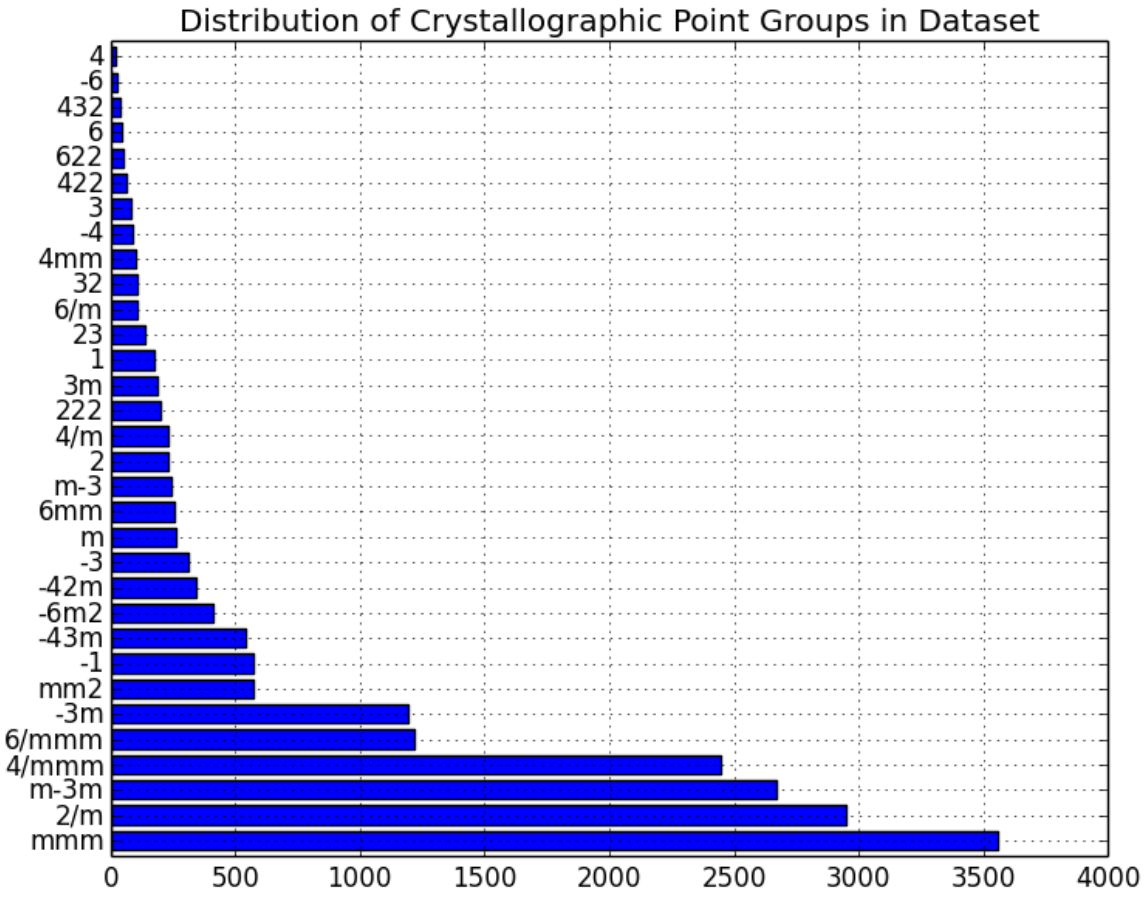
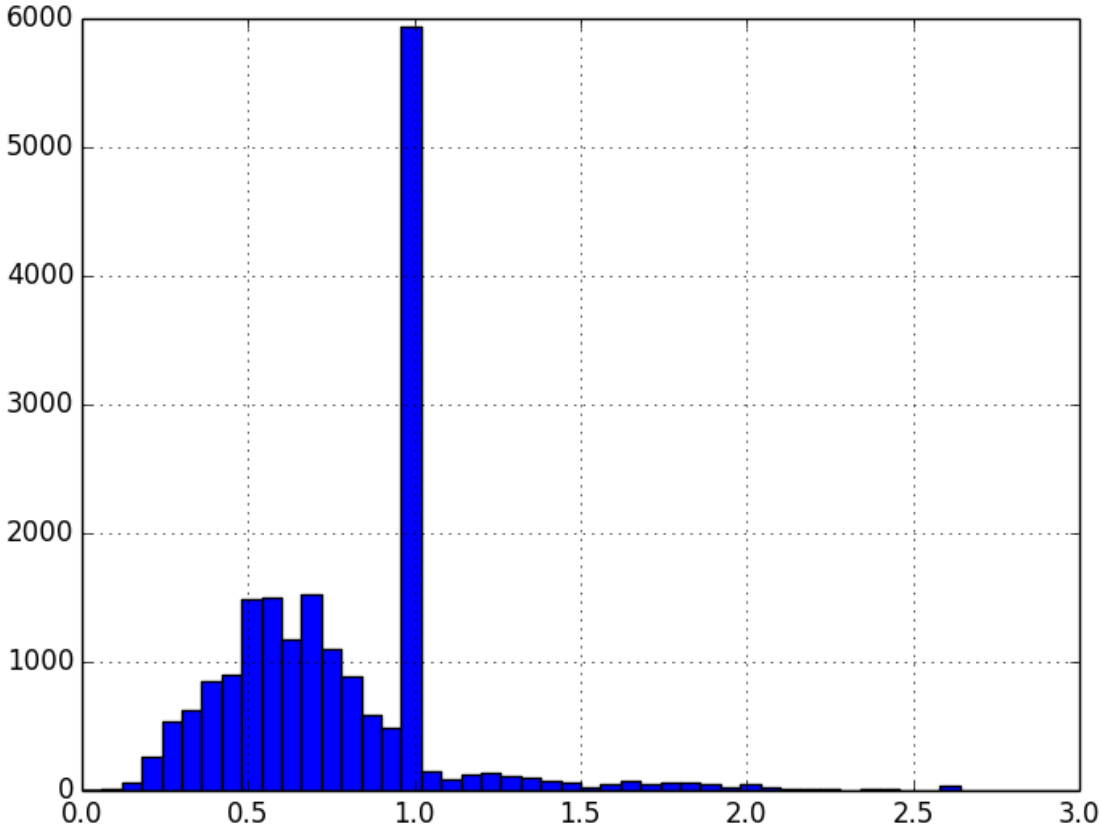
matprojgeom Ryan Henderson

Overview: The goal of this project is to predict structural characteristics of solid-state compounds given only compositional information. Specifically, given a composition $A_xB_yC_z$ we want to guess the crystallographic point group, c/a ratio (if applicable), volume per site, and average coordination for each species.

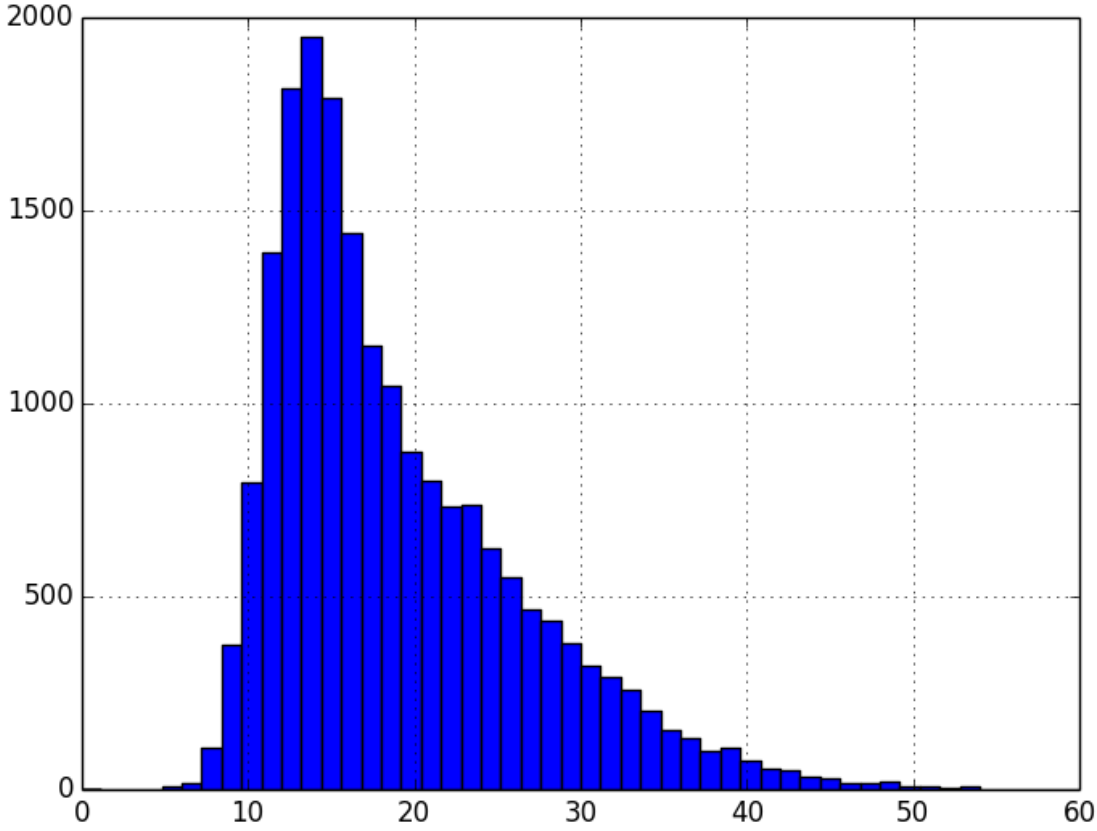
To guide these predictions, I scraped, cleaned, and extracted features from 19,428 crystal structures available from the Materials Project database. Source code is available at https://github.com/rhsimplex/matprojgeom .

The Data: Noble gas compounds were excluded, along with structures possessing extreme c/a ratios (these caused the symmetry finder in the pymatgen package to error out).





Distribution of c/a in Dataset: the large peak at c/a=1 corresponds to the cubic structures



Distribution of volume/per site (A^3) in Dataset

Predicting Point Group -- Random Forest Classifier: The sklearn implementation of the Random Forest Classifier with default parameters was used to predict point group directly. An average accuracy of 0.47 was realized (accuracy = TP/(TP+FP) measured by the k folds method, with k=3. All subsequent accuracies were obtained similarly).

When grouped by crystal system (cubic, hexagonal, etc.) a slightly higher accuracy of 0.52 is realized (Table 1 below). The classifier is very accurate when classifying point groups when the crystal system is already correctly predicted (Tables 3-9, next page). All confusion matrices are drawn from a subset withheld for testing.

	cubic	hexagonal	trigonal	tetragonal	orthorhombic	monoclinic	triclinic
cubic	551	70	50	100	104	44	8
hexagonal	57	286	39	40	65	17	3
trigonal	37	35	198	45	54	51	13
tetragonal	65	34	38	445	113	69	11
orthorhombic	77	76	61	116	521	148	24
monoclinic	59	30	78	76	214	468	89
triclinic	7	7	10	16	35	55	41

Table 1: Confusion Matrix for Crystal System (Accuracy = 0.52)

The Random Forest Classifier ranks features by relative importance.

Feature	Relative Imp
electronegativityStd	0.14
radiiStd	0.11
electronsPerAtom	0.11
rowStd	0.09
electronegativityRange	0.09
fracTransitionMetal	0.08

Table 2: Feature Importances for Point Group Classifier

	m-3m	-43m	432	m-3	23
m-3m	414	13	0	3	1
-43m	11	68	0	4	0
432	0	0	1	0	0
m-3	3	1	0	18	0
23	5	0	0	0	9

Table 3: Confusion Matrix For Cubic Groups (Acc.=0.93)

	6/mmm	-6m2	6mm	622	6/m	-6	6
6/mmm	165	7	5	0	1	0	0
-6m2	4	56	2	0	0	2	0
6mm	6	2	13	0	0	0	0
622	0	0	0	3	0	0	0
6/m	2	0	0	0	11	0	1
-6	1	2	0	0	0	1	0
6	0	0	0	0	0	0	2

Table 4: Confusion Matrix for Hexagonal Groups (Acc.=0.88)

	-3m	3m	32	-3	3
-3m	134	4	3	5	0
3m	5	5	0	0	0
32	0	0	6	1	2
-3	6	0	0	20	1
3	0	0	0	1	5

Table 5: Confusion Matrix for Trigonal Groups (Acc.=0.86)

	4/mmm	-42m	4mm	422	4/m	-4	4
4/mmm	366	3	2	1	8	0	0
-42m	5	26	1	0	0	1	0
4mm	2	0	5	0	0	0	0
422	1	0	0	1	0	0	0
4/m	1	0	0	0	13	0	0
-4	0	0	0	0	0	7	0
4	1	0	0	0	0	0	1

Table 6: Confusion Matrix for Tetragonal Groups (Acc.=0.94)

	mmm	mm2	222
mmm	422	16	5
mm2	31	31	1
222	7	2	6

Table 7: Confusion Matrix for Orthorhombic Groups (Acc.=0.88)

	2/m	m	2
2/m	379	8	16
m	19	12	0
2	21	1	12

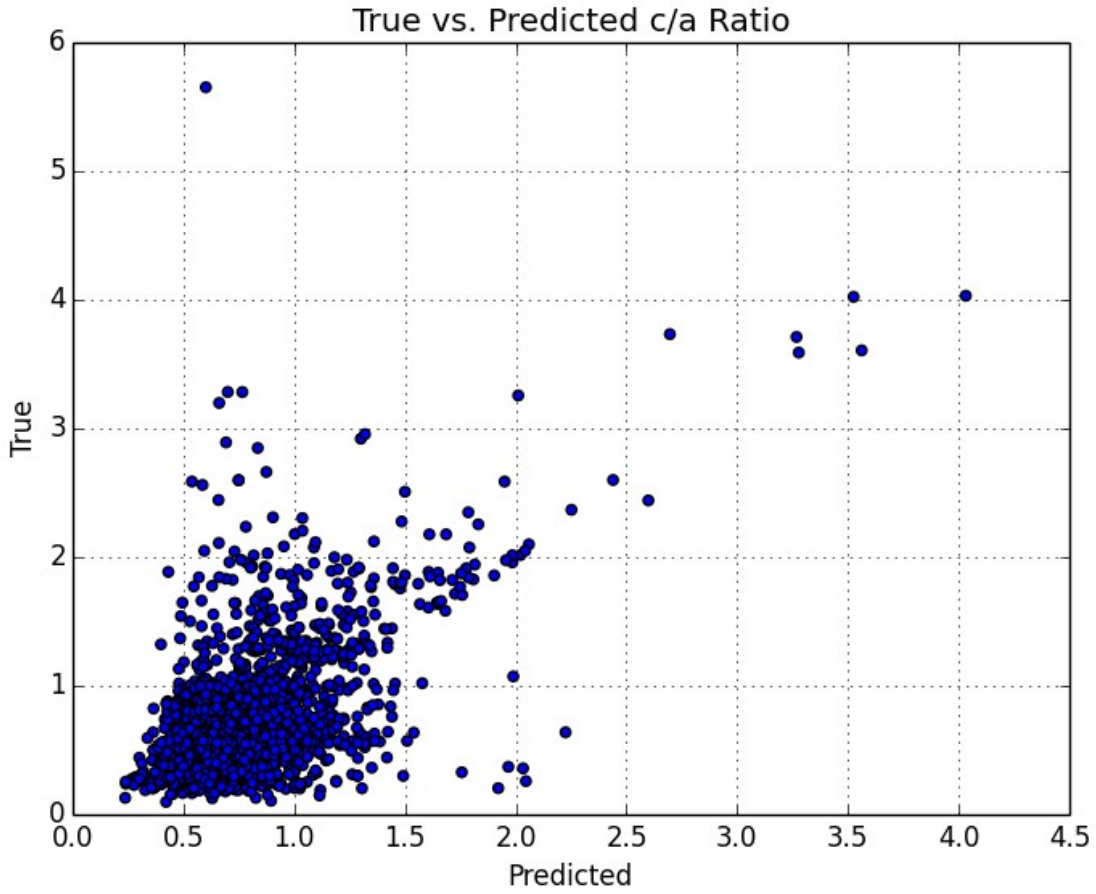
Table 8: Confusion Matrix for Monoclinic Groups (Acc.=0.93)

	-1	1
-1	26	3
1	0	12

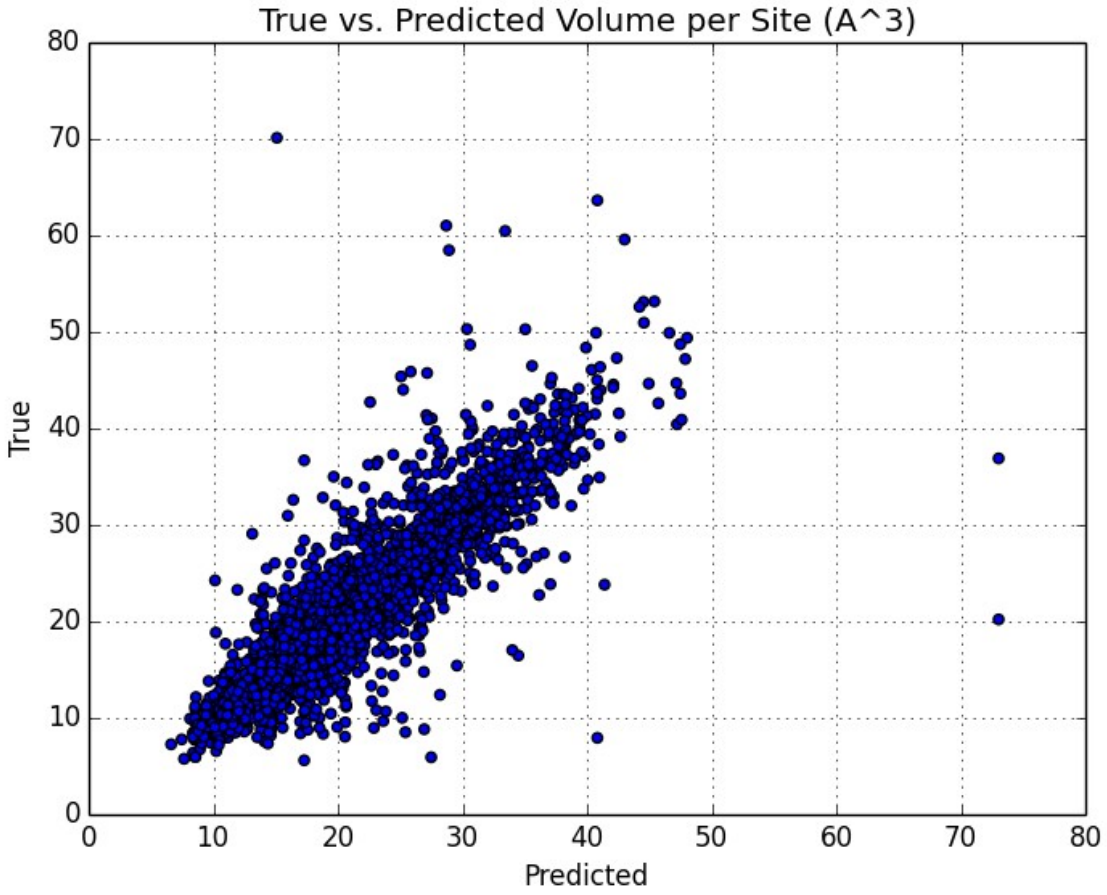
Table 9: Confusion Matrix for Triclinic Groups (Acc.=0.93)

The most relevant features are not surprising. Differences in electronegativity and radius are basic to our intuition of solid-state structure. Electrons per atom (defined here as the sum over the composition of the fraction of each species times its most common oxidation state) also makes a prominent appearance.

Predicting c/a and volume per site – Random Forest Regressor: The sklearn implementation of the Random Forest Regressor with default parameters was used to predict c/a and volume per site. The mean average error (MAE) in c/a was 0.20. The MAE for volume per site was 1.86 A^3.



Structures with c/a = 1.00 have been excluded



Feature	Relative Import
electronsPerAtom	0.47
fracTransitionMetal	0.16
fracMetalloid	0.14
fracRareEarth	0.05
radiiRange	0.05

Important Features for c/a prediction

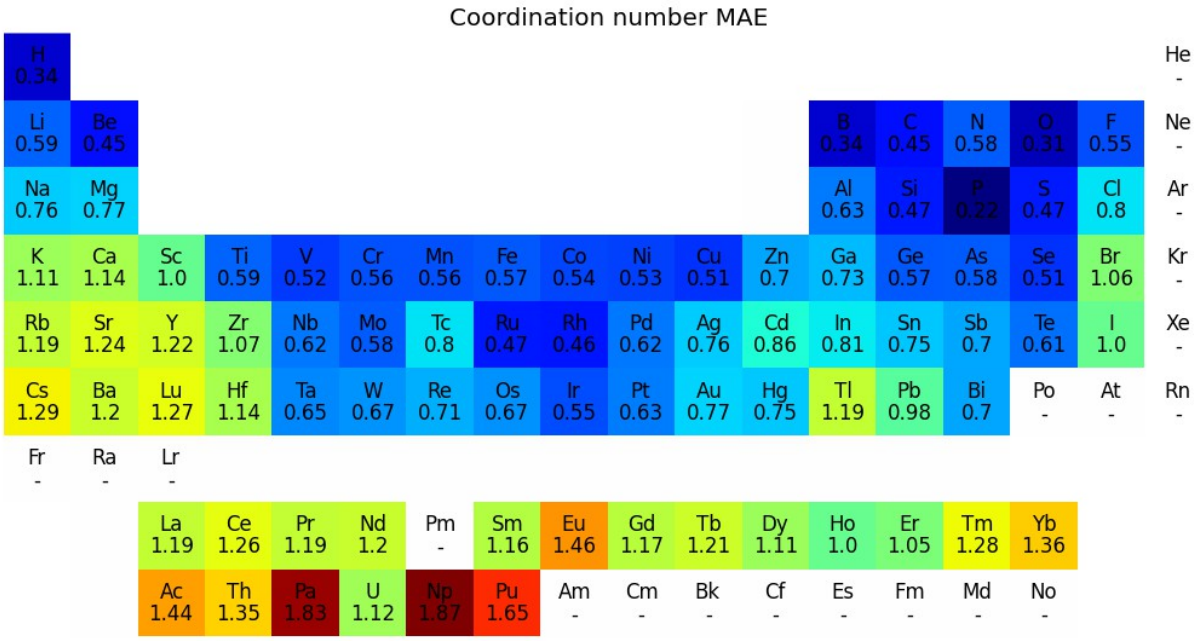
Feature	Relative Importance
fracChalcogen	0.39
fracTransitionMetal	0.23
electronegativityRange	0.14
fracHalogen	0.07
fracAlkali	0.05

Important Features for volume/site prediction

Compositional features are more prominent here, especially for volume per site.

Coordination – Random Tree Regressor: Average coordination for each element were generated from individual datasets. Each elemental dataset contained every structure containing that element, all relevant compositional features, and the average coordination number (as given in the Materials Project database, not computed directly from the structure). The regressor was trained against the given coordination numbers.

Relevant features varied by element. The performance (MAE) for each element is summarized below. For reference, typical coordination numbers range from 1 (e.g. terminal hydrogens) to 12-16 (tetrahedrally closest-packed metals).



Mean Average Error in Coordination Prediction

Conclusion: A random forest classifier/regressor can make structural predictions reasonably well with our features. There are many improvements possible (e.g. leveraging group/subgroup relations when scoring the classifier).

Although my program does not *directly* predict a structure given a composition, it should be straightforward to generate reasonable structures given the composition, point group, c/a ratio, and coordination of each species. This is the obvious next step.