Software Requirements Specification (SRS) for TransPolymer

Prepared by: G392
Project Name:TransPolymer

1. Introduction to Polymers

Polymers are large molecules composed of repeating structural units known as monomers. They are widely used in industries such as packaging, healthcare, electronics, and construction. Predicting polymer properties is crucial for designing new materials with desired characteristics.

2. Challenges in Polymer Prediction

- **Complex Molecular Structures:** Polymers have diverse and intricate molecular structures.
- **Experimental Cost:** Traditional property testing is expensive and time-consuming.
- **Limited Data Availability:** High-quality polymer datasets are scarce.
- **Lack of Standardized Models:** Existing predictive models often fail to generalize.

3. Scope of TransPolymer

TransPolymer is designed to predict polymer properties using transformer-based deep learning models. It aims to improve efficiency, reduce experimental dependency, and enhance material discovery.

4. Dataset and Data Sources

- **Data Sources:** Public and proprietary polymer datasets.
- **Data Types:** Molecular descriptors, structural properties, thermal properties.
- **Data Challenges:** Handling missing values, feature selection, and normalization.

5. Drawbacks of Existing Models

- **Poor Generalization:** Many ML models lack robustness across different polymer classes.
- **Limited Interpretability:** Black-box models make it difficult to understand predictions.
- **High Computational Cost:** Some deep learning models require excessive training time and resources.

6. Methodology & Workflow

6.1 Preprocessing

- Data Cleaning: Removing missing values and inconsistencies.
- Feature Extraction: Computing polymer descriptors.
- Normalization: Scaling numerical features for model compatibility.

6.2 Model Training
- **Architecture:** Transformer-based deep learning model.
- **Training Process:** Supervised learning on labeled polymer datasets.
- **Evaluation Metrics:** RMSE, MAE, and R^2 scores.

6.3 Model Deployment
- **Inference Pipeline:** Accepts polymer descriptors as input.
- **Prediction Output:** Provides estimated polymer properties.
- **API Integration:** Future scope for web-based access.

7. Technology Used & Libraries
- **Programming Language:** Python
- **Deep Learning Framework:** PyTorch
- **Data Processing:** Pandas, NumPy
- **Feature Engineering:** RDKit
- **Evaluation & Visualization:** Scikit-learn, Matplotlib

8. Business Use Cases
- **Material Science Research:** Faster discovery of new polymers.
- **Pharmaceuticals:** Designing drug delivery polymers.
- **Automotive & Aerospace:** Lightweight, high-strength materials.
- **Electronics:** Developing conductive and insulating polymers.

9. Roles of Users & End Users

9.1 Roles of Users
- **Developers:** Implement and maintain the model.
- **Researchers:** Validate predictions and integrate with experiments.
- **Data Scientists:** Optimize model accuracy and interpret results.

9.2 End Users
- **Scientists & Engineers:** Utilize predictions for material discovery.
- **Industry Professionals:** Apply insights to manufacturing and product development.
- **Academics & Students:** Conduct research in polymer science and AI.

10. System Architecture Overview
The system follows a pipeline approach:
1. **Data Collection & Preprocessing** → Cleaning and transforming raw polymer data.
2. **Feature Engineering** → Extracting molecular descriptors.
3. **Model Training** → Transformer-based deep learning model.
4. **Evaluation & Prediction** → Assessing model accuracy and generating results.
5. **Deployment** → API integration and potential user interface development.

11. Proposed Methodology

The methodology involves:

1.**Polymer Tokenization**: Encoding repeating units of polymers using SMILES and additional descriptors (e.g., polymerization degree, composition).

2.**Pretraining on Large Data**: MLM pretraining on ~5M unlabeled polymer sequences.

3.**Fine-tuning**: Fine-tuning on multiple benchmark datasets to enhance predictive accuracy.

4.**Data Augmentation**: Generating non-canonical SMILES for improved learning.

5.**Transformer Encoder**: Using self-attention mechanisms to capture chemical insights.

12. Summary of the Research Paper

The research paper presents TransPolymer, a 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.

#Key Contributions:

1.Introduced a chemically-aware tokenizer for polymer sequences.

2.Pretrained on ~5M augmented polymer sequences from the PI1M database.

3.Fine-tuned and evaluated on 10 benchmark datasets covering conductivity, bandgap, crystallization tendency, and dielectric constant.

4.Achieved state-of-the-art (SOTA) performance, outperforming baseline models like Random Forest, GNNs, and LSTM.

5.Demonstrated the impact of pretraining, data augmentation, and self-attention mechanisms for learning polymer representations.

ARCHITECTURE DIAGRAM: