

**Supporting Information for “Why Physics Still Matters:  
Improving Machine Learning Prediction of Material Properties  
with Phonon-Informed Datasets”**

Pol Benítez and Cibrán López

*Department of Physics, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain and  
Research Center in Multiscale Science and Engineering,  
Universitat Politècnica de Catalunya, 08019 Barcelona, Spain*

Edgardo Saucedo

*Research Center in Multiscale Science and Engineering,  
Universitat Politècnica de Catalunya, 08019 Barcelona, Spain and  
Department of Electronic Engineering,  
Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*

Teruyasu Mizoguchi

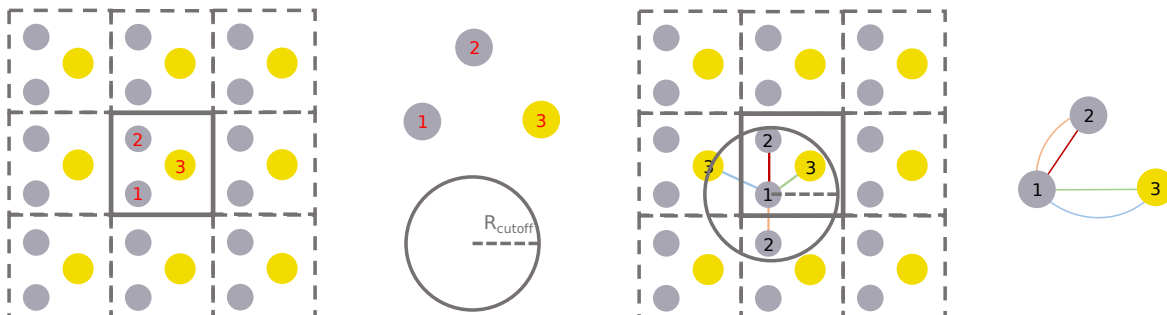
*Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan*

Claudio Cazorla

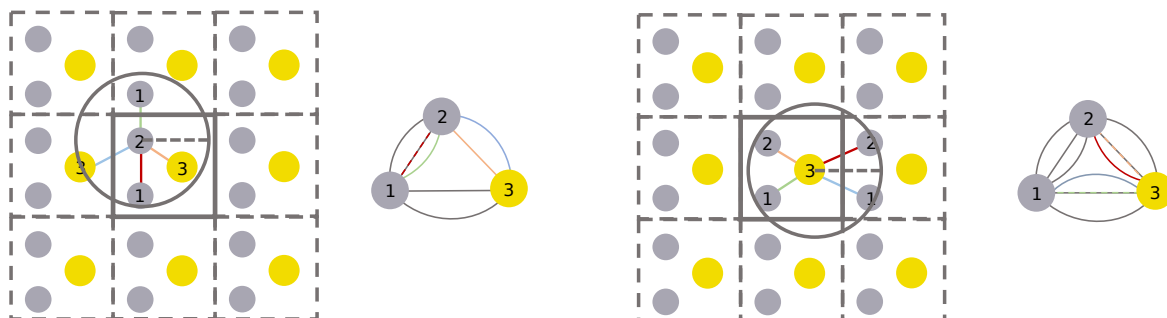
*Department of Physics, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain  
Research Center in Multiscale Science and Engineering,  
Universitat Politècnica de Catalunya, 08019 Barcelona, Spain and  
Institució Catalana de Recerca i Estudis Avançats (ICREA),  
Passeig Lluís Companys 23, 08010 Barcelona, Spain*

1. Take your structure and generate one node for each atom in the unit cell. Each node will include a list of features describing the atom type (such as mass, ionic radius, ...). Define a cutoff radius, which sets the maximum distance between atoms to consider a chemical bond.

2. For each atom, draw a sphere with the cutoff radius centered on it. Ensure the supercell is large enough to contain this sphere. Create edges between the centered atom and all atoms within the cutoff radius. Multiple edges may exist between the same two atoms, representing bonds to periodic images, this preserves the periodicity of the crystal. Each edge should include as a feature the Euclidean distance between the connected nodes.

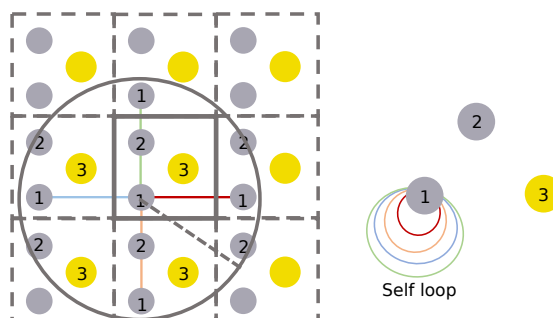
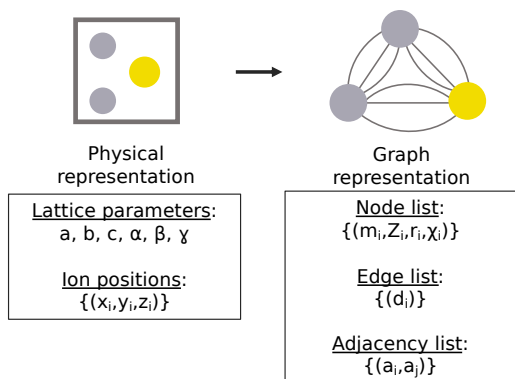


3. Repeat this process for all atoms in the unit cell. Avoid counting the same bond twice; for instance, if atom 1-2 (red connection) was already considered when centering on atom 1, skip it when centering on atom 2.

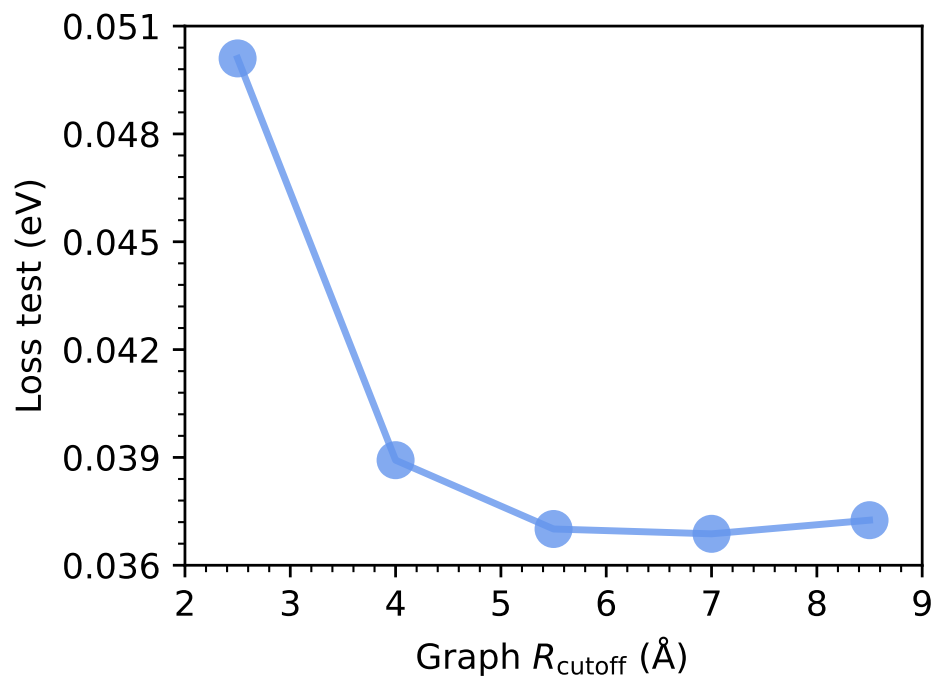


4. This process converts the physical representation (lattice parameters and ion positions) into a graph representation with node, edge, and adjacency tensors.

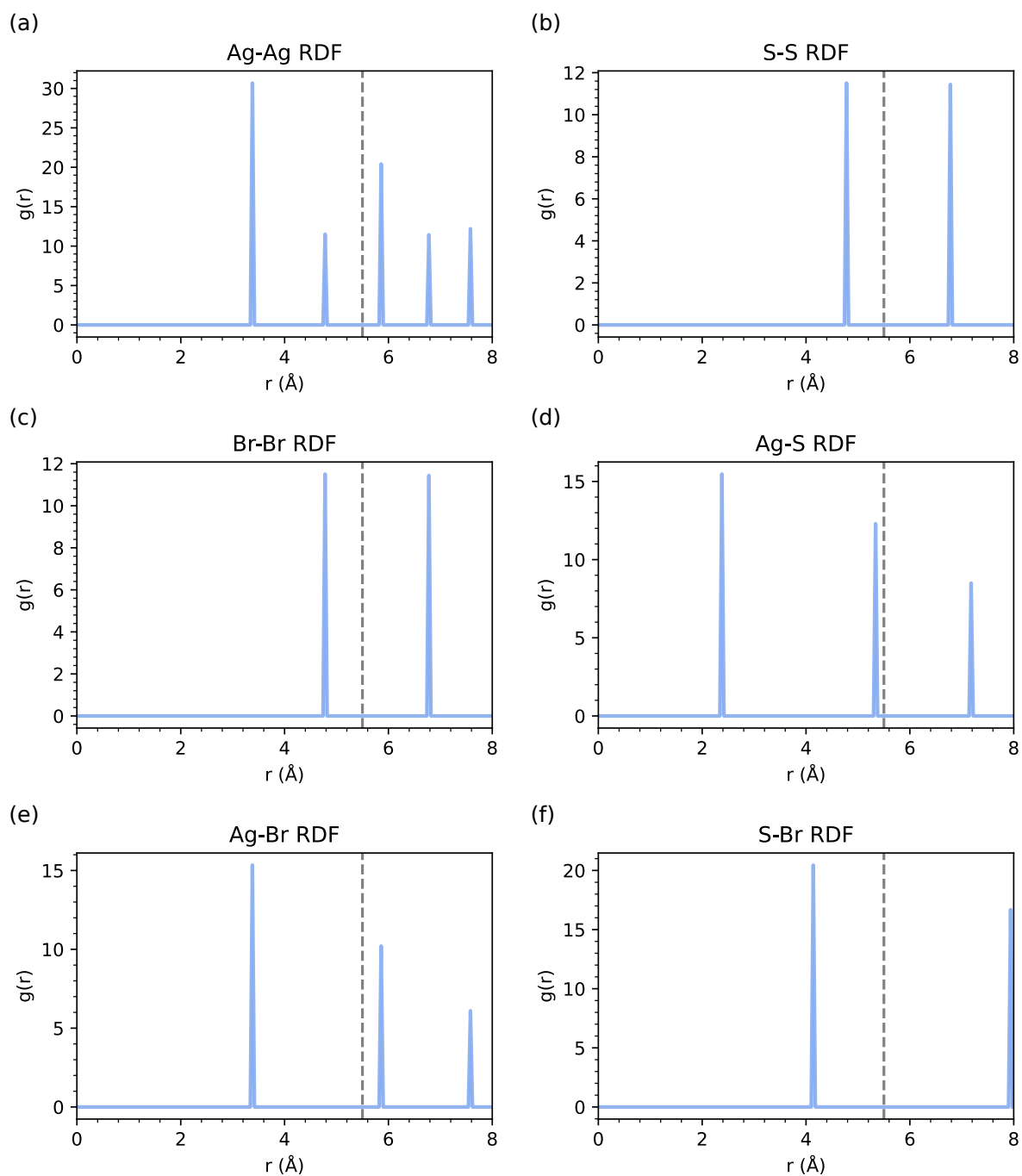
**Extra.** If a periodic image of the centered atom lies within the cutoff radius, treat it as a self-loop (an edge that starts and ends at the same node).



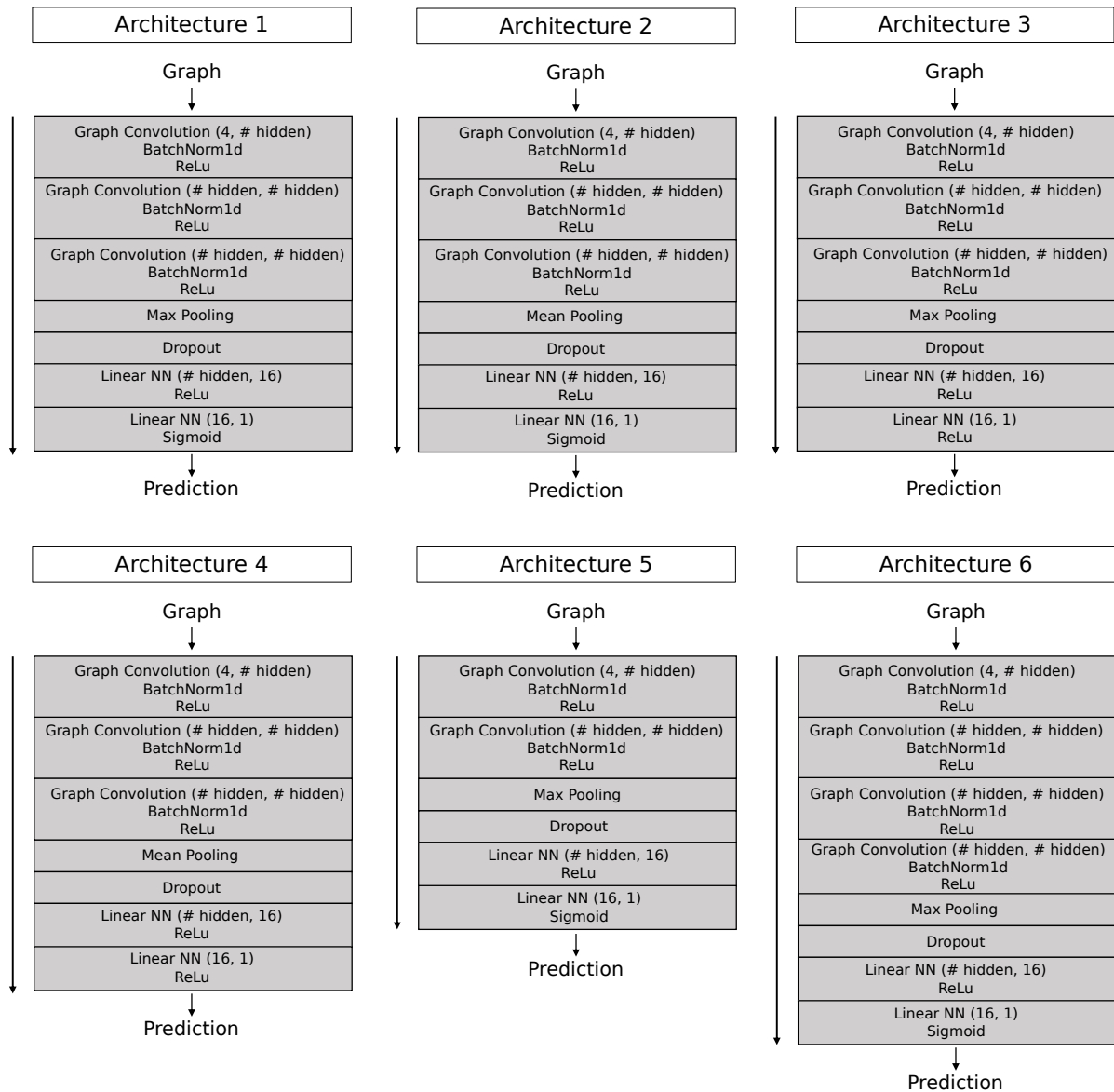
Supplementary Figure 1: Generation of a graph for a given crystal structure.



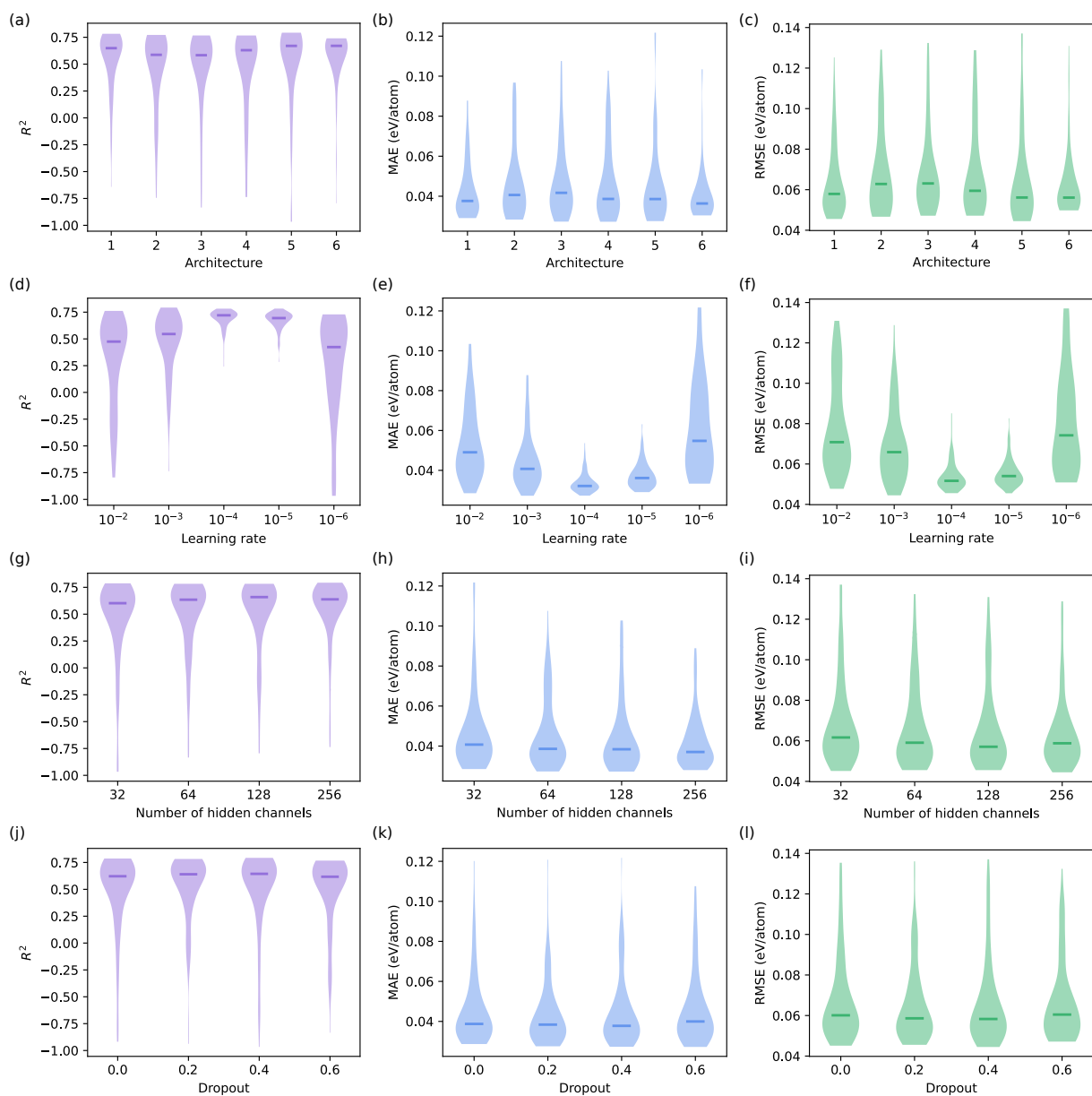
**Supplementary Figure 2:** Band-gap model loss as a function of the graph radius cutoff.



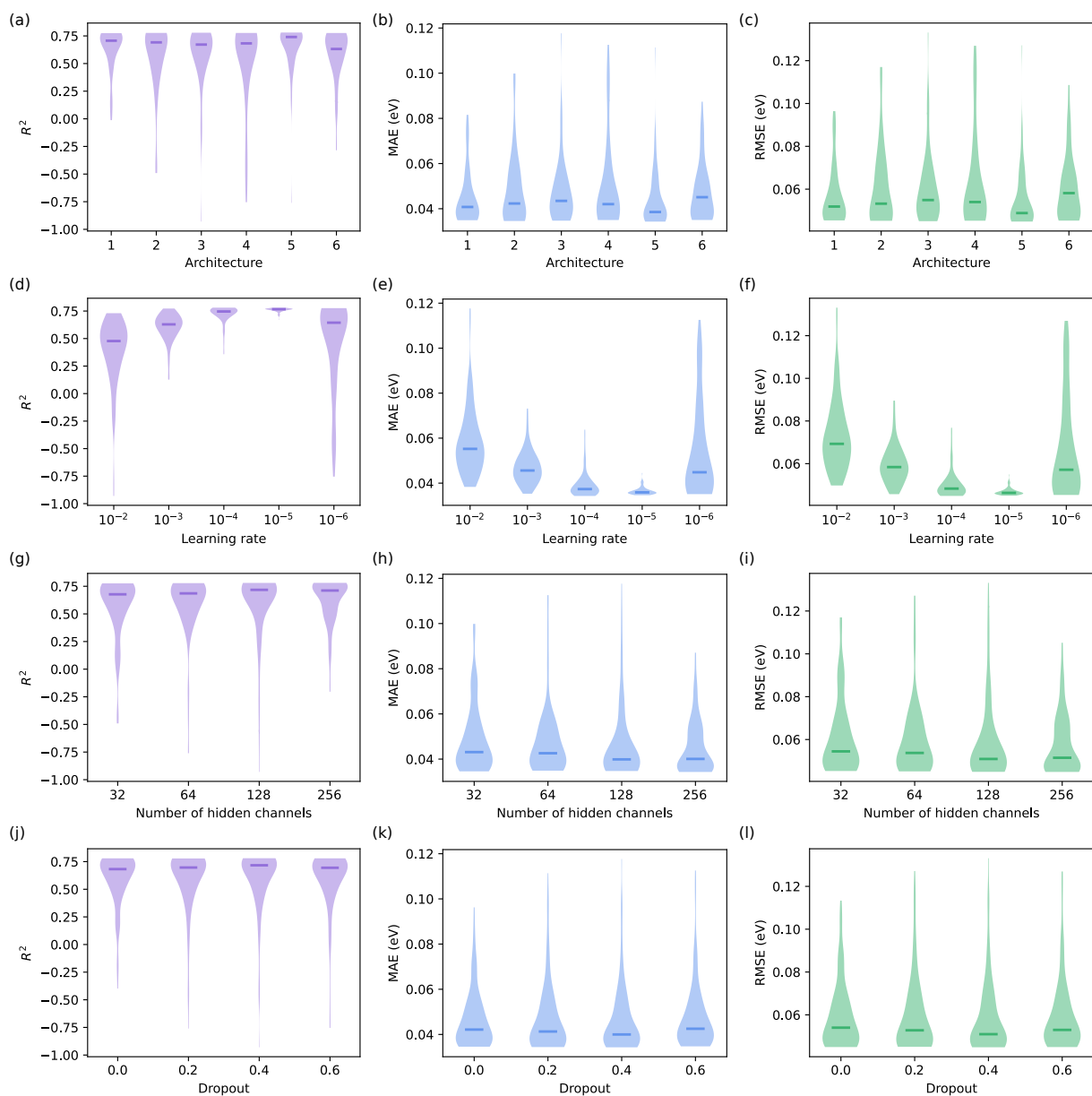
**Supplementary Figure 3:** Radial pair distribution function for equilibrium  $\text{Ag}_3\text{SBr}$ : (a) Ag–Ag, (b) S–S, (c) Br–Br, (d) Ag–S, (e) Ag–Br, and (f) S–Br. The dashed vertical line at 5.5 Å represents the selected cutoff radius for graph generation.



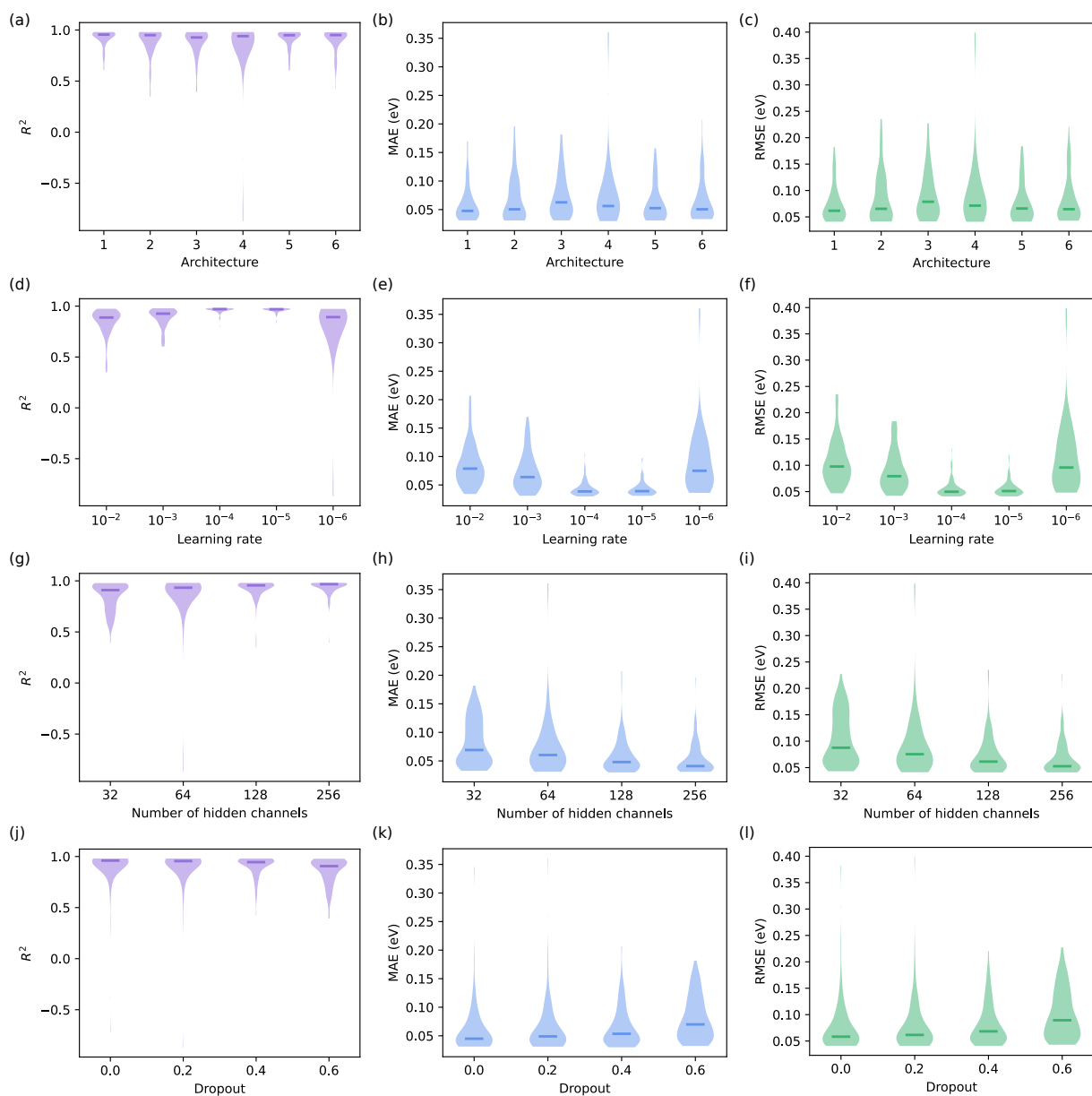
**Supplementary Figure 4:** Representation of the six architectures used, differing in the number of convolution and pooling layers and in the activation functions.



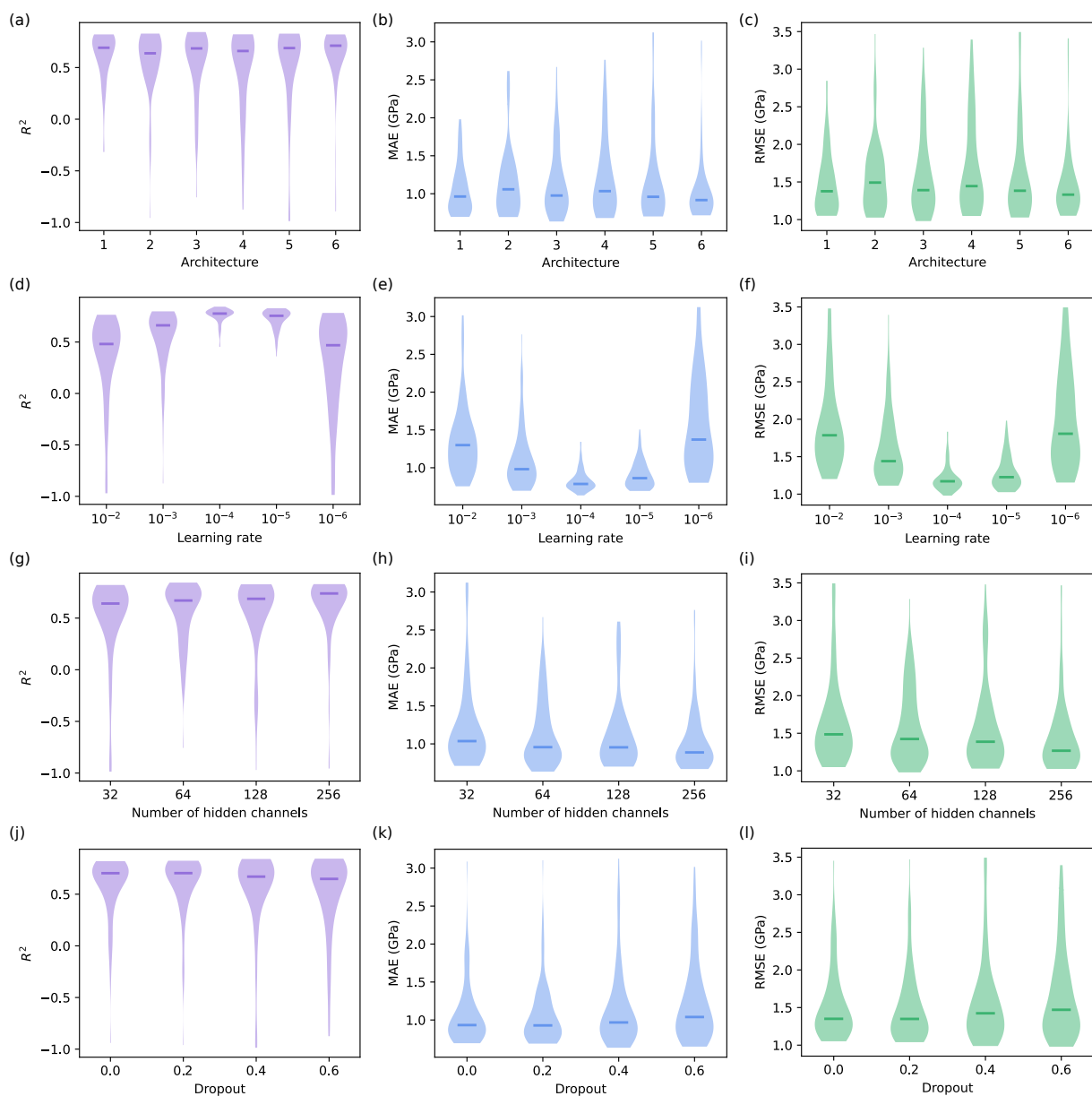
**Supplementary Figure 5:** Violin plots of GNN performance metrics ( $R^2$ , MAE, MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) dropout values, when predicting the energy per atom.



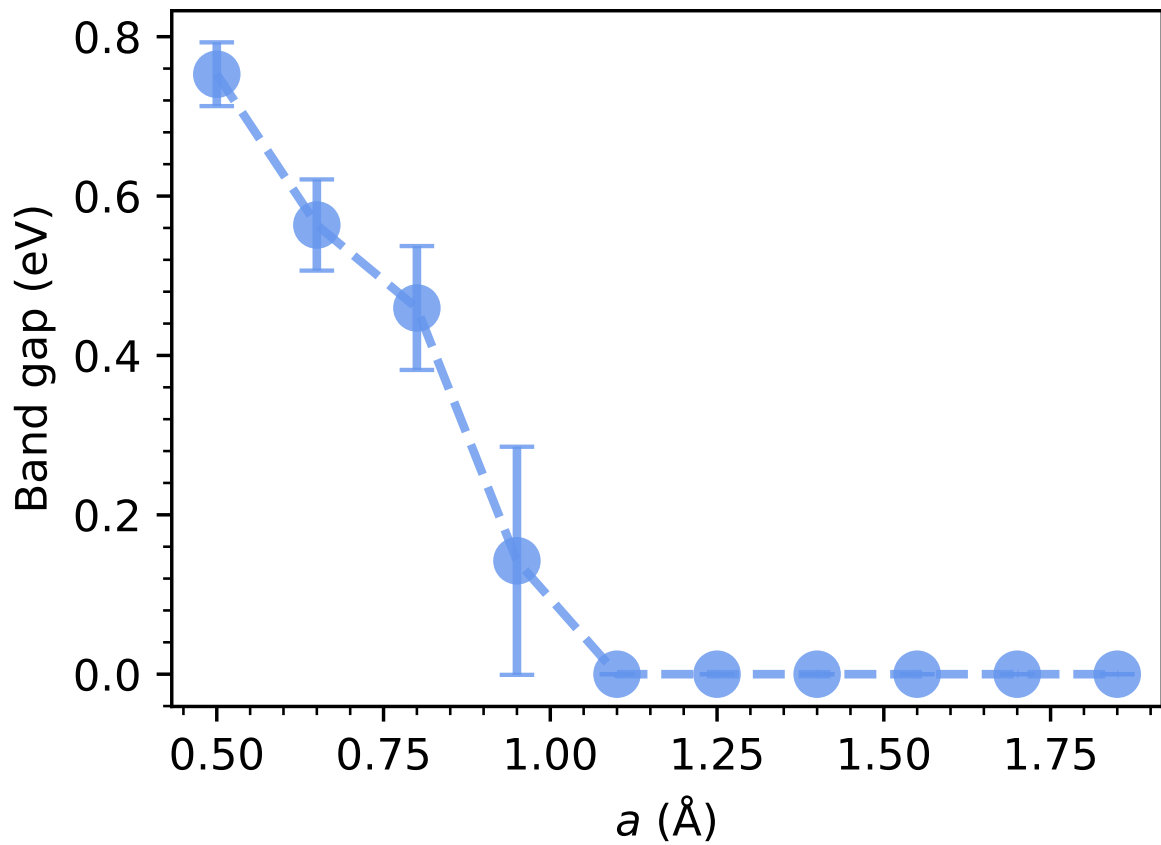
**Supplementary Figure 6:** Violin plots of GNN performance metrics ( $R^2$ , MAE, MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) dropout values, when predicting the band gap.



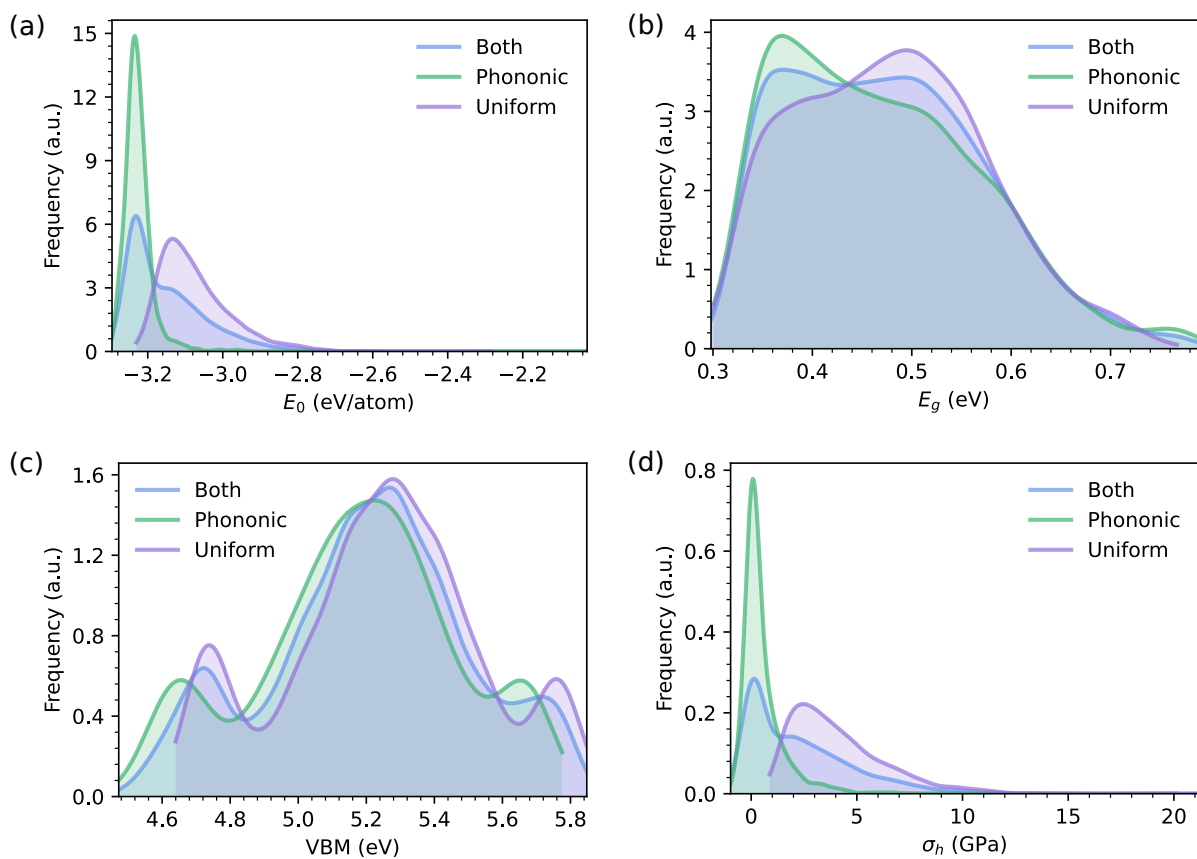
**Supplementary Figure 7:** Violin plots of GNN performance metrics ( $R^2$ , MAE, MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) dropout values, when predicting the valence band maximum.



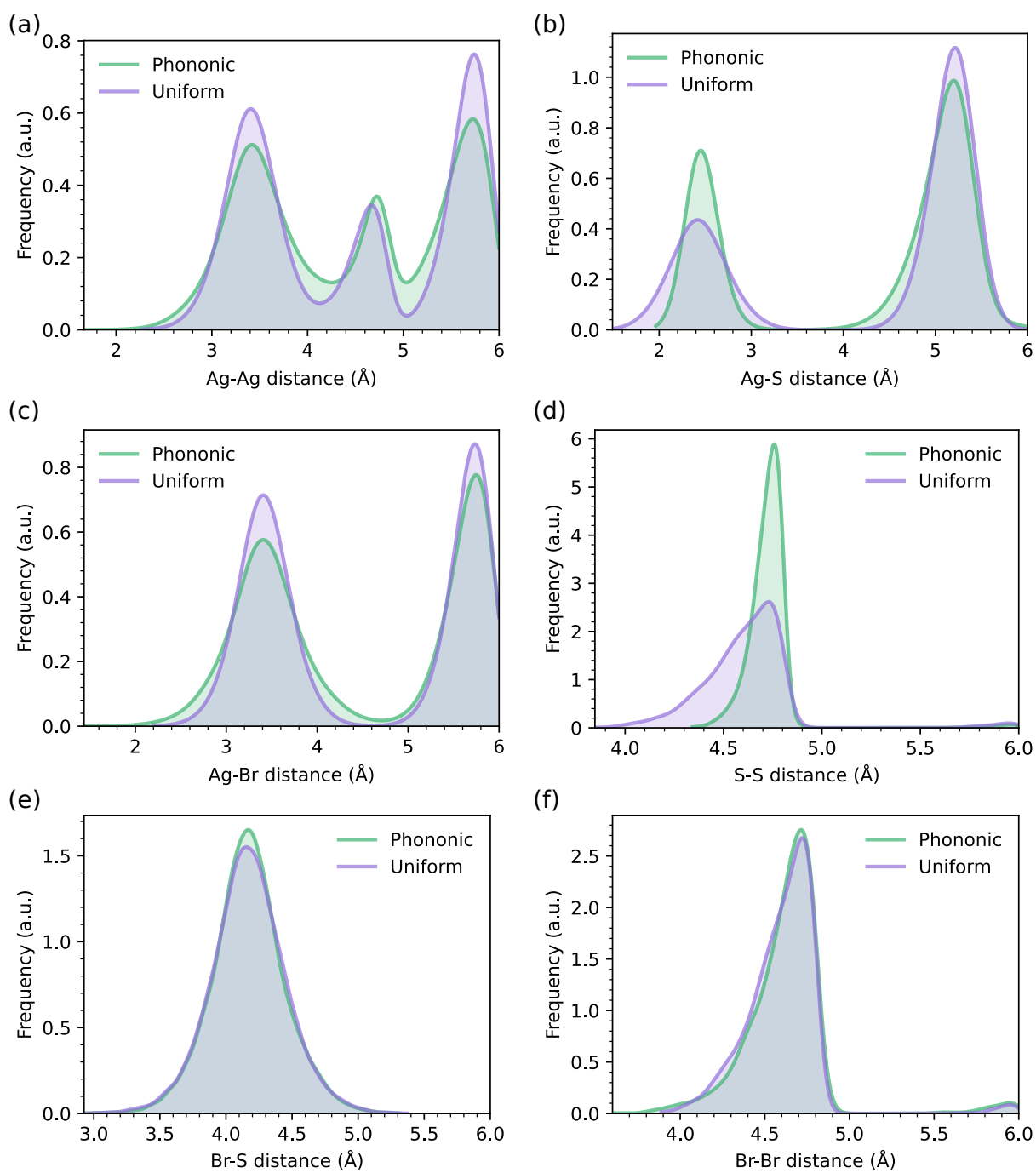
**Supplementary Figure 8:** Violin plots of GNN performance metrics ( $R^2$ , MAE, MSE; left to right) for different (a–c) architectures, (d–f) learning rates, (g–i) number of hidden channels, and (j–l) dropout values, when predicting the hydrostatic stress.



**Supplementary Figure 9:** Band gap calculated for randomly generated disordered configurations as a function of the atomic displacement amplitude. Results were obtained with DFT methods.



**Supplementary Figure 10:** 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).



**Supplementary Figure 11:** 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.

|               | Phonon-based |           |       | AIMD-based |           |       |
|---------------|--------------|-----------|-------|------------|-----------|-------|
|               | MAE (eV)     | RMSE (eV) | $R^2$ | MAE (eV)   | RMSE (eV) | $R^2$ |
| Train         | 0.029        | 0.037     | 0.89  | 0.036      | 0.045     | 0.79  |
| Validation    | 0.028        | 0.035     | 0.90  | 0.037      | 0.046     | 0.78  |
| Complementary | 0.044        | 0.063     | 0.68  | 0.058      | 0.071     | 0.53  |

**Supplementary Table I:** Performance metrics for GNN models trained on phonon and AIMD-based datasets. Performances were evaluated on the corresponding train and validation datasets, as well as on the other model’s dataset (“Complementary”).

## SUPPLEMENTARY DISCUSSION

Supplementary Fig. 1 describes in detail the method employed to generate graphs for crystal structures. Using a cutoff radius, we constructed graphs in which each node corresponds to one atom in the unit cell and stores relevant chemical features: atomic number, atomic mass, atomic radius, and electronegativity. A supercell was then generated large enough to contain all cutoff spheres centered on the atoms of the unit cell. An edge was defined whenever the Euclidean distance between two atoms was smaller than the chosen cutoff radius, and this distance was stored as the edge feature. Our algorithm ensures that each bond is counted only once. When an atom forms a bond with one of its periodic images, this connection is added as a self-loop (or self-edge).

The final graph consists of three tensors:

- Node tensor: stores node information with dimensions  $(n, 4)$ , where  $n$  is the total number of nodes.
- Edge tensor: stores edge information with dimensions  $(m, 1)$ , where  $m$  is the total number of edges.
- Adjacency tensor: stores the indices of nodes connected by each edge, with dimensions  $(m, 1)$ . Each element of this list is a tuple  $(i, j)$ , where  $i$  and  $j$  range from 1 to  $n$ .

In PyTorch Geometric, to make the edges undirected (i.e., without a preferred orientation), each edge is stored twice with reversed order in the adjacency list,  $(i, j)$  and  $(j, i)$ , both sharing the same Euclidean distance feature. This duplication is unnecessary for self-edges, which are undirected by definition,  $(i, i)$ .

Supplementary Fig. 2 shows the convergence test for the cutoff radius. To ensure that the chosen cutoff was sufficiently large, GNN models were trained using graphs generated with different cutoff radii, and the corresponding validation losses were evaluated. We observed convergence of the loss function for models trained with graphs constructed using a cutoff radius of 5.5 Å, which was therefore selected for all subsequent calculations.

Supplementary Fig. 3 displays the radial pair distribution function calculated for different pairs of atoms in equilibrium  $\text{Ag}_3\text{SBr}$ . The dashed vertical line indicates the selected cutoff radius for graph generation. From the figure it is clear that at least the first coordination shell for each atomic pair lies within the selected cutoff distance of 5.5 Å.

Supplementary Fig. 4 displays the six different GNN model architectures used in this work. These architectures differ in both the number and type of layers employed.

Supplementary Figs. 5–8 present the results of the hyperparameter study for each of the four predicted properties. A total of 480 different GNN models were trained per property, combining six architectures with various hyperparameter values: learning rate ( $10^{-2}$ ,  $10^{-3}$ ,  $10^{-4}$ ,  $10^{-5}$ ,  $10^{-6}$ ), number of hidden channels (32, 64, 128, 256), and dropout rate (0, 0.2, 0.4, 0.6). Results are shown using violin plots grouped by hyperparameter value, displaying model performance for three metrics: MAE, RMSE, and  $R^2$ . The horizontal dark line in each violin represents the mean metric value. Models yielding  $R^2 < -1$  were excluded, as such cases correspond to predictions worse than random guessing, typically indicating a failed training process.

Supplementary Fig. 9 shows the convergence test for the maximum atomic amplitude displacement of uniform random structures. The selected value for this study, 0.8 Å, ensures that all configurations contained in the randomly-generated dataset exhibit insulating behaviour.

Supplementary Fig. 10 shows the distributions of computed magnitudes for structures generated with the uniform and phonon-based distortion approaches, as well as for the combined dataset. Magnitudes such as the band gap exhibit similar distributions over the same domain, whereas others, such as the energy per atom, differ significantly, uniform distortions lead to higher-energy (less stable) configurations compared to phonon-based distortions.

Supplementary Fig. 11 presents the pairwise atomic distance distributions for atomically disordered  $\text{Ag}_3\text{SBr}$  configurations generated using both the phonon-based and random approaches. The results show similar ranges and density distributions for both datasets, indicating that although the two schemes sample different regions of the configuration space, the interatomic distances remain comparable.

Supplementary Table I reports the performance metrics for the GNN model trained on a phonon-based dataset and on an alternative dataset consisting of configurations extracted from AIMD simulations. Each dataset contains around 1,000 configurations obtained for  $\text{Ag}_3\text{SBr}$  and  $\text{Ag}_3\text{SI}$ . The AIMD simulations were conducted at  $T = 200, 400, \text{ and } 600$  K and lasted for about 60–70 ps each.

All scripts used in this work are available in the open-access GitHub repository: <https://github.com/polbeni/GNN-materials>.