Leveraging LLMs for Property Prediction in Polymorphic Organic Semiconductors

Shreya Pagaria (Student)[1,2], Mei-Yu Wang (Advisor)[2], Dana O'Connor (Advisor)[2], Julian Uran (Advisor)[2] , Paola Buitrago (Advisor)[2]
[1]Carnegie Mellon University, Pittsburgh, PA, USA; [2]Pittsburgh Supercomputing Center, Pittsburgh, PA, USA

PROBLEM STATEMENT

Organic semiconductors (OSCs) offer tunable, low-cost, and sustainable electronics, making accurate property prediction crucial for materials discovery. However, existing challenges include:

- Modeling polymorphism in organic semiconductors is computational expensive.
- Property prediction from crystal structures is challenging and experiments are resource-intensive.
- No existing baseline for text encodings.

Proposed Solution

We design a scientific workflow compatible with Pegasus running across heterogeneous hardware (PSC Bridges-2 + PSC Neocortex) to evaluate the performance of LLMs on multiple text representations of crystal structures and downstream property prediction.

DATASET & MODELS

A curated polymorph-rich dataset containing experimental and computational crystal structures (e.g. OCELOT) with energy bandgaps as target properties.

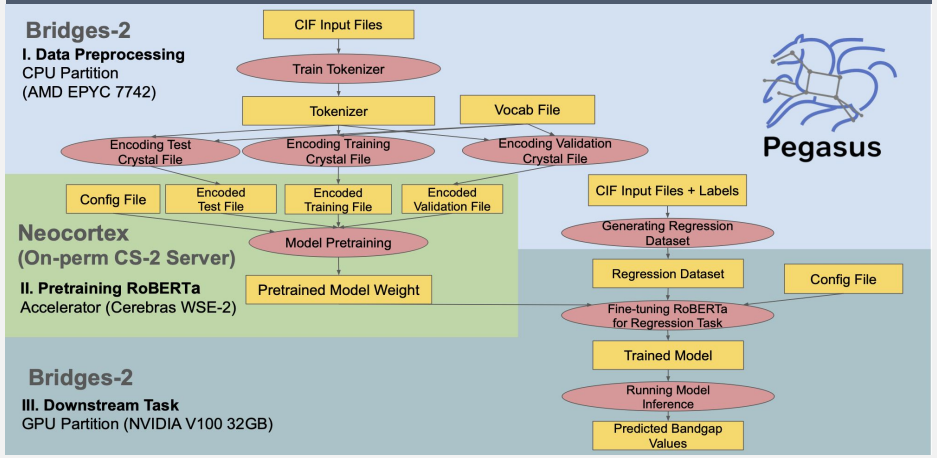
Evaluate against unseen polymorphic-rich datasets: **PAH101** and **ROY**.

Three text representations for crystals:

- Material String
- SLICES
- SLICES-PLUS

Model: RoBERTa Base Model (pretraining+downstream regression task)
Baseline: TF-IDF (text feature extraction) + XGBoost

WORKFLOW



RESULTS

Pretraining Model Performance		Regression Performance (MAE ↓ / R^2 ↑)	
Text Representation	Next Token Prediction Accuracy [%] ↑	TF-IDF + XGBoost	RoBERTa
Materials String	**95.3**	0.41/0.09	**0.25/0.52**
SLICES	91.9	0.40/0.18	0.39/0.21
SLICES-PLUS	86.1	0.40/0.18	0.39/0.19

Table 1. Next Token Prediction Accuracy of RoBERTa Pretrained Models; Regression Task Performance of XGBoost & RoBERTa Models

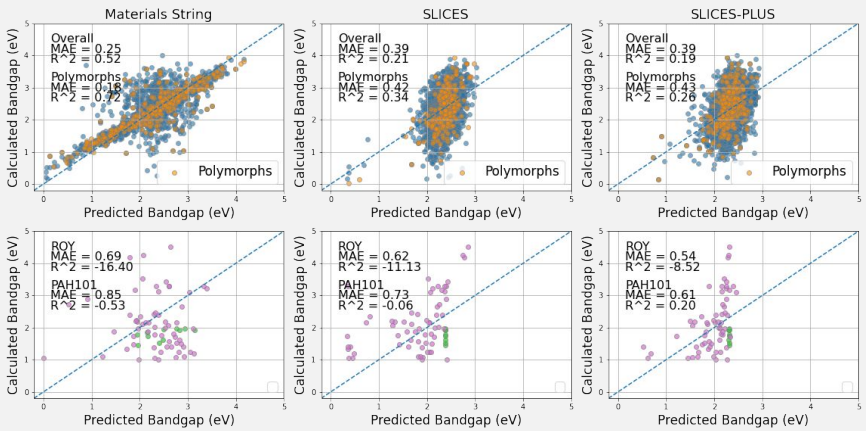


Fig1. Calculated bandgap v.s. predicted bandgap on the test dataset (top panels) and on the unseen validation set (PAH101 and ROY, lower panels)

CONCLUSION & FUTURE WORK

- We design a scientific workflow with Pegasus running across heterogeneous hardware (PSC Bridges-2 + PSC Neocortex) for a material science application.
- Materials String surpasses other crystal text representations and baseline predictions, yielding higher pretraining efficiency, property prediction accuracy, and effectively capturing polymorphic properties. Baseline XGBoost Regressor model underperforms, confirming the advantage of transformer-based models.
- Plan to enhance performance through hyperparameter tuning, alternative model architectures, and scaling strategies.
- Extend the work to other properties (e.g., density) and encoding (e.g, Robocrystallographer).

REFERENCE

- D. O'Connor et al. 2025. Text Representations for Property Prediction of Organic Molecules Using RoBERTa. In Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration (PEARC '25). https://doi.org/10.1145/3708035.3736093
- D. O'Connor D, Buitrago P. Examining the Influence of Graph Representation on Property Prediction of Polymorphic Organic Molecular Crystals. ChemRxiv. 2025
- A., Bhat, et al. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. J Chem Phys. 2021;154(17):174705.
- S. T. Brown et al. 2021. Bridges-2: A Platform for Rapidly-Evolving and Data Intensive Research. In Practice and Experience in Advanced Research Computing. 1–4. https://dl.acm.org/doi/10.1145/3437359.3465593

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 at Pittsburgh 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.