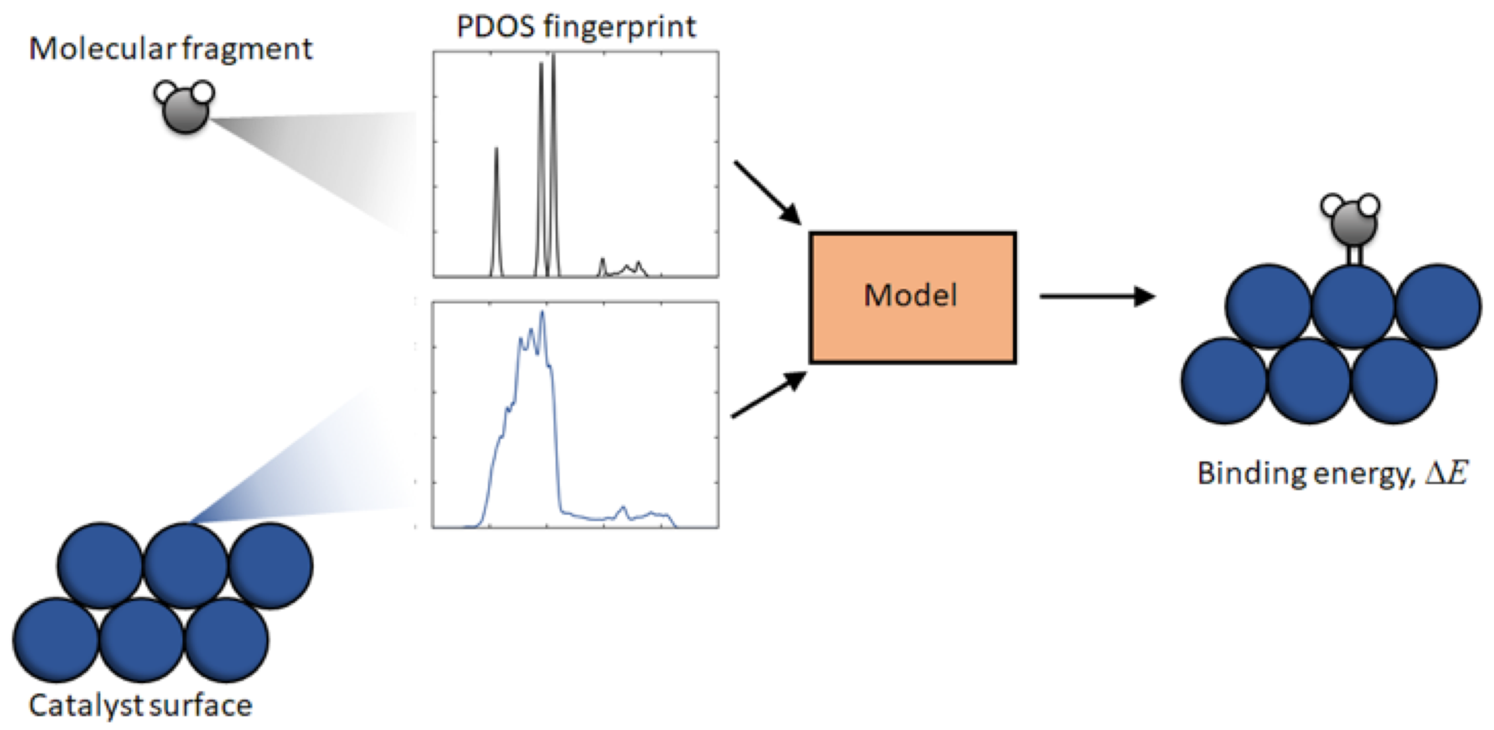
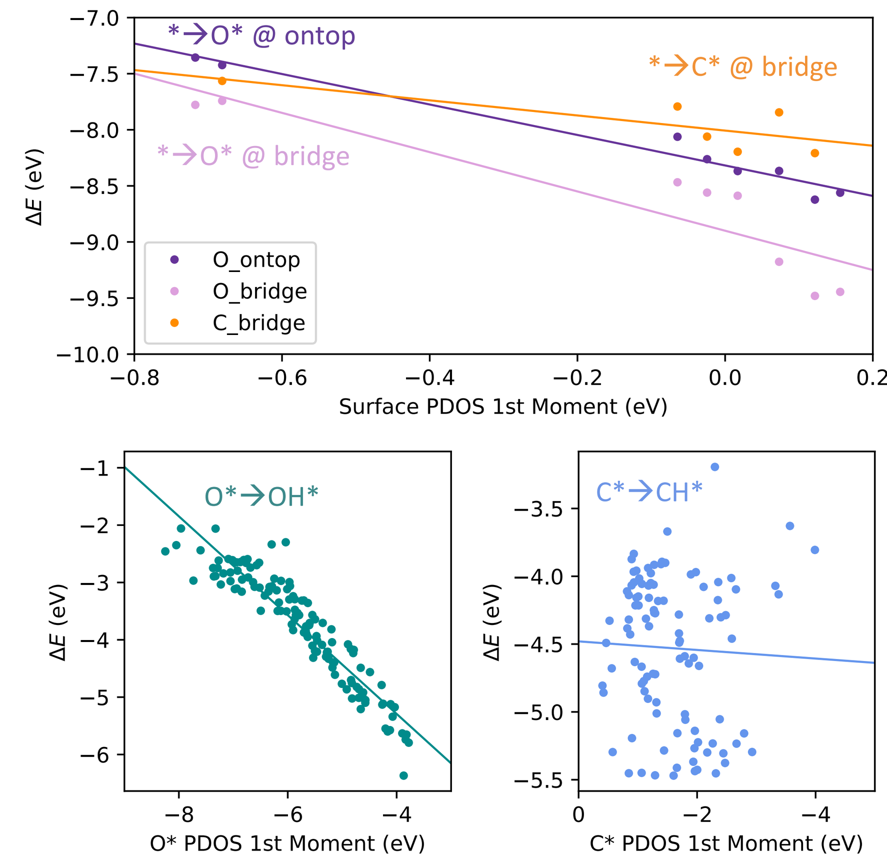
Learning Chemistry from Moment to Moment

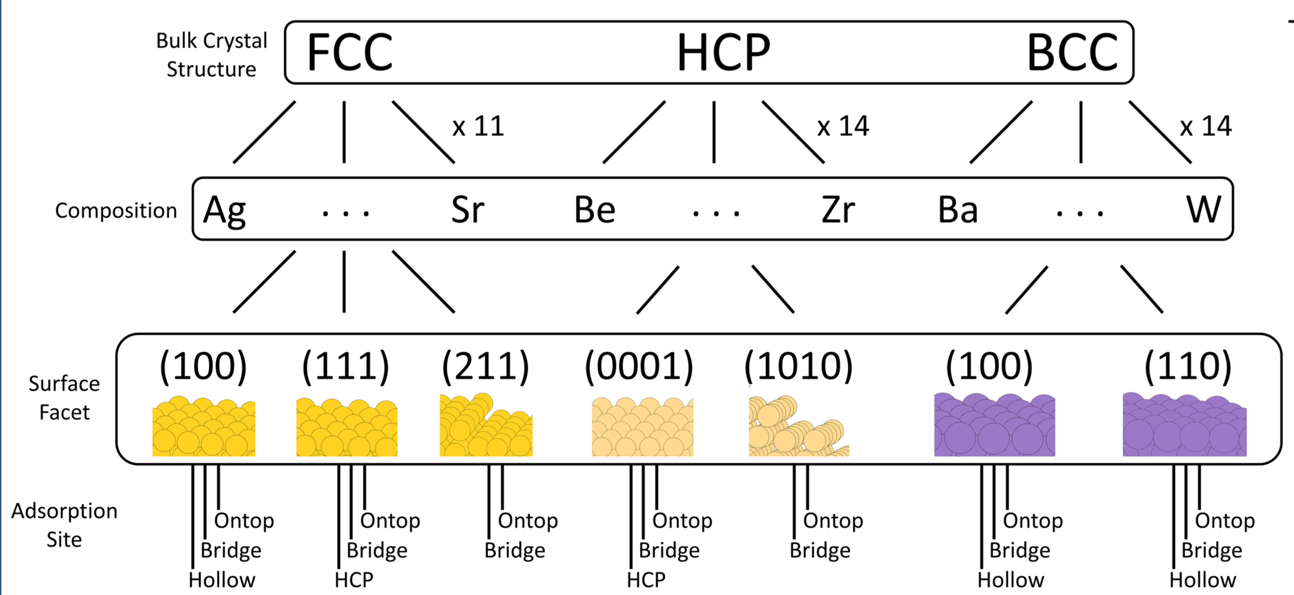
Colin Forest Dickens and Allegra Latimer
Machine Learning (CS229) Final Project: December 11, 2018

Motivation





Data Generation

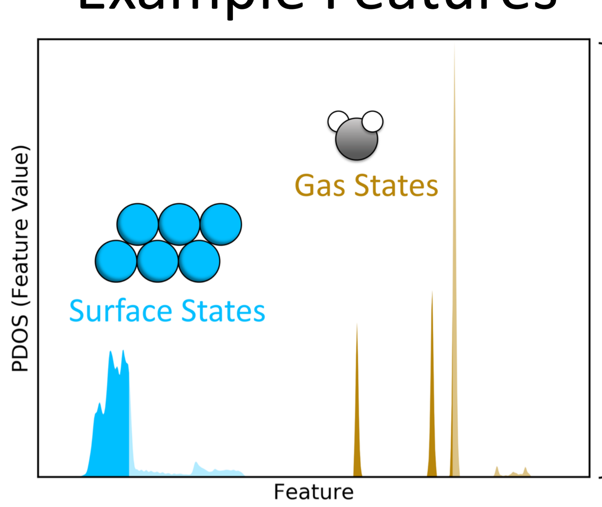


242 Unique Surface Sites
×
17 Unique Surface Reactions
=
4114 Possible Reactions

cleaning

2000 Considered Reactions

Example Features

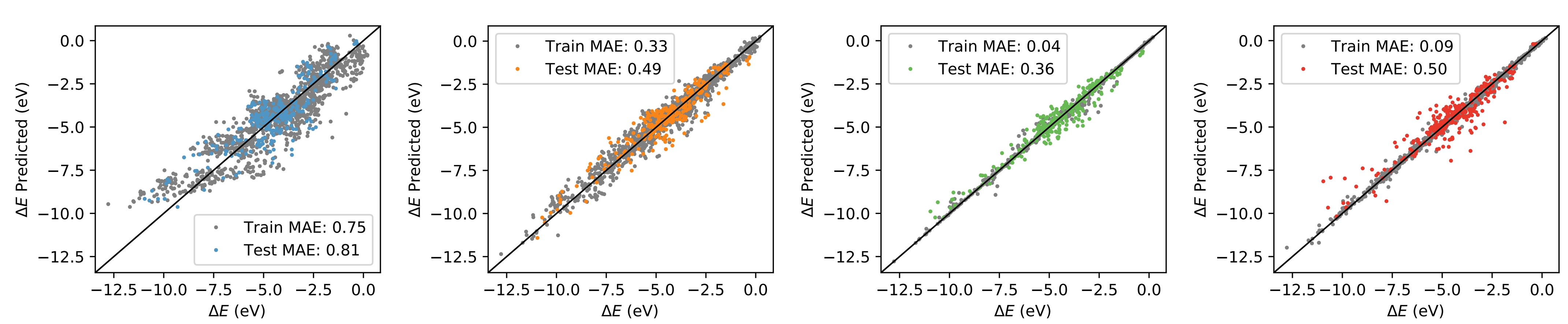


PDOS Moment	Surface	Gas
0	9.5	6.8
1	-0.2	-2.0
2	72.3	251.5
3	-973.2	1057.4

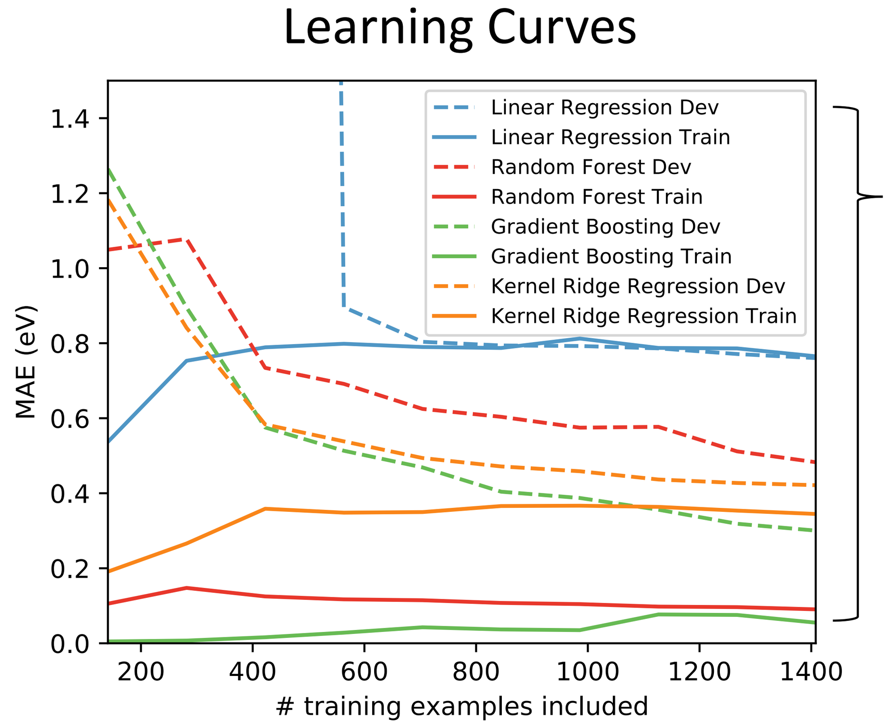
⋮

$$m_i = \frac{\int (\rho(\varepsilon) - m_1)\, \varepsilon^i}{m_0}$$

Feature Engineering: Moments

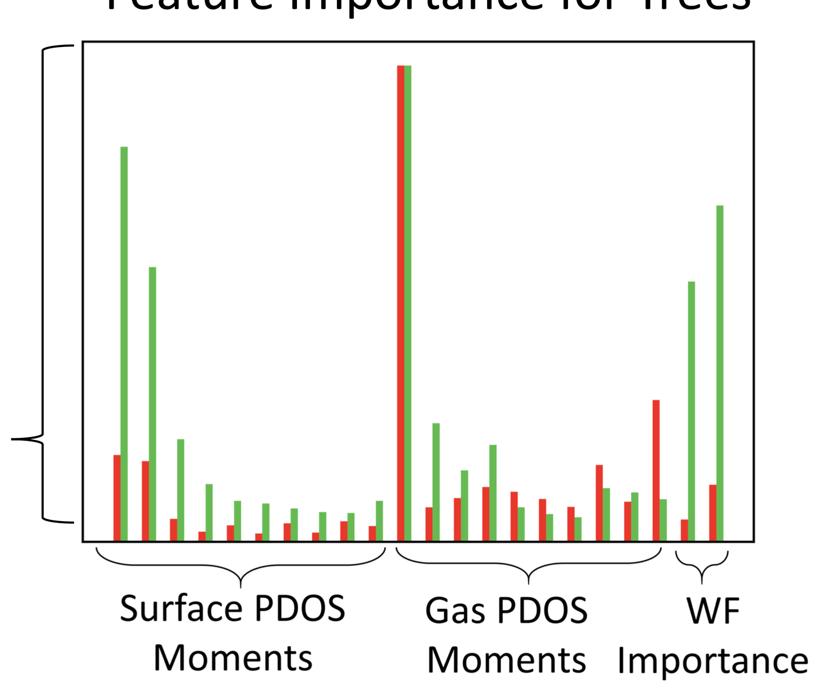


Learning Curves

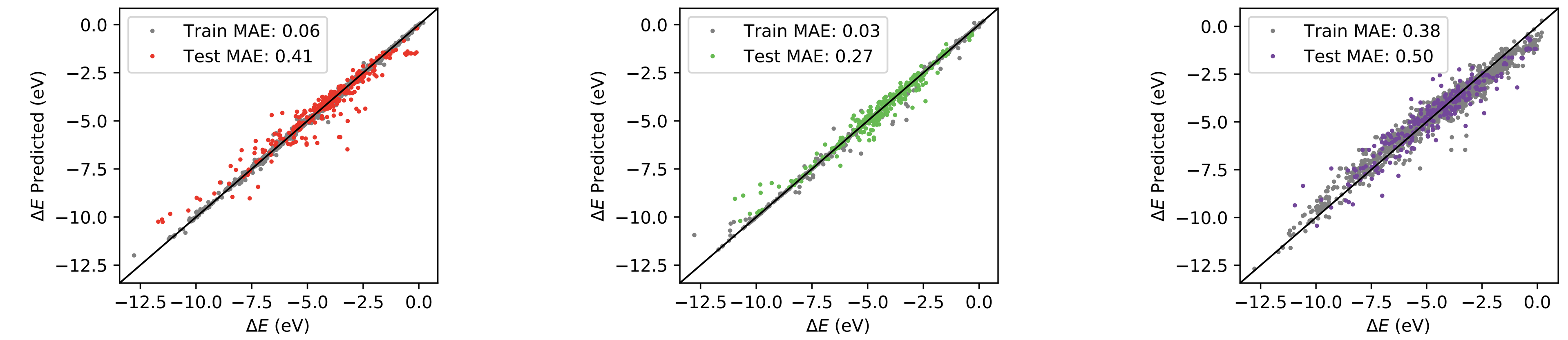


- LR performs poorly, nonlinear models do better
- GB and RF may show increased accuracy with more data
- RF and GB show similar features to be important

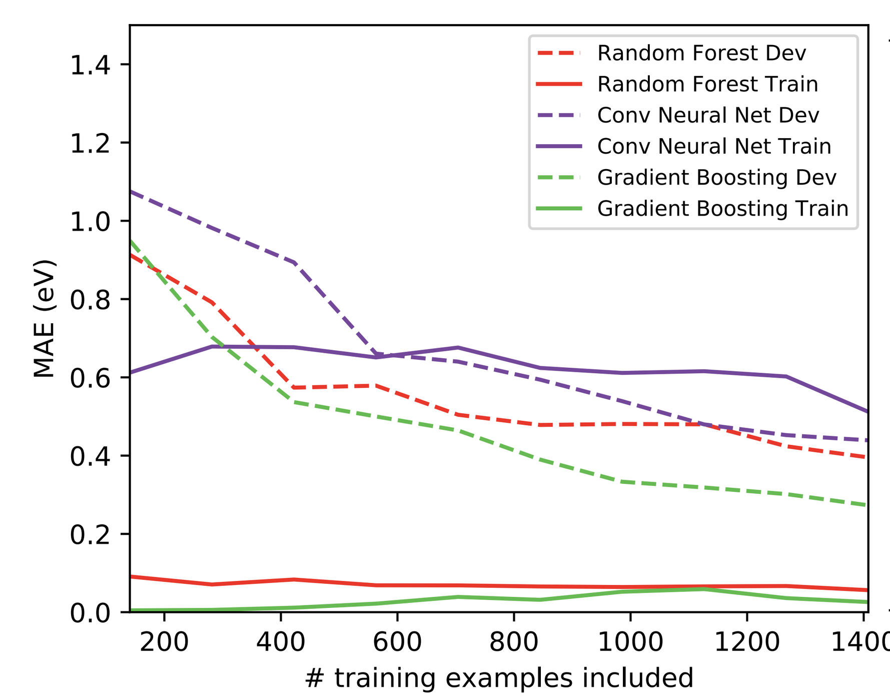
Feature Importance for Trees



Feature Engineering: Full Spectrum



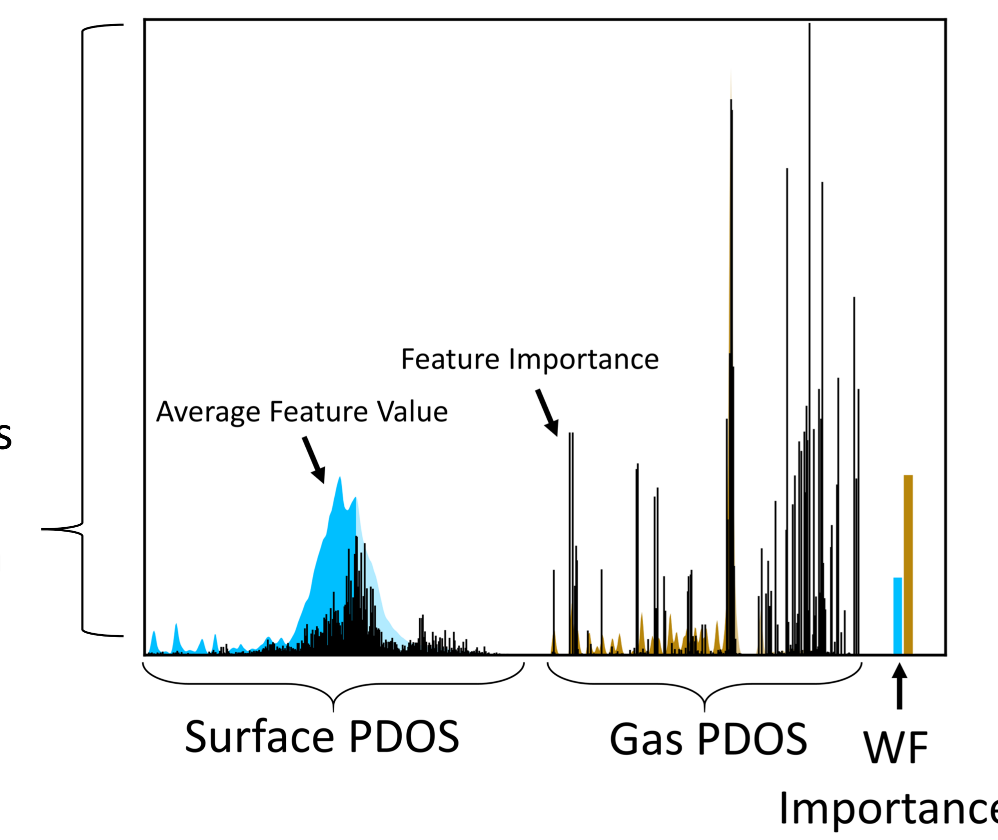
Learning Curves



- Gradient Boosting performs best
- Collecting more data may still help

Feature Importance for Random Forest

- Many important features near fermi level
- Gas PDOS features seem chaotic



Alternative Train Test Splits

- Depending on the application, the model may have to make a prediction about a reaction of surface it hasn't seen before

MAE (eV)	Moments & WF				PDOS (& WF)		
	LR	KRR	RF	GB	RF	GB	CNN
Reaction	5.15	1.31	0.75	0.46	0.94	1.14	0.67
Composition	1.26	0.55	0.6	0.57	0.5	0.44	0.44
Composition + Reaction	0.81	0.49	0.5	0.36	0.41	0.27	0.50
Random	0.8	0.44	0.38	0.29	0.41	0.28	0.35

Future Work

- Consideration of surfaces beyond pure metals, such as oxides, sulfides, alloys, etc.
- Exploration of effects of splitting on model performance:
    - How similar must new examples be to examples in the training set to obtain accurate predictions?
    - How do different random splits affect model performance, i.e. are certain reactions/surfaces necessary to describe many examples?

Acknowledgements:

We are grateful to the instructors of CS229, Andrew Ng and Ron Dror, as well as all of the teaching assistants for their help and support. We also are grateful to our lab-mate, Brian Rohr, who built the convolutional neural networks and provided valuable guidance and advice. Finally, we thank the Stanford Sherlock Computing Resources.

References:

1. Hammer, B. and Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surface Science* 343.3 (1995): 211-220.
2. Vojvodic, A., Nørskov, J. K., & Abild-Pedersen, F. Electronic Structure Effects in Transition Metal Surface Chemistry. *Topics in Catalysis*, *57*(1–4) (2014): 25–32.
3. Dickens, C. F., Montoya, J. H., Bajdich, M., Kulkarni, A., Nørskov, J. K. An electronic structure descriptor for oxygen reactivity at metal and metal-oxide surfaces. *Surface Science 681* (2018): 122-129