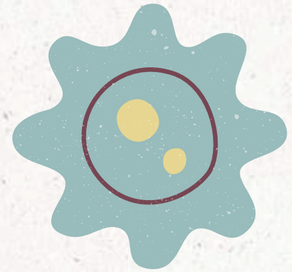
TRANSPOLYMER

~G392

INTRODUCTION:



The TransPolymer model is a Transformer-based AI system designed to predict polymer properties more accurately and efficiently. It uses a special tokenization method to better understand polymer structures and is trained on a large dataset using Masked Language Modeling (MLM). Because of this, TransPolymer performs better than other models across multiple datasets.
In conclusion, this model can be a powerful tool for polymer research and virtual material discovery, helping scientists find new materials faster and more effectively

METHODOLOGY:

Polymer Tokenization:
Breaks down polymer structures into meaningful units using SMILES notation and extra details like polymerization degree and composition.

PRETRAINING ON LARGE DATA:
trains on ~5 million polymer sequences to learn useful chemical patterns.

Fine-tuning: Adapts the model to specific datasets for better accuracy in predicting polymer properties.

Data Augmentation: Creates variations of SMILES representations to improve learning and generalization.

Transformer Encoder: Uses self-attention to understand chemical relationships within polymer structures.

CHALLENGES IN PREDICTING POLYMER PROPERTIES:

SMILES Notation :The usual way of writing polymer structures (SMILES) doesn't fully show how polymers connect or grow.
Not Enough Data: There aren’t many big polymer datasets, so we need to create extra training data to help the model learn better.
Complex Polymer Structures: Some polymers, like copolymers and mixtures, have complicated designs that are hard for models to understand.
Traditional Models Struggle: Older models often memorize data instead of truly learning, making them less reliable for new polymers

EVALUATION PARAMETERS USED:

Root Mean Square Error (RMSE): Measures prediction error.
Coefficient of Determination (R^2): Evaluates the model's predictive performance.
Cross-validation: 5-fold cross-validation is used for most datasets

LIMITATIONS:

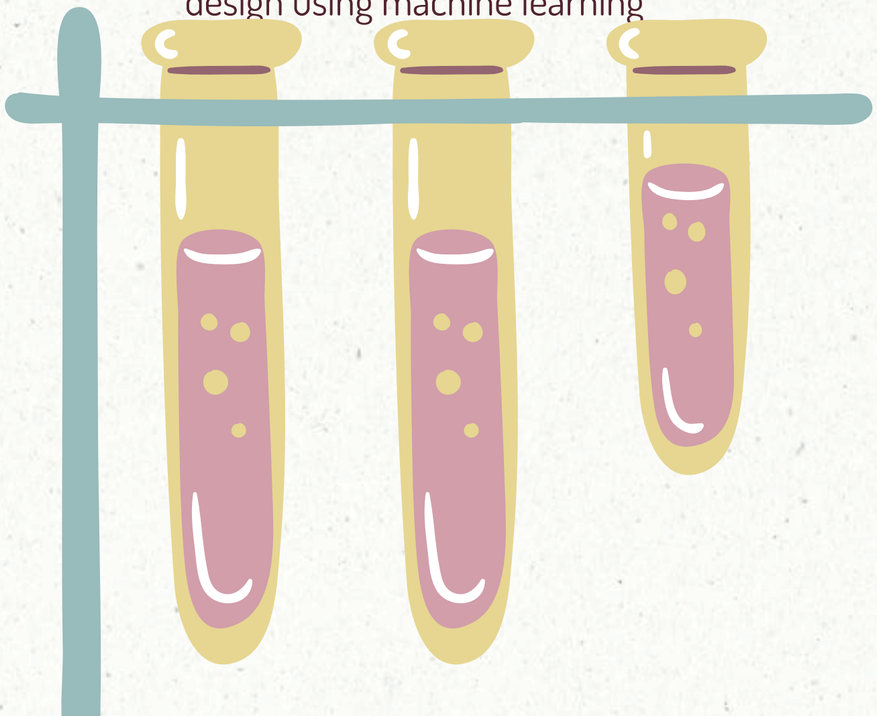
- Dependence on Pretraining Data: Model performance is highly influenced by the amount of pretraining data.
- Computational Cost: Pretraining requires significant resources.
- Handling of Noisy Data: Performance on some datasets is affected by inherent data noise.
- Limited Interpretability: While attention mechanisms highlight important features, deep learning models still lack full interpretability

SCOPE OF PROJECT:

TransPolymer uses AI to revolutionize polymer research , making it easier , faster , and more accurate to predict how different polymers will behave . This helps scientists design new materials more efficiently and opens the door to future breakthroughs in materials science and chemistry.

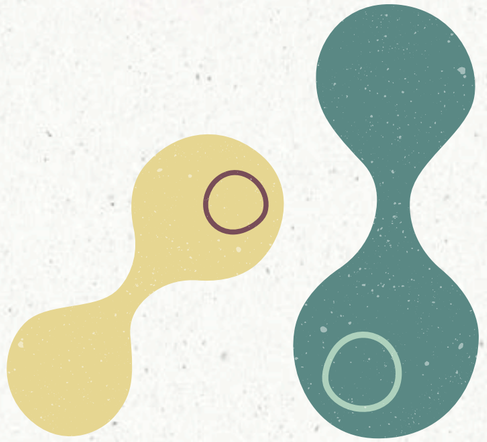
BASIC UNDERSTANDING ON THE SCOPE OF THE PROJECT

The TransPolymer project focuses on using deep learning, specifically Transformer-based models, to predict polymer properties. Traditional methods for polymer property evaluation rely on costly and time-consuming experiments or simulations. The project leverages SMILES representations, Transformer encoders, and self-attention mechanisms to develop a data-driven approach for predicting polymer properties such as conductivity, bandgap, crystallization tendency, and dielectric constant. The ultimate goal is to enable faster, more efficient, and accurate polymer design using machine learning



WHAT BUSINESS CASE DOES THE PROJECT ADDRESS?

- Cuts Costs & Saves Time – Traditional lab testing is slow and expensive. This AI model helps predict polymer properties faster, reducing the need for trial-and-error experiments.
- Speeds Up Material Discovery – Helps industries like batteries, electronics, and medicine find the best polymers more quickly.
- Better Decisions with Data – Provides accurate predictions, so companies can choose the right materials without guesswork.
- Gives Companies an Edge – Businesses can develop new materials faster, staying ahead of competitors.

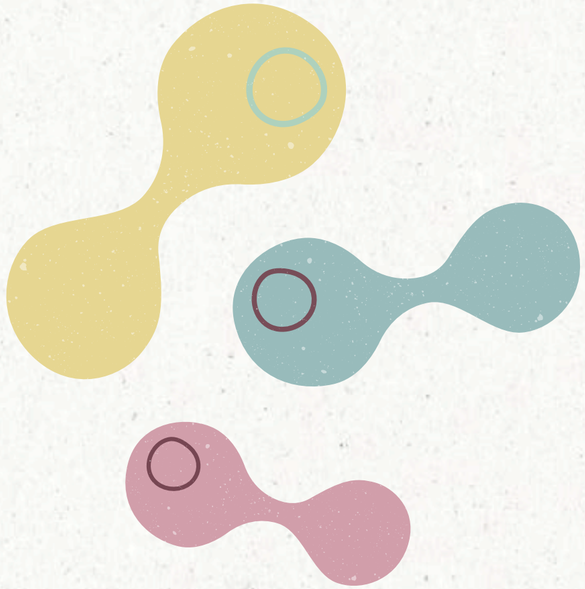


WHO USES TRANSPOLYMER

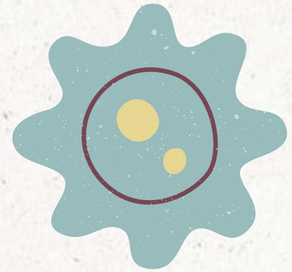
- Scientists & Chemists – Helps researchers predict polymer properties before running costly lab tests.
- R&D Teams – Companies making batteries, coatings, and adhesives can use it to find the best materials.
- University Researchers – Supports data-driven polymer research at universities and institutes.
- AI & Data Teams – Helps teams working on AI for materials science to improve their models.

HOW TRANSPOLYMER HELPS USERS

- Speeds Up Polymer Design – Predicts polymer properties accurately, so fewer lab tests are needed.
- More Accurate Predictions – Works better than older models like Random Forest and GNNs.
- Understands Complex Polymers – Can learn from polymer sequences without needing full molecular structures.
- Boosts Innovation – Helps scientists discover new materials for batteries, electronics, and sustainable products.



Productivity Gains:



- Fast & accurate polymer property predictions
- Reduces lab testing & costs
- Scalable & automated analysis
- Helps design sustainable polymers
- Easy integration via API/UI

Pain Points:

- Needs high-quality data
- External factors (temperature, aging) not always captured
- High computational power required
- Struggles with unseen polymers

Solutions:

- Improve data & use simulations
- Combine AI with traditional chemistry methods
- Optimize model for faster performance
- Continuous retraining with new dat

Roles Involved in the Application:

Admin: Manages database, user roles.
Researcher/User: Uploads polymer data, views predictions.
End Users (Polymer Engineers/Chemists)
Developers
Quality Assurance (QA) Testers

Users Interaction with the Application:

Input: Polymer structure (SMILES), other molecular descriptors.
Output: Predicted polymer properties (e.g., conductivity, thermal stability).

Purpose of Data Preprocessing :

- **Verification (Check if the Data is Correct):**

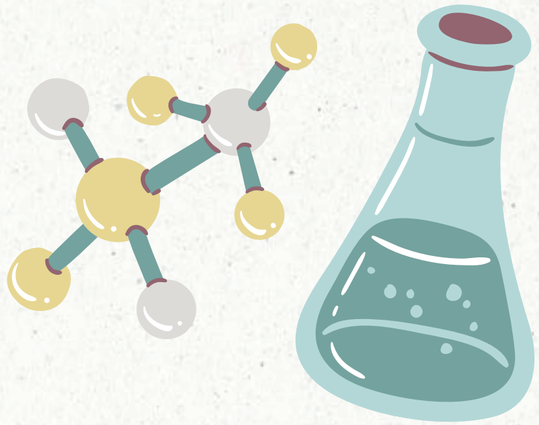
  Check if the dataset is loaded correctly.
  Verify if column names and data formats are as expected.
  Identify missing or inconsistent values.

- **Validation (Ensure Data is Meaningful & Usable):**

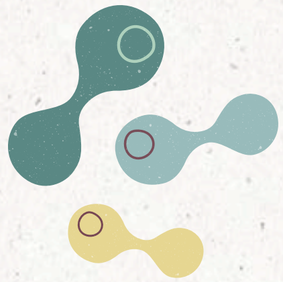
  Validate if all polymers have a valid SMILES string.
  Check if numerical values (e.g., Tg, melting point) are in a realistic range.
  Ensure no duplicate entries.

- **Normalization (Standardizing Data for Better Model Performance):**

  Convert all numerical values to a standard scale (e.g., between 0 and 1).
  Helps avoid bias due to large value differences.



TECHNICAL ARCHITECTURE



FRONTEND

MERN Stack is a full-stack web development framework that includes:

1. MongoDB – NoSQL database for storing data.
2. Express.js – Backend framework for handling API requests.
3. React.js – Frontend framework for building interactive user interfaces.

BACKEND

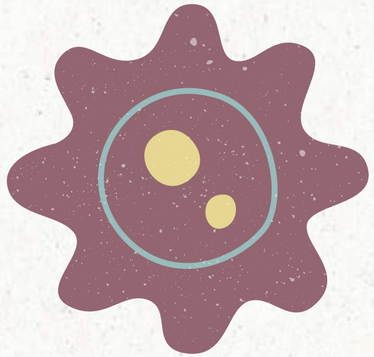
Technology Choices: Python-based backend using FastAPI, Flask, or Django. Purpose: Handles user requests and processes polymer sequence data. Interfaces with the AI model for predictions. Manages authentication, job queueing, and API integrations. in easy way

DATABASES

- PostgreSQL or MySQL – for structured polymer property data.
- MongoDB – for flexible storage of polymer sequences and descriptors.
- Cloud Storage (AWS S3, Google Cloud Storage) – for large-scale dataset management.

AI MODELS

- Primary Model: Transformer-based Model (RoBERTa-like architecture).
- Frameworks Used: PyTorch / TensorFlow for deep learning implementation.
- Training Method: Masked Language Modeling (MLM) pretraining.



SUMMARY OF THE RESEARCH PAPER

Transformer-based deep learning model designed for polymer property prediction. Traditional polymer research relies on costly and time-consuming experiments, but TransPolymer enables data-driven predictions using self-attention mechanisms and Masked Language Modeling (MLM) pretraining.

SUMMARY IN TABULAR FORMAT

ASPECT	DETAILS
MODEL USED	TransPolymer (Transformer-based RoBERTa-like model with self-attention)
PRETRAINING	Masked Language Modeling (MLM) with ~5M polymer sequences
BASELINE MODELS	Random Forest (ECFP), GNN, LSTM, ANN
DATASETS USED	10 datasets, including PE-I, PE-II (polymer conductivity), Egc, Egb (bandgap), Eea (electron affinity), Ei (ionization energy), Xc (crystallization tendency), EPS (dielectric constant), Nc (refractive index), and OPV (organic photovoltaic efficiency).
RESULTS	TransPolymer outperforms all baselines, achieving lower RMSE and higher R^2 scores on all datasets.
BEST PERFORMANCE GAIN	PE-I dataset (R^2 improved from 0.32 to 0.69 over previous best model)
LIMITATIONS	requires large-scale pretraining, computationally expensive, and has interpretability challenges despite attention visualizations.

THANK YOU!