PREDICTING DENSITY OF STATES VIA MULTI-MODAL TRANSFORMER

Namkyeong Lee[1†], **Heewoong Noh**[1†], **Sungwon Kim**[1], **Dongmin Hyun**[2], **Gyoung S. Na**[3], **Chanyoung Park**[1*]
[1] KAIST [2] POSTECH [3] KRICT
{namkyeong96,heewoongnoh,swkim,cy.park}@kaist.ac.kr
dm.hyun@postech.ac.kr,ngs0@krict.re.kr

ABSTRACT

The density of states (DOS) is a spectral property of materials, which provides fundamental insights on various characteristics of materials. In this paper, we propose a model to predict the DOS by reflecting the nature of DOS: *DOS determines the general distribution of states as a function of energy.* Specifically, we integrate the heterogeneous information obtained from the crystal structure and the energies via multi-modal transformer, thereby modeling the complex relationships between the atoms in the crystal structure, and various energy levels. Extensive experiments on two types of DOS, i.e., Phonon DOS and Electron DOS, with various real-world scenarios demonstrate the superiority of DOSTransformer. The source code for DOSTransformer is available at https://github.com/HeewoongNoh/DOSTransformer.

1 INTRODUCTION

Despite the recent progress of machine learning (ML) in materials science, most ML models developed in the field have been focused on material properties consisting of single-valued properties Kong et al. (2022), e.g., band gap energy Lee et al. (2016), formation energy Ward et al. (2016), and Fermi energy Xie & Grossman (2018). On the other hand, spectral properties are ubiquitous in materials science, characterizing various properties of materials, e.g., X-ray absorption, dielectric function, and electronic density of states Kong et al. (2022) (See Figure 1(a)).

The density of states (DOS), which is the main focus of this paper, is a spectral property that provides fundamental insights on various characteristics of materials, even enabling direct computation of single-valued properties Fung et al. (2022). For example, DOS is utilized as a feature of materials for analyzing the underlying reasons for changes in electrical conductivity Deringer et al. (2021). Moreover, band gaps and edge positions, which can be directly derived from DOS, are utilized to discover new photoanodes for solar fuel generation Singh et al. (2019); Yan et al. (2017). Consequently, investigating the ML capability for DOS prediction moves it one step closer to the fundamentals of materials science, thereby accelerating the materials discovery process However, credible computation of DOS requires expensive time/financial costs of exhaustively conducting experiments with expertise knowledge Chandrasekaran et al. (2019); Del Rio et al. (2020). Therefore, alternative algorithmic approaches for DOS calculation are necessary, whereas ML capabilities for learning such spectral properties of the crystal structure are relatively under-explored.

Existing studies for DOS prediction with ML models mainly focus on obtaining high-quality representations of crystal structures. Specifically, Chandrasekaran et al. (2019); Del Rio et al. (2020) predict DOS with multi-layered perceptrons (MLPs) given rule-based fingerprints of each grid point and atom, respectively. Inspired by the recent success of graph neural networks (GNNs) on a variety of tasks in biochemistry Gilmer et al. (2017); Stokes et al. (2020); Jiang et al. (2021), Fung et al. (2022) leverages GNNs to encode crystal structures to predict DOS with additional physical properties. Moreover, Chen et al. (2021) predicts phonon-DOS with euclidean neural networks Thomas et al. (2018); Kondor et al. (2018); Weiler et al. (2018), which by construction are equivariant to 3D rotations, translations, and inversion, aiming to capture a full crystal symmetry.

Despite their success, existing ML methods for DOS prediction overlook the nature of DOS calculation: *DOS determines the general distribution of states as a function of energy*. That is, DOS of the

*Corresponding author. † These authors contributed equally.

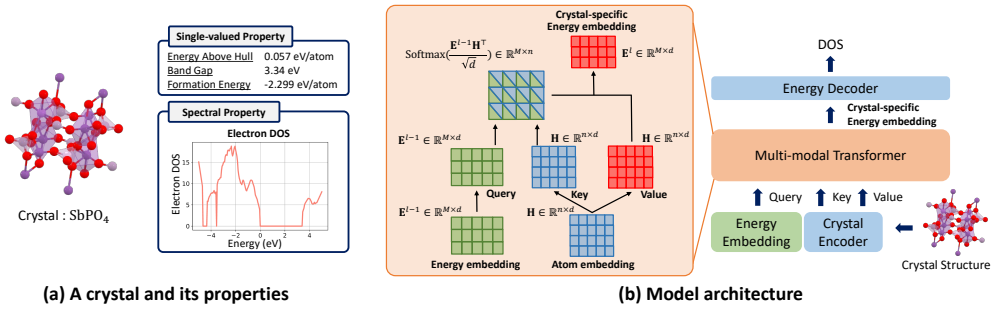

Figure 1: (a) Crystal structure and its various types of properties. (b) Overall model architecture.

crystal structure is determined by not only the structure itself but also the *energy levels*. Therefore, integrating heterogeneous signals from both the crystal structure and the energy is crucial for DOS prediction, which however has been overlooked by existing studies Chandrasekaran et al. (2019); Del Rio et al. (2020); Fung et al. (2022); Chen et al. (2021).

In this paper, we formulate the DOS prediction problem as a multimodal learning problem, which recently got a surge of interest from ML researchers in various domains thanks to its capability of extracting and relating information from heterogeneous data types Lin et al. (2015); Wang et al. (2016); Bayoudh et al. (2020); Baltrušaitis et al. (2018). Specifically, we propose a multimodal transformer model for DOS prediction, named DOSTransformer, which incorporates the crystal structure and the energy as heterogeneous modalities. Distinguished from exising studies, DOSTransformer learns embeddings of energy that are used for modeling complex relationships between the atoms in crystal structure and various energy levels through a cross-attention mechanism. By doing so, DOSTransformer obtains multiple representations for a single crystal structure according to various energy levels, enabling the prediction of a single DOS value on each energy level.

Our extensive experiments on two types of DOS, i.e., Phonon DOS and Electron DOS, and three data split strategies for real-world materials discovery, i.e., one in-distribution split (random split), and two out-of-distribution splits (split according to the number of atom species, and the crystal systems), demonstrate the superiority of DOSTransformer compared with previous methods. To the best of our knowledge, this is the first work to model the complex relationship between the crystal structure and various energy levels for predicting DOS of the crystal structure.

2 Preliminaries

Notations. Let $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ denote a crystal structure, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ represents the set of atoms, and $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of edges connecting the atoms in the crystal structure. Moreover, $\mathcal{G}$ is associated with a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times F}$ and an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ where $\mathbf{A}_{ij} = 1$ if and only if $(v_i, v_j) \in \mathcal{A}$ and $\mathbf{A}_{ij} = 0$ otherwise.

Task: Density of States Prediction. Given a set of crystals $\mathcal{D}_{\mathcal{G}} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ and a set of energies $\mathcal{D}_{\mathcal{E}} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_M\}$, our goal is to train a model $\mathcal{M}$ that predicts the DOS of a crystal structure given a set of energies, i.e., $\mathbf{Y}^i = \mathcal{M}(\mathcal{G}_i, \mathcal{D}_{\mathcal{E}})$, where $\mathbf{Y}^i \in \mathbb{R}^M$ is an M dimensional vector containing the DOS values of a crystal structure $\mathcal{G}_i$ at each energy $\mathcal{E}_1, \ldots, \mathcal{E}_M$, and $\mathbf{Y}^i_j \in \mathbb{R}$ is the DOS value of $\mathcal{G}_i$ at energy level $\mathcal{E}_j$.

3 Methodology

In this section, we introduce our proposed method named DOSTransformer, a novel DOS prediction framework that learns the complex relationship between the atoms in the crystal structure and various energy levels by utilizing a cross-attention mechanism of the multi-modal transformer. The overall model architecture is depicted in Figure 1 (b).

3.1 Crystal Encoder

Before modeling the pairwise interaction between the atoms and the energies, we first encode the crystal structure with GNNs to learn the representation of each atom, which contains not only the feature information but also the structural information. Formally, given a crystal structure $\mathcal{G} =$

$(\mathbf{X}, \mathbf{A})$, we generate an atom embedding matrix for the crystal structure as follows:

$$\mathbf{H} = \text{GNN}(\mathbf{X}, \mathbf{A}), \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{n \times d}$ is an atom embedding matrix for $\mathcal{G}$, whose i-th row indicates the representation of atom v_i, and we stack L' layers of GNNs. Among various GNNs, we adopt graph networks Battaglia et al. (2018) as our crystal encoder, which is a generalized and extended version of various GNNs.

3.2 Multi-modal Transformer

After obtaining the atom embedding matrix $\mathbf{H}$, we model the relationship between the atoms and various energy levels via a cross-attention mechanism of the multi-modal transformer. Specifically, we expect the multi-modal transformer to generate the energy-specific representation of the crystal by repeatedly reinforcing the energy representation with the crystal structure. To do so, we first introduce a learnable embedding matrix $\mathbf{E}^0 \in \mathbb{R}^{M \times d}$, whose j-th row, i.e., $\mathbf{E}^0_j$, indicates the embedding of energy $\mathcal{E}_j \in \mathcal{D}_{\mathcal{E}}$. Then, we present a cross-modal attention for fusing the information from the crystal structure into energy as follows:

$$\begin{aligned} \mathbf{E}^l &= \text{Cross-Attention}(\mathbf{Q}_{\mathbf{E}^{l-1}}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{M \times d} \\ &= \text{Softmax}(\frac{\mathbf{E}^{l-1}\mathbf{H}^\top}{\sqrt{d}})\mathbf{H}, \end{aligned} \tag{2}$$

where $l = 1, \ldots, L$ indicates the index number of the transformer layer. In contrast to conventional Transformers, which use learnable weight matrices for query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$, we directly employ the previously obtained energy embedding matrix $\mathbf{E}^{l-1}$ as the query matrix and the atom embedding matrix $\mathbf{H}$ as the key and value matrices. Based on the above cross-attention mechanism, we obtain the crystal-specific energy embedding $\mathbf{E}^l \in \mathbb{R}^{M \times d}$ by aggregating the information regarding the atoms in the crystal structure that was important at the given energy level. Consequently, the model learns the crystal-specific energy embedding matrix $\mathbf{E}^L \in \mathbb{R}^{M \times d}$ that reflects the complex relationship between the atoms in the crystal structure and various energy levels.

3.3 Energy Decoder

After obtaining the crystal-specific energy embedding matrix $\mathbf{E}^{L,i}$ of a crystal structure $\mathcal{G}_i$, the DOS value at each energy level $\mathcal{E}_j$, i,e., $\hat{\mathbf{Y}}^i_j$, is given as follows:

$$\hat{\mathbf{Y}}^i_j = \phi(\mathbf{E}^{L,i}_j + \alpha \cdot \mathbf{g}_i), \tag{3}$$

where $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^1$ is a parameterized MLP for predicting DOS from the given crystal-specific energy embedding of crystal structure $\mathcal{G}_i$ at energy level j, i.e., $\mathbf{E}^{L,i}_j$ and $\mathbf{g}_i \in \mathbb{R}^d$, which is a sum pooled representation of crystal $\mathcal{G}_i$, and α is a learnable parameter. Note that $\mathbf{E}^{L,i}_j$ indicates j-th row of energy embedding matrix $\mathbf{E}^{L,i}_j$. Finally, DOSTransformer is trained to minimize the root mean squared error loss $\mathcal{L}$ between the predicted target value $\hat{\mathbf{Y}}^i_j$ and the ground truth target value $\mathbf{Y}^i_j$, i.e., $\mathcal{L} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \sqrt{(\hat{\mathbf{Y}}^i_j - \mathbf{Y}^i_j)^2}$.

4 Experiments

4.1 Experimental Setup

Datasets. We use two datasets to comprehensively evaluate the performance of DOSTransformer, i.e., Phonon DOS and Electron DOS. We provide more details on datasets in Appendix A.1.

Evaluation Protocol. For Phonon DOS, we evaluate DOSTransformer with given data splits in a previous work Chen et al. (2021). For Electron DOS, we evaluate DOSTransformer in three data splits, i.e., one in-distribution split and two out-of-distribution splits. For in-distribution, we randomly split the dataset into train/valid/test of 80/10/10%. On the other hand, for out-of-distribution, we evaluate the model performance on the crystal structures that 1) contain a different number of atom species with the training set, and 2) belong to different crystal systems that were not included in the training set. Moreover, we predict the Fermi energy of the crystal structure based on the predicted DOS to evaluate how much physically meaningful DOS is predicted by the proposed method. We provide further details on data split and evaluation on Fermi energy in Appendix A.2.

Methods Compared. We mainly compare DOSTransformer to recently proposed state-of-the-art method, i.e., E3NN Chen et al. (2021). We also compare DOSTransformer to simple baseline methods, i.e., MLP and Graph Network Battaglia et al. (2018), which predicts the entire DOS sequence directly from the learned representation of the crystal structure. Moreover, to evaluate the effectiveness of the transformer layer that considers the relationship between the atoms and various energy levels, we integrate energy embeddings into baseline methods for DOS prediction as done in Equation 3. We provide more details on the implementation and compared methods in Appendix A.3 and A.4, respectively.

Evaluation Metrics. The performance of DOSTransformer is mainly evaluated in terms of RMSE and MAE following previous work Chen et al. (2021).

Table 1: Overall model performance.

	Energy	In-Distribution					Out-of-Distribution (Electron DOS)					
		Phonon DOS		Electron DOS			Scenario 1: # Atom species			Scenario 2: Crystal System		
		RMSE	MAE	RMSE	MAE	Fermi E.	RMSE	MAE	Fermi E.	RMSE	MAE	Fermi E.
MLP	✗	0.1719 (0.0006)	0.1131 (0.0001)	0.2349 (0.0008)	0.1829 (0.0009)	2.1314 (0.0908)	0.2655 (0.0025)	0.2043 (0.0022)	2.3852 (0.0916)	0.2584 (0.0026)	0.1984 (0.0028)	2.4863 (0.3264)
Graph Network	✗	0.1650 (0.0030)	0.1067 (0.0033)	0.1529 (0.0011)	0.1152 (0.0008)	1.5693 (0.0321)	0.2225 (0.0005)	0.1676 (0.0010)	1.9237 (0.1501)	0.2041 (0.0013)	0.1523 (0.0009)	1.7790 (0.0832)
E3NN	✗	0.1356 (0.0019)	0.0809 (0.0014)	0.1514 (0.0013)	0.1108 (0.0009)	1.5790 (0.0415)	0.2104 (0.0007)	0.1524 (0.0007)	1.7308 (0.0357)	0.1858 (0.0006)	0.1349 (0.0004)	1.8642 (0.0169)
MLP	✓	0.1445 (0.0000)	0.0965 (0.0000)	0.1604 (0.0011)	0.1228 (0.0009)	1.8488 (0.0636)	0.2080 (0.0006)	0.1566 (0.0004)	1.9343 (0.0440)	0.1905 (0.0010)	0.1445 (0.0008)	2.2635 (0.0401)
Graph Network	✓	0.1316 (0.0016)	0.0900 (0.0008)	0.1344 (0.0006)	0.1000 (0.0007)	1.5459 (0.0276)	0.1958 (0.0008)	0.1451 (0.0007)	1.7543 (0.0568)	0.1759 (0.0009)	0.1297 (0.0008)	1.7548 (0.0889)
E3NN	✓	**0.1262** (0.0005)	**0.0765** (0.0008)	0.1498 (0.0008)	0.1109 (0.0008)	1.6158 (0.0311)	0.2072 (0.0005)	0.1540 (0.0016)	1.9004 (0.1281)	0.1842 (0.0005)	0.1348 (0.0007)	1.8819 (0.0732)
DOSTransformer	✓	0.1283 (0.0017)	0.0786 (0.0013)	**0.1283** (0.0005)	**0.0918** (0.0006)	**1.4387** (0.0221)	**0.1918** (0.0006)	**0.1373** (0.0005)	**1.6159** (0.0608)	**0.1722** (0.0013)	**0.1231** (0.0012)	**1.7267** (0.0672)

4.2 Experimental Results

The experimental results on two datasets with various evaluation protocols are given in Table 1. We have the following observations: **1)** Comparing the baseline methods that overlook the energy levels (i.e., Energy ✗) with their counterparts that incorporate the energy and the crystal structure as heterogeneous modalities through the energy embeddings (i.e., Energy ✓), we find out that using the energy embeddings consistently enhances the model performance. This indicates that making predictions on each energy-level is crucial for DOS prediction, which also aligns with the domain knowledge of materials science, i.e., DOS determines the general distribution of states as a function of energy. **2)** On the other hand, DOSTransformer outperforms previous methods that do not consider the complex relationships between the atoms in crystal structure and various energy levels. This implies that naively integrating the energy information cannot fully benefit from the energy information. We further analyze the model performance in terms of the out-of-distribution scenarios, i.e., different atom species numbers and the crystal systems in Appendix A.5.1. **3)** Moreover, regarding the capability of predicting the Fermi energy, we observe that predicting the Fermi energy based on the DOS predicted by DOSTransformer consistently outperforms that of baseline methods. This indicates that DOSTransformer predicts physically meaningful DOS, which can further accelerate the materials discovery process. **4)** In the case of Phonon DOS, DOSTransformer performs on par with E3NN with energy embeddings. This is because the dataset contains a limited number of crystals (1,522 species) compared with the Electron DOS dataset (38,889 species). However, considering that Electron DOS is much more complex than Phonon DOS in a variety of ways Kong et al. (2022), and that we are interested in a general prediction method that can be applied to various types of crystals, we argue that DOSTransformer is practical in the real-world application.

5 Conclusion

In this paper, we propose DOSTransformer, which predicts the DOS of a crystal structure by modeling the complex relationships between the atoms in the crystal structure and various energy levels. Extensive experiments verify that incorporating energy information is crucial in predicting the DOS of a crystal structure, and modeling the complex relationship via cross-attention can further improve the model performance.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00077), and the core KRICT project from the Korea Research Institute of Chemical Technology (SI2051-10).

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Khaled Bayoudh, Fayçal Hamdaoui, and Abdellatif Mtibaa. Hybrid-covid: a novel hybrid 2d/3d cnn based on cross-domain adaptation approach for covid-19 screening from chest x-ray images. *Physical and engineering sciences in medicine*, 43(4):1415–1431, 2020.

Anand Chandrasekaran, Deepak Kamal, Rohit Batra, Chiho Kim, Lihua Chen, and Rampi Ramprasad. Solving the electronic structure problem with machine learning. *npj Computational Materials*, 5(1):1–7, 2019.

Zhantao Chen, Nina Andrejevic, Tess Smidt, Zhiwei Ding, Qian Xu, Yen-Ting Chi, Quynh T Nguyen, Ahmet Alatas, Jing Kong, and Mingda Li. Direct prediction of phonon density of states with euclidean neural networks. *Advanced Science*, 8(12):2004214, 2021.

Beatriz G Del Rio, Christopher Kuenneth, Huan Doan Tran, and Rampi Ramprasad. An efficient deep learning scheme to predict the electronic structure of materials and molecules: The example of graphene-derived allotropes. *The Journal of Physical Chemistry A*, 124(45):9496–9502, 2020.

Volker L Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A Drabold, and Stephen R Elliott. Origins of structural and electronic transitions in disordered silicon. *Nature*, 589(7840):59–64, 2021.

Victor Fung, Panchapakesan Ganesh, and Bobby G Sumpter. Physically informed machine learning prediction of electronic density of states. *Chemistry of Materials*, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

F Illas, I de PR Moreira, JM Bofill, and M Filatov. Extent and limitations of density-functional theory in describing magnetic systems. *Physical Review B*, 70(13):132414, 2004.

Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. 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):1–23, 2021.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B Neaton, Carla P Gomes, and John M Gregoire. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nature communications*, 13(1):1–12, 2022.

Joohwi Lee, Atsuto Seko, Kazuki Shitara, Keita Nakayama, and Isao Tanaka. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B*, 93(11):115104, 2016.

Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5007–5015, 2015.

Guido Petretto, Shyam Dwaraknath, Henrique PC Miranda, Donald Winston, Matteo Giantomassi, Michiel J Van Setten, Xavier Gonze, Kristin A Persson, Geoffroy Hautier, and Gian-Marco Rignanese. High-throughput density-functional perturbation theory phonons for inorganic materials. *Scientific data*, 5(1):1–12, 2018.

Arunima K Singh, Joseph H Montoya, John M Gregoire, and Kristin A Persson. Robust and synthesizable photocatalysts for co2 reduction: a data-driven materials discovery. *Nature communications*, 10(1):1–9, 2019.

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5005–5013, 2016.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Qimin Yan, Jie Yu, Santosh K Suram, Lan Zhou, Aniketa Shinde, Paul F Newhouse, Wei Chen, Guo Li, Kristin A Persson, John M Gregoire, et al. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proceedings of the National Academy of Sciences*, 114(12):3040–3043, 2017.

A APPENDIX

A.1 DATASETS

In this section, we provide further details on the dataset used during training.

A.1.1 PHONON DOS

We use the **Phonon DOS** dataset following the instructions of the official Github repository[1] of a previous work Chen et al. (2021). This dataset contains 1,522 crystals whose phonon DOS is calculated from density functional perturbation theory (DFPT) by a previous work Petretto et al. (2018). We use the data splits provided in the Github repository to evaluate model performance in the Phonon DOS dataset.

[1] `https://github.com/zhantaochen/phonondos_e3nn`

A.1.2 ELECTRON DOS

We also use **Electron DOS** dataset that contains a further variety of crystal structures during training. Electron DOS dataset consists of the materials and its electron DOS information that are collected from Materials Project (MP) [2].

Data Preprocessing. In MP dataset, we exclude the materials that are tagged to include magnetism because the DOS of magnetism materials is not accurate to be directly used for training machine learning models Illas et al. (2004). We consider an energy grid of 201 points ranging from -5 to 5 eV with respect to the band edges with 50 meV intervals and the Fermi energy is all set to 0 eV on this energy grid. Moreover, we normalize the DOS of each material to be in the range between 0 and 1. That is, the maximum and minimum value for each DOS is 1 and 0, respectively, for all materials. Moreover, we smooth the DOS values with the Savitzky-Golay filter with the window size of 17 and polyorder of 1 using scipy library following a previous work Chen et al. (2021).

Data Statistics. As described in the main manuscript, we further evaluate the model performance in two out-of-distribution scenarios: **Scenario 1**: regarding the number of atom species, and **Scenario 2**: regarding the crystal systems. We provide detailed statistics of the number of crystals for each scenario in Table 2 and Table 3.

Table 2: The number of crystals according to the number of atom species (Scenario 1).

	Unary (1)	Binary (2)	Ternary (3)	Quaternary (4)	Quinary (5)	Senary (6)	Septenary (7)	Total
# Crystals	386	9,034	21,794	5,612	1,750	279	34	38,889

Table 3: The number of crystals according to different crystal systems (Scenario 2).

	Cubic	Hexagonal	Tetragonal	Trigonal	Orthorhombic	Monoclinic	Triclinic	Total
# Crystals	8,385	3,983	5,772	3,964	8,108	6,576	2,101	38,889

A.2 EVALUATION PROTOCOL

Phonon DOS. As described in the main manuscript, we evaluate the model performance based on the data splits given in a previous work Chen et al. (2021).

Electron DOS. On the other hand, for the Electron DOS dataset, we use different dataset split strategies for each scenario. For the in-distribution setting, we randomly split the dataset into train/valid/test of 80/10/10%. On the other hand, for the out-of-distribution setting, we split the dataset regarding the structure of the crystals. For both scenarios, we generate training sets with simple crystal structures and a valid/test set with more complex crystal structures. More specifically, in the scenario 1 (different number of atom species, i.e., # Atom species in Table 1), we use binary and ternary crystals as training data and Unary, Quaternary, and Quinary crystals as valid and test data. In the scenario 2 (different crystal systems, i.e., Crystal System in Table 1), we use Cubic, Hexagonal, Tetragonal, Trigonal, and Orthorhombic crystals as training set and Monoclinic and Triclinic as valid and test set. Please refer to Table 2 and Table 3 for detailed statistics for each type of crystal structure.

Fermi Energy. We predict the Fermi energy of the crystal structures based on the DOS predicted by the proposed method to evaluate how much physically meaningful DOS is predicted. To do so, given the ground truth DOS, we first train a four-layered MLP with a non-linearity in each layer to predict the Fermi energy of a crystal structure. Then, based on the obtained MLP weights, we predict the Fermi energy given the predicted DOS as the input, and calculate the RMSE. By doing so, we can evaluate how physically meaningful DOS is obtained from each model.

A.3 IMPLEMENTATION DETAILS

In this section, we provide implementation details of DOSTransformer.

[2] `https://materialsproject.org/`

Graph Neural Networks. Our graph neural networks consist of two parts, i.e., encoder and processor. Encoder learns the initial representation of atoms and bonds, while the processor learns to pass the messages across the crystal structure. More formally, given an atom v_i and the bond e_{ij} between atom v_i and v_j, node encoder ϕ_{node} and edge encoder ϕ_{edge} outputs initial representations of atom v_i and bond e_{ij} as follows:

$$\mathbf{h}_i^0 = \phi_{node}(\mathbf{X}_i), \quad \mathbf{b}_{ij}^0 = \phi_{edge}(\mathbf{B}_{ij}), \tag{4}$$

where $\mathbf{X}$ is the atom feature matrix whose i-th row indicates the input feature of atom v_i, $\mathbf{B} \in \mathbb{R}^{n \times n \times F_e}$ is the bond feature tensor with F_e features for each bond. With the initial representations of atoms and bonds, the processor learns to pass messages across the crystal structure and update atoms and bonds representations as follows:

$$\mathbf{b}_{ij}^{l+1} = \psi_{edge}^l(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{b}_{ij}^l), \quad \mathbf{h}_i^{l+1} = \psi_{node}^l(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \mathbf{b}_{ij}^{l+1}), \tag{5}$$

where $\mathcal{N}(i)$ is the neighboring atoms of atom v_i, ψ is two layer MLPs with non-linearity, and $l = 0, \ldots, L'$. Note that $\mathbf{h}_i^{L'}$ is equivalent to the i-th row of the atom embedding matrix $\mathbf{H}$ in Equation 1.

Model Training. In all our experiments, we use the AdamW optimizer for model optimization. For all the tasks, we train the model for 1,000 epochs with early stopping applied if the best validation loss does not change for 50 consecutive epochs.

Hyperparameter Tuning. Detailed hyperparameter specifications are given in Table 4. For the hyperparameters in DOSTransformer, we tune them in certain ranges as follows: number of message passing layers in GNN L' in $\{2, 3, 4\}$, number of transformer layers for cross attention L in $\{2, 3, 4\}$, hidden dimension d in $\{64, 128, 256\}$, learning rate η in $\{0.0001, 0.0005, 0.001\}$, and batch size B in $\{1, 4, 8\}$. We use the sum pooling to obtain the crystal i's representation, i.e., $\mathbf{g}_i$. We report the test performance when the performance on the validation set gives the best result.

Table 4: Hyperparameter specifications of DOSTransformer.

		# Message Passing Layers (L')	# Transformer Layers (L)	Hidden Dim (d)	Learning rate (η)	Batch Size (B)
Phonon DOS		3	2	256	0.001	1
Electron DOS	Random	3	2	256	0.0001	8
	Scenario 1: # Atom species	3	2	256	0.0001	8
	Scenario 2: Crystal Systems	3	2	256	0.0005	8

A.4 METHODS COMPARED

In this section, we provide further details on the methods that are compared with DOSTransformer during experiments.

MLP. We first encode the atoms in a crystal with an MLP. Then, we obtain the representation of crystal i, i.e., $\mathbf{g}_i$, by sum pooling the representations of its constituent atoms. With the crystal representation, we predict DOS with an MLP predictor, i.e., $\hat{\mathbf{Y}}^i = \phi'(\mathbf{g}_i)$, where $\phi' : \mathbb{R}^d \to \mathbb{R}^{201}$.

On the other hand, when we incorporate energy embeddings into the MLP, we predict DOS for each energy j with a learnable energy embedding $\mathbf{E}_j^0$ and obtained crystal representation $\mathbf{g}_i$, i.e., $\hat{\mathbf{Y}}_j^i = \phi(\mathbf{E}_j^0 + \alpha \cdot \mathbf{g}_i)$, where $\phi : \mathbb{R}^d \to \mathbb{R}^1$ is a parameterized MLP.

Graph Network. We first encode the atoms in a crystal with a graph network. As done for MLP, we obtain the representation of crystal i, i.e., $\mathbf{g}_i$, by sum pooling the representations of its constituent atoms. With the crystal representation, we predict the DOS with an MLP predictor, i.e., $\hat{\mathbf{Y}}^i = \phi'(\mathbf{g}_i)$, where $\phi' : \mathbb{R}^d \to \mathbb{R}^{201}$. Note that the only difference with MLP is that the atom representations are obtained through the message passing scheme. We also compare the vanilla graph networks with the one that incorporates the energy information by integrating the energy information with initialized energy embeddings as we have done in MLP.

E3NN. For E3NN, we use the official code published by the authors[3], which implements equivariant neural networks with E3NN python library[4]. After obtaining the crystal representation $\mathbf{g}_i$, all other procedures have been done in the same manner with other baseline models, i.e., MLP and Graph Network.

A.5 Additional Experiments

A.5.1 Model Performance Analysis

In this section, we provide detailed analyses on the model's prediction in the out-of-distribution scenarios. We have following observations: **1)** We observe that DOSTransformer consistently outperforms in both out-of-distribution scenarios, which demonstrates the superiority of DOSTransformer. **2)** The performance of all the compared models generally degrades as the crystal structure gets more complex. That is, models perform worse in Quinary crystals than in Quarternary crystals, and worse in Triclinic crystals than in Monoclinic crystals. **3)** On the other hand, it is not the case in Unary crystal. This is because only one type of atom repeatedly appears in the crystal structure, which cannot give enough information to the model. However, DOSTransformer also makes comparably accurate predictions in the Unary materials by modeling the complex relationship between the atoms and various energy levels.

Table 5: Model performance in Out-of-Distribution scenarios.

Data split strategy		Scenario 1: # Atom species						Scenario 2: Crystal System			
		Unary		Quarternary		Quinary		Monoclinic		Triclinic	
	Energy	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
MLP	✗	0.3502 (0.0062)	0.2992 (0.0063)	0.2531 (0.0027)	0.1925 (0.0023)	0.2850 (0.0037)	0.2192 (0.0032)	0.2550 (0.0022)	0.1959 (0.0022)	0.2699 (0.0029)	0.2070 (0.0027)
Graph Network	✗	0.2717 (0.0039)	0.2274 (0.0034)	0.2124 (0.0014)	0.1587 (0.0019)	0.2464 (0.0012)	0.1847 (0.0014)	0.1996 (0.0010)	0.1485 (0.0008)	0.2194 (0.0011)	0.1649 (0.0013)
E3NN	✗	0.2013 (0.0035)	0.1610 (0.0034)	0.2013 (0.0009)	0.1443 (0.0011)	0.2426 (0.0014)	0.1767 (0.0012)	0.1805 (0.0003)	0.1305 (0.0004)	0.2021 (0.0009)	0.1488 (0.0006)
MLP	✓	0.2020 (0.0007)	0.1685 (0.0005)	0.1997 (0.0014)	0.1492 (0.0009)	0.2365 (0.0006)	0.1782 (0.0009)	0.1856 (0.0007)	0.1406 (0.0003)	0.2067 (0.0010)	0.1578 (0.0008)
Graph Network	✓	0.1913 (0.0023)	0.1585 (0.0021)	0.1880 (0.0011)	0.1379 (0.0004)	0.2237 (0.0016)	0.1663 (0.0012)	0.1709 (0.0004)	0.1257 (0.0003)	0.1911 (0.0013)	0.1423 (0.0011)
E3NN	✓	0.1937 (0.0021)	0.1562 (0.0020)	0.1985 (0.0005)	0.1462 (0.0018)	0.2383 (0.0016)	0.1785 (0.0023)	0.1787 (0.0005)	0.1302 (0.0005)	0.2012 (0.0013)	0.1491 (0.0010)
DOSTransformer	✓	**0.1792** (0.0037)	**0.1461** (0.0034)	**0.1846** (0.0011)	**0.1311** (0.0014)	**0.2188** (0.0008)	**0.1569** (0.0014)	**0.1668** (0.0006)	**0.1188** (0.0008)	**0.1880** (0.0020)	**0.1363** (0.0014)

A.5.2 Model Training and Inference Time

In this section, to verify the efficiency of DOSTransformer, we compare the training and inference time of the methods during the experiment in Table 6. DOSTransformer and E3NN take similar time per training epoch on the Phonon DOS dataset, while E3NN requires much more time per training epoch on the Electron DOS dataset. This is because the Electron DOS dataset contains a much more diverse and complex crystal structure compared to the Phonon DOS dataset, requiring more time to learn equivariant representations for the structure. This demonstrates the practicality of DOSTransformer in real-world applications compared with E3NN.

[3] https://github.com/ninarina12/phononDoS_tutorial
[4] https://docs.e3nn.org/en/latest/index.html

Table 6: Training and inference time per epoch for each dataset (sec/epoch).

	Energy	Training		Inference	
		Phonon DOS	Electron DOS	Phonon DOS	Electron DOS
MLP	✗	4.12	23.79	1.36	2.66
Graph Network	✗	16.20	59.98	1.73	3.48
E3NN	✗	21.14	140.06	3.61	9.05
MLP	✓	4.73	27.74	1.50	2.84
Graph Network	✓	17.37	66.72	1.95	3.99
E3NN	✓	22.68	149.39	3.83	9.84
DOSTransformer	✓	22.09	82.00	2.11	4.88