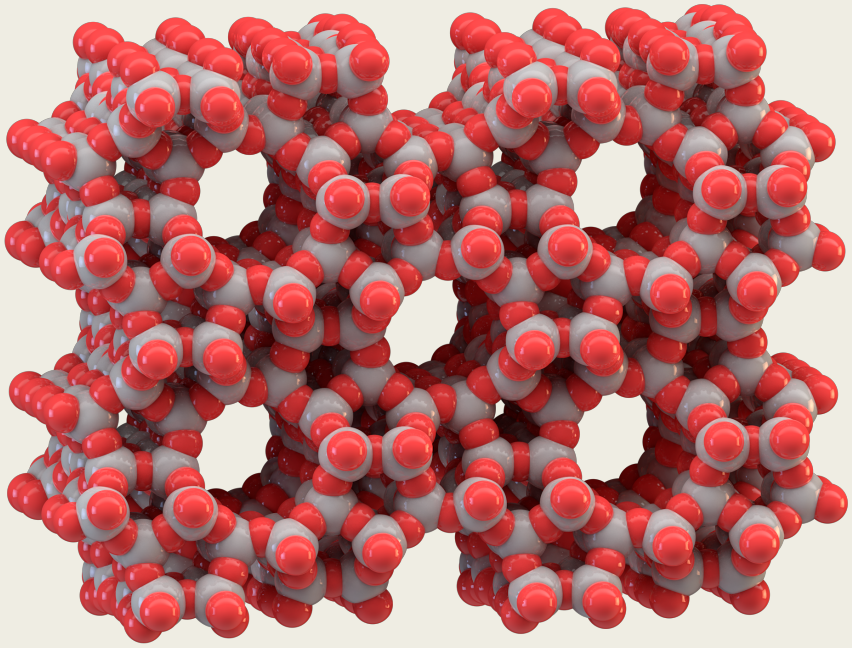
LEARNING CHEMICAL TRENDS IN HETEROGENEOUS CATALYSIS



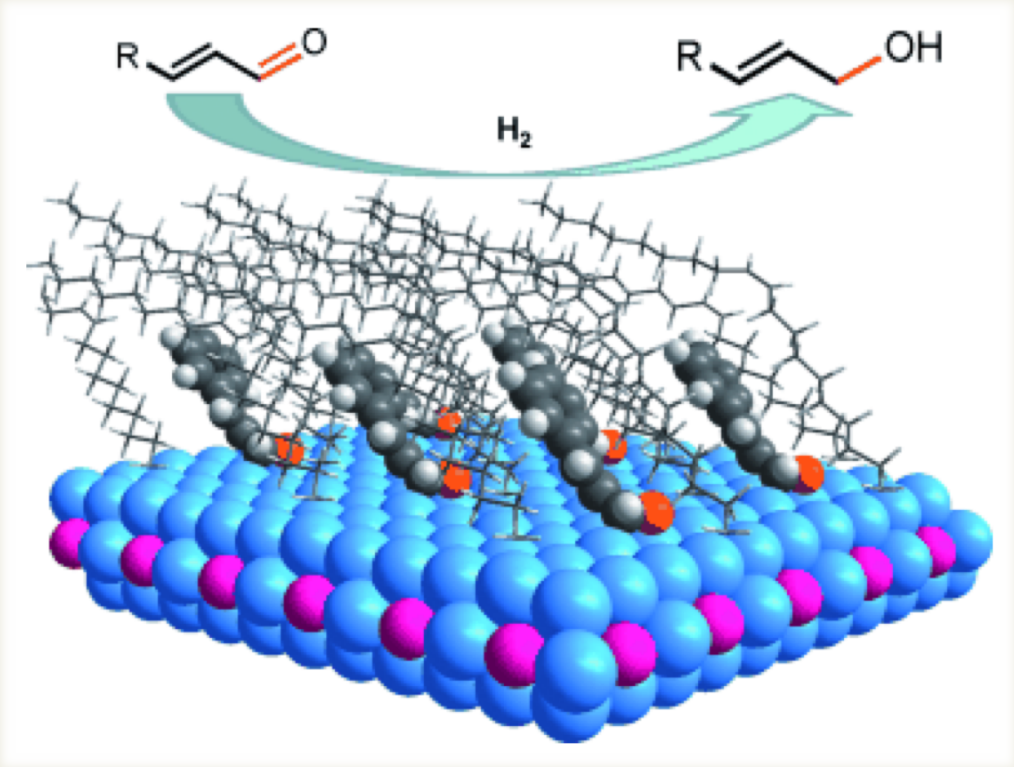
Xinyan Liu, Leo Shaw, Charlie Tsai
Department of Chemical Engineering, Stanford University

December 8, 2015

Abstract

- Despite enormous advances in computational power, the discovery of improved materials for catalytic reactions is still a rare event.
- Even the most systematic studies involving the most well-described systems require a large number of costly and repetitive calculations.
- Herein, we propose a learning model for obtaining energetic parameters relevant to catalysis using widely accessible bulk chemical data drawn from CatApp and the Materials Project.
- Tree-based methods (bagging and random forest) produced the best fit
- SVR could potentially be improved with larger datasets

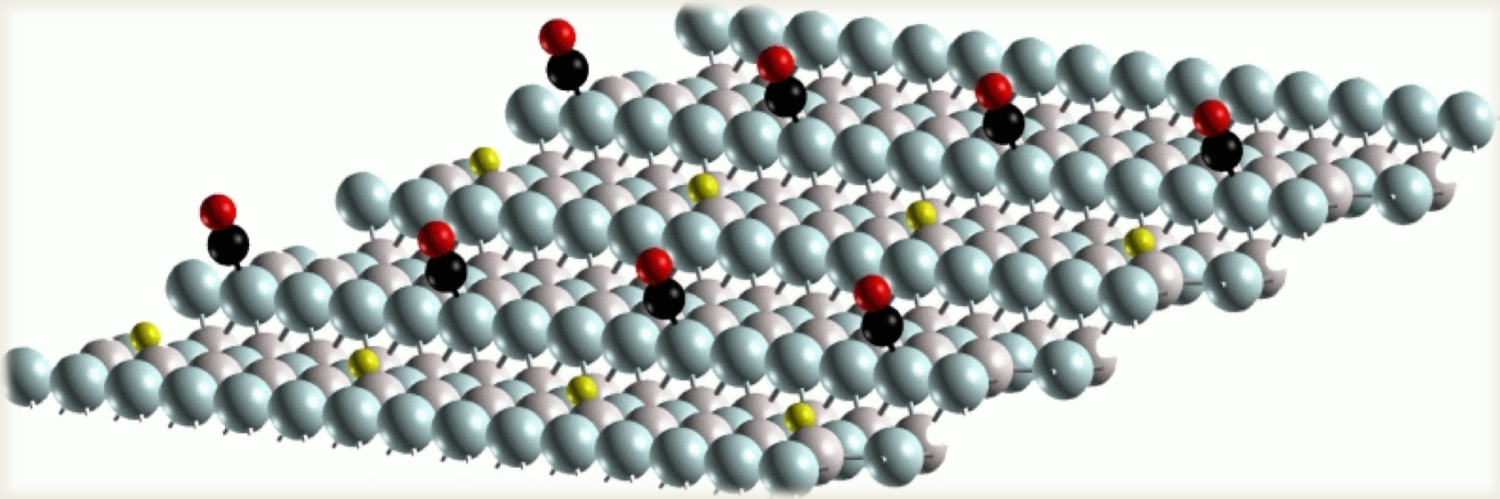
Motivation



- Improved, sustainable energy technologies suffer from a lack of cheap and efficient materials.
- Laboratory discovery is an inefficient and slow process.
- Detailed simulations can be performed to help predict new materials, but a vast number of calculations is needed.
- Statistical methods and learning can reduce the need for expensive computations by providing accurate estimates of desired quantities.

Heterogeneous Catalysis

- Some solid materials are able to enhance the rate of a chemical reaction occurring in a contacting phase.
  - *These catalysts have important applications in renewable energy, the petrochemical industry, agriculture, and chemical synthesis*
- Chemical reactions occur at specific interfaces for which detailed chemical data usually does not exist.
- Although data for surfaces are lacking, properties for bulk materials are available in several large databases.
- Bulk properties do not contain information about the surface and are poor at directly describing the chemical reactivity of materials.
- In this project, we propose to train a model that bridges the gap between bulk properties and surface phenomena.

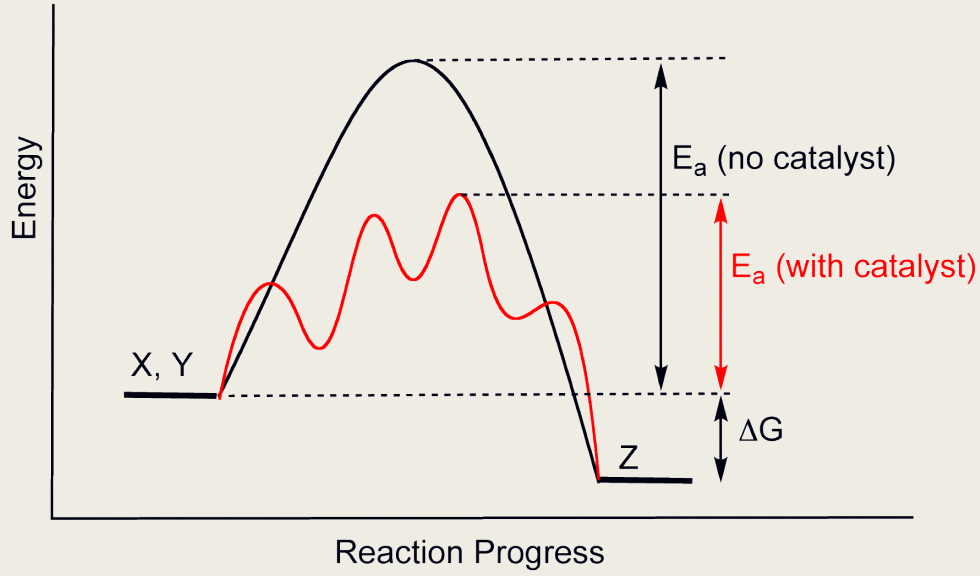


Methods and Input Data

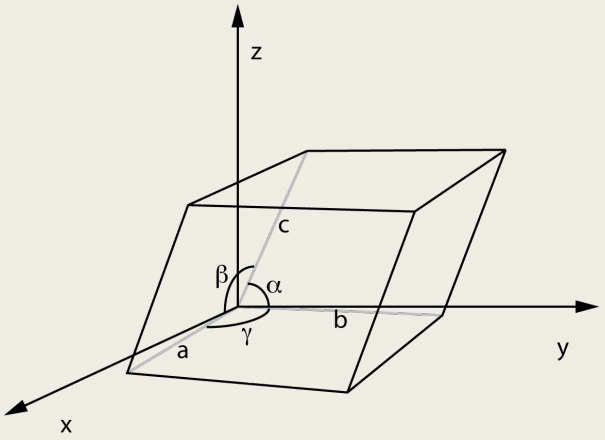
- All reaction steps are related to the stabilities of the reactant species on the catalytic surface.
  - *Usually determined as adsorption energies -- the energy required to stabilize the adsorbate molecule on the catalyst, or how strongly the catalyst binds the adsorbate.*
- Reaction energy data: the CatApp database from SUNCAT
- Bulk properties of the catalysts: the Materials Project
- To simplify the initial evaluation of models, we've restricted the catalysts to transition metals and their binary alloys.
- For the chemical reactions, we consider the adsorption and reaction energies involving C, N, H, and O containing compounds (~2000 data points)

Response

- The response variables are either:
  - *the reaction energy ΔE_{rxn}: the energy difference between the final and initial state*
  - *the activation energy ΔE_a: the reaction barrier*
- These parameters can be directly plugged into kinetic rate equations to determine the activity of a given catalytic site.
- The response is typically between -2 and 2 eV.
- 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.

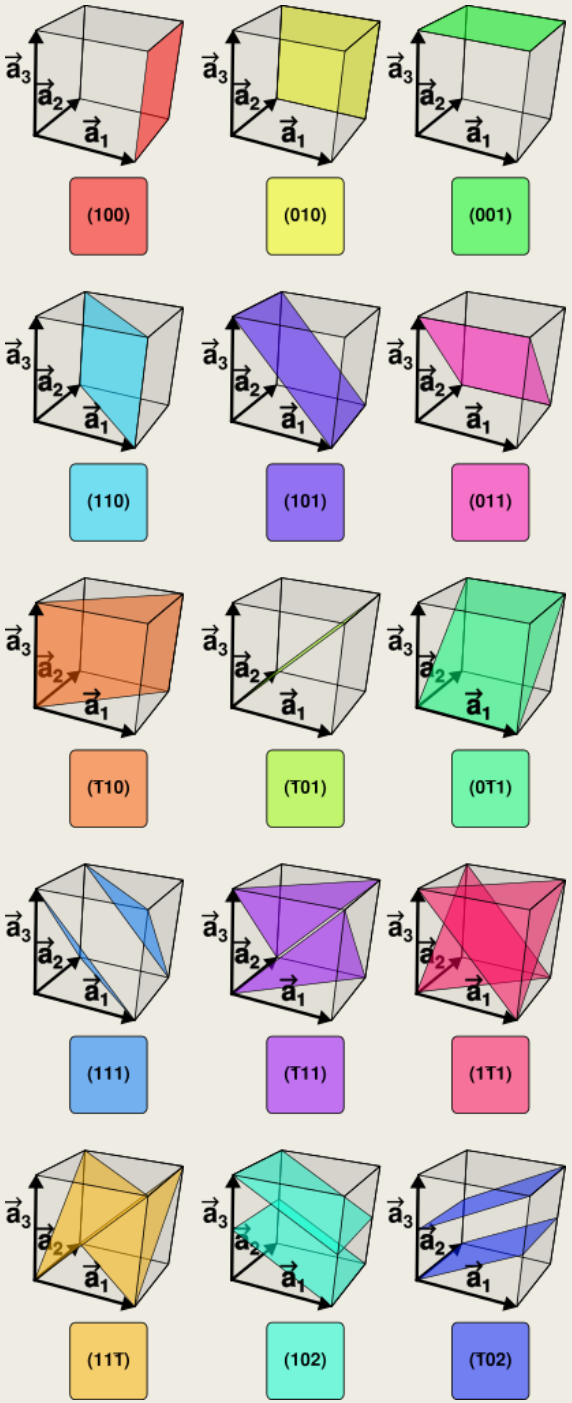


Predictors

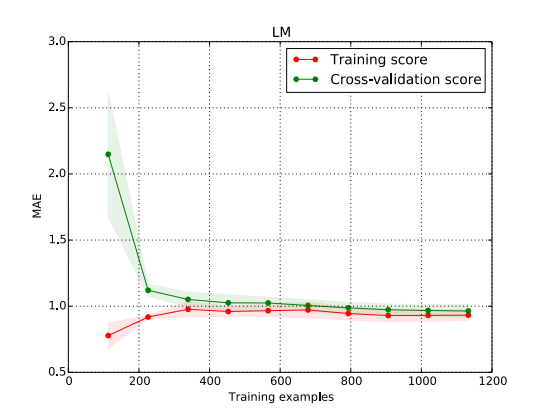


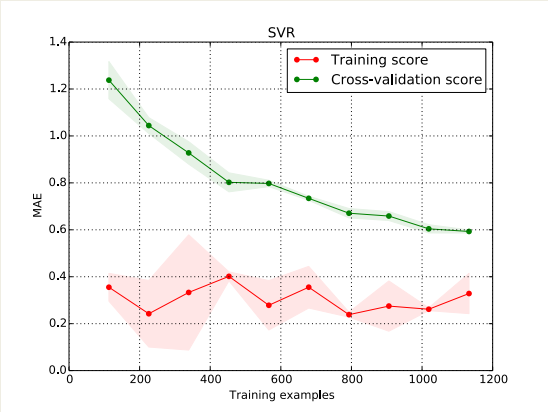
- 23 factors physically and chemically important during a catalytic reaction:
  - *Identity of surface*
  - *Identity of material*
  - *Crystalline structure of catalyst*
  - *Identity of molecule involved in the reaction*

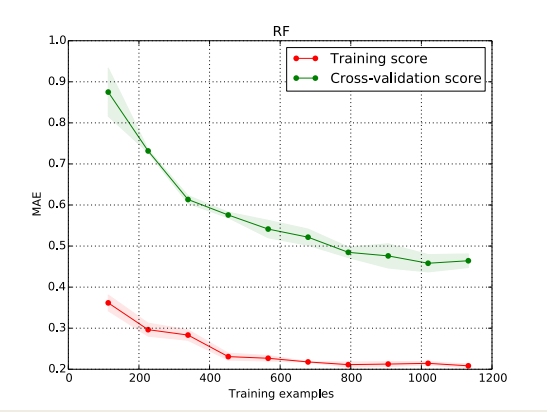
	Predictor	Type
1	Miller Index *h*	Discrete
2	Miller Index *k*	Discrete
3	Miller Index *l*	Discrete
4	Stoichiometry for Metal 1	Discrete
5	Stoichiometry for Metal 2	Discrete
6	Energy of Formation	Continuous
7	Density	Continuous
8	Unit Cell Length *a*	Continuous
9	Unit Cell Length *b*	Continuous
10	Unit Cell Length *c*	Continuous
11	Unit Cell Angle α	Continuous
12	Unit Cell Length β	Continuous
13	Unit Cell Length γ	Continuous
14	Metal 1 s Electrons	Discrete
15	Metal 1 *p* Electrons	Discrete
16	Metal 1 *d* Electrons	Discrete
17	Metal 1 *f* Electrons	Discrete
18	Metal 2 s Electrons	Discrete
19	Metal 2 *p* Electrons	Discrete
20	Metal 2 *d* Electrons	Discrete
21	Metal 2 *f* Electrons	Discrete
22	Max Adsorbate Bonds	Discrete
23	Adsorbate Bonds	Discrete



Learning Curves





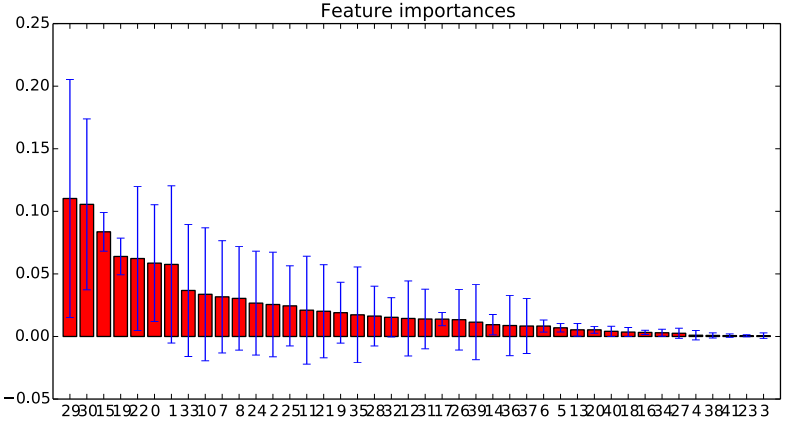


- Training error and cross-validation error were determined as a function of the training set size.

- Linear regression (the least flexible model)
  - *Both errors converge before using the full training set*
  - *Errors are much too large in both cases*
- Support vector regression (with a Gaussian kernel)
  - *The training error is close to the target 0.2 eV, but the cross-validation error is very large*
  - *Errors do not converge for our training set size.*
  - *Model is over-fitting the data.*
- Random forest
  - *Training error is again in an acceptable range, while the cross-validation error is slightly higher.*
  - *Like the SVR model, increasing the number of training examples could help continue to decrease the generalization error further.*

Feature Selection

- Two approaches
  - *An L_1-norm penalty for the linear model (i.e. the LASSO)*
    - Yielded comparably small yet non-negative and non-zero coefficients for the majority of variables
    - Training and cross validation errors were similar or worse than standard linear regression.
    - The LASSO may not be appropriate for feature selection.
  - *RSS decrease to quantify variable importance in the tree-based method (random forest)*
    - Results suggest that the following are the 6 more important predictors:
      - *Bond order*
      - *Total bonds to the adsorbate*
      - *d-electrons for each metal*
      - *Surface termination*
      - *Stoichiometry of the first metal*



    - There is a large drop in importance for the remaining variables.
    - This result agrees with known physical concepts in catalysis
    - Uncertainty associated with feature importance of one lattice parameter, stoichiometry of the first metal, and stoichiometry of the second metal is relatively large most likely due to the binary-nature of the features
    - Many of the features are redundant so far because of the similar catalytic materials considered so far.

Conclusions

Methods		Cross Validation Error (eV)	
		w/o feature selection	with feature selection
Linear Regression (Regularization)	Linear Regression	0.96	/
	LASSO	0.96	
	Ridge	0.96	
Tree-based Methods	Random Forest	0.46	0.54
	Bagging	0.45	
	Boosting	0.57	0.59
Kernel-based Methods	Kernel Ridge Regression	0.62	0.55
	Support Vector Regression	0.59	0.56

- In this project, we used machine learning techniques to bridges the gap between bulk properties and surface phenomena.
- Tree-based methods exhibit the best performances
- Feature selection is shown to moderately improve kernel-based methods.
- Predictions can be made, but the error still needs to be improved