Leveraging Large Language Models for Property Prediction in Polymorphic Organic Semiconductors

Shreya Pagaria
spagaria@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

Mei-Yu Wang
mwang7@psc.edu
Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

Dana O'Connor
oconnord6518@gmail.com
Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

Julian A. Uran
julian@psc.edu
Pittsburgh Supercomputing Center
Pittsburgh, PA, USA

Paola Buitrago
paola@psc.edu
Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania, USA

Abstract

Organic semiconductors (OSCs) are promising for next-generation electronics, but polymorphism complicates accurate property prediction and makes traditional methods costly. We investigate transformer-based large language models (LLMs) for predicting energy gaps in polymorphic OSC crystals. A Pegasus-managed workflow is deployed across heterogeneous hardware (PSC Bridges-2 and Neocortex Cerebras CS-2) to evaluate three crystal text encodings: Materials String, SLICES, and SLICES-PLUS against a baseline XGBoost Regressor model. The results show that the LLM-analyzed Materials String achieves the highest accuracy, particularly in polymorph-rich datasets, outperforming other representations in both pretraining efficiency and downstream tasks, as well as the baseline XGBoost results. These findings highlight the potential of LLM-driven crystal encodings to accelerate materials discovery and enable the scalable, data-driven design of organic semiconductors.

ACM Reference Format:
Shreya Pagaria, Mei-Yu Wang, Dana O'Connor, Julian A. Uran, and Paola Buitrago. 2025. Leveraging Large Language Models for Property Prediction in Polymorphic Organic Semiconductors. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Problem Statement and Proposed Solution

Organic semiconductors combine the electronic properties of semiconductors with the versatility of organic compounds, enabling low-cost, sustainable electronics. Accurate property prediction is essential but is hindered by polymorphism, making traditional methods costly and inefficient. We investigate transformer-based LLMs to predict energy gaps of polymorphic OSC crystals, developing a Pegasus-compatible workflow across heterogeneous hardware (PSC Bridges-2, Neocortex CS-2). The workflow integrates tokenizer training, LLM pretraining, and regression tasks, establishing baselines for crystal text encodings and supporting future model comparisons.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

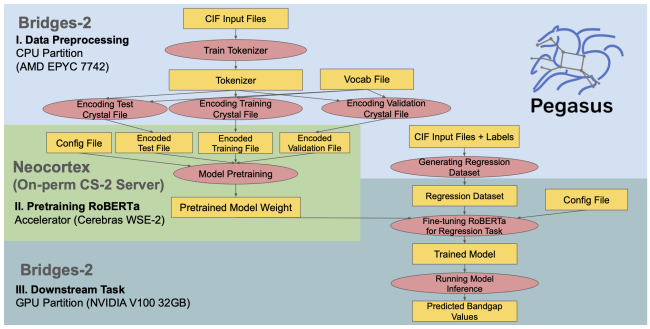

Figure 1: A scientific workflow with Pegasus running across heterogeneous hardware (PSC Bridges-2 + PSC Neocortex Cerebras CS-2)

2 Dataset & Models

We work with a curated polymorph-rich dataset containing experimental and computational crystal structures (see [1, 2]) with energy bandgaps as target properties. We also evaluate against unseen polymorphic-rich datasets such as PAH101[4] and ROY (highly polymorphic 5-methyl-2-[(2-nitrophenyl)amino]-3- thiophenecarbonitrile).

We investigate three different crystal text representations:

- Materials String [5]
- SLICES [8]
- SLICES-PLUS [7]

we applied the RoBERTa approach [6], which performs unsupervised mask-language modeling pretraining for the BERT Base model with dynamic masking, to analyze the crystal text representation. We also compare the performance with a traditional machine learning approach, which applies XGBoost regressor with TD-IDF generated feature from crystal text representations.

3 Workflow

We integrate our workflow with `Pegasus`[3], which is an open scientific workflow management software. It helps in orchestrating

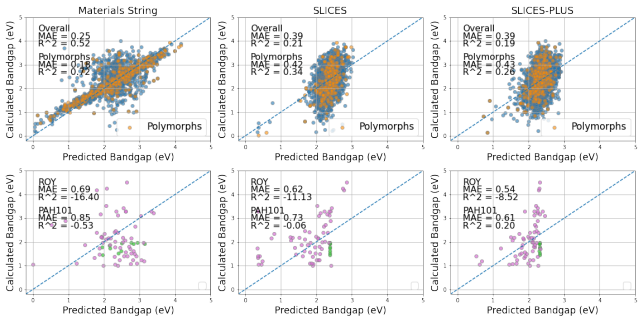

Figure 2: Predicted Energy bandgaps versus the DFT calculated bandgaps.

the various Slurm jobs and data management across the Bridges-2 and Neocortex systems. Pegasus executes a sequence of jobs, combining with parallel data processing steps, starting from data preprocessing through model pretraining to downstream tasks. Pegasus helps identify job dependency and track the workflow:

- **Data preprocessing**: we encode the crystal structure with four text representations and trained individual custom tokenizers for each text representation.
- **Pretraining RoBERTa**: We applied RoBERTa [6], which pretrains BERT Base via unsupervised masked-language modeling with dynamic masking.
- **Downstream Tasks**: we deploy a regression task to predict the energy bandgap with pretrained model weights derived from the previous step.

4 Result

Table 1: Next Token Prediction Accuracy of RoBERTa Pretrained Models

Text Representation	Next Token Prediction Accuracy [%]
Materials String	**95.3**
SLICES	91.9
SLICES-PLUS	86.1

Table 2: Regression Task Performance of XGBoost & RoBERTa Models

Text Representation	TF-IDF + XGBoost (MAE/R^2)	RoBERTa (MAE/R^2)
Materials String	0.41/0.09	**0.25/0.52**
SLICES	0.40/0.18	0.39/0.21
SLICES-PLUS	0.40/0.18	0.39/0.19

Table 1 reports the accuracy of the next-token prediction obtained from RoBERTa after pretraining. Among the evaluated text representations, the materials string achieves the highest accuracy, reaching 95.3%. In Table 2 amd Figure 2 we show the comparison of mean absolute error (MAE) and R^2 for the XGBoost baseline and our RoBERTa results. The Materials String consistently demonstrates superior performance on downstream tasks (MAE = 0.25, R^2 = 0.52) compared to SLICES and SLICES PLUS(MAE = 0.39, R^2 = 0.19-0.21). Noticeably, the polymorphs, which are marked in yellow, exhibit markedly improved prediction accuracy for Materials String, which achieves MAE = 0.18 and R^2 = 0.72, relative to non-polymorphic crystals. However, when evaluated on unseen validation datasets such as ROY and PAH101, the models fail to generalize effectively, yielding substantially worse MAE and R^2 values for both XGBoost regressor baseline and the RoBERTa models.

5 Conclusion and Future Work

We developed a scientific workflow using Pegasus to orchestrate execution across heterogeneous hardware resources, including PSC Bridges-2 and PSC Neocortex, for a materials science application. Our results demonstrate that the materials string outperforms other crystal text encodings, achieving higher pretraining efficiency and accuracy, and effectively capture polymorphic properties. Our transformer-based approach outperforms the XGBoost baseline, demonstrating the applicability of large language models for materials, and highlighting potential for generalizable representations of crystal structure data.

We will extend our benchmark by adding Robocrys, optimize hyperparameters and target normalization to improve generalization on datasets such as ROY and PAH101, and broaden applications to properties like charge-carrier mobility and thermodynamic stability, ultimately advancing inverse design of high-performance materials.

Acknowledgments

This material is based upon work supported by Neocortex, through funding provided by the National Science Foundation under Grant 2005597. This work used Bridges-2 atPittsburgh Supercomputing Center through allocation CIS250525 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services Support (ACCESS) program,which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296. Pegasus is funded by The National Science Foundation under Office of Advanced Cyberinfrastructure (OAC) grant 2138286. Previously, NSF has funded Pegasus under OAC SI2-SSI program grant 1664162, OCI SDCI program grant 0722019 and OCI SI2-SSI program grant 1148515.

References

[1] Qianxiang Ai, Vinayak Bhat, Sean M. Ryno, Karol Jarolimek, Sornberger, and et al. 2021. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *The Journal of Chemical Physics* 154, 17 (05 2021), 174705. doi:10.1063/5.0048714

[2] Imanuel Bier, Dana O'Connor, Yun-Ting Hsieh, Wen Wen, Anna M. Hiszpanski, T. Yong-Jin Han, and Noa Marom. 2021. Crystal structure prediction of energetic materials and a twisted arene with Genarris and GAtor. *CrystEngComm* 23 (2021), 6023–6038. Issue 35. doi:10.1039/D1CE00745A

[3] Ewa Deelman, Karan Vahi, Gideon Juve, and et al. 2015. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems* 46 (2015), 17–35. doi:10.1016/j.future.2014.10.008

[4] Siyu Gao, Xingyu Liu, Yiqun Luo, Xiaopeng Wang, Kaiji Zhao, Vincent Chang, Bohdan Schatschneider, and Noa Marom. 2025. PAH101: A GW+BSE Dataset of 101 Polycyclic Aromatic Hydrocarbon (PAH) Molecular Crystals. *Scientific Data* 12, 1 (2025), 679. doi:10.1038/s41597-025-04959-0

[5] W. Gao, C. Wang, S. Li, and et al. 2024. Accurate prediction of synthesizability and precursors of 3D crystal structures via large language models. *Nature Communications* 15 (2024), 2901. doi:10.1038/s41467-024-47209-1

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* (July 2019). arXiv:1907.11692

[7] Baoning Wang, Zhiyuan Xu, Zhiyu Han, Qiwen Nie, Hang Xiao, and Gang Yan. 2024. SLICES-PLUS: A Crystal Representation Leveraging Spatial Symmetry. arXiv:2410.22828 [physics.comp-ph] https://arxiv.org/abs/2410.22828

[8] Hang Xiao, Jun Li, Zhibo Yan, Yang Zhao, Qiushi Yang, Zihan Xu, Yifan Li, Xin Deng, and Liangcai Zhang. 2023. An invertible, invariant crystal representation

for inverse design of solid-state materials using generative deep learning. *Nature Communications* 14, 1 (2023), 7830.