Learning Chemical Trends in Heterogeneous Catalysis

Xinyan Liu, Leo Shaw, Charlie Tsai
Department of Chemical Engineering, Stanford University
{xinyanl,leoshaw,ctsai89}@stanford.edu

Abstract

Despite tremendous advances in the understanding of catalysis in the past few decades, the discovery of improved materials for energy transformation processes has been guided by physical intuition and trial-and-error approaches. Even though computational simulations have provided a more rational approach, even the most systematic studies involving the most well-described systems require a large number of costly and repetitive calculations. We attempt to mitigate this problem by proposing a learning model for directly obtaining energetic parameters relevant to catalysis using widely accessible bulk chemical data. Among the methods surveyed, our tree-based models achieved a cross-validation error of about 0.45 eV, which approaches the error obtained from simulations (about 0.2 eV).

1. Introduction and Background

One of the major bottlenecks in developing improved and sustainable energy technologies is the lack of cheap and efficient materials. For applications ranging from batteries to fuel cells to inexpensively produced fertilizers, the materials needed to run these chemical reactions are either prohibitively costly or missing entirely. New materials are typically discovered in the laboratory through trial-and-error, an inefficient and slow process. With significant advances in computing power, the rational design of new materials has the potential to be a more efficient approach, where detailed simulations are performed to help predict new materials. Despite this, however, the successful design of a new and improved material is still a rare occurrence, partly due to the vast amount of calculations needed to adequately sample the space of candidate materials. Thousands of possibilities exist for even the most well-understood systems.

For example, in heterogeneous catalysis a further complication arises from the fact that chemical reactions occur at specific interfaces (i.e. surface terminations of the crystal), for which detailed chemical data usually does not exist. Vast quantities of calculations specific to each surface of each material are needed. Although specific data for material surfaces are lacking, properties for bulk materials are widely available through several large databases. In this project, we train a model that bridges the gap between surface phenomena and bulk properties, which do not contain information about the surface and are poor at directly describing the chemical reactivity of materials. Bulk properties can be combined with *a priori* information about the surface, such as the crystallographic planes, and training data from existing computations and experiments. A model can then be developed to directly predict surface properties from widely available bulk data without the need for computationally expensive quantum chemical calculations. In terms of surface properties, we will be focusing on thermochemical data, which most directly describes how efficient a material will be for a given chemical process. As the possible predictors are numerous, variable selection will be used to determine their relative importance.

2. Related Work

Mean-field microkinetic modeling has long been a powerful and efficient means of analyzing catalytic reactions.[1] Inputs related to the energetics on the catalyst surface are needed in order to solve the rate equations and determine important metrics like the turnover frequency of the desired product. In order for predictions to be made about the reaction rates, the solutions to the rate equations need to be expressed as a function of the energetic inputs. Since a vast number of possible mechanisms may be involved, the relation among the different energetic inputs needs to be known in order to reduce the parameter space and make the problem tractable.[2] The adsorption strengths of different reaction species has been found to scale linearly with one another on a wide range of heterogeneous as well as homogeneous catalysts.[3–5] As long as the linear relations are known, the rate can be determined as a function of the binding energies for a few species. Simple bond-counting principles have been used to determine the slopes of these linear scaling lines, but calculations are still needed to fit the intercept. An improved catalyst can then be identified by finding the required binding energies for maximizing the rates. Although this ”descriptor-based” analysis has led to the discovery of many improved catalysts, they have so far been restricted to the simplest types of systems and for the most well-studied reactions. Furthermore, detailed calculations related to each catalytic site need to be made for a single prediction. Even the simplest transition metal nanoparticles have a multiplicity of sites and surface terminations. Recent efforts have focused on understanding the role of surface coordination number[6] in determining adsorption strength but costly calculations are still required to make the predictions. Most efforts are concerned with detailed simulations of known active sites, and there is so far no model for predicting the relevant energetics using only *a priori* information about the surface and the material. If reaction energetics at the interface can be directly predicted, then the inputs needed for estimating reaction rates can be obtained at negligible computational cost.

3. Data Set and Features

A significant simplification in describing catalytic activity arises from the fact that all reaction steps are related to the stabilities of the reactant species on the catalytic surface. These are usually determined as adsorption energies – the energy required to stabilize the adsorbate molecule on the catalyst, or how strongly the catalyst binds the adsorbate. We retrieved data for the reaction energies from the CatApp database[7] from the SUNCAT research group at Stanford University and the bulk properties of the catalysts were taken from the Materials Project database.[8] To simplify the initial evaluation of models, we've restricted the catalysts to transition metals and their binary alloys. For the chemical reactions, we considered elementary reactions of the form AB → A + B. There were approximately 2000 data points for these reaction energetics on transition metals and their binary alloys. These results can be directly used in kinetic models describing catalytic reactions such as higher alcohols conversion and CO_2 reduction.

3.1. Response

The response variables are either the reaction energy ΔE_{rxn}, which is the energy difference between the final and initial state of each chemical reaction step, or the activation energy ΔE_a, which is the reaction barrier associated with each reaction step. These parameters can be directly plugged into kinetic rate equations to determine the activity of a given catalytic site. For all reactions, the magnitude of the response is typically between -2 eV and 2 eV. State-of-the-art density functional theory calculations are typically accurate within 0.2 eV of the experimental values, so our model should have at most a generalized error of that order of magnitude. Otherwise, it could not be a viable alternative to the calculations.

Index	Predictor	Type	Units
0	Miller Index h	Discrete	–
1	Miller Index k	Discrete	–
2	Miller Index l	Discrete	–
3	Stoichiometry for Metal 1	Discrete	–
4	Stoichiometry for Metal 2	Discrete	–
5	Energy of Formation	Continuous	eV
6	Density	Continuous	g/cm^3
7	Unit Cell Length a	Continuous	Å
8	Unit Cell Length b	Continuous	Å
9	Unit Cell Length c	Continuous	Å
10	Unit Cell Angle α	Continuous	degrees
11	Unit Cell Length β	Continuous	degrees
12	Unit Cell Length γ	Continuous	degrees
13	Metal 1 s Electrons	Discrete	–
14	Metal 1 p Electrons	Discrete	–
15	Metal 1 d Electrons	Discrete	–
16	Metal 1 f Electrons	Discrete	–
17	Metal 2 s Electrons	Discrete	–
18	Metal 2 p Electrons	Discrete	–
19	Metal 2 d Electrons	Discrete	–
20	Metal 2 f Electrons	Discrete	–
21	Max Adsorbate Bonds (AB)	Discrete	–
22	Adsorbate Bonds (AB)	Discrete	–
23	Intramolecular Bonds (AB)	Discrete	–
24	# C atoms (AB)	Discrete	–
25	# H atoms (AB)	Discrete	–
26	# O atoms (AB)	Discrete	–
27	# N atoms (AB)	Discrete	–
28	Max Adsorbate Bonds (A)	Discrete	–
29	Adsorbate Bonds (A)	Discrete	–
30	Intramolecular Bonds (A)	Discrete	–
31	# C atoms (A)	Discrete	–
32	# H atoms (A)	Discrete	–
33	# O atoms (A)	Discrete	–
34	# N atoms (A)	Discrete	–
35	Max Adsorbate Bonds (B)	Discrete	–
36	Adsorbate Bonds (B)	Discrete	–
37	Intramolecular Bonds (B)	Discrete	–
38	# C atoms (B)	Discrete	–
39	# H atoms (B)	Discrete	–
40	# O atoms (B)	Discrete	–
41	# N atoms (B)	Discrete	–

Table 1. Summary of physical parameters describing the crystal structure, surface termination, and molecular identity of the reactants.

3.2. Predictors

Our choice of 42 predictors was based on both the availability of bulk material data and properties suspected of being physically and chemically important during a catalytic reaction. The identity of the surface (i.e. the specific crystallographic plane) involved with the reaction is encoded with the discrete integers of the corresponding Miller indices (0-2). For binary alloys, the stoichiometry (3-4) and number of electrons (13-20) uniquely identify the two individual elements and the overall composition. For pure metals, the metal's values were duplicated for metal 2 to maintain the same number of features. The crystalline structure of a catalyst is represented by the 3 lattice constants

(7-9) and angles (10-12) that uniquely describe the crystallographic unit cell of the material. To identify the molecule involved in the reaction, we have used the maximum number of bonds in the atom closest to the surface (based on bond-order analyses from previous work[3]), the number of bonds directly made to the surface, as well as those made within the molecule (21-23, 28-30, 35-37). This must be done for the initial state (AB) and the final states (A and B). Furthermore, adsorbates could bind to the surface through the same type of atom, but have different chain lengths or other constituents further from the surface. We encoded these possibilities by explicitly including the number of each type of element present (24-27, 31-34, 38-41).

4. Methods

For our regression problem, we chose (1) linear regression methods, (2) tree-based methods, and (3) kernel-based methods. This range of flexibilty allowed us to negotiate the bias-variance trade-off inherent in constructing a useful model.

Regularization and feature selection are important in removing features with little predictive value. For example, LASSO regression and ridge regression are modifications of simple linear regression whereby the cost function is supplemented with a penalty on the coefficients themselves. Specifically, the model coefficient vector β is

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \|\beta\|_k$$

where $k = 2$ for ridge, $k = 1$ for LASSO, and the tuning parameter λ is chosen by cross-validation. The left term is the residual sum of squares (RSS), and the term on the right – the "shrinkage penalty" – causes the coefficients to tend toward zero for large values of λ, with less important features having smaller coefficients than more important ones.

Subset selection can also be used to determine which predictors are most relevant. Rather than evaluate models for all 2^{42} possible subsets of the predictors, we chose backward stepwise selection. The initial set of model predictors M_0 is the entire set P, and new sets M_i for each subsequent iteration removes one predictor by evaluating the models resulting from the removal of that predictor. The optimal model gives us the set of the most important predictors.

Tree-based methods allow for regularization as well. Decision trees partition the p-dimensional predictor space into n distinct, non-overlapping boxes, whereby splits in a predictor domain are evaluated by calculating the RSS among the set of all splits for all predictors. In other words, for predictor X_i and cutpoint j, a split creates the half-planes

$$R_1 = \{X | X_i < j\}, R_2 = \{X | X_i \geq j\}$$

For each split, we find

$$\underset{i,j}{\operatorname{argmin}} \sum_{k: x_k \in R_1(i,j)} (y_k - \hat{y}_{R_1})^2 + \sum_{k: x_k \in R_2(i,j)} (y_k - \hat{y}_{R_2})^2$$

where $\hat{y}_{R_m}$ is the prediction for region R_m, i.e. the mean of the training observations in that region. This process is done iteratively, except each round after the first produces a split not in the whole predictor space, but only in one of the regions. The most important features can be determined from this process.

Several methods have been developed to improve upon this basic algorithm. For example, Boosting involves slowly creating a model by sequentially adding trees in order to avoid overfitting. Bagging involves creating bootstrap training sets – sets of random samples (with replacement) from the original set of training points – and then growing decision trees, which are averaged together. Random forest methods were developed as a way to improve the performance of bagged decision trees by specifically decorrelating the individual boot-strapped trees by shrinking the set of p predictors considered at each split – typically $\sqrt{p}$ predictors are used.

Lastly, we used support-vector regression (SVR)[9] to model our data. An extension of support-vector classification, the method with soft margins solves for w and b:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right)$$

with the following conditions on the tuning parameter C, tolerance ϵ, and slack variables ξ_i and ξ_i^*:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

5. Results

5.1. Learning Curves

To evaluate each model's performance, the training error and cross-validation error were determined as a function of the training set size. These results are summarized in Fig. 1. Starting with linear regression, the least flexible model, both the training error and cross-validation errors converge before the full training set is used (at around 400 samples), indicating that increasing the training set size will not further improve the model accuracy. The CV error stabilizes at about 1.0 eV and is much too large, suggesting that either more flexibility in the model are needed. In the random forest model, the training error is in an acceptable range for larger training sets, while the cross-validation error is at

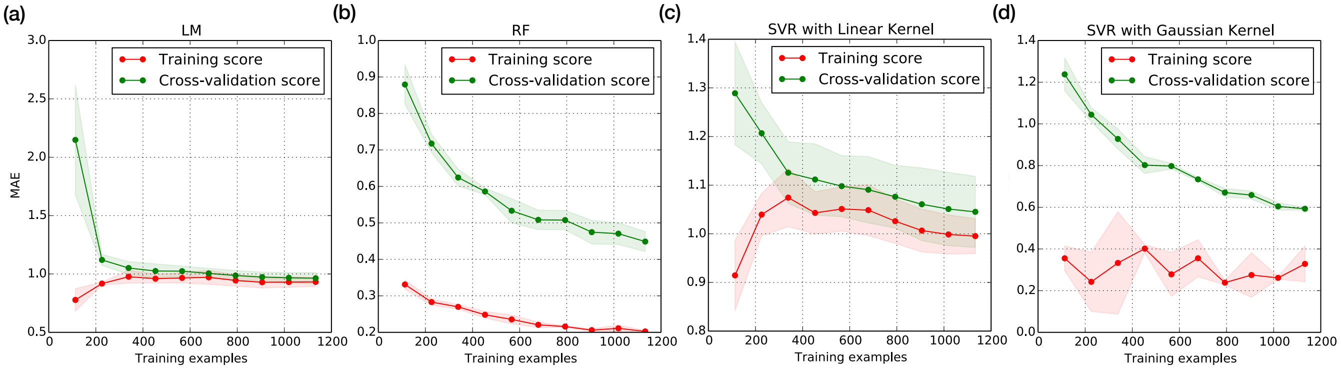

Figure 1. Learning curves for the linear regression (LM), random forest (RF), and support vector regression (SVR) with a linear and Gaussian kernel. The training error and the cross validated generalization error is shown as a function of the training set size.

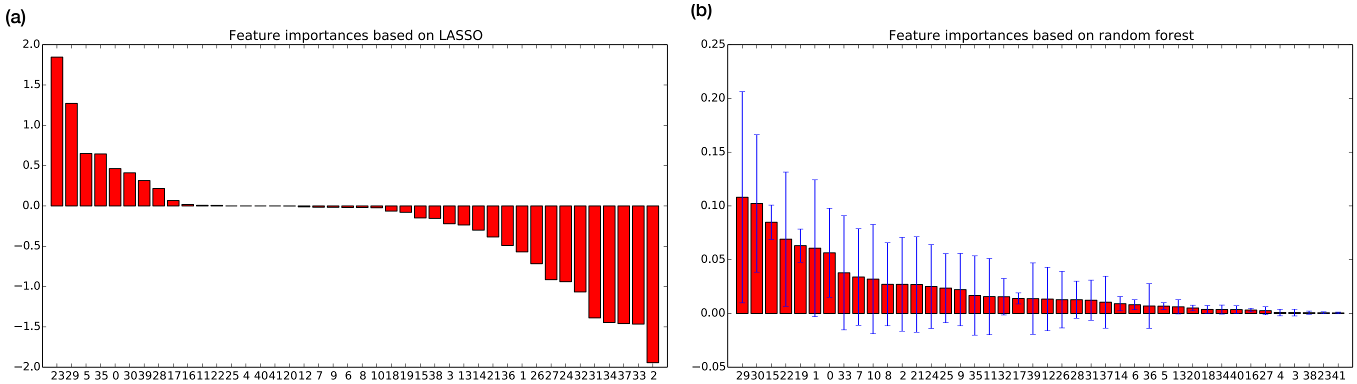

Figure 2. (a) size of the coefficients from a linear fit using the L1-norm penalty; (b) Ranking of the importance of the various predictors in the random forest model. The numbering of the features corresponds to Table 1.

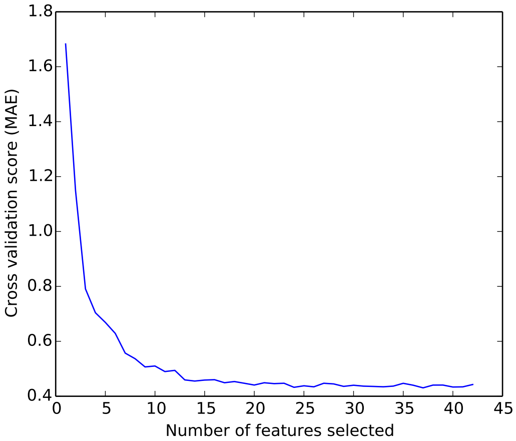

Figure 3. Backward subset selection for the features used in the Random Forest model.

about 0.45 eV when all data points are used. For SVR, a linear and a Gaussian kernel were chosen. The training errors are about 1.0 and 0.3 eV respectively, with the linear kernel performing similarly overall to the linear regresssion. The CV error for the Gaussian kernel is about 0.6 eV and, like the RF model, could be improved with additional samples to decrease the generalization error further.

5.2. Feature Selection and Model Performance

We applied several approaches for performing feature selection on parametric and non-parametric models: a norm penalty for the linear model (i.e. the LASSO and ridge regression) and comparing the RSS decrease of each variable in the tree-based method (random forest). Both the LASSO and ridge regression yield comparably small yet non-zero coefficients for many of variables. The training and cross validation errors were typically worse than standard linear regression, suggesting that these regularization methods may not be appropriate for feature selection.

Since random forest already shows reasonable performance on the training data (Fig. 1), it is used as a means of determining the importance of each feature. The results shown in Fig. 2 suggest that (29) adsorbate bonds of A to the surface, (30) the intramolecular bonds in A, (15, 19) the d- electrons for each metal of the catalyst, (22) adsorbate bonds of the initial material AB, and (0, 1) the Miller indi-

Table 2. Summary of Model Performances

Methods			Cross Validation MAE (eV)	
			w/o feature selection	with feature selection
Linear Regression (Regularization)	Linear Regression		0.96	1.11
	LASSO		0.96	1.15
	Ridge		0.96	1.16
Tree-based Methods	Random Forest		0.46	–
	Bagging		0.45	0.54
	Boosting		0.57	0.59
Kernel-based Methods	Kernel Ridge Regression	Linear Kernel	0.96	–
		Polynomial Kernel	0.73	–
		Gaussian Kernel	0.59	0.65
	Support Vector Regression	Linear Kernel	1.05	–
		Polynomial Kernel	1.41	–
		Gaussian Kernel	0.59	0.56

cies h and k are the seven most important features. There is a clear drop in importance for the remaining variables. This result agrees with known physical concepts in catalysis: the magnitude of the adsorption energy is determined mostly by the bonds of the reactant molecule, the d-electrons are the most important in correlating binding strength with the identity of the metal catalyst, and the reaction energies are a strong function of the stabilization of reactants via surface bonding. The larger errors bars unfortunately indiciate that the uncertainties associated with each of the feature importance metrics can change the relative importance of these predictors. Many of the features may be redundant or non-negligibly correlated. For example, the large variance for the importance of the lattice parameters and angles (7-12) of the catalyst is likely due to the similar influence they each have on the models. However, a few predictors, such as the metal s electrons (13, 17) and the energy of formation of the metal (5) are reliably unimportant. Backwards recursive feature selection with cross-validation was also performed to corroborate our results. We simulated the feature ranking with recursive feature elimination and cross-validated selection of the best number of features. In Fig. 3, the cross-validation error curve decreases monotonically until about 7 predictors are selected, at which point the error plateaus. The predictors are in good agreement with the feature importance plot from before. In summary, only a few of the features have a large influence on the response, and they also have relevant physical significance.

The performance of all models are summarized in Table 2. Feature selection for the tree and kernel-based models used the features selected by the random forest. Only features whose importance is greater than or equal to the mean feature importance were kept. In almost all cases, the CV error with feature selection performed mostly slightly worse than for the models with all predictors. Again, all forms of linear regression performed poorly, while the Gaussian kernel performed better for the other methods. The tree-based methods were still able to achieve the lowest CV error, with the random forest and bagging having a MAE less than 0.5 eV.

6. Conclusions and Future Work

As a first step, our work demonstrates that it is possible to make predictions on surface phenomena using *a priori* information about the surface in addition to bulk chemical data. The regularization analyses confirm the importance of certain parameters (i.e. the d-electrons, the bond-order of the adsorbates, etc.) established in previous physical models. We have so far restricted our data set to binary transition-metal alloys and elementary reactions of the form $AB \rightarrow A + B$. Even by considering just C, N, H, and O containing species, the results account for the vast majority of catalytic reactions of industrial importance. However, more features would be needed to generalize our models to other catalytic reactions or materials other than transition metals. Currently, only the most flexible models are able to reach the target accuracy of ≤ 0.2 eV, and only for the training set. Additional data can be obtained, but they will also be increasingly dissimilar from the current data set, as we have focused on the most consistent systems in this study. Beyond increasing the number of features, performing regularization, and increasing the number of training examples, course grain calculations may be needed as an additional inputs. Calculations can be obtained at a negligible computational cost if the accuracy is sufficiently low. However, this would provided additional structure for the underlying data.

References

[1] Honkala, K.; Hellman, A.; Remediakis, I. N.; Logadóttir, Á.; Carlsson, A.; Dahl, S.; Christensen, C. H.; Nørskov, J. K. *Science* **2005**, *307*, 555–558.

[2] Medford, A. J.; Lausche, A. C.; Abild-Pedersen, F.; Temel, B.; Schjødt, N. C.; Nørskov, J. K.; Studt, F. *Top. Catal.* **2013**,

[3] Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. *Phys. Rev. Lett.* **2007**, *99*, 016105.

[4] Fernández, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martínez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T. *Angew. Chem.* **2008**, *120*, 4761–4764.

[5] Bligaard, T.; Nørskov, J. K.; Dahl, S.; Matthiesen, J.; Christensen, C. H.; Sehested, J. *J. Catal.* **2004**, *224*, 206–217.

[6] Calle-Vallejo, F.; Loffreda, D.; Koper, M. T. M.; Sautet, P. *Nat. Chem.* **2015**, *7*, 403–410.

[7] Hummelshøj, J. S.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nørskov, J. K. *Angew. Chem.* **2012**, *124*, 278–280.

[8] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013**, *1*, 011002.

[9] Smola, A. J.; Schölkopf, B. *Statistics and Computing* **2004**, *14*, 199–222.