Advanced Intelligent Discovery

ADVANCED INTELLIGENT DISCOVERY Open Access

RESEARCH ARTICLE OPEN ACCESS

Why Physics Still Matters: Improving Machine Learning Prediction of Material Properties With Phonon-Informed Datasets

Pol Benítez[1,2] | Cibrán López[1,2] | Edgardo Saucedo[2,3] | Teruyasu Mizoguchi[4] | Claudio Cazorla[1,2,5]

[1]Department of Physics, Universitat Politècnica de Catalunya, Barcelona, Spain | [2]Research Center in Multiscale Science and Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain | [3]Department of Electronic Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain | [4]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan | [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Correspondence: Pol Benítez (pol.benitez@upc.edu) | Claudio Cazorla (claudio.cazorla@upc.edu)

Received: 19 November 2025 | **Revised:** 29 January 2026 | **Accepted:** 9 March 2026

Keywords: energy materials | graph neural networks | materials informatics | model explainability | physics-informed machine learning

ABSTRACT

Machine learning (ML) methods have become powerful tools for predicting material properties with near first-principles accuracy and vastly reduced computational cost. However, the performance of ML models critically depends on the quality, size, and diversity of the training dataset. In materials science, this dependence is particularly important for learning from low-symmetry atomistic configurations that capture thermal excitations, structural defects, and chemical disorder, features that are ubiquitous in real materials but underrepresented in most datasets. The absence of systematic strategies for generating representative training data may therefore limit the predictive power of ML models in technologically critical fields such as energy conversion and photonics. In this work, we assess the effectiveness of graph neural network (GNN) models trained on two fundamentally different types of datasets: one composed of randomly generated atomic configurations and another constructed using physically informed sampling based on lattice vibrations. As a case study, we address the challenging task of predicting electronic and mechanical properties of a prototypical family of optoelectronic materials under realistic finite-temperature conditions. We find that the phonons-informed model consistently outperforms the randomly trained counterpart, despite relying on fewer data points. Explainability analyses further reveal that high-performing models assign greater weight to chemically meaningful bonds that control property variations, underscoring the importance of physically guided data generation. Overall, this work demonstrates that larger datasets do not necessarily yield better GNN predictive models and introduces a simple and general strategy for efficiently constructing high-quality training data in materials informatics.

1 | Introduction

Over the past decade, machine learning (ML) methods have transformed nearly every field of engineering and natural and social sciences. The exponential increase in available data, combined with advances in computational power and algorithmic innovation, has allowed ML models to achieve remarkable successes in domains once considered uniquely human or prohibitively complex. From natural language processing, exemplified by large language models [1] capable of generating coherent text, reasoning about problems, and accelerating scientific discovery, to breakthroughs in biology such as DeepMind's AlphaFold [2], which revolutionized protein structure prediction, ML has demonstrated an unprecedented ability to learn from data and extract patterns beyond human intuition.

In the context of materials science, ML developments have been equally transformative [3]. The field has traditionally relied on a combination of experimental synthesis, characterization, and

theoretical modeling to understand and design materials. However, these approaches can be time and resources intensive, especially when exploring vast compositional and structural spaces. ML offers a complementary paradigm, one in which predictive models can rapidly screen candidate materials [4], infer hidden correlations between compositions, structure, and properties [5, 6], and guide experiments toward promising regions of compositional space [7, 8]. As a result, ML has become a central component of the emerging discipline of materials informatics, enabling accelerated discovery and optimization across applications such as catalysis, energy storage, and structural alloys.

Yet, despite the growing success of ML techniques in materials discovery, the accuracy and reliability of ML models ultimately depend on the quality, diversity, and physical representativeness of the underlying dataset [9]. In practice, materials datasets are often scarce, noisy, or biased toward well-studied chemical systems, resulting in models that generalize poorly when extrapolated to unexplored compositions and structures. Consequently, dataset curation has become as critical as model architecture itself, and balancing data quantity with data quality remains a central challenge in materials informatics.

Many ML approaches implicitly assume that enlarging the dataset will systematically improve model accuracy [10]. However, producing large-scale datasets through high-fidelity first-principles simulations can be computationally prohibitive, often surpassing the cost of the calculations that these models aim to replace. This difficulty raises a key question: how can one achieve reliable predictions from limited high-quality data? Addressing this problem requires careful assessment of model convergence with respect to dataset size and an understanding that well-designed and diversified datasets in principle may outperform much larger scattered ones [11, 12].

Actually, to overcome data scarcity and improve generalization, physics-informed ML has emerged as a promising paradigm [13, 14]. By embedding fundamental physical principles (e.g., symmetry constraints, conservation laws, and thermodynamic relations) into the learning process, such approaches enforce physical consistency, enhance interpretability, and reduce the need for extensive data. Incorporating domain knowledge in this way allows ML models to learn not only from data correlations but also from the governing laws of nature, bridging the gap between statistical learning and physical insight.

One domain where these considerations are particularly pressing is the study of thermal behavior of functional materials [15, 16]. Thermal lattice vibrations, or phonons, strongly influence electronic and mechanical properties by stabilizing specific crystal phases at elevated temperatures [17] or modifying the band structure and band gap of crystals [18]. Capturing such effects with first-principles methods typically requires computationally demanding finite-temperature simulations, which severely limit the size of available datasets. Consequently, there is a strong motivation to develop ML models that can serve as accurate surrogates for these high-cost simulations, provided the training data faithfully represent thermally disordered atomic configurations.

In this work, we explore how fundamental properties of technologically relevant materials, specifically, antiperovskites used in photovoltaics, electrochemistry, and catalysis [19–23], evolve with temperature using graph neural networks (GNNs). We systematically compare GNN models trained on two different types of low-symmetry, non-equilibrium structures: (i) randomly disordered atomic configurations that broadly sample the configurational space and (ii) physics-informed phonon displacements that selectively probe the low-energy subspace accessible to ions in crystals. Our results show that models trained on phonon-informed datasets achieve higher accuracy and robustness with significantly fewer data points. Explainability analyses further reveal that these models assign greater importance to chemically meaningful bonds governing band-gap variations, thereby linking predictive performance to physical interpretability. Overall, this study demonstrates that embedding physical knowledge in dataset construction can substantially enhance ML performance in materials science and provides a pathway toward more efficient and physically grounded ML-driven materials discovery.

2 | Results

GNN models were trained to reproduce the physical behavior of anti-perovskite materials under realistic $T \neq 0$ K conditions. The predictive capabilities of the developed GNN frameworks are systematically assessed as follows. First, the performance of the ML models trained on a large dataset of non-equilibrium atomic configurations is presented, focusing on relevant electronic and mechanical properties. Next, the influence of dataset quality and physical relevance on model accuracy is examined by comparing models trained under distinct atomic displacement-generation schemes. Finally, explainability analyses are employed to identify the atomic-scale features that most strongly govern the predictive behavior of the developed GNN models, establishing direct physicochemical connections between atomic position fluctuations and finite-temperature optoelectronic properties.

2.1 | Dataset and Graph Generation

Crystals with the general chemical formula ABX_3 typically adopt a centrosymmetric cubic $Pm\bar{3}m$ structure at moderate and high temperatures, as exemplified by archetypal materials such as $BaTiO_3$ [24] and Ag_3SBr [25]. In this reference crystal structure, BX_6 octahedra are formed with B ions located at the octahedral centers, X ions at the vertices, and A ions occupying the corners of the cubic unit cell (Figure 1a). When A and B species are cations and X species are anions, the resulting centrosymmetric phase is known as a perovskite. Conversely, when A and B are anions and X are cations, the structure is referred to as an antiperovskite.

Antiperovskite materials are of great technological relevance for energy applications operating at ambient and higher temperatures, including solar cells and electrochemical batteries [19–23]. Recent studies have revealed that the optoelectronic properties of silver chalcohalide antiperovskites, Ag_3XY (X = S, Se; Y = Br, I), may exhibit a pronounced dependence on temperature [26, 27]. In particular, quantities such as the band gap and light absorption coefficient may vary by large percentages ($\sim$10%) due to temperature-induced volume expansion and local symmetry breaking. Owing to these pronounced temperature renormalization effects, and their technological relevance, these materials constitute representative examples of anharmonic functional compounds, characterized by the presence of low-energy phonon

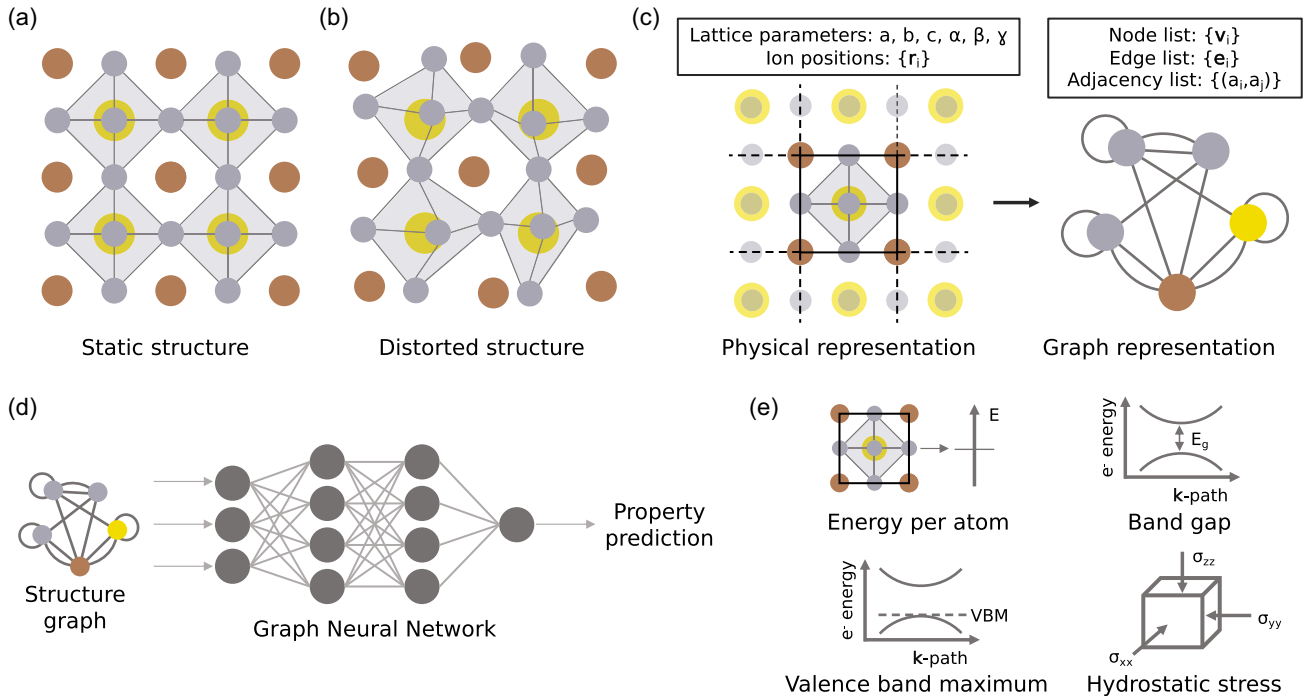

FIGURE 1 | Physical and ML aspects in materials informatics covered in this study. Atomic structure of the reference anti-perovskite system with ions located (a) at the equilibrium lattices sites and (b) around the equilibrium lattice sites. (c) Mapping of a real-space atomic configuration into its graph representation. Crystal periodicity is indicated by dashed lines in the physical representation and self-loops and multiple edges in the graph representation. (d) Property prediction using GNNs. (e) The selected quantities for ML prediction: energy per atom, band gap, valence band maximum, and hydrostatic stress.

modes that are decisive for their thermal and optoelectronic behavior [27].

In this work, we generated a comprehensive first-principles dataset for the silver chalcohalide antiperovskites Ag_3SBr, Ag_3SI, and $Ag_3SBr_xI_{1-x}$. Specifically, the energy per atom, band gap, valence band maximum, and hydrostatic stress were accurately calculated using density functional theory (DFT) for a total of 4, 500 non-equilibrium configurations (Section 2.3 and Figure 1e). These configurations capture the thermal motion of atoms at $T \neq 0$ K and are therefore suitable for training ML models aimed at predicting the physical properties of silver chalcohalide antiperovskites under realistic thermal conditions.

For crystalline materials, particular attention must be paid to properly preserve translational invariance when mapping atomic structures onto graph representations. The strategy that we employ to incorporate crystal periodicity into our GNN models is detailed in the Methods section and Supplementary Information (Supplementary Fig. 1) and is schematically illustrated in Figure 1c. While a crystal structure is defined by its lattice vectors (a, b, c, α, β, γ) and atomic positions $\mathbf{r}_i$, the corresponding graph representation is described by the node tensor $\mathbf{v}_i$ (atomic features), the edge tensor $\mathbf{e}_{ij}$ (bonding features), and the adjacency tensor (a_i, b_j) (connectivity information). Once a graph is constructed, either for an equilibrium configuration or for perturbed atomic positions, it is fed into a GNN model (Figure 1d).

As described in the Methods section and Supplementary Information, for the graph generation, we introduced a cutoff radius to determine which atomic pairs are connected in the graph. Supplementary Fig. 2 presents the model performance as a function of this cutoff radius. Shorter cutoff values produce simpler graphs and faster training but capture fewer interatomic interactions, whereas larger radii increase computational cost without necessarily enhancing predictive accuracy. We found that a cutoff radius of 5.5 Å yields well-converged results and therefore this value was adopted throughout this work. It is worth noting that the radial pair distribution function calculated for Ag_3SBr clearly shows that the first coordination shell for any pair of atomic species lies within this cutoff distance (Supplementary Fig. 3).

2.2 | GNN Model Training

A comprehensive hyperparameter study was performed to identify the optimal architecture and training parameters for each target property. Six distinct model architectures, differing in the number and type of layers, were evaluated (Supplementary Fig. 4). The graph convolution layer was implemented following the approach of Morris et al. [28], in which the edge weights, corresponding to the Euclidean distances between connected atoms, modulate the message-passing operation. This transformation can be expressed as:

$$\mathbf{x}'_i = \mathbf{W}_1 \mathbf{x}_i + \mathbf{W}_2 \sum_{j \in \mathcal{N}(i)} e_{j,i} \cdot \mathbf{x}_j, \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}'_i$ denote the node feature vectors before and after the convolution layer, respectively, $\mathbf{W}_1$ and $\mathbf{W}_2$ are trainable weight matrices, and $e_{j,i}$ is the edge weight between nodes i and j. The following hyperparameters were explored: (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}) for learning rate, (32, 64, 128, 256) for number of hidden channels, and (0.0, 0.2, 0.4, 0.6) for dropout.

As a result, a total of 480 GNN models were trained for each property.

The full dataset comprised 4, 500 configurations generated for antiperovskite materials (Section 2.1) using two different schemes for sampling thermal ionic motion, which are discussed in detail in the next subsection. The dataset was then divided into 70% for training, 15% for validation, and 15% for testing. To ensure consistent and meaningful comparisons, all models were trained on the same training set and evaluated on the same validation set. The best-performing model for each target property was subsequently assessed on the independent test set.

Figure 2 presents the performance of the GNN models that best reproduce the DFT reference values across the training, validation, and test sets. Results are shown for the energy per atom (Figure 2a–c), band gap (Figure 2e–g), valence band maximum (Figure 2i–k), and hydrostatic stress (Figure 2m–o). The corresponding distributions of prediction errors are displayed in Figure 2d,h,l,p. In the scatter plots, the color intensity reflects the point density, with darker colors indicating higher density. Overall, the GNN models exhibit strong predictive performance across all datasets and properties, as most points cluster near the diagonal dashed line, indicating close agreement with DFT. The error distributions are centered around zero, showing no systematic over- or underestimation. Among the studied quantities, the band gap displays the largest prediction dispersion, consistent with previous reports [29], reflecting its higher intrinsic complexity.

Table 1 summarizes key performance metrics for the best GNN models, including the mean absolute error (MAE), root mean squared error (RMSE), and squared Pearson correlation

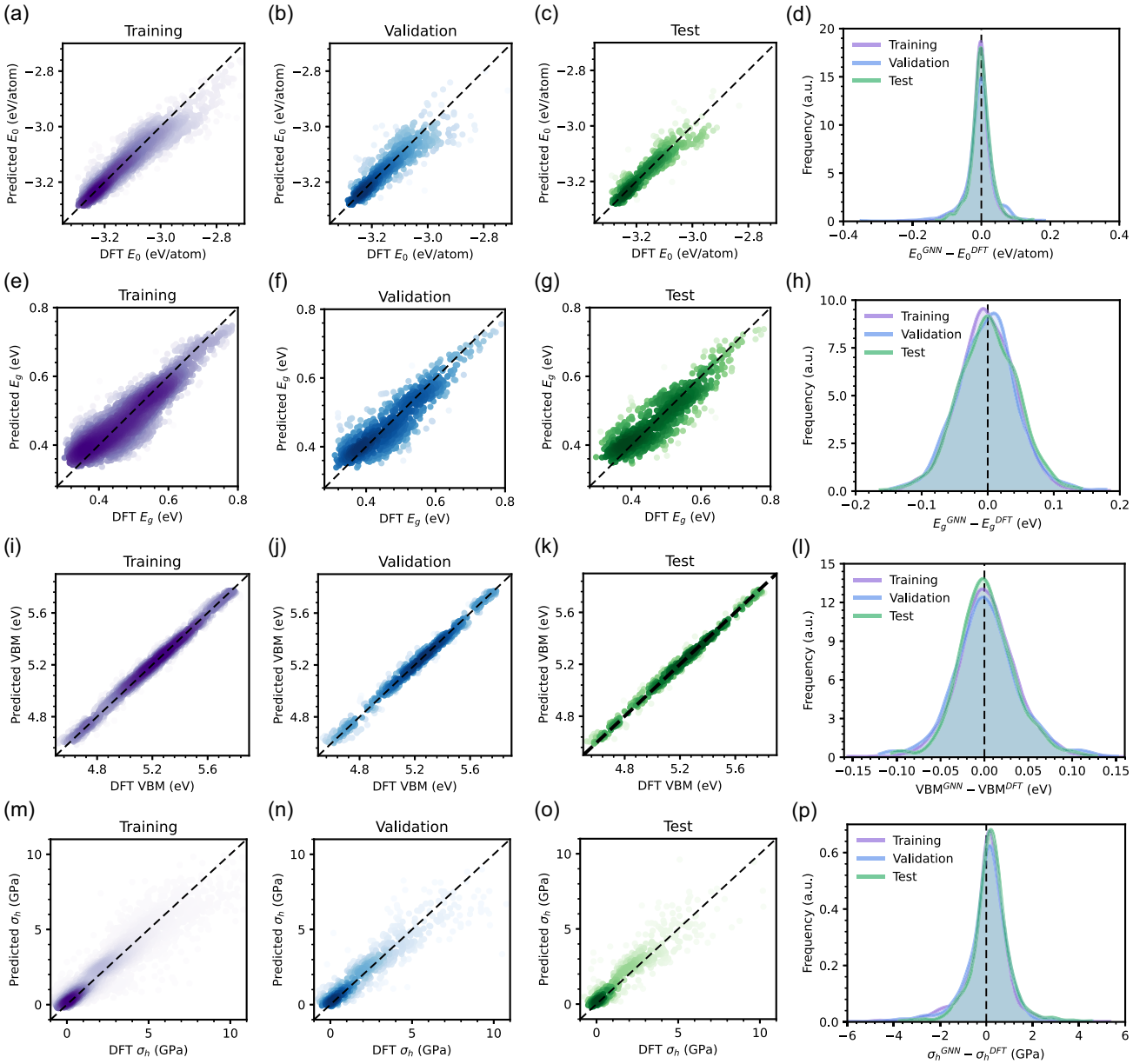

FIGURE 2 | Performance of best GNN models. DFT values versus GNN model predictions for the training, validation, and test datasets. Normalized distribution of errors in ML predictions are shown in (d),(h),(l),(p). Results are shown for (a–d) energy per atom, (e–h) band gap, (i–l) valence band maximum, and (m–p) hydrostatic stress.

TABLE 1 | Best model performance metrics.

	E_0, eV/atom			E_g, eV			VBM, eV			σ_h, GPa		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Train	0.024	0.036	0.89	0.034	0.043	0.81	0.026	0.035	0.99	0.65	1.01	0.84
Validation	0.028	0.044	0.79	0.035	0.045	0.79	0.028	0.037	0.98	0.65	0.99	0.84
Test	0.022	0.032	0.85	0.035	0.044	0.79	0.024	0.033	0.99	0.60	0.88	0.82

Note: MAE, RMSE, and R^2 values for the training, validation, and test sets of the best GNNs model in predicting energy per atom, band gap, valence band maximum, and hydrostatic stress. MAE and RMSE are reported in the units of the corresponding property, while R^2 is dimensionless.

coefficient (R^2). The similarity of results across the training, validation, and test sets confirms the absence of overfitting. The MAE values were approximately 0.025 eV/atom for the energy, 0.035 eV for both the band gap and valence band maximum, and 0.65 GPa for the hydrostatic stress; these figures are comparable to typical DFT accuracies and to results reported for other GNN-based models. The R^2 values were around or above 0.80 in all cases, confirming strong correlations between GNN predictions and DFT calculations. Overall, these results demonstrate that GNN models can effectively and efficiently reproduce DFT-level predictions for atomically disordered antiperovskite systems.

Finally, Supplementary Figs. 5–8 summarize the performance of all trained models (excluding those with $R^2 < -1$, which perform worse than random guessing) over the validation set. The results are displayed as violin plots, showing the distribution of model performance metrics. The observed trends are consistent across the four target properties: while some architectures generally yield slightly better results, the most critical hyperparameter is the learning rate, with optimal performance achieved for values of 10^{-4} and 10^{-5}.

2.3 | GNN Model Performance

The dependence of the performance of our GNN models on dataset type and dataset size were thoroughly analyzed. For this part of our study, we focused on the prediction of the band gap since this quantity exhibits the largest dispersion and lowest R^2 among all studied cases (Figure 2 and Table 1). As for the dataset, atomically disordered antiperovskite configurations were generated using two different approaches, described in detail next, one consisting of random uniform displacements and the other being physically motivated (Figure 3a,b). To ensure good configurational space sampling, all displacements were generated following a Monte Carlo procedure (Figure 3c).

Starting with a perfect 40-atoms antiperovskite simulation cell, that is, with all the ions located at their equilibrium lattice positions, we applied random atomic displacements according to the expression:

$$u_{j,\alpha}^{\text{rand}} = \epsilon \cdot a \tag{2}$$

where j is an atom index, α represents Cartesian direction, ϵ is a random number sampled from the uniform distribution $\epsilon \in \mathcal{U}[-1, 1]$, and a is a characteristic displacement length that was set to different values ranging from 0.1 to 0.8 Å. This a value was chosen to ensure adequate sampling of electronically insulating configurations (Supplementary Fig. 9).

For the physically informed dataset, the atomic displacements were generated according to the phonon eigenvectors and amplitudes calculated for the selected antiperovskite systems. In particular, we used the well-known expression obtained within the quasi-harmonic approximation [30]:

$$u_{j,\alpha}^{\text{phon}} = \epsilon \cdot \zeta_{j,\alpha}$$
$$\zeta_{j,\alpha}^2 = \frac{\hbar}{2Nm_j} \sum_{\mathbf{q},\nu} \frac{1}{\omega_\nu(\mathbf{q})} (1 + 2n_\nu(\mathbf{q}, T)) |e_\nu^\alpha(j, \mathbf{q})|^2 \tag{3}$$

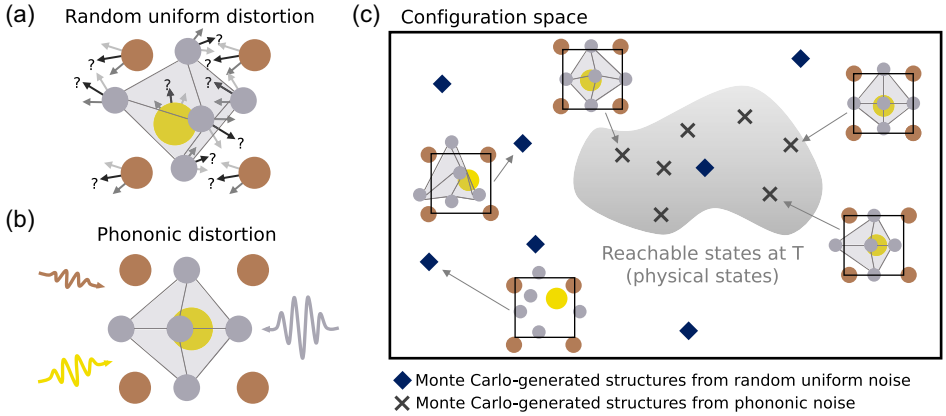

FIGURE 3 | Atomically perturbed antiperovskite structures and configuration space. Reference antiperovskite structure perturbed following (a) random- and (b) phonon-based atomic displacement schemes. Wavy arrows represent phonon modes, where the color indicates the displaced atom. (c) Configurational space of possible structures with fixed lattice parameters, where each point corresponds to a specific atomic configuration. The gray region indicates thermally accessible states starting from the reference structure. Diamonds and crosses represent nonequilibrium structures generated by Monte Carlo sampling following random- and phonon-based schemes, respectively.

where $\hbar$ is the reduced Planck constant, N is the total number of atoms in the unit cell, m_j is the atomic mass of atom j, $\mathbf{q}$ denotes a reciprocal-space vector, ν is a phonon mode branch index, and ω_ν, n_ν, and e_ν correspond to the phonon frequency, Bose-Einstein occupation distribution, and phonon eigenvector, respectively. Anharmonic lattice effects were included by using a normal-mode-decomposition scheme implemented in the `DynaPhoPy` software [31], following a methodology similar to that employed in previous works [25, 27].

Two independent band-gap datasets were generated by performing DFT calculations (Methods) on 2,250 structures obtained separately using the random and phonon-informed approaches described above. Independent GNN models were then trained on each dataset, with 250 configurations initially reserved for testing. The convergence of the final GNN models with respect to dataset size was systematically monitored, employing multiple random seeds (up to 10) to ensure statistically meaningful averages and uncertainties. Each model was subsequently evaluated on three distinct datasets: (1) the dataset generated using the same atomic-displacement scheme as in training, (2) the dataset obtained with the complementary displacement approach, and (3) a combined dataset containing structures generated using both the random and phonon-informed methods. The resulting model performance analysis is presented in Figure 4.

For both atomic-displacement schemes, the GNN models converge at approximately 1,000 training structures (Figure 4a,d). However, the phonon-informed model exhibits superior accuracy, achieving $R^2 \approx 0.85$ and MAE ≈ 0.030 eV, compared to the random-distortion model, which yields $R^2 \approx 0.72$ and MAE $\approx$ 0.036 eV. Remarkably, the phonon-based model even outperforms the model trained on the combined dataset (Table 1), which includes twice as many structures, indicating that physically consistent data is more valuable than simply increasing dataset size. This observation can be rationalized as follows: although random distortions sample a broader configurational space, they may include nonphysical atomic arrangements that hinder the model's ability to learn meaningful relationships between bonding geometry and band-gap variations. We will return to this important point, related to the explainability of GNN models, in the next section.

Cross-testing further confirms the superior performance of the GNN model trained on the physics-informed dataset. When evaluated on their complementary datasets (Figure 4b,e), both models show a decrease in accuracy, consistent with a reduced extrapolation capability. Nevertheless, the phonon-informed model generalizes better to the randomly—displaced dataset, as evidenced by its cross-test MAE, which is approximately 0.010 eV lower than that of the model trained on the random dataset. A similar trend is observed for the R^2 correlation coefficient, which reaches its highest value for the physics-informed model. These results strongly indicate that training on physically meaningful data yields more transferable and robust representations.

Finally, both GNN models were tested on a combined dataset containing configurations generated using the two atomic-displacement schemes (Figure 4c,f). Once again, the phonon-informed model outperforms the random-distortion model, achieving a MAE of <0.040 eV, compared with >0.045 eV for the latter. These findings further reinforce the conclusion that data quality and physical relevance outweigh dataset size in determining GNN performance for predicting the band gaps of antiperovskite materials.

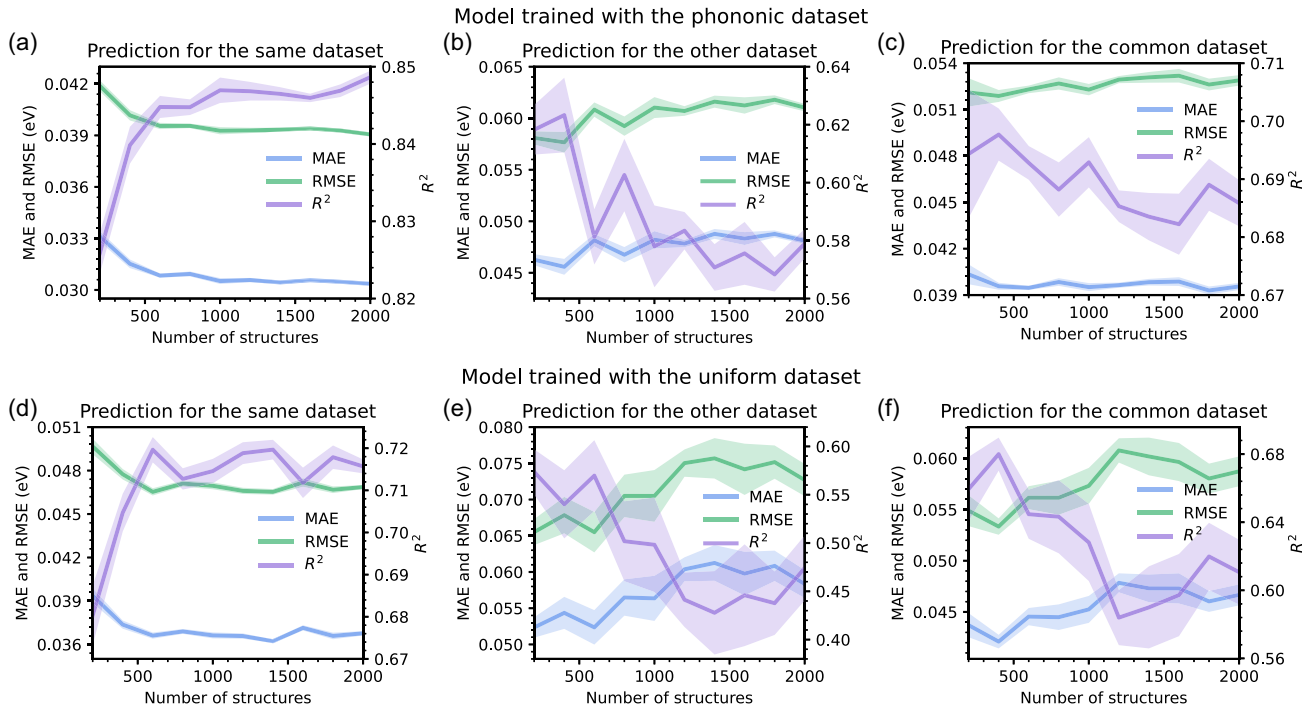

FIGURE 4 | Performance of GNN models trained on different datasets for the prediction of band gaps. Training was performed on datasets generated according to (a,b,c) phonon-based and (d,e,f) random displacement-based schemes. Models performance tests done on (a,d) their corresponding dataset, (b, e) their complementary dataset, and (c,f) a combined dataset comprising both atomic-displacement generation schemes. Solid lines indicate average values and shaded areas statistical errors.

It is instructive to analyze the distributions of the four target physical quantities (i.e., atomic energy, band gap, valence band maximum, and hydrostatic stress) computed with DFT for the three datasets used to assess GNN model performance. As shown in Supplementary Fig. 10, the energy per atom and hydrostatic stress distributions differ substantially among the datasets: the randomly generated dataset contains numerous high-energy, nonphysical configurations, whereas the phonon-informed dataset comprises lower-energy, thermally accessible structures. These pronounced differences complicate direct comparisons between models trained to predict atomic energies or hydrostatic stress on distinct datasets, since their underlying property distributions vary significantly. In contrast, the band-gap distributions are nearly identical across all three datasets (Supplementary Fig. 10), enabling a fair and consistent comparison of model performance.

Furthermore, we analyzed the distributions of pairwise atomic distances within the configurations forming the random and phonon-informed datasets (Supplementary Fig. 11). Only interatomic distances shorter than 6.0 Å were considered, as the graph cutoff radius was set to 5.5 Å, rendering larger distances irrelevant for model construction. For all atomic pairs in the antiperovskite structure, the distance distributions are nearly identical between the two datasets. This structural similarity further ensures that comparisons between GNN models trained for band-gap prediction on the two datasets are physically meaningful.

Interestingly, we compared the performance of our band-gap-predictive GNN model when trained on a phonon-informed dataset and on a second physics-informed dataset composed of configurations extracted from *ab initio* molecular dynamics (AIMD) simulations (Supplementary Information and Supplementary Table 1). The phonon- and AIMD-informed datasets are comparable in size (approximately 1,000 configurations each). The AIMD dataset was generated from simulations performed at three different temperatures (200, 400, and 600 K), each with a duration of about 60–70 ps (Supplementary Information), which is typical of AIMD studies [32–34].

We found that the model trained on the phonon-informed dataset generally performs slightly better than the one trained on the AIMD-informed dataset. Specifically, the former yields marginally smaller MAE and RMSE values and a slightly larger R^2 score (Supplementary Table 1). This behavior can be attributed to the fact that configurations in the AIMD-based dataset are more correlated and less diverse than those in the phonon-based dataset. While extending the duration, and hence the size, of the AIMD simulations would likely lead to comparable predictive performance between the two models, this improvement would come at the cost of a substantially higher computational effort. In contrast, generating the phonon-informed dataset entails a negligible additional computational cost.

2.4 | GNN Model Explainability

To gain insight into why certain GNN models outperform others, we employed graph explainability techniques as implemented in PyTorch Geometric [35, 36]. These methods quantify the relative importance assigned by the GNN to each edge during the message-passing process, thus effectively identifying which chemical bonds the model considers most relevant for predicting the band gap. Specifically, we analyze the edge importance difference between a well trained GNN model with good metrics and a GNN model that did not achieve a good performance after the training.

Figure 5a displays the ideal five-atom cubic antiperovskite unit cell, providing a structural reference for the subsequent analysis. Figure 5b–d illustrates the relationship between the atomic structure and its corresponding graph representation. Large spheres represent the atoms within the unit cell, while smaller ones denote their periodic images. The color scheme differentiates atomic species, and the unit cell boundaries (solid lines) are shown for clarity. The most influential edges, as determined by the explainability algorithm, are highlighted as red dashed lines for the best-performing model (Figure 5b) and for a poorly performing model (Figure 5c). For reference, all graph edges are shown as graey dashed lines in Figure 5d. Edges with an importance score above 0.86 (on a scale from 0 for least relevant to 1 for most relevant) were classified as significant. The best-performing model identifies several highly important bonds, predominantly Ag—S connections, whereas the low-performing model recognizes only a single relevant edge, corresponding to an Ag—Ag bond.

Figure 5e,f present the edge-importance density distributions for all bonds and for each bond type, comparing the best and a poor-performing models. The optimal model assigns low importance (below 0.4) to most bonds but very high importance (above 0.8) to a few, primarily Ag–S bonds. In contrast, the poorly performing model exhibits a more uniform importance distribution, with most bonds receiving moderate scores below 0.6. This behavior indicates that the best model effectively distinguishes between chemically and electronically relevant bonds and those that play a minor role in determining the band gap, whereas the weaker model fails to make such distinctions, effectively treating all bonds as equally relevant.

The prominent emphasis placed on Ag—S bonds by the best-performing model is physically well justified. Previous first-principles studies [26, 27] demonstrated that Ag—S interactions are primarily responsible for the temperature dependence of the band gap in Ag_3SBr. Figure 5g shows the DFT-calculated electronic density of states (eDOS) for the static paraelectric structure, revealing pronounced Ag and S orbital contributions near the conduction band edge. Hybridization between Ag-*s* and S-*s* orbitals generates bonding and antibonding states, with the conduction band associated with the bonding state (Figure 5h). When soft phonon modes distort the lattice, the energy splitting between bonding and antibonding states increases, lowering the energy of the bonding state, and consequently of the conduction band, resulting in a reduced band gap (Figure 5i). Thus, the temperature-induced band-gap variation in silver chalcohalide antiperovskites is predominantly governed by Ag—S bonding.

The ability of the best GNN model to assign high importance to Ag—S bonds therefore demonstrates that it has effectively learned the underlying physics controlling the thermal evolution of the band gap. By contrast, Br—Br bonds, which contribute negligibly near the band edges, receive very low importance values (below 0.2). The poorly performing model, however, does not clearly distinguish between bond types, assigning similar importance to all of them, evidence that it has not internalized the relevant chemical–electronic relationships.

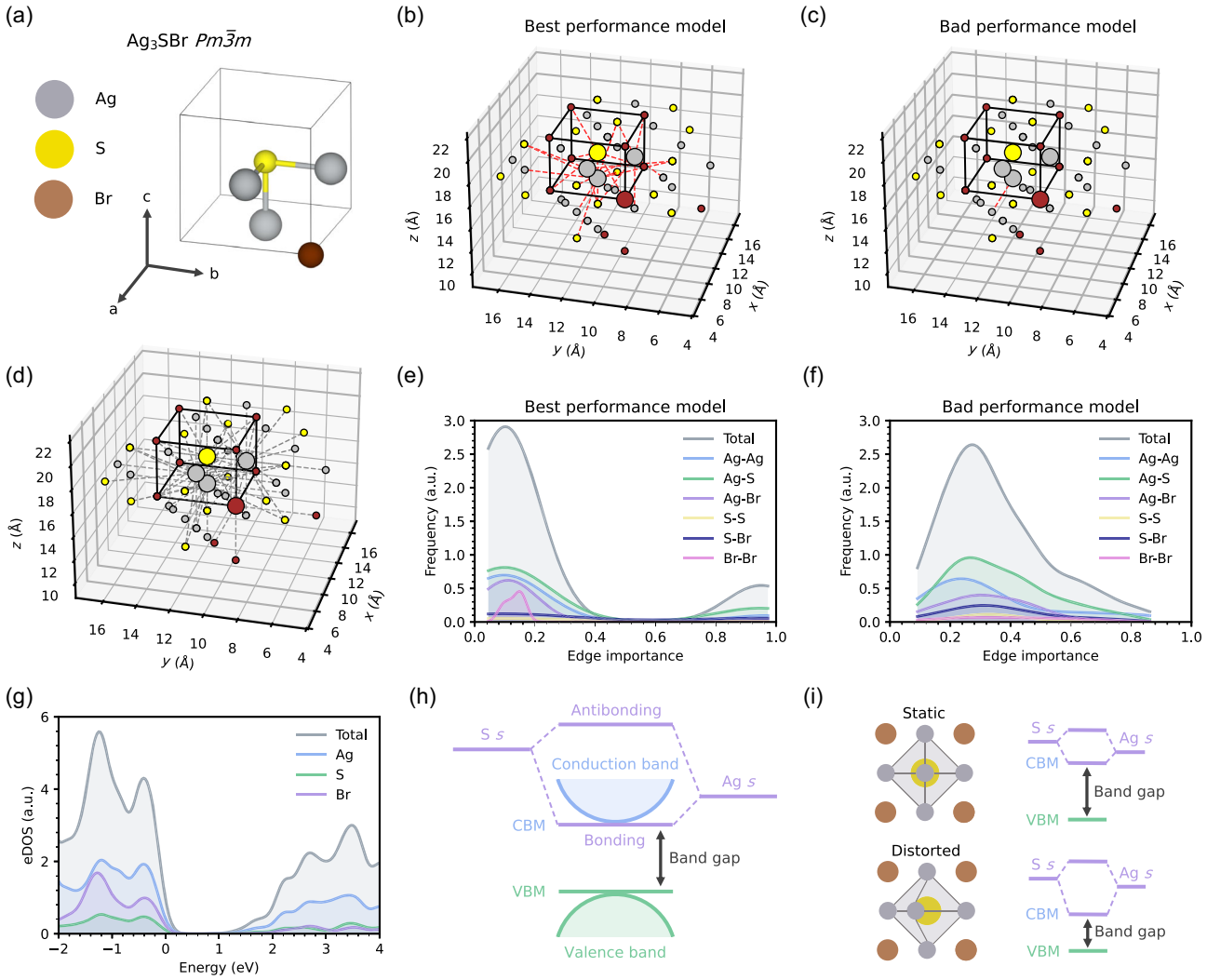

FIGURE 5 | GNN model explainability for band-gap prediction. (a) Reference unit cell for Ag_3SBr. (b,c) Mixed structure-graph representation of the system, where large circles represent unit-cell atoms and small circles their periodic images. Red dashed lines mark edges with importance greater than 0.86, for (b) the best-performing GNN and (c) a poorly performing GNN. The unit cell is highlighted in black. (d) Graph representation showing all edges as gray dashed lines. (e,f) Edge-importance density distributions for the best-performing and poorly-performing GNN, compared with the total and pairwise edge densities. (g) Electronic density of states near the band gap computed with DFT methods. (h) Electronic structure of Ag_3SBr around the band gap, highlighting valence and conduction bands and the orbital hybridization of Ag and S *s* electrons. (i) Band-gap reduction resulting from atomic position perturbations in Ag_3SBr.

3 | Discussion

The results presented in this work highlight the critical importance of physically informed data generation in the development of ML models for materials prediction. By comparing models trained on datasets constructed using random uniform distortions and phonon-informed distortions, we demonstrated that data quality and physical relevance often outweigh dataset size. GNN models trained on phonon-informed data achieved superior performance in predicting band gaps, even with fewer training structures. This finding underscores that targeted sampling of physically meaningful regions of configurational space enhances both generalization and interpretability.

The superior accuracy of the phonon-based models can be rationalized by noting that phonon distortions capture the thermally accessible structural configurations of materials (Figure 3c). In contrast, random distortions probe unphysical regions of configurational space, including unrealistically short interatomic distances (<0.1 Å) or completely disordered structures, which obscure the true structure-property relationships the model must learn. These results indicate that incorporating physics through phonons, symmetry constraints, or the curvature of the potential-energy surface into data generation guides the model toward the most relevant degrees of freedom, effectively embedding prior physical knowledge within the dataset.

Explainability analyses further support this interpretation. The best-performing GNN models correctly identified chemically relevant bonds, particularly Ag–S, as the dominant contributors to band-gap evolution under finite-temperature conditions. Poorly performing models, by contrast, exhibited diffuse and uncorrelated bond-importance distributions, reflecting a lack of understanding of the underlying bonding physics. This clear link between interpretability and predictive accuracy reinforces that explainability tools are essential not only for model transparency but also for verifying whether

ML models have genuinely learned meaningful structure–property correlations.

These insights are also directly relevant to the development of machine learning interatomic potentials (MLIPs) [37]. In MLIP construction, the quality and representativeness of the training dataset determine both the model's transferability and its physical reliability. Our findings suggest that building training datasets using phonon-based or otherwise physically constrained atomic displacements can yield more accurate and stable MLIPs, particularly for describing anharmonic effects, finite-temperature dynamics, and phase transitions.

In light of these findings, the use of atomic configurations extracted from high-quality finite-temperature simulations, such as AIMD simulations, instead of randomly generated structures, is fully justified and highly recommended for MLIP and GNN model training. However, as noted in Sections 1 and 2.3, AIMD simulations are computationally demanding, and their limited sampling times can result in redundant exploration of configurational space. In contrast, the dataset generation approach proposed in this study, based on phonon eigenvectors and amplitudes, requires minimal computational effort and is straightforward to implement with standard methods. This work therefore provides an efficient and physically grounded pathway toward accelerated, interpretable, and reliable ML-driven materials discovery.

4 | Conclusions

This study demonstrates that incorporating physics into data generation, here through phonon-informed atomic displacements, substantially enhances the accuracy, generalization, and interpretability of graph-based ML models for materials prediction. We show that physically meaningful data can outperform much larger random datasets, highlighting that dataset design grounded in fundamental principles is as crucial as model architecture. Beyond modeling the optoelectronic properties of anti-perovskite systems under realistic $T \neq 0$ K conditions, our findings establish a broadly applicable strategy for accelerating and rationalizing ML-driven materials discovery, offering a path towards more efficient and explainable models across diverse material classes.

5 | Methods

5.1 | DFT Calculations

DFT calculations [38, 39] were performed with VASP software [40–42] and semilocal PBEsol exchange-correlation functional [43]. Wave functions were represented in a plane-wave basis set truncated at 700 eV. We selected a dense k-point grid, with $12 \times 12 \times 12$ points for the reciprocal-space Brillouin zone sampling for the cubic anti-perovskite unit cell. We obtained zero-temperature energies converged to within 0.5 meV per formula unit. For geometry relaxations, a force tolerance of 0.005 eV $Å^{-1}$ was imposed in all the atoms.

5.2 | Graph Generation

Several methods exist for generating graph representations of crystal structures, including radius cutoffs, k-nearest neighbors (k-NN), and Voronoi-based approaches [44]. In this work, we employed the widely used radius-cutoff method. Each atom in the unit cell is represented as a node, characterized by four atomic features: atomic number, atomic mass, atomic radius, and electronegativity. A cutoff radius R_{cutoff} defines the maximum interatomic distance considered for forming edges (i.e., chemical bonds). Two atoms are connected by an edge if their Euclidean separation is smaller than R_{cutoff}, and the corresponding distance is stored as the edge feature. To ensure periodicity, we build a supercell large enough to fully contain all cutoff spheres centred on each atom. If an atom forms a bond with a periodic image of itself, this connection is represented as a self-edge. Further details about the graph construction, including sensitivity tests with respect to R_{cutoff}, are provided in the Supplementary Information (Supplementary Figs. 1–3 and Supplementary Discussion). The python scripts used for graph generation are openly available in our GitHub repository: https://github.com/polbeni/GNN-materials. Please note that this repository also contains additional scripts and utilities not directly discussed in this work.

5.3 | GNN Models

Graph neural network models were implemented using the PyTorch Geometric framework [45], which is built on top of the PyTorch machine learning library [46]. Details regarding the specific model architectures, retraining procedures, and hyperparameter optimization are provided in the Supplementary Information.

Acknowledgments

P.B. acknowledges support from the predoctoral program AGAUR-FI ajuts (2024 FI-1 00070) Joan Oró, which is backed by the Secretariat of Universities and Research of the Department of Research and Universities of the Generalitat of Catalonia, as well as the European Social Plus Fund. C.L. acknowledges support from the Spanish Ministry of Science, Innovation and Universities under an FPU grant. C.C. acknowledges support by MICIN/AEI/10.13039/501100011033 and ERDF/EU under the grants PID2023-146623NB-I00 and PID2023-147469NB-C21 and by the Generalitat de Catalunya under the grants 2021SGR-00343, 2021SGR-01519 and 2021SGR-01411. Computational support was provided by the Red Española de Supercomputación under the grants FI-2024-1-0005, FI-2024-2-0003, FI-2024-3-0004,FI-2024-1-0025, FI-2024-2-0006, and FI-2025-1-0015. This work is part of the Maria de Maeztu Units of ExcellenceProgramme CEX2023-001300-M funded by MCIN/AEI (10.13039/501100011033). E.S. acknowledges the European Union H2020 Framework Program SENSATE project: Low dimensional semiconductors for optically tuneable solar harvesters (grantagreement Number 866018), Renew-PV European COST action (CA21148) and the Spanish Ministry of Science and Innovation ACT-FAST (PCI2023-145971-2). E.S. is grateful to the ICREA Academia program.

Funding

This work was supported by the Ministerio de Asuntos Económicos y Transformación Digital, Gobierno de España (CEX2023-001300-M).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All relevant scripts, including model training workflows and data preprocessing routines, are freely accessible in the GitHub repository: https://github.com/polbeni/GNN-materials.

References

1. W. X. Zhao, K. Zhou, J. Li, et al., "*A Survey of Large Language Models*," 1, no. 2 (2023), ArXiv Preprint arXiv: 2303.18223.

2. J. Jumper, R. Evans, A. Pritzel, et al., "Highly accurate protein structure prediction with AlphaFold," *Nature* 596, no. 7873 (2021): 583–589.

3. C. Malica, K. S. Novoselov, A. S. Barnard, et al., "Artificial Intelligence for Advanced Functional Materials: Exploring Current and Future Directions," *Journal of Physics: Materials* 8, no. 2 (2025): 021001.

4. W. Xia, M. Sakurai, B. Balasubramanian, et al., "Accelerating the Discovery of Novel Magnetic Materials Using Machine Learning–guided Adaptive Feedback," *Proceedings of the National Academy of Sciences* 119, no. 47 (2022): e2204485119.

5. Q. Liu, W. Chen, V. Yakubov, J. J. Kruzic, C. H. Wang, and X. Li, "Interpretable Machine Learning Approach for Exploring Process-Structure-Property Relationships in Metal Additive Manufacturing," *Additive Manufacturing* 85 (2024): 104187.

6. N. Karimitari, W. J. Baldwin, E. W. Muller, et al., "Accurate Crystal Structure Prediction of New 2d Hybrid Organic–inorganic Perovskites," *Journal of the American Chemical Society* 146, no. 40 (2024): 27392–27404.

7. F. Ren, L. Ward, T. Williams, et al., "Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-Throughput Experiments," *Science Advances* 4 (2018): 4.

8. C. Zeni, R. Pinsler, D. Zügner, et al., "A Generative Model for Inorganic Materials Design," *Nature* 639, no. 8055 (2025): 624–632.

9. L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-Driven Materials Science: Status, Challenges, and Perspectives," *Advanced Science* 6, no. 21 (2019): 1900808.

10. J. Schmidt, T. F. T. Cerqueira, A. H. Romero, et al., "Improving Machine-Learning Models in Materials Science through Large Datasets," *Materials Today Physics* 48 (2024): 101560.

11. Y. Zhang, and C. Ling, "A Strategy to Apply Machine Learning to Small Datasets in Materials Science," *Npj Computational Materials* 4, no. 1 (2018): 25.

12. P. Xu, X. Ji, M. Li, and W. Lu, "Small Data Machine Learning in Materials Science," *Npj Computational Materials* 9, no. 1 (2023): 42.

13. G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-Informed Machine Learning," *Nature Reviews Physics* 3, no. 6 (2021): 422–440.

14. D. Jha, L. Ward, A. Paul, et al., "Elemnet: Deep Learning the Chemistry of Materials from Only Elemental Composition," *Scientific Reports* 8, no. 1 (2018): 17593.

15. M. Cardona, T. A. Meyer, and M. L. W. Thewalt, "Temperature Dependence of the Energy Gap of Semiconductors in the Low-Temperature Limit," *Physical Review Letters* 92, no. 19 (2004): 196403.

16. R. Masuki, T. Nomoto, R. Arita, and T. Tadano, "Ab Initio Structural Optimization at Finite Temperatures Based on Anharmonic Phonon Theory: Application to the Structural Phase Transitions of $Batio_3$," *Physical Review B* 106, no. 22 (2022): 224104.

17. R. Rurali, C. Escorihuela-Sayalero, J. L. D. S. Tamarit, J. Íñiguez-González, and C. Cazorla, "Giant Photocaloric Effects across a Vast Temperature Range in Ferroelectric Perovskites," *Physical Review Letters* 133, no. 11 (2024): 116401.

18. B. Monserrat, "Electron–phonon Coupling from Finite Differences," *Journal of Physics: Condensed Matter* 30, no. 8 (2018): 083001.

19. I. Caño, J. W. Turnley, P. Benítez, et al., "Novel Synthesis of Semiconductor Chalcohalide Anti-Perovskites by Low-Temperature Molecular Precursor Ink Deposition Methodologies," *Journal of Materials Chemistry C* 12, no. 9 (2024): 3154–3163.

20. Z. Deng, D. Ni, D. Chen, et al., "Anti-Perovskite Materials for Energy Storage Batteries," *InfoMat* 4, no. 2 (2022): e12252.

21. Y. Wang, H. Zhang, J. Zhu, et al., "Antiperovskites with Exceptional Functionalities," *Advanced Materials* 32, no. 7 (2020): 1905007.

22. W. Xia, Y. Zhao, F. Zhao, et al., "Antiperovskite Electrolytes for Solid-State Batteries," *Chemical Reviews* 122, no. 3 (2022): 3763–3819.

23. D. Kalita, P. Sahu, and U. Manju, "Anti-Perovskites for Photovoltaics: Materials Development and Challenges," *Journal of Physics D: Applied Physics* 57, no. 34 (2024): 343002.

24. M. B. Smith, K. Page, T. Siegrist, et al., "Crystal Structure and the Paraelectric-to-Ferroelectric Phase Transition of Nanoscale $Batio_3$," *Journal of the American Chemical Society* 130, no. 22 (2008): 6955–6963.

25. P. Benítez, C. López, C. Liu, et al., "Crystal Structure Prediction and Phase Stability in Highly Anharmonic Silver-Based Chalcohalide Antiperovskites," *PRX Energy* 4, no. 2 (2025): 023002.

26. P. Benítez, S. Chen, R. Jiang, et al., "Giant Thermally Induced Band-Gap Renormalization in Anharmonic Silver Chalcohalide Antiperovskites," *Journal of Materials Chemistry C* 13, no. 20 (2025): 10399–10412.

27. P. Benítez, R. Jiang, S. Chen, et al., "Band-Gap Tunability in Anharmonic Perovskite-Like Semiconductors Driven by Polar Electron–phonon Coupling," *Journal of the American Chemical Society* 147(2025): 37506.

28. C. Morris, M. Ritzert, M. Fey, et al., "Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 01 (2019): 4602–4609.

29. T. Xie, and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," *Physical Review Letters* 120, no. 14 (2018): 145301.

30. A. Togo, "First-Principles Phonon Calculations with Phonopy and phono3py," *Journal of the Physical Society of Japan* 92, no. 1 (2023): 012001.

31. A. Carreras, A. Togo, and I. Tanaka, "Dynaphopy: A Code for Extracting Phonon Quasiparticles from Molecular Dynamics Simulations," *Computer Physics Communications* 221 (2017): 221–234.

32. A. K. Sagotra, D. Chu, and C. Cazorla, "Influence of Lattice Dynamics on Lithium-Ion Conductivity: A First-Principles Study," *Physical Review Materials* 3 (2019): 035405.

33. C. López, A. D. Emperador, E. Saucedo, R. Rurali, and C. Cazorla, "Universal Ion-Transport Descriptors and Classes of Inorganic Solid-State Electrolytes," *Materials Horizons* 10 (2023): 1757–1768.

34. C. López, R. Rurali, and C. Cazorla, "How Concerted Are Ionic Hops in Inorganic Solid-State Electrolytes?," *Journal of the American Chemical Society* 146, no. 12 (2024): 8269–8279.

35. Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, Gnnexplainer: Generating Explanations for Graph Neural Networks, *Advances in Neural Information Processing Systems* (2019): 32.

36. K. Amara, R. Ying, Z. Zhang, et al., "*Graphframex: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks*," (2022): arXiv preprint arXiv:2206.09677.

37. R. Jacobs, D. Morgan, S. Attarian, et al., "A practical guide to machine learning interatomic potentials – Status and future," *Current Opinion in Solid State and Materials Science* 35 (2025): 101214.

38. C. Cazorla, and J. Boronat, "Simulation and Understanding of Atomic and Molecular Quantum Crystals," *Reviews of Modern Physics* 89, no. 3 (2017): 035003.

39. P. E. Blöchl, "Projector Augmented-Wave Method," *Physical Review B* 50, no. 24 (1994): 17953.

40. G. Kresse, and J. Hafner, "Ab Initio Molecular Dynamics for Liquid Metals," *Physical Review B* 47, no. 1 (1993): 558.

41. G. Kresse, and J. Furthmüller, "Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set," *Computational Materials Science* 6, no. 1 (1996): 15–50.

42. G. Kresse, and J. Furthmüller, "Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set," *Physical Review B* 54, no. 16 (1996): 11169.

43. J. P. Perdew, A. Ruzsinszky, G. I. Csonka, et al., "Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces," *Physical Review Letters* 100, no. 13 (2008): 136406.

44. V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, "Benchmarking Graph Neural Networks for Materials Chemistry," *Npj Computational Materials* 7, no. 1 (2021): 84.

45. M. Fey and J. E. Lenssen, "*Fast Graph Representation Learning with Pytorch Geometric*," (2019), arXiv preprint arXiv:1903.02428.

46. A. Paszke, S. Gross, F. Massa, et al.," Pytorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems* (2019): 32.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supporting Figure S1:** Generation of a graph for a given crystal structure. **Supporting Figure S2:** Band-gap model loss as a function of the graph radius cutoff. **Supporting Figure S3:** Radial pair distribution function for equilibrium Ag3SBr: (a) Ag–Ag, (b) S–S, (c) Br–Br, (d) Ag–S, (e) Ag–Br, and (f) S–Br. The dashed vertical line at 5.5 Å represents the selected cutoff radius for graph generation. **Supporting Figure S4:** Representation of the six architectures used, differing in the number of convolution and pooling layers and in the activation functions. **Supporting Figure S5:** Violin plots of GNN performance metrics (R2,MAE,MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) drop out values, when predicting the energy peratom. **Supporting Figure S6:** Violin plots of GNN performance metrics (R2,MAE,MSE; lefttoright) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) drop out values, when predicting the band gap. **Supporting Figure S7:** Violin plots of GNN performance metrics (R2,MAE,MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels ,and (j–l) drop out values, when predicting the valence band maximum. **Supporting Figure S8:** Violin plots of GNN performance metrics (R2,MAE,MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) drop out values, when predicting the hydrostatic stress. **Supporting Figure S9:** Band gap calculated for randomly generated disordered configurations as a function of the atomic displacement amplitude. Results were obtained with DFT methods. **Supporting Figure S10:** Distribution of (a) energy per atom, (b) band gap, (c) valence band maximum, and (d) hydrostatic stress, for datasets generated with phonon-based atomic displacements (green), random atomic displacements (violet), and combining both schemes (blue). **Supporting Figure S11:** Distribution of pair atomic distances for (a) Ag–Ag, (b) Ag–S, (c) Ag–Br, (d) S–S, (e) Br–S, and (f) Br–Br. **Supporting Table SI:** Performance metrics for GNN models trained on phonon and AIMD-based datasets. Perfomances were evaluated on the corresponding train and validation datasets, as well as on the other model's dataset ("Complementary").