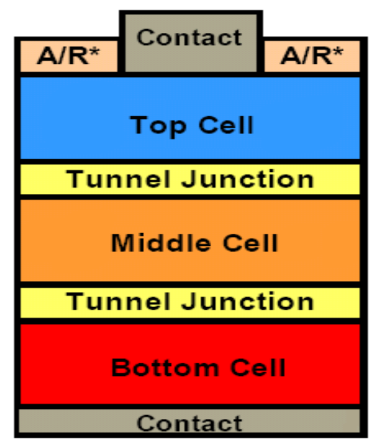
Data-Driven Prediction of Band Gap of Materials

Fariah Hayee‡, Isha Datye‡, Rahul Kini†
fariah@stanford.edu, idatye@stanford.edu, rkini11@stanford.edu
Departments of Electrical Engineering‡, Material Science and Engineering†, Stanford University

Motivation

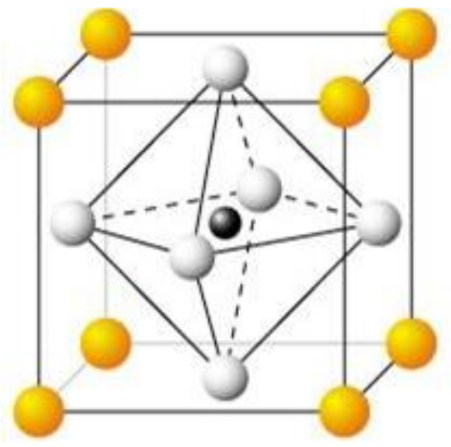
- Development of tandem and perovskite solar cells and battery electrodes is largely constrained by new material discovery and design
- Prediction of material properties using computational methods like density functional theory (DFT) and molecular dynamics (MD) is computationally expensive

Goal: Predict band gaps of materials from element composition using machine learning techniques



*A/R: Anti-Reflective Coating

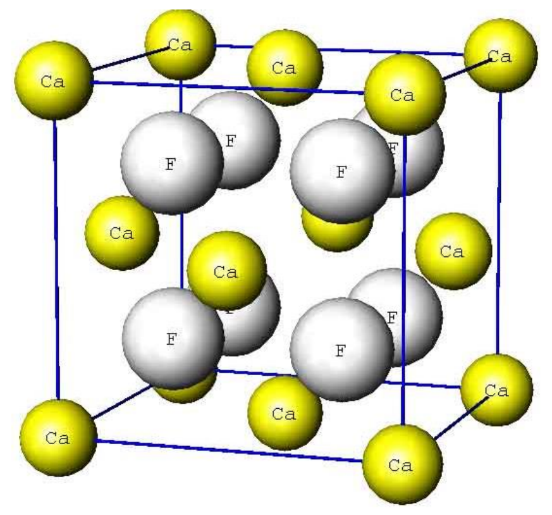
Tandem solar cell



Perovskite structure

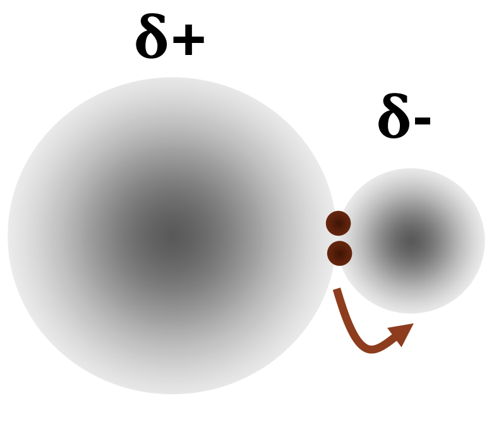
Data and Features

- The dataset contains 2067 samples with DFT-calculated band gap values ranging from 24 meV to 11.5 eV
- 75 features include stoichiometric, elemental, and electronic structural and ionic attributes

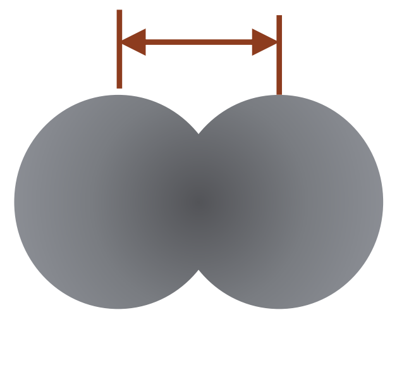


Unit cell

Source: Materials Project



Electronegativity



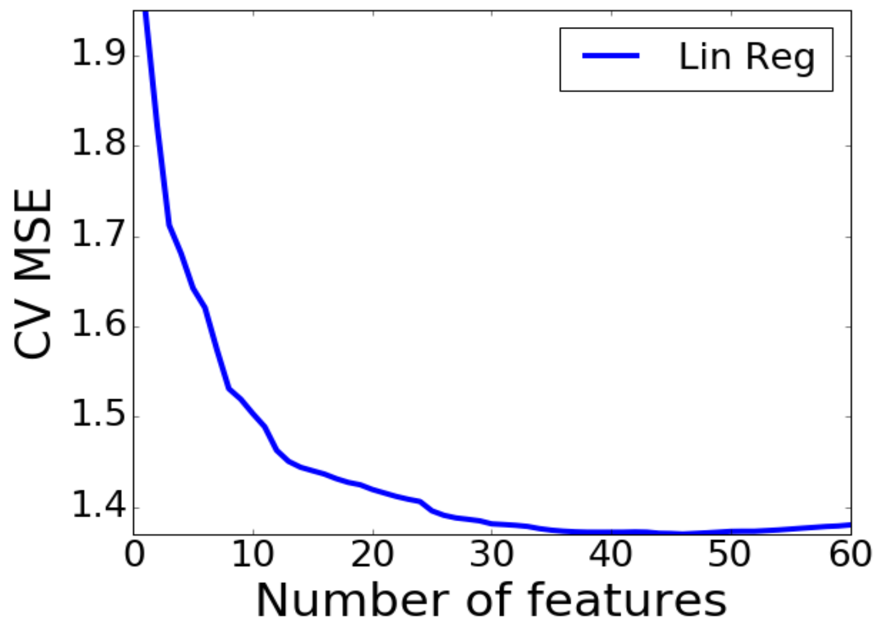
Covalent radius

- **Preprocessing:** Small variance features removed and standardization applied to get a distribution of mean zero and unit variance
- Designed complex features such as:

$$\%p = \frac{\text{Avg p electrons in valence shell}}{\text{Avg electrons in valence shell}}, \quad \Delta \mathrm{Pe}_{p-d} = \max\{0, (\%p * Np - \%d * Nd)^2\},$$

  where Np (Nd) = maximum number of valence p (d) electrons

- **Forward selection method:** Important features include $\%p$, ΔPe_{p-d}, electronegativity difference (ΔEN), covalent radius (R_{CV}), number of f electrons, and periodic table row and column numbers



Linear Regression

- In OLS, Ridge, and Lasso, the fitting parameter θ is calculated by :

OLS:

$$\theta = argmin||X\theta - y||_2^2$$

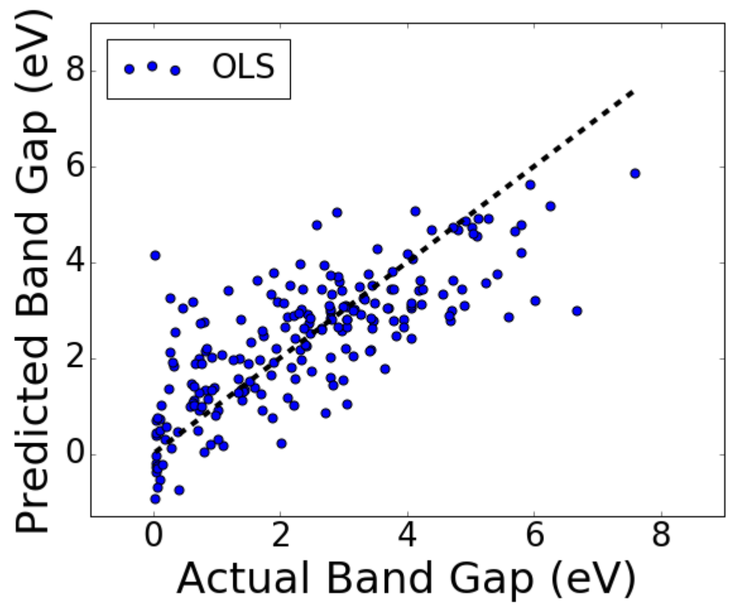
Ridge:

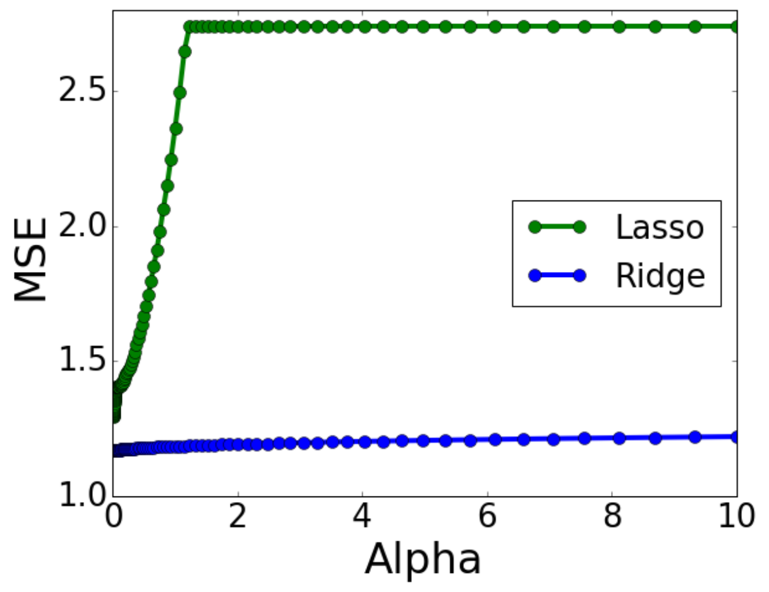
$$\theta = argmin||X\theta - y||_2^2 + \alpha||\theta||_2^2$$

Lasso:

$$\theta = argmin\,\frac{1}{2n}||X\theta - y||_2^2 + \alpha||\theta||_1$$

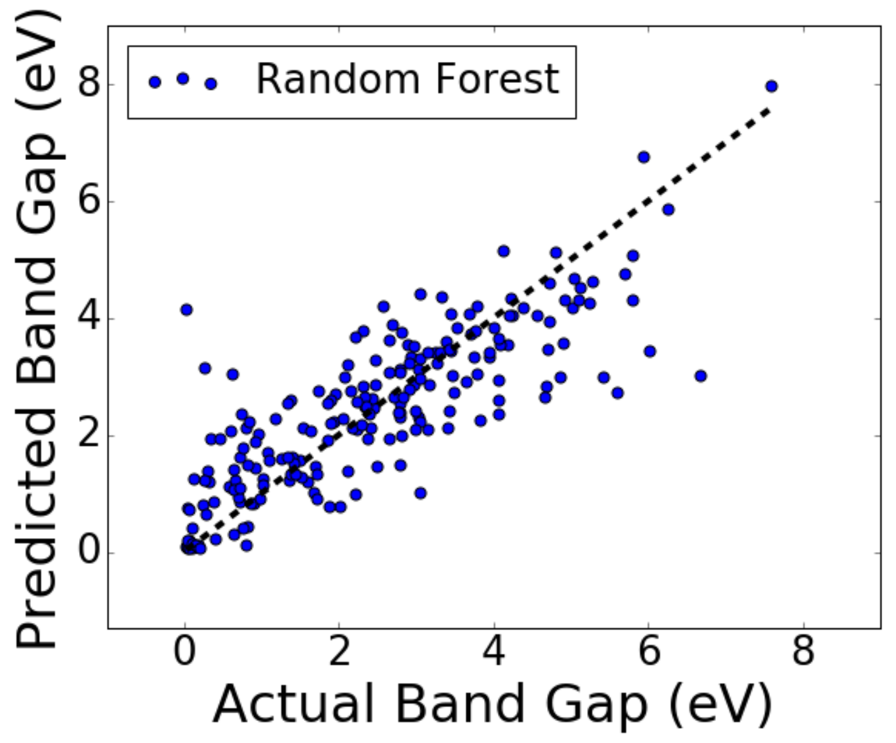
- Ridge and Lasso: no improvement over OLS since optimal α is close to 0

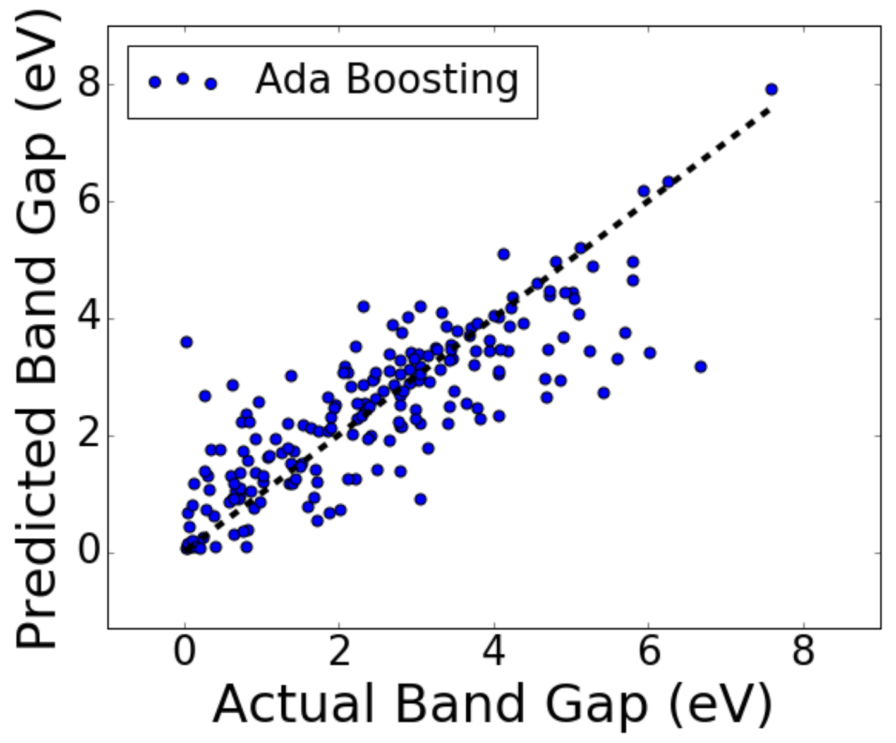




Random Forest and Ada Boosting

- **Random Forest:** Fits classifying decision trees on subsets of the data, uses averaging to improve accuracy and prevent over-fitting
- **Ada Boosting:** Fits series of weak learners on repeatedly modified versions of the data, with higher weight placed on incorrectly predicted examples





Method	Training Data (1860 samples)		Test Data (207 samples)		
	MSE	Score (r^2)	MSE	Score (r^2)	CV MSE
OLS	1.30	0.62	1.17	0.57	1.38
Random Forest	0.16	0.95	0.86	0.68	1.18
Ada Boosting	0.01	1.00	0.81	0.70	1.18
MLP	0.38	0.89	1.22	0.63	1.28

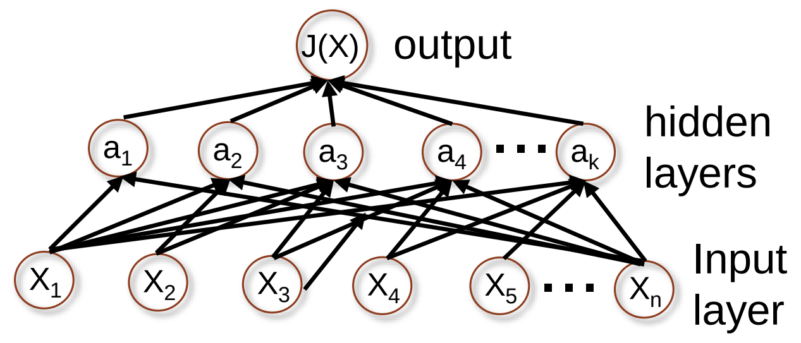
Neural Network

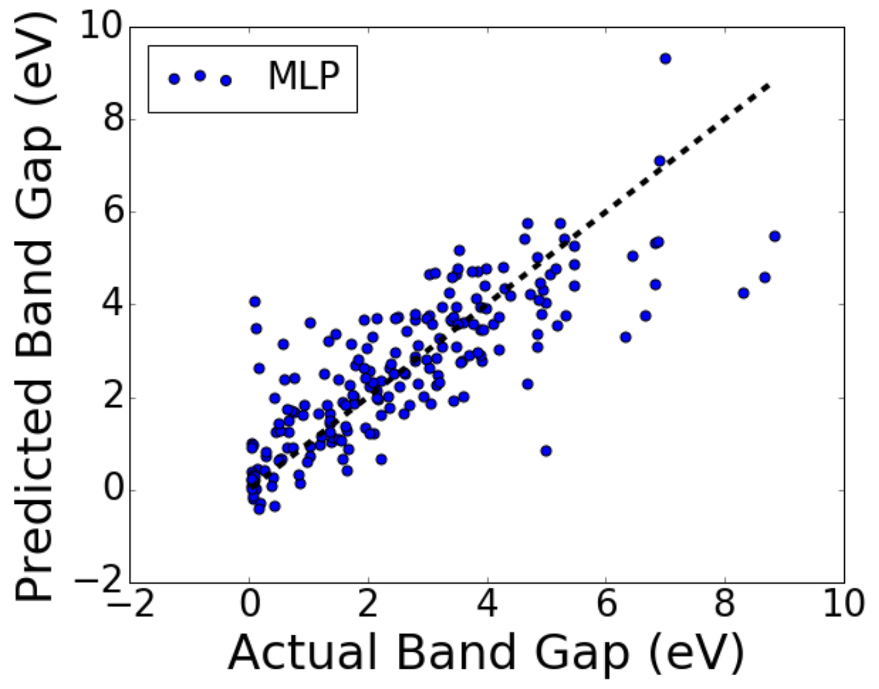
- **Multi-layer Perceptron (MLP):**

  Update parameter using SGD:

$$w \leftarrow w - \eta\left(\alpha\frac{\delta R}{\delta w} + \frac{\delta Loss}{\delta w}\right)$$

- Capable of learning non-linear models
- **Optimal parameters:** Hidden layers = 500, activation function = ‘relu’, initial learning rate = 0.0027599, regularization = 0.00033





Conclusion

- Band gap is positively correlated to ΔEN and %p and negatively correlated to R_{CV} and ΔPe_{p-d}, which are indicative of ionicity and hybridization of the bonds in the compounds
- Linear regression has high bias; Random Forest, Ada boosting, and MLP perform better but tend to over-fit data
- Lowest test MSE achieved with Ada Boosting, suggesting that the error due to bias is reduced more than the error due to variance

Future Work

- Add a partitioning algorithm to our model—train a neural network to partition the dataset into k groups of similar materials and then use the subsets to train regression
- Train a convoluted neural network or a deep learning network on a much larger database (Materials Project or OQMD) to more accurately predict material band gaps

References

[1] Richard King *et al.*, *MRS Bulletin*, March 2016. [2] E.F. Shubert, *Light Emitting Diodes* (Camb. Univ. Press), 2006.

Acknowledgments

We are grateful to Bryce Meredig and the Citrine Informatics team who provided us with the material data sets necessary for this project, as well as the CS 229 teaching assistants (especially Prof. Ng, Prof. Duchi, and Rishabh) who helped guide the selection of appropriate algorithms for this project.