Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Machine learning enabled discovery of application dependent design principles for two-dimensional materials

MACHINE
LEARNING
Science and Technology

Machine learning enabled discovery of application dependent design principles for two-dimensional materials

OPEN ACCESS

RECEIVED
13 April 2020

REVISED
16 June 2020

ACCEPTED FOR PUBLICATION
25 June 2020

PUBLISHED
21 August 2020

Victor Venturi[1], **Holden L Parks**[1], **Zeeshan Ahmad**[1] **and Venkatasubramanian Viswanathan**[1,2]

[1] Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States of America
[2] Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States of America

E-mail: venkvis@cmu.edu

Keywords: machine learning, 2D materials, design principles

Abstract

The unique electronic and mechanical properties of two-dimensional (2D) materials make them promising next-generation candidates for a variety of applications. Large-scale searches for high-performing 2D materials are limited to calculating descriptors with computationally demanding first-principles density functional theory. In this work, we alleviate this issue by extending and generalizing crystal graph convolutional neural networks to systems with planar periodicity and train an ensemble of models to predict thermodynamic, mechanical and electronic properties. We carry out a screening of nearly 45,000 structures for two separate applications: mechanical strength and photovoltaics. By collecting statistics of the screened candidates, we investigate structural and compositional design principles that impact the properties of the structures surveyed. Our approach recovers some well-accepted design rules: hybrid organic-inorganic perovskites with lead and tin tend to be good candidates for solar cell applications and titanium based MXenes usually have high stiffness coefficients. Interestingly, other members of the group 4 elements also contribute to increasing the mechanical strength of MXenes. For all-inorganic perovskites, we discover some compositions that have not been deeply studied in the field of photovoltaics and thus open up paths for further investigation. We open-source the code-base to spur further development in this space.

1. Introduction

Two-dimensional (2D) materials have emerged as attractive candidates for energy applications due to their unique electronic, mechanical, chemical, optoelectronic and magnetic properties [1–4]. Among their different prototype structures, MXenes have been explored for applications in battery electrodes, water purification, catalysis, lubrication, etc. [5–7]. Their structure and composition allows the careful tuning of properties for these applications [8]. Most device applications of 2D materials require mechanical integration with the substrate, promoting an interest in their mechanical properties. MXenes are known to be mechanically stronger compared to other 2D materials resulting in applications in protective coatings, composites and membranes [9]. 2D materials offer a new way of tuning the properties of their 3D counterparts like band gaps through exfoliation [10]. 2D counterparts of perovskites have shown promise for solar cell applications [11] (figure 1).

The existence of a variety of 2D structures and atoms to populate their sites imply that a purely experimental or computational approach based on first-principles calculations to identify materials for desired applications is infeasible. For example, an MXene structure of the form $M_{n+1}X_nT_x$ (X = C, N) when provided with m metal possibilities, t functional group possibilities combinatorially explodes: the upper bound on the number of materials would be $\sim 2^n m^{n+1} t^x$; for example, the case, $n = 2$ (i.e. M_3X_2) with $m = 10$ metal possibilities and $t = 3$ different terminations (e.g. F, OH, O) on either side of the structure would yield approximately 35 000 materials, while just doubling the number of metal options from 10 to $m = 20$ gives ~ 280 000 possibilities. With the generation of massive amounts of materials data [12], data-driven techniques offer a new avenue to tackle this problem [13]. Data-driven methods have shown the promise of not only furthering our fundamental understanding of materials [14] but also provide a platform

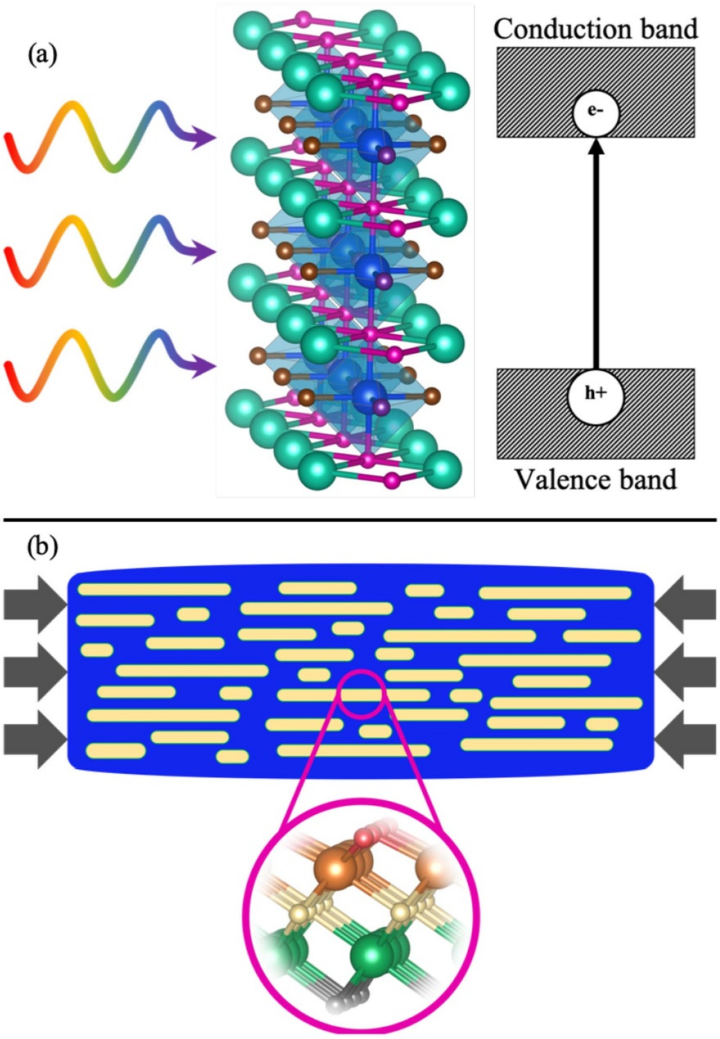

Figure 1. Illustration of 2D materials applications: (a) 2D perovskites whose bandgaps are in the range of 1.5–3 eV have been demonstrating promise in the field of photovoltaics, ; (b) MXenes can be used in composites to increase their mechanical strength, here indicated by the material's response to an external mechanical stress (in grey).

for performing large-scale computational screening through the development of accurate structure–property relationships [15–19]. Machine learning methods can bypass the use of expensive first-principles calculations and help accelerate the often time consuming discovery and optimization of materials for various applications [20–22].

Recently, graph convolution based machine learning models have shown promising generalization capability for predicting the properties of crystals and molecules [18, 23–25]. These methods encode the structure of a material as a graph based on the position and coordination of atoms, thus circumventing the need for carefully handcrafted or engineered structural features. This enables them to be used in a variety of applications. Here, we extend crystal graph convolutional neural networks (CGCNN) to study materials with planar periodicity. Using 100 different, randomly generated training sets, we trained an ensemble of CGCNN models to predict thermodynamic, mechanical and electronic properties. The ensemble of neural networks shows errors comparable to those from highly accurate first-principles calculations, such as density functional theory (DFT), as discussed in appendix A. We use this ensemble to screen $\sim$ 45 000 2D monolayer materials with focus on mechanically strong MXenes ($c_{11}, c_{22} \geq 175$ N/m) and on perovskites whose band gaps fall within an acceptable range ([1.5, 3] eV) for solar cell applications. The two applications chosen are quite different from each other and aim to demonstrate the generalizability of our model predictions. With these models, we recover some well-accepted design rules: for instance, hybrid organic-inorganic perovskites with either tin or lead as metal components are useful for photovoltaics, as are the organic cations fomamidinium, imidazolium, or azetidinium. Similarly, we find that titanium based MXenes tend to be mechanically robust. In addition, incorporating the other main elements of group 4 of the periodic table, such as zirconium and hafnium, aids in increasing the stiffness of this class of structures. Interestingly, we also identify some less commonly investigated principles: all-inorganic perovskites whose band gaps are most likely to fall within a desirable range for solar cell applications have zirconium, niobium, hafnium, scandium, or vanadium as metal components.

2. Methods

2.1. Databases

The present work utilizes computational data and material structures from (1) the Computational 2D Materials Database [26] (C2DB), (2) a database of hybrid organic-inorganic perovskites generated by Kim *et al* [27] (HOIP), (3) a database of cubic perovskites generated by Castelli *et al* [28] (Castelli) and (4) a database of 2D MXenes generated by Rajan *et al* [29] (aNANt). While all four databases contain DFT calculations, only values from the C2DB database were used to train the CGCNN models. Structures derived from the other three databases were screened after training the initial models.

We used data from the C2DB database to train the CGCNN model. As of August 2019, C2DB consists of over 3500 structural, thermodynamic, elastic, electronic, magnetic and optical properties calculated using DFT. Each structure was combinatorially generated from a series of prototype structures that differ in space group, stoichiometry and thickness. Some example prototypes include BN (space group P$\bar{3}$m2), BiI_3 (P$\bar{3}$m1), or PbSe (P4/mmm). DFT calculations were performed using the Perdew, Burke and Ernzerhof (PBE) exchange correlation functional [30] in the projector augmented wave code GPAW [31]. The stability of each material is evaluated by predicting enthalpy of formation, the elastic constants, and the phonon frequencies. If a material is stable, its electronic structure and other properties, such as polarizability, are also calculated. The most relevant properties for screening 2D MXenes and perovskites are the heat of formation; bandgap; and the c_{11}, c_{12} and c_{22} components of the elastic tensor. Using the CGCNN models trained from the C2DB data, we predict the heat of formation, bandgap, and in-plane elastic tensor components of approximately 20000 2D perovskites and 25 000 2D MXenes. The perovskite structures were taken from two sources: the HOIP dataset [27] and the Castelli database [28]. The HOIP database contains 1346 structures that were combinatorially generated from a series of 135 prototypes. The perovskite prototypes were obtained using the minima-hopping method outlined by Goedecker [32]. The prototypes were then optimized using a combination of molecular dynamics simulations and DFT calculations, coupled with the vdW-DF2 [33] exchange correlation functional as implemented in the Vienna *Ab Initio* Simulation (VASP) package [34]. Each structure has stoichiometry ABX_3, where A is one of 16 organic cations, B is one of {Ge, Pb, Sn} and X is one of {F, Cl, Br, I}. The Castelli database contains nearly 19 000 cubic perovskites. The structures were generated combinatorially with each perovskite having stoichiometry ABX_3, where A and B are each one of 52 different metals and X_3 is one of seven different anion groups, then optimized using the RPBE [35] exchange correlation functional as implemented in GPAW [31]. Both the HOIP and Castelli databases contain only bulk structures. To create 2D lattices for screening, we exfoliate the bulk structures to generate a (001) monolayer.

The 2D MXenes structures are taken from the aNANt database [29]. This database contains combinatorially generated 23 870 MXenes with five-layer structures of the form T-M-X-M'-T' (that is, the T/T' occupy the outermost layers in the structure), where T and T' are each one of 14 termination functional groups, M and M' are each one of 11 early transition metals and X is one of {C, N}. Structures were optimized using the PBE exchange correlation functional in VASP [34].

2.2. Model training

In order to screen the ~ 20 000 perovskites and ~ 24 000 MXenes 2D monolayer structures, as well as to uncover the underlying design principles for their respective applications, a technique that can predict properties accurately at a computational cost much lower than DFT is required. We use the CGCNN framework [23] as a surrogate technique for predicting material properties. This method provides the accuracy of DFT calculations (discussed in appendix A) but at a fraction of the associated computational cost: while it can take up to 500 CPU hours to compute the c_{11} coefficient for one structure with DFT, CGCNNs can predict the same property for roughly 25 000 structures in under 20 GPU minutes once trained. This framework has been successfully used in a variety of applications, from selecting solid electrolyte candidates [18] to screening catalytic materials [36]. At the foundation of the CGCNN is the undirected multigraph representation of the crystal structure, in which nodes represent atoms by their respective features and edges encode interatomic bond distances [23]. Iterative convolution layers update atomic feature vectors based on neighbor information, as further explained in appendix C. A simplified depiction of the CGCNN can be seen in figure 2.

After optimizing the network architecture (as discussed in appendix B), we used an ensemble of 100 CGCNN models, each trained on a random set of 70% of the C2DB data to predict the properties of interest:

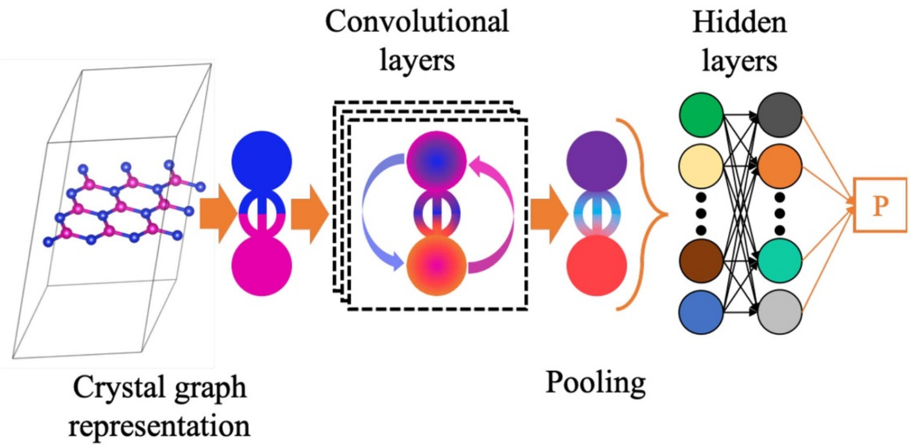

Figure 2. A simple representation of the CGCNN architecture. The atomic structure is first converted into its crystal graph representation. This graph is then passed as input for the convolutional layers, where the atomic feature vectors are updated based on neighbor information and bond lengths. Next, a pooling function is employed to produce an overall vector representation of the crystal, guaranteeing invariance with respect to number of primitive unit cells used in the creation of the original structure. Finally, a set of fully connected hidden layers maps the simplified vector-represented crystal structure to the property of interest.

band gap, $\log(c_{11})$, $\log(c_{22})$, c_{12}, conduction band minimum (CBM), valence band maximum (VBM) and heat of formation (H_{form}).

3. Results and discussion

3.1. Structure screening

In order to evaluate the accuracy of the ensemble of 100 models, we used them to predict the properties of all the structures in the C2DB database. The results for the ensemble predictions of some of the main properties of interest can be seen in the parity plots shown in figure 3. An analysis of the usefulness of the uncertainty quantification of models, as well as an investigation of the outliers of each model, can be found in appendix A.

As discussed in the Introduction, we screened MXenes in search of structures that are strong mechanically, with both $c_{11}, c_{22} \geq 175$ N/m, thus exceeding those of graphene oxide [9] and perovskites whose bandgaps fall in the range [1.5, 3] eV, appropriate for solar cell applications. Since it is also required that these structures be stable, as well as synthesizable, our filtering procedure included the additional requirement that $H_{form} \leq -2$ eV/atom for MXenes and inorganic perovskites and $H_{form} \leq -0.5$ eV/atom for hybrid organic–inorganic perovskites. The difference in treatment for the latter stems from the fact that hybrid perovskites are known to be relatively less stable than their inorganic counterparts [37]. These threshold values of H_{form} were chosen as they represent the average of the lowest heats of formation of the structures in their respective datasets.

We quantify the confidence of our predictions for a given structure s by its c-value (confidence value) [38] representing the fraction of models in the ensemble that predict structure s to be useful for the application, based on the aforementioned criterion. It is calculated in the following manner:

$$c(s) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{M}_i(s), \tag{1}$$

where $N = 100$ is the number of models used and

$$\mathcal{M}_i(s) = \begin{cases} 1 & \text{if the } i\text{th model predicts structure s to be useful} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This enables us to determine the 2D structures with the highest likelihood of being useful for their applications, as shown in table 1. It is important to note, however, that the training sets for the band gap prediction models contained only materials with non-zero band gap, meaning that a further metallic versus insulator filtering, with subsequent update of the c-values, is needed. The reason for this is that, when given a conducting material, the CGCNN models predict a positive band gap, since they have only been trained on insulators or semiconductors.

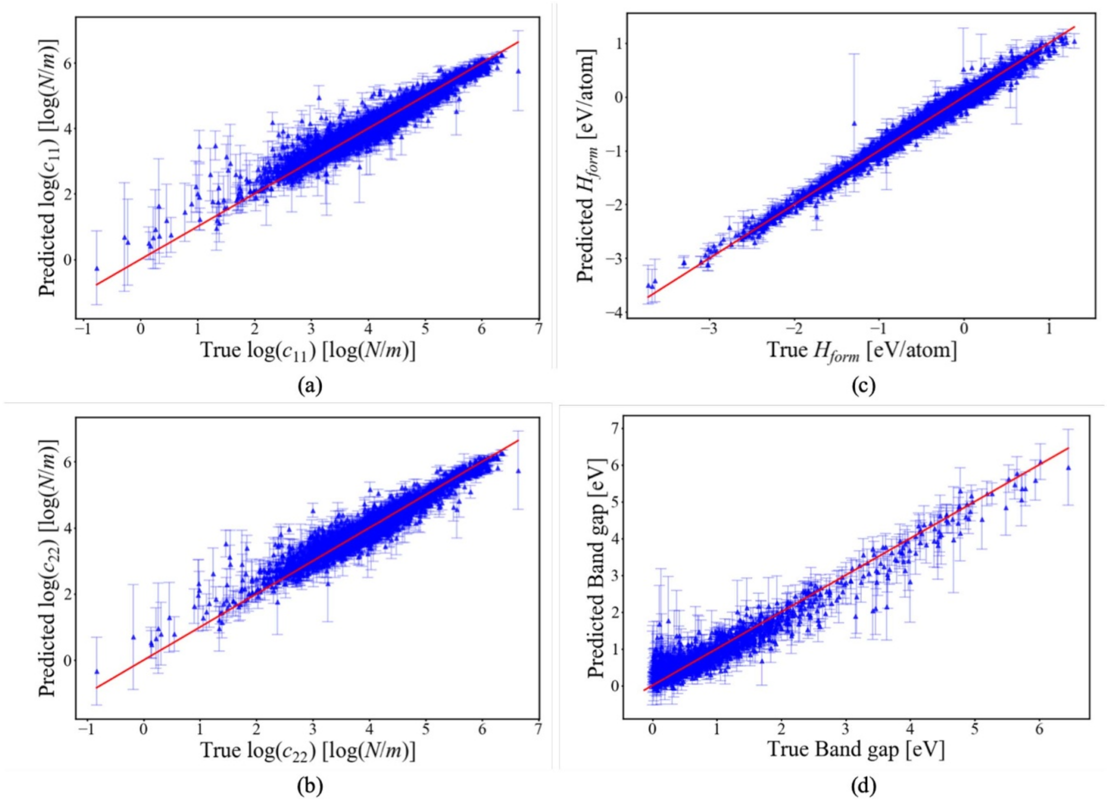

Figure 3. Comparison of predicted and DFT-calculated (a) log(c_{11}), (b) log(c_{11}), (c) H_{form} and (d) band gap on C2DB data. The predictions are made using an ensemble of 100 CGCNN models, each trained on randomly selected training data.

Using the same techniques employed in the generation of the C2DB dataset [26], we performed DFT calculations of the stiffness coefficients of all MXenes with a c-value of 1. All of these structures have both c_{11} and c_{22} greater than 175 N m^{-1}. Furthermore, when comparing the logarithm of these coefficients with those predicted by our model, we get an MAE of 0.117 log(N/m) for c_{11} and of 0.085 log(N/m) for c_{22}, with RMSE of 0.128 and 0.093 log(N/m), respectively.

3.2. Identifying design principles

To uncover the compositional and structural commonalities of useful candidates, we applied an analogous concept to study the design principles that can increase the c-values of different MXene and perovskite materials. First, for each dataset used, we establish the following functions of the design principle (DP): the subset of all structures satisfying the DP, $\mathcal{D}_{DP} = \{$structures that satisfy the DP$\}$; the proportion of the dataset that contains the DP, $P_{DP} = N_{DP}/N_{dataset}$, where $N_{DP} = |\mathcal{D}_{DP}|$ is the cardinality of set $\mathcal{D}_{DP}$ (the number of elements in this set) and $N_{dataset}$ is the total number of structures in the dataset; and the average of c-values of all structures in $\mathcal{D}_{DP}$

$$c_{DP} = \frac{1}{N_{DP}} \sum_{s \in \mathcal{D}_{DP}} c(s), \tag{3}$$

which can be interpreted as the chance of an arbitrary model predicting that a random structure in $\mathcal{D}_{DP}$ is a useful candidate.

Next, we introduce a minimum threshold, c_{cut}, to distinguish the best candidate structures from the others. The subset composed of these materials can be expressed as a function of c_{cut} as $\mathcal{B}(c_{cut}) = \{$structures with c-value $\geq c_{cut}\}$. From this definition, we can examine how the presence of a specific design rule in a material influences its existence among the best candidates in set $\mathcal{B}(c_{cut})$, as well as the c-values of these structures. For this purpose, we will define a few quantities, all functions of c_{cut}. One of the simplest indicators that a given DP is effective at making a structure useful for the application in a combinatorially generated dataset is the proportion of the set of best candidates $\mathcal{B}$ that is comprised of materials satisfying the DP, $P_{DP|best} = |\mathcal{B} \cap \mathcal{D}_{DP}|/|\mathcal{B}|$ and how it compares with P_{DP}. Additionally, it is helpful to examine the difference between the likelihood of a random material being amongst the best candidates $P_{best|All} = |\mathcal{B}|/N_{dataset}$ and the chance of that happening given that the structure contains the DP,

Table 1. Materials with highest five *c*-values. For MXenes, '*' indicates a bond between a metallic atom and a termination. For the perovskites, the site occupations have been specified for clarity. Values reported are the mean of the ensemble predictions.

	Structure	*c*-value	$\langle H_{form} \rangle$ [eV/atom]	$\langle c_{11} \rangle$ [N/m]	$\langle c_{22} \rangle$ [N/m]	$\langle bandgap \rangle$ [eV]
MXenes	Hf*O-N-Hf*O	1.0	−2.62	273.88	261.60	-
	Sc*O-N-Hf*O	1.0	−2.64	210.15	224.16	-
	Sc*F-N-Hf*O	1.0	−2.63	240.97	220.28	-
	Hf*O-C-Zr*F	1.0	−2.23	266.41	249.99	-
	Hf*O-N-Zr*Cl	1.0	−2.16	211.11	227.78	-
Inorganic Perovskites	$NbZrO_3$; A = Zr, B = Nb	0.98	−2.48	-	-	2.28
	$HfVO_3$; A = Hf, B = V	0.98	−2.33	-	-	2.33
	$MoZrO_3$; A = Zr, B = Mo	0.94	−2.25	-	-	2.38
	$HfNbO_3$; A = Nb, B = Hf	0.94	−2.20	-	-	1.99
	$NbTiO_3$; A = Nb, B = Ti	0.92	−2.18	-	-	2.08
Organic Perovskites	$C_3H_5F_7N_2Sn_2$; A = $C_3H_5N_2$	0.99	−1.21	-	-	2.38
	$C_3H_8F_7NPb_2$; A = C_3H_8N	0.94	−1.13	-	-	2.67
	$C_3H_5F_7N_2Pb_2$; A = $C_3H_5N_2$	0.93	−1.19	-	-	2.63
	$C_2H_7F_7N_2Sn_2$; A = $CH_3C(NH_2)_2$	0.92	−1.37	-	-	2.67
	$CH_5F_7N_2Sn_2$; A = $HC(NH_2)_2$	0.91	−1.60	-	-	2.36

Table 2. Design principles with five highest values of ratio $\frac{P_{DP|best}}{P_{DP}}$. The cutoff c-values used in computing values displayed for MXenes was $c_{cut} = 0.95$ and for both inorganic and organic perovskite cases, $c_{cut} = 0.80$.

	Design principle	c_{DP}	P_{DP} [%]	$P_{DP\|best}$ [%]	$P_{best\|All}$ [%]	$P_{best\|DP}$ [%]	$P_{DP\|best}P_{DP}$
MXenes	Hf*O-O-N	0.729	0.046	12.069	0.243	63.636	261.897
	Ti*O-O-N	0.683	0.046	10.345	0.243	54.545	224.483
	Hf*O-O	0.560	0.092	18.966	0.243	50.000	205.776
	Zr*O-F-C	0.448	0.046	8.621	0.243	45.455	187.069
	Hf*O-F	0.570	0.092	17.241	0.243	45.455	187.069
Inorganic Perovskites	A = Sc-B = Cr	0.264	0.037	10.526	0.100	28.571	284.632
	A = Zr-B = Sc	0.406	0.037	10.526	0.100	28.571	284.632
	A = Sc-B = V	0.320	0.037	5.263	0.100	14.286	142.316
	A = Sc-B = Nb	0.274	0.037	5.263	0.100	14.286	142.316
	A = Hf-B = V	0.294	0.037	5.263	0.100	14.286	142.316
Organic Perovskites	A = HC(NH$_2$)$_2$-X = F	0.817	0.446	13.333	2.229	66.667	29.911
	A = C$_3$H$_5$N$_2$-X = F	0.790	1.560	43.333	2.229	61.905	27.775
	A = C$_3$H$_8$N-X = F	0.629	2.675	30.000	2.229	25.000	11.217
	A = C$_3$H$_5$N$_2$-B = Sn	0.282	2.452	26.667	2.229	24.242	10.877
	A = C$_3$H$_5$N$_2$-B = Pb	0.255	1.783	16.667	2.229	20.833	9.347

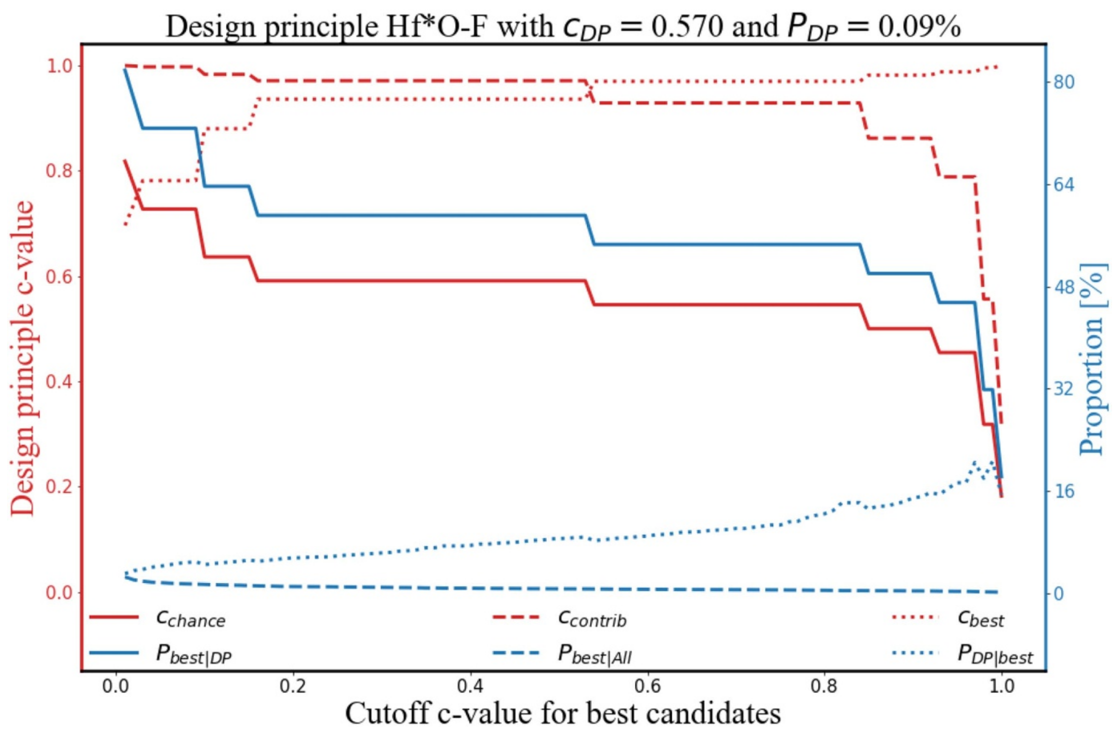

Figure 4. One of the top design principles for MXenes: hafnium (Hf) bonded with oxygen (O) termination and with a fluorine (F) termination bonded to any other metal. The likelihood that any arbitrary model predicts that a random material satisfying this DP is a useful candidate is of ~ 57%, as indicated by c_{DP}. Choosing $c_{cut} = 1.0$ shows that the chance of a structure satisfying the DP being among the best candidates is of nearly $c_{chanceDP} = P_{best|DP} \sim 20\%$ and these candiates contribute to approximately 30% of c_{DP}, as shown by $c_{contribDP}$. Note that, for this specific DP, $P_{DP|best}$ almost steadily increases with the value of c_{cut}, indicating that, the more confident we want to be in our set of useful candidates, in general, the more prevalent this DP becomes in this $\mathcal{B}$ set.

$c_{chanceDP} = P_{best|DP} = |\mathcal{B} \cap \mathcal{D}_{DP}|/|\mathcal{D}_{DP}|$. Besides these quantities, it is also important to measure our confidence in these candidates, members of $\mathcal{B} \cap \mathcal{D}_{DP}$, by averaging their c-values:

$$c_{bestDP} = \frac{1}{|\mathcal{B} \cap \mathcal{D}_{DP}|} \sum_{s \in \mathcal{B} \cap \mathcal{D}_{DP}} c(s). \tag{4}$$

Note that, by construction, c_{bestDP} is a monotonically increasing function of c_{cut} while the set $\mathcal{B} \cap \mathcal{D}_{DP}$ is not empty. Finally, although redundant with all previously described measures, we also studied, for

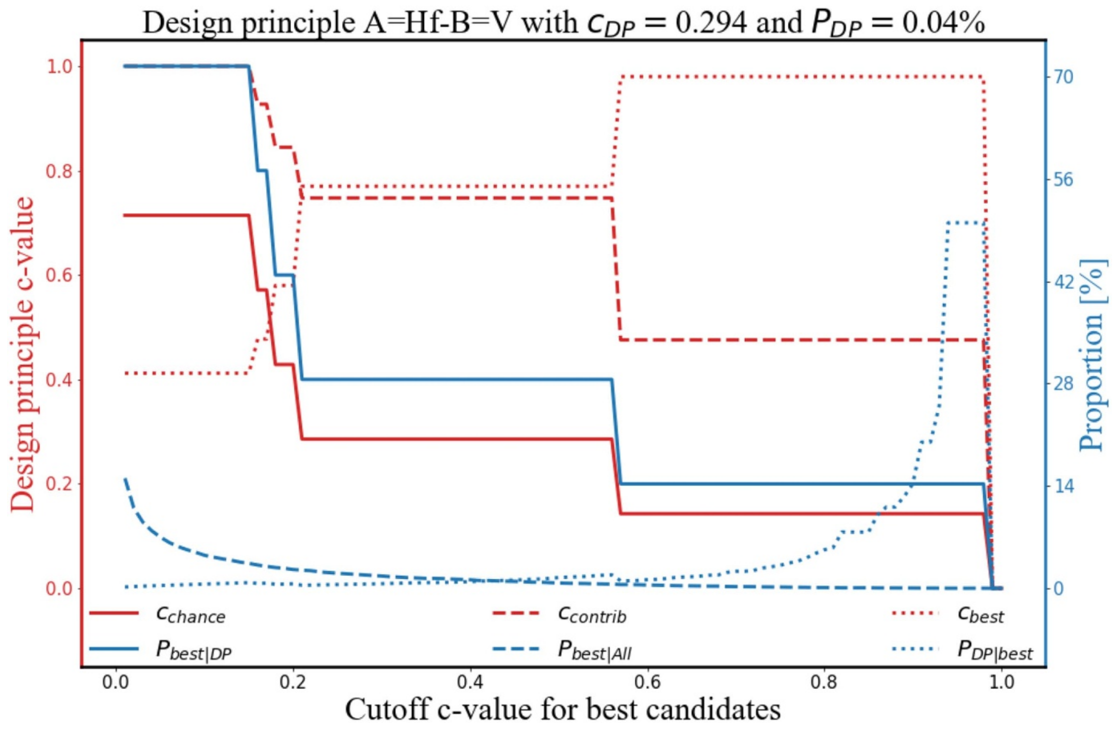

Figure 5. One of the top design principles for inorganic perovskites: A-site occupied by hafnium (Hf), with vanadium (V) in the B-site. The likelihood that any arbitrary model predicts a random material satisfying this DP is a useful candidate is of ~ 29%, as indicated by c_{DP}. Choosing $c_{cut} = 0.80$ shows that the chance of a structure from set $\mathcal{D}_{DP}$ to be among the best candidates is of $c_{chanceDP} = P_{best|DP} = 14\%$ and these candidates contribute to approximately 50% of c_{DP}, as shown by $c_{contribDP}$. Additionally, for this specific DP, $P_{DP|best}$ almost steadily increases with the value of c_{cut}, indicating that, the more confident we want to be in our set of useful candidates, in general, the more prevalent this DP becomes in this $\mathcal{B}$ set. Finally, an examination of the behavior of c_{best} reveals that, for values of the cutoff $c_{cut} > 0.6$, all of the structures in set $\mathcal{B} \cap \mathcal{D}_{DP}$ have a c-value of nearly 1.0.

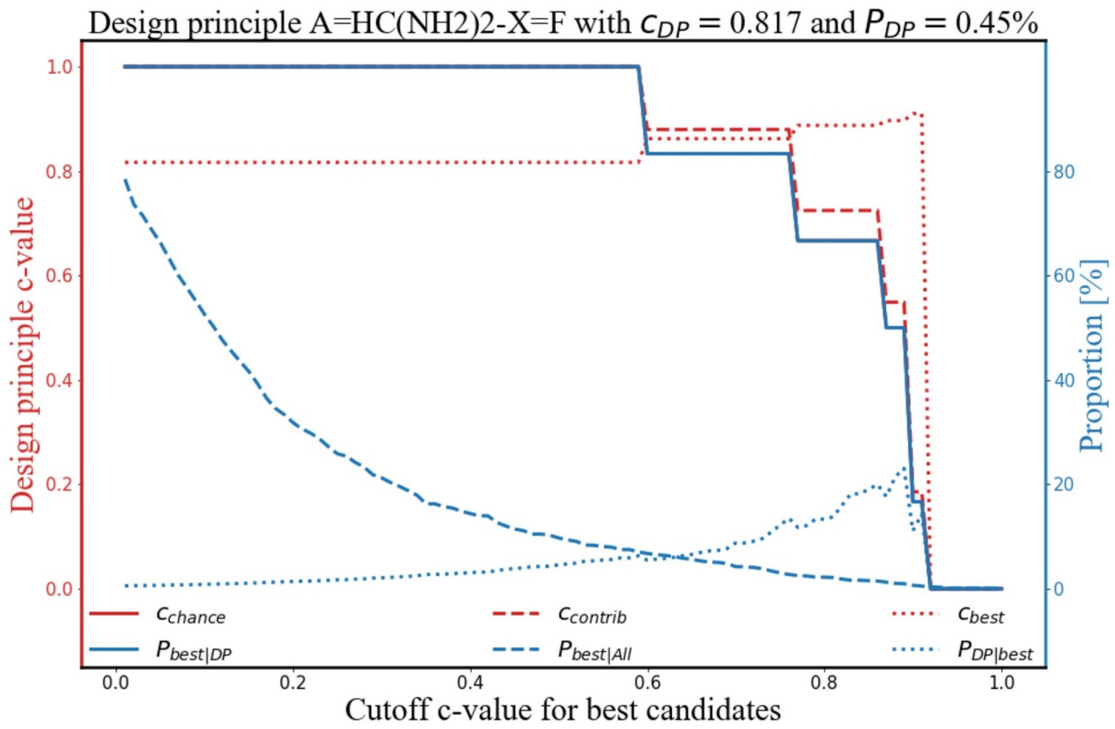

Figure 6. Top design principle for Khazana perovskites: A-site occupied by fomamidinium ($HC(NH_2)_2$), with fluorine (F) in all X-sites. The likelihood that any arbitrary model predicts a random material satisfying this DP is a useful candidate is of ~ 82%, as indicated by c_{DP}. Choosing $c_{cut} = 0.80$ shows that the chance of a structure from set $\mathcal{D}_{DP}$ to be among the best candidates is of $c_{chanceDP} = P_{best|DP} =\sim 65\%$ and these candidates contribute to approximately ~ 70% of c_{DP}, as shown by $c_{contribDP}$.

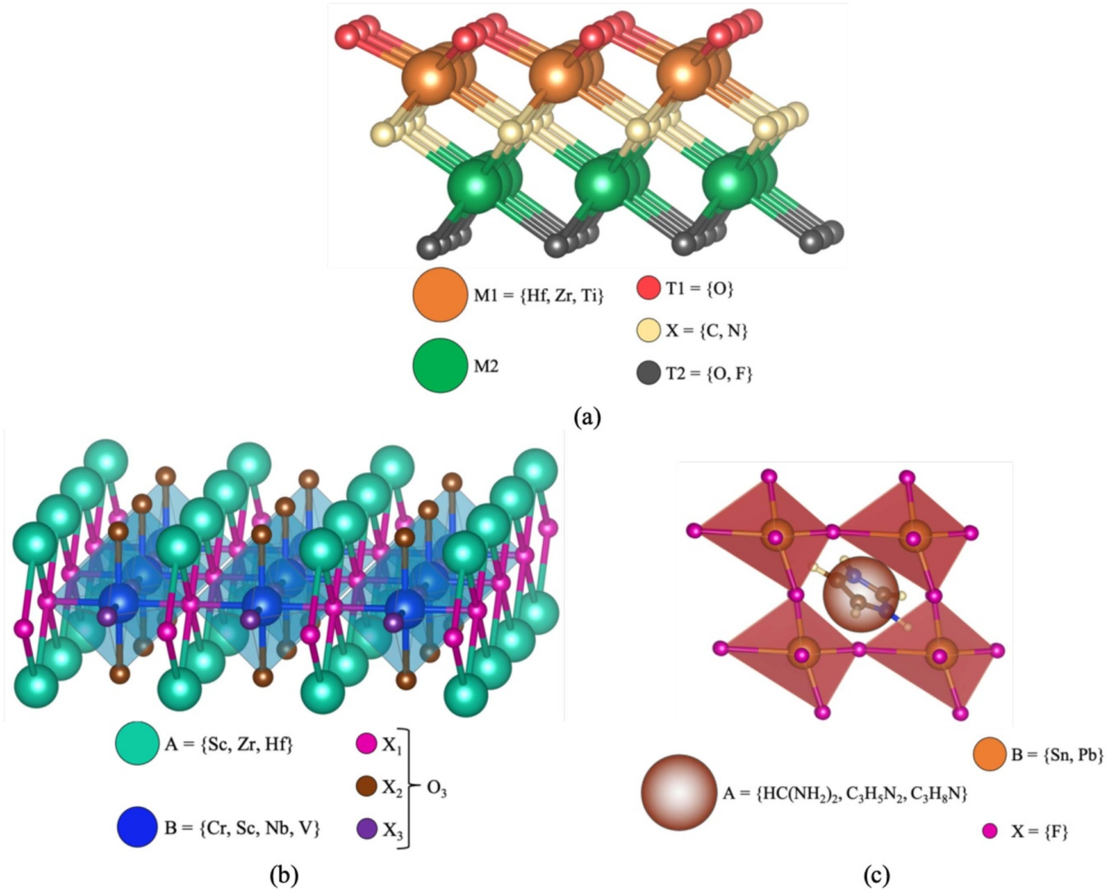

Figure 7. Summary of uncovered design rules for: (a) MXenes, where hafnium (Hf), zirconium (Zr), or titanium (Ti) must be bonded to an oxygen termination, regardless of the central X component or the metal M2 on the opposite side of the structure, which should have either an oxygen or a fluorine termination; (b) inorganic perovskites, in which the A-site must be occupied by scandium (Sc), zirconium (Zr), or hafnium (Hf), the B-site has to contain chromium (Cr), scandium (Sc), niobium (Nb), or vanadium (V) and, as discussed in the main text, all X-sites should be occupied by oxygen; (c) organic-inorganic hybrid perovskites, where all X-sites must be occupied by fluorine, the B-site can be occupied by either tin (Sn) or lead (Pb) and the A-site possibilities are fomamidinium ($HC(NH_2)_2$), imidazolium ($C_3H_5N_2$) and azetidinium (C_3H_8N). Note: in the inorganic perovskite case, even though all X-site occupations should be the same, we represented them by different colors for completeness, since changing the value of c_{cut} used in the analysis allows for other possibilities where all three distinct types of X-sites can be populated by a different atom.

completeness, how the elements from $\mathcal{B} \cap \mathcal{D}_{DP}$ contribute to c_{DP}:

$$c_{contribDP} = \frac{\sum\limits_{s \in \mathcal{B} \cap \mathcal{D}_{DP}} c(s)}{\sum\limits_{s' \in \mathcal{D}_{DP}} c(s')} = \frac{|\mathcal{B} \cap \mathcal{D}_{DP}| c_{bestDP}}{|\mathcal{D}_{DP}| c_{DP}} = P_{best|DP} \frac{c_{bestDP}}{c_{DP}}. \qquad (5)$$

Since the higher the value of the cutoff, the fewer elements are in the set $\mathcal{B} \cap \mathcal{D}_{DP}$, both $c_{contribDP}$ and $c_{chanceDP}$ are monotonically decreasing with c_{cut}. A full dependency of all these variables with the value of the cutoff c_{cut}, for chosen design principles, can be seen in figures 4, 5 and 6. Note that, for all three of the design rules chosen, $P_{DP|best}$ is almost a monotonically increasing function of c_{cut}, indicating that, the more confident one wants to be on the $\mathcal{B}$ set, the more predominant these design principles become in this set. Additionally, prior to the value of c_{cut} for which none of the materials in $\mathcal{B}$ contains the design principles, the chance of finding a member of $\mathcal{B}$ among the set $\mathcal{D}_{DP}$ is of roughly 15% for all three design principles. (Note: values for $c_{cut} = 0$ omitted in the interest of ease of graphical visualization.) This approach of understanding the effect of design principles is best suited for combinatorially generated datasets. Therefore, we used it to study all of the data mentioned in the Datasets section, including that from HOIP [27]. We constructed a list of all possible design principles using the same combinatorics applied in the creation of the respective datasets. These design principles were then ordered by highest to lowest $P_{DP|best}/P_{DP}$ ratio at a cutoff value of $c_{cut} = 0.95$ for MXenes and $c_{cut} = 0.80$ for both inorganic and organic perovskites. Our choice for cutoff values was guided by the results from table 1: we wanted to make sure that the set of best candidates $\mathcal{B}$ was sizeable enough for a meaningful analysis of the design principles. Furthermore, for the MXenes, we excluded the design rules whose $P_{DP} \leq 0.008\%$ due to their high specificity. We would like to note the following interesting equality, which establishes a relationship between the two most intuitive criteria of gauging the effectiveness of a given DP discussed previously:

$$\frac{P_{DP|best}}{P_{DP}} = \frac{|\mathcal{B} \cap \mathcal{D}_{DP}|}{|\mathcal{B}|} \frac{N_{dataset}}{|\mathcal{D}_{DP}|} = \frac{|\mathcal{B} \cap \mathcal{D}_{DP}|}{|\mathcal{D}_{DP}|} \frac{N_{dataset}}{|\mathcal{B}|} = \frac{P_{best|DP}}{P_{best|All}} \tag{6}$$

The results of this analysis are shown in table 2. We have also chosen one of the top design principles for MXenes (figure 4), inorganic (figure 5) and organic perovskites (figure 6) to represent the dependency between the metrics discussed above and the cutoff c_{cut}, which determines the minimum confidence level of the structures in the set of best candidates $\mathcal{B}$.

3.3. Interpreting design principles

Our study was able to identify some known design rules: for example, we see that titanium (Ti) based MXenes tend to have high stiffness coefficients, as suggested by figure 4 and table 1 in [8]. Interestingly, however, we discovered that the other main elements of group 4 of the periodic table, namely, zirconium (Zr) and hafnium (Hf), can also increase the mechanical strength of this class of materials, as long as these elements are bonded with oxygen and the opposite side of the monolayer is either oxygen or fluorine terminated.

Similarly, our model was able to recognize that, in order for hybrid organic-inorganic perovskites to have a band gap in the range of [1.5, 3] eV, the B-sites should be occupied by either lead (Pb) or tin (Sn), a relatively well-known design principle in the photovoltaics community. [39–42] We have also found that the organic A-sites should be composed of fomamidinium ($HC(NH_2)_2$), imidazolium ($C_3H_5N_2$), or azetidinium(C_3H_8N). Curiously, for these hybrid perovskites, our method suggests that the X-sites be populated by fluorine, rather than the usual iodine. A deeper investigation suggests that the reason for this is the enhanced stability of the fluorinated structures: while fluorinated structures have an average band gap of ~ 3 eV and $\langle H_{form} \rangle \approx -1.1$ eV/atom, iodined perovskites have an average band gap of 2.6 eV, but a much higher $\langle H_{form} \rangle \approx -0.3$ eV/atom.

Finally, for purely inorganic perovskites, we find that the A-sites should be occupied by scandium (Sc), hafnium (Hf), or zirconium (Zr). When analyzing the atomistic features used by the CGCNN model for these elements, we find that all of them have a covalent radius of ~ 170 pm, a first ionization potential in the neighborhood of 640 kJ mol^{-1} and a 1.3 electronegativity in Pauling units. At the same time, the B-sites should be populated with vanadium (V), niobium (Nb), or chromium (Cr), all with atomic radii of ~ 130 pm, first ionization potential of roughly 650 kJ mol^{-1} and an electronegativity of approximately 1.6 in the Pauling scale. Surprisingly, in the field of all-inorganic perovskites for photovoltaics applications, none of these compositions has been deeply investigated; focus has been more directed towards caesium-lead systems ($CsPbX_3$, where X can be I, Br, or Cl) [43], indicating our work may contain new potential directions for further research in this area of science. Figure 7 contains a graphical summary of the top design principles uncovered in this work.

The collection of design rules we uncovered shows the capability of our model to both identify established criteria to attaining material performance, as well as to find new, unexplored avenues for application-focused material discovery, since it can be considered a basis for reverse engineering of 2D structures. We believe that machine learning methods such as CGCNN, coupled with a study of structural and compositional design rules, can open up paths for material innovation in a myriad of fields, including photovoltaics, electrochemistry, batteries, mechanically robust materials, among others. In the interest of further accelerating the discovery and screening of more 2D monolayer materials, we have open-sourced our code base on GitHub.

4. Conclusions

In this work, we have extended the CGCNNs to describe materials with planar symmetry. Using this model, we screened large combinatorially generated datasets of MXene and perovskite materials in search of those with high likelihood of having properties of interest, as determined by the ensemble of trained CGCNN models. Using the results from the screening process, we were able to uncover the underlying molecular design principles for their respective applications.

Some of the identified design rules have already been recognized in other literature and are well-accepted, further demonstrating the robustness of the developed methodology. One such example is identifying hybrid organic–inorganic perovskites with lead or tin as good candidates for solar cell applications [39–42], while titanium based MXenes as mechanically robust materials [8]. On the other hand, other design principles identified could open up new avenues for material exploration. One such design rule is that MXene monolayers with elements from group 4 of the periodic table are likely to have high stiffness coefficients.

Finally, the design rules we uncovered can be used as guidance for both experimental and computational testing at different confidence levels. By combining design principles together, as well as by combinatorially populating their unspecified structural sites, datasets of potential high-performance materials can be created. This reverse engineering approach of using design rules as a generative basis can lead to the discovery of even better materials, along with more effective design principles.

5. Data availability

The CGCNN modified code base can be found on GitHub, as well as instructions on how to download it, set up a virtual environment and run it. Further details that support the findings of this study are available from the corresponding author, V Viswanathan, upon reasonable request.

Acknowledgments

The authors thank T Xie & R Kurchin for helpful discussions. This work was supported in part by the Advanced Research Projects Agency-Energy Integration and Optimization of Novel Ion Conducting Solids (IONICS) program under Grant No. DE-AR0000774. V Venturi was supported in part by the Richard King Mellon Foundation Presidential Fellowship in Energy from the College of Engineering at Carnegie Mellon University. H P acknowledges support from the Office of Naval Research under Award No. N00014-19-1-2172. Acknowledgment is also made to the Extreme Science and Engineering Discovery Environment (XSEDE) for providing computational resources through Award No. TG-CTS180061.

Appendix A. CGCNN Ensemble Performance

In order to measure the performance of our model ensemble, we apply some of the metrics introduced by Kuleshov *et al* [44] and Tran *et al* [45], namely, calibration and sharpness, besides mean absolute error (MAE) and root mean square error (RMSE).

In their work, the authors institute the concept of a calibration plot, which compares, for each predicted data point, the standard deviation of the ensemble predictions (y-axis) and the residual between the mean of the predictions and the true value of the data point—the mean error of the predictions (x-axis). In the regions where the observed estimation interval is greater than the expected interval (green), the model ensemble is called underconfident, since the true value falls within the error bars of the ensemble prediction. On the other hand, in the regions where the observed estimation interval is smaller than the expected interval (red), the model is called overconfident, since the standard deviation of the ensemble predictions does not encompass the true data value. The calibration plot for our 100 model ensemble trained to predict conduction band maximum is shown in figure A1.

However, measuring calibration is not sufficient, though necessary, for effective uncertainty quantification. For instance, a well calibrated model with large uncertainty estimates is less useful than a similarly calibrated model with smaller uncertainties. Thus, the concept of sharpness is introduced: a sharper model is that whose prediction standard deviations are smaller. Alternatively, sharpness can be interpreted as a measure of the precision of the model (while MAE/RMSE can be understood as the accuracy); the smaller its value, the more precise the model is. This metric is measured in the following manner [45]:

$$sharpness = \sqrt{\frac{1}{N_{dataset}} \sum_{structure} Var[\mathcal{M}(structure)]}$$

Figure A2 illustrates histograms of the mean prediction errors (blue) and ensemble standard deviations (red) of the models trained over band gap data. The values of prediction mean absolute error (MAE) and root mean square error (RMSE), as well as ensemble sharpness, are also represented. Table A1 contains some of these metrics for all predicted properties. Taking as example our band gap and heat of formation predictions, one can see that our approach performs better than some first-principle simulations: for band gap, the accepted DFT errors are between 0.25 and 0.4 eV [46, 47], while both our MAE and sharpness fall on the lower end of this range; for H_{form}, DFT errors are of usually 0.1 eV atom^{-1} [23], while all our uncertainty metrics are below this value by a safe margin.

CGCNN is a direction-agnostic framework for machine learning and since the material stiffness only depends on the relative positions of the atoms in the crystal, we can use them to predict the elastic constants as seen from the low MAE and RMSE (table A1). We regressed over the log of c_{11} and c_{22} since they are positive and want to avoid overweighing elastic constants of very stiff materials.

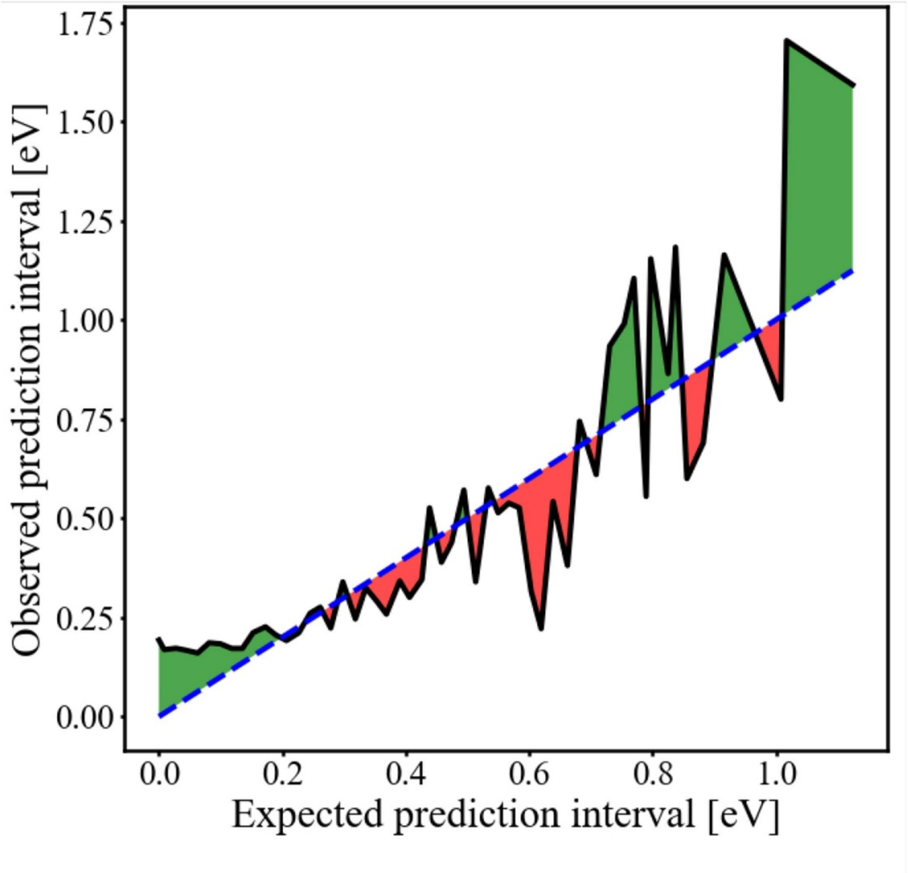

Figure A1. Calibration plot [44, 45] of CGCNN ensemble CBM prediction. In general, the ensemble of model predictions captures the true values of CBM within one standard deviation.

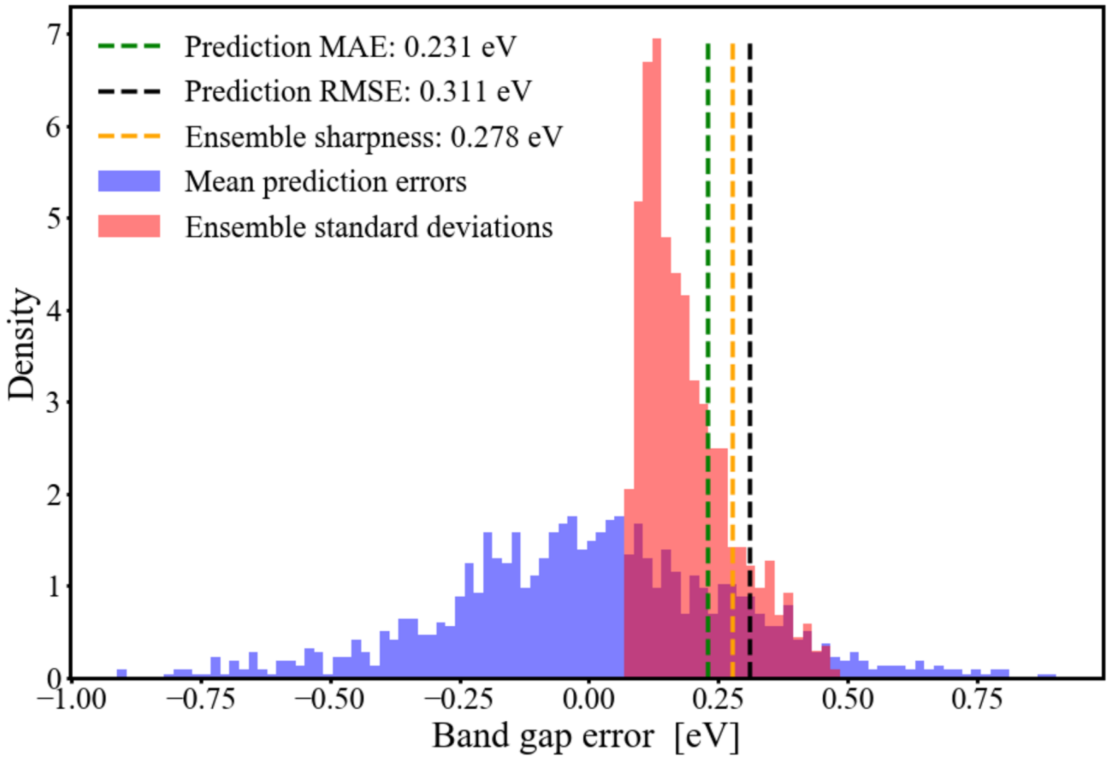

Figure A2. Histograms of errors on band gap prediction and of standard deviation of ensemble predictions.

Table A1. Metrics for uncertainty quantification of model ensembles.

Property	MAE	RMSE	Sharpness	Unit
$\log(c_{11})$	0.182	0.263	0.188	$\log(N/m)$
c_{12}	8.241	12.497	14.079	N/m
$\log(c_{22})$	0.174	0.250	0.172	$\log(N/m)$
CBM	0.193	0.264	0.310	eV
VBM	0.180	0.251	0.286	eV
Band gap	0.231	0.311	0.278	eV
H_{form}	0.066	0.090	0.072	eV/atom
Speed of sound x	385.703	552.147	366.810	m s^{-1}
Speed of sound y	372.015	548.619	351.624	m s^{-1}

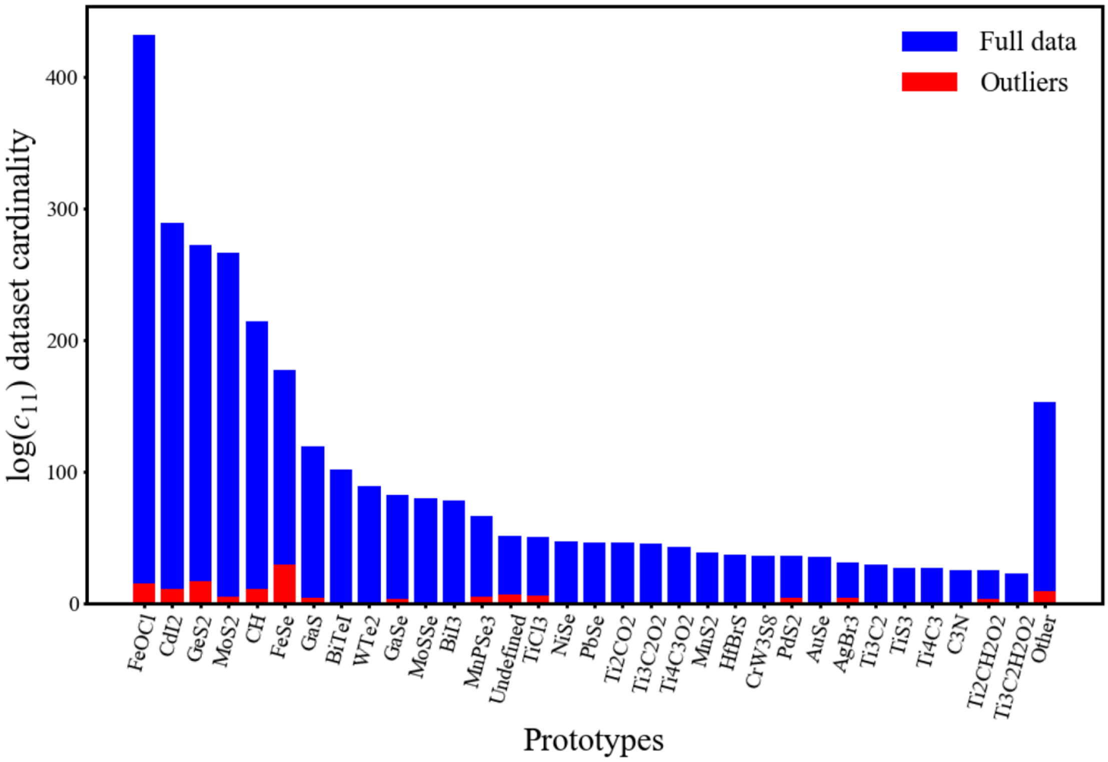

Figure A3. Distribution of C2DB structures by prototype for $\log(c_{11})$. All prototypes whose total representation accounts for less than 5% of the entire C2DB dataset are binned together under the label 'Other'.

Indirectly related to the ensemble performance are its outliers. For each property of interest, we considered the structures whose mean prediction error was more than two standard deviations away from the average of all residuals, that is, those who would be at the trailing ends of the blue distribution on figure A2. We then categorized these outliers by their prototypes in the C2DB dataset and studied how the outlier prototype distribution compares to the prototype distribution in the entire dataset, as shown in figure A3 (for the property $\log(c_{11})$). As one might expect, for all properties, the outlier prototype distribution somewhat follows the data distribution: the prototypes with higher representation in the data also have higher representation among the outliers. While that is true in general, there exist, however, a few prototypes that deviate from this rule, which can be seen by normalizing the number of outliers in each prototype bin by the total number of structures of said prototype in the data, as can be seen in figure A4. In doing so, we were able to evaluate what are the most problematic prototypes for each property. Some trends emerged from our analysis: for nearly all mechanical properties (namely, c_{11}, c_{22} and speed of sound on x and y directions), the FeSe prototype proved to be the one with highest ratio of outliers ($\sim$ 16%), while, for electronic properties (CBM, VBM and band gap), 'Other' prototypes were always among the five prototypes with highest outlier percentage. The latter is not particularly surprising, since 'Other' contains all prototypes with little representation among the data, which makes them more difficult for a neural network to learn.

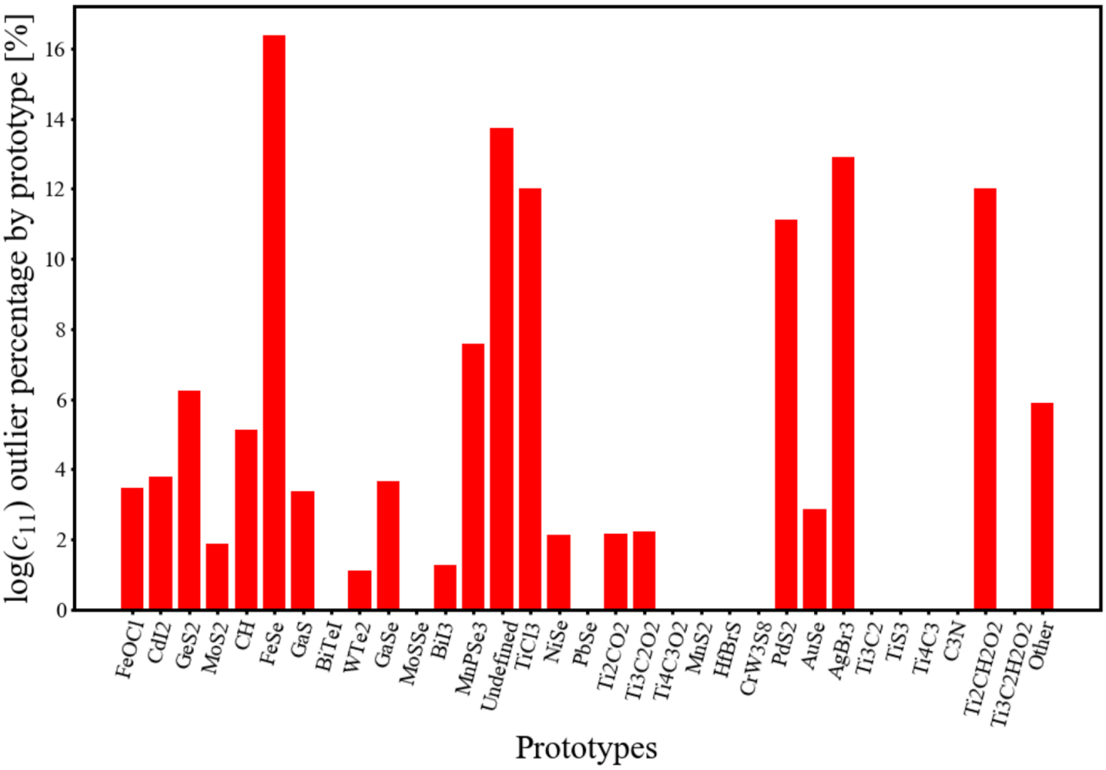

Figure A4. Percentage of log(c_{11}) outliers by prototype. For most mechanical properties, FeSe is the prototype with highest ratio of outliers.

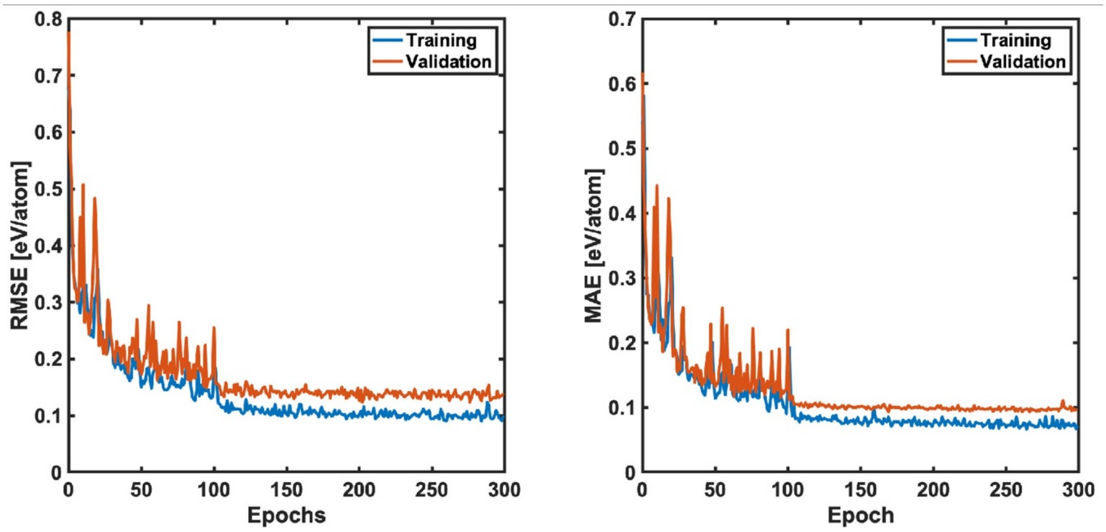

Figure B1. Example of root mean square (RMSE) and mean absolute error (MAE) curves used in optimization of network architecture and hyperparameters. These curves were used to evaluate the prediction performance of models using different number of convolution layers, hidden layers, epochs, number of neighbors used in convolution operations, among others. Represented here, we have the results from training a model to predict H_{form} using a mean pooling function, 300 epochs, 1 hidden and 2 convolution layers.

Appendix B. CGCNN Network Optimization

In this work, we apply CGCNN's power of accurately predicting properties of periodic materials to investigate 2D materials, namely, MXenes and perovskites. Using a 70:15:15 training:validation:test split ratio on the C2DB database for the heat of formation property (H_{form}), we first optimized the network architecture, including number of convolution and hidden layers, learning rate and number of epochs to be used in training and the models' performances were evaluated as shown in figure B1. For example, keeping

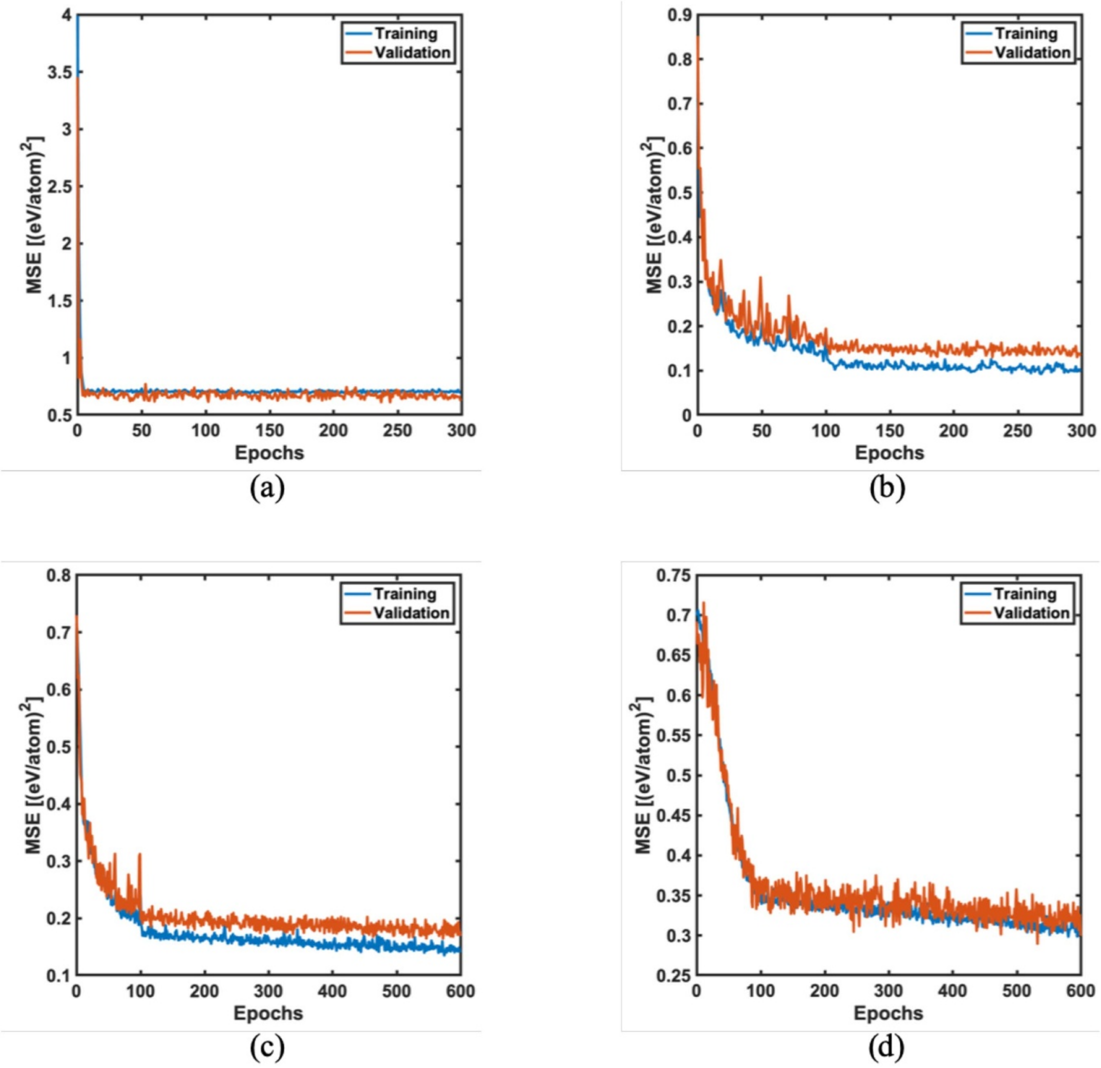

Figure B2. Mean square error (MSE) of H_{form} predictions during training for learning rates of (a) 0.1, (b) 0.01, (c) 0.001 and (d) 0.0001. The lowest errors are obtained with a learning rate of 0.01.

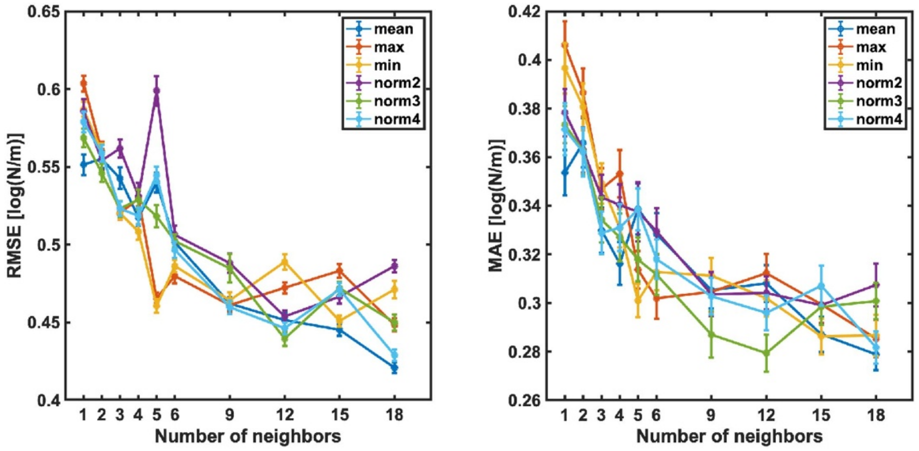

Figure B3. Evaluation of how pooling functions, as well as number of neighbors used in convolution operations affect RMSE and MAE in log(c_{11}) prediction. We also implemented higher order norm functions as pooling operators.

all other hyperparameters fixed, learning rates of 0.1, 0.01, 0.001, 0.001 were used for training and, of those, the one that yielded the best model performance was chosen, as can be seen in figure B2. The number of epochs and of both convolution and hidden layers were optimized in a similar manner. Different possibilities of pooling functions (mean, max and min) were also tested, as shown in figure B3. The final architecture used in the models was composed of 2 convolution layers and 1 hidden layer post-pooling. The networks were trained with a learning rate of 0.01 and a mean pooling function over 300 epochs.

Appendix C. Graph convolution description

As briefly described in the main text, the core of CGCNN is in the representation of a crystal structure as an undirected graph $\mathcal{G}$, with a set of nodes V and edges U. Each node $v_i \in V$ corresponds to an atom in the crystal structure and is represented by the atom's feature vector, which includes properties such as group number, period number, electronegativity, number of valence electrons, among others. Similarly, each edge $u_{(i,j)_k} \in U$ corresponds to a bond in the crystal structure between atoms i and j and is represented by a bond feature vector. Here, the subscript k indicates the possibility of there being multiple bonds between atoms i and j. We denote the atomistic feature vectors at convolution step t by $v_i^{(t)}$ and the bond vectors by $u_{(i,j)_k}^{(t)}$. From convolution layer t to $t+1$, the atomistic feature vectors are updated in the following manner:

$$v_i^{(t+1)} = v_i^{(t)} + \sum_{j,k} \sigma\left(z_{(i,j)_k}^{(t)} \cdot W_f^{(t)} + b_f^{(t)}\right) \odot g\left(z_{(i,j)_k}^{(t)} \cdot W_s^{(t)} + b_s^{(t)}\right),$$

where $z_{(i,j)_k}^{(t)} = v_i^{(t)} \oplus v_j^{(t)} \oplus u_{(i,j)_k}$ ($\oplus$ indicating concatenation between vectors), σ is a sigmoid function, $\odot$ denotes element wise multiplication, g is a non-linear activation function and the $W_f^{(t)}$, $W_s^{(t)}$ and $b_f^{(t)}$, $b_s^{(t)}$ are the weights and biases of the convolutional operation from layer t to layer $t+1$.

ORCID iDs

Victor Venturi https://orcid.org/0000-0001-5672-7549
Holden L Parks https://orcid.org/0000-0002-9530-0764
Zeeshan Ahmad https://orcid.org/0000-0001-9758-8952
Venkatasubramanian Viswanathan https://orcid.org/0000-0003-1060-5495

References

[1] Novoselov K S, Mishchenko A, Carvalho A and Castro Neto A H 2016 2D materials and van der Waals heterostructures *Science* **353** aac9439

[2] Zhang X, Hou L, Ciesielski A and Samorì P 2016 2D materials beyond graphene for high-performance energy storage applications *Adv. Energy Mater.* **6** 1600671

[3] Quesnel E *et al* 2015 Graphene-based technologies for energy applications, challenges and perspectives *2D Mater.* **2** 030204

[4] Ge M, Cao C, Huang J, Shuhui Li, Chen Z, Zhang K-Q, Al-Deyab S S and Lai Y 2016 A review of one-dimensional TiO$_2$ nanostructured materials for environmental and energy applications *J. Mater. Chem.* A **4** 6772–801

[5] Pang J, Mendes R G, Bachmatiuk A, Zhao L, Ta H Q, Gemming T, Liu H, Liu Z and Rummeli M H 2019 Applications of 2D MXenes in energy conversion and storage systems *Chem. Soc. Rev.* **48** 72–133

[6] Dequan E, Junwen Li, Naguib M, Gogotsi Y and Shenoy V B 2014 Ti$_3$C$_2$ MXene as a high capacity electrode material for metal (Li, Na, K, Ca) ion batteries *ACS Appl. Mater. Interfaces* **6** 11173–9

[7] Chaudhari N K, Jin H, Kim B, Baek D S, Joo S H and Lee K 2017 MXene: an emerging two-dimensional material for future energy conversion and storage applications *J. Mater. Chem.* A **5** 24564–79

[8] Anasori B, Lukatskaya M R and Gogotsi Y 2017 2D metal carbides and nitrides (MXenes) for energy storage *Nat. Rev. Mater.* **2** 1–17

[9] Lipatov A, Haidong L, Alhabeb M, Anasori B, Gruverman A, Gogotsi Y and Sinitskii A 2018 Elastic properties of 2D Ti$_3$C$_2$T$_x$ MXene monolayers and bilayers *Sci. Adv.* **4** eaat0491

[10] Mounet N *et al* 2018 Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds *Nat. Nanotechnol.* **13** 246–52

[11] Xiao Z *et al* 2016 Photovoltaic properties of two-dimensional (CH$_3$NH$_3$)$_2$Pb(SCN)$_2$I$_2$ perovskite: A combined experimental and density functional theory study *J. Phys. Chem. Lett.* **7** 1213–18

[12] Himanen L, Geurts A, Foster A S and Rinke P 2019 Data-driven materials science: Status, challenges and perspectives *Adv. Sci.* **6** 1900808

[13] Agrawal A and Choudhary A 2016 Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science *APL Mater.* **4** 053208

[14] Umehara M, Stein H S, Guevarra D, Newhouse P F, Boyd D A and Gregoire J M 2019 Analyzing machine learning models to accelerate generation of fundamental materials insights *npj Comput. Mater.* **5** 1–9

[15] Coley C W, Barzilay R, Jaakkola T S, Green W H and Jensen K F 2017 Prediction of organic reaction outcomes using machine learning *ACS Cent. Sci.* **3** 434–43

[16] Evans J D and Coudert Fçois-X 2017 Predicting the mechanical properties of zeolite frameworks by machine learning *Chem. Mater.* **29** 7833–9

[17] Sendek A D, Yang Q, Cubuk E D, Duerloo K-A N, Cui Y and Reed E J 2017 Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials *Energy Environ. Sci.* **10** 306–20

[18] Ahmad Z, Xie T, Chinmay Maheshwari, Jeffrey C Viswanathan V 2018 Machine learning enabled computational screening of inorganic solid electrolytes for suppression of dendrite formation in lithium metal anodes *ACS Cent. Sci.* **4** 996–1006

[19] Mazaheri T, Sun B, Scher-Zagier J, Thind A S, Magee D, Ronhovde P, Lookman T, Mishra R and Nussinov Z 2019 Stochastic replica voting machine prediction of stable cubic and double perovskite materials and binary alloys *Phys. Rev. Mater.* **3** 063802

[20] Pilania G, Wang C, Jiang X, Rajasekaran S and Ramprasad R 2013 Accelerating materials property predictions using machine learning *Sci. Rep.* **3** 2810

[21] Fujimura K *et al* 2013 Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms *Adv. Energy Mater.* **3** 980–5

[22] Kim K, Ward L, Jiangang H, Krishna A, Agrawal A and Wolverton C 2018 Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds *Phys. Rev. Mater.* **2** 123801

[23] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301

[24] Schütt K T, Sauceda H E, Kindermans P-J, Tkatchenko A and Müller K-R 2018 SchNet–a deep learning architecture for molecules and materials *J. Chem. Phys.* **148** 241722

[25] Chen C, Weike Y, Zuo Y, Zheng C and Ong S P 2019 Graph networks as a universal machine learning framework for molecules and crystals *Chem. Mater.* **31** 3564–72

[26] Haastrup S *et al* 2018 The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals *2D Mater.* **5** 042002

[27] Kim C, Huan T D, Krishnan S and Ramprasad R 2017 A hybrid organic-inorganic perovskite dataset *Sci. Data* **4** 170057

[28] Castelli I E, Landis D D, Thygesen K S, Dahl Søren, Chorkendorff I, Jaramillo T F and Jacobsen K W 2012 New cubic perovskites for one- and two-photon water splitting using the computational materials repository *Energy Environ. Sci.* **5** 9034–43

[29] Rajan A C, Mishra A, Satsangi S, Vaish R, Mizuseki H, Lee K-R and Singh A K 2018 Machine-learning-assisted accurate band gap predictions of functionalized mxene *Chem. Mater.* **30** 4031–8

[30] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8

[31] Enkovaara J *et al* 2010 Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method *J. Phys. Condens. Matter* **22** 253202

[32] Goedecker S 2004 Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems *J. Chem. Phys.* **120** 9911–17

[33] Murray Eamonn D, Lee K and Langreth D C 2009 Investigation of exchange energy density functional accuracy for interacting molecules *J. Chem. Theory Computat.* **5** 2754–62

[34] Kresse G and Furthmüller J 1996 Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set *Comput. Mater. Sci.* **6** 15–50

[35] Hammer B, Hansen L B and Nrskov J K 1999 Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals *Phys. Rev.* B **59** 7413–21

[36] Back S, Yoon J, Tian N, Zhong W, Tran K and Ulissi Z W 2019 Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts *J. Phys. Chem. Lett.* **10** 4401–8

[37] Meng L, You J and Yang Y 2018 Addressing the stability issue of perovskite solar cells for commercial applications *Nat. Commun.* **9** 1–4

[38] Houchins G and Viswanathan V 2017 Quantifying confidence in density functional theory predictions of magnetic ground states *Phys. Rev.* B **96** 134426

[39] Hao F, Stoumpos C C, Cao D H, Chang R P H and Kanatzidis M G 2014 Lead-free solid-state organic–inorganic halide perovskite solar cells *Nat. Photonics* **8** 489

[40] Noel N K *et al* 2014 Lead-free organic–inorganic tin halide perovskites for photovoltaic applications *Energy Environ. Sci.* **7** 3061–8

[41] Zijun Y, Ladi N H, Shai X, Hao Li, Shen Y and Wang M 2019 Will organic–inorganic hybrid halide lead perovskites be eliminated from optoelectronic applications? *Nanoscale Adv.* **1** 1276–89

[42] Toshniwal A and Kheraj V 2017 Development of organic-inorganic tin halide perovskites: A review *Sol. Energy* **149** 54–9

[43] Nadege Ouedraogo N A, Chen Y, Xiao Y Y, Meng Q, Han C B, Yan H and Zhang Y 2020 Stability of all-inorganic perovskite solar cells *Nano Energy* **67** 104249

[44] Kuleshov V, Fenner N and Ermon S 2018 Accurate uncertainties for deep learning using calibrated regression. *Proc. of the 35th Int. Conf. on Machine Learning* vol 80 pp 2796–804

[45] Tran K, Neiswanger W, Yoon J, Xing E and Ulissi Z W 2020 Methods for comparing uncertainty quantifications for material property predictions *Mach. Learn.: Sci. Technol.* **1** 025006

[46] Crowley J M, Tahir-Kheli J and Goddard W A 2016 Resolution of the band gap prediction problem for materials design *J. Phys. Chem. Lett.* **7** 1198–203

[47] Moussa J E, Schultz P A and Chelikowsky J R 2012 Analysis of the Heyd-Scuseria-Ernzerhof density functional parameter space *J. Chem. Phys.* **136** 204117