Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. C, 2021, 9, 11153

Monitoring the role of site chemistry on the formation energy of perovskites *via* deep learning analysis of Hirshfeld surfaces†

Logan Williams, Arpan Mukherjee, Aparajita Dasgupta and Krishna Rajan*

Received 28th April 2021,
Accepted 21st July 2021

DOI: 10.1039/d1tc01972d

rsc.li/materials-c

This paper presents a new approach for predicting thermodynamic properties of perovskites that harnesses deep learning and crystal structure fingerprinting based on Hirshfeld surface analysis. It is demonstrated that convolutional neural network methods capture critical features embedded in two-dimensional Hirshfeld surface fingerprints that enable a quantitative assessment of the formation energy of perovskites. Building on our recent work on lattice parameter prediction from Hirshfeld surface calculations, we show how transfer learning can be used to speed up the training of the neural network, allowing multiple properties to be trained using the same feature extraction layers. We also predict formation energies for various perovskite polymorphs, and our predictions are found to give generally improved performance over a well-established graph network method, but with the methods better suited to different types of datasets. Analysis of the structure types within the dataset reveals the Hirshfeld surface-based method to excel for the less symmetric and similar structures, while the graph network performs better for very symmetric and similar structures.

Introduction

The perovskite crystal structure is a rich family of materials, capable of hosting most elements in the periodic table on either the A site, B site, or both. This flexibility in chemical species on its A and B sites creates a huge number of possible combinations, many of which have technologically interesting properties for use as photovoltaics, or ferroelectrics, *etc.*[1–3] Many, but not all, of these combinations are thermodynamically stable, making the prediction of stable perovskites a valuable tool in the search for new materials with desirable properties.

The rigid sphere model proposed by Goldschmidt[4] in 1926 is still used in perovskite prediction to this day, providing an easy to calculate first order assessment of a composition's ability to form the perovskite structure. Using only the ionic radii of the three elements, two geometric ratios called the tolerance factor and the octahedral factor can be calculated. A region in tolerance factor–octahedral factor space can be defined either empirically or theoretically[5] where perovskite formation is possible. However, the rigid sphere model is an approximation, ionic radii are dependent upon local coordination environment, and chemistry is too complex for Goldschmidt's method to be accurate enough for reliable prediction.

While other, much more complex, methods such as Density Functional Theory (DFT) can provide relatively accurate formation energies at an acceptable computational cost these days, faster methods are always sought to allow for more thorough exploration of potential new materials. Direct experimental measurement is possible, but more time-consuming.[6,7] There have been many studies using a variety of machine learning and transfer learning techniques that focus on either formation energy specifically[8–11] or for general property prediction.[12–15] In a previous study,[16] we have shown that crystal 'fingerprints' based on Hirshfeld surfaces can be used as input to a Convolutional Neural Network (CNN) to accurately predict the equilibrium lattice constant of cubic perovskites. In this work, we show that the same technique can be used to predict formation energies of cubic and non-cubic perovskites. Furthermore, we show that transfer learning can be used to accelerate the CNN training process, as the feature extraction layers can be preserved between different property predictions.

Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY 14260-1660, USA. E-mail: krajan3@buffalo.edu

† Electronic supplementary information (ESI) available: A schematic of the CGCNN network architecture, the details of the neural network architecture, information on the compositions included in both datasets, statistical plots of the datasets and network results, listings of major outliers, statistical data on multiple re-trainings of both the cubic and non-cubic datasets, confusion plots for the full cubic and non-cubic datasets, a qualitative visualization of the nonlinearity of formation energy prediction *via* a t-SNE flattening of the neural networks latent space at various layer depths, formation energy distributions based on relaxed structure symmetry and starting A and B site atoms, a datafile containing the data curation information for the cubic perovskite dataset, and a datafile containing the data and structural classification for the cubic and non-cubic perovskite dataset are available. See DOI: 10.1039/d1tc01972d

Transfer learning is a concept within machine learning where the knowledge gained from one model (the source task) can be used or transferred towards learning another model (the target task). It can be used to speed the training of a different property of a given dataset, or allow more reliable training upon a small dataset that shares some correlation overlap with an available, larger dataset.[17,18] In deep learning models, the number of parameters to be optimized can easily be in the order of millions, requiring huge computational times for model training. Transfer learning can reduce the number of parameters that need to be optimized by keeping some constant from the earlier model and/or speed convergence on the optimal solution by initializing parameters close to the optimal values. Deep learning models usually consist of two parts: feature extraction and task-specific fully connected layers. The feature extractors in a CNN model comprise of the convolution layers and are responsible for pre-processing the image, identifying high-level geometric features, and spatial correlations. Thus, transfer learning can be used to borrow the 'knowledge' from the feature extraction layers of a different network while fine-tuning the fully connected layers suited to a given classification or regression task. However, such transfer learning requires similarity between the datasets of the pre-trained network and the new network. Without such similarity, transfer learning can give poor results.[19,20]

In this study, our CNN is based on image recognition techniques applied to crystal 'fingerprints' created from the Hirshfeld Surface Analysis of the crystal structure.[21] Hirshfeld Surface Analysis has seen extensive use in the field of molecular crystals,[21–25] as it is an efficient way to analyze molecular packing and shape, close contact points, and inter-molecular interactions. In this study, as in our previous work,[16] we modify the traditional Hirshfeld surface fingerprint plot when moving from molecular crystals to inorganic crystals: instead of defining a single Hirshfeld surface about the entire molecule or unit cell, we define one around each unique atomic site in the unit cell, as shown in Fig. 1, taking inspiration from the field of Atom In Molecule (AIM) research using Hirshfeld surfaces.[26–29] This creates a fingerprinting method that equally characterizes all atomic sites within a structure and reflects crystal structure, stoichiometry, defects, and lattice distortions.

By characterizing each unique atomic site separately and then combining the data from each atomic site together, the atomic Hirshfeld surfaces fingerprint plot characterizes a crystal structure in a way that is invariant with respect to translations, rotations, or other arbitrary choices in unit cell selection, which is vital for robust learning and prediction upon varied crystal structures. As shown in Fig. 1 and 2, our neural networks are built from a 3-dimensional description of each atom and its local environment, measuring the radial size (d_i) and distance to the nearest external atom (d_e) for each point upon the atom's Hirshfeld surface.

The organization of the remainder of the paper is as follows. The methodological details for all calculations and data curation are listed in the Methodology section. We then report formation energy results using transfer learning and the cubic perovskites dataset used in our previous paper,[16] analyzing performance and outlier trends. Next, we report and analyze results for a new network trained upon a dataset of cubic and non-cubic perovskites. As this dataset consists of DFT calculations that allowed for structural relaxation during calculation, we then perform a structural analysis of the dataset, recategorizing them from the four perovskite structure prototypes they started in to the actual structure at the end of each calculation. We use this categorization to analyze the results of our tested methods upon different types of structural data. Finally, we include some descriptive statistics of our cubic and non-cubic dataset, analyzing common perovskite forming elements and

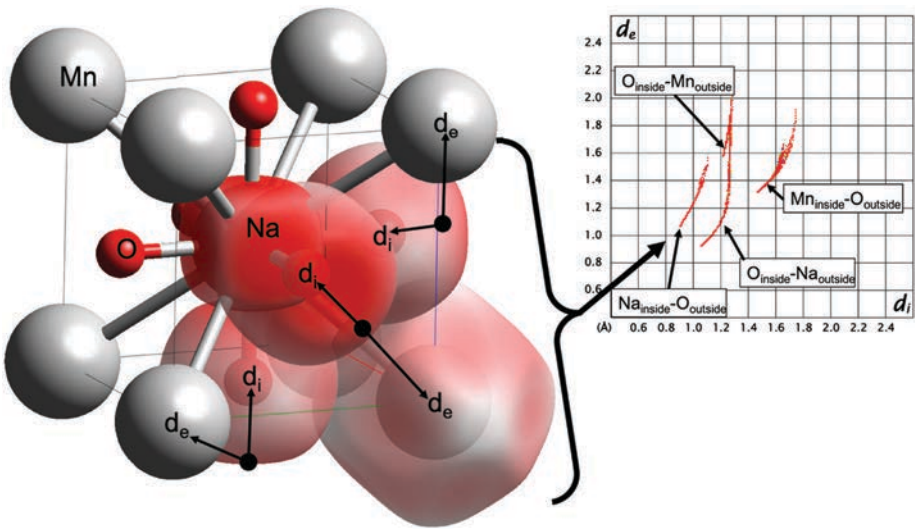

Fig. 1 Schematic showing the creation of the fingerprint plot from the atomic Hirshfeld surfaces. For each point on each Hirshfeld surface, the distance to the nearest atom inside the surface, d_i, and the distance to the nearest atom outside the surface, d_e, are calculated. These (d_i, d_e) pairs are then binned into a two-dimensional histogram to form the fingerprint plot.

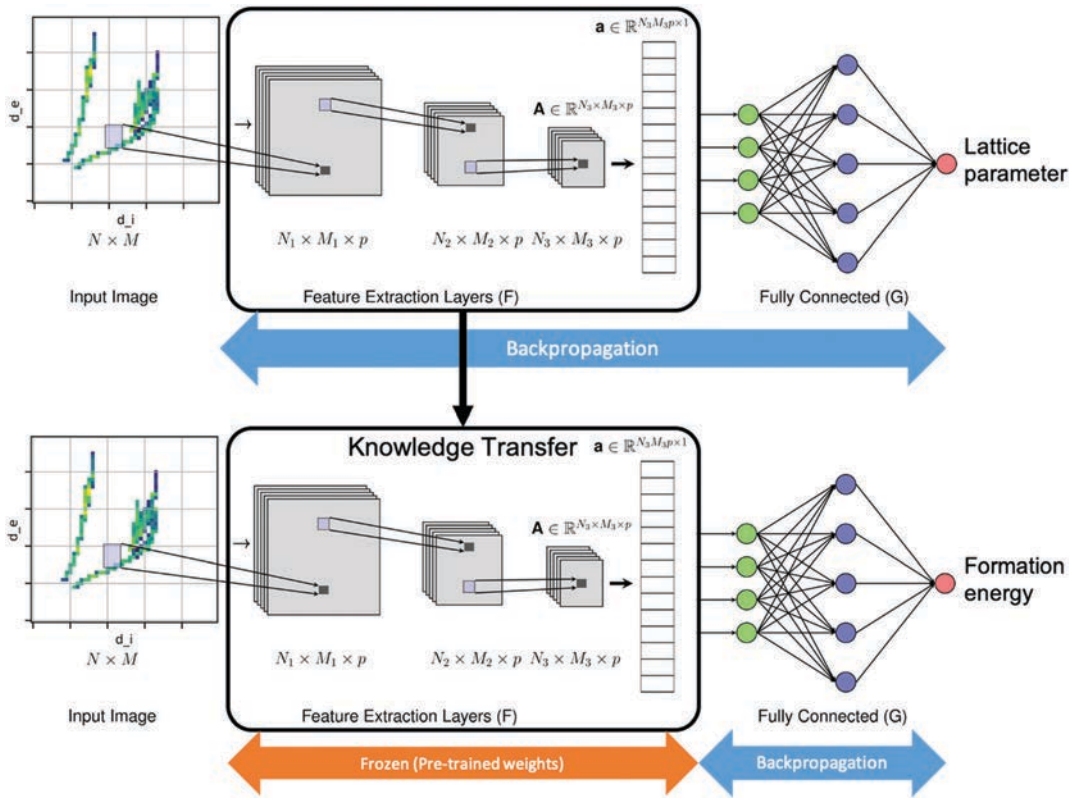

Fig. 2 The feed-forward propagation of our neural networks consists of feature extraction layers (F) and fully connected layers (G). The feature extraction layers reduce the N by M crystal fingerprint into a flattened set of feature maps that are interpretable by the fully connected layers. The fully connected layers perform the final property prediction. In this paper, the feature extraction layers for our cubic dataset network are taken from the prior network[16] used to predict lattice parameter and frozen at those values. The fully connected layers are the only ones whose weights are updated, significantly reducing training time.

the dataset through the lens of the octahedral and tolerance factors.

Methodology

To make our modified fingerprint plots, for each unique atomic site in the structure, the Hirshfeld surface is calculated, using the open-source Tonto[30] software package, by placing the pre-calculated[31] spherically averaged electron density of the neutral, isolated element around each atom and locating the 3D surface where 50% of the electron density comes from the selected atomic site. For each point upon these Hirshfeld surfaces, the distance to the nearest atom inside, d_i, and the nearest atom outside, d_e, are measured, as shown in Fig. 1. These (d_i, d_e) pairs are then collected from all atomic Hirshfeld surfaces in the crystal structure and binned into a 2-Dimensional (2D) histogram called the fingerprint plot, which has the benefit of being rotationally-invariant with respect to the crystal system orientation. This 'fingerprint' plot of the atomic Hirshfeld surfaces characterizes the nature and environment of each atom within a crystal structure. The resultant image is a 2D tensor similar to a depth map or single-channel image, making it suitable for analysis by machine learning methods designed for image processing, such as CNNs.[32,33]

Our CNN architecture consists of feature extraction layers followed by fully connected layers used for property prediction, as shown in Fig. 2. In this particular case of transfer learning with the cubic perovskite dataset, the source task was the task of predicting the lattice parameter, while the target task is predicting the formation energy. The feature extraction layers are the same as in our previous work, whose network produced highly accurate predictions.[16] The fully connected layers have been slightly modified. The complete architecture is described in the ESI.† The feature extraction layers serve to decode chemistry and geometric relationships between the atoms in the crystal structure. It produces a collection of reduced-order matrices, each accentuating different sub-domain inside the fingerprint plot that differ geometrically than the complementary region. For example, if we use p kernels or filters in the last convolution layer, it then produces p distinct such images or feature maps (see Fig. 2). We are using single-source homogeneous inductive transfer learning where the source and the target input domain are the same, while the target tasks are different property parameters. We can assume an implicit but nonlinear correlation between the formation energy and the lattice parameter.[34,35] We implement the concept of transfer learning by assuming that the features identified by our source

CNN model will be sufficient to predict the formation energy with the same level of accuracy. The feature maps produced in the source CNN model and subsequently their vectorized union, also called a flattened layer, can thus act as descriptors for our current target model. Therefore, the feature extraction layers from our previous work[16] can be used unaltered in our current model for the energy of formation. This allows for a drastic reduction in the number of parameters that require optimization, down to only those in the fully connected layers used for property prediction. Additionally, by restricting the descriptor space to the pre-defined feature maps obtained from the source model, we narrow down the hypothesis space for the fully connected layers for the target model.

Our cubic perovskites dataset is the same as from our previous publication, with one additional restriction.[16] To generate our dataset, the 5321 ABO3 cubic perovskites from the Open Quantum Materials Database, OQMD,[35,36] were initially selected. These include all elements up to $Z = 94$ except for the noble gases, the halogens, H, C, N, O, P, S, Se, and Po. The exact compositions included in the final dataset are shown in the ESI.† The dataset was reduced to 5250 structures by removing cases fitting either of two criteria. First, if the relaxed lattice parameter was greater than 5 Å or more than 2% larger than the (generous) unrelaxed lattice parameter used by OQMD, they were removed as these are likely to be unstable structures. Second, if the relaxed lattice parameter was equal to the unrelaxed lattice parameter, they were removed as these may be unnoticed failed calculations. Finally, the dataset was trimmed down to 5206 compounds by discarding structures without a converged non-relaxation energy calculation. The Hirshfeld surface of each atom in every structure was then calculated using the Tonto software package, an open-source tool for Hirshfeld surface and other analyses.[30] For both Hirshfeld surface calculation and comparison CGCNN initialization, initial lattice constants were assigned to each structure as a random value in the range of 3.5–5.5 Å, the range of most oxide cubic perovskites. To achieve smooth fingerprint plots, the atomic Hirshfeld surfaces were interpolated using 10 points between each vertex, and then the fingerprint plot for each structure was created by binning the (d_i, d_e) pairs of all interpolated surfaces in the structure into 50×50 bin histograms (bin size = 0.04 Å) ranging 0.76–2.8 Å for both di and de (coarser than shown in Fig. 1). The neural network was trained using the open-source library Keras with Tensorflow v1.8.0 backend.[37] The exact details of the architecture are provided in the ESI.†

For the cubic and non-cubic perovskites dataset, the 5911 tetragonal, orthorhombic, and rhombohedral ABO3 perovskites from OQMD were combined with the 2501 cubic perovskites from the cubic perovskite dataset that had the same composition as one of the non-cubic structures. This was done to create a dataset that had a roughly normal distribution of the target property. Many higher formation energy cubic perovskites did not get their non-cubic variants calculated by OQMD, which skews the formation energy distribution of the dataset unevenly if all the cubic perovskites are included. The compositions included/excluded and the distribution of formation energies are shown in the ESI.† The lattice parameters of all the structures were all kept at their DFT relaxed values. Hirshfeld surface calculations were performed the same as for the cubic set. A new neural network was trained from scratch for the new dataset, with the same architecture but without transfer learning. The ~70/30 train/test split was done by chemical composition, with all structural polymorphs possessing the same chemistry being assigned to either the test or training set as a group, to prevent artificially high results from too much similarity between test and training materials. This resulted in 5869 and 2543 structures in the training and testing sets, respectively. Pymatgen[38] was used to calculate atomic site coordination numbers when classifying the relaxed structures as detailed in the discussion section and to calculate expected valence states for compositions in the tolerance and octahedral factor calculations.

Results and discussion

The results for our CNN model based on Hirshfeld surfaces fingerprints are shown in Table 1 and Fig. 3. For comparison, a model using the Crystal Graph Convolutional Neural Network (CGCNN)[15] was also trained using the same train/test split and is shown alongside it. The CGCNN method uses a graph network representation of the crystal structure, with elemental data for each atomic node and a function of the atom–atom distance connecting neighboring atoms. Our model achieves better performance than the CGCNN model on both the test and training sets and is highly accurate over a wide chemical space with only a few outliers with large magnitudes.

Despite the overall high accuracy of the model's fit, there are some notable outliers in both the testing and training sets. Of the 11 outliers with greater than 0.5 eV per atom magnitude residual in the combined testing and training sets for the Hirshfeld surfaces plus CNN model, 8 possess either Ba, K, Rb, or Sr in the B site, as shown in Table S2 (ESI†). All of these elements are group 1 or 2 elements with very similar atomic radii (~215–248 pm). As the Hirshfeld surface is built upon spherically-averaged neutral atom electron densities, it appears that these similarly-sized s-block elements present the most challenge to the neural network, and it is likely that the network is sometimes misidentifying one of these elements as another. The network does make accurate predictions for most of these compounds, as can be seen in Fig. S8 (ESI†). Also, it does not

Table 1 The R^2 value, Mean Absolute Error (MAE), and Mean Squared Error (MSE) from the CGCNN[15] model and our Hirshfeld surface fingerprint + CNN model for both training and testing set prediction for the DFT-calculated system energies of the 5206 cubic oxide perovskites with converged calculations. In both training and test sets, our model based on Hirshfeld surface fingerprint outperforms the CGCNN model

Method	Split	R^2	MSE	MAE
CGCNN	Training	0.9824	0.01753	0.0968
CGCNN	**Test**	**0.9780**	**0.02248**	**0.1129**
Hirshfeld surfaces + CNN	Training	0.9926	0.00733	0.0502
Hirshfeld surfaces + CNN	**Test**	**0.9899**	**0.01030**	**0.0567**

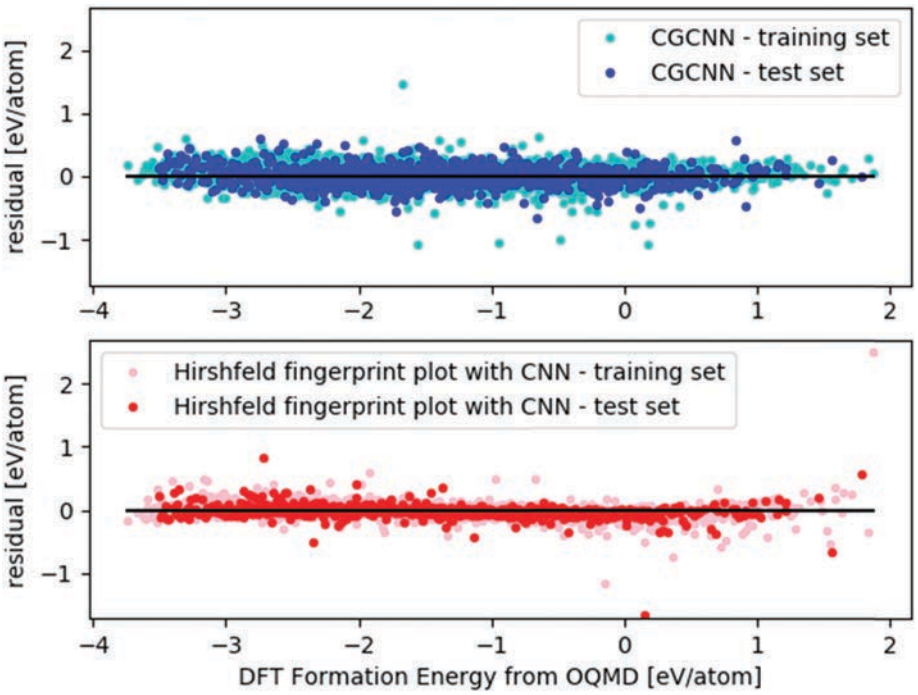

Fig. 3 Training and test set residuals for the prediction of cubic perovskites DFT relaxed formation energies from OQMD using input structures using randomized lattice parameters using (top) the CGCNN technique[15] or (bottom) the atomic Hirshfeld fingerprint plot with a convolutional neural network. The CNN used the same feature extraction layers as the one previously used to predict relaxed lattice parameters.[16] The fingerprint plot plus CNN method shows superior predictive performance compared to the CGCNN method.

seem impossible for a network to use these feature extraction layers to tell these atoms apart, as the previous network for lattice parameter[16] did not show large errors on those elements. The previous network had a bias towards estimating towards the mean that caused its largest residuals to be for small elements such as B, Be, and Li, shown in Table S3 (ESI†). Subtracting out that bias, there is no clear trend to the outliers in the residuals, as shown in Table S4 (ESI†). For comparison, the CGCNN had 24 outliers of greater than 0.5 eV per atom magnitude residual (over twice as many), shown in Table S5 (ESI†). Eight of those contained either Li, Be, or B in the B site. Twenty-one contained a row 6 or row 7 element, indicating that the CGCNN method had the largest errors on high and low atomic number elements. Despite these outliers, the Mean Absolute Error (MAE) for our model is still <0.06 eV per atom for both the testing and training sets.

The ability to accurately predict the DFT formation energy of a cubic perovskite based on its unrelaxed structure should not be too surprising. The features which control both, and which are identified by the shared feature extraction layers, are the same. Atomic species, which determines the number of electrons and contribute to the potentials they exist within, and the local environment of each atom, control both the lattice constant and the formation energy. With a structure near the relaxed lattice constant/energy minima, the formation energy *vs.* the system volume shares a roughly parabolic relationship. In effect, predicting the lattice constant is locating the x position (volume) of the energy minima, while predicting the formation energy is predicting the magnitude, or y-value, of that same energy minima.

The results for our CNN model built on Hirshfeld surface fingerprints and a comparison model using the CGCNN[15] model upon the cubic and non-cubic perovskites dataset are shown in Table 2 and Fig. 4. Our model achieves better performance upon the training set than the CGCNN model and comparable, but slightly worse results on the test set. As can be seen in Fig. 4, the residuals for our model have a small but notable bias of estimating towards the mean. Thus, a single step gradient boosting was applied to the model by fitting a simple linear regression to the training set predictions and target values, then applying it to all the model's predictions. This acts to flatten the linear bias seen

Table 2 The R^2 value, Mean Absolute Error (MAE), and Mean Squared Error (MSE) from the CGCNN[15] model, our Hirshfeld surface fingerprint + CNN model, and our model plus a simple linear residual boosting scheme for both training and testing set prediction for the DFT-calculated system energies of the 8412 cubic and non-cubic oxide perovskites with converged calculations. In both training and test sets, our model based on Hirshfeld surface fingerprint outperforms the CGCNN model after the boosting technique is applied. The CGCNN method results are negligibly changed by the boosting method

Method	Split	R^2	MSE	MAE
CGCNN	Training	0.9781	0.01711	0.0952
CGCNN	**Test**	**0.9648**	**0.02589**	**0.1131**
Hirshfeld surfaces + CNN	Training	0.9878	0.00951	0.0719
Hirshfeld surfaces + CNN	**Test**	**0.9636**	**0.02674**	**0.1029**
HFS + CNN + boosting	Training	0.9923	0.00602	0.0565
HFS + CNN + boosting	**Test**	**0.9663**	**0.02476**	**0.0719**

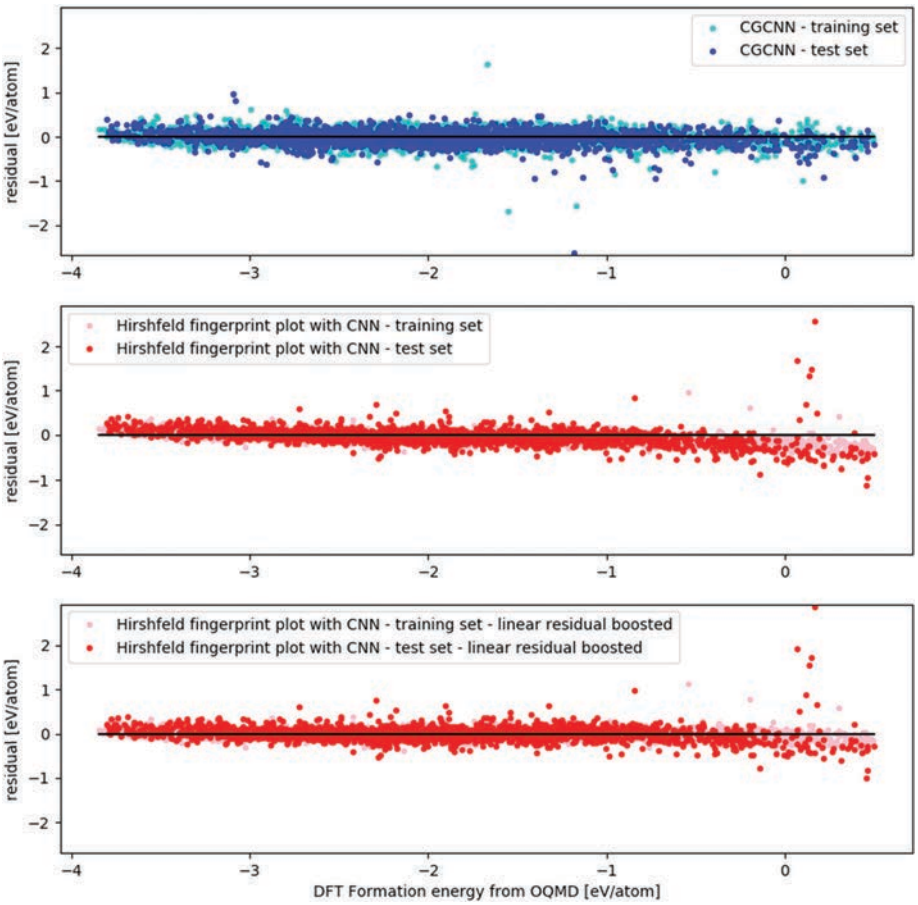

Fig. 4 Training and test set residuals for the prediction of cubic and non-cubic perovskites DFT relaxed system energies from OQMD using (top) the CGCNN technique[15] or (middle) the atomic Hirshfeld fingerprint plot with a convolutional neural network, or (bottom) the middle network with a simple linear residual boost applied to reduce the systematic bias towards estimating towards the mean.

in Fig. 4. With the linear boost applied, our model displays improved predictive performance over the CGCNN method on the test set as well. Application of the same boosting technique to the CGCNN produces a minuscule difference, as the CGCNN model did not display the same bias towards the mean.

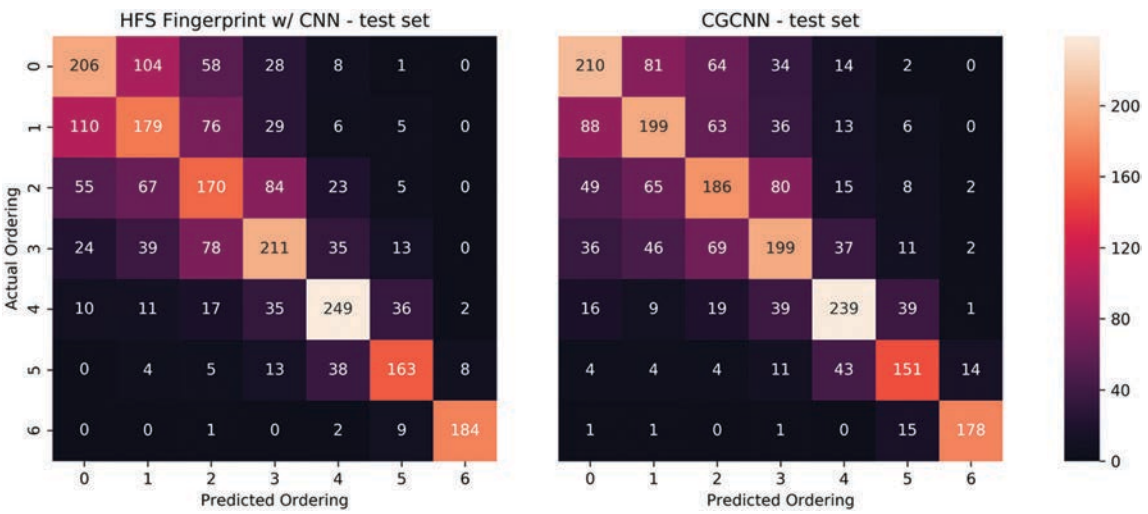

Fig. 5 Confusion plot for the relative ordering of same composition phases in the test set of the cubic and non-cubic perovskite dataset as produced by (left) our model built using a CNN and Hirshfeld surface fingerprint plots and (right) the CGCNN[15] model. The linear boosting technique has no effect on relative orderings, as it scaled linearly based on the predicted value from the CNN.

The outliers in the predictions for this dataset are listed in the ESI† and share some trends with the former. The 'spike' of 7 large positive outliers around formation energies of ~0.1 eV per atom is caused by several polymorphs of $KRbO_3$ and K_2O_3, and 5 more (12 total) of the 23 outliers with residual magnitude ≥0.5 eV per atom contain either K, Rb, or Ba (none contain Sr). Additionally, all but 2 of the large outliers were for cubic or rhombohedral structures. For the CGCNN model, row 6 and 7 elements make up 29 of the 40 outliers with residual magnitude ≥0.5 eV per atom. Additionally, half of the large outliers were for tetragonal structures, and only 2 were orthogonal structures.

When predicting formation energies of multiple phases of the same composition, the relative energy ordering is important for the determination of phase stability or metastability. Fig. 5 shows confusion plots for the predicted *vs.* actual ordering of the test set for the CNN and CGCNN[15] models, respectively. The two methods predict nearly the same number of ground states correctly. The CGCNN method predicts more 2nd and 3rd most stable states correctly, and the CNN predicts more higher-energy states correctly. However, the CNN method is notably closer in its incorrect orderings, with fewer predictions far from the correct placement. The confusion plots for the full datasets are shown in the ESI,† and the CNN significantly outperforms the CGCNN method when including the training set due to its far better fit upon the training data.

The non-cubic perovskite calculations performed by OQMD do not enforce a final structure or symmetry upon the calculation. While all structures began DFT structural relaxation in one of four perovskite structure prototypes (cubic – based on $SrFeO_3$'s structure, tetragonal – based on $PbTiO_3$, rhombohedral – based on $NdAlO_3$, or orthorhombic – based on $GdFeO_3$), many of the compounds underwent significant structural rearrangement during relaxation. To classify the relaxed structures, we took any structure with a coordination number of 8 through 12 for one of the cations and a coordination number of 6 for the other cation species to be a perovskite class structure. Using this criterion, ~46% of the orthorhombic structure, ~55% of the tetragonal structures, and ~58% of the rhombohedral structures relaxed into a non-perovskite form. All 2501 cubic perovskites remained cubic perovskites due to enforced cubic symmetry during their calculations. We then categorized the perovskite structures by their lattice symmetry. Of the 1947 orthorhombic structures ~43% remained orthorhombic while ~5% became tetragonal and ~5% became cubic. Of the 1852 tetragonal structures ~31% remained tetragonal while the remaining ~14% became cubic. Of the 2112 rhombohedral structures ~34% remained rhombohedral while the remaining ~9% became cubic.

This categorization of the structures within the dataset allows us to examine trends in the performance of the two methods tested based on structural distortion and symmetry. As shown in Table 3, the CGCNN method struggled with the heavily distorted structures (the triclinic and non-perovskite) in both the training and testing sets, as well as had significantly worse results upon the tetragonal test set structures than the training set.

Table 3 The R^2 values for the CGCNN[15] and Hirshfeld surfaces + CNN models upon separating the dataset by the symmetry and type (perovskite or non-perovskite) of the relaxed crystal structures from OQMD. The first five categories (cubic, tetragonal, *etc.*) are all referring to perovskites of that symmetry, while the final is for all non-perovskite structures of any symmetry. The CGCNN method had its worst overall performance on the most distorted structures, triclinic and non-perovskite, and the worst loss of performance between the training and test sets upon the tetragonal structures. The HFS + CNN method had the most difficulty upon the most symmetric structures, cubic and rhombohedral

Structure type	Test/train split	# of structures	CGCNN R^2	HFS + CNN R^2	HFS + CNN w/ boosting R^2
Cubic	Train	2126	0.9788	0.9840	0.9905
	Test	922	0.9753	0.9442	0.9481
Tetragonal	Train	471	0.9841	0.9882	0.9930
	Test	210	0.9264	0.9722	0.9757
Orthorhombic	Train	591	0.9819	0.9902	0.9932
	Test	244	0.9711	0.9781	0.9811
Rhombohedral	Train	497	0.9826	0.9881	0.9895
	Test	218	0.9714	0.9470	0.9349
Triclinic	Train	6	0.9400	0.9852	0.9920
	Test	1	—	—	—
Non-perovskite	Train	2177	0.9685	0.9892	0.9931
	Test	947	0.9490	0.9790	0.9835

Contrastingly, the Hirshfeld surfaces fingerprint plot with CNN method struggled the most upon the most symmetrical structures, the cubic and rhombohedral perovskites. These trends are sensible when considering the features used in the respective methods. The CGCNN method uses a graph network and bond distances between atoms, but does not include full 3D characterization of atomic environment. The Hirshfeld surfaces with CNN method does include 3D characterization of the atomic environment, but does not encode atomic composition as directly as the CGCNN method. The highly symmetric crystal systems have the most similar fingerprint plots with the fewest differences for the convolutional neural network to work with, making the Hirshfeld surfaces relatively better suited for learning the properties of the more distorted structures of the dataset and the CGCNN method relatively better suited for the most symmetric structures of the dataset.

Using the structure categorizations described above, we perform some descriptive statistics to identify trends in favored structure types based on chemical composition. Some atoms in the A or B site largely favor non-perovskite structures for all chemistries included in the dataset. Structures starting with Li in the A site relax into a non-perovskite structure as the lowest formation energy structure for over 50% of the studied compositions. Pa and Si do the same for structures with them starting in the B site. For structures started with them in the A site Ac, Ba, Bi, Ca, Cd, Ce, Cs, Dy, Eu, Gd, K, La, Na, Nd, Np, Pb, Pm, Pr, Pu, Rb, Sm, Sr, Tb, Th, Y, and Yb formed a perovskite as the lowest formation energy structure over 50% of the studied compositions. For structures started with them in the B site Al, Au, Co, Cr, Cu, Fe, Ga, Hf, Ir, Lu, Mg, Mn, Mo, Nb, Ni, Os, Pd, Pt, Re, Rh, Ru, Sc, Ta, Tc, Ti, V, W, Nz, and Zr formed a perovskite as the lowest formation energy structure over 50% of the studied compositions. Fig. 6 shows a categorical breakdown

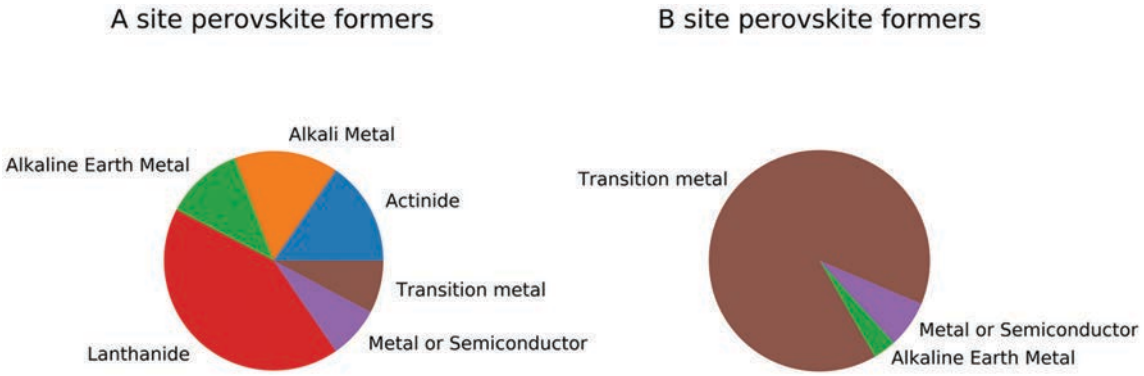

Fig. 6 Categorical breakdown of the elements found to form a perovskite structure as the lowest formation energy structure for over half of their compositions when placed into the (left) A site or (right) B site.

of the strongly perovskite-forming elements. Compared to a study[5] predicting the perovskite formability of ternary and quaternary compositions, we identify similar trends, with transition metals in the B site and lanthanides in the A site favoring perovskite formation. However, as we are analyzing the reliability of elements on a given site towards forming perovskite structures rather than the total breakdown of the dataset, we notably do not find any actinides or lanthanides to reliably form perovskites in the B site, while compounds with a lanthanide or actinide composed roughly 1/3 of their[5] predicted oxide perovskites. The ESI† contains additional plots of formation energy broken down by initial A and B atom species as well as the final, relaxed structure type.

Finally, we analyze the dataset for the tendency to form or not form a perovskite structure in the lens of geometrical analysis using the octahedral and tolerance factors. When plotting the perovskites according to their tolerance $\left[t = (r_A + r_X)/\left(\sqrt{2}(r_B + r_X)\right)\right]$ and octahedral $[\mu = r_B/r_X]$ factors (see Fig. 7), compositions where the most stable structure calculated was a perovskite mostly adhere to the region defined by the hard spheres and no-rattling rules.[5] These rules are derived by assuming each atom is a rigid sphere with radius equal to their ionic radius, and determining the geometric limits of atomic size differences where atoms can be placed into any distorted perovskite structure without leaving atoms in any voids too large for them to be prevented from 'rattling' by their neighbors. Using the notation of Filip and Giustino,[5] there are four types of limits imposed by the no-rattling rules. The Stretch Limit (SL) is defined where the A atom is so large as to touch all 12 anions around it in a cubic perovskite. In this situation, the tolerance factor is always 1. The Octahedral Limit (OL) is where the anions in the same BX_6 octahedron are in contact with one another. In this situation, μ always equals $\sqrt{2} - 1$. The third type of limit are derived by considering tilting of the BX_6 octahedra and a resultant displacement of the A atom. This produces two limiting cases, (TL1) $t = (1.366 + 0.422\mu)/\sqrt{2}\mu + 1)$ for $\mu < 0.8$ and (TL2) $t = (1.125 + 0.732\mu)/(\sqrt{2}\mu + 1)$ for $\mu > 0.8$.

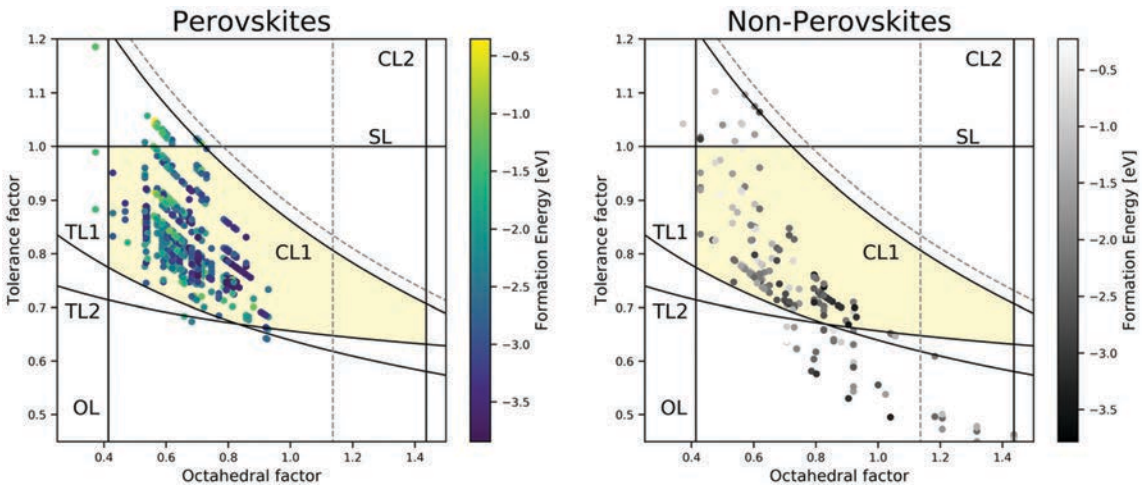

Fig. 7 The perovskite structures from the cubic and non-cubic perovskites dataset plotted according to their octahedral and tolerance factors, with color corresponding to formation energy. (left) Structures where a perovskite was the most stable calculated structure for the composition. (right) Structures where a non-perovskite structure was calculated to have the lowest formation energy for the composition. The lines and curves labeled OL, SL, CL1, CL2, TL1, and TL2 correspond to the no-rattling hard spheres model limits,[5] with the dashed gray curves being the chemical limits from the reference and the black curves recalculated for our dataset to account for the different range of elements included.

Finally, the last limits are from considering the largest and smallest ionic radii in the elements within the dataset, rather than from the geometry. The largest cation and smallest anion in the dataset create the first limit, CL1, as $t = (r_A/r_X + 1)/(\sqrt{2}(\mu + 1))$, with $r_A = r_{Cs} = 1.81$ Å and $r_X = r_O = 1.26$ Å for our dataset. The largest cation and smallest anion also define the largest octahedral factor (CL2) for our dataset as $\mu = r_B/r_X = r_{Cs}/r_O = 1.81/1.26 = \sim 1.44$. Many of the compounds in the dataset had a lower tolerance factor than allowed by the Tilt Limits (TL1 and TL2), but only a few very near the limit formed stable perovskites, corroborating Filip and Giustino's findings on the importance of the Tilt Limits.[5] However, many compounds containing Tl, Rb, Cs, or Ba on the A site exceeded the tolerance factor predicted by the Stretch Limit (SL), beyond which the hard sphere model would predict rattling of the B site. Finally, three perovskites with Mn on the B site (orthorhombic $TlMnO_3$, $NaMnO_3$, and $LiMnO_3$) had lower octahedral factors than predicted by the Octahedral Limit (OL). These exceptions are further evidence that the hard sphere model, while a reasonable approximation, does not strictly bound the potential perovskite formation region. However, there were also many compositions located within the predicted stability region where the lowest energy structure calculated was a non-perovskite structure (right plot of Fig. 7), showing that a composition's presence within the region predicted by the hard sphere model design rules does not guarantee the perovskite structure to be the most favorable structure.

Conclusion

In this study, we have further demonstrated that crystal fingerprinting based on atomic Hirshfeld surfaces is a powerful tool in combination with image processing techniques and presents a highly-effective method of using deep learning for the prediction of multiple inorganic crystal properties.[16] Combining atomic and three-dimensional geometric data, fingerprints based on atomic Hirshfeld surfaces are data-rich descriptors for machine learning. We have shown that transfer learning can be utilized to speed the training of models on new properties by taking advantage of the implicit relationships between material properties. We show generally improved predictive performance over the CGCNN[15] method on the studied dataset, but note that our method performs relatively better on the more distorted structures and relatively worse for highly symmetric subsets of the data compared to the CGCNN method. The results and transfer learning process described in this paper establish this method in the toolbox for machine learning of material properties, as the technique is easily generalizable to other crystal systems and material properties and is especially suited to datasets with more complexity in crystal structures. We also analyze perovskite formation trends, identifying transition metals on the B site and lanthanides and actinides on the A site as having a high likelihood to form perovskites. We also note that while the tilt limits in the hard spheres model of perovskites are quite effective at determining which materials will not form perovskites, there still exist numerous compositions within the expected stability window that prefer a non-perovskite form.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

The authors acknowledge the support from NSF Award# 1640867 – DIBBs: EI: Data Laboratory for Materials Engineering and the Collaboratory for a Regenerative Economy (CoRE center) in the Dept of Materials Design and Innovation – University at Buffalo.

References

1 S. D. Stranks and H. J. Snaith, Metal-Halide Perovskites for Photovoltaic and Light-Emitting Devices, *Nat. Nanotechnol.*, 2015, **10**(5), 391–402.
2 K. J. Choi, Enhancement of Ferroelectricity in Strained BaTiO3 Thin Films, *Science*, 2004, **306**(5698), 1005–1009.
3 M. V. Kovalenko, L. Protesescu and M. I. Bodnarchuk, Properties and Potential Optoelectronic Applications of Lead Halide Perovskite Nanocrystals, *Science*, 2017, **358**(6364), 745–750.
4 V. M. Goldschmidt, Die Gesetze Der Krystallochemie, *Naturwissenschaften*, 1926, **14**(21), 477–485.
5 M. R. Filip and F. Giustino, The Geometric Blueprint of Perovskites, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**(21), 5397–5402.
6 H. Zhang, N. Li, K. Li and D. Xue, Structural Stability and Formability of AB O 3 -Type Perovskite Compounds, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2007, **63**(6), 812–818.
7 G. P. Nagabhushana, R. Shivaramaiah and A. Navrotsky, Direct Calorimetric Verification of Thermodynamic Instability of Lead Halide Hybrid Perovskites, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**(28), 7717–7721.
8 D. Jha, L. Ward, A. Paul, W. Liao, A. Choudhary, C. Wolverton and A. Agrawal, ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition, *Sci. Rep.*, 2018, **8**(1), 17593.
9 X. Li, Y. Dan, R. Dong, Z. Cao, C. Niu, Y. Song, S. Li and J. Hu, Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning, *Appl. Sci.*, 2019, **9**(24), 5510.
10 W. Li, R. Jacobs and D. Morgan, Predicting the Thermodynamic Stability of Perovskite Oxides Using Machine Learning Models, *Comput. Mater. Sci.*, 2018, **150**, 454–463.
11 W. Ye, C. Chen, Z. Wang, I.-H. Chu and S. P. Ong, Deep Neural Networks for Accurate Predictions of Crystal Stability, *Nat. Commun.*, 2018, **9**(1), 3800.
12 D. Jha, K. Choudhary, F. Tavazza, W. Liao, A. Choudhary, C. Campbell and A. Agrawal, Enhancing Materials Property Prediction by Leveraging Computational and Experimental

Data Using Deep Transfer Learning, *Nat. Commun.*, 2019, **10**(1), 5316.
13 H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, Predicting Materials Properties with Little Data Using Shotgun Transfer Learning, *ACS Cent. Sci.*, 2019, **5**(10), 1717–1730.
14 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials, *npj Comput. Mater.*, 2016, **2**(1), 16028.
15 T. Xie and J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, *Phys. Rev. Lett.*, 2018, **120**(14), 145301.
16 L. Williams, A. Mukherjee and K. Rajan, Deep Learning Based Prediction of Perovskite Lattice Parameters from Hirshfeld Surface Fingerprints. *J. Phys*, *Chem. Lett.*, 2020, **11**(17), 7462–7468.
17 M. Oquab, L. Bottou, I. Laptev and J. Sivic, Learning and Transferring Mid-Level Image Representations Using Convolutional Neural Networks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.
18 J. Yosinski, J. Clune, Y. Bengio and H. Lipson, How Transferable Are Features in Deep Neural Networks?, *Advances in neural information processing systems*, 2014, pp. 3320–3328.
19 M. Raghu, C. Zhang, J. Kleinberg and S. Bengio, Transfusion: Understanding Transfer Learning for Medical Imaging, *Advances in neural information processing systems*, 2019, pp. 3347–3357.
20 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *J. Mach. Learn. Res.*, 2020, **21**(140), 1–67.
21 M. A. Spackman and D. Jayatilaka, Hirshfeld Surface Analysis, *CrystEngComm*, 2009, **11**(1), 19–32.
22 M. M. Jotani, S. M. Lee, K. M. Lo and E. R. T. Tiekink, 1-Chloro-4-[2-(4-Chlorophenyl)Ethyl]Benzene and Its Bromo Analogue: Crystal Structure, Hirshfeld Surface Analysis and Computational Chemistry, *Acta Crystallogr., Sect. E: Crystallogr. Commun.*, 2019, **75**(5), 624–631.
23 C. Jelsch and Y. Bibila Mayaya Bisseyou, Atom Interaction Propensities of Oxygenated Chemical Functions in Crystal Packings, *IUCrJ*, 2017, **4**(2), 158–174.
24 P. R. Spackman, S. P. Thomas and D. Jayatilaka, High Throughput Profiling of Molecular Shapes in Crystals, *Sci. Rep.*, 2016, **6**(1), 22204.
25 S. L. Tan, M. M. Jotani and E. R. T. Tiekink, Utilizing Hirshfeld Surface Calculations, Non-Covalent Interaction (NCI) Plots and the Calculation of Interaction Energies in the Analysis of Molecular Packing, *Acta Crystallogr., Sect. E: Crystallogr. Commun.*, 2019, **75**(3), 308–318.
26 P. Bultinck, C. Van Alsenoy, P. W. Ayers and R. Carbó-Dorca, Critical Analysis and Extension of the Hirshfeld Atoms in Molecules, *J. Chem. Phys.*, 2007, **126**(14), 144111.
27 A. Tkatchenko and M. Scheffler, Accurate Molecular van der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data, *Phys. Rev. Lett.*, 2009, **102**(7), 073005.
28 A. Ambrosetti, A. M. Reilly, R. A. DiStasio and A. Tkatchenko, Long-Range Correlation Energy Calculated from Coupled Atomic Response Functions, *J. Chem. Phys.*, 2014, **140**(18), 18A508.
29 T. Bučko, S. Lebègue, J. G. Ángyán and J. Hafner, Extending the Applicability of the Tkatchenko-Scheffler Dispersion Correction via Iterative Hirshfeld Partitioning, *J. Chem. Phys.*, 2014, **141**(3), 034114.
30 D. Jayatilaka and D. J. Grimwood, Tonto: A Fortran Based Object-Oriented System for Quantum Chemistry and Crystallography, in *Computational Science—ICCS 2003*, ed. P. M. A. Sloot, D. Abramson, A. V. Bogdanov, Y. E. Gorbachev, J. J. Dongarra and A. Y. Zomaya, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. pp. 142–151.
31 T. Koga, K. Kanayama, T. Watanabe, T. Imai and A. J. Thakkar, Analytical Hartree-Fock Wave Functions for the Atoms Cs to Lr, *Theor. Chim. Acta*, 2000, **104**(5), 411–413.
32 R. Garg, B. G. Vinay Kumar, G. Carneiro and I. Reid, in *Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue BT - Computer Vision – ECCV 2016*, ed. B. Leibe, J. Matas, N. Sebe and M. Welling, Springer International Publishing, Cham, 2016, pp. 740–756.
33 X. Zhang and R. Wu, Fast Depth Image Denoising and Enhancement Using a Deep Convolutional Network, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2499–2503.
34 Z. Zeng, F. Calle-Vallejo, M. B. Mogensen and J. Rossmeisl, Generalized Trends in the Formation Energies of Perovskite Oxides, *Phys. Chem. Chem. Phys.*, 2013, **15**(20), 7526.
35 A. A. Emery and C. Wolverton, High-Throughput DFT Calculations of Formation Energy, Stability and Oxygen Vacancy Formation Energy of ABO3 Perovskites, *Sci. Data*, 2017, **4**(1), 170153.
36 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), *JOM*, 2013, **65**(11), 1501–1509.
37 M. Abadi; A. Agarwal; P. Barham; E. Brevdo; Z. Chen; C. Citro; G. S. Corrado; A. Davis; J. Dean and M. Devin, Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv Prepr. arXiv1603.04467*, 2016.
38 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

