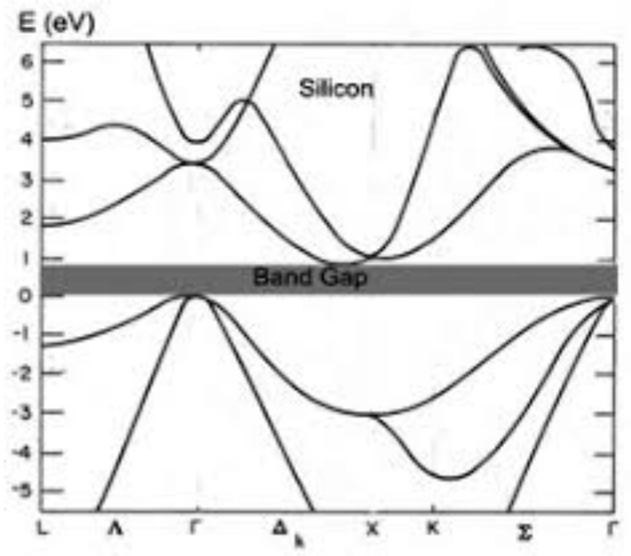
Machine Learning for Materials Band Gap Prediction

Jacob Marks (cs229a), Jason Qu (cs229), and Ilan Rosen (cs229)

Introduction

Problem: A material's electronic properties—and technological utility—depend on its band gap. Band gaps are notoriously difficult to compute from first principles and computationally intense to approximate, so their prediction represents a challenging yet consequential application for ML. We set out to **predict band gap size with only elemental composition and atomic positions** by training learning models on computationally generated datasets.

Material Type:	*Gap Size:*
Metals	Small (0 or negligible)
Nonmetals	
Semiconductors	Intermediate
Insulators	Large (> 3.2 eV)



Energy Landscape of Silicon. The band gap is shaded.

Challenges:

- Domain knowledge for feature engineering
- Large space of possible materials
  - differing crystal structures
  - differing # of atoms/unit cell
- Size/consistency of available datasets

Data set: JARVIS Density Functional Theory database of 3D materials (**14752 non-metals** and **8703 metals**)

Features and Input Encoding

The model input for each material was a list of the atoms in the material's unit cell and their positions. This information is not a suitable feature set for machine learning, as positions are degenerate in coordinate axis.

Encodings tested:

— One-hot representation of element
— One-hot representation of atomic group
— Coulomb Matrix
— Singular values of the Coulomb Matrix
— Coulomb matrix AND one-hot representation of group
— Coulomb matrix augmented by its non-degenerate permutations
— Singular values of Coulomb matrix AND one-hot representation of element

The Coulomb matrix

$$C_{ii} = Z_i^{2.4}$$

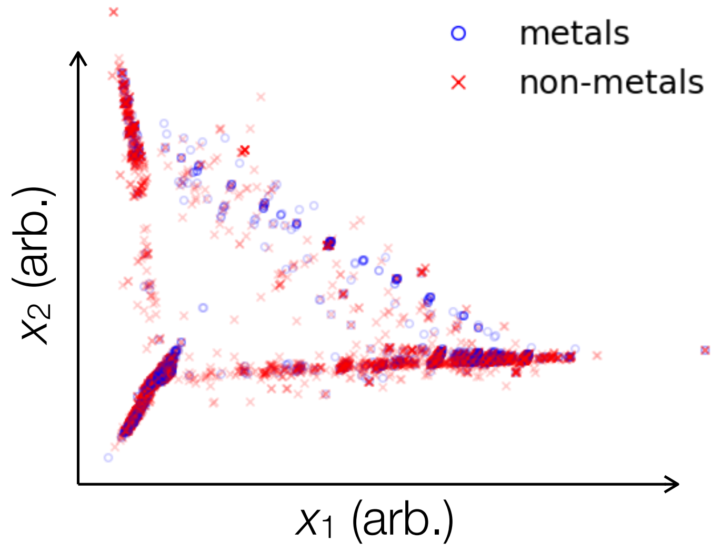
$$C_{ij} = \frac{Z_i Z_j}{|r_i - r_j|}, \quad i \neq j$$

r_i : atomic position

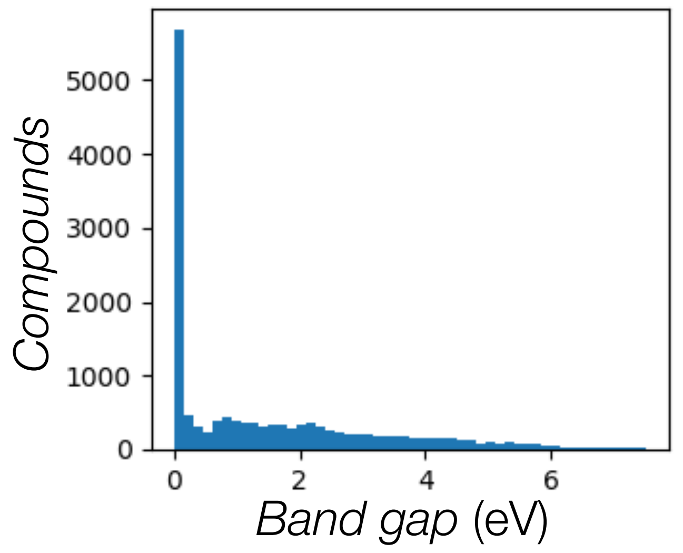
Z_i : atomic number

Encodes the Coulomb (electrostatic) potential between atoms

Dataset principle components (element one-hot encoding)



Nonmetals gap size histogram



Learning Models

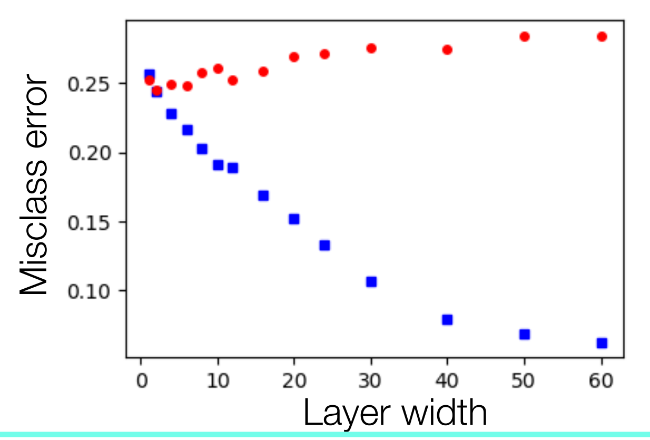
Metal — nonmetal classification

Metrics: misclassification error; error under receiver operating curve

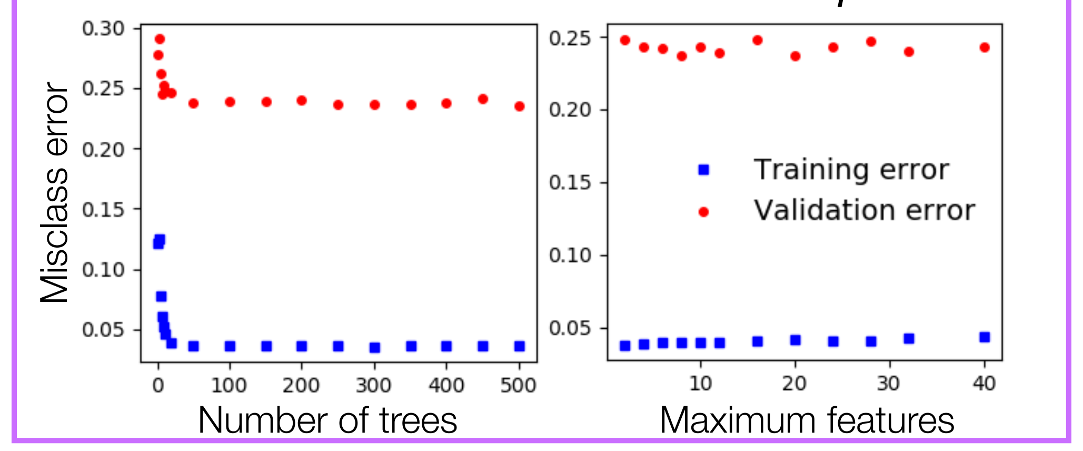
	Encoding	Element one-hot	Group one-hot	Coulomb Matrix	Coulomb svals	Coulomb + group one-hot	Augmented Coulomb
LogReg	Misclass Error	25.7%	27.8%	39.5%	42.8%	34.0%	46.5%
	ROC Area	0.801	0.776	0.650	0.597	0.720	0.553
Neural Net	Misclass Error	**24.8%**	26.6%	42.3%	44.9%	44.4%	42.1%
	ROC Area	**0.822**	0.808	0.606	0.569	0.578	0.600
Random Forest	Misclass Error	**23.8%**	24.9%	31.1%	30.9%	27.6%	31.2%
	ROC Area	**0.842**	0.828	0.748	0.754	0.794	0.745

Performance of the feature encodings
neural nets have depth 1, width 10; random forests have 500 trees

Neural net development



Random forest classifier development



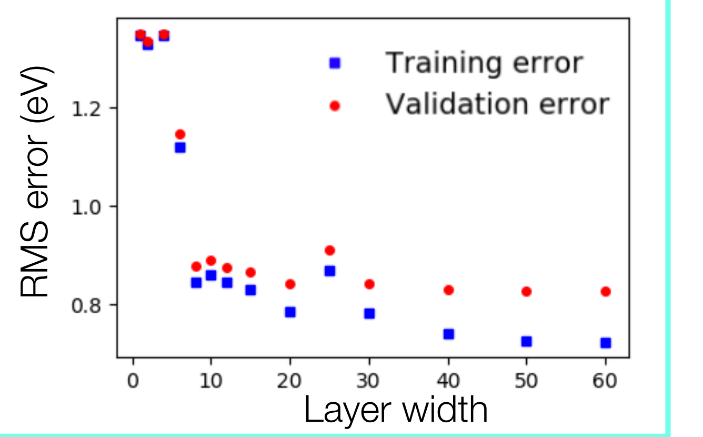
Gap prediction for nonmetals

Metrics: root mean square error (eV); median normalized error

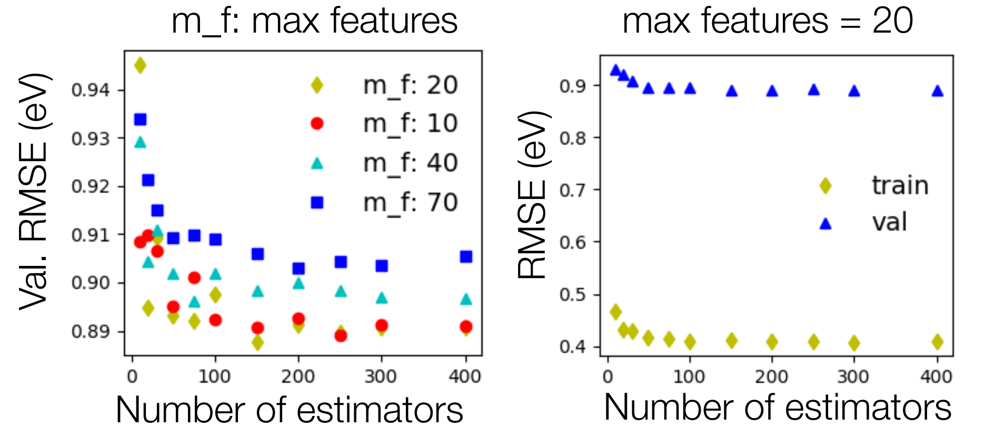
	Encoding	Element one-hot	Group one-hot	Coulomb Matrix	Coulomb svals	Coulomb + group one-hot	Augmented Coulomb	C svals + elem 1-hot
LinReg	RMS Error (eV)	1.348	1.492	1.486	1.539	1.223	1.454	1.119
	Median Norm. Error	0.648	0.801	7.521	8.198	3.977	6.845	3.773
Neural Net	RMS Error (eV)	**0.956**	1.29	1.86	1.77	1.39	1.36	1.81
	Median Norm. Error	**0.484**	0.654	1.53	2.32	3.26	6.79	1.94
Random Forest	RMS Error (eV)	**0.910**	1.18	1.07	1.03	0.900	0.955	0.922
	Median Norm. Error	**0.363**	0.486	0.802	0.779	0.598	1.51	0.493

Performance of the feature encodings
neural nets have depth 1, width 10; random forests have 200 trees

Neural net development



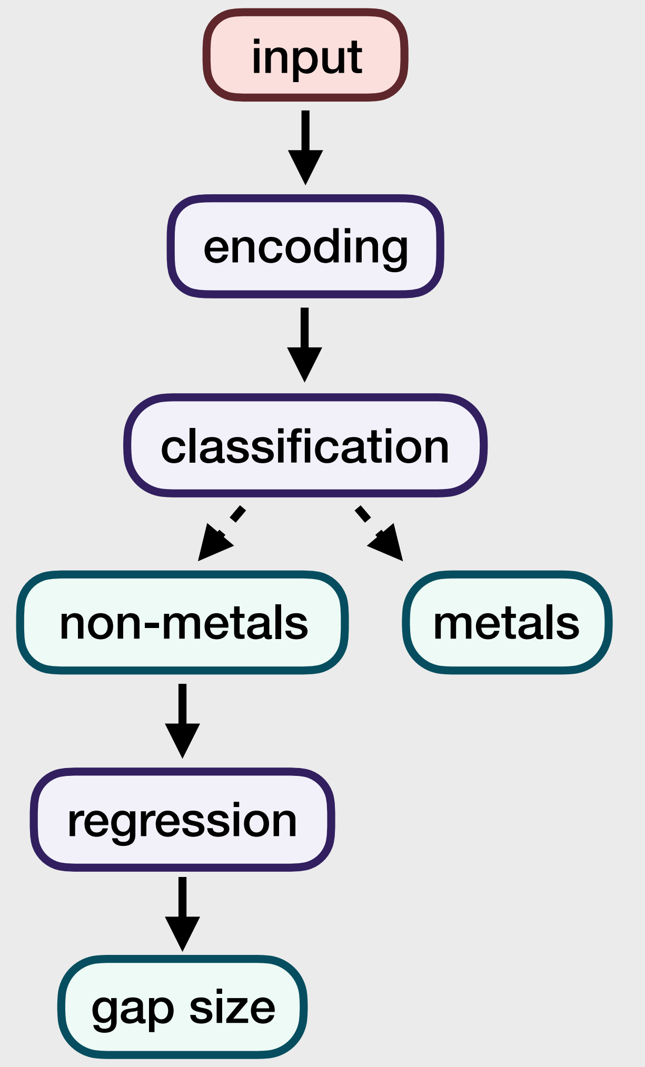
Random forest regressor development



Results

Pipeline: the regression stage operated only on predicted nonmetals from the classification stage. Both stages used a **one-hot element encoding** as features. A tuned **random forest classifier** was chosen for the 1st stage, and a tuned **neural network** (ReLU activation; linear output) for the 2nd stage.

Pipeline for predicting gaps



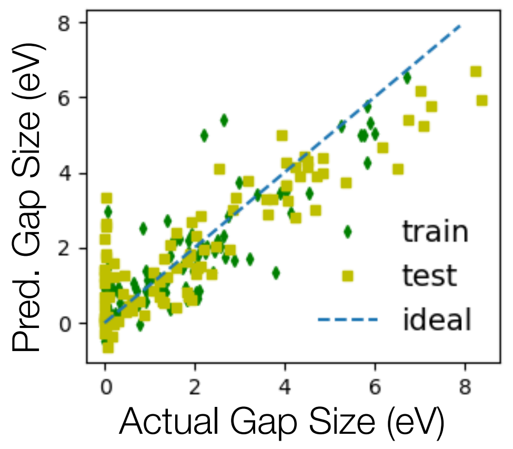
Classification stage results

True / Pred	Metal	Nonmetal
Metal	(True neg. rate) **0.694**	(False neg. rate) **0.188**
Nonmetal	(False pos. rate) **0.306**	(True pos. rate) **0.812**

F1 score: 0.767

Regression stage results

Reported on true positive examples



RMS Error: 0.924 eV

Median Normalized Error: 0.364

Discussion

Performance: following literature, we used RMS error as a metric for the regression stage performance; we chose a neural network accordingly. A random forest regressor outperformed the neural net in median normalized error (0.318 versus 0.544) but had higher RMS absolute error (0.948 eV versus 0.881 eV).

Small-gap insulators: nearly half of the nonmetals in the dataset had gaps between 0.01 eV and 0.1 eV. The classifier model struggled with these materials; removing them decreased the misclassification error to 10.6%.

Feature encoding: the one-hot representation of constituent elements in compounds performed best in both stages.

A one-hot representation of element's groups performed well for classification but not for regression. Physically, an atom's group determines its valance, which is important for predicting its metallicity, whereas the gap magnitude depends on the atomic number (because of electric screening)—information that the group encoding removes.

The Coulomb matrix's singular values contains this information, explaining why the Coulomb matrix singular values + group one-hot encoding performed reasonably well in the regression stage.

References

IIT Delhi, NPTEL Online Course Lecture Notes, Fundamental concepts of semiconductors (2013).
S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Materials **12**, 191 (2013).
K. Choudhary, I. Kalish, R. Beams, and F. Tavazza, Scientific Reports **7**, 5179 (2017).
K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).