D. E.; Miedaner, A.; Curtis, C. J.; Noll, B. C.; Rakowski DuBois, M. C.; DuBois, D. L. Free-Energy Relationships between the Proton and Hydride Donor Abilities of [HNi(Diphosphine) $_2$] $^+$ Complexes and the Half-Wave Potentials of Their Conjugate Bases. *Organometallics* **2001**, *20* (9), 1832–1839. https://doi.org/10.1021/om0100582.
(36) Jeletic, M. S.; Hulley, E. B.; Helm, M. L.; Mock, M. T.; Appel, A. M.; Wiedner, E. S.; Linehan, J. C. Understanding the Relationship Between Kinetics and Thermodynamics in CO2 Hydrogenation Catalysis. *ACS Catal.* **2017**, *7* (9), 6008–6017. https://doi.org/10.1021/acscatal.7b01673.
(37) Ngo, A. H.; Do, L. H. Structure–Activity Relationship Study of Half-Sandwich Metal Complexes in Aqueous Transfer Hydrogenation Catalysis. *Inorg. Chem. Front.* **2020**, *7* (3), 583–591. https://doi.org/10.1039/C9QI01310E.
(38) Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2* (11), 559–572. https://doi.org/10.1080/14786440109462720.
(39) James, G. M.; Witten, D.; Hastie, T. J.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*, Corrected at 6th printing 2015.; Springer texts in statistics; Springer Springer Science+Business Media: New York, 2013.
(40) Wolcott, C. A.; Medford, A. J.; Studt, F.; Campbell, C. T. Degree of Rate Control Approach to Computational Catalyst Screening. *J. Catal.* **2015**, *330*, 197–207. https://doi.org/10.1016/j.jcat.2015.07.015.
(41) Campbell, C. T. The Degree of Rate Control: A Powerful Tool for Catalysis Research. *ACS Catal.* **2017**, *7* (4), 2770–2779. https://doi.org/10.1021/acscatal.7b00115.
(42) Becke, A. D. Density-Functional Thermochemistry: 3. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652. https://doi.org/10.1063/1.464913.
(43) Dunning, T. H. GAUSSIAN-BASIS SETS FOR USE IN CORRELATED MOLECULAR CALCULATIONS .1. THE ATOMS BORON THROUGH NEON AND HYDROGEN. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023. https://doi.org/10.1063/1.456153.
(44) Balabanov, N. B.; Peterson, K. A. Basis Set Limit Electronic Excitation Energies, Ionization Potentials, and Electron Affinities for the 3d Transition Metal Atoms: Coupled Cluster and Multireference Methods. *J. Chem. Phys.* **2006**, *125* (7), 074110. https://doi.org/10.1063/1.2335444.
(45) Balabanov, N. B.; Peterson, K. A. Systematically Convergent Basis Sets for Transition Metals. I. All-Electron Correlation Consistent Basis Sets for the 3d Elements Sc–Zn. *J. Chem. Phys.* **2005**, *123* (6), 064107. https://doi.org/10.1063/1.1998907.

(46) Weigend, F.; Köhn, A.; Hättig, C. Efficient Use of the Correlation Consistent Basis Sets in Resolution of the Identity MP2 Calculations. *J. Chem. Phys.* **2002**, *116* (8), 3175–3183. https://doi.org/10.1063/1.1445115.
(47) Hill, J. G.; Platts, J. A. Calculating Interaction Energies in Transition Metal Complexes with Local Electron Correlation Methods. *J. Chem. Phys.* **2008**, *129* (13), 134101. https://doi.org/10.1063/1.2982790.
(48) Truhlar, D. G. Basis-Set Extrapolation. *Chem. Phys. Lett.* **1998**, *294* (1), 45–48. https://doi.org/10.1016/S0009-2614(98)00866-5.
(49) Neese, F.; Hansen, A.; Liakos, D. G. Efficient and Accurate Approximations to the Local Coupled Cluster Singles Doubles Method Using a Truncated Pair Natural Orbital Basis. *J. Chem. Phys.* **2009**, *131* (6), 064103. https://doi.org/10.1063/1.3173827.
(50) Back, S.; Kim, H.; Jung, Y. Selective Heterogeneous CO $_2$ Electroreduction to Methanol. *ACS Catal.* **2015**, *5* (2), 965–971. https://doi.org/10.1021/cs501600x.
(51) Muller, R. P.; Philipp, D. M.; Goddard, W. A. Quantum Mechanical-Rapid Prototyping Applied to Methane Activation. *Top. Catal.* **2003**, *23* (1–4), 81–98. https://doi.org/10.1023/a:1024872320512.
(52) Singh, A. R.; Montoya, J. H.; Rohr, B. A.; Tsai, C.; Vojvodic, A.; Nørskov, J. K. Computational Design of Active Site Structures with Improved Transition-State Scaling for Ammonia Synthesis. *ACS Catal.* **2018**, *8* (5), 4017–4024. https://doi.org/10.1021/acscatal.8b00106.
(53) Yeo, B. C.; Nam, H.; Nam, H.; Kim, M.-C.; Lee, H. W.; Kim, S.-C.; Won, S. O.; Kim, D.; Lee, K.-Y.; Lee, S. Y.; Han, S. S. High-Throughput Computational-Experimental Screening Protocol for the Discovery of Bimetallic Catalysts. *Npj Comput. Mater.* **2021**, *7* (1), 137. https://doi.org/10.1038/s41524-021-00605-6.
(54) Nørskov, J. K.; Abild-Pedersen, F.; Studt, F.; Bligaard, T. Density Functional Theory in Surface Chemistry and Catalysis. *Proc. Natl. Acad. Sci.* **2011**, *108* (3), 937–943. https://doi.org/10.1073/pnas.1006652108.
(55) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*, 1st ed.; Wiley, 2003. https://doi.org/10.1002/3527602658.
(56) Jones, C. J.; Taube, D.; Ziatdinov, V. R.; Periana, R. A.; Nielsen, R. J.; Oxgaard, J.; Goddard, W. A. Selective Oxidation of Methane to Methanol Catalyzed, with C-H Activation, by Homogeneous, Cationic Gold. *Angew. Chem.-Int. Ed.* **2004**, *43* (35), 4626–4629.
(57) Young, K. J. H.; Oxgaard, J.; Ess, D. H.; Meier, S. K.; Stewart, T.; Goddard, W. A.; Periana, R. A. Experimental Realization of Catalytic CH4 Hydroxylation Predicted for an Iridium NNC Pincer Complex, Demonstrating Thermal, Protic, and Oxidant Stability. *Chem. Commun.* **2009**, No. 22, 3270–3272. https://doi.org/10.1039/b823303a.
(58) Pahls, D. R.; Allen, K. E.; Goldberg, K. I.; Cundari, T. R. Understanding the Effect of Ancillary Ligands on Concerted Metalation–Deprotonation by (dmPhebox)Ir(OAc)2(H2O) Complexes: A DFT Study. *Organometallics* **2014**. https://doi.org/10.1021/om500752m.
(59) Wang, D. Y.; Choliy, Y.; Haibach, M. C.; Hartwig, J. F.; Krogh-Jespersen, K.; Goldman, A. S. Assessment of the Electronic Factors Determining the Thermodynamics of "Oxidative Addition" of C–H and N–H Bonds to Ir(I) Complexes. *J. Am. Chem. Soc.* **2016**, *138* (1), 149–163. https://doi.org/10.1021/jacs.5b09522.
(60) The "Methanol Economy": General Aspects. In *Beyond Oil and Gas: The Methanol Economy*; 2009; pp 179–184. https://doi.org/10.1002/9783527627806.ch10.

(61) Davies, H. M. L.; Morton, D. Recent Advances in C–H Functionalization. *J. Org. Chem.* **2016**, *81* (2), 343–350. https://doi.org/10.1021/acs.joc.5b02818.
(62) Daugulis, O.; Roane, J.; Tran, L. D. Bidentate, Monoanionic Auxiliary-Directed Functionalization of Carbon–Hydrogen Bonds. *Acc. Chem. Res.* **2015**, *48* (4), 1053–1064. https://doi.org/10.1021/ar5004626.
(63) Reid, J. P.; Sigman, M. S. Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. *Nat. Rev. Chem.* **2018**, *2* (10), 290–305. https://doi.org/10.1038/s41570-018-0040-8.
(64) Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L. Towards the Online Computer-Aided Design of Catalytic Pockets. *Nat. Chem.* **2019**, *11* (10), 872–879. https://doi.org/10.1038/s41557-019-0319-5.
(65) Neese, F. Software Update: The ORCA Program System—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1606. https://doi.org/10.1002/wcms.1606.