TRANSPOLYMER

-A Transformer based Model for polymer property predictions

~S.KAVYA REDDY
23P81A0557
G392

ABSTRACT

Develop a Transformer-based model (Transpolymer) for predicting polymer properties.

Traditional methods for evaluating polymer properties are expensive and time-consuming, requiring experiments or simulations

CONCLUSION

Transpolymer uses a smart tokenizer that understands chemical structures. It learns patterns using a self-attention mechanism.

It is first pretrained on a huge dataset of polymers using a technique called Masked Language Modeling (MLM) to improve accuracy.

WHY TRADITIONAL MODELS FALL SHORT?

- CNNs (Convolutional Neural Networks) – Best for image-based tasks but Struggle with sequential data like polymers and fail to capture long-range dependencies effectively.
- RNNs (Recurrent Neural Networks) – Process data sequentially, making them slow, hard to parallelize, and prone to the vanishing gradient problem.
- LSTMs (Long Short-Term Memory Networks) – While better than RNNs, they are computationally expensive, inefficient for very long sequences, and still struggle with parallelization.
- GNNs (Graph Neural Networks) – Work well for molecular graphs but struggle with polymer sequences, as they rely on local neighborhood information rather than capturing global dependencies.

WHY TRANSFOMERS ARE BETTER?

- Use self-attention, allowing them to capture long-range dependencies efficiently.
- Process sequences in parallel, making them faster and more scalable.
- More generalizable across different polymer structures.

WORKFLOW OF TRANSPOLYMER

1. Data Preparation

Polymer Tokenization: Converts polymer structures into meaningful sequences using SMILES representation and chemical descriptors.
Data Augmentation: Generates non-canonical SMILES to improve model robustness and generalization.

2. Pretraining Phase (Unlabeled Data)

Transformer Encoder Setup: Initializes a RoBERTa-based transformer to learn polymer representations.
Masked Language Modeling (MLM):The model is trained to predict the original masked tokens based on the surrounding context.
Masking Strategy: 15% of tokens are randomly masked.
Prediction Mechanism: The model predicts masked tokens using a Softmax layer with Cross-Entropy Loss.
Optimization: Weights are updated using the AdamW optimizer.

3.Fine-tuning phase(Labeled data)

What is Fine-Tuning?

Adapting the pretrained TransPolymer model for specific polymer property predictions using labeled data.

Process:

1. Load the Pretrained Model (trained on unlabeled polymer data).
2. Replace MLM Head with Regression Head (MLP) for property prediction.
3. Train on Labeled Data using Mean Squared Error (MSE) loss.
4. Optimize with AdamW optimizer.
5. Evaluate using RMSE and R^2 scores.

4.Model Evaluation

Performance Metrics: Computes Root Mean Squared Error (RMSE) and R^2 scores to assess accuracy.

Baseline Comparison: The model is compared with traditional approaches like GNNs, LSTMs, and Random Forests (RF).

Ablation Studies: Evaluates the effect of pretraining and data augmentation on final performance.

Comparing Transpolymer with Base lines

MODEL	RMSE	R^2
Random Forest (ECFP)	1.00	0.32
LSTM	1.36	-0.25
GNN	0.97	0.16
Transpolymer (Pre-trained)	0.67	0.69

LIMITATIONS:

1. High computational cost – Transformers require significant GPU/TPU resources, making them expensive to train and deploy.

2. Quadratic complexity – Self-attention scales poorly with input length, limiting efficiency for long sequences.

3. Large memory consumption – Storing activations and parameters demands high memory, restricting deployment on low-resource devices.

4. Data dependency – Transformers need vast amounts of high-quality data to generalize well.

5. Slow inference – Large models lead to high latency, making real-time applications challenging.

6. Context length limitation – Fixed token limits cause issues with very long inputs.

7. High energy consumption – Training and running large models have significant environmental impacts.

Scope of Project

The TransPolymer project focuses on developing a Transformer-based AI model for polymer property prediction. It leverages self-attention mechanisms and chemically aware tokenization to learn meaningful representations from polymer sequences. By pretraining on a large dataset and fine-tuning on specific polymer properties, it enhances material discovery and rational polymer design, reducing reliance on costly experiments.

Business Case Addressed by Transpolymer

1. Accelerating Polymer Discovery – Reduces the time needed for identifying new polymers with desired properties.

2. Minimizing R&D Costs – Lowers expenses by replacing costly and time-consuming experiments with AI-driven predictions.

3. Enhancing Material Performance – Improves polymer design for applications in electronics, energy storage, and materials science.

4. Data-Driven Innovation – Leverages large-scale polymer datasets to improve prediction accuracy and material optimization.

5. Competitive Advantage – Enables industries to develop better-performing polymers faster, giving an edge in the market.

6. Scalability & Adaptability – Can be fine-tuned for various polymer-related tasks, making it useful across different industries.

Typical users of Transpolymers

1. Materials Scientists – Use it to discover and design new polymers with desired properties.

2. Chemical Engineers – Apply it for optimizing polymer formulations in industries like plastics, coatings, and composites.

3. Pharmaceutical Companies – Utilize it for designing polymer-based drug delivery systems.

4. Electronics Manufacturers – Leverage it to develop advanced polymers for semiconductors, flexible electronics, and displays.

5. Energy Sector Experts – Use it for designing polymer electrolytes in batteries and energy storage materials.

6.AI & Data Scientists – Work on improving the model and applying machine learning to polymer informatics.

How Does the Solution Help the Users?

- **Materials Scientists & Chemical Engineers** → Speeds up research and reduces trial-and-error.
- **Pharmaceutical Companies** → Helps design biocompatible polymer materials.
- **Electronics & Energy Experts** → Enhances materials for batteries, sensors, and circuits.
- **R&D Teams** → Reduces time and costs in new material development.
- **AI & Data Scientists** → Provides a benchmark dataset and framework for ML research in polymer informatics.

Technologies Used in the Project

Frontend:

React.js: For a smooth and interactive web app
HTML, CSS, JavaScript : For basic UI design
What It Does:

- A webpage where users can input polymer sequences
- Displays predicted polymer properties
- Shows graphs/visualizations (using Chart.js)

Backend:

FastAPI or Flask : To create an API that runs the Trans polymer model
PyTorch : Since Trans polymer is Transformer-based
MongoDB : If you need to store data
What It Does:

- Receives polymer data from the front-end
- Runs the Trans polymer model to predict properties
- Sends results back to the front-end

Databases:

-The project relies on a custom chemically aware polymer tokenizer to convert polymer sequences into meaningful representations.
-Large datasets of polymer sequences from PI1M and other benchmark datasets were used to train and test the model.
-Data augmentation techniques were applied to improve learning from polymer SMILES representations.

AI Models:

-The core of TransPolymer is a Transformer-based deep learning model, specifically based on RoBERTa architecture.
-It uses Masked Language Modeling (MLM) pretraining and fine-tuning for polymer property prediction.
-Various machine learning models were used for comparison, including Random Forest, GNNs, CNNs, LSTMs

Summary

- TransPolymer is a Transformer-based model designed for accurate polymer property prediction.
- It introduces a chemically aware tokenizer that encodes polymer structures, including repeating units and descriptors.
- Data augmentation through non-canonical SMILES generation further enhances generalization.
- The model is pretrained on ~5 million polymer sequences using Masked Language Modeling (MLM), enabling it to learn meaningful representations.
- TransPolymer outperforms Graph Neural Networks (GNNs), LSTMs, and traditional models like Random Forest. Ablation studies confirm that MLM pretraining, fine-tuning, and data augmentation significantly enhance accuracy.
- The self-attention mechanism improves interpretability by highlighting key chemical interactions.
- Transpolymer can be applied in active-learning frameworks for

Overview

Aspect	Details
Model Used	**TransPolymer** (Transformer-based model)
Baseline Models	Graph Neural Networks (GNNs), Long Short-Term Memory (LSTM), Random Forest, Artificial Neural Networks(ANN)
Pretraining	Masked Language Modeling (MLM)
Properties Predicted	Conductivity, Bandgap, Electron Affinity, Ionization Energy, Crystallization Tendency, Dielectric Constant, Refractive Index.
Performance	Outperforms **all baseline models**
Key Findings	MLM pretraining, fine-tuning Transformer layers, and **data augmentation** significantly improve accuracy.
Interpretability	Self-attention mechanism helps capture **key chemical interactions** for better predictions.
Future Applications	Can be extended to **copolymer predictions, reaction modeling, and multi-task learning.**

Thank You