UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS
DOUTORADO EM CIÊNCIA DOS MATERIAIS

EXPLORING $Cs_3Sb_2X_9$-TYPE PEROVSKITES (X = Cl, Br, I) FOR OPTOELECTRONIC APPLICATIONS: A THEORETICAL INVESTIGATION USING DENSITY FUNCTIONAL THEORY (DFT) AND MACHINE LEARNING

By

ROGÉRIO ALMEIDA GOUVÊA

Porto Alegre, May 2024

EXPLORING $Cs_3Sb_2X_9$-TYPE PEROVSKITES (X = Cl, Br, I) FOR OPTOELECTRONIC APPLICATIONS: A THEORETICAL INVESTIGATION USING DENSITY FUNCTIONAL THEORY (DFT) AND MACHINE LEARNING

By

ROGÉRIO ALMEIDA GOUVÊA

Doctoral Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Materials Science in the Graduate Program of Materials Science (PPGCIMAT)

Supervisor: Dr. Marcos José Leite Santos

Co-supervisor: Dr. Mário Lúcio Moreira

Porto Alegre, May 2024

CIP - Catalogação na Publicação

Gouvêa, Rogério Almeida
EXPLORING Cs3Sb2X9-TYPE PEROVSKITES (X = Cl, Br, I)
FOR OPTOELECTRONIC APPLICATIONS: A THEORETICAL
INVESTIGATION USING DENSITY FUNCTIONAL THEORY (DFT)
AND MACHINE LEARNING / Rogério Almeida Gouvêa. --
2024.
319 f.
Orientador: Marcos José Leite Santos.

Coorientador: Mário Lúcio Moreira.

Tese (Doutorado) -- Universidade Federal do Rio
Grande do Sul, Instituto de Física, Programa de
Pós-Graduação em Ciência dos Materiais, Porto Alegre,
BR-RS, 2024.

1. metal halide perovskites. 2. density functional
theory. 3. 2D perovskites. 4. machine learning. 5.
inorganic perovskites. I. Leite Santos, Marcos José,
orient. II. Moreira, Mário Lúcio, coorient. III.
Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Thesis examination panel

Prof. Gian-Marco Rignanese – Université Catholique de Louvain

Dr. Amauri Jardim de Paula – Universidade Federal do Ceará

Dr. Fabiano Bernardi – Universidade Federal do Rio Grande do Sul

Prof. Naira Maria Balzaretti – Universidade Federal do Rio Grande do Sul

Supervisor Dr. Marcos José Leite Santos – Universidade Federal do Rio Grande do Sul

Co-supervisor Dr. Mário Lúcio Moreira – Universidade Federal de Pelotas

In memory and longing for my mother, Sônia Marisa.

ACKNOWLEDGEMENTS

I would like to express my gratitude firstly to my family for their encouragement and support, without which I could not have dedicated myself to an academic career with its constant uncertainties and challenges. To my supervisor, I am thankful for their proactivity and continuous support of my development, encouraging me to devise projects, enhance my writing, and broaden my perspective as a researcher. I am also grateful to my previous mentors who believed in my potential and helped me mature professionally to face the challenges of the doctoral journey. To my colleagues at UFRGS, besides the pleasant company, I appreciate the opportunity to collaborate and learn in various projects. My gratitude also extends to all the members of the examining committee for their availability and contributions.

I would also like to express my gratitude to Prof. Gian-Marco for accepting me at UCLouvain, which was instrumental in enabling me to develop this thesis to the extent I had envisioned. Additionally, I extend my thanks to Pierre-Paul, my colleague at UCLouvain, for providing guidance on the inner workings and application of MODNet. I am also grateful for the warmth of the crew at MODL/UCLouvain, which made enduring the Belgian winter much easier.

Finally, I thank CNPq for funding this work in Brazil, CAPES for the CAPES/Print Scholarship for the work in Belgium, and the computational resources of CESUP/UFRGS, essential for the development of this work.

ABSTRACT

The most characteristic subgroup of 2D metal halide inorganic perovskites (MHIPs) is composed of perovskites with ordered vacancies along the <111> direction, with the chemical formula $A_3B_2X_9$ (where A is a monovalent cation, B is a trivalent cation such as Bi^{3+} or Sb^{3+}, and X is a halide anion). Due to their low toxicity, remarkable optoelectronic properties, and long-term stability, these structures have attracted considerable attention. They may potentially replace lead halide perovskites, which are highly toxic and sensitive to moisture, while also addressing the challenge of limited carrier generation and transport in 2D MHIPs with organic spacers. This thesis delves into a comprehensive theoretical investigation of the most studied representatives of this material class, $Cs_3Sb_2X_9$ (X= Cl, Br, I) (space group: $P\bar{3}m1$). Through our investigation, we reveal that halide mixing can significantly influence band gap variations and structural shifts, presenting potential ordered structures. We also found that in these materials (1000) surfaces retain beneficial electronic properties for photovoltaics, while (0001) surfaces exhibit reactivity suitable for photocatalysis. Additionally, the band alignments of $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ interface and defect tolerance in $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ interface highlight potential applications in LEDs and photovoltaics, respectively. Expanding our study, we examined transition metal and halogen doping in both polymorphs of $Cs_3Sb_2I_9$ (space groups: $P\bar{3}m1$ and $P6_3/mmc$), the lowest band gap perovskite in this group. We discovered that indium doping enhances optical absorption and stability, while scandium doping stabilizes the lattice with minimal band gap increase, suggesting methods to reduce Urbach energy and improve device performance. Utilizing the capabilities of the machine learning model, Materials Optimal Descriptor Network (MODNet), augmented with a new featurizer for enhanced accuracy, we conducted an extensive exploration of the chemical space for this material class. This included multi-element doping, predicting the formability of new compounds, and identifying stabilizing elements. Our machine learning workflow screened over 100 million candidate structures, identifying promising ternary compounds including $Cs_3Ga_2Br_9$ and $Rb_3Cr_2Br_9$ with lower band gaps than commonly studied perovskites, and suggesting mixed A-cations and anions as potential stabilizers.

Keywords: metal halide inorganic perovskites, 2D perovskites, density functional theory, machine learning.

RESUMO

O subgrupo mais característico das perovskitas inorgânicas de haleto metálico 2D (PIHM) é composto por perovskitas com vacâncias ordenadas ao longo da direção <111>, com a fórmula química $A_3B_2X_9$ (onde A é um cátion monovalente, B é um cátion trivalente como Bi^{3+} ou Sb^{3+}, e X é um ânion haleto). Devido à sua baixa toxicidade, notáveis propriedades optoeletrônicas e estabilidade, essas estruturas têm atraído considerável atenção. Elas podem potencialmente substituir as perovskitas de haleto de chumbo, que são tóxicas e sensíveis à umidade, abordando o desafio da geração e transporte limitados de portadores em PIHMs 2D com espaçadores orgânicos. Esta tese investiga teoricamente os materiais, $Cs_3Sb_2X_9$ (X= Cl, Br, I) (grupo espacial: $P\bar{3}m1$). Através da nossa investigação, revelamos que a mistura de haletos pode influenciar significativamente as variações de *band gap* e mudanças estruturais, apresentando estruturas ordenadas potenciais. Também descobrimos que, nesses materiais, as superfícies (1000) mantêm propriedades eletrônicas benéficas para fotovoltaicos, enquanto as superfícies (0001) exibem reatividade adequada para fotocatálise. O alinhamento das bandas de $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ e a tolerância a defeitos de $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ sugerem aplicações em LEDs e fotovoltaicos, respectivamente. Também examinamos a dopagem de metais de transição e halogênios em ambos os polimorfos de $Cs_3Sb_2I_9$ (grupos espaciais: $P\bar{3}m1$ e $P6_3/mmc$), a perovskita de menor *band gap* deste grupo. A dopagem com índio aumenta a absorção óptica e a estabilidade, e a dopagem com escândio estabiliza a rede cristalina com aumento mínimo do *band gap*, sugerindo métodos para melhorar o desempenho do dispositivo. Usando o modelo de aprendizado de máquina MODNet, aprimorado com um novo gerador de descritores, exploramos extensivamente o espaço químico desta classe de materiais. Isso incluiu dopagem multi-elemento, previsão da formabilidade de novos compostos e identificação de elementos estabilizadores. Nosso fluxo de trabalho de aprendizado de máquina analisou mais de 100 milhões de estruturas, identificando compostos ternários promissores, incluindo $Cs_3Ga_2Br_9$ e $Rb_3Cr_2Br_9$, com *band gaps* mais baixos do que as perovskitas comumente estudadas, e sugerindo cátions A mistos e ânions como potenciais estabilizadores.

Palavras-chave: perovskitas inorgânicos de haleto de metal, perovskitas 2D, teoria do funcional da densidade, aprendizado de máquina.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACBN0 – Agapito-Curtarolo-Buongiorno Nardelli

AI – Artificial intelligence

AIMD – Ab-Initio Molecular Dynamics

AL – Active Learning

AUCROC – Area Under the Receiver Operating Curve

CB – Conduction Band

c.r. – Compression ratio

VB – Valence Band

CBM – Conduction Band Minimum

VBM – Valence Band Maximum

CHGNet – Crystal Hamiltonian Graph Neural Network

CIF – Crystallographic Information File

DFT – Density Functional Theory

(P)DOS – (Partial) Density of States

GBRV – Garrity, Bennett, Rabe, Vanderbilt

GGA – Generalized Gradient Approximation

GCN – Graph Convolution Network

GNN – Graph Neural Network

HSE06 – Heyd, Scuseria, Ernzerhof hybrid functional

HF – Hartree-Fock

HK – Hohenberg-Kohn

HOMO – Highest Occupied Molecular Orbital

HT – High-throughput (HT)

HT-DFT – High-throughput density functional theory calculations

IBZ – Irreducible Brillouin Zone

KS – Kohn-Sham

LDA – Local Density Approximation

LED – Light Emitting Diode

LIHP – Lead-free Inorganic Halide Perovskite

LUMO – Lowest Unoccupied Molecular Orbital

MA – Methylammonium, CH_3NH_3

M3GNet – Material Graph with Three-body Interactions Neural Network

MODNet – Materials Optimal Descriptor Network

MEGNet – MatErials Graph Network

MHIP – metal halide inorganic perovskite

ML – Machine Learning

MLP – Multi-layer perceptron

MP – Materials Project

NSCF – Non-self-consistent Field

OFM – Orbital Field Matrix

OMEGA – Encoded OFM + Pre-trained MEGNet + Adjacent MEGNet Models

OQMD – Open Quantum Materials Database

PAW – Projector Augmented-Wave

PBE – Perdew, Burke, Ernzerhof

PCA – Principal Component Analysis

PCE – Power Conversion Efficiency

PLQY – Photoluminescence Quantum Yield

PP – Pseudopotential

PSC – Perovskite Solar Cell

RF – Random Forest

ROSA – Robust One-Shot Ab-initio

SCF – Self-consistent Field

SHAP – SHapley Additive exPlanations

SOAP – Smooth Overlap of Atomic Positions

VASP – Vienna Ab initio Simulation Package

XC – Exchange-correlation

SUMMARY

CHAPTER 1 — INTRODUCTION

Currently, the world faces a crisis of energy insecurity with aggravating trends due to an increase in energy consumption, increasing fossil fuel prices, and geopolitical conflicts. Furthermore, the environmental and social effects of global warming have further stimulated the search for clean and renewable sources of energy. However, despite strides in renewable energy, projections indicate that about 1.2 billion people could face displacement by more frequent natural disasters in a world 2°C hotter by 2050 (Bellizzi et al. 2023; Diffenbaugh and Barnes 2023). Of all renewable energy sources, solar energy is the most prominent player in ensuring long-term energy security and mitigating the effects of global warming by offering a solution to fossil fuel emissions. Most commercial solar panels use silicon as a light collector, however, alternative absorber materials with perovskite crystal structure have emerged with great potential due to low cost, lightweight and ease of processing (J. Yu et al. 2022; Mohammad and Mahjabeen 2023).

Perovskite solar cells (PSCs) have already achieved photoconversion efficiencies of 26.1% in 2023, comparable to the best silicon technologies (NREL 2023). However, these results are based on hybrid organic-inorganic methylammonium lead iodide ($CH_3NH_3PbI_3$ or $MAPbI_3$) perovskite and these are still not able to replace silicon modules because they suffer from low stability to heat and humidity arising from the organic components (Conings et al. 2015; B.W. Park and Seok 2019; Miyasaka et al. 2020) and rely on the presence of lead in their structure, which has high toxicity (Xin Li et al. 2021). Lead can cause severe damage to ecosystems, soil, water sources, and human health, leading to functional disorders in the nervous, digestive, and blood systems (M. Wang et al. 2021; Hailegnaw et al. 2015). Moreover, the efficient entry of lead from perovskite materials into the food chain emphasizes the need for stricter safety standards regarding lead content in perovskite-based solar cells (S.-Y. Bae et al. 2019; Junming Li et al. 2020).

By replacing organic cations with inorganic counterparts, such as cesium, all-inorganic perovskites demonstrate higher intrinsic stability, making them more resistant to environmental factors (Tai, Tang, and Yan 2019). For the substitution of lead, theoretical calculations based on Density Functional Theory (DFT) have provided evidence linking the exceptional optoelectronic properties of hybrid lead-halide perovskites to the presence of the $5s^2$ lone pair in Pb^{2+} (Fabini, Seshadri, and

Kanatzidis 2020; Brandt et al. 2015; Filippetti and Mattoni 2014). Therefore, materials that contain a lone $6s^2$ or $5s^2$ pair of electrons in the cation can potentially share the high dielectric constant, low effective masses, and valence band (VB) antibonding character yielding defect tolerant transport properties. These compounds, which fall under a broad category, are created from partially oxidized post-transition metals and are arranged in ascending order of the relative stability of the lone-pair *s* orbitals, as follows: $In^+ < Tl^+ < Sn^{2+} < Pb^{2+} < Sb^{3+} < Bi^{3+} < Te^{4+} < Po^{4+}$ (Brandt et al. 2015; Fabini, Seshadri, and Kanatzidis 2020). These elements may constitute BX_6 octahedra, which are characteristic of perovskites. The X anion in these octahedra is one of the halogens, typically Cl, Br, or I, and can adopt various structural arrangements depending on relative size of the ions (Fakharuddin et al. 2019).

The first report of a lead-free inorganic halide perovskite (LIHP) solar cell used Sn^{2+} as an homovalent replacement to Pb^{2+} in the compound $CsSnI_3$ which presented better phase stability than the inorganic lead-based $CsPbI_3$ and also an ideal bandgap of approximately 1.3 eV (Kumar et al. 2014). However, Sn^{2+} readily oxidizes to Sn^{4+} which severely limits their stability and performance causing large concentrations of vacancies in the perovskite films and encouraging faster degradation (M. Liu et al. 2020; Fakharuddin et al. 2019).

Alternatively, +4 oxidation states cations can also work as substitute for Pb^{2+} in LIHPs. For example, in $A_2B^{4+}X_6$ type perovskites, Pb^{2+} is replaced by the combination of a B-vacancy and a B^{4+} cation. Thus, the crystal structure is a double perovskite consisting of two sublattices where the octahedral centers are occupied by vacancies and B^{4+} cations and thus are termed “vacancy ordered” perovskites. Most popular compound in this category is $Cs_2SnI_{6-x}Br_x$ which band gap can be tuned from 1.3 to 2.9 eV by increasing Br content and has significantly improved stability compared to $CsSnI_3$, however, their light absorption is quite limited leading to unsatisfactory performance as a photovoltaic material (Jin Zhang et al. 2023; Umedov et al. 2021). Similar problem also occurs with Cs_2TeI_6 although Te^{4+} preserves the ns^2 lone pair (Maughan et al. 2016).

Another variation of the double perovskite structure involves $A_2B^{1+}B^{3+}X_6$ compounds, where Pb^{2+} is substituted with a combination of a monovalent and a trivalent cation, maintaining the 3D structure seen in conventional $AB^{2+}X_3$ perovskites (refer to *Figure 1*). In these perovskites, B^{1+} is frequently occupied by Cu^+, Ag^+, or Au^+, while a trivalent element with a lone-pair feature, such as Bi^{3+}, Sb^{3+}, or In^{3+}, typically takes the B^{3+} positions. Ab-initio calculations have proven instrumental in predicting possible B^{1+} and B^{3+} combinations and comprehending their properties unveiling highly adaptable carrier effective masses and optical gaps within the visible spectrum (Volonakis et al. 2016; 2017), some well-studied materials in this group include Cs_2AgBiX_6 and $Cs_2InAgCl_6$. However, despite their initial promise, they exhibit indirect bandgaps or parity-forbidden direct gaps and diminished electronic dimensionality, making them unsuitable for solar cell applications (Fakharuddin et al. 2019). $Cs_2InBiCl_6$ and $Cs_2CuInBr_6$ also showcased favorable characteristics such as low bandgaps, small carrier effective masses, and high absorption coefficients. However, the inherent instability of In^+ and Cu^+, tending to transition to In^{3+} and Cu^{2+} oxidation states, hinders their practical application (Xiao et al. 2017; Bala and Kumar 2021). To this day, no experimentally confirmed double perovskite material has been identified as a promising candidate for solar cells. Nevertheless, recent studies offer hope for enhancing stability and achieving suitable direct band gaps by reducing the dimensionality of these perovskites with organic spacers to slice the 3D structure in 2D confined layers (Bala and Kumar 2021; Connor et al. 2023).

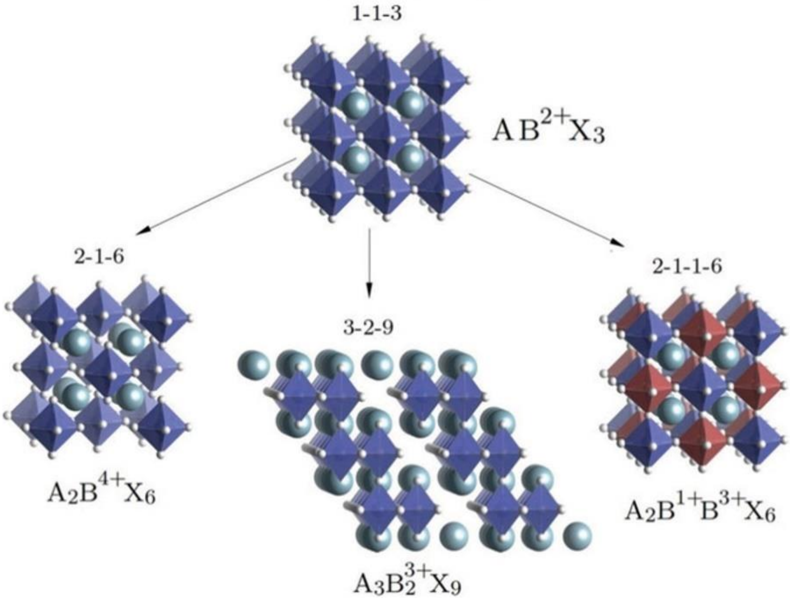


Figure 1 – Schematic relation between the crystal structures of Pb-perovskites ($AB_2^+X_3$) and lead-free perovskite derivatives. Source: (Giustino and Snaith 2016)

A more recently explored variation of LIHP for photovoltaics are the $A_3B^{3+}{}_2X_9$ type compounds, where B is typically Sb^{3+} or Bi^{3+}. These compounds naturally exhibit several structural dimensionalities to incorporate the trivalent cations. Their crystal structures can range from 0D dimer units to 1D chain-like motifs and 2D layered networks (Hoefler, Trimmel, and Rath 2017). Unlike traditional perovskites, where the Goldschmidt tolerance factor is used to assess formability, it cannot be applied to these compounds due to the gradual relaxation of ionic size restrictions as dimensionality decreases (Saparov and Mitzi 2016). The structural versatility of these bismuth and antimony compounds gains further appeal when considering that these perovskites present low toxicity, outstanding stability in ambient atmospheric conditions, long carrier diffusion lifetime, and large light absorption coefficient (Z. Jin et al. 2020). Particularly, antimony perovskites demonstrate energy levels most similar to Pb^{2+} (Xiao et al. 2017), along with higher absorption, smaller effective masses, and lower exciton binding energies compared to bismuth perovskites (B.-W. Park et al. 2015; Chonamada, Dey, and Santra 2020). With its substantial reserves and an annual production of 53,000 tons (Tan et al. 2019), Sb-based perovskites emerge as a highly promising, eco-friendly alternative for lead-based systems in various optoelectronic applications (Thomas 2022).

A prime example of the potential of Sb-based perovskites is the all-inorganic $Cs_3Sb_2X_9$ (X = Cl, Br, or I), which achieved a power conversion efficiency (PCE) of 3.25% (Singh et al. 2021) utilizing $Cs_3Sb_2I_9$ as an absorber, outperforming the more widely researched $Cs_3Bi_2I_9$ (Z. Jin et al. 2020; A. Wang et al. 2023). These perovskites have proven valuable also in optoelectronic applications such as LEDs (light-emitting diode) (A. Wang et al. 2023), with $Cs_3Sb_2Br_9$ quantum dots (QDs) developed by Ma et al. (Ma et al. 2019) emerging as a leading solution for short-wavelength violet emission. Concerning the iodine perovskite $Cs_3Sb_2I_9$, two distinct polymorphs exist: a 2D layered form (space group $P\overline{3}m1$), which is more popular due to its superior transport properties, and a 0D dimeric form (space group $P6_3/mmc$) (Saparov et al. 2015), as illustrated in *Figure 2*. However, the presence of localized charge carriers in 0D $Cs_3Sb_2I_9$, coupled with its efficient self-trapped exciton (STE) emission, renders it an exceptional material for optoelectronic devices (Saidaminov et al. 2016; D. Chen et al. 2016). The dimeric $Cs_3Sb_2I_9$ also presented the best results for random-access memory (ReRAM) devices among other 696 investigated compounds (Y. Park et al.

2021). Moreover, combining perovskites with different dimensionalities has shown to enhance efficiency and stability in PSCs, including the lead-free 0D $Cs_3Bi_2I_9$/2D BiI_3 (Jena, Kulkarni, and Miyasaka 2019; Masawa et al. 2022). As of now, there are no known reports on mixed dimensionality Sb-based perovskites. These findings collectively underscore the potential of $Cs_3Sb_2X_9$ compounds, despite their recent exploration, as materials for the next generation of eco-friendly photovoltaic and optoelectronic devices.

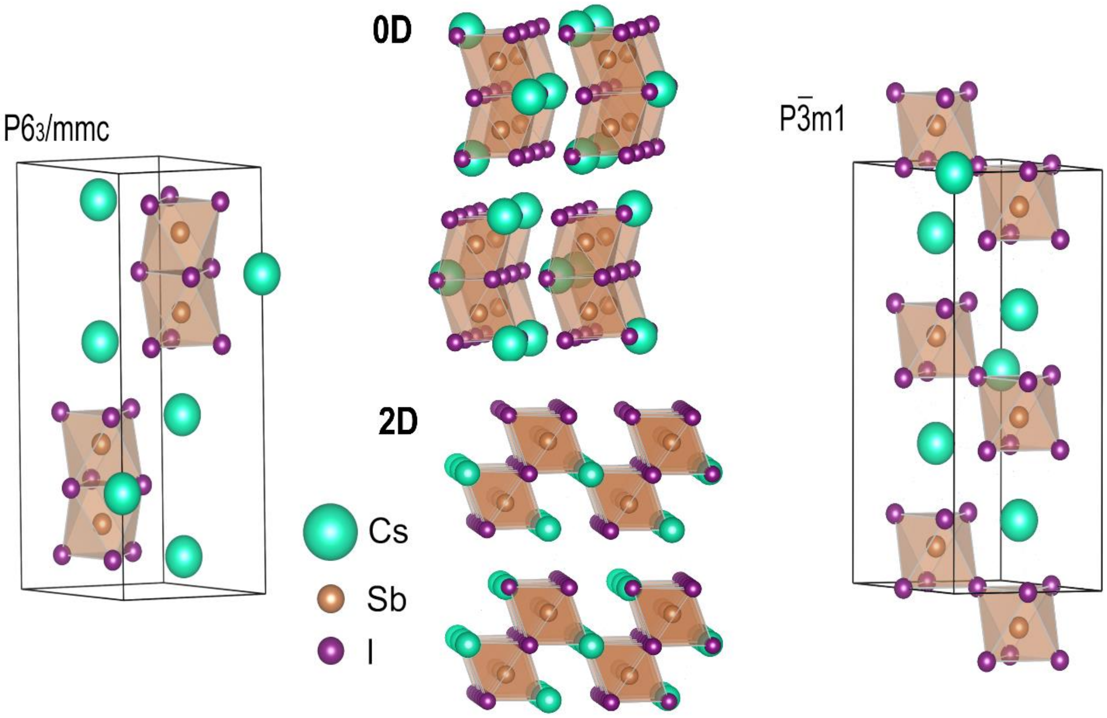


Figure 2 – Crystal structures of both dimeric (0D) and layered (2D) polymorphs of $Cs_3Sb_2I_9$. Layered form is also present for $Cs_3Sb_2Br_9$ and $Cs_3Sb_2Cl_9$ compounds.

In the course of this thesis, a collection of computational studies has been conducted on $Cs_3Sb_2X_9$ compounds and related structures, the research aimed to explore the optoelectronic properties, long-term stability, and potential for replacing lead halide perovskites, while addressing challenges such as limited carrier generation and transport in 2D MHIPs. Additionally, the study sought to investigate the effects of transition metal and halogen doping, utilize machine learning models to extensively explore the chemical space of these materials, and identify new promising compounds and stabilizing elements to optimize the performance of optoelectronic devices. Firstly, a theoretical investigation was performed on $Cs_3Sb_2X_9$ (X= Cl, Br, I) perovskites systematically exploring halogen doping, surface properties, and quantum confinement, as discussed in *Chapter 3*. Subsequently, the investigation delved into metal doping within the $Cs_3Sb_2I_9$ compound, considering both polymorphs, as detailed in *Chapter 4*. Finally, the power of machine learning and the vast materials databases available today were harnessed to develop a multi-model methodology. This methodology was used to explore potential lead-free structures following the same layered structure as $Cs_3Sb_2X_9$ while allowing the investigation of doping in multiple sites, as described in *Chapter 6*. This investigation is preluded, on Chapter 5, by the implementation of additional features on the machine learning framework MODNet (Materials Optimal Descriptor Network). These features could enhance the accuracy of the network in various tasks and were therefore incorporated into the perovskite screening process conducted.

In the upcoming chapter, the theoretical framework necessary to comprehend the methodology and results of each of the conducted investigations will be established. The interplay of material properties, essential for achieving high-performance photovoltaic and optoelectronic devices, will be elucidated. We will examine how ab-initio methods, particularly Density Functional Theory (DFT), can assist in estimating these properties and explain their underlying origins, thus promoting the development of novel materials through a deeper understanding. Finally, an introduction to how machine learning operates will be provided, and the important machine learning models employed in this work within the context of materials science will be described.

CHAPTER 2 — THEORETICAL FRAMEWORK

2.1 Semiconductors, halide perovskites and optoelectronic properties

This section discusses key ideas in semiconductor physics that underpin the whole inquiry presented in this thesis, linking atomic structure to material properties and device optimization for practical applications. We also explore the concepts of stability and synthesizability, which are critical for materials discovery, before closing with a general overview of halides perovskites properties and what yields their status as a leading-edge materials class in optoelectronics.

2.1.1 Condensed matter and semiconductor physics

The fundamental quantum description is essential to comprehend the properties of materials. Quantum mechanics sheds light on the interplay of features in the atomic scale to physical properties of technological interest by virtue of the operator-observable relationship. This principle is a fundamental aspect of quantum mechanics, as it establishes the connection between the abstract mathematical entities that model the behavior of the subatomic particles, named wavefunctions, and the measurable properties of physical systems (Griffiths and Schroeter 2018). In this framework, a material is a system comprising M nuclei, N electrons and their interactions, all described by the many-body Hamiltonian (Cramer 2013) given by:

$$\begin{aligned}\hat{H} = \sum_{i=1}^{N} -\frac{1}{2}\nabla^2_{\mathbf{r}_i} + \sum_{\alpha=1}^{M} -\frac{1}{2M_\alpha}\nabla^2_{\mathbf{R}_\alpha} + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \\ + \sum_{\alpha=1}^{M}\sum_{\beta=\alpha+1}^{M}\frac{Z_\alpha Z_\beta}{|\mathbf{R}_\alpha - \mathbf{R}_\beta|} - \sum_{\alpha=1}^{M}\sum_{i=1}^{N}\frac{Z_\alpha}{|\mathbf{R}_\alpha - \mathbf{r}_i|},\end{aligned} \tag{1}$$

in this case, the system is described using atomic units, where fundamental physical quantities such as electron mass, electron charge, reduced Planck constant, and Coulomb constant ($^1/_{4\pi\epsilon_0}$) have a unit value. The Hamiltonian (Eq. 1) consists of several terms: the first and second terms represent the kinetic energy of electrons ($\hat{T}_e$) and nuclei ($\hat{T}_N$) respectively, while the third, fourth, and fifth terms account for the Coulomb interactions between electrons ($\hat{V}_{ee}$), between nuclei ($\hat{V}_{NN}$), and between nuclei and electrons ($\hat{V}_{Ne}$), respectively. At the heart of quantum mechanics, the Schrödinger equation stands as the fundamental equation directing the evolution of quantum systems, expressed as follows:

$$i\hbar\frac{\partial\Psi(\{\mathbf{r}_i\},t)}{\partial t}=\hat{H}\Psi(\{\mathbf{r}_i\},t). \quad (2)$$

Here, $\Psi(\{\mathbf{r}_i\},t)$ represents the wavefunction of the system, which depends on the spatial coordinates of every particle in the system ($\{\mathbf{r}_i\}$) and time (t). In various scenarios, such as when examining ground-state properties, the stationary version can be utilized by factoring out the temporal component from equation 2, resulting in:

$$\hat{H}\Psi(\{\mathbf{r}_i\})=E\Psi(\{\mathbf{r}_i\}), \quad (3)$$

The wavefunction stands as a complete descriptor of a quantum system's state, encoding not only the probability distribution of particle positions over time (given by $|\Psi(\{\mathbf{r}_i\},t)|^2$) but also of any observable properties Q. These observables are derived through the action of their respective operators ($\hat{Q}$) on the wavefunction, as follows:

$$\hat{Q}\ |\Psi\rangle\ \rightarrow\ q_j\ |\phi_j\rangle. \quad (4)$$

Here, the ket symbol ($|\rangle$) denotes quantum states within the Dirac notation, representing vectors in the complex Hilbert space of all possible system states (Griffiths and Schroeter 2018). When measuring $\hat{Q}$, the system's wavefunction, initially in a superposition of eigenstates $|\phi_i\rangle$ for the observable Q, collapses into a specific eigenstate $|\phi_j\rangle$.This collapse yields the eigenvalue q_j as the measured property's outcome. Thus, even if the wavefunction operates at the quantum level, it ultimately shapes and dictates the macroscopic physical properties.

Because quantum mechanics is intrinsically statistical, identical particles, like electrons, are fundamentally indistinguishable in the sense that they cannot be uniquely labeled or differentiated by their intrinsic properties. This indistinguishability shapes their behavior, introducing statistical properties absent in classical physics. Particles such as electrons, protons, and neutrons fall into the category of fermions. These adhere to Fermi-Dirac statistics, governed by the Pauli exclusion principle. This principle asserts that identical fermions cannot occupy the same quantum state simultaneously, rooted in the antisymmetric property of their wavefunction. Mathematically, this is expressed as the requirement for the total wavefunction of these particles to change sign when any two particles are exchanged, such as:

$$\Psi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_i,\ldots,\vec{x}_j,\ldots,\vec{x}_N)=-\Psi(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_j,\ldots,\vec{x}_i,\ldots,\vec{x}_N), \quad (5)$$

where $\vec{x}=(\ \vec{r},\sigma\ )$ has the spatial and spin degrees of freedom. This property enforces the exclusion of certain quantum states for identical fermions, leading to the stability of matter as it prevents multiple fermions from occupying the same quantum state simultaneously. As a result, electrons in atoms must occupy different energy levels, forming a discrete spectrum of atomic states (Griffiths and Schroeter 2018).

When atoms are brought together to form a solid, the atomic energy levels interact and split into many closely spaced levels, forming a continuous spectrum of energy bands. The energy bands are separated by gaps where no states are allowed. The distribution of electrons among the bands depends on how the electrons of different chemical species interact and determines many physical properties of the solid, such as its electrical conductivity, optical absorption, and magnetic behavior (Kittel 2004). Depending on the size and position of the band gap separating occupied and unoccupied states, we can classify solids into three types: insulators, semiconductors, and metals. Each of these types is illustrated in *Figure* 3. Insulators are solids that have a large band gap between the highest occupied band (called the valence band) and the lowest unoccupied band (called the conduction band). This means that electrons in insulators are tightly bound to their atoms and cannot move freely under an applied electric field. Semiconductors are solids that have a small band gap between the valence and conduction bands. This means that electrons in semiconductors can be excited from the valence band to the conduction band by thermal energy or light, creating free charge carriers that can conduct electricity. Metals are solids that have no band gap or a partially filled conduction band. This means that electrons in metals are free to move within the conduction band under an applied electric field (Kittel 2004; Shur 2005).

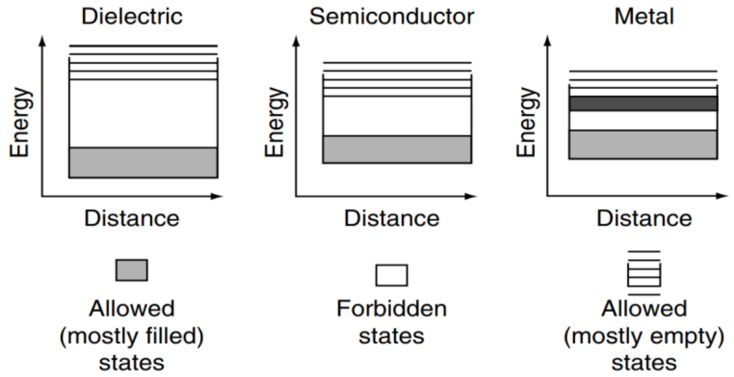


Figure 3 – Metal, semiconductor, and dielectric (insulator) band structures. Shaded patches show filled energy levels, each accommodating two electrons with opposite spins due to the Pauli exclusion principle. Source: (Shur 2005)

Beyond the influence of band gap size lies another crucial aspect: the crystalline structure. The arrangement of the atoms within solids profoundly impacts their optoelectronic traits, and for most materials, this arrangement follows a long-range periodic pattern defining a crystal. The crystal structure is defined as the combination of a periodic lattice and a base, which consists of repeating units, in this case, atomic positions (Kittel 2004). This lattice, described by vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, retains its structure under translation by a vector $\mathbf{T} = n_1 \cdot \mathbf{a}_1 + n_2 \cdot \mathbf{a}_2 + n_3 \cdot \mathbf{a}_3$, where n_i are integers. In three-dimensional space, symmetry operations limit the valid sets of lattice vectors to 14 types, known as Bravais lattices, shown in Figure 4. These lattice vectors enable the modeling of infinite systems with a small primitive cell containing only a few atoms. Within the space defined by the translation vectors, as for all vectors **T** in the Bravais lattice, the potential is periodic: $v(\boldsymbol{r}) = v(\boldsymbol{r} + \boldsymbol{T})$.

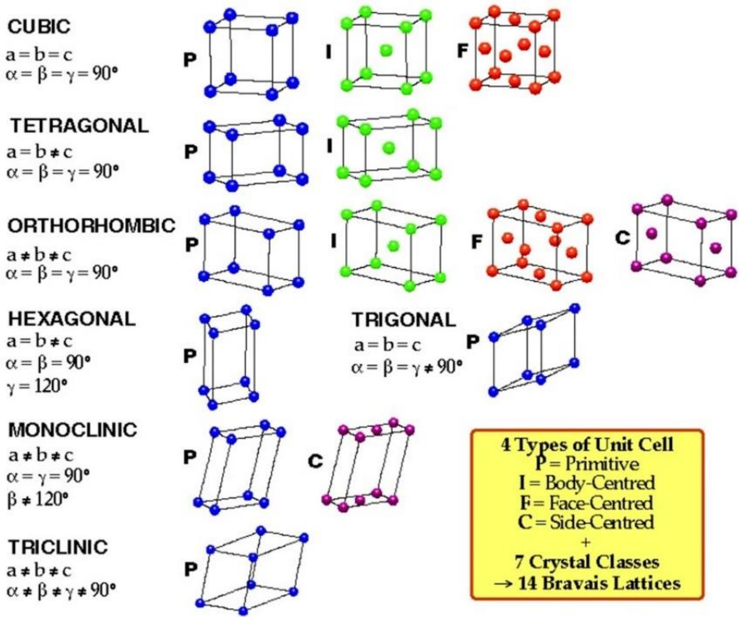


Figure 4 – The 14 Bravais lattices which compose seven crystalline systems. Source: (Mascarenhas 2020)

The periodic potential in crystals exerts a profound influence on electron behavior as the wavefunction's periodicity aligns with the crystal lattice. According to the Bloch theorem (Bloch 1929), solutions to the Schrödinger equation within a periodic potential can be expressed as plane waves modulated by periodic functions. Mathematically, a wavefunction $\Psi(\boldsymbol{r})$ for a particle, in this case, can be written as:

$$\Psi_k(r) = e^{ik\cdot r}u_k(r), \tag{6}$$

Here, $\Psi_k(r)$ represents the electrons' wavefunction, $e^{ik\cdot r}$ is a plane wave where the wave vector k defines the crystal momentum ($\hbar k$), and $u_k(r)$ is a periodic function describing the crystal lattice periodicity. Crystal momentum is a consequence of electron interaction with the periodic potential of the crystal lattice; therefore, it creates another dependency for the electron energy that can be described in band theory in the form of a dispersion relation, $E(k)$, for the single-electron states.

The dispersion relation reveals allowed energy states for electrons across the lattice, crucial for understanding electron transitions between valence and conduction bands in semiconductors. These transitions dictate material properties governing light emission, absorption, and conductivity in optoelectronic devices like solar cells, LEDs, and lasers. Two special quanta, photons and phonons, drive these transitions. Photons, as carriers of electromagnetic radiation, enable electronic transitions through absorption or emission. However, due to the lack of rest mass, photons convey very small momenta. Conversely, phonons, originating from lattice vibrations facilitate crystal momentum transmission, aiding transitions between states with different momenta. However, while phonons efficiently transmit crystal momentum, their energy remains relatively small compared to photons. In combination, photons and phonons serve as crucial mediators, allowing electrons within a semiconductor to transition between the conduction and valence band states while upholding conservation laws (V. K. Jain 2022; Kittel 2004).

Depending on how the transitions in a semiconductor are mediated the band gap can be defined as direct or indirect. In direct bandgap materials, the energy of the conduction band minimum (CBM) and the valence band maximum (VBM) coincide at the same momentum (k). This alignment facilitates efficient absorption and emission of photons with energy of the band gap (E_g) due to their energy and momentum matching. On the other hand, indirect bandgap materials exhibit different momenta for the CBM and VBM, typically requiring the involvement of phonons to facilitate the transition, as shown in *Figure 5*. This mismatch limits the efficacy of photon absorption or emission, limiting their application in optoelectronic devices (Garrillo 2018).

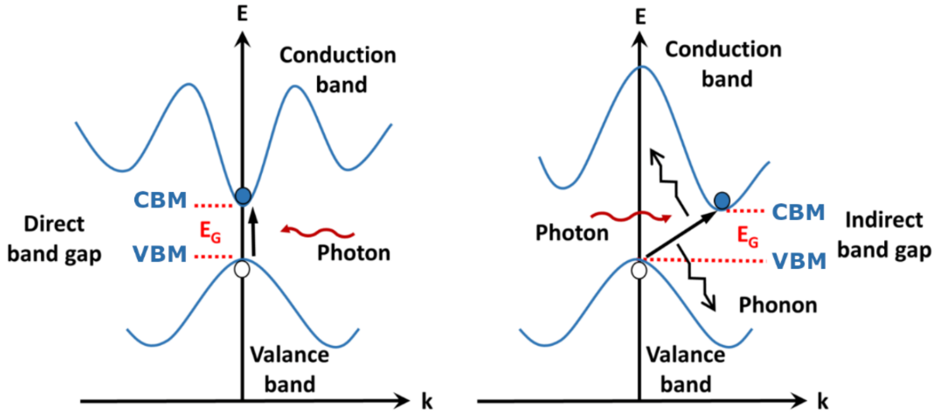


Figure 5 – Comparison of photon absorption in direct (left) vs. indirect (right) bandgap semiconductors. Adapted from: (Garrillo 2018)

This process in which an electron is promoted from the valence band to conduction band, leaving a corresponding hole in the valence band is named photovoltaic effect. This electron-hole pair usually remains bound by the Coulomb attraction force forming an exciton. Excitons resemble excited hydrogen atoms, with the electron orbiting the positively charged hole with a discrete energy spectrum positioned near the conduction band. The exciton binding energy typically ranges a few meV and the neutral exciton traverses the lattice, interacting with phonons, impurities, and imperfections. These interactions may result in either recombination, which restores the ground state and emits energy as light, or decomposition, which produces free carriers (electrons and holes) contributing to photoconductivity (V. K. Jain 2022).

The free-electron dispersion follows the equation $E = \frac{\hbar^2 k^2}{2m}$ where k represents the wavevector of a planewave describing the free electron’s eigenstate. However, within the lattice, the periodic potential changes this behavior. It can be shown that free carriers in the vicinity of CBM or VBM will exhibit characteristics akin to free electrons, except the particle's inertia becomes inversely proportional to the curvature of the dispersion relation. This results in an effective mass, which is determined by:

$$m^* = \frac{1}{\frac{1}{\hbar^2}\frac{d^2E}{dk^2}}. \tag{7}$$

The effective mass will determine the transport of quasiparticles that are the carriers involved in the electronic excitation of semiconductors. Since the derivative in Eq. 7 is related to the "curvature" of the dispersion relation, bands with high curvature correspond to small effective masses (light quasiparticles), while flatter bands, i.e., bands with low curvature, represent "heavier" quasiparticles. Furthermore, at the VBM the curvature will have negative values, and therefore, those electrons will present negative effective masses (the group velocity changes in opposition to the direction of the electromagnetic force). This condition is resolved by considering the equivalent condition of a positively charged quasiparticle, referred as "hole", thereby restoring proper signs to the transport properties (Wasserman 2005; Kittel 2004). Typically, holes also tend to have larger effective masses than electrons because valence band orbitals usually exhibit a more localized nature. In general, for optoelectronic applications, a small effective mass is desirable as it boosts the dynamics of free carriers.

It is crucial to note that electrons in the conduction band are only temporarily stable and will eventually shift to a lower energy level in the valence band, moving into an empty valence band state and removing a hole in the process. The energy difference between the electron's starting and final positions is released in this process, known as photo-carrier recombination. In inorganic semiconductors, there are three main types of this recombination: radiative (transferring energy as photons), non-radiative (transferring energy as phonons), and Auger (transferring energy as kinetic energy to another electron in the conduction band), as illustrated in *Figure 6*. Radiative recombination tends to happen more frequently in materials with direct band gaps. On the other hand, non-radiative recombination typically prevails in indirect band gap materials, such as Si-based devices, favored by point defects and dislocations. In scenarios with high carrier density, Auger recombination emerges as a significant contributor to energy loss. This is particularly noticeable in highly doped materials and confined structures like quantum dots (F. Wang, Liu, and Gao 2019; Garrillo 2018).

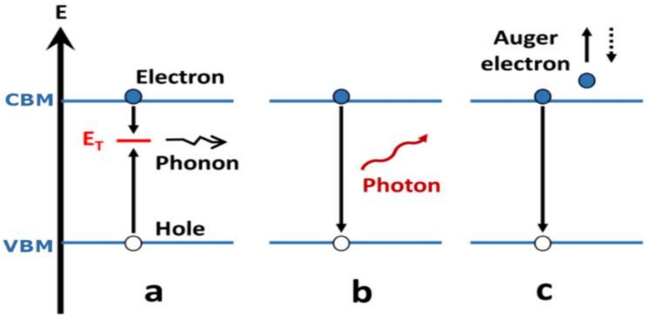


Figure 6 – Non-radiative (a), radiative (b), and auger recombination (c) mechanisms in inorganic semiconductors. The energy level of a defect or impurity in the material, commonly referred to as a trap, is indicated by E_T in (a). Adapted from: (Garrillo 2018)

Structural imperfections within a crystal significantly impact carrier dynamics and recombination in semiconductors. These imperfections including point defects (vacancies, anti-sites, interstitials, etc.), dislocations, grain boundaries, and impurities (either unintentionally or intentionally introduced through doping), generate localized states with energies within the band gap, functioning as trapping sites for charge carriers. Two main categories of trap states exist: shallow level traps and deep level traps.

Shallow traps typically present energies closer to the conduction or valence band and thus can readily capture or release charge carriers, significantly influencing conductivity and carrier concentrations. They can act as either donors or acceptors, introducing free carriers to the semiconductor. Acceptor defects lack electrons to bond with neighboring atoms, essentially introducing holes. At sufficient temperatures, these holes can ionize, moving deeper into the valence band as free carriers. After ionization, the acceptor carries a negative charge. Conversely, donor defects possess an extra electron, resulting in a localized extra negative charge within the bandgap. Thermal energy is usually enough to move this electron to the conduction band, turning the donor site into a positively charged site. In shallow traps, the wavefunction is fairly delocalized with size on the order of the Bohr exciton radius of the material (Grundmann 2010).

In contrast, deep level traps have strongly localized wavefunctions and energy levels well within the bandgap, although there are exceptions. Due to their greater distance from the band edges, deep trap states are inefficient at providing free electrons or holes. Instead, they tend to capture free carriers, reducing conductivity. As a result, these centers typically establish routes for nonradiative recombination by directing electrons through the deep levels into the valence band. A material whose structural defects predominantly cause the formation of shallow states is referred to as defect-tolerant material (Kang and Wang 2017). However, there exist numerous situations where controlled deep traps are leveraged, for example, to modulate photoluminescence in semiconductors (Grundmann 2010; Hussain et al. 2022).

Another crucial aspect for application is the optical response of semiconductors. The responses of periodic systems to an externally applied electric field, such as that induced by the electromagnetic waves in light, are described by a complex dielectric function:

$$\varepsilon(\omega) = \varepsilon_r(\omega) + i\varepsilon_i(\omega), \quad (8)$$

The real part is related to the polarization of the material due to the applied electric field. It describes the phase shift of the field and is intimately related to the material's refractive index for light. The imaginary part arises from the dissipation of energy as the material absorbs the electric field for electronic transitions and interacts with the lattice. Therefore, the imaginary part of the dielectric function is closely related to the absorption spectra, which are crucial for applications like photovoltaics and photocatalysis. For a derivation of the dielectric function from first principles and subsequent derivation of the absorption coefficient for direct and indirect transitions, the reader is referred to *Appendix A.1*.

The dielectric function plays a crucial role in understanding nonradiative recombination losses in semiconductors, especially through defect-assisted processes. For instance, when electrons encounter positively charged defects, their capture is driven by Coulomb attraction, described by a capture cross-section equation as:

$$\sigma_{-} = \frac{q^4}{16\pi(\varepsilon_s k_{\mathrm{B}} T)^2}. \quad (9)$$

Here, q represents the elementary charge, ε_s is the dielectric constant in the static limit ($\omega \rightarrow 0$), k_b is the Boltzmann's constant and T is the temperature. This equation reveals that at a constant temperature, σ_{-} decreases as the dielectric constant increases. This suggests that boosting the dielectric constant within semiconductors can reduce the defect capture cross-section. This change acts as a dielectric-screening effect, weakening the trapping of charge carriers by defects and enhancing carrier transport. This principle applies similarly to holes (Su et al. 2021; Peter and Cardona 2010).

In addition to affecting trap carriers, the dielectric function plays a role in shaping the energy levels of excitons, characterized by hydrogen-like states indicated by $E_n \propto n^{-1/2}$ as per the expression:

$$E_X^n = -\frac{m_r^*}{m_0}\frac{1}{\varepsilon_s{}^2}\frac{m_0 e^4}{2(4\pi\hbar)^2}\frac{1}{n^2}, \quad (10)$$

where $\mathrm{m}_{\mathrm{r}}^{*}$ denotes the exciton's reduced effective mass. The exciton binding energy relates to this equation by $E_X^b \ = - E_X^1$. The term $\frac{m_r^*}{m_0}\frac{1}{{\varepsilon_s}^2}$ scales to approximately 10^{-3}, yielding values in the meV range that vary inversely with the dielectric constant (Grundmann 2010). This dependence underscores the importance of a high dielectric constant in the generation of free carriers by the dissociation of excitons.

2.1.2 Semiconductors for optoelectronic devices

Having established a foundation for the general properties of semiconductors, our attention now shifts to exploring the specific devices that use these materials and the required properties for each. Our focus lies particularly on optoelectronic applications associated with visible light, spanning photovoltaics, luminescence, and photocatalysis. The sought-after properties will be regularly discussed within the results presented in the forthcoming chapters.

- **Photovoltaics:** Solar cells are devices that convert light energy into electrical energy through the photovoltaic effect. They consist of layers with an absorber material—like silicon in traditional cells or a perovskite compound in perovskite cells—sandwiched between charge transport layers. Sunlight hitting the absorber creates excitons which separate into electrons and holes. These flow in opposite directions due to the built-in potential on the device, generating an electric current. Direct band gap materials are best suited for efficient energy conversion and ideal band gaps vary from around 1.2 eV for single junction cells to 1.75 eV for tandem cells with silicon. A high dielectric constant enhances performance by reducing charge trapping and recombination. Moreover, lower effective masses of the charge carrier, increase high carrier mobility, preventing recombination and improving overall efficiency in photovoltaic devices (Marongiu et al. 2019; Su et al. 2021).

- **Luminescence:** Electro- and photoluminescence are the most common light emission processes. Electroluminescence, exemplified by the popular light-emitting diodes (LEDs), showcases the functionality of semiconductor junctions in emitting light. Here, the application of voltage induces electron movement within the semiconductor which recombine with holes releasing energy as photons ($\hbar\omega_{ph} = E_g$). Conversely, photoluminescence relies on external photon excitation to generate the electron-hole pairs, serving various applications such as sensors, imaging, displays, and special light sources (Mousavi et al. 2014). Since the photon frequency is proportional to the band gap, the material used determines the color of the emitted light. The current leading technology for luminescence with semiconductors employs quantum dots (QDs) or nanocrystals (Marongiu et al. 2019). The spatial confinement causes a significant decrease in the dielectric constant, leading to increased exciton binding energy which prevents excitons to dissociate prior to radiative decay (Zheng et al. 2015). Moreover, the spatial confinement increases the recombination probability of electron-hole pairs. This leads to exceptionally high photoluminescence quantum yield with sharp emission lines (Elward and Chakraborty 2013; Marongiu et al. 2019). However, confinement has its downsides, such as heightened Auger recombination and an abundance of surface states, causing non-radiative recombination. To mitigate these issues, enveloping particles in a core-shell structure helps balance electron/hole injections and quench surface states (W. K. Bae et al. 2013). Additionally, localized and heavy holes (high effective masses) with high mobility electrons also aid emission efficiency (W. H. Guo et al. 2020a; Chichibu et al. 2006).

- **Photocatalysis:** Photocatalysis relies on semiconductors called photocatalysts, triggering or speeding up chemical reactions when exposed to light. These catalysts absorb photons, creating excitons that produce reactive species—such as free radicals or charged particles—on their surfaces. This process, crucially influenced by surface states, reduces the activation energy of reactions (W. D. Kim et al. 2016). Photocatalysts are integral in green technologies, utilizing renewable energy sources like sunlight to facilitate chemical transformations, promising advancements in clean energy production and environmental remediation. For instance, they are utilized in degrading pollutants during environmental remediation and in water-splitting reactions for hydrogen fuel production from water. An effective photocatalyst exhibits several key characteristics: (i) the capacity to absorb radiation across a broad spectrum of light, (ii) appropriate alignment of the semiconductor's energy bands concerning the redox reaction potentials for the aimed application, (iii) high mobility and extended diffusion paths for charge carriers (thus, low effective masses), (iv) thermodynamic and photoelectrochemical stability (Jiangtian Li and Wu 2015). Moreover, nanometer-sized materials are usually advantageous as they offer a high surface area (more reactive sites) and enable tunable band gaps through particle size control (Feliczak-Guzik 2023).

2.1.3 Stability and synthesizability

In the quest to optimize semiconductor properties for cutting-edge optoelectronic applications, tools that help scientists discern stable and synthesizable compounds are the cornerstone bridging the gap between theoretical exploration and practical technology (Malyi et al. 2020; Zunger 2018). While the understanding of compound stability can be derived from first-principles, a single tool for its evaluation remains elusive. Instead, a suite of techniques and corresponding criteria are combined to increase confidence in a compound's stability. This study strongly focuses on using thermodynamic stability to measure stability, discussed in detail in this section. Dynamical stability is another important criterion, but presents cost constraints to be evaluated using first-principles as discussed on *Appendix A.2*. Specific applications might necessitate additional stability evaluations, such as photostability and electrochemical stability, yet these could involve multiple mechanisms, making prediction even more challenging (Chonamada, Dey, and Santra 2020; Jiangtian Li and Wu 2015).

A material is deemed thermodynamically stable under a given set of conditions (temperature, pressure, chemical potentials, etc.) if its energy cannot be lowered by rearranging its atoms. Energy lowering can occur through two distinct cases: (1) phase transition to an alternative crystal structure (polymorph) at a fixed composition, or (2) phase separation (decomposition) into competing materials sharing the same average composition (Bartel 2022). The total energy of the material ($E_{compound}$) can be calculated from first-principles and subsequently used to compute the formation energy of the compound (E_f) from its constituent atoms by:

$$E_f = E_{compound} - \sum_i n_i E_i \, , \tag{11}$$

where n_i is the number of atoms of element *i* in the compound, and E_i is the energy of the atom of element *i*, usually in its standard state. If the elemental energies considered are from isolated atoms, *Equation 11* becomes the binding energy of the compound (E_b). A negative formation energy indicates only that the compound is stable against decomposition into the constituent pure phases. However, the difference in formation energy enables direct comparison between two polymorphs, the first case for energy lowering, to determine the ground-state polymorph under given conditions.

However, for the second option to lower energy, a more general and useful metric of thermodynamic stability is the decomposition energy (E_{stab} or E_d) or convex hull distance. E_{stab} is computed against all ground-state polymorphs across the relevant chemical space of interest using the convex hull formalism. Constructing the convex hull involves obtaining E_f for all ground-state polymorphs in the system, typically as a function of the normalized molar composition of N - 1 elements in the chemical space of interest with N unique elements. The hyperplane connecting all these ground-states polymorphs across compositions forms the convex hull. The convex hull analysis assesses whether a given material can lower its energy by decomposing into a linear combination of materials having the same average composition as the material of interest. If we consider compounds in a hypothetical chemical space A−X (e.g., A_2X, AX, A_2X_7, etc.), an example convex hull is provided in *Figure 7*.

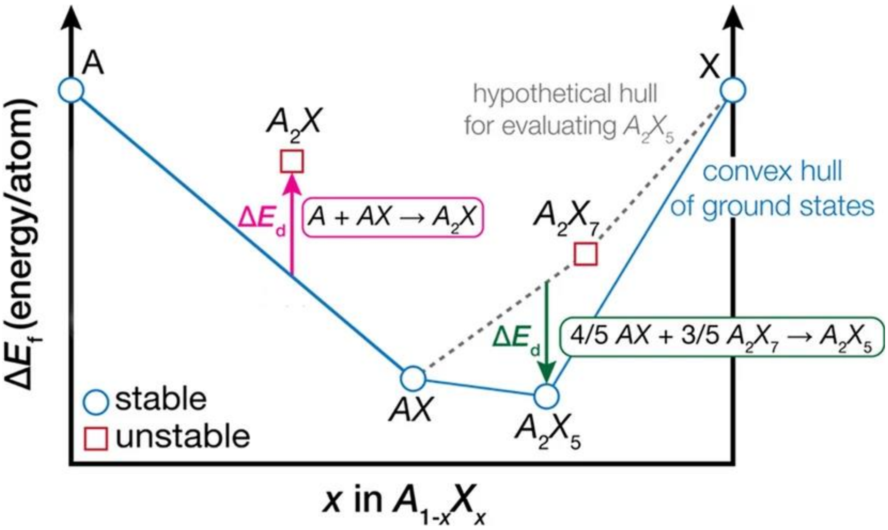


Figure 7 – Convex hull phase diagram for hypothetical A−X system. Blue circles on the solid line represent thermodynamically stable phases and points above the hull are thermodynamically unstable with respect to phase separation. The dashed gray line refers to the hypothetical convex hull used to determine the decomposition reaction and energy for the stable phase, A_2X_5. Adapted from: (Bartel 2022)

The convex hull is extremely helpful because it identifies materials lying on it as thermodynamically stable in terms of phase separation. Materials located above the hull are considered unstable because decomposition to a linear combination of alternate compositions can reduce their energy. On the other hand, materials with a negative decomposition energy are positioned below the convex hull and are stable against decomposition. Hence, these compounds become part of the convex hull for subsequent calculations.

A few thermodynamic considerations must be addressed concerning the decomposition energy. Energy calculations from first-principles, using traditional methods such as Density Functional Theory (*Section 2.2*), typically involve isolated systems in a vacuum at 0 K. However, E_f can be transformed into a formation enthalpy, ΔH_f, at a given temperature, T, by incorporating the zero-point energy correction and integrating the constant volume specific heat from 0 K to T. Both quantities, attainable via first principles through the phonon density of states (Togo and Tanaka 2015). It has been shown that formation energies generally remain unaffected by this conversion from 0 K energies to 298 K enthalpies, owing to error cancellation (Bartel et al. 2019).

However, at very high temperatures, the appropriate thermodynamic potential is the Gibbs free energy, G = H - TS, directly linked to temperature through entropy and potentially influenced significantly by vibrational entropy (Fultz 2010). Another contributing factor to G is configurational entropy, typically prominent in disordered materials like solid solutions with various species, further contributing to stabilization of these compounds (Bartel 2022). With some exceptions, however, these contributions are not significant in ambient temperature and were not considered in this work.

Finally, despite being the ultimate goal, evaluating synthesizability requires incorporating kinetic factors and experimental dependencies, making it challenging to address solely through theoretical physical chemistry. Typically, assessing synthesizability involves assuming that synthesizable materials do not possess thermodynamically stable decomposition products, essentially extrapolating from stability evaluations. However, this approach has limitations as it overlooks aspects like kinetic stabilization. Very recently, tools have emerged aiming for a broader evaluation of synthesizability beyond basic thermodynamics, yet their practical use remains somewhat restricted by extrapolating solely on composition (Antoniuk et al. 2023).

2.1.4 Halide perovskites and doping

Metal halide perovskites have become a very promising class of solution-processable semiconductor materials for high-performance optoelectronic devices. They follow the traditionally defined general formula, ABX_3, where A and B are cations, and X is a halide anion. However, more loose definitions considering the presence of organized BX_6 octahedra coordinated by larger A cations has been more used recently to include perovskites with multiple dimensionalities which may deviate from the traditional stochiometric formula as illustrated in *Figure 1* (Fakharuddin et al. 2019).

The exponential rise of perovskite technology in photovoltaics, rivaling the established efficiency benchmarks set by traditional silicon, showcases the technological potential of these materials. In the case of solar cells, the hybrid organic-inorganic $MAPbI_3$ perovskite is the forerunner combining an unprecedented set of properties, namely (Frost et al. 2014):

- A high absorption coefficient due to the strong overlap between the valence band and the conduction band, which are mainly composed of Pb 6*s* and I 5*p* orbitals. This results in a direct band gap and a large oscillator strength for optical transitions.
- A highly tunable band gap, characteristic of most halide perovskites, can be attributed to the hybridization between the Pb 6*s* and I 5*p* orbitals, which can be modulated by changing the size and orientation of the organic cation, the halide anion, or the metal cation. For example, replacing MA with formamidinium (FA) or cesium (Cs) can lower the band gap, while replacing I with Br or Cl can increase the band gap.
- A long carrier lifetime due to the low density of trap states and the high dielectric constant of the material. The low density of trap states is attributed to the self-healing mechanism of the perovskite structure, which can accommodate various defects and distortions without breaking the Pb-I bonds. The high dielectric constant is attributed to the polarizability of the organic cation and the lattice vibrations, which can screen the Coulomb interactions between the carriers and the defects.
- High defect tolerance due to the low formation energy and the low ionization energy of the intrinsic defects, such as vacancies and interstitials. These

defects can act as shallow donors or acceptors, which can be easily compensated by the Fermi level or the external electric field. Moreover, the defects can also enhance the carrier mobility and the conductivity of the material by creating additional hopping sites or pathways.

Many of these great properties are present in other halide perovskites and underscore the functional versatility of this material class that find applications in various fields besides solar cells. The tunable band gap allows for color control and high photoluminescence quantum yields can be obtained due to the high defect tolerance and dielectric function, this coupled with long carrier lifetimes, makes them suitable for optoelectronic devices, such as light-emitting diodes, lasers, and photodetectors (He and Liu 2023). Moreover, some halide perovskites exhibit fascinating magnetic and superconducting properties (A. Banerjee and Paul 2020; Siyuan Zhou et al. 2024), which open up new possibilities for spintronics, memory devices, and quantum computing (H. Kim et al. 2018; John et al. 2022). Furthermore, halide perovskites have shown promise in energy storage (L. Zhang et al. 2020), such as batteries, capacitors, and thermoelectrics (Haque et al. 2020), due to their high ionic conductivity, large capacitance, and low thermal conductivity.

Exploring the structural and chemical versatility reveals the expansive potential of this formidable class of materials. Halide perovskites can adopt different dimensionalities, compositions, and properties by manipulating the A, B, and X components, as well as the crystallographic directions and the interplay between them. For example, by making cuts along different crystallographic directions, one can obtain perovskites with zero-dimensional (0D), one-dimensional (1D), two-dimensional (2D), from three-dimensional (3D) structures, each with distinct electronic, optical, and magnetic properties. By substituting different cations and anions, one can tune the band gap, charge transport, spin-orbit coupling, and lattice distortion of the perovskites, which affect their performance in various devices. By intercalating guest molecules, such as water, ammonia, or organic solvents, one can induce phase transitions, modulate the dielectric constant, or enhance the stability of the perovskites (Smith, Connor, and Karunadasa 2019; Kahwagi et al. 2020; Hoye et al. 2022). While around 3500 perovskites are present in experimental databases (J. Liang et al. 2022), it is estimated that due to the structural and chemical flexibility the potential number of perovskites can easily exceed 10^7 (Q. Tao et al. 2021; C. Li et al. 2020).

Since the dawn of semiconductor physics, doping has proved to be an effective way to modulate the fundamental properties of semiconductors. Due to their ionic structure with low formation energies, doping is comparatively easier and interesting in halide perovskites than other conventional semiconductors. In the case of perovskites, doping usually means partially replacing the original constituent elements with targeted ions even when concentrations are substantially higher than usual for semiconductors (L. Xu et al. 2019; G. Chen et al. 2020; Kumawat et al. 2019). Doping with appropriate ions effectively contributes towards stabilizing the crystal structure, tuning the optoelectronic properties, and enhancing the device performance (Parveen and K. Giri 2022; C.-H. Lu et al. 2020). Following is an overview of how doping affects each site within the halide perovskite individually (L. Xu et al. 2019; C.-H. Lu et al. 2020):

- *A-site doping:* can affect the dimensionality, stability, and band gap of the perovskites. For example, replacing organic A cations with inorganic ones, such as Cs^+, can increase the thermal and moisture stability of the perovskites, as well as reduce the band gap and enhance the light absorption. In inorganic perovskites, Cs^+ can also be substituted by Rb^+ or K^+ to modulate the structural dimensionality and band gap (Lehner et al. 2015).
- *B-site doping:* can alter the electronic structure, carrier concentration, and defect density of the perovskites. For instance, replacing Pb^{2+} with other metal cations, such as partially oxidized post-transition metals (In^+, Tl^+, Sn^{2+}, Sb^{3+}, Bi^{3+}, etc.), can modulate the band gap and the carrier effective mass of the perovskites, as well as reduce the toxicity of Pb-based perovskites. Moreover, doping with transition metal cations, such as Mn^{2+}, Fe^{2+}, or Ni^{2+}, can introduce localized energy levels and magnetic moments in the perovskites, which can enable spintronic and multiferroic applications (Amerling et al. 2021).
- *X-site doping:* can influence the lattice constant, crystal phase, and optical properties of the perovskites. For example, changing the ratio or composition of the halide anions, such as Cl^-, Br^-, or I^-, can adjust the lattice constant and the crystal phase of the perovskites, which can affect the strain, stability, and band gap of the materials. Furthermore, doping with pseudohalide or superhalogen anions, such as N_3^-, SCN^- and BH_4^- can introduce new optical

features, enhance the photoluminescence and also improve stability (Lin et al. 2021).

The growing literature on tuning compositions in perovskites highlights limitless possibilities for exploration. As research delves deeper into the vast chemical space of halide perovskites, ab-initio simulations and machine learning have become indispensable for comprehending and predicting the diverse effects of these alterations (Q. Tao et al. 2021). These methodologies enable a faster, more systematic, and insightful exploration of material modifications and their potential impacts, surpassing the limitations of experimental investigations alone. The upcoming sections will explore these methodologies in detail.

2.2 Density Functional Theory

Over the past few decades, Density Functional Theory (DFT) has undergone a remarkable evolution, transforming from a theoretical tool with limited applicability to a cornerstone of modern materials science. Its emergence as a powerful computational method has revolutionized our understanding of electronic structure and properties of materials, offering unprecedented insights into their behavior at the atomic level. In this section, we provide an overview of the theory and computational implementation of DFT, highlighting its relevance in predicting materials properties.

2.2.1 Hohenberg-Kohn and Kohn-Sham formalism

The N-electron wave function solution for the Schrödinger equation shown in Equation 3 can be computationally demanding, especially for large systems. DFT arises as an alternative approach, which uses electron density, $n(\boldsymbol{r})$, to describe the many-electron system in which electron and nuclei coordinates were decoupled by Born-Oppenheimer approximation (see Appendix A.3). The Hohenberg-Kohn (HK) theorems provide the basis for DFT. The first theorem states that the ground state electron density uniquely determines the external potential. From the first HK theorem the energy functional of a system in a particular external potential v_0 can be written as

$$E_{v_0}[n] = \langle \Psi[n] | \hat{T} + \hat{V}_{ee} + \hat{V}_0 | \Psi[n] \rangle. \tag{12}$$

The notation $|\Psi[n]\rangle$ signifies that the quantum state of the system is explicitly dependent on the electronic density. The second theorem offers a variational approach to obtain the ground state electron density by minimizing the energy functional. This is expressed in the following equation:

$$E_0 = \min_n E_{v_0}[n]. \tag{13}$$

The energy functional, denoted as $E_{\mathrm{HK}}[n]$, can be separated into two parts: one dependent of the external potential and a universal functional, which yields:

$$E_{\mathrm{HK}}[n] = E_{v_0}[n] \; = F_{\mathrm{HK}}[n] \; + \int v_0(\boldsymbol{r}) n_0(\boldsymbol{r}) \, d^3\boldsymbol{r} \tag{14}$$

The universal functional, $F_{\mathrm{HK}}[n]$, is defined as:

$$F_{\mathrm{HK}}[n] = \langle \Psi[n] | \hat{T} + \hat{V}_{ee} | \Psi[n] \rangle \tag{15}$$

The Hohenberg-Kohn theorems, while formally defining F_{HK}, lack a practical calculation scheme. To address this limitation, Kohn and Sham introduced an efficient methodology for practical application of DFT.

To describe the electron system, an auxiliary system consisting of a non-interacting electron gas is introduced in Kohn-Sham's approach (Kohn and Sham 1965). This auxiliary system has the same ground state electron density as the actual system. The HK functional is expressed as:

$$F_{\mathrm{HK}}[n] = \frac{1}{2} \iint \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}^3\mathbf{r} \mathrm{d}^3\mathbf{r}' + T_0[n] + E_{\mathrm{xc}}[n] \tag{16}$$

The first term represents the Coulomb repulsion between electrons, while the second term accounts for the kinetic energy of a non-interacting electron gas with the same density as the real system. The last term, known as the exchange-correlation (XC) energy, incorporates contributions that reconcile the limitations of the simplified system with the characteristics of the actual physical system, ensuring a formally equivalent representation. These contributions include: (i) correcting the kinetic energy to describe the real interacting system, (ii) correcting the self-interaction energy resulting from the Coulomb term, (iii) accounting for exchange energy due to the required exchange anti-symmetry of the electron wave function, and (iv) considering correlation energy that captures the interdependence of electron dynamics. By varying the total energy expression (14) and (16) with respect to a set of one-electron wavefunctions, denoted as ϕ_i(r), the Kohn-Sham (KS) equations are derived. These wavefunctions define the density for an N-particle system as:

$$n(\boldsymbol{r}) = \sum_{i}^{N} |\phi_i(\mathrm{r})|^2. \tag{17}$$

This leads to the following equation for the Kohn-Sham potential,

$$v_{\mathrm{eff}}(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}^3\mathbf{r}' + \frac{\delta E_{\mathrm{xc}}[n(\mathbf{r})]}{\delta n(\mathbf{r})} + v_{ext}(\mathbf{r}), \tag{18}$$

the Kohn-Sham equations are then given by the following equation:

$$\left[-\frac{1}{2}\nabla_{\mathbf{r}}^2 + v_{\mathrm{eff}}(\mathbf{r})\right]\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}). \tag{19}$$

The one-electron wavefunctions introduced in this approach are referred to as KS orbitals ($\phi_i(\mathbf{r})$). They are utilized to construct the total wave function using a Slater determinant and to determine the ground state electron density through equation (17). The total energy of the system can be obtained as:

$$\begin{aligned} E_{tot} = \sum_{i=1}^{N} \epsilon_i - \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}^3\mathbf{r}\, \mathrm{d}^3\mathbf{r}' - \int v_{xc}(\mathbf{r})n(\mathbf{r})\, \mathrm{d}^3\mathbf{r} + \\ E_{xc}[n(\mathbf{r})] + \sum_{\alpha=1}^{M}\sum_{\beta=\alpha+1}^{M} \frac{Z_\alpha Z_\beta}{|\mathbf{R}_\alpha - \mathbf{R}_\beta|}. \end{aligned} \tag{20}$$

Since the effective potential in the equation necessary to solve for the KS orbitals relies on the electron density, which is obtained from the KS orbitals themselves, the solution of the KS equations requires a self-consistent approach. The process begins with a trial set of KS orbitals, from which the electron density is computed using equation (17). Subsequently, the effective potential is determined using equation (18) and then utilized in equation (19) to calculate new KS orbitals. This iterative process of obtaining new electron density and effective potential continues until a predefined convergence criterion is met, such as a negligible change in the total energy.

A crucial aspect to consider is spin-polarization, which is necessary for describing materials that exhibit unbalanced spins such as magnetic materials, excited states and systems containing transition metals. This also typically extends to structures that exhibit dangling bonds, such as crystalline defects, heterovalent doping, as well as surfaces and clusters, all of which are covered in this study. To incorporate the KS method within a spin-polarized framework, the electron density is defined with its spin components for occupied states. Spin-polarization can be introduced as follows:

$$n_{\sigma}(\mathbf{r}) = \{n^{\uparrow}(\mathbf{r}), n^{\downarrow}(\mathbf{r})\} = \sum_{\sigma=\uparrow,\downarrow i,occ} \sum_{\sigma i}^{N} |\phi_{\sigma i}(\mathbf{r})|^2, \tag{21}$$

where σ represents the spin channel, and the curly brackets indicate the set of two spin components, in which *σ* indicates the spin channel and the curly brackets indicate the set considering two spin components, ↑ *(up)* and ↓ *(down)*, used to express the electron density. Each KS orbital, $\phi_{\sigma i}(\mathbf{r}_i)$, can be scaled by a factor that modulates the spin components. The set of single-electron KS equations can then be written in the form:

$$\left[-\frac{1}{2}\nabla^2_{\mathbf{r}_i} + v^{\sigma}_{eff}\left[n^{\uparrow}(\mathbf{r}), n^{\downarrow}(\mathbf{r})\right]\right]\phi_{\sigma i} = \varepsilon_{\sigma,i}\phi_{\sigma i}, \tag{22}$$

in this equation v^{σ}_{eff} denotes that the effective potential may have different values for different spin channels, leading to eigenvalues that depend on spin polarization.

2.2.3 Local and semi-local XC functionals

The KS scheme has the potential to offer an exact solution for electron systems under any potential. However, the precise form of the exchange-correlation energy remains unknown. To address this, exchange-correlation functionals of the density have been developed and their accuracy for different systems and properties varies. The simplest level of approximation is the local density approximation (LDA), proposed in the original Kohn-Sham paper (Kohn and Sham 1965). LDA considers the electron density at each point in space and uses the Hartree-Fock (HF) exchange energy for a uniform electron distribution, yielding the expression:

$$E^{LDA}_{xc} = \int n(\mathbf{r}).\,\epsilon^{LDA}_{x}\big(n(\mathbf{r})\big)\,\mathrm{d}^3\mathbf{r} + \int n(\mathbf{r}).\,\epsilon^{LDA}_{c}\big(n(\mathbf{r})\big)\,\mathrm{d}^3\mathbf{r},$$

$$\text{where } \epsilon^{LDA}_{x}\big(n(\mathbf{r})\big) = -\frac{3}{4}\left(\frac{3n(\mathbf{r})}{\pi}\right)^{\frac{1}{3}}. \tag{23}$$

Nevertheless, the correlation term lacks an analytical form, prompting the development of parametrizations such as VWN (Vosko, Wilk, and Nusair 1980), PZ81 (J P Perdew and Zunger 1981) and PW92 (John P. Perdew and Wang 1992) by fitting numerical results obtained from Monte Carlo calculations on the homogeneous electron gas (Ceperley and Alder 1980). While LDA offers computational efficiency, it often demonstrates limitations in predicting cohesive energies and tends to underestimate lattice parameters in solids. This shortfall is largely attributed to its inability to effectively capture electron correlations. Despite its initial popularity, LDA's efficacy relies substantially on error cancelation mechanisms (Becke 2014).

To enhance the capabilities of LDA, the generalized gradient approximation (GGA) introduces a dependence on the electron density gradient in formulating semi-local functionals. Various functionals, like PW91 (John P Perdew, Ziesche, and Eschrig 1991) and PBE (J P Perdew, Burke, and Ernzerhof 1996), have emerged within the framework of GGA. PBE is an evolution from PW91 and offers a formulation less reliant on fitting parameters, making it more accessible and widely adopted, especially in the study of solids. The functional is again divided in exchange and correlation parts as follows:

$$\begin{aligned} E_{xc}^{PBE} &= \underbrace{\int n(\mathbf{r})\,.\,\epsilon_x^{LDA}\big(n(\mathbf{r})\big)\,.\,F_x(n(\mathbf{r}),|\nabla n(\mathbf{r})|)\,\mathrm{d}^3\mathbf{r}}_{E_x^{PBE}} + \\ &+ \underbrace{\int n(\mathbf{r})\,.\,\epsilon_c^{PW}\big(n(\mathbf{r})\big)\,.\,H(n(\mathbf{r}),|\nabla n(\mathbf{r})|)\,\mathrm{d}^3\mathbf{r}}_{E_c^{PBE}} \end{aligned} \tag{24}$$

where $\epsilon_c^{PW}\big(n(\mathbf{r})\big)$ is PW91 parametrization for LDA correlation energy, F_x and H are parametrized analytical functions designed to satisfy energetically relevant constraints on the exchange correlation functional such as behavior on slowly (rapidly) varying density, strong (low) correlation regions, translational invariance, etc. Spin polarization is usually treated in these functionals including the local relative spin-polarization, ζ given by

$$\zeta(\mathbf{r}) = \frac{n^{\uparrow}(\mathbf{r}) - n^{\downarrow}(\mathbf{r})}{n^{\uparrow}(\mathbf{r}) + n^{\downarrow}(\mathbf{r})}. \tag{25}$$

This requires that the parametrization functions obey new constraints such as the spin-scaling relationship (Oliver and Perdew 1979) and update the uniform electron gas quantities following the local spin-density approximation (LSDA) (Barth and Hedin 1972).

GGAs systematically improve the atomization or cohesive energies of a wide range of molecules and solids and correct the LDA’s overbinding. However, they are not a universal improvement over LDA as in both cases their accuracy relies substantially on system dependent cancellation of error (J P Perdew et al. 1992;

Hasnip et al. 2014). As of latest assessments (Swart Lab 2023), the PBE functional remains the most popular GGA and DFT functional in general, especially for solids. Its widespread adoption stems from its low computational cost and demonstrated accuracy across a wide range of compounds and properties (Lejaeghere et al. 2014). Throughout our investigations, after careful validation against other well-known functionals for the $Cs_3Sb_2X_9$ compounds (detailed in Table B1), we predominantly employed the PBE functional due to its established efficacy and prevalence assuring comparability of our findings with existing research in the field.

2.2.4 The band gap problem and self-interaction corrections

Though LDAs and GGAs functionals have found success in characterizing various structural and chemical properties across a broad range of materials, they exhibit limitations in accurately predicting specific electronic properties, notably the band gap (Sham and Schlüter 1983). When utilizing the Kohn–Sham valence and conduction band eigenvalues to calculate the band gap, these approaches persistently underestimate band gaps of semiconductors and insulators, for strongly correlated systems, this tendency may even extend to mistakenly predicting metallic ground states (Hasnip et al. 2014).

A major contribution to the band-gap error arises from the Hartree energy E_H in the Hamiltonian, as given in the first term on the right-hand side of Eq. 18. By using the total density, it also includes a Coulomb repulsion between an electron and its own charge density. This spurious self-interaction is exactly cancelled by the exchange term in some non-DFT methods, such as Hartree-Fock theory (discussed below), but it is only partially cancelled by LDA or GGA exchange. Because the self-interaction energy is always positive, the energy of localized states is raised favoring delocalization what leads to a lower band gap or even spurious metallization (Tu et al. 2007). This delocalization also manifests in the inability of semi-local DFT to reproduce the discontinuous potential change when electrons are transferred, known as the derivative discontinuity problem, which is quite relevant for reaction barriers (Mori-Sánchez and Cohen 2014).

Several approximations beyond GGA can mitigate the band gap problem. Commonly used ones are: (i) Meta-GGAs, which incorporate the Laplacian of the density (expressed as the kinetic energy density) providing more flexibility to the functional (J. Tao et al. 2003); (ii) DFT + U, involving an on-site Hubbard-U potential to enhance electron localization, often applied to more localized d or f shells, thereby improving also magnetic properties (Anisimov, Aryasetiawan, and Lichtenstein 1997); (iii) Hybrid functionals, which include an empirical fraction of Hartree–Fock exchange to alleviate the band-gap problem (Becke 1993). For this work, the last two approaches were applied to study the $Cs_3Sb_2X_9$ compounds since they provide a more reliable and systematic improvement to the band gap compared to Meta-GGAs (Borlido et al. 2020).

In the Hartree–Fock (HF) method, the Fock (exact) exchange energy is not a density functional; instead, it relies on single-particle states, expressed as:

$$E_x^{HF} = -\frac{1}{2}\sum_i\sum_j \int d^3\boldsymbol{r} \int d^3\boldsymbol{r}'\, \phi_i^*(\boldsymbol{r})\phi_i(\boldsymbol{r}')\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\phi_j(\boldsymbol{r})\phi_j^*(\boldsymbol{r}'), \tag{26}$$

where the sums run over occupied orbitals. This exchange is inherently non-local and exactly cancels the spurious self-interaction from Hartree term. However, HF neglects electronic correlation completely. Since exchange is very long ranged, decaying only as 1/r, due to lack of correlation screening, HF yields excessively high excitation energies and greatly overestimates the band gap (John P. Perdew, Ernzerhof, and Burke 1996; Hasnip et al. 2014).

Based on the realization that while HF exaggerates the fundamental gap, GGA (or LDA) functionals tend to underestimate it, hybrid XC functionals emerged, combining HF with GGA (or LDA) functionals. The PBE0 functional, for instance (John P. Perdew, Ernzerhof, and Burke 1996), retains the PBE functional's correlation term while blending the PBE exchange term with HF exchange at a 3:1 ratio,

$$E_{xc}^{PBE0} = \frac{1}{4}E_x^{HF} + \frac{3}{4}E_x^{PBE} + E_c^{PBE}. \tag{27}$$

The mixing ratio was derived via perturbation theory from the adiabatic connection theorem, aiming to optimize molecule atomization energies. These hybrid functionals strike a balance between HF and DFT, resulting in more realistic gap predictions. Modern functionals like the widely used HSE06 introduce a partition between long- and short-range contributions and are named screened hybrid functionals (Heyd, Scuseria, and Ernzerhof 2003).

Screened hybrids retain most of the benefits of global hybrids but significantly reduce the computational cost in extended systems (Vydrov et al. 2006). This method involves splitting the Coulomb operator into short (SR) and long ranges (LR), with the LR exchange only including PBE exchange, as follows:

$$\frac{1}{r} = \underbrace{\frac{erf(\omega r)}{r}}_{LR} + \underbrace{\frac{erfc(\omega r)}{r}}_{SR}. \tag{28}$$

The range separation utilizes the error function (*erf*(x)) and its complement (*erfc*(x) = 1 - *erf*(x)). At ω → 0, the long-range term diminishes, while the short-range term mirrors the complete Coulomb operator. Tests with various ω values indicate that ω = 0.11 bohr^{-1} = 0.206 Å^{-1} strikes a favorable balance between computational efficiency and accuracy through a wide range of compounds using HSE06 (Krukau et al. 2006). Despite significantly improved results for equilibrium geometry, band gap and heats of formation (Gerber et al. 2007) compared to standard (semi-)local DFT calculations employing hybrid functionals for periodic systems demands about an order of magnitude more time due to the computational expense associated with computing exact exchange via HF (Duchemin and Gygi 2010; Hasnip et al. 2014).

The DFT+U method is a less computationally expensive alternative to hybrid functionals, only slightly pricier than (semi-)local functionals. It minimizes self-interaction errors by substituting intra-atomic interactions in chosen subshells with empirically parameterized Coulomb (*U*) and exchange integrals (*J*) (Anisimov, Aryasetiawan, and Lichtenstein 1997). The widely used expression for DFT+U follows a rotationally invariant approach (Dudarev and Botton 1998), offering an effective U (U_{eff} = $\bar{U} - \bar{J}$, overbar denoting spherical average) and employs the formula:

$$E_{DFT+U} = E_{DFT} + \frac{U_{eff}}{2}\sum_{\sigma}\left[\left(\sum_{m_1} n^{\sigma}_{m_1,m_1}\right) - \left(\sum_{m_1,m_2} \hat{n}^{\sigma}_{m_1,m_2}\hat{n}^{\sigma}_{m_1,m_2}\right)\right]. \tag{29}$$

Here, $n^{\sigma}_{m_1,m_2}$ represents on-site occupancy matrix elements for corresponding mth states in spin-channel σ. This equation adds a penalty functional to the total energy expression, promoting localization in orbitals by steering the on-site occupancy matrix towards idempotency. The similarity of the U correction in DFT+U to Hubbard's model (Hubbard 1964) for realistic treatment of on-site interactions has earned it the nickname "Hubbard correction".

In practice, U_{eff} is usually adjusted to match experimental results, similar to how hybrid functionals may require tweaking the Hartree–Fock exchange percentage (Verma and Truhlar 2016). Both methods' optimal values vary based on the system and property. In this study, different U values were screened for $Cs_3Sb_2X_9$ (X = Cl, Br, I) in *Chapter 3*, considering band gap and energy level of electronic states from previous HSE06 calculations. More details on DFT+U application is provided in appendix *B.2 Determination of Hubbard U parameters*. Nonetheless, approaches to determine Hubbard values from first-principles exist such as the linear response method (Cococcioni and de Gironcoli 2005) and the more recent ACBN0 method (Agapito, Curtarolo, and Nardelli 2015) which iteratively calculates Hubbard values and approaches HSE06 for many semiconductors, detailed implementation provided in *Appendix C.2.5.* The ACBN0 method was applied in our investigation on *Chapter 4* to provide better estimates of the electronic structure in doped structures in large supercells which can be prohibitively expensive to evaluate with hybrid functionals.

So far, our discussion has been centered on providing a more precise depiction of solids through their first-principles interactions. However, actually calculating total energy, wavefunctions, and material properties requires a few critical elements: expressing KS orbitals using a finite basis set $\{\phi\}$ for the infinitely-many electrons in a solid, using efficient methods for ion representation, and optimized tools to extract meaningful data from wavefunctions. These considerations aim to capture a material's setup at its core while balancing computational expenses and will be explained in detail in subsequent sections.

2.2.5 Periodic boundary conditions and the plane-wave basis set

The resolution of the Kohn-Sham equations (19) can be simplified by taking advantage of this periodicity and exploiting properties of the reciprocal space as expressed in Bloch's theorem (Equation 6). All solutions to the Kohn-Sham equation can be expressed in the form of a plane wave function multiplied by a function $u_{n\boldsymbol{k}}$, the Bloch orbital, conforming to the crystal's periodicity:

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{kr}}u_{n\boldsymbol{k}}(\boldsymbol{r}). \tag{30}$$

As a result of the periodicity, the distinguishable **k** vectors are confined to a primitive cell of the reciprocal lattice, the first Brillouin Zone (BZ) (Jensen 2017).

It is worth noting that the quantum state label *i* from the single-electron orbital has been substituted by ***nk**,* where *n* represents the band index and ***k*** is the wave vector in the first BZ. At first sight, replacing the infinite number of electrons in a crystal with an infinite number of wave vectors ***k*** in the first BZ may seem equivalent. However, the wave functions at k-points that are sufficiently close exhibit significant similarity, and a sampling method can be applied. Numerous methods have been developed to identify specific sets of k-points for effectively sampling the BZ. In this study, the Monkhorst-Pack scheme (Monkhorst and Pack 1976), widely recognized and employed in literature, was utilized for all simulations.

Treating the electronic structure of solids in reciprocal space allows to compute several quantities, such as the electronic density, by simply integrating across the BZ. For instance:

$$f(\boldsymbol{r}) = \frac{\Omega}{(2\pi)^3}\int_{BZ} F(\boldsymbol{k})\, d\boldsymbol{k} = \sum_j w_j F(\boldsymbol{k}_j). \tag{31}$$

Here, $F(\boldsymbol{k})$ denotes the Fourier transform of real-space function *f*(**r**), Ω is the real space cell volume and w_j represents the weighting factors that collectively sum up to one. Moreover, handling point group symmetry becomes simpler, allowing for a reduction in the number of k-points needed for sampling. This reduction is achieved by adjusting the weights and sampling points solely within the irreducible wedge of the first BZ.

The Bloch orbital given by equation (30) can be expanded into plane waves as shown in the equation:

$$u_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_G c_{n\boldsymbol{k}}(\boldsymbol{G})e^{i\boldsymbol{G}\boldsymbol{r}}. \tag{32}$$

where c_{nk}(**G**) represents the coefficient for the plane wave $e^{i\boldsymbol{G}\boldsymbol{r}}$, with **G** representing the reciprocal lattice vectors defined by **G**·**R** = 2πn, where n is an integer. The complete Bloch functions, can then be expressed as a discrete plane-wave expansion:

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_G c_{n\boldsymbol{k}}(\boldsymbol{k}+\boldsymbol{G})e^{i(\boldsymbol{k}+\boldsymbol{G})\boldsymbol{r}}. \tag{33}$$

While an infinite number of reciprocal lattice vectors are needed for precise description of the orbitals, it has been observed that for sufficiently large |**G**|, the contribution of plane waves diminishes exponentially, allowing for truncation of the series without significant information loss. Therefore, a finite basis set, determined by a chosen cutoff energy (E_{cut}), optimizes computational resources.

Determining a suitable E_{cut} involves convergence tests on benchmark quantities usually total energy or the cohesive energy for solids. These tests involve systematically increasing the cutoff energy until the calculated properties converge to a stable value that does not significantly change with further increases in E_{cut}. This convergence ensures that the chosen E_{cut} captures the essential physics of the system while balancing computational efficiency. Similarly, the procedure extends to determining the number of sampled k-points within the irreducible Brillouin Zone (BZ), employing a predefined sampling scheme (Sholl and Steckel 2009).

The representation of one-electron orbitals in plane waves offers advantages owing to its completeness and simplicity. It enables the utilization of optimized numerical libraries for Fourier transforms and allows for a high level of parallelization. These features also facilitate analytical calculations for energies and their derivatives, such as forces and stresses, in comparison to localized basis sets. However, when dealing with surfaces or isolated structures, the use of periodic boundary conditions demands supercells, necessitating the inclusion of empty space (vacuum) in the system to prevent spurious interactions between periodic replicas at the cost of increased computational resources. Typical values might range from 10 to 15 Å, but this can vary significantly depending on the specific system and a convergence test is advised. Similarly, for doping studies, employing a supercell approach in which the doping site can be considered sufficiently localized is essential and depends on the specific case (Martin 2020; Sholl and Steckel 2009).

In this study, plane-wave basis sets were consistently employed in all simulations. Prior to each investigation, converged energy cutoffs and grid sizes for k-point sampling were predetermined for the base compounds, choosing the most rigorous convergence criteria to be uniformly applied to all other structures. The criterion employed in this selection process aimed for a 0.001 eV/atom difference relative to the extrapolated cohesive energy for both k-point sampling and plane-wave energy cutoff. Additionally, all simulations used Γ-centered k-grids which preserves symmetry of the hexagonal and trigonal lattices (Patel, Dabhi, and Vora 2022). When dealing with surfaces and isolated systems, as addressed in *Chapter 3*, convergence calculations were performed to define the required size of the vacuum layer by monitoring the total energy variation. Moreover, high-concentration doping or alloying were considered when introducing heteroatoms. This approach is suitable because the chemical flexibility of halide perovskites generally allows for high concentrations of substituent atoms (X. Zhang et al. 2019). Hence, we are not concerned in generating very localized states to study the effects of doping on the host material.

2.2.6 Pseudopotentials and projector-augmented wave method

The representation of tightly bound core states' sharp peaks and the oscillating valence states within the core region, due to the orthogonalization constraint, requires extremely short wavelength plane waves (high E_{cut}) for an accurate description of the atomic core. However, these highly localized core states have minimal impact on the material properties since they are mostly inert to the chemical environment. Therefore, for an efficient computation using plane waves, the pseudo-ion containing the nucleus and the core electrons is better approximated by a screened pseudopotential interacting with the valence electrons. The design of this pseudopotential aims for its scattering characteristics with valence electrons to mirror those of the all-electron potential but yielding a smooth wavefunction without nodes (Ψ_{pseudo}) that decays exactly like the all-electron wavefunction (Ψ_{AE}) outside a cutoff radius r_c.

There are two primary methods for constructing pseudopotentials: norm-conserving and ultrasoft. Norm-conserving pseudopotentials prioritize accuracy and transferability between different systems by ensuring the same norm of the true and pseudo-wavefunctions within the pseudized core region ($r \leq r_c$), albeit resulting in relatively harder potentials. On the other hand, ultrasoft pseudopotentials relax the norm-conserving condition, resulting in smoother potentials that are still highly transferable but with reduced plane-wave cutoffs. Implementing these pseudopotentials in the Kohn-Sham equations involves replacing the ionic potential ($v_{ext}(\mathbf{r})$ in equation (18)) with the pseudopotential, $v_{ps}(\mathbf{r})$, while keeping other terms unchanged.

Alternative to the pseudopotential approach, the projector augmented-wave (PAW) formalism (P. E. Blöchl 1994) is an extension of the former that preserves the core orbitals. This is achieved through a linear transformation from pseudo-orbitals, $\tilde{\phi}_{n\boldsymbol{k}}$:

$$\phi_{\boldsymbol{nk}} = \tilde{\phi}_{\boldsymbol{nk}} + \sum_i \left[(\varphi_i - \tilde{\varphi}_i) \underbrace{\left(\int \tilde{p}_i^*(\boldsymbol{r}') \phi_{\boldsymbol{nk}}(\boldsymbol{r}') d^3\boldsymbol{r}' \right)}_{c_i} \right]. \tag{34}$$

Where the all-electron partial waves are represented by φ_i and are solutions of the radial Schrödinger equation for a non-spin-polarized reference atom at a specific energy and momentum. $\tilde{\varphi}_i$ serves as a smoother version of φ_i in the augmentation region (r ≤ r_c), matching exactly outside of it. In the interstitial region between the PAW spheres, the orbitals $\tilde{\phi}_{\boldsymbol{nk}}$ are identical to the exact orbitals $\phi_{\boldsymbol{nk}}$. However, inside the spheres, the pseudo-orbitals serve merely as a computational tool and offer an inaccurate approximation to the true orbitals. The last equation is required to map the auxiliary quantities $\tilde{\phi}_{\boldsymbol{nk}}$ onto the corresponding exact orbitals through the projectors $\tilde{p}_i$, which are fitted to yield the appropriate coefficients c_i. In practice, core electrons remain fixed in the configuration used to generate the PAW dataset. Similarly, different configurations necessitate the production of new projectors, much like in the case of pseudopotentials (P. E. Blöchl, Kästner, and Först 2005).

Selecting the appropriate pseudopotential or PAW dataset for simulations involves weighing computational cost against accuracy. Tools like the standard solid-state pseudopotentials (SSSP) website aid in comparing various implementations across chemical elements (Prandini et al. 2018). In this work, we employed norm-conserving pseudopotentials from the PseudoDojo project (van Setten et al. 2018) for the study in *Chapter 3*. For *Chapter 4*, we used ultrasoft GBRV pseudopotentials (Garrity et al. 2014), except for optical property calculations, which required norm-conserving pseudopotentials. In those cases, we applied Vanderbilt pseudopotentials from the SG15 collection (Schlipf and Gygi 2015). These selections were the result of testing, considering the distinct computational demands and elemental compositions pertinent to each project. Finally, the PBE PAW pseudopotentials provided in the Vienna Ab initio Simulation Package (VASP) *v.*5.4 (Kresse and Joubert 1999; Kresse and Furthmüller 1996) were applied in the calculations described on the project in *Chapter 6* to enable direct comparison of total energies to materials databases.

2.2.7 Computational implementation

The investigations in this thesis employed ab-initio electronic-structure calculations using plane waves and pseudopotentials within the density functional theory (DFT) framework as implemented in the widely used ab initio packages Quantum ESPRESSO (QE) (Giannozzi et al. 2009) for the studies on *Chapters 3* and 4, and VASP (Kresse and Furthmüller 1996) for calculations in *Chapter 6*. Despite implementation differences, both codes follow a similar underlying procedure, which will be sequentially discussed, with VASP notably optimized for PAW formalism.

In essence, the primary role of a DFT code involves computing a system's energy and pertinent properties in its ground state. The suite of integrated codes revolves around core executables, such as `pw.x` in QE, focusing on tasks such as geometric configuration optimization and self-consistent potential evaluation (as defined in equation (18), resulting in the determination of charge density and total energy for a relaxed ground state structure.

For solid-state applications, electron wave functions expanded using plane waves (equation (33), simplify the KS differential equations (equation (19) into the following eigenvalue problem (Fiolhais, Nogueira, and Marques 2003):

$$\sum_{\boldsymbol{G}'}\left[\frac{1}{2}|\boldsymbol{k}+\boldsymbol{G}|^2\delta_{\boldsymbol{G}\boldsymbol{G}'}+\tilde{v}_{ext}(\boldsymbol{G}-\boldsymbol{G}')+\underbrace{4\pi\frac{n(\boldsymbol{G}-\boldsymbol{G}')}{(\boldsymbol{G}-\boldsymbol{G}')^2}}_{\tilde{v}_{Hartree}(\boldsymbol{G}-\boldsymbol{G}')}+\tilde{v}_{xc}(\boldsymbol{G}-\boldsymbol{G}')\right]c_{n\boldsymbol{k}}(\boldsymbol{G}')= \\ =\epsilon_{n\boldsymbol{k}}c_{n\boldsymbol{k}}(\boldsymbol{G}), \tag{35}$$

where $\tilde{v}$ refers to the Fourier transform of the respective potential[†]. For each k-point included in the BZ sampling, there are as many equations as the number of plane waves coupled through the self-consistent electron density, given by:

$$n(\boldsymbol{r}) = \sum_{n,\boldsymbol{k}} w_{\boldsymbol{k}} \sum_{\boldsymbol{G},\boldsymbol{G}'} f(\epsilon_{n\boldsymbol{k}}) c^*_{n\boldsymbol{k}}(\boldsymbol{G}') c_{n\boldsymbol{k}}(\boldsymbol{G}) e^{i(\boldsymbol{G}-\boldsymbol{G}')\boldsymbol{r}},$$

or, the equivalent in reciprocal space:

$$n(\boldsymbol{G}) = \sum_{n,\boldsymbol{k}} w_{\boldsymbol{k}} \sum_{\boldsymbol{G}'} f(\epsilon_{n\boldsymbol{k}}) c^*_{n\boldsymbol{k}}(\boldsymbol{G}'-\boldsymbol{G}) c_{n\boldsymbol{k}}(\boldsymbol{G}').$$

[†] Notice that the case G,G’= 0 is a special case of the equation since the Hartree potential and ion-ion interactions will diverge, but calculating the limit the divergence disappears and a constant value is obtained, refer to Eq. 6.31 from *A Primer in Density Functional Theory* (Fiolhais, Nogueira, and Marques 2003).

Where $f(\epsilon_{n\boldsymbol{k}})$ represents the occupation number of the KS state $n\boldsymbol{k}$, each weighted by the corresponding contribution from BZ sampling, $w_{\boldsymbol{k}}$. Reciprocal space calculations prove useful for certain aspects of the effective potential due to their computational advantages. For example, the kinetic energy is diagonal in reciprocal space and the Hartree potential becomes a simple product as shown in (35). However, when it comes to external and exchange-correlation potentials, real space computation is more effective. This requires a seamless conversion of data between these spaces, efficiently achieved by leveraging the Fast Fourier Transform algorithm (Frigo and Johnson 2005).

The usual method to solve these equations involves matrix diagonalization, such as the block Davidson diagonalization algorithm (Davidson 1975), with the matrix size determined by the chosen energy cutoff E_{cut}. Solving via diagonalization scales with N_e^3, where N_e is the number of electrons in a unit cell (Levitt and Torrent 2015). Next, the KS equations are solved to find the single-particle eigenvalues and wave functions for a specific nuclear configuration. A new electron density is then calculated from these wave functions. Self-consistent changes in total energy or electron density from the previous step are verified at this point. Total energy differences below 10^{-6} eV/atom is an usual threshold for SCF cycle convergence, in QE this value is passed on the keyword `conv_thr`. If self-consistency is not reached, the current electron density is combined with the previous cycle's density to produce a new one. The mixing algorithm can be tuned and plays an important role on achieving convergence (A. S. Banerjee, Suryanarayana, and Pask 2016). When self-consistency is achieved, various quantities such as total energy, atomic forces, stress within the unit cell, and electronic band structures can then be computed for the atomic arrangement.

If forces and stresses exceed a set tolerance, atomic positions and cell parameters are adjusted and electronic iterations restart. Frequently a quasi-Newton relaxation algorithm like BFGS (Billeter, Curioni, and Andreoni 2003) is used for the geometrical optimization. Every atomic iteration contains multiple electronic iterations and after several atomic iterations, the system should reach equilibrium, concluding the calculation. For geometrical optimization, a convergence threshold in total energy and another one for forces must be satisfied simultaneously, usual thresholds for these quantities are below 10^{-6} eV/atom and below 0.01 eV/Å, respectively. These correspond to the keywords `etot_conv_thr` and `forc_conv_thr`, respectively, in QE.

Once v_{eff} is known, the system's electronic density and the corresponding hamiltonian are established. `pw.x` can then be employed to solve the KS equations in a non-self-consistent manner, generating KS eigenvalues for specific scenarios — like along a specified path in the Brillouin Zone (BZ) for band structure calculations or on a denser k-point grid for density of states (DOS) and projected density of states (PDOS) calculations. From the converged electronic density, charge density plots can be generated, and Bader charge analysis performed. Moreover, the final wavefunctions and density of the ground-state structure serve as starting point to compute multiple dynamical properties such as phonon and optical spectra.

2.2.8 From ground-state DFT simulations to materials properties

Moving past the process of acquiring the ground-state wavefunction and electronic density of a material via Density Functional Theory (DFT), our focus now shifts to extracting materials properties that align with our outlined interests in section 2.1.2.

Within DFT framework, the total energy emerges as a cornerstone for computing several key properties such as binding energies and formation enthalpy. The formation enthalpies enable a comparative analysis of formability of a given compound and may help assess their thermodynamic stability via the convex hull approach (Barber, Dobkin, and Huhdanpaa 1996; Bartel 2022). In the context of perovskites this is particularly useful to evaluate heteroatom doping as done for halide alloying in *Chapter 3* and *4* and for B-site doping in *Chapter 4*. In fact, the science of point defects is deeply rooted on evaluating energies (Freysoldt et al. 2014). Total energies can be also used in surfaces to estimate surface free energy as described in Eqs. (B3) and (B4) and were applied to compare the surface formability between different directions for each of the $Cs_3Sb_2X_9$ perovskites on *Chapter 3*.

Band structure and DOS/PDOS diagrams are indispensable tools in ab-initio simulations, revealing the core elements of a material's electronic structure. In semiconductors, these analyses are critical for determining band gaps, distinguishing between direct and indirect band gaps, and assessing effective masses from the curvature of the bands. PDOS curves elucidate contributions of different atomic species and orbitals within both valence and conduction bands, pivotal knowledge for tailoring materials for specific applications. Moreover, exploring the spatial projection of Kohn-Sham (KS) orbitals, such as in the examination of distributions of highest occupied/lowest unoccupied states, aids in understanding optoelectronic transitions. This is particularly useful for studying the impact of heteroatoms on band structure alterations, as demonstrated in *Chapter 4*, or for comparative analysis, as shown in the cluster investigation in *Chapter 3*. Spin-density plots, obtainable through similar projections, offer deeper insights into magnetic properties. Most discussions in this thesis stem from interpreting these curves and plots, which detailed obtention is described in *Appendix A.4*.

Electronic structure changes can be further understood by analyzing the charge transfer through the charge density plots and charge analysis by methods such as Bader and Lowdin charge analysis, formalism detailed on *Appendix A.5*. In *Chapter 4*, this approach was frequently employed to assess the impact of B-site doping on surrounding halogens and comprehend difference in formation energies. Changes in charge distribution follow geometric alterations which result from proper geometric optimization of the initial structure. Consequently, the interplay among these properties offers deeper insights into atomic-level processes, enriching discussions and enhancing comprehension of material properties. The role of geometry and charge transfer was considered in all studied structures in this thesis, as an example, the role of geometry and charge transfer was crucial to trace the atomistic origin of the PDOS distribution and band gap for $Cs_3Sb_2X_9$ interfaces and clusters on *Chapter 3*.

While valuable insights into material properties derive from ground-state DFT calculations, these represent merely a fraction of the whole picture. Once the ground-state density and wavefunction are acquired, leveraging them as inputs for higher-level calculations, such as many-body perturbation theory (MBPT) or density functional perturbation theory (DFPT), opens avenues to approximate a myriad of dynamical properties (Onida, Reining, and Rubio 2002; Yip 2005). These encompass response and spectroscopic properties such as phonon frequencies, elastic constants, thermal conductivity, dielectric tensors, electron energy-loss spectra, electronic excitations, optical absorption spectra, among others. Calculating these properties typically involves high computational costs. This is due to the need for computing derivatives of the wavefunction in reciprocal space, which requires solutions in denser k-grids to ensure adequate accuracy, alongside the standard computation of observables (P.

Giannozzi et al. 2017). In this work, optical absorption spectra within the independent-particle approximation (Del Sole and Girlanda 1993) has been computed to better assess the viability of studied metal-doped $Cs_3Sb_2I_9$ polymorphs as solar cell absorbers, the detailed methodology for this method is presented in *Appendix A.6*.

The atomic-level insights gained from simulations highlight why DFT is prevalent in materials science, aiding in understanding structure-property relationships often elusive in experiments. These simulations help optimize material properties across various compositions and applications (Frauenheim et al. 2002). Today, the field benefits from large materials databases and high-throughput calculations, rapidly screening materials for diverse compositions and applications (Saal et al. 2013). The fusion of streamlined simulations with databases, alongside the integration of machine learning algorithms, marks a significant breakpoint in materials exploration, fundamentally enhancing predictive capabilities and expediting the discovery of novel materials (Y. Liu et al. 2017). These transformative advancements will be further clarified in the upcoming sections.

2.3 High-throughput calculations and large materials databases

With the advances in simulation methods combined with the large increase in computational capacity results in a major reduction in time used to perform calculations, so a relatively larger time is spent on simulations setup and analysis, as illustrated in *Figure 8*. This changed the theoretical workflow of the computational materials scientist and led to new strategies. Rather than conducting numerous manually crafted simulations, there's now the capability to automate input generation and execute hundreds of simulations concurrently and in sequence. This evolution represents what is commonly referred to as a high-throughput (HT) workflow (Schleder et al. 2019).

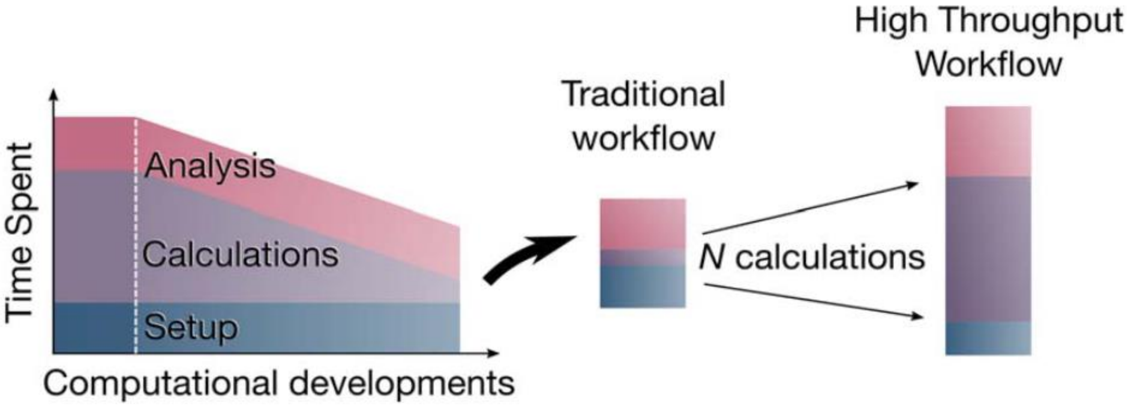


Figure 8 – The duration required for computations varies in relation to technological progress. As computer technology advances, the computational phase may become less time-intensive compared to the setup's construction and the subsequent analysis of results. Source: (Schleder et al. 2019)

These high-throughput DFT calculations (HT-DFT) methods are typically executed through a tripartite process: (i) conducting electronic structure computations for numerous synthesized and hypothetical materials; (ii) methodically storing information in databases; and (iii) screening and data mining: typically involve verifying stability and identifying potentially innovative materials. Subsequently, new physical insights are derived through further calculations or experiments (Schleder et al. 2019; Körbel, Marques, and Botti 2016). *Table 1* showcases popular HT-DFT tools, each with unique functionalities and varying complexities. Yet, they commonly support tasks like manipulating crystallographic structures, managing input/output for different DFT software, and conducting basic material property analysis.

Table 1 – Popular HT-DFT tools used in materials science. Adapted from: (Song et al. 2020)

Name	Function	URL
Pymatgen	Robust, open-source python library for materials analysis.	https://pymatgen.org
AFLOWπ	Minimalist framework for high-throughput first principles calculations.	http://aflowlib.org/src/aflowpi
FireWorks	Open-source code for defining, managing, and executing workflows.	https://materialsproject.github.io/fireworks
AiiDA	Workflow to automate complex numerical procedures.	http://www.aiida.net
Pymatflow	Workflow simplifier for materials science research.	http://pymatflow.readthedocs.org
ASE	Setup, steering, and analysis for atomistic simulations.	https://wiki.fysik.dtu.dk/ase
Atomate	Workflow built on top of pymatgen, custodian, and FireWorks.	https://atomate.org
Custodian	Simple, robust, and flexible just-in-time job management framework.	https://pypi.org/project/custodian

The screening or mining process involves applying specific criteria to a database to choose the best candidates based on desired attributes. This process filters materials in a step-by-step manner, eliminating those that do not meet the constraints. Top candidates are then assessed to understand why they excel and to predict potential further improvements. Materials meeting the criteria can be ranked based on defined merits, allowing further investigation or application (Y. Wu et al. 2013; Curtarolo et al. 2013).

The constraints can serve as filters guided by prior knowledge of phenomena and properties, or as descriptors derived from machine-learning processes, as will be discussed later. Typically, the filtering process begins with an analysis of thermodynamic stability to pinpoint a subset of potentially stable materials. Subsequent filters are then tailored to the specific application being sought, available resources and research design following a funnel-type model as illustrated in *Figure 9*. For example, in fields like photovoltaics and optoelectronics, a desirable attribute is usually a band gap within the visible-light absorption/emission range and suitable effective masses. Materials excelling in these aspects might undergo further evaluation using more costly methods, such as hybrid functional calculations for more accurate band structure prediction. This assessment could also involve acquiring optical absorption curves and confirming the material is stable. Dynamic and thermal stability may be verified through phonon dispersion curves and ab-initio molecular dynamics simulations (Q. Li et al. 2021; Cai et al. 2022). Ultimately, the materials acquired through this method are open to experimental testing. Present-day literature brims with successful cases of new materials discovered through the high-throughput screening approach (Sanvito et al. 2017; Schlexer Lamoureux et al. 2019; J. Yang and Mannodi-Kanakkithodi 2022; H. Luo et al. 2023).

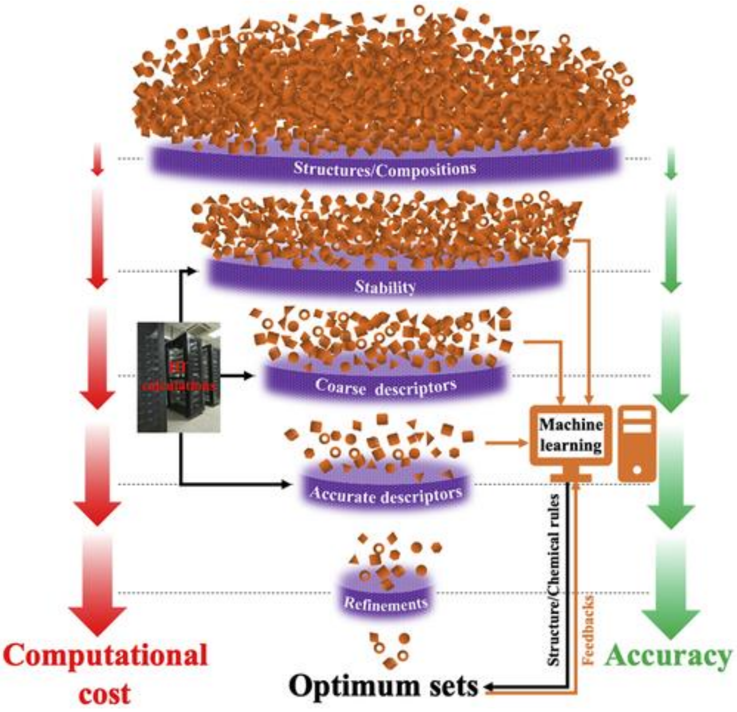


Figure 9 – The funnel type model of high-throughput computational screening.
Source: (S. Luo et al. 2021)

Methods for discovering novel materials using high-throughput (HT) techniques are closely tied to managing extensive datasets. The accessibility of this data, often available in theoretical databases, fosters collaboration within the scientific community, an important aspect in advancing innovative applications within this rapidly expanding field. Theoretical and experimental databases serve multiple purposes, such as enhancing battery technologies, exploring new catalysts, designing efficient thermoelectric materials, and creating high-performance optoelectronic devices (Schleder et al. 2019; Song et al. 2020).

Table 2 highlights some of the largest materials databases, encompassing both experimental and computational data. Notably, the most popular for computational data include the Materials Project, a result of the multimillion-dollar Materials Genome Initiative (A. Jain et al. 2013), alongside the subsequently launched AFLOWLIB (Curtarolo et al. 2012) and Open Quantum Materials Database (OQMD) (Saal et al. 2013). The Materials Project contains over 150,000 materials, while OQMD features data on more than 1 million materials, and AFLOWLIB boasts over 3.5 million entries. All three databases share a core collection of over 50,000 experimentally obtained materials sourced from the widely used Inorganic Crystal Structure Database (ICSD) (Belsky et al. 2002; Nosengo 2016).

Table 2 – Popular material databases. Multiple stands for inorganic and organic materials also for mixed experimental and computational data. Adapted from: (Song et al. 2020)

Name	Data type	URL	Free
Materials Project	Multiple	https://materialsproject.org	√
ICSD	Inorganic & Experimental	https://icsd.fiz-karlsruhe.de	×
AFLOWLIB	Inorganic & Computational	http://aflowlib.org	√
COD	Multiple & Experimental	http://crystallography.net	√
OQMD	Multiple & Computational	http://oqmd.org	√
NOMAD	Multiple	https://nomad-repository.eu	√
JARVIS	Computational	https://jarvis.nist.gov	√
Materials Cloud	Multiple	https://www.materialscloud.org	√
Materials Commons	Computational	https://materialscommons.org	√
CSD	Multiple	https://www.ccdc.cam.ac.uk	×

Their differences lie in the hypothetical materials they include: the Materials Project focuses on materials with a reasonable chance of being synthesized, whereas AFLOWlib and OQMD loosen this restriction, accommodating compounds that may not ever be synthesized but offer significant insights into compound formability (Nosengo 2016; Balachandran et al. 2018). AFLOWlib specializes in providing extensive data on alloys and disordered materials (Toher and Curtarolo 2023), while OQMD offers particularly wide coverage of perovskites (Shen et al. 2022). Additionally, OQMD stands out as the most open among the three: users can download the entire database onto their computer, not just individual search results, fostering broader accessibility and utilization. For these reasons, OQMD was our choice of dataset to model perovskite properties with machine learning models as explored on *Chapter 6*.

In addition to using a standard set of pseudopotentials in their computations, all large materials databases employ fixed parameters for k-point sampling and plane wave energy cutoff. This speeds up screening while preserving adequate accuracy for fundamental properties (formation energies and crystal volume) (Hegde et al. 2023). Thus, a direct comparison of total energies is made possible by using the same simulation settings as those used for the database. We employed this strategy in our perovskite HT simulations with *Atomate2* (Mathew et al. 2017; Ganose et al. 2024), replicating OQMD’s simulation settings. By analyzing the convex hull distance relative to OQMD entries, we were able to estimate the thermodynamic stability and decomposition products of the calculated structures to screen for viable compounds.

While materials databases offer immense power, acknowledging their limitations is important. The limitations of materials datasets encompass two primary challenges. Firstly, there's a scarcity of high-quality data due to an imbalance between vast computational datasets, primarily derived from cost-effective methods like DFT with Generalized Gradient Approximation (GGA) or classic molecular dynamics, and a lack of comprehensive experimental data or data from more expensive, accurate computational methods. This results in a significant disparity in data points available for properties like formation enthalpies, band gaps, or thermal conductivity (Gong et al. 2022). Secondly, these datasets often exhibit biases that affect their representativeness across the materials space. Biases range from favoring specific elements or compound types to excluding certain structural motifs or limiting primitive cell sizes. These biases hinder the broader applicability of these datasets, impacting their utility in comprehensive material understanding and machine learning applications (S. Kim et al. 2020; Muy et al. 2019). Nevertheless, the literature is teeming with successful machine learning applications in materials science and is effectively revolutionizing the field while these challenges are mitigated. In the next section, we thoroughly explore the implementations and implications of this research field.

2.4 Machine Learning for Materials Science

The evolution of materials science mirrors the broader evolution of science and technology through history. Initially, empirical observations, particularly in metallurgy across different ages (stone, bronze, iron, steel), formed the foundation. Then, a few centuries ago, theoretical models and generalizations emerged, represented by mathematical laws like the laws of thermodynamics in materials science. However, the complexity of theoretical models grew over time, making analytical solutions impractical for many scientific problems. The advent of computers introduced a third paradigm—computational science. This paradigm enabled simulations of intricate real-world phenomena based on the theoretical models from the second paradigm. Examples in materials science include density functional theory (DFT) and molecular dynamics (MD) simulations. Each scientific paradigm has contributed to advancing its predecessor, leading to the popularization of theory, experiment, and computation across various scientific fields (Agrawal and Choudhary 2016).

Recently, the burgeoning volume of data generated by experiments and HT simulations has birthed the fourth paradigm—data-driven science. This paradigm integrates theory, experiment, and computation/simulation, unifying the earlier paradigms as illustrated in *Figure 10*. Because of the vast amounts of data collected in materials databases, this trend has gained traction in materials science, resulting in the emergence of materials informatics as a new field within the discipline (Agrawal and Choudhary 2016; Schleder et al. 2019).

Materials informatics aims to discover the connection between materials attributes and their properties. Due to the complexity of the patterns across vast materials landscapes, machine learning (ML) models are generally utilized since they are essentially function approximators (Takeshima 2022). In this context, we seek to answer for a material $\mathbf{x}_i$, represented by appropriate descriptors as a vector or a graph (Damewood et al. 2023), what is its property $y_i = \hat{f}(\boldsymbol{x_i})$. There are three types of machine learning algorithms: supervised learning, unsupervised learning and reinforcement learning (Alloghani et al. 2020). Our focus in materials science is usually on supervised learning algorithms, which are models that map inputs to outputs, and attempt to extrapolate patterns learned in past data on unseen data. Supervised learning algorithms can be either regression models, in which we attempt to predict a continuous variable, such as the band gap — or classification models, where we try to predict a binary or multi-class variable, such as whether a material is a metal or semiconductor/insulator (Dunn et al. 2020). Machine learning (ML) models offer a significant advantage for predicting material properties, as they can reduce the computational cost by orders of magnitude compared to ab initio simulations (Tawfik and Russo 2022).

Following what is known as the supervised approach, the ML algorithm will tackle this problem by learning the patterns on a given dataset, this phase is denominated *training* and results in a ML model with the appropriate parameters that hopefully generalizes to other materials, outside the dataset, and provides predictions. The ML approach is expected to reveal feature-property connections that are not apparent to human observation. This contrasts with theoretical models, which are rooted in the underlying physical theories behind the data to make predictions.

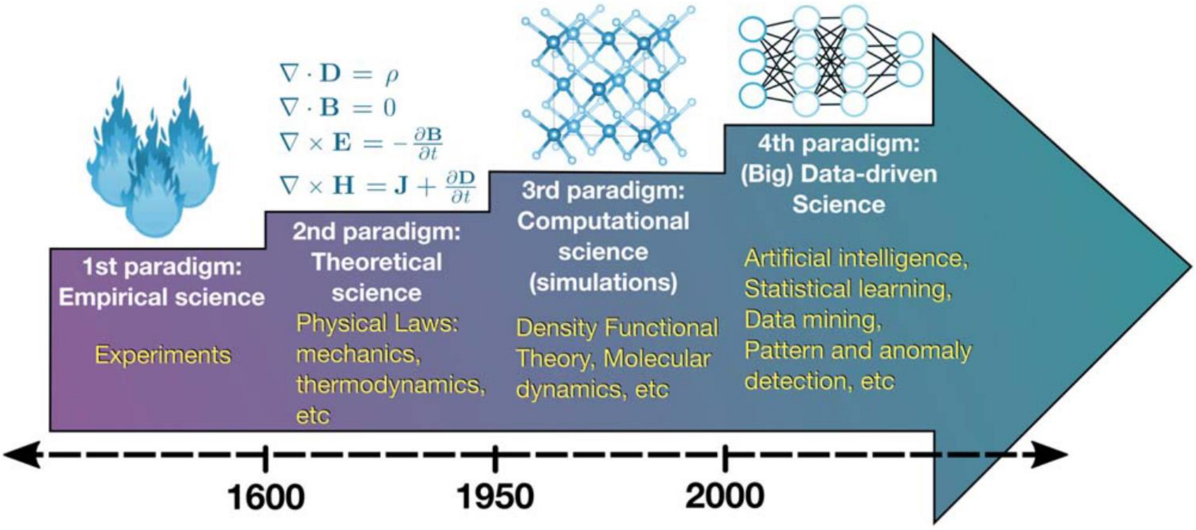

$\nabla \cdot \mathbf{D} = \rho$
$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$
1st paradigm: Empirical science
Experiments
2nd paradigm: Theoretical science
Physical Laws: mechanics, thermodynamics, etc
3rd paradigm: Computational science (simulations)
Density Functional Theory, Molecular dynamics, etc
4th paradigm: (Big) Data-driven Science
Artificial intelligence, Statistical learning, Data mining, Pattern and anomaly detection, etc
1600
1950
2000

Figure 10 – Evolution of science through the four paradigms. Source: (Schleder et al. 2019)

2.4.1 Machine learning model training: strategies and tradeoffs

There are numerous machine learning algorithms, each of which is better suited to a certain issue and/or dataset. This is consistent with the “No Free Lunch Theorem”, which states that no ML algorithm can be considered universally superior. This means that the goal of ML is not to find the best learning algorithm. Instead, we must determine what type of distribution is relevant to our specific application in materials science and which ML algorithm performs best on that data. As a result, we can try a variety of algorithms to train a model, each with a distinct speed-interpretability-accuracy tradeoff (Murphy 2012). In *Figure 11* some of the most common ML algorithms and their tradeoffs are presented. An overview of common supervised and unsupervised ML algorithms is given in *Appendix A.7* for the sake of brevity. The figure reveals that simple interpretable algorithms like classification rules or linear regression are often inaccurate due to limited parameters, while flexible deep neural networks achieve high accuracy but are often "black boxes" in interpretability.

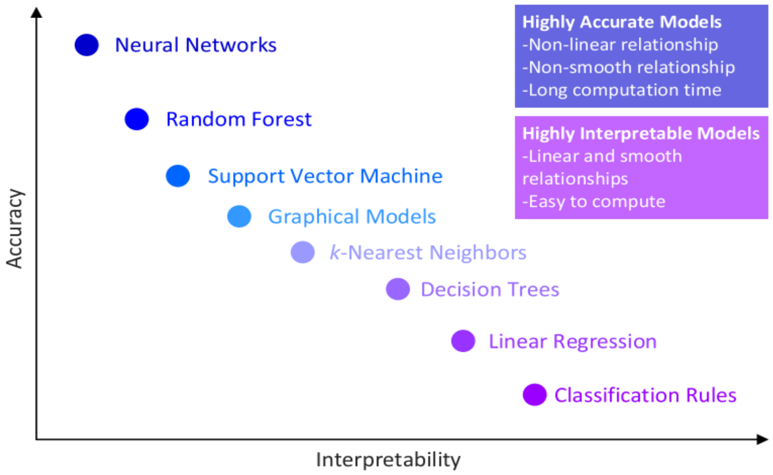


Figure 11 – The trade-off between interpretability and accuracy of some relevant ML models. Source: (Morocho-Cayamcela, Lee, and Lim 2019)

This compromise between interpretability and accuracy underlines the broader bias-variance tradeoff inherent in function approximators. (Vapnik 2000; Geman, Bienenstock, and Doursat 1992). Complex models, with a high number of parameters, possess the capability to capture nonlinear patterns across a high-dimensional space, incorporating the contributions of multiple descriptors. However, these models have several limitations, including the difficulty to extract meaningful relationships between predictions and descriptors, greater computing expense during training, and the inherent risk of overfitting — a situation where the model fittingly memorizes noise as if it were signal. This phenomenon tends to amplify variance in predictions for new data, as slight alterations in input data can lead to wide fluctuations in forecasts, although there are exceptions (Neal 2019). Conversely, simpler models with fewer parameters tend to suffer from underfitting. These models inadequately capture the essential relationships between descriptors and properties, resulting in systematic errors, or bias, within their predictions. This presents a fundamental challenge: striking a balance between complexity and accuracy in machine learning models (Rashidi et al. 2019).

To effectively address this challenge, it is crucial to evaluate a model's performance beyond its training data. Typically, this involves dividing the dataset into three separate subsets: training, validation, and test sets. Ensuring these subsets present similar statistical distribution is imperative for proper assessment of performance. The validation set becomes instrumental in optimizing the model's *hyperparameters* — parameters not altered during the training process. These hyperparameters often fine-tune the model's complexity, controlling the risks of underfitting or overfitting. Examples encompass the depth of a decision tree or the number of layers in a neural network, as well as parameters governing training speed and optimization capability, such as learning rates in neural networks. By training multiple models with distinct hyperparameters and assessing their performance on the validation set, the optimal hyperparameters can be identified. Subsequently, with fixed hyperparameters, the model undergoes training, and its predictions on the test set are compared against actual labels to evaluate performance. The separation into validation and test sets serves to prevent hyperparameter tuning that artificially inflates test set results, potentially compromising the model's generalizability during deployment. This meticulous process aims to strike an optimal balance, as depicted in *Figure 12*, ensuring that the model's performance holds true during external testing.

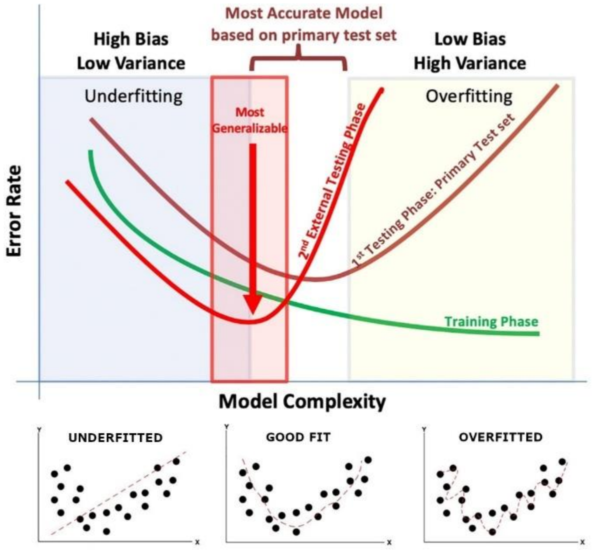


Figure 12 – Illustration of the bias-variance trade-off in machine learning. Training data can be fit to arbitrary precision using complex models, but the problem lies in generalizing to test data. Underfitting produces less variable predictions but high error rate and bias, while overfitting results in low bias and high variance. The ideal zone lies between overfitting and underfitting zones, requiring multiple adjustments to generalize well to validation and testing data. Source: (Rashidi et al. 2019)

Moreover, presenting and effectively utilizing sufficient data is crucial for training a model capable of delivering reliable predictions in applications. This is especially relevant in materials science where data collection can be very costly; experimental data is expensive to obtain, and relying on DFT theoretical calculations still presents a significant cost, particularly for properties that are not directly obtained from ground-state calculations (Rodrigues et al. 2021; Pilania 2021). Fortunately, methodologies have emerged to efficiently leverage available data, such as cross-validation and ensemble methods (elaborated upon in *Appendices A.8* and *A.9*). Equally essential in model training is the selection of appropriate descriptors for prediction and the careful curation of these features to prevent under- or overfitting, as will be explored in the following section.

2.4.4 Descriptors and feature selection

Descriptors, interchangeably referred to as features or variables, encapsulate the characteristics of data points within a dataset, constituting the feature vector $\boldsymbol{x_i}$ used to predict a corresponding property, or set of properties, y_i, through the ML approximated function $\hat{f}$. In the field of materials informatics, features encode information regarding chemical compositions, crystal structures, bonding patterns, and more (Seko, Togo, and Tanaka 2018). Even with advanced algorithms, poor descriptors will consistently result in unsatisfactory ML models. Hence, using effective descriptors which can correlate with the target property is essential for accurate predictions. Nonetheless, apart from predictive power, three other critical elements determine the quality of descriptors, as outlined by Tawfik and Russo (2022):

- *Meaningfulness*: Descriptors should align with physical or chemical principles. This is important to preserve the interpretability of the results and help guide design principles.
- *Computational efficiency*: The computational cost of deriving a descriptor should be substantially lower than that of calculating the target property.
- *Number of entries in the descriptor*: When the descriptor is calculated for the material, it should provide suitable number of entries to add to the feature vector (e.g., no more than a few hundreds). This simplifies the ML model, since numerous descriptors strain storage, processing, and result in opaque, "black-box" ML models.

These four criteria—Meaningful, Efficient, small Number of descriptors, Accurate (MENA)—comprise the benchmarks for ML descriptors in materials science. When it comes to predict DFT-calculated properties in materials science, descriptors can be categorized into four classes, namely, elemental, geometry-based, electronic structure and ab-initio based features. This categorization reflects a gradual progression, where each category elevates accuracy with the tendency to also introduce increasing levels of complexity and computational demands. Let's now examine each category in detail:

- *Elemental Descriptors*: Represent the simplest category, swiftly calculated and intimately connected to elemental traits within a material's structure, such as atomic numbers or elemental melting points. However, their lack of uniqueness can compromise accuracy, especially when detailed structural information is crucial, such as in dealing with polymorphs.
- *Geometry-Based Descriptors:* Drawing from material geometry these features encompass both translationally-invariant geometric and elemental properties. They include symmetry groups, property-labelled materials fragments (PLMFs) (Isayev et al. 2017), geometrical fingerprints, and symmetry functions. While some descriptors in this class, like symmetry functions,

involve mathematically complex operations, they are generally computationally feasible.

- *Electronic Structure Descriptors:* By delving into electronic properties at the atomic level, these features provide valuable insights. However, since they encode information locally their ability to grasp physical properties of the entire structure is still limited, examples of this class include the electronic structure attributes (Ward et al. 2016), molecular orbital attributes (Welborn, Cheng, and Miller 2018) and methods combining local structure with atomic electronic information such as smooth overlap of atomic positions (SOAP) (Bartók, Kondor, and Csányi 2013) and orbital field matrix (OFM) (Lam Pham et al. 2017).
- *Ab Initio-Based Descriptors*: Since they correspond directly to physically-computed quantities such as total electronic energy and molecular orbital energies, they are highly meaningful. However, these descriptors require ab initio calculations either partial or complete at a lower level of theory what entails a heightened computational cost. A prime example are the ROSA (Robust One-shot Ab-initio) descriptors by Tawfik and Russo (Isayev et al. 2015; Tawfik and Russo 2022).

When selecting which features will be in our model, it is crucial to weigh the computational cost of descriptors against their predictive value for the property or properties of interest. An excessive number of features can trigger a problem known as "curse of dimensionality", causing sparsity of the dataset that hampers the algorithm's capacity to learn meaningful patterns in the data. Moreover, high-dimensionality amplifies the impact of fluctuations and outliers, promoting overfitting and undermining generalization. To address this issue, we should first consider how to represent local features of a material.

A compound, denoted as ξ, can be represented by a collection of atomic descriptions, each encapsulating its elemental and structural details within the unit cell. This setup allows to create a matrix where a total of $N_a^{(\xi)}$ rows represent the different atoms in the unit cell of the compound, and a total of N_x columns represent the elemental, structural and electronic local features (i.e., $N_x = N_{x,elem} + N_{x,st} + N_{x,elec}$), the final representation of this feature matrix becomes:

$$\boldsymbol{X}^{(\xi)} = \begin{pmatrix} x_1^{(\xi,1)} & x_2^{(\xi,1)} & \cdots & x_{N_x}^{(\xi,1)} \\ x_1^{(\xi,2)} & x_2^{(\xi,2)} & \cdots & x_{N_x}^{(\xi,2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(\xi,N_a^{(\xi)})} & x_2^{(\xi,N_a^{(\xi)})} & \cdots & x_{N_x}^{(\xi,N_a^{(\xi)})} \end{pmatrix}, \tag{36}$$

where $x_n^{(\xi,i)}$ denotes the nth representation of atom i in compound ξ. However, this representation is not agnostic to the number of atoms in the unit cell. To use this representation effectively in machine learning algorithms on datasets containing materials with varying numbers of atoms in the unit cell, one might consider capping the representation by the largest unit cell and padding the others with zeros. Nevertheless, this approach would introduce sparsity and bias the algorithm towards correlating with the number of atoms rather than focusing on the chemical and structural characteristics. A more effective approach considers the distribution of the local features among the constituent atoms, effectively calculating basic statistics such as mean, standard deviation, maximum, minimum, range, etc., across the columns of the matrix $\boldsymbol{X}^{(\xi)}$ to generate the feature vector **x**$_i$ (Seko, Togo, and Tanaka 2018), this is illustrated in *Figure 13*.

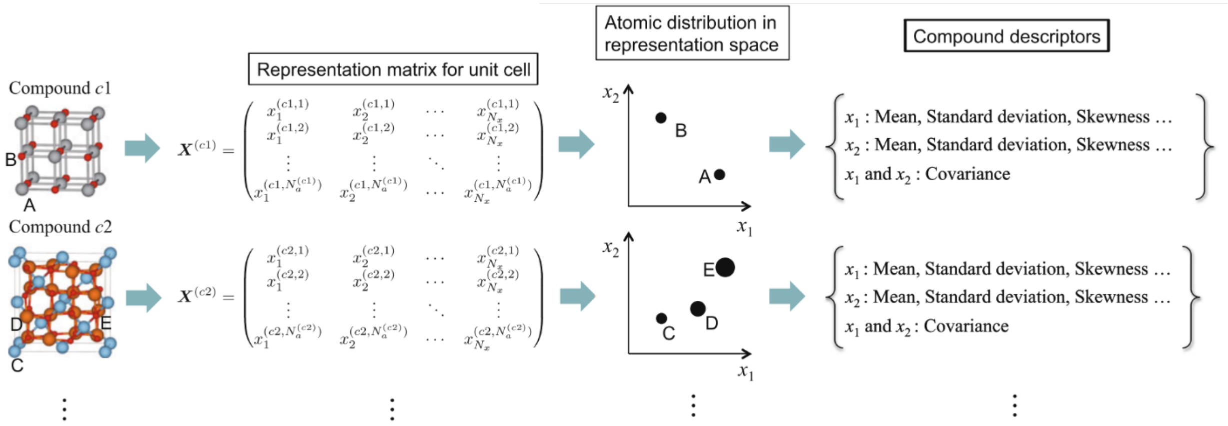

Representation matrix for unit cell
Atomic distribution in representation space
Compound descriptors
Compound c1
B
A
$X^{(c1)} =$
x_2
x_1
B
A
x_1 : Mean, Standard deviation, Skewness ...
x_2 : Mean, Standard deviation, Skewness ...
x_1 and x_2 : Covariance
Compound c2
D
E
C
$X^{(c2)} =$
x_2
x_1
E
C
D
x_1 : Mean, Standard deviation, Skewness ...
x_2 : Mean, Standard deviation, Skewness ...
x_1 and x_2 : Covariance

Figure 13 – Schematic illustration of the generation of generalizable compound descriptors from local features. Source: (Seko, Togo, and Tanaka 2018).

The generation of descriptors has become highly automated in today's context, facilitated by packages such as MatMiner (Ward et al. 2018) which can generate thousands of individual descriptors, roll descriptive statistics and combine them into mathematical functions. An overview of the numerous featurizers included in MatMiner is presented in *Figure 14*. However, an abundance of descriptors may not bolster our machine learning (ML) model. Instead, it could exacerbate the curse of dimensionality, cluttering the model with numerous features that introduce noise. Therefore, we commence the first step in the process—known as feature engineering—by undertaking feature selection.

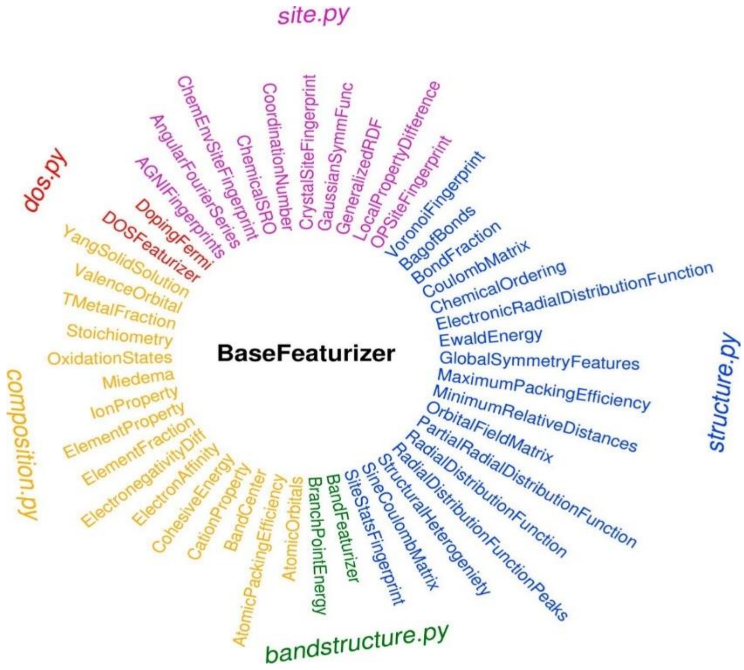


Figure 14 – Matminer includes several featurizers across five modules: composition, site, structure, bandstructure, dos. Each featurizer produces numerous features, enabling MatMiner to generate thousands of unique features. Source: (Ward et al. 2018).

In feature selection, the primary goal is to retain the most relevant and informative features that significantly contribute to the model's predictive power, while excluding irrelevant or redundant ones. Techniques like statistical-based and model-based selection aid in this process. Model-based selection involves using interpretable machine learning models, typically decision tree ensembles, to estimate feature importance. Conversely, statistical-based selection assesses the relationship between features and the target variable or within features themselves to identify redundancy (Venkatesh and Anuradha 2019).

A prime example of statistical-based selection is integrated into the MODNet framework, further explored on *Section 2.4.9.3*, which utilizes *normalized mutual information* (NMI) to select optimal features to the model. NMI is calculated as follows:

$$NMI(X,Y) = \frac{MI(X,Y)}{\left(\frac{H(X)+H(Y)}{2}\right)}. \tag{37}$$

Here, MI denotes the mutual information (Kraskov, Stögbauer, and Grassberger 2004), and H represents the information entropy ($H(X) = MI(X,X)$). NMI yields a normalized value and offers greater flexibility and resistance to outliers to capture associations between variables compared to the Pearson correlation coefficient which assumes linearity. MODNet's feature selection computes NMI of all features $f \in \mathcal{F}$ with the target variable (y) to select the first optimal feature, and subsequently implements a *relevance and redundancy* (RR) score which is repeatedly computed with the already selected features set $f_S \in \mathcal{F}_S$:

$$RR(f) = \frac{NMI(f,y)}{[max_{f_S \in \mathcal{F}_S}(NMI(f,f_S))]^p + c}, \tag{38}$$

where (p, c) are hyperparameters determining the balance between RR which vary with the number of features and were benchmarked to be $p = max(0.1, 4.5 - n^{0.4})$ and c=$10^{-6}n^3$. The selection proceeds until the number of features reaches a threshold, which can be fixed arbitrarily or, ideally, optimized to minimize model error (De Breuck, Hautier, and Rignanese 2021).

While MODNet's process excels in comparison to other model-based selections when benchmarked against various frameworks, the cross-NMI computation becomes intensive as it scales with n^2 where n is the number of features in the initial set. Therefore, in this study, when the initial descriptors in the feature vector $\boldsymbol{x}_i$ exceeded 1500, they were reduced to 1500 using the feature importance score derived from the decision tree ensemble model, XGBoost (T. Chen and Guestrin 2016).

After the critical phase of feature selection, other tools in the feature engineering toolbox ought to be considered before the machine learning model is trained. These include important data preprocessing steps such as normalization, one-hot enconding, missing data imputation, and dimensionality reduction. Also, an essential step is the selection of an appropriate error metric for the problem at hand. The topic of data preprocessing methods is covered in detail in *Appendices A.10* and *A.11*, while the subsequent section elaborates on the choice of the error metric.

2.4.7 Error metrics

Once the method and model type are established and the data properly processed, choosing an error metric aligned with the property measured is the last important decision. The error metric forms the core of the loss function (see Equation

A30), affecting the optimization of model's hyperparameters and weights. Selection depends on factors like property nature, acceptable error, prediction variance, and interpretability preferences. There are numerous error metrics in ML, and customized metrics for specific applications are frequently developed. However, we will focus on the metrics traditionally used in materials science problems. These metrics fall into two primary categories: regression and classification metrics, each named according to the nature of the task they address. Beginning with the regression metrics, we have:

- *Mean Squared Error (MSE):* The most common to estimate fitting accuracy in regression. $MSE = (\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)$, where $\hat{y}_i$ is the predicted value of the i-th example and y_i is the actual value. It is closely linked to the estimation of a distribution parameter (θ) through $MSE = E[(\hat{\theta} - \theta)^2\,] = \ Bias(\hat{\theta})^2 + Var(\hat{\theta})$.
- *Root Mean Squared Error (RMSE):* Derived by taking the square root of MSE, it recovers the original unit, facilitating model accuracy interpretation.
- *Mean Absolute Error (MAE):* This is also a very common metric given by $MAE = (\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|)$, it is less sensitive to outliers than MSE and preserves the original unit, commonly used in materials science (Dunn et al. 2020).
- *Mean Absolute Percentage Error (MAPE):* A normalized version of MAE expressed as a percentage; MAPE = $(100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|})$.
- *Coefficient of Determination (R²):* Measures the proportion of variance explained by the model, mathematically $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where the total sum of squares is $SS_{tot} = \sum_i (y_i - \bar{y})^2$ and the residual sum of squares is given by $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$. When $R^2\ = 1$ a perfect fit of the actual data is obtained, explaining all variance in the dependent variable. However, R² can mislead in cases of overfitting or a high feature-to-sample ratio (Schleder et al. 2019).

In the case of classification tasks, the confusion matrix is a very usual visualization method and gives insight on important quantities for the classification task. In this matrix, illustrated in *Figure 15*, each row and column corresponds to predicted and actual values, allowing a clear representation of outcomes. The resulting matrix contains four distinct cells: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP denotes instances where both actual and predicted values are positive, TN signifies instances where the actual value is positive while the model predicts a negative value, FP represents cases where the actual value is negative, but the model predicts a positive value, and FN indicates situations where both actual and predicted values are negative.

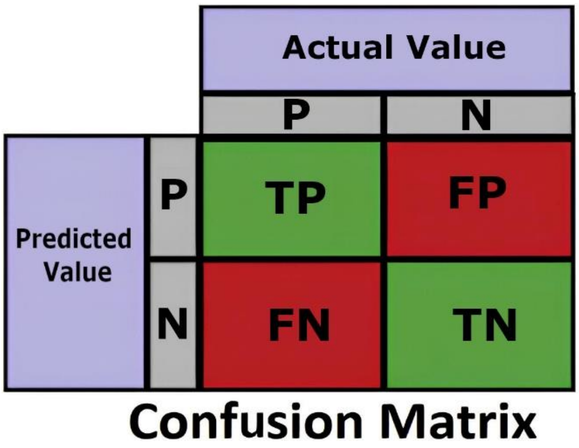


Figure 15 – The confusion matrix showcases the number of correctly predicted elements in the diagonal entries and incorrectly predicted ones in the off-diagonal entries. Adapted from: (Das, Sahoo, and Pradhan 2022).

From these quantities we can extract important evaluation metrics for classification tasks:

- *Accuracy:* The ratio of correct predictions to the total predictions, calculated as (TP + TN) / (TP + TN + FP + FN). It is valuable for balanced classes but can mislead with imbalanced classes.
- *Precision:* The ratio of true positives to the sum of true positives and false positives, calculated as TP / (TP + FP). Useful when the cost of false positives is high.
- *Recall:* The ratio of true positives to the sum of true positives and false negatives, calculated as TP / (TP + FN). Valuable when the cost of false negatives is high.
- *F1 score:* The harmonic mean of precision and recall, calculated as 2 * (precision * recall) / (precision + recall). Important when both precision and recall matter, this metric is more reliable for unbalanced classes.

The receiver operating curve (ROC) is also routinely used to understand model's ability to differentiate classes, being the plot of the true (T) positive rate TPR = $(\frac{\mathrm{TP}}{\mathrm{TP+FN}})$ versus the false positive rate FPR = $(\frac{FP}{FP+TN})$ with changing threshold (Schleder et al. 2019; Pedregosa et al. 2011). The ROC curve is illustrated on *Figure 16*, this leads to another metric the *Area under the ROC curve (AUCROC).* In the ideal case of AUC = 1, we have a perfect classifier that achieves ideal separability between classes, exhibiting no false positives or false negatives.

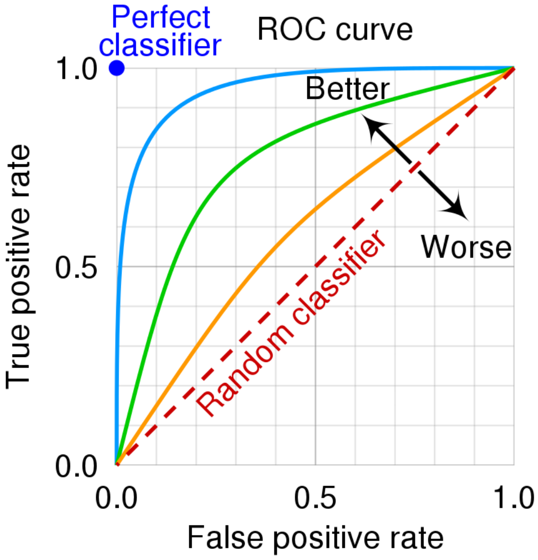


Figure 16 – The area under the ROC curve is an indicator of a model's ability to accurately classify data. Source: (‘Receiver Operating Characteristic’ 2023).

Cross-entropy is another versatile classification error metric well-suited for neural networks due to its differentiability. It can be either binary or categorical cross-entropy, depending on whether there are two classes or more, respectively. For binary cross-entropy (BCE), a sigmoid function converts logits to probabilities, given by:

$$BCE(i) = -y_i \log(p_i) - (1 - y_i)\log(1 - p_i)), \tag{39}$$

where y_i is the true label (0 or 1) and p_i is the predicted probability. In categorical cross-entropy (CCE), the Softmax function transforms logits (z_i) for each class into probabilities, defined as:

$$p_i = Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}},$$
$$CCE(i) = -\sum_{j=1}^{C} y_{ij}\log\left(p_{ij}\right), \tag{40}$$

where C is the number of classes, y_{ij} is the indicator function (1 if sample i belongs to class j), and p_{ij} is the predicted probability of sample i in class j.

In this work, MAE was the primary choice for regression tasks due to its ease of interpretation and popularity. However, for tasks like training autoencoders as addressed in *Chapter 5*, MSE was preferred because it places a heavier penalty on outliers. For classification, we opted for the AUCROC metric, implemented using the *Scikit-learn* package (Pedregosa et al. 2011). We also applied cross-entropy as the loss function for optimization, primarily due to its compatibility with neural networks.

2.4.8 Neural networks and deep learning

ML methods like regression, random forests, and support vector machines have long been staples in materials science (Carr et al. 2009; Madden and Howley 2009; Podolyan, Walters, and Karypis 2010; Majid, Khan, and Choi 2011; Carrete et al. 2014). These models excel when datasets are small. However, to harness the vast repositories of materials information and the increasing output data in materials research, more advanced algorithms capable of capturing complex interactions within extensive chemical spaces become essential. Artificial neural networks (ANNs) and their more sophisticated progression, deep neural networks (DNNs), lead this current surge of ML frameworks in materials research (Choudhary et al. 2022). ANNs are pervasive and influential machine learning algorithms to model input-output correlations. Their architecture, depicted in *Figure 17*, begins with an input layer that houses the input data within its nodes. Subsequently, hidden layers, composed of multiple nodes, also called neurons, fully connect with nodes in successive layers. These interconnections are visually represented by connecting lines, culminating in the final output layer. This parallelism to the intricate connectivity of biological neural networks (Barrett, Morcos, and Macke 2019) empowers ANNs to decipher nuanced patterns and correlations necessary for the complex materials science problems. In fact, neural networks can potentially estimate any function to arbitrary accuracy, according to the "universal approximation theorem" (Hornik, Stinchcombe, and White 1990).

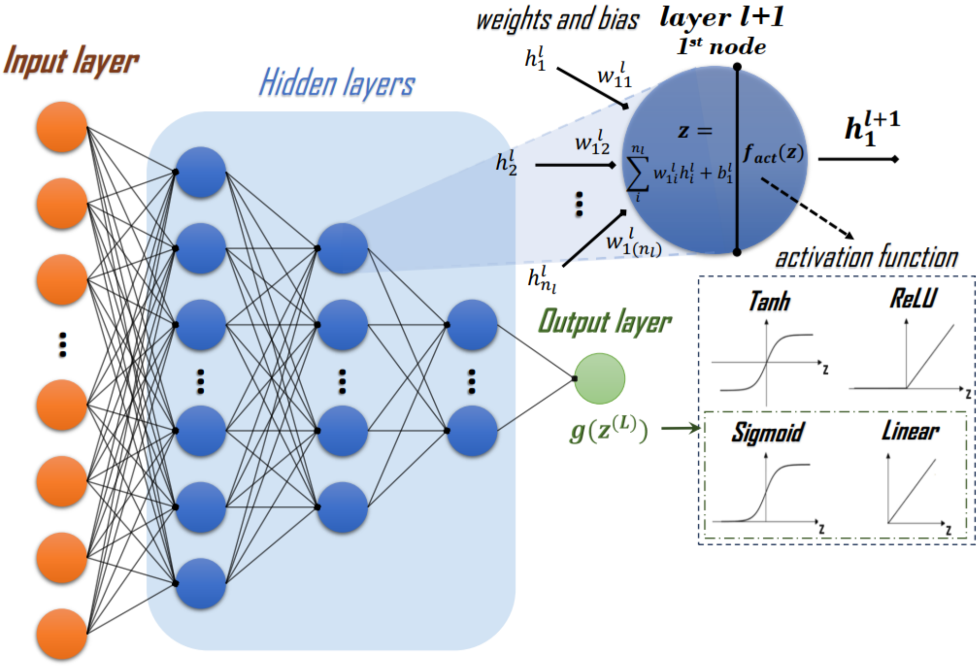


Figure 17 – Diagram of a neural network that shows how inputs are processed through the network's layers to produce an output. The diagram also illustrates how activation functions like Tanh, ReLU, Sigmoid, and Linear are applied in the hidden layers and final output. Source: author.

A *Deep Neural Network* (DNN) is formed when an ANN includes multiple hidden layers. The most common type of DNN is the *Feedforward Neural Network* or *Multilayer Perceptron* (MLP), where information flows unidirectionally from input to output. Discussion on other models is left on *Appendix A.14*. In a MLP, each layer receives an output denoted as $\boldsymbol{h}^l = [h_1^l,\ h_2^l,\ \ldots,\ h_{n_l}^l]$, corresponding to the n_l nodes in that layer. This output multiplies a weight matrix that contains entries for every combination of neurons between the current and next layers, forming an $n_{l+1} \times n_l$ matrix explicitly defined as:

$$\boldsymbol{W}^{(l)} = \begin{pmatrix} w_{11}^l & w_{21}^l & \cdots & w_{(\mathrm{n}_l)1}^l \\ w_{12}^l & w_{22}^l & \cdots & w_{(\mathrm{n}_l)2}^l \\ \vdots & \vdots & \ddots & \vdots \\ w_{1(\mathrm{n}_{l+1})}^l & w_{2(\mathrm{n}_{l+1})}^l & \cdots & w_{(\mathrm{n}_l)(\mathrm{n}_{l+1})}^l \end{pmatrix}, \tag{41}$$

Each node in the layer contains a bias, forming $\boldsymbol{b}^l = [b_1^l,\ b_2^l,\ \ldots,\ b_{\mathrm{n}_l}^l]$, which is added to the multiplication results. This addition generates the subsequent $\boldsymbol{h}^{l+1}$ following the formula:

$$\boldsymbol{h}^{(l+1)} = f_{act}(\underbrace{\boldsymbol{h}^{(l)} \cdot \boldsymbol{W}^{(l)} + \boldsymbol{b}^{(l)}}_{\boldsymbol{z}^{(l)}}). \tag{42}$$

Here, f_{act} represents the activation function applied to the output of the previous step, $\boldsymbol{z}^{(l)}$. Activation functions play a crucial role by introducing non-linearity into neuron outputs. Without them, the network would be constrained to linear regression models, incapable of capturing the intricate non-linear relationships present in data. *ReLU*, *Tanh* and *Sigmoid* functions, illustrated in *Figure 17*, are commonly used in hidden layers to capture these complex relationships (Dubey, Singh, and Chaudhuri 2021).

The final layer in the network is the output layer, representing the modeled quantity. The final output (**y**) of an ANN with L layers is expressed as:

$$\boldsymbol{y} = g(\boldsymbol{h}^{(L)} \cdot \boldsymbol{W}^{(L)} + \boldsymbol{b}^{(L)}), \tag{43}$$

where $\boldsymbol{h}^{(L)}$ is the output of the last hidden layer with n_L nodes. This output multiplies the weight matrix $\boldsymbol{W}^{(L)}$ of the final layer. For output layers, the activation function g usually employs the sigmoid activation for classification tasks due to its bounded nature, and linear activation for regression tasks. At this point, the final result is compared to the actual values in the supervised learning approach, utilizing the predefined loss function for the problem. Subsequently, the gradients of this loss function drive iterative updates to the network's weights and biases through a process known as backpropagation.

Backpropagation is a crucial algorithm in training neural networks, operating in two main phases. First, during the forward pass, input data travels through the network, generating predictions. Second, in the backward pass (backpropagation), it computes the gradients of the loss function with respect to the network's weights and biases, propagating errors backward through the layers. These gradients guide the adjustment of weights to minimize the difference between predicted and actual outputs, refining the network's performance in each iteration (epoch). The efficacy of this process is significantly influenced by the choice of activation function, as it shapes the model's learning behavior and adaptability. For instance, *ReLU* helps mitigate the problem of vanishing gradients, while *Tanh* aids when the optimization problem benefits of bounded outputs and symmetry (Goodfellow, Bengio, and Courville 2016; Murphy 2012). Additionally, hyperparameters like learning rate, batch size, batch normalization, and regularization techniques play pivotal roles in shaping the neural network's training (a comprehensive explanation of these hyperparameters is provided in *Appendix A.13*). Advances in the backpropagation algorithm (Hinton and

Salakhutdinov 2006; Baydin et al. 2018) and refinement of the hyperparameters gave rise to the popularity of Deep Neural Networks (DNNs) that are widely used today (Awad and Khanna 2015).

Traditionally, these networks operate with structured data presented as a feature vector. Ideally, this vector contains entries that significantly correlate with the desired output function. While deep learning models have shown exceptional success in handling speech, images, or time-series data—where an inherent linear structure can be harnessed—managing unstructured data, such as atomic structures, demands different tools. Properties arising from atomic structures do not merely result from the spatial arrangement of atoms in Euclidean space; they also encompass the nature of their bonds, functional groups, and overall connectivity (Bronstein et al. 2017; Choudhary et al. 2022). As a result, using descriptors to indirectly capture these interactions for input into neural networks, as described in section 2.4.4, has inherent limitations in capturing these properties.

Fortunately, deep learning techniques offer an alternative approach by representing atomic structures using graphs. These graphs, denoted as $\mathcal{G}$=($\mathcal{V}$,$\mathcal{E}$,$\mathcal{U}$), consist of nodes or vertices ($\mathcal{V}$) holding atomic element information, edges ($\mathcal{E}$) store bond attributes and capture structural connectivity through adjacency lists, and a global attribute vector $\mathcal{U}$ which serves as a master node or context vector, bridging information transmission among all nodes and edges (Sanchez-Lengeling et al. 2021). An illustration of a graph for a material fragment is presented below in *Figure 18*. These graphs are subsequently fed into graph neural networks (GNNs), specialized deep neural networks tailored for graphs. GNNs excel at capturing underlying connectivity patterns within atomic structures, typically through message passing.

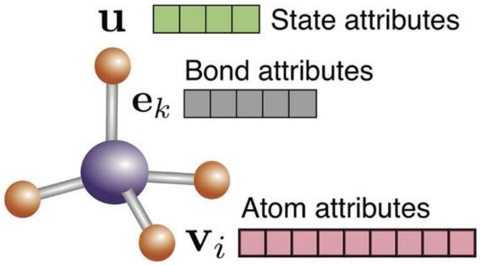


Figure 18 – Schematic of how material information is usually encoded in a graph. Adapted from: (C. Chen et al. 2019).

Message passing involves two key steps: aggregating information from neighboring nodes and subsequently updating the state of the receiving nodes and edges based on this aggregated data. This process, illustrated in *Figure 19*, applies pooling to aggregate information from neighboring nodes while also transferring information along the edges. Importantly, aggregation remains permutation invariant, enabling the GNN to operate consistently regardless of the order in which neighboring nodes are embedded. This ensures that the network learns from the graph structure rather than node sequence or arrangement. MLPs are then employed to transform these vectors and update the graph information. This form of GNN utilizing aggregation is termed a graph convolutional network (GCN), which stands as the most widely applied form, although other variations exist based on the specific transformations during message passing.

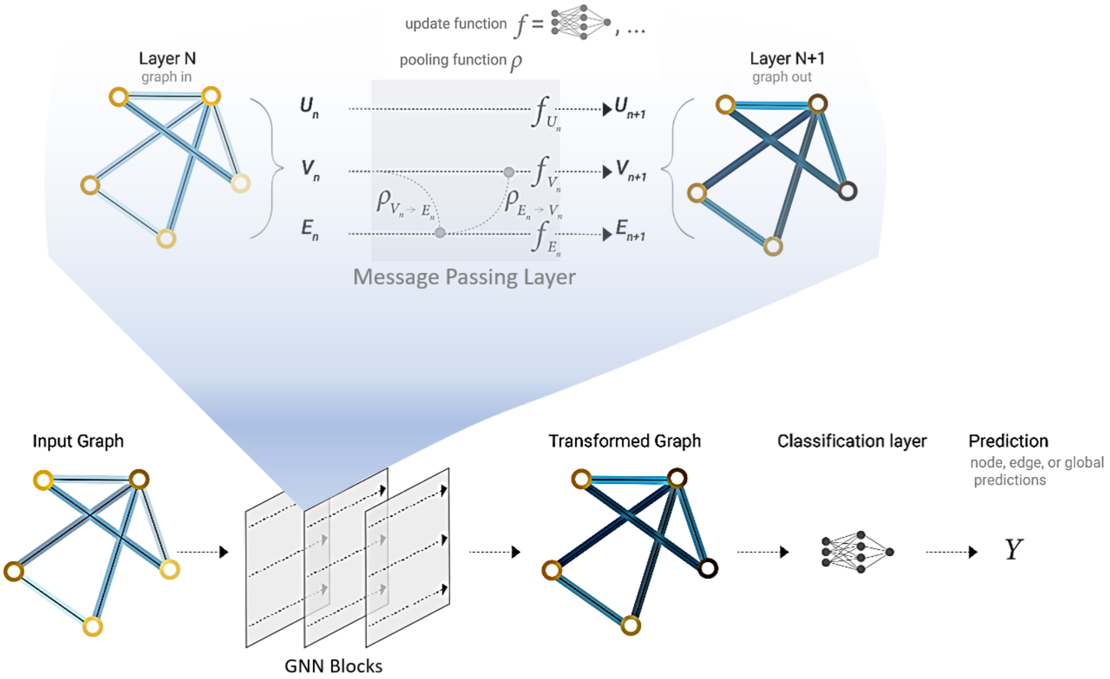


Figure 19 – Diagram of a GCN architecture that pools neighboring nodes within a degree's distance to update node representations of a graph. Final graph is transformed in predictions via MLP. Adapted from: (Sanchez-Lengeling et al. 2021)

Several message passing layers can be stacked to propagate information more extensively through the nodes. There are also additional pooling layers that facilitate the transfer of information between nodes and edges to the master node. GNNs are highly adaptable, allowing it to effectively capture structural connectivity within graphs. Moreover, another advantage of the GNN lies in its seamless ability to make predictions for individual nodes and edges. This capability holds significant potential in materials science, enabling predictions not only for global properties but also for local chemical information within the structure. However, this flexibility comes with the need for an extensive dataset to train all their internal parameters, which is not the usual case in materials science.

An insightful summary of deep learning's integration into the artificial intelligence (AI) ecosystem is given in *Figure 20*, which also shows the variety of materials science data sources and potential neural network architectures. The implementation of many of these new technologies was possible by the development of libraries like *Tensorflow* (Abadi et al. 2016), *PyTorch* (Paszke et al. 2019) and MXNet (T. Chen et al. 2015), these libraries provide a high-level interface for building and training DNNs. *Tensorflow*, Google's powerful open-source framework, was our tool of choice to construct, train and deploy DNN models in this work. *Tensorflow* is integrated to the high-level application programming interface (API) *Keras* (Chollet 2015), which simplifies neural network development with its user-friendly interface and rapid prototyping features.

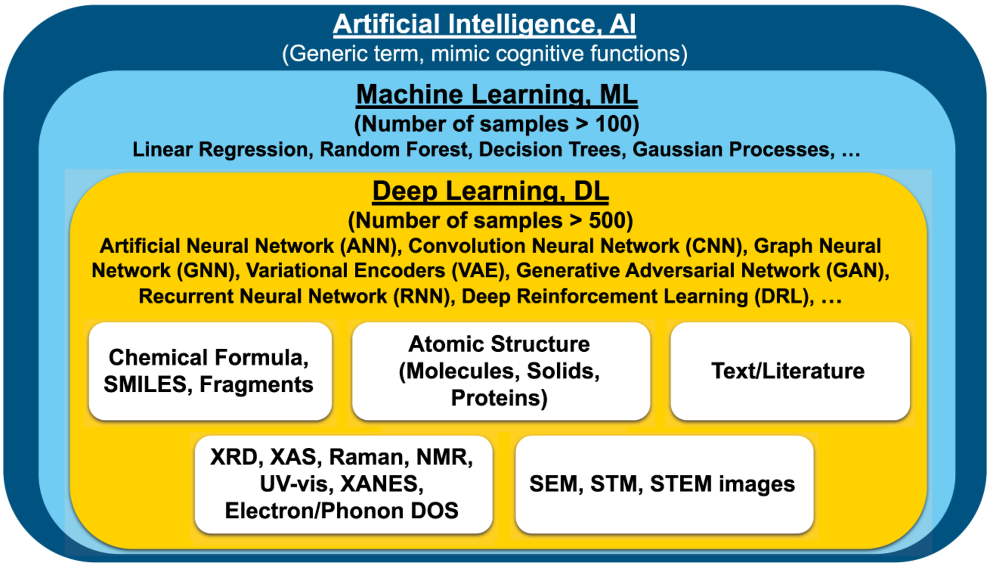

Artificial Intelligence, AI
(Generic term, mimic cognitive functions)
Machine Learning, ML
(Number of samples > 100)
Linear Regression, Random Forest, Decision Trees, Gaussian Processes, …
Deep Learning, DL
(Number of samples > 500)
Artificial Neural Network (ANN), Convolution Neural Network (CNN), Graph Neural Network (GNN), Variational Encoders (VAE), Generative Adversarial Network (GAN), Recurrent Neural Network (RNN), Deep Reinforcement Learning (DRL), …
Chemical Formula, SMILES, Fragments
Atomic Structure (Molecules, Solids, Proteins)
Text/Literature
XRD, XAS, Raman, NMR, UV-vis, XANES, Electron/Phonon DOS
SEM, STM, STEM images

Figure 20 – Schematic overview of deep learning (DL) methods within the field of artificial intelligence. Instances of DL application on materials science are shown. Source: (Choudhary et al. 2022).

2.4.9 Deep learning solutions in materials science

In the field of materials science, DNNs have been used to predict the properties of materials, including their optoelectronic, magnetic, and thermo-mechanical properties (Choudhary et al. 2022). To cite a few recent examples, DNNs were used to predict the electronic density of states for materials classes of arbitrary compositional and structural diversity (Fung, Ganesh, and Sumpter 2022), as well as their thermal transport (Qian and Yang 2021), and even phonon structure (Gurunathan, Choudhary, and Tavazza 2023), all with state-of-the-art results.

When using material structural and chemical information to train deep learning models, two primary approaches emerge. The first employs feature-based models which prioritize meaningful descriptors and efficient selection to enhance prediction accuracy. While proficient at correlating chemical data with target properties, these models might struggle to capture complex relationships between structural features and properties. Nevertheless, with appropriate descriptors, they exhibit notable accuracy, even with limited datasets (D. Jiang et al. 2021).

The second approach involves graph-based models, particularly Graph Neural Networks (GNNs), which leverage structural information represented as graphs. These models excel in capturing complex relationships, achieving state-of-the-art performance. However, their effectiveness often hinges on large datasets to reach their full potential (De Breuck, Hautier, and Rignanese 2021; Shunning Li et al. 2022). Despite this requirement, GNNs currently stand as the prevailing and highly accurate AI method for predicting various material properties based on structural information (Choudhary et al. 2022). This method differs in being less reliant on highly engineered descriptors but sacrifices interpretability due to the multitude of parameters and message passing blocks.

In the upcoming subsections, we will delve into some of the most relevant ML frameworks, both graph-based and feature-based, and explore their contributions as utilized in this thesis.

2.4.9.1 MEGNet

MatErials Graph Network (MEGNet) is a machine learning framework based on the GCN architecture aimed at predicting molecular and crystal properties (C. Chen et al. 2019). It relies as input the atomic numbers, coordinates, and cell information in the case of crystals. Through its graph convolution layers, MEGNet models grasp the essence of atoms, bonds, and structures by learning embeddings. These models stand out for their exceptional performance in various properties such as formation energy, band gap, and elastic modulus in both molecular and crystal domains.

One significant breakthrough of MEGNet models lies in their pioneering approach to efficiently predict multiple targets, especially when these targets share a physical relationship, like in thermodynamic potentials. This achievement is realized by integrating suitable global state attributes. Moreover, MEGNet models demonstrate that the learned element embeddings, representing the unique chemical characteristics of each element, encapsulate periodic chemical trends. These embeddings can be transferred from a property model trained on a larger dataset, such as formation energies, to enhance property models with limited data, such as band gaps and elastic moduli.

MEGNet has shown remarkable effectiveness in predicting general materials properties, achieving state-of-the-art accuracy in formation energies. However, a significant limitation of MEGNet is its reliance on precise atomic positions for accurate results. Traditional GNNs like MEGNet lack physical constraints to maintain energy continuity with variations in atomic positions. This limitation hinders the computation of forces and stresses necessary for proper geometry optimization. As a result, the atomic configuration can only be derived via DFT structural relaxations or experiments, contradicting the objectives of materials discovery, where the pursuit of an equilibrium geometry is often the primary focus rather than the starting point (Choudhary et al. 2022).

2.4.9.2 M3GNet and CHGNet

The three-body interactions neural network (M3GNet) differs from traditional GNNs by explicitly incorporating two- and three-body interactions into its framework (C. Chen and Ong 2022). This unique feature enables the model to be trained on and generate forces and stresses, essentially functioning as a universal machine-learning interatomic potential (MLIP) describing the structure's potential energy surface. Consequently, M3GNet can seamlessly obtain the equilibrium configuration from an initial structure.

This accomplishment relies on employing a graph $\mathcal{G}$=($\mathcal{V}$,$\mathcal{E}$,$\mathcal{X}$,$\mathcal{M}$,$\mathcal{U}$), in which besides node ($\mathcal{V}$), bond ($\mathcal{E}$) and global attributes ($\mathcal{U}$), the coordinates for each atom are passed on $\mathcal{X}$, and also the 3×3 lattice matrix is passed on $\mathcal{M}$, essential for obtaining tensorial quantities such as forces and stresses. The model architecture resembles the traditional GCN, except for a bond update function. Considering a generic bond $e_{ij} \in \mathcal{E}$, the update function will consider all atoms k in the neighborhood of atom i, using their attributes $v_k \in \mathcal{V}$, every distance r_{ij} and three-body angles θ_{ijk} to generate the updated e'_{ij}. Subsequently, this updated bond follows typical graph convolution with $\mathcal{V}$ and $\mathcal{U}$ attributes.

This process is sequential and repeats for the specified number of GNN blocks defined in the architecture, culminating in final values for the vertex attributes. These attributes undergo processing via a gated multilayer perceptron to produce individual atomic energies, aggregated to the final energy of the structure. Using auto-differentiation (Bücker et al. 2006), forces and stresses are derived as $\mathbf{f} = -\partial E/\partial \mathbf{x}$ and $\boldsymbol{\sigma} = V^{-1}\partial \text{E}/\partial \boldsymbol{\epsilon}$, where **x** are the atomic coordinates, V represents the volume, and $\boldsymbol{\epsilon}$ denotes the strain.

M3GNet training utilized the large dataset of structural relaxations by Materials project comprising of more than 187,000 energies, 16,000,000 forces and 1,600,000 stresses. M3GNet demonstrates notably superior accuracy and consistency in its formation energies compared to MEGNet. This is evident from the substantial performance leap of M3GNet over MEGNet in the Matbench "Materials Discovery" task (Riebesell et al. 2023), reflected in the MAE on the convex hull distance: 0.07 eV/atom for M3GNet versus 0.13 eV/atom for MEGNet.

More recently, a new model was developed for MLIP named CHGNet (B. Deng et al. 2023), standing for Crystal Hamiltonian Graph Neural Network, which builds upon the original architecture of M3GNet and is trained in an even more extensive dataset. CHGNet incorporates magmoms (the initial magnetic moment of individual atoms) as a proxy for inferring the atomic charge in atomistic simulations, thereby substantially improving the regularization of the MLIP. CHGNet showcases a MAE of 0.06 eV/atom on the Matbench discovery task. It distinguishes itself from competing models by reaching an R² value of 0.69, while the determination coefficient of M3GNet remains at 0.58, marking its superior performance.

CHGNet is recognized for its high accuracy and robustness across various chemical contexts, qualifying it for practical application in high-throughput materials discovery, as evaluated by Matbench (Riebesell 2024). By enabling the rapid production of equilibrium structures, it accelerates the optimization and evaluation processes for DFT calculations. CHGNet also demonstrates promising accuracy in estimating the ab-initio demanding phonon band structures and phonon density of states for dynamical stability. However, it is important to analyze results on a case-by-case basis for this application. Alternatively, optimized structures can also be used to assess properties using traditional graph-based models or to generate structural descriptors for the subsequent prediction of properties in feature-based models.

2.4.9.3 MODNet

The Material Optimal Descriptor Network (MODNet) is a feature-based machine learning framework designed to predict materials properties from composition or atomic structure (De Breuck, Hautier, and Rignanese 2021). It utilizes a feedforward neural network fed with a limited number of descriptors derived from chemical, physical, and geometrical considerations, typically a subset of MatMiner descriptors. MODNet's design aims to maximize data efficiency, especially for tasks where obtaining large training sets is challenging or costly, as commonly seen in experimental datasets or computationally demanding ab-initio properties.

MODNet consistently outperforms MEGNet, especially in scenarios with a small number of training samples, typically below ~4,000 samples, even when leveraging transfer learning from larger models. MODNet’s performance also outshines random forest (RF) algorithms and neural network models lacking built-in feature selection. In Table 3, the scaled MAE for different tasks in Matbench is presented for MODNet, MEGNet and RF.

Table 3 – Scaled errors for MODNet, MEGNet and a random forest framework (RF-SCM/MagPie) in different MatBench tasks with progressively larger dataset size. Data highlighted in green denotes the best algorithm in the corresponding task. Source: (Dunn 2024)

Task*	**Dataset size**	**General Purpose Algorithm / Scaled Error****		
		MODNet	MEGNet	RF-SCM/MagPie***
E_x 2D materials (regression)	636	**0.4939**	0.8061	0.7476
Refractive index (regression)	4,764	**0.3353**	0.4193	0.5189
$Log_{10}K^{VRH}$ (regression)	10,987	**0.1890**	0.2306	0.2830
E_f perovskites (regression)	18,928	0.1603	**0.0621**	0.4160
E_g DFT (regression)	106,113	0.1657	**0.1457**	0.2127
Metallicity DFT (classification)	106,113	0.1924	0.1957	**0.1814**
E_f DFT (regression)	132,752	0.044	**0.025**	0.1157

**Specifics on each task can be found on matbench.materialsproject.org*

***Regression tasks used scaled MAE, defined as the ratio of mean absolute error to mean absolute deviation, and classification used (1-AUCROC)/0.5 as error metrics.*

****This algorithm employs Sine Coulomb Matrix and MagPie descriptors within a random forest framework.*

MODNet's superior performance in smaller datasets stems from its sophisticated feature-selection process (explained in section 2.4.4) and a comprehensive hyperparameter optimization facilitated through a genetic algorithm (De Breuck, Heymans, and Rignanese 2022). Additional points to highlight are MODNet's flexible architecture enabling joint-learning which enhances prediction in related properties and the efficiency of the selection algorithm in grasping the underlying physics driving the predictions.

2.4.9 Machine learning for materials discovery

Material science can be viewed as a combinatorial puzzle of mixing and arranging atoms to create new sets of properties (Riebesell et al. 2023). Davies et al. (2016) identified astounding 10^{10} possible quaternary materials by electronegativity and charge-balancing rules with even more unexplored quinternary and higher combinations, representing a vast realm of untapped potential in materials discovery. Uncovering new materials propels technological advancements, particularly in optimizing optoelectronic devices, the main subject herein discussed. Determining stability and properties via DFT is accurate but too costly for such a broad exploration.

ML then emerges as the prime tool for navigating this expansive chemical space efficiently as illustrated in *Figure 21*.

High-throughput ML screening in materials science follows two main steps: creating diverse candidate structures and employing ML frameworks to filter for stability and other crucial properties. Given the vast search space, integrating active learning techniques, in which the ML model selects the most informative data points to learn from, is practically mandatory (*Appendix A.15* elaborates on active learning). Very recently, Google's Graph Networks for Materials Exploration (GNoME) utilized large-scale active learning and identified 2.2 million new structures, increasing the catalog of thermodynamically stable materials by about an order of magnitude (Merchant et al. 2023). Following similar principles, we conducted a ML high-throughput screening employing CHGNet and MODNet in an active learning cycle, specifically targeting doping in 2D layered perovskites with the prototype formula $A_3B_2X_9$, as discussed in *Chapter 6*.

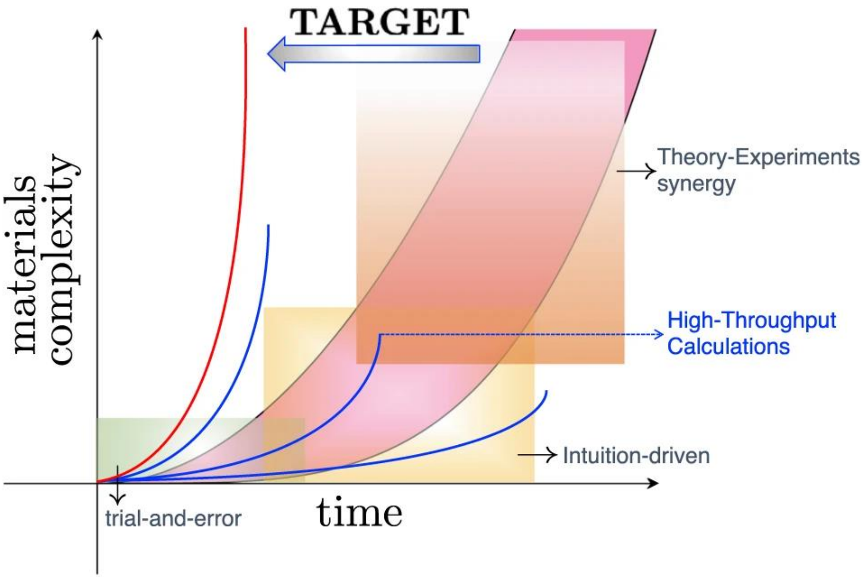


Figure 21 – Illustration of capacity of different methods to discover materials with increasing complexity as a function of time, transitioning from trial and error to high-throughput calculations and machine learning aided by statistical methods (TARGET). Source: (Lookman et al. 2019).

ML models excel in predicting DFT formation energies but struggle with decomposition enthalpy, a critical factor in the discovery of stable materials (Bartel 2022). This underscores the importance of proper choice and optimization of the ML models used for materials discovery. Both feature-based and graph-based models strive to approximate a universal materials featurizer, ideally capturing all physical properties within descriptors to correlate with any target property of interest. Graph-based models, while proficient in learning atom and bond embeddings, lack transferability and interpretability. In contrast, feature-based models prioritize interpretability and, with a thoughtful choice of descriptors, perform well even with limited data. However, they underperform compared to GNN models with extensive datasets. In *Chapter 5*, we pinpoint areas for improvement in feature-based models, delving into the integration of advanced electronic structure descriptors and GNN models features in a suitable framework for materials discovery.

CHAPTER 3 — Lead-Free $Cs_3Sb_2X_9$ (X = Cl, Br, I) Perovskites: Halide Alloying, Surfaces, Interfaces, and Clusters

doi.org/10.1007/s10853-023-09228-2

3.1 RESEARCH PROBLEM

Halide alloying is a promising strategy to achieve significant improvements in the power conversion efficiency, stability, and color tunability of lead-free perovskite solar cells and LEDs (A. Wang et al. 2023; Wei et al. 2023). This approach has already been applied to tune the band gap tuning and improve electronic properties of $Cs_3Sb_2X_9$ compounds. For example, Br/I mixed halide perovskite $Cs_3Sb_2Br_{9-n}I_n$ demonstrated tunable optical band gaps and enhanced photocatalytic efficiency for CO_2 photoreduction (Malavasi et al. 2023; D. Wu et al. 2022). Chlorine alloying in $Cs_3Sb_2I_9$ stabilized its layered phase (F. Jiang et al. 2018) also increasing solar cell power conversion efficiency (Paul, Pal, and Larson 2020; Jihong Li et al. 2022). Cl/Br alloyed $Cs_3Sb_2Cl_{9-n}Br_n$ exhibited a band gap transition from indirect to direct with Br substitution, as studied by Pradhan et al. (Pradhan, Jena, and Samal 2022). Despite some theoretical insights into halogen alloying effects in $Cs_3Sb_2X_9$ perovskites (F. Jiang et al. 2018; Pradhan, Jena, and Samal 2022), a systematic comparison of single halogen substitutions in all compounds of this group is still lacking.

Moreover, surface studies of $Cs_3Sb_2X_9$ perovskites are scarce, with investigations mainly focused on $Cs_3Sb_2Br_9$ surfaces such as (0001), (1000) and $(20\bar{2}1)$ (C. Lu et al. 2020; P. Liu et al. 2020). Chlorine-doped $Cs_3Sb_2I_9$ (0001) and $(20\bar{2}1)$ planes were studied only to understand thiourea adsorption (Pradhan, Jena, and Samal 2022). Thus, a detailed investigation of surfaces for each $Cs_3Sb_2X_9$ halide perovskite is necessary. Additionally, heterostructures and their interfaces play a crucial role in optoelectronic applications, but to our knowledge only $Cs_3Sb_2X_9/Cs_3Bi_2X_9$ has been recently investigated theoretically (Long, Zhang, and Cheng 2022). No study has yet addressed interfaces of different halogens, despite the material's success in optoelectronics and extensive heterostructure investigations in lead-based counterparts (Xiaoming Li et al. 2016; G. Zhang et al. 2020). Furthermore, DFT simulations in confined structures of the lead-free perovskite could clarify, for instance, the factors contributing to a significantly higher photoluminescence quantum yield (PLQY) in $Cs_3Sb_2Br_9$ perovskite compared to chlorine and iodine analogues (Ma et al. 2019).

In order to systematically address the knowledge gaps identified in our research, the results of this study are presented in four sections: Halide alloying, Surface properties, Band alignment and interfaces, and Perovskite clusters. The first subsection, 'Halide alloying' will concentrate on a thorough examination of the geometry, electronic structure, and energetics of the $Cs_3Sb_2X_{9-n}Y_n$ systems (X, Y = Cl, Br, I). This analysis aims to shed light on the impact of halide alloying on the properties of the perovskite structure. The subsequent subsection, 'Surface properties,' will delve into the characteristics of low index surfaces, specifically (0001) and (1000), for each halide perovskite. Understanding how these surfaces interact and influence the perovskite's properties is essential for applications in various fields. Moving forward, the third subsection, 'Band alignment and interfaces' will explore the band alignment and interface properties of two key interfaces: $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ and $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$. This investigation will offer insights into the behavior of interfaces and their role in device performance. Lastly, the 'Perovskite clusters' subsection will involve calculations on perovskite clusters, simulating confined systems that resemble the conditions found in nanocrystals. Understanding the behavior of these clusters is critical for applications in nanotechnology and materials science.

3.2 METHODOLOGY

Crystal structure for cesium antimony halide perovskites with formula $Cs_3Sb_2X_9$ (X = Cl, Br, I) were based on the description of Arakcheeva et al.(Arakcheeva et al. 1999) deriving from the traditional ABX_3-type perovskites with two-thirds of occupancies of B site. The trigonal phase (space group $P\bar{3}m1$, no. 164), exists for each halogen variant of $Cs_3Sb_2X_9$, and their corresponding crystallographic data was obtained from literature (Arakcheeva et al. 1999; Kun et al. 1993; Kihara and Sudo 1974). This is the base structure for all our simulations and is also most studied phase due to its layered 2D form enabling better transport properties (Y. L. Liu et al. 2019).

First principles plane-wave density functional theory (DFT) calculations were performed in the QUANTUM-ESPRESSO code package (Giannozzi et al. 2009). Exchange-correlation effects were characterized using the generalized gradient approximation (GGA) Perdew-Burke-Ernzerhof (PBE) functional (J P Perdew, Burke, and Ernzerhof 1996). Norm-conserving, full-relativistic pseudopotentials from the PseudoDojo project (van Setten et al. 2018) were used to represent all elements. The plane-wave cutoff energy for the calculations was set at 1200 eV. To obtain the electronic properties and optimize the structures, Brillouin zone integrations were conducted using a Monkhorst-Pack grid (Monkhorst and Pack 1976) of 5 × 5 × 3 k-points for self-consistent calculation of bulk systems, 5 × 5 × 1 for slabs and interfaces and only Γ point sampling for clusters. The BFGS quasi-newton algorithm (Billeter, Curioni, and Andreoni 2003) was employed for ion and cell parameter relaxation, with convergence thresholds for energy and forces set to 10^{-6} eV/atom and 10^{-5} eV/Å, respectively, for slabs and clusters the corresponding values are increased to 10^{-5} eV/atom and 10^{-4} eV/Å. Density of states, band structures and Bader charges are then obtained in the optimized structures.

In addition, hybrid functional calculations using the HSE06 functional (Heyd, Scuseria, and Ernzerhof 2003) were performed for the bulk structures. To attain similar precision in supercell calculations with viable computational cost, a correction model based on Hubbard correction +U was implemented to reproduce the main features of HSE calculation, particularly the band gap (detailed procedure in *Supporting Information, Appendix B.2*), U values determined for the pristine perovskites were then transferred to other structures. The structures examined after the bulk calculations, along with their corresponding calculation details, are outlined in the following:

- Halide alloying: to investigate the effect of alloying on band gap and cell parameters, all $Cs_3Sb_2X_{9-n}Y_n$ (X,Y = Cl, Br, I) structures with integer values of n ranging from 0 to 9 were calculated. The selection of substitution sites is discussed in detail in the *Supporting Information (Appendix B.3)*. Furthermore, thermodynamic properties such as interaction parameters and miscibility gap temperature were evaluated.
- Surfaces: low-index surfaces (0001) and (1000) were selected to study the surface energetics of the $Cs_3Sb_2X_9$ system. For each halogen, calculations were performed considering Cs-X termination, as AX-termination has consistently been found to be the most stable in halide perovskites (Y. Yang et al. 2018; Nazari, Azar, and Doroudi 2020; Di Liberto, Fatale, and Pacchioni 2021). A model comprising 16 atomic layers was utilized, and a vacuum

distance of 13 Å proved sufficient to eliminate electrostatic interactions between periodic replicas. Total energies were calculated for both relaxed and unrelaxed slabs.

- Band alignment and interface: an interface was created along the [0001] direction, which is the most commonly observed growth direction (Jihong Li et al. 2022), between $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ and $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ structures. This interface was computed to determine the average potential difference, establish band offsets between the structures, and calculate electronic properties. This method has been previously described (Weston et al. 2018).
- Perovskite clusters: For each halide perovskite, clusters were constructed using the NanoCrystal tool (Chatzigoulas et al. 2018). The smallest CsX-terminated structure was constructed and consisted of a non-stoichiometric $Cs_{13}Sb_6X_{30}$ cluster with 49 atoms. In these clusters, the Sb atoms were not exposed, and six complete SbX_6 octahedra were observed. The clusters exhibited both longitudinal (0001) and lateral (1000) surfaces in similar proportions. Surface passivation was not deemed necessary as the clusters displayed well-defined bands resembling bulk perovskites without localized midgap states. Cl doping on the iodine cluster was explored on both longitudinal faces and edges to assess the preferential doping site in the nanostructure.

3.4 RESULTS AND DISCUSSION

3.4.1 Bulk perovskites

Our study commenced by replicating the existing findings on $Cs_3Sb_2X_9$ (X = Cl, Br, I) perovskites, and simultaneously introducing the DFT+U approach. Band gap and lattice parameters for each of the perovskites were calculated using PBE exchange-correlation functional (tests with other functionals in *Supporting Information, Appendix B.1*) and are presented in *Table 4*. Errors between the calculated lattice parameters of $Cs_3Sb_2X_9$ and their experimental values (Kihara and Sudo 1974; Jian Zhang et al. 2017; Yamada et al. 1997) are below 3% and are also consistent with previous theoretical values (Saparov et al. 2015; Y. L. Liu et al. 2019) endorsing the present theoretical level and parameters for calculation. Band gap values are significatively smaller than experimental ones (Yamada et al. 1997; Blasse 1983; Jian Zhang et al. 2017), as expected for traditional DFT functionals, on the other hand, hybrid HSE functional yields a better approximation of this property due to higher theoretical level including electron exchange explicitly. To achieve similar accuracy for band gap, Hubbard +U values were determined for each perovskite following the methodology which is detailed on *Supporting Information (Appendix B.2)* attaining similar atomic charges and electronic structure to HSE calculation when applying in the corresponding halogens U_{Cl} = 4.5 eV, U_{Br} = 2.5 eV and U_{I} = 3 eV.

Table 4 – Band gap and lattice parameters with different functionals compared to experimental values for $Cs_3Sb_2X_9$ (X = Cl, Br, I). Deviations from experimental lattice constants for the different functionals are given in percentages under $\Delta_{a,b}$ and Δ_c.

	Band gap (eV)	Lattice constants (Å)		$\Delta_{a,b}$ (%)	Δ_c(%)
		a,b	c		
$Cs_3Sb_2Cl_9$					
Expt. (Blasse 1983)	3.09	7.633	9.345		
HSE	3.20	-	-		
HSE (Y. L. Liu et al. 2019)	3.11	-	-		
PBE	2.45	7.836	9.476	+2.662	+1.401
PBE+U (U_{Cl}=4.5eV)	3.08	7.841	9.532	+2.725	+2.001
PBE (Y. L. Liu et al. 2019)	2.44	7.827	9.472	+2.541	+1.359
$Cs_3Sb_2Br_9$					
Expt. (Jian Zhang et al. 2017)	2.36	7.930	9.716		
HSE	2.34	-	-		
HSE (Y. L. Liu et al. 2019)	2.60	-	-		
PBE	2.00	8.144	9.932	+2.698	+2.227
PBE+U (U_{Br}=2.5eV)	2.37	8.167	9.898	+2.988	+1.873
PBE (Y. L. Liu et al. 2019)	2.01	8.138	9.943	+2.623	+2.336
$Cs_3Sb_2I_9$					
Expt. (Yamada et al. 1997)	2.06	8.420	10.386		
HSE	2.10	-	-		
HSE (Y. L. Liu et al. 2019)	2.04	-	-		
HSE (Saparov et al. 2015)	2.06	-	-		
PBE	1.58	8.660	10.647	+2.853	+2.517
PBE+U (U_{I}=3.0 eV)	2.04	8.641	10.641	+2.624	+2.455
PBE (Saparov et al. 2015)	1.55	8.661	10.625	+2.862	+2.301
PBE (Y. L. Liu et al. 2019)	1.56	8.664	10.633	+2.898	+2.378

(*) All HSE calculations used PBE lattice parameters.

Electronic structure of the different perovskites calculated with Hubbard correction are shown in *Figure 22*, total and partial density of states (PDOS) reveal that both conduction (CB) and valence bands (VB) are composed of Sb 5p, Sb 5s and halogen outermost p orbitals. Halogen and Sb p orbitals overlap on most of the CB of these materials forming antibonding orbitals and are also prevalent on lower energy levels of VB. Although, highest energy levels on VB are composed of hybridized Sb 5s and halogen p orbital producing antibonding orbitals at valence band maximum (VBM) as well, thus the electronic transition from valence to conduction band minimum (CBM) occurs between antibonding orbitals which is associated with better photovoltaic properties (Brandt et al. 2015). The obtained results and band structures are in agreement with previous theoretical and experimental studies (Y. L. Liu et al. 2019; Saparov et al. 2015) which demonstrate a direct band gap in the bromine perovskite and an indirect, very close to direct, band gap in $Cs_3Sb_2I_9$ and $Cs_3Sb_2Cl_9$.

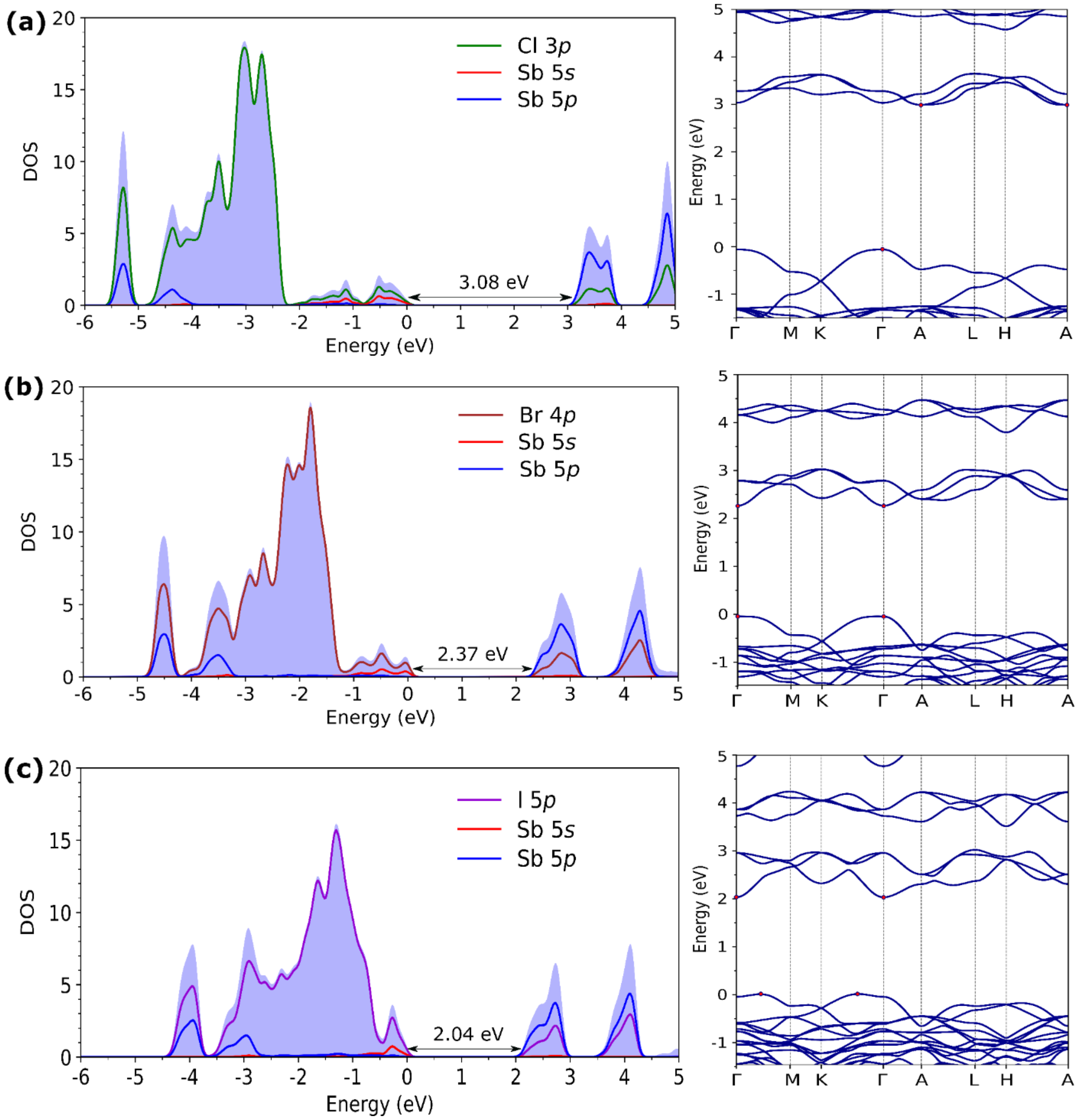


Figure 22 – Density of states and band structure of $Cs_3Sb_2X_9$ perovskites after Hubbard correction (a) $Cs_3Sb_2Cl_9$ perovskite with U_{Cl} = 4.5 eV, (b) $Cs_3Sb_2Br_9$ perovskite U_{Br} = 2.5 eV and (c) $Cs_3Sb_2I_9$ perovskite U_I = 3 eV.

3.4.2 Halide alloying

To assess how halide alloying influences geometry, stability and electronic properties of $Cs_3Sb_2X_9$, quantities such as lattice parameters, band gaps, binding energy and formation enthalpy were evaluated for the range of compositions of $Cs_3Sb_2X_{9-n}Y_n$ (X, Y = Cl, Br or I), from n = 0 to n = 9. To maintain conciseness, we provide a comprehensive methodology for calculating these parameters in the

Supporting Information (*Appendix B.3*) and proceed directly to the discussion of the results.

Figure 23 results for lattice constants and band gap are presented, lattice constants do not vary linearly with composition and a second-order Vegard's law was used for fitting the data and obtain bowing parameters. Values for fitting parameters are presented in *Table B4*, a and b lattice parameter bowing is larger than for c parameter as expected due to most metal-halogen bonds laying on the ab plane, the bowing parameter also increases progressively following the trend of anion radius difference, Cl^- (1.67 Å), Br^- (1.84 Å), and I^- (2.07 Å). For c lattice parameter, bowing is almost negligible for Cl-Br and Br-I alloys and becomes significant only for Cl-I alloy in which radius difference is larger. Regarding band gap of the alloys, a negative bowing in band gap is observed for $Cs_3Sb_2Cl_{9-n}I_n$ and a positive bowing is observed for $Cs_3Sb_2Br_{9-n}I_n$, therefore these two alloys present similar band gap in the heavy-halogen-rich region (n≈9) despite large difference in anion radius between them.

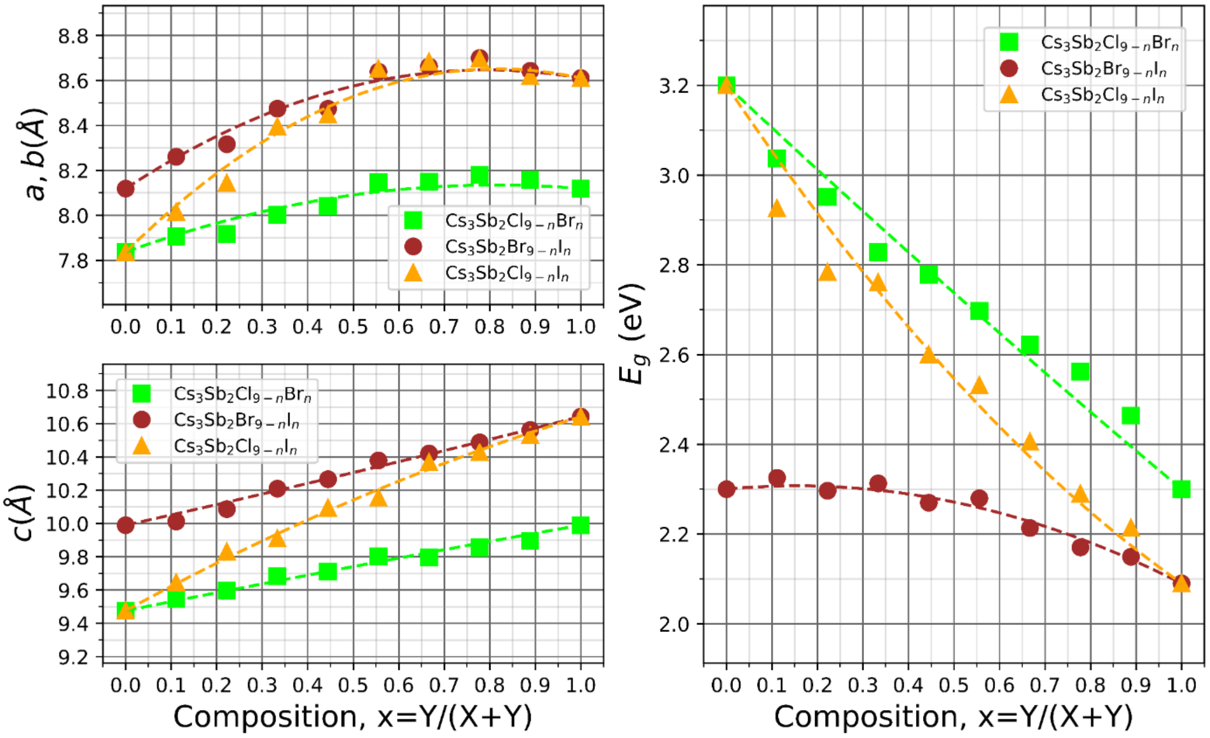


Figure 23 - Lattice parameters and band gap of $Cs_3Sb_2X_{9-n}Y_n$ solid solutions as a function of composition (x=Y/(X+Y)).

The calculated binding energies of $Cs_3Sb_2X_{9-n}Y_n$ (X,Y = Cl, Br or I) solid solutions are shown in *Figure 24* as function of relative halogen composition, *x* = Y/(X+Y), thus x=0 corresponds to pure $Cs_3Sb_2X_9$ and x=1 corresponds to pure $Cs_3Sb_2Y_9$. For all solid solutions, one can observe a linear relationship for binding energy as function of halogen composition, and the fitting line correlates with a coefficient over 0.99, as given in *Table B4*. For pure perovskites, $Cs_3Sb_2Cl_9$ has the higher binding energy (3.67 eV/atom) and a decrease is observed along with increasing the atomic number of the halogen yielding 3.02 eV/atom for $Cs_3Sb_2Br_9$ and 2.71 eV/atom for $Cs_3Sb_2I_9$. These results can be related to chlorine having the highest electronegativity followed by bromine and iodine, therefore larger charge transfer and more ionic character in the bonds of chlorine perovskite are responsible for its larger energy. The variation of binding energy with composition, resulting from halide alloying, can be fitted to a straight line which connects two pure phases. Considering the effect of increasing heavy halogen composition, $Cs_3Sb_2Br_{9-n}I_n$ presents the smallest binding energy variation (~0.30 eV/atom), followed by $Cs_3Sb_2Cl_{9-n}Br_n$ (~0.66 eV/atom) and $Cs_3Sb_2Cl_{9-n}I_n$ solid solutions which presents the largest variation (~0.96 eV/atom). Therefore, incorporating chlorine in the heavier iodine perovskites increases crystal stability, hence another factor for improved performance seen in $Cs_3Sb_2Cl_{9-n}I_n$ devices (Paul, Pal, and Larson 2020; Jihong Li et al. 2022; Peng et al. 2020).

Binding energy informs about bond strength in solid solutions, while the enthalpy of formation gauges the thermodynamic favorability of their creation from constituent phases. Formation enthalpy of alloys can be used to describe their miscibility from a thermodynamic principle of regular solution formation, formation enthalpy can be estimated from DFT total energies through the following formula:

$$\Delta H_f = E_{Cs_3Sb_2X_{9-n}Y_n} - (1-x)E_{Cs_3Sb_2X_9} - xE_{Cs_3Sb_2Y_9}, \quad (44)$$

where $E_{Cs_3Sb_2X_9}$ and $E_{Cs_3Sb_2Y_9}$ are the total energies of pure $Cs_3Sb_2X_9$ and $Cs_3Sb_2Y_9$ and $E_{Cs_3Sb_2X_{9-n}Y_n}$ is the total energy of the $Cs_3Sb_2X_{9-n}Y_n$ solid. The formation enthalpy $\Delta H_f(x)$ as given in Eq. (44) represents the energy cost of mixing X and Y halogens in the lattice. Changes in formation enthalpy as a function of composition is shown in *Figure 24*. For all $Cs_3Sb_2X_{9-n}Y_n$ solid solutions there is an upward bowing in their ΔH_f dependence on x, implying a preference for decoherent phase separation into $Cs_3Sb_2X_9$ and $Cs_3Sb_2Y_9$ at zero temperature. Comparing these $\Delta H_f(x)$ curves for every solid solution, the formation enthalpies for $Cs_3Sb_2X_{9-n}Y_n$ are in the order $Cs_3Sb_2Cl_{9-n}Br_n$ < $Cs_3Sb_2Br_{9-n}I_n$ < $Cs_3Sb_2Cl_{9-n}I_n$ for the same *x*. Therefore, halogen mixing is easier with halogens with similar ionic size. Larger formation enthalpy energies are concentrated on the heavy-halogen-rich side of the ΔHf(x) curve ($x > 0.5$), demonstrating that limited solubility might occur when a lighter halogen is added to heavier halide perovskites.

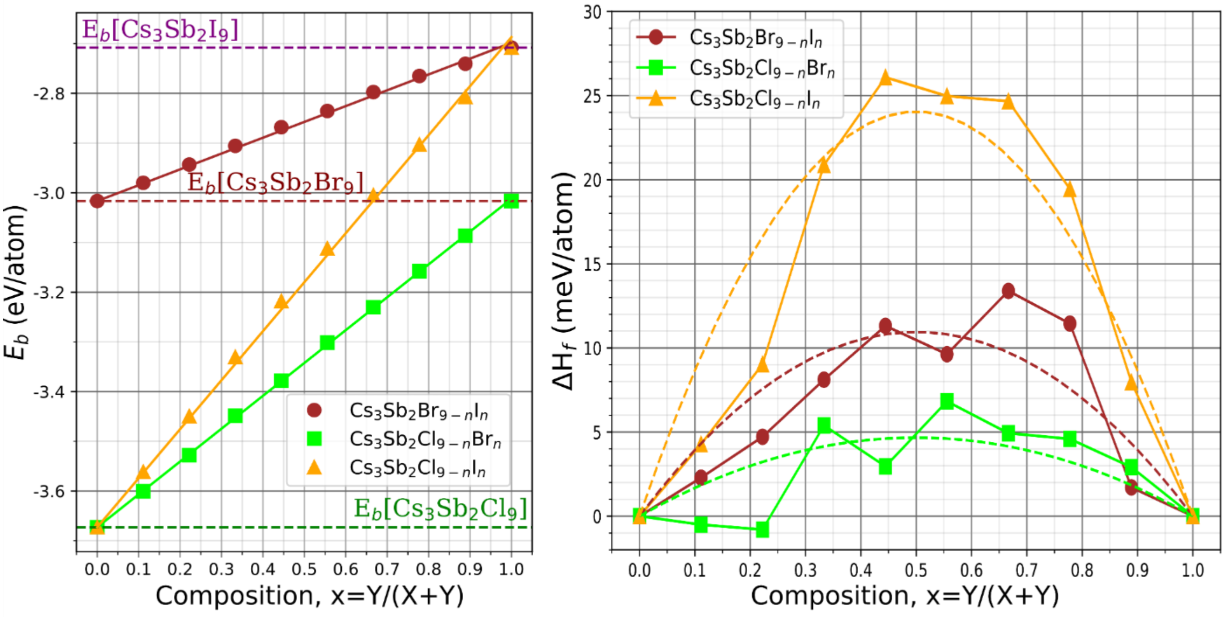


Figure 24 – Binding energy (left) and formation enthalpy (right) of mixed halide perovskites ($Cs_3Sb_2X_{9-n}Y_n$) solid solutions as function of composition, x (=Y/(X+Y)).

Conventional solid-solution theory states that formation enthalpy is a quadratic function of composition (x = Y/(X+Y)) for a binary alloy, hence the following relationship should apply:

$$\Delta H_f = \Omega x(1 - x), \tag{45}$$

where Ω is the interaction parameter, which is smaller for solutions with higher solubility. Temperature influences directly the ability of stable solutions to form and a critical temperature over which alloys are fully mixable, named miscibility gap temperature (T_{MG}), can be estimated from fitted interaction parameter Ω ensuing from regular solution model as $T_{MG} = \Omega/(2k_b)$ (Shu et al. 2013).

Solid solutions for $Cs_3Sb_2X_{9-n}Y_n$ present significant deviations from the regular solution especially due to asymmetry arising from difficulty of incorporating lighter halogens in heavy-halogen perovskites. An additional factor is the presence of peaks on x = 33.3% for $Cs_3Sb_2Cl_{9-n}Br_n$ and x = 66.6% for $Cs_3Sb_2Br_{9-n}I_n$, and negative ΔH_f for x < 33.3% in $Cs_3Sb_2Cl_{9-n}Br_n$, these observations might be a cue on the presence of ordered structures close to these compositions as reported in experiments with other halogen containing structures (Yin, Yan, and Wei 2014; Pramchu, Jaroenjittichai, and Laosiritaworn 2019; Zhao, Liu, and Dai 2016). Some reduction of formation energy and formation of ordered mixed halide compounds has been attributed to an overall Coulomb energy gain in the structure due to diminished repulsion between different halogens (Yin, Yan, and Wei 2014).

Estimated interaction parameters increase from $Cs_3Sb_2Cl_{9-n}Br_n$ ($\Omega \approx$ 19 meV/atom) and $Cs_3Sb_2Br_{9-n}I_n$ ($\Omega \approx$ 44 meV/atom) to $Cs_3Sb_2Cl_{9-n}I_n$ ($\Omega \approx$ 96 meV/atom), for Cl-Br and Br-I alloys interaction parameter can be considered small and yield a miscibility gap temperature (T_{mg}) of only 108 K and 254 K, respectively, suggesting that component-uniform $Cs_3Sb_2Cl_{9-n}Br_n$ and $Cs_3Sb_2Br_{9-n}I_n$ solid solutions can be prepared at the standard growth temperature below 450 K (Singh et al. 2018; Saparov et al. 2015), in fact substitution must be easy since regular solution model tends to overestimate T_{mg} (Shu et al. 2013). In a very recent study, $Cs_3Sb_2Cl_{9-n}Br_n$ solid solution was proved and the same linear relationship of halogen composition with band gap was shown (J. Lee et al. 2023). The Ω value for $Cs_3Sb_2Cl_{9-n}I_n$ is considerably larger resulting in T_{mg} = 558 K, substantially higher as expected due to large lattice mismatch between the pure perovskites. Thus, miscibility is predicted to be very limited for this I-Cl perovskite solid solution, in agreement with experimental reports (Paul, Pal, and Larson 2020).

3.4.3 Surfaces

Surfaces have been investigated for $Cs_3Sb_2Br_9$ (C. Lu et al. 2020; P. Liu et al. 2020) but a comparative study encompassing low-index surfaces for all three $Cs_3Sb_2X_9$ perovskites is currently lacking. *Table 5* presents the surface energy results for CsX-terminated slabs, showcasing both (0001) and (1000) surfaces. Our findings align with previous calculations conducted on $Cs_3Sb_2Br_9$ (C. Lu et al. 2020), revealing that (1000) surfaces exhibit greater stability than (0001) surfaces. Furthermore, our results demonstrate a similar trend observed in halogen surfaces of $CsPbX_3$ (Nazari, Azar, and Doroudi 2020), where surface stability follows the order Cl, Br, I. However, it is noteworthy that in $Cs_3Sb_2X_9$, the reduction in surface free energy is more accentuated for bromide and iodine surfaces than what is observed in $CsPbX_3$ surfaces.

Table 5 – Surface free energy (γ) and cleavage energy (γ_{cle}) for Sb-X terminated slabs for each halide perovskite.

	$\gamma\ (eV/Å^2)$	$\gamma_{\text{cle}}\ (eV/Å^2)$
(0001) surface		
$Cs_3Sb_2Cl_9$	0.0567	0.0691
$Cs_3Sb_2Br_9$	0.0366	0.0479
$Cs_3Sb_2I_9$	0.0303	0.0418
(1000) surface		
$Cs_3Sb_2Cl_9$	0.0111	0.0173
$Cs_3Sb_2Br_9$	0.0073	0.0093
$Cs_3Sb_2I_9$	0.0060	0.0081

To analyze the geometry of the relaxed slabs, *Figure 25* presents a plot illustrating the element counts in each coordinate, starting from the top of the (1000) slabs. By comparing the distribution of atoms in the unrelaxed slabs (bottom) to the relaxed slabs (top) in the figure, noticeable changes can be observed. Specifically, it can be observed that in each of the halide perovskites, the Cs atom is drawn further into the slab. This shift is more pronounced in iodine perovskite and less significant in chlorine perovskite, attributed to their varying electronegativities. In the second perovskite layer, the effects differ depending on the halide. In chlorine perovskite, the Cs and Cl atoms remain in proximity while the Sb atom is pushed away. On the other hand, in iodine perovskite, the I-Sb bond is strengthened while Cs recedes into the slab. In the case of bromine perovskite, due to its intermediate electronegativity between chlorine and iodine, only a moderate distortion occurs in the second layer. As a result, its geometry more closely resembles that of the bulk in contrast to the other halide perovskites.

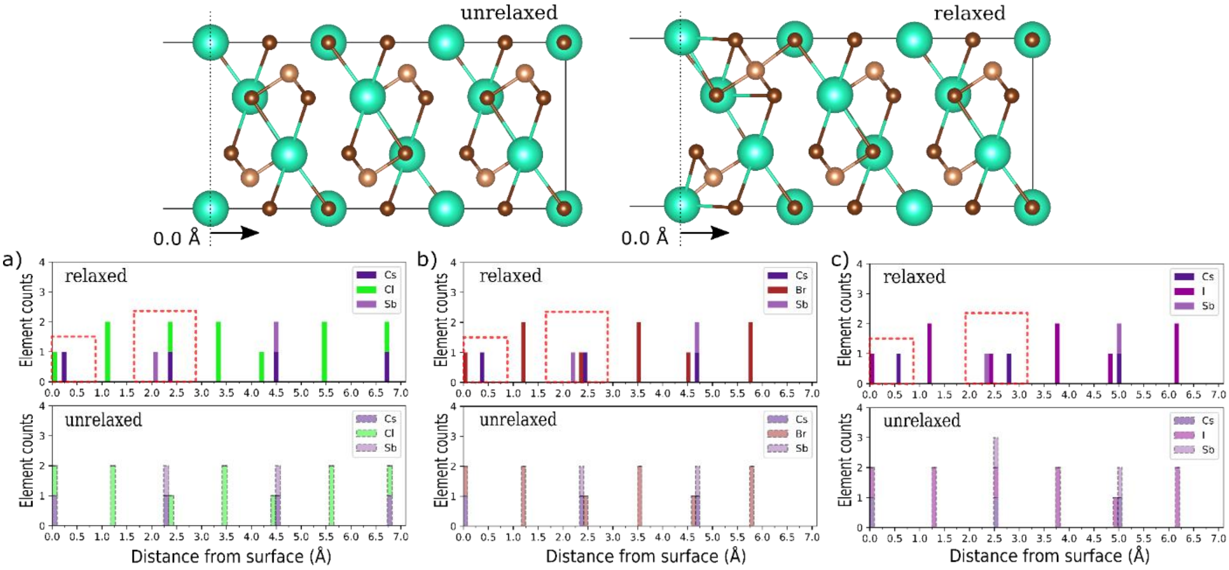


Figure 25 – Geometric analysis of the top layers of (1000) CsX-terminated surface of halide perovskites. Top of figure illustrates unrelaxed and relaxed $Cs_3Sb_2I_9$ slabs, measurements start on the outermost surface atom as shown. (a), (b) and (c) present element counts in given position for both relaxed and unrelaxed slab for $Cs_3Sb_2Cl_9$, $Cs_3Sb_2Br_9$ and $Cs_3Sb_2I_9$, respectively.

The recession of Cs atoms in the first layer following halide electronegativity is also observed in the (0001) slabs, with considerably less distortion, as shown in supplementary *Figure B5*. Furthermore, negligible distortion occurs in subsequent perovskite layers on these slabs. Consequently, the larger surface free energy of the (0001) surface compared to the (1000) surface can be attributed to its greater cleavage energy and reduced capacity to stabilize dangling bonds in surface atoms through geometry relaxation.

Total and partial density of states is presented for each of the Cs-X terminated slabs in *Figure 26*. A substantial difference in the band edge states is observed between the (0001) and (1000) surfaces. Specifically, (0001) surfaces display a significant reduction in the band gap due to the shift of Sb 5s states to higher energies and the presence of unpaired spin states that extend the VB edge. This result indicates a surface that is more active for photocatalysis. Since the (0001) surface is more readily obtained experimentally (Jihong Li et al. 2022), the results observed throughout the literature for photocatalysis (C. Lu et al. 2020; J. Lee et al. 2023; G. Chen et al. 2020) with these materials can be understood based on these surface properties. This idea is reinforced by the work of C. Lu et al. (2020) which showed that $Cs_3Sb_2Br_9$ (0001) surfaces had a lower free energy for CO_2 reduction than (1000). Our findings also indicate that this effect is particularly pronounced in the chlorine perovskite, reducing the band gap by approximately 1 eV compared to the bulk band gap value. The Fermi level appears in the middle of the opposing spin states prior to the band gap, and the lack of a gap between the occupied and unoccupied states implies that this surface may be highly reactive, acting as an electron trap due to dangling halogen bonds. This suggests that, in addition to the bromine and iodine perovskites that have been explored in photocatalysis, $Cs_3Sb_2Cl_9$ and chlorine-doped $Cs_3Sb_2Br_9$ or $Cs_3Sb_2I_9$ could also yield favorable results.

On the other hand, (1000) surfaces display a well-defined band gap that separates the VB and CB, and the band gap energy is not significantly reduced in comparison to the bulk structure. This suggests that the transport properties will not be significantly affected by the presence of such surfaces. The bromide (1000) surface, in particular, displays a remarkably close band gap energy to its bulk counterpart, which is a direct consequence of its lower distortion as analyzed in *Figure 25*(b). Thus, promoting the growth of the (1000) surface of $Cs_3Sb_2Br_9$ may be a viable option for maintaining good transport properties, especially for photovoltaic applications in which surface states can greatly compromise device efficiency. Recently, Sachchidanand et al. (2021) investigated numerically $Cs_3Sb_2Br_9$ for photovoltaic application and their results were promising, this work reinforces their suggestion and also recommends surface control to achieve greater performance.

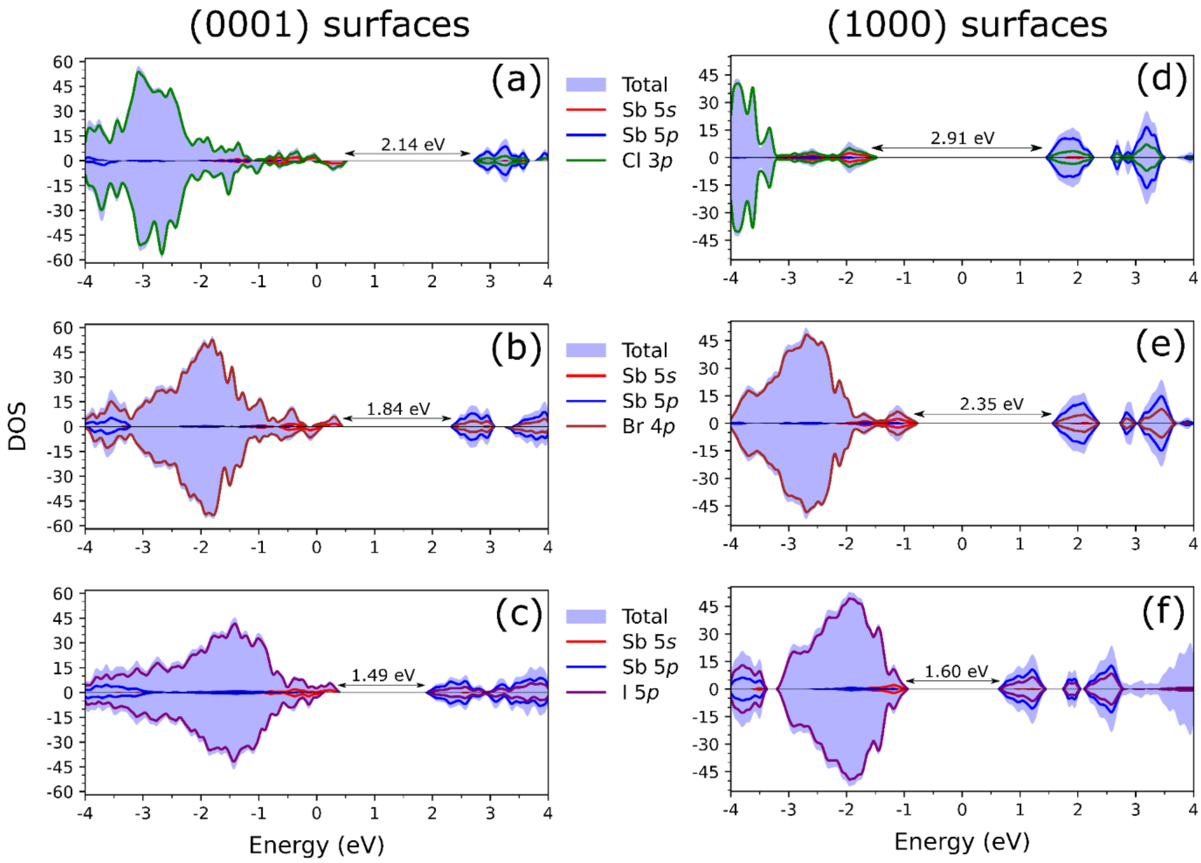


Figure 26 – Partial density of states for each of the halide perovskite Cs-X terminated slabs (0001) surfaces is shown in plots (a-c) and (1000) surface is shown in plots (d-f).

3.4.4 Interfaces and band alignment

Heterostructures and interfaces play a crucial role in optoelectronic applications, but research on lead-free perovskites such as $Cs_3Sb_2X_9$ remains scarce. Theoretical calculations on these systems can offer valuable insights into electronic properties and carrier dynamics, contributing to the optimization of halide perovskite-based devices. To address this, we constructed supercells containing $Cs_3Sb_2Br_9$|$Cs_3Sb_2Cl_9$ and $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ perovskites along the [0001] direction, as presented in *Figure 27(a*,c), to explore interfaces and band alignment within these materials. Geometry and relative positions of VB levels through potential alignment calculation were evaluated in these structures. In *Figure 27*(a) and 27(c) relaxed atomic positions are presented with unrelaxed initial positions presented as contours with same atomic colors. Both interface supercells had their atomic positions and lattice parameters relaxed, supercells relaxed to the average lattice parameter of two pristine lattices with no significant deviation (see *Table B6*), on the other hand, atomic positions presented relevant displacements on both structures.

For $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$, displacements were more concentrated on the $Cs_3Sb_2Br_9$ region and for $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ displacements were appreciable in the $Cs_3Sb_2Cl_9$ region as can be appreciated in *Figure 27*(a,c). This can be understood based on the Coulomb energy difference experienced by the cation-X bonds in the interface, bonds with the least electronegative halogen increase while bonds with the most electronegative element shorten due to a displacement of metal ions toward most electronegative region. Bader charges for Sb presented in *Figure 27*(b,d) show clearly the different charge environment experienced by the cations depending on the bonding halogen.

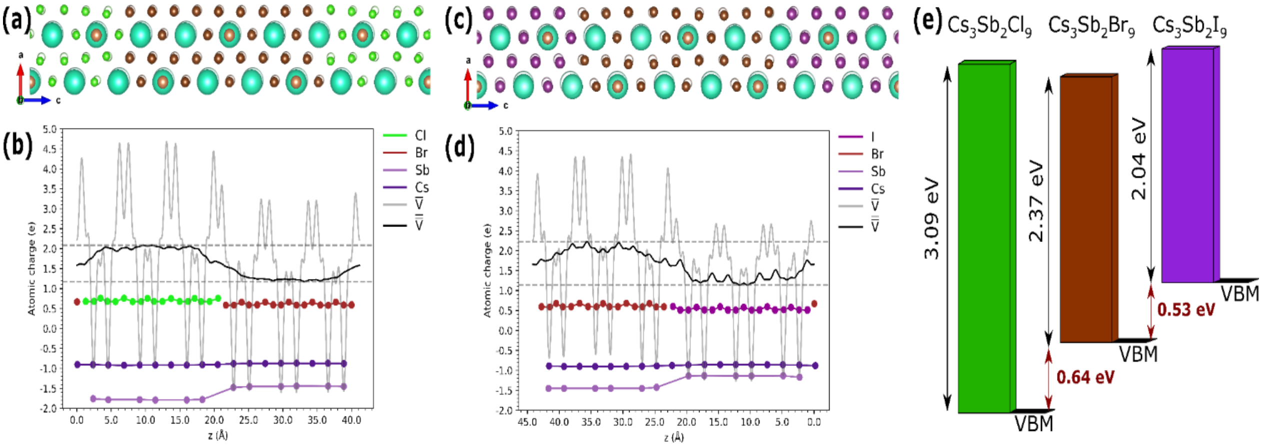


Figure 27 – Interfaces of $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ and $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ in (a) and (c) with relaxed atomic positions (initial positions outlined in the background). In (b) and (d), average planar potential ($\bar{V}$) and its macroscopic average ($\bar{\bar{V}}$) perpendicular to the interface are shown, along with Bader charge per atom in each layer. (e) Illustrates the band alignment of the three halide perovskites based on band offset calculations.

In *Figure 28*, partial density of states for supercell structures is presented for both $Cs_3Sb_2Br_9|Cs_3Sb_2Cl_9$ and $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$, full density of states is presented for each interface in *Figure 28(a)* and *(d)*, these states are also projected for each material. Bromide states present a slightly reduced gap when in contact to chlorine than to iodine, chlorine being more electronegative disturbs bromine-cation states close to VBM to higher energy levels (reducing band gap) as the electron cloud is pulled toward the more electronegative ion. Analyzing the projections for chlorine and iodine in *Figure 28(c)* and *(d)*, respectively, there is also a clear reduction from pristine band gap values for these compounds that can be associated to reduced repulsion between halogens due to Coulomb energy gain (Yin, Yan, and Wei 2014) shifting energy levels to lower energies in the CB.

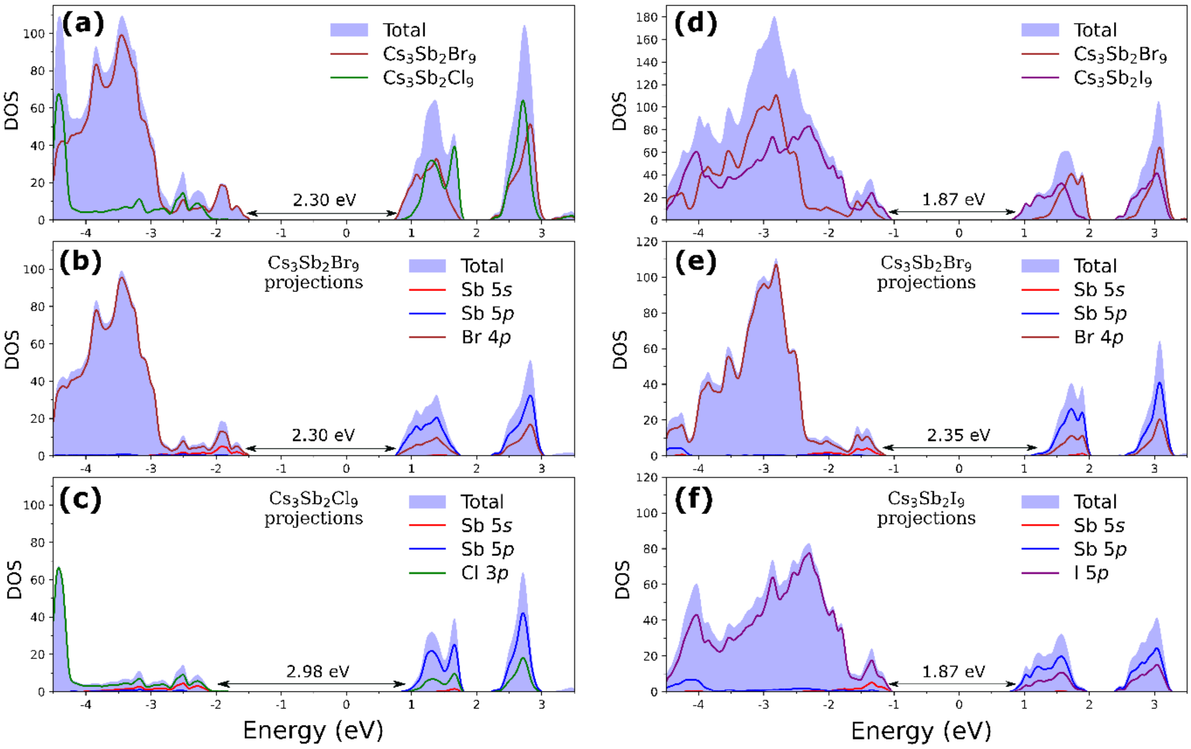


Figure 28 – Partial density of states of supercell interfaces for (a-c) $Cs_3Sb_2Br_9$|$Cs_3Sb_2Cl_9$ and (d-f) $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ along with separated projections for each side on the interface.

The relative changes of DOS when interface is formed also provide information on possible carrier dynamics for these perovskite combinations. In VB there is a balanced contribution of both $Cs_3Sb_2I_9$ and $Cs_3Sb_2Br_9$ in the $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ case, on the other hand, for $Cs_3Sb_2Br_9$|$Cs_3Sb_2Cl_9$ valence states of Cl 3*p* and Br 4*p* are well separated suggesting chlorine perovskite is likely to act as a strong hole injector for $Cs_3Sb_2Br_9$. Therefore, a core-shell structure as $Cs_3Sb_2Br_9$@$Cs_3Sb_2Cl_9$ should be efficient for LEDs with the chlorine shell acting both as a diffusion barrier and carrier injector. Similar systems working with this principle have already been developed for $CsPbBr_{3-x}Cl_x$ perovskites (P. Zhang et al. 2018; Y. R. Park et al. 2021). For the CB, there is a significant tailing of iodine states in $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ that directly influences the band gap, therefore, this combination increases the defect tolerance compared to $Cs_3Sb_2I_9$ only and makes this material an interesting candidate for photovoltaic applications. Similar conclusions have been drawn for $CsPbX_{3-x}Y_x$ structures, in which a high iodine content was linked to larger diffusion lengths making them appropriate for photovoltaic devices (P. Zhang et al. 2018).

The aforementioned findings also corroborate our previous observations on band gap bowing for mixed halide perovskites; mixed iodine and bromine perovskites benefit from a stronger tailing and mixing of halide states, resulting in a larger bowing parameter. In contrast, for mixed chlorine and bromine perovskites, this interaction is significantly smaller, causing the band gap to follow a linear trend with halogen composition which was corroborated recently (J. Lee et al. 2023). The difference in mixing in the different halides can be clearly seen in the potential curves in *Figure 27(b,d)*, where $Cs_3Sb_2Br_9$|$Cs_3Sb_2Cl_9$ reaches bulk potential approximately 8 Å away from the interface, whereas $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ reaches bulk potential in approximately 12 Å away from the interface due to stronger coupling.

To conclude our analysis in the carrier dynamics on these interfaces, accurate band alignment assessment is crucial. To evaluate VB offsets between different $Cs_3Sb_2X_9$ perovskites separate calculations for each bulk material are performed and the VBM with respect to average electrostatic potential in the material is determined. However, simply taking the difference of VBM between different materials is not sufficient to determine the band offset since VBM is ill-defined for bulk calculations with periodic boundary conditions (Kleinman 1981). A more reliable and accurate way to determine band offsets is to correct this difference through a potential alignment performed on supercell interface calculation as given by the formula (Van De Walle and Martin 1987; Weston et al. 2018; Hinuma et al. 2014):

$$\Delta E_v = (E_v^B - E_v^A) + \Delta V, \tag{46}$$

where E_v^A and E_v^B represent the VBM of materials A and B relative to bulk average electrostatic potential, and ΔV is the potential alignment obtained from the superlattice calculation of the interface. For $Cs_3Sb_2X_9$ perovskites, ΔV is determined from planar averaged electrostatic potential ($\bar{V}$) and its macroscopic average ($\bar{\bar{V}}$) for supercell interfaces $Cs_3Sb_2Br_9$|$Cs_3Sb_2Cl_9$ and $Cs_3Sb_2I_9$|$Cs_3Sb_2Br_9$ as presented in *Figure 27(b,d)*. Details on the calculation followed methodology described elsewhere (Weston et al. 2018). Detailed results for the bulk VBM values, average potentials and band offsets are presented in *Supporting Information* (*Table B7*), final results for band alignment including band gap and band offsets are presented in *Figure 27(e)*. VBM of perovskites rise in energy from higher to lower electronegativity meaning that holes will be less energetic in iodine than chlorine. These results agree qualitatively with previous report from Liu et al. (2019) using electron affinity of the materials, although calculation of band alignment is deemed more precise when actual interface calculations are performed (Hinuma et al. 2014) as in this work. CB energy levels are close in energy for all perovskites and bromine perovskite present the lowest CBM level according to our calculations, therefore $Cs_3Sb_2Br_9$ has the greatest electron affinity of the three halide perovskites. Traditional band alignment using HSE normalized VBM values combined with PBE calculated potential alignment (Weston et al. 2018), as presented in *Figure B6*, reach same conclusion of PBE+U band alignment despite small offset differences (*Table B8*).

Previous work from Liu et al. (2019) reports higher electron affinity for $Cs_3Sb_2Cl_9$ perovskite instead of $Cs_3Sb_2Br_9$, this may be due to their large band gap deviation obtained for $Cs_3Sb_2Br_9$ (2.60 eV in HSE vs. 2.30 eV in experiment). In this work, both HSE06 and PBE+U calculations presented a small deviation (~0.1 eV) to experimentally reported gap. If this result is verified by experiment, a heterostructure such as $Cs_3Sb_2Br_9$@$Cs_3Sb_2Cl_9$ would be promising for photoluminescence since $Cs_3Sb_2Cl_9$ would both confine charge carriers and inject electrons and holes for recombination in $Cs_3Sb_2Br_9$. The construction of similar core-shell structures is feasible for $CsPbX_3$ as has been reported for $CsPbBr_3$@$PbBr_x$ (Xiaoming Li et al. 2016) and $CsPbBr_3$@$CsPbBr_{3-x}Cl_x$ (G. Zhang et al. 2020). Moreover, G. Zhang et al. (2020) have proved that despite strong anion exchange in $CsPbX_3$ perovskites stable heterojunctions could still be synthesized in appropriate conditions, thus, similar heterojunctions for $Cs_3Sb_2X_9$ to harness favorable band alignment are presumably possible.

3.4.5 Clusters

Despite the promising applications, theoretical investigation for $Cs_3Sb_2X_9$ perovskites under spatial confinement is still lacking. To fill this gap, a non-stoichiometric $Cs_{13}Sb_6X_{30}$ cluster with 49 atoms was investigated for its geometry and electronic structure properties. In this cluster, the SbX_6 coordination was preserved to avoid dangling Sb bonds. The partial density of states for the halogen $Cs_{13}Sb_6X_{30}$ clusters is presented in *Figure 29*. Spatial confinement produces larger band gaps compared to bulk counterparts, which is expected and agrees with experimental observations (Ma et al. 2019). Since the clusters are non-stoichiometric with excessive

Cs and X, the halogen p orbitals and Sb 5p, which form the CB in bulk perovskite, exhibit some polarization and stretch a few tenths of an eV towards the VB due to stronger binding with available Cs electrons. Nevertheless, the bands are well defined, presenting similar qualities of VB and CB of bulk materials, and it is reasonable to consider the models a good approximation to evaluate the properties of this material under spatial confinement. The band gap differences from bulk perovskites were +0.02 eV and +0.24 eV for chlorine and iodine halogen clusters, respectively, compared to +0.39 eV for the bromide cluster. These findings give atomistic origins for what has been observed in Ma et al.'s work (2019), $Cs_3Sb_2Cl_9$ quantum dots with a size of 5.0 nm and $Cs_3Sb_2I_9$ quantum dots with a size of 5.8 nm presented similar band gaps to the reported gap in bulk perovskite, 3.22 eV (3.09 eV in bulk by Blasse, 1983) and 1.93 eV (1.95 eV in bulk by Saparov et al. 2015), respectively. Similarly, $Cs_3Sb_2Br_9$ quantum dots presented a band gap of 3.03 eV in that work, a relevant difference from the 2.36 eV reported for bulk single-crystal $Cs_3Sb_2Br_9$ (Jian Zhang et al. 2017).

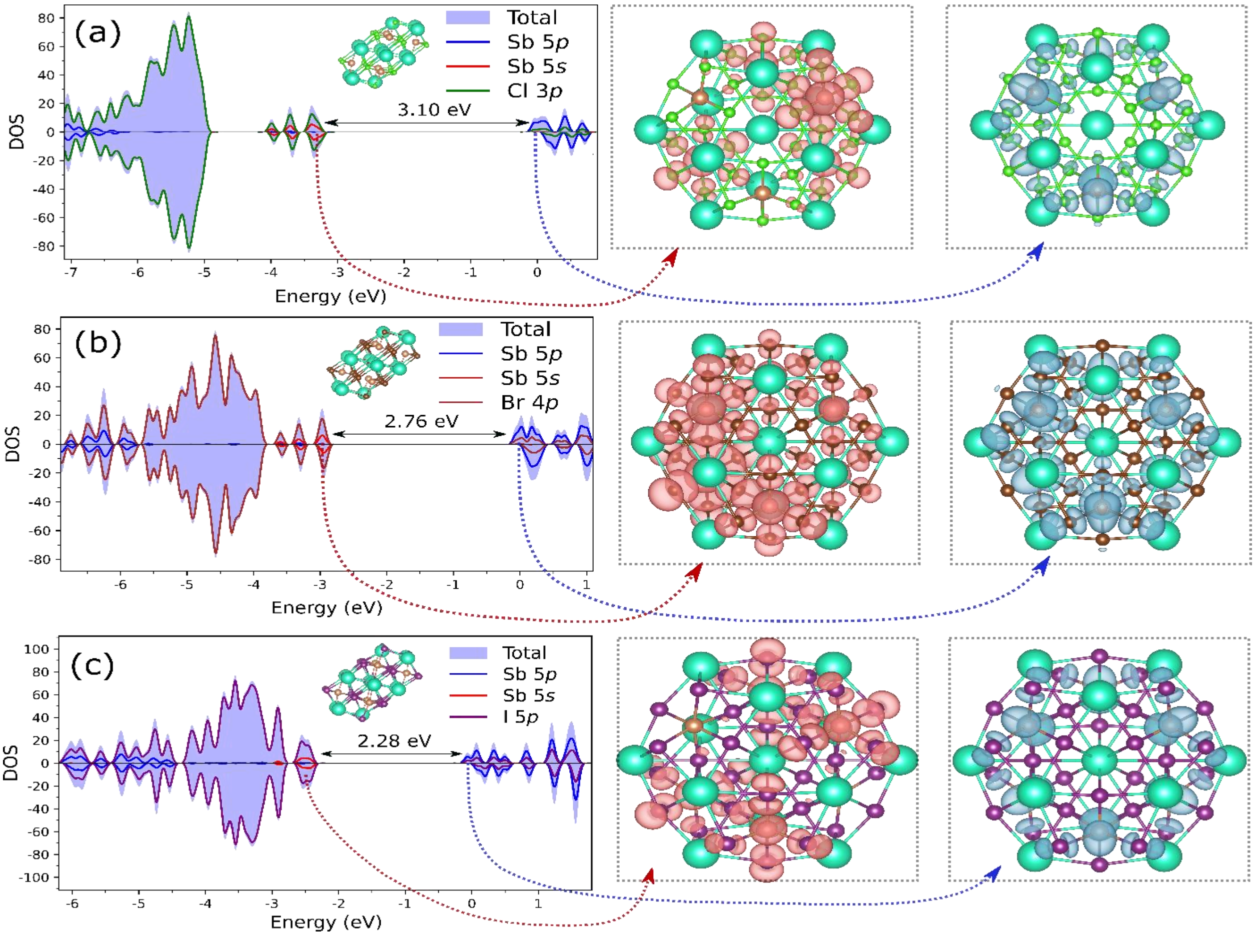


Figure 29 – Partial density of states for clusters of $Cs_{13}Sb_6X_{30}$ for each of the halogens Cl, Br and I. HOMO and LUMO states are also presented for each cluster.

Smaller band gap increase for chlorine and iodine can be understood based on geometrical changes and charge transfer in the clusters. By analyzing the difference in average bond lengths, in *Table 6*, and the average Bader charges in each atom between clusters and bulk, in

Table 7, we notice that $Cs_3Sb_2Br_9$ cluster has the lowest differences in bond length and in the charge of Sb and Cs. For the $Cs_3Sb_2Cl_9$ cluster, as in the case of the (0001) CsCl-terminated surface of $Cs_3Sb_2Cl_9$, Sb ions accumulate charge due to stronger Cs-Cl bonds, leading to higher energy states in the VB and a lower band gap. In the case of the $Cs_3Sb_2I_9$ cluster, similar charge accumulation in Sb ions occurs, but the effect is dominated by exposed Cs atoms moving inward to bond more strongly with iodine. This distortion in SbI_6 octahedra results in elongated Sb-I bonds. Thus, for both iodine and chlorine clusters, there is an attraction that can be seen clearly when we investigate geometry changes in the cluster (more details in *Supporting Information B.5*), showing that Sb is repelled in favor of Cs-Cl or Cs-I bond formation. Conversely, the bromine cluster, presumably due to intermediate electronegativity, exhibits a balance that avoids strong contraction of Cs-Br, leading to lower distortion. Therefore, based on previous results, the small band gap increments of iodine and chlorine clusters compared to bulk can be understood based on larger Sb-X bonds that induce lower interatomic potentials, counteracting the effect of spatial confinement in the band gap. In addition to our primary investigations, we explored the effects of halogen alloying within the clusters. Our focus was on the substitution of iodine with chlorine, a process conducted at both the longitudinal face and edge sites. Notably, our findings reveal the most significant influence of substitution occurring at the edge sites. For more detailed information, including specific data and figures (*Figure B9*), refer to the *Supporting Information (Appendix B.5 Clusters)*.

Table 6 – Average bond length of Cs-X and Sb-X bonds for bulk $Cs_3Sb_2X_9$ perovskites and $Cs_{13}Sb_6X_{30}$ clusters. Δd indicates the difference between average bond lengths in cluster and bulk structures.

		Average bond length (Å)		
material	bonds	bulk	cluster	Δd (bulk-cluster)
$Cs_3Sb_2Cl_9$	Cs-Cl	3.879	3.640	-0.239
	Sb-Cl	2.609	2.739	+0.130
$Cs_3Sb_2Br_9$	Cs-Br	4.059	3.876	-0.182
	Sb-Br	2.782	2.863	+0.081
$Cs_3Sb_2I_9$	Cs-I	4.341	3.968	-0.372
	Sb-I	2.988	3.092	+0.103

Table 7 – Average Bader charge of elements in bulk $Cs_3Sb_2X_9$ perovskites and corresponding $Cs_{13}Sb_6X_{30}$ clusters. Δe indicates the difference between average Bader charges in cluster and bulk structures.

		Average Bader charges (e)		
	Element	bulk	Cluster	Δe (bulk-cluster)
$Cs_3Sb_2Cl_9$	Cs	-0.9116	-0.9004	-0.0112
	Sb	-1.7993	-1.7155	-0.0838
	Cl	+0.7037	+0.7332	-0.0295
$Cs_3Sb_2Br_9$	Cs	-0.8866	-0.8858	-0.0008
	Sb	-1.4626	-1.4127	-0.0499
	Br	+0.6206	+0.6663	-0.0457
$Cs_3Sb_2I_9$	Cs	-0.8745	-0.8524	-0.0221
	Sb	-1.1759	-1.1210	-0.0549
	I	+0.5529	+0.5936	-0.0407

The lower distortion of $Cs_{13}Sb_6Br_{30}$ is reflected in the charge density distribution corresponding to states below and above band gap, as presented on the right-side of *Figure 29* for each halogen cluster. Meanwhile, iodine and chlorine clusters exhibit localization of charge, leaving some SbX_6 octahedra with negligible contributions. $Cs_{13}Sb_6Br_{30}$, on the other hand, shows a more homogeneous charge distribution for these states. A less localized distribution of HOMO-LUMO states throughout the structure has been associated with stronger optical transitions in halide perovskite clusters (Koliogiorgos et al. 2018). This seems to corroborate experimental observations reported by Ma et al. (Ma et al. 2019), in which $Cs_3Sb_2Br_9$ QDs presented larger PLQY than other halide perovskite QDs.

Jian Zhang et al. (2017) also reported exceptional PLQY for colloidal $Cs_3Sb_2Br_9$ QDs, which was attributed to high exciton binding energy and good surface passivation. High exciton binding energy in inorganic semiconductors is linked to greater valence electron localization from reduced electronic screening (Dvorak, Wei, and Wu 2013), indicating lower electronic screening in $Cs_3Sb_2Br_9$ compared to $Cs_3Sb_2Cl_9$ and $Cs_3Sb_2I_9$. Our calculations also hint at a lower screening in bromine perovskite, as a lower Hubbard U value (U_{Br} = 2.5 eV) is necessary to reproduce bulk electronic structure at HSE level than iodine and chlorine perovskites (U_I = 3 eV and U_{Cl} = 4.5 eV, respectively). For instance, a surprisingly small Hubbard value for transition metals in oxide perovskites of 4d series, compared to 5d series, has been linked to weaker screening effects and larger exciton binding energy (Vaugier, Jiang, and Biermann 2012; Varrassi et al. 2021). This may be the case in $Cs_3Sb_2Br_9$ and future studies should include explicit many-body corrections to further investigate this matter. Additionally, calculating larger structures could offer clearer insights into confinement effects and surface properties in these materials.

3.5 CONCLUSION

Our exploration of halide mixing provided insights into band gap variations, structural shifts, and potential ordered structures. Investigation into enthalpy of formation revealed potential uniform solid solutions for $Cs_3Sb_2Cl_{9-n}Br_n$ and $Cs_3Sb_2Br_{9-n}I_n$, with higher temperatures required for full alloying in $Cs_3Sb_2Cl_{9-n}I_n$. (1000) surfaces retained electronic properties advantageous for photovoltaics hindering recombination. Conversely, (0001) surfaces exhibited significant band gap reduction, suggesting reactivity suitable for photocatalysis. These findings underscore the impact of surface orientation on electronic properties. Regarding interfaces, more efficient LEDs are suggested to be obtained from $Cs_3Sb_2Br_9$@$Cs_3Sb_2Cl_9$ harnessing the chlorine shell as a diffusion barrier and carrier injector. Defect tolerance of $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ was indicated, making it valuable for photovoltaics exploration. Cluster simulations estimated $Cs_3Sb_2X_9$ nanocrystal properties, suggesting geometry's role in superior photoluminescence observed in prior experiments with $Cs_3Sb_2Br_9$ nanocrystals. Halogen substitution's impact on cluster sites unveiled edge sites' importance for band gap tuning. In summary, the present study underscores the potential of lead-free $Cs_3Sb_2X_9$ perovskites for stable and efficient solar cells and optoelectronic devices. The study bridges knowledge gaps in halogen alloying, surface analysis, heterostructures, and confined structures within these materials. The results lay groundwork for further optimizing and developing $Cs_3Sb_2X_9$ perovskites to enhance their efficiency across diverse optoelectronic applications.

CHAPTER 4 — Doping effects on the optoelectronic properties and the stability of $Cs_3Sb_2I_9$: Density Functional Theory insights on photovoltaics and light-emitting devices

doi.org/10.1016/j.jsamd.2024.100700

4.1 RESEARCH PROBLEM

$Cs_3Sb_2I_9$ polymorphs possess great potential as lead-free materials across a wide range of technologies. Literature shows that the unprecedented success of the pioneering lead halide perovskites in applications relies heavily on doping, which enables fine-tuning of key properties such as bandgap, photoluminescence (PL) intensity, carrier lifetime, charge mobility, and the induction of catalytic active sites (Saliba et al. 2016; Kumawat et al. 2019; Raza et al. 2021). Moreover, doping plays a vital role in controlling defect density and enhancing the stability of halide perovskites (S. Chen et al. 2023). It is noteworthy that these strides in doping have often been preceded or run parallel with comprehensive theoretical simulations. Nonetheless, research on halogen or cation doping in $Cs_3Sb_2I_9$ perovskites although on the rise (Malavasi et al. 2023; G. Chen et al. 2020; X. Wang et al. 2020; F. Jiang et al. 2018; Paul, Pal, and Larson 2020; Jihong Li et al. 2022; Singh et al. 2019) still leave a wide avenue for further theoretical and experimental investigations, particularly in the under-explored dimer-phase.

Seeking to bridge this research gap, we employed Density Functional Theory (DFT) to assess the effects of specific metal and halogen dopants on the optoelectronic properties of $Cs_3Sb_2I_9$ perovskite polymorphs. The metal dopants included ions of similar ionic radius to Sb^{3+}, namely Ag, In, Mo, Nb, and Sc, while the halogens were chlorine and bromine, as substitutes for iodine. For each specific dopant, we analyzed the resulting changes in electronic structure, geometry, and absorption coefficients. Additionally, we examined defect formation energies and convex hull distances to gauge the stability and viability of these modifications. Our findings revealed significant potential for strategic metal and halogen doping to tailor the optoelectronic properties of $Cs_3Sb_2I_9$ polymorphs underscoring the distinctions between them.

4.2 METHODOLOGY

In this study, we employed a comprehensive methodology to investigate the properties of metal doped and halogen doped $Cs_3Sb_2I_9$ perovskite considering both layered (space group $P\overline{3}m1$) and dimeric (space group $P6_3/mmc$) polymorphs. The key components of our methodology are as follows:

- *Ab-initio calculations*: we conducted Density Functional Theory (DFT) calculations using Quantum ESPRESSO (Giannozzi et al. 2009) with the PBE exchange-correlation functional for all studied structures. Ultrasoft GBRV pseudopotentials (Garrity et al. 2014) were utilized to describe electron-ion interactions. Kohn-Sham orbitals were expanded in a plane-wave basis set with energy cutoffs of 50 Ry for wave functions and 300 Ry for charge density. Brillouin zone integration was performed using a 4×4×2 Γ-centered Monkhorst-Pack grid. We ensured self-consistency in total energy with tolerances of less than 10^{-8} Ry/atom for electronic energy and 10^{-6} Ry/atom for ionic minimization. For structural relaxation, the BFGS quasi-Newton algorithm was employed (Billeter, Curioni, and Andreoni 2003), and atomic positions were relaxed until residual forces on each atom were less than 10^{-4} Ry/Bohr. We evaluated various electronic properties, including the density of states, band structure, band gap energy, effective masses, and charge analysis.

- *Metal and halogen doped structures*: To investigate the effects of metal doping, we considered transition metals, M = Ag, In, Mo, Nb, Sc, which presented +3 oxidation states with similar ionic radii to Sb, based on the data of Shannon (Ouyang 2020; Shannon 1976) as shown in *Table C1* on *Supporting Information*, for substitution in $Cs_3Sb_2I_9$ perovskite structures. Additionally, we explored bismuth (Bi) doping for comparison purposes due to its isoelectronic configuration to Sb^{3+}. Doped 1×1×2 supercells were created by replacing one Sb atom with a metal cation — yielding $Cs_3Sb_{1.5}M_{0.5}I_9$ as composition. In the case of halogen doping two iodine atoms of $Cs_3Sb_2I_9$ were replaced by either Br or Cl — producing a $Cs_3Sb_2I_7Y_2$ (Y = Cl, Br) composition — since there are two distinct halogen sites 6*h* and 12*k* for $Cs_3Sb_2I_9$ ($P6_3/mmc$) and 6*i* and 3*e* for

$Cs_3Sb_2I_9$ ($P\overline{3}m1$) energy analysis was conducted to determine the substitution sites.

- *Formation energy calculations*: we calculated the formation energy for defect substitution using *Equation (*47*)* (Freysoldt et al. 2014), as follows:

$$E_f[D] = E_{tot}[D] - E_{tot}[Cs_3Sb_2X_9] \pm \sum_{i} n_i \mu_i \tag{47}$$

where the formation energy of a neutral charge defect $E_f[D]$ is calculated by the difference in total energy of a supercell with defect, $E_{tot}[D]$, and the energy of pristine supercell of $Cs_3Sb_2I_9$ added to the energy corresponding to the removal (−) or addition (+) of elements to form the defect proportionally to number of ions, n_i, multiplied by μ_i, corresponding to the chemical potential of this element.

- *Optical properties calculations*: to assess the optical properties of the materials, we computed the dielectric function, ε(ω), using the SIMPLE code (Prandini et al. 2019) in Quantum ESPRESSO. We employed Vanderbilt pseudopotentials from SG15 (Schlipf and Gygi 2015) databases and considered relativistic effects at the scalar level. Real and imaginary components of the dielectric function were used to determine optical constants. These calculations were conducted for both polymorphs of the $Cs_3Sb_2I_9$ structure and the respective doped systems, utilizing a 10×10×4 grid of k-points which yielded well converged results for the 28 atoms supercells.
- *ACBN0 calculations*: To address the underestimation of band gap prediction by the PBE exchange-correlation functional, we conducted an analysis employing the pseudo-hybrid ACBN0 method (Agapito, Curtarolo, and Nardelli 2015), which applies the DFT+U ansatz with iteratively calculated Hubbard values, detailed implementation provided in *Appendix C.2.5*. These additional calculations were performed on pristine structures and selected metal-doped structures to provide better band gap approximations.

4.3 RESULTS AND DISCUSSION

4.3.1 $Cs_3Sb_2I_9$ polymorphs

The initial coordinates for $Cs_3Sb_2I_9$ polymorphs were obtained from the literature (Kihara and Sudo 1974) and are shown in *Table C2*. The deviations in the relaxed lattice parameters from the experimental values consistently remained under 4%, surpassing the level of accuracy observed in previous first-principles calculations as depicted in *Table C2*. The halogen-antimony bond sizes also closely matched the experimental values, with an error margin of less than 0.02 Å, demonstrating the accuracy of the theoretical method in reproducing the system geometry.

Figure 30 shows the band structures and projected density of states. For both polymorphs, the valence band is mainly composed of the I 5*p* and Sb 5*s* orbitals, while the conduction band has contributions from the iodine 5*p* and antimony 5*p* orbitals. The presence of antibonding coupling between Sb lone-pair 5*s* orbital and I 5p in the higher levels of the valence band resembles the Pb lone-pair 6s and I 5p orbital antibonding coupling that is attributed to a better defect tolerance in the $CH_3NH_3PbI_3$ perovskite (Y. L. Liu et al. 2019). In both polymorphs, the valence band maximum lies in between two high symmetry points, K and Γ, indicated as K* in Figure 30, while the conduction band minimum lies on Γ the point. The trigonal structure presents an indirect band gap of 1.52 eV, yet the direct transition at Γ-Γ requires only slightly higher energy photons, resulting in a band gap of 1.59 eV which enables faster transitions in this structure. The hexagonal structure presents an indirect band gap of 1.81 eV, while the lower energy direct transition (M–M) occurs with a band gap energy of 2.14 eV. Indeed, DFT systematically underestimates band gap compared to experimental values, nevertheless, the disparities in band gaps observed in this study align with previous theoretical reports (see Table C3) and fall within the expected range for halide perovskites (Leppert, Rangel, and Neaton 2019). Analysis of the charge density plots, average Bader charges and binding energies are depicted for both polymorphs in the *Supporting Information (Appendix C.2.1)*, reinforcing the enhanced thermodynamical stability of the hexagonal structure, which is obtained in a facile approach in a solution based synthesis (Singh et al. 2018; Saparov et al. 2015). Additionally, from the effective mass analysis (Table C5), a lower electron and hole effective masses in the $P\bar{3}m1$ structure can be observed, compared to the $P6_3/mmc$ polymorph, due to the higher structural dimensionality in the former.

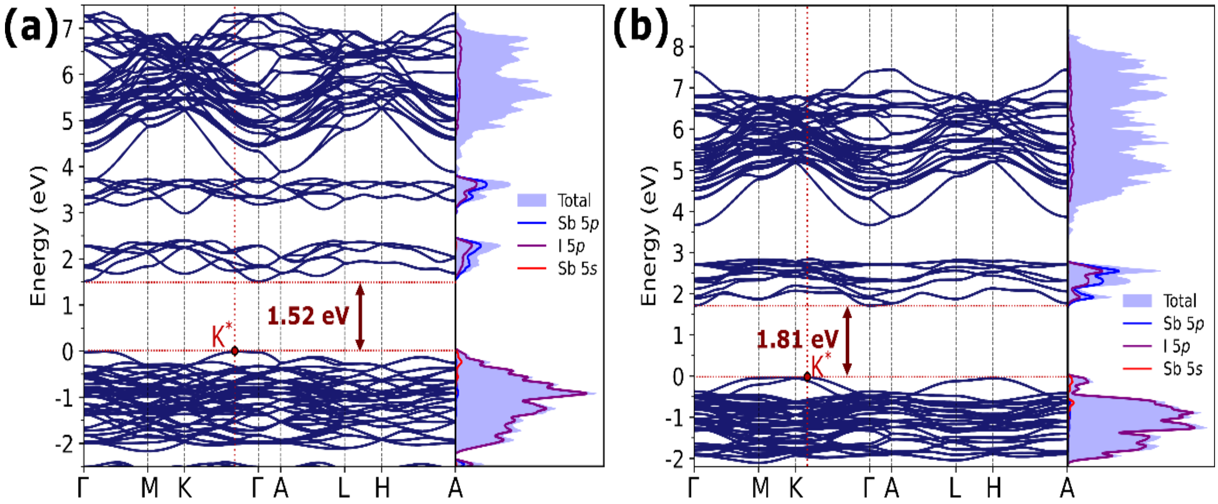


Figure 30 – Band structures and projected density of states for (a) $Cs_3Sb_2I_9$ ($P\bar{3}m1$), and (b) $Cs_3Sb_2I_9$ ($P6_3/mmc$). Fermi level is at 0 eV.

4.3.2 Doping with transition metals

Figure 31 shows the $Cs_3Sb_2I_9$ polymorphs with the replacement of one Sb atom by one dopant metal atom (M) in the respective unit cell. The influence of substitutional doping at the Sb site, on the lattice parameters for both polymorphs is presented in Table 8. Even with the high doping concentration explored, the lattice parameters display only minor deviations, amounting to less than 1.3%. This remarkable result demonstrates that structures with the chosen dopants, which possess a comparable ionic radius to Sb^{3+}, can be synthesized. The I–M bond distance is calculated considering the average of these bonds in the octahedra. As one can observe the values obtained from doped structures are slightly shorter than those from pristine structures, except for Bi doping. In the trigonal structure doped with indium, the slight lattice expansion is related to the larger ionic radius of indium compared to antimony. Surprisingly, this effect is absent in the hexagonal structure, strongly suggesting a more effective accommodation of the dopant within the 0D hexagonal $P6_3/mmc$ polymorph. This is better illustrated by the wider distribution of the highest occupied state density in the hexagonal structure compared to the trigonal (*Figure C6*) and the smaller deviations on the octahedra InI_6 compared to SbI_6 in each polymorph (*Table C7* and C8, check $\boldsymbol{\Delta}a_{axis},$ $\boldsymbol{\Delta}a_{min}$ *and* $\boldsymbol{\Delta}a_{max}$). Except for Bi doping, which exhibits a larger ionic radius, and the unexpected cases of Nb and Sc doping in the 0D polymorph, the lattice parameter *c* generally decreases for most dopants. These results can be attributed to the larger rotations that the NbI_6 and ScI_6 octahedra present in these structures (*Table C8*, check $\Delta\varphi$, $\Delta\theta$ and $\Delta\psi$) and their larger M–X–M angles (*Table C9*).

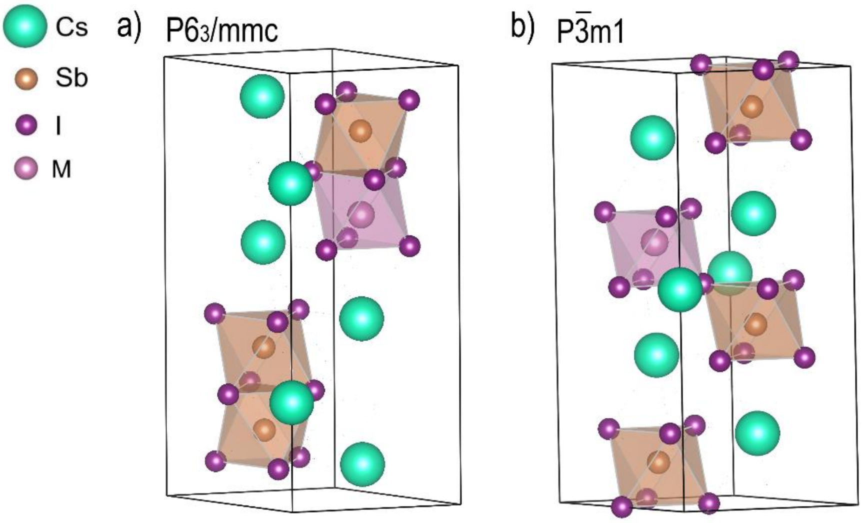


Figure 31 – Generic structures of (a) $Cs_3Sb_2I_9$ ($P6_3/mmc$), and (b) $Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with metal (M).

Table 8 – Distortion of lattice parameters according to the metal doping in the structures and average bond distance for the halogen-dopant bond, D_{I-M}.

Lattice and M atom	Lattice parameters (Å)		Lattice deviations (Å)		D_{I-M} (Å)
	a, b	c	$\Delta_{a,b}$ (%)	Δ_c (%)	
$Cs_3Sb_2I_9$ ($P6_3/mmc$)					
Pristine (Sb)	8.543	21.642	–	–	3.050
Ag	8.496	21.721	−1.21	−0.52	2.995
In	8.554	21.713	−0.51	−0.48	3.009
Mo	8.479	21.604	−1.09	−0.99	2.848
Nb	8.476	21.846	−0.93	0.74	2.895
Sc	8.525	21.706	−0.72	0.55	2.951
Bi	8.664	21.840	0.49	2.80	3.118
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)					
Pristine (Sb)	8.622	21.246	–	–	3.009
Ag	8.522	21.027	−1.16	−1.03	2.970
In	8.629	21.175	0.08	−0.33	2.984
Mo	8.585	21.115	−0.42	−0.62	2.836
Nb	8.595	21.151	−0.31	−0.45	2.880
Sc	8.619	21.162	−0.02	−0.40	2.921
Bi	8.678	21.373	0.65	0.59	3.047

Figure 32 presents the projected density of states for the doped $Cs_3Sb_2I_9$ structures in the $P6_3/mmc$ and $P\bar{3}m1$ polymorphs. The band structures for each of the doped $Cs_3Sb_2I_9$ structures in the $P6_3/mmc$ and $P\bar{3}m1$ polymorphs are shown in Figure C10 and *C11*, respectively.

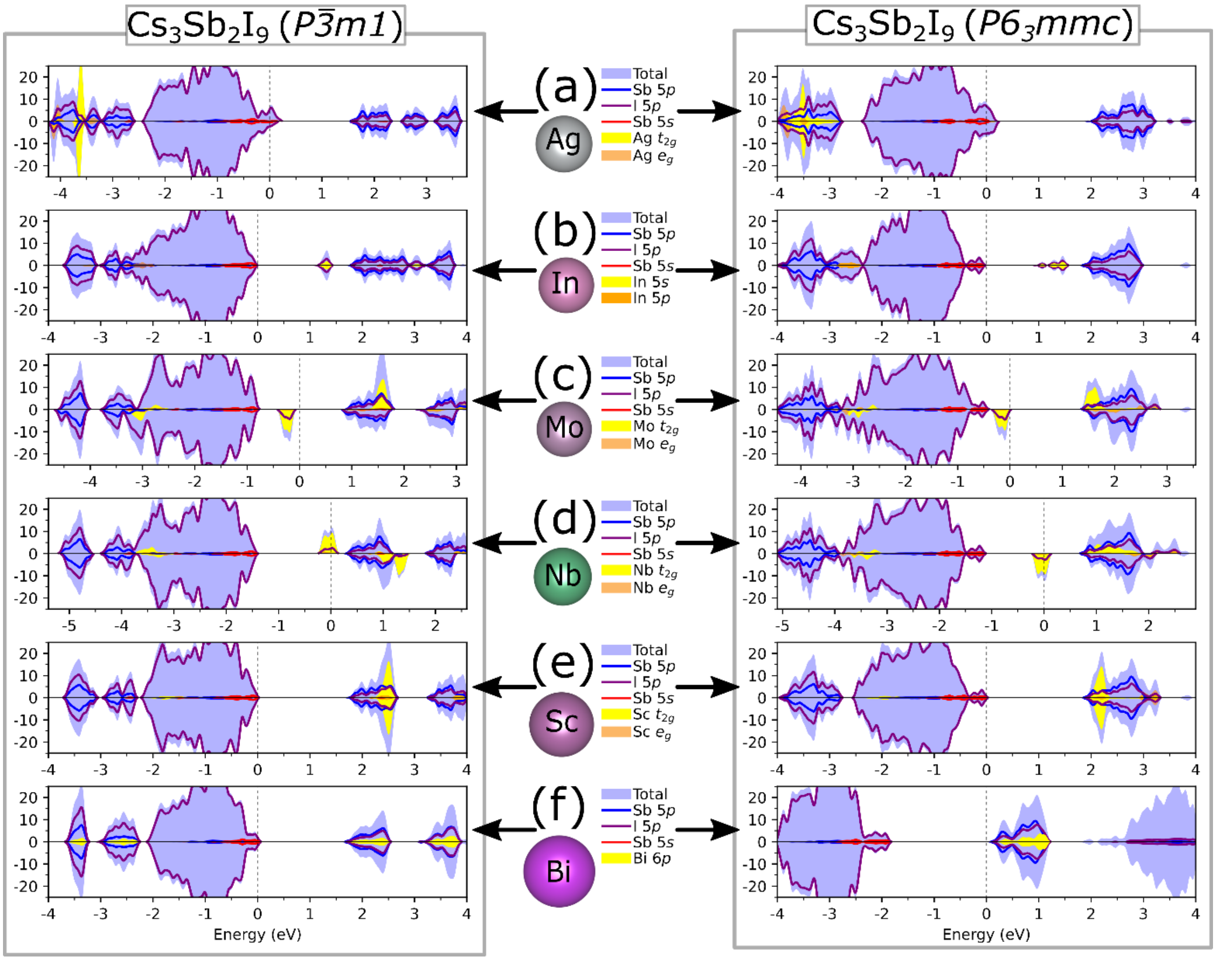


Figure 32 – Projected density of states for $Cs_3Sb_2I_9$ ($P\bar{3}m1$) (left) and $Cs_3Sb_2I_9$ ($P6_3/mmc$) (right) doped with (a) Ag, (b) In, (c) Mo, (d) Nb, (e) Sc, and (f) Bi, respectively. Fermi level at 0 eV.

- Ag doping: the difference in band structures in $Cs_3Sb_{1.5}Ag_{0.5}I_9$ is attributed to the interaction between Ag $4d_{xz}$ orbitals and I 5*p* orbitals at the band edge in $P\bar{3}m1$ structure, while $P6_3/mmc$ shows Ag $4dz^2$ orbitals along with I 5*p* orbitals in the valence band maximum (VBM). The distinct orbital contributions can be clearly seen in the highest occupied state density in *Figure C6*. Ag doping decreases the band gap in both polymorphs through the involvement of partially filled Ag 4*d* and I 5*p* orbitals at the top of valence band, creating a slight imbalance in up and down spin populations. This introduces unoccupied states, resulting in metallic behavior. The Ag doping of the

$Cs_3Sb_2I_9$ $P6_3$/mmc polymorph, results in decreased band gap with a direct gap (Γ–Γ), making it promising candidate for optoelectronic applications.

● In doping: doping with In maintains well-defined band gaps, with I 5*p* and In 5*s* orbitals in the conduction band and I 5*p* and Sb 5*s* orbitals in the valence band, resulting in a reduction of the band gap and the preservation of Sb 5*s* orbitals at the top of the valence band. This is an interesting observation, as indium is the only element among the selected dopants with a partially filled *p* orbital in the valence level (similar to Pb, Sb, and Bi). $Cs_3Sb_{1.5}In_{0.5}I_9$ shows a significant decrease in the band gap, particularly in the $P6_3$/mmc polymorph, where the band with the highest In contribution is more dispersed compared to the trigonal $P\overline{3}m1$ polymorph. This is evident in the highest occupied state density (*Figure C6*), with $P\overline{3}m1$ exhibiting localized density within the InI_6 layer, while In-doped $P6_3$/mmc perovskites show a more pronounced spread to the SbI_6 layers. Indium doping in both polymorphs leads to a more indirect transition, which can be attributed to the distinct electronic configuration of indium ([Kr] $5s^2$ $4d^{10}$ $5p^1$), introducing $5s^2$ orbitals in the conduction band, while the pristine antimony perovskite primarily relies on overlapping *p* orbitals in the conduction band.

● Mo and Nb doping: Mo doping results in a fully occupied mid-gap state composed of I 5*p* and Mo 4*d* orbitals and the conduction band is composed of Mo 4*d*, I 5*p*, and Sb 5*p* orbitals. Nb-doped structures present similar metallic character of the Ag-doped structure due to a partially filled mid-gap state formed by Nb 4*d* and I 5*p* orbitals. It can also be noticed in the Mo and Nb-doped band structures that mid-gap states shift more towards the conduction band in the $P\overline{3}m1$ polymorph than in the $P6_3$/mmc perovskite, this can be understood from the formation of overlapping Sb-M states in the $P6_3$/mmc structure, which can be seen in the highest occupied charge density of these structures in *Figure C6*. Therefore, due to a favorable orientation of octahedra in the 0D structure, which allows Sb–M overlapping, there is a lowering of the energy of the mid-gap states in comparison to the layered 2D structure. The presence of electronic structure with mid-gap states in Mo- and Nb-doped $Cs_3Sb_2I_9$ suggests that the dopants may act as recombination centers and their particular d-d transition can be exploited for luminescence applications as has been done recently with Mn-doped $Cs_3Sb_2Cl_9$ (X. Wang et al. 2020). Additionally, a quite localized spin polarization for these structures is shown in *Figure C4*, which may potentially be manipulated to create spintronic devices or for catalysis.

● Sc and Bi doping: doping with Sc and Bi, due to their isoelectronic valence to Sb, does not alter substantially the density of states and only an increase in band gap is observed. Sc and Bi states are mostly concentrated in the conduction band, preserving pristine transitions between valence and conduction band. Shifts towards higher energy levels of Sc and Bi dopant states can also be seen in the trigonal structure compared to the hexagonal one, which is again attributed to favorable M–Sb interaction enabled by the 0D structure. The Sc-doped structure also presented a narrowing in the difference between the direct and indirect gap in the hexagonal structure, favoring a direct transition. The more direct band gap in Sc doped structure is explained by the electronic configuration of [Ar] $4s^2$ $3d^1$ in Sc that introduces 3*d* states just above the CBM mediating transitions from the VBM.

Table 9 presents a comparison of effective masses in the $k_{[001]}$ direction between doped and pristine structures. A detailed discussion of these results is provided in the *Supporting Information (Appendix C.2.3)*. The doped structures exhibited increased effective electron masses in the $k_{[001]}$ direction for all dopants except Bi, which showed a slight decrease. The most significant increase in electron effective masses was observed in the In- and Mo-doped $P6_3/mmc$ and In- and Nb-doped $P\bar{3}m1$ structures. The $Cs_3Sb_2I_9$ doped structures generally maintained similar hole effective masses, except for Ag and Mo-doped structures due to the introduction of localized *d* states in the valence band. For Sc and Bi-doped structures, effective masses resembled those of the pristine hexagonal polymorph, but there was a significant increase in hole masses in the trigonal polymorph. This effect was particularly pronounced in the Bi-doped $Cs_3Sb_{1.5}Bi_{0.5}I_9$ $P\bar{3}m1$ due to the overlap of Bi 6*s* states in the valence band.

Table 9 – Calculated effective masses, direct and indirect band gap for the metal-doped and pristine $Cs_3Sb_2I_9$ polymorphs.

Structure	**Effective mass (m*) k[100] direction**		**Indirect gap (eV)**	**Direct gap (eV)**	**Δ(direct-indirect) gap**
	electrons	***holes***			
$Cs_3Sb_2I_9$ ($P6_3/mmc$)					
Pristine (Sb)	0.32	1.10	1.81 (**K*–Γ**)	1.94 (**M–M**)	0.13
Ag	0.38	-	-	1.67 (**Γ–Γ**)	-
In	0.82	1.18	0.96 (**K–Γ**)	1.38 (**Γ–Γ**)	0.42
Mo	0.61	-	1.32 (**L–Γ**)	1.38 (**M–M**)	0.06
Nb	0.46	1.23	0.51 (**K–Γ**)	0.57 (**M–M**)	0.06
Sc	0.44	1.08	1.86 (**K–Γ**)	1.94 (**M–M**)	0.08
Bi	0.29	1.35	1.94 (**K–Γ**)	2.10 (**M–M**)	0.16
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)					
Pristine (Sb)	0.31	1.09	1.52 (**K*–Γ**)	1.54 (**Γ–Γ**)	0.02
Ag	0.36	1.46	1.27 (**K*–A**)[a]	1.38 (**A–A**)[a]	0.11
In	0.72	1.03	1.14 (**K*–Γ**)	1.25 (**A–A**)	0.11
Mo	0.33	1.94	1.01 (**L–H**)	1.07 (**H–H**)	0.06
Nb	0.94	1.63	1.43 (**K*– Γ**)[a]	1.54 (**Γ–Γ**)[a]	0.11
Sc	0.42	1.36	1.77 (**K*–A**)	1.84 (**A–A**)	0.07
Bi	0.29	-	1.62 (**K*–Γ**)	1.65 (**Γ–Γ**)	0.03

***** K* is a point in the K - Γ high-symmetry line. [a] the structure is metallic, the value shown is estimated considering a slight change in the Fermi level.

It is important to remark that while the Ag and Nb doped structures may not be ideal for optoelectronic applications due to their predicted metallic nature, it is crucial to note that the presence of intrinsic defects, which were not considered in this study, can alter the Fermi level. Consequently, in an experimental setting, the semiconductor behavior could potentially be restored. Therefore, investigating these compounds remains relevant and worthwhile (Freysoldt et al. 2014).

The calculated formation energies for the insertion of dopant atoms into the structures, along with the average Bader charge of iodine atoms (bonded to M) and doping metals (M), are presented in *Table 10*. A trend of increasing formation energy can be observed; Sc < In < Nb < Ag < Mo. It is worth noting that for Sc and In, the obtained energy value suggests a spontaneous substitution of Sb. Considering that dopants are bonded to halogens, charge transferring to the halogen is expected and would stabilize the structure more effectively, reducing the formation energy. However, the descending order of charge donation for the halogen is as follows: Sc > Nb > In > Mo > Ag. Despite Nb donating more charge to the halogen than In, Nb exhibits higher formation energy, and a similar trend is observed for Mo and Ag. Therefore, factors beyond charge donation, such as the presence of unpaired electrons, can influence the formation energy. This is likely the case for Nb, Mo, and Ag, which have highest formation energies and unpaired electrons. Additional insights about the formation energies obtained from charge density plots and a multilinear regression model confirm the importance of charge transfer ability of the dopant. These findings are detailed in the *Supporting Information (Appendices C.2.1 and C.2.4)*.

The introduction of Ag, In, Sc, and Bi in the dimer polymorph leads to a decrease in formation energy compared to the layered polymorph. This result can be attributed to better coordination with the halogen atoms and a structure that can withstand more deformation. Conversely, in the case of Mo- and Nb-doped perovskites, the formation energy increases in the dimer polymorph. This can be attributed to the destabilization of the adjacent SbI_6 octahedra by the high-coordination metals.

Table 10 – Formation energies for metal doping ($E_f[D]$), average Bader charge of iodine in the dopant-halogen bonds and average Bader charge for the dopant (M).

Lattice and M atom	***$E_f[D]$* (eV)**	**Avg. Bader charge (I–M)**	**Avg. Bader charge (M)**
$Cs_3Sb_2I_9$ ($P6_3/mmc$)			
Pristine (Sb)	-	-0.492	0.947
Ag	1.929	-0.411	0.364
In	-0.492	-0.522	1.058
Mo	2.050	-0.513	1.025
Nb	0.598	-0.553	1.313
Sc	-3.627	-0.644	1.719
Bi	-0.551	-0.526	1.098
$Cs_3Sb_2I_9$ ($P\overline{3}m1$)			
Pristine (Sb)	-	-0.496	0.944
Ag	1.995	-0.424	0.322
In	-0.245	-0.538	1.058
Mo	1.218	-0.539	1.010
Nb	0.407	-0.581	1.287
Sc	-3.345	-0.646	1.675
Bi	-0.476	-0.516	1.041

To assess the thermodynamic stability of these materials and determine their potential for experimental synthesis, we performed a convex hull distance calculation for each composition (Barber, Dobkin, and Huhdanpaa 1996). The corresponding convex hull compounds to each doped structure stoichiometry was obtained from Open Quantum Materials Database (OQMD) and the grand canonical linear programming (GCLP) method (Saal et al. 2013; Kirklin et al. 2015). The algorithm considers energies of all phases in the Cs–Sb–I–M quaternary phase diagram at 0 K and zero pressure, providing the most stable linear combination of phases for a given stoichiometry. By comparing the total energy of the doped structures to those of the competing phases in the convex hull, we obtained the convex hull energies (E_{hull}) which are used to evaluate stability of the doped structure, $Cs_3Sb_{1.5}M_{0.5}I_9$ using the following formula for the convex hull distance or stability, E_{stab}:

$$E_{stab} = E_{\mathrm{Cs_3Sb_{1.5}M_{0.5}I_9}} - E_{hull} \tag{48}$$

A “negative” convex hull distance indicates stability, suggesting a spontaneous formation of the considered compounds. Alternatively, a slightly looser definition (Emery and Wolverton 2017) considers E_{stab} below 25 meV per atom (approximately kT at room temperature) as viable accounting for nearly-stable structures and possible uncertainties/errors associated with DFT.

Convex hull stability trends closely align with those observed for formation energies as shown in *Table 11*. The greater stability of the doped hexagonal structures compared to their trigonal counterparts is not solely due to $Cs_3Sb_2I_9$ ($P\bar{3}m1$) having a 5.6 meV/atom higher formation energy than $Cs_3Sb_2I_9$ ($P6_3/mmc$). The atomic arrangement of the hexagonal structure promotes more favorable bonding, indicated by larger binding energies (Table C4), and greater flexibility, as seen when the lattice parameters of the two polymorphs are compared (*Table 8*). As one can observe, Ag doped is the only structure exceeding the stability threshold of 25 meV, indicating that it should not be easily synthesized unless an appropriate defect or dopant is introduced during the synthesis. On the other hand, Sc and Bi doped structures exhibited negative convex hull distance, indicating higher stability. Therefore, Sc and Bi may be considered as dopants to improve the resistance of degradation and reduce the Urbach energy of these compounds, which is currently an issue (Chonamada, Dey, and Santra 2020). In the case of Sc doping, the greater stability should not decrease effective masses substantially and the band gap may become less indirect. Moreover, co-doping Sc and Bi along with the other studied dopants (In, Nb, Mo, and Ag) may offer ways to further stabilize these compounds, resulting in crystals with smaller band gaps.

Table 11 – Convex hull compositions and distance of the convex (E_{stab}) hull for each considered dopant in both polymorphs of $Cs_3Sb_2I_9$.

Dopant	Convex hull composition	E_{stab} (meV/atom)	
		$Cs_3Sb_2I_9$ ($P6_3/mmc$)	$Cs_3Sb_2I_9$ ($P\bar{3}m1$)
-	$Cs_3Sb_2I_9$ ($P6_3/mmc$)	0	5.58
Ag	0.25 Cs_2AgI_3 + 0.75 AgI + 1 CsI_3 + 1.5 $Cs_3Sb_2I_9$	31.24	38.44
In	0.5 CsI + 1 $CsInI_4$ + 1.5 $Cs_3Sb_2I_9$	7.04	9.58
Mo	1 CsI + 0.5 CsI_3 + 1.5 $Cs_3Sb_2I_9$ + 1 MoI_2	19.97	24.96
Nb	0.9 Cs_2NbI_6 + 1.4 $Cs_3Sb_2I_9$ + 0.1 $NbSb_2$	13.68	18.78
Sc	1.5 CsI + 1.5 $Cs_3Sb_2I_9$ + 1 ScI_3	-5.03	-1.41
Bi	0.5 $Cs_3Bi_2I_9$ + 1.5 $Cs_3Sb_2I_9$	-0.97	2.80

The absorption coefficients for the pristine and doped structures for both polymorphs were calculated and are shown in *Figure 33*. Among the range of greatest interest for photovoltaic applications (between 1 and 2 eV approximately), the structure doped with In presented the highest values of absorption coefficient. In the same range, the second higher absorption is obtained from Ag doping, followed by Mo and Nb. On the other hand, absorption below 1 eV is observed for Mo and Nb doping, in agreement to their smaller band gaps. Due to mid-gap states and unfavorable Fermi level these structures may not be so efficient for charge carrier generation.

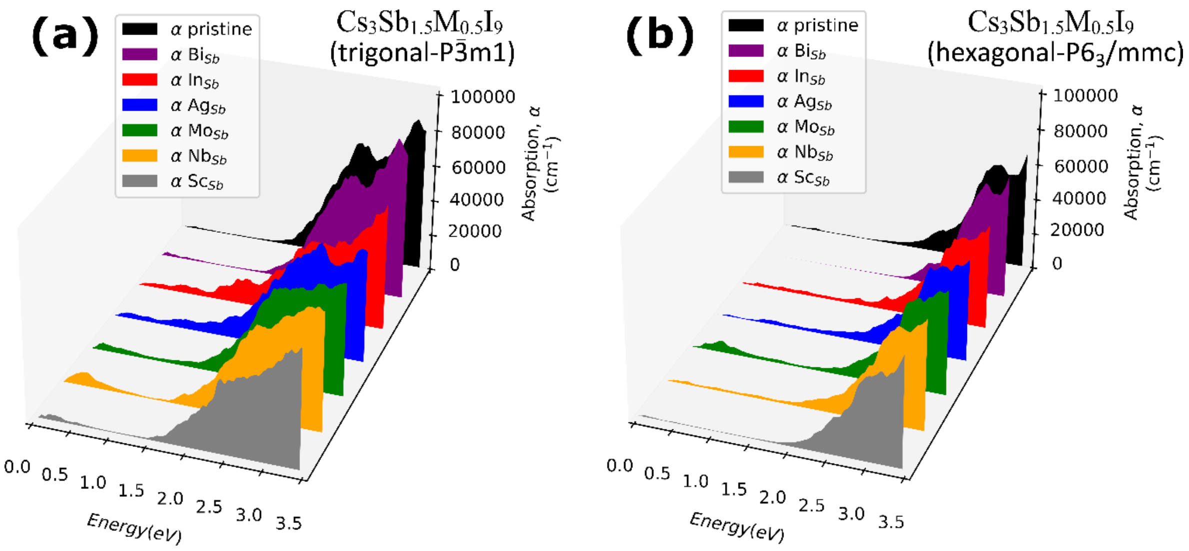


Figure 33 – Absorption coefficient for (a) $Cs_3Sb_2I_9$ ($P\bar{3}m1$), and (b) $Cs_3Sb_2I_9$ ($P6_3/mmc$) pristine and doped with Ag, In, Mo, Nb, and Sc, considering one unit cell and one dopant atom.

Taking into consideration the results discussed, the prospects of each doped structure are compiled as follows:

● Ag doping: Ag doping results in a lower band gap, enhancing absorption and making the gap direct in the dimer form. However, introducing Ag into the $Cs_3Sb_2I_9$ structure requires high energy. Therefore, Ag must be considered in co-doping with an element that further stabilizes the lattice such as In or Sc. Previous reports demonstrate the potential of Ag doping in inorganic halide perovskites. $Cs_2AgInBr_{6(1-x)}Cl_{6x}$ structured phases have been reported (Y. Liang 2021), while Ag^+ doping on $CsPbBr_3$ nanocrystals improves conductivity and charge-carrier mobility (Shu Zhou et al. 2019). In addition, Ag doping in $MAPbI_3$ have been found to enhance charge transport efficiency, reduce trap states, and improve PCE (Hao et al. 2021). These findings suggest the potential application of these doped structures in LEDs and photovoltaics.

● In doping: indium doped structures presented promising results for photovoltaic applications with improved absorption coefficients, a well-defined and small band gap, and small hole effective masses. This result is in alignment with literature observations, which predicts $Cs_3In_2I_9$ to be a very efficient light absorber with an optimal band gap of only 1.25 eV, although, to the best of our knowledge, no experimental results have been reported up to now for this material (W. H. Guo et al. 2020a). The 2D $Cs_3In_2I_9$ presents an ideal band structure for solid-state lighting with a direct band gap at the Γ point, strong electron dispersion at the CBM and weak hole dispersion at the VBM, and light (heavy) electron (hole) effective mass. The studied $Cs_3Sb_{1.5}In_{0.5}I_9$ shows potential for similar applications but has shown viability considering convex hull energy and is worth candidate for synthesis.

● Mo and Nb doping: with an increased light absorption in the visible range due to mid-gap states introduced by the dopants, Mo- and Nb-doped $Cs_3Sb_2I_9$ may be useful for luminescence applications. These dopants are better incorporated in the layered form which has better transport properties, and Nb substitution has lower formation energy. The d-d transition introduced by the dopants may yield bright sharp emissions as seen for similar materials (X. Wang et al. 2020). Additionally, there has been reports of Mo-doping in perovskites for photocatalysis and sensors (Z. Zhang et al. 2019; Kwak et al. 2019), Nb-doping in halide perovskites is also reported for photocatalysis (Z. Guo et al. 2019) and even for improved PCE and reduced hysteresis (Patil, Mali, and Hong 2020).

● Sc and Bi doping: Although these dopants do not lead to significant improvements in optical absorption in $Cs_3Sb_2I_9$ due to a slightly higher band gap, they can play a crucial role in stabilizing the lattice, particularly in the dimer phase. This is of great importance because reducing the Urbach energy is a key factor in approaching efficiencies closer to the theoretical estimates for these materials in photovoltaics (Singh et al. 2018). Reports on Sc-doping in perovskites are limited, however a few studies indicate its potential for enhancing electrochemical catalysis (Jeong et al. 2018; M. Xu et al. 2020).

Finally, ACBN0 calculations were then performed on both pristine and doped structures (In-doped and Sc-doped) for each polymorph, as illustrated in *Figure 34*. In-doped and Sc-doped structures were chosen based on their respective optical properties and enhanced stability. Notably, our results exhibit remarkable agreement with experimental band gaps for the pure structures, akin to findings from prior research where PBE+U was applied to $Cs_3Sb_2X_9$ (X = Cl, Br, I) structures (Gouvêa et al. 2024). In the case of In-doped structure the calculated band gap is 1.76 eV for the 2D structure and 1.47 eV for the 0D polymorph indicative of its promising potential for optoelectronic applications. Conversely, the band gap of the Sc-doped structure increases compared to pristine structures, aligning with the trends observed in PBE calculations. Furthermore, our analysis reveals no significant disparities in direct and indirect band gaps when employing the more accurate functional, as detailed in *Table C11*.

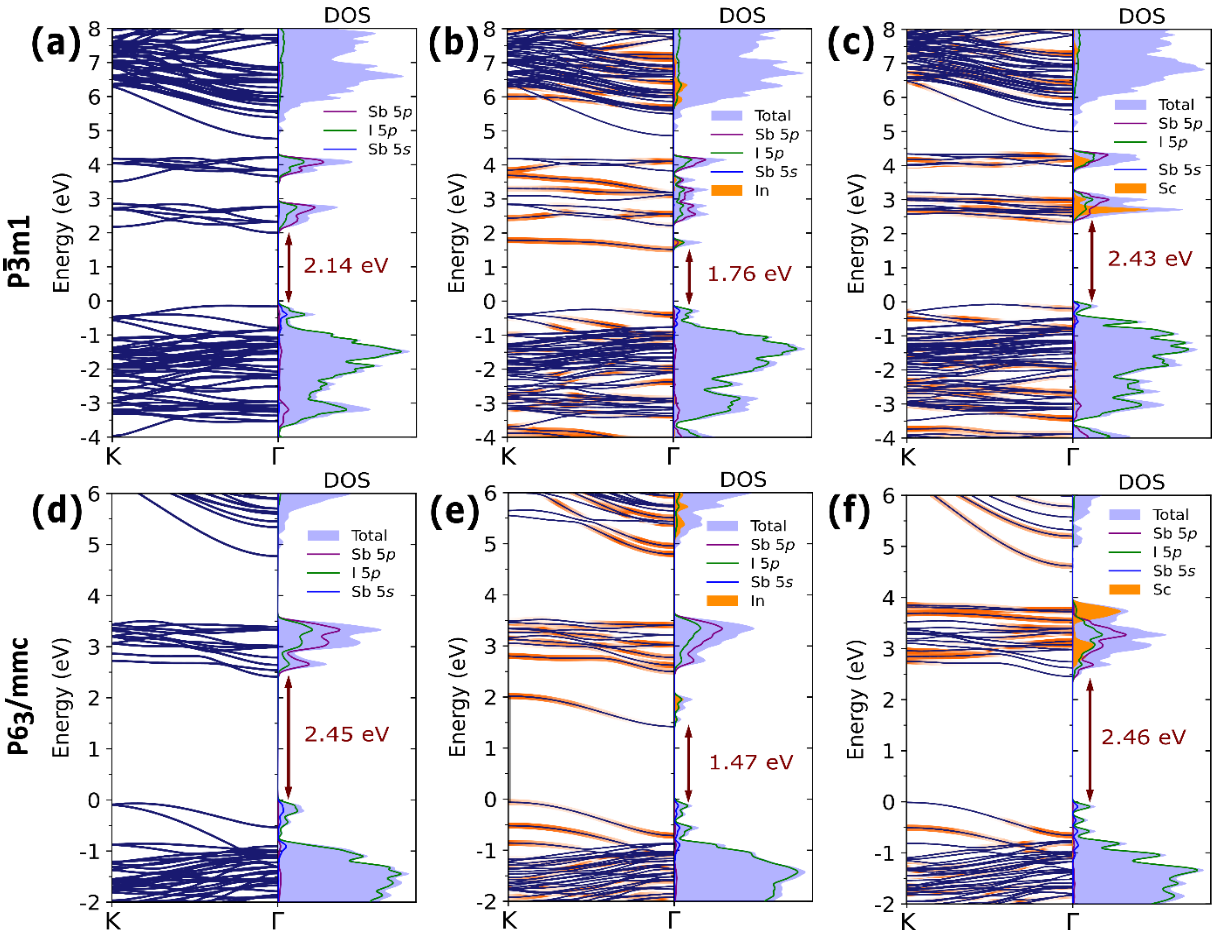


Figure 34 – Band structure and projected density of states for pristine, In- and Sc-doped $Cs_3Sb_2I_9$, obtained via the ACBN0 method. Panels (a-c) display the results for the P$\bar{3}$m1 polymorph, while panels (d-f) show those for the P6_3/mmc polymorph.

4.3.3 Doping with halogens

To the best of our knowledge, there are no previous reports on the effects of halogen doping in both polymorphs of $Cs_3Sb_2I_9$ while evaluating the differences between the structures. Since, homogeneous solid solutions have been reported for the layered polymorph (Ma et al. 2019), we deemed it appropriate to perform a 22% halogen substitution using chlorine and bromine. This concentration is sufficiently high to assess the differences but not so high that significant structural changes are likely to occur. Both polymorphs exhibit two distinct sites for halogen incorporation: a bridging site (3*e* in P$\bar{3}$m1, 6*h* in P6_3/mmc) and a terminal site (6*i* in P$\bar{3}$m1, 12*k* in P6_3/mmc). We analyzed the energy differences for substitutions in each of these sites (*Table C12*).

Our results indicate that the energy difference between the sites is lower than the thermal energy at room temperature (kT ≅ 25 meV) for bromine substitution in both structures. This suggests that bromine substitution should occur simultaneously in both sites for both structures. In the case of chlorine substitution, the energy difference is significant, particularly for the trigonal structure, favoring substitution in the terminal site. Experimental evidence has shown that in $Cs_3Sb_2Br_9$ ($P\bar{3}m1$), halogen substitutions tend to saturate the terminal sites first (Pradhan, Jena, and Samal 2022). Therefore, we have chosen to incorporate both terminal and bridging substitutions in the hexagonal form and only terminal substitutions in the trigonal structure (*Figure C14*).

Despite the relatively high levels of halogen doping, the percent deviation from the pristine lattice parameters, due to the introduction of smaller and more electronegative elements, was found to be less than 2.0%, as shown in *Table 12*. Furthermore, halogen doping was found to induce a symmetry breaking in $Cs_3Sb_2I_9$ by promoting *a ≠ b*.

Table 12 – Distortion of the lattice parameters for the structures with halogen doping.

Structure	Lattice parameters (Å)			Lattice deviations (Å)		
	a	b	c	Δ_a (%)	Δ_b (%)	Δ_c (%)
$Cs_3Sb_2I_9$ ($P6_3/mmc$)						
Pristine	8.543	8.543	21.642	-	-	-
$Cs_3Sb_2Br_2I_7$	8.494	8.470	21.365	-0.57	-0.85	-1.28
$Cs_3Sb_2Cl_2I_7$	8.499	8.378	21.342	-0.51	-1.92	-1.39
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)						
Pristine	8.622	8.622	21.246	-	-	-
$Cs_3Sb_2Br_2I_7$	8.577	8.542	21.120	-0.52	-0.92	-0.59
$Cs_3Sb_2Cl_2I_7$	8.541	8.481	21.019	-0.93	-1.63	-1.07

As observed in the projected density of states for the structures with halogen doping (*Figure C15 and B15*), the conduction band is mainly formed by the orbitals I 5*p*, Sb 5*p* in addition to contributions from the dopants Cl 3*p* or Br 4*p*. The valence band is composed of orbitals I 5*p*, Sb 5*s*, and Cl 3*p* or Br 4*p* according to the respective doping. In the band structures, presented in *Figure 35*, one can observe that the halogen doping leads to an increase in band gap energy, a trend that is particularly evident in the Cl-doped structures. This results was already expected since

1D orthorhombic $Cs_3Sb_2Cl_9$ (Pmcn) and 2D $Cs_3Sb_2Cl_9$ (P$\bar{3}$m1) present band gaps of ~ 3.4 eV and ~ 3 eV, respectively (B. Pradhan et al. 2018; Vargas et al. 2017), explaining also a larger relative increase of the band gap for the dimer perovskite which has lower dimensionality. Moreover, more electronegative halogens tend to form lower energy VB states widening the band gap. In the case of Br doping, the impact on the band dispersion is milder since its size and electronegativity are closer to that of iodine.

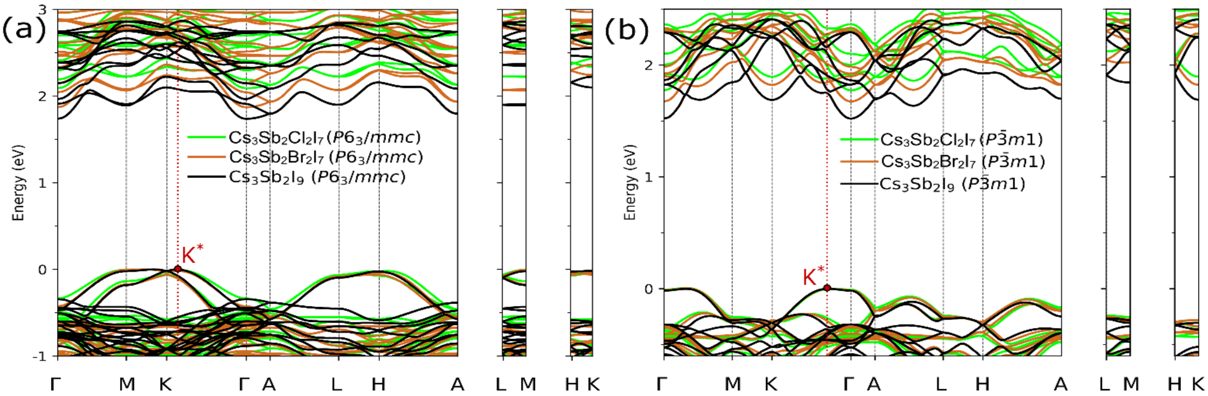


Figure 35 – Band structures for (a) $Cs_3Sb_2I_9$ ($P6_3/mmc$), and (b) $Cs_3Sb_2I_9$ (P$\bar{3}$m1) doped with Cl, and Br. Fermi level at 0 eV.

Table 13 shows the transitions responsible for the direct and indirect gap in each structure, considering the influence of the doping in comparison to the pristine structures. In all cases, the lowest energy transitions are indirect, with the VBM localized in a reciprocal space point, represented by K*, in the high-symmetry line K–Γ. Surprisingly, Br doping in the dimer structure results in a more indirect gap compared to Cl doping. This behavior can be associated with the splitting of the halide p-states, which has been linked to indirect-direct transition in $Cs_3Sb_2Cl_{9-x}Br_x$ in a recent study (Pradhan, Jena, and Samal 2022). This idea is further supported by the observation of energy level splitting in the high and low mass bands of $Cs_3Sb_2I_9$ ($P6_3/mmc$) at the K point and at the segment L–H. Conversely, the direct-indirect transitions in the layered polymorph remain consistent, with no relative changes, regardless of the dopants.

There is an increase in hole effective mass with doping, especially for Cl-doped structure as shown in *Table 13*. The Br-doped structure presents a more dispersed band structure than the Cl-doped structure and in a few cases a slightly lower effective mass is seen for Br-doped structures compared to pristine for electron and hole effective masses in the $k_{[001]}$ of the hexagonal polymorph and for the electron and hole effective mass in the $k_{[100]}$ of the trigonal structure. These can be associated with a splitting of the halide states caused by the Br incorporation.

Our findings suggest that despite the substantial contractions observed in the lattice parameters (*Table 12*), the deviations in effective masses remain minimal. Therefore, the halogen-doped structures hold promise for incorporating additional dopants, such as Bi, which tend to expand the unit cell, without causing significant disruptions to transport properties.

Table 13 – Calculated effective masses, direct and indirect band gap for halogen-doped $Cs_3Sb_2I_9$ polymorphs compared to the pristine structures.

Structure	Effective mass (m*)				Indirect E_g (eV)	Direct E_g (eV)	ΔE_g (direct-indirect)
	electrons		holes				
	$k_{[100]}$	$k_{[001]}$	$k_{[100]}$	$k_{[001]}$			
$Cs_3Sb_2I_9$ ($P6_3/mmc$)							
Pristine	0.32	1.33	1.10	1.05	1.81 (**K*–Γ**)	1.94 (**M–M**)	0.13
$Cs_3Sb_2Br_2I_7$	0.39	1.20	1.16	0.96	1.95 (**K*–Γ**)	2.13 (**M–M**)	0.18
$Cs_3Sb_2Cl_2I_7$	0.39	1.88	1.44	1.31	2.18 (**K*–Γ**)	2.31 (**M–M**)	0.13
$Cs_3Sb_2I_9$ ($P\overline{3}m1$)							
Pristine	0.31	0.40	0.80	0.33	1.52 (**K*–Γ**)	1.53 (**Γ–Γ**)	0.01
$Cs_3Sb_2Br_2I_7$	0.30	0.51	0.79	0.41	1.67 (**K*–Γ**)	1.68 (**Γ–Γ**)	0.01
$Cs_3Sb_2Cl_2I_7$	0.32	0.63	0.87	0.48	1.76 (**K*–Γ**)	1.77 (**Γ–Γ**)	0.01

The analysis of the calculated Bader charges demonstrated that Br^- and Cl^- deplete more electronic charge from Sb as compared to I^- (see Table C13). Namely, a shorter Sb–X bond distance and increased structural distortion is observed when doping with Cl^- (density plots in Figure C17 and C17). Further discussion over the charge analysis can be found in the Appendix C.3. Finally, the optical absorption of the halogen-doped $Cs_3Sb_2I_9$ polymorphs was also calculated to evaluate their differences, as illustrated in Figure C19. Other than the expected shift in the absorption edge due to variations in the band gap, there was a slight decrease in the absorption coefficient in the visible-UV range with halogen doping in the 2D perovskite compared to the pristine form. This decrease was slightly more apparent for the 0D perovskite in the UV region, where there was also a small shift towards higher energies.

Our findings suggest that the more electronegative halogens induce changes in band gap energy, band gap transition, and effective masses in the dimer polymorph, while the impact on the layered form is comparatively less pronounced. Despite the substantial contractions observed in the lattice parameters with halogen doping, the deviations in effective masses remain minimal. Therefore, coupled with the negative formation energy for the substitution, Cl and Br-doped $Cs_3Sb_2I_9$ hold promise for incorporating additional dopants which tend to expand the unit cell (such as Bi) without causing significant disruptions to transport and optical properties.

4.4 CONCLUSION

Indium-doped structures demonstrated well-defined small band gaps and the highest absorption coefficient within the desired range of interest for photovoltaic applications, in addition to small hole effective masses, and stability attested by the convex hull distance. Silver doping led to a lower band gap, direct band gap in the dimer structure and enhanced absorption, making it a candidate for LED and photovoltaic applications. Doping with molybdenum and niobium increased light absorption in the visible range due to the introduction of mid-gap states, suggesting potential use in luminescence applications. Scandium and bismuth doping played a crucial role in stabilizing the lattice, although optical absorption slightly decreased. Sc-doping may aid in reducing the high Urbach energy of these materials, thereby improving their performance, particularly in photovoltaics. The $P6_3/mmc$ structure demonstrated superior flexibility and stability for dopant incorporation, and its high band gap and indirect transition could be addressed through doping, highlighting the potential of this often-overlooked polymorph. Halogen doping, as evaluated in this study, had a more pronounced impact on the dimer polymorph, resulting in increased band gaps but also improving structure stability. The changes in effective mass caused by halogen doping were small, accompanied by a contraction of lattice parameters that could facilitate the incorporation of larger dopants like bismuth. This report sheds light on the potential of lead-free $Cs_3Sb_2I_9$ polymorphs for a broad range of applications. It emphasizes the opportunities for improvement in these materials through fine-tuning their properties via doping.

CHAPTER 5 — Boosting feature-based machine learning models for materials science: encoding descriptors and graph-based features for enhanced accuracy and faster featurization in MODNet

5.1 RESEARCH PROBLEM

Feature-based models, representing materials by descriptors crafted based on empirical knowledge, and graph-based models, using graph representations of the structure leveraging the power of graph neural networks (GNNs) are popular choices for machine learning in materials science (Choudhary et al. 2022; De Breuck, Hautier, and Rignanese 2021; Zhang et al. 2022). It has been shown for molecular properties that descriptor-based models can perform comparably well to GNNs in various chemical endpoints (Jiang et al. 2021). Arguably, feature-based models represent the most interpretable approach, guiding researchers in developing design principles to optimize a given property (or set of properties) and remaining effective even with small datasets.

General-purpose feature-based algorithms, such as MODNet (De Breuck, Hautier, and Rignanese 2021) and Automatminer (Dunn et al. 2020), utilize an extensive set of features from suites like MatMiner (Ward et al. 2018). Unfortunately, *general electronic structure featurizers* (Tawfik and Russo 2022) like Orbital Field Matrix (OFM) and Smooth Overlap of Atomic Positions (SOAP) (Pham et al. 2017; Bartók, Kondor, and Csányi 2013), despite enhancing predictions in ML models, including GNNs (X. Yang 2022), can be challenging to apply for large-scale screening. Their featurization is time-consuming and introduces numerous descriptors, complicating their application alongside other featurizers due to the curse of dimensionality. On the other hand, graph-based models achieve the highest accuracies when a sufficiently large dataset (> 10,000) is provided (Dunn 2024). Due to these limitations, graph-based models perform better when screening large chemical spaces, as is typically the case in materials discovery tasks.

In this context, we propose employing graph-based models as auxiliary tools to promptly generate features that complement feature-based models, aiming to enhance accuracy. This involves harnessing pre-trained graph-based models, leveraging their knowledge acquired from properties with large available datasets, and utilizing a graph-based model trained to generate a latent-space version of general electronic structure descriptors. This latter approach enables us to simultaneously avoid lengthy featurization and the introduction of numerous descriptors from these featurizers. Utilizing feature-based models offers the crucial benefit of interpretability, which can be efficiently harnessed through the SHAP method (Lundberg and Lee 2017) to explore and advance the domain knowledge in materials science. Moreover, since feature-based models can readily offer interpretable design rules, they are highly suitable for guiding active learning for materials discovery with statistically robust features, a challenge currently being addressed in the field (Choubisa et al. 2023).

We observed a significant boost in accuracy in predictions for the task of heat of formation in perovskites utilizing the new features, approaching the benchmarked values for GNN models. The strategy holds up when it is further tested for generalizability in more difficult tasks, such as predicting band gap and stability for the subset of compounds in the Open Quantum Materials Database (OQMD) that contain halogens. Chemical insights are drawn for each of these tasks, and the role of the different included features in the proposed approach is thoroughly investigated.

5.2 METHODOLOGY

To harness the power and speed of GNNs into feature-based models we propose the construction of GNN featurizers following three general approaches:

1. Compress the patterns of general electronic structure descriptors such as OFM and SOAP with an autoencoder (definition on *Appendix A.12*), a GNN can then be trained to promptly produce the encoded representation avoiding the expensive featurization.
2. Integrate pre-trained GNN models trained on properties with abundant datasets (formation energy, band gaps, etc.) into the feature-based model. The model is truncated to extract values from a hidden layer of the final MLP instead of the target property. These values serve as descriptors in the feature dataset.
3. Train an adjacent GNN model on the training data, again take the values from a hidden layer of the final MLP and use as input for the feature-based model.

The general process is depicted in *Figure 36*. The procedures were implemented in MODNet extending the default MatMiner featurizer included in MODNet v.0.1.13 from the original MODNet publication (De Breuck, Hautier, and

Rignanese 2021). OFM was chosen as the general electronic structure featurizer, and MEGNet was the preferred GNN framework for the investigation. This newly extended featurizer was named the OMEGA featurizer, standing for “encoded OFM + pre-trained MEGNet + Adjacent MEGNet models”. This procedure is versatile, and alternative implementations could have been considered, such as using SOAP instead of OFM and more recent GNNs like ALIGNN (Choudhary and DeCost 2021). Furthermore, an alternative version of the featurizer, named the OMEGAfast featurizer, also offers a GNN model to quickly obtain a latent-space representation of the default MatMiner features, thereby further streamlining the featurization process.

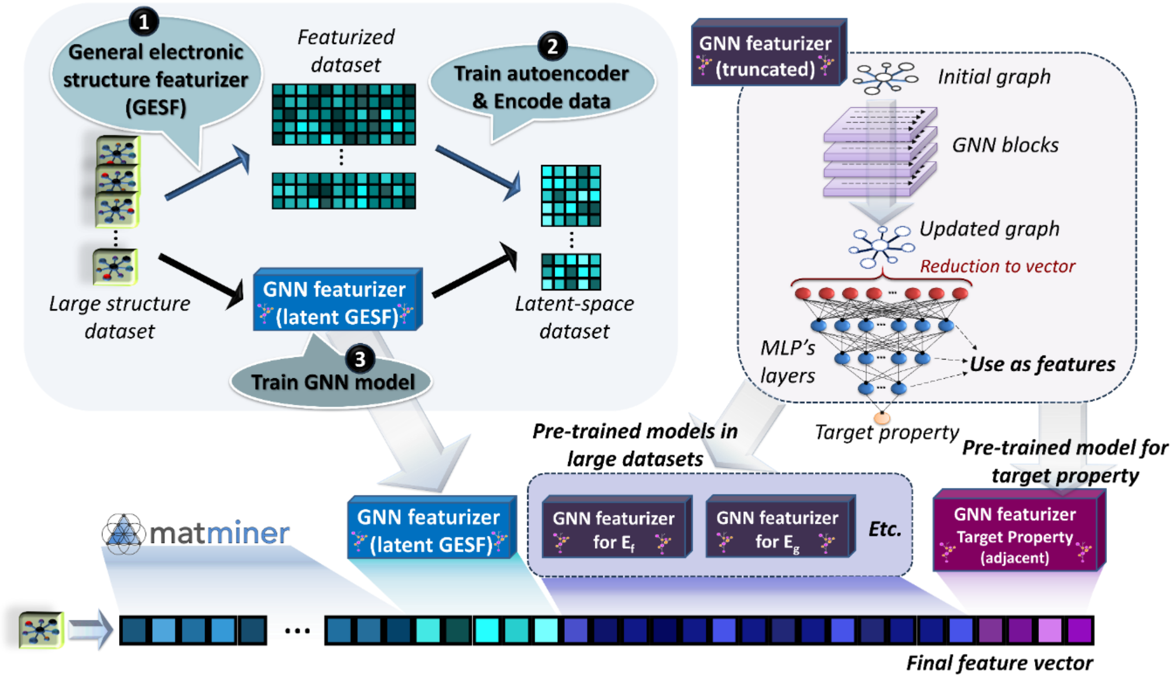

1
General electronic structure featurizer (GESF)
Featurized dataset
2
Train autoencoder & Encode data
Large structure dataset
GNN featurizer (latent GESF)
Latent-space dataset
3
Train GNN model
GNN featurizer (truncated)
Initial graph
GNN blocks
Updated graph
Reduction to vector
MLP’s layers
Use as features
Target property
Pre-trained models in large datasets
Pre-trained model for target property
matminer
GNN featurizer (latent GESF)
GNN featurizer for E_f
GNN featurizer for E_g
Etc.
GNN featurizer Target Property (adjacent)
Final feature vector

Figure 36 – Illustration of the proposed procedure to harness GNN models as fast featurizers for general electronic structure descriptors. Information from pre-trained GNN models on larger datasets or on the training data can also be leveraged by taking the values on hidden layers before the target as features.

The dataset of formation energy for perovskites named `matbench_perovskites` *(MatBench v.0.1)* was used as a proof-of-concept for all tests to implement the new featurizers. This dataset was selected because it represents the smallest benchmark task, comprising 18,928 samples, where the deviation between graph-based and feature-based models becomes significant. The benchmarked mean absolute errors (MAE) for MEGNet, MODNet, AutoMatMiner and RF/SCM-MagPie on this task are shown in *Table 14*.

Table 14 – Mean absolute errors for MatBench task of heat of formation of perovskites (`matbench_perovskites`) with different algorithms. Source: (Dunn 2024)

General Purpose Algorithm	MAE on task `matbench_perovskites` (eV)
MEGNet	$0.0352\ (\pm 0.0016)$
MODNet	$0.0908\ (\pm 0.0028)$
AutoMatMiner	$0.2005\ (\pm 0.0085)$
RF-SCM/MagPie	$0.2355(\pm 0.0034)$

Our approach was to compare results of the default MatMiner featurizer on MODNet on this task with the implementations of additional features and substituting the MatMiner features with latent-space representations. We evaluated the effects of different latent-space sizes and considered reconstruction loss on the latent-space representation. The following outlines the detailed implementation and evaluation procedures for each of the GNN featurizers:

● **OFM-encoded GNN featurizer:** The OFM featurizer captures valence electron interactions at each site of the structure by employing a weighted vector outer product of one-hot encoded valence orbitals for every atom (further details can be found in *Appendix D.1* Orbital field matrix featurizer). To represent the entire structure, the average of all local OFMs is calculated. To create the OFM GNN featurizer, we initially featurize the `matbench_mp_gap` dataset (MatBench v.0.1) comprising 106,113 structures. All structures are featurized using OFM, and an autoencoder is trained to discover a latent space representation of the obtained features. A few compressions are tested to assess the impact on performance in the `matbench_perovskites` task. Subsequently, the latent OFM features are taken as target and a GNN model is trained to generate the latent OFM features directly from the initial structures.

● **MatMiner-encoded GNN featurizer:** Following the same procedure done with the OFM featurizer, we encoded the features obtained by applying the default MatMiner featurizer of MODNet v.0.1.13 on the `matbench_mp_gap` dataset, this featurizer includes a total of 1336 MatMiner features. We assessed the results of different levels of compression in the `matbench_perovskites` and `matbench_mp_gap` tasks. Subsequently, the latent MatMiner features are used at the chosen level of compression as the target to train a GNN model to generate the latent MatMiner features directly from the initial structures.

● **Pre-trained MEGNet models as featurizers:** Five pre-trained MEGNet models released by Materials Virtual Lab (C. Chen et al. 2019) were used as featurizers for MODNet by taking the values produced in the MLP layers preceding the output value. Specifically, the models trained for formation energy, fermi energy and the elastic constants K^{VRH} and G^{VRH} on the 2019.4.1 Materials Project crystals dataset, as well as the band gap regression model trained on the 2018.6.1 Materials Project crystals dataset. The default MEGNet architecture comprises MEGNet blocks followed by an MLP with two dense layers—one with 32 neurons and another with 16 neurons—before producing the target property (see *Figure D2* in *Appendix D.2*). We experimented by using each of these layers as features for the model, identified in this work as MEGNetPreL32 and MEGNetPreL16, respectively. The resulting values are concatenated and added to the final feature vector, resulting in 160 descriptors for the MEGNetPreL32 featurizer and 80 descriptors for the MEGNetPreL16 featurizer. The models were then evaluated on the `matbench_perovskites` task.

● **Pre-trained adjacent MEGNet models as featurizer:** By training a MEGNet model beforehand on the training dataset, we can leverage its flexibility to enhance accuracy in MODNet, functioning as an adjacent model. Adjacent MEGNet models are trained for each fold of the train-test split and elemental embeddings were transferred from the MP-crystals-2018.6.1 (C. Chen et al. 2019) heat of formation task. All adjacent models hyperparameters employed were the default values as of MEGNet v.1.3.2, as described in *Appendix D.3*. We utilized the same optimal MLP layer defined for the pre-trained models as the output for this featurizer.

Finally, the features generated with the GNN featurizers are all tested, and their results are discussed in comparison to the pristine features and the dimensionally reduced ones. The synergic effect of including all these features is evaluated for performance on the `matbench_perovskites` task. Additional tests are also conducted in two tasks with comparable dataset sizes, predicting the stability and band gap of halides obtained from the OQMD database.

To identify the key features utilized by the MODNet model in models incorporating GNN/latent-space features, we employed SHAP value analysis. Subsequently, we trained surrogate models to obtain the most relevant GNN/latent-space features from the interpretable features. Applying SHAP analysis on these models, we uncovered the correlation between the most relevant GNN/latent-space features and more intuitive chemical descriptors, contributing to restore model's interpretability (detailed description of method on *Appendix D.4*).

5.3 RESULTS AND DISCUSSION

5.3.1 GNN featurizers for latent-space representations

Using the default MatMiner featurizer in MODNet, the MAE obtained for the *matbench_perovskites* task on perovskites' heat of formation was 0.0888 eV. This result falls within the error margin of 0.0028 eV from the benchmarked MAE for this task with MODNet (refer to *Table 14*). This value serves as the reference for all subsequent implementations to evaluate performance gain on this task.

The first implementation was the addition of OFM features to the original MatMiner features, as presented in *Table 15*. The inclusion of OFM features led to a performance boost compared to using solely the default MatMiner features for prediction. Despite nearly doubling the number of features, the inclusion of more specific chemical information on the orbital interactions in each structure proved beneficial to the model. It is noteworthy that the curse of dimensionality is mitigated through MODNet's robust feature selection algorithm and subsequent hyperparameter optimization through genetic algorithm to choose an optimal subset of features for the given problem.

An autoencoder is then trained to compress the OFM features from the larger `matbench_mp_gap` dataset to obtain a latent space representation. Different compressions are evaluated after hyperparameter tuning, and detailed results are provided in *Appendix D.3*. In Table 14, two of these autoencoders, with compression ratios (c.r.) of 20% and 10%, are used on the previous OFM features, replacing them with a latent-space representation to train and perform new predictions. The latent-space representation at a 20% c.r. further enhances the improvement compared to the original OFM features. This improvement can be attributed to an indirect transfer learning of chemical patterns in the larger chemical space sampled by the autoencoder. It may also benefit from a more compact number of features that better capture the chemical properties. When the latent space is reduced to 10% of the original dimensions, the boost compared to the original MatMiner featurizer can still be observed, but this smaller representation doesn't capture enough chemical information to surpass the addition of the original OFM features.

Finally, we validate the usage of the autoencoder to reduce the dimensionality of the OFM features by comparing the performance of the PCA-reduced features (also in the `matbench_mp_gap` dataset) with dimensions matching the 20% c.r. autoencoder, which produced the best results. These latent features are henceforth called ℓ-OFM for brevity. We observe a smaller but not significant performance reduction when using the PCA-reduced features compared to the encoder, as highlighted in the entries on *Table 15*.

Table 15 – Mean absolute errors for MODNet models on `matbench_perovskites` *task including pristine OFM features and different latent space reductions on top of the default MatMiner features.* n *represents the number of features after removing constant features across the dataset. Shaded rows highlight the chosen latent-space representation using autoencoder and the PCA-reduced representation with same dimensions for comparison. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet.*

Features	n	**MAE (eV)**
Default MatMiner (MM)	1020	0.0888
MM + original OFM	$1020 + 943$	0.0751 (-15.3%)
MM + latent OFM 20% c.r. (ℓ-OFM)	$1020 + 188$	0.0743 (-16.2%)
MM + latent OFM 10% c.r.	$1020 + 94$	0.0777 (-12.4%)
MM + PCA reduced OFM $(n = 188)$	$1020 + 188$	0.0748 (-15.7%)

We now proceed to evaluate the impact of using a latent-space representation of the MatMiner features compared to their original implementation. In this case, we also analyzed results in both `matbench_perovskites` and `matbench_mp_gap` to solidify the choice of the best encoder, as shown in *Table 16*. In the task of perovskite heat of formation, a general improvement in results is observed when using the latent-space features. This improvement is attributed to transfer learning from the larger dataset used for encoding. Transitioning to the band gap prediction task, which uses the same dataset employed to encode the features, an initial improvement is seen when converting to a latent-space of the same size. This improvement is expected due to the encoding being practically lossless and better capturing the chemical information in the entire dataset. However, as the compression increases, performance degrades, and in both datasets, a sharp decline is evident when going from a 60% to a 40% compression ratio. This supports our conjecture that the observed improvement in the task with a smaller dataset is attributed to the autoencoder functioning as a proxy for transfer learning to the feature-based model, akin to the role played by transferring elemental embeddings trained on larger datasets in graph-based models (C. Chen et al. 2019). It is plausible that training the autoencoder on a larger dataset could capture more robust chemical correlations, thereby potentially enhancing the results for band gap prediction as well.

Comparing the 60% c.r. with PCA reduction at the same size on the highlighted cells in *Table 16*, unlike the case of OFM features, there is a significant difference in favor of the encoder. The improved results observed for the encoded features can be attributed to the enhanced capacity of the autoencoder to capture non-linearity in its representation of the latent space. This is particularly important for the highly heterogeneous features in the MatMiner featurizer. This result reinforces our choice of the 60% c.r. encoder for MatMiner features, henceforth called ℓ-MM, by providing a lower number of features with relatively low degradation and advocating for the use of encoders in general instead of PCA-reduced features for latent-space representation.

Table 16 – Evaluation of the effects of dimensionality reduction on default MatMiner features used on MODNet model on Matbench tasks `matbench_perovskites` *and* `matbench_mp_gap`. n *is the number of features (constant features across the dataset removed) for the respective model and* N *the number of samples comprised in the dataset. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet for each task.*

Features used	**Task**			
	matbench perovskites (N=18,928)		**matbench mp_gap (N=106,113)**	
	n	**MAE (eV)**	n	**MAE (eV)**
Default MatMiner	1020	0.0888 ±0.0028	1264	0.2724 ±0.0052
Latent MatMiner without compression (1:1 latent space)	1264	0.0767 (−13.6%)	1264	0.2542 (−6.7%)
Latent MatMiner 80% c.r.	1011	0.0788 (−11.3%)	1011	0.2809 (+3.1%)
Latent MatMiner 60% c.r. (ℓ-MM)	758	0.0793 (−10.7%)	758	0.2911 (+6.8%)
Latent MatMiner 40% c.r.	505	0.0844 (−4.9%)	505	0.3280 (+20.4%)
PCA reduced MatMiner ($n = 758$)	758	0.0816 (−8.1%)	758	0.2968 (+8.9%)

The latent-space features, while providing a boost for the smaller dataset, still require the featurization of the original features to be subsequently encoded. To circumvent the expensive featurization process, we train MEGNet models to directly derive features from the structures, serving as featurizers. The detailed implementation and hyperparameter tuning are presented in *Appendix D.3*. It was observed that with proper hyperparameter tuning, the errors remained relatively low for the reconstruction of the original latent space considering the number of targets. The mean absolute error (on data normalized to the unit interval) was approximately 0.03 for the 758 latent-space features using the MatMiner GNN featurizer and about 0.01 for the 108 latent-space features using the OFM GNN featurizer.

In *Table 17*, the results for predicting perovskite heat of formation with MEGNet featurizers are presented. We observe a contrasting decline in efficiency when obtaining the latent MatMiner features with MEGNet compared to the original latent features. The error introduced by the MEGNet model, albeit seemingly small, was large enough to increase the MAE by over 0.025 eV. However, it is important to note that this result is still significantly better than benchmarked values from Automatminer and random forest, as presented in *Table 14*. The situation is much more favorable for the case of OFM features, in which the latent features obtained from the MEGNet model only decrease the performance by 0.0051 eV compared to the pristine latent features. Using latent features from both MEGNet models in combination reduces the error obtained from the application of MEGNet-derived latent MatMiner features alone, as expected, but not enough to boost performance compared to the default MODNet implementation.

These results underscore the importance of controlling errors in the GNN model using this approach, as there is a cumulative reconstruction error from the latent-space representation and the derivation from the MEGNet model that may lead to a substantial loss of chemical information. Training on a larger and more diverse dataset with careful curation of features is recommended based on these results, especially in heterogeneous featurizers such as the MatMiner featurizer applied here.

Table 17 – Mean absolute errors for MODNet models on `matbench_perovskites` *task comparing the inclusion of latent features originally obtained from the autoencoder and through the MEGNet featurizers. n represents the number of features after removing constant features across the dataset. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet.*

Features	**MAE (eV)**
Default MatMiner (MM)	0.0888
ℓ-MM	0.0793 (−10.7%)
MEGNet ℓ-MM	0.1052 (+18.5%)
MM + ℓ-OFM	0.0743 (−16.2%)
MM + MEGNet ℓ-OFM	0.0794 (−10.6%)
MEGNet ℓ-MM + MEGNet ℓ-OFM	0.0973 (+9.6%)

5.3.2 GNN featurizers from pre-trained models

As previously outlined in the methodology, our investigation focuses now on determining the more effective of the two final layers of the MLP in MEGNet pre-trained models, sourced from the Materials Virtual Lab, for use as features in prediction. In *Table 18*, we present a performance comparison for the `matbench_perovskites` task, incorporating the layers with 32 neurons (referred to as MEGNetPreL32) and the layers with 16 neurons (referred to as MEGNetPreL16). Additionally, we conducted assessments on randomly selected subsets comprising 5000 samples and 1000 samples from the initial `matbench_perovskites` dataset to verify the consistency of our findings for smaller datasets.

Our analysis reveals a consistent enhancement in performance with the inclusion of the MEGNetPreL32 featurizer over the MEGNetPreL16 featurizer, irrespective of dataset size. This improvement is attributed to a more general latent-space representation in the earlier layers of the model, which MODNet can effectively leverage. Notably, the percentage reduction in MAE compared to exclusive use of MatMiner features increases as the dataset size decreases. This underscores the transfer learning essence of this technique, transferring pre-acquired chemical knowledge to enhance performance on small datasets.

Table 18 – Mean absolute errors for MODNet models on `matbench_perovskites` task and subsets comparing the inclusion of features from pre-trained MEGNet models distributed by Materials Virtual Lab. N represents the size of the dataset used for the prediction. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet for each task.

Features	**Task**		
	matbench perovskites (N=18,928)	**matbench perovskites** (N=5,000)	**matbench perovskites** (N=1,000)
	MAE (eV)	**MAE (eV)**	**MAE (eV)**
Default MatMiner (MM)	0.0888	0.1667	0.2802
MM + MEGNetPreL16	0.0752 (−15.3%)	0.1202 (−27.9%)	0.1862 (−33.5%)
MM + MEGNetPreL32	0.0726 (−18.2%)	0.1167 (−30.0%)	0.1749 (−37.6%)

We proceed to examine the effects of combining latent-space representations from the MatMiner and OFM featurizers with the best-performing MEGNetPreL32 featurizer, as outlined in *Table 19*. It is evident that, despite the substantial contribution of pre-trained MEGNet models to accuracy, the addition of OFM latent features brings further improvement, highlighting a synergistic effect when these featurizers are integrated. The combined latent representations result in a total percent reduction of 26.5% in MAE over the default MatMiner featurizer in MODNet.

Table 19 – Mean absolute errors for MODNet models on `matbench_perovskites` *task comparing the inclusion of OFM latent features and both OFM latent features and MEGNetPreL32 features. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet.*

Features	MAE (eV)
Default MatMiner (MM)	0.0888
MM + ℓ-OFM	0.0743 (−16.2%)
MM + MEGNetPreL32	0.0726 (−18.2%)
MM + ℓ-OFM + MEGNetPreL32	0.0629 (−29.1%)
ℓ-MM	0.0793 (−10.7%)
ℓ-MM + ℓ-OFM	0.0728 (−18.0%)
ℓ-MM + MEGNetPreL32	0.0729 (−18.0%)
ℓ-MM + ℓ-OFM + MEGNetPreL32	0.0653 (−26.5%)

The latent-space and GNN derived features incorporated in these models, as opposed to the original features derived from chemical principles, lack direct interpretability. *Figure 37* showcases some of the most relevant features in the MODNet model with the set of features "MM + ℓ-OFM + MEGNetPreL32" determined through SHAP plot, and their relationship to interpretable features with highest SHAP value, determined through surrogate models (complete SHAP summary plots and detailed obtention presented in *Appendix D.4.1*). SHAP analysis proved to be a more reliable feature importance assessment than the built-in feature selection algorithm in MODNet, as discussed in *Appendix D.4.1*. The MEGNet pretrained features of the formation energy model were the most relevant for the model prediction. When decomposed into chemical descriptors, we can observe that more electronegative elements (*MagpieData_minimum_Electronegativity*, in *Figure 37*) tend to increase the heat of formation, just as increased difference in the number of electrons in the valence shell (*Magpie_avg_dev_NUnfilled*) or elemental ground-state band gap (*MagpieData_maximum_GSbandgap*). The heat of formation of perovskites also expectantly increases when the interaction of orbitals s^2 and p^4 is present (*OFM: s^2-p^4*), characteristic of many oxide perovskites, conversely to when two pnictogen elements are present (*OFM: p^3-p^3*), which creates weaker chemical bonds. The observation on combining pnictogens was observed before for perovskites in the high-throughput screening performed by Schmidt et al. (2021), in which, no system alloying two pnictogens presented decomposition energy below 100 meV/atom. The same effect is seen when Voronoi distances increase (*VoronoiFingerprint, mean_Voro_dist_minimum*), which corresponds to larger A-site cations, which usually indicate higher stability in perovskites (Sa et al. 2022). Among the MatMiner features ranking higher in the model output, we observe that presence of *d* orbitals tends to reduce the heat of formation, just as the presence of transition metals in general. A higher melting temperature, characteristic of higher binding energy, also correlates with a higher heat of formation. Analyzing the decomposition of the ℓ-OFM features, a couple observations can be drawn, such as mixed anion perovskites inducing weaker bonds (*OFM: p^4-p^3*) just as is the case of halide perovskites in general (*OFM: p^5-s^2*).

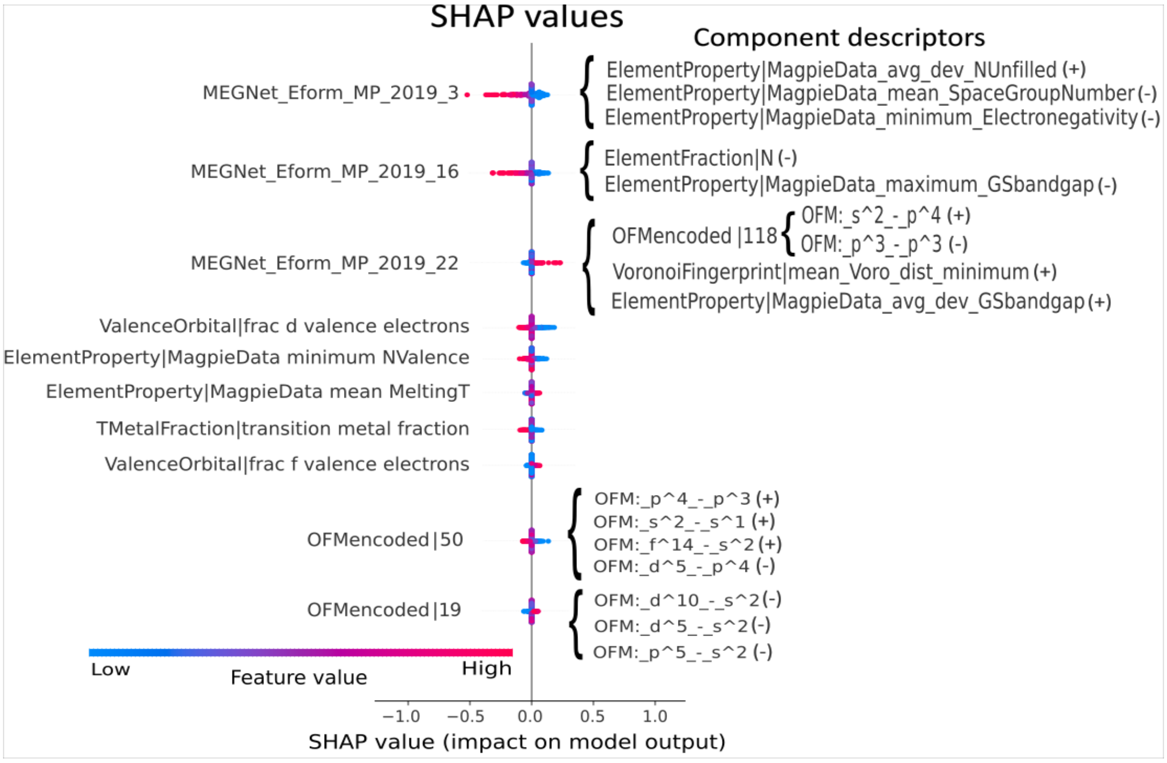


Figure 37 – SHAP analysis of top features in MODNet model for perovskite formation energy with "MM + ℓ-OFM + MEGNetPreL32" features. Encoded MEGNetPreL32 and ℓ-OFM features are decomposed into original MatMiner and OFM features, (+) indicates proportional variation and (-) indicates inversely proportional variation to the encoded features.

5.3.3 Adjacent GNN featurizer and final results

The latest addition to the proposed OMEGA featurizer involves incorporating an adjacent MEGNet model specifically trained for the target property using the same dataset. Although this introduces an additional computational burden due to the need to train an extra model, it offers the advantage of harnessing the flexibility of a GNN to enhance accuracy. Moreover, it still provides a means to partially regain interpretability on these features, as the model utilizes interpretable features that can be linked to those obtained from the GNN featurizer. The adjacent model underwent training with fixed default hyperparameters, utilizing elemental embedding from the formation energy task as recommended in the original MEGNet publication. This model can hence be used as a featurizer by extracting values from the dense layer with 32 neurons in the final MLP. This choice was based on the optimal performance observed for pre-trained models extracting this specific layer.

In *Table 20*, we integrate the adjacent model with other GNN featurizers for perovskite heat of formation. We evaluate both scenarios: retaining the MatMiner features or creating their latent representation using the GNN featurizer. The results highlight that GNN featurizers have an incremental effect on the accuracy of the model. The adjacent model decreases MAE by 0.0188 eV compared to adding only MEGNet ℓ-OFM and MEGNetPreL32 to the original MatMiner features. Additionally, it reduces MAE by 0.0227 eV compared to adding the same components to the latent-space MatMiner features generated by MEGNet, reaching almost same level of accuracy as when MatMiner featurization is performed. This result underscores that the adjacent model dominates the predictions.

When we analyze the most important features to the MODNet OMEGA model in *Figure 38* through SHAP value analysis (full SHAP plots presented in *Appendix D.4.2*), we can observe similar contributions to the "MM + ℓ-OFM + MEGNetPreL32" model appear for the MatMiner, ℓ-OFM and PreMEGNetL32 features. However, the model is dominated by the adjacent GNN model features, from which the top three features are also presented in the corresponding decomposition in MatMiner descriptors in *Figure 38*. Compared to the decompositions of the pre-trained MEGNet models in *Figure 37*, the adjacent model correlates to more subtle patterns, such as geometrical fingerprints and coulomb matrix eigenvalues. This can be attributed to the flexibility of the GNN model, which exploits highly non-linear relationships to enhance accuracy.

Nonetheless, our results reveal that using the proposed design, the intricate patterns leveraged for the enhanced accuracy of the GNN models can be explored through the combined training with easily interpretable chemical descriptors and SHAP value analysis. This provides greater interpretability with an accuracy that approaches benchmarked GNN results for the task (see *Table 14*). Moreover, by applying GNN models to produce latent-space features for MatMiner and OFM chemical descriptors, the computational cost of the featurization process is reduced, making it more viable for high-throughput materials screening.

Table 20 – Mean absolute errors for MODNet models on `matbench_perovskites` *task comparing the inclusion of all GNN featurizers over original MatMiner features and over MEGNet model generated latent-space MatMiner features. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet.*

Features	**MAE (eV)**
Default MatMiner (MM)	0.0888
MM + MEGNet ℓ-OFM	0.0794 (−10.6%)
MM + MEGNet ℓ-OFM + MEGNetPreL32	0.0683 (−23.1%)
OMEGA (MM + MEGNet ℓ-OFM + MEGNetPreL32 + Adjacent)	0.0495 (−44.2%)
MEGNet ℓ-MM	0.1052 (+18.5%)
MEGNet ℓ-MM + MEGNet ℓ-OFM	0.0973 (+9.6%)
MEGNet ℓ-MM + MEGNet ℓ-OFM + MEGNetPreL32	0.0726 (−18.2%)
OMEGAfast (MEGNet ℓ-MM + MEGNet ℓ-OFM + MEGNetPreL32 + Adjacent)	0.0499 (−43.8%)

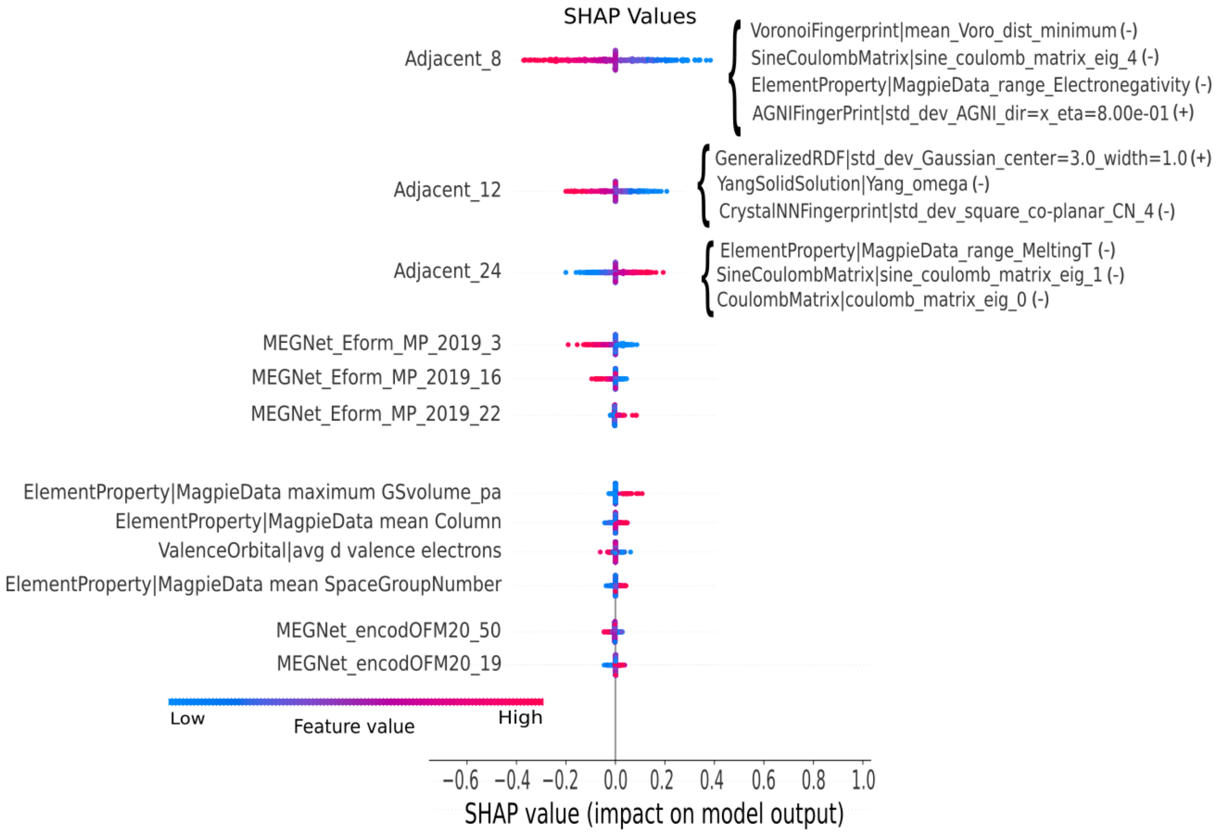


Figure 38 – SHAP analysis of selected top features in MODNet model for perovskite heat of formation with OMEGA features. Adjacent GNN model features are decomposed into original MatMiner and MEGNet ℓ-OFM *features, where a few are shown, (+) indicates proportional variation and (-) indicates inversely proportional variation to the encoded features.*

To further explore the proposed method, two additional tasks were evaluated: predicting the convex hull distance and band gaps of halogen-containing materials from the OQMD dataset. The OQMD dataset was filtered by the presence of halogens, namely F, Cl, Br, and I. After testing, we observed that the models would generalize better after removing structures whose stability (determined from the convex hull) was above the threshold of 2.9 eV/atom, corresponding to 0.1% of the structures, and keeping only those structures with distinct compositions. This resulted in a dataset with 31,271 samples to train for stability. This dataset was further filtered for structures with band gaps above 0.5 eV, resulting in 8,518 structures for training in band gap prediction.

These tasks are more challenging than our proof-of-concept task on perovskites. Since the GNN featurizer for OFM and the pre-trained models from the Materials Virtual Lab were all trained on the Materials Project dataset, which has limited information on halides, we anticipate lower generalization for this dataset (Shen et al. 2022). To estimate the convex hull distance, in particular, is notoriously difficult and presents a current challenge on materials screening (Bartel 2022). The results are presented on *Table 21*, where we observe minimal influence on stability predictions when including the MEGNet ℓ-OFM featurizer. However, significant improvement is seen when including MEGNetPreL32, further enhanced by the addition of the adjacent model. For the band gap task, MEGNet ℓ-OFM still fails to provide significant improvement to the model, but MEGNetPreL32 and adjacent model GNN featurizers compensate with a notable enhancement.

Table 21 – Mean absolute errors for MODNet models on tasks of prediction of convex hull distance and band gap on the subset of halogen-containing materials from OQMD, comparing the inclusion of all GNN featurizers over original MatMiner features. In parentheses, percentage MAE deviation from the default MatMiner featurizer in MODNet on the given task.

Features	**Task**	
	OQMD halogen E_{hull} (N=31,271)	**OQMD halogen E_g** (N=8,518)
	MAE (eV)	**MAE (eV)**
Default MatMiner (MM)	0.0556	0.4557
MM + MEGNet ℓ-OFM	0.0561 (+0.8%)	0.4501 (−1.2%)
MM + MEGNet ℓ-OFM + MEGNetPreL32	0.0538 (−3.2%)	0.3835 (−15.8%)
OMEGA	0.0519 (−6.7%)	0.3784 (−17.0%)

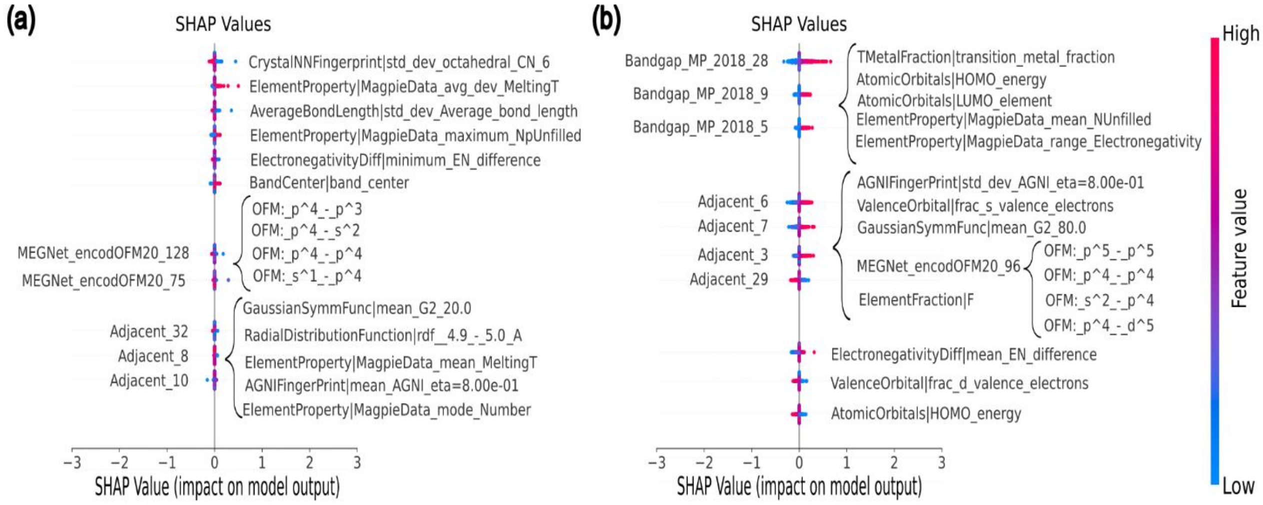


Figure 39 – SHAP analysis of selected top features in MODNet model for OQMD halogen task for stability (a) and band gap (b) with OMEGA features. Adjacent model and MEGNet ℓ-OFM *features are decomposed into chemical descriptors, which a few with highest impact on the group of features are shown.*

In *Figure 39*, the OMEGA features with the highest importance on the corresponding MODNet model output are presented. For the stability task, the selected features are presented in *Figure 39*(a). A stronger influence of the original MatMiner features than any of the additional OMEGA features is observed. This is consistent with the marginal improvement in performance, showing only 6.7% reduction in MAE when all additional OMEGA features are included, in contrast to the results of perovskites' heat of formation model. This reflects the difficulty of predicting stability from materials structures even with state-of-the-art GNN models (Riebesell 2024). The importance of the features determined by SHAP is also quite homogeneous throughout the plot with no clear dominant features. This is more evident when looking at the extended SHAP plot in *Figure D9*. This observation underscores the importance of crafting new materials descriptors that better correlate with this property. From this analysis, we also find significant importance attached to geometric descriptors, such as geometrical fingerprints, symmetry, and bond lengths, alongside traditional chemical descriptors like electronegativity difference, valence orbital filling, and estimated melting temperature. These features are prominent in both the top MatMiner features and adjacent model decompositions. MatMiner's band structure featurizer, which composes of the electronegativity of the elements, also figures in the top features, along with a couple of ℓ-OFM features whose most relevant features are linked to the chalcogenides (linked to the p^4 valence shell contributions). This is supported by the prevalence of chalcogenides in inorganic materials.

In the case of the band gap task for the OQMD halogen, the OMEGA features take precedence as seen in *Figure 39*(b). Not surprisingly, features from the pretrained MEGNet model for band gap regression are positioned at the top. These features can be correlated to chemical descriptors such as electronegativity difference, HOMO/LUMO energies of the atomic orbitals, filling of the valence band, and transition metal presence/d-orbital valence filling. Decomposing the adjacent model features reveals similar chemical descriptors contributing (see extended SHAP plot on *Figure D14*), along with additional geometrical descriptors, aiding in fine-tuning the model and explaining the increase in accuracy as shown in *Table 21*.

5.4 CONCLUSION

Our results validate the approach of integrating pre-trained GNN models as featurizers in feature-based models, enhancing their competitiveness for larger datasets. The final implementation in this investigation, named OMEGA featurizer, employed MEGNet models to produce features based on pretrained models and to swiftly derive latent-space OFM features. We showcase the efficacy of the OMEGA featurizer on the task of perovskite heat of formation prediction. Compared to the default featurizer in MODNet, the OMEGA featurizer reduces the MAE by 44.2%, achieving an accuracy close to benchmarked GNN models for this task. Furthermore, the generalizability of the OMEGA featurizer on additional tasks is demonstrated, including predicting the convex hull distance and band gaps of halogen-containing materials from the OQMD dataset. The results highlight the effectiveness of the OMEGA featurizer, particularly for the band gap prediction task.

Additionally, this novel approach bridges the interpretability gap between easily interpretable feature-based models and highly accurate but less interpretable GNNs. By analyzing feature importance with SHAP plots and employing surrogate models, we can extract relevant chemical information from GNN features used for prediction. This paves the way for exploring these models to screen vast chemical spaces, facilitating chemically guided active learning due to their inherent interpretability. In conclusion, the incorporation of GNN features into feature-based models offers a versatile and powerful method to boost predictions. It enables leveraging pre-trained GNN knowledge, reduces featurization costs, and enhances model accuracy while partially retaining interpretability through dimensionality reduction and decomposition techniques like SHAP analysis. This paves the way for more accurate, efficient, and interpretable materials discovery through feature-based modeling.

CHAPTER 6 — Machine Learning-Assisted Exploration of 111-Type 2D Perovskite Structures for Photovoltaic and Optoelectronic Applications: A High-Throughput Screening Approach

6.1 RESEARCH PROBLEM

Perovskites with ordered vacancies along the <111> direction, with chemical formula $A_3B_2X_9$ (where A is a monovalent cation, B is a trivalent cation such as Bi^{3+} or Sb^{3+}, and X is a halide anion), form the most representative subgroup of all-inorganic 2D metal halide perovskites (MHPs). These structures have drawn tremendous interest due to low toxicity, long-term stability, and remarkable optoelectronic properties (Z. Jin et al. 2020), holding promise to substitute the highly toxic and moisture-sensitive lead halide perovskites and also circumvent the problem of limited carrier generation and transport of 2D MHPs with organic spacers (Acharyya, Kundu, and Biswas 2020; Blancon et al. 2020). Most of these layered materials will present in the trigonal ($P\bar{3}m1$) crystal system, with a few crystallizing in the monoclinic phase (e.g, $Rb_3Bi_2I_9$, space group $P21/n$) (Tomaszewski 1994; S. Y. Kim et al. 2019).

Most investigations in the literature focus on varying the elements within the A-, B-, or X-sites in perovskites to seek compounds that offer improved optoelectronic properties and enhanced stability under environmental conditions. There has also been interest in halogen alloying and B-site doping studies for these (111)-type perovskites (K.-H. Hong et al. 2017; Pradhan, Jena, and Samal 2022; Gouvêa et al. 2024; Exner et al. 2024). However, it is recurrently observed in perovskite literature that combining cation and anion doping is a common strategy to enhance their properties. For instance, top-tier lead halide perovskites are formed by concurrent cation and halogen doping, such as $FA_{0.992}MA_{0.008}PbI_{2.976}Br_{0.024}$, which reaches certified PCE of 25,2% (Mica et al. 2020; Yoo et al. 2021). Therefore, conducting comprehensive studies that screen a wide range of elements and enable mixed-cation mixed-halide (111)-type perovskites is imperative for better understanding the potential of these materials and hopefully enhancing their performance for practical applications.

Recent advancements in computational materials databases and machine learning methods have made it feasible to screen large compositional spaces for stable structures with desired properties (Schleder et al. 2019; J. Yang and Mannodi-Kanakkithodi 2022). However, most of these investigations still focus on permutations of elements for a fixed set of stoichiometries and space groups. They lack flexibility to infer stability for systems under multi-site doping. A significant challenge in predicting these complex configurations lies in the necessity for relaxed structures to accurately predict properties in machine learning models, typically trained on DFT-relaxed structures. A potential solution to this challenge is to develop machine learning models that are invariant to geometrical changes under relaxation (Schmidt et al. 2021; B. Zhang et al. 2022) or to fine-tune the original models for improved predictions on unrelaxed structures (Choubisa et al. 2023). Thankfully, the development of machine-learning interatomic potentials (MLIP) using many-body graph convolution networks, such as M3GNet (C. Chen and Ong 2022) and CHGNet (B. Deng et al. 2023), now enables the determination of energetics and derivation of equilibrium geometry for a given structure with an arbitrary unit cell, facilitating the acquisition of final geometries for individual atom substitutions.

In this study, we introduce a high-throughput screening approach that combines a bond length invariant ML model with MLIPs. This tandem approach, facilitated by an active learning loop, identifies a subset of candidate (111)-type perovskite structures meeting stability and band gap criteria from a pool of over 100 million structures with mixed-cation and mixed-anion compositions. Our work aligns with recent advancements addressing multi-site doping (Choubisa et al. 2020; 2023), and active learning methods for materials discovery such as Google's Graph Networks for Materials Exploration (GNoME), which incorporates tandem frameworks and ab-initio random structure searching techniques (Merchant et al. 2023). A notable distinction of our approach is the utilization of the feature-based Materials Optimal Descriptor Network (MODNet) for predictions (De Breuck, Hautier, and Rignanese 2021). As a feature-based model MODNet offers enhanced interpretability, providing insights into chemical patterns for materials design and further exploration and by simply selecting the appropriate subset of features, MODNet can work with full structure information or only their prototype geometry can be taken into consideration. The framework yields a group of candidate structures, which, upon preliminary verification with ab initio calculations, reveals good qualitative agreement to identify thermodynamically stable structures by convex hull distance calculation and quite good predictions for the band gap value.

6.2 METHODOLOGY

We began with a $1 \times 1 \times 2$ supercell structure of $A_3B_2X_9$ ($P\bar{3}m1$) as a prototype for the (111)-type perovskite. For site A, we considered Rb, Cs, and K; for site B, the following 28 elements: Si, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, Se, Sr, Y, Zr, Nb, Mo, Ag, In, Sn, Sb, Te, Ba, Pb, Bi; and finally, for site X, halogens and chalcogens: Cl, Br, I, O, S. The choice of these elements was based on the literature on perovskite materials (details on *Table E*). We restricted the number of substitutions at each site according to the stoichiometry $A_{6-2x}A'_{2x}B_{4-y}B'_{y}X_{18-2z}X'_{2z}$ where x, y and z are valid integers and at least two-thirds of X sites are occupied by halogens. The base composition allows for multi-site doping with up to two different elements on each elemental site (all the sites occupied by the same element in pristine $A_3B_2X_9$ structure). These structures were then generated for the supercell with up to 6 permutations of elements for each elemental site, resulting in up to 216 elemental rearrangements for each composition. The unit cell dimension of each of these structures is scaled based on the ionic radius of the elements (details on *Appendix E.1*). This leads to a total of 100,627,800 arrangements with 470,610 distinct compositions.

The method proposed follows 5 steps of ML-guided screening, as illustrated in *Figure 40*, the screening applies the CHGNet model as MLIP and custom ensemble MODNet models for stability and band gap, comprehensive details on their implementation is provided in *Supporting Information (Appendix E.4)*.

1. **One-shot CHGNet screening:** a one-shot energy evaluation with CHGNet reduces the initial pool of permutated structures to a single optimal structure for each composition.
2. **Unrelaxed structure screening:** screening is performed with an ensemble MODNet model trained on features invariant to precise structural information obtained on relaxation, through the implementation of a special featurizer (*InvariantMatMiner2023).* This model is trained for stability on OQMD dataset with addition of in-group data on alloyed (111)-perovskites (Gouvêa et al. 2024; Exner et al. 2024), detailed implementation of the featurizer and effects

of in-group data addition are discussed on *Appendix E.4*. After excluding protostructures with stability thresholds > 35 meV/atom, an initial dataset with 15,000 structures for active learning is constructed, taking 70% (10,500 structures) exhibiting lowest upper stability bound and 30% (4,500 structures) that optimize an acquisition function considering estimated uncertainty and entropy.

3. **Active learning cycle:** the initial 15,000 structures previously selected undergo constrained relaxation with CHGNet. The final energy for each structure is used to evaluate their decomposition energy (E_{stab}^{CHGNet}) through OQMD's formation energy convex hull. An ensemble MODNet model is then trained to predict E_{stab}^{CHGNet} in all structures not considered in the initial/updated dataset for active learning. The same thresholds are applied to select a new subset of 2,500 structures to include in the active learning cycle. The active learning cycle ends when only structures with estimated E_{stab}^{CHGNet} > 45 meV/atom remain on the pool.
4. **Advanced model ML screening:** screened structures are featurized with a more advanced MODNet featurizer (*OMEGA+ROSA*, details in *Appendix E.3*) and ensemble MODNet models are now trained on halogen-containing OQMD dataset for stability, band gap classification and band gap prediction. The threshold applied on each of these models is shown in *Figure 40*, resulting in a final set of structures to be evaluated by DFT calculations.
5. **ML phonon frequency screening :** the advent of the more precise MLIP allows to estimate the dynamic stability through vibrational properties even for large structures in a matter of minutes. This tool was applied with CHGNet's MLIP and the detection of negative frequencies in the phonon density of states (PhDOS) below a threshold of -0.35 THz was applied to exclude possibly dynamically unstable structures.

Since a great number of structures was still present after applying the filters, we reduce structures with same elements in B-site to their optimal composition in terms of predicted stability with the OMEGA+ROSA MODNet model. Finally, high-throughput DFT calculations were performed with *Atomate2* (Ganose et al. 2024) employing VASP 5.4.4 and PBE PAW pseudopotentials (Kresse and Joubert 1999; Kresse and

Furthmüller 1996) on a subset of these structures from which total energies were used to determine stability relative to OQMD’s convex hull.

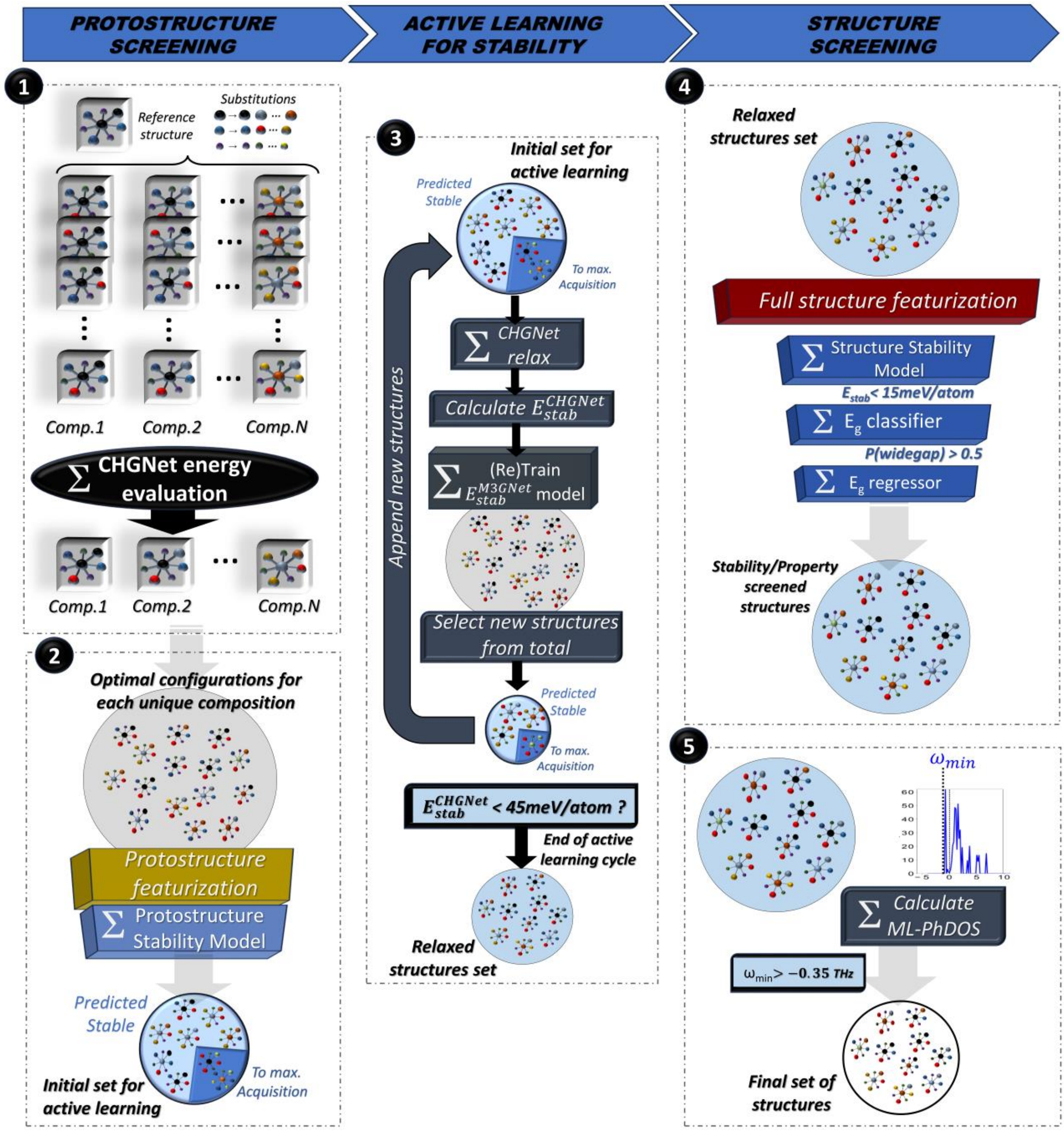

PROTOSTRUCTURE SCREENING
ACTIVE LEARNING FOR STABILITY
STRUCTURE SCREENING
1
Reference structure
Substitutions
Comp.1
Comp.2
Comp.N
∑ CHGNet energy evaluation
Comp.1
Comp.2
Comp.N
2
Optimal configurations for each unique composition
Protostructure featurization
∑ Protostructure Stability Model
Predicted Stable
To max. Acquisition
Initial set for active learning
3
Initial set for active learning
Predicted Stable
To max. Acquisition
∑ CHGNet relax
Calculate E_{stab}^{CHGNet}
∑ (Re)Train E_{stab}^{M3GNet} model
Append new structures
Select new structures from total
Predicted Stable
To max. Acquisition
E_{stab}^{CHGNet} < 45meV/atom ?
End of active learning cycle
Relaxed structures set
4
Relaxed structures set
Full structure featurization
∑ Structure Stability Model
E_{stab}< 15meV/atom
∑ E_g classifier
P(widegap) > 0.5
∑ E_g regressor
Stability/Property screened structures
5
ω_{min}
60
50
40
30
20
10
0
-5
0
5
10
∑ Calculate ML-PhDOS
ω_{min}> −0.35 THz
Final set of structures

Figure 40 – Schematic diagram of the high-throughput screening method assisted by machine learning, divided into five0 steps, namely: (1) reduction to optimal configurations for each composition, (2) simple ML screening to generate initial set for active learning, (3) active learning cycle for stability by CHGNet, (4) full featurization of final active learning set and prediction of finely tuned models to obtain final structures, (5) calculation of ML phonon density of states to screen structures based on most negative phonon frequency.

6.3 RESULTS AND DISCUSSION

In the CHGNet one-shot screening, for each unique composition, an optimal structure with the lowest energy is selected among the set of generated structures. In *Figure 41*, some of these structures are presented for a few selected compositions, exhibiting the selected structure along with some higher energy counterparts. From the analysis, it is evident that CHGNet appropriately captures the main factors contributing to minimizing the total energy. Beginning with the mixed-anion $Cs_3Sb_2(BrCl_2)_3$ in *Figure 41(a)*, it can be observed that structures with bromine more homogeneously distributed exhibit lower energies. This observation is corroborated by DFT calculations involving the same alloy (Gouvêa et al., 2024). CHGNet also successfully captures the tendency of halogen alloying on these perovskites to saturate terminal sites first, as observed experimentally (Pradhan, Jena, and Samal, 2022). This is evident by the highest energy among candidate structures being observed when bromine is concentrated on the bridging sites.

In the example of $Cs_3GaBiBr_9$ as a B-site mixed-cation structure in *Figure 41(b)*, CHGNet deems the structure presenting intercalating layers of the distinct B atoms in the c-axis direction more stable. This result can be intuitively understood by observing that in this case, the strain caused by the different cation sizes is distributed along the 2D layer structure. A more special case of mixed-anion mixed-cation is seen in *Figure 41*(c) for $Cs_3Sc_{1.5}Nb_{0.5}SBr_8$. Since Nb tends to present an oxidation state of +4 or +5, it is expected that the chalcogen in the structure is attracted towards it. CHGNet follows the expected chemical behavior, attributing the configuration in which S atoms are closer to the Nb site with consistently lower energy. For the last case presented in *Figure 41(d)*, for the mixed A-cation composition $Cs_2RbIn_2Br_9$, CHGNet seems to capture that the smaller cation Rb should be placed in the terminal sites facing the gaps in the layered structure instead of the sites right within the layer. This evaluation can also be intuitively understood in terms of strain minimization, since a smaller cation in the center of the layer would induce an increased tilt in the perovskite octahedra.

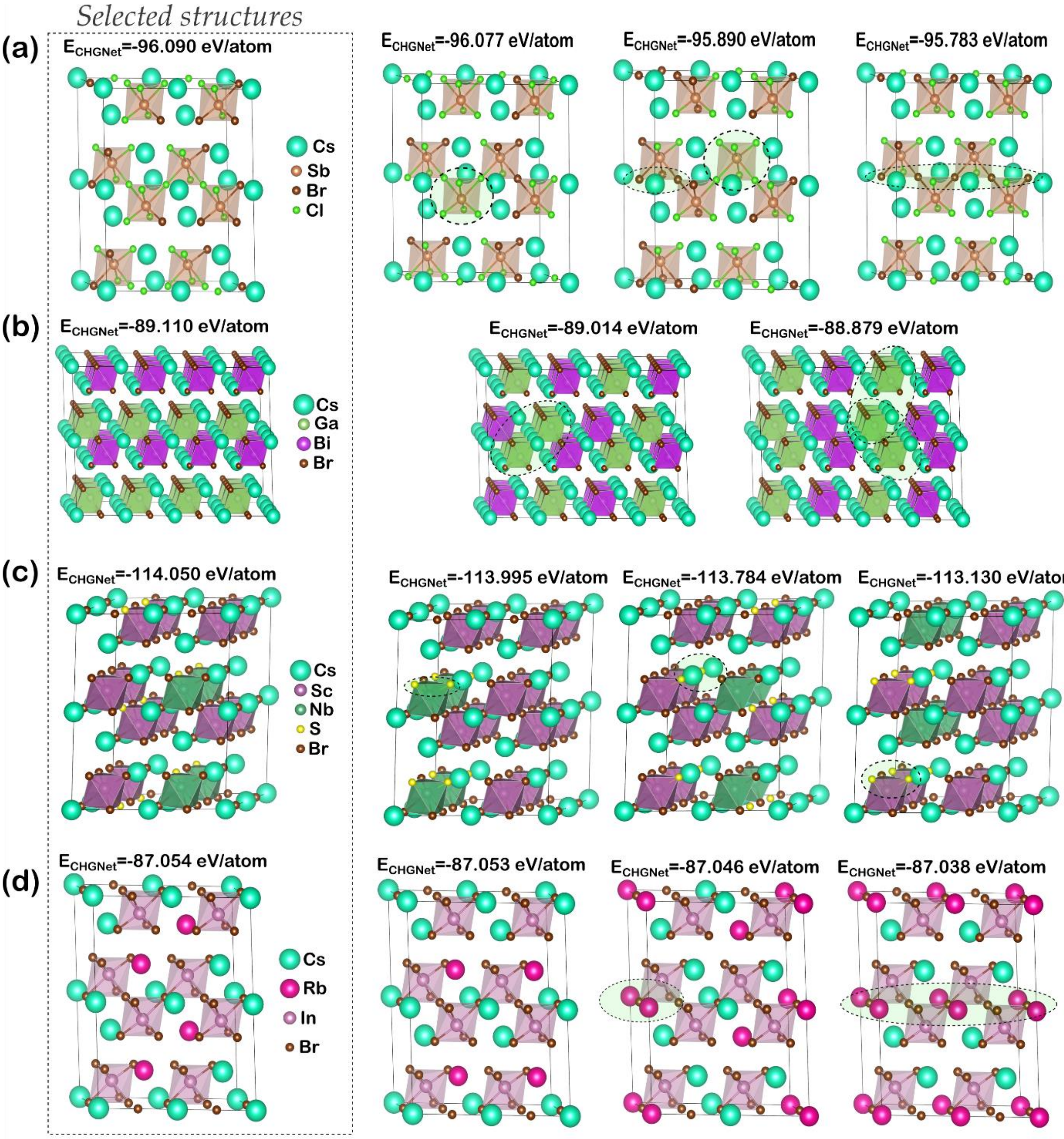


Figure 41 – Selection through the CHGNet one-shot screening of mixed-cation and/or mixed-anion structures presenting optimal and sub-optimal arrangements for their given composition, namely: (a) $Cs_3Sb_2(BrCl_2)_3$, (b) $Cs_3GaBiBr_9$, (c) $Cs_3Sc_{1.5}Nb_{0.5}SBr_8$ and (d) $Cs_2RbIn_2Br_9$. Spots associated with an increase in total energy are highlighted in the figure.

For the second phase of the screening, we trained the protostructure model to select starting structures for the AL cycle. The metrics of the protostructure model are presented on *Table 22* for validation and test metrics (details on training and evaluation in *Appendix E.2*). The metrics are within expectations for a feature-based model which does not take into consideration the full structural information, GNN models with graph-attention for unrelaxed structures in much larger datasets report MAE of 30 meV/atom (Schmidt et al. 2021) and fine-tuning of best performing GNNs for unrelaxed structures report MAE of 34 meV/atom (Choubisa et al. 2023). An important criterion for assessing the accuracy of the model for our application is to verify the estimated stability of experimentally reported (111)-type perovskites, as shown in our results in *Table E3*. By utilizing the upper limit on the decomposition energy (prediction + model uncertainty), all experimentally reported structures were found to fall within a threshold of 15 meV/atom. To account for the possible biases of the model, an additional 20 meV/atom margin was included, setting our threshold at 35 meV/atom to exclude highly unlikely structures from our initial active learning training.

Table 22 – Evaluation metrics for the protostructure-based and the structure-based models to estimate stability.

Model Name **(*Featurizer name*)**	**MAE** **(meV/atom)**		**R²**	
	Validation	Test	Validation	Test
Protostructure-based stability estimator *(InvariantMatminer2023)*	34.5	60.6	0.933	0.808
Structure-based stability estimator *(OMEGA + ROSA)*	28.1	49.0	0.962	0.877
Structure-based band gap estimator *(OMEGA + ROSA)*	0.19	0.37	0.981	0.905
	AUCROC			
	Validation	Test		
Structure-based band gap classifier *(OMEGA + ROSA)*	0.866	0.768		

The active learning cycle begins by relaxing the 15,000 selected structures screened from the previous step through CHGNet. The active learning model appears to reach a plateau in accuracy with just 15,000 structures. Detailed information on the metrics with an increasing dataset is provided in *Appendix E.4*. The active learning cycle was halted when only structures with E_{stab}^{CHGNet} > 45 meV/atom were being added to the next cycle, resulting in 6 cycles of AL. This threshold was determined based on previous criteria involving the stability of experimental structures and an additional margin to account for a tendency of our estimated decomposition energy from CHGNet relaxed structures to shift towards higher values. This shift can be verified in *Table 23*. Furthermore, we observed a significant number of structures with large negative decomposition energy through CHGNet energies. This reflects the low coverage of current materials databases for multinary materials and has only recently been addressed by large-scale simulation efforts, including quaternary and quintenary materials (Merchant et al. 2023). The MODNet OMEGA+ROSA stability estimator provides a more reasonable estimation for these structures since it is based on general chemical descriptors that are not directly influenced by an incomplete convex hull, as demonstrated in *Figure 42*.

Table 23 – Estimated decomposition energy (E_{stab}), probability of being semiconductor (p_{semi}) and band gap (E_g) from structure-based MODNet models along with CHGNet estimated stability (E_{stab}^{CHGNet}) and most negative phonon frequency (ω_{min}) for selected (111)-type perovskites. Theoretical band gap (DFT E_g) is also presented for experimentally reported structures.

Composition of (111)-type structures	Predictions from structure-based MODNet models			E_{stab}^{CHGNet} (meV/atom)	$\omega_{\min}$ (Thz)	DFT E_g (eV)
	E_{stab} (meV/atom)	p_{semi}	E_g (eV)			
Experimentally reported						
$Cs_3Sb_2I_9$ (ICSD: #39822)	-5.5	0.80	1.73	9.4	-0.04	1.72
$Cs_3Sb_2Br_9$ (ICSD: #39824)	-17.5	0.88	2.01	5.8	-0.25	1.98
$Cs_3Sb_2Cl_9$ (ICSD: #22075)	-8.6	0.84	2.54	9.7	-0.16	2.47
$Cs_3Sb_2BrCl_8$ (ref: §1)	-4.5	0.80	2.34	9.4	-0.27	2.38
$Cs_3Sb_2Br_2Cl_7$ (ref: §1)	1.6	0.92	2.34	8.8	-0.26	2.34
$Cs_3Sb_2(BrCl_2)_3$ (ref: §1)	0.0	0.96	2.32	9.1	-0.30	2.29
$Cs_3Fe_2Cl_9$ (ICSD: #22074)	13.6	0.60	0.92	11.1	-0.15	0.53
$Rb_3Sb_2Br_9$ (ICSD: #39823)	-24.3	0.84	2.04	3.6	-0.22	2.07
$Cs_3Bi_2Br_9$ (ICSD: #1142)	-31.5	0.76	2.58	12.8	-0.28	2.60

(continues)

Table 23 – (continued)

Composition of (111)-type structures	**Predictions from structure-based MODNet models**			E_{stab}^{CHGNet} (meV/atom)	ω_{min} (Thz)
	E_{stab} (meV/atom)	p_{semi}	E_g (eV)		
Oxygen containing					
$CsRb_2CaNbI_8O$	13.2	0.92	1.69	21.7	-0.46
$Cs_3TiMnBr_8O$	11.0	0.72	1.36	22.7	-0.83
Cs_3CaNbI_8O	12.1	0.88	1.82	0.1	-0.37
$Cs_2KY_{1.5}Se_{0.5}Br_8O$	7.8	0.88	2.56	26.0	-0.58
$Cs_3Y_{1.5}Se_{0.5}Br_8O$	8.2	0.88	2.79	3.0	-0.35
$Cs_2RbY_{1.5}Se_{0.5}Br_8O$	8.6	0.92	2.66	15.5	-0.45
Sulphur containing (12 most stable and 1 containing two S per formula)					
$Cs_3Sc_{1.5}Nb_{0.5}SBr_8$	-15.5	0.88	2.44	14.6	-0.35
$CsRb_2Sc_{1.5}Nb_{0.5}SBr_8$	-5.6	0.80	2.38	34.1	-0.22
$Rb_3Sc_{1.5}Nb_{0.5}SBr_8$	-3.1	0.88	2.43	32.3	-0.08
$Cs_3CaVSBr_8$	-1.0	0.76	1.65	16.5	-0.16
$Cs_2RbCaVSBr_8$	-0.5	0.76	1.58	24.1	-0.31
$Cs_3CrInSBr_8$	0.0	0.88	1.27	25.1	-0.10
$CsRb_2CaVSBr_8$	1.7	0.72	1.50	26.4	-0.36
$Cs_3Y_{1.5}V_{0.5}SBr_8$	2.6	0.68	2.01	26.0	-0.26
$Cs_3Sc_{1.5}VSBr_8$	3.9	0.72	1.87	2.5	-0.17
$Cs_3YTiSBr_8$	4.5	0.76	1.61	23.3	-0.41
$Cs_3YFeSBr_8$	4.8	0.84	1.20	19.4	-0.27
$CsRb_2TiSbSBr_8$	4.9	0.76	1.48	29.2	-0.27
$Cs_3Sc_{1.5}MnS_2Br_7$	7.7	0.8	1.75	20.9	-0.20

(ref: §1) - (Pradhan, Jena, and Samal 2022)

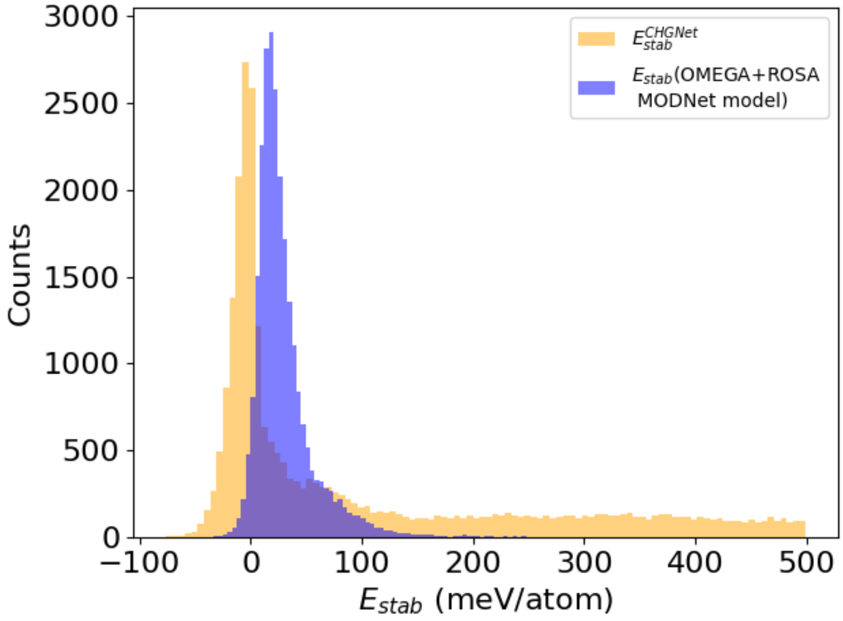


Figure 42 – Histograms of estimated decomposition energies (E_{stab}) for machine-learning relaxed structures in the active learning cycle. Comparison of E_{stab} calculated from CHGNet total energy and predicted with the structure-based MODNet model.

Finally, the resulting structures are evaluated using models trained with the advanced featurizer, which considers full structural information. The estimated stability of experimentally reported structures using these models is also presented in *Table 23*. In this case, significant improvements in estimates compared to protostructure-based models (Table E3) are observed, with values closer to the convex hull (E_{stab} = 0 meV/atom), particularly for halogen-alloyed structures. The evaluation metrics of these models are presented in *Table 22*, where a significant improvement is observed for the model employing the OMEGA+ROSA featurizer, although it is trained only on the halogen-containing dataset, compared to the protostructure-based model, trained on a much larger dataset (details in *Appendix E.2*). Training the models using the OMEGA+ROSA featurizer on a larger dataset should improve the accuracy of the model, as it appears to lose accuracy due to overfitting to the training data, thereby degrading accuracy in the test set. Efforts in this direction are deferred to future work, as the model has proven useful for the proposed application despite this limitation.

Since MODNet models are feature-based, the importance and effects of each feature on the model output is straightforward via SHAP analysis. *Figure 43* showcases the SHAP analysis for stability and band gap estimators. The

OMEGA+ROSA featurizer contains GNN features and ab-initio based features (details in *Appendix E.3*) which may not be directly interpretable. These features are then correlated with interpretable chemical or geometrical descriptors included in the model training, via SHAP plots from surrogate models. *Appendix E.5* provides a comprehensive version of each model's SHAP plots (*Figure E3* and *E6*), along with the decomposition into interpretable features for the most important groups (*Figure E4*, *E5*, *E7* and *E8*).

For the stability model, in *Figure 43*(*a*), no single dominant feature is observed with many high-ranking features presenting similar impact on the model output. The model heavily relies on GNN features for capturing complex patterns, as anticipated due to the inherent difficulty in predicting material stability. SHAP analysis reveals a correlation of the exchange-correlation contribution in total energy (*ROSA|e_xc_per_atom*) and E_{stab}. Additionally, ROSA's estimated kinetic and entropy energy contributions (*ROSA|e_kinetic_per_atom* and *ROSA|e_entropy_per_atom*) are relevant to the model. Expected chemical descriptors related to estimated melting temperature (*Magpie_data_avg_dev_Melting*T and *Magpie_data_mean_MeltingT*) also rank highly. GNN adjacent model features collectively contribute significantly, correlating with chemical descriptors such as electronegativity, presence of transition metals, and various geometrical descriptors. In the SHAP analysis of the band gap estimator, shown in *Figure 43*(*b*), the PBE band gap estimated by the ROSA featurizer (*ROSA|Band_Gap_PBE*) stands out in predictions. This feature correlates with explicit descriptors such as electronic entropy contribution (*ROSA|e_entropy_per_atom*), presence of d valence electrons (*frac_d_valence_electrons)* and transition metals (*transition_metal_fraction*), which help distinguish metallic from semiconductor materials. ROSA's eigenvalues above and below the Fermi level (*Eigenvalue+1* and *Eigenvalue-1*), directly linked to the PBE band gap, also rank among the top features, offering additional insight through their decompositions. Furthermore, pre-trained GNN features for band gap prediction in the OMEGA featurizer significantly impact output, correlating with valence shell information. Meanwhile, adjacent GNN model features enhance flexibility in capturing the influence of multiple geometrical descriptors on predictions.

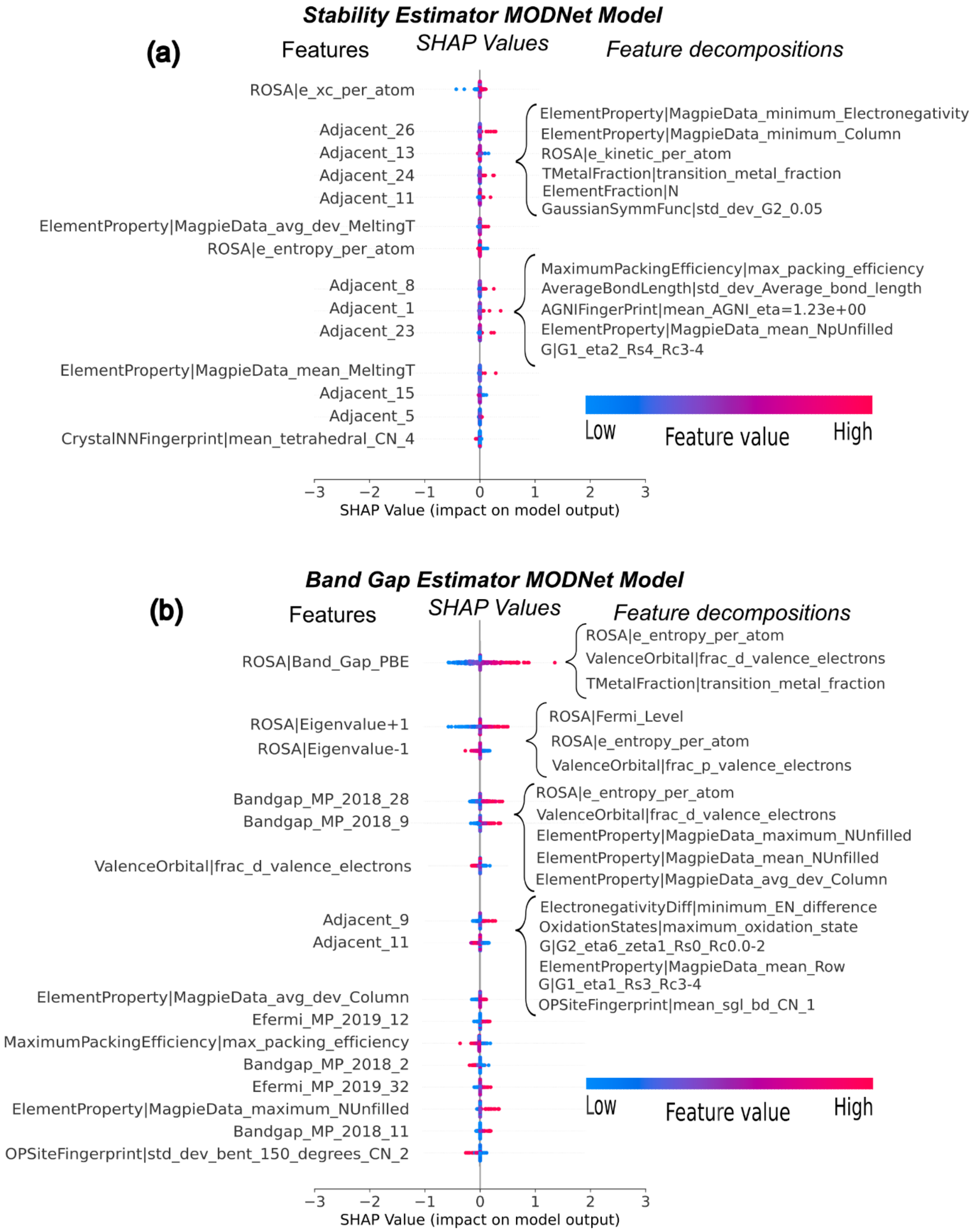


Figure 43 – SHAP analysis of selected top features in structure-based MODNet models for (a) stability and (b) band gap. Groups of features based on GNN models and ROSA features such as PBE band gap and eigenvalues are decomposed into interpretable chemical/geometrical descriptors, which a few with highest impact on the group of features are shown.

The ML screening proceeds by applying the structure-based models to the final pool of candidate materials, selecting those with predicted E_{stab} below 15 meV/atom, classified as semiconductors ($p_{semi} > 0.5$), and with a predicted band gap less than 3.5 eV. These thresholds were determined based on the results of the experimentally reported (111)-type perovskite structures in *Table 23*. A total of 4432 structures are screened in this process, and an overview of their distribution and estimated properties is presented in *Figure 44*. *Figure 44*(a) illustrates the frequency of each B-site cation in the screened structures. It indicates that Sc, Y, and Sb-containing (111)-type perovskite structures are likely the most frequent on the convex hull. Additionally, In, Mn, and Ga also exhibit a high frequency. These observations are sensible since all these elements frequently appear in the +3 oxidation state in perovskites, as expected for the prototype formula of these materials. Bismuth, which forms well-known $A_3B_2X_9$ perovskites, is not frequent in the filtered structures. This is expected and can be attributed to the larger cation size, making it unfit to form these 2D layered structures with most other cations considered in the screening, leading to a lower count.

However, when observing the B-cation combinations with the lowest E_{stab} in *Figure 44(b)*, we notice that Bi-containing perovskites, although fewer in number, tend to form quite stable structures, aligning with experimental observations. Sb-containing perovskites also present quite stable structures, but even lower E_{stab} structures appear in compounds containing Y and Sc. Previous work (Exner et al. 2024) has shown that Sc doping presents a stabilizing effect in the $Cs_3Sb_2I_9$ lattice. The model also perceives the chemically similar yttrium as a stabilizer for (111)-type perovskites. Other elements previously considered for Sb substitution due to similar cation size also appear as potentially stable structures, such as Ag, In, Mo, and Nb. Additionally, the model identifies Ga, Ge, and Cr as potentially stable in the B-site of these perovskites, a novel observation to the best of our knowledge. When comparing these observations with the lowest band gap predicted for every B cation combination in *Figure 44(c)*, we observe that perovskites containing transition metals V, Mn, Fe, and Cu, although predicted to produce the lowest band gaps, are not favored in stability. A better compromise is observed for Ag, Cr, and In-containing materials.

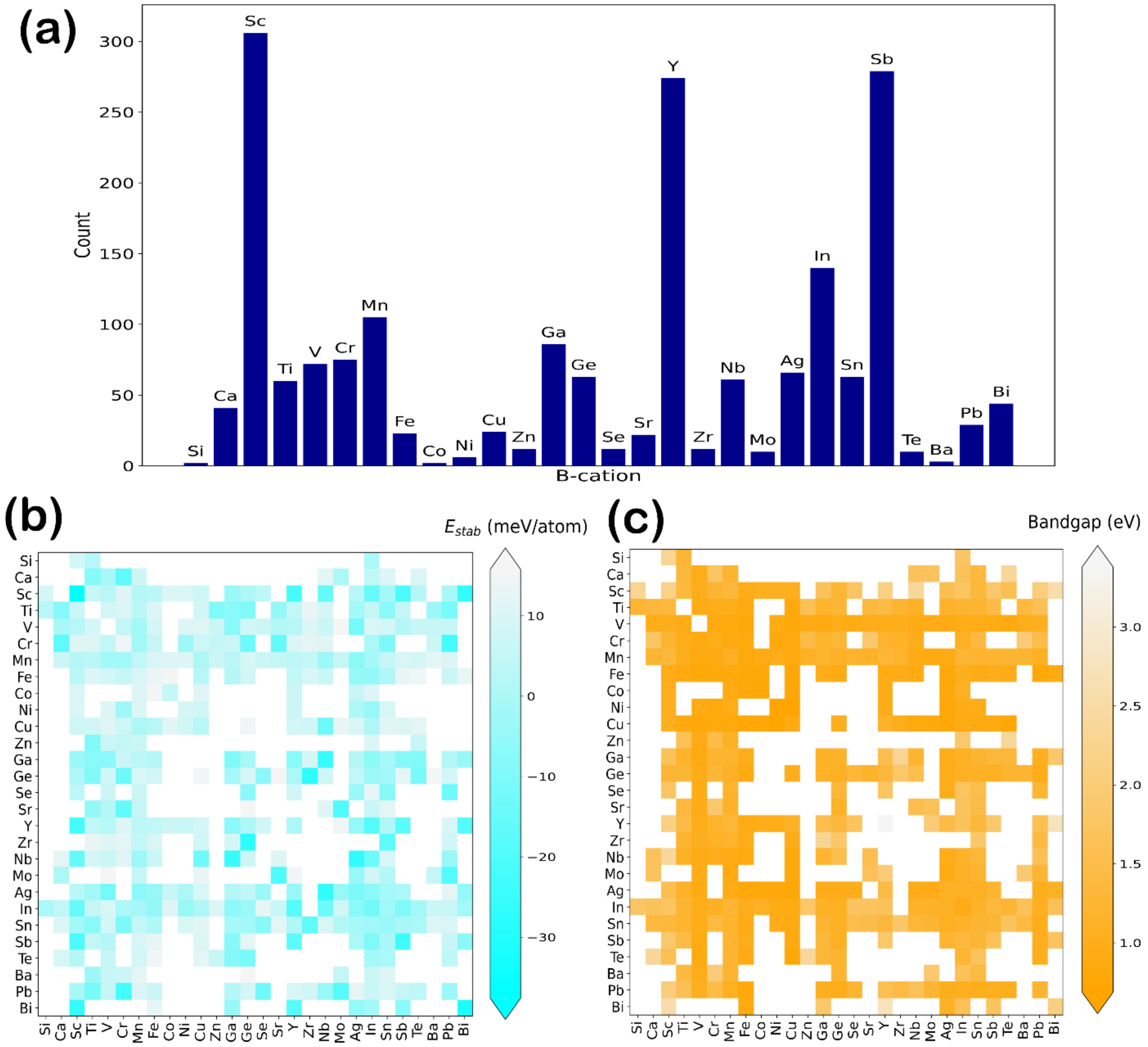


Figure 44 – Overview of distribution and estimated properties for the structures selected by screening through the structure-based ML models. Panel (a) shows *B-site element frequency for most stable structures (predicted E_{stab} < 5 meV/atom). Panels (b) and (c) present the lowest E_{stab} and band gap values, respectively, for structures in each possible combination of B cations.*

Another interesting observation concerns the presence of chalcogens in the screened structures. Only 30 structures passing the screening contained sulfur, and merely 6 structures contained oxygen. Moreover, although the screening encompassed structures containing up to one-third of the anion sites occupied by chalcogens (3 out of 9 X-sites in the $A_3B_2X_9$ unit formula), the final pool predominantly contained chalcogen-containing perovskites with a single chalcogen element per unit formula. In fact, there was a single structure containing two sulfur atoms per unit formula (see *Table 23*). These observations align with the challenging incorporation of these elements to form chalcohalides, which is a well-documented challenge in the literature (F. Hong et al. 2016; Theofylaktos et al. 2019).

As a final screening step, CHGNet was employed to obtain PhDOS which were analyzed for all structures from the previous phase. We observed that experimentally reported structures exhibited negative frequencies with absolute values consistently lower than those of oxygen-containing structures and structures with multiple substitutions, as shown in *Table E6*. This observation is indicative of improved dynamical stability, as multiple substitutions and oxygen incorporation are common destabilizing factors for halide perovskites (Chonamada, Dey, and Santra 2020; Aristidou et al. 2017). Therefore, based on the values of the minimum phonon frequency (ω_{min}) for the experimentally reported structures in *Table 23*, a threshold of -0.35 THz was established to filter structures by estimated dynamical stability. This process resulted in a final pool of 2991 candidate structures.

The final pool of materials still contains an extensive number of structures for high-throughput DFT calculations. However, considering the current precision of employed machine learning methods and the unaccounted kinetic stabilization, there's no specific justification for imposing stricter thresholds. Therefore, we adopted a strategy to sample structures for ab-initio calculation based on their respective group of B-site compositions. This procedure is adopted because the transition metal in the B-site typically plays the most defining role in perovskite properties. We illustrate this method of grouping the structures in *Table 24*, presenting structures for various B-site compositions.

Table 24 – Predicted stability and band gap for selected (111)-type structures screened in this work, results are presented for specific groups of B-site composition sorted in order of increasing predicted E_{stab}.

Composition of (111)-type structures	Predictions from structure-based MODNet models			Composition of (111)-type structures	Predictions from structure-based MODNet models		
	E_{stab} (meV/atom)	p_{semi}	E_g (eV)		E_{stab} (meV/atom)	p_{semi}	E_g (eV)
B-site: (Bi,Sb)				**B-site: (In), 6 most stable**			
$Cs_3BiSbBr_9$	-7.3	0.76	2.02	$Cs_3In_2Cl_9$	-12.6	0.72	2.64
$Cs_3Bi_{1.5}Sb_{0.5}Br_9$	-7.1	0.80	2.18	$Cs_3In_2Br_9$	-12.6	0.72	1.81
$Rb_3Bi_{1.5}Sb_{0.5}Br_9$	-2.5	0.68	2.29	$Rb_3In_2Br_9$	-11.1	0.52	1.63
$Cs_2RbBi_{1.5}Sb_{0.5}Br_9$	-1.3	0.80	2.22	$Rb_3In_2Cl_9$	-8.62	0.80	2.83
$Cs_3Bi_{0.5}Sb_{1.5}Br_9$	-0.8	0.80	2.01	$Cs_2RbIn_2Br_9$	-5.45	0.84	1.72
$Cs_3Bi_{1.5}Sb_{0.5}I_9$	-0.8	0.84	1.76	$Cs_3In_2I_9$	-1.33	0.88	0.92
$Cs_3Bi_{0.5}Sb_{1.5}I_9$	-0.6	0.88	1.73	**B-site: (In,Sc), most stable**			
$Cs_2RbBiSbBr_9$	3.4	0.72	2.13	$Cs_3Sc_{0.5}In_{1.5}Br_9$	-15.2	0.92	2.26
B-site: (Bi,Ga)				**B-site: (In,Y), 2 most stable**			
$Cs_3GaBiBr_9$	-4.3	0.84	2.16	Cs_3YInBr_9	-19.1	0.84	2.76
$Cs_3Ga_{1.5}Bi_{0.5}Br_9$	-0.2	0.88	1.85	$Cs_3Y_{0.5}In_{1.5}Br_9$	-18.2	0.84	2.24
B-site: (Ga), most stable				**B-site: (In,Sb), 2 most stable**			
$Cs_3Ga_2Br_9$	-9.2	0.72	1.67	$Cs_3In_{1.5}Sb_{0.5}Br_9$	-9.2	0.84	1.81
B-site: (Cr), most stable				$Cs_3InSbBr_9$	-9.0	0.80	1.85
$Rb_3Cr_2Br_9$	14.1	0.60	0.95	$Cs_3In_{0.5}Sb_{1.5}Br_9$	-7.8	0.80	1.72
B-site: (Co), most stable				**B-site: (Sc,Ni), most stable**			
$Rb_3Co_2Cl_9$	5.1	0.88	0.90	$Cs_2RbSc_{1.5}Ni_{0.5}(BrCl_2)_3$	10.0	0.88	2.05
B-site: (Fe,Sb), most stable				**B-site: (Y,Sn), 2 most stable**			
$Cs_3FeSb(Br_2Cl)_3$	12.8	0.76	0.98	$Cs_3Y_{1.5}Sn_{0.5}Br_5Cl_4$	0.9	0.92	3.4
B-site: (Sc,Nb), most stable				$Cs_2RbY_{1.5}Sn_{0.5}Br_5Cl_4$	2.4	0.92	3.3
$Cs_3Sc_{1.5}Nb_{0.5}SBr_8$	-15.5	0.88	2.43	**B-site: (Y,Fe), 2 most stable**			
B-site: (Sc,Ag), most stable				$Cs_3Y_{1.5}Fe_{0.5}Br_4Cl_5$	4.5	0.76	1.73
$Cs_3ScAgBr_7Cl_2$	2.4	0.80	1.86	$Cs_3YFeSBr_8$	4.8	0.84	1.20

In *Table 24*, the structures screened for the B-site composition containing Bi and Sb reveal the system of structures $Cs_3Bi_xSb_{1-x}Br_9$ as quite stable, which finds corroboration in recent experimental literature for these materials (Giovilli et al. 2023).

Additionally, we can observe the presence of the lowest band gap representative, $Cs_3Bi_{0.5}Sb_{1.5}I_9$, which also agrees with experiments for these iodine perovskites (G. Chen et al. 2020). Moreover, the screening ruled out the presence of $Cs_3Bi_2I_9$ which cannot form the layered perovskite structure and appears in experiments solely as a 0D perovskite. Indium-based perovskites are also suggested to be quite stable. For example, the $Cs_3In_2X_9$ (X= Cl, Br, I) compounds, which were previously reported in a ab-initio screening for (111)-type perovskites (W. H. Guo et al. 2020b). However, due to the greater flexibility, our method also suggests $Cs_3In_{1.5}Sb_{0.5}Br_9$, Cs_3YInBr_9, $Cs_3Sc_{0.5}In_{1.5}Br_9$, which may have better chance of being accomplished experimentally since no reports on the synthesis of $Cs_3In_2Br_9$ ($P\bar{3}m1$) are known to us. The role of Sc and Y as stabilizers is reinforced and aligns with previous experiments which have shown benefit in incorporating these elements in lead-halide perovskites. For example, the Sc addition improved the morphology and carrier lifetimes of $MAPbI_{3-x}Cl_x$ films (Shufang Li et al. 2020) and Y has been found to improve crystallinity and power conversion efficiencies when added to $CsPbBr_3$ and $CsPbI_3$ to enhance the PL of perovskite LEDs (Q. Wang et al. 2019). Moreover, the Ag-containing $Cs_3ScAgBr_7Cl_2$ also ranks high in stability and low in band gap this aligns with Ag and In being B cations consistently investigated for low band gap double perovskites (Menedjhi et al. 2021; Z. Liu et al. 2021).

Novel compositions are seen involving gallium and the frequently studied antimony and bismuth for these 2D perovskites. Compounds such as $Cs_3Ga_{1.5}Bi_{0.5}Br_9$ and $Cs_3Ga_{1.5}Sb_{0.5}Br_9$ are predicted to be stable with fairly low band gaps, aligning with very recent reports incorporating gallium with success in the inorganic $Cs_2AgBiBr_6$ to increase optical absorption (Ihtisham-ul-haq et al. 2024). We also highlight other structures with predicted low band gaps following the prototype $A_3B_2X_9$ structure such as $Cs_3Ga_2Br_9$, $Rb_3Cr_2Br_9$ and $Rb_3Co_2Cl_9$, to be investigated. Another important trend observed in the model predictions is the presence of mixed anions and also mixed-A cations promoting stability when the B-cations present in the structure differ in their usual oxidation state and/or ionic radius. This is a common mechanism to engineer perovskites (Hu et al. 2019) which the models are able to grasp and is demonstrated throughout the results in *Table 24*. Examples include $Cs_3FeSb(Br_2Cl)_3$, $Cs_3Y_{1.5}Sn_{0.5}Br_5Cl_4$, $Cs_2RbSc_{1.5}Ni_{0.5}(BrCl_2)_3$, $Cs_3Y_{1.5}Fe_{0.5}Br_4Cl_5$ and $Cs_3Sc_{1.5}Nb_{0.5}SBr_8$.

Overall, the proposed screening method filtered a significant number of B-cation element combinations deemed unfavorable, as indicated by the missing spots in *Figure 44(b),* but it still suggests numerous structures and systems worth further investigation. In *Figure 45*, we present ab-initio calculated band structures for a preliminary subset of the screened structures, whose stability against OQMD's convex hull is detailed in *Table 25*. These calculations revealed that the ternary compounds $Cs_3Ga_2Br_9$ and $Rb_3Cr_2Br_9$ exhibit band gaps lower than those of the more commonly studied $Cs_3Sb_2Br_9$ and $Cs_3Bi_2Br_9$, along with direct/nearly-direct band gaps. Notably, $Cs_3Ga_2Br_9$ demonstrates more dispersive valence and conduction bands, implying superior transport properties compared to $Rb_3Cr_2Br_9$. However, $Cs_3Ga_2Br_9$ has a decomposition energy of 32.6 meV/atom, beyond the expected range for stability, while $Rb_3Cr_2Br_9$ exhibits a decomposition energy of 14.3 meV/atom, within the usual error margin attributed to DFT-assessed stability. Despite this, analyzing $Cs_3GaBiBr_9$ reveals a negative decomposition energy with a relatively small band gap, further highlighting the potential of Ga-containing (111)-type perovskites for exploration. Additionally, $Cs_3Bi_{0.5}Sb_{1.5}Br_9$, $Cs_3Sc_{0.5}In_{1.5}Br_9$ and $Cs_3GaBiBr_9$ exhibit direct/nearly-direct band gaps close to the values predicted by the band gap regressor model, with the first two demonstrating more dispersive conduction and valence bands. These structures also display negative decomposition energies, indicating thermodynamic stability within the given convex hull. Another noteworthy candidate is $Cs_3Y_{0.5}Fe_{1.5}Br_9$, boasting a direct band gap of 1 eV, deviating 0.7 eV from the predicted value of 1.7 eV. Furthermore, this structure exhibits a localized state with spin inversion in the conduction band, suggesting potential applications in magnetic storage devices or spintronics if synthesizable. Additional ab-initio band structure calculations for selected compounds are presented in *Figure E9* and *E10.*

Although the ab-initio evaluation is preliminary, it already confirms the trends of enhanced stability for Y and Sc-containing (111)-type perovskites and offers perspective on novel chemical systems containing Ga and Cr for this class of materials. Moreover, from the structures sampled in these preliminary calculations, the predicted stability and the stability evaluated with ab-initio methods, presented in *Table 25,* show good qualitative agreement, with only a few structures wrongly predicted within the stability range. Namely, $Cs_3Ni_2Br_9$, $Cs_2RbSc_{1.5}Ni_{0.5}(BrCl_2)_3$,

$Cs_2RbSc_{1.5}Ag_{0.5}(Br_2Cl)_3$, and the previously mentioned $Cs_3Ga_2Br_9$. This observation underscores the power of the proposed framework for materials discovery.

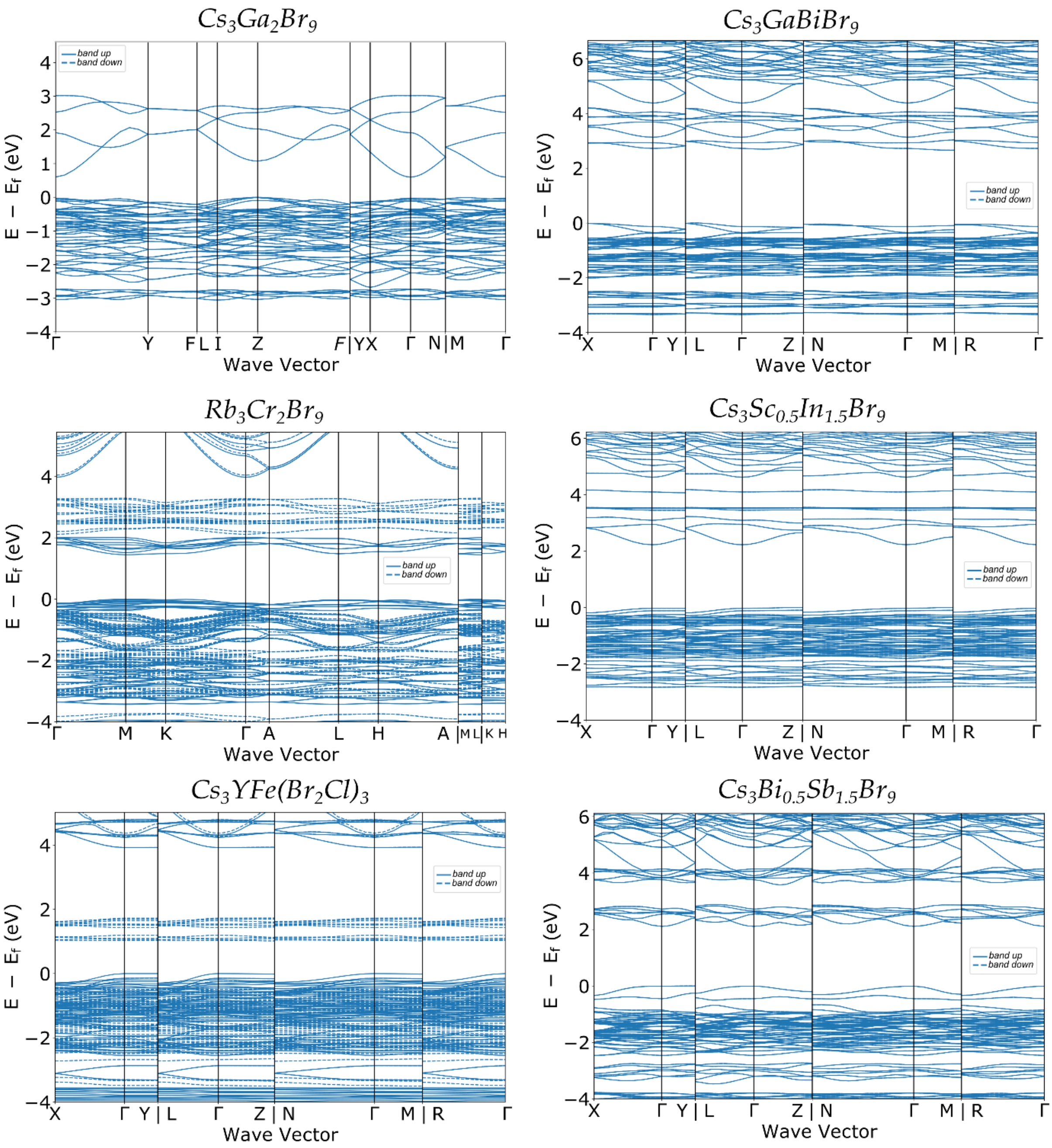


Figure 45 – Electronic band structure of a selection of the ML screened structures obtained in this work.

Table 25 – Decomposition energy (E_{stab}) determined from ab-initio calculations utilizing OQMD's convex hull for selected (111)-type structures screened in this work. The predicted decomposition energy from our structure-based model is presented for comparison as E_{stab}^{ML}.

Composition of (111)-type structures	E_{stab} (meV/atom)	E_{stab}^{ML} (meV/atom)
$Cs_3Ga_2Br_9$	32.6	−9.1
$Rb_3Cr_2Br_9$	14.3	14.1
$Cs_3YFe(Br_2Cl)_3$	−34.3	6.3
$Cs_3GaBiBr_9$	−17.7	−4.3
$Cs_3Sc_{0.5}In_{1.5}Br_9$	−28.0	−15.2
$Cs_3Bi_{0.5}Sb_{1.5}Br_9$	−2.7	−0.8
$Rb_3Sc_2Cl_9$	−6.4	3.2
$Rb_3Cr_2Br_9$	−43.0	−20.3
$Cs_3Co_2Br_9$	3.1	−3.0
$Rb_3Co_2Br_9$	−8.0	−2.0
$Cs_3Ni_2Br_9$	53.0	−41.0
$Cs_3Sc_{1.5}Nb_{0.5}SBr_8$	7.2	14.6
$Cs_3Sc_{1.5}V_{0.5}SBr_8$	34.8	22.0
$Cs_2RbSc_{1.5}Ni_{0.5}(BrCl_2)_3$	49.8	−27.0
$Cs_3Sc_{1.5}Co_{0.5}(Br_2Cl)_3$	3.9	−13.1
$Cs_2RbSc_{1.5}Ag_{0.5}(Br_2Cl)_3$	41.0	−21.0
$Cs_3In_{1.5}Ge_{0.5}Br_9$	6.9	−47.2
$Cs_3Y_{0.5}Sb_{1.5}Br_9$	−8.5	−8.2
$Cs_3FeSb(Br_2Cl)_3$	12.1	0.0

6.4 CONCLUSION

Starting from over 100 million candidate structures for (111)-type halide perovskites the framework proposed on this work leveraged the power of MODNet model enhanced with an advanced featurizer and CHGNet MLIP to reach a final pool of structures that reflected experimentally observed trends of perovskite formability and band gap. The selection of structures which were evaluated with DFT calculations revealed possibly missed ternary structures such as $Cs_3Ga_2Br_9$ and $Rb_3Cr_2Br_9$ which present lower band gaps than the more explored $Cs_3Sb_2Br_9$ and $Cs_3Bi_2Br_9$. Structures containing Sc and Y were predicted as stabilizers for (111)-type perovskites. Mixed A-cations and anions also figured to be potential stabilizers of these structures, contributing also to tuning band gap and carrier effective masses.

Overall, the results obtained showcase the ability of our method to successfully screen a large compositional space considering multiple atomic configurations by leveraging the power of machine learning and big data. Several interesting chemical systems can be found much more easily using this approach allowing for accelerated materials discovery. Through intuitive chemical descriptors and enhanced flexibility, this method provides several candidate compositions to stabilize perovskites through multi-site doping and can be used as a well-informed guide for experimentalists to achieve tailored material properties.

CHAPTER 7

CONCLUSION AND FUTURE PROSPECTS

This thesis investigated the promising class of lead-free 2D MHIPs using computational methods based on ab-initio properties. DFT was harnessed to unravel the atomistic origins of experimentally observed MHIP properties and identify pathways for optimizing these materials through compositional tuning and structural manipulation. Furthermore, by leveraging cutting-edge ML tools trained on DFT data, our research predicted novel MHIPs with targeted optoelectronic properties, accounting for the possibility of multi-element alloying/doping for fine-tuning these properties. This combined approach opens avenues for exploring advanced MHIP materials with precise control over functionalities. Additionally, it provides valuable guidance to experimentalists for their discovery through intuitive chemical descriptors and decomposition energy estimates. We review our main findings and offer prospects for further research on this theme in the next paragraphs.

Chapter 3 investigated the currently most representative subgroup of MHIPs, the cesium antimony halide perovskites $Cs_3Sb_2X_9$ (X= Cl, Br, I) (space group: $P\bar{3}m1$). Our investigation into halide mixing revealed insights into band gap variations and structural shifts, indicating potential ordered structures. We found that (1000) surfaces retain electronic properties beneficial for photovoltaics, while (0001) surfaces exhibit significant band gap reduction, suggesting reactivity suitable for photocatalysis. Efficient LEDs may be obtained from $Cs_3Sb_2Br_9$@$Cs_3Sb_2Cl_9$ interfaces, utilizing the chlorine shell as a diffusion barrier. Defect tolerance was observed in $Cs_3Sb_2I_9|Cs_3Sb_2Br_9$ interface, valuable for photovoltaics. Additionally, cluster simulations suggested geometry's role in photoluminescence observed experimentally, with edge sites crucial for band gap tuning. Overall, our simulations showcase the predictive power of DFT for in-depth comprehension of this group of materials exploring the effects of compositional tuning, surface engineering, interface formation and dimensionality reduction. This investigation has practical implications for further optimization of $Cs_3Sb_2X_9$ perovskites for stable and efficient solar cells and optoelectronic devices.

In *Chapter 4,* the effects of metal and halogen doping on both polymorphs of $Cs_3Sb_2I_9$ (space groups $P\bar{3}m1$ and $P6_3/mmc$) were explored. The goal was to enhance stability and optical absorption of this material, which has the lowest band gap among $Cs_3Sb_2X_9$ perovskites. Only substitutional metals with a +3 oxidation state ion and comparable ionic radius to Sb^{3+} were considered. In-doped $Cs_3Sb_{1.5}In_{0.5}I_9$ showed potential stability with increased optical absorption, yielding band gaps of approximately 1.5 eV ($P6_3/mmc$) and 1.7 eV ($P\bar{3}m1$), as per ACBN0 calculated band structures. Additionally, Sc-doped $Cs_3Sb_{1.5}Sc_{0.5}I_9$ contributed to lattice stabilization with only a slight increase in band gap, suggesting a method to reduce Urbach energy and improve device performance. A similar observation is seen on halogen doping which exhibited negative formation energy with a significant lattice shrinkage compared to the pristine $Cs_3Sb_2I_9$ structure. Halogen doping also led to lattice shrinkage and negative formation energy, indicating potential for combining halogen and Sc doping to incorporate other metals such as Ag, Mo, Nb, and Bi, which exhibited positive decomposition energies when incorporated alone. Collectively, these observations triggered our interest in investigating concomitant doping in the elemental sites of these perovskites to optimize their performance. Our exploration of the vast chemical space of doping within these halide perovskites has been facilitated by a novel methodology integrating state-of-the-art machine learning techniques, culminating in our subsequent articles.

Machine learning plays a crucial role in materials discovery, with two prominent approaches: feature-based and graph-based models. Feature-based models are interpretable and efficient for small datasets, but their electronic structure featurizers can be computationally expensive and introduce high dimensionality. Conversely, graph-based models excel with large datasets but lack interpretability. The research devised in *Chapter 5* addresses this by proposing a method to bridge the gap. We leverage pre-trained graph models to generate informative features for feature-based models. This approach aims to achieve high accuracy while retaining interpretability, vital for guiding materials design. The newly developed OMEGA featurizer utilizes pre-trained graph models to generate features, consistently improving prediction accuracy. Furthermore, the interpretability of feature-based models is partially retained through techniques like SHAP analysis, allowing researchers to extract chemical insights from the generated features. This paves the way for a more efficient and interpretable materials discovery process, facilitating large-scale chemical space exploration and chemically interpretable active learning.

In *Chapter 6*, a ML workflow was developed to screen over 100 million candidate structures for (111)-type halide perovskites, leveraging MODNet model with an advanced featurizer and CHGNet MLIP. This approach successfully identified a final pool of structures reflecting experimentally observed trends in perovskite formability and band gap. The selection of structures, evaluated through DFT calculations, revealed potentially overlooked ternary compounds like $Cs_3Ga_2Br_9$ and $Rb_3Cr_2Br_9$ with lower band gaps than commonly studied $Cs_3Sb_2Br_9$ and $Cs_3Bi_2Br_9$. Furthermore, the prediction of Sc and Y as stabilizers for (111)-type perovskites, along with the identification of mixed A-cations and anions as potential stabilizers, showcased the method's ability to tune band gaps and carrier effective masses. This study demonstrates the effectiveness of ML and big data in accelerating materials discovery, providing valuable insights for designing tailored perovskite materials for various applications.

Future prospects of this work include a more comprehensive exploration of the (111)-type perovskite structures identified by the proposed machine learning screening method, particularly those involving gallium and chromium in their composition. A focused evaluation of ab-initio properties of these structures, including alloying, is necessary to guide experimental efforts for their obtention. While our machine learning screening framework has demonstrated success, we anticipate improved accuracy through training the models on larger datasets utilized in state-of-the-art GNN models. Additionally, enhancing the proposed featurizer for the MODNet model including more recent GNN featurizers beyond MEGNet and encoding electronic descriptors in datasets larger than the Materials Project would be beneficial. Following the same rationale, utilizing adjacent GNN models including attention mechanisms may provide improved accuracy and facilitate the interpretability of GNN features. Another option is to train elemental embeddings alongside the chemical descriptors to provide an additional layer of interpretable information for the MODNet models.

REFERENCES

‘3.1. Cross-Validation’. 2023. Scikit-Learn. 2023. https://scikit-learn/stable/modules/cross_validation.html.

‘6.3. Preprocessing Data’. 2023. Scikit-Learn. 2023. https://scikit-learn/stable/modules/preprocessing.html.

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. ‘TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems’. arXiv. https://doi.org/10.48550/arXiv.1603.04467.

Acharyya, Paribesh, Kaushik Kundu, and Kanishka Biswas. 2020. ‘2D Layered All-Inorganic Halide Perovskites: Recent Trends in Their Structure, Synthesis and Properties’. *Nanoscale* 12 (41): 21094–117. https://doi.org/10.1039/D0NR06138G.

Agapito, Luis A., Stefano Curtarolo, and Marco Buongiorno Nardelli. 2015. ‘Reformulation of DFT + U as a Pseudohybrid Hubbard Density Functional for Accelerated Materials Discovery’. *Physical Review X* 5 (1): 1–16. https://doi.org/10.1103/PhysRevX.5.011006.

Agapito, Luis A., Andrea Ferretti, Arrigo Calzolari, Stefano Curtarolo, and Marco Buongiorno Nardelli. 2013. ‘Effective and Accurate Representation of Extended Bloch States on Finite Hilbert Spaces’. *Physical Review B* 88 (16): 165127. https://doi.org/10.1103/PhysRevB.88.165127.

Agrawal, Ankit, and Alok Choudhary. 2016. ‘Perspective: Materials Informatics and Big Data: Realization of the “Fourth Paradigm” of Science in Materials Science’. *APL Materials* 4 (5): 053208. https://doi.org/10.1063/1.4946894.

Alex Ganose, Janosh Riebesell, J. George, Jimmy Shen, Andrew S. Rosen, Aakash Ashok Naik, nwinner, et al. 2024. ‘Materialsproject/Atomate2: V0.0.13’. [object Object]. https://doi.org/10.5281/ZENODO.10677081.

Alloghani, Mohamed, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J. Aljaaf. 2020. ‘A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science’. In *Supervised and Unsupervised Learning for Data Science*, edited by Michael W. Berry, Azlinah Mohamed, and Bee Wah Yap, 3–21. Unsupervised and Semi-Supervised Learning. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1.

Alzubaidi, Laith, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. ‘Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions’. *Journal of Big Data* 8 (1): 53. https://doi.org/10.1186/s40537-021-00444-8.

Amerling, Eric, Haipeng Lu, Bryon W. Larson, Annalise E. Maughan, Alan Phillips, Evan Lafalce, Luisa Whittaker-Brooks, et al. 2021. ‘A Multi-Dimensional Perspective on Electronic Doping in Metal Halide Perovskites’. *ACS Energy Letters* 6 (3): 1104–23. https://doi.org/10.1021/acsenergylett.0c02476.

Anisimov, Vladimir I, F Aryasetiawan, and A I Lichtenstein. 1997. ‘First-Principles Calculations of the Electronic Structure and Spectra of Strongly Correlated Systems: The **LDA** + *U* Method’. *Journal of Physics: Condensed Matter* 9 (4): 767–808. https://doi.org/10.1088/0953-8984/9/4/002.

Antoniuk, Evan R., Gowoon Cheon, George Wang, Daniel Bernstein, William Cai, and Evan J. Reed. 2023. ‘Predicting the Synthesizability of Crystalline Inorganic Materials from the Data of Known Material Compositions’. *Npj Computational Materials* 9 (1): 1–11. https://doi.org/10.1038/s41524-023-01114-4.

Arakcheeva, A. V., M. S. Novikova, A. I. Zaitsev, G. U. Lubman, Russian Academy, and Siberian Branch. 1999. ‘Perovskite-like Modification of $Cs_3Sb_2I_9$ as a Member of the 0D Family $A_3B_2X_9$’. *Journal of Structural Chemistry* 40 (4): 572–79.

Ardabili, Sina, Amir Mosavi, Majid Dehghani, and Annamária R. Várkonyi-Kóczy. 2020. ‘Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review’. In *Engineering for Sustainable Future,* edited by Annamária R. Várkonyi-Kóczy, 52–62. Lecture Notes in Networks and Systems. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36841-8_5.

Aristidou, Nicholas, Christopher Eames, Irene Sanchez-Molina, Xiangnan Bu, Jan Kosco, M. Saiful Islam, and Saif A. Haque. 2017. ‘Fast Oxygen Diffusion and Iodide Defects Mediate Oxygen-Induced Degradation of Perovskite Solar Cells’. *Nature Communications* 8 (May): 1–10. https://doi.org/10.1038/ncomms15218.

Aryasetiawan, F., K. Karlsson, O. Jepsen, and U. Schönberger. 2006. ‘Calculations of Hubbard U from First-Principles’. *Physical Review B - Condensed Matter and Materials Physics* 74 (12): 125106. https://doi.org/10.1103/PhysRevB.74.125106.

Awad, Mariette, and Rahul Khanna. 2015. ‘Deep Neural Networks’. In *Efficient Learning Machines*, by Mariette Awad and Rahul Khanna, 127–47. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_7.

Bacalis, N., E. N. Economou, and M. H. Cohen. 1988. ‘Simple Derivation of Exponential Tails in the Density of States’. *Physical Review B* 37 (5): 2714–17. https://doi.org/10.1103/PhysRevB.37.2714.

Bader, R. F. W. 1990. ‘Atoms in Molecules: A Quantum Theory’. *Clarendon Press Oxford AE Reed, LA Curtiss F Weinhold, Chem Rev* 88: 899.

Bae, Su-Yong, Su Young Lee, Ji-wan Kim, Ha Nee Umh, Jaeseong Jeong, Seongjun Bae, Jongheop Yi, Younghun Kim, and Jinhee Choi. 2019. ‘Hazard Potential of Perovskite Solar Cell Technology for Potential Implementation of “Safe-by-Design” Approach’. *Scientific Reports* 9 (1): 4242. https://doi.org/10.1038/s41598-018-37229-8.

Bae, Wan Ki, Young Shin Park, Jaehoon Lim, Donggu Lee, Lazaro A. Padilha, Hunter McDaniel, Istvan Robel, Changhee Lee, Jeffrey M. Pietryga, and Victor I. Klimov. 2013. ‘Controlling the Influence of Auger Recombination on the Performance of Quantum-Dot Light-Emitting Diodes’. *Nature Communications* 4: 1–8. https://doi.org/10.1038/ncomms3661.

Bala, Anu, and Vijay Kumar. 2021. ‘Direct Band Gap Halide-Double-Perovskite Absorbers for Solar Cells and Light Emitting Diodes: *Ab Initio* Study of Bulk and Layers’. *Physical Review Materials* 5 (9): 095401. https://doi.org/10.1103/PhysRevMaterials.5.095401.

Balachandran, Prasanna V., Antoine A. Emery, James E. Gubernatis, Turab Lookman, Chris Wolverton, and Alex Zunger. 2018. ‘Predictions of New ABO_3 Perovskite Compounds by Combining Machine Learning and Density Functional Theory’. *Physical Review Materials* 2 (4): 043802. https://doi.org/10.1103/PhysRevMaterials.2.043802.

Banerjee, Amartya S., Phanish Suryanarayana, and John E. Pask. 2016. ‘Periodic Pulay Method for Robust and Efficient Convergence Acceleration of Self-Consistent Field Iterations’. *Chemical Physics Letters* 647 (March): 31–35. https://doi.org/10.1016/j.cplett.2016.01.033.

Banerjee, Arnab, and Goutam Paul. 2020. ‘Room-Temperature Magnetoresistance in Hybrid Halide Perovskites: Effect of Spin-Orbit Coupling’. *Physical Review Applied* 14 (6): 064018. https://doi.org/10.1103/PhysRevApplied.14.064018.

Barber, C. Bradford, David P. Dobkin, and Hannu Huhdanpaa. 1996. ‘The Quickhull Algorithm for Convex Hulls’. *ACM Transactions on Mathematical Software* 22 (4): 469–83. https://doi.org/10.1145/235815.235821.

Barrett, David Gt, Ari S Morcos, and Jakob H Macke. 2019. ‘Analyzing Biological and Artificial Neural Networks: Challenges with Opportunities for Synergy?’ *Current Opinion in Neurobiology* 55 (April): 55–64. https://doi.org/10.1016/j.conb.2019.01.007.

Bartel, Christopher J. 2022. ‘Review of Computational Approaches to Predict the Thermodynamic Stability of Inorganic Solids’. *Journal of Materials Science* 57 (23): 10475–98. https://doi.org/10.1007/s10853-022-06915-4.

Bartel, Christopher J., Alan W. Weimer, Stephan Lany, Charles B. Musgrave, and Aaron M. Holder. 2019. ‘The Role of Decomposition Reactions in Assessing First-Principles Predictions of Solid Stability’. *Npj Computational Materials* 5 (1): 1–9. https://doi.org/10.1038/s41524-018-0143-2.

Barth, U Von, and L Hedin. 1972. ‘A Local Exchange-Correlation Potential for the Spin Polarized Case.’ *Journal of Physics C: Solid State Physics* 5 (13): 1629–42. https://doi.org/10.1088/0022-3719/5/13/012.

Bartók, Albert P., Risi Kondor, and Gábor Csányi. 2013. ‘On Representing Chemical Environments’. *Physical Review B* 87 (18): 184115. https://doi.org/10.1103/PhysRevB.87.184115.

Baydin, Atilim Gunes, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2018. ‘Automatic Differentiation in Machine Learning: A Survey’. *Journal of Marchine Learning Research* 18: 1–43.

Becke, Axel D. 1993. ‘A New Mixing of Hartree–Fock and Local Density-Functional Theories’. *The Journal of Chemical Physics* 98 (2): 1372–77. https://doi.org/10.1063/1.464304.

———. 2014. ‘Perspective: Fifty Years of Density-Functional Theory in Chemical Physics’. *The Journal of Chemical Physics* 140 (18): 18A301. https://doi.org/10.1063/1.4869598.

Bellizzi, Saverio, Christian Popescu, Catello M Panu Napodano, Maura Fiamma, and Luca Cegolon. 2023. ‘Global Health, Climate Change and Migration: The Need for Recognition of “Climate Refugees”’. *Journal of Global Health* 13 (March): 03011. https://doi.org/10.7189/jogh.13.03011.

Belsky, Alec, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. 2002. ‘New Developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in Support of Materials Research and Design’. *Acta Crystallographica Section B Structural Science* 58 (3): 364–69. https://doi.org/10.1107/S0108768102006948.

Berri, Saadi. 2020. ‘Theoretical Analysis of the Structural, Electronic, Optical and Thermodynamic Properties of Trigonal and Hexagonal Cs3Sb2I9 Compound’. *European Physical Journal B* 93 (10): 1–12. https://doi.org/10.1140/epjb/e2020-10143-1.

Billeter, Salomon R., Alessandro Curioni, and Wanda Andreoni. 2003. ‘Efficient Linear Scaling Geometry Optimization and Transition-State Search for Direct Wavefunction Optimization Schemes in Density Functional Theory Using a Plane-Wave Basis’. *Computational Materials Science* 27 (4): 437–45. https://doi.org/10.1016/S0927-0256(03)00043-0.

Blancon, Jean-Christophe, Jacky Even, Costas. C. Stoumpos, Mercouri. G. Kanatzidis, and Aditya D. Mohite. 2020. ‘Semiconductor Physics of Organic–Inorganic 2D Halide Perovskites’. *Nature Nanotechnology* 15 (12): 969–85. https://doi.org/10.1038/s41565-020-00811-1.

Blasse, G. 1983. ‘The Luminescence of $Cs_3Bi_2Cl_9$ and $Cs_3Sb_2Cl_9$’ 233: 222–33.

Bloch, Felix. 1929. ‘Über die Quantenmechanik der Elektronen in Kristallgittern’. *Zeitschrift für Physik* 52 (7–8): 555–600. https://doi.org/10.1007/BF01339455.

Blöchl, P. E. 1994. ‘Projector Augmented-Wave Method’. *Physical Review B* 50 (24): 17953–79. https://doi.org/10.1142/9789814365031_0023.

Blöchl, P. E., Johannes Kästner, and Clemens J Först. 2005. ‘Electronic Structure Methods: Augmented Waves, Pseudopotentials and the Projector Augmented Wave Method’. *Handbook of Materials Modeling: Methods,* 93–119.

Blöchl, Peter E., O. Jepsen, and O. K. Andersen. 1994. ‘Improved Tetrahedron Method for Brillouin-Zone Integrations’. *Physical Review B* 49 (23): 16223–33. https://doi.org/10.1103/PhysRevB.49.16223.

Borlido, Pedro, Jonathan Schmidt, Ahmad W. Huran, Fabien Tran, Miguel A. L. Marques, and Silvana Botti. 2020. ‘Exchange-Correlation Functionals for Band Gaps of Solids: Benchmark, Reparametrization and Machine Learning’. *Npj Computational Materials* 6 (1): 96. https://doi.org/10.1038/s41524-020-00360-0.

Born, Max, Kun Huang, and M. Lax. 1955. ‘Dynamical Theory of Crystal Lattices’. *American Journal of Physics* 23 (7): 474–474. https://doi.org/10.1119/1.1934059.

Brandt, Riley E., Vladan Stevanović, David S. Ginley, and Tonio Buonassisi. 2015. ‘Identifying Defect-Tolerant Semiconductors with High Minority-Carrier Lifetimes: Beyond Hybrid Lead Halide Perovskites’. *MRS Communications* 5 (2): 265–75. https://doi.org/10.1557/mrc.2015.26.

Breiman, Leo. 1996. ‘Bagging Predictors’. *Machine Learning* 24: 123–40.

Bronstein, Michael M., Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. ‘Geometric Deep Learning: Going beyond Euclidean Data’. *IEEE Signal Processing Magazine* 34 (4): 18–42. https://doi.org/10.1109/MSP.2017.2693418.

Bücker, Martin, George Corliss, Uwe Naumann, Paul Hovland, and Boyana Norris, eds. 2006. *Automatic Differentiation: Applications, Theory, and Implementations*. Vol. 50. Lecture Notes in Computational Science and Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-28438-9.

Cabello-Solorzano, Kelsy, Isabela Ortigosa De Araujo, Marco Peña, Luís Correia, and Antonio J. Tallón-Ballesteros. 2023. ‘The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis’. In *18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023),* edited by Pablo García Bringas, Hilde Pérez García, Francisco Javier Martínez De Pisón, Francisco Martínez Álvarez, Alicia Troncoso Lora, Álvaro Herrero, José Luis Calvo Rolle, Héctor Quintián, and Emilio Corchado, 750:344–53. Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42536-3_33.

Cai, Xia, Yiming Zhang, Zejiao Shi, Ying Chen, Yujie Xia, Anran Yu, Yuanfeng Xu, et al. 2022. ‘Discovery of Lead-Free Perovskites for High-Performance Solar Cells via Machine Learning: Ultrabroadband Absorption, Low Radiative Combination, and Enhanced Thermal Conductivities’. *Advanced Science* 9 (4): 2103648. https://doi.org/10.1002/advs.202103648.

Calzolari, Arrigo, and Marco Buongiorno Nardelli. 2013. ‘Dielectric Properties and Raman Spectra of {ZnO} from a First Principles Finite-Differences/Finite-Fields Approach’. *Sci. Rep.* 3 (October). https://doi.org/10.1038/srep02999.

Carr, D. Andrew, Mohammed Lach-hab, Shujiang Yang, Iosif I. Vaisman, and Estela Blaisten-Barojas. 2009. ‘Machine Learning Approach for Structure-Based Zeolite Classification’. *Microporous and Mesoporous Materials* 117 (1): 339–49. https://doi.org/10.1016/j.micromeso.2008.07.027.

Carrete, Jesús, Wu Li, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. 2014. ‘Finding Unprecedentedly Low-Thermal-Conductivity Half-Heusler Semiconductors via High-Throughput Materials Modeling’. *Physical Review X* 4 (1): 011019. https://doi.org/10.1103/PhysRevX.4.011019.

Ceperley, D. M., and B. J. Alder. 1980. ‘Ground State of the Electron Gas by a Stochastic Method’. *Physical Review Letters* 45 (7): 566–69. https://doi.org/10.1103/PhysRevLett.45.566.

Chatzigoulas, Alexios, Konstantina Karathanou, Dimitris Dellis, and Zoe Cournia. 2018. ‘NanoCrystal: A Web-Based Crystallographic Tool for the Construction of Nanoparticles Based on Their Crystal Habit’. *Journal of Chemical Information and Modeling* 58 (12): 2380–86. https://doi.org/10.1021/acs.jcim.8b00269.

Chen, Chi, and Shyue Ping Ong. 2022. ‘A Universal Graph Deep Learning Interatomic Potential for the Periodic Table’. *Nature Computational Science* 2 (11): 718–28. https://doi.org/10.1038/s43588-022-00349-3.

Chen, Chi, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. 2019. ‘Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals’. *Chemistry of Materials* 31 (9): 3564–72. https://doi.org/10.1021/acs.chemmater.9b01294.

Chen, Daqin, Zhongyi Wan, Xiao Chen, Yongjun Yuan, and Jiasong Zhong. 2016. ‘Large-Scale Room-Temperature Synthesis and Optical Properties of Perovskite-Related Cs 4 PbBr 6 Fluorophores’. *Journal of Materials Chemistry C* 4 (45): 10646–53. https://doi.org/10.1039/C6TC04036E.

Chen, Guoqiang, Peng Wang, Yaqiang Wu, Qianqian Zhang, Qian Wu, Zeyan Wang, Zhaoke Zheng, Yuanyuan Liu, Ying Dai, and Baibiao Huang. 2020. ‘Lead-Free Halide Perovskite $Cs_3Bi_{2x}Sb_{2-2x}I_9$ (x ≈ 0.3) Possessing the Photocatalytic Activity for Hydrogen Evolution Comparable to That of $(CH_3NH_3)PbI_3$’. *Advanced Materials* 32 (39): 1–7. https://doi.org/10.1002/adma.202001344.

Chen, Shan, Huajie Yin, Porun Liu, Yun Wang, and Huijun Zhao. 2023. ‘Stabilization and Performance Enhancement Strategies for Halide Perovskite Photocatalysts’. *Advanced Materials* 35 (6): 2203836. https://doi.org/10.1002/adma.202203836.

Chen, Tianqi, and Carlos Guestrin. 2016. ‘XGBoost: A Scalable Tree Boosting System’. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco California USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chen, Tianqi, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. ‘MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems’. arXiv. https://doi.org/10.48550/arXiv.1512.01274.

Chichibu, Shigefusa F., Akira Uedono, Takeyoshi Onuma, Benjamin A. Haskell, Arpan Chakraborty, Takahiro Koyama, Paul T. Fini, et al. 2006. ‘Origin of Defect-Insensitive Emission Probability in In-Containing (Al,In,Ga)N Alloy Semiconductors’. *Nature Materials* 5 (10): 810–16. https://doi.org/10.1038/nmat1726.

Chollet, François. 2015. ‘Keras’. https://keras.io.

Chonamada, Trupthi Devaiah, Arka Bikash Dey, and Pralay K. Santra. 2020. ‘Degradation Studies of $Cs_3Sb_2I_9$: A Lead-Free Perovskite’. *ACS Applied Energy Materials* 3 (1): 47–55. https://doi.org/10.1021/acsaem.9b01899.

Choubisa, Hitarth, Mikhail Askerka, Kevin Ryczko, Oleksandr Voznyy, Kyle Mills, Isaac Tamblyn, and Edward H. Sargent. 2020. ‘Crystal Site Feature Embedding Enables Exploration of Large Chemical Spaces’. *Matter* 3 (2): 433–48. https://doi.org/10.1016/j.matt.2020.04.016.

Choubisa, Hitarth, Petar Todorović, Joao M. Pina, Darshan H. Parmar, Ziliang Li, Oleksandr Voznyy, Isaac Tamblyn, and Edward H. Sargent. 2023. ‘Interpretable Discovery of Semiconductors with Machine Learning’. *Npj Computational Materials* 9 (1): 117. https://doi.org/10.1038/s41524-023-01066-9.

Choudhary, Kamal, and Brian DeCost. 2021. 'Atomistic Line Graph Neural Network for Improved Materials Property Predictions'. *Npj Computational Materials* 7 (1): 185. https://doi.org/10.1038/s41524-021-00650-1.

Choudhary, Kamal, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, et al. 2022. 'Recent Advances and Applications of Deep Learning Methods in Materials Science'. *Npj Computational Materials* 8 (1): 1–26. https://doi.org/10.1038/s41524-022-00734-6.

Cococcioni, Matteo, and Stefano de Gironcoli. 2005. 'Linear Response Approach to the Calculation of the Effective Interaction Parameters in the LDA+U Method'. *Physical Review B* 71 (3): 035105. https://doi.org/10.1103/PhysRevB.71.035105.

Conings, Bert, Jeroen Drijkoningen, Nicolas Gauquelin, Aslihan Babayigit, Jan D'Haen, Lien D'Olieslaeger, Anitha Ethirajan, et al. 2015. 'Intrinsic Thermal Instability of Methylammonium Lead Trihalide Perovskite'. *Advanced Energy Materials* 5 (15): 1500477. https://doi.org/10.1002/aenm.201500477.

Connor, Bridget A., Alexander C. Su, Adam H. Slavney, Linn Leppert, and Hemamala I. Karunadasa. 2023. 'Understanding the Evolution of Double Perovskite Band Structure upon Dimensional Reduction'. *Chemical Science* 14 (42): 11858–71. https://doi.org/10.1039/D3SC03105E.

Correa-Baena, Juan Pablo, Lea Nienhaus, Rachel C. Kurchin, Seong Sik Shin, Sarah Wieghold, Noor Titan Putri Hartono, Mariya Layurova, et al. 2018. 'A-Site Cation in Inorganic $A_3Sb_2I_9$ Perovskite Influences Structural Dimensionality, Exciton Binding Energy, and Solar Cell Performance'. *Chemistry of Materials* 30 (11): 3734–42. https://doi.org/10.1021/acs.chemmater.8b00676.

Cramer, Christopher J. 2013. *Essentials of Computational Chemistry: Theories and Models*. Second edition. Chichester, West Sussex, England: Wiley.

Curtarolo, Stefano, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. 2013. 'The High-Throughput Highway to Computational Materials Design'. *Nature Materials* 12 (3): 191–201. https://doi.org/10.1038/nmat3568.

Curtarolo, Stefano, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, et al. 2012. 'AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations'. *Computational Materials Science* 58 (June): 227–35. https://doi.org/10.1016/j.commatsci.2012.02.002.

Damewood, James, Jessica Karaguesian, Jaclyn R. Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. 2023. 'Representations of Materials for Machine Learning'. *Annual Review of Materials Research* 53 (1): 399–426. https://doi.org/10.1146/annurev-matsci-080921-085947.

Das, Chandramouli, Abhaya Kumar Sahoo, and Chittaranjan Pradhan. 2022. 'Multicriteria Recommender System Using Different Approaches'. In *Cognitive Big Data Intelligence with a Metaheuristic Approach*, 259–77. Elsevier. https://doi.org/10.1016/B978-0-323-85117-6.00011-X.

Davidson, Ernest R. 1975. 'The Iterative Calculation of a Few of the Lowest Eigenvalues and Corresponding Eigenvectors of Large Real-Symmetric Matrices'. *Journal of Computational Physics* 17 (1): 87–94. https://doi.org/10.1016/0021-9991(75)90065-0.

Davidson, Ernest R., and Aurora E. Clark. 2022. 'A Viewpoint on Population Analyses'. *International Journal of Quantum Chemistry* 122 (8): e26860. https://doi.org/10.1002/qua.26860.

Davies, Daniel W., Keith T. Butler, Adam J. Jackson, Andrew Morris, Jarvist M. Frost, Jonathan M. Skelton, and Aron Walsh. 2016. 'Computational Screening of All

Stoichiometric Inorganic Materials'. *Chem* 1 (4): 617–27. https://doi.org/10.1016/j.chempr.2016.09.010.

De Breuck, Pierre-Paul, Geoffroy Hautier, and Gian Marco Rignanese. 2021. 'Materials Property Prediction for Limited Datasets Enabled by Feature Selection and Joint Learning with MODNet'. *Npj Computational Materials* 7 (1): 1–8. https://doi.org/10.1038/s41524-021-00552-2.

De Breuck, Pierre-Paul, Matthew L Evans, and Gian-Marco Rignanese. 2021. 'Robust Model Benchmarking and Bias-Imbalance in Data-Driven Materials Science: A Case Study on MODNet'. *Journal of Physics: Condensed Matter* 33 (40): 404002. https://doi.org/10.1088/1361-648X/ac1280.

De Breuck, Pierre-Paul, Grégoire Heymans, and Gian-Marco Rignanese. 2022. 'Accurate Experimental Band Gap Predictions with Multifidelity Correction Learning'. *Journal of Materials Informatics* 2 (3): 10. https://doi.org/10.20517/jmi.2022.13.

Del Sole, R., and Raffaello Girlanda. 1993. 'Optical Properties of Semiconductors within the Independent-Quasiparticle Approximation'. *Physical Review B* 48 (16): 11789–95. https://doi.org/10.1103/PhysRevB.48.11789.

Deng, Bowen, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. 2023. 'CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling'. *Nature Machine Intelligence* 5 (9): 1031–41. https://doi.org/10.1038/s42256-023-00716-3.

Deng, Xiao Yan, Guang Hua Liu, Xi Ping Jing, and Guang Shan Tian. 2014. 'On-Site Correlation of p-Electron in D10 Semiconductor Zinc Oxide'. *International Journal of Quantum Chemistry* 114 (7): 468–72. https://doi.org/10.1002/qua.24593.

Di Liberto, Giovanni, Ornella Fatale, and Gianfranco Pacchioni. 2021. 'Role of Surface Termination and Quantum Size in α-$CsPbX_3$ (X = Cl, Br, I) 2D Nanostructures for Solar Light Harvesting'. *Physical Chemistry Chemical Physics* 23 (4): 3031–40. https://doi.org/10.1039/D0CP06245F.

Dietterich, Thomas G. 2000. 'Ensemble Methods in Machine Learning'. In *Multiple Classifier Systems,* 1857:1–15. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1.

Diffenbaugh, Noah S., and Elizabeth A. Barnes. 2023. 'Data-Driven Predictions of the Time Remaining until Critical Global Warming Thresholds Are Reached'. *Proceedings of the National Academy of Sciences* 120 (6): e2207183120. https://doi.org/10.1073/pnas.2207183120.

Dubey, Shiv Ram, Satish Kumar Singh, and Bidyut Baran Chaudhuri. 2021. 'Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark'. https://doi.org/10.48550/ARXIV.2109.14545.

Duchemin, Ivan, and François Gygi. 2010. 'A Scalable and Accurate Algorithm for the Computation of Hartree–Fock Exchange'. *Computer Physics Communications* 181 (5): 855–60. https://doi.org/10.1016/j.cpc.2009.12.021.

Dudarev, S., and G. Botton. 1998. 'Electron-Energy-Loss Spectra and the Structural Stability of Nickel Oxide: An LSDA+U Study'. *Physical Review B - Condensed Matter and Materials Physics* 57 (3): 1505–9. https://doi.org/10.1103/PhysRevB.57.1505.

Dunn, Alexander. 2024. 'MatBench Leaderboard'. 2024. https://matbench.materialsproject.org/.

Dunn, Alexander, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. 2020. 'Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm'. *Npj Computational Materials* 6 (1): 1–10. https://doi.org/10.1038/s41524-020-00406-3.

Dvorak, Marc, Su Huai Wei, and Zhigang Wu. 2013. ‘Origin of the Variation of Exciton Binding Energy in Semiconductors’. *Physical Review Letters* 110 (1): 1–5. https://doi.org/10.1103/PhysRevLett.110.016402.

Elward, Jennifer M., and Arindam Chakraborty. 2013. ‘Effect of Dot Size on Exciton Binding Energy and Electron–Hole Recombination Probability in CdSe Quantum Dots’. *Journal of Chemical Theory and Computation* 9 (10): 4351–59. https://doi.org/10.1021/ct400485s.

Emery, Antoine A., and Chris Wolverton. 2017. ‘High-Throughput DFT Calculations of Formation Energy, Stability and Oxygen Vacancy Formation Energy of ABO3 Perovskites’. *Scientific Data 2017 4:1* 4 (1): 1–10. https://doi.org/10.1038/sdata.2017.153.

Enkovaara, Jussi, Nichols A Romero, Sameer Shende, and Jens J Mortensen. 2011. ‘GPAW - Massively Parallel Electronic Structure Calculations with Python-Based Software’. In *Procedia Computer Science*, 4:17–25. https://doi.org/10.1016/j.procs.2011.04.003.

Ethem Alpaydd n. 2009. *Introduction to Machine Learning*. 2nd edition. Cambridge, Massachusetts: MIT Press.

Exner, Arthur, Rogério Almeida Gouvêa, Ariadne Köche, Sherdil Khan, Jacqueline Ferreira Leite Santos, and Marcos José Leite Santos. 2024. ‘Doping Effects on the Optoelectronic Properties and the Stability of $Cs_3Sb_2I_9$: Density Functional Theory Insights on Photovoltaics and Light-Emitting Devices’. *Journal of Science: Advanced Materials and Devices*, March, 100700. https://doi.org/10.1016/j.jsamd.2024.100700.

Fabini, Douglas H, Ram Seshadri, and Mercouri G Kanatzidis. 2020. ‘The Underappreciated Lone Pair in Halide Perovskites Underpins Their Unusual Properties’. *MRS BULLETIN* • 45. https://doi.org/10.1557/mrs.2020.142.

Fakharuddin, Azhar, Umair Shabbir, Weiming Qiu, Tahir Iqbal, Muhammad Sultan, Paul Heremans, and Lukas Schmidt-Mende. 2019. ‘Inorganic and Layered Perovskites for Optoelectronic Devices’. *Advanced Materials* 31 (47): 1807095. https://doi.org/10.1002/adma.201807095.

Feliczak-Guzik, Agnieszka. 2023. ‘Nanomaterials as Photocatalysts—Synthesis and Their Potential Applications’. *Materials* 16 (1): 193. https://doi.org/10.3390/ma16010193.

Filippetti, A., and A. Mattoni. 2014. ‘Hybrid Perovskites for Photovoltaics: Insights from First Principles’. *Physical Review B - Condensed Matter and Materials Physics* 89 (12): 1–8. https://doi.org/10.1103/PhysRevB.89.125203.

Fiolhais, Carlos, Fernando Nogueira, and Miguel AL Marques. 2003. *A Primer in Density Functional Theory*. Vol. 620. Springer Science & Business Media.

Flores, Efracio Mamani, Rogério Almeida Gouvea, Maurício Jeomar Piotrowski, and Mário Lucio Moreira. 2018. ‘Band Alignment and Charge Transfer Predictions of ZnO/ZnX (X = S, Se or Te) Interfaces Applied to Solar Cells: A PBE+: U Theoretical Study’. *Physical Chemistry Chemical Physics* 20 (7): 4953–61. https://doi.org/10.1039/c7cp08177d.

Frauenheim, Thomas, Gotthard Seifert, Marcus Elstner, Thomas Niehaus, Christof Köhler, Marc Amkreutz, Michael Sternberg, Zoltán Hajnal, Aldo Di Carlo, and Sándor Suhai. 2002. ‘Atomistic Simulations of Complex Materials: Ground-State and Excited-State Properties’. *Journal of Physics: Condensed Matter* 14 (11): 3015–47. https://doi.org/10.1088/0953-8984/14/11/313.

Freysoldt, Christoph, Blazej Grabowski, Tilmann Hickel, Jörg Neugebauer, Georg Kresse, Anderson Janotti, and Chris G. Van De Walle. 2014. ‘First-Principles Calculations for Point Defects in Solids’. *Reviews of Modern Physics* 86 (1): 253–305. https://doi.org/10.1103/RevModPhys.86.253.

Frigo, M., and S.G. Johnson. 2005. ‘The Design and Implementation of FFTW3’. *Proceedings of the IEEE* 93 (2): 216–31. https://doi.org/10.1109/JPROC.2004.840301.

Frost, Jarvist M., Keith T. Butler, Federico Brivio, Christopher H. Hendon, Mark Van Schilfgaarde, and Aron Walsh. 2014. ‘Atomistic Origins of High-Performance in Hybrid Halide Perovskite Solar Cells’. *Nano Letters* 14 (5): 2584–90. https://doi.org/10.1021/nl500390f.

Fultz, Brent. 2010. ‘Vibrational Thermodynamics of Materials’. *Progress in Materials Science* 55 (4): 247–352. https://doi.org/10.1016/j.pmatsci.2009.05.002.

Fung, Victor, P. Ganesh, and Bobby G. Sumpter. 2022. ‘Physically Informed Machine Learning Prediction of Electronic Density of States’. *Chemistry of Materials* 34 (11): 4848–55. https://doi.org/10.1021/acs.chemmater.1c04252.

Garrillo, Pablo Arturo Fernandez. 2018. ‘Development of Highly Resolved and Photo-Modulated Kelvin Probe Microscopy Techniques for the Study of Photovoltaic Systems’. Phd, Université Grenoble Alpes.

Garrity, Kevin F., Joseph W. Bennett, Karin M. Rabe, and David Vanderbilt. 2014. ‘Pseudopotentials for High-Throughput DFT Calculations’. *Computational Materials Science* 81 (May): 446–52. https://doi.org/10.1016/j.commatsci.2013.08.053.

Gebhardt, Julian, and Andrew M. Rappe. 2018. ‘Doping of BiFeO 3 : A Comprehensive Study on Substitutional Doping’. *Physical Review B* 98 (12): 125202. https://doi.org/10.1103/PhysRevB.98.125202.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. ‘Neural Networks and the Bias/Variance Dilemma’. *Neural Computation* 4 (1): 1–58.

Gerber, Iann C., János G. Ángyán, Martijn Marsman, and Georg Kresse. 2007. ‘Range Separated Hybrid Density Functional with Long-Range Hartree-Fock Exchange Applied to Solids’. *The Journal of Chemical Physics* 127 (5): 054101. https://doi.org/10.1063/1.2759209.

Giannozzi, P., O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, et al. 2017. ‘Advanced Capabilities for Materials Modelling with Quantum ESPRESSO’. *Journal of Physics Condensed Matter* 29 (46). https://doi.org/10.1088/1361-648X/aa8f79.

Giannozzi, Paolo, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L. Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano De Gironcoli, et al. 2009. ‘QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials’. *Journal of Physics Condensed Matter* 21 (39). https://doi.org/10.1088/0953-8984/21/39/395502.

Giannozzi, Paolo, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, et al. 2009. ‘QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials.’ *Journal of Physics. Condensed Matter : An Institute of Physics Journal* 21 (39): 395502. https://doi.org/10.1088/0953-8984/21/39/395502.

Giovilli, Giulia, Benedetta Albini, Virginia Grisci, Sara Bonomi, Marco Moroni, Edoardo Mosconi, Waldemar Kaiser, Filippo De Angelis, Pietro Galinetto, and Lorenzo Malavasi. 2023. ‘Band Gap Tuning through Cation and Halide Alloying in Mechanochemically Synthesized $Cs_3(Sb_{1-x}Bi_x)_2Br_9$ and $Cs_3Sb_2(I_{1-x}Br_x)_9$ Solid Solutions’. *Journal of Materials Chemistry C* 11 (30): 10282–91. https://doi.org/10.1039/D3TC01492D.

Giustino, Feliciano, and Henry J. Snaith. 2016. ‘Toward Lead-Free Perovskite Solar Cells’. *ACS Energy Letters* 1 (6): 1233–40. https://doi.org/10.1021/acsenergylett.6b00499.

Gong, Sheng, Shuo Wang, Tian Xie, Woo Hyun Chae, Runze Liu, Yang Shao-Horn, and Jeffrey C. Grossman. 2022. ‘Calibrating DFT Formation Enthalpy Calculations by Multifidelity Machine Learning’. *JACS Au* 2 (9): 1964–77. https://doi.org/10.1021/jacsau.2c00235.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Gouvêa, Rogério Almeida, Mário Lúcio Moreira, Chandra Veer Singh, and Marcos José Leite Santos. 2024. ‘Lead-Free Cesium Antimony Halide Perovskites: Halide Alloying, Surfaces, Interfaces, and Clusters’. *Journal of Materials Science* 59 (1): 142–60. https://doi.org/10.1007/s10853-023-09228-2.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. ‘Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks’. In *Proceedings of the 23rd International Conference on Machine Learning*, 369–76. ICML ’06. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1143844.1143891.

Griffiths, and Schroeter. 2018. *Introduction to Quantum Mechanics*. Cambridge university press.

Grundmann, Marius. 2010. *Physics of Semiconductors*. Vol. 11. Springer.

Guo, Wen Hui, Jun Jie Shi, Yao Hui Zhu, Meng Wu, Juan Du, Yu Lang Cen, Shi Ming Liu, and Shu Peng Han. 2020a. ‘Two-Dimensional 111-Type in -Based Halide Perovskite $Cs_3In_2X_9$ (X=Cl, Br, I) with Optimal Band Gap for Photovoltaics and Defect-Insensitive Blue Emission’. *Physical Review Applied* 13 (2): 1. https://doi.org/10.1103/PhysRevApplied.13.024031.

———. 2020b. ‘Two-Dimensional 111-Type In-Based Halide Perovskite $Cs_3In_2X_9$ (X=Cl, Br, I) with Optimal Band Gap for Photovoltaics and Defect-Insensitive Blue Emission’. *Physical Review Applied* 13 (2): 1. https://doi.org/10.1103/PhysRevApplied.13.024031.

Guo, Zhanglin, Shuai Zhao, Anmin Liu, Yusuke Kamata, Siowhwa Teo, Shuzhang Yang, Zhenhua Xu, Shuzi Hayase, and Tingli Ma. 2019. ‘Niobium Incorporation into CsPbI2Br for Stable and Efficient All-Inorganic Perovskite Solar Cells’. *ACS Applied Materials and Interfaces* 11 (22): 19994–3. https://doi.org/10.1021/acsami.9b03622.

Gurunathan, Ramya, Kamal Choudhary, and Francesca Tavazza. 2023. ‘Rapid Prediction of Phonon Structure and Properties Using the Atomistic Line Graph Neural Network (ALIGNN)’. *Physical Review Materials* 7 (2): 023803. https://doi.org/10.1103/PhysRevMaterials.7.023803.

Hailegnaw, Bekele, Saar Kirmayer, Eran Edri, Gary Hodes, and David Cahen. 2015. ‘Rain on Methylammonium Lead Iodide Based Perovskites: Possible Environmental Effects of Perovskite Solar Cells’. *The Journal of Physical Chemistry Letters* 6 (9): 1543–47. https://doi.org/10.1021/acs.jpclett.5b00504.

Hao, Jiabin, Zeming Wang, Huiying Hao, Guanlei Wang, Hongcheng Gao, Jianyu Wang, Bing Pan, and Qiang Qi. 2021. ‘Efficient Ag-Doped Perovskite Solar Cells Fabricated in Ambient Air’. *Crystals* 11 (12). https://doi.org/10.3390/cryst11121521.

Haque, Md Azimul, Seyoung Kee, Diego Rosas Villalva, Wee-Liat Ong, and Derya Baran. 2020. ‘Halide Perovskites: Thermal Transport and Prospects for Thermoelectricity’. *Advanced Science* 7 (10): 1903389. https://doi.org/10.1002/advs.201903389.

Hasnip, Philip J, Keith Refson, Matt I J Probert, Jonathan R Yates, Stewart J Clark, and Chris J Pickard. 2014. ‘Density Functional Theory in the Solid State’. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372 (2011): 20130270.

He, Chenlu, and Xiaogang Liu. 2023. ‘The Rise of Halide Perovskite Semiconductors’. *Light: Science & Applications* 12 (1): 15. https://doi.org/10.1038/s41377-022-01010-4.

Hegde, Vinay I., Christopher K. H. Borg, Zachary Del Rosario, Yoolhee Kim, Maxwell Hutchinson, Erin Antono, Julia Ling, Paul Saxe, James E. Saal, and Bryce Meredig. 2023. 'Quantifying Uncertainty in High-Throughput Density Functional Theory: A Comparison of AFLOW, Materials Project, and OQMD'. *Physical Review Materials* 7 (5): 053805. https://doi.org/10.1103/PhysRevMaterials.7.053805.

Henkelman, Graeme, Andri Arnaldsson, and Hannes Jónsson. 2006. 'A Fast and Robust Algorithm for Bader Decomposition of Charge Density'. *Computational Materials Science* 36 (3): 354–60. https://doi.org/10.1016/j.commatsci.2005.04.010.

Heyd, Jochen, Gustavo E. Scuseria, and Matthias Ernzerhof. 2003. 'Hybrid Functionals Based on a Screened Coulomb Potential'. *Journal of Chemical Physics* 118 (18): 8207–15. https://doi.org/10.1063/1.1564060.

Hinton, G. E., and R. R. Salakhutdinov. 2006. 'Reducing the Dimensionality of Data with Neural Networks'. *Science* 313 (5786): 504–7. https://doi.org/10.1126/science.1127647.

Hinuma, Yoyo, Andreas Grüneis, Georg Kresse, and Fumiyasu Oba. 2014. 'Band Alignment of Semiconductors from Density-Functional Theory and Many-Body Perturbation Theory'. *Physical Review B* 90 (15): 155405.

Hoefler, Sebastian F., Gregor Trimmel, and Thomas Rath. 2017. 'Progress on Lead-Free Metal Halide Perovskites for Photovoltaic Applications: A Review'. *Monatshefte Fur Chemie* 148 (5): 795–826. https://doi.org/10.1007/s00706-017-1933-9.

Hong, Feng, Bayrammurad Saparov, Weiwei Meng, Zewen Xiao, David B. Mitzi, and Yanfa Yan. 2016. 'Viability of Lead-Free Perovskites with Mixed Chalcogen and Halogen Anions for Photovoltaic Applications'. *The Journal of Physical Chemistry C* 120 (12): 6435–41. https://doi.org/10.1021/acs.jpcc.6b00920.

Hong, Ki-Ha, Jongseob Kim, Lamjed Debbichi, Hyungjun Kim, and Sang Hyuk Im. 2017. 'Band Gap Engineering of $Cs_3Bi_2I_9$ Perovskites with Trivalent Atoms Using a Dual Metal Cation'. *The Journal of Physical Chemistry C* 121 (1): 969–74. https://doi.org/10.1021/acs.jpcc.6b12426.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1990. 'Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks'. *Neural Networks* 3 (5): 551–60. https://doi.org/10.1016/0893-6080(90)90005-6.

Hoye, Robert L. Z., Juanita Hidalgo, Robert A. Jagt, Juan-Pablo Correa-Baena, Thomas Fix, and Judith L. MacManus-Driscoll. 2022. 'The Role of Dimensionality on the Optoelectronic Properties of Oxide and Halide Perovskites, and Their Halide Derivatives'. *Advanced Energy Materials* 12 (4): 2100499. https://doi.org/10.1002/aenm.202100499.

Hu, Zhaosheng, Zhenhua Lin, Jie Su, Jincheng Zhang, Jingjing Chang, and Yue Hao. 2019. 'A Review on Energy Band-Gap Engineering for Perovskite Photovoltaics'. *Solar RRL* 3 (12): 1900304. https://doi.org/10.1002/solr.201900304.

Hubbard, J. 1964. 'Exchange Splitting in Ferromagnetic Nickel'. *Proceedings of the Physical Society* 84 (4): 455–64. https://doi.org/10.1088/0370-1328/84/4/301.

Hussain, Nisar, Irfan Ayoub, Umer Mushtaq, Rishabh Sehgal, Seemin Rubab, Rakesh Sehgal, Hendrik C. Swart, and Vijay Kumar. 2022. 'Introduction to Phosphors and Luminescence'. In *Rare-Earth-Activated Phosphors,* 3–41. Elsevier. https://doi.org/10.1016/B978-0-323-89856-0.00008-0.

Ihtisham-ul-haq, M. I. Khan, Asad Ullah, Ali Mujtaba, Badriah S. Almutairi, Wajeehah Shahid, Asghar Ali, and Jeong Ryeol Choi. 2024. 'Bandgap Reduction and Efficiency Enhancement in Cs $_2$ AgBiBr $_6$ Double Perovskite Solar Cells through Gallium Substitution'. *RSC Advances* 14 (8): 5440–48. https://doi.org/10.1039/D3RA08965G.

Isayev, Olexandr, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. 2015. 'Materials Cartography: Representing and

Mining Materials Space Using Structural and Electronic Fingerprints'. *Chemistry of Materials* 27 (3): 735–43. https://doi.org/10.1021/cm503507h.

Isayev, Olexandr, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. 2017. 'Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals'. *Nature Communications* 8 (1): 15679. https://doi.org/10.1038/ncomms15679.

Jäger, Sebastian, Arndt Allhorn, and Felix Bießmann. 2021. 'A Benchmark for Data Imputation Methods'. *Frontiers in Big Data* 4 (July): 693674. https://doi.org/10.3389/fdata.2021.693674.

Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, and Gerbrand Ceder. 2013. 'Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation'. *APL Materials* 1 (1): 11002.

Jain, Vimal Kumar. 2022. *Solid State Physics*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-96017-9.

Janotti, Anderson, and Chris G. Van de Walle. 2011. 'LDA + U and Hybrid Functional Calculations for Defects in ZnO, SnO_2, and TiO_2'. *Physica Status Solidi (b)* 248 (4): 799–804. https://doi.org/10.1002/pssb.201046384.

Jena, Ajay Kumar, Ashish Kulkarni, and Tsutomu Miyasaka. 2019. 'Halide Perovskite Photovoltaics: Background, Status, and Future Prospects'. *Chemical Reviews* 119 (5): 3036–3103. https://doi.org/10.1021/acs.chemrev.8b00539.

Jensen, Frank. 2017. *Introduction to Computational Chemistry*. Third edition. Chichester, UK ; Hoboken, NJ: Wiley.

Jeong, Donghwi, Junyoung Kim, Ohhun Kwon, Chaehyun Lim, Sivaprakash Sengodan, Jeeyoung Shin, and Guntae Kim. 2018. 'Scandium Doping Effect on a Layered Perovskite Cathode for Low-Temperature Solid Oxide Fuel Cells (LT-SOFCs)'. *Applied Sciences (Switzerland)* 8 (11). https://doi.org/10.3390/app8112217.

Jiang, Dejun, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. 2021. 'Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models'. *Journal of Cheminformatics* 13 (1): 12. https://doi.org/10.1186/s13321-020-00479-8.

Jiang, Fangyuan, Dongwen Yang, Youyu Jiang, Tiefeng Liu, Xingang Zhao, Yue Ming, Bangwu Luo, et al. 2018. 'Chlorine-Incorporation-Induced Formation of the Layered Phase for Antimony-Based Lead-Free Perovskite Solar Cells'. *Journal of the American Chemical Society* 140 (3): 1019–27. https://doi.org/10.1021/jacs.7b10739.

Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. 2018. 'Junction Tree Variational Autoencoder for Molecular Graph Generation'. In *International Conference on Machine Learning*, 2323–32. PMLR. https://proceedings.mlr.press/v80/jin18a.html.

Jin, Zhixin, Zheng Zhang, Jingwei Xiu, Haisheng Song, Teresa Gatti, and Zhubing He. 2020. 'A Critical Review on Bismuth and Antimony Halide Based Perovskites and Their Derivatives for Photovoltaic Applications: Recent Advances and Challenges'. *Journal of Materials Chemistry A* 8 (32): 16166–88. https://doi.org/10.1039/D0TA05433J.

John, Rohit Abraham, Alessandro Milozzi, Sergey Tsarev, Rolf Brönnimann, Simon C. Boehme, Erfu Wu, Ivan Shorubalko, Maksym V. Kovalenko, and Daniele Ielmini. 2022. 'Ionic-Electronic Halide Perovskite Memdiodes Enabling Neuromorphic Computing with a Second-Order Complexity'. *Science Advances* 8 (51): eade0072. https://doi.org/10.1126/sciadv.ade0072.

K. Kihara and T. Sudo. 1974. 'The Crystal Structures of Beta-$Cs_3Sb_2Cl_9$ and $Cs_3Bi_2Cl_9$'. *Acta Cryst.* B30: 1088–93. https://doi.org/10.1107/S0365110X65003250.

Kahwagi, Rashad F., Sean T. Thornton, Ben Smith, and Ghada I. Koleilat. 2020. ‘Dimensionality Engineering of Metal Halide Perovskites’. *Frontiers of Optoelectronics* 13 (3): 196–224. https://doi.org/10.1007/s12200-020-1039-6.

Kang, Jun, and Lin-Wang Wang. 2017. ‘High Defect Tolerance in Lead Halide Perovskite $CsPbBr_3$’. *The Journal of Physical Chemistry Letters* 8 (2): 489–93. https://doi.org/10.1021/acs.jpclett.6b02800.

Kim, Hyojung, Ji Su Han, Jaeho Choi, Soo Young Kim, and Ho Won Jang. 2018. ‘Halide Perovskites for Applications beyond Photovoltaics’. *Small Methods* 2 (3): 1700310. https://doi.org/10.1002/smtd.201700310.

Kim, Sangtae, Miso Lee, Changho Hong, Youngchae Yoon, Hyungmin An, Dongheon Lee, Wonseok Jeong, et al. 2020. ‘A Band-Gap Database for Semiconducting Inorganic Materials Calculated with Hybrid Functional’. *Scientific Data* 7 (1): 387. https://doi.org/10.1038/s41597-020-00723-8.

Kim, Se Yun, Yeonghun Yun, Seunghak Shin, Joon Hyung Lee, Young Woo Heo, and Sangwook Lee. 2019. ‘Wide Range Tuning of Band Gap Energy of $A_3B_2X_9$ Perovskite-like Halides’. *Scripta Materialia* 166: 107–11. https://doi.org/10.1016/j.scriptamat.2019.03.009.

Kim, Whi Dong, Ji-Hee Kim, Sooho Lee, Seokwon Lee, Ju Young Woo, Kangha Lee, Weon-Sik Chae, et al. 2016. ‘Role of Surface States in Photocatalysis: Study of Chlorine-Passivated CdSe Nanocrystals for Photocatalytic Hydrogen Generation’. *Chemistry of Materials* 28 (3): 962–68. https://doi.org/10.1021/acs.chemmater.5b04790.

Kirklin, Scott, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. 2015. ‘The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies’. *Npj Computational Materials* 1 (September). https://doi.org/10.1038/npjcompumats.2015.10.

Kittel, C. 2004. ‘Introduction to Solid State Physics’. John Wiley & Sons.

Kleinman, Leonard. 1981. ‘Comment on the Average Potential of a Wigner Solid’. *Physical Review B* 24 (12): 7412.

Kohn, W, and L J Sham. 1965. ‘Self-Consistent Equations Including Exchange and Correlation Effects’. *Phys. Rev.* 140 (4A): A1133--A1138. https://doi.org/10.1103/PhysRev.140.A1133.

Koliogiorgos, Athanasios, Christos S. Garoufalis, Iosif Galanakis, and Sotirios Baskoutas. 2018. ‘Electronic and Optical Properties of Ultrasmall ABX_3 (A = Cs, CH3NH3/B = Ge, Pb, Sn, Ca, Sr/X = Cl, Br, I) Perovskite Quantum Dots’. *ACS Omega* 3 (12): 18917–24. https://doi.org/10.1021/acsomega.8b02525.

Koller, David, Fabien Tran, and Peter Blaha. 2011. ‘Merits and Limits of the Modified Becke-Johnson Exchange Potential’. *Physical Review B* 83 (19): 195134. https://doi.org/10.1103/PhysRevB.83.195134.

Körbel, Sabine, Miguel A. L. Marques, and Silvana Botti. 2016. ‘Stability and Electronic Properties of New Inorganic Perovskites from High-Throughput Ab Initio Calculations’. *Journal of Materials Chemistry C* 4 (15): 3157–67. https://doi.org/10.1039/C5TC04172D.

Korjus, Kristjan, Martin N. Hebart, and Raul Vicente. 2016. ‘An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable’. Edited by Chuhsing Kate Hsiao. *PLOS ONE* 11 (8): e0161788. https://doi.org/10.1371/journal.pone.0161788.

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. 2004. ‘Estimating Mutual Information’. *Physical Review E* 69 (6): 066138. https://doi.org/10.1103/PhysRevE.69.066138.

Kresse, G., and J. Furthmüller. 1996. 'Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set'. *Physical Review B - Condensed Matter and Materials Physics* 54 (16): 11169–86. https://doi.org/10.1103/PhysRevB.54.11169.

Kresse, G, and D Joubert. 1999. 'Kresse, Joubert - Unknown - From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method' 59 (3): 11–19.

Krukau, Aliaksandr V., Oleg A. Vydrov, Artur F. Izmaylov, and Gustavo E. Scuseria. 2006. 'Influence of the Exchange Screening Parameter on the Performance of Screened Hybrid Functionals'. *The Journal of Chemical Physics* 125 (22): 224106. https://doi.org/10.1063/1.2404663.

Kumar, Mulmudi Hemant, Sabba Dharani, Wei Lin Leong, Pablo P. Boix, Rajiv Ramanujam Prabhakar, Tom Baikie, Chen Shi, et al. 2014. 'Lead-Free Halide Perovskite Solar Cells with High Photocurrents Realized through Vacancy Modulation'. *Advanced Materials* 26 (41): 7122–27. https://doi.org/10.1002/adma.201401991.

Kumawat, Naresh Kumar, Zhongcheng Yuan, Sai Bai, and Feng Gao. 2019. 'Metal Doping/Alloying of Cesium Lead Halide Perovskite Nanocrystals and Their Applications in Light-Emitting Diodes with Enhanced Efficiency and Stability'. *Israel Journal of Chemistry* 59 (8): 695–707. https://doi.org/10.1002/ijch.201900031.

Kun, S V, V B Lazarev, E Yu Peresh, A V Kun, and Yu V Voroshilov. 1993. 'Phase Equilibria in RbBr-Sb(Bi)Br_3 Systems and Crystal Structure of Compounds of $A_3{}^1B_2{}^5C_9{}^7$ (A^1 - Rb, Cs; B^5 -Sb, Bi; C^7 - Br, I) Type'. *Neorganicheskie Materialy* 29 (3): 410–13.

Kwak, Dongwook, Mengjing Wang, Kristie J. Koski, Liang Zhang, Henry Sokol, Radenka Maric, and Yu Lei. 2019. 'Molybdenum Trioxide (α-MoO3) Nanoribbons for Ultrasensitive Ammonia (NH3) Gas Detection: Integrated Experimental and Density Functional Theory Simulation Studies'. *ACS Applied Materials and Interfaces* 11 (11): 10697–706. https://doi.org/10.1021/acsami.8b20502.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. 'Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles'. *Advances in Neural Information Processing Systems* 30.

Lam Pham, Tien, Hiori Kino, Kiyoyuki Terakura, Takashi Miyake, Koji Tsuda, Ichigaku Takigawa, and Hieu Chi Dam. 2017. 'Machine Learning Reveals Orbital Interaction in Materials'. *Science and Technology of Advanced Materials* 18 (1): 756–65. https://doi.org/10.1080/14686996.2017.1378060.

Lavecchia, Antonio. 2019. 'Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects'. *Drug Discovery Today* 24 (10): 2017–32. https://doi.org/10.1016/j.drudis.2019.07.006.

Lee, Jiale, Wei-Kean Chong, Steven Hao Wan Kok, Boon-Junn Ng, Xin Ying Kong, Siang-Piao Chai, and Lling-Lling Tan. 2023. 'Mixed Halide Formation in Lead-Free Antimony-Based Halide Perovskite for Boosted CO $_2$ Photoreduction: Beyond Band Gap Tuning'. *Advanced Functional Materials*, June, 2303430. https://doi.org/10.1002/adfm.202303430.

Lee, June-Goo, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. 2017. 'Deep Learning in Medical Imaging: General Overview'. *Korean Journal of Radiology* 18 (4): 570–84. https://doi.org/10.3348/kjr.2017.18.4.570.

Lee, Sang Hoon, and Young Woo Son. 2020. 'First-Principles Approach with a Pseudohybrid Density Functional for Extended Hubbard Interactions'. *Physical Review Research* 2 (4). https://doi.org/10.1103/PhysRevResearch.2.043410.

Lehner, Anna J., Douglas H. Fabini, Hayden A. Evans, Claire Alice Hébert, Sara R. Smock, Jerry Hu, Hengbin Wang, Josef W. Zwanziger, Michael L. Chabinyc, and Ram Seshadri. 2015. 'Crystal and Electronic Structures of Complex Bismuth Iodides A3Bi2I9 (A = K, Rb, Cs) Related to Perovskite: Aiding the Rational Design of

Photovoltaics'. *Chemistry of Materials* 27 (20): 7137–48. https://doi.org/10.1021/acs.chemmater.5b03147.

Lejaeghere, K., V. Van Speybroeck, G. Van Oost, and S. Cottenier. 2014. 'Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals'. *Critical Reviews in Solid State and Materials Sciences* 39 (1): 1–24. https://doi.org/10.1080/10408436.2013.772503.

Leppert, Linn, Tonatiuh Rangel, and Jeffrey B. Neaton. 2019. 'Towards Predictive Band Gaps for Halide Perovskites: Lessons from One-Shot and Eigenvalue Self-Consistent G W'. *Physical Review Materials* 3 (10): 103803. https://doi.org/10.1103/PhysRevMaterials.3.103803.

Levitt, Antoine, and Marc Torrent. 2015. 'Parallel Eigensolvers in Plane-Wave Density Functional Theory'. *Computer Physics Communications* 187 (February): 98–105. https://doi.org/10.1016/j.cpc.2014.10.015.

Li, Changjiao, Hua Hao, Ben Xu, Guanghui Zhao, Lihao Chen, Shujun Zhang, and Hanxing Liu. 2020. 'A Progressive Learning Method for Predicting the Band Gap of ABO $_3$ Perovskites Using an Instrumental Variable'. *Journal of Materials Chemistry C* 8 (9): 3127–36. https://doi.org/10.1039/C9TC06632B.

Li, Jiangtian, and Nianqiang Wu. 2015. 'Semiconductor-Based Photocatalysts and Photoelectrochemical Cells for Solar Fuel Generation: A Review'. *Catalysis Science & Technology* 5 (3): 1360–84. https://doi.org/10.1039/C4CY00974F.

Li, Jihong, Yongao Lv, Huifang Han, Jia Xu, and Jianxi Yao. 2022. 'Two-Dimensional $Cs_3Sb_2I_{9-x}Cl_x$ Film with (201) Preferred Orientation for Efficient Perovskite Solar Cells'. *Materials* 15 (8): 2883. https://doi.org/10.3390/ma15082883.

Li, Junming, Hai-Lei Cao, Wen-Bin Jiao, Qiong Wang, Mingdeng Wei, Irene Cantone, Jian Lü, and Antonio Abate. 2020. 'Biological Impact of Lead from Halide Perovskites Reveals the Risk of Introducing a Safe Threshold'. *Nature Communications* 11 (1): 310. https://doi.org/10.1038/s41467-019-13910-y.

Li, Qiuqi, Dan Cao, Xueyin Liu, Xiangyu Zhou, Xiaoshuang Chen, and Haibo Shu. 2021. 'Hierarchical Computational Screening of Layered Lead-Free Metal Halide Perovskites for Optoelectronic Applications'. *Journal of Materials Chemistry A* 9 (10): 6476–86. https://doi.org/10.1039/d0ta10098f.

Li, Shufang, Linna Zhu, Zhipeng Kan, Yong Hua, and Fei Wu. 2020. 'A Multifunctional Additive of Scandium Trifluoromethanesulfonate to Achieve Efficient Inverted Perovskite Solar Cells with a High Fill Factor of 83.80%'. *Journal of Materials Chemistry A* 8 (37): 19555–60. https://doi.org/10.1039/D0TA07567A.

Li, Shunning, Yuanji Liu, Dong Chen, Yi Jiang, Zhiwei Nie, and Feng Pan. 2022. 'Encoding the Atomic Structure for Machine Learning in Materials Science'. *WIREs Computational Molecular Science* 12 (1): e1558. https://doi.org/10.1002/wcms.1558.

Li, Xiaoming, Ye Wu, Shengli Zhang, Bo Cai, Yu Gu, Jizhong Song, and Haibo Zeng. 2016. '$CsPbX_3$ Quantum Dots for Lighting and Displays: Room-Temperature Synthesis, Photoluminescence Superiorities, Underlying Origins and White Light-Emitting Diodes'. *Advanced Functional Materials* 26 (15): 2435–45. https://doi.org/10.1002/adfm.201600109.

Li, Xin, Xupeng Gao, Xiangtong Zhang, Xinyu Shen, Min Lu, Jinlei Wu, Zhifeng Shi, et al. 2021. 'Lead-Free Halide Perovskites for Light Emission: Recent Advances and Perspectives'. *Advanced Science* 8 (4): 1–33. https://doi.org/10.1002/advs.202003334.

Liang, Jiechun, Tingting, Ziwei Wang, Yunduo Yu, Linfeng Hu, Huamei Li, Xiaohong Zhang, Xi Zhu, and Yu Zhao. 2022. 'Accelerating Perovskite Materials Discovery and Correlated Energy Applications through Artificial Intelligence'. *Energy Materials* 2 (January): 200016. https://doi.org/10.20517/energymater.2022.14.

Liang, Yunting. 2021. ‘Exploring Inorganic and Nontoxic Double Perovskites $Cs_2AgInBr_{6(1-x)}Cl_{6x}$ from Material Selection to Device Design in Material Genome Approach’. *Journal of Alloys and Compounds* 862: 158575. https://doi.org/10.1016/j.jallcom.2020.158575.

Lin, Pei-Ying, Aswaghosh Loganathan, Itaru Raifuku, Ming-Hsien Li, Yueh-Ya Chiu, Shao-Tung Chang, Azhar Fakharuddin, et al. 2021. ‘Pseudo-Halide Perovskite Solar Cells’. *Advanced Energy Materials* 11 (28): 2100818. https://doi.org/10.1002/aenm.202100818.

Liu, Maning, Hannu Pasanen, Harri Ali-Löytty, Arto Hiltunen, Kimmo Lahtonen, Syeda Qudsia, Jan-Henrik Smått, Mika Valden, Nikolai V. Tkachenko, and Paola Vivo. 2020. ‘B-Site Co-Alloying with Germanium Improves the Efficiency and Stability of All-Inorganic Tin-Based Perovskite Nanocrystal Solar Cells’. *Angewandte Chemie International Edition* 59 (49): 22117–25. https://doi.org/10.1002/anie.202008724.

Liu, Ping, Yuan Liu, Siwei Zhang, Jingzhou Li, Chunyun Wang, Cong Zhao, Pengbo Nie, et al. 2020. ‘Lead-Free $Cs_3Sb_2Br_9$ Single Crystals for High Performance Narrowband Photodetector’ 2001072: 1–9. https://doi.org/10.1002/adom.202001072.

Liu, Yu Liang, Chuan Lu Yang, Mei Shan Wang, Xiao Guang Ma, and You Gen Yi. 2019. ‘Theoretical Insight into the Optoelectronic Properties of Lead-Free Perovskite Derivatives of $Cs_3Sb_2X_9$ (X = Cl, Br, I)’. *Journal of Materials Science* 54 (6): 4732–41. https://doi.org/10.1007/s10853-018-3162-y.

Liu, Yue, Tianlu Zhao, Wangwei Ju, and Siqi Shi. 2017. ‘Materials Discovery and Design Using Machine Learning’. *Journal of Materiomics* 3 (3): 159–77. https://doi.org/10.1016/j.jmat.2017.08.002.

Liu, Zhenyang, Hanjun Yang, Junyu Wang, Yucheng Yuan, Katie Hills-Kimball, Tong Cai, Ping Wang, Aiwei Tang, and Ou Chen. 2021. ‘Synthesis of Lead-Free Cs_2AgBiX_6 (X = Cl, Br, I) Double Perovskite Nanoplatelets and Their Application in CO2 Photocatalytic Reduction’. *Nano Letters* 21 (4): 1620–27. https://doi.org/10.1021/acs.nanolett.0c04148.

Long, Yaowen, Hong Zhang, and Xinlu Cheng. 2022. ‘Stability, Electronic Structure, and Optical Properties of Lead-Free Perovskite Monolayer Cs3B2X9 (B = Sb, Bi; X = Cl, Br, I) and Bilayer Vertical Heterostructure Cs3B2X9/Cs3X9(B,B′ = Sb, Bi; X = Cl, Br, I)’. *Chinese Physics B* 31 (2): 027102. https://doi.org/10.1088/1674-1056/ac2e5f.

Lookman, Turab, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. 2019. ‘Active Learning in Materials Science with Emphasis on Adaptive Sampling Using Uncertainties for Targeted Design’. *Npj Computational Materials* 5 (1): 1–17. https://doi.org/10.1038/s41524-019-0153-8.

Löwdin, Per-Olov. 1953. ‘Approximate Formulas for Many-Center Integrals in the Theory of Molecules and Crystals’. *The Journal of Chemical Physics* 21 (2): 374–75. https://doi.org/10.1063/1.1698901.

Lu, Chang, Dominique S. Itanze, Alexander G. Aragon, Xiao Ma, Hui Li, Kamil B. Ucer, Corey Hewitt, et al. 2020. ‘Synthesis of Lead-Free $Cs_3Sb_2Br_9$ Perovskite Alternative Nanocrystals with Enhanced Photocatalytic CO_2 Reduction Activity’. *Nanoscale* 12 (5): 2987–91. https://doi.org/10.1039/C9NR07722G.

Lu, Cheng-Hsin, Gill V. Biesold-McGee, Yijiang Liu, Zhitao Kang, and Zhiqun Lin. 2020. ‘Doping and Ion Substitution in Colloidal Metal Halide Perovskite Nanocrystals’. *Chemical Society Reviews* 49 (14): 4953–5007. https://doi.org/10.1039/C9CS00790C.

Lu, Shuaihua, Qionghua Zhou, Xinyu Chen, Zhilong Song, and Jinlan Wang. 2022. ‘Inverse Design with Deep Generative Models: Next Step in Materials Discovery’. *National Science Review* 9 (8): nwac111. https://doi.org/10.1093/nsr/nwac111.

Lundberg, Scott M, and Su-In Lee. 2017. ‘A Unified Approach to Interpreting Model Predictions’. In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Luo, Haoran, Jiangbin Deng, Qianzhi Gou, Omololu Odunmbaku, Kuan Sun, Juanxiu Xiao, Meng Li, and Yujie Zheng. 2023. ‘Accelerated Discovery of Novel High-Performance Zinc-Ion Battery Cathode Materials by Combining High-Throughput Screening and Experiments’. *Chinese Chemical Letters* 34 (8): 107885. https://doi.org/10.1016/j.cclet.2022.107885.

Luo, Shulin, Tianshu Li, Xinjiang Wang, Muhammad Faizan, and Lijun Zhang. 2021. ‘High-throughput Computational Materials Screening and Discovery of Optoelectronic Semiconductors’. *WIREs Computational Molecular Science* 11 (1): e1489. https://doi.org/10.1002/wcms.1489.

Ma, Zhuangzhuang, Zhifeng Shi, Dongwen Yang, Fei Zhang, Sen Li, Lintao Wang, Di Wu, et al. 2019. ‘Electrically-Driven Violet Light-Emitting Devices Based on Highly Stable Lead-Free Perovskite $Cs_3Sb_2Br_9$ Quantum Dots’. *ACS Energy Lett* 2020: 394. https://doi.org/10.1021/acsenergylett.9b02096.

Madden, Michael G., and Tom Howley. 2009. ‘A Machine Learning Application for Classification of Chemical Spectra’. In *Applications and Innovations in Intelligent Systems XVI*, edited by Tony Allen, Richard Ellis, and Miltos Petridis, 77–90. London: Springer. https://doi.org/10.1007/978-1-84882-215-3_6.

Majid, Abdul, Asifullah Khan, and Tae-Sun Choi. 2011. ‘Predicting Lattice Constant of Complex Cubic Perovskites Using Computational Intelligence’. *Computational Materials Science* 50 (6): 1879–88. https://doi.org/10.1016/j.commatsci.2011.01.035.

Malavasi, Lorenzo, Pietro Galinetto, Benedetta Albini, Giulia Giovilli, Marco Moroni, Edoardo Mosconi, Filippo De Angelis, Waldemar Kaiser, and Virginia Grisci. 2023. ‘Band Gap Tuning Through Cation and Halide Alloying in Mechanochemical Synthesized $Cs_3(Sb_{1-x}Bi_x)_2Br_9$ and $Cs_3Sb_2(I_{1-x}Br_x)_9$ Solid Solutions’. Preprint. Chemistry. https://doi.org/10.26434/chemrxiv-2023-62vzx.

Malyi, Oleksandr I., Gustavo M. Dalpian, Xin-Gang Zhao, Zhi Wang, and Alex Zunger. 2020. ‘Realization of Predicted Exotic Materials: The Burden of Proof’. *Materials Today* 32 (January): 35–45. https://doi.org/10.1016/j.mattod.2019.08.003.

Malyi, Oleksandr I., Kostiantyn V. Sopiha, and Clas Persson. 2019. ‘Energy, Phonon, and Dynamic Stability Criteria of Two-Dimensional Materials’. *ACS Applied Materials & Interfaces* 11 (28): 24876–84. https://doi.org/10.1021/acsami.9b01261.

Margatina, Katerina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. ‘Active Learning by Acquiring Contrastive Examples’. arXiv. https://doi.org/10.48550/arXiv.2109.03764.

Marongiu, Daniela, Michele Saba, Francesco Quochi, Andrea Mura, and Giovanni Bongiovanni. 2019. ‘The Role of Excitons in 3D and 2D Lead Halide Perovskites’. *Journal of Materials Chemistry C* 7 (39): 12006–18. https://doi.org/10.1039/C9TC04292J.

Martin, Richard M. 2020. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge university press.

Masawa, Salma Maneno, Jihong Li, Chenxu Zhao, Xiaolong Liu, and Jianxi Yao. 2022. ‘0D/2D Mixed Dimensional Lead-Free Caesium Bismuth Iodide Perovskite for Solar Cell Application’. *Materials* 15 (6): 2180. https://doi.org/10.3390/ma15062180.

Mascarenhas, Yvonne Primerano. 2020. 'Crystallography before the Discovery of X-Ray Diffraction'. *Revista Brasileira de Ensino de Fisica* 42. https://doi.org/10.1590/1806-9126-RBEF-2019-0336.

Mathew, Kiran, Joseph H. Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek-heng Chu, et al. 2017. 'Atomate: A High-Level Interface to Generate, Execute, and Analyze Computational Materials Science Workflows'. *Computational Materials Science* 139 (November): 140–52. https://doi.org/10.1016/j.commatsci.2017.07.030.

Maughan, Annalise E., Alex M. Ganose, Mitchell M. Bordelon, Elisa M. Miller, David O. Scanlon, and James R. Neilson. 2016. 'Defect Tolerance to Intolerance in the Vacancy-Ordered Double Perovskite Semiconductors Cs_2SnI_6 and Cs_2TeI_6'. *Journal of the American Chemical Society* 138 (27): 8453–64. https://doi.org/10.1021/jacs.6b03207.

May, Kevin J., and Alexie M. Kolpak. 2020. 'Improved Description of Perovskite Oxide Crystal Structure and Electronic Properties Using Self-Consistent Hubbard U Corrections from ACBN0'. *Physical Review B* 101 (16): 165117. https://doi.org/10.1103/PhysRevB.101.165117.

McCall, Kyle M., Zhifu Liu, Giancarlo Trimarchi, Constantinos C. Stoumpos, Wenwen Lin, Yihui He, Ido Hadar, Mercouri G. Kanatzidis, and Bruce W. Wessels. 2018. 'α-Particle Detection and Charge Transport Characteristics in the A3M2I9 Defect Perovskites (A = Cs, Rb; M = Bi, Sb)'. *ACS Photonics* 5 (9): 3748–62. https://doi.org/10.1021/acsphotonics.8b00813.

Menedjhi, Adel, Nadir Bouarissa, Salima Saib, and Khaled Bouamama. 2021. 'Halide Double Perovskite $Cs_2AgInBr_6$ for Photovoltaic's Applications: Optical Properties and Stability'. *Optik* 243 (April): 167198. https://doi.org/10.1016/j.ijleo.2021.167198.

Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. 'Scaling Deep Learning for Materials Discovery'. *Nature* 624 (7990): 80–85. https://doi.org/10.1038/s41586-023-06735-9.

Mica, Natalie A., Rui Bian, Pavlos Manousiadis, Lethy K. Jagadamma, Iman Tavakkolnia, Harald Haas, Graham A. Turnbull, and Ifor D. W. Samuel. 2020. 'Triple-Cation Perovskite Solar Cells for Visible Light Communications'. *Photonics Research* 8 (8): A16. https://doi.org/10.1364/PRJ.393647.

Miyasaka, Tsutomu, Ashish Kulkarni, Gyu Min Kim, Senol Öz, and Ajay K. Jena. 2020. 'Perovskite Solar Cells: Can We Go Organic-Free, Lead-Free, and Dopant-Free?' *Advanced Energy Materials* 10 (13): 1902500. https://doi.org/10.1002/aenm.201902500.

Mohammad, Ashif, and Farhana Mahjabeen. 2023. 'Promises and Challenges of Perovskite Solar Cells: A Comprehensive Review'. *BULLET : Jurnal Multidisiplin Ilmu* 2 (5): 1147–57. https://www.journal.mediapublikasi.id/index.php/bullet/article/view/3685.

Monkhorst, H J, and J D Pack. 1976. 'Special Points for Brillouin-Zone Integrations'. *Phys Rev B* 13: 5188.

Mori-Sánchez, Paula, and Aron J. Cohen. 2014. 'The Derivative Discontinuity of the Exchange–Correlation Functional'. *Phys. Chem. Chem. Phys.* 16 (28): 14378–87. https://doi.org/10.1039/C4CP01170H.

Morocho-Cayamcela, Manuel Eugenio, Haeyoung Lee, and Wansu Lim. 2019. 'Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions'. *IEEE Access* 7: 137184–206. https://doi.org/10.1109/ACCESS.2019.2942390.

Mortazavi, Bohayra, Ivan S. Novikov, Evgeny V. Podryabinkin, Stephan Roche, Timon Rabczuk, Alexander V. Shapeev, and Xiaoying Zhuang. 2020. 'Exploring Phononic Properties of Two-Dimensional Materials Using Machine Learning Interatomic

Potentials'. *Applied Materials Today* 20 (September): 100685. https://doi.org/10.1016/j.apmt.2020.100685.

Mousavi, S. H., S. A. Jafari Mohammdi, H. Haratizadeh, P. W. deOliveira, S. H. Mousavi, S. A. Jafari Mohammdi, H. Haratizadeh, and P. W. deOliveira. 2014. 'Light-Emitting Devices – Luminescence from Low-Dimensional Nanostructures'. In *Advances in Optical Communication*. IntechOpen. https://doi.org/10.5772/59103.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT press.

Muy, Sokseiha, Johannes Voss, Roman Schlem, Raimund Koerver, Stefan J. Sedlmaier, Filippo Maglia, Peter Lamp, Wolfgang G. Zeier, and Yang Shao-Horn. 2019. 'High-Throughput Screening of Solid-State Li-Ion Conductors Using Lattice-Dynamics Descriptors'. *iScience* 16 (June): 270–82. https://doi.org/10.1016/j.isci.2019.05.036.

Nazari, Safieh, Yavar T. Azar, and Alireza Doroudi. 2020. 'Surface-Termination-Dependent Stability and Band Alignment in $CsPbX_3$ (X = I, Br, Cl) Perovskites: A First-Principle Study'. *Materials Today Communications* 24 (September): 100961. https://doi.org/10.1016/j.mtcomm.2020.100961.

Neal, Brady. 2019. 'On the Bias-Variance Tradeoff: Textbooks Need an Update'. https://doi.org/10.48550/ARXIV.1912.08286.

'Normalization | Google for Developers'. 2023. Google for Developers. 2023. https://developers.google.com/machine-learning/data-prep/transform/normalization.

Nosengo, Nicola. 2016. 'Can Artificial Intelligence Create the next Wonder Material?' *Nature* 533 (7601): 22–25. https://doi.org/10.1038/533022a.

NREL. 2023. 'Interactive Best Research-Cell Efficiency Chart'. NREL. 2023. https://www.nrel.gov/pv/interactive-cell-efficiency.html.

Oliver, G. L., and J. P. Perdew. 1979. 'Spin-Density Gradient Expansion for the Kinetic Energy'. *Physical Review A* 20 (2): 397–403. https://doi.org/10.1103/PhysRevA.20.397.

Onida, Giovanni, Lucia Reining, and Angel Rubio. 2002. 'Electronic Excitations: Density-Functional versus Many-Body Green's-Function Approaches'. *Reviews of Modern Physics* 74 (2): 601–59. https://doi.org/10.1103/RevModPhys.74.601.

Ouyang, Runhai. 2020. 'Exploiting Ionic Radii for Rational Design of Halide Perovskites'. *Chemistry of Materials* 32 (1): 595–604. https://doi.org/10.1021/acs.chemmater.9b04472.

Park, Byung wook, and Sang Il Seok. 2019. 'Intrinsic Instability of Inorganic–Organic Hybrid Halide Perovskite Materials'. *Advanced Materials* 31 (20): 1–17. https://doi.org/10.1002/adma.201805337.

Park, Byung-Wook, Bertrand Philippe, Xiaoliang Zhang, Håkan Rensmo, Gerrit Boschloo, and Erik M. J. Johansson. 2015. 'Bismuth Based Hybrid Perovskites $A_3Bi\ _2I_9$ (A: Methylammonium or Cesium) for Solar Cell Application'. *Advanced Materials* 27 (43): 6806–13. https://doi.org/10.1002/adma.201501978.

Park, Young Ran, Hong Hee Kim, Sangwon Eom, Won Kook Choi, Hyosung Choi, Bo Ram Lee, and Youngjong Kang. 2021. 'Luminance Efficiency Roll-off Mechanism in $CsPbBr_{3-x}Cl_x$ Mixed-Halide Perovskite Quantum Dot Blue Light-Emitting Diodes'. *Journal of Materials Chemistry C* 9 (10): 3608–19. https://doi.org/10.1039/d0tc05514j.

Park, Youngjun, Seong Hun Kim, Donghwa Lee, and Jang-Sik Lee. 2021. 'Designing Zero-Dimensional Dimer-Type All-Inorganic Perovskites for Ultra-Fast Switching Memory'. *Nature Communications* 12 (1): 3527. https://doi.org/10.1038/s41467-021-23871-w.

Parveen, Sumaiya, and P. K. Giri. 2022. 'Emerging Doping Strategies in Two-Dimensional Hybrid Perovskite Semiconductors for Cutting Edge Optoelectronics Applications'. *Nanoscale Advances* 4 (4): 995–1025. https://doi.org/10.1039/D1NA00709B.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. 'PyTorch: An Imperative Style, High-Performance Deep

Learning Library'. In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Patel, Hiren S, Vishnu A Dabhi, and Aditya M Vora. 2022. 'Adverse Effect of K-Mesh Shifting in Several Crystal Systems: An Analytical Study'. *Materials Today: Proceedings* 57: 275–78. https://doi.org/10.1016/j.matpr.2022.02.599.

Patil, Jyoti V., Sawanta S. Mali, and Chang Kook Hong. 2020. 'Efficient and Stable All-Inorganic Niobium-Incorporated CsPbI2Br-Based Perovskite Solar Cells'. *ACS Applied Materials and Interfaces* 12 (24): 27176–83. https://doi.org/10.1021/acsami.0c04577.

Paul, Goutam, Amlan J. Pal, and Bryon William Larson. 2020. 'Structure, Morphology, and Photovoltaic Implications of Halide Alloying in Lead-Free $Cs_3Sb_2Cl_xI_{9-x}$ 2D-Layered Perovskites'. *Solar RRL* 2000422: 1–8. https://doi.org/10.1002/solr.202000422.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. 'Scikit-Learn: Machine Learning in Python'. *The Journal of Machine Learning Research* 12: 2825–30.

Peng, Yueheng, Fengzhu Li, Yan Wang, Yachen Li, Robert LZ Hoye, Linrun Feng, Kai Xia, and Vincenzo Pecunia. 2020. 'Enhanced Photoconversion Efficiency in Cesium-Antimony-Halide Perovskite Derivatives by Tuning Crystallographic Dimensionality'. *Applied Materials Today* 19: 100637.

Perdew, J. P., K. Burke, and M. Ernzerhof. 1996. 'Generalized Gradient Approximation Made Simple'. *Phys Rev Lett* 77: 3865.

Perdew, J. P., J. A. Chevary, S. H. Vosko, K. A. Jackson, M R Pederson, D J Singh, and C Fiolhais. 1992. 'Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation'. *Phys Rev B* 46 (11): 6671.

Perdew, J P, and A Zunger. 1981. 'Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems'. *Phys Rev B* 23 (10): 5048.

Perdew, John P., Matthias Ernzerhof, and Kieron Burke. 1996. 'Rationale for Mixing Exact Exchange with Density Functional Approximations'. *The Journal of Chemical Physics* 105 (22): 9982–85. https://doi.org/10.1063/1.472933.

Perdew, John P., Adrienn Ruzsinszky, Gábor I. Csonka, Oleg A. Vydrov, Gustavo E. Scuseria, Lucian A. Constantin, Xiaolan Zhou, and Kieron Burke. 2008. 'Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces'. *Physical Review Letters* 100 (13): 1–4. https://doi.org/10.1103/PhysRevLett.100.136406.

Perdew, John P., and Yue Wang. 1992. 'Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy'. *Physical Review B* 45 (23): 13244–49. https://doi.org/10.1103/PhysRevB.45.13244.

Perdew, John P, P Ziesche, and H Eschrig. 1991. 'Electronic Structure of Solids' 91'. Akademie Verlag, Berlin.

Peter, YU, and Manuel Cardona. 2010. *Fundamentals of Semiconductors: Physics and Materials Properties*. Springer Science & Business Media.

Pilania, Ghanshyam. 2021. 'Machine Learning in Materials Science: From Explainable Predictions to Autonomous Design'. *Computational Materials Science* 193. https://doi.org/10.1016/j.commatsci.2021.110360.

Podolyan, Yevgeniy, Michael A. Walters, and George Karypis. 2010. 'Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods'. *Journal of Chemical Information and Modeling* 50 (6): 979–91. https://doi.org/10.1021/ci900301v.

Pradhan, Abinash, Milan Kumar Jena, and Saroj L. Samal. 2022. 'Understanding of the Band Gap Transition in $Cs_3Sb_2Cl_{9–x}Br_x$: Anion Site Preference-Induced Structural

Distortion'. *ACS Applied Energy Materials* 5 (6): 6952–61. https://doi.org/10.1021/acsaem.2c00591.

Pradhan, Bapi, Gundam Sandeep Kumar, Sumanta Sain, Amit Dalui, Uttam Kumar Ghorai, Swapan Kumar Pradhan, and Somobrata Acharya. 2018. 'Size Tunable Cesium Antimony Chloride Perovskite Nanowires and Nanorods'. *Chemistry of Materials* 30 (6): 2135–42. https://doi.org/10.1021/acs.chemmater.8b00427.

Pramchu, Sittichain, Atchara Punya Jaroenjittichai, and Yongyut Laosiritaworn. 2019. 'Effects of Bromine Substitution for Iodine on Structural Stability and Phase Transition of $CsPbI_3$'. *Applied Surface Science* 496: 143593. https://doi.org/10.1016/j.apsusc.2019.143593.

Prandini, Gianluca, Mario Galante, Nicola Marzari, and Paolo Umari. 2019. 'SIMPLE Code: Optical Properties with Optimal Basis Functions'. *Computer Physics Communications* 240: 106–19. https://doi.org/10.1016/j.cpc.2019.02.016.

Prandini, Gianluca, Antimo Marrazzo, Ivano E Castelli, Nicolas Mounet, and Nicola Marzari. 2018. 'Precision and Efficiency in Solid-State Pseudopotential Calculations'. *Npj Computational Materials* 4 (1): 72. https://doi.org/10.1038/s41524-018-0127-2.

Prendergast, David, and Steven G. Louie. 2009. 'Bloch-State-Based Interpolation: An Efficient Generalization of the Shirley Approach to Interpolating Electronic Structure'. *Physical Review B - Condensed Matter and Materials Physics* 80 (23): 235126. https://doi.org/10.1103/PhysRevB.80.235126.

Qian, Xin, and Ronggui Yang. 2021. 'Machine Learning for Predicting Thermal Transport Properties of Solids'. *Materials Science and Engineering: R: Reports* 146 (October): 100642. https://doi.org/10.1016/j.mser.2021.100642.

Rashidi, Hooman H., Nam K. Tran, Elham Vali Betts, Lydia P. Howell, and Ralph Green. 2019. 'Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods'. *Academic Pathology* 6 (January): 2374289519873088. https://doi.org/10.1177/2374289519873088.

Raza, Muhammad Ali, Feng Li, Meidan Que, Liangliang Zhu, and Xi Chen. 2021. 'Photocatalytic Reduction of CO_2 by Halide Perovskites: Recent Advances and Future Perspectives'. *Materials Advances* 2 (22): 7187–7209. https://doi.org/10.1039/D1MA00703C.

'Receiver Operating Characteristic'. 2023. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=1188201158.

Refaeilzadeh, Payam, Lei Tang, and Huan Liu. 2009. 'Cross-Validation'. In *Encyclopedia of Database Systems*, edited by Ling Liu and M. Tamer Özsu, 532–38. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_565.

Ren, Zekun, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, et al. 2022. 'An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties'. *Matter* 5 (1): 314–35. https://doi.org/10.1016/j.matt.2021.11.032.

Riebesell, Janosh. 2024. 'Matbench Discovery'. 2024. https://janosh.github.io/matbench-discovery.

Riebesell, Janosh, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. 2023. 'Matbench Discovery–An Evaluation Framework for Machine Learning Crystal Stability Prediction'. *arXiv Preprint arXiv:2308.14920*.

Rodrigues, Jose F., Larisa Florea, Maria C. F. De Oliveira, Dermot Diamond, and Osvaldo N. Oliveira. 2021. 'Big Data and Machine Learning for Materials Science'. *Discover Materials* 1 (1): 12. https://doi.org/10.1007/s43939-021-00012-0.

Sa, Rongjian, Benlong Luo, Zuju Ma, and Diwen Liu. 2022. ‘The Effect of the A-Site Cation on the Stability and Physical Properties of Vacancy-Ordered Double Perovskites A_2PtI_6 (A = Tl, K, Rb, and Cs)’. *Journal of Solid State Chemistry* 305 (January): 122714. https://doi.org/10.1016/j.jssc.2021.122714.

Saal, James E., Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. 2013. ‘Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)’. *Jom* 65 (11): 1501–9. https://doi.org/10.1007/s11837-013-0755-4.

Sabatini, Riccardo, Tommaso Gorni, and Stefano De Gironcoli. 2013. ‘Nonlocal van Der Waals Density Functional Made Simple and Efficient’. *Physical Review B - Condensed Matter and Materials Physics* 87 (4): 4–7. https://doi.org/10.1103/PhysRevB.87.041108.

Sachchidanand, Vivek Garg, Anil Kumar, and Pankaj Sharma. 2021. ‘Numerical Simulation of Novel Lead-Free $Cs_3Sb_2Br_9$ Absorber-Based Highly Efficient Perovskite Solar Cell’. *Optical Materials* 122 (PA): 111715. https://doi.org/10.1016/j.optmat.2021.111715.

Saidaminov, Makhsud I., Jawaher Almutlaq, Smritakshi Sarmah, Ibrahim Dursun, Ayan A. Zhumekenov, Raihana Begum, Jun Pan, Namchul Cho, Omar F. Mohammed, and Osman M. Bakr. 2016. ‘Pure Cs 4 PbBr 6 : Highly Luminescent Zero-Dimensional Perovskite Solids’. *ACS Energy Letters* 1 (4): 840–45. https://doi.org/10.1021/acsenergylett.6b00396.

Saliba, Michael, Taisuke Matsui, Ji-Youn Seo, Konrad Domanski, Juan-Pablo Correa-Baena, Mohammad Khaja Nazeeruddin, Shaik M. Zakeeruddin, et al. 2016. ‘Cesium-Containing Triple Cation Perovskite Solar Cells: Improved Stability, Reproducibility and High Efficiency’. *Energy & Environmental Science* 9 (6): 1989–97. https://doi.org/10.1039/C5EE03874J.

Sanchez-Lengeling, Benjamin, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. 2021. ‘A Gentle Introduction to Graph Neural Networks’. *Distill* 6 (9): e33. https://doi.org/10.23915/distill.00033.

Sanvito, Stefano, Corey Oses, Junkai Xue, Anurag Tiwari, Mario Zic, Thomas Archer, Pelin Tozman, Munuswamy Venkatesan, Michael Coey, and Stefano Curtarolo. 2017. ‘Accelerated Discovery of New Magnets in the Heusler Alloy Family’. *Science Advances* 3 (4): e1602241. https://doi.org/10.1126/sciadv.1602241.

Saparov, Bayrammurad, Feng Hong, Jon-Paul Sun, Hsin-Sheng Duan, Weiwei Meng, Samuel Cameron, Ian G. Hill, Yanfa Yan, and David B. Mitzi. 2015. ‘Thin-Film Preparation and Characterization of $Cs_3Sb_2I_9$: A Lead-Free Layered Perovskite Semiconductor’. *Chemistry of Materials* 27 (16): 5622–32. https://doi.org/10.1021/acs.chemmater.5b01989.

Saparov, Bayrammurad, and David B. Mitzi. 2016. ‘Organic–Inorganic Perovskites: Structural Versatility for Functional Materials Design’. *Chemical Reviews* 116 (7): 4558–96. https://doi.org/10.1021/acs.chemrev.5b00715.

Schleder, Gabriel R., Antonio C.M. Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. 2019. ‘From DFT to Machine Learning: Recent Approaches to Materials Science - A Review’. *JPhys Materials* 2 (3). https://doi.org/10.1088/2515-7639/ab084b.

Schlexer Lamoureux, Philomena, Kirsten T. Winther, Jose Antonio Garrido Torres, Verena Streibel, Meng Zhao, Michal Bajdich, Frank Abild-Pedersen, and Thomas Bligaard. 2019. ‘Machine Learning for Computational Heterogeneous Catalysis’. *ChemCatChem* 11 (16): 3581–3601. https://doi.org/10.1002/cctc.201900595.

Schlipf, Martin, and François Gygi. 2015. ‘Optimization Algorithm for the Generation of ONCV Pseudopotentials’. *Computer Physics Communications* 196 (November): 36–44. https://doi.org/10.1016/j.cpc.2015.05.011.

Schmidt, Jonathan, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel A. L. Marques. 2021. ‘Crystal Graph Attention Networks for the Prediction of Stable Materials’. *Science Advances* 7 (49): eabi7948. https://doi.org/10.1126/sciadv.abi7948.

Seko, Atsuto, Atsushi Togo, and Isao Tanaka. 2018. ‘Descriptors for Machine Learning of Materials Data’. In *Nanoinformatics*, edited by Isao Tanaka, 3–23. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-7617-6_1.

Setten, M. J. van, M. Giantomassi, E. Bousquet, M. J. Verstraete, D. R. Hamann, X. Gonze, and G. M. Rignanese. 2018. ‘The PSEUDODOJO: Training and Grading a 85 Element Optimized Norm-Conserving Pseudopotential Table’. *Computer Physics Communications* 226 (May): 39–54. https://doi.org/10.1016/j.cpc.2018.01.012.

Sham, L. J., and M. Schlüter. 1983. ‘Density-Functional Theory of the Energy Gap’. *Physical Review Letters* 51 (20): 1888–91. https://doi.org/10.1103/PhysRevLett.51.1888.

Shannon, R D. 1976. ‘Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides’. *Acta Crystallographica Section A* 32 (5): 751–67. https://doi.org/10.1107/S0567739476001551.

Shao, G. 2008. ‘Electronic Structures of Manganese-Doped Rutile TiO(2) from First Principles’. *J Phys Chem C* 112: 18677.

Sharma, Manas, Debabrata Mishra, and Jagadish Kumar. 2019. ‘First-Principles Study of the Structural and Electronic Properties of Bulk ZnS and Small Znn Sn Nanoclusters in the Framework of the DFT+U Method’. *Physical Review B* 100 (4): 045151. https://doi.org/10.1103/PhysRevB.100.045151.

Shen, Jiahong, Sean D Griesemer, Abhijith Gopakumar, Bianca Baldassarri, James E Saal, Muratahan Aykol, Vinay I Hegde, and Chris Wolverton. 2022. ‘Reflections on One Million Compounds in the Open Quantum Materials Database (OQMD)’. *Journal of Physics: Materials* 5 (3): 031001. https://doi.org/10.1088/2515-7639/ac7ba9.

Sherstinsky, Alex. 2020. ‘Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network’. *Physica D: Nonlinear Phenomena* 404 (March): 132306. https://doi.org/10.1016/j.physd.2019.132306.

Shirley, Eric L. 1996. ‘Optimal Basis Sets for Detailed Brillouin-Zone Integrations’. *Physical Review B - Condensed Matter and Materials Physics* 54 (23): 16464–69. https://doi.org/10.1103/PhysRevB.54.16464.

Sholl, David S., and Janice A. Steckel. 2009. *Density Functional Theory: A Practical Introduction*. 1st ed. Wiley. https://doi.org/10.1002/9780470447710.

Shu, Qiang, Ji Hui Yang, Shiyou Chen, Bing Huang, Hongjun Xiang, Xin Gao Gong, and Su Huai Wei. 2013. ‘$Cu_2Zn(Sn,Ge)Se_4$ and $Cu_2Zn(Sn,Si)Se_4$ Alloys as Photovoltaic Materials: Structural and Electronic Properties’. *Physical Review B - Condensed Matter and Materials Physics* 87 (11): 1–6. https://doi.org/10.1103/PhysRevB.87.115208.

Shur, Michael. 2005. ‘Semiconductors’. In *The Electrical Engineering Handbook,* 153–62. Elsevier. https://doi.org/10.1016/B978-012170960-0/50015-3.

Sierepeklis, Odysseas, and Jacqueline M. Cole. 2022. ‘A Thermoelectric Materials Database Auto-Generated from the Scientific Literature Using ChemDataExtractor’. *Scientific Data* 9 (1): 648. https://doi.org/10.1038/s41597-022-01752-1.

Simonyan, Karen, and Andrew Zisserman. 2015. ‘Very Deep Convolutional Networks for Large-Scale Image Recognition’.

Singh, Anupriya, Karunakara Moorthy Boopathi, Anisha Mohapatra, Yang Fang Chen, Gang Li, and Chih Wei Chu. 2018. ‘Photovoltaic Performance of Vapor-Assisted Solution-Processed Layer Polymorph of $Cs_3Sb_2I_9$’. *ACS Applied Materials and Interfaces* 10 (3): 2566–73. https://doi.org/10.1021/acsami.7b16349.

Singh, Anupriya, Nan Chieh Chiu, Karunakara Moorthy Boopathi, Yu Jung Lu, Anisha Mohapatra, Gang Li, Yang Fang Chen, Tzung Fang Guo, and Chih Wei Chu. 2019.

‘Lead-Free Antimony-Based Light-Emitting Diodes through the Vapor-Anion-Exchange Method’. Research-article. *ACS Applied Materials and Interfaces* 11 (38): 35088–94. https://doi.org/10.1021/acsami.9b10602.

Singh, Anupriya, Po-Ting Lai, Anisha Mohapatra, Chien-Yu Chen, Hao-Wu Lin, Yu-Jung Lu, and Chih Wei Chu. 2021. ‘Panchromatic Heterojunction Solar Cells for Pb-Free All-Inorganic Antimony Based Perovskite’. *Chemical Engineering Journal* 419 (September): 129424. https://doi.org/10.1016/j.cej.2021.129424.

Smith, Matthew D., Bridget A. Connor, and Hemamala I. Karunadasa. 2019. ‘Tuning the Luminescence of Layered Halide Perovskites’. *Chemical Reviews* 119 (5): 3104–39. https://doi.org/10.1021/acs.chemrev.8b00477.

Song, Zhilong, Xiwen Chen, Fanbin Meng, Guanjian Cheng, Chen Wang, Zhongti Sun, and Wan-Jian Yin. 2020. ‘Machine Learning in Materials Design: Algorithm and Application’. *Chinese Physics B* 29 (11): 116103. https://doi.org/10.1088/1674-1056/abc0e3.

Soriano, M., and J. J. Palacios. 2014. ‘Theory of Projections with Nonorthogonal Basis Sets: Partitioning Techniques and Effective Hamiltonians’. *Physical Review B* 90 (7): 075128. https://doi.org/10.1103/PhysRevB.90.075128.

Sorzano, C. O. S., J. Vargas, and A. Pascual Montano. 2014. ‘A Survey of Dimensionality Reduction Techniques’. arXiv. http://arxiv.org/abs/1403.2877.

S. Stein, Helge, Dan Guevarra, Paul F. Newhouse, Edwin Soedarmadji, and John M. Gregoire. 2019. ‘Machine Learning of Optical Properties of Materials – Predicting Spectra from Images and Images from Spectra’. *Chemical Science* 10 (1): 47–55. https://doi.org/10.1039/C8SC03077D.

Su, Rui, Zhaojian Xu, Jiang Wu, Deying Luo, Qin Hu, Wenqiang Yang, Xiaoyu Yang, et al. 2021. ‘Dielectric Screening in Perovskite Photovoltaics’. *Nature Communications* 12 (1): 2479. https://doi.org/10.1038/s41467-021-22783-z.

Supka, Andrew R., Troy E. Lyons, Laalitha Liyanage, Pino D’Amico, Rabih Al Rahal Al Orabi, Sharad Mahatara, Priya Gopal, et al. 2017. ‘AFLOWπ: A Minimalist Approach to High-Throughput Ab Initio Calculations Including the Generation of Tight-Binding Hamiltonians’. *Computational Materials Science* 136 (August): 76–84. https://doi.org/10.1016/j.commatsci.2017.03.055.

Swart Lab. 2023. ‘Swart Lab - Theoretical Chemistry’. 2023. https://www.marcelswart.eu/dft-poll/.

Tai, Qidong, Kai-Chi Tang, and Feng Yan. 2019. ‘Recent Progress of Inorganic Perovskite Solar Cells’. *Energy & Environmental Science* 12 (8): 2375–2405. https://doi.org/10.1039/C9EE01479A.

Takeshima, Hidenori. 2022. ‘Deep Learning and Its Application to Function Approximation for MR in Medicine: An Overview’. *Magnetic Resonance in Medical Sciences* 21 (4): 553–68. https://doi.org/10.2463/mrms.rev.2021-0040.

Tan, Zhifang, Manchen Hu, Guangda Niu, Qingsong Hu, Jinghui Li, Meiying Leng, Liang Gao, and Jiang Tang. 2019. ‘Inorganic Antimony Halide Hybrids with Broad Yellow Emissions’. *Science Bulletin* 64 (13): 904–9. https://doi.org/10.1016/j.scib.2019.05.016.

Tang, W, E Sanville, and G Henkelman. 2009. ‘A Grid-Based Bader Analysis Algorithm without Lattice Bias.’ *Journal of Physics. Condensed Matter : An Institute of Physics Journal* 21 (8): 084204. https://doi.org/10.1088/0953-8984/21/8/084204.

Tao, Jianmin, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. 2003. ‘Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids’. *Physical Review Letters* 91 (14): 3–6. https://doi.org/10.1103/PhysRevLett.91.146401.

Tao, Qiuling, Pengcheng Xu, Minjie Li, and Wencong Lu. 2021. ‘Machine Learning for Perovskite Materials Design and Discovery’. *Npj Computational Materials* 7 (1): 1–18. https://doi.org/10.1038/s41524-021-00495-8.

Tavadze, Pedram, Reese Boucher, Guillermo Avendaño-Franco, Keenan X. Kocan, Sobhit Singh, Viviana Dovale-Farelo, Wilfredo Ibarra-Hernández, Matthew B. Johnson, David S. Mebane, and Aldo H. Romero. 2021. ‘Exploring DFT+U Parameter Space with a Bayesian Calibration Assisted by Markov Chain Monte Carlo Sampling’. *Npj Computational Materials* 7 (1): 1–9. https://doi.org/10.1038/s41524-021-00651-0.

Tawfik, Sherif Abdulkader, and Salvy P. Russo. 2022. ‘Naturally-Meaningful and Efficient Descriptors: Machine Learning of Material Properties Based on Robust One-Shot Ab Initio Descriptors’. *Journal of Cheminformatics* 14 (1): 78. https://doi.org/10.1186/s13321-022-00658-9.

Theofylaktos, Lazaros, Kyro Odysseas Kosmatos, Eleni Giannakaki, Helen Kourti, Dimitris Deligiannis, Maria Konstantakou, and Thomas Stergiopoulos. 2019. ‘Perovskites with D-Block Metals for Solar Energy Applications’. *Dalton Transactions* 48 (26): 9516–37. https://doi.org/10.1039/C9DT01485C.

Thomas, Ankit Stephen. 2022. ‘A Review on Antimony-Based Perovskite Solar Cells’. *Equilibrium Journal of Chemical Engineering* 6 (2): 75. https://doi.org/10.20961/equilibrium.v6i2.64322.

Tian, Xinxin, Tao Wang, Lifang Fan, Yuekui Wang, Haigang Lu, and Yuewen Mu. 2018. ‘A DFT Based Method for Calculating the Surface Energies of Asymmetric MoP Facets’. *Applied Surface Science* 427: 357–62. https://doi.org/10.1016/j.apsusc.2017.08.172.

Togo, Atsushi, and Isao Tanaka. 2015. ‘First Principles Phonon Calculations in Materials Science’. *Scripta Materialia* 108 (November): 1–5. https://doi.org/10.1016/j.scriptamat.2015.07.021.

Toher, Cormac, and Stefano Curtarolo. 2023. ‘AFLOW for Alloys’. https://doi.org/10.48550/ARXIV.2310.16769.

Tomaszewski, P. E. 1994. ‘Crystal Structure and Phase Transitions in the $A_3B_2X_9$ Family of Crystals’. *Physica Status Solidi (B)* 181 (1): 15–21. https://doi.org/10.1002/pssb.2221810102.

Toriyama, M. Y., A. M. Ganose, M. Dylla, S. Anand, J. Park, M. K. Brod, J. Munro, K. A. Persson, A. Jain, and G. J. Snyder. 2021. ‘Comparison of the Tetrahedron Method to Smearing Methods for the Electronic Density of States’. https://doi.org/10.48550/ARXIV.2103.03469.

Toriyama, Michael Y., Alex M. Ganose, Maxwell Dylla, Shashwat Anand, Junsoo Park, Madison K. Brod, Jason M. Munro, Kristin A. Persson, Anubhav Jain, and G. Jeffrey Snyder. 2022. ‘How to Analyse a Density of States’. *Materials Today Electronics* 1 (May): 100002. https://doi.org/10.1016/j.mtelec.2022.100002.

Tran, Fabien, and Peter Blaha. 2009. ‘Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential’. *Physical Review Letters* 102 (22): 226401. https://doi.org/10.1103/PhysRevLett.102.226401.

Tu, Guangde, Vincenzo Carravetta, Olav Vahtras, and Hans Ågren. 2007. ‘Core Ionization Potentials from Self-Interaction Corrected Kohn-Sham Orbital Energies’. *The Journal of Chemical Physics* 127 (17): 174110. https://doi.org/10.1063/1.2777141.

Umedov, Shodruz T., Dhruba B. Khadka, Masatoshi Yanagida, Anastasia Grigorieva, and Yasuhiro Shirai. 2021. ‘A-Site Tailoring in the Vacancy-Ordered Double Perovskite Semiconductor Cs_2SnI_6 for Photovoltaic Application’. *Solar Energy Materials and Solar Cells* 230 (September): 111180. https://doi.org/10.1016/j.solmat.2021.111180.

Van De Walle, Chris G., and Richard M. Martin. 1987. ‘Theoretical Study of Band Offsets at Semiconductor Interfaces’. *Physical Review B* 35 (15): 8154–65. https://doi.org/10.1103/PhysRevB.35.8154.

Vapnik, Vladimir N. 2000. *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3264-1.

Vargas, Brenda, Estrella Ramos, Enrique Pérez-Gutiérrez, Juan Carlos Alonso, and Diego Solis-Ibarra. 2017. ‘A Direct Bandgap Copper–Antimony Halide Perovskite’. *Journal of the American Chemical Society* 139 (27): 9116–19. https://doi.org/10.1021/jacs.7b04119.

Varrassi, Lorenzo, Peitao Liu, Zeynep Ergönenc Yavas, Menno Bokdam, Georg Kresse, and Cesare Franchini. 2021. ‘Optical and Excitonic Properties of Transition Metal Oxide Perovskites by the Bethe-Salpeter Equation’. *Physical Review Materials* 5 (7): 1–17. https://doi.org/10.1103/PhysRevMaterials.5.074601.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. ‘Attention Is All You Need’. In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vaugier, Loïg, Hong Jiang, and Silke Biermann. 2012. ‘Hubbard U and Hund Exchange J in Transition Metal Oxides: Screening versus Localization Trends from Constrained Random Phase Approximation’. *Physical Review B - Condensed Matter and Materials Physics* 86 (16): 1–21. https://doi.org/10.1103/PhysRevB.86.165105.

Venkatesh, B., and J. Anuradha. 2019. ‘A Review of Feature Selection and Its Methods’. *Cybernetics and Information Technologies* 19 (1): 3–26. https://doi.org/10.2478/cait-2019-0001.

Verma, Pragya, and Donald G. Truhlar. 2016. ‘Does DFT+U Mimic Hybrid Density Functionals?’ *Theoretical Chemistry Accounts* 135 (8): 182. https://doi.org/10.1007/s00214-016-1927-4.

Volonakis, George, Marina R. Filip, Amir Abbas Haghighirad, Nobuya Sakai, Bernard Wenger, Henry J. Snaith, and Feliciano Giustino. 2016. ‘Lead-Free Halide Double Perovskites via Heterovalent Substitution of Noble Metals’. *Journal of Physical Chemistry Letters* 7 (7): 1254–59. https://doi.org/10.1021/acs.jpclett.6b00376.

Volonakis, George, Amir Abbas Haghighirad, Rebecca L. Milot, Weng H. Sio, Marina R. Filip, Bernard Wenger, Michael B. Johnston, Laura M. Herz, Henry J. Snaith, and Feliciano Giustino. 2017. ‘Cs2InAgCl6: A New Lead-Free Halide Double Perovskite with Direct Band Gap’. *Journal of Physical Chemistry Letters* 8 (4): 772–78. https://doi.org/10.1021/acs.jpclett.6b02682.

Vosko, S. H., L. Wilk, and M. Nusair. 1980. ‘Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis’. *Canadian Journal of Physics* 58 (8): 1200–1211. https://doi.org/10.1139/p80-159.

Vydrov, Oleg A., Jochen Heyd, Aliaksandr V. Krukau, and Gustavo E. Scuseria. 2006. ‘Importance of Short-Range versus Long-Range Hartree-Fock Exchange for the Performance of Hybrid Density Functionals’. *The Journal of Chemical Physics* 125 (7): 074106. https://doi.org/10.1063/1.2244560.

Wang, Aili, Chuantian Zuo, Xiaobin Niu, Liming Ding, Jianning Ding, and Feng Hao. 2023. ‘Recent Promise of Lead-Free Halide Perovskites in Optoelectronic Applications’. *Chemical Engineering Journal* 451 (January): 138926. https://doi.org/10.1016/j.cej.2022.138926.

Wang, Feng, Xiao-Ke Liu, and Feng Gao. 2019. ‘Fundamentals of Solar Cells and Light-Emitting Diodes’. In *Advanced Nanomaterials for Solar Cells and Light Emitting Diodes*, 1–35. Elsevier. https://doi.org/10.1016/B978-0-12-813647-8.00001-1.

Wang, Minghao, Wei Wang, Ben Ma, Wei Shen, Lihui Liu, Kun Cao, Shufen Chen, and Wei Huang. 2021. ‘Lead-Free Perovskite Materials for Solar Cells’. *Nano-Micro Letters* 13 (1): 62. https://doi.org/10.1007/s40820-020-00578-z.

Wang, Qi, Xiaoming Wang, Zhi Yang, Ninghao Zhou, Yehao Deng, Jingjing Zhao, Xun Xiao, et al. 2019. ‘Efficient Sky-Blue Perovskite Light-Emitting Diodes via Photoluminescence Enhancement’. *Nature Communications* 10 (1): 5633. https://doi.org/10.1038/s41467-019-13580-w.

Wang, Xiaoyu, Nasir Ali, Gang Bi, Yao Wang, Qibin Shen, Arash Rahimi-Iman, and Huizhen Wu. 2020. ‘Lead-Free Antimony Halide Perovskite with Heterovalent Mn^{2+} Doping’. *Inorganic Chemistry* 59 (20): 15289–94. https://doi.org/10.1021/acs.inorgchem.0c02252.

Ward, Logan, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. 2016. ‘A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials’. *Npj Computational Materials* 2 (1): 16028. https://doi.org/10.1038/npjcompumats.2016.28.

Ward, Logan, Alexander Dunn, Alireza Faghaninia, Nils E.R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, et al. 2018. ‘Matminer: An Open Source Toolkit for Materials Data Mining’. *Computational Materials Science* 152 (September): 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018.

Wasserman, A.L. 2005. ‘Effective Masses’. In *Encyclopedia of Condensed Matter Physics*, 1–5. Elsevier. https://doi.org/10.1016/B0-12-369401-9/00457-5.

Wei, Qianwen, Mehri Ghasemi, Rongfei Wang, Chong Wang, Juan Wang, Weijie Zhou, Baohua Jia, Yu Yang, and Xiaoming Wen. 2023. ‘Metal Halide Perovskite Alloy: Fundamental, Optoelectronic Properties and Applications’. *Advanced Photonics Research* 4 (2): 2200236. https://doi.org/10.1002/adpr.202200236.

Welborn, Matthew, Lixue Cheng, and Thomas F. Miller. 2018. ‘Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis’. *Journal of Chemical Theory and Computation* 14 (9): 4772–79. https://doi.org/10.1021/acs.jctc.8b00636.

Weston, L., H. Tailor, K. Krishnaswamy, L. Bjaalie, and C. G. Van de Walle. 2018. ‘Accurate and Efficient Band-Offset Calculations from Density Functional Theory’. *Computational Materials Science* 151 (May): 174–80. https://doi.org/10.1016/j.commatsci.2018.05.002.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. ‘Transformers: State-of-the-Art Natural Language Processing’. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by Qun Liu and David Schlangen, 38–45. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Wu, Daofu, Benjun Huo, Yanyi Huang, Xusheng Zhao, Jiayu Yang, Ke Hu, Xinchun Mao, Peng He, Qiang Huang, and Xiaosheng Tang. 2022. ‘Synthesis of Stable Lead-Free $Cs_3Sb_2(Br_xI_{1-x})_9$ ($0 \leq x \leq 1$) Perovskite Nanoplatelets and Their Application in CO_2 Photocatalytic Reduction’. *Small* 18 (12): 2106001. https://doi.org/10.1002/smll.202106001.

Wu, Yabi, Predrag Lazic, Geoffroy Hautier, Kristin Persson, and Gerbrand Ceder. 2013. ‘First Principles High Throughput Screening of Oxynitrides for Water-Splitting Photocatalysts’. *Energy Environ. Sci.* 6 (1): 157–68. https://doi.org/10.1039/C2EE23482C.

Xiao, Zewen, Ke-Zhao Du, Weiwei Meng, Jianbo Wang, David B. Mitzi, and Yanfa Yan. 2017. ‘Intrinsic Instability of Cs_2 In(I)M(III)X_6 (M = Bi, Sb; X = Halogen) Double Perovskites: A Combined Density Functional Theory and Experimental Study’. *Journal of the American Chemical Society* 139 (17): 6054–57. https://doi.org/10.1021/jacs.7b02227.

Xu, L., S. Yuan, H. Zeng, and J. Song. 2019. ‘A Comprehensive Review of Doping in Perovskite Nanocrystals/Quantum Dots: Evolution of Structure, Electronics, Optics, and Light-Emitting Diodes’. *Materials Today Nano* 6 (June): 100036. https://doi.org/10.1016/j.mtnano.2019.100036.

Xu, Meigui, Hainan Sun, Wei Wang, Yujuan Shen, Wei Zhou, Jun Wang, Zhi Gang Chen, and Zongping Shao. 2020. ‘Scandium and Phosphorus Co-Doped Perovskite Oxides as High-Performance Electrocatalysts for the Oxygen Reduction Reaction in an Alkaline Solution’. *Journal of Materials Science and Technology* 39: 22–27. https://doi.org/10.1016/j.jmst.2019.09.007.

Yamada, Koji, Hiroshi Sera, Shigeko Sawada, Hironobu Tada, Tsutomu Okuda, and Haruhiko Tanaka. 1997. ‘Reconstructive Phase Transformation and Kinetics of $Cs_3Sb_2I_9$ by Means of Rietveld Analysis of X-Ray Diffraction and 127 I NQR’ 325 (134): 319–25.

Yang, Jiaqi, and Arun Mannodi-Kanakkithodi. 2022. ‘High-Throughput Computations and Machine Learning for Halide Perovskite Discovery’. *MRS Bulletin* 47 (9): 940–48. https://doi.org/10.1557/s43577-022-00414-2.

Yang, Xiangrui. 2022. ‘Leveraging Orbital Information and Atomic Feature in Deep Learning Model’. https://doi.org/10.48550/ARXIV.2211.11543.

Yang, Yi, Fei Gao, Shiwu Gao, and Su Huai Wei. 2018. ‘Origin of the Stability of Two-Dimensional Perovskites: A First-Principles Study’. *Journal of Materials Chemistry A* 6 (30): 14949–55. https://doi.org/10.1039/c8ta01496e.

Yin, Wan Jian, Yanfa Yan, and Su Huai Wei. 2014. ‘Anomalous Alloy Properties in Mixed Halide Perovskites’. *Journal of Physical Chemistry Letters* 5 (21): 3625–31. https://doi.org/10.1021/jz501896w.

Yip, Sidney, ed. 2005. *Handbook of Materials Modeling*. Dordrecht ; New York: Springer.

Yoo, Jason J., Gabkyung Seo, Matthew R. Chua, Tae Gwan Park, Yongli Lu, Fabian Rotermund, Young-Ki Kim, et al. 2021. ‘Efficient Perovskite Solar Cells via Improved Carrier Management’. *Nature* 590 (7847): 587–93. https://doi.org/10.1038/s41586-021-03285-w.

Yu, Jinna, Yuk Ming Tang, Ka Yin Chau, Raima Nazar, Sajid Ali, and Wasim Iqbal. 2022. ‘Role of Solar-Based Renewable Energy in Mitigating CO_2 Emissions: Evidence from Quantile-on-Quantile Estimation’. *Renewable Energy* 182 (January): 216–26. https://doi.org/10.1016/j.renene.2021.10.002.

Yu, Maituo, Shuyang Yang, Chunzhi Wu, and Noa Marom. 2020. ‘Machine Learning the Hubbard U Parameter in DFT+U Using Bayesian Optimization’. *Npj Computational Materials* 6 (1): 1–6. https://doi.org/10.1038/s41524-020-00446-9.

Zhang, Boyu, Mushen Zhou, Jianzhong Wu, and Fuchang Gao. 2022. ‘Predicting the Materials Properties Using a 3D Graph Neural Network With Invariant Representation’. *IEEE Access* 10: 62440–49. https://doi.org/10.1109/ACCESS.2022.3181750.

Zhang, Gaoqian, Pengjie Song, Zhaohui Shen, Bo Qiao, Dandan Song, Jingyue Cao, Zheng Xu, Wageh Swelm, Ahmed Al-Ghamdi, and Suling Zhao. 2020. ‘$CsPbBr_3$@$CsPbBr_{3-x}Cl_x$ Perovskite Core-Shell Heterojunction Nanowires via a Postsynthetic Method with HCl Gas’. *ACS Omega* 5 (20): 11578–84. https://doi.org/10.1021/ACSOMEGA.0C00824.

Zhang, Jian Min, Qing Pang, Ke Wei Xu, and Vincent Ji. 2008. ‘First-Principles Study of the (001) Surface of Cubic $PbTiO_3$’. *Surface and Interface Analysis* 40 (10): 1382–87. https://doi.org/10.1002/sia.2911.

Zhang, Jian, Ying Yang, Hui Deng, Umar Farooq, Xiaokun Yang, Jahangeer Khan, Jiang Tang, and Haisheng Song. 2017. 'High Quantum Yield Blue Emission from Lead-Free Inorganic Antimony Halide Perovskite Colloidal Quantum Dots'. *ACS Nano* 11 (9): 9294–9302. https://doi.org/10.1021/acsnano.7b04683.

Zhang, Jin, Chen Yang, Yulong Liao, Shijie Li, Pengfei Yang, Yingxue Xi, Changlong Cai, and Weiguo Liu. 2023. 'Investigating Optical Adsorption Properties of Lead-free Double Perovskite Semiconductors $Cs_2SnI_{6-x}Br_x$(x = 0–6) via First Principles Calculation'. *Microwave and Optical Technology Letters* 65 (5): 1017–23. https://doi.org/10.1002/mop.33059.

Zhang, Lei, Juhong Miao, Jingfa Li, and Qingfang Li. 2020. 'Halide Perovskite Materials for Energy Storage Applications'. *Advanced Functional Materials* 30 (40): 2003653. https://doi.org/10.1002/adfm.202003653.

Zhang, Pingli, Gangbei Zhu, Ying Shi, Yunpeng Wang, Jiahua Zhang, Luchao Du, and Dajun Ding. 2018. 'Ultrafast Interfacial Charge Transfer of Cesium Lead Halide Perovskite Films $CsPbX_3$ (X = Cl, Br, I) with Different Halogen Mixing'. *Journal of Physical Chemistry C* 122 (48): 27148–55. https://doi.org/10.1021/acs.jpcc.8b07237.

Zhang, Xinyuan, Lina Li, Zhihua Sun, and Junhua Luo. 2019. 'Rational Chemical Doping of Metal Halide Perovskites'. *Chemical Society Reviews* 48 (2): 517–39. https://doi.org/10.1039/C8CS00563J.

Zhang, Zhenbao, Yubo Chen, Ziyang Dai, Shaozao Tan, and Dengjie Chen. 2019. 'Promoting Hydrogen-Evolution Activity and Stability of Perovskite Oxides via Effectively Lattice Doping of Molybdenum'. *Electrochimica Acta* 312: 128–36. https://doi.org/10.1016/j.electacta.2019.04.163.

Zhao, Zong Yan, Qing Lu Liu, and Wen Wu Dai. 2016. 'Structural, Electronic, and Optical Properties of $BiOX_{1-x}Y_x$ (X, Y = F, Cl, Br, and I) Solid Solutions from DFT Calculations'. *Scientific Reports* 6 (July): 1–12. https://doi.org/10.1038/srep31449.

Zheng, Kaibo, Qiushi Zhu, Mohamed Abdellah, Maria E. Messing, Wei Zhang, Alexander Generalov, Yuran Niu, Lynn Ribaud, Sophie E. Canton, and Tõnu Pullerits. 2015. 'Exciton Binding Energy and the Nature of Emissive States in Organometal Halide Perovskites'. *The Journal of Physical Chemistry Letters* 6 (15): 2969–75. https://doi.org/10.1021/acs.jpclett.5b01252.

Zhou, Jian, Qiang Sun, Qian Wang, and Puru Jena. 2015. 'High-Temperature Superconductivity in Heavily N- or B-Doped Graphene'. *Physical Review B* 92 (6): 064505. https://doi.org/10.1103/PhysRevB.92.064505.

Zhou, Shu, Yaping Ma, Guodong Zhou, Xin Xu, Minchao Qin, Yuhao Li, Yao Jane Hsu, et al. 2019. 'Ag-Doped Halide Perovskite Nanocrystals for Tunable Band Structure and Efficient Charge Transport'. *ACS Energy Letters* 4 (2): 534–41. https://doi.org/10.1021/acsenergylett.8b02478.

Zhou, Siyuan, Hao Tian, Xiaoyu Kuang, Siyu Jin, Miao Yu, Jichao Chen, and Aijie Mao. 2024. 'Piezochromic Effects and Low-Pressure Superconductivity Discovered in Inorganic Halide Perovskite $RbPbI_3$'. *Journal of Materials Chemistry C* 12 (1): 245–53. https://doi.org/10.1039/D3TC03535B.

Zunger, Alex. 2018. 'Inverse Design in Search of Materials with Target Functionalities'. *Nature Reviews Chemistry* 2 (4): 1–16. https://doi.org/10.1038/s41570-018-0121.

APPENDICES

APPENDIX A: Supporting Information for Theoretical Background

A.1 Derivation of dielectric constant and absorption coefficient

Formally, the absorption process can be approximated with a perturbation on the Hamiltonian operator in *Equation 2*. This perturbation, accounting for the electromagnetic field of light, can be simplified to first order as:

$$\widehat{H}' = \widehat{H}_{\text{em}} \approx \mathrm{q}\hat{\boldsymbol{r}} \cdot \widehat{\boldsymbol{E}}, \tag{A1}$$

which represents an electrical dipole. This expression is used to evaluate the transition probability rate (R) as a function of photon energy from electrons from the initial (i) state on the valence band (v) to the final (f) state on the conduction band (c), using Fermi's golden rule:

$$\begin{aligned} R(\hbar\omega) &= \frac{2\pi}{\hbar}\left|\widehat{H}'_{fi}\right|^2 \delta(E_f - E_i - \hbar\omega) = \\ &= \frac{2\pi}{\hbar}\int_{\boldsymbol{k}_c}\int_{\boldsymbol{k}_v}\left|\left\langle c\middle|\widehat{H}_{\text{em}}\middle|v\right\rangle\right|^2 \delta(E_c(\boldsymbol{k}_c) - E_v(\boldsymbol{k}_v) - \hbar\omega)d\boldsymbol{k}_c d\boldsymbol{k}_v, \end{aligned} \tag{A2}$$

where δ denotes the Dirac-delta function. The term $\left\langle c\middle|\widehat{H}_{\text{em}}\middle|v\right\rangle$ is directly associated with momentum matrix elements from k-p theory (Grundmann 2010), denoted p_{cv}. Considering the light-induced momentum is very small, we can limit integration to valid k-points, where $\boldsymbol{k}_{\text{c}} = \boldsymbol{k}_{\text{v}}$. By calculating the lost power of the electric field, the imaginary part of the dielectric function is obtained as follows:

$$\varepsilon_i\,(\omega) = \frac{1}{4\pi\epsilon_0}\left(\frac{2\pi e}{m\omega}\right)^2 |p_{cv}|^2 \int_{\boldsymbol{k}} \delta(E_c(\boldsymbol{k}) - E_v(\boldsymbol{k}) - \hbar\omega)d\boldsymbol{k} \tag{A3}$$

And through the Kramers-Kronig relations (Grundmann 2010), the real part can be derived from (10) as:

$$\varepsilon_r\,(\omega) = 1 + \int_{\boldsymbol{k}} \frac{e^2}{\epsilon_0 m\omega_{cv}^2}\,\frac{2|p_{cv}|^2}{m\hbar\omega_{cv}}\,\frac{1}{1-(\omega^2\,/\omega_{cv}^2)}d\boldsymbol{k}. \tag{A4}$$

At this point, an explicit formulation for the integral concerning the possible transitions can be found. If the density of states (DOS) of a system is defined as the number of allowed states per unit energy lying in the energy range between E and E+dE, represented by:

$$n(E) = \sum_i \delta(E - E_i) \tag{A5}$$

For all eigenstates $|i\rangle$ of the system. Similarly, we can define a joint density of states (JDOS) to measure the number of allowed optical transitions between the occupied valence band electronic states and the unoccupied conduction band electronic states separated by a given photon energy. Assuming a parabolic three dimensional (and doubly degenerated on spin) from a minimum point in band separation (Peter and Cardona 2010), results in :

$$n_j(E) = \begin{cases} A(E-E_g)^{1/2}, E > E_g \\ 0, \quad E < E_g \end{cases} \tag{A6}$$

The equivalence with the integral in Eq. 10 can be observed. By substituting $\int n_j(E_{cv})dE_{cv}$ in Eq. 10 and considering the relationship of the absorption coefficient, α, with ε_i:

$$\alpha = \frac{\omega}{\tilde{n}c}\varepsilon_i , \tag{A7}$$

Where $\tilde{n}$ is the refractive index. This leads to the well-known relationship for the absorption coefficient of direct transitions:

$$\alpha \propto (\hbar\omega - E_g)^{1/2} \tag{A8}$$

In the case of an indirect band gap, extending Fermi's golden rule is necessary to include second-order perturbation, allowing accommodation for both phonon-electron and photon-electron interactions as detailed elsewhere (Peter and Cardona 2010). The indirect transitions can be derived following a similar path as for direct transitions but now results in a quadratic dependence on energy:

$$\alpha \propto (\hbar\omega - E_g - \hbar\omega_{ph})^2 \tag{A9}$$

where ω_{ph} is the phonon frequency. Despite the higher order, the two-particle process is much less probable than simple photon absorption and the coefficient is about 10^{-3} smaller.

Moreover, in experiments, another commonly observed feature is an exponential tail in the absorption coefficient below the band gap, referred to as the Urbach tail, expressed as:

$$\alpha = \alpha_0 . exp\left(\frac{\hbar\omega - E_f}{E_0}\right), \tag{A10}$$

in this $E_f < E_g$ being named Urbach focus and E_0 is the characteristic width of the absorption edge or Urbach energy. The Urbach tail is attributed to transitions between band tails below the band edges. These tails may originate from imperfections in the crystal lattice, such as defects or doping, and fluctuations in electronic energy bands caused by lattice vibrations. For most semiconductors, E_0 is typically around 50 meV or less. However, amorphous materials and some halide perovskites may exceed 100 meV (Bacalis, Economou, and Cohen 1988; S. Y. Kim et al. 2019).

A.2 Dynamical stability

Thermodynamic stability is assessed by considering equilibrium conditions under specific external parameters, like temperature and pressure, to determine a material's formation from its constituent phases. However, once established the candidate material can be formed, an essential consideration emerges: the potential for it to transform into an alternative structure not initially accounted for. At equilibrium, regardless of the atomic movements, the system's potential energy consistently increases. Thermodynamic stability, crucial as it is, does not encompass how a system responds in real-time to external changes or kinetic influences. This limitation prompts the necessity to explore dynamical stability (Malyi, Sopiha, and Persson 2019; Bartel 2022).

Dynamical stability measures system's resilience to perturbations such as vibrations or small displacements. Phonon analysis and ab-initio molecular dynamics (AIMD) are two methods gauge this stability by examining a material's vibrational modes or simulating atomic motion using quantum mechanics. These approaches offer deeper insights than thermodynamics alone but often demand substantial computational resources. Phonon analysis calculates the dynamical matrix to describe atomic forces and interactions, requiring extensive computation due to the need for high precision, especially for systems with many atoms. Meanwhile, AIMD simulates real-time dynamics, demanding numerous cost intensive steps to model the quantum mechanics of electrons and ions accurately. Thus, both techniques demand significant computational resources and time, making their application usually resource-prohibitive for high-throughput calculations (Mortazavi et al. 2020; Bartel 2022).

Dynamical stability assessment wasn't considered in this work because the examined structures were either known to be stable (halogen-alloyed $Cs_3Sb_2X_9$ on *Chapter 3*) or we focused on doping the structure, aiming for lower concentrations (transition metal-doped $Cs_3Sb_2I_9$ on *Chapter* 4 and multiple composition screening on *Chapter 6*). In both cases, thermodynamic stability holds more significance as it offers a reliable means to compare structures within the same chemical system or infer stability at lower concentrations in doping scenarios (Gebhardt and Rappe 2018; Xiao et al. 2017; Pramchu, Jaroenjittichai, and Laosiritaworn 2019).

For instance, structures doped at a 25% concentration, as explored in the investigation conducted in *Chapter 4* for $Cs_3Sb_2I_9$ polymorphs, if deemed stable by thermodynamic criteria, are highly likely to maintain stability at lower concentrations. Conversely, an imaginary phonon frequency in the same structure and concentration might imply instability or suggest an alternative structure that cannot be inferred for lower concetration. This dependency arises because phonon analysis is significantly concentration dependent (J. Zhou et al. 2015). Additionally, imaginary modes can sometimes vanish when considering phonon-phonon interactions in temperature-dependent phonon spectra calculations or might result from approximations in the chosen density functional (Bartel 2022). Similarly, AIMD simulations imply even higher computational cost and cannot be used to infer stability on lower concentrations.

Nevertheless, exploring the dynamical stability of specific doped structures across various concentrations—particularly those with most negative decomposition energies in our investigations—has the potential to yield valuable insights. This possibility is being considered for future research.

A.3 Born-Oppenheimer approximation

The ground-state properties of a physical system are obtained by solving the time-independent Schrödinger equation to determine the wavefunction $\Psi(\{\mathbf{R}_\alpha\},\{\mathbf{r}_i\})$ as given by:

$$\hat{H}\Psi(\{\mathbf{R}_\alpha\},\{\mathbf{r}_i\}) = E\Psi(\{\mathbf{R}_\alpha\},\{\mathbf{r}_i\}), \qquad \text{(A11)}$$

in this context, $\{\mathbf{R}_\alpha\}$ represents the positions of nuclei, while $\{\mathbf{r}_i\}$ represents the positions of electrons. The total energy of the system is denoted by E. However, the Schrödinger equation has analytical solutions only for simple systems, therefore for many-body problems, numerical methods and approximations are necessary. One widely used approximation is the Born-Oppenheimer approximation (Born, Huang, and Lax 1955), which decouples the dynamics of electrons and nuclei. This approximation relies on the difference in time scales between nuclear and electron motion, allowing electrons to quickly adapt to changes in atomic positions. By separating the electronic and nuclear motions, the wave function can be expressed as the product of a nuclear and an electronic component, for example:

$$\Psi(\{\mathbf{R}_\alpha\},\{\mathbf{r}_i\}) = \chi_n(\{\mathbf{R}_\alpha\})\psi_n(\{\mathbf{R}_\alpha\}',\{\mathbf{r}_i\}), \tag{A12}$$

where $\{\mathbf{R}_\alpha\}'$ indicates that the dependence on the nuclear positions is parametric in the electronic function. Consequently, equation (2) is separated in a nuclear part and an electronic part, simplifying to

$$\hat{H}_{el}\psi_n(\{\mathbf{r}_i\},\{\mathbf{R}_\alpha\}') = \epsilon_n(\{\mathbf{R}_\alpha\})\psi_n(\{\mathbf{r}_i\},\{\mathbf{R}_\alpha\}'), \tag{A13}$$

where the electronic Hamiltonian is given by $\hat{H}_{el} = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{Ne}$. The total energy for fixed nuclear positions, $E_n(\{\mathbf{R}_\alpha\})$, incorporates a constant nuclear repulsion term and defines a potential energy surface for the nuclear dynamics which can be solved separately. While this approach remains valid for a wide range of systems, it may encounter limitations in cases where there is significant coupling between electronic excitations and nuclear vibrations. In the context of the present study, the properties investigated in the materials did not require consideration of such coupling and could be effectively examined using the Born-Oppenheimer approximation.

A.4 (Projected) Density of states

Density of states (DOS) for a solid is defined as the number of one-electron levels between energies E and E+dE. Within the Kohn-Sham formalism, the equation is given by (Martin 2020):

$$g(E) = \frac{1}{N_k}\sum_n \sum_{\boldsymbol{k}} \delta(\varepsilon_{n\boldsymbol{k}} - E), \tag{A14}$$

where N_k is the number of sampled k-points and $\varepsilon_{n\boldsymbol{k}}$ denotes the energy of an electron, the DOS has units of inverse energy. Throughout the literature, a duplication of each dimension of the k-grid (yielding an eightfold denser grid), ensures well-converged Density of States (DOS) curves. We applied the same procedure in this work, passing the converged electronic density as input for a non-self-consistent field (NSCF) calculation with the denser k-point grid (Michael Y. Toriyama et al. 2022).

Considering the case of the projected density of states (PDOS), the same integration is performed with a further projection of the wavefunctions onto orthogonalized atomic wavefunctions (Soriano and Palacios 2014) producing the following expression:

$$g_{\mu}(E) = \frac{1}{N_k}\sum_{n}\sum_{\boldsymbol{k}}\sum_{\nu,\mu} c^{*}_{n\nu,\boldsymbol{k}} c_{n\mu,\boldsymbol{k}} S_{\nu\mu} \delta(\varepsilon_{n\boldsymbol{k}} - E), \tag{A15}$$

represents the density of states on the projection orbital μ, where $S_{\nu\mu}$ represents the overlap matrix of the atomic basis and the coefficients correspond to the projections of the KS orbitals $\phi_{n\boldsymbol{k}}$.

Additionally, an appropriate handling of occupations is necessary to evaluate the DOS. Throughout this thesis, the smearing method with a gaussian smearing of 0.005 eV has been proven successful to converge the electronic density even in the presence of dopants and improved the geometric optimization cycle. However, to generate DOS and PDOS curves the NSCF calculation is performed applying the tetrahedron method (Peter E. Blöchl, Jepsen, and Andersen 1994) which is proven to provide more accurate curves with lower number of k-points (M. Y. Toriyama et al. 2021).

A.5 Bader and Löwdin charge analysis

Bader charge analysis (Bader 1990) was utilized to partition charges within the atoms of the structures investigated in this work. This method involves dividing space into Bader volumes using surfaces where the gradient of electron density equals zero. Therefore, at each point on these surfaces, $\boldsymbol{\nabla} n(r) = 0$. Each Bader volume encompasses the maximum of electron density associated with an ion's position. Consequently, this method enables the breakdown of electron density contributions from each atom by integrating density across these volumes, resulting in Bader charges, q_i^{Bader} for the atoms. The Bader charges were all computed using the Bader Charge Analysis code (Henkelman, Arnaldsson, and Jónsson 2006), while the Bader effective charges q_i^{eff} for each atom were derived as $q_i^{eff} = Z_i^{val} - q_i^{Bader}$, where Z_i^{val} represents the number of valence electrons explicitly included in the pseudopotential of the atomic species i in the DFT calculation.

Considering the projection of the wavefunctions onto orthogonalized atomic wavefunctions, represented with greek letters, we can define the density matrix **P** and the overlap matrix **S** for the system as:

$$P_{\nu\mu} = \sum_{n}\sum_{\boldsymbol{k}}\sum_{\nu,\mu} f(\varepsilon_{n\boldsymbol{k}}) c^{*}_{n\nu,\boldsymbol{k}} c_{n\mu,\boldsymbol{k}}, \tag{A16}$$

and,

$$S_{\nu\mu} = \langle \nu | \mu \rangle = \int \nu^{*}.\mu \; d^{3}\boldsymbol{r} \tag{A17}$$

where $f(\varepsilon_{n\boldsymbol{k}})$ is the occupation of the quantum state $n\boldsymbol{k}$, it is straightforward that the number of electrons N_e is equal to:

$$N_e = \sum_{\nu,\mu} (\mathbf{PS})_{\nu\mu} \tag{A18}$$

One can then decide to partition the electron population by associating non-intersecting subsets of the basis set to atoms, typically by taking those centered on atom A as belonging to A. We will denote this as ν∈A and define the Mulliken charge on A as:

$$q_A^{mulliken} = Z_A - \sum_{\nu \in A}\sum_{\mu} (\mathbf{PS})_{\nu\mu} \tag{A19}$$

where Z_A is the atomic number of atom A (Jensen 2017). Löwdin charges are a direct refinement of the Mulliken method aiming to conserve the dipole moment in a two-center charge distribution (Löwdin 1953) and are obtained from the following transformation:

$$q_A^{mulliken} = Z_A - \sum_{\nu \in A}\sum_{\mu} (\mathbf{S}^{\frac{1}{2}}\mathbf{PS}^{-\frac{1}{2}})_{\nu\mu} \tag{A20}$$

Löwdin charges are straightforwardly obtained along the projection performed in the PDOS calculation in the DFT codes considered here. These types of analysis are particularly susceptible to the basis set and therefore may not reflect real charge distribution when using simple atomic wavefunctions. Additionally, intramolecular basis set superposition error is also an important factor that is hard to measure on these methods (Jensen 2017).

Bader charge analysis stands out as a more dependable and theoretically robust method compared to projections on atomic basis sets for assessing charge transfer. Its strength lies in its inherent topological nature, whereas techniques like Löwdin charges often face issues due to the absence of rotational invariance (Davidson and Clark 2022). Consequently, Löwdin charges found less frequent application on our study, despite their valuable projection information. For instance, they were useful in analysing spin-density distribution among dopants, as illustrated in Table C6.

A.6 Optical properties calculation

A natural physical quantity to study in first-principles simulations of optical properties is the full dielectric function, a complex frequency-dependent quantity. In the following we will show the basic equations that relate this macroscopic property of a material to its underling microscopic electronic structure. The general expression of the independent particle dielectric function1 is simplified in the optical limit to the IP dielectric function into two separate contributions, an intraband Drude-like term to the conduction electrons at the Fermi surface and an interband term due to vertical transitions between occupied and unoccupied bands:

$$\varepsilon_{IP}(\hat{\boldsymbol{q}},\omega)=\varepsilon_{IP}^{inter}(\hat{\boldsymbol{q}},\omega)+\varepsilon_{IP}^{intra}(\hat{\boldsymbol{q}},\omega) \tag{A21}$$

where

$$\varepsilon_{IP}^{inter}(\hat{\boldsymbol{q}},\omega)=1-\frac{4\pi}{V}\sum_{\boldsymbol{k}}\sum_{n\neq n'}\frac{|\langle\phi_{n\boldsymbol{k}}|\hat{\boldsymbol{q}}\cdot\boldsymbol{v}|\phi_{n\boldsymbol{k}}\rangle|^2}{(E_{n'\boldsymbol{k}}-E_{n\boldsymbol{k}})^2}\frac{f_{n\boldsymbol{k}}-f_{n'\boldsymbol{k}}}{\omega-(E_{n'\boldsymbol{k}}-E_{n\boldsymbol{k}})+i\eta}, \tag{A22}$$

and,

$$\varepsilon_{IP}^{intra}(\hat{\boldsymbol{q}},\omega)=\frac{\omega_D^2(\hat{\boldsymbol{q}})}{\omega(\omega+i\gamma)}, \tag{A23}$$

where the Drude plasma frequency is obtained from:

$$\omega_D^2(\hat{\boldsymbol{q}})=\frac{4\pi}{V}\sum_{\boldsymbol{k}}\sum_{n}|\langle\phi_{n\boldsymbol{k}}|\hat{\boldsymbol{q}}\cdot\boldsymbol{v}|\phi_{n\boldsymbol{k}}\rangle|^2\left(-\frac{\partial f}{\partial E}\right). \tag{A24}$$

This last contribution is only relevant when metallic behavior is presented. The velocity operator term[‡] is **v** = −i$[\boldsymbol{r},H^{KS}]$, η and γ are empirical broadening terms, and $f_{n\boldsymbol{k}}$ gives the occupation according to the Fermi-Dirac distribution of the KS Bloch state $\phi_{n\boldsymbol{k}}$.

In this work the SIMPLE code (Prandini et al. 2019) distributed in the Quantum ESPRESSO package was applied to obtain the dielectric function as defined above, this implementation makes use of the Shirley interpolation method which obtains a set of basis functions best suited for integrations in the Brillouin zone reducing the computational cost expressively (Shirley 1996; Prendergast and Louie 2009). The basic idea of this optimal basis (OB) method is to obtain a reduced set of basis

[‡] $[\boldsymbol{r},H^{KS}]$ designates a commutator operation, which translates in this case to: $(\boldsymbol{r}H^{KS}-H^{KS}\boldsymbol{r})$

functions, indicated with the notation $\{b_i\}$, to represent the periodic part of the Bloch wavefunctions at any k-point inside the BZ. This basis set is constructed starting from the periodic KS states $\{u_{n\boldsymbol{k}}\}$ calculated on an initial grid of N_k k-points, holding the following relationship:

$$u_{n\boldsymbol{k}}(\boldsymbol{r}) \cong \sum_{i=1}^{N_b} \tilde{b}_i^{n\boldsymbol{k}}\, b_i(\boldsymbol{r}). \tag{A25}$$

Once the OB is constructed it is possible to obtain the periodic part of the Bloch wavefunctions at a generic k-point following a interpolation procedure (Prendergast and Louie 2009). This allows to perform the fine samplings of the BZ required for the calculations. The dimension N_b of the OB is directly dictated by the threshold s_b passed by the user, in our case s_b = 0.1 bohr^3 yielded well converged curves.

The calculation of the matrix elements of the Hamiltonian in terms of OB and the subsequent diagonalization of the matrix for each k-point gives the coefficients $\tilde{b}_i^{n\boldsymbol{k}}$ and the band energies $E_{n\boldsymbol{k}}$ for all the bands included in the calculation. Finally, one only needs to compute the matrix elements of the k-dependent velocity operator which becomes:

$$\boldsymbol{v}(\boldsymbol{k}) = -i[\boldsymbol{r}, H^{KS}] \;=\; -i\nabla + \boldsymbol{k} \;-i\;[r, V^{nl}(\boldsymbol{k})]. \tag{A26}$$

The first two two terms are easily obtained from the first diagonalization of the transformed Hamiltonian, the last term involving the commutator of the non-local part of the pseudopotentials requires additional computation.

Since the SIMPLE code supports only norm-conserved pseudopotentials, Vanderbilt pseudopotentials of SG15 (Schlipf and Gygi 2015) databases were used for all computations of the dielectric function in this work. The pseudopotentials included relativistic effects at the scalar level while the dielectric function was calculated including non-local contribution from the pseudopotentials to the velocity matrix elements. Once established the dielectric function, it is possible to compute the reflectivity, introducing the refractive index *n* and extinction *k*, as follows:

$$[n(\omega)+ ik(\omega)]^2 =\varepsilon(\omega). \tag{A27}$$

The reflectivity formula is then given by:

$$R(\omega) = \frac{[n(\omega)-1]^2 + k(\omega)^2}{[n(\omega)+1]^2 + k(\omega)^2}, \tag{A28}$$

while the absorption coefficient is given by:

$$\alpha(\omega) = \frac{4.\pi.k(\omega)}{\lambda}. \qquad \text{(A29)}$$

A.7 Overview of machine learning algorithms

Here is a brief overview of some of the most popular supervised learning algorithms (Ethem Alpaydd n. 2009; Pedregosa et al. 2011):

- ***Linear Regression:*** a simple algorithm that models a linear relationship between inputs and a continuous numerical output variable. Easily interpretable results by its output coefficient, faster to train than other machine learning models. However, assumes linearity between inputs and output usually underfitting for small or high-dimensional datasets. It is also very sensitive to outliers.
- **Logistic Regression:** models a linear relationship between inputs and categorical outputs using a sigmoid curve. It is easy to interpret but the assumed linearity frequently leads to overfitting with small or high-dimensional data.
- **K-Nearest Neighbors (KNN):** a non-parametric algorithm that stores all training data with labels. It predicts by finding the k nearest neighbors in feature space based on distance metrics, assigning a class via majority vote. Simple and distribution-agnostic, but time-consuming for large datasets, sensitive to noise and outliers, and requires proper feature scaling for accuracy.
- **Support Vector Machines (SVM)**: a parametric algorithm that finds the optimal hyperplane that maximizes the margin between two classes in the feature space. Effective in high dimensional spaces and robust to outliers, SVM can also handle non-linear data with kernel functions. However, it is prone to overfitting with noisy data and can be quite expensive and sensitive to parameter choices.
- **Decision Trees:** this non-parametric algorithm utilizes a tree-like structure to partition data based on informative features, supporting numerical and categorical data without the need for feature normalization. Known for its interpretability, easy visualization, and swift training, it also adeptly manages non-linearities. However, large trees may lead to overfitting, and its sensitivity to noise and missing values can bias trees, particularly if certain classes dominate.
- **Decision Trees Ensemble:** involves combining multiple decision trees and aggregating their predictions through methods like majority voting or averaging.

Techniques like *random forests* and *gradient boosting* fall under this category, addressing overfitting and variance issues associated with individual trees to enhance accuracy. *Random forests* randomly select subsets of features and samples to build multiple decision trees, they aggregate predictions through voting or averaging. *Gradient Boosting Machines* construct decision trees sequentially, correcting errors of preceding trees. They excel in accuracy by refining predictions iteratively but are more prone to overfitting and demand longer training times than random forests. While effective for large, high-dimensional datasets, drawbacks of ensemble decision trees include reduced interpretability of individual trees and increased computational resources needed for training.

- **Neural Networks (NNs):** a parametric algorithm that consists of multiple layers of interconnected nodes that can learn complex non-linear functions from data through forward and backward propagation. Present higher accuracy for large datasets and can be easily adapted to other problems (transfer learning). Neural networks are, however, very difficult to interpret and in the case of dense neural networks they essentially become a black box. These networks demand substantial computational resources during training and present numerous hyperparameters to tune. Additionally, their versatility results in a whole class of possible models based on the type of layers employed.

When the objective involves capturing and processing the inherent structure and patterns within the data, unsupervised ML algorithms can be very useful. Their primary objective involves discerning hidden relationships, clustering similar data points, and condensing the information into a more manageable and meaningful representation.

There are two main types of unsupervised learning algorithms: clustering and dimensionality reduction. Clustering algorithms partition the data into groups (clusters) based on some measure of similarity or distance. Dimensionality reduction algorithms reduce the number of features or dimensions of the data while preserving its essential information. Here is a brief overview of some of the most popular unsupervised machine learning algorithms (Ethem Alpaydd n. 2009; Pedregosa et al. 2011):

- **K-Means Clustering:** a simple algorithm that assigns each data point to one of K clusters based on the distance to the cluster centroid. Easy to implement and interpret, this method is scalable to large datasets. Requires previous knowledge or screening to determine the best value for K. Sensitive to outliers and initial conditions, assumes spherical clusters around the centroid.

- **Hierarchical Clustering:** forms cluster hierarchies by merging or splitting clusters based on their sizes. It is versatile with no fixed cluster count requirement. However, can be computationally intensive, sensitive to outliers, and optimal clustering level selection can be challenging to determine.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** DBSCAN identifies clusters by density fluctuations, useful for spatial data analysis and outlier detection. It is adaptive with varied cluster shapes, robust against outliers, but requires parameter tuning and may struggle with varying density clusters.
- **Principal Component Analysis (PCA):** PCA reduces data dimensionality, retaining maximum variance. Common for data visualization, and feature extraction, it reduces noise and enhances efficiency. Yet, it assumes linear feature relationships, reduces interpretability compared to the original features, and might not retain local structure.
- **Kernel PCA (KPCA):** Kernel PCA is an advanced version of PCA designed for handling complex, nonlinear data relationships. It uses a "kernel trick" to map data into a higher-dimensional space, making it easier to identify intricate patterns and connections that linear techniques might miss. However, interpreting results might be more challenging compared to the original features, and like PCA, KPCA might not fully retain the local structure of the data.
- **Autoencoders:** Autoencoders are a type of artificial neural network primarily used for unsupervised learning tasks, focused on data reconstruction and representation learning. They consist of an encoder-decoder architecture designed to compress and then reconstruct input data. In contrast to PCA, can learn complex, non-linear transformations, capturing intricate data relationships. However, most of the interpretability of the original features is lost in the encoding. More details on autoencoder architecture are provided on section A.12.

A.8 Cross-validation

K-fold cross-validation is a technique for making near-optimal use of available data by repeating model training and validation on different subsets of data, often using a large training set and a small validation set in each iteration. The train set is divided into K subsets, and the model is trained using $K-1$ of the subsets and validated using the set that was not utilized for training (Refaeilzadeh, Tang, and Liu 2009). This process is repeated K times, with the average of all validations used to calculate overall performance as:

$$CV(\hat{f}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{\mathrm{i}=1}^{n_k} L(\hat{y}_k^{(i)}, y^{(i)}) \qquad \text{(A30)}$$

where L is some loss function appropriate to the problem, and $\hat{y}_k^{(i)}$ is the predicted label of the i-th training example of the model $\hat{f}$ trained using the subset of the training data excluding subset k, which is of size n_k. Cross-validation is repeated with different hyperparameter combinations until the best ones are found, at which point the model is trained with the chosen hyperparameters on the whole cross-validation data and applied to the separate test set, as illustrated in *Figure A1*. This approach, known as *cross-validation and testing* is the most popular data partitioning method in ML (Refaeilzadeh, Tang, and Liu 2009). Regarding the number of splits, employing a five to tenfold cross-validation usually achieves a favorable balance between preventing overfitting and maintaining an adequately sized training dataset (Korjus, Hebart, and Vicente 2016).

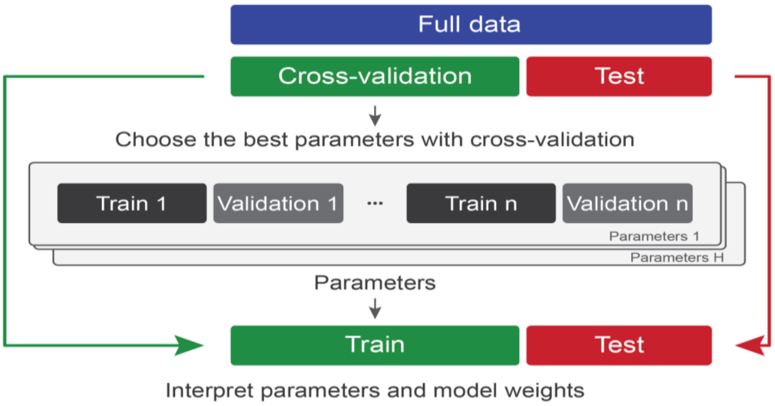


Figure A1 – Cross-validation and testing involves splitting data into a test set and cross-validation set where models are trained and validated to find optimal hyperparameters, the prediction accuracy is evaluated on the holdout test set. Source: (Korjus, Hebart, and Vicente 2016)

Alternatively, one can also apply *nested cross-validation (NCV)*, a more costly alternative in which K-fold split is first performed to divide in “training + validation” and test sets, subsequently, each of these folds undergo K-fold cross-validation, as shown in *Figure A2*. This method guarantees that each data point appears once in the test set and provides more robust metrics. For this reason the nested cross-validation is used in the famous benchmarking initiative for ML models on materials science, the MatBench (Dunn et al. 2020). Although NCV maximizes data efficiency, it also comes with limitations. It introduces favorable bias when evaluating unseen data because each data point appears frequently in the training data during evaluation (K-1 times). This hampers the interpretability of hyperparameters and model weights. However, NCV excels in tasks such as understanding statistical dependence and serves as a valuable benchmarking tool by removing test set bias.

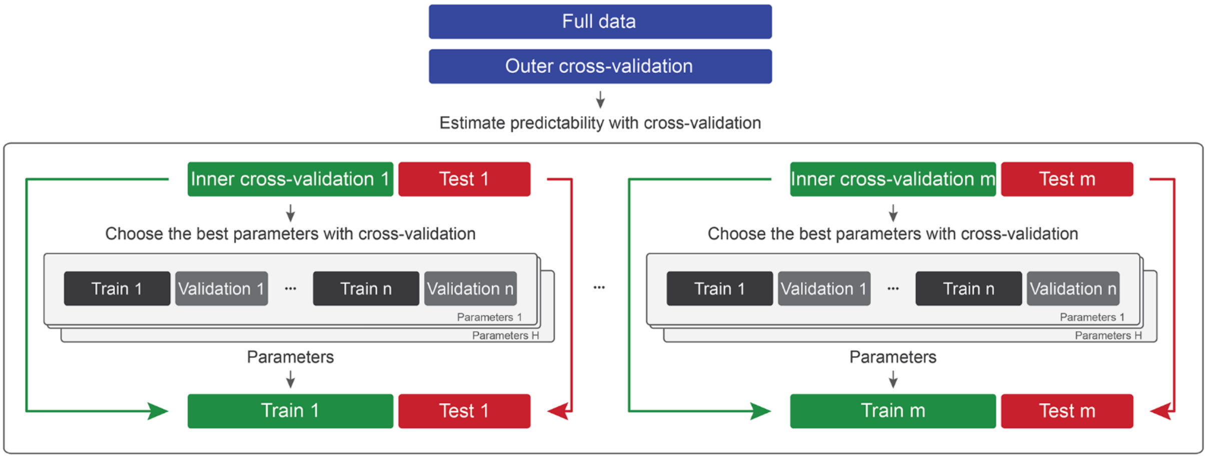


Figure A2 – In nested cross-validation training and validation sets in inner folds determine optimal hyperparameters and performance is evaluated across multiple folds for robustness. Source: (Korjus, Hebart, and Vicente 2016)

In this work, NCV was applied only when comparing to metrics obtained in MatBench as done in the study conducted in *Chapter* 5. Otherwise, the traditional cross-validation and testing approach was adopted since it provides more meaningful hyperparameters and weights to interpret, as well as performance metrics more likely to hold in unseen data. To compensate for its lower data efficiency, we used a 5-fold cross-validation or 10-fold in the case the dataset was small, we also used a small test dataset of 2% to 5%, separating roughly 10.000 samples that retained the statistical distribution of the features.

Data partitioning is trivially done using the *Scikit-learn* package (Pedregosa et al. 2011) through the functions `train_test_split` and `KFold` from the `model_selection` module as described on the official documentation ('3.1. Cross-Validation' 2023). Additionally, for classification tasks, Scikit-learn offers a variant called `StratifiedKFold`. This particular method ensures that the percentage of samples in each class is maintained across validation and test sets, effectively preventing class imbalance issues.

A.9 Ensemble methods

Ensemble methods are a powerful machine learning tool where multiple models are combined to improve predictive performance. These methods work by aggregating predictions from several base models to produce a more accurate and robust final prediction. They can significantly enhance the overall performance by leveraging the strengths of individual models and minimizing their flaws (Pedregosa et al. 2011; Dietterich 2000). Ensemble methods encompass various techniques, including but not limited to:

4. *Aggregation:* this method constructs multiple models using different subsets of the training data, then combines their predictions by averaging or voting to reduce overfitting and variance. Aggregating models based on bootstrap resampling is termed *bagging* (Breiman 1996).
5. *Boosting:* it involves sequentially training models where each subsequent model focuses on the examples that previous models found difficult, thereby improving overall predictive accuracy.
6. *Stacking:* here, predictions from diverse models are used as inputs to a meta-model that learns how to best combine these predictions to generate the final output.

Ensemble methods are particularly advantageous in situations where individual models might struggle due to the complexity or noise in the data. By leveraging diverse models, they can often yield superior performance compared to using a single model. Nevertheless, they do incur increased computational costs because combining and training multiple models can require a lot of resources.

In our study, ensemble aggregation was systematically used on our ensemble MODNet models based on the Deep Ensemble framework (Lakshminarayanan, Pritzel, and Blundell 2017). This method enhances the robustness of predictions and allows for the construction of confidence intervals and quantification of uncertainty in individual predictions (De Breuck, Evans, and Rignanese 2021). Additionally, we aggregated models from each k-fold in cross-validation through averaging, forming an ensemble to efficiently evaluate performance and enhance uncertainty predictions for active learning (see section A.15), saving considerable time compared to retraining on the entire cross-validation set.

A.10 Data preprocessing: normalization and one-hot encoding

Normalization is an essential step in data preprocessing for continuous variables in ML algorithms. It ensures that the features are on the same scale, preventing larger-magnitude features from overshadowing others. Normalization can also reduce outliers and improve overall quality and consistency of the data, which in turn improves the ability of predictive models to detect patterns and make accurate predictions.

Scale consistency is essential for faster convergence in ML algorithms, particularly those reliant on gradient descent such as neural networks, linear regression, and gradient boosting machines (defined on *Appendix A.4*). A comparative study of normalization methods (Cabello-Solorzano et al. 2023) for a range of machine learning algorithms demonstrated that, irrespective of the normalization method used, very few algorithms are essentially unaffected by normalization, highlighting its significance.

The most common normalization techniques are: range scaling, feature clipping, log scaling, and z-score scaling (‘Normalization | Google for Developers’ 2023), presented in detail below:

1. *Range scaling*: This technique converts the feature values from their natural range into a standard range, usually 0 and 1 (or sometimes -1 to 1). The formula for scaling to a range is:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{A31}$$

Scaling to a range is a good choice when the feature values are approximately uniformly distributed across the range, and there are no extreme outliers that cannot be safely removed by clipping. For example, bulk modulus or band gaps from semiconductors.

2. *Log scaling*: This technique computes the logarithm of the feature values to compress a wide range to a narrow range. The formula for scaling follows:

$$x_{log} = log(x + 1) \tag{A32}$$

Log scaling is useful when the feature values follow a power law distribution, meaning that a few values have many points, while most values have few points. For example, thermal and electrical conductivity distributions (Sierepeklis and Cole 2022). The log scaling will change the data distribution making it more suitable for linear models.

3. *Feature clipping:* This technique caps the feature values above (or below) a certain threshold to a fixed value. The threshold is defined based on domain knowledge and the application. A traditional clipping strategy is to clip by z-score to ±Nσ (limit to ±3σ, for example) where σ is the standard deviation. In materials science, it is typical to employ clipping techniques, aiming to remove outliers or restrict data within a defined range. This helps eliminate abnormal data points that could arise from theoretical or experimental limitations, human error, or when the model's objective is to predict within a predetermined range.

4. *Z-score scaling* or *Standardization*: This technique represents the feature values as the number of standard deviations away from the mean. The formula for z-score is:

$$x_z = \frac{x - \mu}{\sigma} \tag{A33}$$

where μ is the feature's mean value and σ its standard deviation. Z-score is a good option when the feature values have a normal (or Gaussian) distribution, meaning that most values are close to the mean, and the distribution is symmetric. Z-score ensures that the feature values have a mean of 0 and a standard deviation of 1, which can make the model more robust and less sensitive to outliers.

The distribution and properties of the features have an impact on the choice of the normalization method. Therefore, it is essential to experiment with different approaches and evaluate how they affect model quality and performance. For all ML model training conducted normalization of the features was performed through the package *Scikit-learn* (Pedregosa et al. 2011), more precisely through the module `sklearn.preprocessing`. This module offers the functions `MinMaxScaler` for a range scaling of [0,1] by default, `StandardScaler` for z-score normalization, and log scaling (or any other function transformation) can be easily applied through `FunctionTransformer`. Clipping outliers can be achieved through straightforward data manipulation using *NumPy,* the fundamental python package for numerical computing.

For categorical features, like the crystal system of a material (one of the seven Bravais lattices) or the orbital character of HOMO (s, p, d, or f), preprocessing is essential to align them with the matrix structure required to optimize the $\hat{f}$ function approximated by the ML algorithm. There are primarily two approaches for handling these features.

One method involves using an ordinal encoder, which transforms each categorical feature into an integer. However, this numerical representation may imply an order among categories to the algorithm (similar to positions in a race), often misrepresenting the inherent meanings of these categories. Therefore, the preferred alternative, especially when categories carry individual meanings, is to utilize a one-hot encoder. This encoder assigns a '1' to indicate a match with a category and '0' otherwise.

For instance, consider the crystal system example. The feature vector would encompass seven entries, each corresponding to one of the Bravais lattices. If a material belongs to the tetragonal crystal system, the feature vector would have a '1' in the tetragonal entry, while the rest would hold '0's. This method preserves the inherent meanings of categories without implying any order among them.

Similar to the normalization functions, the `sklearn.preprocessing` module includes a function called `OneHotEncoder` that automatically creates a new feature vector and assigns suitable values to each data point by learning from the entire training dataset. Examples of this fitting process are demonstrated in the official documentation for continuous and categorical features (‘6.3. Preprocessing Data’ 2023).

A.11 Data preprocessing: imputation and dimensionality reduction

Imputation is an essential step in machine learning preprocessing that helps maintain dataset's integrity by addressing missing data. Techniques for handling this issue involve estimating and substituting missing values to enhance model accuracy and rectify biases caused by non-random missing data. Traditional imputation methods, such as mean/median/mode imputations, replace missing values with the corresponding statistics, yet they fail to preserve dataset variance and will bias results when data is not missing at random. More sophisticated methods train machine learning models to predict missing values based on other features. These methods vary from simple regression models to random forests and complex neural networks. However, they perform better when some randomness in the missing values exists (Jäger, Allhorn, and Bießmann 2021).

In datasets used for predicting properties based on structures and composition from theoretical materials databases, missing data in the feature dataset often arises from limitations in the descriptors used and is therefore not random. Tools like MatMiner, during batch featurization, might apply featurizers to structures and compositions not originally intended for them, resulting in missing data for certain descriptors. This occurs, for instance, with Miedema descriptors, which lack elemental data on halogens and many semimetals, rendering them unable to compute descriptors for materials containing these elements. Another common scenario involves electronic and structural descriptors that rely on accurate bonds to generate materials fingerprints. At times, parameters used to compute nearest-neighbors, such as cutoff radius, fail for specific structures, leading to missing descriptors.

When training a broad model on a sizable dataset, these problematic featurizers often perform well on most data. However, in cases where they fail, descriptors can be imputed by assigning a value beyond the typical feature range as a placeholder (e.g. using -1 in the normalized features) for those materials. Similarly, for categorical features, a new placeholder class is created for the missing values. This preserves information from cases that work as expected, while containing the bias introduced by missing features. This was the approach to handle missing values used in this work, however, it is crucial to note, remains a compromise. The ideal treatment would involve redesigning the featurizer to handle each case properly, which is not always feasible.

The final step in preprocessing involves dimensionality reduction, ideally suited after prior steps have taken place. The concept here is to convert features into a lower-dimensional space by crafting new features that retain crucial information. This differs from feature selection, which targets the most relevant subset of features based on their importance to the target variable (Sorzano, Vargas, and Montano 2014). However, dimensionality reduction might not be necessary (or advisable) when employing more complex ML algorithms like neural networks.

Dimensionality reduction serves various purposes in machine learning. While it can mimic feature selection by extracting pertinent information, its primary aim lies in enhancing ML model performance. By reducing the number of features while preserving essential information, these techniques address computational complexity, overfitting, and the curse of dimensionality. This usually leads to refining the speed and accuracy of machine learning models. However, it is also possible to degrade results, especially when the features with low variance are the most informative. Moreover, dimensionality reduction lowers the interpretability of the results by obscuring the contributions of the original features behind the reduced components.

Several methods for dimensionality reduction exist, they are usually categorized as unsupervised ML algorithms (see appendix A.4), not requiring labeled data. The most famous method for reducing dimensionality is Principal Component Analysis (PCA) (Sorzano, Vargas, and Montano 2014). When provided with a set of observations x in an M-dimensional space ($\mathbb{R}^M$), PCA serves to identify the most optimal subspace of a specific dimension m, based on the least-square error criteria. This algorithm is based on the search of orthogonal directions explaining as much variance of the data as possible. In terms of dimensionality reduction, it can be formulated as the problem of finding m orthonormal directions $\boldsymbol{w}_i$ minimizing the error:

$$\mathrm{J}_{PCA} = E\left\{\left\|\mathbf{x} - \sum_{\mathrm{i}=1}^{\mathrm{m}} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle \boldsymbol{w}_i \right\|^2\right\}, \tag{A34}$$

where $\langle \boldsymbol{w}_{\mathrm{i}}, \boldsymbol{x} \rangle$ denotes the projection of x on the i-th reduced vector $\boldsymbol{w}_{\mathrm{i}}$. The approximation to the original vectors is given simply by $\hat{\boldsymbol{x}} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \langle \boldsymbol{w}_{\mathrm{i}}, \boldsymbol{x} \rangle \boldsymbol{w}_{\mathrm{i}}$. As depicted in equation (36), the PCA method assumes linearity in projections, thus not performing optimally when the data's inherent structure includes intricate non-linear dependencies. In such cases, non-linear dimensionality reduction techniques such as Kernel PCA (KPCA) or autoencoders would be better suited. A more advanced alternative for dimensionality reduction to capture non-linear patterns is to use autoencoders which rely on neural networks to find a latent space representation of the data and are discussed specifically on the subsequent *Appendix A.12*.

A.12 Autoencoders

An autoencoder is a specific class of artificial neural network employed in unsupervised learning. Its primary objective is to encode input data into a lower-dimensional representation and subsequently decode it back (Hinton and Salakhutdinov 2006). This architectural framework comprises two main components: an encoder, responsible for mapping input data to a compact latent space, and a decoder, which reconstructs the original data, as illustrated in *Figure A3*. During training, the network seeks to minimize the mean squared error in the reconstruction using backpropagation.

The distinctive advantage of autoencoders over other dimensionality reduction techniques lies in their heightened capacity to capture intricate, non-linear patterns present in the input data. Autoencoders have found noteworthy applications in materials science, particularly for feature learning and materials representation (S. Stein et al. 2019; W. Jin, Barzilay, and Jaakkola 2018; Damewood et al. 2023). Their utilization facilitates the extraction of meaningful features from raw data, enabling more effective representation and interpretation of material properties. More recently, variational autoencoders (VAEs) were introduced which during encoding impose a constraint to obtain a regularized latent space. This regularization allows to obtain valid input by sampling the latent space, making them valuable for solving inverse design problems in materials science (S. Lu et al. 2022; Ren et al. 2022).

In this work, traditional autoencoders were used to learn encoded representations of the general electronic descriptor Orbital Field Matrix and also for general descriptors from MatMiner featurizers. The encoded representation, due to lower number of descriptors, is much easier to be generated by a machine learning model in a high-throughput screening.

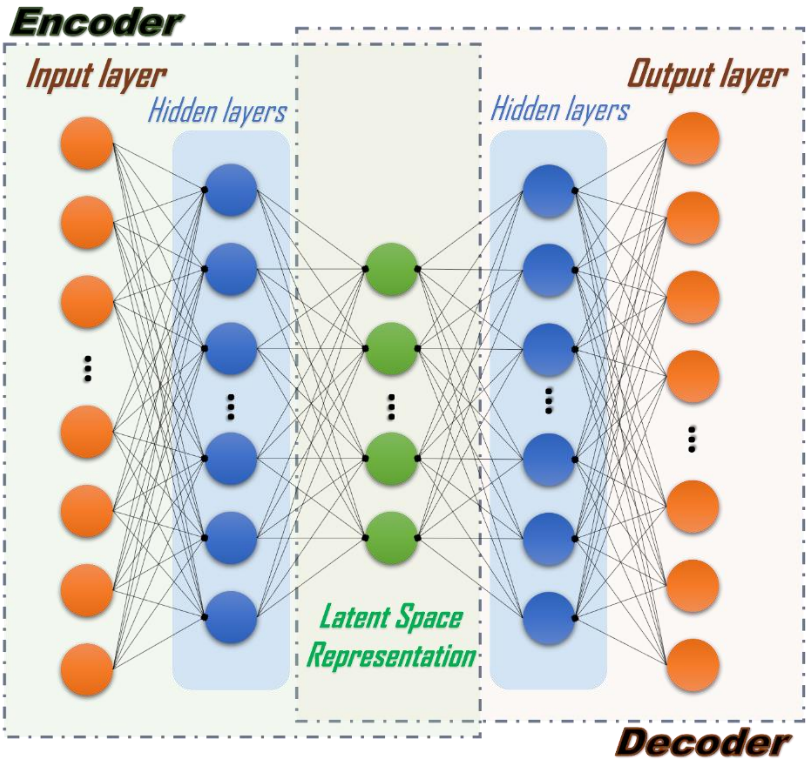


Figure A3 – Illustration depicting an autoencoder, a bottleneck architecture that converts a high-dimensional input into a latent low-dimensional representation (encoder) and then reconstructs the input using this latent representation (decoder). Source: the author.

A.13 Hyperparameters in neural networks

Neural networks present several hyperparameters for tuning, here is presented an overview of the more traditional ones (Goodfellow, Bengio, and Courville 2016; Choudhary et al. 2022):

● **Learning rate (LR):** Learning rate determines the step size taken during optimization. A higher LR can speed up convergence but may lead to overshooting. A lower LR might converge slowly but could help in reaching a more optimal solution. Some common techniques to improve over a fixed LR include:

> ***- Learning rate scheduling***: This method involves altering the LR during training according to a predefined schedule or pattern. The aim is to improve convergence or performance by adjusting the LR dynamically. It may be through a gradual increase of the LR at the beginning of training (warm-up) or a reduction when a number of epochs has passed (step decay).
>
> ***- Learning rate callback:*** For this case the LR is adapted based on the validation loss through a callback function, when the validation loss stops improving for a given number of epochs the learning rate is reduced. This is implemented through the `ReduceLROnPlateau` callback in Keras, for example.
>
> ***- Adaptive learning rates:*** These are improvements over the plain stochastic gradient descent and include the algorithms that may be used to update the weights in an adaptive manner.

● **Stochastic Gradient Descent (SGD):** During backpropagation network weights are iteratively updated via SGD algorithms to minimize the loss function until the desired accuracy is achieved, some of the most usual algorithms for SGD include:

> ***- RMSprop:*** Maintains per-parameter learning rates that are adapted based on the average of recent magnitudes of the gradients.
>
> ***- Adagrad:*** Adjusts the learning rate for each parameter based on the historical gradients for that parameter.
>
> ***- Adam Optimizer:*** The most popular choice currently, this optimizer adjusts the learning rate adaptively for each parameter in the model based on the history of gradients calculated for that parameter effectively combining benefits of the previous methods.

● **Batch size:** Choosing batch size involves trade-offs, larger batches offer computational efficiency and stable updates but may hinder generalization, while smaller batches promote generalization with more noise in updates, working similarly to a regularization technique, but might be computationally inefficient. Using powers of 2 for the batch size (8, 16, 32, etc.) is common practice since it aligns with hardware optimizations to use GPUs. Batch size selection requires testing to ensure it suits the model, dataset, and available computational resources.

● **Batch normalization:** Batch normalization is a method crucial for enhancing the efficiency and stability of neural network training. It works by standardizing the inputs for each layer, effectively managing the flow of gradients throughout the network. This mitigates issues like vanishing or exploding gradients, which can impede learning progress. By ensuring gradient stability, batch normalization accelerates the learning process, enabling the use of higher learning rates resulting in quicker and more effective network training.

● **Regularization techniques:** These strategies keep neural networks from overfitting, which is typical given their enormous capacity to approximate the training data. Most common techniques include:

> ***- L1 & L2 Regularization:*** These techniques involve adding penalty terms to the loss function based on either the absolute (L1) or squared (L2) values of the model's weights. They coerce the model to favor smaller weights, effectively promoting simpler solutions and reducing sensitivity to noise.
>
> ***- Dropout:*** This technique randomly deactivates some neurons during training, forcing the network to learn more robust and generalized features. By preventing co-adaptation among neurons, dropout enhances the network's resilience to overfitting.
>
> ***- Early stopping:*** This method halts the training process once the model's performance on a validation dataset starts deteriorating, thereby preventing the network from overly fitting the training data and improving its ability to generalize to new examples. In Keras, the implementation of early stopping involves using the `EarlyStopping` callback method. This method requires specifying a `patience` value, which determines the number of epochs the

model can go without seeing an improvement in the validation loss before stopping the training process.

A.14 Alternative DNNs in materials science

Deep neural networks (DNNs) have revolutionized the field of machine learning by enabling the development of more accurate models for complex tasks. They have been used to achieve state-of-the-art performance in many applications, including image recognition, speech recognition, and natural language processing (Simonyan and Zisserman 2015; Graves et al. 2006; Wolf et al. 2020). DNNs have also been used to develop models for drug discovery, medical image analysis, and climate science (Lavecchia 2019; J.-G. Lee et al. 2017; Ardabili et al. 2020). Besides the simple feedforward neural networks (FNN) other types of artificial networks exist differing from the simple form by application of other types of transformation in each layer and more elaborate relationships between the layers, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks (Alzubaidi et al. 2021; Sherstinsky 2020). There are also neural networks suited to create generative models which excel at generating data samples that resemble those in the training set, making them particularly useful for inverse design problems. The most common examples include variational encoders (VAE) and generative adversarial networks (GAN). Recently, attention mechanisms have also been used to improve performance on many of these examples (Vaswani et al. 2017). While these models are utilized in materials science, their adoption is still not as extensive as FNNs and GNNs. Typically, they are employed in conjunction with these more conventional models (Choudhary et al. 2022).

A.15 Active learning

Active learning is a machine learning paradigm in which a model is trained on a dataset that is dynamically expanded by iteratively selecting the most informative examples for labeling. Unlike traditional supervised learning, where a fixed and fully labeled dataset is used for training, active learning actively chooses which instances from an unlabeled dataset should be labeled and added to the training set. The goal is to maximize the model's performance with a minimal number of labeled examples. This approach can be seen as a specific application of adaptive experimental design, where statistical inference is facilitated through machine learning models (Lookman et al. 2019). *Figure A4* illustrates this concept, particularly in the context of materials science.

In active learning, the acquisition function is the component guiding the selection of instances from the unlabeled pool for labeling and inclusion in the training set. Common acquisition functions, such as uncertainty of the surrogate model or maximum entropy sampling, aim to select difficult-to-predict or diverse data points from the unlabeled dataset (Margatina et al. 2021). These metrics prioritize exploration, enhancing the model's generalization. In optimization tasks like materials discovery, achieving a balance is crucial. Thus, the acquisition function must navigate between exploring uncertain regions (exploration) and leveraging existing model knowledge to optimize the objective function (exploitation). This balance ensures effective active learning in scenarios where data labeling is resource-intensive, and optimizing the model's performance is paramount.

The active learning cycle usually looks like:

1. *Initial Model Training:* train a model on the initial labeled dataset.
2. *Uncertainty Estimation:* Utilize the current model to predict labels for unlabeled instances, incorporating uncertainty estimates.
3. *Instance Selection:* Provide relevant variables (such as predictions, uncertainty, entropy) to the acquisition function. Identify instances that maximize the acquisition function, as these are likely to provide the most benefit to the model.
4. *Labeling:* Manually label the selected instances or obtain labels through some external means.

5. *Model Update:* Add the newly labeled instances to the training set and retrain the model.
6. *Repeat:* Iterate through the process by going back to step 2 until a satisfactory model performance is achieved or a certain budget for labeling is exhausted.

This cyclic process ensures an iterative and dynamic approach to learning, where the model progressively improves its performance with minimal labeled data.

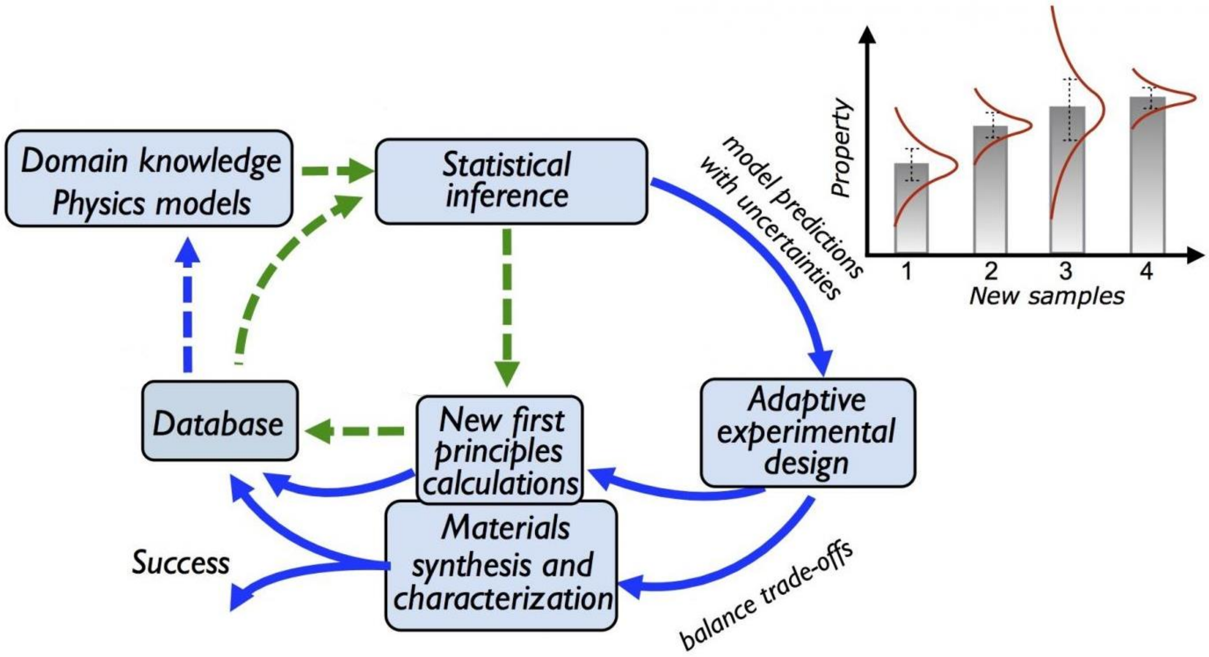

Domain knowledge
Physics models
Statistical inference
model predictions with uncertainties
Property
1
2
3
4
New samples
Database
New first principles calculations
Adaptive experimental design
Materials synthesis and characterization
Success
balance trade-offs

Figure A4 - The adaptive design paradigm to iteratively learn a surrogate model and use acquisition function to balance exploitation and exploration of the search space of unexplored materials to select the next best experiment or calculation. Adapted from: (Lookman et al. 2019)

APPENDIX B: Supporting Information for “Lead-Free $Cs_3Sb_2X_9$ (X = Cl, Br, I) Perovskites: Halide Alloying, Surfaces, Interfaces, and Clusters”

B.1 Functional testing

Table B1 presents lattice parameters and band gap for PBE (J P Perdew, Burke, and Ernzerhof 1996), PBEsol (John P. Perdew et al. 2008), LDA (PZ) (J P Perdew and Zunger 1981) and rVV10 (Sabatini, Gorni, and De Gironcoli 2013) functionals. Smallest deviations in geometry were obtained for PBEsol and rVV10 functionals, van der Walls corrected functional is especially better on the *c* parameter of the material and are most suitable to analyze properties highly dependent on geometry. LDA functional failed in reproduce lattice parameters and band gap as GGA parametrizations. Despite success of PBEsol and rVV10 for geometry, band gap is closer to experiment in traditional PBE functional. Additionally, PBE is shown to be more suitable than PBEsol for Hubbard correction using the Dudarev approximation of U_{eff}=U-J (Dudarev and Botton 1998) due to larger correlation between U and J parameter (Tavadze et al. 2021). Therefore, since deviations in lattice parameters remain small, PBE was chosen to proceed the electronic calculations and as base functional to apply Hubbard correction in this work.

Table B1 - Band gap and lattice parameters with different functionals compared to experimental values for $Cs_3Sb_2X_9$ (X = Cl, Br, I). Deviations from experimental lattice constants for the different functionals are given in percentages under $\Delta_{a,b}$ and Δ_c.

Species	Band gap (eV)	Lattice constants (Å)		$\Delta_{a,b}$ (%)	Δ_c(%)
		a,b	c		
$Cs_3Sb_2Cl_9$					
Expt.[1]	3.09	7.633	9.345		
PBE	2.45	7.836	9.476	+2.662	+1.401
PBEsol	2.21	7.594	9.232	-0.505	-1.208
PZ	2.02	7.276	9.052	-4.669	-3.131
rVV10	2.27	7.484	9.305	-1.952	-0.428
$Cs_3Sb_2Br_9$					
Expt.[2]	2.30	7.930	9.716		
PBE	2.00	8.144	9.932	+2.698	+2.227
PBEsol	2.00	7.856	9.634	-0.935	-0.844
PZ	1.87	7.597	9.442	-4.200	-2.823
rVV10	1.75	7.823	9.703	-1.348	-0.129

Table B1 – (continued)

Species	Band gap (eV)	Lattice constants (Å)		$\Delta_{\mathrm{a,b}}$ (%)	Δ_c(%)
		a,b	c		
$Cs_3Sb_2I_9$					
Expt.[3]	2.06	8.420	10.386		
PBE	1.58	8.660	10.647	+2.853	+2.517
PBEsol	1.32	8.363	10.284	+2.850	+2.512
PZ	1.16	8.161	10.096	-3.079	-2.786
rVV10	1.28	8.410	10.393	-0.120	+0.068

1 - Jian Zhang et al. 2017 ; 2 - Jian Zhang et al. 2017; 3 - Yamada et al. 1997

B.2 Determination of Hubbard U parameters

A point of note when applying Hubbard corrections is that such corrections are mainly intended to localized states such as *d* orbitals which are unsatisfactorily modelled by traditional DFT and can be adequately justified in this case. The value of Hubbard potential can even be calculated from first-principles based on several approximations for highly localized orbitals (Aryasetiawan et al. 2006). When Hubbard values are determined empirically and applied on *p* or *s* orbitals only to correct the band gap underestimation from DFT it may lead to controversial physical results and therefore careful study is required. Literature presents both exceedingly good physical descriptions using empirical U on non-localized orbitals (Calzolari and Nardelli 2013; X. Y. Deng et al. 2014; Sharma, Mishra, and Kumar 2019; Flores et al. 2018) as well as failed approaches (Shao 2008; Janotti and Van de Walle 2011) and the quality of the results has to be assessed in a case-by-case basis. Moreover, several approaches considering U on non-localized orbitals have produced excellent results recently, this is the case of Bayesian optimization for machine learning Hubbard U values (M. Yu et al. 2020) and also the more sophisticated pseudo-hybrid functionals such as ACBN0 (Agapito, Curtarolo, and Nardelli 2015; May and Kolpak 2020).

In this study to obtain an electronic structure more coherent to hybrid functional calculations, Hubbard +U term was supplemented in the PBE functional acting on the 5p orbital of Sb and p orbitals of the halogen which were the most significant on valence band and conduction band of the perovskites. The Hubbard values applied on these orbitals are denoted respectively, as U_{Sb} and U_X henceforth. The values of U were determined empirically varying the U_{Sb} in steps of 2 eV and varying the U_X in steps of 0.5 eV keeping the structure fixed in the experimental atomic positions and cell parameters, band gap and total forces on the cell were measured for each combination. The combinations showing lower forces and band gap closer to experiment were selected to a second phase in which the structures were fully relaxed and the U_X was tuned in steps of 0.1 eV to obtain the experimental gap. Finally, the final structures had the projected density of states (PDOS) calculated and compared to the PDOS obtained with pure PBE functional and the hybrid HSE functional. The Hubbard values which could provide better description of the band gap, projected density of states, geometry, and Bader charge in the atoms, taking experimental and HSE results as reference, were chosen to proceed the simulations.

Hubbard values of Sb of 4 eV and 6 eV presented considerably larger forces in first screening in $Cs_3Sb_2Br_9$ and $Cs_3Sb_2I_9$ perovskites leading to lattice parameters much larger than experiment, therefore only structures of U_{Sb} of 0 and 2 eV were studied further for consistency. The final parameters for each structure, fully relaxed and with optimal halogen Hubbard value, are shown in *Table B2* along with the band gap value. HSE06 calculations were then performed using the lattice parameters relaxed with PBE of each perovskite to compare with PBE+U electronic structure.

Table B2 - Hubbard parameters for halogen with U_{Sb} = 0 eV and U_{Sb} = 2 eV for $Cs_3Sb_2X_9$ relaxed structure yielding experimental band gap values, lattice parameters for structure also shown.

	U_{Sb}	$U_{X=Cl, Br or I}$	Band gap	Lattice constants (Å)	
	(eV)	(eV)	(eV)	a,b	c
$Cs_3Sb_2Cl_9$					
Expt.[1]			3.09	7.633	9.345
HSE	0	0	3.20	-	-
PBE	0	0	2.45	7.876	9.476
PBE+U	0	4.5	3.08	7.881	9.532
PBE+U	2	5.5	3.12	7.868	9.513
$Cs_3Sb_2Br_9$					
Expt.[2]			2.30	7.930	9.716
HSE	0	0	2.34	-	-
PBE	0	0	2.00	8.144	9.932
PBE+U	0	2.5	2.37	8.168	9.898
PBE+U	2	2.8	2.30	8.139	9.875
$Cs_3Sb_2I_9$					
Expt.[3]			2.06	8.420	10.386
HSE	0	0	2.10	-	-
PBE	0	0	1.58	8.660	10.647
PBE+U	0	3	2.04	8.641	10.641
PBE+U	2	4	2.07	8.617	10.559

We observe that when U_{Sb} is increased the localization of Sb 5*p* orbital causes a further reduction of the band gap which, in turn, increases optimal U_X values to keep the experimental band gap. Larger U_{Sb} also reduce the lattice parameters slightly. Band gaps for all Hubbard corrected structures could be made close to the available experimental and theoretical data in the literature with moderate Hubbard values.

Since merely obtaining experimental band gap does not justify applying Hubbard corrections, partial density of states of the structures were compared to the well-established HSE06 functional for every perovskite in *Figure B1*. Hubbard corrections shifted the states to positions much closer to those calculated by HSE06, it is also clear that when U_{Sb} is implemented the results depart from HSE06 suggesting a worse approximation and therefore U_{Sb} = 0 eV was elected. In order to understand how interatomic charge distributions would vary between functionals, Bader analysis based on the atom-in-molecule (AIM) theory was conducted (Tang, Sanville, and Henkelman 2009). Average ionization charge on each element was calculated for PBE, PBE+U (U on halogen only) and HSE functionals and results are presented in *Table B3*. It is clear from the results in *Table B3* and *Figure B1* that the atomic charges with PBE+U functional yield results closer to what is calculated by hybrid functional HSE06 suggesting that the Hubbard correction adopted here is successful in reproducing a higher level of theory.

Table B3. Average ionization charge for each ion on the perovskites $Cs_3Sb_2X_9$ (X = Cl, Br, I) calculated with Bader charge theory. Hubbard correction on PBE+U are, respectively for chlorine, bromine and iodine perovskites, U_{Cl} = 4.5 eV, U_{Br} = 2.5 eV and U_I = 3 eV.

Species		Functional		
		PBE	PBE+U	HSE
$Cs_3Sb_2Cl_9$	Cs charge	-0.8934	-0.9116	-0.9132
	Sb charge	-1.5813	-1.7993	-1.8265
	Cl charge	+0.6491	+0.7037	0.7102
$Cs_3Sb_2Br_9$	Cs charge	-0.8746	-0.8866	-0.8868
	Sb charge	-1.3158	-1.4626	-1.4412
	Br charge	+0.5840	+0.6206	+0.6159
$Cs_3Sb_2I_9$	Cs charge	-0.8628	-0.8745	-0.8773
	Sb charge	-0.9596	-1.1759	-1.1747
	I charge	+0.5009	+0.5529	+0.5534

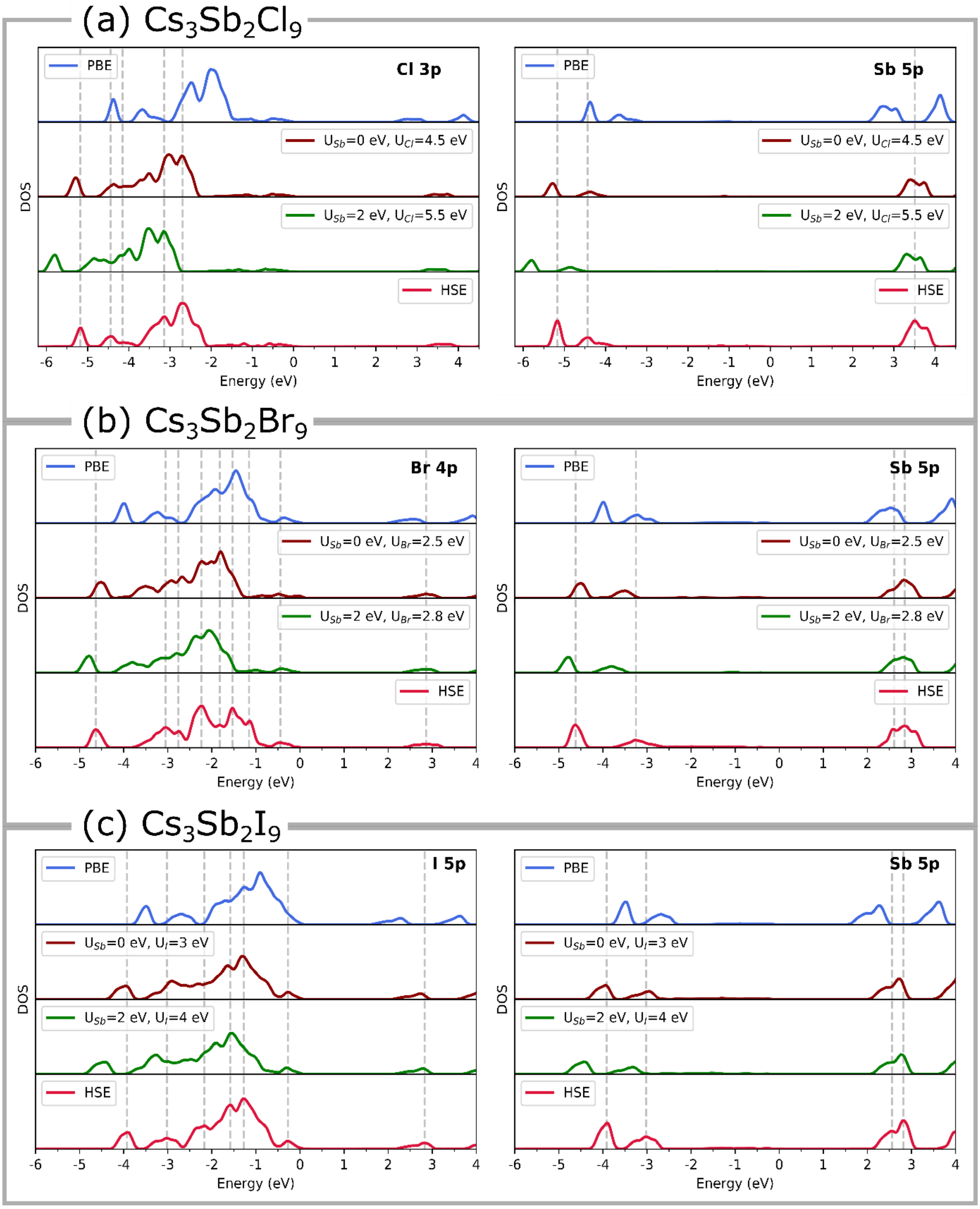


Figure B1 - Comparison of density of states of $Cs_3Sb_2X_9$ (X = Cl, Br, I) with different functionals, vertical lines in maximums of HSE functional DOS are shown to compare with lower level functionals PBE and PBE+U.

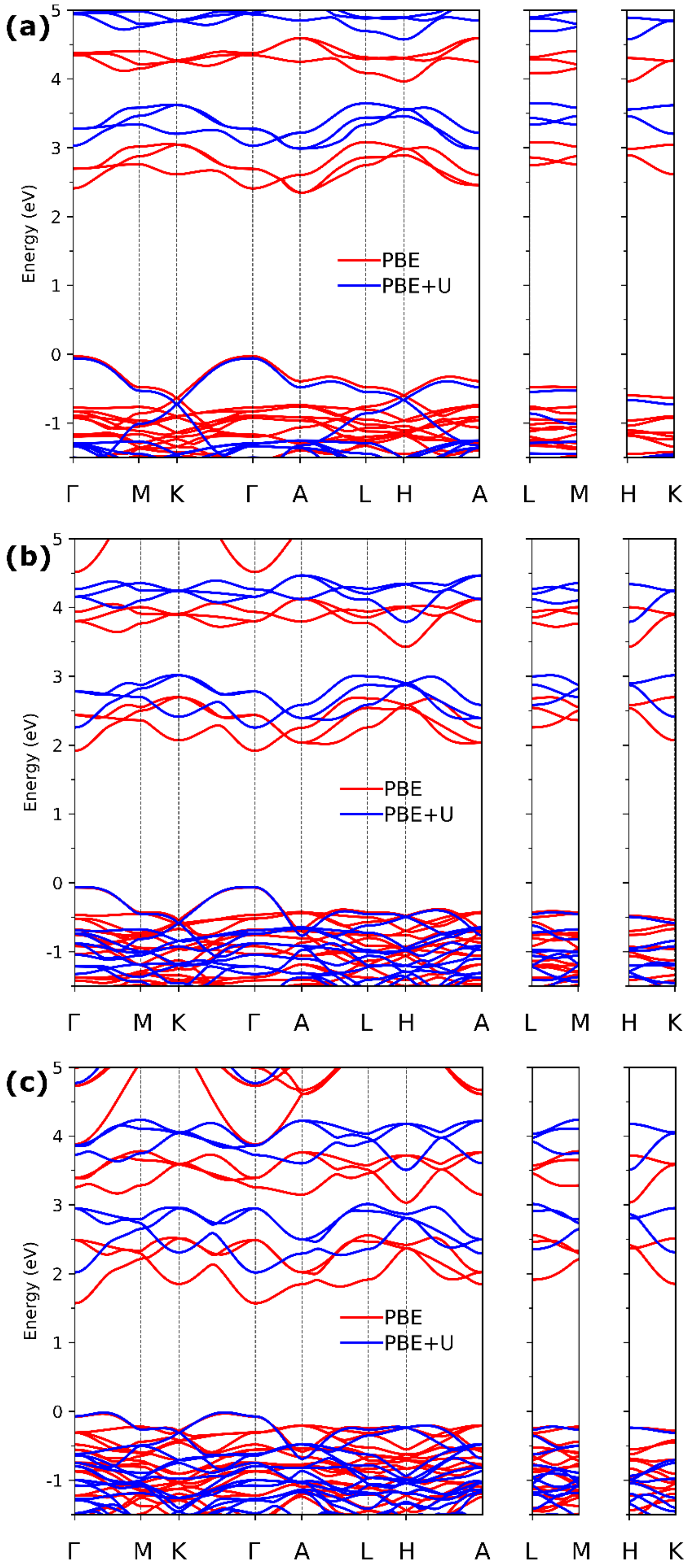


Figure B2 - Band structures of (a) $Cs_3Sb_2Cl_9$, (b) $Cs_3Sb_2Br_9$, (c) $Cs_3Sb_2I_9$ calculated by PBE and PBE+U functional. The red and blue solid lines represent the PBE and the PBE+U bands, respectively. The energy 0 eV represents the Fermi level.

B.3 Halogen alloying

$Cs_3Sb_2X_9$ ($P\overline{3}m1$) features two inequivalent halogen sites: a bridging site (3e) and a terminal site (6i). We calculated the formation energy for Cl substitution in $Cs_3Sb_2I_9$ to assess the cost of substitution at these sites. The results showed a 0.139 eV difference in formation energy, with Cl(i) substitution having lower energy. Thus, terminal sites are preferred for substitution, consistent with Pradhan et al.'s findings (A. Pradhan, Jena, and Samal 2022). However, as the difference is easily accessible in experiments and decreases for halogens of similar size, we adopted a more realistic approach: substituting two terminal sites, followed by one bridging site, until all 9 sites were fully substituted. We also maintained the maximum distance between the minority halogens during substitutions.

Crystal structures present smaller total energy than isolated atoms implying that crystals are more stable and will require energy to break their bonds and decompose. This energy is referred as binding energy (E_b) and can be determined from DFT total energy through the following formula, adapted for the case of $Cs_3Sb_2X_{9-n}Y_n$ solid solutions:

$$E_b = \frac{E[\mathrm{Cs_3Sb_2X_{9-n}Y_n}] - \sum_i n_i E[i]}{\sum_i n_i}, \tag{B1}$$

where $E[Cs_3Sb_2X_{9-n}Y_n]$ is the total energy of $Cs_3Sb_2X_{9-n}Y_n$ solid solution, n_i and *E(i)* are the number of *i* atoms in the cell and the energy of an isolated atom, respectively. E_b is negative for any stable compound and larger (more negative) binding energies values result in a more stable solid solution.

Variation of lattice parameters with compositions was modelled according to Vegard's law as given by the formula:

$$\begin{cases} a_{Cs_3Sb_2X_{9-n}Y_n}(x) = (1-x)a_{Cs_3Sb_2X_9} + xa_{Cs_3Sb_2Y_9} + \theta x(1-x) \\ c_{Cs_3Sb_2X_{9-n}Y_n}(x) = (1-x)c_{Cs_3Sb_2X_9} + xc_{Cs_3Sb_2Y_9} + \theta x(1-x) \end{cases} \tag{B2}$$

Where θ represents the bowing parameter for the lattice constants. Band gap variation with composition was also modelled with second-order Vegard's law to determine their bowing parameter, referred as b_g.

Results of calculations considering the $Cs_3Sb_2X_{9-n}Y_n$ (X,Y = Cl, Br, I) structures with *n* integer varying from 0 to 9 are shown in *Figure B3* presenting band gap, area enclosed by $\vec{a}$ and $\vec{b}$ lattice vectors (A_{ab}) and c lattice parameter for both PBE and

PBE+U case. PBE+U as applied in this work also proves successful in modelling the alloys and this can be validated when comparison is made with HSE and PBE calculations. Lattice parameters applying Hubbard correction presented a slightly increase in comparison to pristine PBE as shown by the trends on A_{ab} and c in the graph. Regarding band gaps, Hubbard corrections applied on the individual halogens in the composite structures were successful in predicting band gaps in accordance with HSE results with less than 2% error (*Table B5*) which increases trust in applying PBE+U method to investigate halogen alloyed structures. Convergence test was performed for n=4 in the three solid solutions considered, convergence showed that the increase of energy cutoff and k-points changes the band gap by less than 0.01 eV and the alloy formation energy by less than 0.1 meV/atom.

Table B4 – Fitting parameters for binding energy, lattice parameters, band gap and formation enthalpy of the $Cs_3Sb_2X_{9-n}Y_n$ solid solutions. Pearson correlation coefficient (c_p) is shown for fitting curve.

Solid solution	Property and fitting parameter				
	E_b : m [1]	a, b : θ [2]	c : θ [3]	E_g : b_g [4]	ΔH_f : Ω [5]
$\mathrm{Cs_3Sb_2Cl_{9-n}}Br_{\mathrm{n}}$	0.6560	0.6525	0.032	-0.0498	18.682
	(c_p=0.999)	(c_p=0.981)	(c_p=0.991)	(c_p=0.9921)	(c_p=0.7188)
$\mathrm{Cs_3Sb_2Br_{9-n}I_n}$	0.3087	0.8419	-0.041	0.3049	43.796
	(c_p=0.998)	(c_p=0.982)	(c_p=0.995)	(c_p=0.9915)	(c_p=0.8877)
$\mathrm{Cs_3Sb_2}Cl_{9-\mathrm{n}}I_{\mathrm{n}}$	0.9650	1.4222	0.327	-0.3950	96.209
	(c_p=0.989)	(c_p=0.989)	(c_p=0.996)	(c_p=0.9844)	(c_p=0.9472)

(1) $E_b = E_b[Cs_3Sb_2X_9](1-x) + E_b[Cs_3Sb_2Y_9]x = mx + E_b[Cs_3Sb_2X_9]$

(2) $a_{Cs_3Sb_2X_{9-n}Y_n}(x) = (1-x)a_{Cs_3Sb_2X_9} + xa_{Cs_3Sb_2Y_9} + \theta x(1-x)$

(3) $c_{Cs_3Sb_2X_{9-n}Y_n}(x) = (1-x)c_{Cs_3Sb_2X_9} + xc_{Cs_3Sb_2Y_9} + \theta x(1-x)$

(4) ${E_g}_{Cs_3Sb_2X_{9-n}Y_n(x)} = (1-x){E_g}_{Cs_3Sb_2X_9} + x{E_g}_{Cs_3Sb_2Y_9} + b_g x(1-x)$

(5) $\Delta H_f = \Omega x(1-x)$

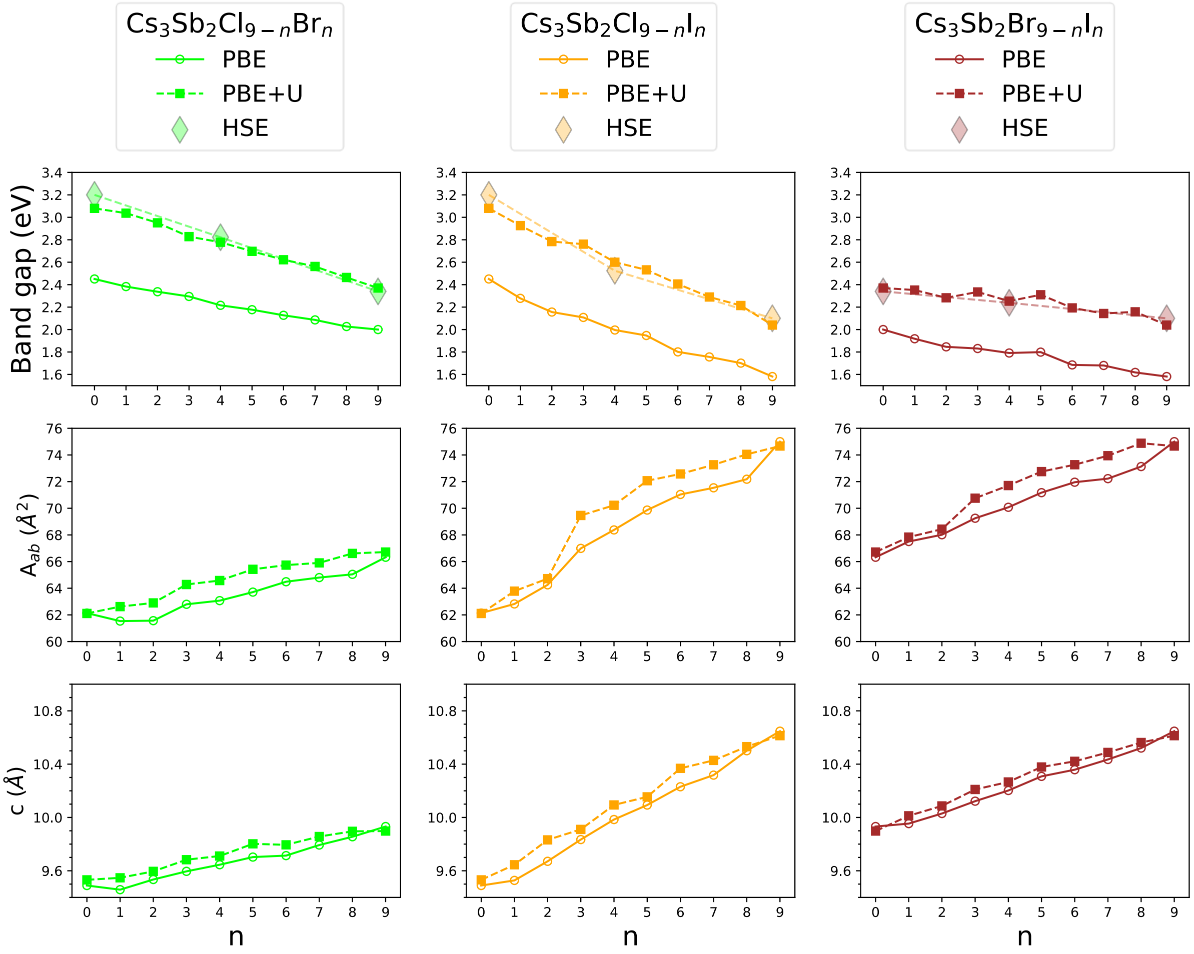


Figure B3 - Band gap, area enclosed by $\vec{\boldsymbol{a}}$ and $\vec{\boldsymbol{b}}$ lattice vectors ($\boldsymbol{A_{ab}}$) and $\boldsymbol{c}$ lattice parameter of $Cs_3Sb_2X_{9\text{-}n}Y_n$ (X,Y = Cl, Br, I) perovskites with n integer varying from 0 to 9. Calculations were performed for both PBE, PBE+U (U_{Cl}=4.5 eV, U_{Br}=2.5 eV, U_I=3 eV). HSE06 calculations are also presented for $Cs_3Sb_2Cl_5Br_4$, $Cs_3Sb_2Cl_5I_4$ and $Cs_3Sb_2Br_5I_4$ to obtain band gap.

Table B5 – HSE and PBE+U results on band gap of $Cs_3Sb_2Cl_5Br_4$, $Cs_3Sb_2Cl_5I_4$ and $Cs_3Sb_2Br_5I_4$.

	HSE	PBE+U	Error %
$Cs_3Sb_2Cl_5Br_4$	2.822	2.778	1.76
$Cs_3Sb_2Cl_5I_4$	2.524	2.600	1.63
$Cs_3Sb_2Br_5I_4$	2.238	2.253	1.26

In *Figure B4* density of states for the intermediate composition perovskite $Cs_3Sb_2Cl_5Br_4$ is shown, we can see the agreement is not only on band gap but DOS composition in general is very similar between the PBE+U and HSE especially on first 2 eV of valence band and all of the conduction band.

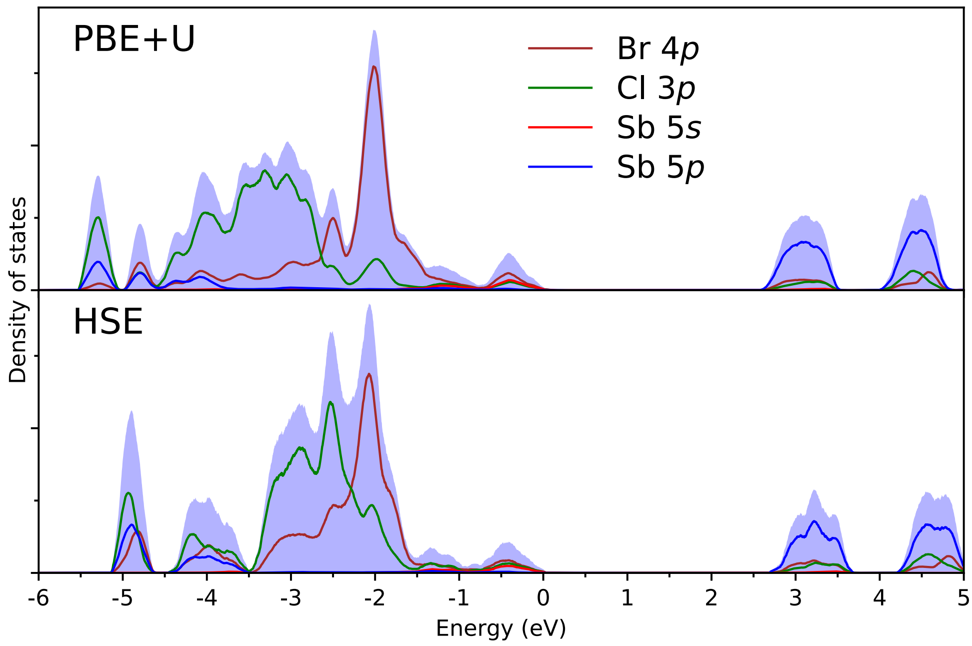


Figure B4 - Density of states of $Cs_3Sb_2Cl_5Br_4$ comparing PBE+U corrected and HSE.

B.4 Surfaces and interface calculations

Surface energy of stoichiometric symmetric slabs can be easily calculated by the traditional equation:

$$\gamma = \frac{(E_{total} - nE_{bulk})}{2A} \qquad \text{(B3)}$$

Where γ is the surface energy of one facet, E_{total} is the total energy of the relaxed surface slab, E_{bulk} is the total energy of bulk perovskite material per formula, n is the number of bulk perovskite formula in the slab, and A is the surface area of the slab model. To expand this definition to non-stoichiometric and asymmetric slabs, one must first consider that surface energy consists of cleavage (E_{cle}) and relaxation (E_{rel}) energies (Tian et al. 2018; J. M. Zhang et al. 2008), and therefore can be written as:

$$\gamma = \frac{(E_{cle} + E_{rel})}{A} \qquad \text{(B4)}$$

When cleaving, two surface terminations are formed and the cleavage energy is divided between them equally. In $Cs_3Sb_2X_9$, Sb-X termination corresponds to a stoichiometric slab and these can be used to find cleavage energy through:

$$E_{cle} = \frac{(E_{unrelax} - nE_{bulk})}{2} \tag{B5}$$

where $\boldsymbol{E_{unrelax}}$ is the total energy of the unrelaxed stoichiometric slab.

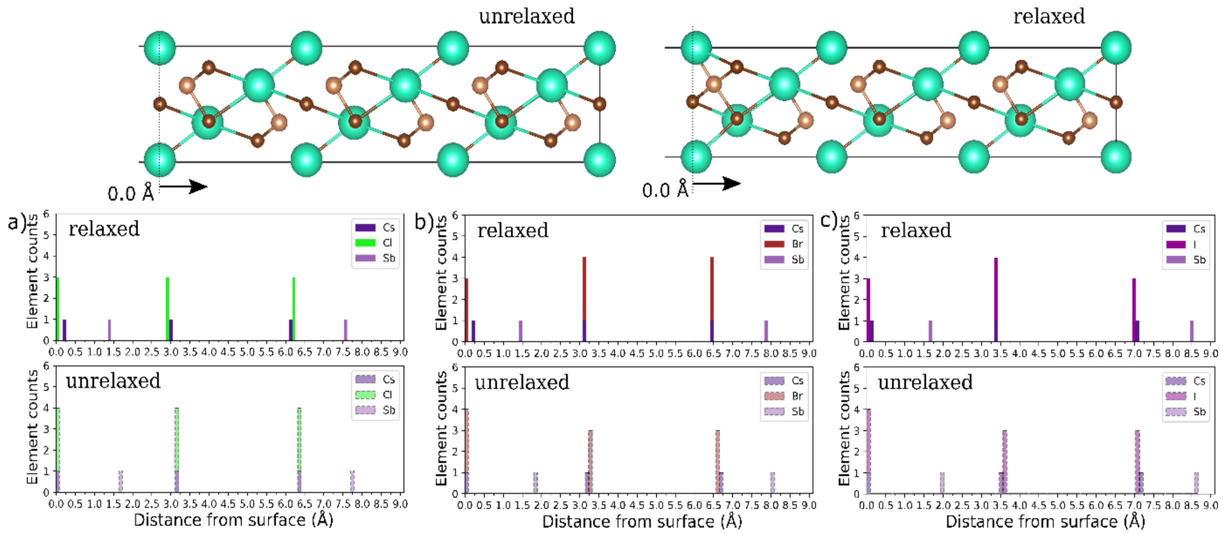


Figure B5 – Geometric analysis of the top layers of (0001) CsX-terminated surface of halide perovskites. Top of figure illustrates unrelaxed and relaxed $Cs_3Sb_2Br_9$ slabs, measurements start on the outermost surface atom as shown. (a), (b) and (c) present element counts in given position for both relaxed and unrelaxed slab for $Cs_3Sb_2Cl_9$, $Cs_3Sb_2Br_9$ and $Cs_3Sb_2I_9$, respectively.

Table B6 - Supercell parameters of interfaces (1000) compared to original surfaces parameters.

	a (Å)	b (Å)
(1000) surface		
$Cs_3Sb_2Cl_9$	7.883	9.489
$Cs_3Sb_2Br_9$	8.144	9.932
$Cs_3Sb_2I_9$	8.660	10.647
(1000) interface		
$Cs_3Sb_2Cl_9$/ $Cs_3Sb_2Br_9$	8.013	9.710
$Cs_3Sb_2Br_9$/ $Cs_3Sb_2I_9$	8.402	10.290

Table B7 – Valence band maximum obtained from bulk $Cs_3Sb_2X_9$ and calculated potential alignment (ΔV) calculated to determine valence band offsets (ΔE_v) for $Cs_3Sb_2Cl_9$/$Cs_3Sb_2Br_9$ and $Cs_3Sb_2Br_9$/$Cs_3Sb_2I_9$ interfaces. All quantities determined from PBE+U calculations.

Pristine bulk	VBM	Interfaces	ΔV	ΔE_v
$Cs_3Sb_2Cl_9$	2.923	$Cs_3Sb_2Cl_9$/ $Cs_3Sb_2Br_9$	0.915	0.645
$Cs_3Sb_2Br_9$	2.654	$Cs_3Sb_2Br_9$/ $Cs_3Sb_2I_9$	1.086	0.530
$Cs_3Sb_2I_9$	2.098			

Table B8 – Valence band maximum obtained from bulk $Cs_3Sb_2X_9$ in HSE theory level and calculated potential alignment ($\boldsymbol{\Delta V}$) calculated from supercells using PBE functional. Valence band offsets ($\boldsymbol{\Delta E_v}$) for $Cs_3Sb_2Cl_9$/$Cs_3Sb_2Br_9$ and $Cs_3Sb_2Br_9$/$Cs_3Sb_2I_9$ interfaces are calculated from these quantities.

Pristine bulk	VBM	Interfaces	ΔV	ΔE_v
$Cs_3Sb_2Cl_9$	3.024	$Cs_3Sb_2Cl_9$/ $Cs_3Sb_2Br_9$	0.873	0.550
$Cs_3Sb_2Br_9$	2.702	$Cs_3Sb_2Br_9$/ $Cs_3Sb_2I_9$	1.119	0.588
$Cs_3Sb_2I_9$	2.170			

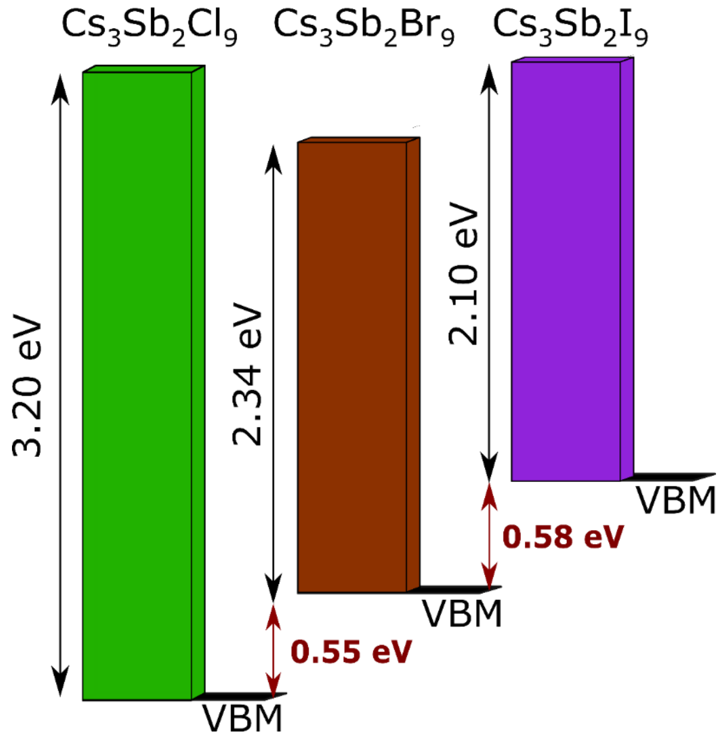


Figure B6 - Band alignment illustration of the three halogen perovskites based on calculations of band offsets using HSE VBM and gaps along with PBE potential alignment offset.

B.5 Clusters

To evaluate the energy of formation of the substitutional Cl defect, the following formula based on total DFT energies was performed (Freysoldt et al. 2014):

$$E_f[D] = E_{tot}[\mathrm{Cs}_{30}\mathrm{Sb}_6 I_{29} Cl] - (\, E_{tot}[\mathrm{Cs}_{30}\mathrm{Sb}_6 I_{30}] - \mu_I + \mu_{Cl}\,) \qquad \text{(B6)}$$

Where $E_{tot}[\mathrm{Cs}_{30}\mathrm{Sb}_6 I_{29} Cl]$ is the total energy of the chlorine doped structure, $E_{tot}[\mathrm{Cs}_{30}\mathrm{Sb}_6 I_{30}]$ is the total energy of the pristine $Cs_{13}Sb_6I_{30}$ cluster and μ_I and μ_{Cl} represents the potential energy reservoir of chlorine and iodine atoms, respectively, calculated from the diatomic gas phase.

The pattern of larger gap difference for bromide cluster is reproduced for PBE only calculations, as shown in *Figure B7*, and therefore the possibility of Hubbard correction inducing this behavior observed with PBE+U is dismissed although the gap difference is slightly larger with Hubbard correction.

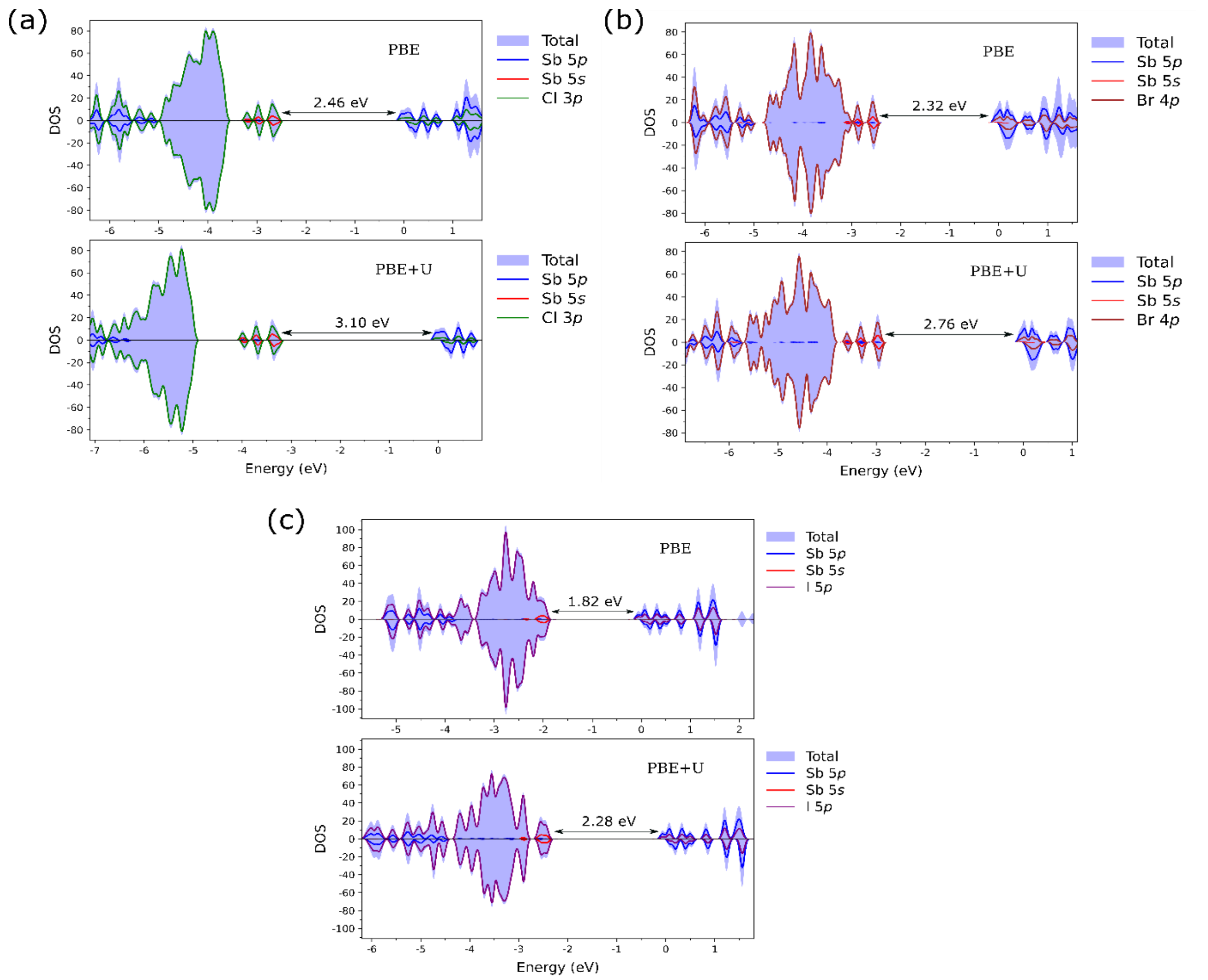


Figure B7 – Partial density of states for different $Cs_3Sb_2X_9$ clusters for X = Cl (a), Br (b) and I (c) comparing results with PBE and PBE+U.

Figure B8 presents the position of every atom in the clusters as a function of distance from the cluster center. The relaxed geometry is compared to the unrelaxed geometry derived directly from the corresponding bulk perovskite positions. For the iodine perovskite cluster, external Cs atoms suffer the strongest contraction compared to other halogen clusters, almost 1 Å. This contraction reaches for inner iodine atoms and is responsible for shortening the Cs-I bond seen in the clusters. For the chlorine perovskite cluster, we see that external Cs atoms suffer less contraction due to the strong bonding between chlorine and Cs. Close to the center, we also observe that the Cs atoms become more involved in the bonding with chlorine, as the Sb-Cl bond increases, reinforcing the Cs-Cl bond.

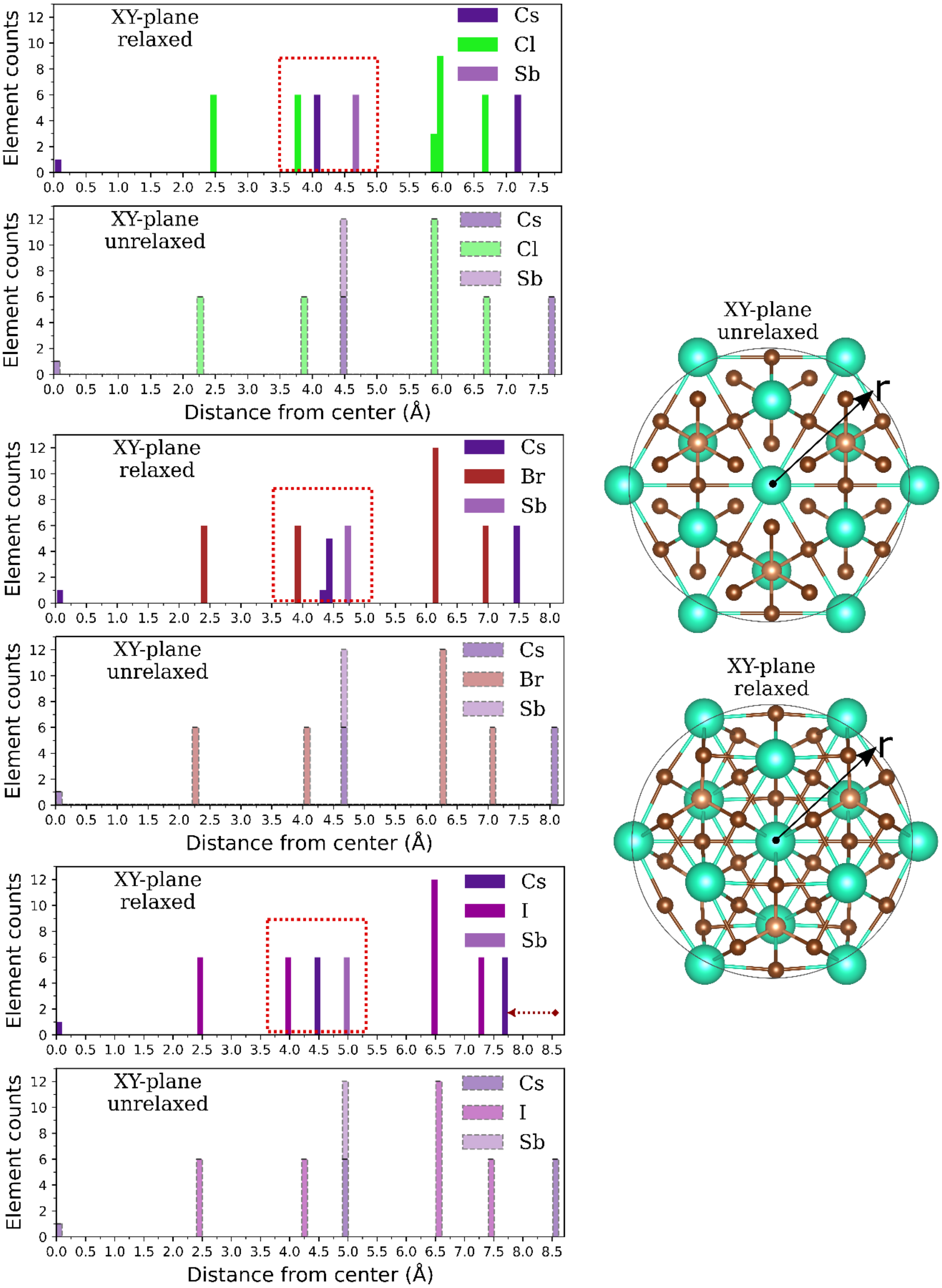


Figure B8 – Geometric analysis of halogen perovskite clusters $Cs_{13}Sb_6X_{30}$. On the right, an illustration of unrelaxed and relaxed clusters is presented, along with the convention used for measuring the atomic positions radially in the XY-plane. On the left, the element counts are provided for both relaxed and unrelaxed clusters at a given position, as measured in the radius from the XY-plane, where 0 represents the center of the cluster.

To assess halide alloying effects in these clusters, we tested substituting iodine with Cl in the $Cs_{13}Sb_6I_{30}$ cluster, at both the longitudinal face and edge sites, as shown in *Figure B9*(a). Face site substitution resulted in a higher Bader charge transfer from the bonding Sb atom, -1.276e, compared to edge site substitution, -1.264e. As Cl maintains an overall charge of +0.70e in both sites, this stems from a more favorable geometric orientation enhancing charge transfer at the face site. Formation energy for Cl substitution was -0.90 eV for the face site, indicating spontaneous substitution. A slightly larger formation energy of -0.85 eV was found for edge site substitution. Examining the substitutional site's effects on density of states and spin polarization in *Figure B9*(b), Cl in edge sites significantly impacts conduction and VB edges and spin polarization.

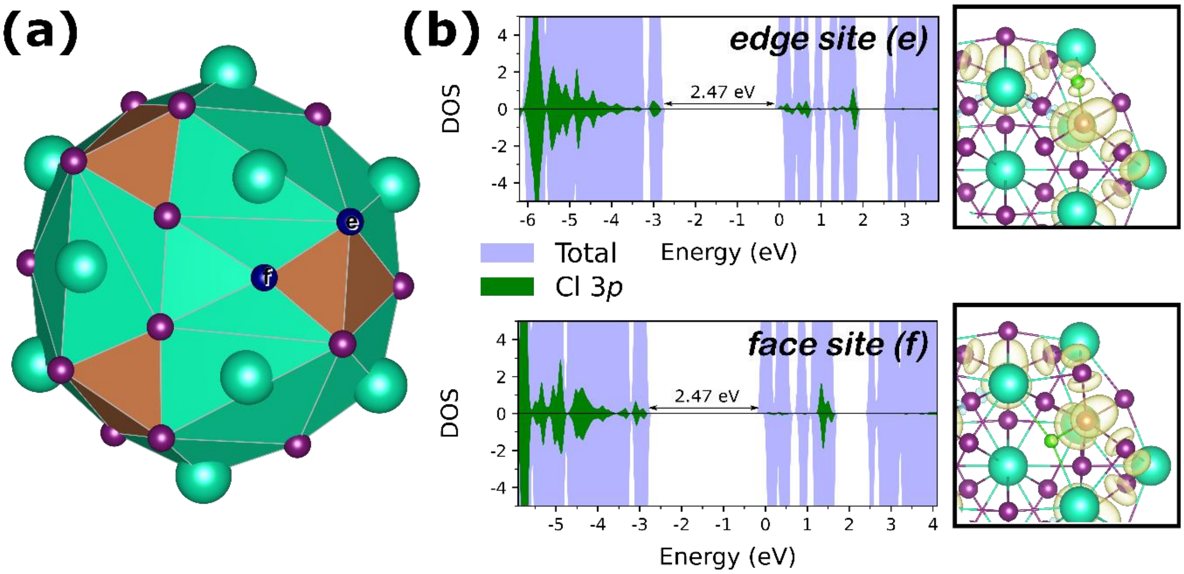


Figure B9 – Face and edge sites are highlighted as f and e, respectively, in $Cs_{13}Sb_6I_{30}$ cluster (a), along with density of states plots and spin polarization for each site (b).

APPENDIX C: Supporting Information for "Doping effects on the optoelectronic properties and the stability of $Cs_3Sb_2I_9$: Density Functional Theory insights on photovoltaics and light-emitting devices"

C.1 Pristine structures

Table C1 - Atomic and ionic radii for Sb and selected elements for Sb substitution in this work, ionic radius difference to Sb in percentage is also presented.

	Sb	Ag	In	Mo	Nb	Sc	Bi
Atomic radius (Å)	1.33	1.65	1.56	1.90	1.98	1.84	1.43
Effective ionic radius (Å)[a]	0.76	0.75	0.80	0.69	0.72	0.74	1.03
Ionic radius difference to Sb (%)	0.00	-1.71	+5.26	-9.21	-5.26	-2.63	+35.5

[a] ionic radius for oxidation state +3 and coordination number (CN) = 6.

Table C2 – Reference atomic coordinates for both polymorphs of $Cs_3Sb_2I_9$ compound used in this work, obtained from (Yamada et al. 1997).

	Atom	Site	X	Y	Z
Dimer	Cs	2b	0	0	0.25
	Cs	4f	0.3333	0.6667	0.0850
	Sb	4f	0.3333	0.6667	0.8453
	I	6h	0.4929	0.9858	0.25
	I	12k	0.1653	0.3306	0.9189
Layered	Cs	1a	0	0	0
	Cs	2d	0.6667	0.3333	0.672
	Sb	2d	0.6667	0.3333	0.196
	I	3e	0.5	0.5	0
	I	6i	0.149	0.851	0.646

Table C3 - Lattice parameters a, b, and c, I–Sb bond length and band gap energy E_g of for the $Cs_3Sb_2I_9$ structure.

Structure	Lattice parameters (Å)		Lattice deviations (Å)		$D_{I\text{-}Sb}$ (Å)§	E_g (eV)
	a, b	c	$\Delta_{a,b}$ (%)	Δ_c (%)		
$Cs_3Sb_2I_9$ ($P6_3/mmc$)						
Calculated	8.543	21.642	2.32	3.47	3.213	1.81
Theoretical (PBE/PW)^{*1}	8.682	21.763	3.99	4.05	-	2.00
Experimental	8.349^{*3}	20.936^{*3}	-	-	3.198^{*3}	2.43^{*4}
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)						
Calculated	8.622	10.623	2.40	2.28	3.18	1.52
Theoretical (PBE)^{*2}	8.664	10.633	2.90	2.38	3.18	1.55
Experimental	8.420^{*3}	10.386^{*3}	-	-	3.164^{*3}	2.00^{*5}

**Sources: 1* (Berri 2020) *, 2* (Y. L. Liu et al. 2019)*, 3* (Yamada et al. 1997),
4 (Correa-Baena et al. 2018) *, 5* (Saparov et al. 2015).

§ Bond lengths I-Sb are measured in the 6h and 3e iodine atoms to match experimental data on the respective polymorphs.

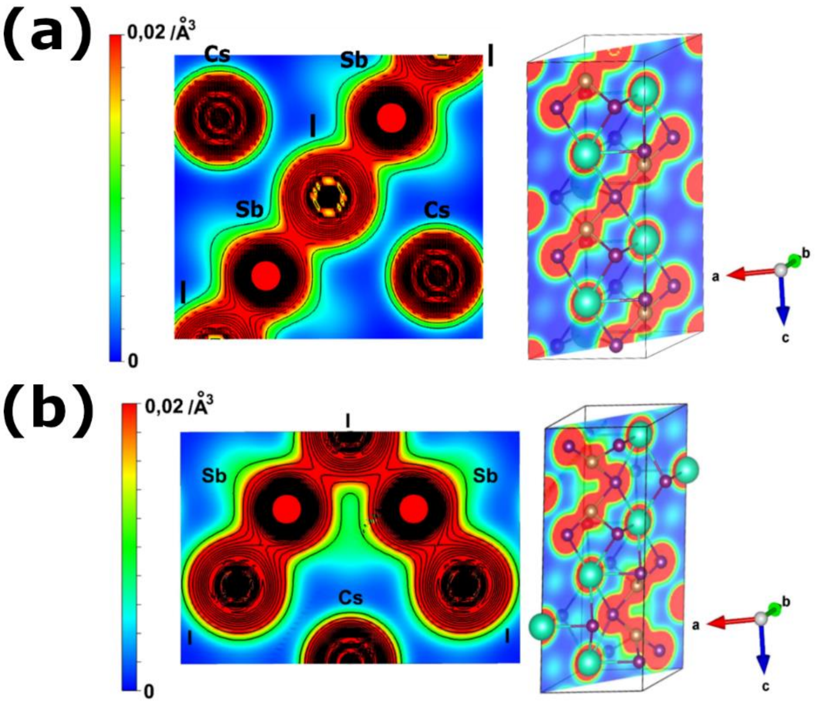


Figure C1 - (a) Charge density plots and contour maps (lines in black) for (a) $Cs_3Sb_2I_9$ ($P6_3/mmc$) and (b) $Cs_3Sb_2I_9$ ($P\bar{3}m1$) in the direction shown in the plane represented in the corresponding unit cells.

The charge density plots for $Cs_3Sb_2I_9$ structures shown in *Figure C1* shows a distinctly higher electronic density along I-Sb bonds in $P\bar{3}m1$ polymorph due to slightly shorter I-Sb bonds. Moreover, the average Bader charges per atom (in units of the electron charge, *e*) for the pristine structure are shown in Table C4. The hexagonal structure presents a slightly larger charge transference suggesting a more ionic character which is associated to a higher binding energy compared to the trigonal structure (5 meV/atom binding energy difference).

Table C4 - Average Bader charges and binding energies for pristine $Cs_3Sb_2I_9$ structures.

Structure	E_b (eV/atom)	Element	Average Charge (*e*)
$Cs_3Sb_2I_9$ ($P6_3/mmc$)	-1.040	Cs	0.860
		Sb	0.947
		I	-0.503
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)	-1.035	Cs	0.858
		Sb	0.944
		I	-0.496

The effective masses of electrons and holes in different directions for two polymorphs of $Cs_3Sb_2I_9$ are displayed in Table C5. The electrons have lower effective mass values in the $k_{[100]}$ direction, which is parallel to the octahedra array, making it the preferred transport direction for both polymorphs. The results are similar to previous findings (McCall et al. 2018) regarding the presence of a nearly flat band in the hexagonal $Cs_3Sb_2I_9$ structure. The electron effective mass is significantly different between the $k_{[100]}$ (0.32 m_e) and $k_{[001]}$ (1.33 m_e) directions, demonstrating the high anisotropy of the dimer polymorph. The layered polymorph exhibits a more isotropic band structure for the conduction band at the Γ point, with low electron effective masses in both directions. The electron and hole effective masses in the $P\bar{3}m1$ structure are lower compared to the $P6_3/mmc$ structure due to the higher dimensionality in the former.

Table C5 - Calculated effective masses for the pristine $Cs_3Sb_2I_9$ structures.

Structure	Effective mass (m*)			
	electrons		holes	
	$k_{[100]}$	$k_{[001]}$	$k_{[100]}$	$k_{[001]}$
$Cs_3Sb_2I_9$ ($P6_3/mmc$)	0.32	1.33	1.10	1.05
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)	0.31	0.40	0.80	0.33

C.2 Doping with transition metals

C.2.1 Electronic states and charge analysis

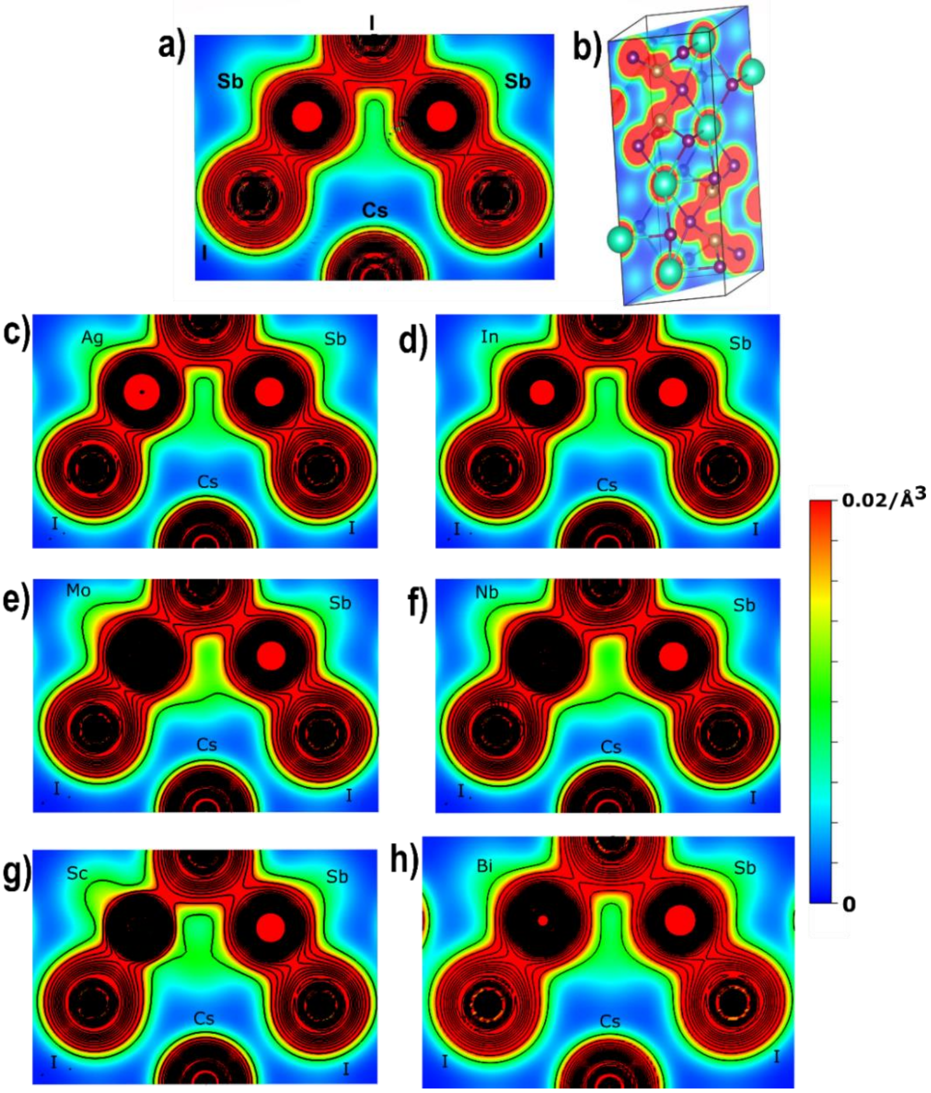


Figure C2 - Charge density and contour map (in black) of (a) pristine $Cs_3Sb_2I_9$ ($P6_3/mmc$) in the direction shown in the plane represented in (b) the unit cell, and doped with (c) Ag, (d) In, (e) Mo, (f) Nb, (g) Sc, and (h) Bi, respectively.

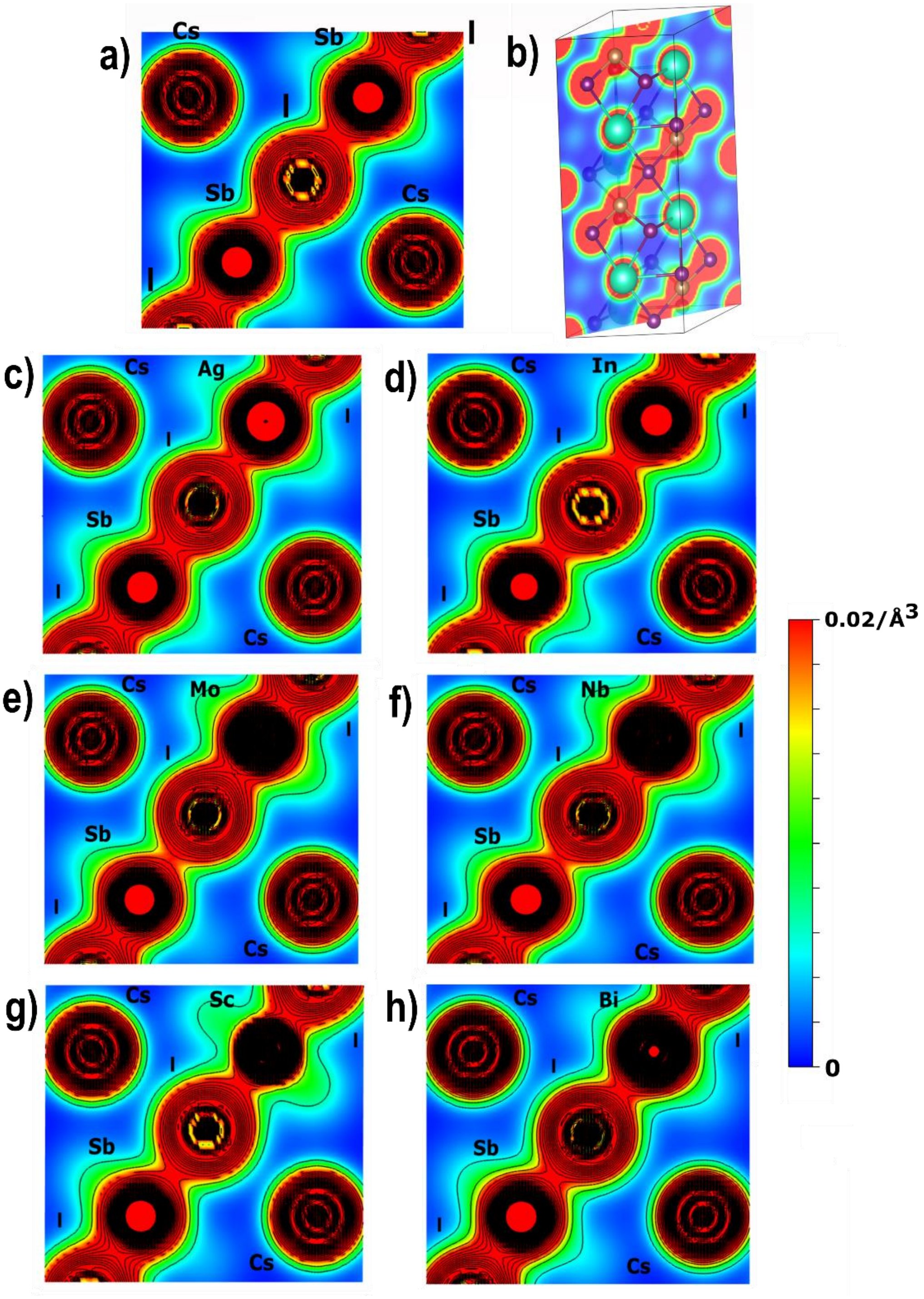


Figure C3 - Charge density and contour map (in black) of (a) pristine $Cs_3Sb_2I_9$ ($P\bar{3}m1$) in the direction shown in the plane represented in (b) the unit cell, and doped with (c) Ag, (d) In, (e) Mo, (f) Nb, (g) Sc, and (h) Bi, respectively.

Charge density plots for pristine and metal-doped $Cs_3Sb_2I_9$ perovskites for both polymorphs (*Figures C2 and C3*). Notably, Mo- and Nb-doped shows significant charge accumulation around the dopants, deviating from the expected +3 ionic radius configuration, indicating poor coordination with iodine. Conversely, despite Sc having a comparable atomic radius to Mo and Nb, Sc-doped perovskite exhibits a spread of charge reaching for the iodine atoms in the octahedra suggesting a better coordination. Moreover, Bi containing perovskites, despite bismuth's larger radius, display a charge distribution similar to Sb due to their isoelectronic valence shell.

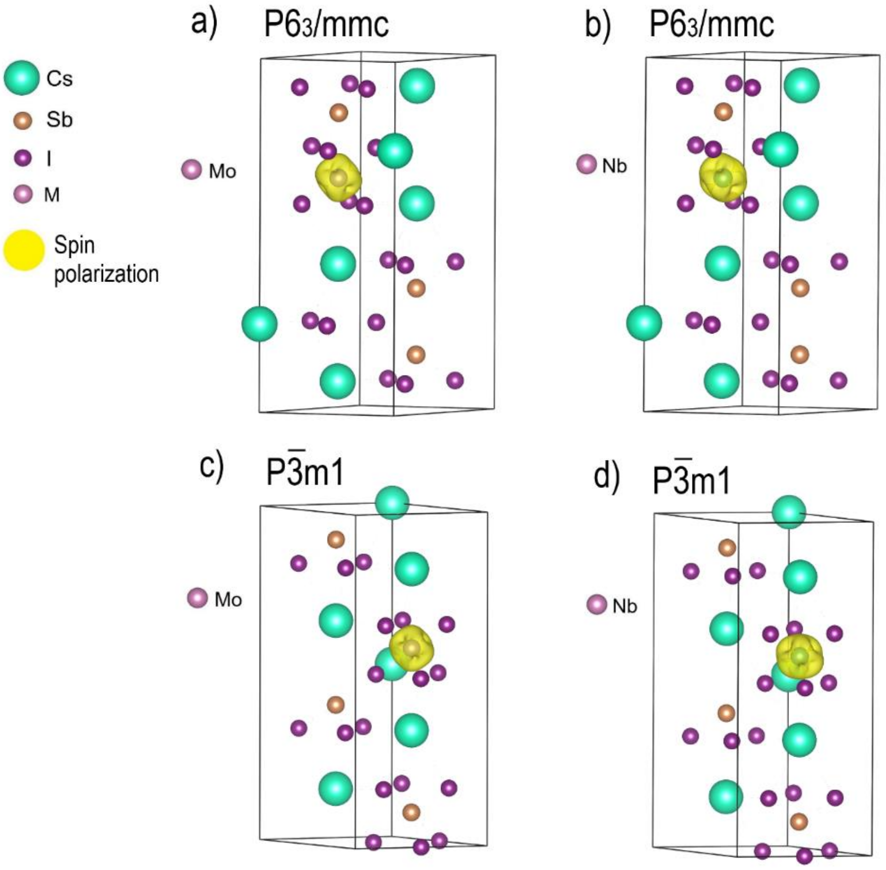


Figure C4 - Spin polarization for $Cs_3Sb_2I_9$ ($P6_3/mmc$) doped with (a) Mo, and (b) Nb, respectively, and for $Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with (c) Mo, and (d) Nb, respectively.

Based on the Löwdin charges presented in *Table C6* indicating the spin polarization per orbital of the dopant atom, we could observe that Mo contributes with a higher total and partial polarization than Nb for both $Cs_3Sb_2I_9$ structures. This can be explained from the Ligand Field Theory viewpoint, considering that the octahedral field formed by the halogen ligands — weak field ligands — is a high spin configuration. Therefore, in comparison with the Nb d^3 configuration, the Mo d^4 electronic structure is expected to present higher total polarization.

Table C6 - Löwdin charges for Mo and Nb as dopants.

Structure	Dopant	Polarization			Total polarization
		s	p	d	
$Cs_3Sb_2I_9$ (P6$_3$/mmc)	Mo	-0.0428	-0.0354	-2.7816	-2.8598
	Nb	-0.0324	-0.0263	-1.9032	-1.9619
$Cs_3Sb_2I_9$ (P$\bar{3}$m1)	Mo	-0.0432	-0.0361	-2.8044	-2.8836
	Nb	-0.0326	-0.0260	-1.8935	-1.9521

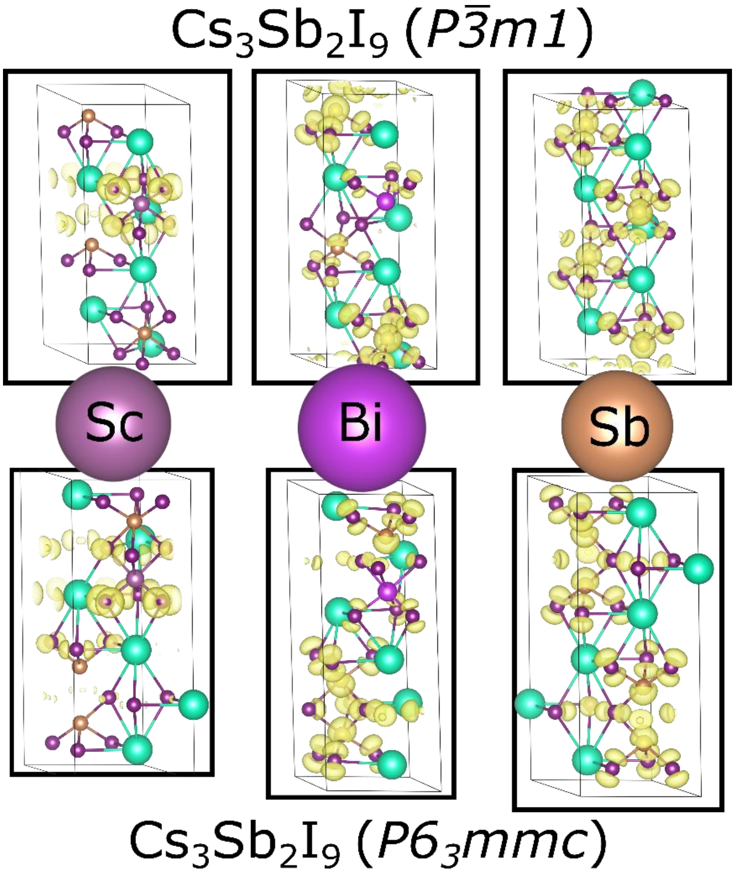


Figure C5 – Charge density of the highest occupied Kohn-Sham state $Cs_3Sb_{1.5}M_{0.5}I_9$ (M=Sc,Bi) and pristine $Cs_3Sb_2I_9$, both P$\bar{3}$m1 (top) and P6$_3$/mmc (bottom) polymorphs.

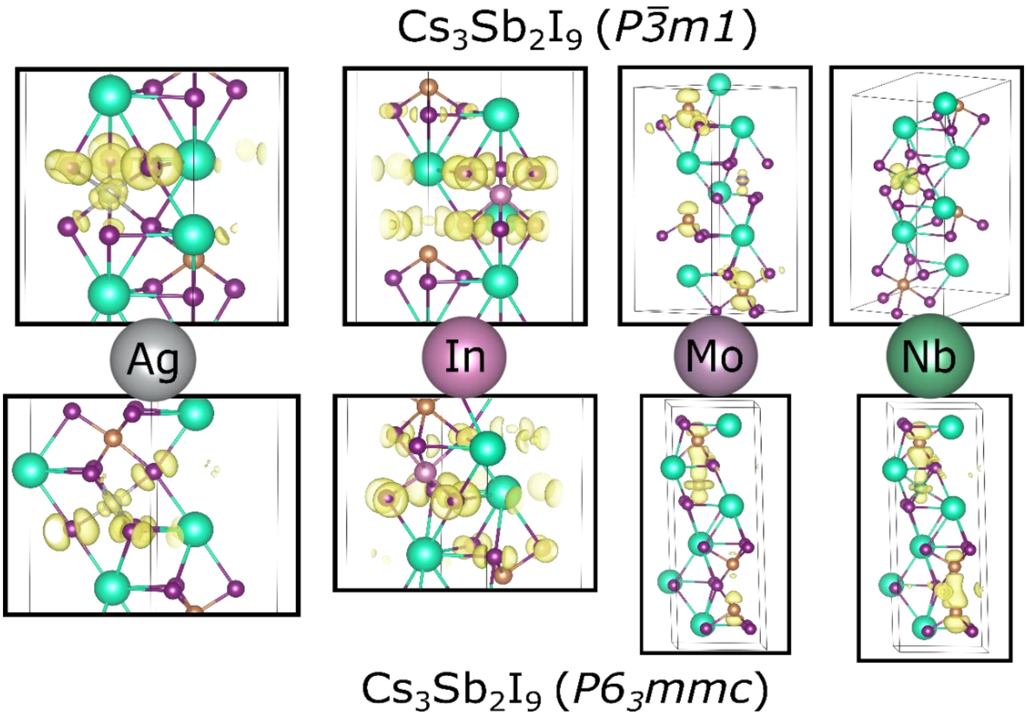


Figure C6 – Charge density of the highest occupied Kohn-Sham state in $Cs_3Sb_{1.5}M_{0.5}I_9$ (M=Ag, In, Mo, Nb), both P$\bar{3}$m1 (top) and P6$_3$/mmc (bottom) polymorphs.

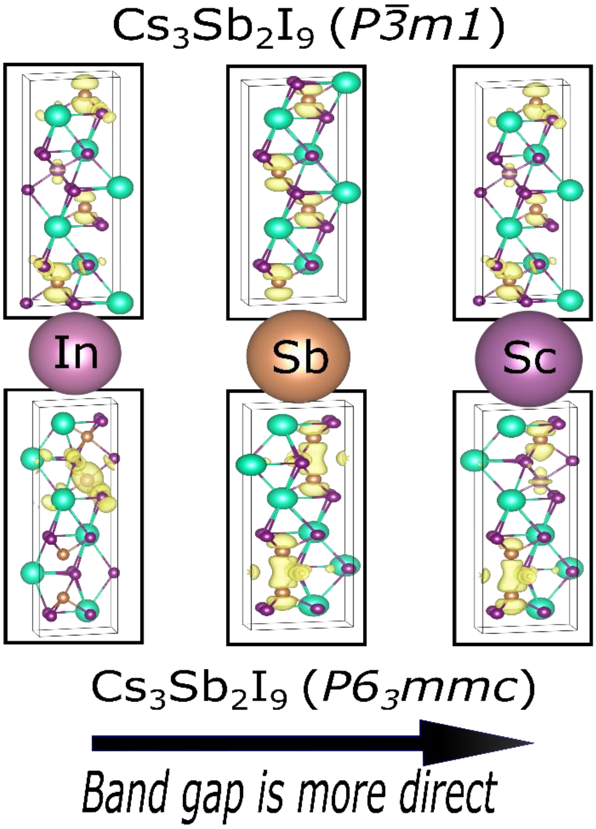


Figure C7 – Charge density of the lowest unoccupied Kohn-Sham state in $Cs_3Sb_{1.5}M_{0.5}I_9$ (M=In, Sb, Sc), both P$\bar{3}$m1 (top) and P6$_3$/mmc (bottom) polymorphs.

C.2.2 Geometric analysis

To verify the geometrical changes that the metal or halogen substitution yielded to the metal-halogen octahedra a careful analysis was carried out involving isolation of the octahedra in the structure and standardization through appropriate projections. The orientation of the octahedra in space is characterized by the rotation angles φ, θ and ψ, measurements of the laterals and axis of the octahedra are given by a_{min}, a_{max}, *and* a_{axis}, respectively. Finally, displacements of the central atom to the mass center of the octahedra are given by $\boldsymbol{\delta_{X/Y/Z}}$ for coordinate axis aligned to the octahedra and $\boldsymbol{\delta^{abs}_{X/Y/Z}}$ for the absolute coordinate axis for the structure.

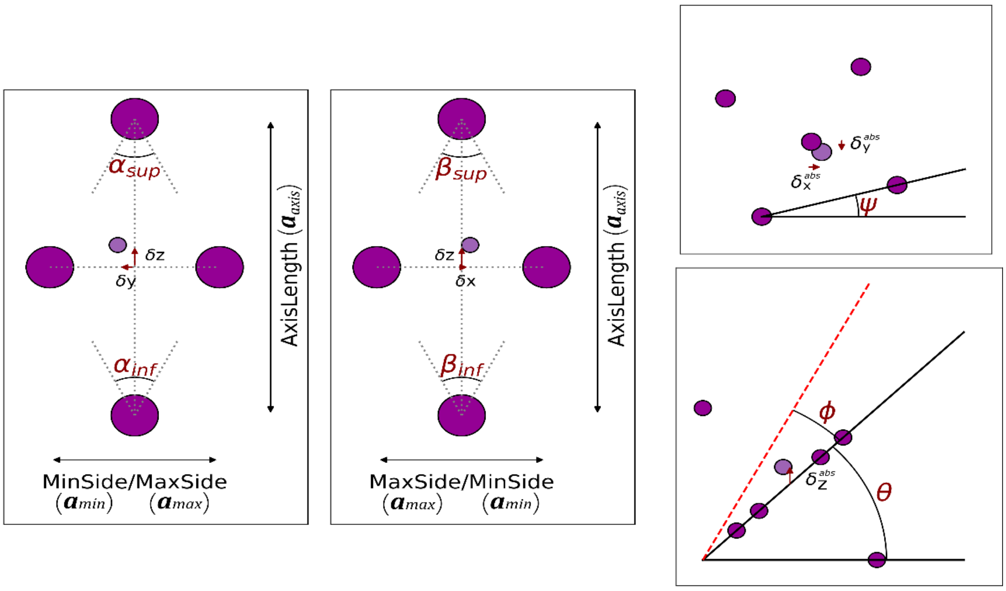


Figure C8 – Geometrical parameters considered for the corresponding MX_6 octahedra in the $Cs_3Sb_2I_9$ perovskite polymorphs in each of the studied structures both pristine and Sb or I substituted.

Table C7 - Deviation of geometrical parameters for the corresponding MX_6 octahedra in the doped trigonal $Cs_3Sb_2I_9$ in relation to pristine.

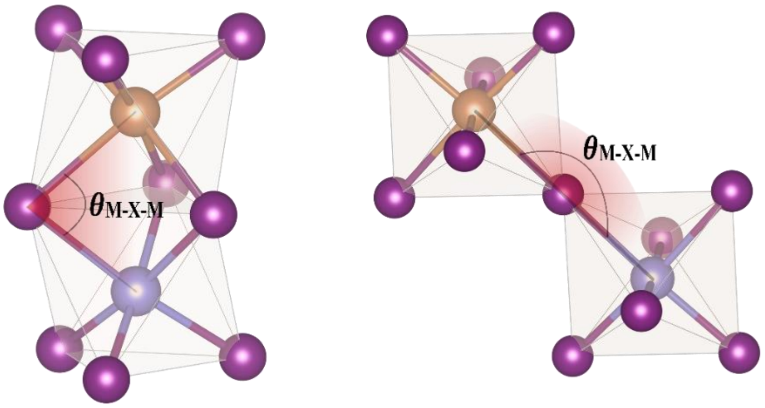
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)	a_{axis}	a_{min}	a_{max}	δ_X	δ_Y	δ_Z	δ_X^{abs}	δ_Y^{abs}	δ_Z^{abs}
Pristine	6.10	4.31	4.34	0.0	0.16	0.12	0.0	0.0	0.20
$Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with	Δa_{axis}	Δa_{min}	Δa_{max}	$\Delta\delta_X$	$\Delta\delta_Y$	$\Delta\delta_Z$	$\Delta\delta_X^{abs}$	$\Delta\delta_Y^{abs}$	$\Delta\delta_Z^{abs}$
Ag	-0.16	-0.17	-0.13	0.0	0.0	0.0	0.0	0.0	0.0
In	-0.15	-0.15	-0.14	0.0	+0.06	+0.04	0.0	0.0	+0.07
Mo	-0.43	-0.30	-0.30	0.0	-0.1	-0.08	0.0	0.0	-0.13
Nb	-0.34	-0.25	-0.27	0.0	-0.05	-0.04	0.0	0.0	-0.07
Sc	-0.27	-0.21	-0.22	0.0	0.0	-0.01	0.0	0.0	0.0
Bi	+0.10	+0.04	+0.07	0.0	+0.01	0.0	0.0	0.0	+0.01
Br (22.2%)	-0.01	-0.18	-0.11	+0.06	+0.05	-0.01	0.0	+0.07	+0.04
Cl (22.2%)	-0.03	-0.17	-0.08	+0.11	+0.08	-0.02	-0.01	+0.11	+0.06

$Cs_3Sb_2I_9$ ($P\bar{3}m1$)	φ	θ	ψ	α_{sup}	α_{inf}	β_{sup}	β_{inf}
Pristine	23.93	39.78	71.61	71.22	70.53	70.52	69.83
$Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with	$\Delta\varphi$	$\Delta\theta$	$\Delta\psi$	$\Delta\alpha_{sup}$	$\Delta\alpha_{inf}$	$\Delta\beta_{sup}$	$\Delta\beta_{inf}$
Ag	+0.05	-0.18	+0.26	+0.03	-0.25	+0.26	-0.02
In	+0.21	-0.71	-0.76	-1.47	+0.57	-0.56	+1.48
Mo	+0.15	-0.5	+0.43	-0.18	-0.45	+0.46	+0.19
Nb	+0.15	-0.49	-0.24	-0.8	+0.18	-0.18	+0.81
Sc	+0.17	-0.57	+0.18	-1.01	+0.29	-0.28	+1.01
Bi	+0.01	-0.04	-0.46	-0.48	+0.44	-0.44	+0.48
Br (22.2%)	-0.08	+0.05	+1.63	-0.62	-0.9	-0.77	-1.05
Cl (22.2%)	-0.18	+0.07	+3.14	-1.23	-1.32	-1.61	-1.7

Table C8 - Deviation of geometrical parameters for the corresponding MX_6 octahedra in the doped hexagonal $Cs_3Sb_2I_9$ in relation to pristine.

$Cs_3Sb_2I_9$ ($P6_3/mmc$)	a_{axis}	a_{min}	a_{max}	δ_X	δ_Y	δ_Z	δ_X^{abs}	δ_Y^{abs}	δ_Z^{abs}
Pristine	6.13	4.32	4.39	0.0	-0.17	-0.13	0.0	0.0	-0.21
$Cs_3Sb_2I_9$ ($P6_3/mmc$) doped with	Δa_{axis}	Δa_{min}	Δa_{max}	$\Delta\delta_X$	$\Delta\delta_Y$	$\Delta\delta_Z$	$\Delta\delta_X^{abs}$	$\Delta\delta_Y^{abs}$	$\Delta\delta_Z^{abs}$
Ag	-0.15	-0.16	-0.10	0.0	0.0	+0.01	0.0	0.0	+0.01
In	-0.13	-0.12	-0.11	0.0	-0.06	-0.04	0.0	0.0	-0.07
Mo	-0.44	-0.38	-0.32	0.0	+0.07	+0.05	0.0	0.0	+0.08
Nb	-0.35	-0.28	-0.26	0.0	+0.03	+0.03	0.0	0.0	+0.04
Sc	-0.24	-0.23	-0.18	0.0	-0.05	-0.03	0.0	0.0	-0.06
Bi	+0.09	+0.04	+0.07	0.0	0.0	+0.25	+0.21	0.0	+0.21
Br (22.2%)	+0.01	-0.10	-0.03	+0.02	+0.02	+0.01	0.0	+0.03	+0.02
Cl (22.2%)	+0.02	-0.26	-0.14	+0.01	+0.06	+0.01	-0.02	+0.04	+0.05

$Cs_3Sb_2I_9$ ($P6_3/mmc$)	φ	θ	ψ	α_{sup}	α_{inf}	β_{sup}	β_{inf}
Pristine	23.79	40.23	70.56	70.71	71.61	69.43	70.33
$Cs_3Sb_2I_9$ ($P6_3/mmc$) doped with	$\Delta\varphi$	$\Delta\theta$	$\Delta\psi$	$\Delta\alpha_{sup}$	$\Delta\alpha_{inf}$	$\Delta\beta_{sup}$	$\Delta\beta_{inf}$
Ag	-0.02	+0.07	+0.33	+0.26	-0.18	+0.18	-0.27
In	0.09	-0.29	+0.57	+0.17	-0.54	+0.54	-0.16
Mo	0.07	-0.25	-0.01	-0.37	+0.07	-0.06	+0.38
Nb	0.08	-0.26	+0.88	+0.09	-0.41	+0.42	-0.09
Sc	0.07	-0.22	+1.04	-0.08	-0.19	+0.20	+0.09
Bi	-0.04	+9.42	+0.38	+0.35	-0.21	+0.20	-0.36
Br (22.2%)	1.39	-0.89	+3.32	-1.77	-0.75	-1.07	-0.08
Cl (22.2%)	0.51	-0.01	+5.00	-2.05	-1.36	-2.46	-1.83



$\theta_{M\text{-}X\text{-}M}$
$\theta_{M\text{-}X\text{-}M}$

Figure C9– Illustration of the metal-halogen-metal angle in the octahedra of hexagonal (left) and trigonal (right) structures.

Table C9 – Deviation of metal-halogen-metal angle (M-X-M) for MX_6 octahedra in the doped hexagonal and trigonal $Cs_3Sb_2I_9$ in relation to pristine.

$Cs_3Sb_2I_9$ ($P\bar{3}m1$)

Dopant	Pristine	Ag	In	Mo	Nb	Sc	Bi
$\boldsymbol{\theta}_{\mathbf{M\text{-}X\text{-}M}}$	180.0	177.59	177.39	178.78	179.67	178.76	179.85
$\boldsymbol{\Delta\theta}_{\mathbf{M\text{-}X\text{-}M}}$	0.0	-2.41	-2.61	-1.22	-0.33	-1.24	-0.15

$Cs_3Sb_2I_9$ ($P6_3/mmc$)

Dopant	Pristine	Ag	In	Mo	Nb	Sc	Bi
$\boldsymbol{\theta}_{\mathbf{M\text{-}X\text{-}M}}$	78.17	79.10	80.33	80.34	80.94	81.14	78.52
$\boldsymbol{\Delta\theta}_{\mathbf{M\text{-}X\text{-}M}}$	0.0	+0.93	+2.16	2.17	+2.77	+2.97	+0.35

C.2.3 Band structures and effective mass analysis

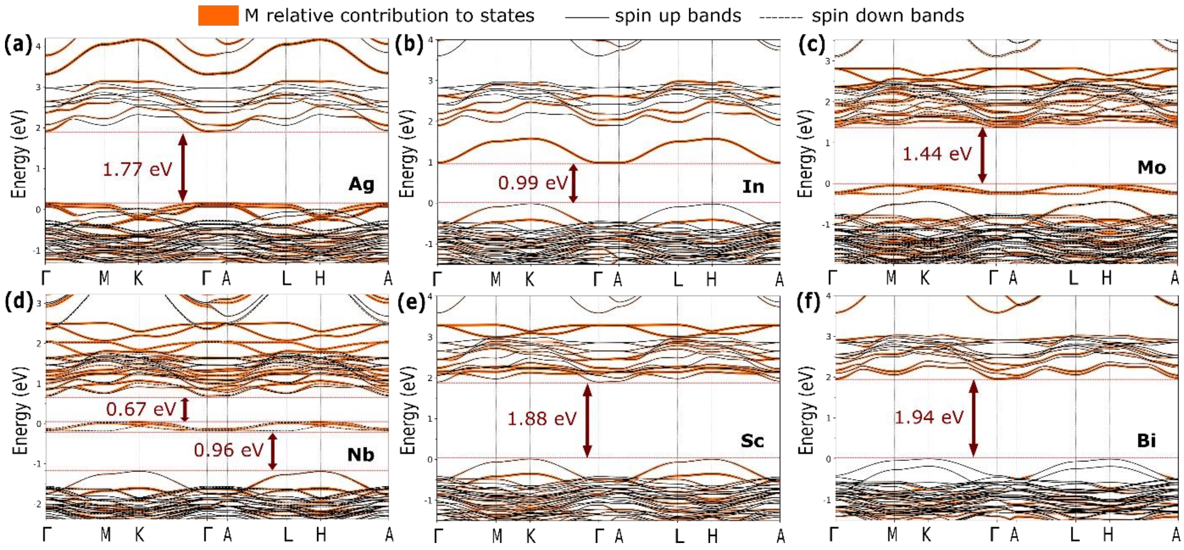


Figure C10– $Cs_3Sb_{1.5}M_{0.5}I_9$ (P6$_3$/mmc) band structure for M = (a) Ag, (b) In, (c) Mo, (d) Nb, (e) Sc, and (f) Bi, respectively. Fermi level at 0 eV.

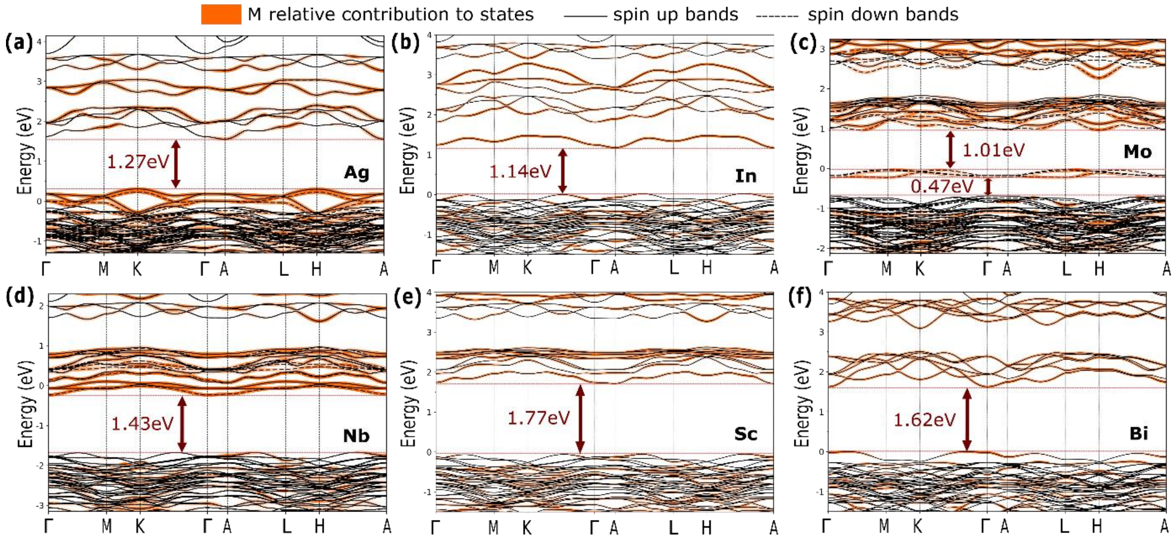


Figure C11 – $Cs_3Sb_{1.5}M_{0.5}I_9$ (P$\bar{3}$m1) band structure for M = (a) Ag, (b) In, (c) Mo, (d) Nb, (e) Sc, and (f) Bi, respectively. Fermi level at 0 eV.

The data on effective mass calculated for each doped structure is presented on Table 9 of the manuscript. The values for the $k_{[001]}$ direction are not presented since they were not representative, with nearly flat bands in the Γ–A direction for all doped structures (see *Figure C10* and *C11*). For both P6$_3$/mmc and P$\bar{3}$m1 structures, the effective masses for electrons increased in the $k_{[100]}$ direction for all doping cases, except for Bi-doped structures, with a slight decrease.

The doped $Cs_3Sb_2I_9$ ($P6_3/mmc$) structures maintained the hole effective mass in the $k_{[100]}$ direction approximately constant (around 1.1 m_e), except for the Ag and Mo-doped structures in which representative effective masses could not be obtained due to very flat band structures. This can be attributed, in both cases, to the introduction of localized *d* states in the valence band that overlap the pristine Sb 5*s* orbitals reducing dispersion in the valence band. Furthermore, the anisotropic structure combined with the introduction of s^2 electron lone pair from In electronic structure can lead to a higher hole effective masses (Brandt et al. 2015). The doped $Cs_3Sb_2I_9$ ($P\overline{3}m1$) structures present similar trend for their hole effective masses with significant increase in the cases of Ag, Mo and Nb doped, again due to the localized *d* orbitals introduced in the VBM. In this case, the flat band structure resulting from doping with Bi hindered obtaining a representative hole effective mass.

In the case of electron effective masses, the increase is more prominent in the In- and Mo-doped $P6_3/mmc$ and In- and Nb-doped $P\overline{3}m1$ structures, which double or triple effective mass value decreasing carrier mobility substantially. Nb and Mo contribute substantially in the CBM of their corresponding doped structures and since *d* states are fairly localized and are not present in the pristine structure this increase is expected. In the case of indium doped structures, however, this effect is due to In 5*s* states in the conduction band disrupting the *p*-*p* character of interactions from the pristine structure. The composition of $Cs_3Sb_{1.5}In_{0.5}I_9$ CBM is illustrated in *Figure C7*.

Sc and Bi-doped structures present similar effective masses to the pristine structure in the case of the hexagonal polymorph, however for the trigonal polymorph the hole masses increase significantly creating a flat band in the case of $Cs_3Sb_{1.5}Bi_{0.5}I_9$ $P\overline{3}m1$, this is due to a spurious contribution of Bi 6s states in the VBM that can be seen in *Figure 32* on the main text. This contribution is absent in the case of the hexagonal structure and the geometric analysis in the BiI_6 octahedra of both structures suggests that Bi stabilization in the hexagonal structure happens through octahedra tilting and central ion displacement (in Table C7 and C8, check $\Delta\delta_Z$ and $\Delta\theta$). It is important to point out that since Bi is a heavy element, spin-orbit coupling which was not considered in our calculations would certainly influence the obtained values. However, the point of introducing Bi is to evaluate the difference of this dopant in the distinct polymorphs which is likely to hold in a demanding HSE+SOC calculation necessary to obtain more accurate values.

C.2.4 Formation energy data analysis

A multilinear regression model in R with the "caret" library and 10-fold cross-validation was used to analyze the impact of certain predictors on the formation energy of substitutions with metal-doped structures. The predictors included atomic properties and Bader charge values for Sb and dopant metals, as well as bond distances. The results showed that Bader charge values for dopant metals and iodine had a strong influence on formation energies, with electronegativity and bond distance playing a role. *Figure C12* presents a correlation matrix to assess covariances and their impact on model accuracy. It is clear that efficiently distributed charge and higher ionic character of the bond lead to lower formation energy.

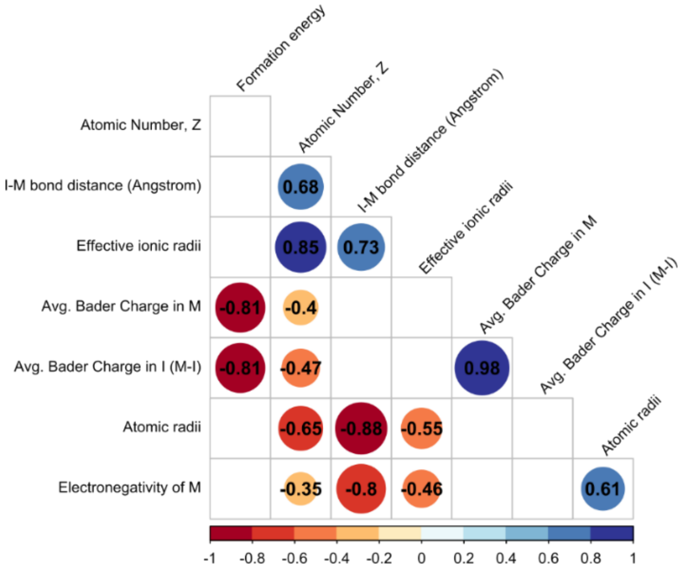


Figure C12 – Correlogram of the numeric variables on the created dataset for evaluating influences in the defect formation energy for the M-doped structures. Shown values have significance level of 95%.

C.2.5 ACBN0 calculations of the metal-doped $Cs_3Sb_2I_9$

The ACBN0 method (Agapito, Curtarolo, and Nardelli 2015) operates on the premise of self-consistently determining the local Coulomb repulsion parameter, U, within the DFT+U framework (Anisimov, Aryasetiawan, and Lichtenstein 1997). This compensates for the overdelocalization inherent in LDA and GGA exchange-correlation functionals. In this approach, U values for distinct atomic sites are derived from the bare Coulomb and exchange interactions, computed via a renormalized occupation matrix resulting from the projection of DFT Kohn-Sham wave functions onto the minimal PAO-3G basis set (Agapito et al. 2013). These U values are iteratively converged through successive DFT+U calculations and projections. In our computations, convergence was achieved when changes in U values for each Hubbard site fell below 0.1 eV, adhering to the default implementation of ACBN0 in the AFLOWπ package (Supka et al. 2017). This method offers notable advantages, including its adaptability to unique Hubbard sites and its significantly reduced computational expense compared to hybrid functionals. Hence, it emerges as one of the few viable methods for accurately predicting band gaps in data-driven research and investigations involving supercells, such as the present study. Notably, the ACBN0 method approaches the band gap accuracy attained by GW and HSE levels of theory, similar to the alternative low-cost mBJLDA (Tran and Blaha 2009) but with a better description of materials in which orbital-dependent potentials are important (S. H. Lee and Son 2020; Koller, Tran, and Blaha 2011).

The calculation of Hubbard values with ACBN0 as implemented in AFLOWπ required norm-conserving pseudopotentials, we verified if the U values determined for the pristine structures would result in similar band structure when applied to the GBRV ultrasoft pseudopotentials. The results are presented in *Figure C13* for both polymorphs and we cannot observe significant deviation on the band structure, this is in accordance to previous observations regarding transferability of ACBN0 Hubbard values (S. H. Lee and Son 2020).

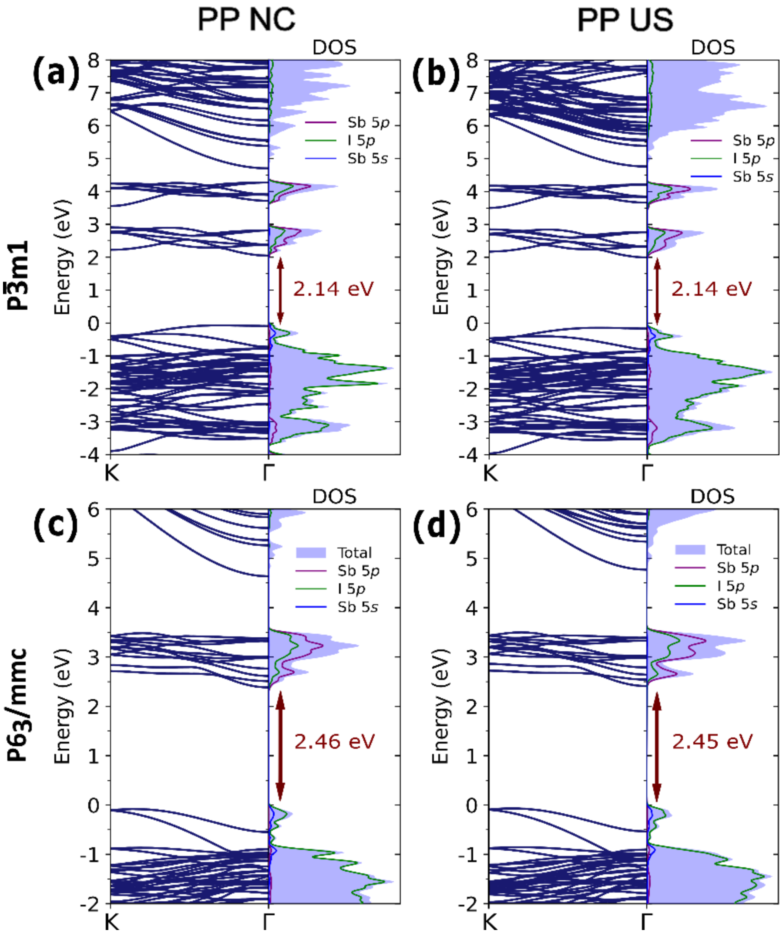


Figure C13 - Band structure and projected density of states of pristine $Cs_3Sb_2I_9$ utilizing the ACBN0 method. Norm-conserving and ultrasoft pseudopotentials are displayed on the right and left, respectively, for the P$\bar{3}$m1 polymorph (a-b) and P6$_3$/mmc polymorph(c-d). The Fermi level is marked at 0 eV.

We present in *Table C10* the ACBN0 converged Hubbard U values for pristine and doped structures, we can see that the pristine P$\bar{3}$m1 structure presented U_{I5p} value is in close agreement with our recent work (Gouvêa et al. 2024) applying DFT+U on $Cs_3Sb_2X_9$ (X = Cl, Br, I) which presented U_{I5p} = 4 eV.

Table C10 – ACBN0 converged Hubbard U values for pristine and doped $Cs_3Sb_2I_9$ polymorphs.

Structure	Hubbard values (eV) in given orbital				
	Cs 5s	Sb 5p	I 5p	In 5s	Sc 3d
$Cs_3Sb_2I_9$ ($P6_3/mmc$)					
Pristine (Sb)	0.00	0.06	4.21	-	-
In-doped	0.00	0.05	4.30	15.56	-
Sc-doped	0.00	0.05	4.25	-	0.00
$Cs_3Sb_2I_9$ ($P\overline{3}m1$)					
Pristine (Sb)	0.00	0.06	4.27	-	-
In-doped	0.00	0.04	4.32	15.56	-
Sc-doped	0.00	0.05	4.27	-	0.00

Table C11 presents the indirect and direct gap values for both pristine and the selected doped structures. Notably, all structures exhibit an increase in the band gap compared to PBE calculations, with values closely approaching experimental data for the pristine structures. Furthermore, the observed trend of the band gap shifting towards indirect for In-doped structures and slightly more direct for Sc-doped structures remains consistent. This indicates agreement between the plain PBE and ACBN0 methods regarding the overall spatial distribution of the orbitals. However, differences from the PBE results are more pronounced in the hexagonal structure, suggesting a stronger localization that is better captured with Hubbard corrections.

Table C11 – Direct and indirect band gap for ACBN0 calculated $Cs_3Sb_2I_9$ pristine and doped structures for both polymorphs.

Structure	Indirect gap (eV)	Direct gap (eV)	Δ(direct-indirect) gap
$Cs_3Sb_2I_9$ ($P6_3/mmc$)			
Pristine (Sb)	2.45 (**K*–Γ**)	2.61 (**M–M**)	0.16
In-doped	1.47 (**K–Γ**)	2.08 (**Γ–Γ**)	0.61
Sc-doped	2.46 (**K–Γ**)	2.60 (**M–M**)	0.14
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)			
Pristine (Sb)	2.14 (**K*–Γ**)	2.16 (**M–M**)	0.02
In-doped	1.76 (**K*–Γ**)	1.88 (**A–A**)	0.12
Sc-doped	2.43 (**K–Γ**)	2.53 (**M–M**)	0.10

K* is a point in the K - Γ high-symmetry line.

C.3 Halogen doping

Table C12 – Halogen substitution formation energy for $Cs_3Sb_2I_9$ ($P6_3/mmc$) and $Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with Cl or Br in each of the inequivalent iodine sites of the corresponding perovskite.

Structure	Dopant	Wyckoff position	$E_f[D]$ (eV)	$\Delta E_f[D]$ (eV)
$Cs_3Sb_2I_9$ ($P6_3/mmc$)				
	Cl	k	-0.534	0.081
		h	-0.452	
	Br	k	-0.469	0.013
		h	-0.457	
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)				
	Cl	i	-0.571	0.139
		e	-0.432	
	Br	i	-0.467	0.009
		e	-0.458	

Table C13 – Formation energies for halogen doping, average Bader charge of Iodine in the Sb–I bond, and average Bader charge for the dopant halogen (X) in the Sb–X bond.

Lattice and doped compound	$E_f[D]$ (eV)*	Avg. Bader charge (I–Sb)	Avg. Bader charge (X–Sb)	Avg. Bader charge (Sb)
$Cs_3Sb_2I_9$ ($P6_3/mmc$)				
$Cs_3Sb_2Br_2I_7$	-0.469	-0.493	-0.592	1.041
$Cs_3Sb_2Cl_2I_7$	-0.534	-0.494	-0.649	1.115
$Cs_3Sb_2I_9$ ($P\bar{3}m1$)				
$Cs_3Sb_2Br_2I_7$	-0.458	-0.506	-0.562	1.039
$Cs_3Sb_2Cl_2I_7$	-0.571	-0.509	-0.624	1.109

*The formation energy was calculated with the substitution of one halogen atom.

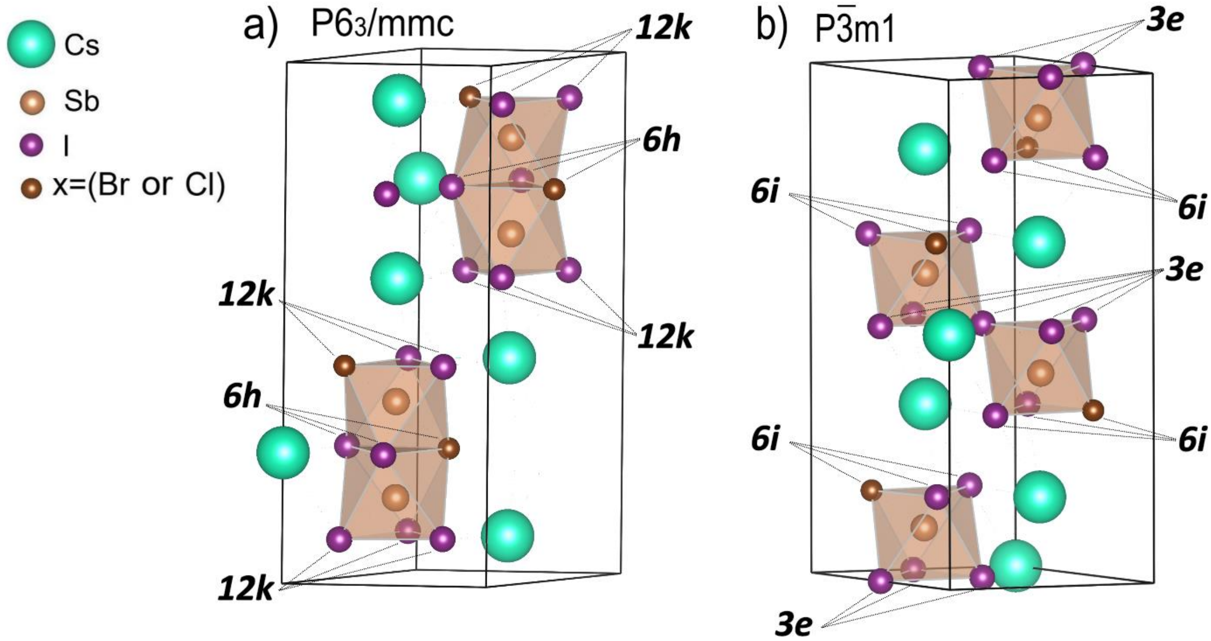


Figure C14 - Structures of the polymorphs (a) $Cs_3Sb_2I_9$ ($P6_3/mmc$), and (b) $Cs_3Sb_2I_9$ ($P\bar{3}m1$) doped with halogen (X = Br or Cl). Wyckoff positions for each bridging and terminal halogen is indicated.

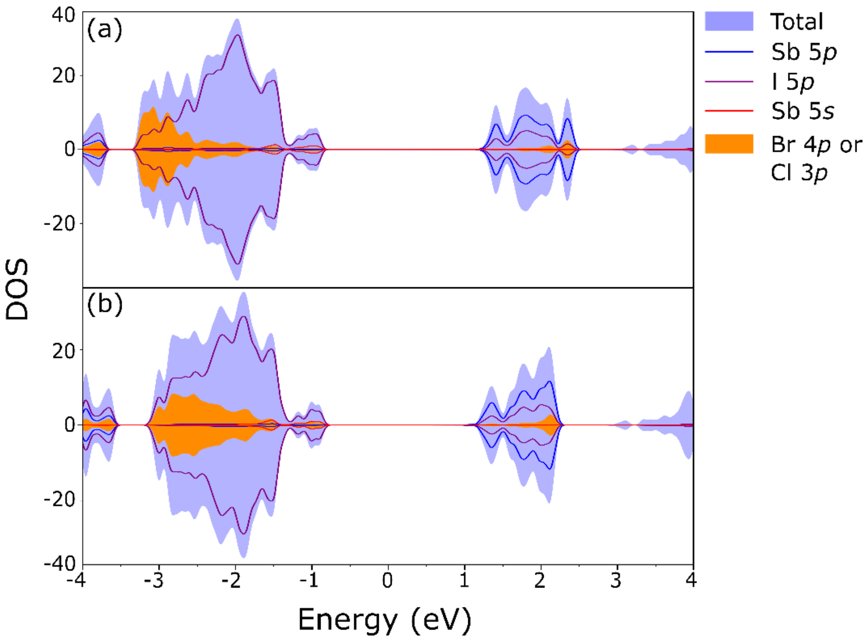


Figure C15 - Projected density of states for $Cs_3Sb_2I_9$ (P6$_3$/mmc) doped with (a) Br, and (b) Cl, respectively.

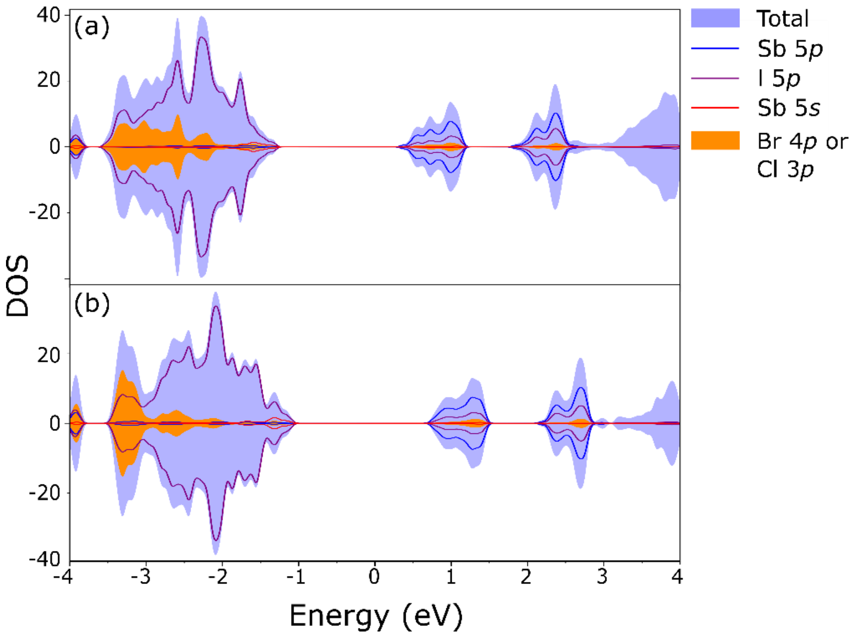


Figure C16 - Projected density of states for $Cs_3Sb_2I_9$ (P$\bar{3}$m1) doped with (a) Br, and (b) Cl, respectively.

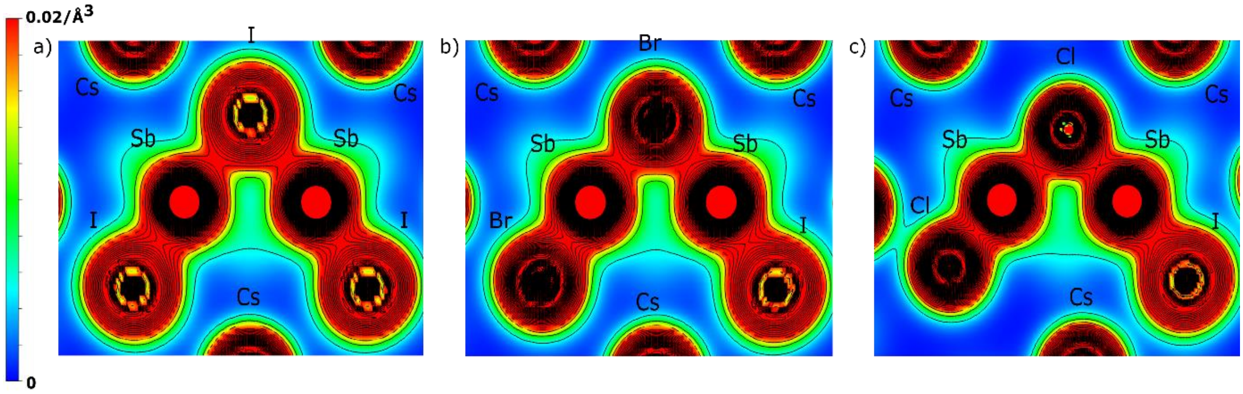


Figure C17 - Charge density of $Cs_3Sb_2I_9$ (P6$_3$/mmc) (a) pristine and doped with (b) Br, and (c) Cl, respectively.

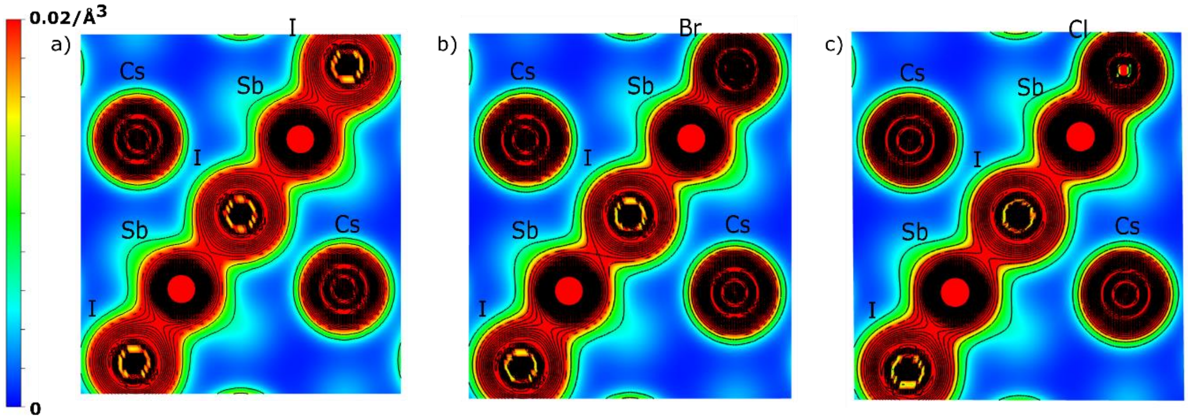


Figure C18 - Charge density of $Cs_3Sb_2I_9$ (P$\bar{3}$m1) (a) pristine and doped with (b) Br, and (c) Cl, respectively.

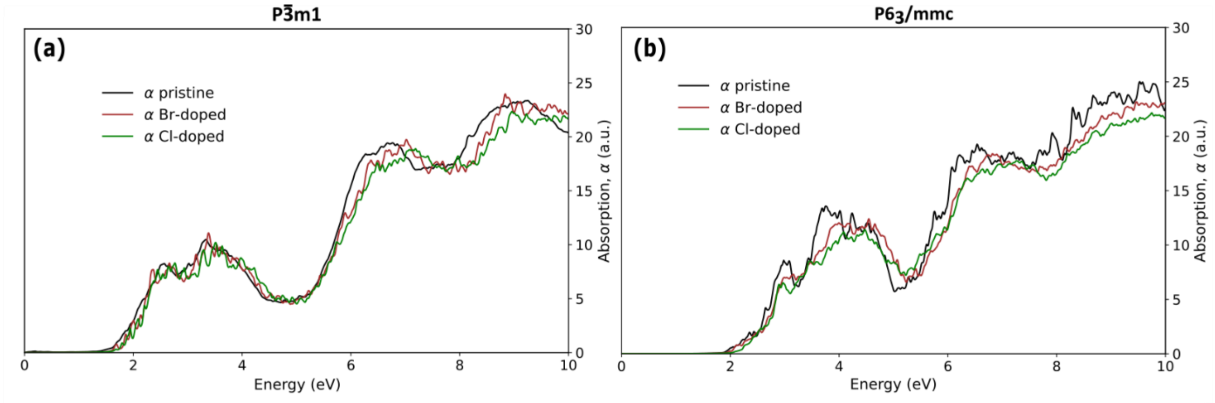
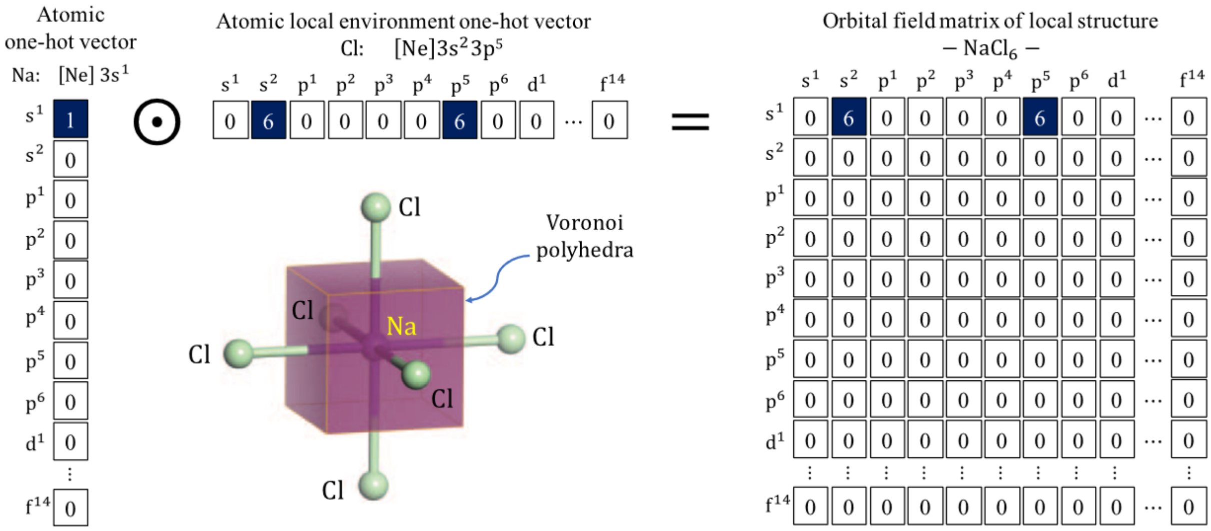


Figure C19 - Absorption coefficient for pristine and halogen-doped (a) $Cs_3Sb_2I_9$ (P$\bar{3}$m1), and (b) $Cs_3Sb_2I_9$ (P6$_3$/mmc).

APPENDIX D: Supporting Information for “Boosting feature-based machine learning models for materials science: encoding descriptors and graph-based features for enhanced accuracy and faster featurization in MODNet”

D.1 Orbital field matrix featurizer

This study follows the original Orbital Field Matrix (OFM) implementation from Lam Pham et al. (2017), as also found in the MatMiner featurizer. The neutral valence shell electronic configurations of elements can be represented as one-hot encoded vectors using an ordered dictionary, D = {s^1, s^2, p^1, p^2, ..., p^6, d^1, d^2, ..., d^{10}, f^1, f^2, ..., f^{14}}. For example, Na and Cl have electronic configurations [Ne]3s^1 and [Ne]3$s^2$3p^5. Sodium can then be represented by a one-hot encoded vector with position s^1 set to 1, while chlorine's vector has positions s^2 and p^5 set to 1 (remaining entries are zeros). If we consider these elements within a crystal structure, as illustrated in Figure D1, the OFM descriptor aims to capture the valence shell interactions in each site.



Atomic one-hot vector
Na: [Ne] 3s¹
Atomic local environment one-hot vector
Cl: [Ne]3s²3p⁵
s¹ s² p¹ p² p³ p⁴ p⁵ p⁶ d¹ f¹⁴
0 6 0 0 0 0 6 0 0 ··· 0
Orbital field matrix of local structure
– NaCl₆ –
Cl
Na
Voronoi polyhedra

Figure D1 - OFM representation for an Na atom in a regular octahedral site surrounded by six Cl atoms. Source: (Lam Pham et al. 2017)

It is important that the descriptor captures site coordination and element distance from neighboring atoms. Therefore, the OFM for a central atom in a site (X^p) is defined as the weighted outer vector product of one-hot encoded atomic vectors, such as:

$$X_{ij}^{p} = \sum_{k=1}^{n_p} o_i^p o_j^k \frac{\theta_k^p}{\theta_{max}^p} \frac{1}{r_{pk}}. \tag{D1}$$

Here, *i, j* ∈ D, *k* is the index of nearest-neighbor atoms, n_p is the number of such atoms around site *p*, $\theta_k^p/\theta_{max}^p$ represents the weight of atom k in the coordination of the central atom at site p, θ_k^p is the solid angle determined by the Voronoi polyhedron face separating *k* and *p*, and θ_{max}^p is the maximum among n_p of them. r_{pk} captures the distance separating atoms *p* and *k*, also distinguishing elements with the same valence configuration. To construct the OFM for a crystal structure local OFMs are summed, and the values are averaged by the number of sites:

$$F_{ij} = \frac{1}{N_p} \sum_{p}^{N_p} X_{ij}^{p} \tag{D2}$$

D.2 MEGNet framework and pre-trained models

Figure D2 illustrates the architecture of the MEGNet framework based on a graph convolutional network. As depicted in the figure, the final MLP of the model preceding the output contains two sequential dense layers of 32 and 16. These values can be tuned for hyperparameter optimization as elaborated on the next section, particularly the default architecture corresponds to $h_1 = 64,\ h_2\ =\ 32$, and h_3 = 16. In which h_1 influences the MLPs inside the MEGNet blocks.

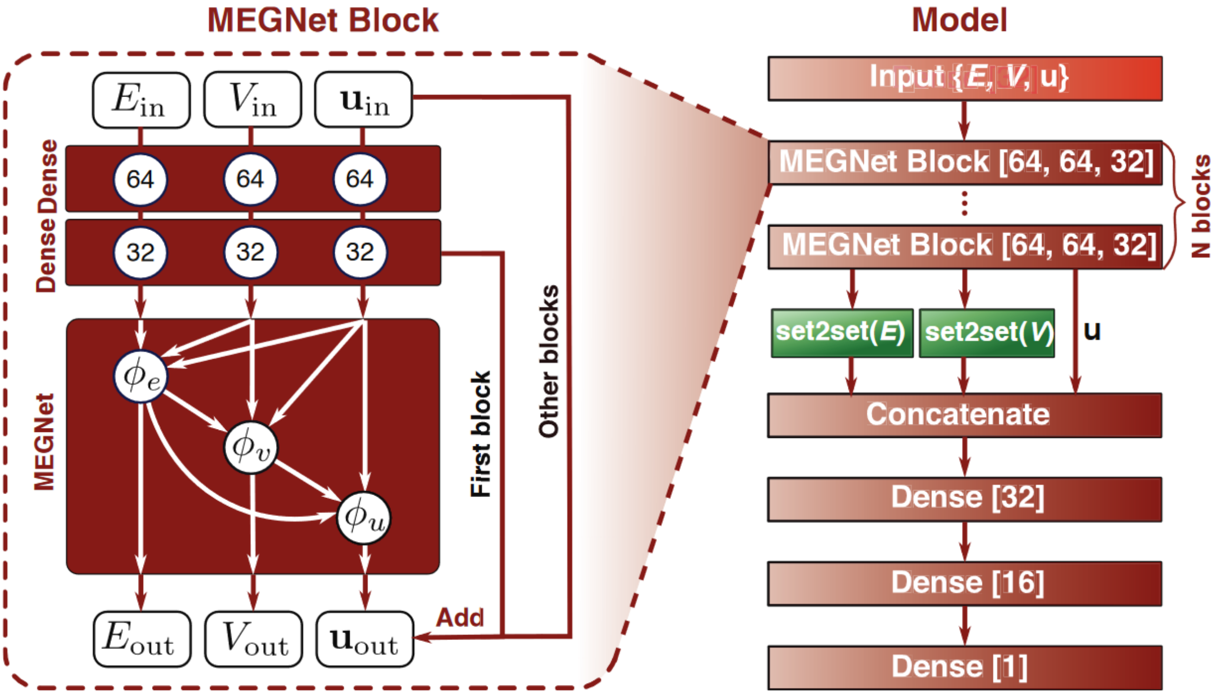


Figure D2 - Architecture for the MEGNet model. In the pre-trained models used in this work the same architecture was present with three MEGNet blocks. The numbers in brackets are the number of neurons for each layer. Source: (C. Chen et al. 2019)

D.3 Hyperparameter tuning

D.3.1 Autoencoders' hyperparameters

The autoencoder architecture employed in this study consisted of a feedforward neural network constructed with the Keras framework (Chollet 2015) consisting of a single hidden layer for both the encoder and decoder. The number of neurons in the hidden layer was initialized at 2 times the number of features in the featurizer (n), whether OFM or general MatMiner features. Architectures with two hidden layers were excluded in the preliminary tests, as were hidden layers with a number of neurons smaller than n, which yielded poorer results. Hyperparameter tuning was conducted in two steps. Initially, the features' compression was fixed at 50% (approximately $n/2$ resulting features), and the optimal configuration was sought, considering the following possibilities, shown in *Table D1*. Adam optimizer was utilized for weight optimization during backpropagation (see Appendix A.13) For these combinations, the configurations with the smallest average reconstruction errors over three runs, employing a train-test split of 9:1, are presented in *Table D2*.

Table D1 – Hyperparameters and corresponding values considered for the autoencoder optimization.

Hyperparameter	Possible Values
Batch Size	$16, 32, 64, 128$
Number of Epochs	$50, 100, 200, 300$
Learning Rate	$0.0005, 0.001, 0.002$

Table D2 – Best hyperparameters for autoencoders in this work considering a 50% compression.

Encoded featurizer	Batch size	Number of epochs	Learning rate
OFM	64	300	0.001
MatMiner MODNet v.0.1.13	64	200	0.0005

Based on these parameters, we proceeded with a similar approach to vary the number of neurons in the dense layer, ranging from 1.5n to 2.5n in increments of 0.1n. This time, we tested compressions of 20%, 50% and 80%. The combined loss for these compressions was assessed to identify the optimal architecture. As a result, the hidden layer sizes were determined to be 2.5n for the OFM featurizer and 2.2n for the MatMiner featurizer. The final architecture for each autoencoder is depicted in Figure D3.

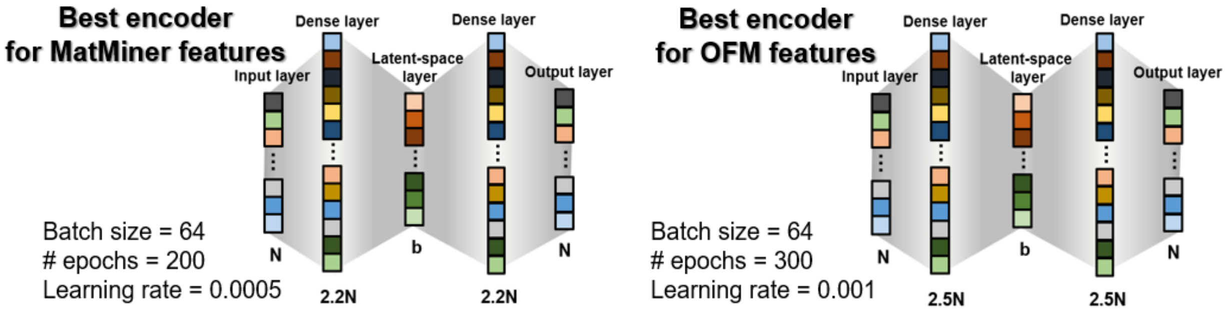


Figure D3 – Best autoencoder architectures found for MatMiner and OFM featurizers trained on matbench_v.0.1_mp_gap dataset.

Subsequently, the reconstruction loss was assessed for various levels of compression in each autoencoder, employing the same 9:1 train-test split. The results are outlined in Table D3 and Table D4. The encoder for MatMiner features consistently maintained the reconstruction error below 1%, even up to a compression to a latent-space size of 10% of the initial features. In the case of OFM, the compression was highly efficient, remaining below 0.1% mean absolute error (MAE) for most tested latent space sizes. Consequently, the reconstruction error is not anticipated to significantly impact predictions. Nonetheless, the most suitable latent space size must be determined by evaluating their performances in prediction tasks.

Table D3 – Reconstruction errors with different compression ratios for the autoencoder for MODNet's v.0.1.13 MatMiner featurizer. Errors in data normalized to the interval 0 to 1, metric for losses is MSE.

Compression ratio	**Latent n**	***Train Loss***	***Validation Loss***	***Test MAE***
*1.0**	1264	7.91e-05	7.69e-05	0.004789
0.9	1137	8.66e-05	8.52e-05	0.005098
0.8	1011	8.59e-05	8.04e-05	0.005010
0.7	884	8.60e-05	9.20e-05	0.005309
0.6	758	9.27e-05	9.45e-05	0.005411
0.5	631	9.79e-05	1.06e-04	0.005733
0.45	568	1.02e-04	1.13e-04	0.005880
0.4	505	1.09e-04	1.14e-04	0.005929
0.35	442	1.14e-04	1.28e-04	0.006269
0.3	379	1.29e-04	1.44e-04	0.006624
0.25	316	1.53e-04	1.64e-04	0.006962
0.2	252	1.82e-04	1.85e-04	0.007387
0.15	189	2.38e-04	2.32e-04	0.008094
0.1	126	3.26e-04	3.24e-04	0.009452
0.05	63	5.92e-04	5.87e-04	0.012396

** A compression ratio of 1.0 indicates a remapping to a latent space with the same dimensions. Note the number of dimensions may not precisely match the original featurizer's number of descriptors as some descriptors remain constant (0) throughout the dataset.*

Table D4 – Reconstruction errors with different compression ratios for the autoencoder for OFM featurizer. Errors in data normalized to the interval 0 to 1, metric for losses is MSE.

Compression ratio	Latent n	*Train Loss*	*Validation Loss*	*Test MAE*
*1.0**	943	2.50e-05	3.26e-05	0.000898
0.9	848	1.45e-05	1.55e-05	0.000718
0.8	754	5.09e-06	6.69e-06	0.000534
0.7	660	3.80e-06	5.10e-06	0.000518
0.6	565	8.59e-06	1.04e-05	0.000915
0.5	471	3.51e-06	4.80e-06	0.000474
0.45	424	5.34e-06	6.53e-06	0.000507
0.4	377	7.25e-06	1.02e-05	0.000608
0.35	330	3.26e-06	5.01e-06	0.000442
0.3	282	4.82e-05	5.38e-05	0.001278
0.25	235	1.56e-05	1.61e-05	0.000750
0.2	188	4.70e-06	8.52e-06	0.000742
0.15	141	2.06e-05	2.66e-05	0.000790
0.1	94	1.45e-05	2.10e-05	0.000821
0.05	47	1.00e-05	1.13e-05	0.000837

** A compression ratio of 1.0 indicates a remapping to a latent space with the same dimensions. Note the number of dimensions may not precisely match the original featurizer's number of descriptors as some descriptors remain constant (0) throughout the dataset.*

D.3.2 MEGNet models' hyperparameters

MEGNet models were trained to generate latent-space representations of encoded features (OFM and MatMiner features) and, in the case of the adjacent model, to produce general features based on the target property. No hyperparameter tuning was performed for the adjacent model, and the selected parameters are detailed in *Table D5*, relying on suggested values from MEGNet v1.3.2. All other parameters adhered to the default values in the MEGNet model function, including the utilization of 3 MEGNet blocks.

Table D5 – Hyperparameters applied for the adjacent MEGNet model training. Parameters not referred in the table follow the default values as of MEGNet's version 1.3.2.

Hyperparameter	Values
`nfeat_bond`	100
`r_cutoff`	5
`gaussian_width`	0.5
Number of epochs	100
MLP architecture $(h_1 x\ h_2 x\ h_3)$	$64x64x128$
Batch size	128
Learning Rate	0.001

For the MEGNet models used to generate latent space features, hyperparameter tuning played a crucial role and was executed in three steps. Initially, the number of epochs was varied across three different MLP architectures. Subsequently, the batch size (initially set at 32) and learning rate (default value of 0.001) were adjusted, with a new screening for the optimal number of epochs. Finally, a verification step was undertaken to assess whether increasing h_1 in the MLP architecture from 64 to 128 would yield improvement. This process resulted in a total of 37 trained models, all evaluated on the same train-test split, with 20% of the dataset reserved for testing. All hyperparameter values considered for the respective optimization cases are presented in *Table* D6.

Table D6 – Considered hyperparameter values for MEGNet models to generate encoded features for OFM and MatMiner featurizers.

Hyperparameter		Possible Values
Number of epochs		$10, 15, 20, 25, 30, 50, 70, 100$
MLP architecture	h_1	$64, 128$
$(h_1 x\ h_2 x\ h_3)$	$h_2\ x\ h_3$	$(16x32), (32x64), (64x128)$
Batch size		$16, 32, 64, 128$
Learning Rate		$0.0005, 0.001, 0.002$

A MEGNet model was trained to generate the latent OFM representation (20% compression), producing 188 features, and another MEGNet model to generate the latent representation of Matminer features (60% compression), producing 758 features. A few selected results for both MEGNet models considered are shown in *Table D7*. We can observe the relevance of hyperparameter tuning on the final loss of these models. Despite the substantial number of features, the MEGNet framework was very successful in reproducing the latent space features directly from the structure. Even for the more heterogeneous and large set of MatMiner features, the error was about 0.03, which corresponds to 3% of the total variation within each normalized feature.

Table D7 – MEGNet models' hyperparameters and reconstruction loss for generation of latent space features. Evaluation conducted on normalized features (range 0 to 1), highlighted in grey was the best obtained model on the hyperparameter screening.

Encoded featurizer	Hyperparameters				Reconstruction Loss (MAE)	
	Number of epochs	Batch size	Learning rate	MLP architecture $(h_1 x\ h_2 x\ h_3)$	Training	Test
Latent OFM, 20% compression (188 features)	15	32	0.0005	$64\ x\ 64\ x\ 32$	0.0180	0.0182
	25	64	0.001	$64\ x\ 64\ x\ 32$	0.0164	0.0166
	15	128	0.001	$64\ x\ 64\ x\ 32$	0.0137	0.0138
	25	32	0.0005	$64\ x\ 128\ x\ 64$	0.0131	0.0132
	25	32	0.001	$64\ x\ 128\ x\ 64$	0.0126	0.0127
Latent MatMiner MODNet v.0.1.13, 60% compression (758 features)	50	16	0.001	$64\ x\ 32\ x\ 16$	0.0671	0.0671
	20	64	0.0005	$64\ x\ 128\ x\ 64$	0.0484	0.0486
	30	16	0.001	$64\ x\ 32\ x\ 16$	0.0393	0.0393
	20	128	0.001	$128\ x\ 128\ x\ 64$	0.0324	0.0326
	50	128	0.0005	$64\ x\ 128\ x\ 64$	0.0306	0.0308

D.4 SHAP values analysis

In understanding complex machine learning models, SHAP (SHapley Additive exPlanations) emerges as a robust tool for revealing feature contributions (Lundberg and Lee 2017). SHAP values (ϕ) provide a clear view of how each feature influences predictions, employing Shapley values from cooperative game theory obtained through the formula,

$$\phi_i(f) = \frac{1}{N} \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \, [f(S \cup \{i\}) - f(S)], \tag{D3}$$

ensures a fair distribution of contributions, capturing the unique impact of each feature on model predictions. In the equation, the factorial terms in the denominator are crucial for normalization. The factorial function, denoted by the exclamation mark, represents the product of all positive integers up to a given integer n. Specifically, the terms $|S|!\,(|N| - |S| - 1)!$ and $|N|!$ ensure that contributions from each feature are appropriately scaled relative to the size of subsets (S) and the total number of features (N). Normalization plays a pivotal role in ensuring a fair and unbiased distribution of feature contributions. By accounting for the varying sizes of feature subsets and the entire set of features, the formula effectively weights each feature's contribution. This weighting ensures that the impact of individual features on model predictions is accurately reflected, without being overshadowed by the influence of larger feature sets.

Across a wide range of machine learning models, SHAP analysis serves as a valuable tool for assessing feature contributions. Implementation of SHAP analysis in MODNet neural networks was seamlessly achieved in this work using the *shap* python library (Lundberg and Lee 2017). However, it is important to note that the computational cost is considerably higher compared to simpler tree-based models. For instance, computing our SHAP summary plots containing 300 samples with 500 perturbations each took approximately 1 hour on 24 CPU cores. Nonetheless, we believe this computational expense is reasonable and worthwhile for recovering model interpretability.

When applied to tree-based models like XGBoost (T. Chen and Guestrin 2016) the inherent additivity and independence within tree ensembles streamline the SHAP calculation process, enabling fast parallel computation of contributions from individual trees. Consequently, XGBoost was selected as the model for generating surrogate models to decompose the contributions of chemical descriptors on the encoded GNN/latent-space features. Remarkably, these computations only required a few minutes in the same setup applied to the neural networks.

In all GNN and encoded features, such as those from Adjacent GNN, MEGNetPreL32 GNN and ℓ-OFM the encoded features can be correlated with other interpretable features derived from the same structures. For the latent-space OFM this can be done by direct comparison with the original OFM features. For the MEGNetPreL32 and Adjacent model features, we can still extract chemical information since the model also contains other interpretable features. To achieve this, we trained an XGBoost model for latent OFM features, predicting each feature from the original OFM features using the initial dataset employed for autoencoder training. Similarly, for MEGNetPreL32/Adjacent features, we trained an XGBoost model to predict the respective feature using both the original MatMiner features and the latent-space OFM features incorporated into the model. The XGBoost models were base to calculate the SHAP values offering clarity on the relationship of the selected features with interpretable properties.

D.4.1 SHAP analysis of MODNet model with "MM + ℓ-OFM + MEGNetPreL32" for matbench_perovskites

MODNet's feature selection algorithm, applied to MatMiner features, demonstrates a strong correspondence with the most important features for model prediction, aligning well with chemical intuition. However, when both ℓ-OFM and MEGNetPreL32 are included, although the algorithm correctly incorporates ℓ-OFM features, resulting in an increase in final accuracy, their attributed importance is relatively low, with these features appearing only in the second half of the selected features list to train the neural network. The significant accuracy boost observed when ℓ-OFM features are included highlights a limitation of the algorithm, likely related to a high degree of redundancy between ℓ-OFM and MEGNetPre32 features. This limitation is circumvented for our models by employing the more precise SHAP value analysis on the MODNet model. As illustrated in *Figure D4*, both ℓ-OFM and MEGNetPreL32 emerge now among the top 20 features in the model, as expected.

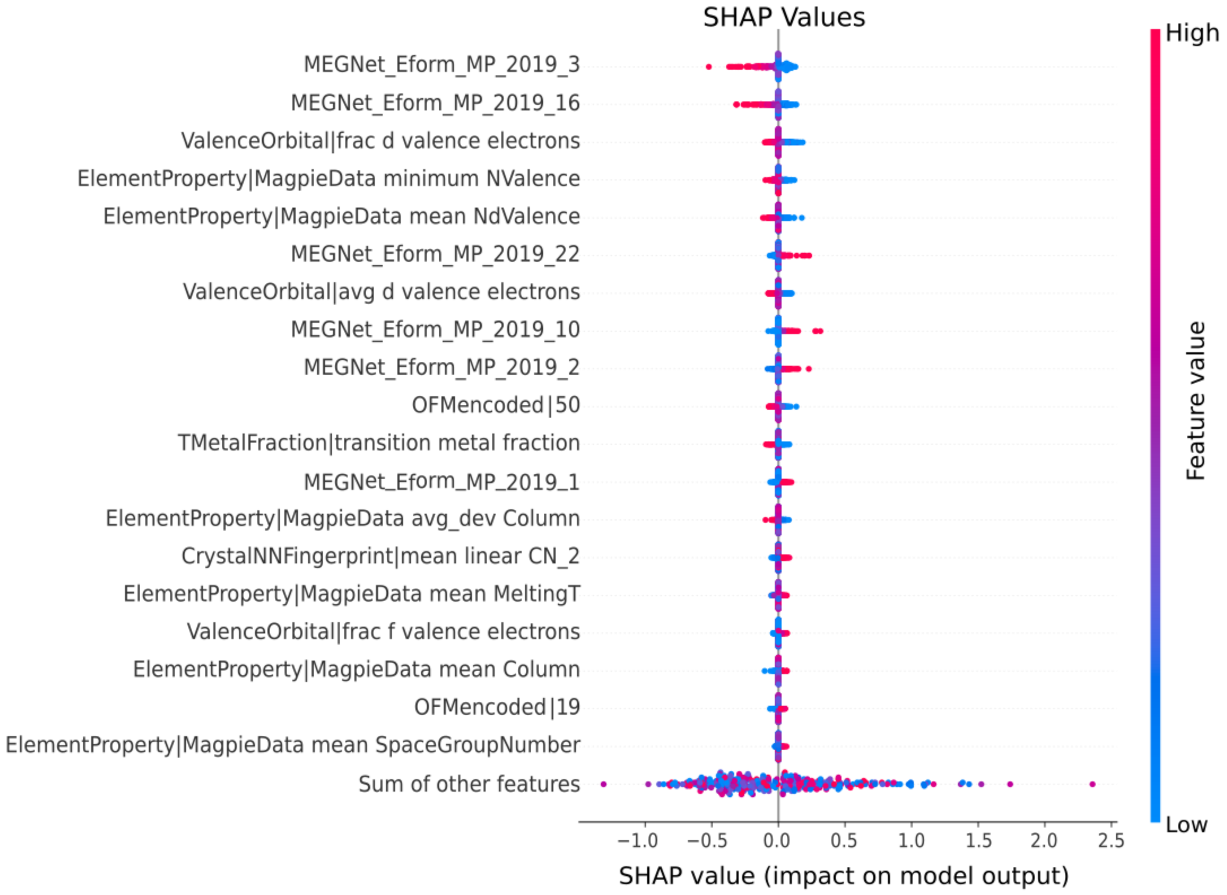


Figure D4 – SHAP analysis plot of the MODNet model with the features MM + ℓ-OFM + MEGNetPreL32.

The ℓ-OFM features are decomposed in chemical descriptors in *Figure D5* through the XGBoost surrogate models. Similarly, the three most relevant pre-trained MEGNet model features are decomposed in MM and ℓ-OFM descriptors in *Figure D6*.

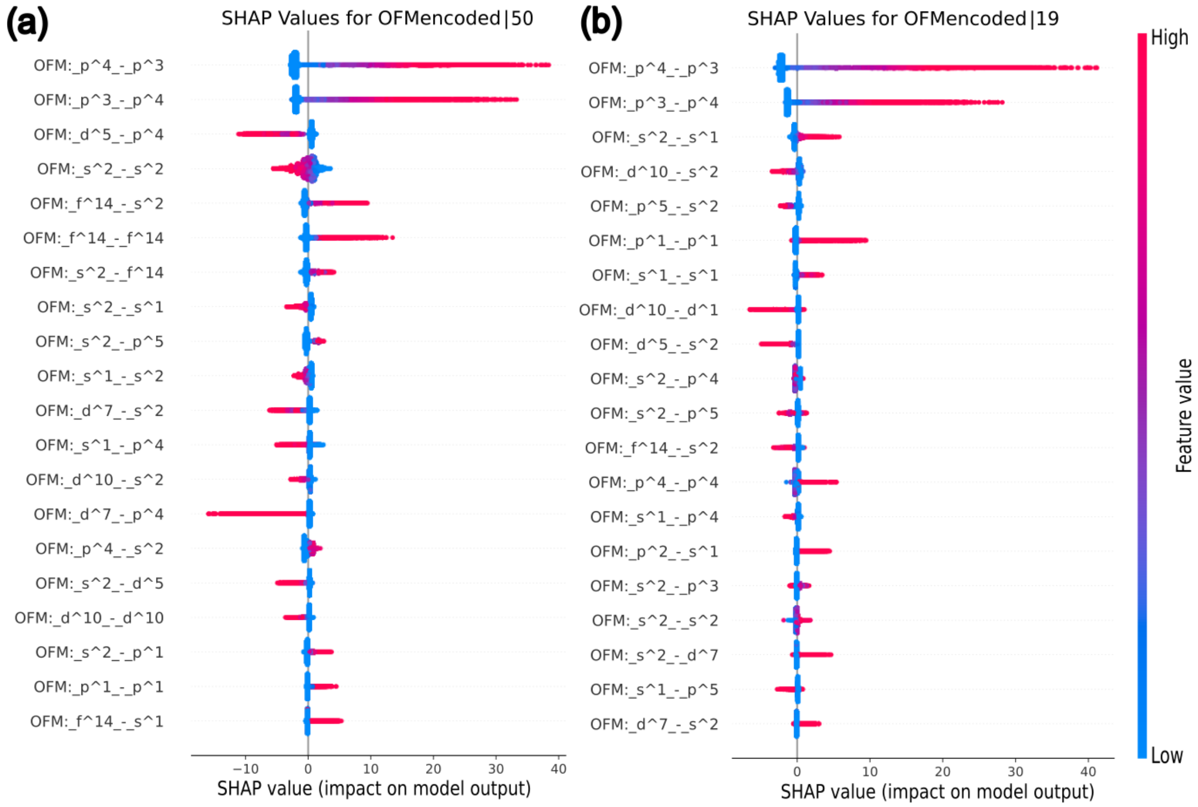


Figure D5 – SHAP analysis plot presenting original OFM contributions to the most relevant ℓ-OFM features in the MODNet model with "MM + ℓ-OFM + MEGNetPreL32" features. On the left (a), for the 50th ℓ-OFM component and, on the right (b), for the 19th ℓ-OFM component.

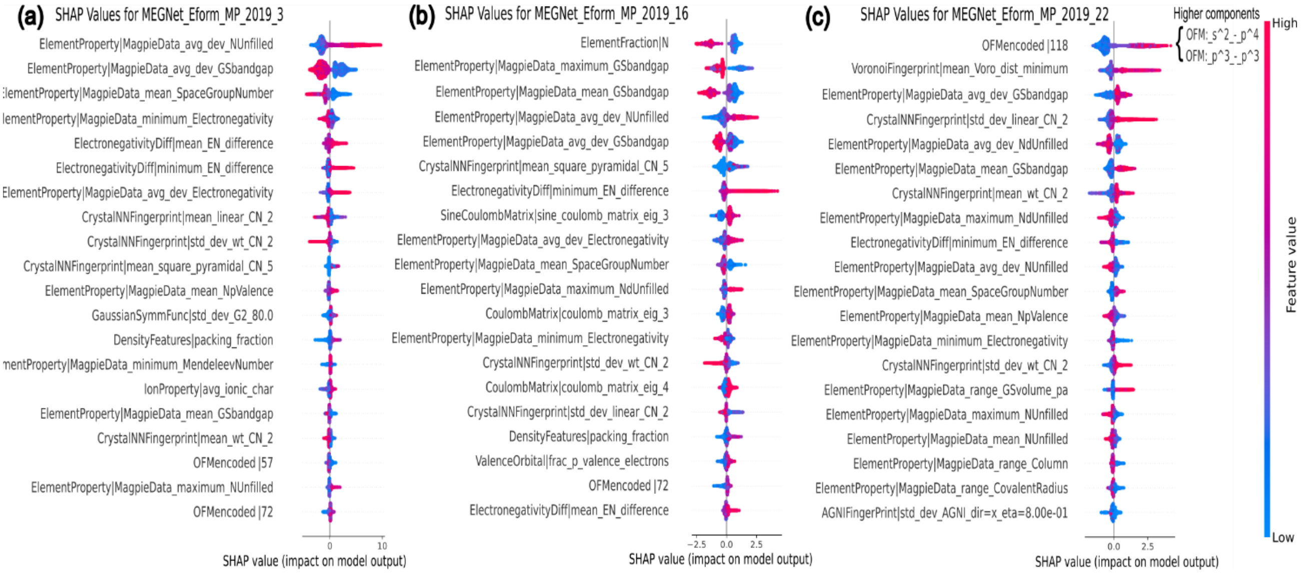


Figure D6 – SHAP analysis plot presenting the decomposition in MatMiner chemical descriptors and ℓ-OFM of the most relevant MEGNetPreL32 features in the MODNet model with "MM + ℓ-OFM + MEGNetPreL32" features. From left to right, the components for neurons #3 (a), #16 (b) and #22 (c).

D.4.2 SHAP analysis of MODNet model with OMEGA features for matbench_perovskites

The most important features of the MODNet model with OMEGA features are presented in *Figure D7* on the SHAP analysis plot. The adjacent features take precedence in the prediction, followed by the features of the pre-trained MEGNet formation energy model and MEGNet ℓ-OFM encoded features corresponding well to the results previously seen on the 'MM + ℓ-OFM + MEGNetPreL32' model.

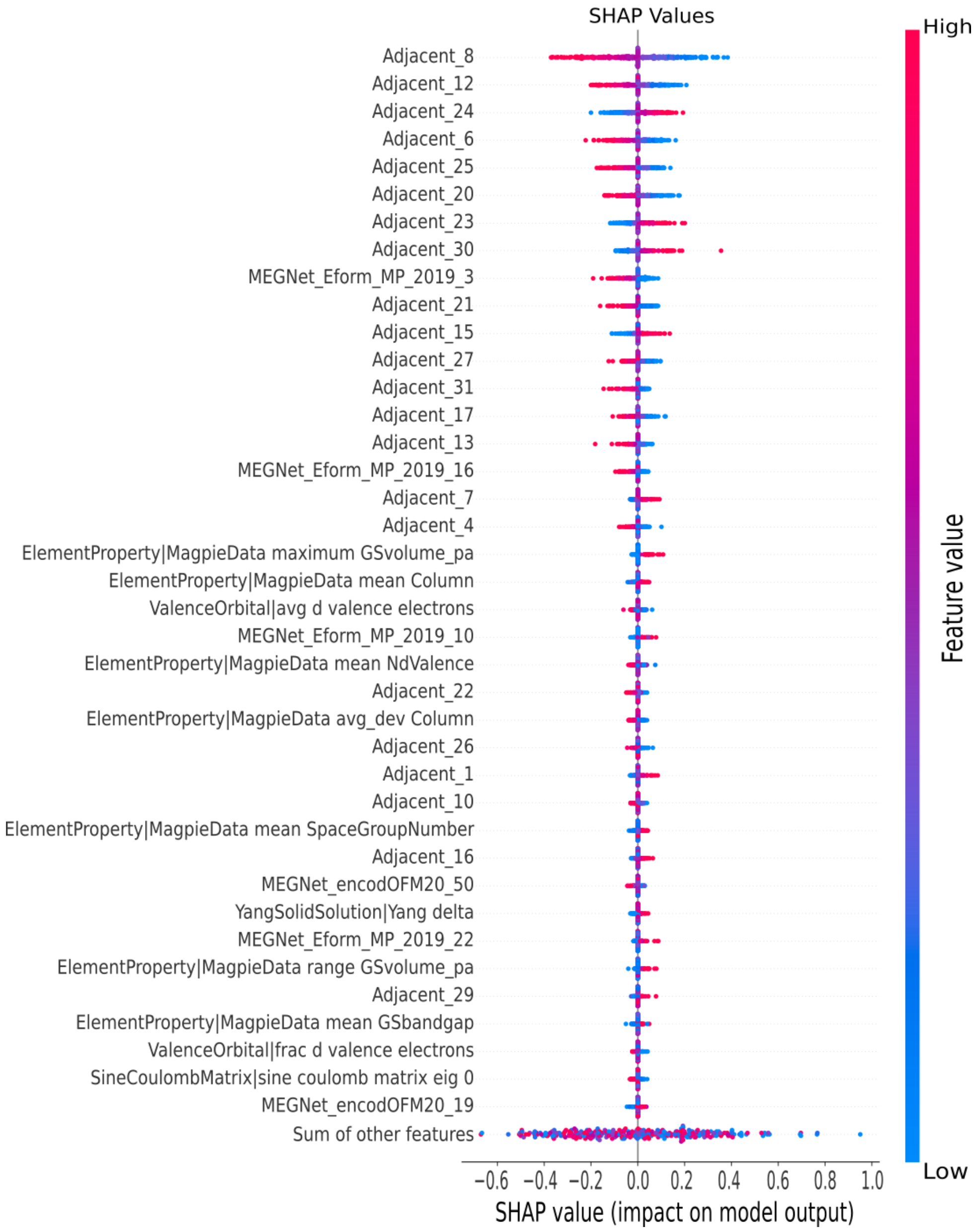


Figure D7 – SHAP analysis plot of the MODNet model for the matbench_perovskites task with the OMEGA features.

The decomposition of the adjacent model into MatMiner features is shown in the SHAP analysis plot in *Figure D8*. We can observe that, compared to the pre-trained MEGNet models, the adjacent model captures more subtle patterns such as geometrical fingerprints and sine Coulomb matrix eigenvalues. These nuances may be associated with its improved performance.

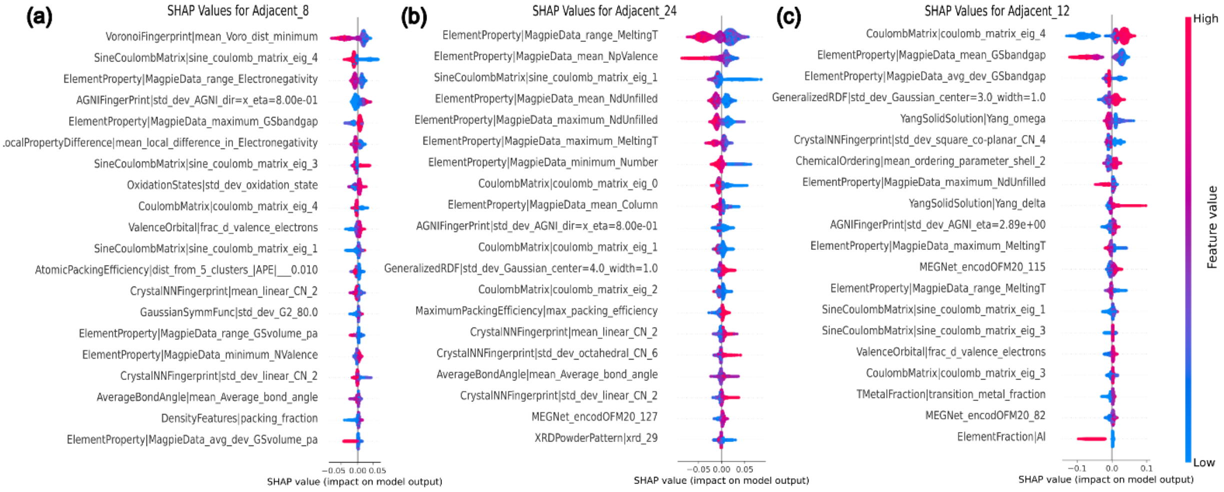


Figure D8 – SHAP analysis plot presenting the decomposition in MatMiner chemical descriptors and ℓ-OFM of the most relevant adjacent model features in the MODNet model with OMEGA features. From left to right, the components for neurons #8 (a), #24 (b) and #12 (c).

D.4.3 SHAP analysis of MODNet model with OMEGA features for OQMD halogen stability task

The most important features of the MODNet model with OMEGA features for the OQMD halogen task on stability are presented in *Figure D9* on the SHAP analysis plot. A notably uniform distribution of SHAP values across the features for this task is evident. The adjacent features ranking higher in the SHAP plot were correlated to chemical descriptors in *Figure D10*. Similarly, the encoded ℓ-OFM entries within the top features in the SHAP plot were analyzed for the associated original OFM components in *Figure D11*.

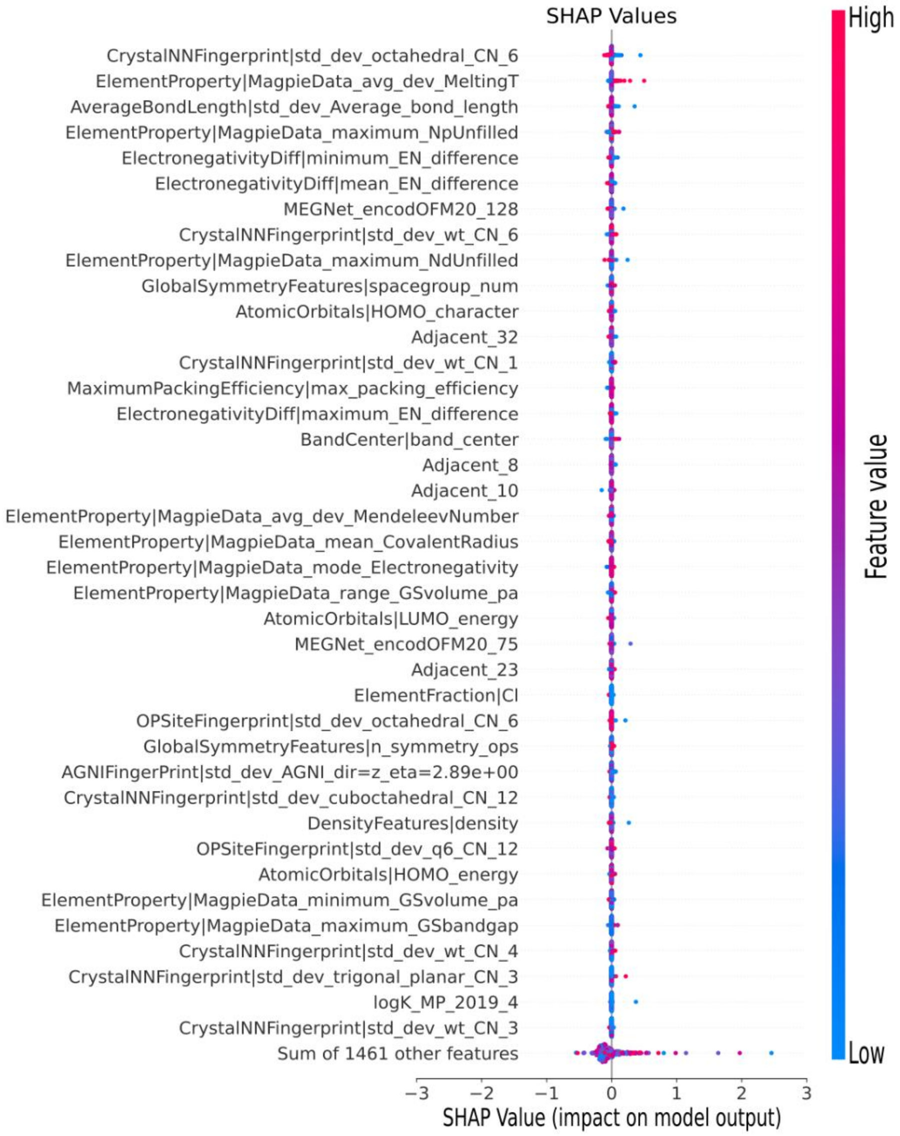


Figure D9 – SHAP analysis plot of the MODNet model with the OMEGA features for the stability task on the OQMD halogen dataset.

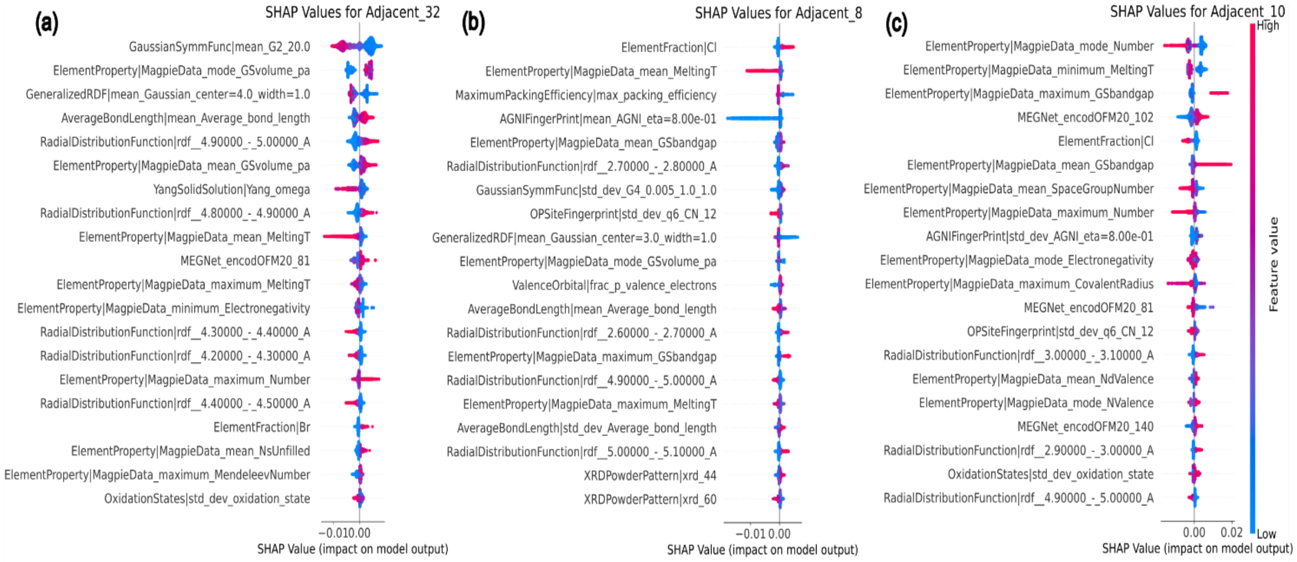


Figure D10 – SHAP analysis plot presenting the decomposition in MatMiner chemical descriptors and ℓ-OFM of the most relevant adjacent model features in the MODNet model with OMEGA features. From left to right, the components for neurons #32 (a), #8 (b) and #10 (c).

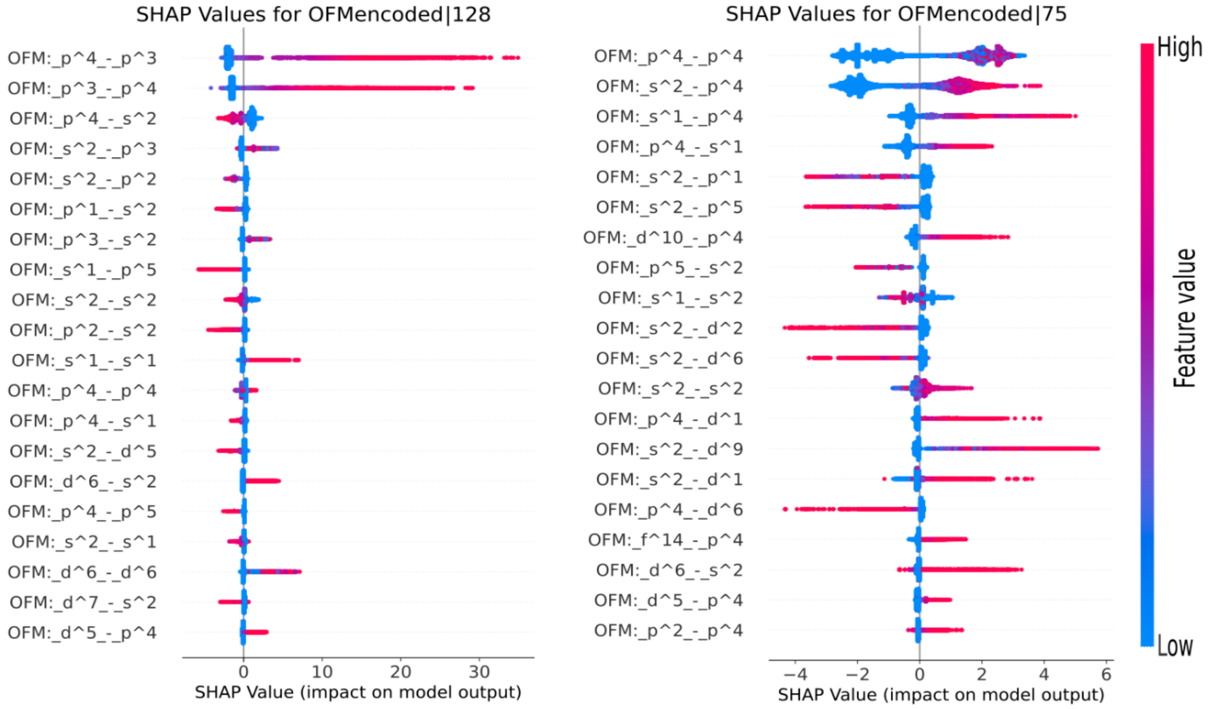


Figure D11 – SHAP analysis plot presenting original OFM contributions to the most relevant ℓ-OFM features, produced with the MEGNet model, in the MODNet model with OMEGA features. On the left (a), for the 128th ℓ-OFM component and, on the right (b), for the 75th ℓ-OFM component.

D.4.4 SHAP analysis of MODNet model with OMEGA features for OQMD halogen band gap task

The most important features of the MODNet model with OMEGA features for the OQMD halogen task on band gap are presented in *Figure D12* on the SHAP analysis plot. It is clear that the pretrained MEGNet model features on band gap regression (*Bandgap_MP_2018*) as well as the adjacent model features dominate the prediction. The pretrained MEGNet band gap features ranking higher in the SHAP plot were correlated to chemical descriptors in *Figure D13*. This procedure was repeated for the adjacent model features within the top features in the SHAP in *Figure D14*.

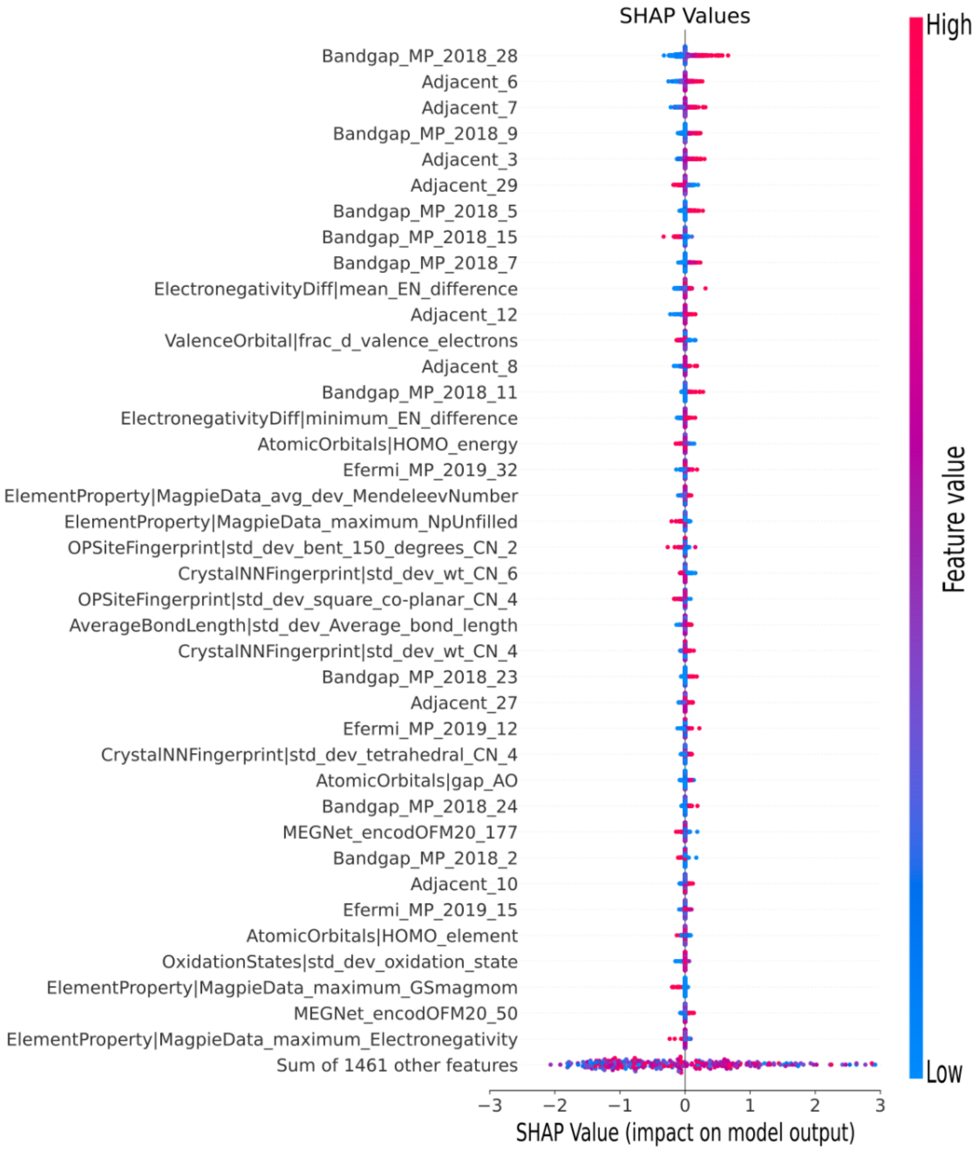


Figure D12 – SHAP analysis plot of the MODNet model with the OMEGA features for the band gap task on the OQMD halogen dataset.

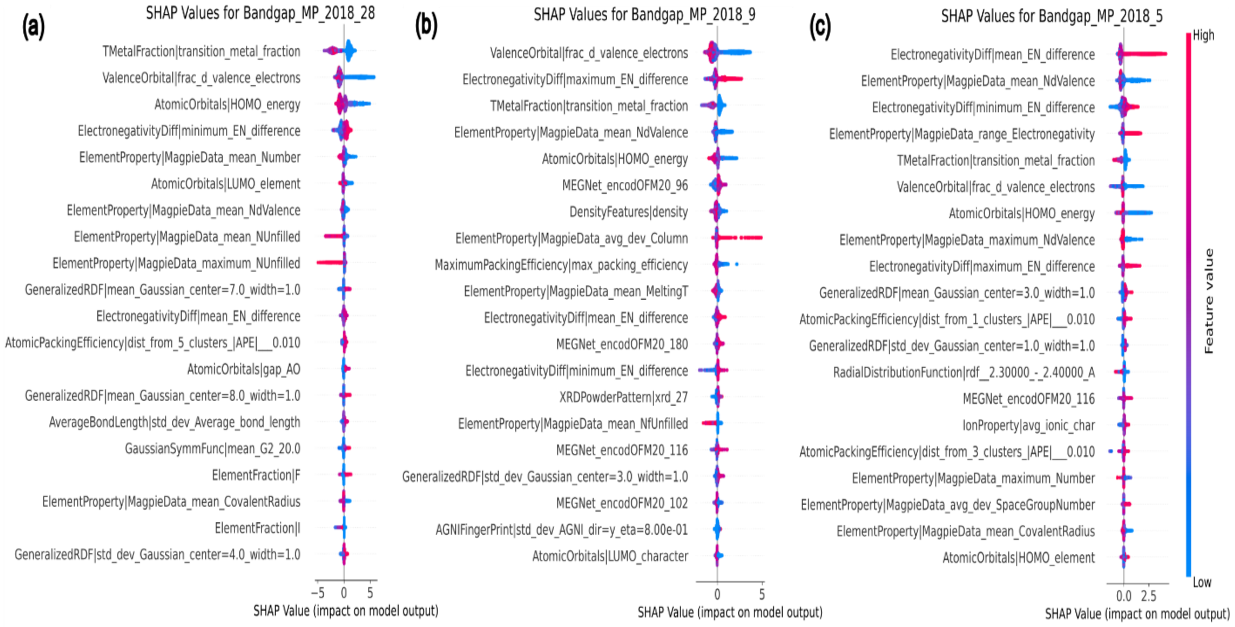


Figure D13 – SHAP analysis plot presenting the decomposition in MatMiner chemical descriptors and ℓ-OFM of the most relevant MEGNetPreL32 features in the MODNet model with OMEGA features for band gap prediction. From left to right, the components for neurons #28 (a), #9 (b) and #5 (c).

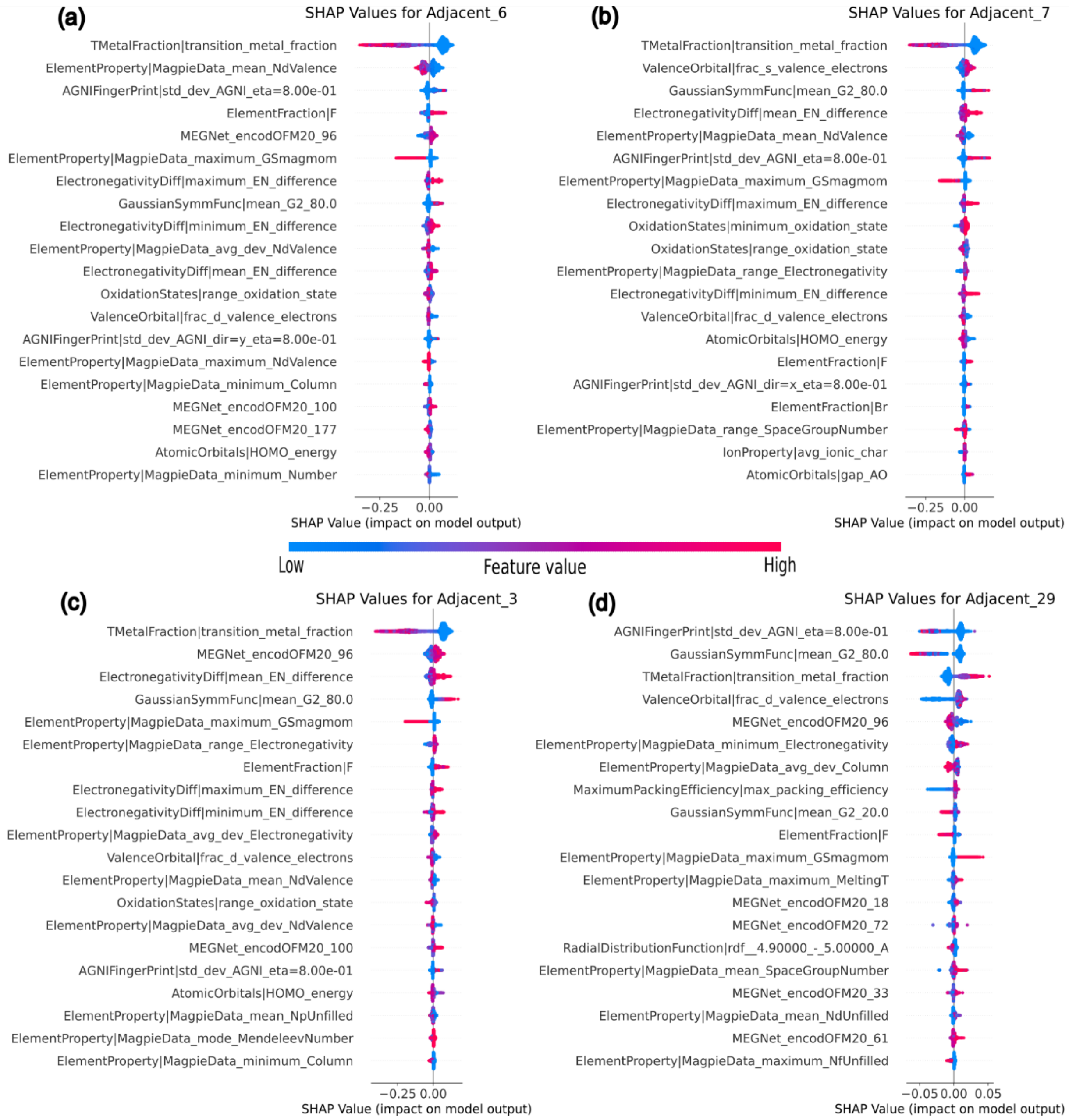


Figure D14 – SHAP analysis plot presenting the decomposition in MatMiner chemical descriptors and ℓ-OFM of the most relevant adjacent model features in the MODNet model with OMEGA features for band gap prediction. The adjacent features include the neurons #6 (a), #7 (b), #3 (c) and #29 (d).

APPENDIX E: Supporting Information for "Machine Learning-Assisted Exploration of 111-Type 2D Perovskite Structures for Photovoltaic and Optoelectronic Applications: A High-Throughput Screening Approach"

E.1 (111)-type perovskites structure generation

To obtain our properly scaled, decorated (111)-type perovskites, we start with the prototype $1 \times 1 \times 2$ supercell structure of $A_3B_2X_9$ ($P\overline{3}m1$). We chose to use 6 permutations for the atoms in each elemental site to guarantee all possible B-site arrangements are included, notably for the case of 2 different species occupying each of the 4 B-sites in the supercell, we have 6 distinct arrangements. This supercell possesses minimal dimensions of a = b = 4 Å and c = 10 Å, ensuring that all nearest neighbor atoms are close together. Following the substitution of elements from *Table E1* onto this prototype supercell, we use the *pymatgen* functions to obtain average cationic and anionic radius of each corresponding ion in the structure, by comparing the distances between ions and the sum of their ionic radius, the unit cell is gradually scaled until all distances are greater than the sum of ionic radius of the involved elements. Finally, the axes are optimized individually, to determine the smallest unit cell which attends the previous condition. This optimization aims to assist CHGNet to provide better estimates for the energies, by avoiding presenting the model geometries too dissimilar from the configurations presented during training.

Table E1 – Digital object identifiers for the works used for the selection of the elements in the initial set.

Chosen elements	Reference DOI
A-site	
K	10.1021/acs.chemmater.7b02013
Rb	10.1021/acs.chemmater.7b02013
Cs	10.1021/acs.chemmater.7b02013
B-site	
Si	10.1021/acs.chemmater.7b02013
Ca	10.1002/solr.202000616
Sc	similar ionic radius to Sb^{3+}, 10.1016/j.jsamd.2024.100700
Ti	10.1039/D2NR02761E
V	10.1039/D2NR02761E
Cr	similar to Mn and Fe, 10.1039/D0RA09270C
Mn	10.1039/D0RA09270C
Fe	ICSD #22074: $Cs_3Fe_2Cl_9$
Co	Similar to Mn and Fe, 10.1002/smtd.202300095
Ni	10.1021/acs.chemmater.7b02013
Cu	10.1002/solr.202000616
Zn	10.1002/solr.202000616
Ga	10.1021/acs.jpcc.7b02221, 10.1002/er.7929
Ge	10.1021/acs.chemmater.7b02013
Se	10.1021/acs.chemmater.7b02013
Sr	10.1002/solr.202000616
B-site	
Y	chemically similar to stabilizer Sc^{3+}, 10.1016/j.jsamd.2024.100700
Zr	10.1002/er.7929
Nb	similar ionic radius to Sb^{3+}, 10.1016/j.jsamd.2024.100700
Mo	similar ionic radius to Sb^{3+}, 10.1016/j.jsamd.2024.100700
Ag	10.1002/solr.202000616, 10.1016/j.jsamd.2024.100700
In	10.1021/acs.jpcc.7b02221, 10.1103/PhysRevApplied.13.024031
Sn	10.1021/acs.chemmater.7b02013
Sb	Well known $A_3Sb_2X_9$ (X=Cl,Br,I)
Te	10.1021/acs.chemmater.7b02013
Ba	10.1002/solr.202000616
Pb	10.1021/acs.chemmater.7b02013
Bi	Well known $A_3Bi_2X_9$ (X=Cl,Br,I)
X-site	
F	10.1007/s12034-023-02890-x
Cl	-
Br	-
I	-
O	10.1016/j.cogsc.2022.100669
S	10.1021/acs.jpcc.6b00920, 10.3390/nano10112284

E.2 Ensemble MODNet model training and evaluation

The base dataset for all our evaluations was based on the Open Quantum Materials Database (OQMD), version v1.5, this dataset was however filtered removing structures whose stability was above the threshold of 2.9 eV/atom, corresponding to 0.1% of the structures. This filtering was shown to improve model generalization in preliminary tests. For every model trained the dataset was divided in training and test set with a 95:5 split, the test set presented similar statistics to the training set (mean, maximum and minimum values) regarding the values of the target property and the training set provided good generalization for prediction on test data. This test set is isolated while the training dataset undergoes k-fold splitting and ensemble MODNet models are trained on each k-fold. We used a k of 5 for all models except for the band gap regressor, which due to the smaller number of samples generalized better with k=10. Finally, these MODNet models are combined forming a deep ensemble model. Evaluation metrics reported throughout the paper are mean absolute error (MAE), coefficient of determination (R^2) for regression tasks of stability and band gap. For the classification model on band gap the evaluation metric was the area under the receiving operator curve (AUCROC). We report validation metrics, evaluated as the mean of the ensemble in each k-fold, and test metrics, evaluated on the isolated test set. In *Table E2*, specific descriptions for each of the considered dataset and models are reported.

Table E2 – Description of Ensemble MODNet models which were trained in this work presenting number of samples in training and test set, as well as k-value for k-fold splitting to train the ensemble MODNet models.

Model name	Dataset description	Number of samples in dataset	k value
General protostructure stability estimator	OQMD dataset* filtered from stability outliers, featurized with InvariantMatMiner2023	1,020,487	5
Active learning M3GNet stability estimator for (111)-type perovskites	M3GNet relaxed (111)-type perovskites with estimated stability on OQMD dataset*, featurized with InvariantMatMiner2023	Variable (15,000 -30,000)	5
Halogen containing stability estimator	OQMD dataset* filtered from stability outliers keeping only halogen-containing structures, featurized with OMEGA+ROSA	30,645	5
Halogen containing band gap estimator	OQMD dataset* filtered from stability outliers keeping only halogen-containing structures, featurized with OMEGA+ROSA	8,347	10
Halogen containing band gap classifier	OQMD dataset* filtered from stability outliers keeping only halogen-containing structures, featurized with OMEGA+ROSA	30,645	5

*OQMD dataset includes in-group data with additional 36 datapoints on (111)-type antimony perovskites.

E.3 OMEGA + ROSA MODNet featurizer

Since the prediction of stability is a challenging task for ML models, we employed a more encompassing featurizer which could make use of GNN models flexibility, and descriptors derived from a one-shot ab-initio calculation. This featurizer utilizes three sets of features, namely:

1. General geometric, electronic, and chemical descriptors: these features all derive from the default MODNet featurizer, MatMiner2023.
2. GNN/latent-space electronic descriptors: we developed a set of descriptors all based on GNN models to harness their flexibility and fast calculation. These are included on top of *MatMiner2023* features in what we called *OMEGA* featurizer standing for "encoded OFM + pre-trained MEGNet + Adjacent MEGNet models". This featurizer can significantly improve MODNet predictions in most tasks and its detailed derivation is provided elsewhere (see *Chapter 5*).
3. Ab-initio derived descriptors: inspired by the work of Tawfik and Russo (2022) we included in our featurizer their Robust One-Shot Ab-initio (ROSA) features which contain computations of eigenvalues and energies for the structure through DFT but from a simple initialization of linear combination of atomic orbitals (LCAO). These features can be easily computed with a script provided by the authors which uses the open-source DFT code GPAW to calculate the one-shot quantities (Enkovaara et al. 2011).

We also included some geometrical functions from the work of Tawfik and Russo (2022), namely the *G* symmetry functions, including a total of about 600 features in the model. The final dataset after the application of the OMEGA + ROSA featurizer includes over 5000 features; to perform feature selection we employ XGBoost to reduce the set of features to 1500 with its fast assessment of feature importance and subsequently, we employ MODNet's advanced feature selection algorithm to determine the final set of features to be utilized by the model.

E.4 Detailed workflow for our machine-learning screening method

One-shot M3GNet screening:

The first step in the workflow was to run each of the decorated prototype (111)-type perovskite structures through CHGNet (v. 0.3.3) and evaluate their total energy; this is a very fast computation because force computation is not required. We reduce to a single optimal structure for each composition based on the assumption that, although they are not in their equilibrium positions, CHGNet should be able to distinguish the most advantageous atomic arrangements for each structure. The 100,627,800 structures are then reduced to 470,610 structures with unique compositions, this forms the global pool of candidate structures for subsequent screening in the active learning cycle.

Protostructure ML screening:

The subsequent step involves creating descriptors for the entire global pool of structures which is done with a modification of the default pre-defined MatMiner2023 featurizer MODNet v0.4.1. Particularly, the composition-only features are used in their original form while the structure-based featurizer is reduced to only features invariant under relaxation by considering only bond fractions and essential statistics related to coordination number, geometrically determined through nearest neighbor analysis. This approach allows for comparison between our unrelaxed decorated structures and fully relaxed structures present in materials databases since the model effectively considers only structure prototypes. This featurizer is referred as *InvariantMatMiner2023* throughout this work. With the structures properly described, we evaluate them with an ensemble MODNet model trained on the outlier filtered OQMD dataset for stability prediction. Detailed description of the datasets and training is provided on *Appendix E.2*.

After evaluating the decorated (111)-type perovskites with the model, a threshold of 35 meV/atom in decomposition energy is established to filter unlikely compositions. This threshold takes into account the estimated decomposition of various structures reported experimentally and is presented in *Table E3*. Additionally, upon observing the table, it becomes evident that the inclusion of our in-group data (Gouvêa et al. 2024; Exner et al. 2024) on alloyed (111)-perovskites (36 data points) significantly improves predictions, especially for halogen-alloyed perovskites. Henceforth, when mentioning the OQMD dataset, it refers to the dataset filtered for stability outliers that includes our in-group data.

Since the MODNet models employed are ensembles, they can easily provide uncertainty estimates for predictions. Utilizing the predicted stability and uncertainty, we establish the upper limit for the decomposition energy of each structure. Subsequently, we sort the structures by their lower decomposition energy upper limit. Finally, we extract 15,000 samples from this sorted dataset, with 70% (10,500) presenting the lowest E_{stab} upper limits, and the remaining 30% (4,500) comprising structures that optimize the acquisition function. This set of structures forms the initial pool for the active learning cycle. Entropy is normalized and averaged across all features in a sample in the pool in relation to the samples with lowest E_{stab} included. Entropy and uncertainty are then normalized according to the range of values in the pool, and their product results in the final acquisition value for each sample. Therefore, entropy and uncertainty were given the same weight in the acquisition function.

Table E3 – Estimated stability from considered protostructure-based MODNet model in experimentally reported (111)-type perovskites.

ICSD reported structure	Protostructure stability estimator, Trained on original OQMD + our group data*			Protostructure stability estimator, trained on original OQMD		
	prediction	std	upper limit (prediction+std)	prediction	std	upper limit (prediction+std)
$Cs_3Sb_2I_9$ (ICSD: #39822)	-10.8	14.7	3.8	-8.5	11.4	2.9
$Cs_3Sb_2Br_9$ (ICSD: #39824)	-17.7	13.5	-4.2	-18.6	17.1	-1.5
$Cs_3Sb_2Cl_9$ (ICSD: #22075)	-22.5	19.5	-3.0	-21.8	22.4	0.6
$Cs_3Sb_2BrCl_8$ (ref: §1)	7.6	6.6	14.2	17.3	6.5	23.8
$Cs_3Sb_2Br_2Cl_7$ (ref: §1)	5.7	4.3	10.0	7.3	4.6	11.9
$Cs_3Sb_2(BrCl_2)_3$ (ref: §1)	4.4	3.9	8.3	4.6	7.1	11.7
$Cs_3Fe_2Cl_9$ (ICSD: #22074)	1.6	4.4	6.0	-17.5	3.1	-14.5
$Rb_3Sb_2Br_9$ (ICSD: #39823)	-17.7	13.5	-4.2	-6.1	6.7	0.5
$Cs_3Bi_2Br_9$ (ICSD: #1142)	-21.2	7.4	-13.9	-7.6	17.2	9.6

(ref: §1) - (A. Pradhan, Jena, and Samal 2022)

Active learning (AL) cycle:

The initial pool of structures for AL undergo relaxation with CHGNet imposing the criteria for relaxation of maximum force of 0.00001 eV/Å and maximum number of steps 1000. This procedure determines the optimal volume and lattice parameters before ionic relaxation and is constrained to small steps to avoid instability. The decomposition energy or stability (E_{stab}) can be determined from the total energy per atom of these relaxed structures along with their composition. By deducting the reference energy of the constituent atoms from the total energy (see *Chapter 2*, *equation 11*) the formation energy of the compound is obtained. The convex hull compounds corresponding to each structure composition is obtained from OQMD database and the grand canonical linear programming (GCLP) method (Saal et al.

2013; Kirklin et al. 2015). Finally, E_{stab}^{CHGNet} an estimate of the stability using the CHGNet relaxed structure energy is obtained. Notice the different notation to underscore that the total energy of the candidate structure was obtained with the MLIP and may be subject to diverse biases on the training data and will be only an approximation to the total energy explicitly obtained with DFT (E_{stab}).

Once E_{stab}^{CHGNet} is obtained for all structures in the initial pool, an ensemble MODNet model is trained with the *InvariantMatMiner2023* featurizer to predict the MLIP stability. This model guides the selection of a new set of structures from the global pool of candidate structures, which had already been featurized with *InvariantMatMiner2023*. Based on the same criteria applied to obtain the initial pool for AL, structures are excluded if estimated decomposition energy is above 35 meV/atom and they are then selected based on the lowest upper limit of the decomposition energy (70% of the total) and those structures that maximize the acquisition function (30% of the total). We applied a step of 2,500 structures added for each AL cycle. The size of the initial dataset (15,000) and the increment (2,500) in each AL cycle were defined arbitrarily aiming for a compromise on improving model generalization and sensible number of structures for the cost of fully relaxing structures, evaluating E_{stab}^{CHGNet} and featurizing each structure for the ML models.

We noticed later however that the initial dataset was quite extensive, and the model could generalize well with about 9,000 structures, this was verified comparing the evaluation metrics of the original model and a model starting with 3,000 structures model, these results are shown in *Table E4*. The AL cycle ends when the predicted most stable structures remaining in the pool present E_{stab}^{CHGNet} > 45 meV/atom. In our case, this led to 6 active learning steps, resulting in a total set of 30,000 structures. All structures in the set with E_{stab}^{CHGNet} < 45 meV/atom form the final AL pool of candidate materials.

Table E4 – Mean absolute error for the test set for protostructure MODNet models implemented during the active learning cycle for E_{stab}. Models are identified as "production" for the models actually used for the active learning screening, and "testing" for the test on the effect of the number of samples.

Identification of the model	*Number of samples*	*MAE (meV/atom)*
Testing	3,000	25.0
Testing	6,000	18.2
Testing	9,000	16.5
Testing	12,000	15.1
Production	15,000	14.9
Production	17,500	15.1
Production	20,000	16.1
Production	22,500	17.8
Production	25,000	14.9
Production	27,500	15.5
Production	30,000	14.3

Structure ML screening:

The final AL pool of candidate materials are now described with a more complex and demanding descriptor, MODNet's *OMEGA+ROSA* featurizer which is fully detailed on *Appendix E.3*. The halogen-containing structures from OQMD are also featurized with *OMEGA+ROSA* featurizer and the ensemble MODNet models are trained for stability and band gap classification, where a band gap < 0.5 eV is considered metallic and above 0.5 eV considered semiconductor. The use of 0.5 eV band gap instead of 0 eV to classify the structures introduced a bias that improved the ability of the model to differentiate semiconductors in our tests. Utilizing again the structures with band gap > 0.5 eV a new subset of structures is produced to train for band gap regression. Detailed information on the model training and datasets is provided on *Appendix E.2*. The use of the OMEGA+ROSA features is justified when the evaluation metrics are compared to employing the default *MatMiner2023* featurizer or the *OMEGA* featurizer to generate descriptors for MODNet, as shown in *Table E5*. After applying the models on the final AL pool of candidate materials, we select those with predicted decomposition energy below 35 meV/atom, materials classified as semiconductors, and a band gap less than 3.5 eV. In the end, our machine learning screening identifies 4336 structures that meet these requirements.

Table E5 – Evaluation metrics for structure-based MODNet models with different featurizers. N represents the number of samples provided for the task.

Featurizer	**Stability** (N=30,645)		**Band gap classifier** (N=30,645)		**Band gap estimator** (N=8,347)	
	MAE (meV/atom)		**AUCROC**		**MAE (eV)**	
	Validation	*Test*	*Validation*	*Test*	*Validation*	*Test*
Default MatMiner2023	30.1	55.7	0.799	0.714	0.218	0.472
OMEGA	28.8	52.5	0.904	0.764	0.198	0.404
OMEGA+ROSA	28.1	49.6	0.866	0.768	0.177	0.376

ML phonon screening:

The structures which passed our criteria of stability and band gap are checked now with a dynamical stability estimate by a phonon band structure calculation through CHGNet relaxation (3000 relaxation steps or maximum force below 0.0001 eV/Å) and frozen phonon method. We defined a threshold to filter structures which negative phonon frequency (ω_{min}) was below -0.35 THz since a more negative frequency is usually associated with more unstable structures. This is based on general observation of experimentally reported structures present in our dataset and groups of structures with multiple substitutions which show a consistent pattern as presented on *Table E6*. Applying these criteria, we could reduce the dataset from 4336 to 2991 structures.

Table E6 – Average minimum phonon frequency for ML screened structures considering different group of structures.

Group of ML screened structures	Average ω_{min}
Experimentally reported	-0.205
3 distinct elements	-0.255
4 distinct elements	-0.272
5 distinct elements	-0.290
Oxygen-containing	-0.506

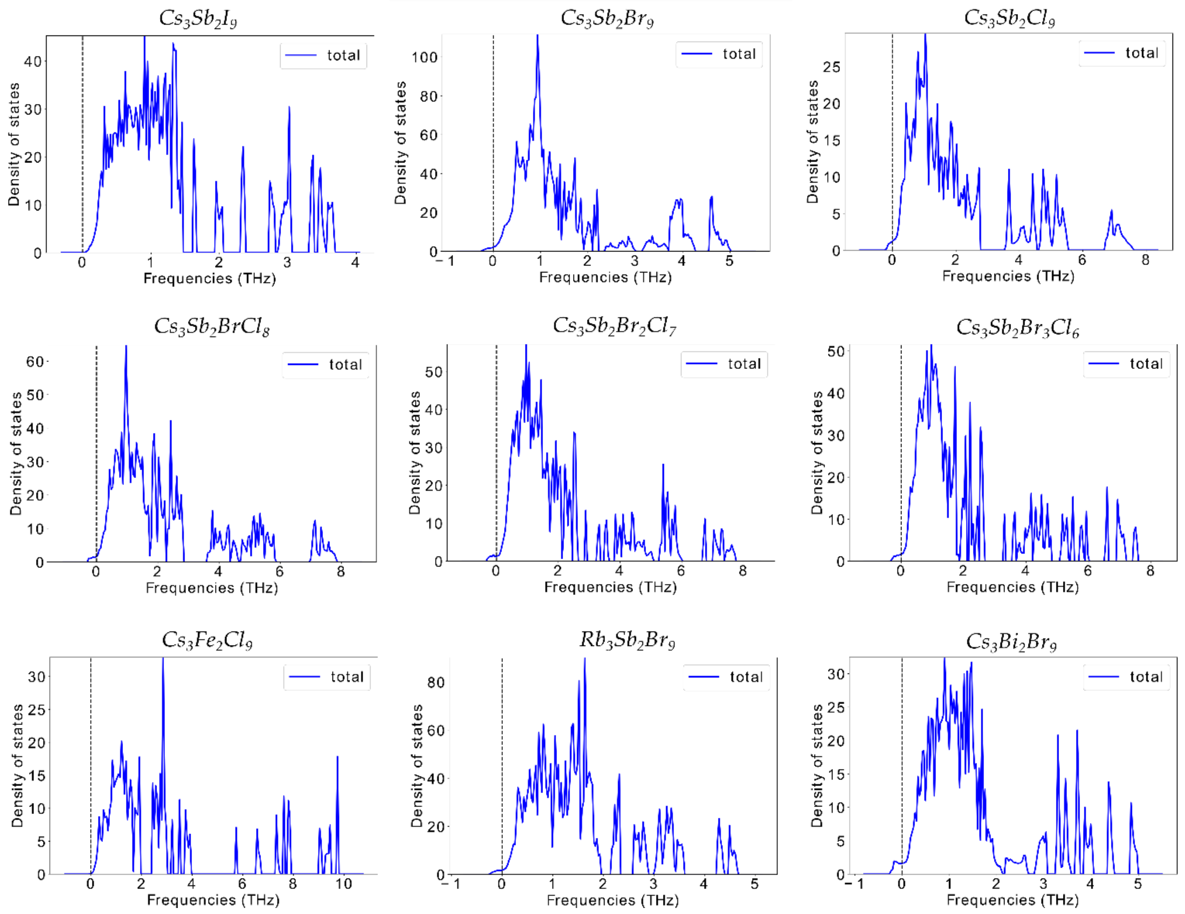


Figure E1 – Phonon density of states calculated with CHGNet for the set of experimentally reported (111)-type perovskite structures considered in this work.

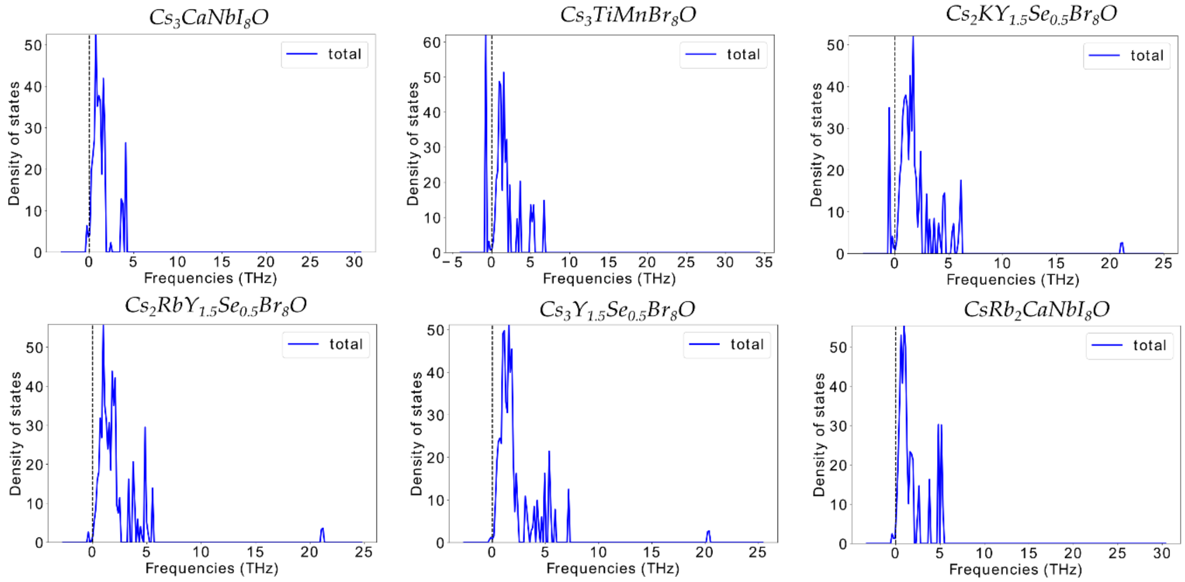


Figure E2 – Phonon density of states calculated with CHGNet for O-containing (111)-type perovskite structures screened by the ML models within thresholds considered in this work.

E.5 SHAP analysis of structure-based MODNet models

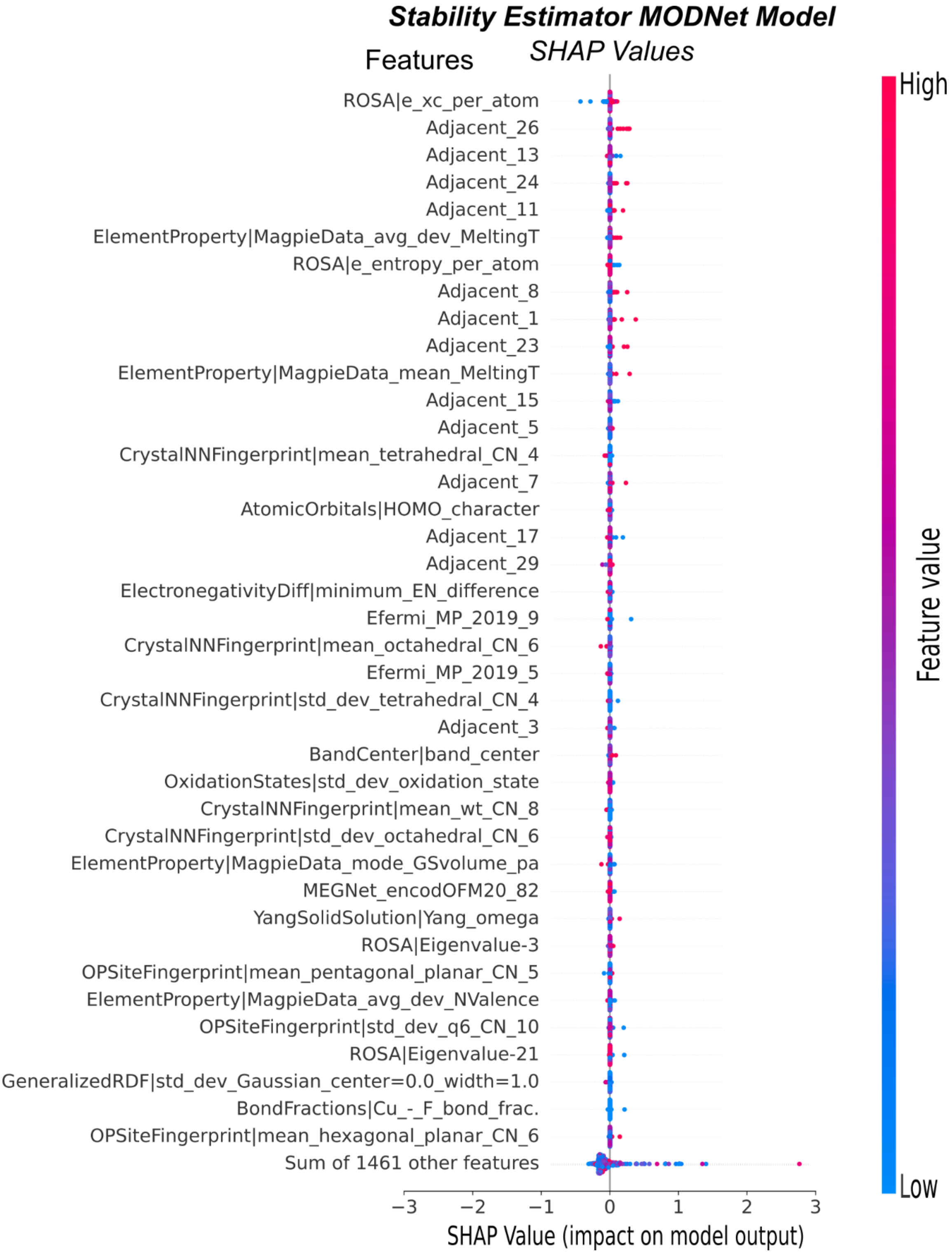


Figure E3 – SHAP analysis plot of the structure-based MODNet model for the stability task trained on the OQMD halogen-containing dataset.

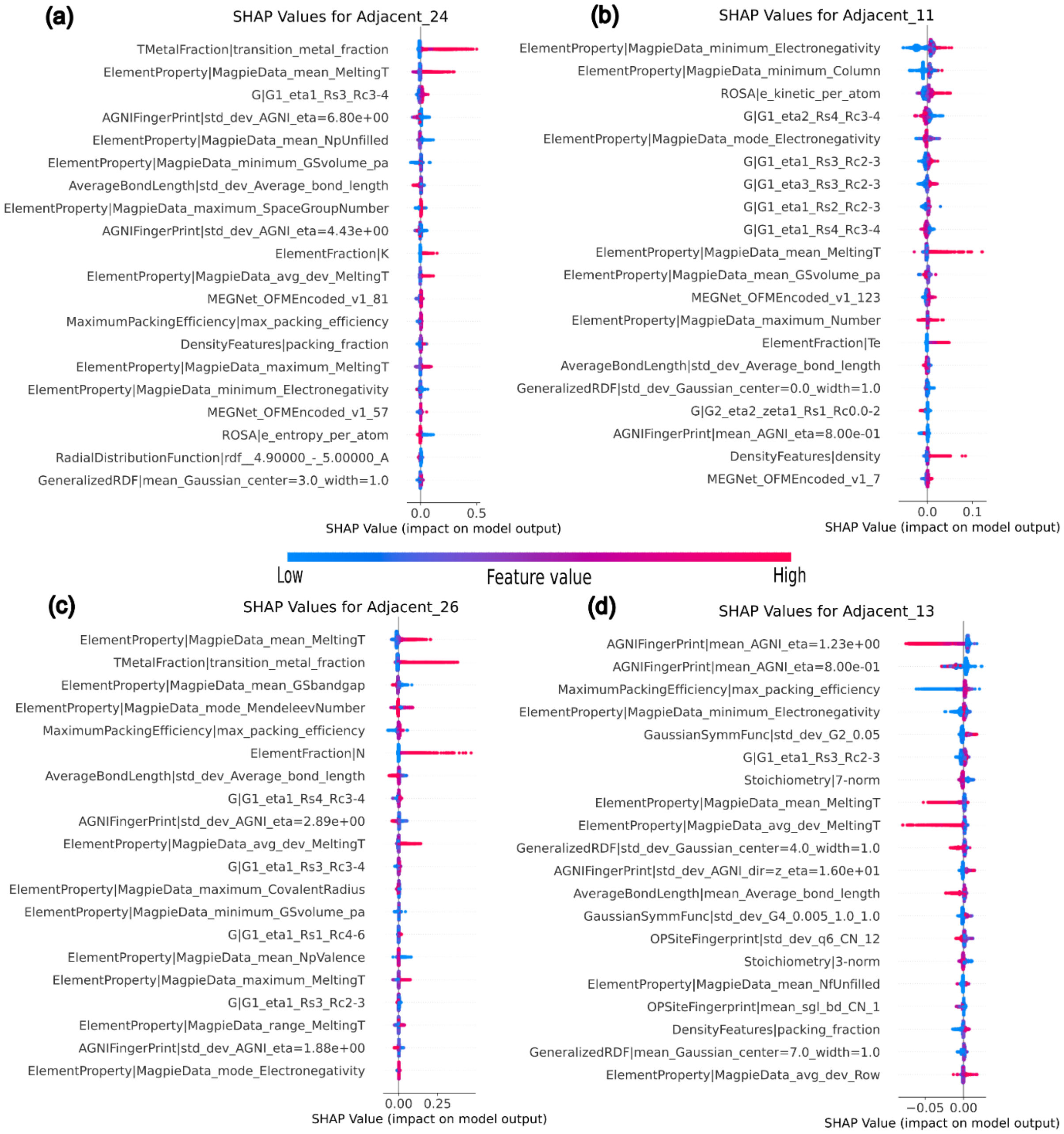


Figure E4 – SHAP analysis plot presenting the decomposition in interpretable chemical and geometrical descriptors of the most relevant GNN adjacent model features in the structure-based MODNet model for stability prediction. The adjacent features include the neurons #24 (a), #11 (b), #26 (c) and #13 (d).

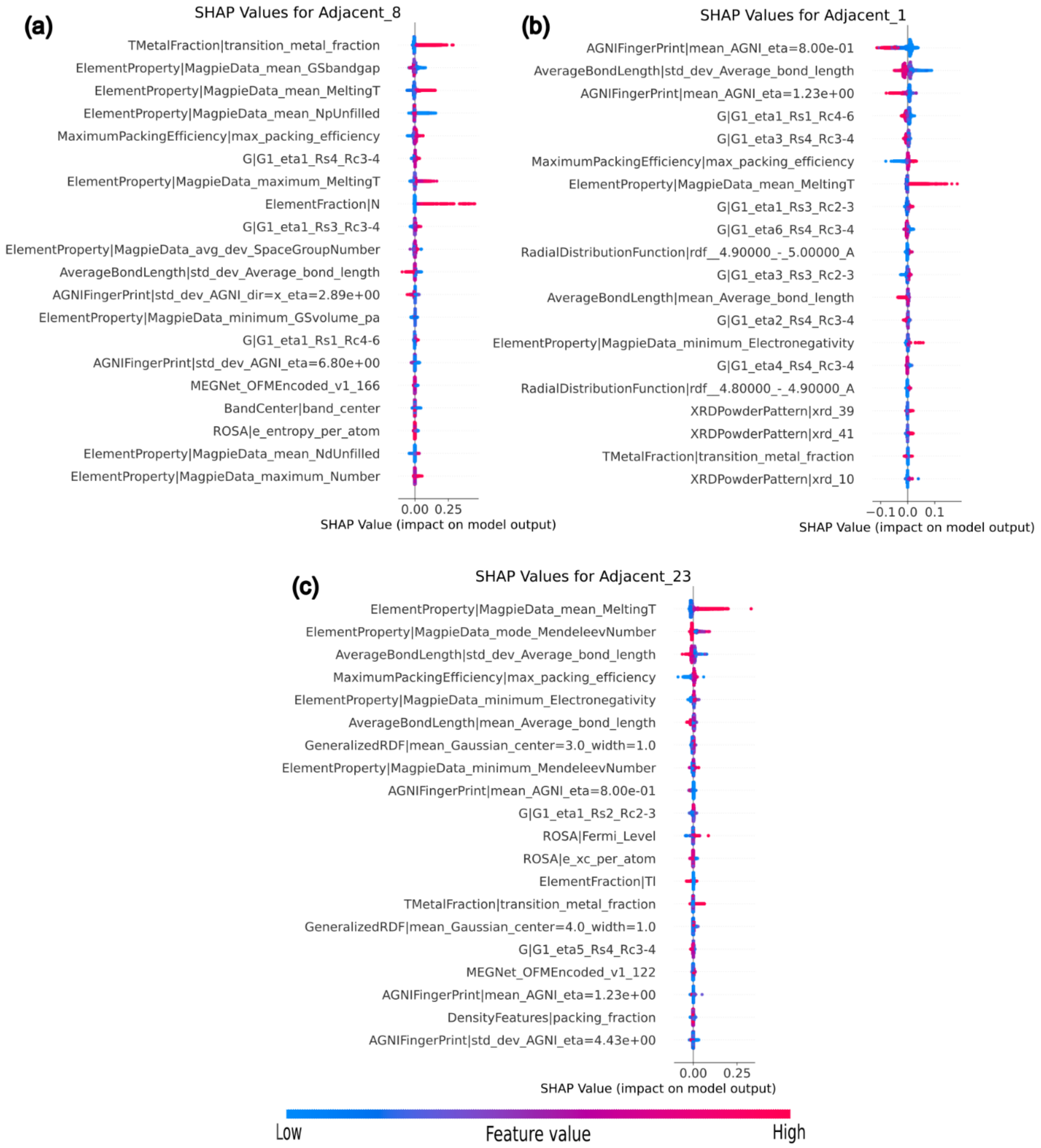


Figure E5 – SHAP analysis plot presenting the decomposition in interpretable chemical and geometrical descriptors of the most relevant adjacent GNN model features in the structure-based MODNet. The features include the adjacent GNN model neurons #8 (a), #1 (b) and #23 (c).

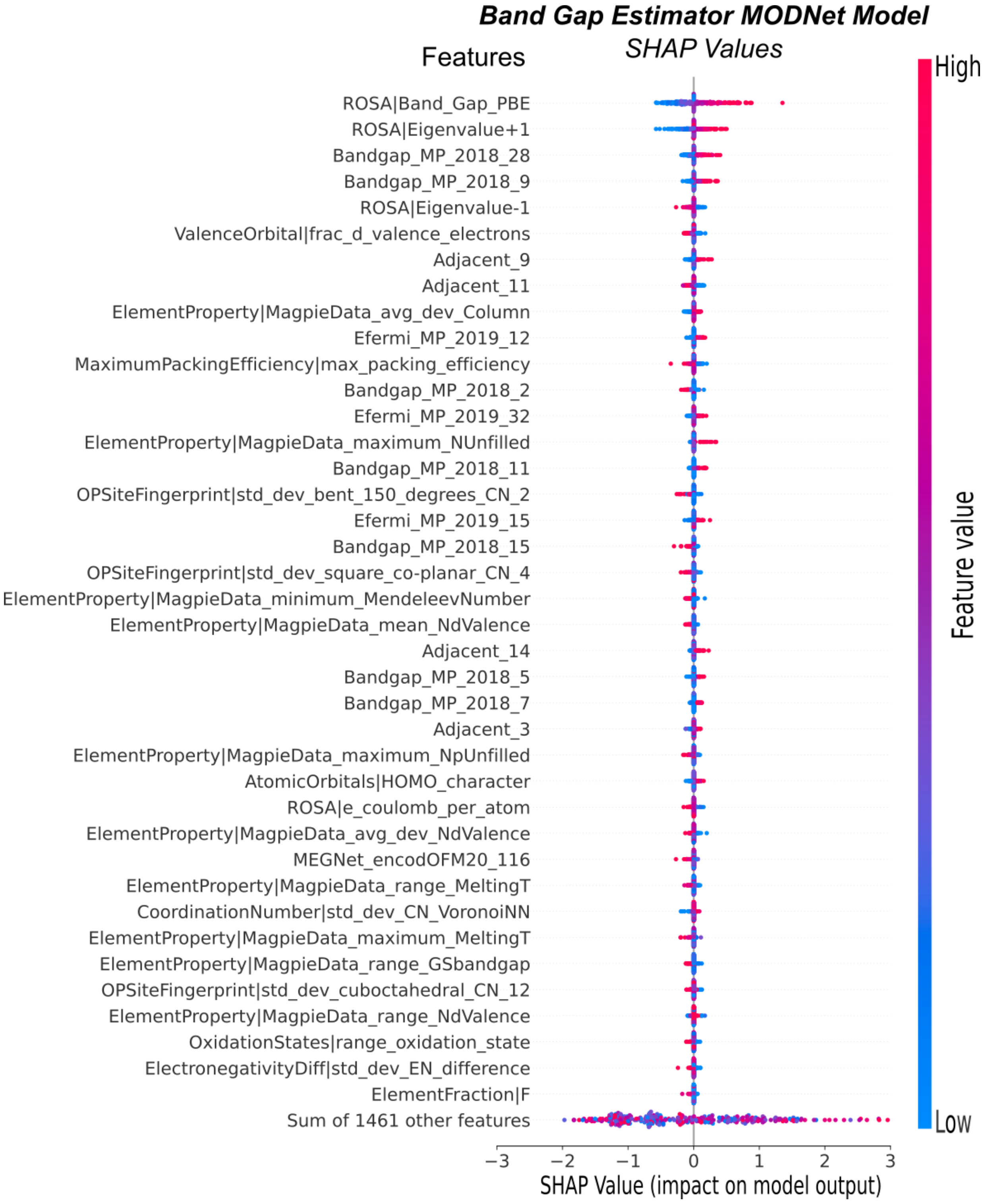


Figure E6 – SHAP analysis plot of the structure-based MODNet model for the band gap task trained on the OQMD halogen-containing dataset.

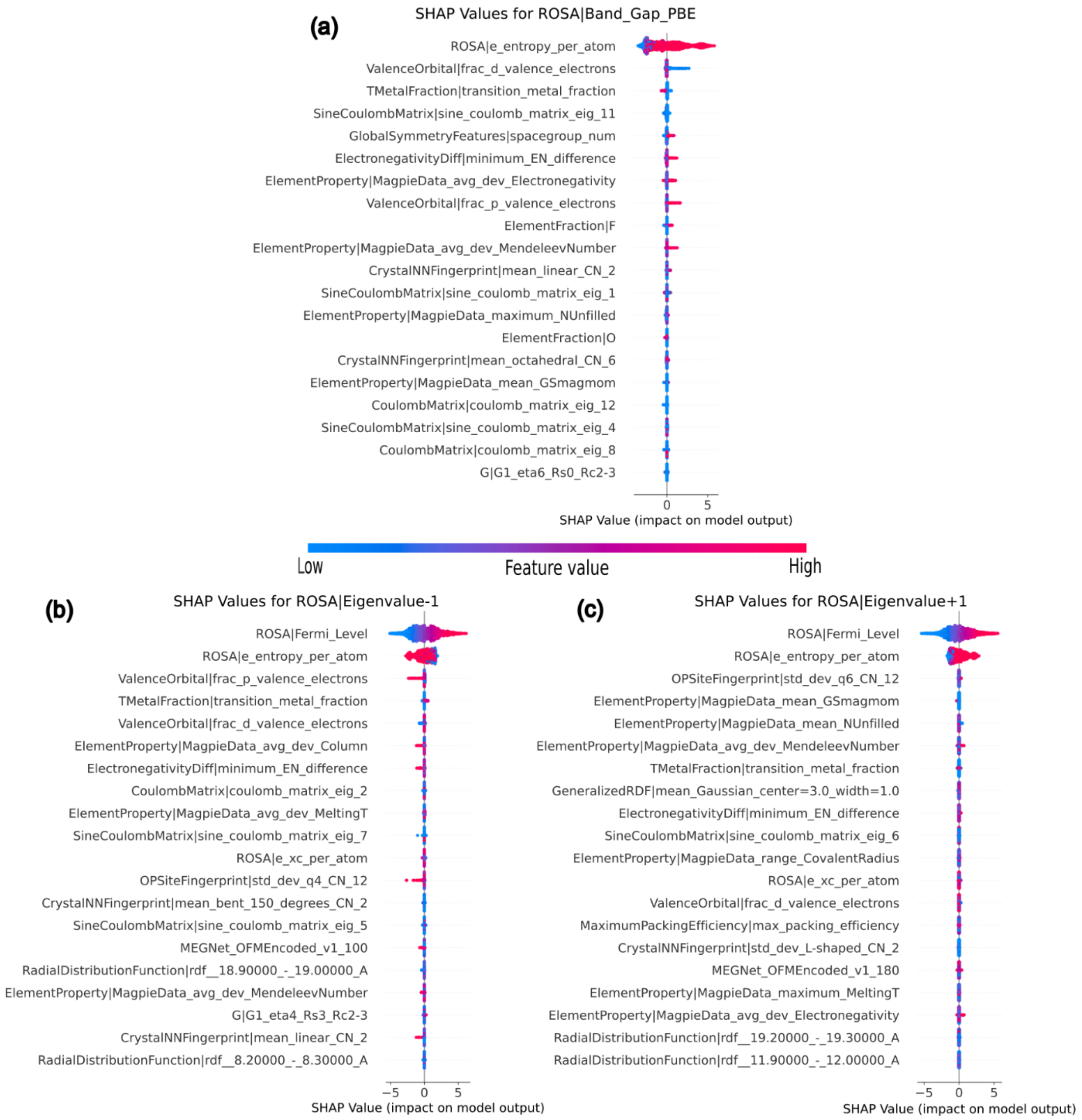


Figure E7 – SHAP analysis plot presenting the decomposition in interpretable chemical and geometrical descriptors of the most important ROSA features in the structure-based MODNet model with OMEGA+ROSA features for band gap estimation. The features decomposed are ROSA's (a) PBE band gap, (b) Eigenvalue -1 and (c) Eigenvalue +1.

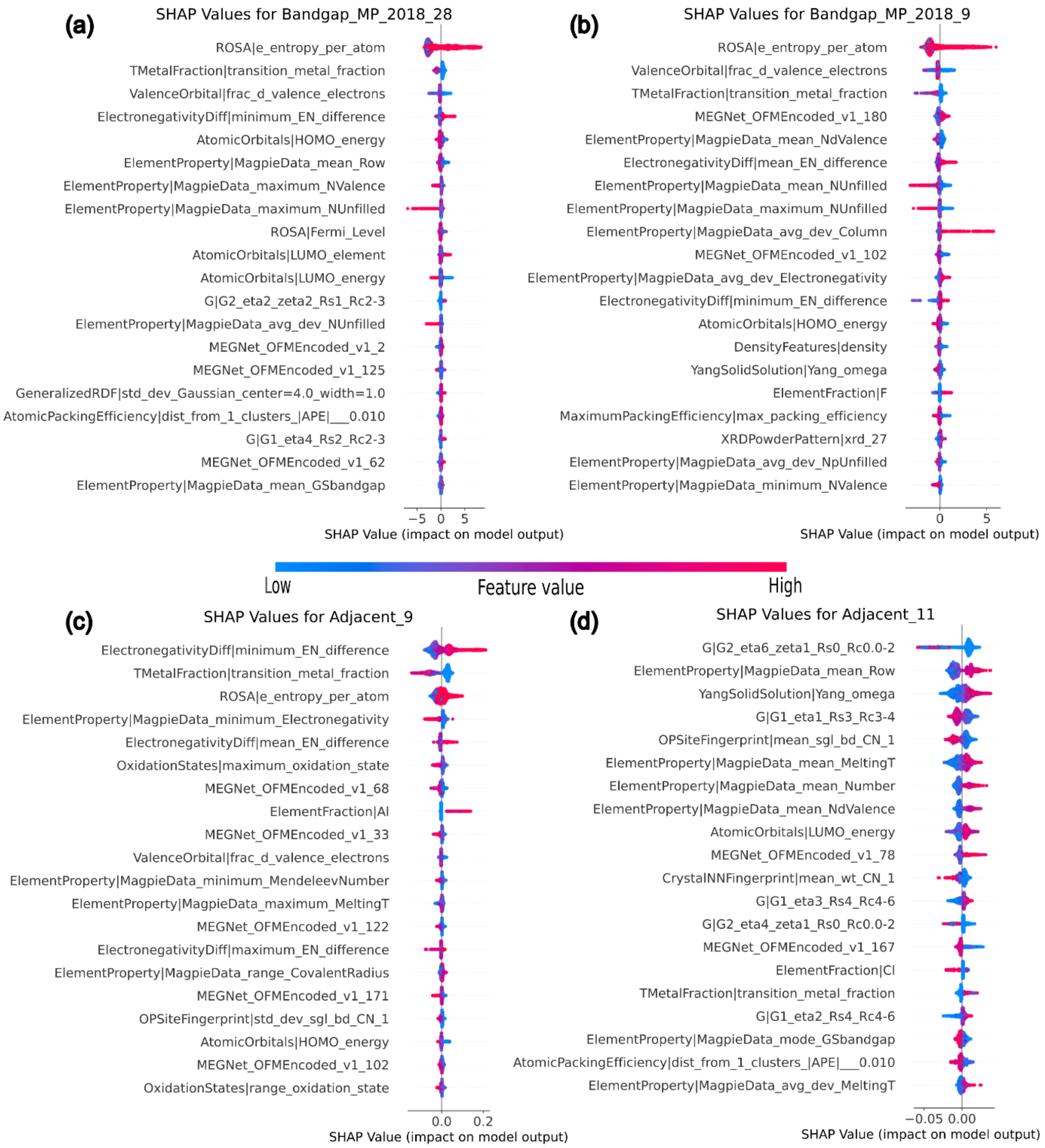
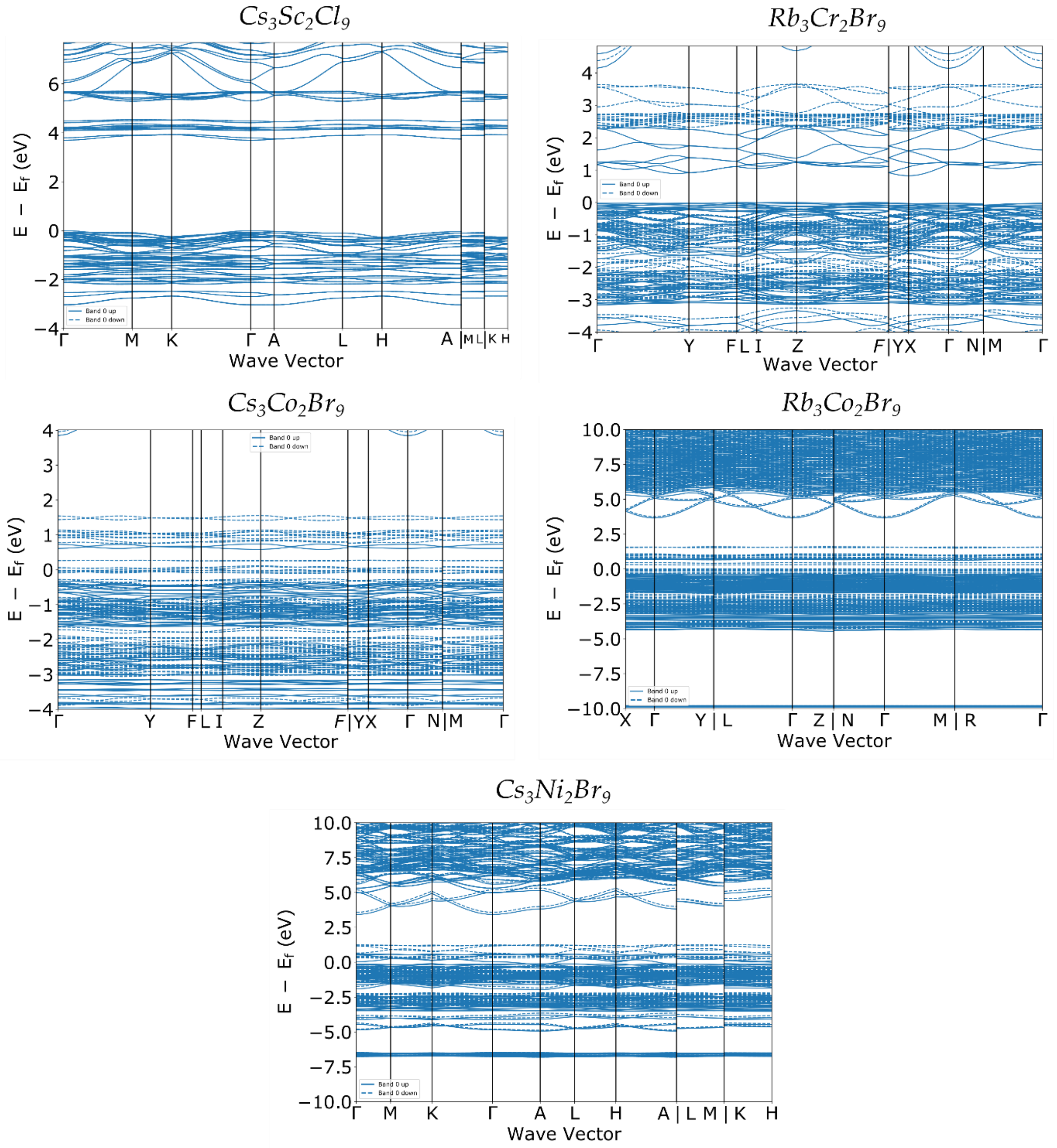


Figure E8 – SHAP analysis plot presenting the decomposition of interpretable chemical and geometrical descriptors for the most important GNN features in the structure-based MODNet model with OMEGA+ROSA features for band gap estimation. The decomposed features correspond to neurons #29 (a) and #9 (b) from the pre-trained band gap model, and neurons #9 (c) and #11 (d) from the adjacent GNN model.

E.6 Complementary data for ab-initio evaluation of screened (111)-type perovskites



$Cs_3Sc_2Cl_9$

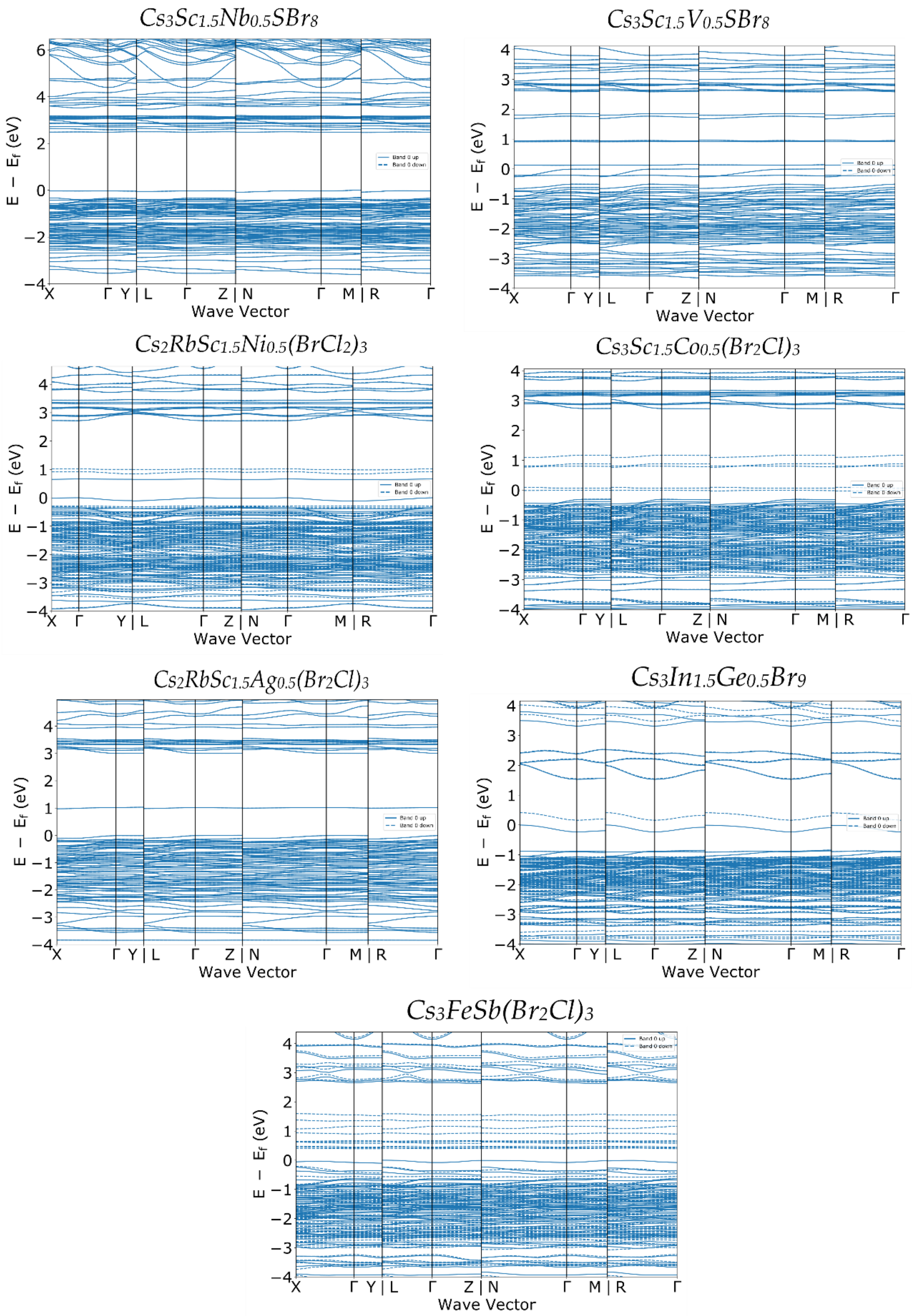
$Rb_3Cr_2Br_9$

$Cs_3Co_2Br_9$

$Rb_3Co_2Br_9$

$Cs_3Ni_2Br_9$

Figure E9 – Band structures for selected ternary (111)-type perovskites passing the ML screening phase.



$Cs_3Sc_{1.5}Nb_{0.5}SBr_8$

$Cs_3Sc_{1.5}V_{0.5}SBr_8$

$Cs_2RbSc_{1.5}Ni_{0.5}(BrCl_2)_3$

$Cs_3Sc_{1.5}Co_{0.5}(Br_2Cl)_3$

$Cs_2RbSc_{1.5}Ag_{0.5}(Br_2Cl)_3$

$Cs_3In_{1.5}Ge_{0.5}Br_9$

$Cs_3FeSb(Br_2Cl)_3$

Figure E10 – Band structures for selected (111)-type perovskites passing the ML screening phase.