Polymer Property Prediction via Fusion of Language Models and Probabilistic Edge Graph Neural Networks

Araki Wakiuchi [1] **Arifin** [1] **Ryo Yoshida** [2]

[1]*JSR Corporation, Japan* [2]*The Institute of Statistical Mathematics, Japan*. Correspondence to: Araki Wakiuchi Araki_Wakiuchi@jsr.co.jp.

1. Introduction

Accurate polymer-property prediction supports material design, optimization, and high-throughput screening, yet probabilistic copolymerization of diverse monomers produces complex structures whose descriptions vary widely across literature and experiments, hindering reproducible analysis.

Large-scale MD databases such as RadonPy[1] provide controlled property data, but integrating them with experimental records is difficult because representation formats differ; surrogate models that ingest heterogeneous inputs—text, graphs, and numeric descriptors—are therefore required for reliable, transferable predictions.

String notations like BigSMILES[2] enable language models[3] to embed domain knowledge, while graph neural networks (GNNs) capture local topology; however, their cost escalates for chains with large repeating units and high molecular weight. Stochastic edge schemes have been proposed to encode polymer connectivity more compactly[4].

We address these issues with a hybrid architecture that couples MatSciBERT[5] with a GNN augmented by probabilistic edges. A cross-attention layer unifies textual and structural embeddings, allowing the network to learn from both simulated and experimental datasets (Fig. 1). The resulting model is robust to representational diversity and scales to realistic polymer sizes.

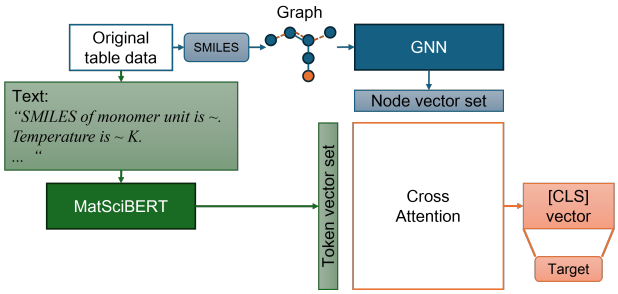

Fig. 1: The illustration of our model architecture as the fusion of language model and graph model.

2. Methods

2.1 Model architecture

The language model component uses *MatSciBERT*, a domain-specific pretrained BERT model optimized for materials science. The textual input is a single sequence obtained by concatenating (i) the SMILES string of the polymer repeating unit, (ii) MD simulation conditions, and (iii) quantum-chemical descriptors (QM values) of the monomers.

The graph component is built from the same repeating-unit SMILES as follows.

- **Nodes** — All heavy atoms in one repeating unit are included, and two additional methyl-carbon nodes are appended to represent the start and end termini of the chain.
- **Node feature (4 dim)** — ⟨atomic number, degree, formal charge, aromaticity⟩, zero-padded when necessary.
- **Edge feature (5 dim)** — the first four dimensions are a one-hot encoding of SINGLE, DOUBLE, TRIPLE, or AROMATIC bonds; the fifth dimension stores a continuous probability for virtual (inter-unit or terminal) connections.

The repeating-unit SMILES contains two "*" dummy atoms that indicate the connection sites. Their neighboring atoms are designated as *head* and *tail*. Given a target average chain length n (provided by the MD dataset), we encode the stochastic chain-length distribution as

$$p_{\text{terminal}} = \frac{1}{n}, \qquad p_{\text{repeat}} = \frac{n-1}{n}.$$

1. The virtual edge between the start-methyl carbon and the first head atom, and that between the last tail atom and the end-methyl carbon, receive p_{terminal} in their fifth feature dimension.
2. Every virtual edge that bridges the tail of unit i and the head of unit $i+1$ receives p_{repeat} in the same slot.

This scheme embeds the expected chain-length distribution directly into the graph, allowing the network to learn polymer-specific stochastic connectivity without enlarging the node or edge set.

In the fusion block, token embeddings from the final layer of MatSciBERT and node embeddings from the final graph layer are combined through a cross-attention mechanism. The resulting `[CLS]` vector is fed to a linear head to predict the target properties.

2.2 Dataset

We used the publicly available RadonPy PI1070.csv dataset, derived from large-scale MD simulations of polymers. The target properties are density, heat capacity at constant pressure (C_p), and refractive index. As each property has a different dynamic range, we apply standard scaler based

on the training set's distribution prior to model training.

Input formats comprise English text with SMILES, MD condition and QM values and a molecular graph constructed from the repeating unit SMILES. The dataset was split into 80 percent training and 20 percent test sets, with the following four-stage training protocol:

1. Fine-tuning MatSciBERT alone (10 epochs)
2. Pre-training the graph model alone (10 epochs)
3. Freezing the language model part, training the graph and cross-attention layers (20 epochs)
4. Unfreezing the language model part and then jointly training the entire network (200 epochs)

3. Results and Discussion

Table 1 presents the final predictive performance for density, C_p, and refractive index (R^2). Figure 2 shows prediction vs observation plots of the test. The fused model achieves $R^2 > 0.96$ across all properties, with minimal difference between the training and test results, indicating a strong generalization.

For comparison, we trained single-modal models with the same protocol using only fine-tuned MatSciBERT (200 epochs). Our cross-attention fusion model outperformed the baseline for all properties. This highlights the strength of using both domain-specific language and molecular structure representations.

Table 1: R^2 results for our model and only BERT model for training and test data

R^2	BERT		Ours	
	Train	Test	Train	Test
Density	0.990	0.957	0.991	0.974
C_p	0.985	0.952	0.996	0.972
Refractive index	0.983	0.903	0.995	0.961

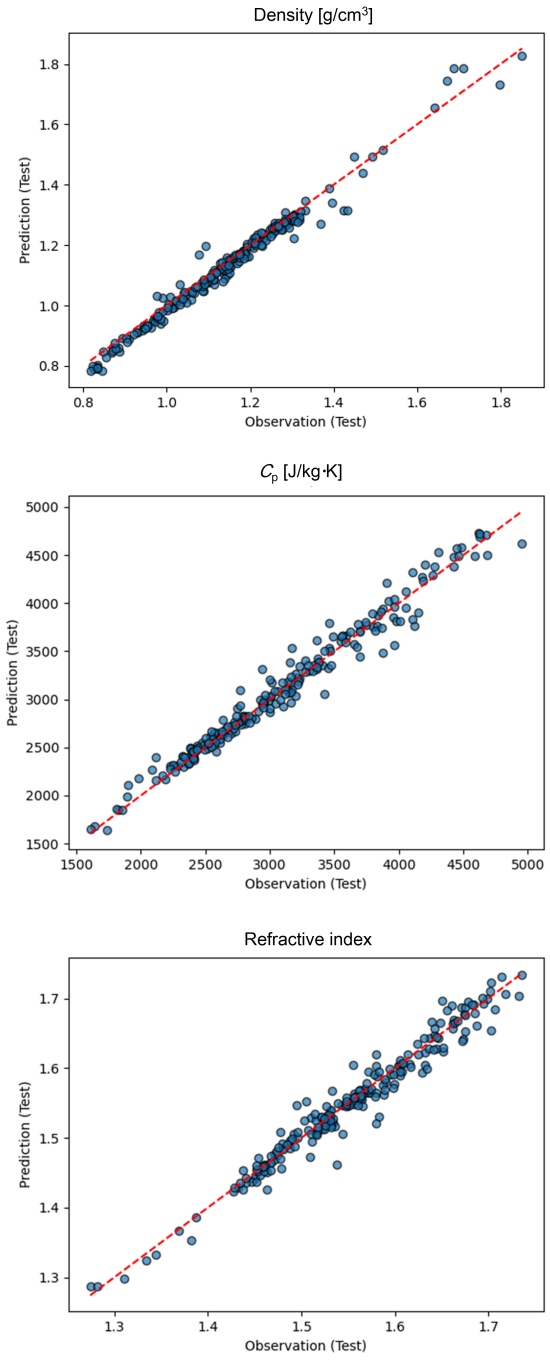

Fig. 2: Prediction versus observation plots for the test dataset.

References

[1] Hayashi Y. et al. RadonPy: Automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Computational Materials*, 8:222, 2022.

[2] Lin T. S. et al. BigSMILES: A structurally-based line notation for describing macromolecules. *ACS Central Science*, 5(9):1523–1531, 2019.

[3] Xu C. et al. TransPolymer: A Transformer-based language model for polymer property prediction. *npj Computational Materials*, 9(1):64, 2023.

[4] Aldeghi M. et al. A graph representation of molecular ensembles for polymer property prediction. *Chemical Science*, 13(35):10486–10498, 2022.

[5] Gupta A. et al. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8:222, 2022.