Combining density functional theory and machine learning for optimization of multicomponent oxide electrocatalysts

by

Jessica Karaguesian

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Masters of Science in Computational Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author . . . .
Center for Computational Science and Engineering
August 22, 2022

Certified by . . . .
Rafael Gómez-Bombarelli
Assistant Professor of Materials Science
Thesis Supervisor

Accepted by . . . .
Youssef M. Marzouk
Professor of Aeronautics and Astronautics
Co-Director, Department of Computational Science and Engineering

Combining density functional theory and machine learning for optimization of multicomponent oxide electrocatalysts

by

Jessica Karaguesian

Submitted to the Center for Computational Science and Engineering
on August 22, 2022, in partial fulfillment of the
requirements for the degree of
Masters of Science in Computational Science and Engineering

Abstract

Multicomponent metal oxides, such as perovskite oxides, hold promise for use as sustainable alternatives to Ir-, Ru-, and Pt-based electrocatalysts at scale. Perovskites can accommodate a wide variety of elements in their A- and B-sites, enabling tuning of their structural and electronic properties through compositional alloying. These properties, which are obtainable from density functional theory (DFT) calculations, can be used as low-dimensional descriptors that correlate with experimental stability and activity in, for example, the oxygen evolution reaction (OER). Established descriptors of stability include energy above convex hull and energy above Pourbaix hull, while those for catalytic activity include oxygen 2p- and B-site metal d-band centers, for example. We are therefore presented with a combinatorial problem of determining which A- and B-site compositions optimize such descriptors. The compositional search space of $A_xA'_{1-x}B_yB'_{1-y}O_3$ perovskites with up to two different elements in A- and B-sites is at least $O(10^6)$, making it intractable to calculate descriptors exhaustively using DFT. We therefore combine high-throughput DFT calculations with crystal-based graph neural networks to screen multicomponent perovskites.

Using a high-throughput virtual screening platform, a DFT-simulated dataset of over 5,000 multicomponent perovskites was generated, with varied A- and B-site alloying ratios and over 3,000 unique cationic combinations. Leveraging this dataset, alongside calculations available in the literature, graph convolutional neural networks (GNNs) were trained to predict the aforementioned crystal descriptors from unrelaxed cubic structures and used to predict descriptors for $O(10^6)$ $A_xA'_{1-x}B_yB'_{1-y}O_3$ perovskites. GNNs were also combined with baseline estimates of multicomponent perovskite properties calculated as interpolations of constituent ABO_3 perovskites, thereby achieving improved model performance. Moreover, impacts of varied cationic ordering were modelled, showing that different decorations of cations within the perovskite lattice can modulate resulting properties to the same degree as—or more than—varying compositional ratios. Equivariant message passing neural networks were thus implemented to achieve cation decoration-aware property predictions. Lastly, GNNs predicting per-site properties were established, encoding local chemical environments to provide physical insights about each atom in a crystal lattice.

The presented work provides the community with a benchmark multicomponent perovskite dataset, improved machine learning models, and physical insights to be used in further studies of alloyed perovskites, and thus lays groundwork for improved design of multicomponent oxide electrocatalysts.

Thesis Supervisor: Rafael Gómez-Bombarelli
Title: Assistant Professor of Materials Science

Acknowledgments

First and foremost, I would like to express my deepest appreciation for Prof. Rafael Gómez-Bombarelli, who gave me the opportunity to explore an entirely new field of scientific research. His unparalleled support, mentorship, and enthusiasm were integral to my graduate experience at MIT. I also am thankful for the research culture he cultivated—one of curiosity, collaboration and camaraderie.

I am extremely thankful to all group members, whose professional and personal guidance was invaluable throughout my time at MIT. I will cherish these strong friendships long after my departure. To my office mates, thank you for fostering an environment full of laughter.

I am particularly beholden to James Damewood, Jackie Lunger, and Daniel Schwalbe-Koda, with whom the computational aspects of this large-scale collaborative electrocatalysis project were conducted.

Thank you to James for continually sharing his vast scientific knowledge with me, for our fun brainstorming sessions, and for his thoughtful graduate school advice. The studies presented in Sections 3.2 and 3.3 were conducted in complete collaboration with James.

Thank you to Jackie for her incredible partnership as we learned alongside one another—our complimentary research styles and lighthearted friendship made the experience absolutely delightful. The studies presented in Sections 3.4 were conducted in complete collaboration with Jackie.

Thank you to Daniel for helping me find my footing upon first joining the group. I appreciate his mentorship and patience as I navigated unknown scientific territory. I would also like to acknowledge his previous work setting up DFT functionalities for perovskite systems.

I would also like to thank Simon Axelrod and Gavin Winter, who had contributed to the PAINN codebase upon which my work was built.

Lastly, thank you to my family and friends—both in Cambridge and afar—for their continued kindess and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Successful adoption of sustainable energy technologies hinges on the development of improved energy storage systems to ensure energy from intermittent natural sources can be accessed on demand [Fabbri and Schmidt, 2018]. Naturally-derived electricity—such as solar—can be stored in chemical bonds by electrochemically splitting water into hydrogen and oxygen. The oxygen evolution reaction at the anode is a bottleneck in the process, however, due to large overpotentials and slow kinetics [Fabbri and Schmidt, 2018, Guo et al., 2020, Rossmeisl et al., 2007]. The development of efficient OER electrocatalysts is thus critical to developments in sustainable energy technology. Use of current best-performing Ru- or Ir-based oxide catalysts is not feasible at-scale due to the scarcity, high-cost, and moderate stability of these materials [Vazhayil et al., 2021]. Non-platinum group metal electrocatalysts are therefore desired. Multicomponent oxides, such as perovskite oxides, have been identified as promising candidates due to their low-cost, availability, and tunable properties [Beall et al., 2021].

1.1 Multicomponent perovskites as electrocatalysts

Perovskites oxides—with formula ABO_3—are a class of compounds with crystal structures analogous to that of $CaTiO_3$. The A- and B-atoms are 6- and 12-fold oxygen coordinated cations, respectively, as shown in Figure 1-1. The promise of perovskites

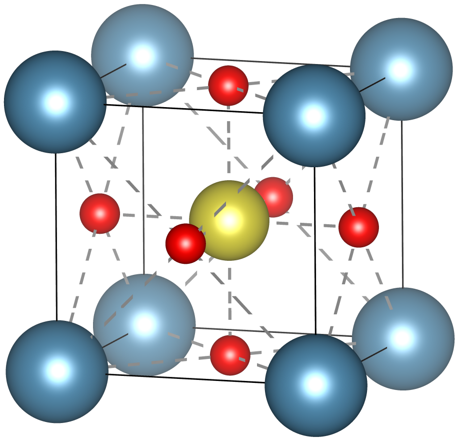

Figure 1-1: Perovskite structure. The structure of one 5-atom perovskite unit cell is shown with A-sites, B-sites, and oxygen atoms coloured in blue, yellow, and red, respectively.

as electrocatalysts stems from their ability to accommodate a wide variety of cations in their A- and B-sites. Moreover, these sites can be alloyed, forming what will be referred to here as multicomponent perovskite oxides: $AA'BO_3$, $ABB'O_3$, $AA'BB'O_3$, etc. By varying A- and B-site compositions, the properties of such pervoskites are tunable, thus presenting a combinatorial problem of determining which cationic compositions optimize electrocatalytic performance [Beall et al., 2021].

1.2 Descriptors of catalytic stability and activity

To accelerate the discovery of new functional materials, high-throughput virtual screening (HTVS) can be employed to model material behaviour *in silico* prior to experimental testing of promising candidates [Emery et al., 2016, Emery and Wolverton, 2017, Castelli et al., 2012b, Jacobs et al., 2018]. For heterocatalysts such as perovskites, density functional theory (DFT) simulations can used to calculate a variety of structural and electronic properties that have been mapped to catalytic behaviour, yielding quantitative structure-activity relationships [Liao et al., 2022]. In theory, improved perovskite electrocatalysts can thus be designed by tuning cationic compositions to obtain DFT-calculated descriptors indicative of desired catalytic performance [Lee et al., 2011, Jacobs et al., 2019]. Established descriptors of catalytic stability include energy above convex hull (ΔE_{hull}) and energy above Pourbaix hull (ΔG_{pbx}) [Bartel, 2022, Shinde et al., 2017, Singh et al., 2017]. Catalytic activity descriptors, meanwhile, can include oxygen $2p$-band centers ($O_{2p} - E_V$) and B-site metal d-band centers ($B_d - E_V$)—as well as their difference, $B_d - O_{2p}$ [Grimaud et al., 2013, Lee et al., 2020, Jacobs et al., 2018, Mueller et al., 2015, Hong et al., 2017].

1.2.1 Energy above convex hull

Energy above the convex hull provides a measure of thermodynamic stability, denoting the energy needed to form a given phase—here, the perovskite phase ABO_3—from the lowest-energy phase(s) with the same overall composition. Energies above hull are calculated as the difference between a material's energy per atom and the energy at its compositional makeup in the associated convex hull phase diagram (PD) [Bartel, 2022].

In theory, compositional PDs present the ground state polyform(s) at all possible ratios of constituent species. The ternary PD of A, B, O, for example, is a Gibbs triangle with each pure component—A, B, and O—on a vertex, all binary combinations of the constituent components on its edges (e.g. $A_{0.5}B_{0.5}$ is equidistant on the edge between A and B vertices), and all linear combinations of the three components in the intervening space [Ong et al., 2013, Jain et al., 2013]. Each point on the PD is associated with the lowest energy polyforms(s) of the given composition, yielding an energetic convex hull because $E(PD[A, B, 3O]) \leq E(A) + E(B) + 3E(O)$ by definition.

The energetic distance of a material above this convex hull at its compositional makeup is thus its ΔE_{hull}. For a perovskite with N atoms,

$$\Delta E_{hull}(ABO_3) = E_{DFT}(ABO_3) - E_{DFT}(PD[A, B, 3O]), \tag{1.1}$$

where $E_{DFT}(ABO_3)$ is the energy per atom of the ABO_3 perovskite and $E_{DFT}(PD[A, B, 3O])$ is the energy per atom of the lowest energy material(s) with overall composition of $A + B + 3O$. For stable materials, $\Delta E_{hull} = 0$, indicating that the material is the lowest energy phase at the given composition—i.e. if $\Delta E_{hull}(ABO_3) = 0$ then $PD[A, B, 3O] = ABO_3$. Meanwhile if $\Delta E_{hull} > 0$, it is thermodynamically favourable for the material to decompose into a different polymorph or a linear combination of its components [Bartel, 2022]. For instance, if the lowest energy material system with overall composition $A + B + 3O$ is $2AO + 2BO_2$ then the decomposition $2ABO_3 \rightarrow 2AO + 2BO_2$ is thermodynamically favoured and $\Delta E_{hull}(ABO_3) = E_{DFT}(ABO_3) - 2E_{DFT}(AO) - 2E_{DFT}(BO_2)$.

Energy above convex hull is therefore an important descriptor of catalytic stability. Optimizing ΔE_{hull} to be close to 0 is useful in screening for materials likely to be thermodynamically stable [Li et al., 2018].

1.2.2 Energy above Pourbaix hull

Energy above Pourbaix hull provides an analogous measure of aqueous stability, at a given pH and potential [Singh et al., 2017]. ΔG_{pbx} values are calculated as the Gibbs free energy difference between a material and the electrochemical equilibrium combination of possible decomposition products [Singh et al., 2017, Persson et al., 2012]. In general, it has been shown that materials with $\Delta G_{pbx} < 0.5$ eV/atom can be stable against corrosion [Singh et al., 2017].

Pourbaix diagrams present the aqueous phase electrochemical equilibria—water-stable phase(s)—across a range of potentials and pH values. In the Materials Project, these diagrams are computed from the free energies of solid phases and of aqueous ions [Jain et al., 2013].

1.2.3 Electronic density of states band centers

Studies have shown that perovskites exhibit scaling relations between catalytic surface kinetics and electronic density of states (DOS) band centers [Lee et al., 2011, Grimaud et al., 2013]. $O_{2p} - E_V$ is the canonical descriptor of catalytic activity with higher values having been correlated with decreased overpotentials and oxygen migration barriers, for example [Grimaud et al., 2013, Mayeshiba and Morgan, 2016]. Although

$O_{2p} - E_V$ is the most commonly used band center descriptor of activity, correlations between other band center-derived descriptors and, for example, current density have also been reported [Hong et al., 2017]. Therefore, tailoring band center values—namely, maximizing O_{2p} in stable perovskites—may be used to optimize catalytic performance.

1.3 Data-driven property prediction

Although DFT calculations make screening thousands of perovskite oxides feasible, allowing us to identify those with promising descriptor values, such calculations become intractable for large material search spaces. The DFT calculations used to simulate $AA'BB'O_3$ perovskite structures in this work, described in Section 2.1, have runtimes on the order of days. Considering 20 A- and 20 B-site atoms, there are $O(10^6)$ $A_xA'_{1-x}B_yB'_{1-y}O_3$ structures, making it infeasible to exhaustively compute descriptors with DFT. Moreover, this search space becomes significantly larger if varied cationic arrangements within the lattice are considered—as will be described in Section 3.3.

To overcome the need for expensive DFT calculations, machine learning techniques have been increasingly implemented to predict descriptors of catalytic stability and activity [Xie and Grossman, 2018, Li et al., 2018, Tao et al., 2021].

1.4 Per-site properties

The bulk descriptors described above are very useful to guide compositional tuning towards optimized catalytic performance. Nevertheless, they have limitations in that they do not capture atomic level structure-function relationships. Catalysis occurs on individual active sites, and thus per-site properties influence catalytic activity. It may therefore be useful to also model site-level descriptors based on local electronic structure (Bader charges, site-projected O2p- and d-band centers), local magnetic structure, and local vibrational structure (site-projected phonon band centers). Like the descriptors outlined above, these properties can be calculated from DFT simulations and partitioned into per-site contributions.

This thesis work builds upon perovskite datasets and property predictions models from the literature to (a) generate a broad set of DFT-calculated multicomponent perovskite data; (b) predict catalytic descriptors from unrelaxed perovskites using graph neural networks, mitigating the need for DFT calculations; (c) screen $O(10^6)$ multicomponent catalysts, identifying promising candidates to be tested experimentally; (d) explore and model cationic decorations in perovskite lattices, and their impacts on structural and electronic properties; and (e) extend bulk crystal graph neural networks to predict per-site properties. Together, the presented work sets the stage for subsequent inverse design of multicomponent perovskites with optimized properties.

Chapter 2

Methodology

2.1 Density functional theory

Density functional theory (DFT) is a quantum mechanical method used to calculated structural and electronic properties of many-electron systems using functionals of spatially-dependent electron density [Hafner, 2008]. The ability to predict properties of crystalline solids from first principles enables us to simulate material behaviour *in silico*, providing fundamental physical insights and reducing experimental demands. Indeed, it is feasible to screen thousands of structures in a high-throughput manner using DFT to identify promising candidates to test experimentally [Kirklin et al., 2015].

2.1.1 High-throughput DFT workflow

Our high-throughput workflow to calculate electronic properties from first-principles consisted of two DFT calculation types, 1) structure optimization and 2) electronic structure calculation. All calculations were conducted with parameters compatible with the Materials Project [Jain et al., 2013, 2011], enabling us to leverage this expansive database to calculate properties such as energy above hull—as described below. Previous benchmarking comparisons between our DFT calculations and Materials Project-derived data confirmed this compatability. Simulations were conducted using the Vienna Ab Initio Simulation Packaged (VASP) [Kresse and Hafner, 1993, Kresse et al., 1994, Kresse and Furthmüller, 1996a,b]. The Projector Augmented Wave (PAW) approach was used to describe core electrons and Perdew-Burke-Enzerhof (PBE) PAW pseudopotentials were implemented. A PBE Generalized Gradient Approximation (GGA) exchange-correlation functional was used, with or without the +U correction following the Materials Project conventions [Kresse and Joubert, 1999, Jain et al., 2013]. All DFT calculations were performed at 0K and 0atm with spin polarization and the Materials Projects high-spin ferromagnetic default initializations [Jain et al., 2013].

Initial, unrelaxed perovskite structures for input to DFT simulations were generated using Atomic Simulation Environment (ASE) and Python Materials Genomics (pymatgen) [Larsen et al., 2017, Ong et al., 2013]. To accommodate varied cationic alloying in the A- and B-sites, cubic 2x2x2 $A_8B_8O_{24}$ supercells were used to initialize most DFT calculations. Calculations for reference binary ABO_3 perovskites, however, were initialized with a single unit cell, 1x1x1 supercell, for consistency with literature data. Using modules from the ASE and pymatgen packages, a workflow for unique structure generation was also developed. Given the composition of a multicomponent perovskite (or other mulitcomponent oxide), this functionality generates all possible symmetrically-inequivalent crystals structures.

2.1.2 DFT-derived descriptors

Many DFT-calculated properties can provide insights into the relationship between material structures and electrochemical performance [Liao et al., 2022]. In this work, energies above the convex hull (ΔE_{hull}) and above the Pourbaix hull (ΔG_{pbx}) were used has descriptors of thermodynamic and aqueous stability, while electronic denstiy of state (DOS) band centers were used as descriptors of catalytic activity.

Energy above convex hull

To calculate its energy above the convex hull, a material's energy per atom is compared to all other polyform(s) with the same overall composition. The accuracy of ΔE_{hull} calculations thus depends largely on the dataset with which the convex hull is generated. Though it is not feasible in practice to construct the hull from *all* possible polyform(s), having as many as possible provides improved approximations of the ground state at a given composition. To expand the dataset on which convex hulls were generated, we thus extracted all metal oxide structures from the Materials Project [Jain et al., 2013]. Leveraging the compatibility of our DFT calculations with the Materials Project, ΔE_{hull} values were then calculated from phase diagrams built upon the oxides dataset and our perovskite dataset (from both in-house and literature-derived calculations) using the pymatgen phase diagram module [Ong et al., 2013].

Energy above Pourbaix hull

Analogously to ΔE_{hull}, energies above Pourbaix hull were computed as the free energy difference between a material and the electrochemical equilibrium combination of decomposition products with the same overall composition [Jain et al., 2013]. The accuracy of ΔG_{pbx} thus also depends on the dataset upon which the Pourbaix hull is generated. To maximize the amount of data used to construct the hull, we therefore employed the Pourbaix modules in pymatgen, which can draw on all Materials Project data [Ong et al., 2013, Jain et al., 2013]. ΔG_{pbx} values were then calculated by, again, leveraging the compatibility of our DFT calculations with the Materials Project. ΔG_{pbx} is computed at a given pH and potential—here, we considered pH 13.5 and 1.6 V vs. RHE to reflect conditions for alkaline OER.

Density of state band centers

Electronic densities of state, as well as the Fermi level, can be obtained from electronic structure DFT calculations. Using these results, we calculate band centers as

$$\text{Band center} = \frac{\int_E D(E)EdE}{\int_E D(E)dE} - E_V, \tag{2.1}$$

where $D(E)$ is the density of states at energy E and E_V is the Fermi level. $O_{2p} - E_V$ and $B_d - E_V$ were calculated as the band centers of the p DOS and d DOS combined over all oxygen atoms and B-site metal atoms, respectively.

Per-site O_{2p} and B_d band centers, as well as per-site phonon band centers, were computed as the band centers of the respective site-projected DOS—again using Equation 2.1.

2.2 Bulk crystal graph neural networks

High-throughput virtual screening (HTVS) enables exploration of chemical space for materials with desired properties. Although traditional approaches have calculated material properties using first-principles simulations such as DFT, these methods are expensive and often are not scalable to exhaustively screen the desired space [Jain et al., 2011, Emery et al., 2016, Emery and Wolverton, 2017]. Machine learning (ML)-based property prediction is thus becoming increasingly used towards materials discovery [Unterthiner et al., 2014, Gómez-Bombarelli et al., 2016]. Initial approaches leveraged fixed-dimensional manually-engineered molecular fingerprints as input to machine learning architectures [Unterthiner et al., 2014, Ramsundar et al., 2015], but differentiable graph-based fingerprints have since improved property prediction and interpretability [Duvenaud et al., 2015]. In such methods, chemical structures are represented graphically as a set of atoms (nodes) connected by bonds (edges) upon which neural network training methods optimize an end-to-end mapping between structure and property. Message passing over such graphs has become a widely-used architecture wherein convolutions between nodal features of neighbouring atoms, and optionally their associated bonds, generate representations of local chemical environment. Subsequent application of feed-forward neural networks then map these local environments to global properties [Yang et al., 2019, Xie and Grossman, 2018].

Although these graph convolutional neural network (GNN) methods were originally developed for non-periodic molecular systems, they have since been extended to crystalline systems through integration of periodic-boundary conditions. In such architectures, the molecular graphs can accommodate multiple edges between a given pair of nodes to represent periodicity. Moreover, since crystalline systems do not have discrete covalent bonds, edges instead connect each atomic node with all neighbouring atoms within a defined cutoff radius [Xie and Grossman, 2018].

2.2.1 CGCNN: Crystal graph convolutional neural networks

The crystal graph convolutional neural network (CGCNN) architecture developed by Xie and Grossman [2018], illustrated in Figure 2-1, takes a three-dimensional crystal structure as input and represents it as an undirected multigraph $\mathcal{G}$. Each site i in the crystal is designated by a feature vector $\mathbf{v}_i$ encoding its atomic properties. Each graph edge $(i,j)_k$—connection k between neighbouring atoms i and j—is featurized with the distance between atoms and represented by a feature vector $\mathbf{u}_{(i,j)_k}$. Neighbour distances are computed in a periodicity-aware manner and thus the crystal graph can have multiple edges between a given pair of nodes.

The crystal graph is passed through message passing layers, which iteratively update the nodal feature vectors with information from their associated neighbors and edges, thus automatically learning crystal site representations informed by their unique chemical environments. In each convolutional layer t, each edge vector and their associated neighbour vectors are concatenated as $\mathbf{z}^t_{(i,j)_k} = \mathbf{v}^t_i \oplus \mathbf{v}^t_j \mathbf{u}_{(i,j)_k}$ and convolved as

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \sum_{j,k \in \mathcal{G}} \left[\sigma\left(\mathbf{z}^t_{(i,j)_k} \mathbf{W}^t_1 + \mathbf{b}^t_1 \right) \odot g\left(\mathbf{z}^t_{(i,j)_k} \mathbf{W}^t_2 + \mathbf{b}^t_2 \right) \right], \tag{2.2}$$

where $\oplus$ and $\odot$ denote concatenation and element-wise multiplication, respectively,

$\mathbf{W}^t$ and $\mathbf{b}^t$ denote weight and bias matrices for the tth convolution layer, $\sigma(\cdot)$ denotes a sigmoid function, and $g(\cdot)$ denotes a non-linear activation function. The terms within the summation denote the two-body correlation between neighbour pairs wherein $\sigma(\cdot)$ acts to differentiate interaction strengths between $\mathbf{v}_i$ and its various neighbours [Xie and Grossman, 2018, Sanyal et al., 2018, Park and Wolverton, 2020].

After the specified number of convolutions, the updated nodal feature vectors are then pooled to yield a crystal feature vector $\mathbf{v}_c$,

$$\mathbf{v}_c = \frac{1}{N} \sum_{i \in \mathcal{G}} \mathbf{v}_i, \tag{2.3}$$

where N is the total number of atomic nodes in the crystal. Note that the outlined formulation to construct crystal features from crystal graphs ensures that the resulting crystal feature vector is permutationally invariant with respect to atom indexing and size invariant with respect to supercell size. The crystal feature vector is then passed through several fully connected hidden layers,

$$\mathbf{v}_c^{l+1} = f\left(\mathbf{v}_c^l \mathbf{W_h}^l + \mathbf{b}_\mathbf{h}^l\right), \tag{2.4}$$

where $\mathbf{W_h}^l$ and $\mathbf{b}_\mathbf{h}^l$ denote weight and bias matrices for the lth hidden layer. The final output layer maps $\mathbf{v}_c$ to a scalar output $\hat{y}$, the predicted target property [Xie and Grossman, 2018].

The model is trained by minimizing the difference, defined by a loss function $L(y, \hat{y})$, between target (DFT-calculated) property y and predicted property $\hat{y}$, respectively, computed as $\hat{y} = \text{CGCNN}(\mathcal{G}; \mathbf{W})$, where CGCNN is the neural network, $\mathcal{G}$ is the crystal graph, and $\mathbf{W}$ is the set of weights that parametrizes the model layers [Xie and Grossman, 2018]. Here, a mean average error loss function was used. During training, model weights $\mathbf{W}$ were optimized to minimize $L(y, \hat{y} = \text{CGCNN}(\mathcal{C}; \mathbf{W}))$ through iterative updates calculated by backpropagation and stochastic gradient descent or an Adam optimization algorithm [Xie and Grossman, 2018, Kingma and Ba, 2014].

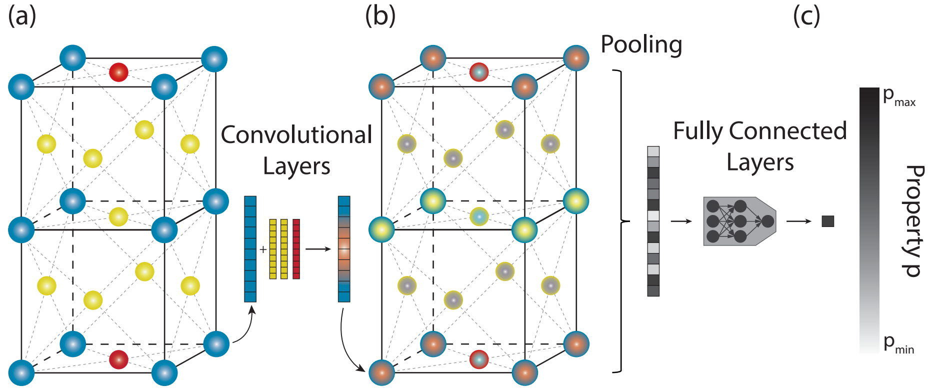

Figure 2-1: Crystal graph convolutional neural network architecture. (a) Three-dimensional crystal structure is converted to a graph, with nodes representing atoms and edges representing connections between neighbouring atoms. Atomic feature vectors from neighbouring atoms are passed through convolutional layers to obtain (b) nodal feature vectors encoding local chemical environment. This graphical representation is pooled into a crystal-wide vector, which is passed through a feed-forward neural network and mapped to (c) a scalar output value.

Crystal graph neural networks on unrelaxed crystals

The CGCNN model developed by Xie and Grossman [2018] has traditionally been trained on datasets of relaxed structures. Such models act as a surrogate for DFT electronic structure property calculation. To mitigate DFT structure optimization calculations—the more computationally expensive task—models must be trained on unrelaxed crystal structures, however. In the presented work, we therefore trained and tested CGCNN—as well as the PAINN model described below—on unrelaxed perovskite structures. For crystals derived from our own HTVS-generated dataset, the unrelaxed structures were available. This was not the case for literature-derived structures and therefore these structures were *unrelaxed* by 1) identifying the component A- and B-site atoms, 2) matching each atom with the nearest A- or B- site, respectively, in an unrelaxed template structure, and 3) removing from the training dataset any crystals for which uncertainty in matching sites between the template and relaxed structure was beyond a designated threshold. All unrelaxed crystals were scaled to have the same lattice parameters—namely, structures were scaled to have a 4 Å cubic unit cell and thus a 8 Å cubic 2x2x2 supercell for multicomponent perovskites. It follows that each perovskite was represented by the same graphical structure, with only the atomic feature vectors differing to represent cationic composition. Models trained to predict DFT-calculated properties from unrelaxed structures thereby act as a surrogate for both the DFT structure relaxation and electronic structure property calculations.

2.2.2 PAINN: Polarizable atom interation neural networks

Although crystals are represented as 2D graphs in the CGCNN formulation described above, the interactions of atoms indeed occur in continuous 3D space. The relative arrangement of atomic nodes is represented using only the scalar distance between atoms. Moreover, CGCNN and other such commonly-used graph-based message passing neural networks pass messages with rotationally invariant filters, and thus there can be loss of relevant directional information, see Figure 2-2 [Xie and Grossman, 2018, Yang et al., 2019]. To account for this, equivariant directional message passing was introduced in the directional message passing neural network (DimeNet) formulation, wherein message embeddings are transformed not only by the distance between atoms but also by the directions to neighbouring atoms [Klicpera et al., 2020]. Nevertheless, angular information in DimeNet is restricted to messages while the molecular graph representations remain rotationally invariant. The polarizable atom interation neural network (PAINN) architecture thus extends this formulation to include rotationally equivariant representations alongside this equivariant message passing, see Figure 2-2 [Schütt et al., 2021].

In the PAINN formulation, the molecular graph $\mathcal{G}$ is embedded in 3D space with edges denoted by the vector $\vec{r_{ij}} = \vec{r_i} - \vec{r_i}$ between associated nodes i, j, see Figure 2-2. Indeed, this contrasts the invariant approach in CGCNN, which featurizes edges only by the scalar distance between nodes i, j. Moreover, PAINN allows nodal representations to be both scalar and vectorial, $\mathbf{v}_i$ and $\vec{\mathbf{v}}_i$. Generally, the tth message

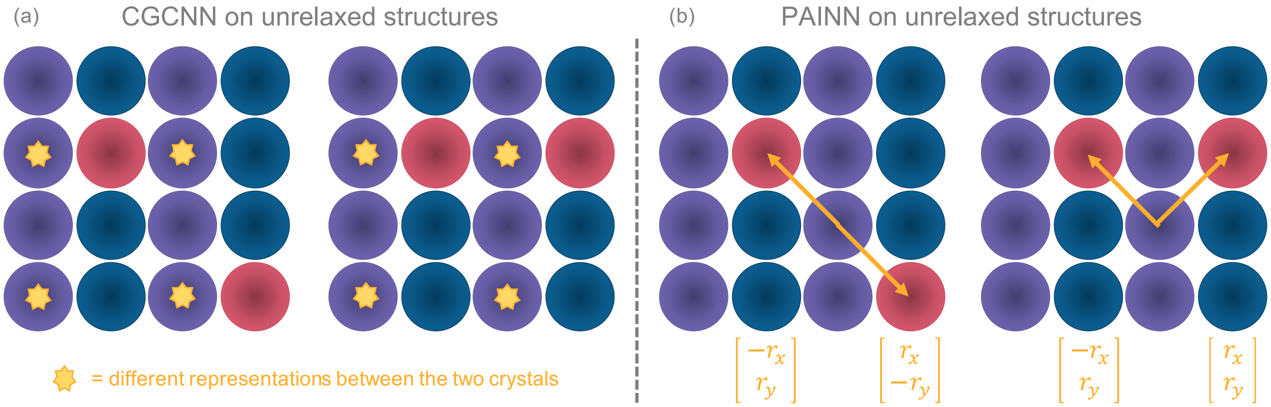
(a)
CGCNN on unrelaxed structures
= different representations between the two crystals
(b)
PAINN on unrelaxed structures
$\begin{bmatrix}-r_x\\ r_y\end{bmatrix}$
$\begin{bmatrix}r_x\\ -r_y\end{bmatrix}$
$\begin{bmatrix}-r_x\\ r_y\end{bmatrix}$
$\begin{bmatrix}r_x\\ r_y\end{bmatrix}$

Figure 2-2: Capabilities of CGCNN and PAINN to differentiate atomic ordering differences in unrelaxed crystalline systems. The two toy crystal structures shown here, with sites coloured by arbitrary atomic type, have the same chemical composition but different atomic ordering. After one message passing layer between the first sphere of nearest neighbours (a) all atoms in the CGCNN-generated representation would have the same featurization between the two structures except those starred, despite the non-starred purple atoms being in differing local environments. Meanwhile, (b) the PAINN-generated representation captures the differing local environments between the structures for all purple atoms. The purple atom second from the left and second from the bottom, for example, would have the same (a) CGCNN representation because edges with the two diagonal red neighbours have the same distance regardless of position. The (b) PAINN-generate representation, however, distinguishes the positional difference of the red atoms through differing edge vectors, shown in yellow.

passing update is given by

$$\begin{aligned}\mathbf{v}_i^{t+1} &= \mathbf{U}_t\left(\mathbf{v}_i^t, \sum_{j,ij\in\mathcal{G}} \mathbf{M}_t\left(\mathbf{v}_i, \mathbf{v}_j, \vec{r}_{ij}\right)\right)\\ \vec{\mathbf{v}}_i^{t+1} &= \vec{\mathbf{U}}_t\left(\vec{\mathbf{v}}_i^t, \sum_{j,ij\in\mathcal{G}} \vec{\mathbf{M}}_t\left(\mathbf{v}_i, \mathbf{v}_j, \vec{\mathbf{v}}_i, \vec{\mathbf{v}}_j, \vec{r}_{ij}\right)\right),\end{aligned} \tag{2.5}$$

for scalar and vectorial representations $\mathbf{v}_i^{t+1}$ and $\vec{\mathbf{v}}_i^{t+1}$, respectively, wehre $\mathbf{U}_t$ and $\vec{\mathbf{U}}_t$ are update functions, and $\mathbf{M}_t$ and $\vec{\mathbf{M}}_t$ are message functions [Schütt et al., 2021]. For any rotation matrix $R \in \mathbb{R}^{3x3}$, rotational invariance of $\mathbf{v}_i^{t+1}$ can be achieved with rotationally invariant functions $\mathbf{M}_t(\vec{\mathbf{x}}) = \mathbf{M}_t(R\vec{\mathbf{x}})$ and $\mathbf{U}_t(\vec{\mathbf{x}}) = \mathbf{U}_t(R\vec{\mathbf{x}})$ while rotational equivariance of $\vec{\mathbf{v}}_i^{t+1}$ can be achieved with rotationally equivariant functions $R\vec{\mathbf{M}}_t(\vec{\mathbf{x}}) = \vec{\mathbf{M}}_t(R\vec{\mathbf{x}})$ and $R\vec{\mathbf{U}}_t(\vec{\mathbf{x}}) = \vec{\mathbf{U}}_t(R\vec{\mathbf{x}})$. It follows from these constraints that non-linear functions can be applied to the scalar features $\mathbf{v}_i^{t+1}$ while vectorial features

$\vec{\mathbf{v}}_i^{t+1}$ are only transformed linearly [Schütt et al., 2021].

Following message passing layers, a feed-forward neural network is applied to the learned molecular representation, with the final layer mapping to a scalar output $\hat{y}$ for each target property, similar to the formulation described above for CGCNN. As above, during training model weights $\mathbf{W}$ were optimized to minimize $L(y, \hat{y} = \text{PAINN}(\mathcal{C}; \mathbf{W}))$, here using an Adam optimizer and a mean squared error loss function [Schütt et al., 2021, Kingma and Ba, 2014].

Though the original formulation of PAINN by Schütt et al. [2021] was for non-periodic molecular systems, we have adapted the method for use in crystalline systems by implementing periodic boundary conditions. Displacement vectors between atoms are computed in a periodicity-aware manner and thus the resulting graph can have multiple edges between a given pair of nodes. Also like CGCNN, edges are defined to pair each atomic node i with all neighbouring atoms j within a defined cutoff radius.

Schütt et al. [2021] implemented PAINN with nuclear charges $Z_i \in \mathbb{N}$ and positions $\vec{r}_i \in \mathbb{R}^3$ as inputs for each atom i. We found, however, that model performance was improved by instead using the atom initialization vectors from Xie and Grossman [2018] as input alongside positions $\vec{r}_i \in \mathbb{R}^3$.

2.2.3 Interpolation of multicomponent perovskite properties

In this work, we are aiming to predict the properties of multicomponent perovskites, with composition $AA'BB'O_3$ (including ternary perovskites $ABB'O_3$ or $AA'BO_3$), which constitute a $O(10^6)$ search space. Given the considerably smaller $O(10^3)$ search space of binary ABO_3 perovskites, however, we propose to leverage their calculated properties to generate a baseline prediction of multicomponent perovskite properties. This smaller search space, in combination with the ability to calculate DFT descriptors for binary ABO_3 perovskites using a 1x1x1 supercell, makes it tractable to calculate target properties from first-principles. As is discussed in the results below, this has been done, in the literature or in this work, for over 3,000 ABO_3 perovskites [Emery et al., 2016, Emery and Wolverton, 2017, Castelli et al., 2012a,b, Jacobs et al., 2018].

For a given $A_xA'_{1-x}B_yB'_{1-y}O_3$ perovskite, we propose to calculate a baseline property approximation $\tilde{P}_{A_xA'_{1-x}B_yB'_{1-y}O_3}$ as as linear combination of constituent binary perovskites properties:

$$\begin{aligned}\tilde{P}_{AA'BB'O_3} = x \cdot y \cdot P_{ABO_3} + (1-x) \cdot y \cdot P_{A'BO_3} + \\ x \cdot (1-y) \cdot P_{AB'O_3} + (1-x) \cdot (1-y) \cdot P_{A'B'O_3}\end{aligned} \tag{2.6}$$

We propose to use this interpolated property approximation $\tilde{P}_{AA'BB'O_3}$ as a baseline to be improved upon using crystal GNNs. We calculate the deviation from the linear approximation as $\Delta P_{AA'BB'O_3} = P^{DFT}_{AA'BB'O_3} - \tilde{P}_{AA'BB'O_3}$ and train a GNN to predict $\Delta P_{AA'BB'O_3}$. The steps for interpolation-guided property prediction are thus:

(1) Calculate interpolation baseline $\tilde{P}_{AA'BB'O_3}$

(2) Use GNN to predict $\Delta P^{GNN}_{AA'BB'O_3}$, minimizing $L(\Delta P_{AA'BB'O_3}, \Delta P^{GCN}_{AA'BB'O_3})$

(3) Compute property prediction as $P^{\text{predicted}}_{AA'BB'O_3} = \tilde{P}_{AA'BB'O_3} + \Delta P^{GCN}_{AA'BB'O_3}$

2.3 Per-site crystal graph neural networks

In this work, we extend the capabilities of crystal graph neural networks to predict per-site crystal properties, assigning an output property to each site in the crystal lattice. This task is a natural extension of predicting bulk properties—one output property for each crystal—and has been implemented as a modification to the CGCNN architecture [Xie and Grossman, 2018].

2.3.1 Per-site CGCNN

The crystal graph generation and message passing layers implemented in the per-site CGCNN model are the same as described above for the bulk CGCNN. After the convolutional layers, however, no pooling is conducted. Instead, the learned nodal feature vectors are passed through several fully connected hidden layers, then finally to an output layer yielding a vector of predicted properties for each site, see Figure 2-3. The model is trained by minimizing the difference, defined by a loss function $L(\mathbf{p}, \tilde{\mathbf{p}})$, between target (DFT-calculated) and predicted per-site property vectors, $\mathbf{p}$ and $\tilde{\mathbf{p}}$, respectively. As in the bulk model, a mean average loss function was used. The vector of per-site predictions is computed as $\tilde{\mathbf{p}} = \mathrm{CGCNN}(\mathcal{C}; \mathbf{W})$. Analogously to the bulk CGCNN training, model weights $\mathbf{W}$ were optimized to minimize $L(\mathbf{p}, \tilde{\mathbf{p}} = f(\mathcal{C}; \mathbf{W}))$ [Xie and Grossman, 2018].

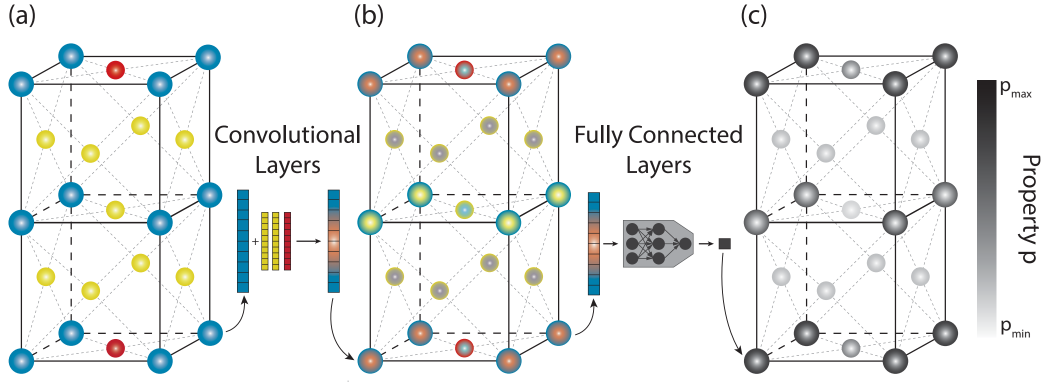

Figure 2-3: Per-site crystal graph convolutional neural network architecture. (a) Three-dimensional crystal structure is converted to a graph, with nodes representing atoms and edges representing connections between neighbouring atoms. Atomic feature vectors from neighbouring atoms are passed through convolutional layers to obtain (b) nodal feature vectors encoding local chemical environment. The resulting feature vectors for each atom are passed through a feed-forward neural network and mapped to a (c) property vector whose entries correspond to each site in the crystal.

2.3.2 Per-site multilayer perceptron

A multi-layer perceptron (MLP) was implemented as a baseline for performance comparison with per-site CGCNN. The MLP takes as input all the same structural data as the CGCNN, but does not leverage graph structure nor convolutions. Each atom in the crystal is featurized as its CGCNN atomic encoding, along with the atomic encoding of and radial distance to its nearest neighbours, up to the same cutoff used in the CGCNN model. The resulting atomic feature vectors were passed through through several fully connected hidden layers, with the final layer mapping to a vector of predicted properties for each site. This MLP model formulation is memory-inefficient compared to the per-site CGCNN because instead of linking atom featurizations together in a graph, it repeats them in each associated neighbour list.

2.4 Neural network training and hyperparameter optimization

All models presented here were trained, validated, and tested on 60%, 20%, and 20% of the data, respectively. Randomly-generated train-validation-test splits were applied and kept consistent between all models using a given dataset. All presented results reflect performance on the test set. Reported averages and standard deviations reflect statistics from initializing three replicate models with different random model weights. Final model weights were chosen based on optimal validation set performance. Hyperparameter tuning was performed using SigOpt [Clark and Hayes, 2019], optimizing hyperparameters including the number of convolutional layers, activation functions, number of hidden layers, hidden feature dimensions, learning rate.

Chapter 3

Results

3.1 Datasets

In this work, we leverage density functional theory (DFT) datasets of perovskite oxides both from the literature and calculated using our in-house high-throughput virtual screening (HTVS) platform. Previous high-throughput screening efforts in perovskites have largely focused on ABO_3 perovskites, without alloying of the A- and B-sites, or on narrow compositional ranges of perovskites with A- and/or B-site alloying. We have therefore used our HTVS capabilities to generate a dataset of highly-alloyed multicomponent perovskites over a wide compositional space, which is then leveraged to train graph neural networks for rapid property prediction.

As shown in Table 3.1, perovskite datasets available in the literature from the Open Quantum Materials Database (OQMD) [Kirklin et al., 2015, Saal et al., 2013, Emery et al., 2016, Emery and Wolverton, 2017] and the K. Jacobsen research group [Castelli et al., 2012a,b] contain over 3,000 binary ABO_3 perovskites. Together, these datasets include perovskites with 72 and 73 different A- and B-site cations, respectively. Despite their lack of cationic alloying, these datasets thus cover a broad chemical space.

In contrast, the dataset from the D. Morgan research group [Jacobs et al., 2018] includes perovskites with 19 and 23 different A- and B-site cations, respectively, but explores the impacts of alloying—it contains both ternary $AA'BO_3$, $ABB'O_3$ and quaternary $AA'BB'O_3$ perovskites.

The HTVS-generated dataset presented contains over 5,000 multicomponent perovskites, focusing primarily on quaternary $AA'BB'O_3$ structures, with 22 and 25 different A- and B-site cations, respectively. The numbers reported in this section do not include additional series of calculations completed for several $A_xA'_{1-x}B_yB'_{1-y}O_3$ families, however, which are discussed in Section 3.3.1.

Table 3.1: DFT-calculated perovskite datasets

Dataset	Total Perovskites	Binary Perovskites ABO_3	Ternary Perovskites $AA'BO_3$ or $ABB'O_3$	Quaternary Perovskites $AA'BB'O_3$
Literature	**5964**	**3611**	**1384**	**956**
OQMD[1]	1135	1135	0	0
Jacobsen et al.[2]	2363	2363	0	0
Morgan et al.[3]	2466	113	1384	956
HTVS (this work)	**5797**	**160**	**1029**	**4608**

[1] [Emery et al., 2016, Emery and Wolverton, 2017]
[2] [Castelli et al., 2012a,b]
[3] [Jacobs et al., 2018]

We performed electronic structure calculations on all datasets outlined here—including those from the literature—to obtain DFT-derived descriptors of catalytic activity and stability, including energy above convex hull (ΔE_{hull}), energy above Pourbaix hull (ΔG_{pbx}), differences between oxygen 2p-band centers and the Fermi level ($O_{2p} - E_V$), and differences between B-site metal d-band centers and the Fermi level ($B_d - E_V$). The property coverage space of the literature and in-house HTVS datasets are shown in Figure 3-1. We find that the distributions of ΔE_{hull} and ΔG_{pbx} values found in our HTVS dataset are more narrowly concentrated near 0 eV/atom, indicating that the dataset achieves better thermodynamic and aqueous stability overall. Instability of certain perovskites in the literature datasets is likely attributed to the presence of atypical A- and B-site cations in some binary perovskites. Meanwhile, despite containing significantly less cationic diversity than the binary perovskite literature data, our dataset displays similar distributions of electronic density of state (DOS) descriptors—lacking only upper-limit $O_{2p} - E_V$ values near 0 eV and both lower- and upper-limit $B_d - E_V$ values. This highlights the capacity of alloying in A- and B-sites to tune properties.

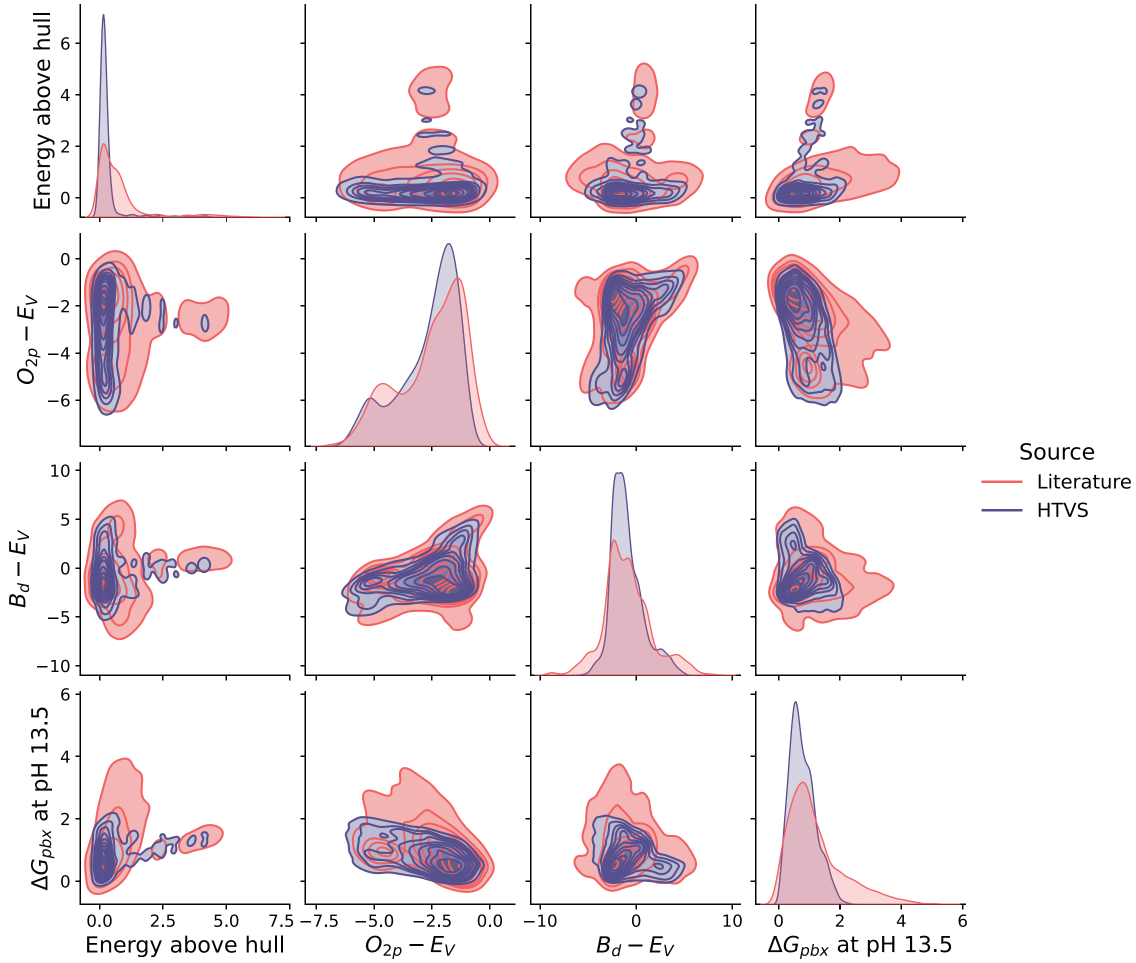

Figure 3-1: Distribution and correlation of DFT-calculated properties in perovskite datasets. Distributions of bulk descriptors of catalytic stability—energy above convex hull (ΔE_{hull}) and above Pourbaix hull (ΔG_{pbx})—and catalytic stability—oxygen $2p$-band centers ($O_{2p} - E_V$) and B-site metal d-band centers ($B_d - E_V$) are compared between literature and HTVS datasets. The literature values here reflect data from Emery et al. [2016], Emery and Wolverton [2017], Castelli et al. [2012a,b], Jacobs et al. [2018].

3.1.1 Multicomponent perovskite datasets

This work mainly focuses on multicomponent perovskite systems, wherein compositional alloying of A- and B-sites facilitates tuning of catalytically-relevant properties. The HTVS efforts presented here thus focused on generating an improved multicomponent perovskite dataset. All data in this section are derived from DFT simulations of 40-atom 2x2x2 perovskite supercells.

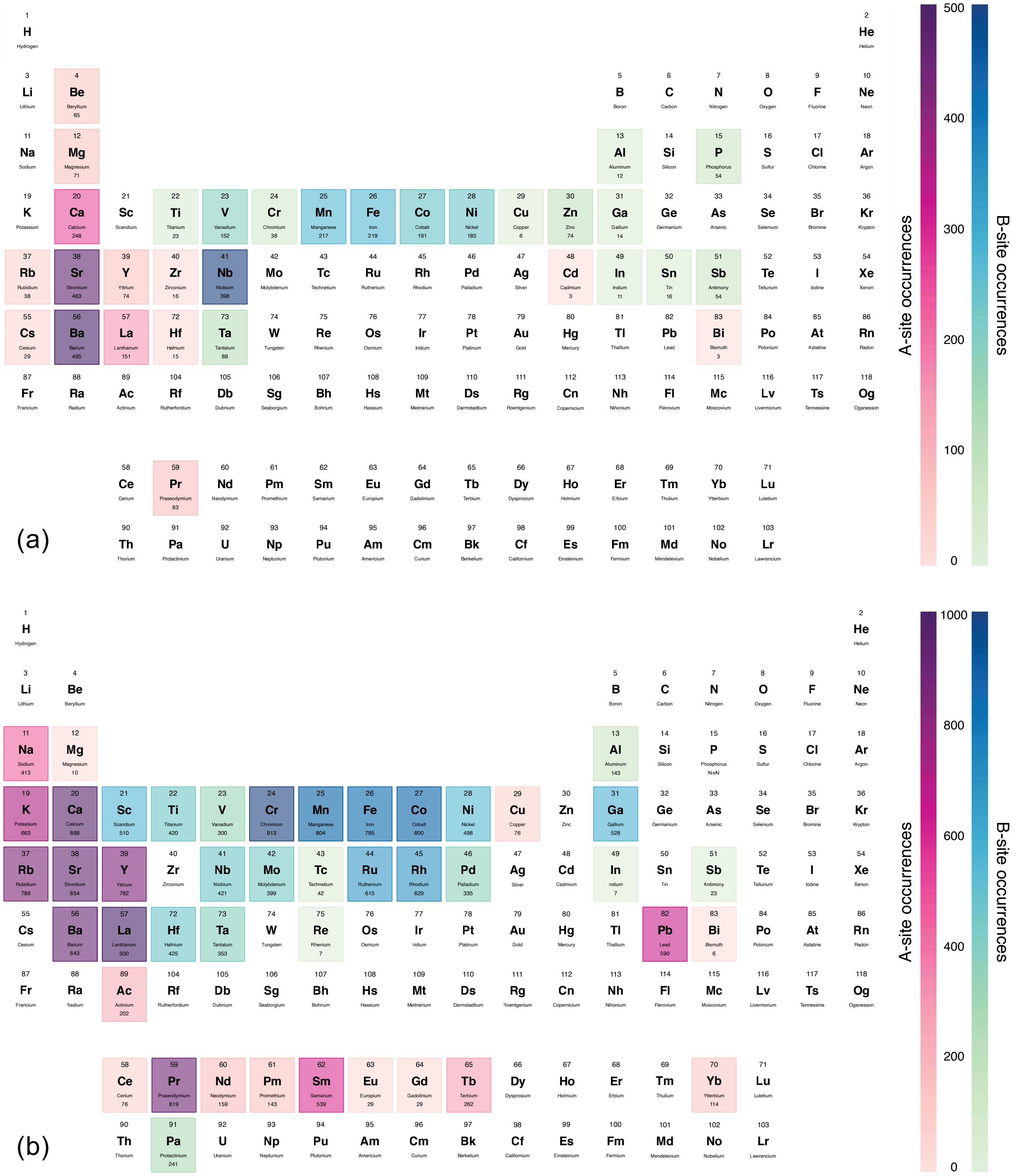

Figure 3-2: Occurrences of A- and B-site cations in multicomponent perovskite datasets. Cations found in multicomponent perovskites from (a) literature and (b) HTVS datasets are shown. Note that both the literature and HTVS datasets include certain elements—Ti, Mg, Be, Zr, Sn, Ta, Nb, Hf, Zn, Y and Mg, Cu, Y, respectively—in both A- and B-sites. In these cases, elemental occurrences are presented for the site where the element was most commonly found.

Figure 3-2 shows the distribution of A- and B-site cations in literature and HTVS multicomponent perovskite datasets. Our HTVS dataset has greater spread over different cations whereas the dataset from the Morgan research group [Jacobs et al., 2018] focuses much of its data on a small subset cations—namely, Ba, Sr, Ca, La in the A-site and Nb, Fe, Mn, Co, Ni, V in the B-site.

The diversity of compositions in the HTVS dataset is further highlighted in Figure 3-3a. Despite containing slightly fewer ternary perovskites, the HTVS dataset covers slightly more A,A',B and A,B,B' compositions. This broader coverage of compositional space is more pronounced for quaternary perovskites, with our HTVS dataset containing over ten-times the number of unique A,A',B,B' groupings compared to the literature dataset. The HTVS dataset also contains more thorough sampling of compositional ratios, shown in Figure 3-3b. This dataset is more evenly spread over possible x and y values in $A_xA'_{1-x}BO_3$, $AB_yB'_{1-y}O_3$, and $A_xA'_{1-x}B_yB'_{1-y}O_3$.

Indeed, as shown in Figure 3-4, the HTVS dataset covers property regions not present in the literature dataset—e.g. perovskites with both low $O_{2p}-E_V$ and B_d-E_V values, perovskites with low ΔG_{pbx} and low $O_{2p} - E_V$ or $B_d - E_V$ values. All the while, HTVS-derived structures exhibit similar stability to that found in the literature dataset—this is reflected in the distributions of ΔE_{hull} and ΔG_{pbx} near 0 eV/atom.

Beyond its broadened compositional coverage, our multicomponent HTVS dataset's novelty lies in the decorational diversity of cations within the perovskite lattice. In the literature dataset, cations were arranged in largely consistent configurations—e.g. with B-site alloying in a rock salt arrangement. Although B-sites typically do order into rock salt arrangements when ordering occurs, many alloyed perovskites do not have regular ordering [King and Woodward, 2010]. We therefore require datasets with varied cationic arrangements to analyze the impacts of decorational differences. Moreover, such datasets are required to build machine learning models that capture property differences across perovskites of the same composition but with different cationic decorations. Therefore, when generating input structures for the presented HTVS dataset, the desired combinations of A- and B-cations were populated into random sites within their respective sublattices.

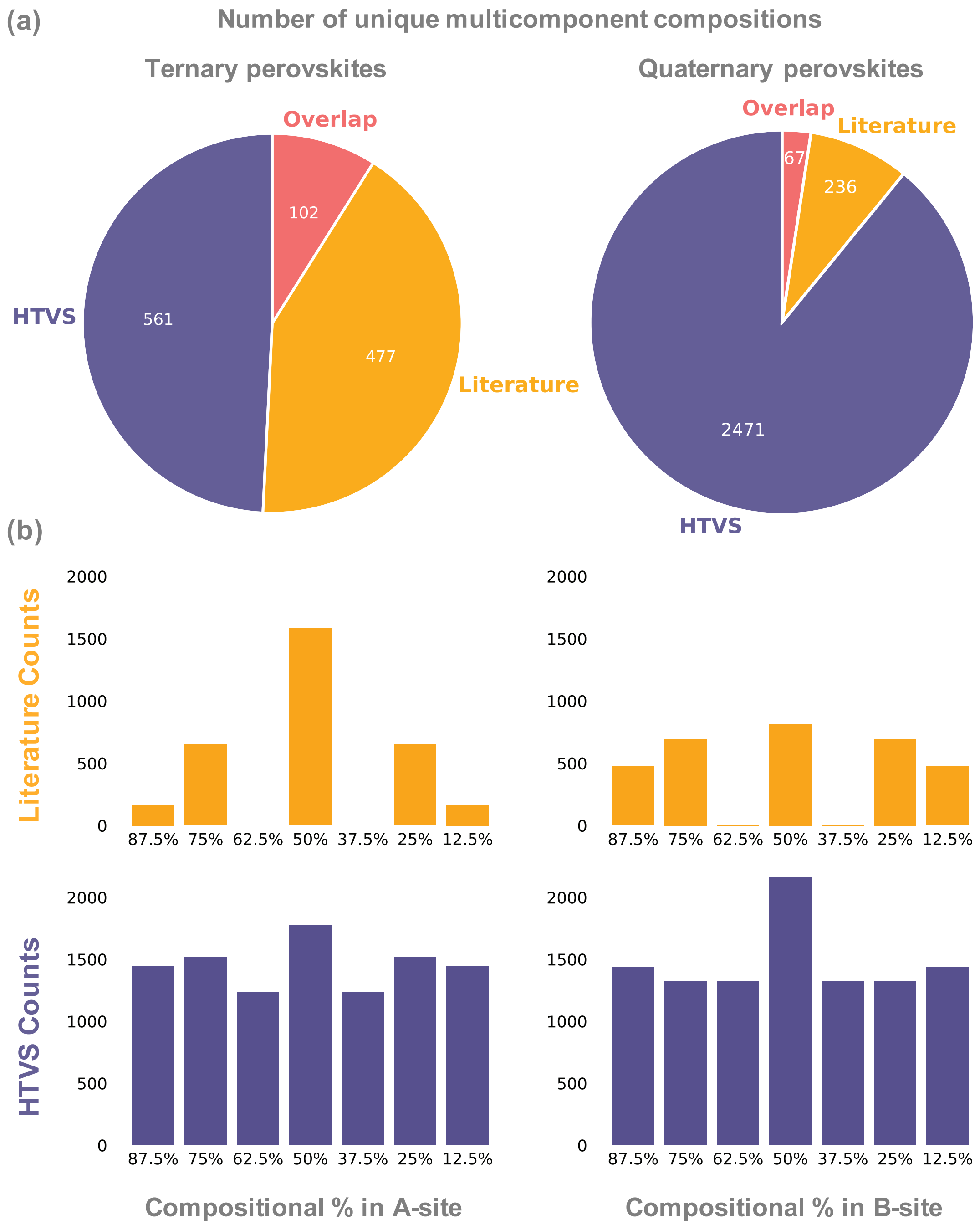

Figure 3-3: Compositional makeup of multicomponent perovskite datasets. The number of unique ternary $AA'BO_3$, $ABB'O_3$ compositions (a, left) and quaternary $AA'BB'O_3$ compositions (a, right) are compared between literature and HTVS datasets. This analysis ignores the ratios of A- and B-site alloying, counting only the number of unique A,A',B,B' groupings. Overlap denotes compositions present in both datasets. Separately, distributions of A- and B-site alloying ratios are compared between literature (b, top) and HTVS (b, bottom) datasets. Presented counts were combined from both ternary and quaternary perovskites. Note that the symmetry in these histograms arises from the reciprocal occurrences of 87.5%/12.5%, 75%/25%, 62.5%/37.5%, and 50%/50%.

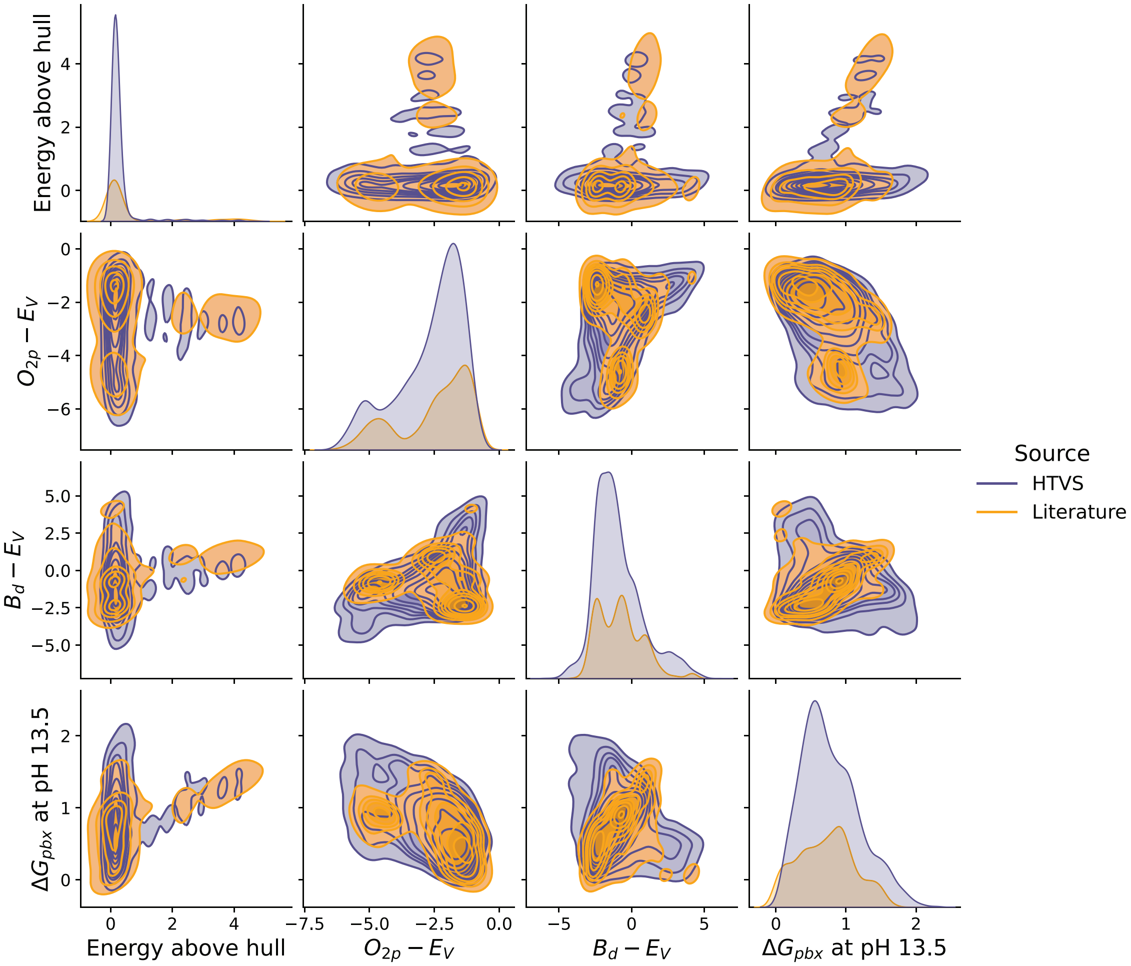

Figure 3-4: Distribution and correlation of DFT-calculated properties in multicomponent perovskite datasets. Distributions of bulk descriptors of catalytic stability—energy above convex hull (ΔE_{hull}) and above Pourbaix hull (ΔG_{pbx})—and catalytic stability—oxygen $2p$-band centers ($O_{2p} - E_V$) and B-site metal d-band centers ($B_d - E_V$) are compared between literature and HTVS datasets of multicomponent perovskites. The literature values here reflect data from Jacobs et al. [2018].

For quaternary perovskites alone, the presented HTVS dataset includes over 4,000 $AA'BB'O_3$ structures with highly varied A- and B-site alloying ratios and over 2,000 unique cationic combinations. Compared to the $O(100)$ unique combinations of cations in the literature, our dataset significantly increases the diversity of available highly-alloyed perovskite DFT data. This broader compositional coverage provides improved training data for machine learning models described in the following section. Moreover, the varied cationic arrangements present in the HTVS dataset enable us to study the impact of decorational differences—as is discussed in Section 3.3. Upon publication, this HTVS dataset will therefore provide the research community with a more thorough benchmark dataset for study of alloyed perovskites and analyses of compositional trends.

3.2 Bulk crystal property prediction

DFT-derived electronic structure properties are useful as descriptors of catalytic stability and activity for candidate materials, but screening highly-alloyed systems with DFT becomes intractable due the requirement for large supercells and the size of the compositional search space. To achieve the A- and B-site alloying in the multicomponent perovskites described above, for example, 40-atom 2x2x2 supercells ($A_8B_8O_{24}$) were used. Geometry optimization and electronic structure calculations thus required DFT calculations with runtimes on the order of days. The search space of these $A_xA'_{1-x}B_yB'_{1-y}O_{24}$ structures, considering $x, y \in [0, 8]$, is $O(10^6)$ and thus not tractable to simulate exhaustively.

We therefore aimed to implement data-driven approaches with which screening $O(10^6)$ multicomponent $AA'BB'O_3$ structures becomes tractable. To do so, we must be able to predict properties from unrelaxed crystal structures, thus mitigating the need for expensive structure optimization DFT calculations. Here, we implemented graph neural networks (GNNs) to predict energies above hull (ΔE_{hull}), energies above Pourbaix hull (ΔG_{pbx}), differences between the oxygen $2p$-band center and the Fermi level ($O_{2p} - E_V$), differences between the B-site metal d-band center and the Fermi level ($B_d - E_V$), as well as differences between the B-site atom d-band center and the oxygen 2p-band center ($B_d - O_{2p}$).

Unlike that of multicomponent $AA'BB'O_3$ perovskites, the search space of binary ABO_3 perovskites is tractable to screen exhaustively with DFT as it is $O(10^3)$ and simulations can be performed using a 5-atom 1x1x1 supercell ($A_1B_1O_3$) with runtimes on the order of hours instead of days. Leveraging this accessiblity, we propose interpolating preliminary estimates of multicomponent perovskite descriptors from calculated properties of constituent binary perovskites. GNNs were then used to predict deviations from these baseline estimates, improving property prediction performance on unrelaxed bulk perovskite structures.

The mulitcomponent perovskite datasets presented in the previous section were used to train and test the models—a combination of data from the literature and from calculations with our in-house high-throughput virtual screening (HTVS) platform. Each presented model was validated on ∼1,400 structures and tested on ∼1,400 structures, consistent between all models. Models trained to predict properties directly were trained on ∼4,700 perovskites while those trained to predict deviations from interpolated property estimates were trained on a subset of ∼4,000 crystal structures. Statistics reflect performance on the held-out test set, reported over three replicate models initialized with different random weights.

Direct property prediction models, without interpolation priors, were also trained on a dataset of both multicomponent and binary ABO_3 perovskites. The multicomponent perovskite training data (∼4,700) was supplemented with all binary perovskites (∼3,700) for a total of >8,000 training structures. The same validation and testing data as above were used to evaluate the models. Performance on these multicomponent perovskite datasets was decreased in all cases, however, and thus these models were not pursued further. This drop in performance may be attributed to model training over the broader elemental variety (∼70 A- and B-site elements, respectively, instead of ∼20) and property distribution present in the binary perovskite dataset, see Figure 3-1. It is conceivable that training on this more general dataset caused increased attention toward the presence/absence of individual elements and less toward the effects of alloying in multicomponent systems.

3.2.1 Property prediction from unrelaxed perovskite structures

The crystal graph convolutional neural network (CGCNN) developed by Xie and Grossman [2018] has shown success in predicting electronic structure properties from crystal structure. From the input structure, a graph representing atoms as nodes and atom connectivity as edges is constructed, with atom connectivity defined as all pairs of neighbouring atoms within a cutoff radius. The nodal feature vectors encode the element present at each site, while the edge features encode the distance between each atom in connected pairs. Given a relaxed crystal structure these interatomic distances encode information about structural distortions, which influence bulk properties of a crystalline system. Expensive geometry optimization DFT simulations are required to determine the relaxed structure of a perovskite system, however, thereby negating the use of machine learning to bypass *ab initio* calculations and make large search spaces tractable.

Table 3.2: CGCNN performance on relaxed and unrelaxed datasets

		MAE	
Property	Units	Relaxed structures	Unrelaxed structures
ΔE_{hull}	eV/atom	0.073 ± 0.003	0.114 ± 0.003
ΔG_{pbx}[1]	eV/atom	0.059 ± 0.004	0.087 ± 0.004
$O_{2p} - E_V$	eV	0.239 ± 0.001	0.273 ± 0.005
$B_d - E_V$	eV	0.422 ± 0.009	0.450 ± 0.005
$B_d - O_{2p}$	eV	0.331 ± 0.001	0.353 ± 0.003

[1] ΔG_{pbx} was predicted at pH 13.5 and 1.6 V vs. RHE

We therefore performed property prediction on unrelaxed structures, scaling all inputs to be cubic perovskites with the same lattice parameter. Performance was compared to that of models trained on relaxed structures. The mean average error (MAE) of properties predicted with these models on test sets are reported in Table 3.2. For all properties, performance decreased upon training on unrelaxed structures, albeit to varying degrees. MAE values of stability descriptors, ΔE_{hull} and ΔG_{pbx}, were increased by $45 - 55\%$ while those of DOS descriptors were increased by only $5 - 15\%$, suggesting that distortional information encoded in the edge features of relaxed structure graphs is more important in calculating energetic information.

We also implemented polarizable atom interaction neural network (PAINN) models for property prediction on unrelaxed structure, mainly aiming to improve the ability to distinguish between different cationic decorations, as described in Section

3.3. PAINN performance, as reported in the first column of Table 3.3, was superior than CGCNN for all properties. Further, models for ΔE_{hull}, ΔG_{pbx}, $B_d - E_V$, and $B_d - O_{2p}$ achieved MAE values within standard deviation or lower than respective CGCNN models on relaxed structures.

3.2.2 Interpolation of multicomponent perovskite properties

Properties of multicomponent $AA'BB'O_3$ perovskites are inherently related to those of binary perovskites of constituent cations—ABO_3, $AB'O_3$, $A'BO_3$, $A'B'O_3$. We thus sought to leverage binary perovskite data, with thousands of such DFT calculations available in the literature [Emery et al., 2016, Emery and Wolverton, 2017, Jacobs et al., 2018], to predict multicomponent perovskite properties. As described in Section 2.2.3, we calculated approximations of multicomponent perovskite properties, $\tilde{P}_{AA'BB'O_3}$, as a weighted linear combination of constituent binary perovskite properties. Constituent binary perovskites of interest not found in the literature were simulated using our HTVS DFT workflow.

Table 3.3: Interpolation-PAINN performance

		MAE		
Property	Units	PAINN	Interpolation	Interpolation + PAINN
ΔE_{hull}	eV/atom	0.058 ± 0.006	0.063	**0.034 ± 0.001**
ΔG_{pbx}[1]	eV/atom	0.064 ± 0.005	0.080	**0.041 ± 0.001**
$O_{2p} - E_V$	eV	0.281 ± 0.024	0.417	**0.237 ± 0.004**
$B_d - E_V$	eV	0.341 ± 0.021	0.574	**0.294 ± 0.005**
$B_d - O_{2p}$	eV	0.244 ± 0.023	0.393	**0.191 ± 0.006**

[1] ΔG_{pbx} was predicted at pH 13.5 and 1.6 V vs. RHE. Models were also trained on data at pH 13 and 14, achieving performance statistics within t.

The MAE values of these interpolation estimates, evaluated on the same test set used to evaluate GNN performance, are reported in Table 3.3. The interpolation estimate for ΔE_{hull} was particularly successful in predicting multicomponent perovskite properties, achieving MAE values <10% higher than and within standard deviation of those from PAINN predictions—and less than those from CGCNN predictions on relaxed structures. The quality of interpolation results for ΔE_{hull} perhaps owes to energy additivity [Huggins and Sun, 1946]. Indeed, the ΔE_{hull} of all constituent binary perovskites is calculated from a subset of the compositional phase diagram used to calculate ΔE_{hull} for $AA'BB'O_3$. This may also explain the relative success of ΔG_{pbx} interpolations as compared to those for band centers.

3.2.3 Prediction of deviations from interpolation estimates

Aiming to synergize the promising performance of the interpolation estimates and PAINN predictions from unrelaxed structure, we implemented a PAINN model to learn the deviations of multicomponent perovskite properties from the interpolation. As described in Section 2.2.3, interpolated property estimates were calculated for all training/validation/test data and compared to DFT-calculated values to compute deviations, $\Delta P_{AA'BB'O_3} = P^{DFT}_{AA'BB'O_3} - \tilde{P}_{AA'BB'O_3}$, from ideal mixing. PAINN models were trained to predict these deviations and model outputs were summed with interpolation estimates to obtain an updated property prediction. As reported in Table 3.3, this combined interpolation-PAINN approach achieved better performance than any other method presented here. For all properties, the interpolation-PAINN approach outperformed CGCNN on relaxed structures and decreased MAE values achieved by PAINN or interpolation alone by $14-41\%$ and $46-51\%$, respectively.

3.2.4 Screening multicomponent perovskites

The bulk crystal property prediction methods described in Sections 3.2.1-3.2.3 make it possible to screen the $O(10^6)$ compositional search space of 40-atom 2x2x2 supercell multicomponent $AA'BB'O_3$ perovskites. Unrelaxed structures of $A_xA'_{8-x}B_yB'_{8-y}O_{24}$ for all possible combinations of the 20 A- and 20 B-site elements shown in Figure 3-5 were generated, with $x, y \in [0, 8]$. For each composition, a random decoration of A- and B-site atoms within the crystal lattice was used. All properties listed in Table 3.3 were predicted for these 1.2 million multicomponent perovskite structures. The direct PAINN model, without interpolation as a prior, was used because not all constituent binary perovskite DFT simulations had been completed. Results will be repeated using the improved interpolation-PAINN approach.

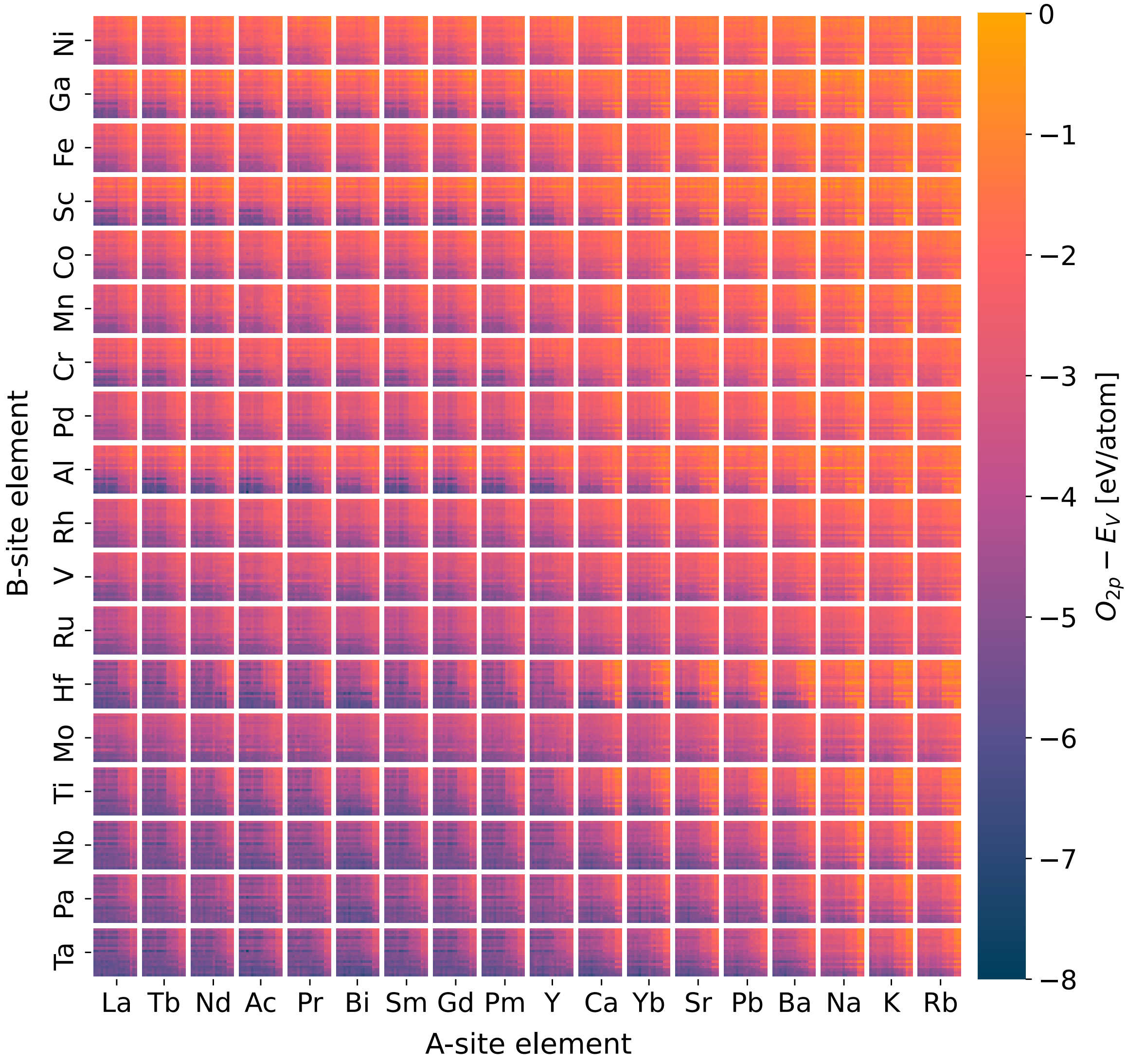

Figure 3-5: Predicted $O_{2p}-E_V$ of 1.2 million multicomponent $AA'BB'O_3$ perovskites. PAINN-predicted $O_{2p} - E_V$ values of $A_xA'_{8-x}B_yB'_{8-y}O_{24}$ perovskites were averaged over all $x, y \in [0, 8]$ to obtain an average value for each $AA'BB'O_3$ composition shown here. The outer and inner horizontal axes indicate A-site elements while the vertical axes indicate B-site elements. Elements on each set of axes are sorted by increasing $O_{2p}-E_V$ value, averaged across all structures containing the element. The order of A, A' and of B, B' here is meaningless and thus the $O_{2p} - E_V$ value of each $AA'BB'O_3$ is displayed at four corresponding pixels in the heatmap.

Averaged predictions of differences between the O_{2p} band center and the Fermi level for each $AA'BB'O_3$ combination are shown in Figure 3-5, giving an overview of elemental contributions to $O_{2p} - E_V$ in multicomponent systems. To obtain these $O_{2p} - E_V$ values for each $AA'BB'O_3$ composition, predicted $O_{2p} - E_V$ values for each $A_x A'_{8-x} B_y B'_{8-y} O_{24}$ were averaged over all $x, y \in [0, 8]$. These results are now being investigated to identify materials with a confluence of desirable properties for catalysis of the oxygen evolution reaction (OER) under alkaline conditions—optimizing for both catalytic stability (low ΔE_{hull} and ΔG_{pbx} at pH ~13.5) and activity (e.g. minimizing $O_{2p} - E_V$).

3.3 Modelling cationic decorational differences

In Section 3.2, analyses focused on the ability of GNNs to predict bulk properties of different $AA'BB'O_3$ compositions. In this section, we will extend our analyses to also consider the impacts of different cationic decorations—the relative arrangements of A- and B-site elements within the perovskite lattice.

For a given $A_x A'_{8-x} B_y B'_{8-y} O_{24}$ composition, A, A' and B, B' atoms can have different relative arrangements within the alloyed A- and B-site sublattices, respectively. For $x, y \in \{1, 8\}$ there is a single symmetrically inequivalent sublattice while there are 3 and 6 inequivalent sublattice arrangements for $x, y \in \{2, 3, 5, 6\}$ and $x, y = 4$, respectively. Symmetry inequivalence increases when two alloyed sublattices are combined in the full $A_x A'_{8-x} B_y B'_{8-y} O_{24}$ perovskite structure. While there is only one inequivalent structure for $x = y = 1$, this number increases to 6 for $x = 1, y = 4$; 26 for $x \in \{2, 3, 6, 7\}, y = 4$; and 52 for $x = 4, y = 4$, for example. Note that these numbers are the same upon interchanging x and y.

3.3.1 Impact of decorational differences on DFT-calculated properties

We sought to analyze whether cationic decorations significantly impact DFT-calculated descriptors—and thus whether they are an important consideration when modelling perovskite properties. We therefore performed DFT calculations for all symmetrically-inequivalent structures of several perovskite families, three of which are presented in Figure 3-6. For each $A_xA'_{8-x}B_yB'_{8-y}O_{24}$ series, $y = 4$ ($y = 0.5$) is held constant in while x ranges from 1 (0.125) to 7 (0.875). We found that decorational differences impacted DFT-calculated properties to varying degrees. Energy above hull distributions calculated for three representative series are shown in the first row of Figure 3-6.

In the case of $La_xPr_{1-x}Y_4Ni_4O_{24}$, ΔE_{hull} differences between structures of a given x composition are larger (∼0.2 eV/atom on average) than those between average ΔE_{hull} values for different x concentrations (up to ∼0.06 eV/atom). Similar results are seen for $Y_xLa_{1-x}In_4Mg_4O_{24}$, with ΔE_{hull} ranges of over 0.3 eV/atom between decorations of a given x concentration and only ∼0.15 eV/atom differences in average ΔE_{hull} values between different x compositions. These findings demonstrate that cationic decorations play a crucial role in determining perovskite properties. To successfully determine whether a given $AA'BB'O_3$ composition is stable as a perovskite, for example, we must consider ΔE_{hull} of the ground state cationic arrangement—or of a Boltzmann distribution of decorational states. If DFT results—or machine learning model predictions—do not account for decorations, properties may be calculated for structures that are not thermodynamically accessible, making them meaningless towards the goal of developing experimentally useful perovskites.

In contrast to the first two families, however, we find that decorational differences in $K_xBa_{1-x}Ti_4Al_4O_{24}$ have very little impact on ΔE_{hull} values at a given x concentration. While mean ΔE_{hull} values differ by 0.15 eV/atom between $x = 0.125$ and $x = 0.875$, ΔE_{hull} differs by <0.03 eV/atom between cationic arrangements of any given x. This suggests that these compositions have high entropy, making various arrangements thermodynamically accessible. For $K_xBa_{1-x}Ti_4Al_4O_{24}$, composition—not cationic arrangement—is thus the main determinant of DFT-calculated properties.

We note that similarly varied distributions of other properties, such as $O_{2p} - E_V$, are observed and thus decorations should be considered in all cases.

Together, these results highlight the need to consider cationic decorations when predicting the catalytic stability and activity of perovskites. To identify promising compositions, property optimization should be conducted over distributions of thermodynamically-accessible structures. The probability p_i that a perovskite will be in a cationic arrangement i may be given by the Boltzmann distribution

$$p_i = \frac{1}{Z} g_i e^{-\frac{E_i}{k_B T}}, \tag{3.1}$$

where g_i is the degeneracy of decoration i, E_i is the energy, k_B is the Boltzmann constant, T is temperature, and $Z = \sum_i g_i e^{-\frac{E_i}{k_B T}}$ is the partition function. Note that degeneracy g_i is the number of symmetrically equivalent structures corresponding to a given decoration i. We can then consider properties of multicomponent perovskite systems, $P_{AA'BB'O_3}$, to be Boltzmann-weighted averages

$$P_{AA'BB'O_3} = \sum_i p_i P_i, \tag{3.2}$$

where P_i is the property computed for cationic decoration i.

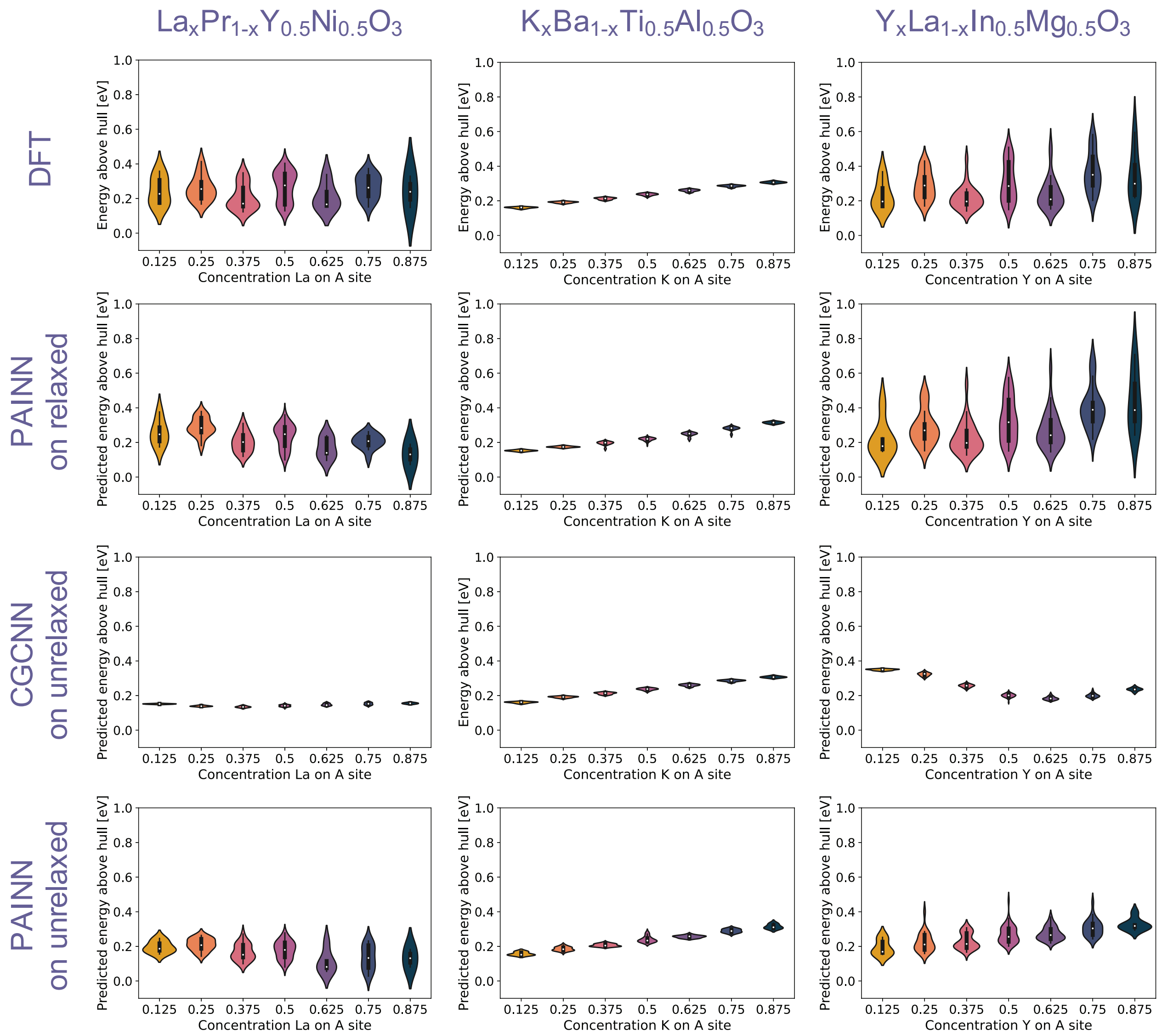

Figure 3-6: DFT-calculated and GNN-predicted distributions of ΔE_{hull} for $A_xA'_{8-x}B_4B'_4O_{24}$ perovskites with varied cationic decorations. All symmetrically-inequivalent structures of $A_xA'_{8-x}B_4B'_4O_{24}$ perovskites with the denoted compositions and $x \in [1, 7]$ were simulated with DFT. The distributions of ΔE_{hull} values, shown in the top row, across different cationic decorations vary among A, A', B, B' compositions. Models trained on relaxed perovskite structures reproduce these distributions, as shown in the second row. Note that performance of PAINN and CGCNN is approximately the same on relaxed structures. On unrelaxed data, CGCNN predicts approximately constant values for all decorations. PAINN on unrelaxed structures, by contrast, can capture impacts of decorational differences, albeit less successfully than the relaxed equivalent.

3.3.2 Capturing decorational differences with graph neural networks

Given the impacts of cationic arrangement differences presented in Section 3.3.1, accurate perovskite property prediction requires machine learning methods capable of capturing decorational effects. We therefore tested the performance of models described in Section 3.2 on the $A_xA'_{8-x}B_4B'_4O_{24}$ series shown in Figure 3-6. All CGCNN and PAINN models employed here were trained to predict deviations from ideal interpolation, with properties then computed using the steps outlined in Section 2.2.3. All models were trained on data described in Section 3.2, with a maximum of one $A_xA'_{8-x}B_4B'_4O_{24}$ structure included in the training data for any x.

CGCNN models trained on relaxed perovskite structures were able to reproduce distributions of DFT-calculated properties across different cationic decorations. Results for ΔE_{hull} are shown in Figure 3-7 and the second row of Figure 3-6, noting that results from CGCNN on relaxed structures generated approximately the same distributions as those shown from PAINN.

On unrelaxed perovskites, however, CGCNN predictions collapsed to a single value for all symmetrically-inequivalent structures of a given composition, seen for ΔE_{hull} predictions in Figure 3-6. Predictions typically corresponded approximately to the mean of the predicted property at a given composition. The failure of CGCNN to capture decoration-dependent variations is also seen as horizontal clustering in parity plots comparing DFT-calculated and CGCNN-predicted values. This clustering at constant ΔE_{hull} values predicted by CGCNN is displayed in Figure 3-7. This drastic decrease in performance may be attributed to the use of only scalar distance edge features in CGCNN. With relaxed structures, distorted distances between neighbouring atoms appear to encode sufficient information for CGCNN to differentiate decorational differences. With unrelaxed structures, however, the edge features are the same for every perovskite—the distances between a pair of atoms is the same across neighbours and across all structures regardless of decorational and/or compositional differences.

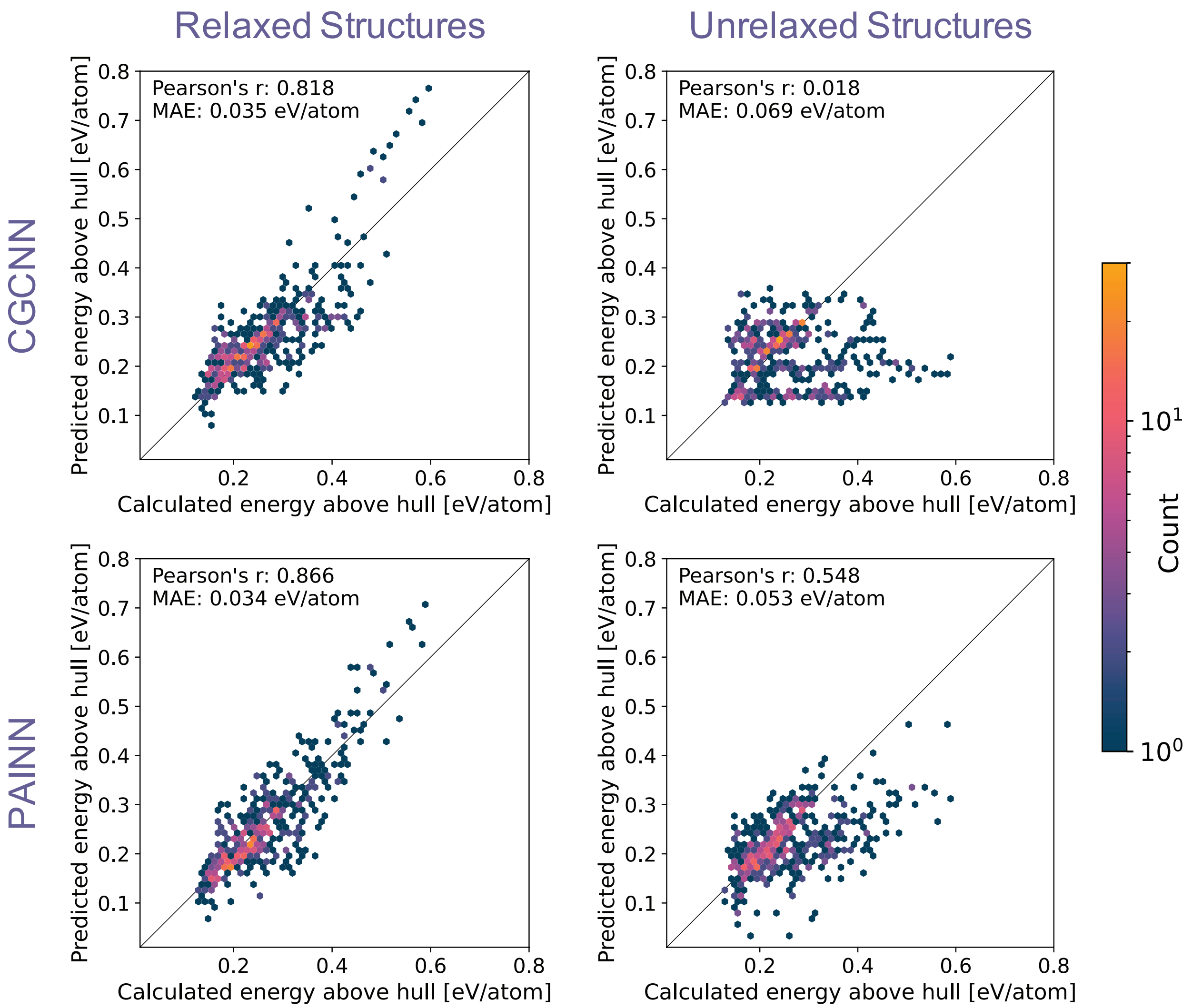

Figure 3-7: Capabilities of GNN models to predict ΔE_{hull} distributions across varied cationic decorations. CGCNN and PAINN models were trained to predict ΔE_{hull} from relaxed and from unrelaxed perovskite structures. Parity plots show the performance of each model on series of $A_xA'_{8-x}B_4B'_4O_{24}$ perovskites with varied symmetrically-inequivalent cationic arrangements. CGCNN and PAINN exhibit similar performance on relaxed structures, but PAINN predictions on unrelaxed perovskites outperform those of CGCNN. Horizontal clustering shown in the parity plot of CGCNN on unrelaxed structures reflects convergence to a constant prediction for all decorations of a given composition.

The failure of CGCNN models to capture decorational effects from unrelaxed structures motivates our implementation of PAINN models in perovskite systems. In contrast to CGCNN, PAINN models trained on unrelaxed structure were able to capture effects of decorational differences. As shown in the last row of Figure 3-6, PAINN successfully reflects that $La_xPr_{1-x}Y_4Ni_4O_{24}$ and $Y_xLa_{1-x}In_4Mg_4O_{24}$ compositions display differences in ΔE_{hull} across structures with different cationic arrangements while $K_xBa_{1-x}Ti_4Al_4O_{24}$ compositions do not. Although PAINN on unrelaxed structures does not achieve the same performance as models on relaxed structures, the horizontal clustering observed with CGCNN is not seen in the PAINN parity plots, see Figure 3-7.

These results demonstrate successful use of an equivariant message passing GNN to capture 3D information in crystal structures. Encoding directional information as graph edge features greatly improved the ability of PAINN models to differentiate cationic decorations in unrelaxed structures compared to CGCNN, whose edge features only encode scalar distances. Despite these promising initial results, however, further model improvement is desired to achieve results equivalent or superior to those obtained here with relaxed structures.

The ability to predict decorational-based distributions with the unrelaxed PAINN model makes it feasible to begin representing perovskite systems as thermodynamic ensembles and predicted properties as ensemble averages. In doing so, we aim to make our computational perovskite property modelling increasingly indicative of experimental results in the future.

3.4 Per-site property prediction

Models described in Sections 3.2 and 3.3 were used to predict bulk descriptors of catalytic stability and activity. It is often useful to obtain per-site descriptors as well, however, because in reality catalytic reactions on localized active sites. Moreover, understanding contributions to catalytic properties at an atomic level opens to door to more fine-tuned understanding and design of crystalline materials. We therefore implemented GNNs to predict properties at each site in a crystal, in contrast to the bulk models that yield one scalar output for each crystal.

Datasets used to train, validate, and test per-site property prediction models were compiled from the Materials Project [Jain et al., 2013] and the in-house HTVS calculations described in Section 3.1. Bader charge, magnetic moment, atomic vibration frequency (site-projected phonon band center), and site-projected O_{2p}- and metal d-band centers were considered due to both their data availability and promise as descriptors of catalytic activity.

A wide range of different crystal structures for each property was included in the training data, making the models applicable to a wide range of structures and stoichiometries.

The Bader charge dataset consists of all ~120,000 structures in the Materials Project with available charges and is therefore not limited to a given materials class. Here, Bader charges refer to the partial atomic charges on metal centers as calculated by Bader analysis [Tang et al., 2009]. Similarly, all ~10,000 structures in the Materials Project with available phonon calculations were used to generate the atomic vibration frequency dataset, where atomic vibration frequency refers to the band center of the site-projected phonon density of states.

The magnetic moment dataset is comprised of ~35,000 magnetic oxides from both the Materials Project and the perovskite datasets presented in Section 3.1. Structures with unphysical magnetic moments—those greater than 5 for d-band valence—were removed from the dataset. Moreover, crystals containing any atoms with f-band valence were left out, given the scarcity of data and limited accuracy of f-element

DFT pseudopotentials. Only ferromagnetic structures were considered, taking the absolute value of magnetic moments. Finally, the site-projected O_{2p}- and metal d-band center dataset is comprised of the ~10,000 perovskites described in Section 3.1, from both the literature and our HTVS calculations. These band centers are predicted simultaneously for the metal and oxygen atoms, respectively, in each structure.

Table 3.4: Per-site model performance

			MAE		
Property	# Train Crystals	Units	Per-site CGCNN	Per-site MLP	Per-element average
Atomic vibration frequency	5,899	THz	**0.817 ± 0.003**	1.025 ± 0.011	1.571
Metal d-band center	6,024	eV	**0.581 ± 0.002**	1.266 ± 0.017	1.281
O 2p-band center	6,024	eV	**0.303 ± 0.004**	0.579 ± 0.003	1.227
Magnetic moment	21,113	$\lvert\mu_B\rvert$	**0.185 ± 0.002**	0.377 ± 0.002	0.553
Bader charge	71,787	q_e	**0.068 ± 0.001**	0.147 ± 0.001	0.578

Performance of the per-site CGCNN models is summarized in Table 3.4. Results are compared to those from the per-site multilayer perceptron (MLP) model described in Section 2.3.2 as well as to a per-element average—the mean property value for each element over the respective train dataset. Per-site CGCNN was found to have the best performance for each property. The superior performance of per-site CGCNN compared to the MLP model, which takes all the same structural information as input but does not leverage graph representations or convolutions, reflects the power of message passing GNNs in encoding local environments.

Comparison of DFT-calculated and CGCNN-predicted properties on an element-wise basis demonstrates the ability of per-site CGCNN to explicitly learn physical principles dictated by local environments. For all row 4 transition metals in the test set, distributions of calculated and predicted Bader charge, magnetic moment, and per-site d-band center are compared in Figure 3-8a-c. Per-site CGCNN accurately captures both periodic trends across the row—obtainable from basic physical principles—as well as property distributions across each element in different materials—dictated by local chemical environments. Similar comparisons of O_{2p}-band centers are shown in Figure 3-8c and of per-site atomic vibration frequencies for the first 10 elements in the periodic table in Figure 3-8d.

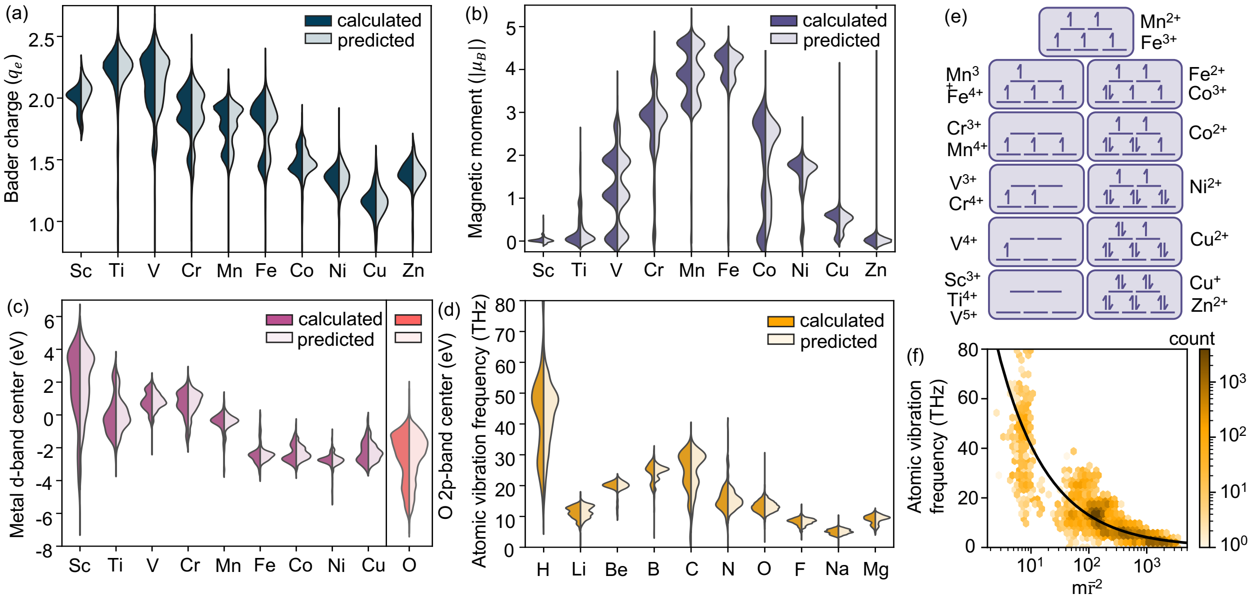

Figure 3-8: Elemental distributions of calculated and predicted per-site properties. Distributions of DFT-calculated values are compared to per-site CGCNN predictions for a variety of materials properties. Bader charges (a), magnetic moments (b), and site-projected metal d-band centers (c) are compared for each row 4 transition metal. Per-site atomic vibration frequencies are compared for the first ten elements of the periodic table (d). Calculated and predicted site-projected O_{2p}-band centers (c) are also compared. Per-site CGCNN predictions capture periodic trends and property distributions across each element in different chemical structures. Schematic (e) depicts the relation between d-band filling, oxidation states, and magnetic moments. Plot (f) models calculated per-site atomic vibration frequencies using a simple mass-on-a-spring analogy, as a function of mass m and average bonding length $\bar{r}$.

While periodic trends generally follow well-known physical principles, per-element distributions of material properties result from local chemical environments and are not always obvious from basic physical principles alone—thus highlighting the ability of per-site CGCNN capture local environments. Insights captured by the models are discussed for each property:

Bader charges

Per-site CGCNN captures the trend of decreasingly positive charges across the periodic table, from Sc to Zn, as electronegativity increases. The model also reproduces nodes within Bader charge distributions for each element, indicative of different oxidation states. The oxidation site of site i is traditionally estimated by the bond valence method as $\sum_j e^{\frac{R_0 - R_{ij}}{b}}$ over all neighbouring sites j, where R_{ij} is the distance between sites i and j, R_0 is a tabulated material system-specific bond valence parameter, and b is an empirical constant [O'Keefe and Brese, 1991]. Moreover, previous data-driven efforts have learned oxidation states from hand crafted features [Jablonka et al., 2021], a requirement which the per-site CGCNN mitigates by predicting Bader charges directly from crystal structure.

Magnetic moments

Broadly, per-site predictions capture band filling across the $3d$ row of the periodic table. Namely, as unpaired electrons fill the d-band from Sc to Mn, magnetic moments increase in discrete intervals of one Bohr magneton (μ_B). From Mn to Zn, however, the magnetic moment decreases as further d-band filling causes paired electrons, thus cancelling spins. Schematic 3-8e summarizes the relationship between magnetic moment, d-band filling, and oxidation state. Similarly to Bader charge model results, the magnetic moment CGCNN model learns oxidation state effects to reproduce the varied discrete magnetic moments for each element, as modulated by local chemical environments.

Site-projected electronic density of states

Per-site CGCNN reproduces known trends in d-band centers: as the number of $3d$ electrons increases across the periodic table, the d-band widens to maintain the Fermi level, thus pushing the d-band center to be more negative [Nørskov et al., 2014]. Indeed, the model also captures the distributio of O_{2p}-band centers in different environments.

Atomic vibration frequency

For the first 10 elements in the periodic table shown here, per-site atomic vibration frequency generally decays as atomic mass increases. In Figure 3-8f, a simple mass-on-a-spring analogy is used to model calculated vibration frequencies as a function of $m\bar{r}^2$, where m is site mass and $\bar{r}$ is average bonding length—indicative of bonding strength. A best fit line across the dataset is shown in black. Deviations from the best fit line reflect modulation of bond strengths by nearest neighbour environments, which are learned by the per-site CGCNN.

Together, these results underscore the ability of our model to learn per-site properties from local chemical environments, thus providing physical insights. Note, however, that all datasets used in this section consist of DFT-relaxed structures and performance on unrelaxed structures has yet to be tested. Moreover, following the success of PAINN in predicting bulk crystal properties, we will train models to predict per-site properties using the equivariant message passing architecture.

Chapter 4

Conclusions and Outlook

This thesis work developed improved computational modelling of multicomponent perovskites, leveraging DFT calculations, graph convolutional neural networks, and physical chemistry insights. Ultimately, this work contributes to the broader goal of developing inverse design frameworks for the discovery of improved multicomponent oxide electrocatalysts.

We presented a new DFT dataset of over 5,000 multicomponent perovskites, covering a wide compositional space and studying how cationic arrangements impact perovskite properties. This dataset was not only useful for our analyses and machine learning model development, but will also provide the research community with a significantly larger and broader benchmark dataset of highly-alloyed perovskite data. We are now also supplementing this dataset with calculations of defected structures, containing oxygen vacancies or oxynitride substitutions, to further our understandings of perovskite systems and expand the search space for useful electrocatalysts.

Graph neural networks previously developed in the literature were implemented and tailored to predict (a) bulk crystal properties from unrelaxed perovskites, and (b) per-site properties in a variety of crystalline systems. The former mitigates the need for expensive DFT structure optimization calculations, which enabled our predictions of catalytic descriptors for 1.2 million multicomponent perovskites. Using available binary ABO_3 data to interpolate baseline property estimates, PAINN equivariant message passing neural networks on unrelaxed structures outperformed the popular

CGCNN model on relaxed data. Meanwhile, per-site property prediction models allowed us to capture atomic level insights of DFT-derived properties and are now being used to model and design individual catalytic sites on perovskite surfaces.

Lastly, our analyses and models of the impacts of cation decorations in perovskite lattices enable us to consider perovskite systems as ensembles and thus predict properties of thermodynamically-relevant structures. Nevertheless, further optimization of GNNs capturing decorational effects from unrelaxed structure is needed, perhaps by employing E(3)-equivariant graph neural networks [Batzner et al., 2022, Geiger and Smidt, 2022]. The ability to capture decorational effects paves the way for inverse design frameworks that populate perovskite lattices site-by-site. For example, by representing perovskite structure as a decision tree, we have implemented a Monte Carlo tree search algorithm that assigns atoms to each cationic site with the objective of optimizing desired catalytic descriptors.

Bibliography

C. J. Bartel. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *Journal of Materials Science*, pages 1–24, 2022.

S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):1–11, 2022.

C. E. Beall, E. Fabbri, and T. J. Schmidt. Perovskite oxide based electrodes for the oxygen reduction and evolution reactions: the underlying mechanism. *ACS Catalysis*, 11(5):3094–3114, 2021.

I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, and K. W. Jacobsen. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy & Environmental Science*, 5 (10):9034–9043, 2012a.

I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science*, 5(2):5814–5819, 2012b.

S. Clark and P. Hayes. SigOpt Web page. `https://sigopt.com`, 2019. URL `https://sigopt.com`.

D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-

Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

A. A. Emery and C. Wolverton. High-throughput dft calculations of formation energy, stability and oxygen vacancy formation energy of abo3 perovskites. *Scientific data*, 4(1):1–10, 2017.

A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton. High-throughput computational screening of perovskites for thermochemical water splitting applications. *Chemistry of Materials*, 28(16):5621–5634, 2016.

E. Fabbri and T. J. Schmidt. Oxygen evolution reaction—the enigma in water electrolysis, 2018.

M. Geiger and T. Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15 (10):1120–1127, 2016.

A. Grimaud, K. J. May, C. E. Carlton, Y.-L. Lee, M. Risch, W. T. Hong, J. Zhou, and Y. Shao-Horn. Double perovskites as a family of highly active catalysts for oxygen evolution in alkaline solution. *Nature communications*, 4(1):1–7, 2013.

Q. Guo, J. Mao, J. Huang, Z. Wang, Y. Zhang, J. Hu, J. Dong, S. Sathasivam, Y. Zhao, G. Xing, et al. Reducing oxygen evolution reaction overpotential in cobalt-based electrocatalysts via optimizing the “microparticles-in-spider web” electrode configurations. *Small*, 16(8):1907029, 2020.

J. Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

W. T. Hong, K. A. Stoerzinger, Y.-L. Lee, L. Giordano, A. Grimaud, A. M. Johnson, J. Hwang, E. J. Crumlin, W. Yang, and Y. Shao-Horn. Charge-transfer-energy-dependent oxygen evolution reaction mechanisms for perovskite oxides. *Energy & Environmental Science*, 10(10):2190–2200, 2017.

M. L. Huggins and K.-H. Sun. Energy additivity in oxygen-containing crystals and glasses. *The Journal of Physical Chemistry*, 50(4):319–328, 1946.

K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit. Using collective knowledge to assign oxidation states of metal cations in metal–organic frameworks. *Nature Chemistry*, 13(8):771–777, 2021.

R. Jacobs, T. Mayeshiba, J. Booske, and D. Morgan. Material discovery and design principles for stable, high activity perovskite cathodes for solid oxide fuel cells. *Advanced Energy Materials*, 8(11):1702708, 2018.

R. Jacobs, J. Hwang, Y. Shao-Horn, and D. Morgan. Assessing correlations of perovskite catalytic performance with electronic structure descriptors. *Chemistry of Materials*, 31(3):785–797, 2019.

A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, 50(8):2295–2310, 2011.

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

G. King and P. M. Woodward. Cation ordering in perovskites. *Journal of Materials Chemistry*, 20(28):5785–5796, 2010.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

J. Klicpera, J. Groß, and S. Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science*, 6(1):15–50, 1996a.

G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B*, 54(16):11169, 1996b.

G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Physical review B*, 47(1):558, 1993.

G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b*, 59(3):1758, 1999.

G. Kresse, J. Furthmüller, and J. Hafner. Theory of the crystal structures of selenium and tellurium: the effect of generalized-gradient corrections to the local-density approximation. *Physical Review B*, 50(18):13181, 1994.

A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, K. W. Jacobsen, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

H. Lee, O. Gwon, K. Choi, L. Zhang, J. Zhou, J. Park, J.-W. Yoo, J.-Q. Wang, J. H. Lee, and G. Kim. Enhancing bifunctional electrocatalytic activities via metal d-band center lift induced by oxygen vacancy on the subsurface of perovskites. *ACS Catalysis*, 10(8):4664–4670, 2020.

Y.-L. Lee, J. Kleis, J. Rossmeisl, Y. Shao-Horn, and D. Morgan. Prediction of solid oxide fuel cell cathode activity with first-principles descriptors. *Energy & Environmental Science*, 4(10):3966–3970, 2011.

W. Li, R. Jacobs, and D. Morgan. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Computational Materials Science*, 150:454–463, 2018.

X. Liao, R. Lu, L. Xia, Q. Liu, H. Wang, K. Zhao, Z. Wang, and Y. Zhao. Density functional theory for electrocatalysis. *Energy & Environmental Materials*, 5(1): 157–185, 2022.

T. T. Mayeshiba and D. D. Morgan. Factors controlling oxygen migration barriers in perovskites. *Solid State Ionics*, 296:71–77, 2016.

D. N. Mueller, M. L. Machala, H. Bluhm, and W. C. Chueh. Redox activity of surface oxygen anions in oxygen-deficient perovskite oxides during electrochemical reactions. *Nature communications*, 6(1):1–8, 2015.

J. K. Nørskov, F. Studt, F. Abild-Pedersen, and T. Bligaard. *Fundamental concepts in heterogeneous catalysis*. John Wiley & Sons, 2014.

M. O'Keefe and N. Brese. Atom sizes and bond lengths in molecules and crystals. *Journal of the American Chemical Society*, 113(9):3226–3229, 1991.

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

C. W. Park and C. Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6):063801, 2020.

K. A. Persson, B. Waldwick, P. Lazic, and G. Ceder. Prediction of solid-aqueous equilibria: Scheme to combine first-principles calculations of solids with experimental aqueous states. *Physical Review B*, 85(23):235438, 2012.

B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

J. Rossmeisl, Z.-W. Qu, H. Zhu, G.-J. Kroes, and J. K. Nørskov. Electrolysis of water on oxide surfaces. *Journal of Electroanalytical Chemistry*, 607(1-2):83–89, 2007.

J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65(11):1501–1509, 2013.

S. Sanyal, J. Balachandran, N. Yadati, A. Kumar, P. Rajagopalan, S. Sanyal, and P. Talukdar. Mt-cgcnn: Integrating crystal graph convolutional neural network with multitask learning for material property prediction. *arXiv preprint arXiv:1811.05660*, 2018.

K. Schütt, O. Unke, and M. Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9377–9388. PMLR, 18–24 Jul 2021.

A. Shinde, S. K. Suram, Q. Yan, L. Zhou, A. K. Singh, J. Yu, K. A. Persson, J. B. Neaton, and J. M. Gregoire. Discovery of manganese-based solar fuel photoanodes via integration of electronic structure calculations, pourbaix stability modeling, and high-throughput experiments. *ACS Energy Letters*, 2(10):2307–2312, 2017.

A. K. Singh, L. Zhou, A. Shinde, S. K. Suram, J. H. Montoya, D. Winston, J. M. Gregoire, and K. A. Persson. Electrochemical stability of metastable materials. *Chemistry of Materials*, 29(23):10159–10167, 2017.

W. Tang, E. Sanville, and G. Henkelman. A grid-based bader analysis algorithm without lattice bias. *Journal of Physics: Condensed Matter*, 21(8):084204, 2009.

Q. Tao, P. Xu, M. Li, and W. Lu. Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(1):1–18, 2021.

T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pages 1–9, 2014.

A. Vazhayil, L. Vazhayal, J. Thomas, N. Thomas, et al. A comprehensive review on the recent developments in transition metal-based electrocatalysts for oxygen evolution reaction. *Applied Surface Science Advances*, 6:100184, 2021.

T. Xie and J. C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018.

K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.