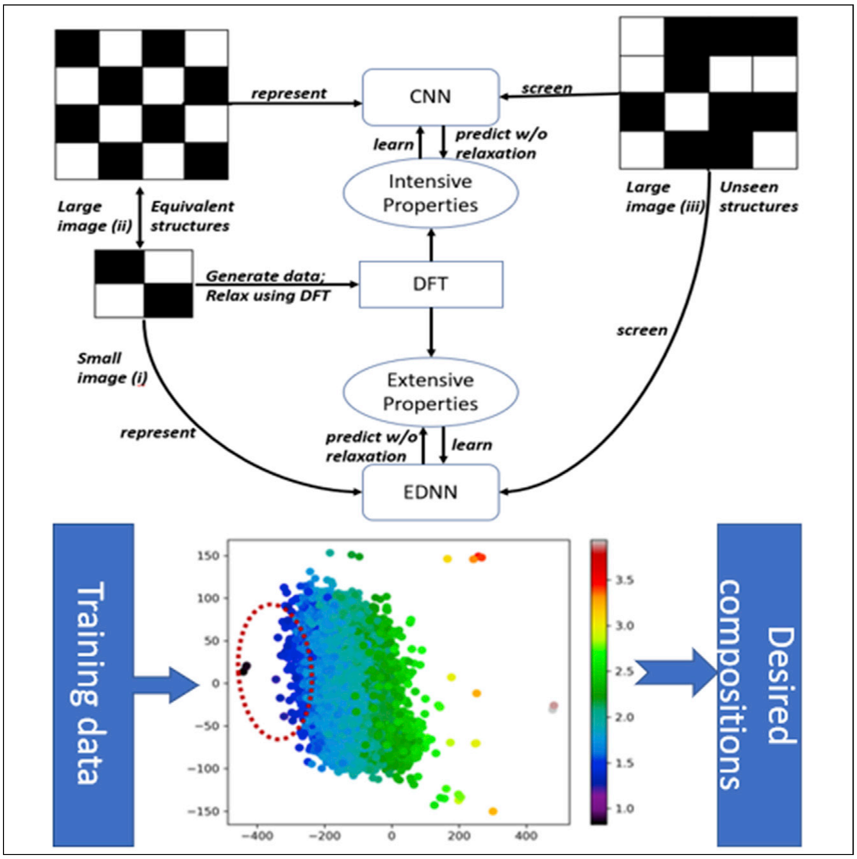
Matter

CellPress

Article

Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces



Here, we report crystal site feature embedding (CSFE), a representation for machine learning of materials that achieves low mean absolute errors for density functional theory band gaps and formation energies. Using CSFE with CNNs and EDNNs, we explored chemical families and doping fractions beyond those present in the training dataset. CSFE allowed us to sample large chemical spaces for materials of interest using autoencoders. We demonstrate the application of the representation by finding perovskite compositions for the ultraviolet and infrared.

Hitarth Choubisa, Mikhail Askerka, Kevin Ryczko, Oleksandr Voznyy, Kyle Mills, Isaac Tamblyn, Edward H. Sargent

isaac.tamblyn@nrc.ca (I.T.)
ted.sargent@utoronto.ca (E.H.S.)

HIGHLIGHTS

The representation enables efficient and targeted exploration of chemical families

It provides exploration of chemical spaces beyond those used in the training process

It leads to autoencoder-powered exploration and sampling of large chemical spaces

3 **Understanding**
Dependency and conditional studies on material behavior

Choubisa et al., Matter 3, 1–16
August 5, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.matt.2020.04.016

Please cite this article in press as: Choubisa et al., Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces, Matter (2020), https://doi.org/10.1016/j.matt.2020.04.016

Matter

CellPress

Article

Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces

Hitarth Choubisa,[1,6] Mikhail Askerka,[1,6] Kevin Ryczko,[2,5] Oleksandr Voznyy,[1] Kyle Mills,[3,5] Isaac Tamblyn,[2,3,4,5,*] and Edward H. Sargent[1,7,*]

SUMMARY

Mapping materials science problems onto computational frameworks suitable for machine learning can accelerate materials discovery. Combining proposed crystal site feature embedding (CSFE) representation with convolutional and extensive deep neural networks, we achieve a low mean absolute test error of 3.7 meV/atom and 0.069 eV on density functional theory energies and band gaps of mixed halide perovskites. We explore how a small amount of cadmium doping can potentially be applied in solar cell design and sample the large chemical space by using a variational autoencoder to discover interesting perovskites with band gaps in the ultraviolet and infrared. Additionally, we use CSFE to explore chemical spaces and small doping concentrations beyond those used for training. We further show that CSFE has a mean absolute test error of 7 meV/atom and 0.13 eV for total energies and band gaps for 2D perovskites and discuss its adaptability for exploration of an even wider variety of chemical systems.

Progress and Potential

Density functional theory (DFT) is of interest in modern-day materials discovery. However, DFT is computationally expensive. Here, we develop a new crystal site feature embedding (CSFE) representation that achieves low error in predicting DFT properties and enables predicting properties of chemical families and doping fractions beyond those present in the training datasets. Using CSFE with autoencoders, we present a scheme that enables sampling of large chemical spaces and offers insight into key semiconductor parameters such as band gap. We demonstrate that CSFE works on both 2D and 3D perovskites and identify promising ultraviolet and infrared candidate materials.

INTRODUCTION

Computational screening of physical and electronic properties of materials with doping or partial site occupations demands the exploration of chemical spaces of billions of compounds with hundreds of atoms. Traditional atomistic simulation methods, such as density functional theory (DFT), are often limited to at most tens of thousands of compounds.[1,2] Machine-learning methods can be utilized to increase further the chemical space that one can explore by learning the properties predicted by DFT. The challenge is to be able to cover as many compounds as possible, therefore providing an acceleration compared with pure DFT while keeping the error on the test data as low as possible (Table 1).

One must design a way to represent an atomic system for the machine learning (ML) algorithm to enable a DFT + ML approach.[10,11] Creating an efficient representation is non-trivial and crucial for the success of the approach.[12] For transferability, previously reported representations can be divided into (1) position dependent, i.e., explicitly depending on the atomic coordinates, and (ii) position independent. Position-dependent representations include sine or Coulomb matrices,[12] bond fractions,[13] total or partial radial distribution functions,[12] orbital field matrix,[14] distance-angular and element type distributions,[6] Voronoi tessellation algorithm,[15,5] and graph-representation-based methods.[9,16,17] These representations usually achieve a high ML accuracy (Table 1D) by reflecting the position-dependent physical properties of molecular systems. Additionally, these representations can be applied to compounds with an arbitrary structure; however, their use is greatly limited by the

Table 1. Performance of ML Algorithms Reported in Literature toward DFT Formation Energies and Band Gaps

	System	N_{train}	No. of Atoms in Cell	ML Performance Energies (meV/Atom)	ML Performance Band Gaps (eV)	Acceleration Factor
A (this study)	halide perovskites	~8,500	40–96 or 320–768	MAE = 3.5; RMSE = 1.5	MAE = 0.069; RMSE = 0.090	10^5[a]
B[3]	$A_2BB'X_6$ perovskites	11,000	20	–	RMSE = 0.37	<100
C[4]	binary inorganics	270	<25	–	RMSE = 0.24	<350
D[5]	OQMD database[6]	230,000	<35	MAE = 88[b]	MAE = 0.065	–
E[7]	ABC_2D_6 elpasolites	12,000	10	MAE = 100[b]	–	200
F[8]	$A_2BB'X_6$ perovskites	640	20	–	MAE ~ 0.1[c]	<100
G[9]	materials project	16,458	<60*	MAE = 72	MAE = 0.38	–

*90% of the data points had fewer than 60 atoms.

[a] Refer to section B of Supplemental Information for discussion on the size of chemical space.

[b] Error for formation energy.

[c] Cross-validation error.

need to know the a priori unknown optimal geometry of a structure or depends on certain parameters such as cutoff radius as in graph-based approaches.[18] This means that if one would like to estimate the total energy of a single unknown compound, sampling hundreds or thousands of positions around a guess structure would be required to find out the optimal geometry before properties corresponding to the compound of interest are found.

Position-independent representations, such as atomic feature vector,[3,7] allow for a significantly more efficient exploration of unknown structures; however, they are generally limited to a particular family of compounds that share the same atomic arrangement (e.g., double cubic perovskites in Table 1B or elpasolites in Table 1E), therefore setting an upper limit on the number of compounds in a particular chemical space.

With these representations, one can define an acceleration factor of the DFT + ML approach as

$$A = \frac{t^{DFT}_{\text{entire chemical space}}}{t^{DFT}_{\text{training set}} + t^{ML}_{\text{train}} + t^{ML}_{\text{screen}}},$$

where t is the time required to calculate a given number of compounds with a given method; this reflects the speed-up due to the addition of ML to the pipeline compared with brute-force DFT. Given that

$$t^{DFT}_{\text{training set}} \gg t^{ML}_{\text{train}} > t^{ML}_{\text{screen}},$$

we can often neglect the time required to train an ML algorithm and to use a trained ML algorithm to scan the chemical space; therefore, the acceleration factor

$$A = \frac{t^{DFT}_{\text{entire chemical space}}}{t^{DFT}_{\text{training set}}}$$

is defined as a ratio of the estimated time required to calculate the compounds from the entire chemical space to the time required to accumulate the DFT training data. In the case of all compounds requiring similar CPU time for DFT calculations, this can be further simplified to the ratio of the screening and the training sets:

$$A = \frac{N_{\text{entire chemical space}}}{N_{\text{training set}}}.$$

[1] Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, ON M5S 3G4, Canada

[2] Department of Physics, University of Ottawa, 598 King Edward, Ottawa, ON K1N 6N5, Canada

[3] Ontario Tech University, Oshawa, ON L1G 0C5, Canada

[4] National Research Council of Canada, Ottawa, ON K1A 0R6, Canada

[5] Vector Institute for Artificial Intelligence, Toronto, ON M5G 1M1, Canada

[6] These authors contributed equally

[7] Lead Contact

*Correspondence: isaac.tamblyn@nrc.ca (I.T.), ted.sargent@utoronto.ca (E.H.S.)

https://doi.org/10.1016/j.matt.2020.04.016

Therefore, the atomistic representations that have the most potential for chemical space exploration are (1) position independent, (2) achieve low mean absolute error (MAE) on the properties of interest, and (3) have a high acceleration factor.

At the same time, different representations are likely to yield different acceleration factors due to two main reasons, (1) ease of exploration using a representation and (2) expressive power of the representation. The first factor depends on computational cost and ease of generating the representation, and the second factor depends on adaptability of the representation to various chemical families.

Recently, a highly accurate position-dependent atomistic representation was reported for two-dimensional (extendable to 3D) materials with the example of mixed graphene and boron nitride sheets.[9] In this representation, the 2D atomistic representation of choice (for instance, Gaussian potential centered around nuclei) is mapped onto a real space grid to form $N \times N \times 1$ grayscale images, where N is the number of pixels along a unit cell dimension and the last dimension denotes the magnitude of the potential. Using this representation, together with convolutional neural networks, Ryczko et al.[19] were able to achieve an impressive MAE of ~0.2 meV/atom for DFT energies. This method was further revised in the extensive deep neural network (EDNN) framework,[20] which allowed for inference at arbitrary system size.

EDNNs[20] are deep-learning networks that can efficiently infer extensive parameters (e.g., energy, entropy) of arbitrary large systems, doing so with O(N) scaling. It uses domain decomposition for training and inference, where each subdomain (tile) is composed of a non-overlapping focus region surrounded by an overlapping context region. The size of these regions is motivated by the physical interaction length scales of the problems (Figure 2).

In this report, we utilize the position-independent mapping. We do so by considering the atomic arrangement instead of explicit atomic positions, therefore reconciling for the first time position-dependent and position-independent atomic representations. In our approach we separate the unit cell into sites, each described with a P feature tensor that reflects calculated or experimentally determined properties for that site. The sites are then mapped onto a 3D grid according to their arrangement within the cell (Figure 1). We call this approach crystal site feature embedding (CSFE). The advantages of CSFE are as follows: (1) it handles a variable number of atoms in a unit cell; (2) the representation is compact: $(n_x) \times (n_y) \times (n_z) \times (n_{\mathrm{properties}})$; (3) the representation does not require the knowledge of relaxed atomic positions and hence enables users to skip the expensive DFT relaxations; (4) the inference can be done on large systems; and (5) it offers an advantage of using image-recognition techniques, such as convolutional neural network (CNN) or EDNN, where the features are learned on-the-fly. Since the representation uses predefined positions, it will not work in the case of extremely large DFT relaxations. In such cases, one should identify the proper starting positions and build the representation based on a close approximation.

In the original EDNN work the authors use a naive pseudopotential mapping.[20] Applying this to the present case of 3D/2D perovskites would yield a dense and computationally expensive-to-train representation. While CSFE has its own advantages as already mentioned, combining CSFE with EDNN not only solves the challenges mentioned above but also enables us to achieve high performance, as reported in this paper (see section C of Supplemental Information for extra discussion).

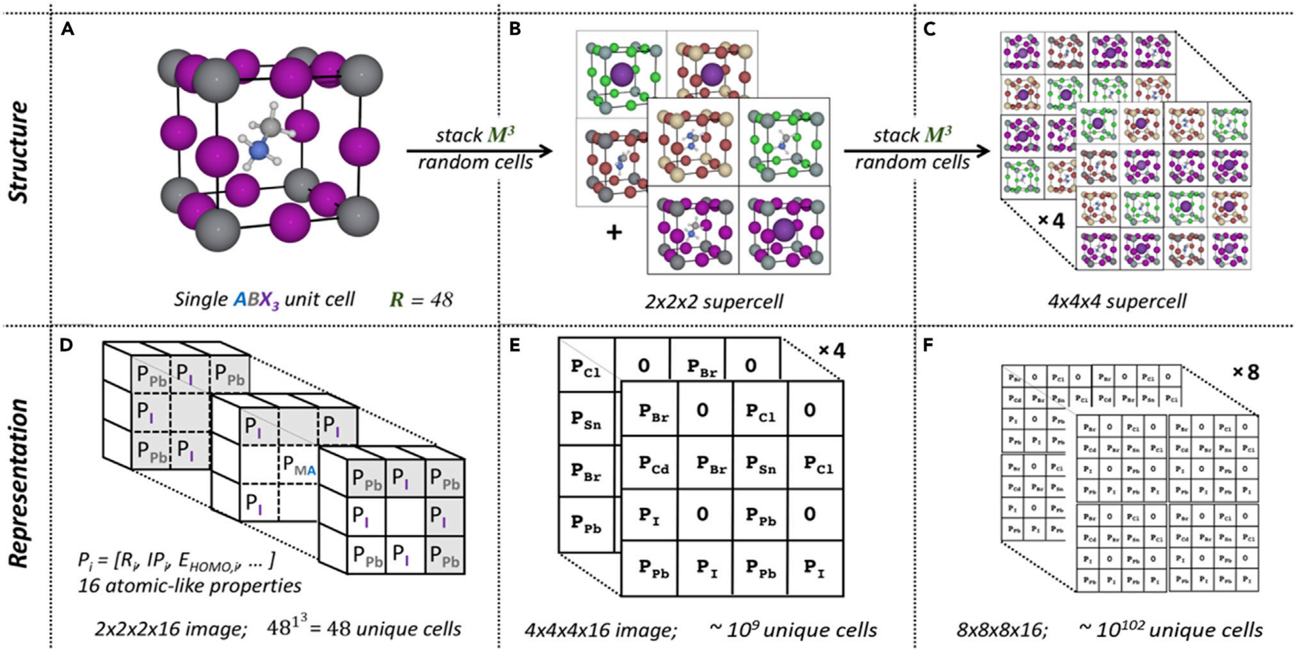

Structure
Representation
A
B
C
D
E
F
stack M^3 random cells
stack M^3 random cells
+
×4
Single ABX_3 unit cell
$R = 48$
2x2x2 supercell
4x4x4 supercell
$P_i = [R_i, IP_i, E_{HOMO,i}, ...]$
16 atomic-like properties
2x2x2x16 image; $48^{1^3} = 48$ unique cells
×4
4x4x4x16 image; ~ 10^9 unique cells
×8
8x8x8x16; ~ 10^{102} unique cells

Figure 1. Illustration of Crystal Site Feature Embedding

General DFT + ML workflow used in this report highlighting the structures (A–C) and representations (D and E) used for DFT calculations, ML training (B and E), and ML screening (C and F). (A) The building blocks for the supercell calculations are single cubic perovskite cells with a lattice constant rescaled to the optimized mean value of 5.90 Å for each of the 48 possible ABX_3 combinations. (D) Each cell is represented with a 2 × 2 × 2 × 16 4D tensor (the 3 × 3 × 3 tensor in A and D is given for illustration; the shaded sites are degenerate by translational symmetry), where the first three dimensions point at the position of a particular site (A, B, or X) and the last dimension has the calculated and experimental properties (P_i) of choice for those sites. (B) Eight random cells are then stacked together to form a 2 × 2 × 2 supercell that is then relaxed with DFT, and the labels of interest are extracted. (E) On the ML side, the 2 × 2 × 2 supercells are represented by 4 × 4 × 4 × 16 tensors. We use about 8,500 PBE calculations to train the algorithm against total energies and about 11,500 GLL-BC + SOC calculations (see Experimental Procedures for detailed description) to train a model against band gaps. (C) A trained ML algorithm is then used to explore The 4 × 4 × 4 ($M = 4$) supercells that are represented (F) by 8 × 8 × 8 × 16 tensors, which allowed for screening of partial perovskite compositions with molar increments of 1/64 = 1.56%. CSFE can be used to explore even larger chemical spaces ($M > 4$) if necessary.

RESULTS

We use halide perovskites to explore the potential of CSFE. Halide perovskites are a family of compounds with a general formula of ABX_3, where A^+ is an alkali (Cs^+ or Rb^+) or an organic (methylammonium, MA^+ or formamidinium, FA^+) monovalent cation; B^{2+} is Pb^{2+}, Sn^{2+}, Cd^{2+} or Ge^{2+} divalent cation; and X^- is a halogen (Cl^-, Br^-, I^-) cation (Supplemental Information sections A, E, and F). Halide perovskites are suited for screening using CSFE because: (1) a large proportion exhibit corner-shared (cubic or partially distorted) 3D structure;[21] (2) by combining different ratios of precursors, the occupations of the sites can be as small as 10%–13%[22,23] or even at the doping level (less than 5%);[24] and (3) the composition changes both the stability[19] and the electronic properties[25–28] of the perovskites.

Using the CSFE representation (refer to section D of Supplemental Information for the features we used for representing elements), we achieve an MAE of 3.5 meV/atom for the perovskite total energies (Figure 3A) using EDNNs and an MAE of 0.069 ± 0.005 eV in 5-fold cross-validation for perovskite band gaps (Figures 3B and S3) (see the corresponding learning curves in Figures S3 and S4). Using the trained ML models, we then explore supercells that are twice as large in each of the x, y, and z directions (Figures 1C and 1F).

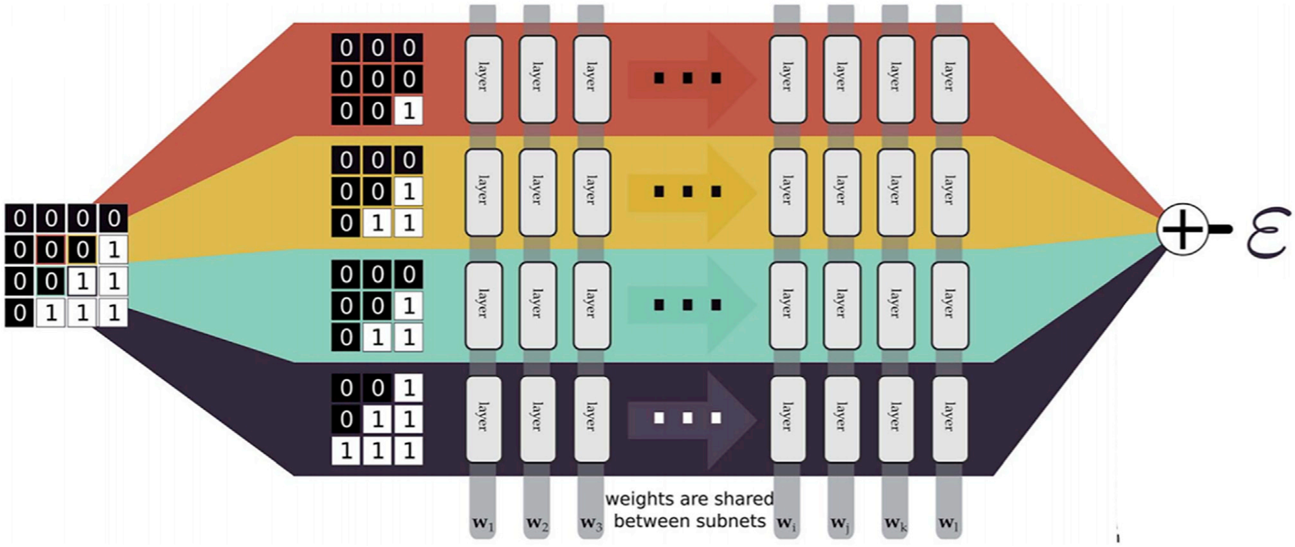


Figure 2. An Input Example Is Decomposed into Four Tiles, with Each Tile Consisting of a Focus and Context Region

Both the focus and the context are unit width, resulting in 3 × 3 tiles. The tiles are simultaneously passed through the same neural network. The individual outputs are summed, producing an estimate of an extensive quantity. Reproduced from Mills et al.,[20] published by The Royal Society of Chemistry.

The chemical space, therefore, increases (refer to section B of Supplemental Information for exact numbers) and allows screening the fractions of ions as small as 1/64 ≈ 1.6%. EDNNs allow for exploration of supercells by the construction of the method as soon as the larger "images" are composed of an integer number of the smaller ones (Figure 4).[20] This applies to any extensive property, and in this report we use this approach for evaluating perovskite total energies. For the band gaps, we propose to take advantage of the Brillouin zone (BZ) folding to artificially increase the chemical space. Indeed, due to the BZ folding, the value of the band gap of an M × M × M supercell with k × k × k-points grid corresponds exactly to the band gap of a 2M × 2M × 2M supercell with k/2 × k/2 × k/2-points grid (Figure S5). At the same time, the former spans a chemical space of R^{M^3} compounds, while the latter spans a chemical space of $R^{(2M)^3}$, which is R^8 times larger (Figure 4). Therefore, one can obtain DFT (or any other reference method) labels in R^{M^3} and use the representation from $R^{(2M)^3}$ space to learn them with CNNs. This allows us to then explore the $R^{(2M)^3}$ space, or more generally $R^{(mM)^3}, m \in N$, at a low cost using CNNs.

Taking advantage of the BZ folding and the ability of EDNNs, we then validate the ability of the ML approach to explore the larger space of halide perovskites. Using two prototypical systems, $MAPb(I_xBr_{(1-x)})_3$ and $MAPb_xSn_{(1-x)}I_3$, we find excellent (within one root-mean-squared error [RMSE]) agreement between the ML predicted and DFT formation energies and band gaps (Figures 3C and 3D). Moreover, this is the first time to our knowledge the U shape of the $MAPb_xSn_{(1-x)}I_3$ band gap was reproduced using ML methods. Since we explore the composition of X or B perovskite sites in fractions of 1.6%, we are also able to capture the variation of the target property due to multiple atomic arrangements that correspond to the same composition. This variation is reflected through the 3σ shaded region in Figure 3 and is particularly significant for the perovskite band gaps. Our algorithm can be confidently used for arbitrary compounds from the $R^{(2M)^3}$ space for the formation of energy predictions (Figure S7); however, more care needs to be taken in the case of band gaps, as the optimal CNN architecture for R^{M^3} space does not necessarily yield

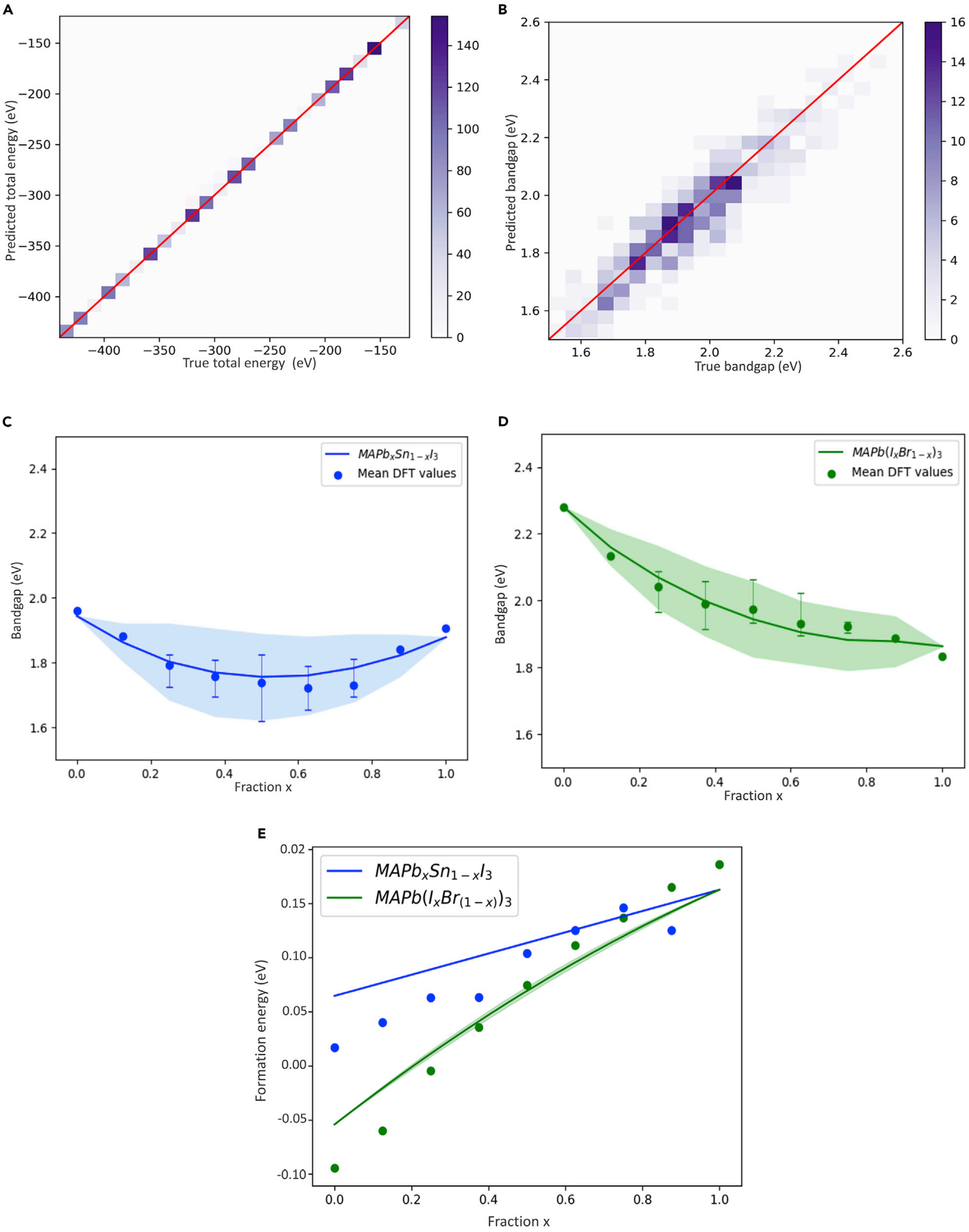

A
Predicted total energy (eV)
−150
−200
−250
−300
−350
−400
−400
−350
−300
−250
−200
−150
True total energy (eV)
140
120
100
80
60
40
20
0
B
Predicted bandgap (eV)
2.6
2.4
2.2
2.0
1.8
1.6
1.6
1.8
2.0
2.2
2.4
2.6
True bandgap (eV)
16
14
12
10
8
6
4
2
0
C
$MAPb_xSn_{1-x}I_3$
Mean DFT values
Bandgap (eV)
2.4
2.2
2.0
1.8
1.6
0.0
0.2
0.4
0.6
0.8
1.0
Fraction x
D
$MAPb(I_xBr_{1-x})_3$
Mean DFT values
Bandgap (eV)
2.4
2.2
2.0
1.8
1.6
0.0
0.2
0.4
0.6
0.8
1.0
Fraction x
E
$MAPb_xSn_{1-x}I_3$
$MAPb(I_xBr_{(1-x)})_3$
Formation energy (eV)
0.02
0.15
0.10
0.05
0.00
-0.05
-0.10
0.0
0.2
0.4
0.6
0.8
1.0
Fraction x

Figure 3. Performance and Benchmarking of ML Models Using CSFE

(A and B) Performance of the ML algorithm on test data for total energies (A) and band gaps (B).

(C–E) DFT benchmarks of the trained ML model. The lines correspond to the mean prediction of direct band gaps (C and D) and formation energies (E) over a 256-random-atom arrangement for a given composition, and the shaded region denotes the $\pm 3\sigma$ (σ : standard deviation) region. The ML inference is done on the chemical space that exceeds the one used for training. Our ML model reproduces the band gaps and formation energies for the $MAPb(I_xBr_{(1-x)})_3$ and $MAPb_xSn_{(1-x)}I_3$. The blue and green dots correspond to the average of the DFT calculated band gaps of ten different arrangements for a given composition with range of DFT values indicated using error bars; see Figure S6 for comparison with experimental data.[25,28] The model reproduces the experimentally observed U shape of the band gap for the $MAPb_xSn_{(1-x)}I_3$ perovskites (see Figures S3–S16 for details on ML models and training).

best results for the $R^{(2M)^3}$ space (Figure S8). We believe that this can be alleviated by using larger datasets for training.

To demonstrate the practicality of our newly developed approach, we apply it to trimetal B-site halide perovskite mixtures. In the design of photovoltaic devices, it is desirable to have the band gap of the active layer in the region of 1.1 eV, while the band gap for halide perovskites is generally higher. We show that by adding Ge or Cd, the band gap of the $MAPb_xSn_{(1-x)}I_3$ system can be further lowered (Figure 5). In the case of Cd, small doping at the level of 5% is expected to be most practical, since a higher percentage of Cd usually leads to the formation of a non-perovskite phase[29] (refer to section G of Supplemental Information for DFT analysis of the reduction in band gap).

To exhaustively search the entire chemical space of the mixed halide perovskites, we trained a variational autoencoder that can reproduce the initial images through a condensed latent space (Figure 6A). In this autoencoder, the initial images are reduced progressively (encoding branch) using reducing convolutions and flattened to form the latent space (hidden dense layer, Figure 6A). This latent space is then used to reconstruct the initial images by applying non-reducing convolutional and upsampling layers (decoding branch). Despite having all the information about the images, however, the latent space constructed using solely the encoding and decoding branches does not have any knowledge about the perovskites' band gap and is therefore inconvenient to sample. To alleviate this, we added a target learning branch that is composed of a series of dense layers and is trained to fit the band gap based on the hidden layer (see performance of the target learning branch compared with the pure band-gap model in Figures S11–S13).

The latent space constructed this way separates data points with different band gaps (Figure 6B). One can, therefore, use a certain region of the latent space (say, the one corresponding to 1.1–1.3 eV) to sample perovskite compositions in the given chemical space and perform inverse design for halide perovskites.

Furthermore, there is interest in perovskite materials that emit in the infrared (IR) (~1 eV) and the UV (~3.2 eV) for IR sensors and UV lasers.[30,31] We screened the space and found stable (negative formation energies) mixed perovskites with band gaps in the UV and IR regions. We then verified these using hybrid HSE06[32,33]-SOC calculations on the candidates. Band gaps calculated using HSE06 functional are shown to be closer to experimental values.[34] Promising materials and their band gaps are presented in Table 2.

We also investigated whether the present ML framework could enable the study of chemical systems beyond the 3D materials explored up to this point. 2D perovskites are composed of a corner-sharing $[MX_6]$ metal halide network sandwiched between organic barriers (Figure 7). We focused on hydrogenated PEA (phenethylamine)

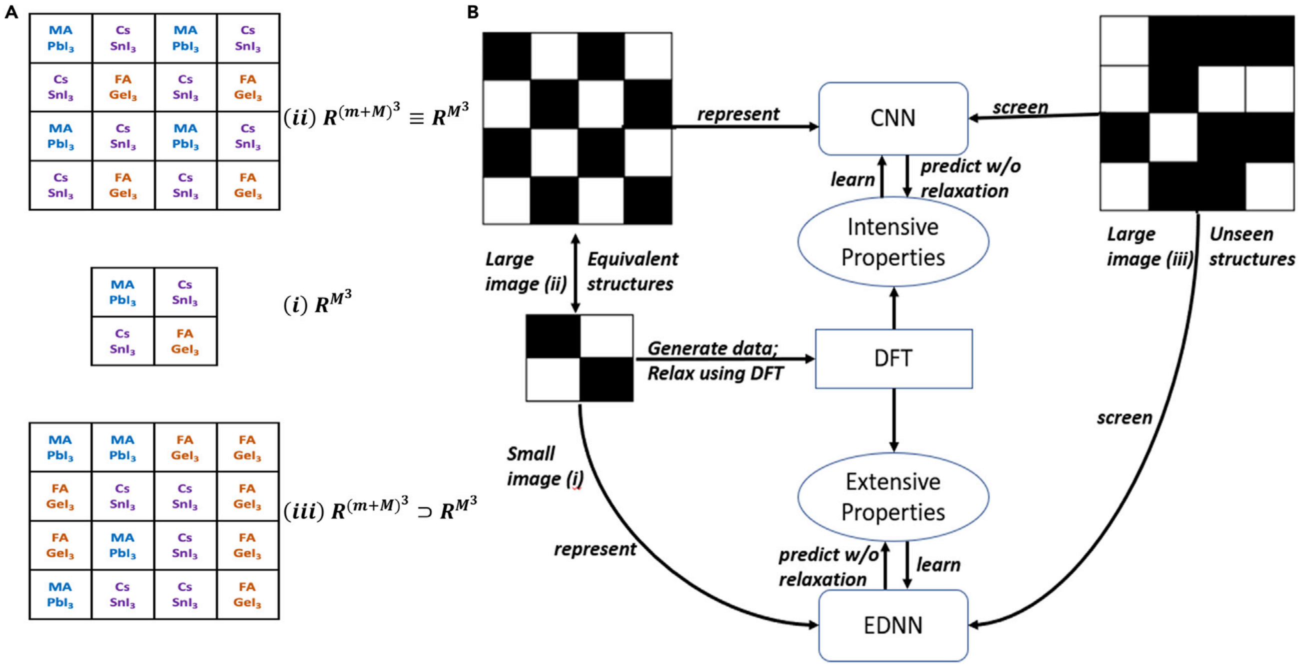


Figure 4. Workflow for Materials Discovery Using CSFE

CSFE reduces the computations at the data generation (DFT) and training stages while simultaneously increasing the size of the search space at the screening stage. A large image representation (ii) can be used to learn the intensive properties of the equivalent small images (i) with CNN and then use large images (iii, $R^{(m\cdot M)^3}$) that do not belong to the original chemical space (R^{M^3}) for screening. EDNN by construction allows the exploration of large images after being trained on the small ones by splitting the images into tiles. CSFE, therefore, allows training in the R^{M^3} space and making inferences in the $R^{(m\cdot M)^3}$ space. In the present work, $m = 2$. This approach was applied to halide perovskites (A, the third dimension is omitted for clarity) but it is also applicable to any periodic systems (B, black and white schematically represent different cell types).

molecules as organic barriers, as these have shown promise in light emission and photovoltaics,[30,35,36] but the associated chemical space is huge, making full DFT exploration impractical.

We substituted B site with the following ions: Cd^{2+}, Ge^{2+}, Pb^{2+}, Sn^{2+}, Ba^{2+}, and halides with Cl^-, Br^-, and I^-. Using DFT-generated data and CSFE representation, we learn band gaps and total energies of the chemical family. Results are shown in Figures 8B and 8C along with band-gap distribution of the materials in Figure 8A. We were able to achieve MAE of 0.008 meV/atom for energies and MAE of 0.13 eV for band-gap prediction. Also, to show that the proposed representation is capable of learning with a different set of descriptors, we train our ML models by one-hot encoding the element type.

In summary, we designed CSFE representation to accelerate the computational exploration of atomistic systems. This representation is position independent, and achieves a low MAE of 3.5 meV/atom for total energies and 0.069 for band gaps. Using EDNNs and BZ folding, we expanded the chemical search space, allowing us to explore doping concentrations as small as 1.6%, which makes it stand out from the previously reported ML approaches. We used it to discover how a small amount of Cd doping can help us move the band gap to 1.1 eV, which is desirable for the active layers of solar cells. By training a CNN-based autoencoder on CSFE, we opened the gates for the generative design of 3D halide perovskites and suggested perovskite compositions with desirable band gaps. We discovered stable IR and UV perovskites, which we then further verified by performing accurate HSE06-SOC

Please cite this article in press as: Choubisa et al., Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces, Matter (2020), https://doi.org/10.1016/j.matt.2020.04.016

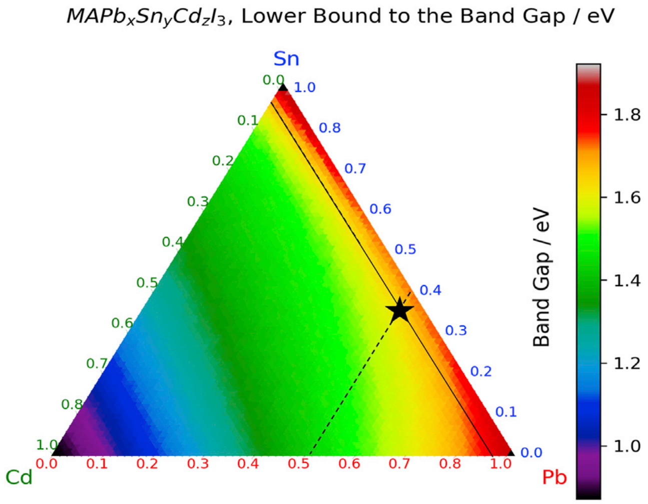


Figure 5. CSFE-Powered Search of Alloyed Perovskite Systems

ML screening of the band gaps of over 4 × 10^5 triple B-site $MAPb_xSn_yCd_zI_3$ perovskite structures shows that the minimum band gap could be achieved by adding the maximum allowed ~5% of Cd (solid line) to a ~50% Pb and ~45% Sn. Each composition on the diagram is sampled by choosing 200 random ion arrangements; the number on the plot corresponds to lower bound, i.e., mean $- 3\sigma$ value (same as lower bound of the shaded region in Figures 3C and 3D). More examples of using CSFE for tri-site halide perovskite compositions are presented in Figure S9, and validation using DFT is in Figure S10.

calculation. Finally, CSFE was applied to 2D perovskites of interest in light emission and photovoltaics.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Edward H. Sargent (ted.sargent@utoronto.ca).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The source code and data information for the CSFE approach is available in Supplemental Information and attached as Data S1.zip. Please refer to section H of Supplemental Information for the instructions on how to execute it.

First-Principles Calculations

We used ground-state DFT with the Perdew-Burke-Ernzerhof (PBE)[22] generalized gradient approximation (GGA) exchange-correlation functional as implemented in the Vienna Ab Initio Simulation Package (VASP)[37] to perform structural optimization of the perovskites. All calculations were non-spin polarized. We used plane wave energy cutoff of 520 eV and Gaussian smearing (0.05 eV wide) to converge the electronic problem. The Monkhorst-Pack k-points mesh of 2 × 2 × 2 (density >900/atom) and the force convergence criterion of 0.0005 eV/atom × N atoms in the unit cell were used as implemented in the MPRelaxSet class of the Pymatgen Python package.[38–40] The atomic-like properties were calculated by placing the isolated A^+, B^{2+}, X^- ions in a 15 × 15 × 15-Å cubic cell. For the features included in the tensor, see Section C of Supplemental Information. The structures of the halide perovskites were prepared using the ASE[41] Python module, and RDKit[42] was used to

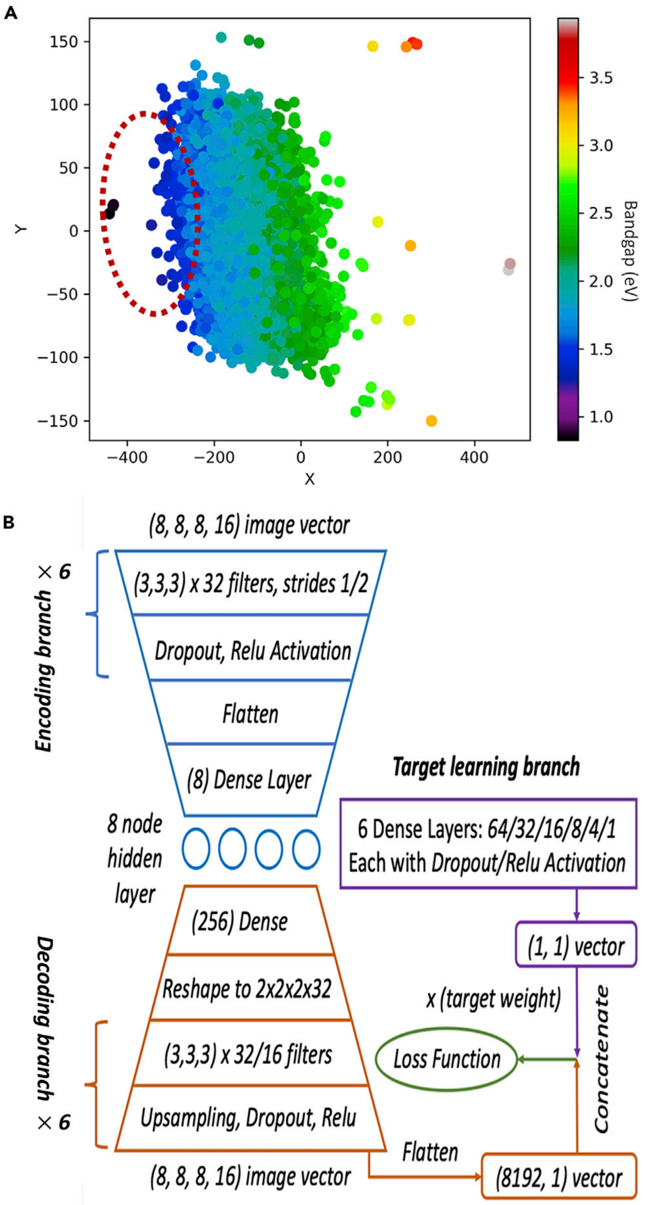


Figure 6. Exhaustive Screening of Chemical Space Using Auto-encoders

(A) The distribution of the data along the two most significant principal components of the hidden layer (denoted by X and Y in the plot) of the autoencoder. The data points are colored according to their band gap.

(B) The architecture of the autoencoder used to exhaustively explore the space of 3D mixed halide perovskites.

hydrogenate (MA, FA) and randomly initialize the dihedral angles (MA) in the organic A^+ cations. The band-gap calculations were performed using the GLL[43]-BC[18] functional with a plane wave basis set (cutoff of 520 eV) as implemented in GPAW[44] electronic structure code. GLL-BC functional partially alleviates the well-known tendency of GGA functionals to underestimate the band gaps by calculating the derivative discontinuity through a response of the hole-to-density variations.[18] Halide perovskite band gaps are also dramatically affected by spin-orbit coupling (reduced by ~0.25 for Sn and ~1.02 for Pb[45]). Spin-orbit coupling was calculated

Table 2. Mixed Halide Perovskite Candidates for IR and UV Emission

Chemical Formula	Band Gap (eV)	Segregated Phase (S) Total Energy (eV)	Unsegregated Phase (U) Total Energy (eV)	Stable Phase
Cs_4Rb_3MA Sn_5Pb_3 $Br_{18}I_6$	0.94	−159.55	−193.90	U
$MAFA_2Cs_5$ Cd_8 I_{24}	0.98	−205.77	−172.04	S
$MAFA_2Cs_3Rb_2$ Cd_6Sn_2 I_{24}	1.20	−211.77	−175.47	S
Cs_8 Pb_5Cd_3 Cl_{24}	3.1	−131.40	−157.86	U
FA_3Cs_4MA Ge_5Cd_3 $Cl_{18}Br_6$	3.08	−288.66	−342.89	U

Three out of our five proposed compositions are more stable in their mixed state than segregated phase.

non-self-consistently through diagonalization of the spin-orbit Hamiltonian on a basis of scalar-relativistic Kohn-Sham eigenstates as implemented in GPAW (Appendix A of Olsen[46]). The approach described above was previously shown to give reasonable estimates to the band gaps of halide perovskites.[45] We also perform an additional study about influence of orientations and reorganizations of organic cations in section F of Supplemental Information.

For performing very accurate estimation of band gaps for our candidate compositions, we used HSE06-SOC[32,33] calculations as implemented in VASP.[37] We used the default value of AEXX = 0.25, which has been shown to yield acceptable results for experimental benchmarking.[47]

For the generation of DFT data for 2D perovskites, we used similar functionals and settings as for the 3D counterpart. All calculations were non-spin polarized and done at gamma point, since each of these structures contains 564 atoms that would otherwise have been quite difficult to perform with a denser k-points mesh.

Supercell Preparation

The single halide perovskite unit cells were prepared by covering all possible combinations of A, B, and X sites, where

- A: Cs^+, Rb^+, MA^+, FA^+,
- B: Pb^{2+}, Sn^{2+}, Cd^{2+}, Ge^{2+},
- X: Cl^-, Br^-, I^-.

This resulted in 48 unique single cells. Each cell was optimized individually with a constraint of maintaining a cubic symmetry, which led to the following distribution of the cell lengths across the 48 points (Figure S1). To be able to construct supercells, we then rescaled the cell length to the mean value of 5.90 Å across the distribution. We then prepared a training set that consisted of 8,500 2 × 2 × 2 supercells by combining eight random single cells. This ensured a uniform distribution of the single cells across the supercells (Figure S2). With the initial structures prepared as mentioned above for all the cases, we relax the supercells while performing DFT calculations to obtain the relevant properties—band gaps and total energies. However, we define our CSFE representation on the unrelaxed structures, which allows us to skip any costly operations associated with DFT relaxation and self-consistent calculations while screening for materials.

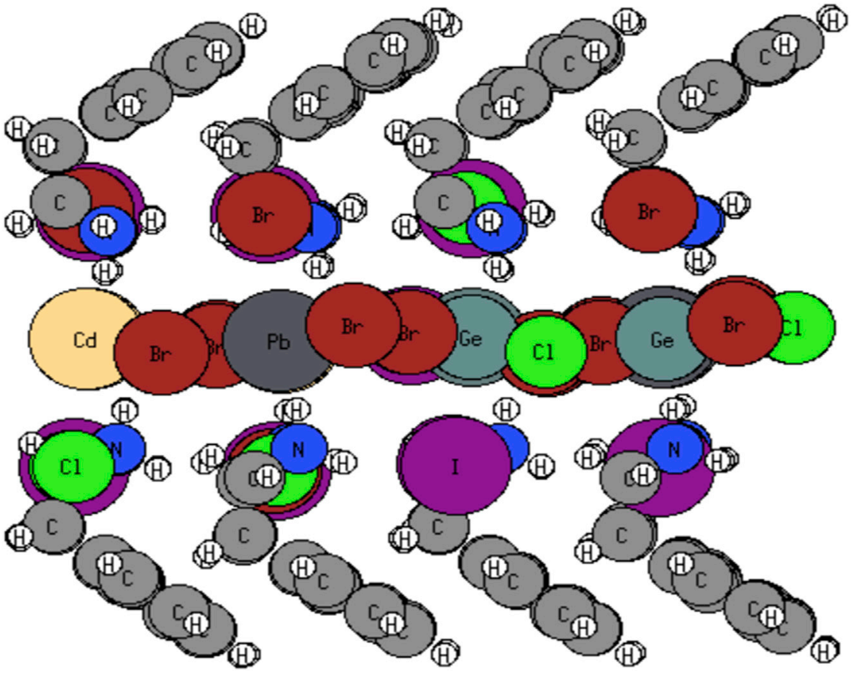


Figure 7. 2D Perovskites
In this example, the chemical formula of the structure is $PEA_{24}Cd_3Ge_3Pb_4Sn_2Cl_{18}Br_{21}I_9$.

The 2D perovskite prototype structure was borrowed from the paper by Yuan et al.[48] The prototype ($PEA_{24}Pb12I_{48}$) is <n = 1> perovskite, more commonly known as 2D perovskite. We randomly substituted Pb^{2+} sites with Ba^{2+}, Sn^{2+}, Ge^{2+}, Pb^{2+}, and Cd^{2+} and halides with Cl^{-}, Br^{-}, and I^{-}. The prototype supercell had 12 B sites and 48 halide sites. The ligand molecule remains the same (PEA) for the chemical family under consideration. Thus, while preparing the representation, we treat it as a constant and it is not considered for representation.

We also calculate formation energies for the compositions with respect to standard states of the constituent elements, consistent with the definition used by the Materials Project.[39]

Deep Learning

We use CNN[19] and EDNN[20] to fit the DFT band gaps and total energies, respectively. We implemented both approaches in the TensorFlow[49]-powered Python module Keras.[50] The previous implementation of both methodologies to atomistic systems was based on a very widely used VGGNet[51] architecture that was successfully used in image classification of the 1000-class ImageNet 2012 database. Following that approach, we optimized the hyperparameters of our VGGNet-like network, namely, the size of the 3D convolutional filters, the number of filters, the presence of a reducing convolutional layer, regularization, dropout, and type of activation function. These parameters were optimized using kopt[52] (forked from the original hyperopt)[36] Python library. Additionally, the EDNN-specific parameters, such as the size of the focus and context regions, were chosen based on the performance on validation data. The features were standardized by moving the distribution to zero mean and scaling it to unit variance before spatial mapping. The results of ML inferences for tri-metal compositions were plotted using the ternary Python

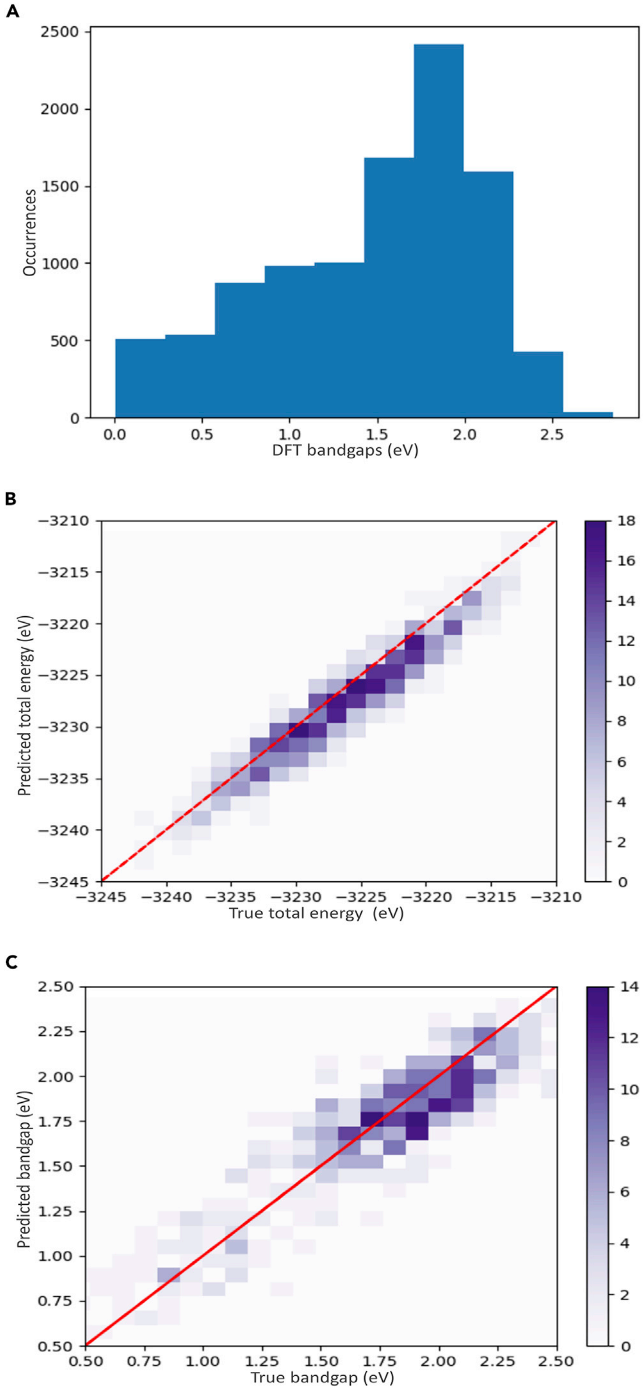


Figure 8. Performance of Trained ML Model for 2D Perovskites Systems

(A) Distribution of band gaps for 2D perovskites. As we can see, there are many compositions in the infrared region.

(B and C) Performance of ML algorithms (B, total energy; C, band gap) on the 2D dataset.

module.[53] For specific neural network architectures used, please refer to Figures 6A and S15–S17.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.matt.2020.04.016.

ACKNOWLEDGMENTS

E.H.S. and all co-authors from the Department of Electrical and Computer Engineering at the University of Toronto acknowledge financial support from the Global Research Outreach program of Samsung Advanced Institute of Technology and the Ontario Research Foundation – Research Excellence Program, and by the Natural Sciences and Engineering Research Council of Canada. DFT computations were performed on the Niagara supercomputer at the SciNet HPC Consortium.[54] SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund – Research Excellence; and the University of Toronto. Training of neural networks was performed on the GPU devices of the SOSCIP (Southern Ontario Smart Computing Innovation Platform, funded by the Federal Economic Development Agency of Southern Ontario, the Province of Ontario, IBM Canada, Ontario Centers of Excellence, Mitacs, and 15 Ontario academic member institutions) GPU accelerate platform. Work at NRC was completed under the auspices of the MCF and AI4D programs.

AUTHOR CONTRIBUTIONS

E.H.S. supervised the physics and chemistry part of the research. I.T. supervised the ML and DFT components of the research. E.H.S. and I.T. supervised the project. H.C. and M.A. wrote the code, performed the calculations, and wrote the manuscript. K.M. provided the starter code for EDNN. O.V., K.R., and K.M. participated in the discussions and provided valuable comments on the manuscript.

Correspondence and requests for additional materials should be addressed to I.T. and E.H.S.

DECLARATION OF INTERESTS

The authors declare no competing financial or non-financial interests.

Received: January 18, 2020
Revised: March 16, 2020
Accepted: April 16, 2020
Published: May 12, 2020

REFERENCES

1. Brandt, R.E., Poindexter, J.R., Gorai, P., Kurchin, R.C., Hoye, R.L.Z., Nienhaus, L., Wilson, M.W.B., Polizzotti, J.A., Sereika, R., Žaltauskas, R., et al. (2017). Searching for “defect-tolerant” photovoltaic materials: combined theoretical and experimental screening. Chem. Mater. *29*, 4667–4674.

2. Sokolov, A.N., Atahan-Evrenk, S., Mondal, R., Akkerman, H.B., Sánchez-Carrera, R.S., Granados-Focil, S., Schrier, J., Mannsfeld, S.C.B., Zoombelt, A.P., Bao, Z., et al. (2011). From computational discovery to experimental characterization of a high hole mobility organic crystal. Nat. Commun. *2*, 1–8.

3. Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B.P., Ramprasad, R., Gubernatis, J.E., and Lookman, T. (2016). Machine learning bandgaps of double perovskites. Sci. Rep. *6*, 1–10.

4. Lee, J., Seko, A., Shitara, K., Nakayama, K., and Tanaka, I. (2016). Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. Phys. Rev. B *93*, 115104.

5. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. Npj Comput. Mater. *2*, 1–7.

6. Faber, F.A., Christensen, A.S., Huang, B., and von Lilienfeld, O.A. (2018). Alchemical and structural distribution based representation for universal quantum machine learning. J. Chem. Phys. *148*, 241717.

7. Faber, F.A., Lindmaa, A., von Lilienfeld, O.A., and Armiento, R. (2016). Machine learning energies of 2 million elpasolite (AB{C}_{2}D_{6}) crystals. Phys. Rev. Lett. *117*, 135502.

8. Pilania, G., Gubernatis, J.E., and Lookman, T. (2017). Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput. Mater. Sci. *129*, 156–163.

9. 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.

10. Back, S., Tran, K., and Ulissi, Z.W. (2019). Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning. ACS Catal. *9*, 7651–7659.

11. Tabor, D.P., Roch, L.M., Saikin, S.K., Kreisbeck, C., Sheberla, D., Montoya, J.H., Dwaraknath, S., Aykol, M., Ortiz, C., Tribukait, H., et al. (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. Nat. Rev. Mater. *3*, 5–20.

12. Schütt, K.T., Glawe, H., Brockherde, F., Sanna, A., Müller, K.R., and Gross, E.K.U. (2014). How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys. Rev. B *89*, 205118.

13. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. (2018). Matminer: an open source toolkit for materials data mining. Comput. Mater. Sci. *152*, 60–69.

14. Pham, T.L., Kino, H., Terakura, K., Miyake, T., Tsuda, K., Takigawa, I., and Dam, H.C. (2017). Machine learning reveals orbital interaction in materials. Sci. Technol. Adv. Mater. *18*, 756–765.

15. Ward, L., Liu, R., Krishna, A., Hegde, V.I., Agrawal, A., Choudhary, A., and Wolverton, C. (2017). Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. Phys. Rev. B *96*, 024104.

16. Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S.P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. *31*, 3564–3572.

17. Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S., and Tropsha, A. (2017). Universal fragment descriptors for predicting properties of inorganic crystals. Nat. Commun. *8*, 1–12.

18. Kuisma, M., Ojanen, J., Enkovaara, J., and Rantala, T.T. (2010). Kohn-Sham potential with discontinuity for band gap materials. Phys. Rev. B *82*, 115106.

19. Ryczko, K., Mills, K., Luchak, I., Homenick, C., and Tamblyn, I. (2018). Convolutional neural networks for atomistic systems. Comput. Mater. Sci. *149*, 134–142.

20. Mills, K., Ryczko, K., Luchak, I., Domurad, A., Beeler, C., and Tamblyn, I. (2019). Extensive deep neural networks for transferring small scale learning to large scale systems. Chem. Sci. *10*, 4129–4140.

21. Kovalenko, M.V., Protesescu, L., and Bodnarchuk, M.I. (2017). Properties and potential optoelectronic applications of lead halide perovskite nanocrystals. Science *358*, 745–750.

22. Zhang, L., Yang, X., Jiang, Q., Wang, P., Yin, Z., Zhang, X., Tan, H., Yang (Michael), Y., Wei, M., Sutherland, B.R., et al. (2017). Ultra-bright and highly efficient inorganic based perovskite light-emitting diodes. Nat. Commun. *8*, 1–8.

23. Zhao, Y., Tan, H., Yuan, H., Yang, Z., Fan, J.Z., Kim, J., Voznyy, O., Gong, X., Quan, L.N., Tan, C.S., et al. (2018). Perovskite seeding growth of formamidinium-lead-iodide-based perovskites for efficient and stable solar cells. Nat. Commun. *9*, 1–10.

24. Zheng, X., Chen, B., Dai, J., Fang, Y., Bai, Y., Lin, Y., Wei, H., Zeng, X.C., and Huang, J. (2017). Defect passivation in hybrid perovskite solar cells using quaternary ammonium halide anions and cations. Nat. Energy *2*, 1–9.

25. Noh, J.H., Im, S.H., Heo, J.H., Mandal, T.N., and Seok, S.I. (2013). Chemical management for colorful, efficient, and stable inorganic-organic hybrid nanostructured solar cells. Nano Lett. *13*, 1764–1769.

26. Eperon, G.E., Leijtens, T., Bush, K.A., Prasanna, R., Green, T., Wang, J.T.-W., McMeekin, D.P., Volonakis, G., Milot, R.L., May, R., et al. (2016). Perovskite-perovskite tandem photovoltaics with optimized band gaps. Science *354*, 861–865.

27. Prasanna, R., Gold-Parker, A., Leijtens, T., Conings, B., Babayigit, A., Boyen, H.-G., Toney, M.F., and McGehee, M.D. (2017). Band gap tuning via lattice contraction and octahedral tilting in perovskite materials for photovoltaics. J. Am. Chem. Soc. *139*, 11117–11124.

28. Hao, F., Stoumpos, C.C., Chang, R.P.H., and Kanatzidis, M.G. (2014). Anomalous band gap behavior in mixed Sn and Pb perovskites enables broadening of absorption spectrum in solar cells. J. Am. Chem. Soc. *136*, 8094–8099.

29. Dunlap-Shohl, W.A., Younts, R., Gautam, B., Gundogdu, K., and Mitzi, D.B. (2016). Effects of Cd diffusion and doping in high-performance perovskite solar cells using CdS as electron transport layer. J. Phys. Chem. C *120*, 16437–16445.

30. Ge, C., Zhai, W., Tian, C., Zhao, S., Guo, T., Sun, S., Chen, W., and Ran, G. (2019). Centimeter-scale 2D perovskite $(PEA)_2PbBr_4$ single crystal plates grown by a seeded solution method for photodetectors. RSC Adv. *9*, 16779–16783.

31. Suzuki, H. (2004). Organic infrared and near-infrared light-emitting materials and devices for optical communication applications. In Organic Photonic Materials and Devices VI, J.G. Grote and T. Kaino, eds. (International Society for Optics and Photonics), pp. 196–209.

32. Heyd, J., Scuseria, G.E., and Ernzerhof, M. (2003). Hybrid functionals based on a screened Coulomb potential. J. Chem. Phys. *118*, 8207–8215.

33. Heyd, J., and Scuseria, G.E. (2004). Efficient hybrid density functional calculations in solids: assessment of the Heyd-Scuseria-Ernzerhof screened Coulomb hybrid functional. J. Chem. Phys. *121*, 1187–1192.

34. Garza, A.J., and Scuseria, G.E. (2016). Predicting band gaps with hybrid density functionals. J. Phys. Chem. Lett. *7*, 4165–4170.

35. Zhang, Y., Liu, Y., Xu, Z., Ye, H., Li, Q., Hu, M., Yang, Z., and Shengzhong, L. (2019). Two-dimensional $(PEA)_2PbBr_4$ perovskite single crystals for a high performance UV-detector. J. Mater. Chem. C *7*, 1584–1591.

36. Lee, J.-W., Dai, Z., Han, T.-H., Choi, C., Chang, S.-Y., Lee, S.-J., De Marco, N., Zhao, H., Sun, P., Huang, Y., et al. (2018). 2D perovskite stabilized phase-pure formamidinium perovskite solar cells. Nat. Commun. *9*, 1–10.

37. Kresse, G., and Furthmüller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys. Rev. B *54*, 11169–11186.

38. Monkhorst, H.J., and Pack, J.D. (1976). Special points for Brillouin-zone integrations. Phys. Rev. B *13*, 5188–5192.

39. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater. *1*, 011002.

40. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. Comput. Mater. Sci. *68*, 314–319.

41. Larsen, A.H., Mortensen, J.J., Blomqvist, J., Castelli, I.E., Christensen, R., Du\lak, M., Friis, J., Groves, M.N., Hammer, B., Hargus, C., et al. (2017). The atomic simulation environment—a Python library for working with atoms. J. Phys. Condens. Matter *29*, 273002.

42. RDKit: Open-Source Cheminformatics Software.https://www.rdkit.org/

43. Gritsenko, O., van Leeuwen, R., van Lenthe, E., and Baerends, E.J. (1995). Self-consistent approximation to the Kohn-Sham exchange potential. Phys. Rev. A. *51*, 1944–1954.

44. Enkovaara, J., Rostgaard, C., Mortensen, J.J., Chen, J., Du\lak, M., Ferrighi, L., Gavnholt, J., Glinsvad, C., Haikola, V., Hansen, H.A., et al. (2010). Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. J. Phys. Condens. Matter *22*, 253202.

45. Castelli, I.E., García-Lastra, J.M., Thygesen, K.S., and Jacobsen, K.W. (2014). Bandgap calculations and trends of organometal halide perovskites. APL Mater. *2*, 081514.

46. Olsen, T. (2016). Designing in-plane heterostructures of quantum spin Hall insulators from first principles: 1T′-MoS_2 with adsorbates. Phys. Rev. B *94*, 235106.

Please cite this article in press as: Choubisa et al., Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces, Matter (2020), https://doi.org/10.1016/j.matt.2020.04.016

47. Ghosh, B., Chakraborty, S., Wei, H., Guet, C., Li, S., Mhaisalkar, S., and Mathews, N. (2017). Poor photovoltaic performance of $Cs_3Bi_2I_9$: an insight through first-principles calculations. J. Phys. Chem. C *121*, 17062–17067.

48. Yuan, M., Quan, L.N., Comin, R., Walters, G., Sabatini, R., Voznyy, O., Hoogland, S., Zhao, Y., Beauregard, E.M., Kanjanaboos, P., et al. (2016). Perovskite energy funnels for efficient light-emitting diodes. Nat. Nanotechnol. *11*, 872–877.

49. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: a system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283.

50. Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

51. Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. ArXiv, 1409.1556.

52. Kopt: Python module, https://github.com/Avsecz/kopt

53. Marc Harper et al. (2015). python-ternary: Ternary Plots in Python. Zenodo. https://doi.org/10.5281/zenodo.34938

54. Loken, C., Gruner, D., Groer, L., Peltier, R., Bunn, N., Craig, M., Henriques, T., Dempsey, J., Yu, C.-H., Chen, J., et al. (2010). SciNet: lessons learned from building a power-efficient top-20 system and data centre. J. Phys. Conf. Ser. *256*, 012026.