Fake It Till You Make It: Multi-Physics Synthesis Breaks the Data Barrier in Chemical Language Models

Naiyu Yin[1]; Ning Liu[2]; Jiuzhou Chen[1]; Brian Y.Lattimer[3]; Jim Lua[2]; Yu Yue[1]
[1]Lehigh University, [2]Global Engineering and Materials Inc., [3] Virginia Tech

Abstract

- Chemical language models (CLMs) show promise for polymer property prediction and generative design, yet progress is limited by small experimental datasets. We propose **Poly4mer**, a multi-physics synthesis framework that bridges polymer representation learning with physics-based data generation.
- The framework integrates group contribution theory and fire-property simulations to create a large corpus of synthetic polymers with physically consistent labels.
- Using this dataset, we train an encoder–decoder architecture CLM tailored for polymer SMILES and a two-phase property predictor combining physics-based pretraining and experimental fine-tuning.
- Poly4mer achieves state-of-the-art fire-property prediction accuracy and enables inverse polymer design, demonstrating that physics-grounded synthesis can overcome data scarcity in CLMs.

Introduction

The Data Barrier: While CLMs like SMI-TED [1] have revolutionized molecular discovery, specialized domains like polymer flammability suffer from severe data scarcity. Experimental measurements are costly, and datasets often contain only tens or hundreds of samples.

The Representation Challenge: Standard CLMs use SMILES designed for small molecules. They struggle with polymer-specific syntax, specifically the asterisk (*) token used to denote open bonds and connectivity, which lacks semantic meaning in models trained only on monomers.

Our Solution: We propose a "multi-physics synthesis" framework. By "faking" data using physics-based rules before reality catches up, we generate ~120k synthetic samples to pretrain a foundation model that is structurally valid and physically consistent.

Methods

Physics-Based Synthetic Data Generation We utilize the Group Contribution (GC) method to bridge data-driven modeling and physical knowledge:

- **Structure Generation:** We identify 48 functional groups and assemble them into ~120k valid homo-, co-, and terpolymers.
- **Property Simulation:** We use the Fire Dynamics Simulator to generate labels for four key flammability metrics: time to ignition (t_{ig}), peak heat release rate (pHRR), smoke extinction area (SEA), and CO yield.

Polymer Chemical Language Model (Poly4mer) We introduce a specialized Encoder-Decoder architecture:

- **Asterisk Encoder:** A novel module specifically designed to handle the * token in polymer SMILES, operating alongside a backbone encoder pretrained on monomers.
- **Unified Autoencoding:** The model is trained to reconstruct polymer tokens (Decoder) and predict properties (Predictor) simultaneously from a shared latent embedding.

Two-Phase Training Strategy:

- **Phase I (Pretraining):** Supervised training on the 120k synthetic samples to learn polymer semantics and physical priors.
- **Phase II (Fine-tuning):** Refinement on a small dataset of 55 experimentally measured polymers to ensure high predictive fidelity.

Results

Reconstruction

- 99.89% accuracy on 125,688 synthetic polymers.
- 98.18% accuracy on 55 experimental polymers.
  → Demonstrates reliable polymer structure encoding and decoding.

Fire Property Prediction

- Evaluation on experimental data in terms of relative MSE (%) for four fire properties: Tig, pHRR, SEA, and CO.
- In Table 1, Poly4mer outperforms state-of-the-art baselines: MoLFormer [2], TransPolymer [3], polyBERT [4], and TrinityLLM [5].
  → Demonstrates accurate fire property prediction.

Inverse Design Case Study

- Starting from PVC *CC(Cl)*, latent embedding optimization yields polymers with improved predicted fire properties.
  → Demonstrates success inverse polymer design.

Table 1. Fire properties prediction performance on experimental data in terms of relative MSE (%) against state-of-the-art baselines.

Methods	Fire Properties			
	t_{ig} ↓	pHRR ↓	SEA ↓	CO ↓
MoLFormer [4]	34.04 ± 22.07	62.01 ± 19.94	41.11 ± 14.70	-
TransPolymer [15]	31.78 ± 5.81	77.88 ± 76.64	38.91 ± 16.95	-
polyBERT [14]	41.39 ± 5.09	61.42 ± 11.26	51.28 ± 10.00	-
TrinityLLM [1]	16.65 ± 5.72	33.64 ± 2.06	22.35 ± 12.24	18.58 ± 8.28
Poly4mer (ours)	**16.32** ± 6.20	**21.33** ± 3.23	**17.47** ± 4.19	**16.13** ± 2.97

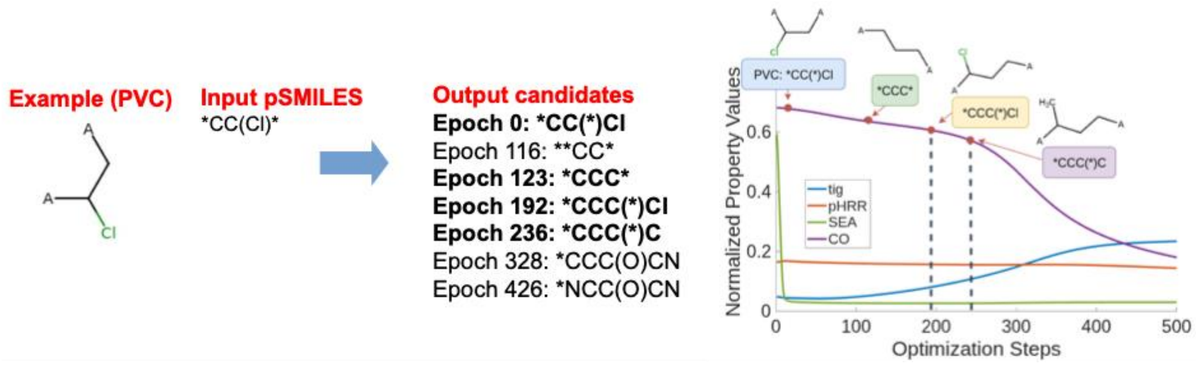


Figure 1. An illustration of the polymer design procedure.

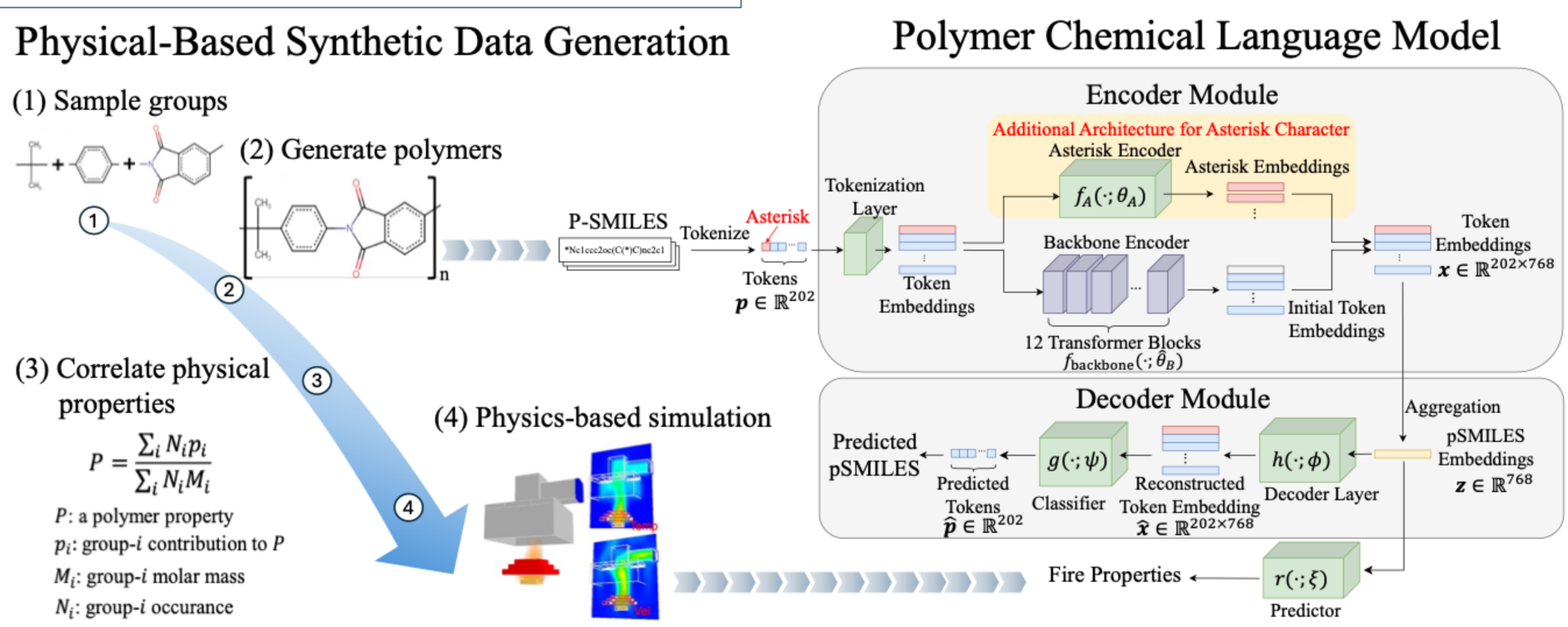
Conclusions and Discussion

Breaking the Data Barrier: The results validate the "fake it till you make it" paradigm. Synthetic data generated via GC and FDS provides a sufficient signal to train complex transformer architectures, which then generalize well to small experimental datasets.

State-of-the-art Performance: The high polymer reconstruction and property prediction accuracy confirms that the learned latent embeddings are to be semantically rich, allowing for meaningful optimization of material properties.

Future Directions:

- Incorporate alternative modalities and non-canonical encodings to improve generalization and distinguish between chemically equivalent polymers.
- Mechanistic interpretability studies to quantitatively verify that the model's latent representations truly capture polymer chemistry.
- Extend the Poly4mer framework to predict a broader range of thermal, mechanical, and physical properties beyond just flammability.



Contact

Naiyu Yin
Lehigh University
Email: nay224@lehigh.edu

Yue Yu
Lehigh University
Email: yuy214@lehigh.edu

References

1. Emilio Vital Brazil, Eduardo Soares, Victor Yukio Shirasuna, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. Smi-ted: A large-scale foundation model for materials and chemistry. In AI for Accelerated Materials Design-ICLR 2025.
2. Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12):1256-1264, 2022.
3. Changwen Xu, Yuyang Wang, and Amir Barati Farimani. Transpolymer: a transformer-based language model for polymer property predictions. npj Computational Materials, 9(1):64, 2023.
4. Christopher Kuenneth and Rampi Ramprasad. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Nature communications, 14(1):4099, 2023.
5. Ning Liu, Siavash Jafarzadeh, Brian Y Lattimer, Shuna Ni, Jim Lua, and Yue Yu. Harnessing large language models for data-scarce learning of polymer properties. Nature Computational Science, 5(3):245-254, 2025.

Code