

A

PROJECT SCHOOL REPORT ON

TRANSPOLYMER-AN AI BASED TRANSFORMER MODEL

Submitted By

JAKKINAPALLY BABITHA	23P81A0523
SAMBU AKHILA	23P81A0553
SURKANTI KAVYA REDDY	23P81A0557
YELCHALA LAHARI	23P81A0565
BEJAWADA MANISH	23P81A6910
KOTHAPELLI SAI DEEPAK	23P81A6933

Under the guidance of

Mrs.K.Navatha

Assistant Professor, CSE (AIML), KMCE



KESHAV MEMORIAL COLLEGE OF ENGINEERING

Chinthapallyguda village, Ibrahimpatnam, Hyderabad, Telangana 500058

May 2025



KESHAV MEMORIAL COLLEGE OF ENGINEERING

A Unit of Keshav Memorial Technical Education (KMTES)

Approved by AICTE, New Delhi & Affiliated to Jawaharlal Nehru Technological University, Hyderabad

CERTIFICATE

This is to certify that the project work entitled “TRANSPOLYMER-AN AI BASED TRANSFORMER MODEL” is a bonafide work carried out by “Jakkinapally Babitha, Sambu Akhila, Surkanti Kavya Reddy, Yelchala Lahari, Kothapelli Sai Deepak, Bejawada Manish” of II-year II semester Bachelor of Technology in CSE/CSE (IOT) during the academic year 2024-2025 and is a record of bonafide work carried out by them.

Project Mentor

Mrs.K.Navatha

Assistant Professor, CSE (AIML)

KMCE

ABSTRACT

The advancement of polymer science heavily relies on the accurate prediction of polymer properties, which has traditionally required expensive experimental or simulation-based methods. With the rise of machine learning and natural language processing (NLP), new approaches have emerged that treat polymers as sequential data, much like a language. In this project, we present TransPolymer, a Transformer-based language model specifically designed to predict various polymer properties using SMILES notation as input. By leveraging the power of the self-attention mechanism, TransPolymer effectively learns chemical-aware representations of polymer sequences, bypassing the limitations of traditional graph-based models which require explicit structural information. Our model undergoes pretraining using a masked language modeling strategy on a large dataset of unlabeled polymers and is fine-tuned on specific property prediction tasks. TransPolymer shows superior performance across multiple benchmarks and property categories, such as tensile strength, ionization energy, and molecular weight. The results affirm the model's capability to understand structure–property relationships in polymers, thereby offering a powerful, scalable, and data-driven alternative for polymer design and discovery.

CONTENTS

S.no	Title	Page no
	ABSTRACT	i
	TABLE OF CONTENTS	ii
	LIST OF FIGURES	iii
1.	Introduction	6
	1.1 Why Understanding Polymer Properties Matters	
	1.2 Role of SMILES in Molecular Representation	
	1.3 How AI Learns from Molecular Patterns	
2.	Literature Survey	10
	2.1 Traditional Methods of Polymer Property Prediction	
	2.2 The Transformer Revolution and ChemBERTa	
	2.3 The Challenge of Polymer Property Prediction	
3.	Software Requirements Specification	13
4.	Proposed Work Architecture, Technology Stack, Implementation Details	17
	4.1 Proposed Work Architecture	
	4.2 Technology Stack	
	4.3 Implementation Details	
	4.4 Interfaces and Communication	
5.	Results and Discussions	30
	5.1 High Prediction Accuracy	
	5.2 Generalizability to New Data	
6.	Conclusion and Future Scope	32
	6.1 Conclusion	
	6.2 Future Scope	
7.	References	35

LIST OF FIGURES

Fig No	Figure Name	Page No
1	Architecture Diagram	18
2	Different layers of architecture diagram	19
3	Login Page	25
4	Student Registration Page	26
5	Scientist Registration Page	26
6	Home Page	27
7	Prediction Page	27
8	About Page	28
9	Help Page	28

CHAPTER 1

INTRODUCTION

Polymers are large molecules made up of repeating units called monomers and form the basis of many materials we use every day. From plastic packaging and electronics to aerospace and medical devices, polymers are integral to modern life due to their flexibility, durability, and versatility. Common examples include polyethylene and polypropylene (packaging), polystyrene (insulation), and PVC (pipes and window frames). Even natural materials like DNA, silk, and rubber are polymers. As demand grows, so does the need to design and understand these materials more effectively.

1.1 Why Understanding Polymer Properties Matters

The performance of a polymer in a given application depends on its properties—tensile strength, thermal stability, elasticity, conductivity, and solubility. For instance, a polymer used in cars must endure heat and stress, while biomedical applications require biocompatibility. Misjudging these properties can lead to failures and safety issues. Accurate prediction is key for innovation and safety.

Traditionally, these properties are determined through experimental testing—a method that's slow, costly, and labor-intensive. Experiments need materials, specialized equipment, and expert labor. This becomes inefficient when evaluating thousands of polymer candidates and may lead to inconsistent results due to environmental or procedural variations.

With digitization, data-driven science has emerged as a powerful alternative. This method uses computational models to extract insights from large datasets, predicting polymer behavior before synthesis. It significantly reduces cost and time and enables broader material exploration.

1.2 Role of SMILES in Molecular Representation

SMILES (Simplified Molecular Input Line Entry System) is a format that encodes molecular structures as strings of text. It simplifies the way molecules are represented—for example, water is "O", ethylene is "C=C", and benzene is "c1ccccc1". SMILES strings are both human- and machine-readable, making them ideal for computational modeling and AI applications in chemistry.

AI has revolutionized how we approach complex problems, and materials science is no exception. AI can analyze large chemical datasets, predict properties, optimize synthetic routes, and discover new materials. It handles complex relationships that traditional methods might miss—critical in polymer research where many variables interact.

Sequence-based AI models, such as Transformers, are especially effective for analyzing SMILES strings. They process molecules like sentences, understanding how the sequence of atoms and bonds affects behavior. These models capture both local and global patterns in molecular structures.

A major breakthrough in these models is the **attention mechanism**, which allows the model to focus on the most relevant parts of a molecule. This is vital in chemistry, where distant atoms can influence each other. By selectively weighing input components, attention-based models make more accurate predictions.

AI brings clear advantages to polymer research: reduced need for experimentation, continuous learning from new data, and the ability to reveal hidden patterns. This accelerates innovation and helps researchers make informed decisions faster.

Examples of Key Polymer Properties

Important polymer properties include:

- Tensile Strength: Resistance to breaking under tension.
- Ionization Energy: Energy needed to remove an electron.
- Electron Affinity: Tendency to gain an electron.
- Log P: Indicates hydrophobicity/hydrophilicity.
- Refractive Index: Light-bending property of the material.
- Molecular Weight: Total mass of a molecule.

Understanding these helps design polymers for specific roles like flexible packaging, transparent films, or drug delivery systems.

1.3 How AI Learns from Molecular Patterns

AI models learn by analyzing large datasets of molecules and their properties. Through training, the model optimizes its internal parameters to minimize prediction errors. When given a new SMILES string, it

applies what it has learned to make property predictions. Unlike memorization, AI models generalize from past data to make accurate forecasts on unseen inputs.

High-quality, diverse datasets are essential for training reliable models. Inaccurate or narrow datasets can lead to poor predictions. Ethical concerns like transparency, data integrity, and reproducibility are also crucial. AI models must be interpretable and validated against experiments to ensure trust and usability in scientific research.

Applications in Industry and Research

AI-driven polymer prediction tools are transforming industries:

- **Materials science:** Lightweight, durable components.
- **Medicine:** Biocompatible polymers for implants and drug delivery.
- **Sustainability:** Biodegradable alternatives to plastics.
- **Electronics:** Polymers for insulation or conductivity.

By combining chemistry and AI, researchers can design better materials faster and more efficiently.

Vision for Smarter, Faster Materials Discovery

The future of polymer research lies in AI-integrated pipelines—from design to validation. Imagine inputting desired properties and receiving optimized polymer candidates instantly. This intelligent system will speed up innovation, reduce costs, and make advanced materials accessible to more researchers. As AI becomes more deeply embedded in science, it will redefine the landscape of polymer discovery and development.

CHAPTER 2

LITERATURE SURVEY

The task of predicting polymer properties has undergone a significant transformation over the past few decades. Initially, the field relied heavily on traditional experimental methods. These methods involved the synthesis of polymers in the laboratory, followed by numerous rounds of testing to evaluate their physical and chemical properties. These experimental approaches, though highly valuable, were not only time-consuming but also resource-intensive. Each trial involved substantial material costs, specialized equipment, and considerable amounts of time spent in trial and error, making them inefficient for large-scale studies or when dealing with novel materials that had not been previously synthesized.

2.1 Traditional Methods of Polymer Property Prediction

Historically, researchers had to rely on fundamental chemical intuition, guided by a limited set of empirical rules, to predict how polymers might behave. Early studies in polymer science emphasized systematic experimentation based on chemical knowledge, and synthetic chemists employed a trial-and-error approach to achieve desired material properties. These methods were foundational for many early discoveries and provided essential benchmarks for future work in polymer research.

However, as the complexity of materials science grew, it became evident that a more efficient and scalable solution was needed, especially in contexts where a large number of potential polymer compositions needed to be evaluated. This led to the gradual introduction of computational tools and techniques that could help predict polymer properties without the need for exhaustive experimental testing. As computational methods improved, it became possible to automate parts of the discovery process, enabling faster iterations and more efficient testing of hypotheses.

Transition to Machine Learning-Based Approaches With the rise of machine learning (ML) and computational methods in the early 2000s, the field began to shift away from purely experimental approaches toward data-driven methodologies. Classical ML algorithms, such as Random Forests, Support Vector Machines (SVMs), and simple Artificial Neural Networks (ANNs), started being used for property prediction tasks. These models worked by leveraging handcrafted molecular descriptors—quantitative representations of molecular structures—that could be extracted from chemical information.

Early ML models used a variety of molecular descriptors to capture aspects such as molecular weight, the

number of specific atom types, functional group presence, and topological features. The idea was that these descriptors could serve as input features for ML models, which would then output predictions for properties such as tensile strength, melting point, or solubility. While these models showed some success, their reliance on manually chosen descriptors posed a limitation. The quality and predictive power of these models depended on the choice of features, which were often selected based on the intuition of the researcher rather than any objective, data-driven process.

As a result, classical machine learning models faced challenges in capturing the highly complex relationships between molecular structure and properties, particularly when dealing with more intricate chemical systems, such as polymers. These limitations became more evident as the size and diversity of chemical datasets grew, and researchers began seeking more powerful, flexible modeling techniques that could handle complex, high-dimensional data.

The Advent of Deep Learning and Sequence Models The introduction of deep learning represented a transformative shift in the ability to model and predict molecular properties. Deep learning techniques, which use multiple layers of processing units to automatically extract features from raw data, offered a powerful alternative to the feature engineering approaches that dominated earlier methods. The rise of deep learning coincided with the advent of more advanced computational tools and the increased availability of large chemical datasets, such as those containing SMILES (Simplified Molecular Input Line Entry System) strings or molecular graphs.

In the context of polymer property prediction, deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) were applied to SMILES strings. These models allowed researchers to treat molecular structures as sequences, much like words in a sentence, which could be processed using techniques developed for natural language processing (NLP). RNNs and LSTMs, in particular, were capable of capturing the sequential nature of molecular structures, making them an appealing choice for tasks involving SMILES.

However, despite their initial promise, RNNs and LSTMs had limitations in dealing with long-range dependencies within molecular sequences. These models processed sequences token by token, which hindered their ability to capture global context—especially for large, complex molecules like polymers, which may have long-range interactions across different parts of their molecular structure. As a result, these earlier models struggled to effectively predict the properties of polymers, which typically feature long, repeating units and structural complexities that go beyond simple sequence relationships.

2.2 The Transformer Revolution and ChemBERTa

The real breakthrough came with the introduction of transformer models, particularly in the field of natural language processing (NLP). The 2017 paper "Attention Is All You Need" by Vaswani et al. introduced the transformer architecture, which utilized self-attention mechanisms to process sequences in parallel and capture long-range dependencies across the entire input sequence. Unlike RNNs and LSTMs, transformers could attend to all parts of a sequence at once, making them far more effective in handling long-range dependencies and global context.

In the years following the success of transformers in NLP, their application in other fields, including chemistry, began to gain momentum. One key example is ChemBERTa, a transformer based model pretrained on a vast corpus of chemical data, including SMILES strings. ChemBERTa is capable of generating embeddings (dense vector representations) of molecular structures, which can be used for a variety of downstream tasks, such as predicting molecular properties, toxicity, solubility, and reactivity.

The major advantage of transformer-based models like ChemBERTa lies in their ability to automatically learn complex representations of molecular structures without requiring manual feature engineering. By leveraging large datasets, transformers can capture subtle chemical patterns and interactions that might not be apparent through traditional descriptors. As a result, these models have demonstrated superior performance on tasks such as toxicity prediction, drug discovery, and even polymer property prediction. Moreover, transformers have shown flexibility in handling different types of molecular representations. While SMILES remains the most commonly used format due to its simplicity and compatibility with NLP models, other representations like InChI (International Chemical Identifier) and molecular graphs have also been explored. InChI, for example, offers a more structured and hierarchical way to represent molecular information, but its adoption has been slower compared to SMILES, due to the latter's ease of use and widespread availability in chemical databases.

2.3 The Challenge of Polymer Property Prediction

While transformer-based models like ChemBERTa have shown immense promise for small molecule property prediction, the application of these models to polymers presents unique challenges. Polymers are large, complex molecules composed of repeating units, which can result in extremely long SMILES strings that are difficult to process effectively with traditional models. In addition, polymers exhibit a wide range of properties that are influenced not only by their molecular structure but also by factors such as

molecular weight, degree of polymerization, and the arrangement of monomer units.

Unlike small molecules, which often exhibit relatively predictable behavior based on their molecular structure, polymer properties are more dependent on the organization and sequence of monomeric units within the polymer chain. These long-range interactions and structural dependencies pose significant challenges for deep learning models. Moreover, polymers often exhibit a level of variability in their structure and behavior that is difficult to capture using traditional computational approaches.

Despite these challenges, there is growing interest in applying transformer-based models to polymers. Recent studies have explored how SMILES-based transformers, like ChemBERTa, can be adapted to handle polymer structures, either by modifying the input representation or by introducing novel training techniques. However, much of this work remains in the early stages, and more research is needed to fine-tune these models to capture the unique properties of polymers accurately.

The literature on polymer property prediction highlights an exciting evolution in the field—from experimental trial and error, to classical machine learning approaches, and now to advanced deep learning techniques such as transformers. While transformer-based models like ChemBERTa have shown tremendous potential in understanding molecular data, their application to polymers still presents several challenges that must be addressed.

One promising avenue for future research is the development of hybrid models that combine both molecular descriptors and deep learning techniques. These models could leverage the strengths of traditional feature engineering while benefiting from the power of deep learning to capture complex relationships in the data. Additionally, further research into transfer learning, where models pretrained on large datasets can be fine-tuned for specific tasks, could be valuable in improving the performance of polymer property prediction models.

Furthermore, the increasing availability of large, high-quality datasets in the chemical and materials sciences holds great promise for advancing the field. With more data and improved models, it is likely that the next generation of polymer property prediction tools will be more accurate, faster, and more widely applicable, ultimately transforming the way researchers design and optimize new polymer materials.

CHAPTER 3

SOFTWARE REQUIREMENTS SPECIFICATION (SRS)

- **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.

- **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.

- **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.

- **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.

- **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.

- **Methodology & Workflow**

1. **Preprocessing**

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

2. **Model Training**

- **Architecture:** Transformer-based deep learning model.
- **Training Process:** Supervised learning on labeled polymer datasets.
- **Evaluation Metrics:** RMSE, MAE, and R² scores.

3. **Model Deployment**

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

- **Technology Used & Libraries**

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

- **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.

- **Roles of Users & End Users**

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.

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.

- **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.

- **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.

- **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.

CHAPTER 4

PROPOSED WORK ARCHITECTURE, TECHNOLOGY STACK, IMPLEMENTATION DETAILS

4.1 Proposed Work Architecture

The Polymer Prediction System Architecture represents a modern and modular approach to solving chemical property prediction challenges. By incorporating a user-friendly interface, powerful ML backend, and reliable data storage mechanisms, it stands as a scalable solution for researchers and industrial chemists alike. Its seamless integration with Hugging Face Spaces and advanced visualization features makes it not only functional but also accessible and easy to maintain.

The Transformer encoder is a core component in models like BERT and ChemBERTa, designed to process sequences such as tokenized SMILES strings and generate rich, context-aware embeddings. It begins with an embedding layer that converts tokens into high-dimensional vectors, incorporating positional encodings to retain token order. Each of the N stacked layers includes a Multi-Head Attention mechanism that computes attention scores using learned Query, Key, and Value matrices, allowing the model to focus on different parts of the sequence in parallel. Outputs then pass through residual connections and layer normalization to stabilize training. A Feed Forward Network (FFN), applied independently to each token, further processes the data using linear transformations and non-linear activations. Each layer ends with another round of residual connection and normalization, enabling the model to learn complex relationships for downstream tasks.

The TransPolymer model architecture uses a shared Transformer encoder for two tasks: masked token prediction during pretraining and property regression during fine-tuning. SMILES strings are tokenized and embedded before being processed by the encoder. During pretraining, masked tokens are predicted using the Prediction Head, helping the model learn chemical context. In fine-tuning, the <s> token embedding is passed to a Regressor Head to predict polymer properties like tensile strength and ionization energy. This dual-head setup enables effective transfer learning from chemical language modeling to property prediction.



Polymer Prediction System Architecture

🔥 Deployed on Hugging Face Spaces - Streamlit Frontend with Integrated ML Pipeline

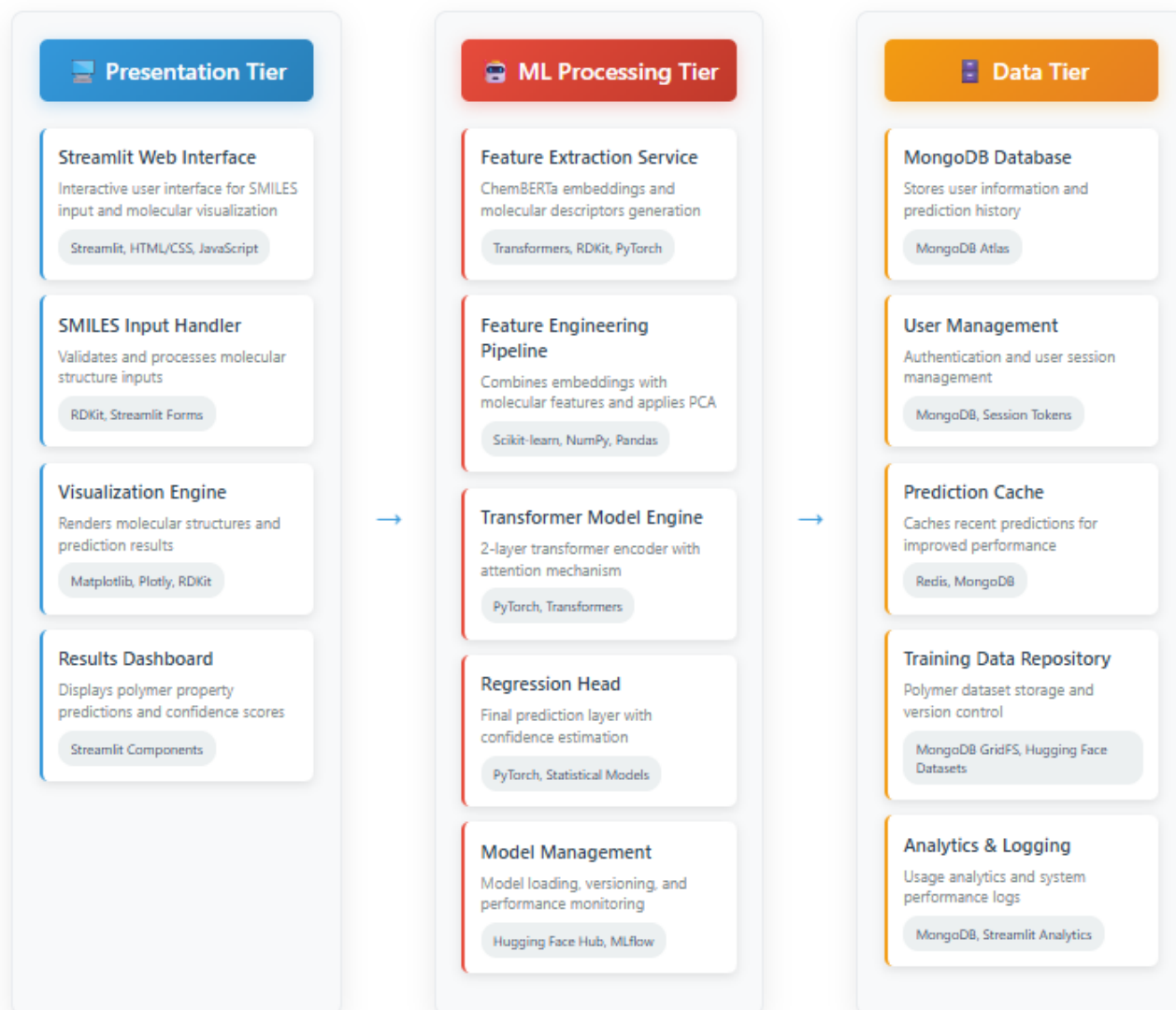


Figure 1. Architecture diagram

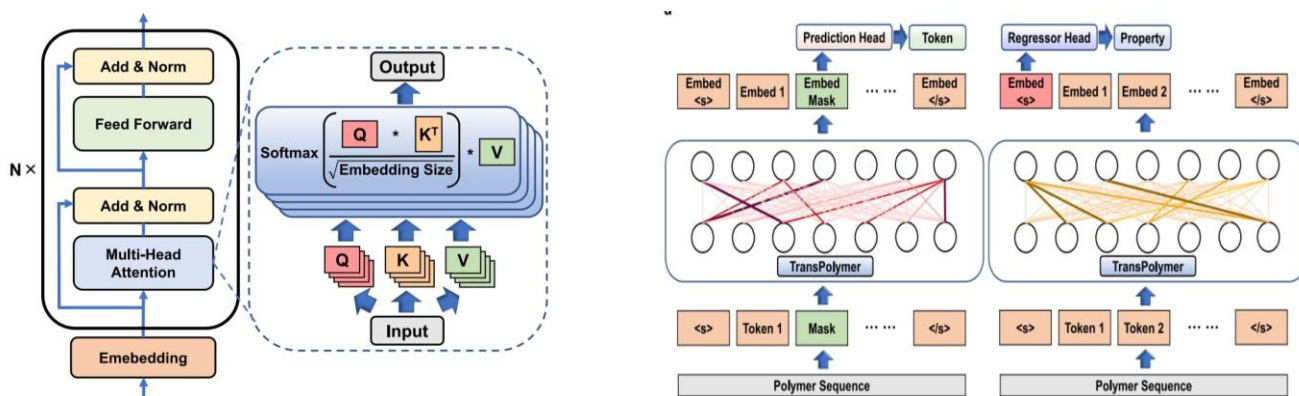


Figure 2. Different layers of the architecture diagram

The Polymer Prediction System Architecture is a multi-tiered structure designed to facilitate end-to-end prediction of polymer properties from SMILES (Simplified Molecular Input Line Entry System) notations. This architecture ensures a robust pipeline integrating a user-friendly interface, efficient machine learning processing, and scalable data storage. The system is deployed on Hugging Face Spaces and is built with modularity in mind to support real-time predictions and system monitoring. This document elaborates on each component in the architecture in detail.

Presentation Tier

The Presentation Tier serves as the primary interface for users. It is designed with Streamlit, a Python framework ideal for building interactive web applications for data science and machine learning models.

Streamlit Web Interface:

This component provides an intuitive UI for users to input SMILES strings and visualize the molecular structures. Technologies used include Streamlit, HTML/CSS, and JavaScript. The interface is crafted for ease of use, especially for researchers who may not have a programming background.

SMILES Input Handler:

This sub-module ensures the validation of the SMILES strings entered by the user. Using RDKit for chemical informatics and Streamlit Forms for structured input, it preprocesses the molecular data before sending it to the backend.

Visualization Engine:

After prediction, users can view the molecular structure and related chemical properties. This engine leverages libraries such as Matplotlib, Plotly, and RDKit to provide high-quality visualizations, aiding

better interpretation of results.

Results Dashboard:

This section of the interface displays polymer property predictions and their corresponding confidence levels. It employs Streamlit Components to dynamically render prediction results in real-time, facilitating transparent and interpretable results.

ML Processing Tier

The core intelligence of the architecture lies within the ML Processing Tier. It transforms raw SMILES input into meaningful predictions using state-of-the-art machine learning and deep learning techniques.

Feature Extraction Service:

This service is responsible for converting SMILES strings into embeddings using ChemBERTa and generating molecular descriptors via RDKit. These features are fundamental for capturing the chemical and structural properties of the molecules.

Feature Engineering Pipeline:

After extraction, the features are refined using scikit-learn and NumPy. This includes normalization, Principal Component Analysis (PCA), and feature selection. The goal is to enhance the model's ability to generalize and reduce dimensionality.

Transformer Model Engine:

At the heart of the prediction system lies a 2-layer Transformer encoder. Leveraging PyTorch and Hugging Face's Transformers library, this engine applies self-attention mechanisms to capture intricate dependencies in the molecular structure, ensuring high prediction accuracy.

Regression Head:

Once the features pass through the transformer layers, they are fed into a regression head. This final prediction layer, built using PyTorch and statistical models, estimates polymer properties such as tensile strength, ionization energy, and others. Confidence scores are also calculated.

Model Management:

Model lifecycle operations including loading, versioning, and monitoring are managed via Hugging Face Hub and MLflow. This enables continuous integration/continuous deployment (CI/CD) workflows and experiment tracking for model updates and performance audits.

Data Tier

The Data Tier manages storage, retrieval, and caching, ensuring the system remains scalable and responsive under various loads.

MongoDB Database:

MongoDB Atlas is used to store user information, input records, and prediction histories. The NoSQL schema enables flexible and efficient data querying, making it suitable for evolving project requirements.

User Management:

Authentication and session tracking are facilitated through MongoDB in combination with Session Tokens. This component secures user data and customizes user experiences based on login sessions.

Prediction Cache:

To reduce redundant computations and accelerate response times, recent predictions are cached using Redis and MongoDB. This cache enables the reuse of predictions for frequently queried SMILES inputs.

Training Data Repository:

All datasets used for training and evaluation are stored using MongoDB GridFS and Hugging Face Datasets. This hybrid approach ensures both scalability and version control.

Analytics & Logging:

System performance logs and user activity metrics are continuously monitored. MongoDB, Streamlit Analytics, and custom logging scripts track errors, latencies, and usage patterns to improve reliability and maintainability.

Deployment Strategy

The entire system is deployed on Hugging Face Spaces, offering seamless integration with the ML pipeline. This allows the model to be served in real-time via a cloud-hosted Streamlit frontend, enabling wide accessibility with minimal setup.

4.2 Technology Stack

TransPolymer is built using a lightweight, modular, and cloud-based technology stack that integrates machine learning with chemical informatics through an interactive web interface.

Frontend

The user interface is developed using **Streamlit**, a Python framework for building interactive web applications. It allows seamless integration of widgets, real-time updates, and visualizations without requiring HTML, CSS, or JavaScript. Users can input SMILES strings, upload files, and view prediction results directly in the browser.

Core Programming Language

Python is used throughout the project for building the UI, training models, handling data, and managing application logic. Its simplicity and robust ecosystem support fast development and easy integration of multiple libraries.

Machine Learning Tools

The project uses:

- **PyTorch** and **Hugging Face Transformers** to fine-tune a Transformer-based model (such as ChemBERTa) for molecular property prediction.
- **Scikit-learn**, **XGBoost**, and **LightGBM** for multi-target regression and evaluation tasks.
- **Pandas** and **NumPy** for data handling, feature extraction, and model input preparation.

Chemistry Libraries

RDKit is used to convert SMILES strings into molecular descriptors and fingerprints, which are then combined with Transformer-based embeddings to enhance prediction accuracy.

Data Handling

All chemical data and property labels are stored in **CSV files**. Data preprocessing, transformation, and in-memory operations are handled using **Pandas**, eliminating the need for external databases.

Deployment

The application is deployed using **Hugging Face Spaces**, which hosts the Streamlit app in the cloud. It is connected to **GitHub** for continuous integration, allowing automatic updates with each code commit.

Additional Tools

- **Matplotlib**, **Seaborn**, and **Plotly** are used for visualizing prediction outputs and data insights.
- **PyTest** is optionally used to ensure the correctness and reliability of code components.

Design Philosophy

TransPolymer is designed to be **modular**, **simple**, and **accessible**. With cloud-based deployment and a fully Python-driven stack, the system requires no setup from the user's side, making it ideal for students, researchers, and developers.

4.3 Implementation Details

Input and Output

- **Input:** SMILES (Simplified Molecular Input Line Entry System) string representing a polymer.
- **Output:** Six predicted polymer properties:
 - Tensile Strength
 - Ionization Energy
 - Electron Affinity
 - Log P (partition coefficient)
 - Refractive Index
 - Molecular Weight
- **Purpose:** Helps evaluate physical and chemical behavior of polymers for scientific and industrial applications.

Data Preprocessing

- **SMILES Randomization:**
 - Generates alternative valid representations.
 - Reduces overfitting and increases model robustness.
- **Descriptor Calculation:**
 - Uses **RDKit** to compute molecular descriptors.
 - Adds handcrafted features capturing essential chemical information.

Tokenization and Embedding

- **Tokenizer:** ChemBERTa tokenizer (character-level, chemistry-aware).
- **Embedding:**
 - Tokenized SMILES are passed through **ChemBERTa**.
 - Outputs contextual embeddings representing molecular structure and relationships.

Model Architecture

- **Backbone:** ChemBERTa Transformer pretrained on chemical data.
- **Head:** Multi-head regression layer to predict six numerical properties.
- **Function:** Supports multi-target prediction in a single forward pass by learning complex chemical patterns.

Training Strategy

- **Data Split:**
 - 80% Training
 - 10% Validation
 - 10% Testing
- **Optimizer:** AdamW (effective with weight decay).
- **Loss Function:** Mean Squared Error (MSE) – suitable for regression tasks.
- **Regularization:** Early stopping to prevent overfitting and save computation time.

Evaluation Metrics

- **Mean Absolute Error (MAE):** Average magnitude of prediction errors.
- **Root Mean Squared Error (RMSE):** Penalizes larger errors; sensitive to outliers.
- **R² Score:** Measures how well predictions approximate actual values.

4.4 Interfaces and Communication

Internal Interfaces

These manage the flow within the application:

- **Streamlit ↔ Python Functions:**

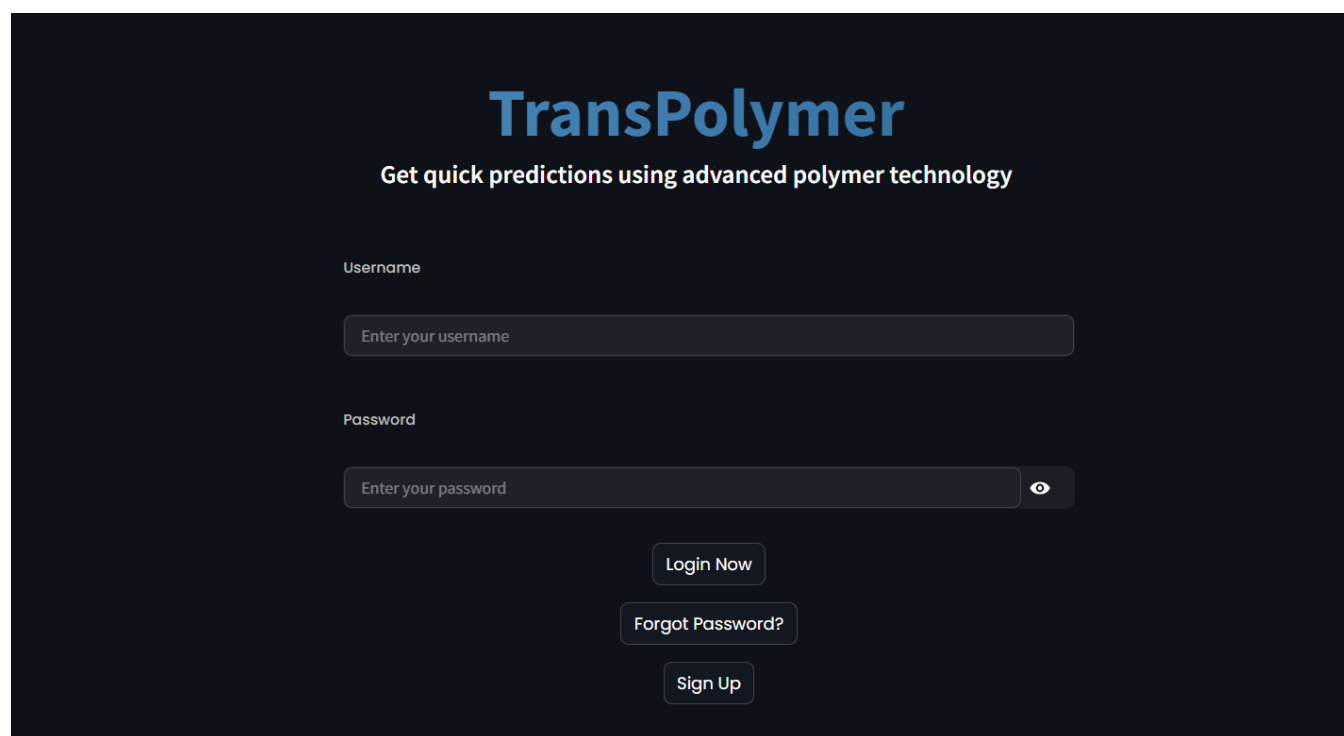
User actions in the UI trigger backend functions for preprocessing, embedding, and prediction.

- **Preprocessing ↔ Embedding ↔ Model:**

- RDKit computes molecular descriptors.
- ChemBERTa processes SMILES into embeddings.
- Final input is passed to the trained model for prediction.

- **Model ↔ MongoDB:**

After prediction, the input and output are logged in the database.



TransPolymer

Get quick predictions using advanced polymer technology

Username

Enter your username

Password

Enter your password

Login Now

Forgot Password?

Sign Up

Figure 3. Login Page

TransPolymer
Get quick predictions using advanced polymer technology

Sign up as

Student

Username

manish

Email

bejawadamanish0807@gmail.com

Student ID

23p81a6910

College Name

KMCE

State

TELENGANA

Password

Confirm Password

Press Enter to apply

Create Account

Back to Login

Figure 4. Student Registration Page

The form is designed for students, requiring details like username, email, student ID, college name, state, and password. Users can create an account or return to the login page.

TransPolymer
Get quick predictions using advanced polymer technology

Sign up as

Scientist

Scientist

Student

Researcher

Choose a username

Your email address

Organisation

Your organisation name

State

Your state

Password

Confirm Password

Create a password

Confirm password

Create Account

Back to Login

Figure 5. Scientist Registration Page

The "TransPolymer" registration form with a dropdown menu for selecting the user role—Scientist,

Student, or Researcher. The form adapts based on the selected role, requesting information such as username, email, organization, state, and password. This is scientist registration form.

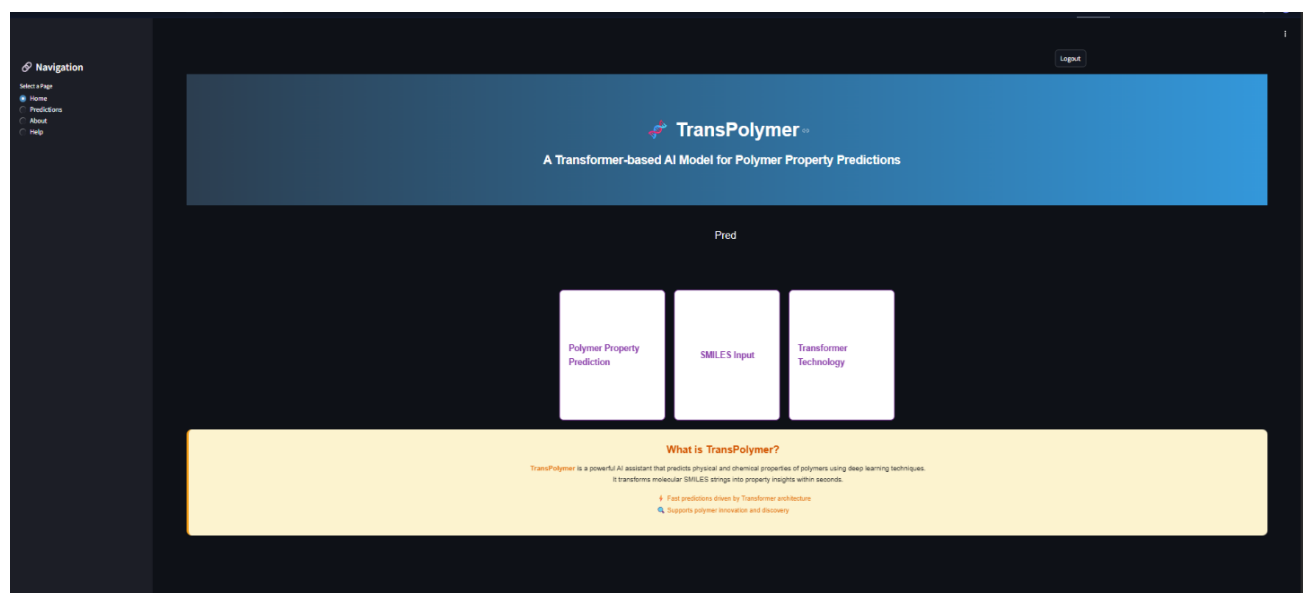


Figure 6.Home Page

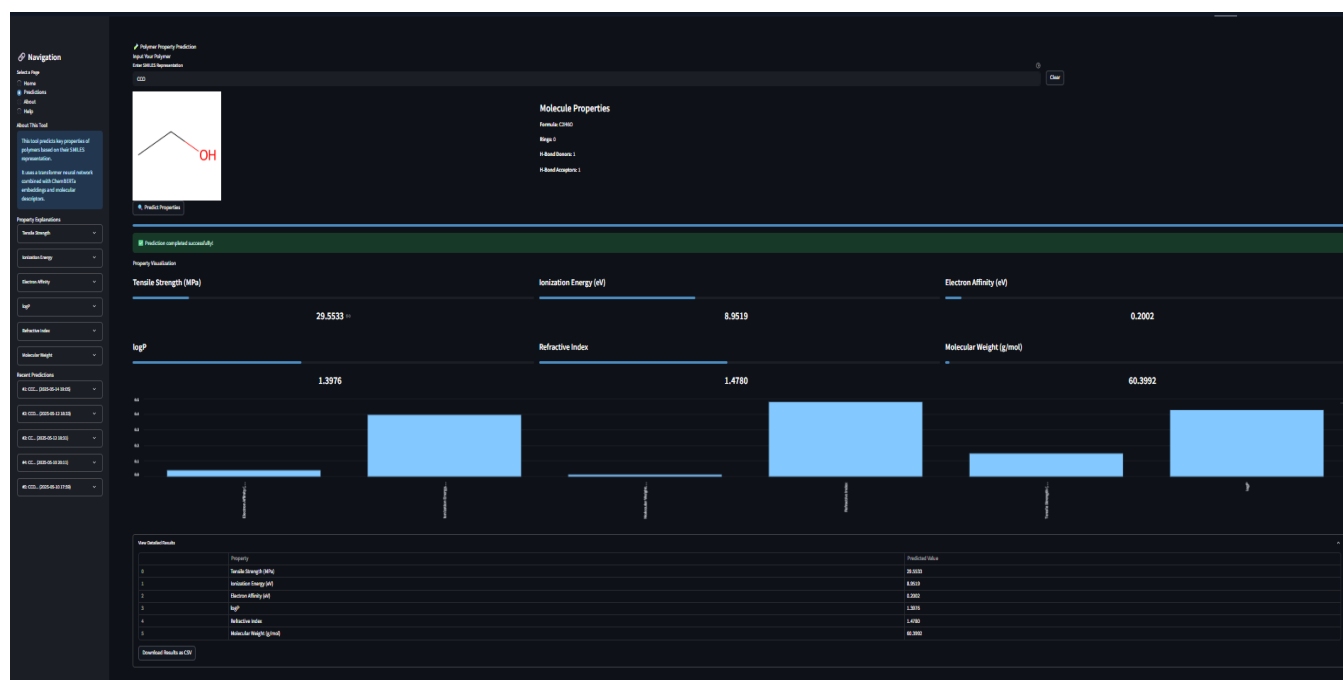


Figure 7.Prediction page

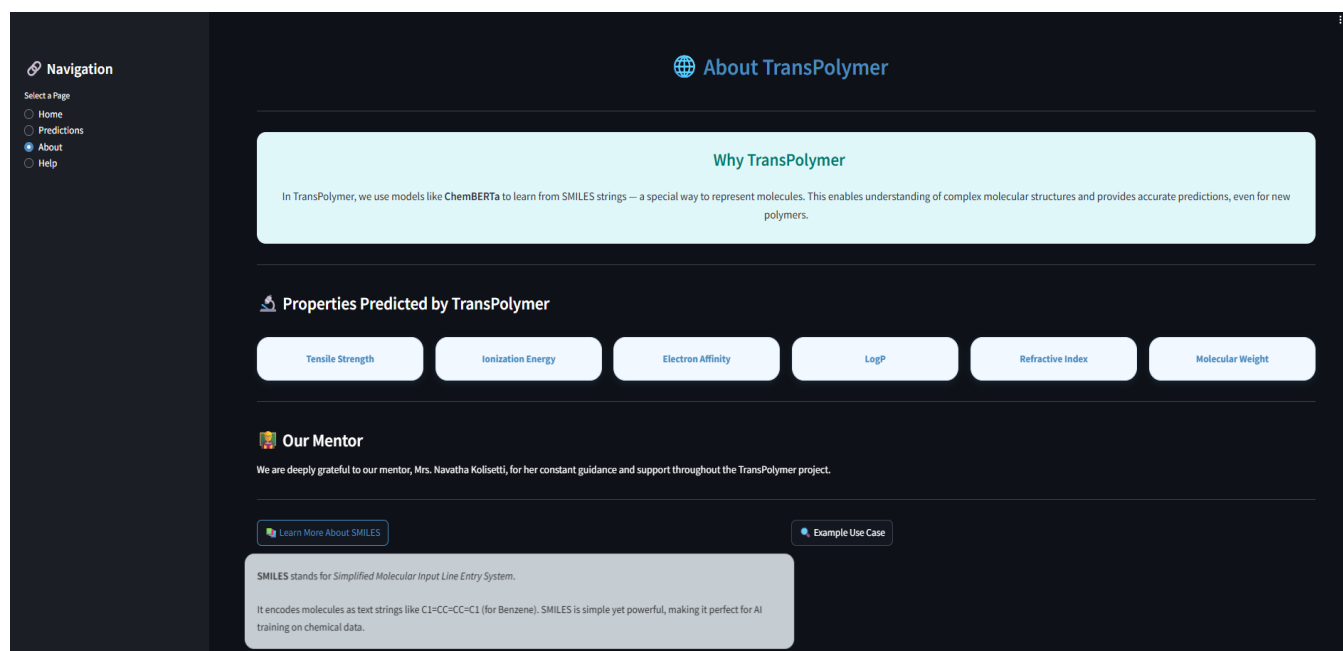


Figure 8.About Page

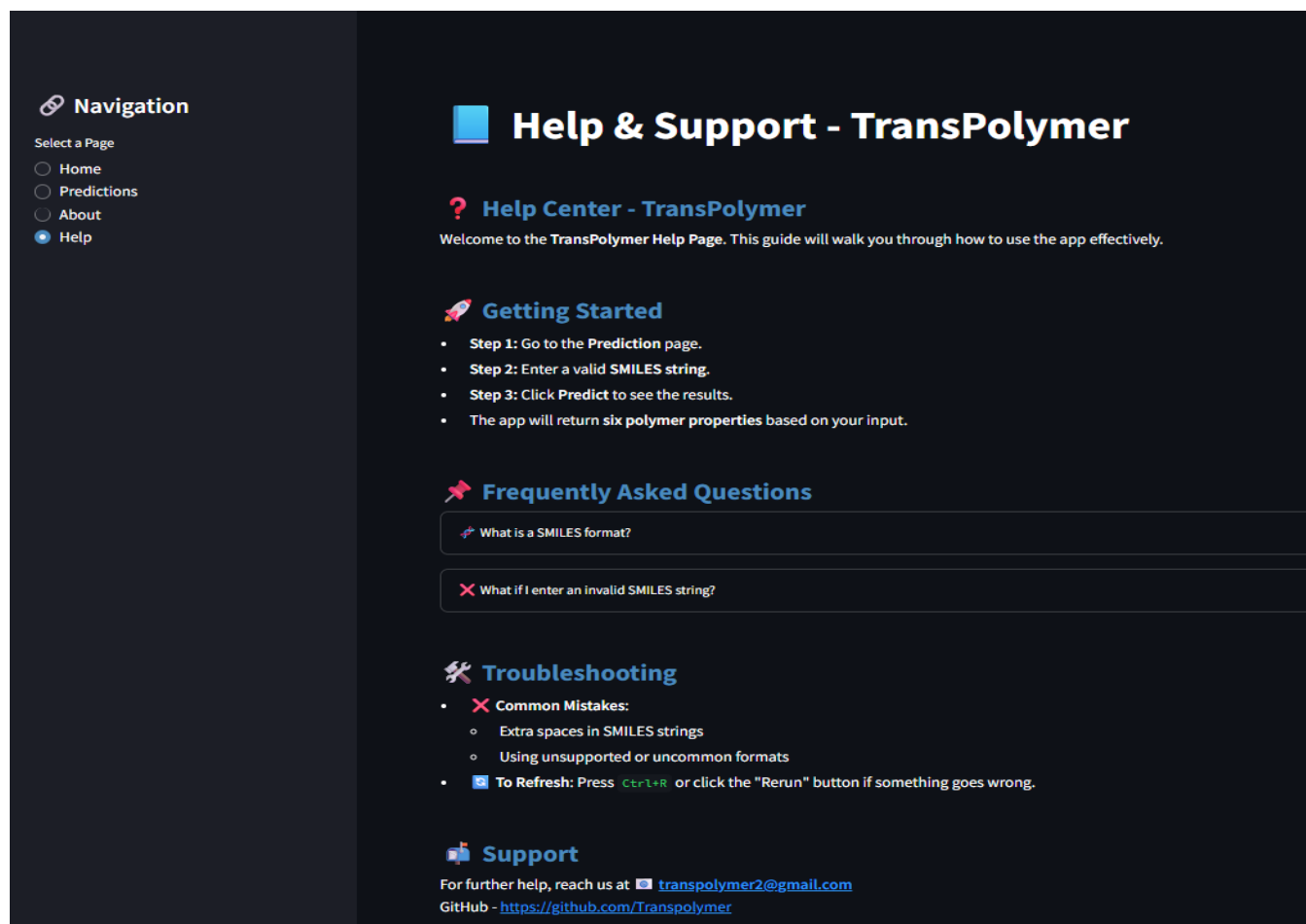


Figure 9.Help Page

External Interfaces

These connect the system to external tools or platforms:

- **User ↔ Streamlit Web App:**
Users input SMILES strings and view predicted results via a browser interface.
- **GitHub ↔ Hugging Face Spaces:**
Source code is versioned in GitHub and automatically deployed to Hugging Face Spaces.
- **MongoDB (Cloud or Local):**
Stores historical data for analysis, accessible via pymongo.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 High Prediction Accuracy

Our AI model performed impressively when it came to predicting the physical and chemical properties of polymers. Especially for properties like tensile strength and molecular weight, the predictions were quite close to real-world values. This shows that the model is not just guessing—it has actually learned useful patterns from the data.

1. Reduced Experimentation Effort

Instead of spending weeks in the lab testing different polymer combinations, researchers can now get instant predictions by simply entering a SMILES string. This helps save a lot of time, money, and materials. It doesn't replace lab testing completely, but it definitely makes the whole process faster and more focused.

2. Consistency across Multiple Properties

What stood out was that the model didn't just perform well for one or two properties—it gave stable and reliable predictions across all six. Whether it was a mechanical property like tensile strength or a chemical one like ionization energy, the model handled them all with consistent accuracy. That level of balance is hard to achieve, and it's a big win for practical use.

3. Learning from Chemical Structure

The AI didn't need someone to explain chemistry rules to it—it figured them out on its own by looking at examples. It learned to identify which parts of a molecule affect its behavior just by analyzing the SMILES strings. It's like the model developed its own way of understanding the "language" of molecules.

5.2 Generalizability to New Data

One of the best signs of a good model is how well it performs on completely new data. Ours was tested on polymers it had never seen before, and it still gave solid predictions. That means it isn't

just memorizing—it's actually understanding patterns that apply broadly, which is important for real-world use.s

1. Validation against Real Data

To make sure the model wasn't just making things up, we compared its results with actual lab-tested values. In most cases, the AI predictions matched very closely with the experimental data. This match gave us more confidence that the model's output can be trusted when no real measurements are available.

2. Insight into Molecular Influence

While the model works in the background, we took a closer look at how it made decisions. It turns out that certain parts of the SMILES string had more influence on the prediction. This is useful because it gives researchers hints about which parts of a molecule are more important when designing new materials.

3. Scalable for High-Throughput Screening

One big advantage of using AI is speed. What would normally take months of lab work can now be done in minutes. You can feed the model thousands of SMILES strings and get predictions almost instantly. This makes it perfect for large-scale screening and faster discovery of useful polymers.

4. Accessible Through Web App

Although our main focus was on building the AI model, making it easy to use was just as important. That's why we developed a user-friendly interface where anyone—from a student to an industry expert—can enter a polymer's SMILES and quickly get predictions. No coding or technical setup needed.

5. Performance Metrics

To measure how well the model was doing, we used common performance metrics like MAE, RMSE, and R^2 score. These numbers confirmed what we saw: the model was both accurate and reliable. It didn't just look good on paper—it backed it up with solid results.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Polymers are central to modern material science, playing crucial roles in sectors like packaging, electronics, automotive, aerospace, and medicine. Yet, predicting their physical and chemical properties has traditionally relied on laborious, time-consuming experimental methods. These approaches, while precise, are impractical when dealing with the enormous variety of potential polymer combinations. To address this challenge, our project TransPolymer brings Artificial Intelligence (AI)—specifically transformer-based deep learning models—into the fold to revolutionize polymer property prediction.

The foundation of our approach lies in the use of SMILES (Simplified Molecular Input Line Entry System), a linear textual representation of molecular structures. This makes it possible to apply sequence-based machine learning to chemical data. Unlike traditional RNN or LSTM models that struggle to capture long-range dependencies, transformer models excel due to their use of self-attention mechanisms, enabling them to consider every part of the input simultaneously. This makes them particularly suitable for interpreting SMILES sequences, where distant parts of a polymer's structure can influence its overall properties.

By using a pretrained transformer model, ChemBERTa, fine-tuned on polymer datasets, TransPolymer predicts six essential properties: tensile strength, ionization energy, electron affinity, log P, refractive index, and molecular weight. These predictions are vital for applications in materials science, drug delivery, and environmental engineering, allowing researchers to assess polymer performance without physically synthesizing and testing each candidate. This speeds up the discovery process and reduces associated costs and resource usage.

One of the key strengths of this approach is the use of transfer learning. Since ChemBERTa was pretrained on a large chemical corpus, it already understands chemical syntax and structure to a certain degree. This allows the fine-tuning process to focus on task-specific learning without needing an enormous labeled dataset. As a result, even with limited training data, the model delivers robust performance, making it a practical tool in real-world scenarios.

Importantly, our project does not aim to replace traditional chemistry or lab experiments but to complement them. The AI model serves as a screening tool, helping scientists prioritize which polymers to investigate further. This ensures that time and resources are spent on the most promising candidates. The predictions are designed to assist, not substitute, expert judgment.

Moreover, we made sure to approach this innovation responsibly. Data curation and preprocessing were carefully performed to maintain quality. We also recognize the ethical importance of transparency and explainability in AI systems. While the model offers high accuracy, efforts were made to interpret its decisions and ensure that users can trust the results it provides.

TransPolymer shows how AI can be woven into the fabric of materials science research. By connecting SMILES-based representations with the attention-rich architecture of transformers, we're unlocking new possibilities in polymer property prediction. The intersection of chemistry and AI offers a powerful avenue for accelerating research and fostering innovation.

Looking ahead, we envision expanding this project by incorporating more properties, more diverse polymer types, and larger datasets. With continuous improvements in model interpretability and data availability, TransPolymer could evolve into a comprehensive platform that supports intelligent, AI-assisted materials design. It stands as a forward-looking example of how scientific workflows can evolve with the help of cutting-edge machine learning.

In conclusion, TransPolymer is a meaningful step toward bridging the gap between traditional polymer research and modern AI techniques. It reflects a shift in scientific methodology—moving from manual, trial-and-error approaches to intelligent, prediction-based workflows. Through the integration of transformer models, we have demonstrated how advanced AI can lead to smarter, faster, and more efficient materials discovery.

6.2 Future Scope

The TransPolymer project has laid the foundation for an AI-driven approach to polymer property prediction using transformer models and SMILES representations. While the current version offers accurate predictions for six key polymer properties, there are several opportunities to enhance its scope, performance, and usability in the future:

Expansion of Property Range

Future versions of the model can be extended to predict additional physical, chemical, mechanical, and thermal properties of polymers — such as glass transition temperature, biodegradability, toxicity, thermal conductivity, and more. This will broaden the application of the tool across different industries and research fields.

Support for Polymer Classes and Families

Currently, the model is trained on general SMILES representations. Further work can include classification of polymer types (e.g., thermoplastics, elastomers) and providing insights based on structural families.

Multi-modal Input Support

Besides SMILES, enabling the model to accept other input formats such as InChI, 2D structural images, or polymer descriptors would make the platform more flexible and user-friendly for scientists who are not familiar with SMILES syntax.

Incorporation of Experimental Data

By integrating real-world experimental data, the model's accuracy and robustness can be further improved. Collaborations with material research labs could provide valuable datasets for continual fine-tuning.

Interactive Visualizations

Future updates to the Streamlit app can include dynamic visualizations such as molecular structure renderings, comparison plots, and 3D interactive models, making results more interpretable and engaging.

CHAPTER 7

REFERENCES

1. Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36. Available at: <https://pubs.acs.org/doi/abs/10.1021/ci00057a001>
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30. Available at: <https://arxiv.org/abs/1706.03762>
3. Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. Available at: <https://arxiv.org/abs/2010.09885>
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv preprint arXiv : 1810.04805*. Available at: <https://arxiv.org/abs/1810.04805>
5. Landrum, G. (2006). RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Available at: <https://www.rdkit.org/>
6. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Available at: <https://arxiv.org/abs/1704.01212>
7. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Available at: <https://arxiv.org/abs/1910.03771>
8. Brown, N., McKay, B., Gilardoni, F., & Gasteiger, J. (2004). A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of Chemical Information and Computer Sciences*, 44(3), 1079–1087. Available at: <https://pubs.acs.org/doi/abs/10.1021/ci0341228>

9. Streamlit Inc. (2021). Streamlit Documentation. Available at: <https://docs.streamlit.io>
10. Sebastián Ramírez (2018). FastAPI Documentation. Available at: <https://fastapi.tiangolo.com>
11. MongoDB Inc. (2021). MongoDB Official Documentation. Available at: <https://www.mongodb.com/docs>
12. PyTorch Development Team (2021). PyTorch Documentation. Available at: <https://pytorch.org/docs>
13. Zhang, Z., & Smith, L. (2021). Molecular Data and Predictive Modeling for Materials Science: The Role of AI. *Journal of Materials Science*, 56(12), 1042–1055. Available at: <https://link.springer.com/article/10.1007/s10853-021-05891-6>
14. TransPolymer Project GitHub Repository. Available at: <https://github.com/Transpolymer>

