Property Prediction of Functional Organic Molecular Crystals with Graph Neural Networks

Dana O'Connor
danao@psc.edu
AI/Big Data Group, Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

Paola A. Buitrago
AI/Big Data Group, Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

ABSTRACT

Predicting the properties of molecular crystals is imperative to the field of materials design. In lieu of alternative methods, advances in machine learning have made it possible to predict the properties of materials before synthesis. This is especially important for organic semiconductors (OSCs) that are prone to exhibit polymorphism, as this phenomenom can impact the properties of a system, including the bandgap in OSCs. While graph neural networks (GNNs) have shown promise in predicting the bandgap in OSCs, few studies have considered the impact of polymorphism on their performance. Using the MatDeepLearn framework, we examine five different graph convolution layers of ALIGNN, GATGNN, CGCNN, MEGNet, and SchNet, which all have graph convolutions implemented in torch geometric. A dataset of functional organic molecular crystals is extracted from the OCELOT database, which has calculated density functional theory (DFT) values for the bandgap as well as several sets of polymorphs. The trained models are then evaluated on several test cases including the polymorphs of ROY. In future work we plan to examine the impact of graph representations on the performance of these models in the case of predicting properties of polymorphic OSCs.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Physical sciences and engineering**.

KEYWORDS

Artifical intelligence, graph neural networks, GNN, Bridges-2, materials science, polymorphs, polymorphism, ALIGNN, CGCNN, GATGNN, MEGNet, SchNet, convolution, property prediction

ACM Reference Format:
Dana O'Connor and Paola A. Buitrago. 2024. Property Prediction of Functional Organic Molecular Crystals with Graph Neural Networks. In *Practice and Experience in Advanced Research Computing (PEARC '24), July 21–25, 2024, Providence, RI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626203.3670584

PEARC '24, July 21–25, 2024, Providence, RI, USA

ACM ISBN 979-8-4007-0419-2/24/07
https://doi.org/10.1145/3626203.3670584

1 INTRODUCTION

Organic semiconductors (OSCs) have a wide array of applications in electronics due to their great versatility, novel properties, and low environmental impact[11]. These OSCs are often deployed in the form of molecular crystals. While these materials offer a rich design space, traditional "trial and error" discovery methods can be costly in terms of time and resources. Furthermore, the resulting candidate may not have desirable properties. In response to this, materials informatics (MI) and machine learning (ML) have shown great promise at enhancing materials design by feeding and training models to screen candidate materials[2, 16].

The field of materials informatics has skyrocketed in the past decade. Databases such as the Organic Materials Database,[3], the Cambridge Structural Database (CSD),[8] Organic Crystals in Electronic and Light-Oriented Technologies (OCELOT),[1] and the Materials Project[9] offer a large coverage of known chemical space and have proven ripe for data mining. Some of these databases also include properties of interest either experimentally measured[8] or calculated with density functional theory (DFT) or quantum chemical calculations[1, 3]. Previous work has shown the success of these databases in MI, with Padula *et al.* mining the CSD to screen for singlet fission candidates using the CSD python API[16]. Additionally, these databases are also being used as fodder for training ML models[2].

Graph neural networks (GNNs) have shown great promise in recent years for property prediction when trained on these databases. GNNs have made a splash in the materials community since the debut of the crystal graph convolution neural network (CGCNN). GNNs have been used in previous work to predict the bandgap of organic molecular crystals, a key property in OSCs.[6, 19] Taniguchi *et al.* recently utilized the MatDeepLearn (MDL)[6] framework to show how different graph convolution layers can impact the performance of GNNs on predicting the bandgap of materials extracted from the OMDB.[19] They found that the MEGNet model exhibited the best performance, with the lowest MAE of 0.240 eV.

However, they found that their model struggled with two polymorphic test cases. This is important because polymorphism, the ability of a molecule to crystallize in more than one solid form, is common in organic molecular crystals. Polymorphism can arise from crystallization conditions, such as solvent, temperature, and pressure, or kinetics. Polymorphism can impact the physichochemical properties of a material, including the bandgap. Hence, it is crucial that any property prediction method used is able to account for and differentiate between polymorphs of OSCs. For the polymorphs of ROY and BP4VA, the MEGNet model had a MAE of 0.40 and 0.52 eV, respectively, which is higher than that of the MAE on the OMDB. Taniguchi *et al.* noted that more recent graph convolution layers, such as ALIGNN or M3GNet, may lead to enhanced prediction accuracy overall. Furthermore, they hypothesized that the model struggled with polymorphism because their training data, extracted from the OMDB, did not contain any polymorphs. They expected that the inclusion of polymorphs in the training data would lead to enhanced prediction accuracy for these systems[19].

In this work, we present preliminary results of our adaptations to the MatDeepLearn framework[6]. First, we implemented two new graph convolution layers available in torch geometric: the edge gated graph convolution of ALIGNN[5] and the graph attention graph convolution of GATGNN[13]. Second, in contrast to Taniguchi *et al.* we train these GNNs on the OCELOT dataset which contains several sets of polymorphs[1]. Third, we evaluate the performance of the trained GNNs on ROY, a well known challenge to the problem of polymorphism. We perform density functional theory calculations on the polymorphs of ROY extracted from the CSD with the same settings as those used in the OCELOT database[1] to obtain the predicted bandgap. The bandgaps of the polymorphs of ROY vary over a narrow range, so this should provide a stringent test case for our models. In future work we hope to examine the impact of graph representations on model performance, both in general and for polymorphic systems.

2 METHODS

2.1 Data

The OCELOT database is made up of molecular crystals extracted from the CSD as well as contributions from the community[1]. In this work, we chose to only look at publicly available data, i.e. those extracted from the CSD. The bandgap of the crystal structures is calculated via density functional theory (DFT) within the Vienna Ab-Initio Simulation package (VASP)[10] using the generalized gradient approximation of Perdew, Burke, and Ernzerhof (PBE)[17] paired with the Grimme D3 dispersion correction[7].

The bandgap distribution of the OCELOT database is shown in Figure 1a. As can be seen, the average bandgap is 2.31 eV, which is fairly similar to that of Olsthorn *et al.*, who created a database of crystal structures mined from the OMDB[15]. The dataset contains the calculated bandgaps as the distance between the minimum energy of the lowest conduction band and the maximum energy of the highest valence band independently. There are 9479 data points as compared to 10472 in the work of Taniguchi *et al.*[19]. The OCELOT dataset also contains 476 sets of polymorphs, with 1248 polymorphs total, as seen in Figure 1b. We used a train:validation:test split of 85:5:10, similar to both Taniguchi *et al.*[19] and Fung *et al.*[6]. This leads to a training, validation, and testing set of 8056, 474, and 948 samples, respectively.

2.1.1 *ROY.*

5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY) has 11 polymorphs that are available in the CSD. These crystals were extracted from the CSD and then optimized with PBE+D3 using the GPAW[14] and D3[7] codes available from the atomic simulation environment[12] with the same settings used to produce the OCELOT database[1].

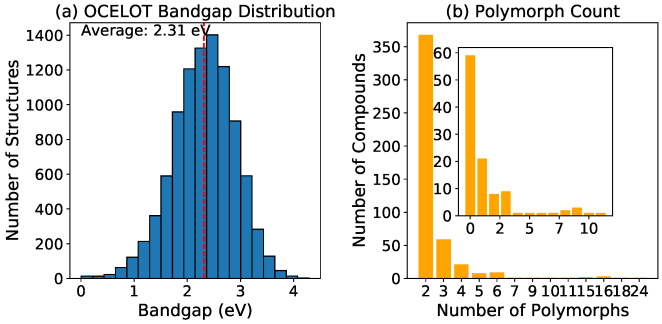

Figure 1: The (a) bandgap distribution and (b) polymorph count of the OCELOT database. For (b), the x-axis tells how many polymorphs are in each set of polymorphs and the y-axis tells how many of the compounds have that number of polymorphs. For example, more than 350 of the crystal structures have 2 polymorphs.

2.2 Graphs

There is a common architecture across most GNNs: an encoding/embedding layer the converts the crystal structures into graphs, a series of convolution layers, a series of pooling layers, a series of dense layers, and then a final output layer. For this work, we use the same encoding/embedding for all the models examined, i.e. the same graph representation. The crystal structures are converted into graphs based on atomic coordinates. The main focus of this work is to examine the performance of different graph convolutions. The creation of edges, which are representative of interatomic distances, relies on two parameters: the distance cutoff and the maximum number of nearest neighbors. Taniguchi *et al.* found that the best performance was obtained with a distance cutoff of 8 rA and 12 maximum nearest neighbors, so we use those values here as well[19]. The node elements were one-hot encoded following the procedure of Xie *et al.*[20] The graph convolutions examined here, those of CGCNN,[20] SchNet,[18] MEGNet,[4] ALIGNN,[5] and GATGNN,[13] were previously developed and later implemented in torch geometric.

In ALIGNN, the residual gated graph convolutional operator is used where node features are updated using a series of gates composed of MLPs using sigmoid activation functions[5]. In CGCNN, the central node features, neighboring node features, and edge features are concatenated before being passed through a series of gated multilayered perceptron (MLP) updates, using sigmoid and softplus activation functions[20]. In GATGNN, the graph attentional operator uses attention coefficients calculated with LeakyReLU activation functions to update node features[13]. In MEGNet, the convolution operates on the concatenated central node features, neighboring features, and edge features. Both the node and edge features are updated in a series of MLPs before updating global attributes in a final convolution[4]. In SchNet, interaction blocks are used for convolutions where interatomic distances between atoms are passed through a series of MLPs[18].

3 RESULTS AND DISCUSSION

The results of the graph convolutions are shown in Figure 2a and Figure 2b, which show the training and validation loss as a function of epoch and the parity between the predicted and DFT calculated values, respectively. As can be seen, the training loss was fairly smooth for all five models while the validation loss was more erratic. This could be attributed to the size of the validation set, which was 474 data points. The model with the lowest training and validation loss was MEGNet, which is in agreement with Taniguchi *et al.*[19] The newer implementations, ALIGNN and GATGNN, had similar train and validation losses to the previous models.

The MEGNet model also has the lowest mean absolute error (MAE) and root mean square error (RMSE) on the testing set. Additionally, MEGNet had the highest R^2 value of the five models, implying that it is not only the most accurate but the most consistent. The performance rankings of the models according to MAE are: MEGNet > CGCNN > GATGNN > SchNet > ALIGNN. Taniguchi *et al.* had predicted that the ALIGNN model would perform better than MEGNet, but in fact we find that ALIGNN performed the worst in terms of MAE and RMSE and had the lowest R^2 value of the models examined here. It could be that the key to the high performance of the ALIGNN model seen elsewhere[5] could be the actual graph representation as opposed to the graph convolution layer. This will be a focus of future work.

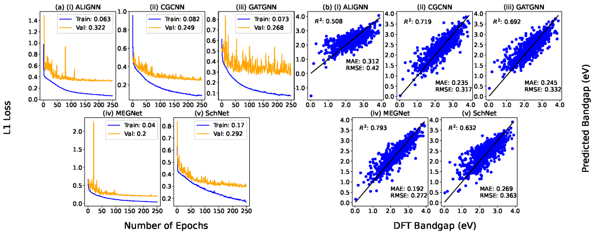

Figure 2: Training and validation error (a) and parity plots (b) for ALIGNN (i), CGCNN (ii), GATGNN (iii), MEGNet (iv), and SchNet (v). The training and validation errors are shown in blue and orange, respectively. The black line in (b) represents the line $y = x$, comparing DFT calculated values to ML predicted values. All values are averaged across three rounds of training.

3.1 Polymorphs

3.1.1 OCELOT. Of the 1421 crystal structures in the test set, 128 of them are polymorphic, with either a polymorph in the test set or the training set. The performance of the 5 models on these polymorphs is shown in Figure 3a. Similar to Figure 2b, MEGNet is the best performing model with an MAE of 0.082 eV and an RMSE of 0.144 eV along with an R^2 value of 0.952. The rankings for the performance on this task are MEGNet > CGCNN > GATGNN > SchNet > ALIGNN.

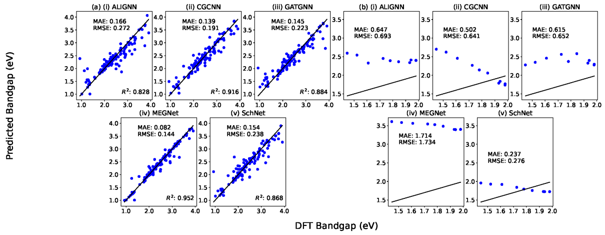

Figure 3: Parity plots on the prediction of bandgaps for polymorphs within (a) the test set of OCELOT and (b) the polymorphs of ROY using the (i) ALIGNN, (ii) CGCNN, (iii) GATGNN, (iv) MEGNet, and (v) SchNet models. The baseline bandgaps of (a) and (b) were calculated with DFT using the settings described in [1]. The black line represents $y = x$.

3.1.2 ROY. The predictions of the 5 models for the polymorphs of ROY are shown in Figure 3b. Surprisingly, the MEGNet model performs the worst with the highest MAE (1.714 eV) and RMSE (1.734 eV). The SchNet model performs the best with an MAE of 0.237 eV and an RMSE of 0.276 eV. All of the models tend to overestimate the bandgap for all polymorphs except for SchNet and CGCNN, which is in agreement with previous work[19]. Taniguchi *et al.* observed an MAE of 0.40 eV using the MEGNet model. So while the SchNet model greatly improves upon that, the MEGNet model in this work performs significantly worse. Since the OCELOT database is smaller than that of Olsthorn *et al.*[15], which had 10472 entries compared to OCELOT's 9479, which Taniguchi *et al.* trained on, it could be that the MEGNet model is more data hungry than the SchNet model. Examining the impact of training size on model performance will be tackled in future work.

4 CONCLUSION

In conclusion, we have presented preliminary results for our work using graph neural networks to predict the bandgap of organic molecular crystals. We have used the graph convolutions implemented in the MatDeepLearn framework and introduced two new graph convolutions from ALIGNN and GATGNN. We trained five models on the OCELOT database, which is a collection of organic molecular crystals with promising optoelectronic properties that contains several sets of polymorphs. In agreement with previous work, we find that the MEGNet model performed best on our test set[19]. However, we found that the MEGNet model performed poorly when predicting the bandgaps of the polymorphs of ROY. In that case, SchNet performed the best. We plan to examine other graph representations and conduct further test cases for property prediction, such as screening the Cambridge Structural Database[8] for promising OSC materials similar to [16, 19].

ACKNOWLEDGMENTS

This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation CIS210015p from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The DFT calculations were performed on the regular memory AMD EPYC

nodes while the GNN models were trained on the NVIDIA V100 GPU nodes.

REFERENCES

[1] Qianxiang Ai, Vinayak Bhat, Sean M. Ryno, Karol Jarolimek, Parker Sornberger, Andrew Smith, Michael M. Haley, John E. Anthony, and Chad Risko. 2021. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *J. Chem. Phys.* 154 (2021), 174705.

[2] Vinayak Bhat, Parker Sornberger, Balaji Sesha Sarath Pokuri, Rebekah Duke, Baskar Ganapathysubramanian, and Chad Risko. 2023. Electronic, redox, and optical property prediction of organic π-conjugated molecules through a hierarchy of machine learning approaches. *Chem. Sci.* 14 (2023), 203–213.

[3] Stanislav S. Borysov, R. Matthias Geilhufe, and Alexander V. Balatsky. 2017. OMDB: An open-access online database for data mining. *PLoS ONE* 12 (2017), 1–14.

[4] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. 2019. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* 31 (2019), 3564–3572.

[5] K. Choudhary and B. DeCost. 2021. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Comput. Mater.* 7 (2021), 185.

[6] Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G. Sumpter. 2021. Benchmarking graph neural networks for materials chemistry. *npj Computat. Mater.* 7 (2021), 84.

[7] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. 2010. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of chemical physics* 132, 15 (2010), 154104.

[8] C.R. Groom, I.J. Bruno, M.P. Lightfoot, and S.C. Ward. 2016. The Cambridge Structural Database. *Acta Cryst.* B72 (2016), 171–179.

[9] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* 1, 1 (2013), 011002.

[10] Georg Kresse and Daniel Joubert. 1999. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b* 59, 3 (1999), 1758.

[11] Christian Kunkel, Johannes T Margraf, Ke Chen, Harald Oberhofer, and Karsten Reuter. 2021. Active discovery of organic semiconductors. *Nature Communications* 12, 1 (2021), 2422.

[12] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. 2017. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* 29, 27 (2017), 273002.

[13] Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Luic, and Jianjun Hu. 2020. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* 22 (2020), 18141–18148.

[14] Jens Jørgen Mortensen, Ask Hjorth Larsen, Mikael Kuisma, Aleksei V Ivanov, Alireza Taghizadeh, Andrew Peterson, Anubhab Haldar, Asmus Ougaard Dohn, Christian Schäfer, Elvar Örn Jónsson, et al. 2024. GPAW: An open Python package for electronic structure calculations. *The Journal of Chemical Physics* 160, 9 (2024), 092503.

[15] Bart Olsthoorn, R Matthias Geilhufe, Stanislav S Borysov, and Alexander V Balatsky. 2019. Band gap prediction for large organic crystal structures with machine learning. *Advanced Quantum Technologies* 2, 7-8 (2019), 1900023.

[16] Daniele Padula, Ömer H Omar, Tahereh Nematiaram, and Alessandro Troisi. 2019. Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy & Environmental Science* 12, 8 (2019), 2412–2416.

[17] John P Perdew, Kieron Burke, and Matthias Ernzerhof. 1996. Generalized gradient approximation made simple. *Phys. Rev. Lett.* 77, 18 (1996), 3865.

[18] K.T. Schutt, F. Arbabzadah, S. Chmiela, K.R. Muller, and A. Tkatchenko. 2017. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* 8 (2017), 13890.

[19] Takuya Taniguchi, Mayuko Hosokawa, and Toru Asahi. 2023. Graph Comparison of Molecular Crystals in Band Gap Prediction Using Neural Networks. *ACS Omega* 8 (2023), 39481–39489.

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

A CODE AND DATA AVAILABILITY

The github repo for this code is located at: https://github.com/danao413/pearc24/tree/main/GNN

Received 26 April 2024; revised 12 June 2024