Electronic Supplementary Material (ESI) for Journal of Materials Chemistry C.

Supporting Information for:

Monitoring the Role of Site Chemistry on the Formation Energy of Perovskites via Deep Learning Analysis of Hirshfeld Surfaces

Logan Williams[1], Arpan Mukherjee[1], Aparajita Dasgupta[1], Krishna Rajan[1*]
[1]Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY14260-1660 USA.
*E-mail: krajan3@buffalo.edu

Network architecture

The complete network architecture used for the transfer learning of the cubic perovskite formation energies in this study is listed below. The neural network was technically two networks: one consisting of layers 0-18 and used only to evaluate and process the input, and one consisting of layers 19-25 that was trained on the training data and tested on the test set. The neural network was created using the open source library Keras with Tensorflow v1.8.0 backend.[1]

Table S 1. Complete architecture of the neural network used in this study for the transfer learning of the cubic perovskite dataset. The network was split into 2 parts. The beginning was taken from the previous paper.[2] It took the atomic Hirshfeld surfaces fingerprint plots as inputs and produced a flattened feature vector as output, without performing any optimization / training. The second part was took the feature vectors as input, and was optimized to produce DFT formation energy as output using a subset of the OQMD dataset as training data.

Layer #	Layer Type	Output Shapes	Parameters	Trained or From Previous
0	InputLayer	(None, 50, 50, 1)	0	From Previous
1	Conv2D	(None, 50, 50, 32)	320	From Previous
2	Conv2D	(None, 50, 50, 32)	9248	From Previous
3	MaxPooling2D	(None, 25, 25, 32)	0	From Previous
4	Dropout	(None, 25, 25, 32)	0	From Previous
5	BatchNormalization	(None, 25, 25, 32)	128	From Previous
6	Conv2D	(None, 25, 25, 64)	18496	From Previous
7	Conv2D	(None, 25, 25, 64)	36928	From Previous
8	Conv2D	(None, 25, 25, 64)	36928	From Previous
9	MaxPooling2D	(None, 12, 12, 64)	0	From Previous
10	Dropout	(None, 12, 12, 64)	0	From Previous
11	BatchNormalization	(None, 12, 12, 64)	256	From Previous
12	Conv2D	(None, 12, 12, 128)	73856	From Previous
13	Conv2D	(None, 12, 12, 128)	147584	From Previous
14	Conv2D	(None, 12, 12, 128)	147584	From Previous
15	Conv2D	(None, 12, 12, 128)	147584	From Previous
16	MaxPooling2D	(None, 6, 6, 128)	0	From Previous
17	BatchNormalization	(None, 6, 6, 128)	512	From Previous
18	Flatten	(None, 4608)	0	From Previous
19	Dense	(None, 128)	589952	Trained
20	Dropout	(None, 128)	0	Trained
21	BatchNormalization	(None, 128)	512	Trained
22	Dense	(None, 128)	16512	Trained
23	Dropout	(None, 128)	0	Trained
24	BatchNormalization	(None, 128)	512	Trained
25	Dense	(None, 1)	129	Trained

Compositions within cubic dataset

Not all cubic perovskite compounds in the OQMD database were included in this study. The figure below shows the compounds in the composition space that were not included as red x's. The excluded compounds met one of the criteria listed in the main paper methodology section that caused the validity of the results to be uncertain: unconverged bandgap calculation, numerically identical unrelaxed and relaxed lattice parameters, or an unreasonably large lattice parameter after relaxation.

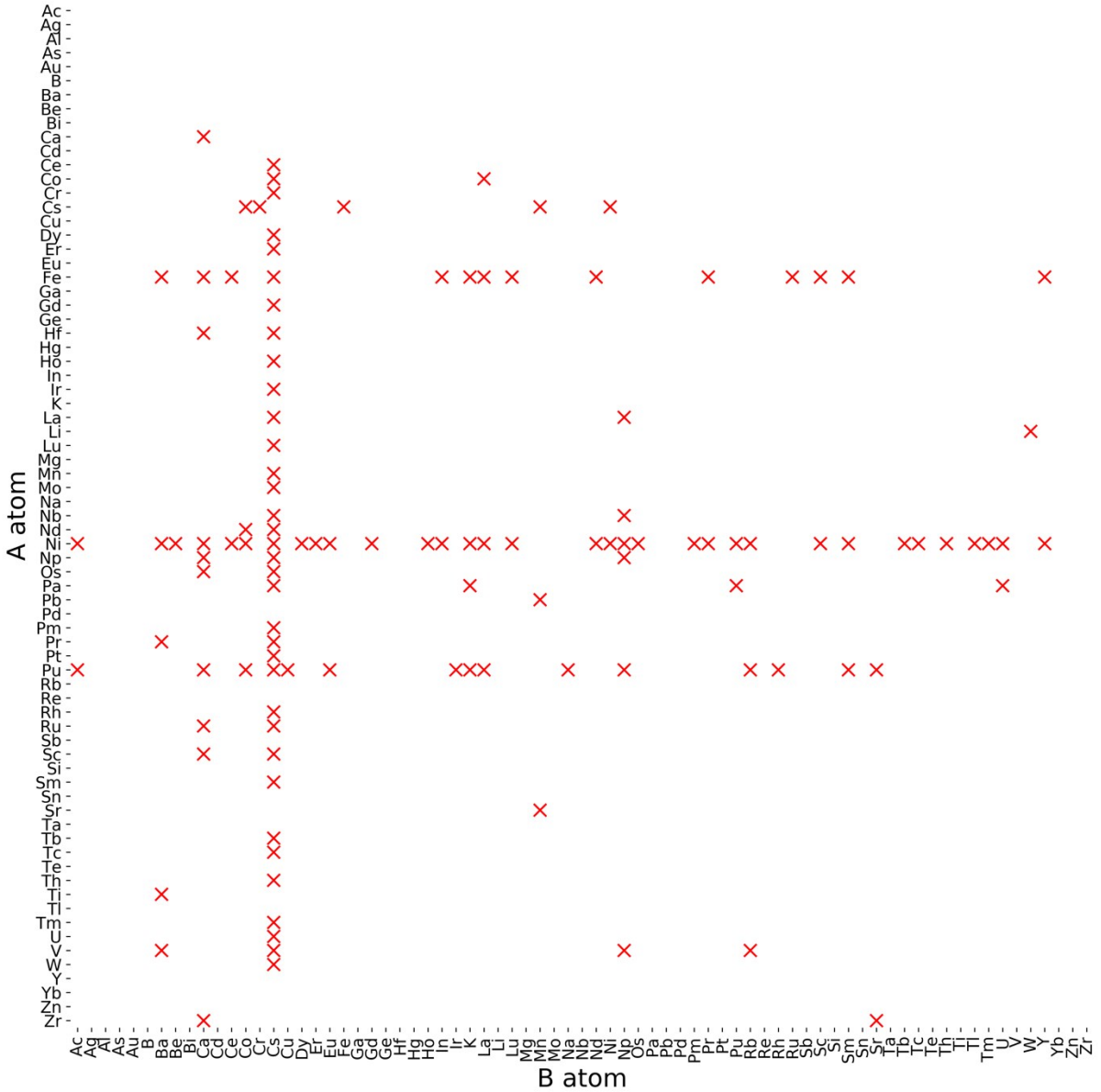

Figure S. 1. The compositions included in the cubic perovskite dataset. X's mark compositions not included in the current dataset. Many Cs compounds were excluded due to identical unrelaxed and relaxed lattice parameters, indicating a possible unnoticed failure of the relaxation calculation. Many Ni compounds failed to produce converged bandgap calculations, as well as some Fe and Pu compounds.

Statistical plots of cubic dataset

Below are included plots showing the distribution of the formation energy in the cubic perovskite dataset with regards to the atom in the A and B sites (Figure S. 2 and Figure S. 3), the observed vs predicted values from the CNN using atomic Hirshfeld surfaces fingerprint plots (Figure S. 5 and Figure S. 6), and the distribution of errors (residuals) from the predictions in regards to the atoms on the A and B sites (Figure S. 7 and Figure S. 8)

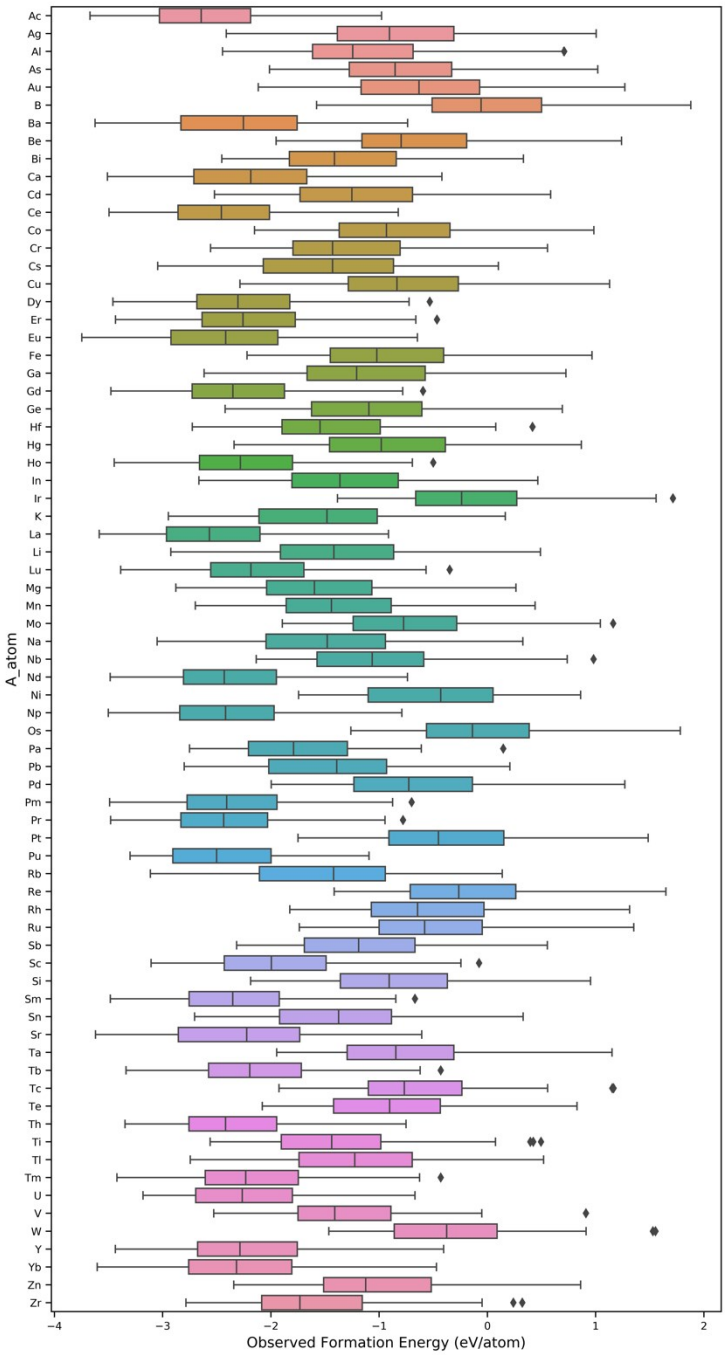

Figure S. 2. Box and whisker plot for the cubic perovskite dataset, showing distribution of DFT formation energies based on the atom in the A site.

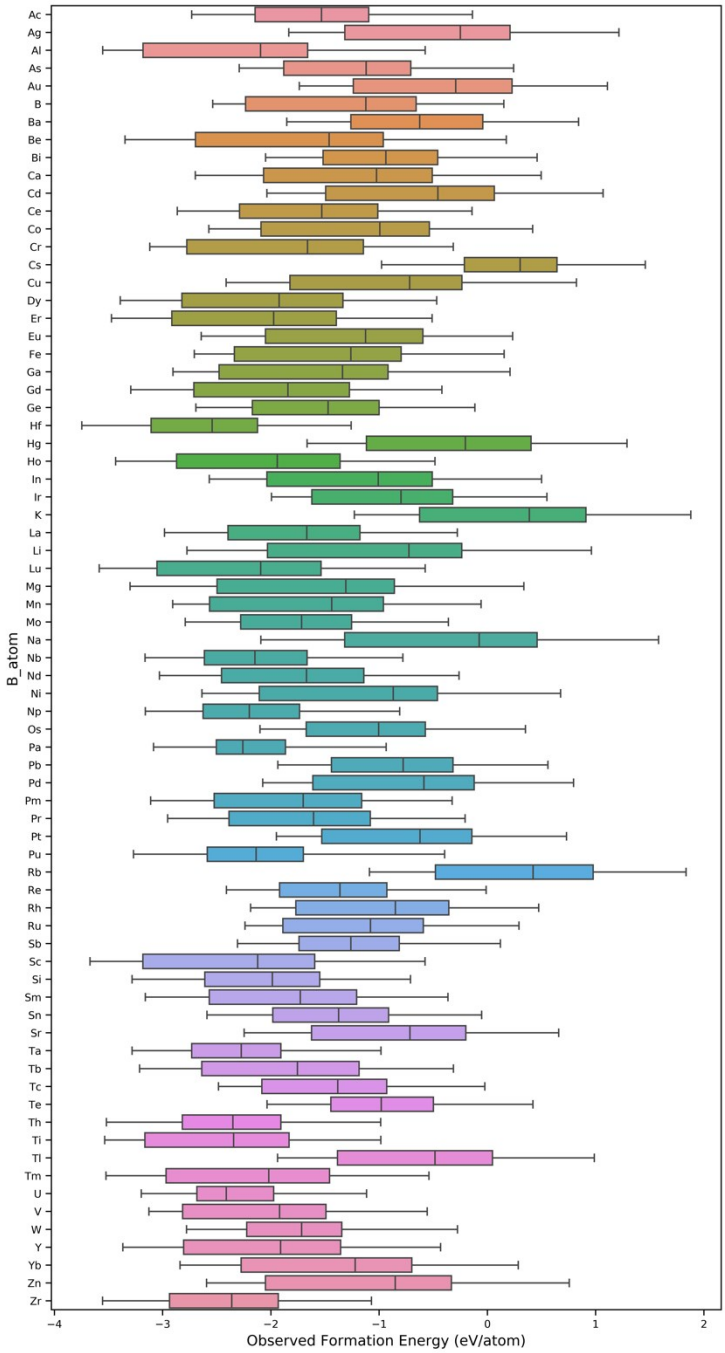

Figure S. 3. Box and whisker plot for the cubic perovskite dataset, showing distribution of DFT formation energies based on the atom in the B site.

Schematic of the CGCNN[3] method used as comparison technique

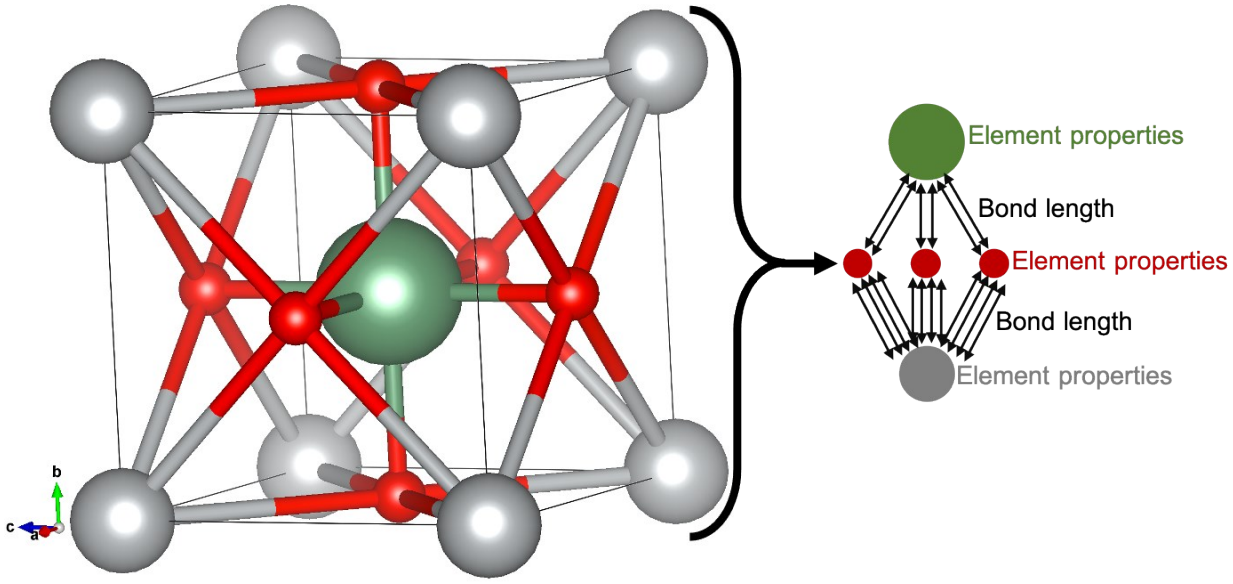

Figure S. 4. Schematic of the features and crystal graph architecture used in the CGCNN[3] method that our results were compared against. The CGCNN method builds a graph network based on atoms and their neighbors within a certain distance. It uses elemental descriptors for the nodes of the graph network and a function of the bond distance for the connections between the nodes of the graph network.

Statistical plots of cubic dataset neural network results

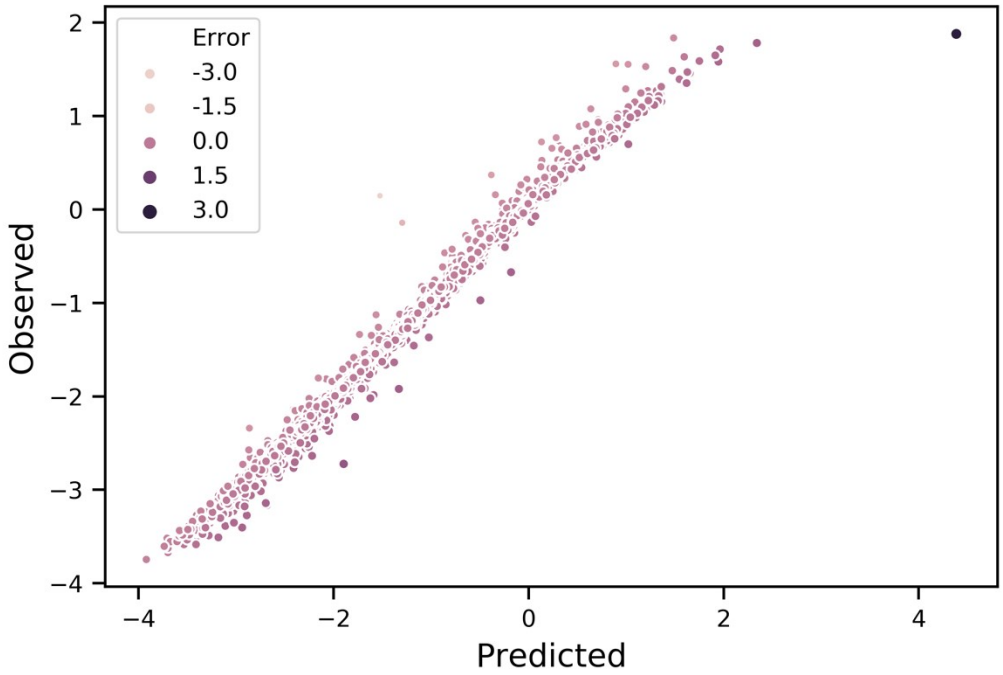

Figure S. 5. Observed vs Predicted plot for the HFS+CNN prediction of cubic perovskite DFT-calculated formation energy, showing the size of the error for each datapoint.

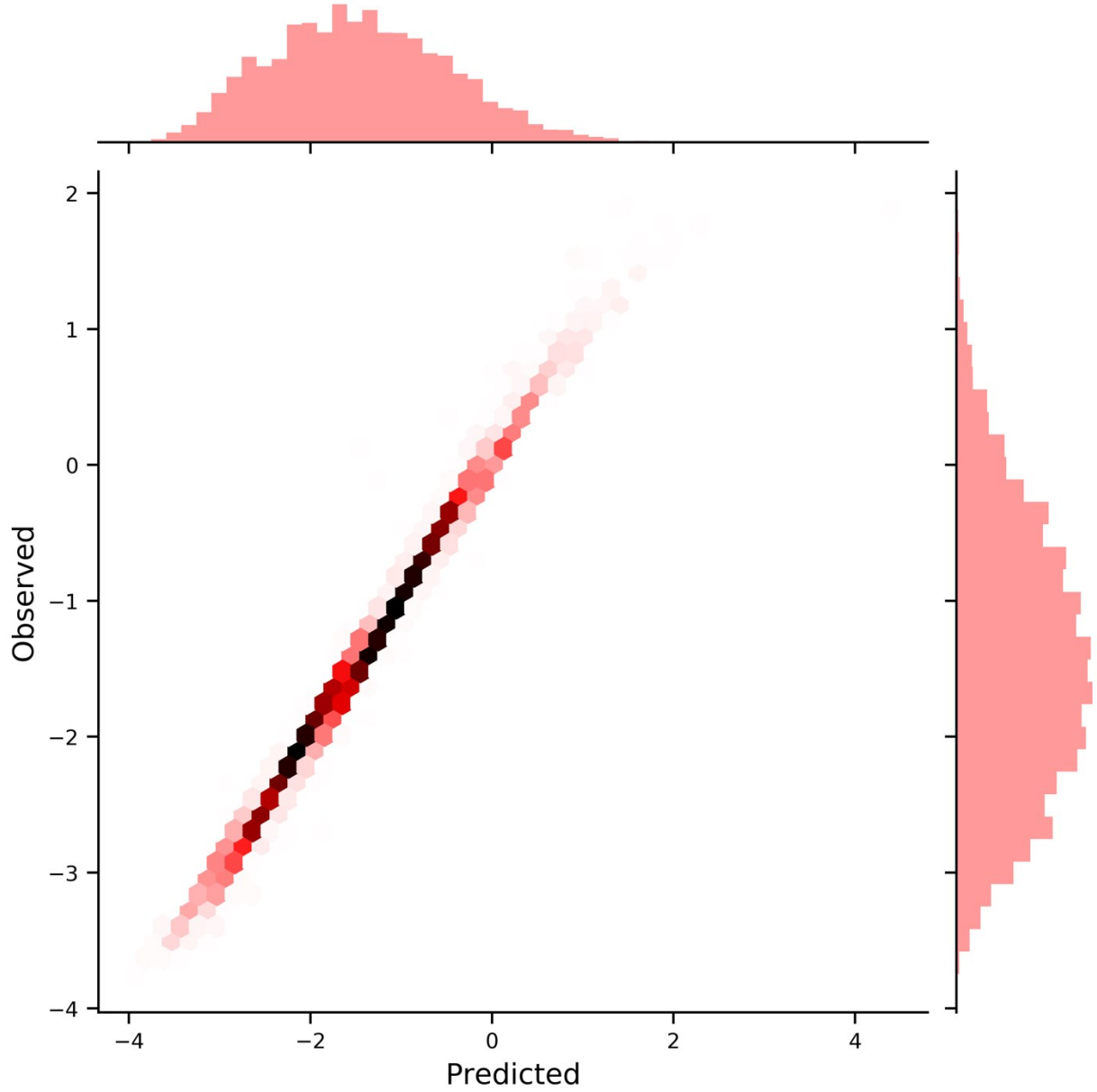

Figure S. 6. Observed vs Predicted plot for the HFS+CNN prediction of cubic perovskite DFT-calculated formation energy, showing the distributions of the dataset and the predictions on the sides of the axes.

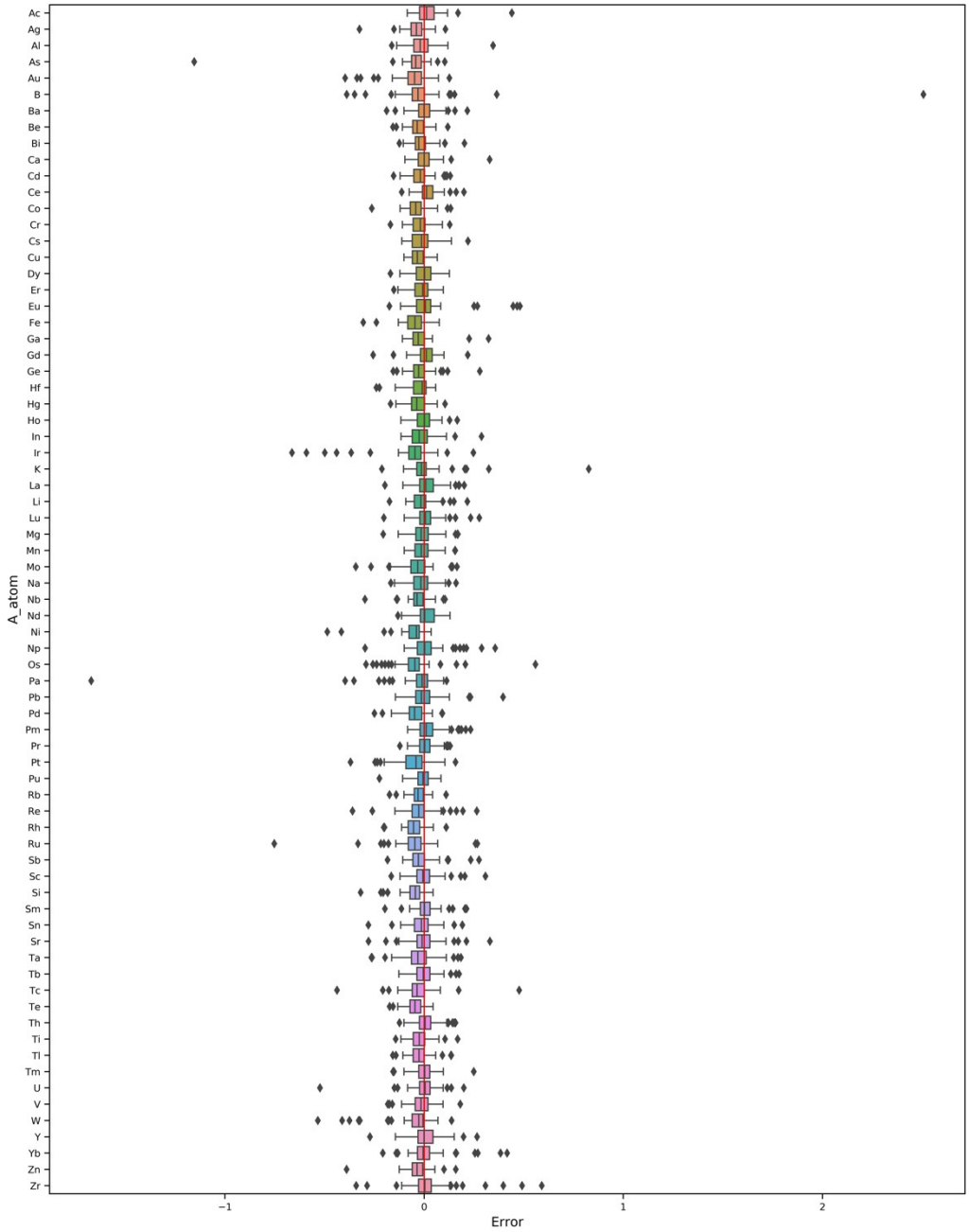

Figure S. 7. Box and whisker plot showing the distribution of the errors from the CNN built on atomic Hirshfeld surface fingerprint plots upon the cubic perovskite dataset based on the atom in the A site. Most systems have highly accurate predictions and large outliers are rare. A few atomic species contain small clusters of small-magnitude outliers.

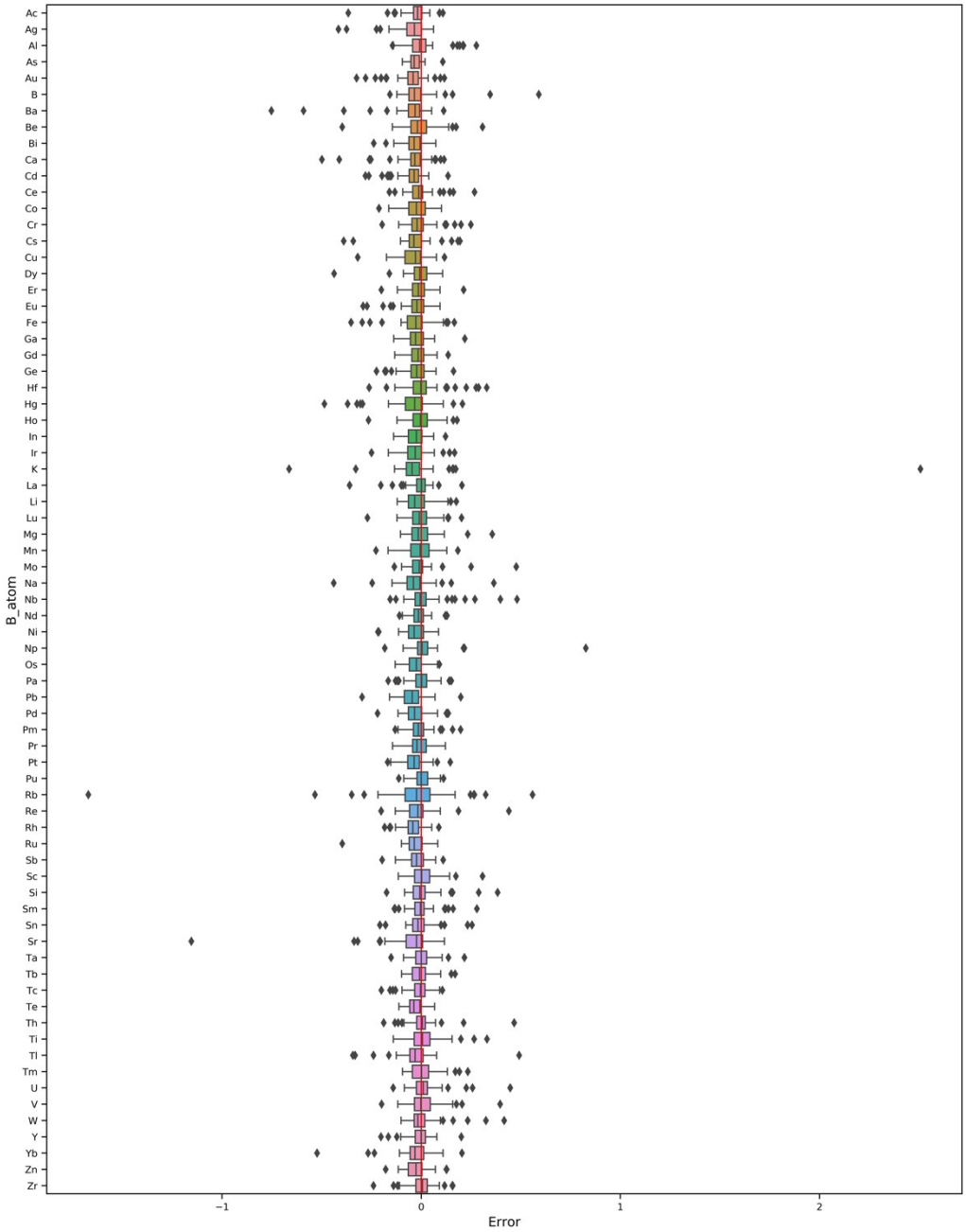

Figure S. 8. Box and whisker plot showing the distribution of the errors from the CNN built on atomic Hirshfeld surface fingerprint plots upon the cubic perovskite dataset based on the atom in the B site. Most systems have highly accurate predictions and large outliers are rare. A few atomic species contain small clusters of small-magnitude outliers.

Tables of outliers for cubic dataset

Below are tables showing the largest outliers for: the CNN based on atomic Hirshfeld surfaces fingerprint plots to predict DFT Formation Energy of the cubic perovskite dataset (Table S 2), the CNN from the prior work[2] based on atomic Hirshfeld surfaces fingerprint plots to predict lattice parameter (Table S 3), that same CNN but with the systematic bias towards estimating towards the mean subtracted out by fitting a linear function to the bias (Table S 4), and the CGCNN prediction of DFT Formation Energy (Table S 5).

Table S 2. Outliers from the CNN prediction of OQMD formation energy of cubic perovskites based on the atomic Hirshfeld surface fingerprint plots that are over 0.5 eV/atom in magnitude. 8 of the 11 outliers contain either Ba, K, Rb, or Sr in the B site. All of these are group 1 or 2 elements and possess similar atomic size (~215-248 pm).

Formula	OQMD ID	A atom	B atom	DFT Energy [eV/atom]	Test or Train	CNN Prediction	CNN Residual
BaIrO3	354509	Ir	Ba	0.72301	Train	0.13193	-0.59108
BaRuO3	351592	Ru	Ba	0.36980	Train	-0.38285	-0.75265
KBO3	354529	B	K	1.87903	Train	4.38580	2.50677
KIrO3	351163	Ir	K	1.55826	Test	0.89468	-0.66358
KNpO3	353363	K	Np	-2.72349	Test	-1.89679	0.82670
RbOsO3	354598	Os	Rb	1.78189	Test	2.34054	0.55864
RbPaO3	350010	Pa	Rb	0.14691	Test	-1.52488	-1.67180
RbWO3	350721	W	Rb	1.55350	Train	1.01992	-0.53358
SrAsO3	353710	As	Sr	-0.14189	Train	-1.29662	-1.15473
YbUO3	352722	U	Yb	-2.34077	Test	-2.86392	-0.52314
ZrBO3	353529	Zr	B	-1.92102	Train	-1.33074	0.59028

The majority (8 of 11) of the large (0.5 eV/atom or greater) outliers from the CNN based on atomic Hirshfeld surfaces fingerprint plots came from compounds with Ba, K, Rb, or Sr in the B site. Although there is significantly more noise in this subset of data than then whole dataset (particularly for the less energetically favorable compounds), it can be seen in Figure S. 9 that the model still makes accurate predictions for many of the compounds containing these elements.

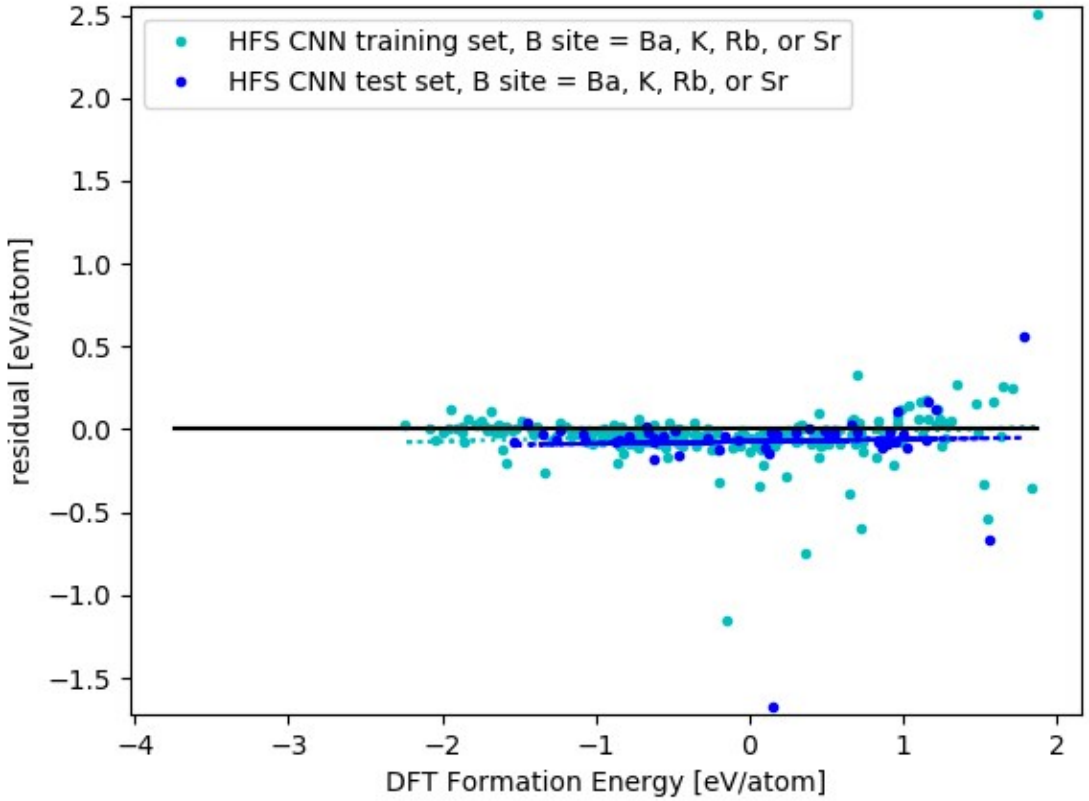

Figure S. 9. Residuals vs DFT calculated system energy for cubic perovskites with either Ba, K, Rb, or Sr in the B site.

Table S 3. Major outliers (error greater than or equal to 0.13 Å in magnitude) for the CNN predicting lattice parameter from our previous work.[2] All are low lattice parameter structures with small elements on the B site.

Formula	OQMD ID	A atom	B atom	Unrelaxed Lattice Parameter	Relaxed Lattice Parameter	CNN Prediction	CNN Residual
BeNbO3	354613	Nb	Be	4.4103	3.4055	3.5672	0.1617
BeOsO3	354294	Os	Be	4.3401	3.4085	3.5419	0.1334
BeSnO3	352081	Sn	Be	4.5602	3.5103	3.6471	0.1368
BiBO3	355209	Bi	B	4.6836	3.5562	3.6887	0.1325
BSbO3	351589	Sb	B	4.6042	3.4912	3.6487	0.1575
BTeO3	353061	Te	B	4.6625	3.5146	3.6640	0.1494
BWO3	354532	W	B	4.3660	3.4006	3.5996	0.1991
ErBeO3	354333	Er	Be	4.6169	3.5176	3.6535	0.1359
EuSiO3	350094	Eu	Si	5.0374	3.5930	3.7297	0.1367
HgBO3	354852	Hg	B	4.5139	3.5175	3.6511	0.1337
LiMgO3	353554	Mg	Li	4.7101	3.5470	3.6903	0.1433
NbBO3	354617	Nb	B	4.4031	3.3997	3.5508	0.1511
ReBO3	350783	Re	B	4.3458	3.3904	3.5278	0.1374
SiRhO3	355285	Rh	Si	4.5384	3.5693	3.7039	0.1346

Table S 4. Major outliers (error greater than or equal to 0.08 Å in magnitude) for the CNN predicting lattice parameter from our previous work[2], after adjustment to correct for a systematic bias towards the norm of the dataset. The results were adjusted by fitting a linear function to the predicted values of the training set vs. the observed (relaxed) values, and applying that linear function to the output from the CNN. No strong trend is seen the composition of the compounds that are the greatest outliers.

Formula	OQMD ID	A atom	B atom	Unrelaxed Lattice Parameter	Relaxed Lattice Parameter	CNN Prediction	Adjusted CNN Prediction	Adjusted CNN residual
AcUO3	353980	Ac	U	4.9045	4.4655	4.5380	4.5501	0.0846
BaSmO3	354376	Ba	Sm	5.3869	4.4759	4.5589	4.5727	0.0968
BeNbO3	354613	Nb	Be	4.4103	3.4055	3.5672	3.5036	0.0981
BSbO3	351589	Sb	B	4.6042	3.4912	3.6487	3.5914	0.1002
BTeO3	353061	Te	B	4.6625	3.5146	3.6640	3.6080	0.0933
BWO3	354532	W	B	4.3660	3.4006	3.5996	3.5385	0.1380
CaBeO3	350609	Be	Ca	4.8057	4.1106	4.0540	4.0284	-0.0822
CaHfO3	351137	Ca	Hf	5.0022	4.0735	4.0199	3.9916	-0.0819
CaMnO3	354167	Mn	Ca	4.8652	4.2177	4.1304	4.1107	-0.1070
EuSiO3	350094	Eu	Si	5.0374	3.5930	3.7297	3.6788	0.0857
GdLuO3	352794	Lu	Gd	4.9635	4.3052	4.3944	4.3953	0.0901
GdPbO3	353224	Pb	Gd	4.9744	4.3756	4.4540	4.4596	0.0840
InWO3	354278	W	In	4.6666	4.0753	4.0219	3.9938	-0.0815
LiMgO3	353554	Mg	Li	4.7101	3.5470	3.6903	3.6363	0.0893
NaSnO3	353611	Sn	Na	5.0139	3.9767	4.0867	4.0636	0.0869
NbBO3	354617	Nb	B	4.4031	3.3997	3.5508	3.4859	0.0862
PuUO3	351925	U	Pu	4.6565	4.4414	4.5137	4.5240	0.0826
ScUO3	351981	U	Sc	4.7254	4.0296	3.9636	3.9309	-0.0987
SiRhO3	355285	Rh	Si	4.5384	3.5693	3.7039	3.6509	0.0817
SmHfO3	353183	Hf	Sm	4.8651	4.3585	4.4394	4.4439	0.0853
SrInO3	352151	Sr	In	5.2032	4.1602	4.0888	4.0659	-0.0943
SrNdO3	351064	Sr	Nd	5.3011	4.4654	4.5449	4.5575	0.0922
YbHoO3	350358	Yb	Ho	5.0926	4.2320	4.3228	4.3181	0.0861

Table S 5. Outliers of magnitude greater than 0.5eV/atom from the CGCNN[3]model's prediction of formation energy trained on the OQMD dataset of cubic perovskites. Of the 24 such outliers, 8 contained either Li, Be, or B in the B site, and 21 contained a row 6 or row 7 element.

Formula	OQMD ID	A atom	B atom	Formation Energy [eV/atom]	Test or Train	CGCNN Prediction	CGCNN Residual
BeFeO3	353229	Fe	Be	0.17434	Train	-0.89083	-1.06517
BeWO3	354408	W	Be	-1.23241	Train	-0.68311	0.54931
CsAuO3	350310	Au	Cs	0.84211	Test	1.41975	0.57765
CsBeO3	353894	Cs	Be	-0.66301	Test	-1.32627	-0.66326
CsBO3	352311	Cs	B	-0.20764	Train	-0.76268	-0.55503
CsSiO3	351473	Cs	Si	-1.43517	Train	-2.00721	-0.57204
GaCoO3	350919	Ga	Co	0.18812	Train	-0.55871	-0.74683
HfBeO3	353628	Hf	Be	-2.72597	Train	-2.21696	0.50900
HfMgO3	355056	Hf	Mg	-2.32129	Test	-1.79176	0.52953
KUO3	352272	K	U	-1.55408	Train	-2.64256	-1.08849
LiHfO3	353246	Hf	Li	-1.96830	Train	-1.39820	0.57010
LiTaO3	352711	Ta	Li	-1.39934	Train	-0.82167	0.57768
LiWO3	352975	W	Li	-0.64769	Train	-0.00977	0.63792
NaCrO3	351362	Na	Cr	-0.47855	Train	-1.47801	-0.99946
NiPtO3	353435	Ni	Pt	-1.66989	Train	-0.19497	1.47492
NiWO3	353281	Ni	W	-0.27606	Train	-0.80590	-0.52984
PaTlO3	351202	Tl	Pa	-2.56301	Test	-2.04409	0.51893
PuWO3	351872	W	Pu	-0.39435	Train	-0.93596	-0.54161
ThMgO3	352136	Th	Mg	-3.30097	Train	-2.71035	0.59062
ThUO3	351718	Th	U	-2.14584	Train	-2.68709	-0.54125
TiIrO3	354340	Ti	Ir	0.07521	Train	-0.68195	-0.75716
UBiO3	350829	Bi	U	-1.71684	Test	-2.25834	-0.54150
UTlO3	352223	Tl	U	-2.74528	Test	-2.13136	0.61392
YbPaO3	352228	Pa	Yb	-0.95121	Train	-1.98884	-1.03763

Reproducibility of cubic dataset results

The network used in the paper was taken from the best of 5 trained networks, as determined by test set r^2 value. The r^2 value for the other networks were 0.966, 0.980, 0.966, and 0.976. The training set r^2 values were 0.981, 0.987, 0.983, and 0.997. The decision to take the best of the 5 trained networks was made to account for the tendency of neural networks to overfit to the training data and perform notably worse upon the test set. For comparison, the CGCNN code trains 3 networks internally with a validation set used to pick the best network.

Fifty different test/train splits than the one used in the main paper were also tested afterwards, using the same initial condition for all trainings. All layers were unfrozen for the retraining. The retrained networks displayed a tendency towards overfitting half of the time, with 25 of the 50 networks overfit, either moderately or strongly. The average test set r^2 of the 50 networks was 0.964 and the average training set r^2 was 0.994. The test set r^2 distribution for the 50 models was roughly trimodal, split into well fit, overfit, and strongly overfit categories. For the 7 networks that were strongly overfit, the average test set r^2 was 0.918 and the average training set r^2 was 0.994. For the 18 models that were moderately overfit, the average test set r^2 was 0.951 and the average training set r^2 was 0.994. For the 25 models that were well fit, the average test set r^2 was 0.986 and the average training set r^2 was 0.995.

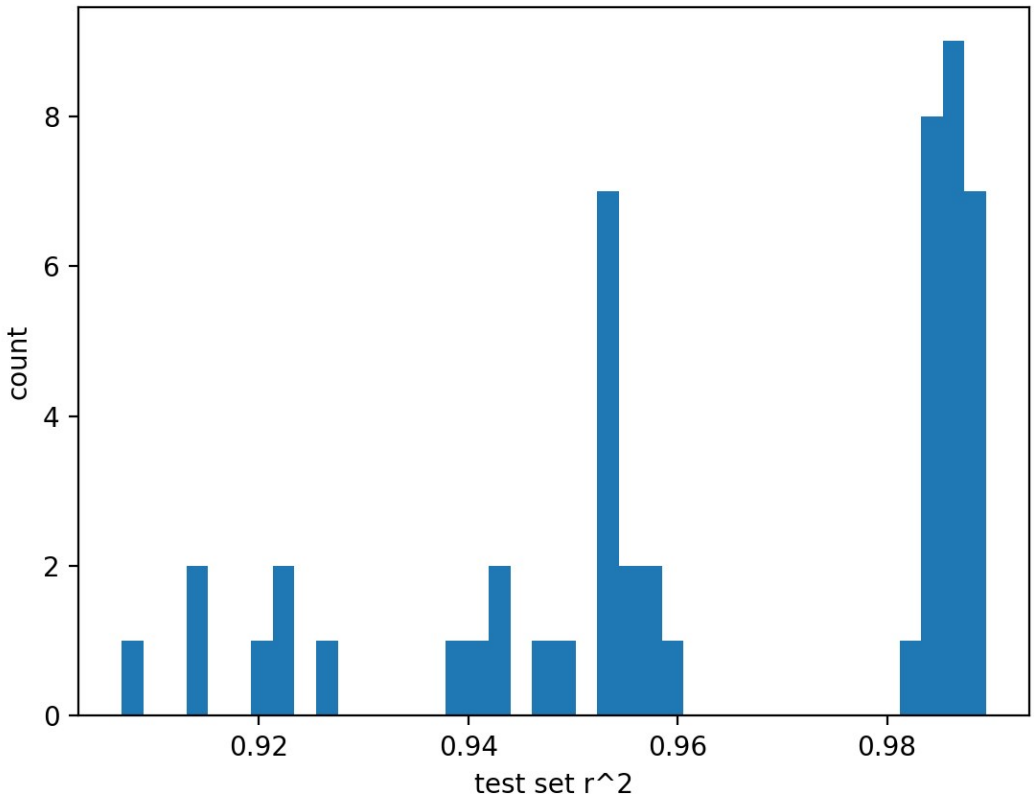

Figure S. 10. Histogram of the distribution of the test set r^2 values for the 50 different test/train split retrainings of the network. The test set results are trimodal, with a cluster well fit around 0.986, another ranging ~0.94-0.96, and a third spread around ~0.92.

When retraining the networks with the feature extraction layers frozen, the average test set r^2 was 0.943 and the average training set r^2 was 0.944, showing that using frozen feature extraction layers from a different split of the total dataset limits the accuracy of the network.

Histogram of perovskites within OQMD by structure prototype

The OQMD database contains more calculated compositions using the cubic perovskite prototype than it does of the other three perovskite prototypes they use. This creates an imbalanced dataset, with many compositions either having data for all four structure types or for only the cubic structure. To create a dataset more suitable for machine learning, the dataset was restricted to only compositions where non-cubic structures were also calculated within OQMD.

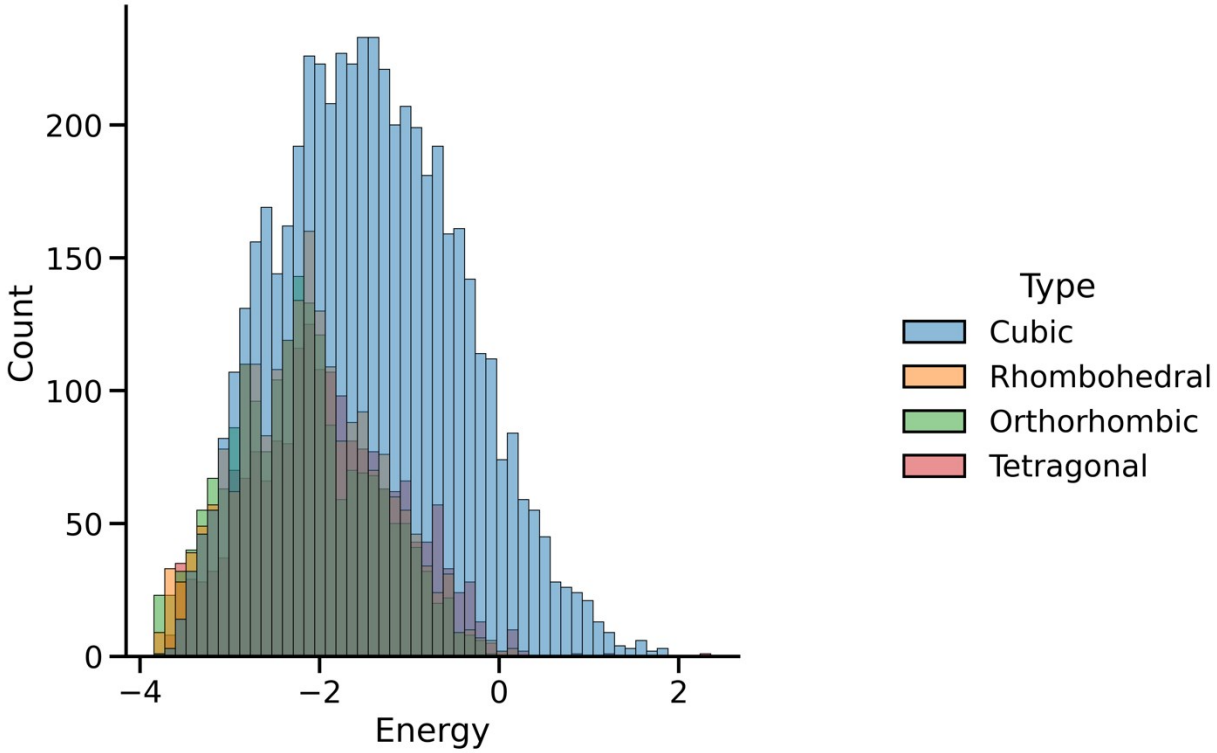

Figure S. 11. The distribution of formation energies amongst the perovskites in OQMD built from the cubic, rhombohedral, orthorhombic, and tetragonal perovskite prototype structures. Approximately half of the cubic perovskites in OQMD have the three non-cubic structures also calculated for their composition, with this split being primarily in favor of the lowest energy cubic perovskites.

Compositions included in the cubic and non-cubic dataset

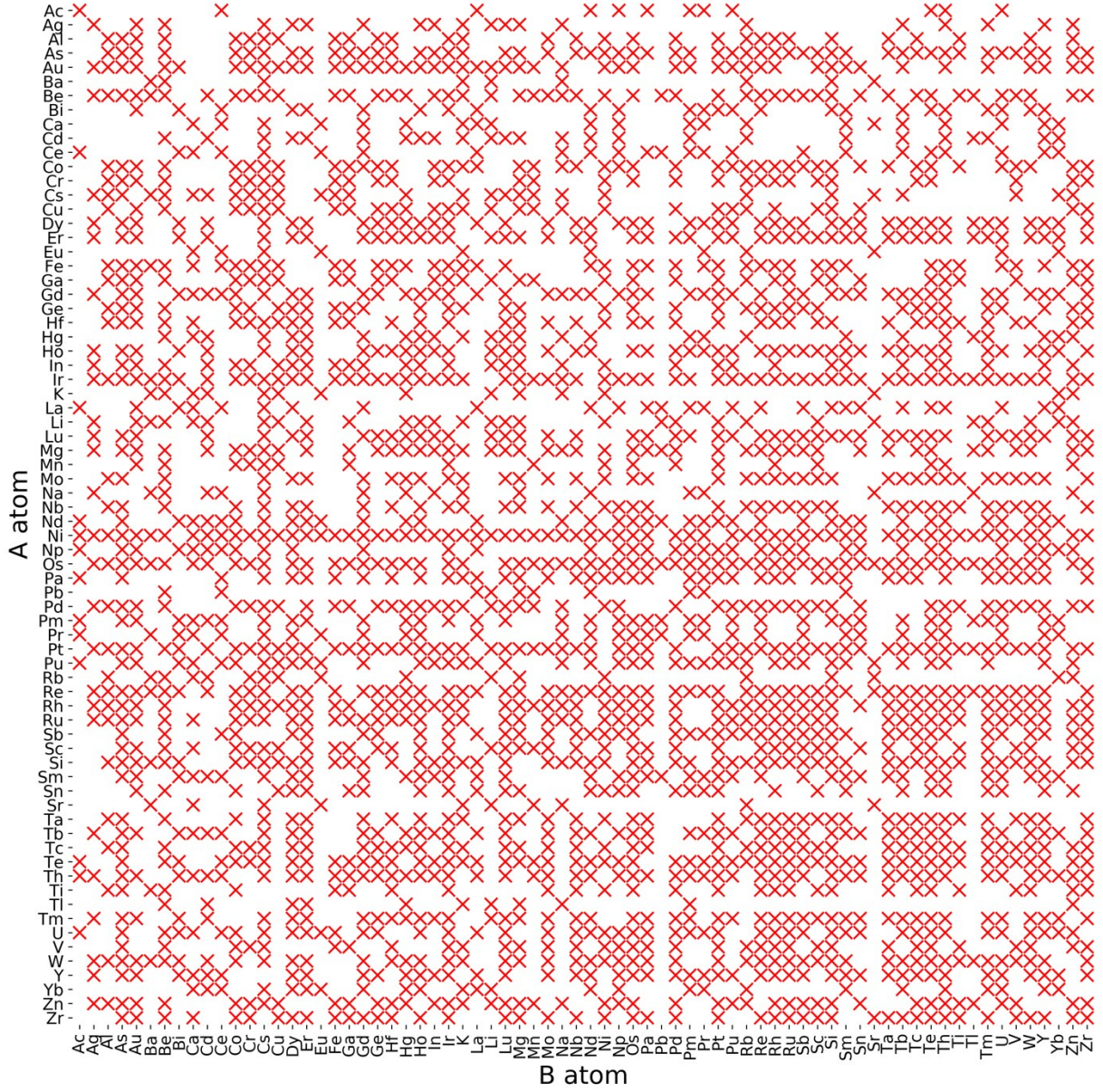

Figure S. 12. The compositions included in the cubic perovskites and non-cubic variants study. X's mark compositions not included in the current dataset. Many cubic compounds with higher formation energies were not included in the list of non-cubic perovskites to calculate by the OQMD[4] project.

Tables of outliers for the cubic and non-cubic dataset

Table S 6. Outliers of magnitude greater than 0.5 eV/atom from the boosted CNN using atomic Hirshfeld surface fingerprint plots on the cubic and non-cubic perovskites dataset. 12 of the 23 outliers contain either K, Rb, or Ba, similar to the outliers from the cubic only dataset. All but two of the major outliers exist for either a cubic or a rhombohedral phase.

OQMD ID	Formation Energy [eV/atom]	Formula	Structure Type	Test or Train	Boosted CNN Prediction	Boosted CNN Residual
680527	-0.53978	NaOsO3	TetraPerovskite_PbTiO3	train	0.59397	1.13375
351071	0.30148	RbNaO3	CubicPerovskite_SrFeO3	train	0.89237	0.59089
827283	-0.19652	UVO3	RhombPerovskite_NdAlO3	train	0.58261	0.77913
352670	-0.13769	AcOsO3	CubicPerovskite_SrFeO3	test	-0.92652	-0.78883
352220	0.01141	AgPdO3	CubicPerovskite_SrFeO3	test	-0.49717	-0.50858
825448	-2.28222	BaEuO3	RhombPerovskite_NdAlO3	test	-2.81651	-0.53428
353778	0.21877	BaMoO3	CubicPerovskite_SrFeO3	test	-0.33410	-0.55287
352432	0.31813	BaTaO3	CubicPerovskite_SrFeO3	test	-0.29747	-0.61560
354832	0.45994	BiOsO3	CubicPerovskite_SrFeO3	test	-0.54963	-1.00957
354078	-1.90141	GdRhO3	CubicPerovskite_SrFeO3	test	-1.27810	0.62331
353567	0.16618	K2O3	CubicPerovskite_SrFeO3	test	3.03919	2.87301
682719	-0.84201	K2O3	OrthoPerovskite_GdFeO3	test	0.13856	0.98057
826181	0.06963	K2O3	RhombPerovskite_NdAlO3	test	1.99372	1.92408
353595	0.15070	KRbO3	CubicPerovskite_SrFeO3	test	1.86324	1.71254
354728	0.13718	KRbO3	CubicPerovskite_SrFeO3	test	1.69408	1.55690
826197	0.11809	KRbO3	RhombPerovskite_NdAlO3	test	0.98956	0.87146
681206	0.17827	KRbO3	TetraPerovskite_PbTiO3	test	0.83954	0.66128
826809	0.08369	KRbO3	RhombPerovskite_NdAlO3	test	0.59725	0.51355
352055	-1.32336	LiNbO3	CubicPerovskite_SrFeO3	test	-0.69393	0.62943
827216	-2.18036	PuTlO3	RhombPerovskite_NdAlO3	test	-1.64255	0.53780
352939	0.46616	TaHgO3	CubicPerovskite_SrFeO3	test	-0.35485	-0.82101
825737	-2.72396	TbCrO3	RhombPerovskite_NdAlO3	test	-2.10865	0.61532
351910	-2.28914	ZrBiO3	CubicPerovskite_SrFeO3	test	-1.53902	0.75011

Table S 7. Outliers of magnitude greater than 0.5 eV/atom from the CGCNN[3] model on the cubic and non-cubic perovskites dataset. Row 6 and 7 elements are in 29 of the 40 outliers. Half of the 40 outliers are for tetragonal structures, and only 2 were orthogonal structures.

OQMD ID	Formation Energy [eV/atom]	Formula	Structure Type	Test or Train	CGCNN Prediction	CGCNN Residual
680866	-0.75628	AlInO3	TetraPerovskite_PbTiO3	train	-1.47412	-0.71783
681887	-2.78652	BaLaO3	OrthoPerovskite_GdFeO3	train	-2.28605	0.50047

826932	-2.99716	CaSiO3	RhombPerovskite_NdAlO3	train	-2.36958	0.62758
680403	-1.74163	CaSiO3	TetraPerovskite_PbTiO3	train	-2.37153	-0.62990
680287	0.10016	CdMoO3	TetraPerovskite_PbTiO3	train	-0.89109	-0.99125
350688	-2.52116	ErNpO3	CubicPerovskite_SrFeO3	train	-3.05845	-0.53730
825501	-2.79764	GdBeO3	RhombPerovskite_NdAlO3	train	-2.20389	0.59375
352272	-1.55408	KUO3	CubicPerovskite_SrFeO3	train	-3.23986	-1.68578
826277	-0.15661	NaLiO3	RhombPerovskite_NdAlO3	train	-0.67279	-0.51618
680803	-1.75433	NbCrO3	TetraPerovskite_PbTiO3	train	-2.42981	-0.67548
353435	-1.66989	NiPtO3	CubicPerovskite_SrFeO3	train	-0.03333	1.63655
352361	-0.83158	PaTeO3	CubicPerovskite_SrFeO3	train	-1.34851	-0.51693
681342	-1.17059	PuCrO3	TetraPerovskite_PbTiO3	train	-2.72533	-1.55474
681194	-1.95281	PuNiO3	TetraPerovskite_PbTiO3	train	-2.62112	-0.66831
351872	-0.39435	PuWO3	CubicPerovskite_SrFeO3	train	-1.17295	-0.77860
826854	-1.73944	RbReO3	RhombPerovskite_NdAlO3	train	-1.21089	0.52854
680792	-1.19144	SrReO3	TetraPerovskite_PbTiO3	train	-1.71811	-0.52667
680149	-1.75642	UCdO3	TetraPerovskite_PbTiO3	train	-2.31174	-0.55532
352228	-0.95121	YbPaO3	CubicPerovskite_SrFeO3	train	-1.78830	-0.83709
679797	0.21709	AgMoO3	TetraPerovskite_PbTiO3	test	-0.70000	-0.91709
681168	-1.40193	AlAgO3	TetraPerovskite_PbTiO3	test	-2.33149	-0.92956
680426	-0.72753	AlBiO3	TetraPerovskite_PbTiO3	test	-1.65915	-0.93162
680292	-0.73329	BaReO3	TetraPerovskite_PbTiO3	test	-1.41438	-0.68109
680890	-1.13465	CaBeO3	TetraPerovskite_PbTiO3	test	-2.04233	-0.90768
680989	0.13304	CsGeO3	TetraPerovskite_PbTiO3	test	-0.46104	-0.59408
826933	-3.08026	EuSiO3	RhombPerovskite_NdAlO3	test	-2.27173	0.80854
352466	-2.54720	HoNpO3	CubicPerovskite_SrFeO3	test	-3.07189	-0.52469
826809	0.08369	KRbO3	RhombPerovskite_NdAlO3	test	-0.47780	-0.56149
350253	-2.28401	LiUO3	CubicPerovskite_SrFeO3	test	-2.84502	-0.56100
680108	-1.18248	LiUO3	TetraPerovskite_PbTiO3	test	-3.81640	-2.63392
680732	-0.96928	NdNiO3	TetraPerovskite_PbTiO3	test	-1.72204	-0.75276
350865	-1.29518	PuPbO3	CubicPerovskite_SrFeO3	test	-2.09440	-0.79922
352717	-0.98233	PuTcO3	CubicPerovskite_SrFeO3	test	-1.49274	-0.51042
680048	-0.56025	RbMoO3	TetraPerovskite_PbTiO3	test	-1.09075	-0.53050
680736	-0.52846	RbWO3	TetraPerovskite_PbTiO3	test	-1.11886	-0.59040
680363	-0.70944	TiHgO3	TetraPerovskite_PbTiO3	test	-1.45476	-0.74532
825404	-2.90772	UAlO3	RhombPerovskite_NdAlO3	test	-3.52657	-0.61885
682925	-2.93972	UAlO3	OrthoPerovskite_GdFeO3	test	-3.49668	-0.55696
681228	-1.47288	YbBeO3	TetraPerovskite_PbTiO3	test	-1.97868	-0.50580
826935	-3.09632	YbSiO3	RhombPerovskite_NdAlO3	test	-2.13958	0.95673

Confusion plots for the cubic and non-cubic dataset neural network results

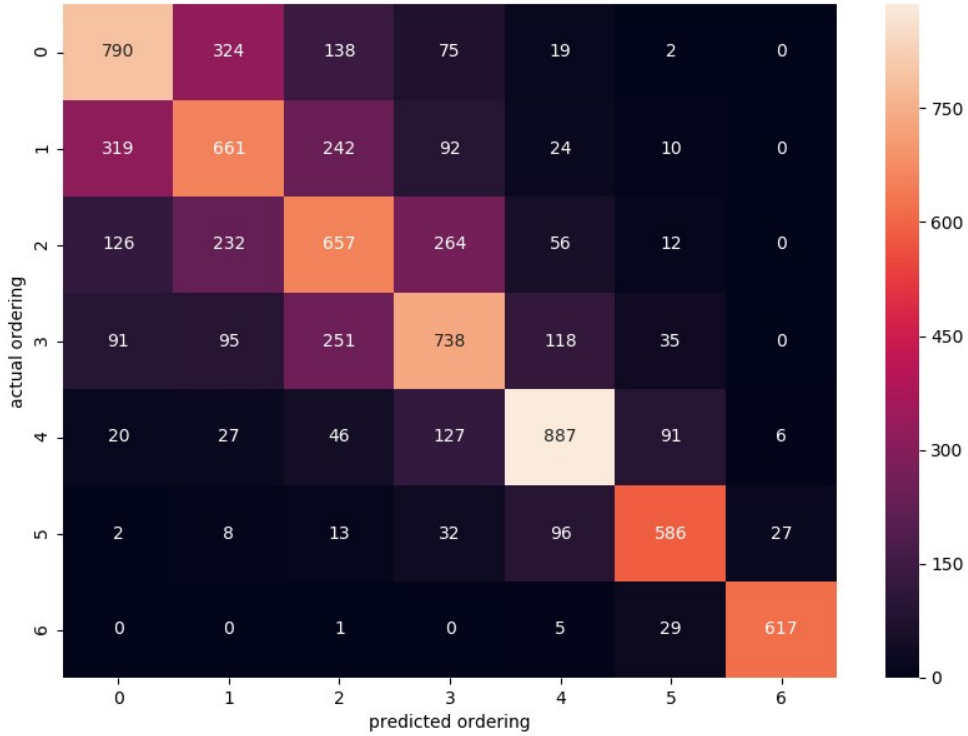

Figure S. 13. Confusion plot for the relative ordering of same composition phases in the full cubic and non-cubic perovskite dataset as produced by our model built using a CNN and Hirshfeld surface fingerprint plots.

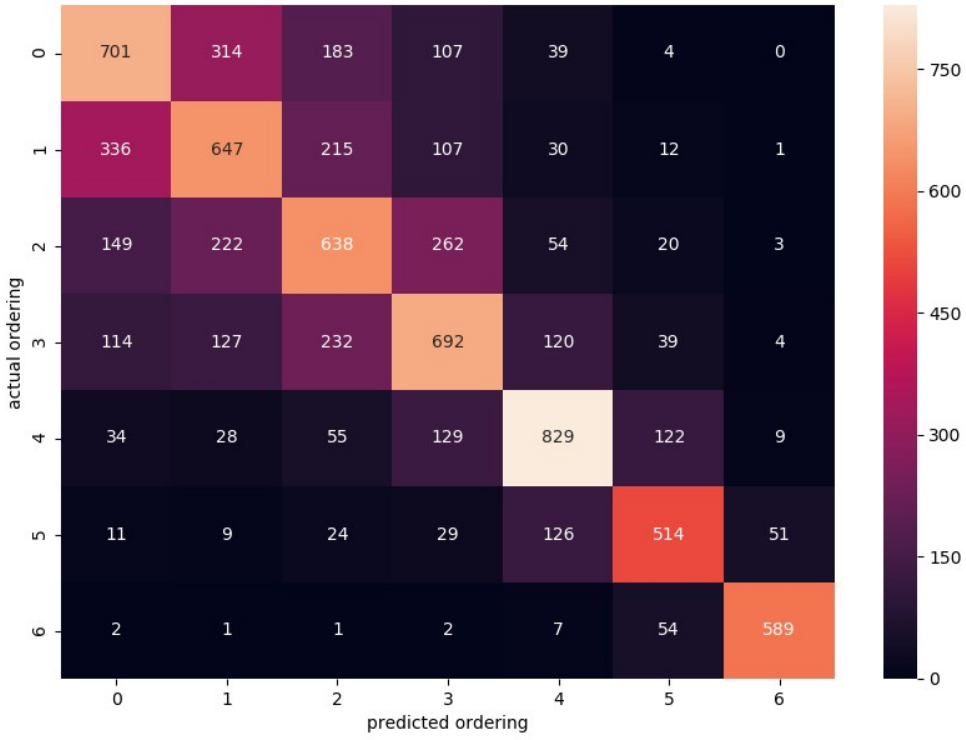

Figure S. 14. Confusion plot for the relative ordering of same composition phases in the full cubic and non-cubic perovskite dataset as produced by the CGCNN[3] model.

Neural Network qualitative visualization

The following figure (Figure S. 15) is included to qualitatively highlight the nonlinearity of the relationship between the features in the input to the neural network (the atomic Hirshfeld surfaces fingerprint plot) and the formation energy of the crystal structure. Each plot shows a visualization of the latent space for the entire dataset at different points through the neural network. The latent space refers to the output of the intermediate layers in a deep neural network that contains the geometric features as captured by the network through progressive layers of abstraction. Here we have performed a t-SNE[5] on the output of the intermediate convolution and fully connected layers to generate a compressed visualization of the geometric features hidden in the fingerprint plots in Figure S. 15. The output of the convolution layers are usually complex sets of multidimensional matrices that are flattened to give a vector before feeding it as an input to the t-SNE framework. The outputs of the dense layers are fed as is. The t-SNE is a manifold learning algorithm that computes the structural similarities of data in high-dimension and produces a lower order representation while preserving the topology of the original data. The latent space visualization gives a global and at times hierarchical structure of the overall dataset as it moves along the forward direction of a deep network. Although the absolute coordinates do not bear any significance, the proximity between the compounds in 2D is a representation of how they would have been in the high-dimensional space. For example, the t-SNE reduces each 50x50 input fingerprint plot to a 2D representation in Figure S. 15 (upper left) by approximately mimicking the similarities between the fingerprint plots for all the compounds. Thus, the compounds with similar looking fingerprint plots sit closer to each other , yet they show only minimal correlation with the formation energy. As the dataset is processed by the neural network, the proximities of the network's internal representations of the compounds change significantly and the global structure evolves into a clear trend in formation energy with respect to the latent space. This plot visualizes the level of nonlinear complexity that the network learns from the beginning till the end. Figure S. 15 (lower right) displays a smooth gradation of compound coloring in terms of the formation energy, showing that the network is able to find out combinations of geometric motifs inside the fingerprint plots that are in near perfect correlation with the output.

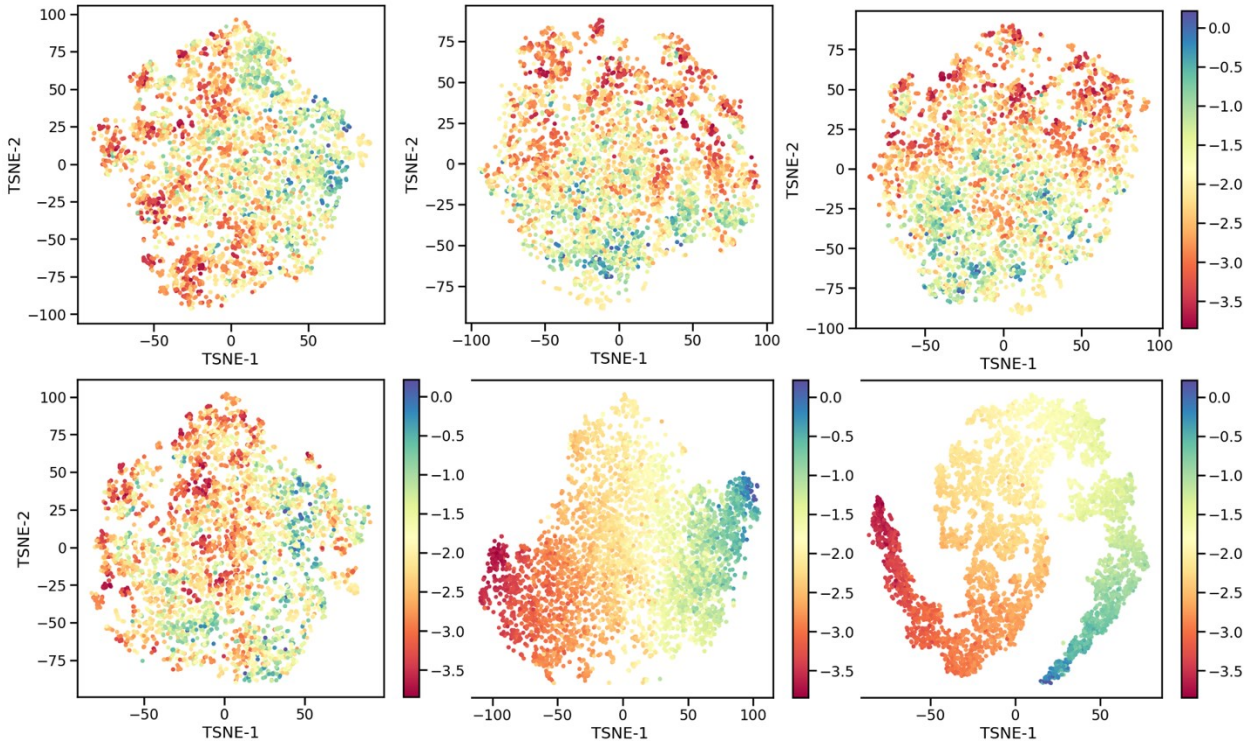

Figure S. 15. Latent Space Visualization using t-SNE for the (upper left) input layer, (upper middle) first convolution layer, (upper right) 5th convolution layer, (lower left), 9th convolution layer, (lower middle) first dense layer, and (lower right) final dense layer. The t-SNE methodology shows the higher-dimensional similarity relationships between the input (upper left) or mid-network layer output (the rest) for all the compounds in the dataset flattened into two dimensions. The relationship is shown to be highly nonlinear with regards to the input, but the final network layer's latent space is easily correlated to the formation energy.

Statistical plots for the cubic and non-cubic dataset based on structure analysis

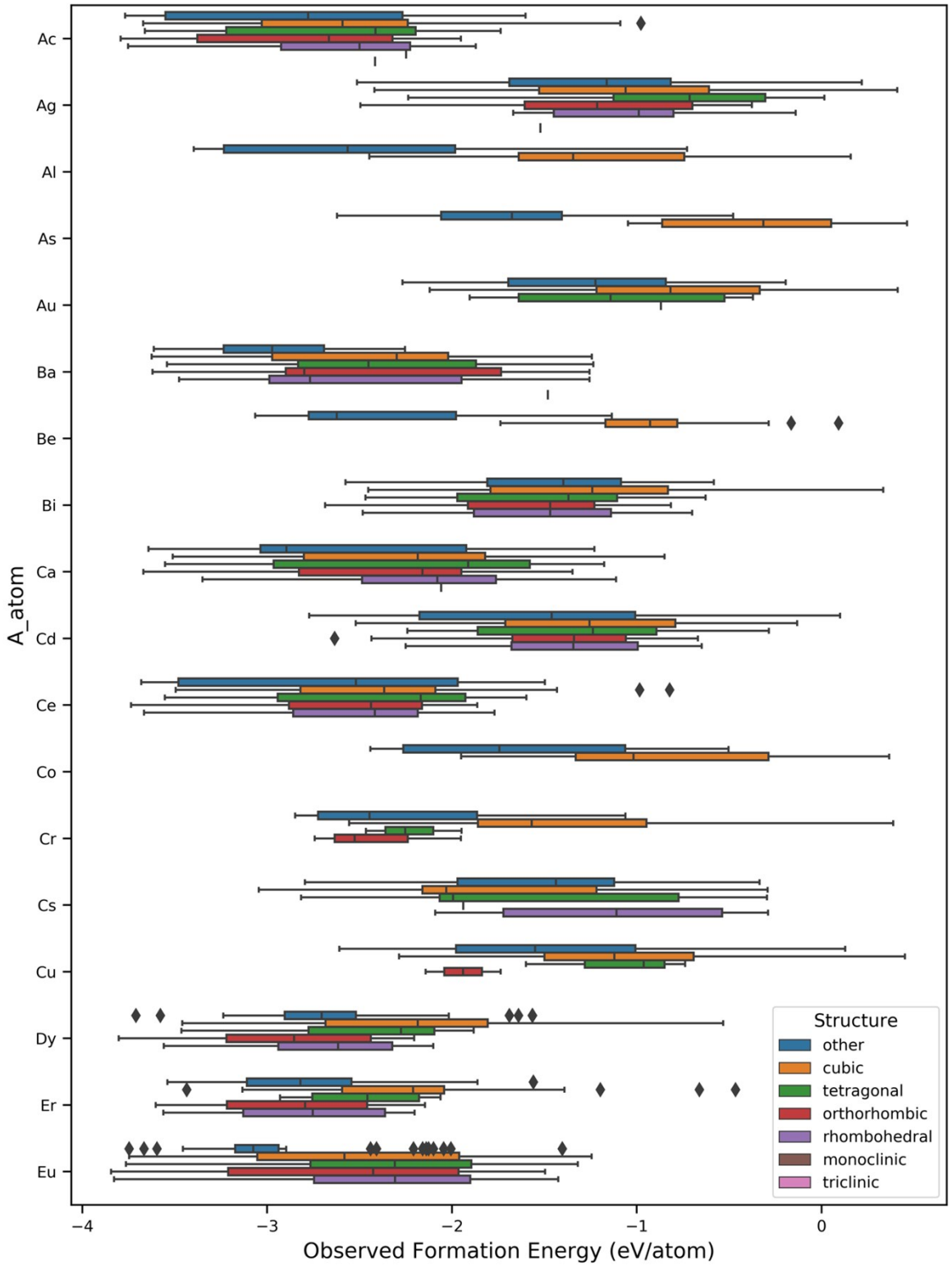

Figure S. 16. Ranges of DFT predicted formation energies vs. the atom placed in the A position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the A-E range.

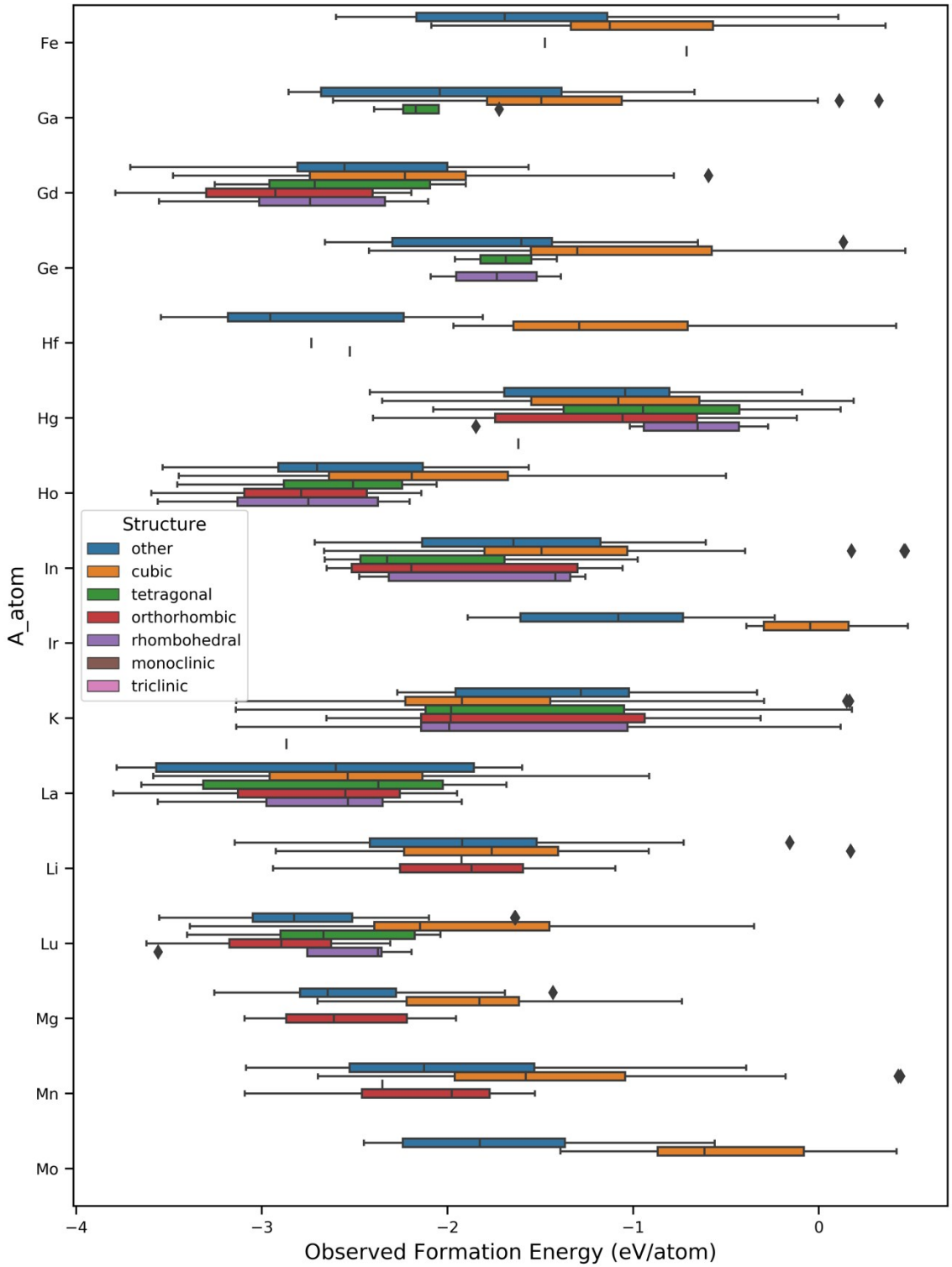

Figure S. 17. Ranges of DFT predicted formation energies vs. the atom placed in the A position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the A-H range. This plot contains elements in the F-M range.

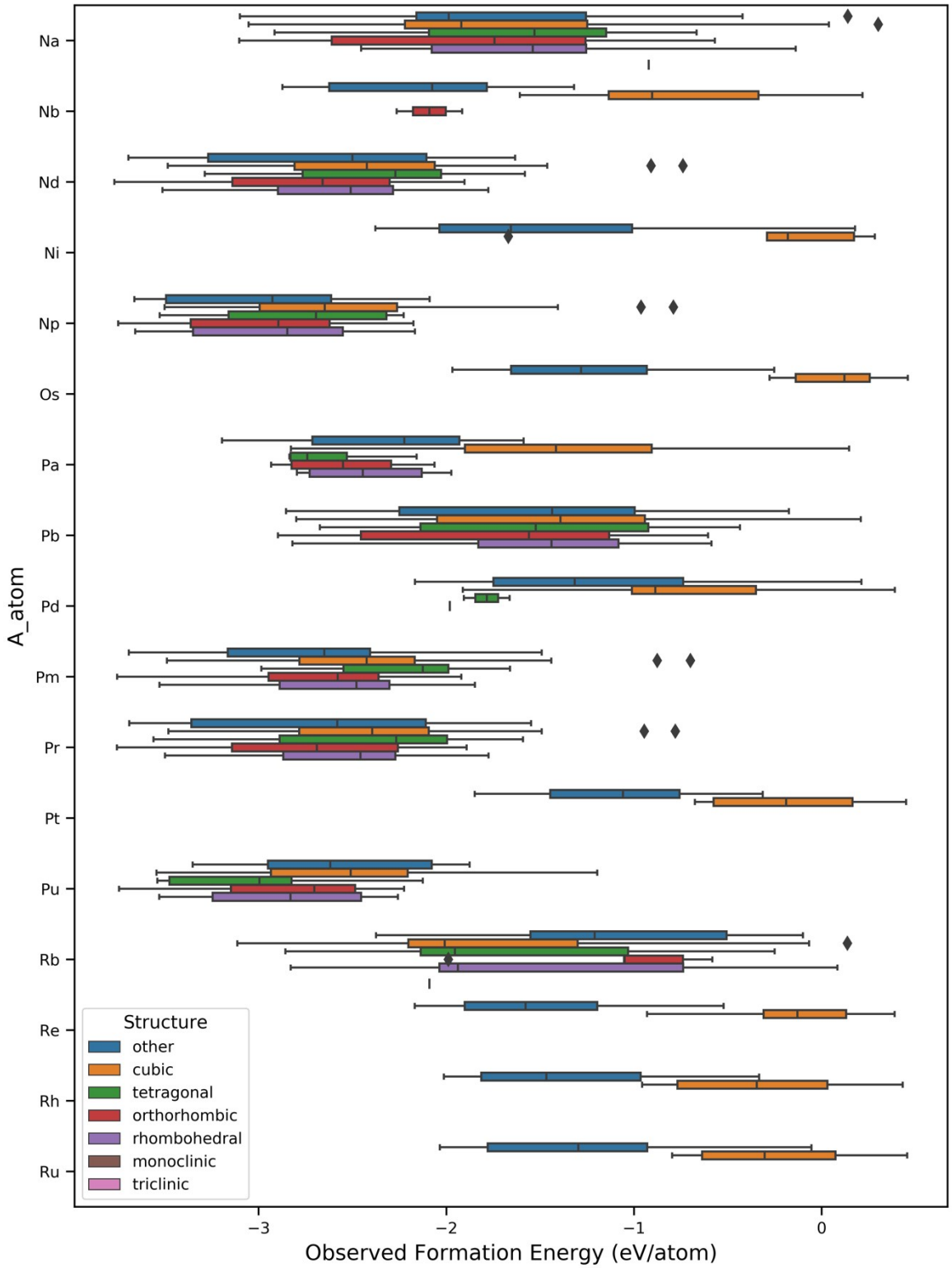

Figure S. 18. Ranges of DFT predicted formation energies vs. the atom placed in the A position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the N-R range.

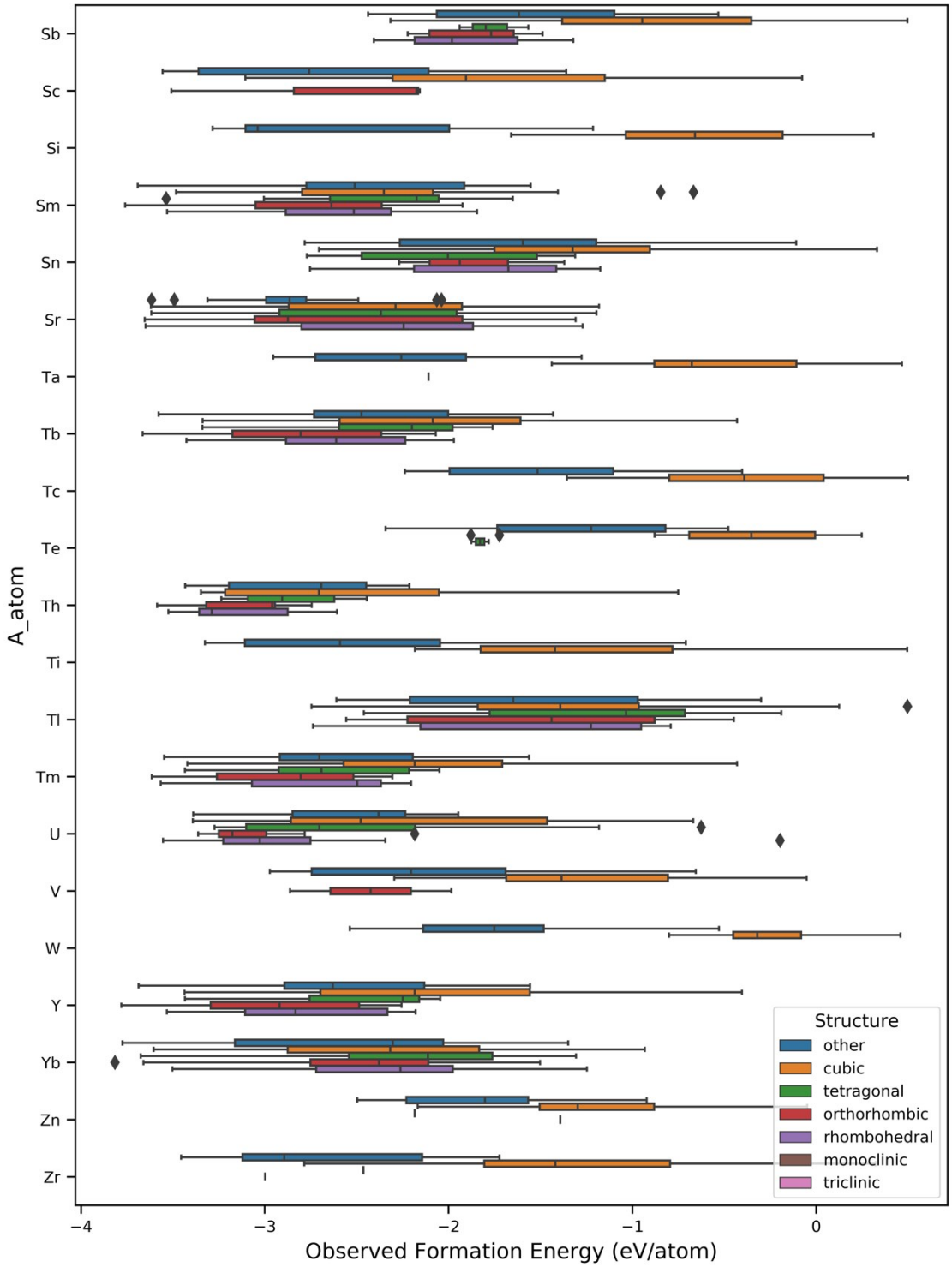

Figure S. 19. Ranges of DFT predicted formation energies vs. the atom placed in the A position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the S-Z range.

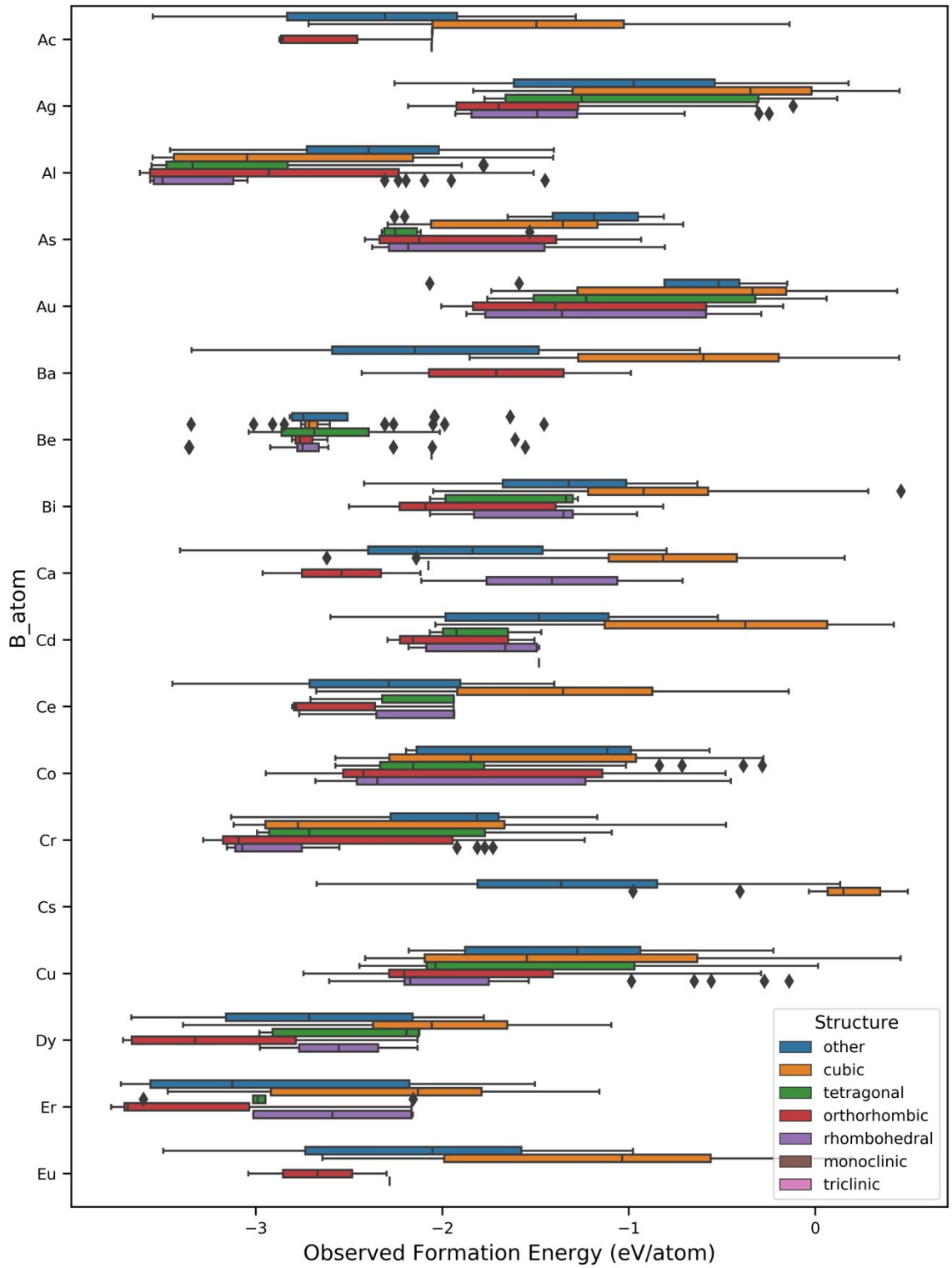

Figure S. 20. Ranges of DFT predicted formation energies vs. the atom placed in the B position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the A-E range.

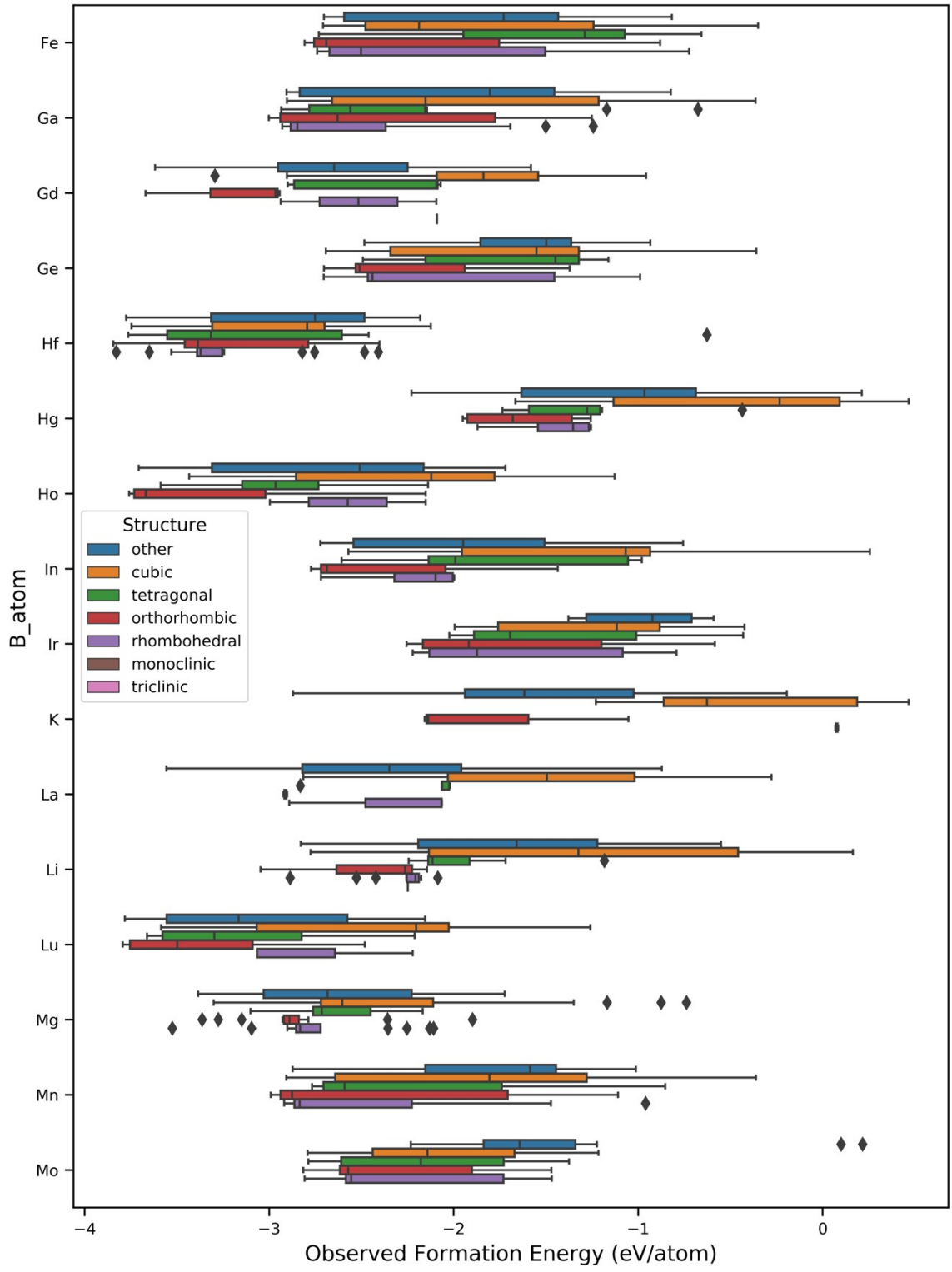

Figure S. 21. Ranges of DFT predicted formation energies vs. the atom placed in the B position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the F-M range.

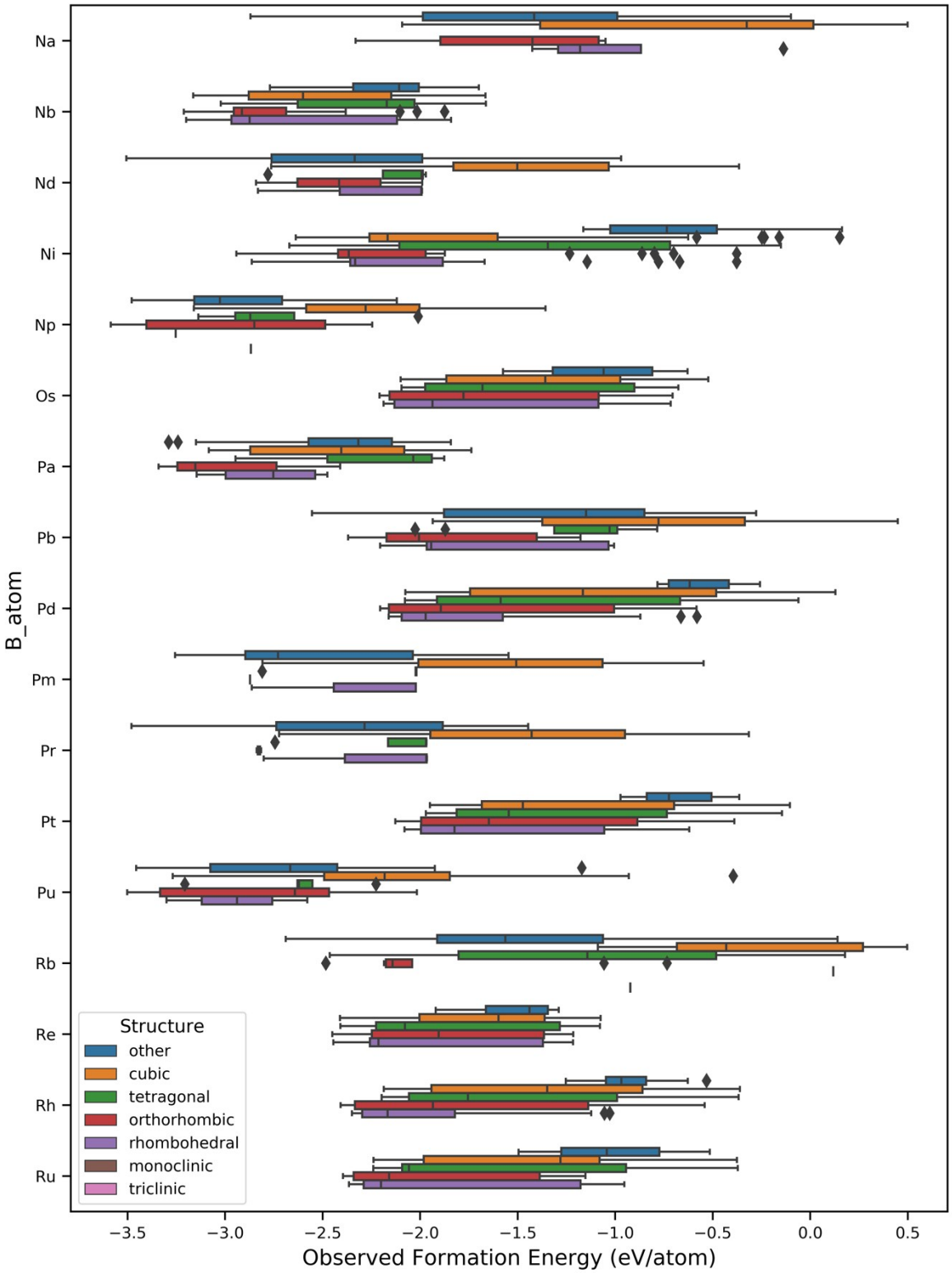

Figure S. 22. Ranges of DFT predicted formation energies vs. the atom placed in the B position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the N-R range.

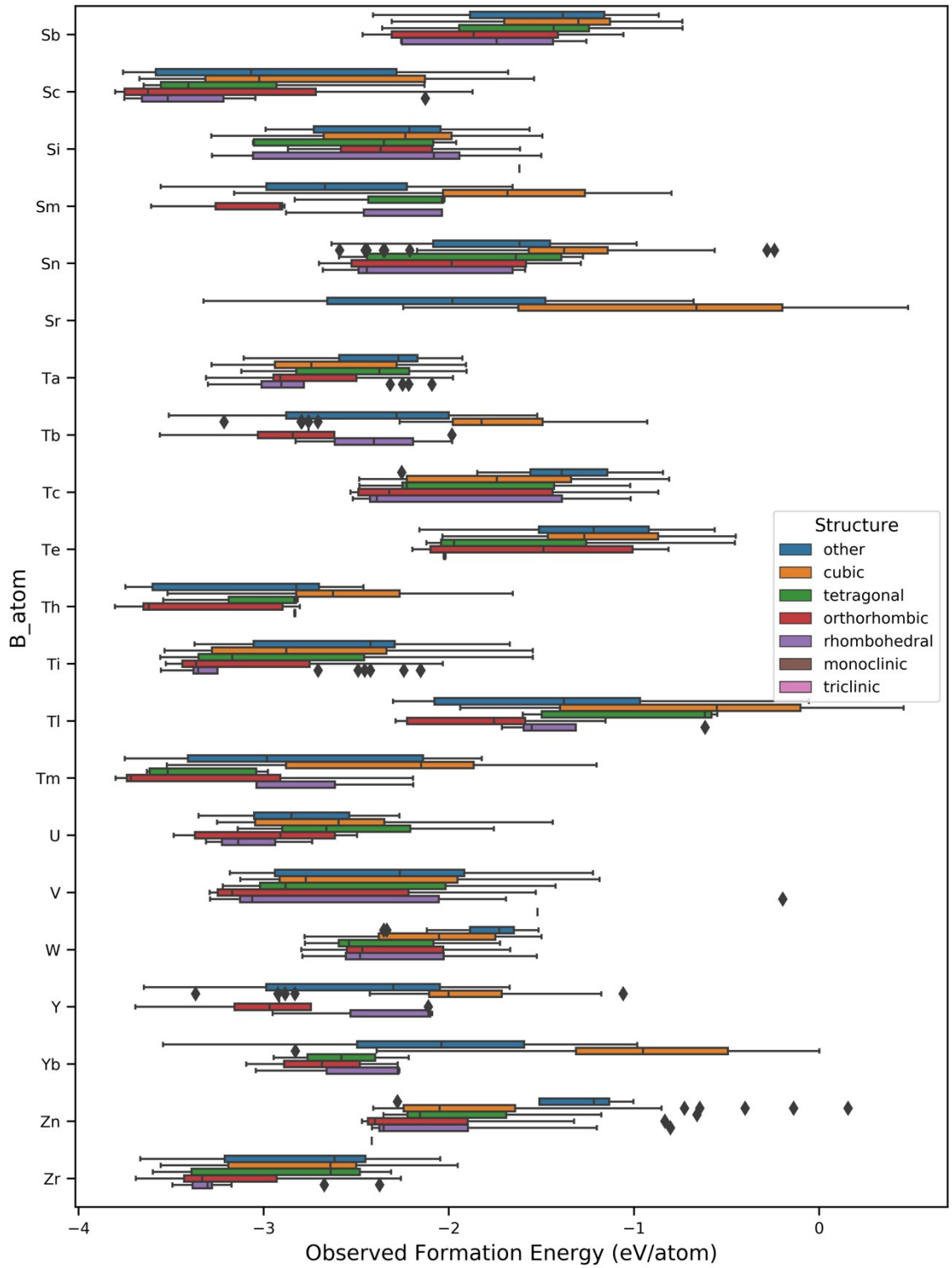

Figure S. 23. Ranges of DFT predicted formation energies vs. the atom placed in the B position of the initial structure for the cubic and non-cubic perovskites dataset from OQMD. The energies are categorized by their lattice symmetry if their final structure has perovskite-like coordination (8,10, or 12 for A site and 6 for B site), or "other" if the structure has different coordination. This plot contains elements in the S-Z range.

Reproducibility of the cubic and noncubic dataset results

To test the robustness of the results for the cubic and noncubic dataset, 20 different test/train splits were generated following the same rules of keeping all structural polymorphs of the same chemical composition in the same side of the split. Of the 20 new networks, 5 displayed significant overfitting, creating an average test set r^2 of 0.959 ± 0.053 (standard deviation) and an average training set r^2 of 0.989 ± 0.002. In the 15 networks that did not display overfitting, the average test set r^2 was 0.987 ± 0.003 and the average training set r^2 was 0.989 ± 0.002. These results are actually better than those reported in the main paper, indicating that our original network was somewhat overfit and even higher performance is fairly reliably able to be obtained from these networks.

References

(1) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv Prepr. arXiv1603.04467* **2016**.

(2) Williams, L.; Mukherjee, A.; Rajan, K. Deep Learning Based Prediction of Perovskite Lattice Parameters from Hirshfeld Surface Fingerprints. *J. Phys. Chem. Lett.* **2020**, *11* (17), 7462–7468.

(3) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

(4) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65* (11), 1501–1509.

(5) Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (11).