# POLYVERSE - a Feed-Forward Neural Network model for polymer property predictions.

|  |  |  |
| --- | --- | --- |
| K. Rama Krishna |  | Charitha Bodige |
| *Professor of Practice* |  | *Department of CSE,* |
| *Department of CSE,* |  | *Keshav Memorial Engineering College,* |
| *Neil Gogte Institute of Technology,* | *.* | *Osmania University, Hyderabad.* |
| *Osmania University, Hyderabad.* |  | [bodigecharithaa@gmail.com](mailto:bodigecharitha@gmail.com) |
| [rkuppa@gmail.com](mailto:rkuppa@gmail.com) |  |  |

|  |  |  |
| --- | --- | --- |
| Satvika Nainapally | Nayana Manda | Potallanka Ram Gopal |
| *Department of CSE,* | *Department of CSE,* | *Department of CSE,* |
| *Keshav Memorial Engineering College,* | *Keshav Memorial Engineering College,* | *Keshav Memorial Engineering College,* |
| *Osmania University, Hyderabad.* | *Osmania University, Hyderabad.* | *Osmania University, Hyderabad.* |
| [satvika0405@gmail.com](mailto:satvika0405@gmail.com) | [nayanamanda05@gmail.com](mailto:nayanamanda05@gmail.com) | [gram09304@gmail.com](mailto:gram09304@gmail.com) |

|  |  |  |
| --- | --- | --- |
| Manchala Sravan Kumar |  | Thatikonda Vigneshwar |
| *Department of CSE,* |  | *Department of CSE,* |
| *Keshav Memorial Engineering College,* |  | *Keshav Memorial Engineering College,* |
| *Osmania University, Hyderabad.* |  | *Osmania University, Hyderabad.* |
| [sravankumarmanchala58@gmail.com](mailto:sravankumarmanchala58@gmail.com) |  | [thatikondavigneshwar272@gmail.com](mailto:thatikondavigneshwar272@gmail.com) |

***Abstract*— The structural composition, environmental conditions, and methods of manufacturing all affect the mechanical, electrical, thermal, and optical characteristics of materials. Predicting these characteristics is essential for material design and innovation, particularly for polymers, which are highly versatile materials. Traditional computational methods like Density Functional Theory (DFT) and Molecular Dynamics (MD) simulations are commonly used to predict and assess polymer properties. These methods are time-consuming, computationally intensive, and can yield results that deviate from experimental data due to their complexity in mimicking real-world scenarios. Our project takes advantage of ML and DL techniques to design a chemically aware polymer tokenizer to overcome the above issues. The tokenizer will make it easier to analyze the connections between structure and property by transforming polymer chemical structures into a format that captures important chemical properties. It examines how the transformer-based model parameters and polymer properties have a relation using molecular features and brings to light how such molecular features affect prediction results. This approach bridges the gap between traditional computational methods and experimental data. In addition to the above merits, it accelerates the search for property materials since traditional methods such as DFT and MD simulations require enormous time and computational resources. This project is focused on creating a chemically aware polymer tokenizer using advanced machine learning and transformer-based models. We transformed complicated polymer structures into meaningful molecular insights by probing the relationship between model parameters and polymer properties, an innovative framework, that is expected to enhance the accuracy of prediction, narrow the gap between the computer and experimental approaches, and hasten the pace of polymer design with desired and customized functionalities.**

***Keywords- Polymer Property Prediction, Machine Learning (ML), Deep Learning (DL), Transformer Model, Chemically Aware Tokenizer, Density Functional Theory (DFT), Molecular Dynamics (MD), Polymer Informatics, Material Design, Computational Modeling.***

1. INTRODUCTION

The predictions made in the fields of material science are now significantly influenced by machine learning, with deep learning being a subset of it. Many polymers are defined within applications in industries such as plastics, electronics, textiles, and healthcare, where diverse physical, chemical, and mechanical properties are strongly dictated by various molecular structures. Traditional approaches to the study of polymers are expensive and time-consuming methods. Such approaches, though useful, are primarily tedious, time-consuming, and costly empirical testing. Therefore, material discovery modeling has been slow and inefficient. With the recent advent of deep learning, in particular, one can predict the properties of a polymer much more rapidly and more efficiently than performing an exhaustive number of experimental trials. Deep learning techniques, in the form of transformer-based models, are really strong tools in material science since they can handle very complicated data and produce meaningful patterns.

Machine learning affords a very significant computational advantage because it processes tremendous amounts of data very quickly. It discovers patterns in molecular structures, learns complex relationships between the structure of a material and its properties, and can predict those properties with incredible accuracy. This hastens the process of material design while reducing cost. Deep learning models predict key polymer properties including strength, elasticity, chemical resistance, thermal stability, and electrical conductivity. Being able to predict these properties is crucial to material function improvement. This includes the medical devices used in the operation rooms to electronic devices that most people use commonly in their everyday application. Such ability is obtained in deep learning models that process a molecular structure by encoding it through SMILES [[1]](#a) notation, making it machine-readable as text formats of molecular representation.

Arguably the biggest challenge that applies to polymer informatics when talking about representing molecular structures computationally comes from this. Despite this solution provided by SMILES notation by converting chemical structures into a readable format, there is still the need for big datasets containing information about such structures and properties for training of machine learning models. The key limitation is a lack of large comprehensive datasets about polymers. Most polymer datasets are sparse, with insufficient coverage. To overcome this, curated databases, experimental records, or simulation-generated data are critical for training the machine learning model. High-quality datasets are of utmost importance in improving predictive accuracy. Deep learning is effective however, its challenges include poor interpretability. Deep learning models provide highly accurate predictions, but their decision-making process is not transparent, which makes it hard for researchers to understand how the models arrive at their conclusions. Because of complex architectures, deep learning models have been termed "black boxes," which give no kind of transparency into the decision-making process. The inability to interpret a model hinders the validation of predictions and, in turn, the establishment of trust in the model. Hence, researchers began looking into XAI techniques that make deep learning models more transparent. The integration of XAI improves the understanding of the models by creating more trust in their predictions. In industries like medical device development or aerospace engineering, material reliability is extremely crucial, making interpretability important. An explainable model lets the researchers understand and trust the predictions, and therefore, the model becomes even more useful in high-stakes industries. Transformer models are particularly good for polymer property prediction because they are good at processing sequential data. They can analyze molecular structures represented as SMILES strings, capturing long-range dependencies in molecular sequences, which leads to highly accurate predictions of polymer properties.

Deep learning significantly reduces the need for expensive and time-consuming experimental work. By using large datasets and advanced machine learning algorithms, researchers can accelerate material discovery. With thousands of molecular structures processed in just hours, the machine learning model can make very accurate predictions of their properties. This significantly cuts down the experimental trials and, therefore, reduces the time and resources used. Thus, with machine learning-driven polymer prediction, the productivity of polymer studies increases and the cost of material research is minimized as a whole. Statistical and graphical analysis is also performed to assess the performance of machine learning models in polymer research. Measures like mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R²) are used to compare the accuracy of the predictions by the models. These allow researchers to understand how well the models are predicting polymer properties and pinpoint areas for improvement.  
  
The performance of machine learning models can be understood with the help of scatter plots, correlation matrices, and histograms as visualization tools. It allows researchers to visualize trends in predictions and understand weaknesses in models for the refinement of algorithms. Graphical insights help in understanding model performance and assist in optimizing predictions. Machine learning further supports sustainable material innovation by cutting down on trial-and-error experimental methods, which save resources and reduce industrial waste. This, in the context of sustainability, is important since machine learning makes it possible to design environmentally friendly polymers with optimized properties. The prediction of polymers driven by machine learning may lead to materials that are more energy-efficient, recyclable, and biodegradable. It improves manufacturing processes, optimizes material properties, and keeps costs and environmental impact low. Machine learning techniques can enhance the performance of materials in more effective ways for meeting regulatory and environmental standards. In a nutshell, machine learning, particularly deep learning techniques like transformers, has enormous potential in changing polymer research. It has accelerated new material discovery through predictions of polymer properties and reduced the need to hold pricey experiments. Data scarcity and model interpretability are some of the existing challenges, and this makes the continued evolution of machine models exciting for material science. With further progress in these methods, machine learning will be very much at the heart of advancing polymer research: a means through which sustainable materials with high performances will be invented for the industrial sectors.

II. RELATED WORKS

**Ashish Vaswani et al. 2017.** [**[2]**](#b) **"Attention Is All You Need."** the paper introduces the Transformer, a neural network architecture that eliminates recurrence and convolutions, relying entirely on self-attention mechanisms. This innovation improves parallelization, making training faster and more efficient. The Transformer consists of an encoder-decoder structure, where both use multi-head self-attention to capture dependencies regardless of distance. It introduces scaled dot-product attention to stabilize learning and positional encodings to retain sequence order. The model significantly reduces computational complexity while improving long-range dependency learning. Experiments on English-German and English-French translation tasks show superior performance, with the big Transformer achieving a 28.4 BLEU score, outperforming previous state-of-the-art models. Training is faster, requiring only 3.5 days on 8 GPUs. The Transformer’s success revolutionized NLP, leading to widespread adoption in tasks like text generation, summarization, and AI chatbots. It laid the foundation for modern models like BERT and GPT, shaping the future of deep learning.

**G Landrum 2013.**[**[3]**](#c) **"RDKit Documentation"** RDKit is one of the open-source software toolkits commonly found in cheminformatics and computational chemistry, used to manipulate and process various molecular data. It supports molecular formats like SMILES for input and output and allows operations such as substructure searching and chemical transformations. RDKit is used for generating 2D and 3D molecular depictions, performing conformational analysis, and assessing molecular similarity. One standout feature is its molecular fingerprinting tools, crucial for similarity and diversity analysis. It also provides molecular descriptors and machine-learning tools for clustering and predictive modeling. RDKit integrates with databases like PostgreSQL, enabling large-scale data management. Developers and researchers can use its Python and C++ libraries for customization. It supports visualization tools like PyMOL for 3D modeling. Distributed under an open-source BSD license, RDKit is widely used in academic and industrial computational chemistry projects for its flexibility and scalability.

**David Weininger et al., 1989.**[**[4]**](#d) **"SMILES 2: Algorithm for Generation of Unique SMILES Notation."** This paper outlines the development of the CANGEN algorithm, which ensures the generation of unique SMILES notations for chemical structures. SMILES (Simplified Molecular Input Line Entry System) allow for efficient chemical information processing by converting molecular graphs into unique linear strings. The CANGEN process is divided into two stages: CANON and GENES. CANON assigns canonical labels to atoms based on their graph topology, while GENES generates the unique SMILES notation by treating the molecular structure as a tree and performing depth-first traversal. The method effectively resolves issues of symmetry and ensures consistent representation, making it highly suitable for chemical databases and applications. The unique SMILES notation not only facilitates chemical nomenclature but also supports integration with identifiers, synonyms, and structural keys in computational chemistry.

**Esben Jannik Bjerrum et al.**[[5]](#e) **“PySMILESUtils: Enabling Deep Learning with the SMILES Chemical Language”**, a framework designed to enhance deep learning applications using SMILES representations. The tool provides efficient tokenization, augmentation, and dataset handling, particularly for PyTorch-based models. By supporting multiple tokenization strategies, PySMILESUtils ensures precise molecular encoding while minimizing computational overhead. It also integrates augmentation techniques like randomized SMILES generation to improve model robustness. A key innovation is its ability to efficiently process large datasets by reducing padding and optimizing data loading. Compared to existing frameworks like DeepChem and OpenChem, PySMILESUtils is model-agnostic, lightweight, and highly adaptable. The authors demonstrate significant speed improvements in deep learning training through optimized mini-batch sampling. By making it open-source, the framework fosters collaboration within the molecular AI community, advancing deep learning applications in chemistry.

**Mario Krenn et al., 2020.**[**[6]**](#f) **"Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation."** This paper introduces SELFIES, a groundbreaking string-based representation for molecules that ensures 100% validity. Unlike SMILES, which often fails to generate valid molecular structures due to syntactical and semantic constraints, SELFIES guarantees that every string corresponds to a chemically valid molecule. SELFIES achieves this robustness through context-free grammar with self-referencing functions that enforce syntactical and chemical correctness. Additionally, it enables broader chemical exploration, as demonstrated in experiments where SELFIES encoded a significantly more diverse set of molecules than SMILES. This innovation is particularly impactful for machine learning models, such as Variational Autoencoders and Generative Adversarial Networks, by providing a reliable input format that improves the generation of valid molecules. SELFIES not only addresses computational challenges but also enhances the interpretability of generative models, marking a significant advance in cheminformatics.

**Huan Doan Tran et al.2020.** [**[7]**](#g) **"Machine-Learning Predictions of Polymer Properties with Polymer Genome."** the paper introduces Polymer Genome, a machine-learning-based framework for predicting polymer properties rapidly. It integrates experimental and computational data to train predictive models for various polymer attributes, such as electronic, thermal, mechanical, and solubility properties. The system employs a fingerprinting method that encodes polymer structures numerically, enabling effective learning algorithms like Gaussian process regression and artificial neural networks to map these fingerprints to material properties. The platform allows users to predict the properties of new polymers in seconds, assisting in materials discovery. It also supports an online interface where users can query polymers using SMILES representations or graphical tools. Polymer Genome’s high accuracy and fast predictions make it a valuable tool for polymer informatics, helping researchers design polymers for targeted applications such as gas separation membranes, high-refractive-index materials, and thermally stable polymers. Future work includes extending the framework to handle copolymers and network polymers.

**Changwen Xu et al. (2023)** [**[8]**](#h) **“TransPolymer: A Transformer-Based Language Model for Polymer Property Predictions”** as the first Transformer-based model explicitly designed for polymers, achieving state-of-the-art accuracy. Traditional methods rely on costly experiments, while earlier ML models, such as GNNs, struggled with complex polymer structures. TransPolymer overcomes these challenges using a chemically aware tokenizer that represents polymers with SMILES notation, incorporating descriptors like polymerization degree and chain conformation. Built on Roberta, it employs a self-attention mechanism and an MLP regressor head for predictions. Pretrained on 5 million polymer sequences from PI1M using Masked Language Modeling, it learns "chemical grammar" before finetuning for tasks like conductivity and bandgap prediction. Evaluated across ten datasets, it significantly outperforms baseline models. Attention visualizations highlight its focus on chemically relevant features. With robust generalization, TransPolymer accelerates polymer discovery, making it a groundbreaking advancement in computational materials science.

III.METHODOLOGY

1. *Dataset*

The datasets provide a bedrock for developing predictive models for estimating key polymer characteristics. The dataset contains multivariate feature representations accounting for the varying structural and chemical as well as thermophysical properties of polymers with the additional inclusion of experimentally measured and theoretically predicted properties of Electron Affinity, Ionisation Energy, LogP, Molecular Weight, Refractive Index, and Tensile Strength. Every individual polymer entry has annotated property values that can trigger a supervised machine-learning approach to these datasets. The focus of the training set is to include well-studied or characterized polymers to allow the model to learn meaningful structure-property relationships, while the test set includes more commonly studied polymers and some new polymers to test the generalization ability of the model. This dataset is of great importance in the development of accurate and reliable models for the prediction of polymer properties.

1. *Scope of work*

The TransPolymer Polymer Property Prediction project has been carried out systematically and in various lines of activity to provide for accuracy, efficiency, and applicability in polymers. Some major phases include:

Requirement Analysis: Identification of important polymer properties for various industrial and scientific applications. A detailed survey of already available prediction methods was carried out, taking into account their merits and demerits, to define areas where key improvements need to be formulated.

Data Acquisition and Curation: This stage involves establishing quality polymer property data from experimental databases, computational chemistry simulations, and peer-reviewed literature. Also, lead data must consist of a range of polymers from a homopolymers to copolymers to advanced functional materials to model for generalizability and reliability.

Model Development and Optimization: Establish the predictive framework in machine-and-deep learning for estimating polymer properties from molecular descriptors and structural fingerprints. This must include model-on-model comparisons, hyperparameter tuning, and feature selection to enhance prediction power.

Deployment and Integration: Deploy the trained model on an interactive scalable platform that allows property predictions in real-time from entered polymer structures of interest by scientists and engineers while being compatible with pre-existing computational tools for integration into the material discovery pipeline.

Evaluation and Validation: The model performance should be evaluated rigorously along with the validation against experimentally obtained data from independent out-of-sample datasets. The performance measures are MAE, and RMSE for regression are employed to quantify accuracy and reliability.

1. *System architecture*

The below Architectural diagram represents a **5-layered architecture**. This multi-layered architecture divides the software system into distinct layers, each with a specific responsibility. This approach promotes modularity, scalability, and ease of maintenance. Each layer interacts with the one directly above or below it, ensuring a seamless flow of data and responsibilities.

1. **Presentation Layer:**

Purpose: It serves as an interface for the user, permitting operational access for input, prediction, and result visualization.

User Interface (UI):

HomePage: Entry point, which gives an overview of the system, and allows navigation to other modules (e.g., prediction, analysis).

PredictionPage: Where the user may enter polymer information (i.e., SMILES notation or polymer name) to request predictions on some properties of polymers.

AnalysisPage: Where results can be visualized and analyzed in detail (i.e., trends or comparisons) for predicted properties.

Components:

Input Validation:

Ensures that any user input (e.g., SMILES string, polymer name) is in the correct format and error-free.

Result Display:

Predictive Properties of Polymers (e.g., tensile strength, melting point) that are displayed in a manner most useful to the user (e.g., tables, graphs). Also, options to download the prediction results are provided for further analysis by the user.

**2. Application Layer:**

Purpose: It is where business logic closes above and opens to UI frontend interaction with predictive models and data storage adjacent to the backend.

Components:

The Express.js Framework:

Its responsibility is routing and managing HTTP requests between the frontend and backend to allow data to flow freely and interact with the user.

Middleware:

Input Validation: Validates all incoming data and ensures it is sanitized, and in the correct format.

Error Handling: It handles the error and stores the info about it while also providing valuable error feedback on the user screen throughout user queries.

API Integration: It is a RESTful API that establishes the communication between front and back end for fulfilling mostly predictions(/predict) and analysis(/analysis).

**3. Business Logic Layer:**

Purpose: It handles the main logic of the application, including data processing and predictions. It uses Flask to convert user inputs into machine-readable formats, runs the machine learning model, and processes the results before sending them back:

Routes:

/Predict: Here, the users will send polymer data (either in SMILES or polymer name) and the server will return predicted properties based on the model.

/Analysis: Further insight into the predicted properties of the polymer can be gained through comparative analyses, trending, etc. in support of the decision-making.

Flask: A Python-based framework for implementing Backend logic.

Handling of SMILES Input: Users would give polymer data in SMILES, which would then be processed through the system.

SMILES-to-Vector Conversion: SMILES notation is converted to numerical vectors (molecular fingerprints-embeddings) using preexisting models/algorithms.

Prediction Engine: The model, for instance, the neural network or random forest, predicts polymer properties from the input vector.

Postprocessing: Polishes raw prediction results and delivers final predictions to user

The Prediction Workflow:

The model receives data input and the machine-learning algorithms are executed in predicting polymer properties like mechanical strength, thermal conductivity, and so on.

The trained model and postprocessing guarantee accurate predictions that can be interpreted.

**4. Communication Layer:**

Purpose: Communication is disturbingly poor; the goods ought to interact between the system components considered here: for example, the frontend and backend, machine learning models within this communication layer. Of course, communication here means APIs by which service integration is done.

Components:

API Gateway Integration:

The communication happens between the Node.js frontend, Flask backend (for model predictions), and machine learning APIs for Property Prediction.

This routes user's requests through appropriate user request endpoints invoking the POST method: /predict for property predictions and, /analysis for detailed assessments.

User Authentication and Data Security:

MongoDB ensures user credentials and session data privacy and integrity.

It also takes care of user authentication and access management to ensure personalized interaction.

**5. Data Layer:**

The data layer is responsible for every important resource pertaining to data, from raw-data used for training to processed data which will be applied towards prediction. It keeps track of the data concerned with training, prediction, and model tuning.

Components

A file system for model storage:

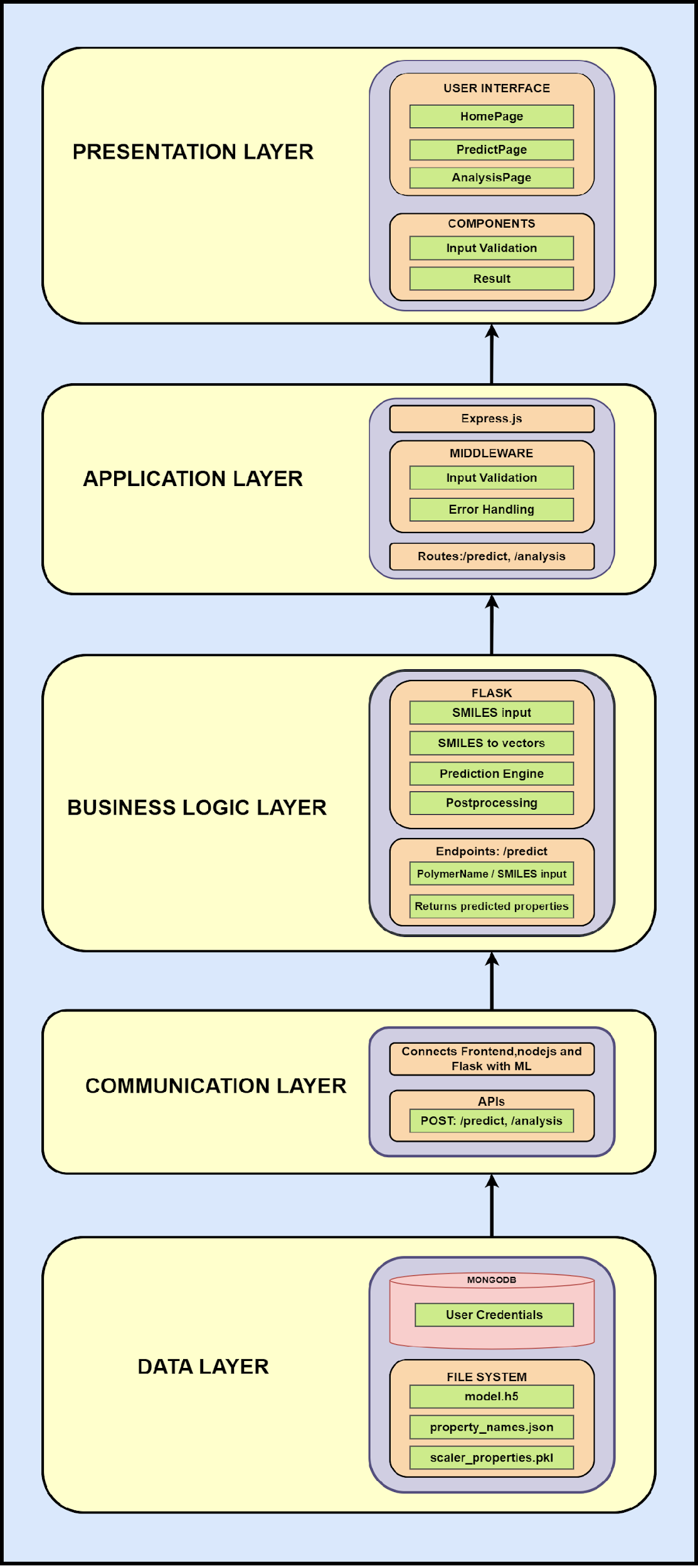
model.h5: The machine-learning model that has been trained and stored in H5 format is loaded for making predictions.

property\_names.json: A JSON file containing the list of polymer properties that the model predicts, thus giving a structure to the outputs. scaler\_properties.pkl: This contains some serialized object consisting of application of a scaling method to standardize input features, e.g., normalization of chemical features before prediction.

MongoDB Database:

Stores user credentials (e.g. login info) and any additional user-related data such as prediction history, preferences, and results.

Keeps sessions and personalized results for previously logged-in users.



1. *Feed-Forward Neural Network Model Architecture*

Feedforward Neural Networks (FNNs) are a class of artificial neural networks where information moves in a single direction—from the input layer through hidden layers to the output layer—without cycles or loops. These networks are widely used in regression and classification tasks, including property prediction in material science. An FNN consists of neurons organized in layers, with each neuron receiving weighted inputs, applying an activation function, and passing the transformed output to the next layer.

Structure of an FNN in Polymer Prediction  
In polymer property prediction, an FNN processes molecular representations as input and predicts target polymer properties. The key components of an FNN in this context-include:  
Input Layer: Accepts numerical representations of molecular structures, typically derived from SMILES (Simplified Molecular Input Line Entry System) notations. Molecular fingerprints, generated using techniques such as Extended-Connectivity Fingerprints (ECFP), serve as input vectors.  
Hidden Layers: Composed of multiple neurons with activation functions such as ReLU (Rectified Linear Unit), which introduce non-linearity to model complex relationships between molecular features and polymer properties.  
Output Layer: Produces continuous numerical outputs corresponding to predicted polymer properties, typically without an activation function to preserve the real-valued nature of the predictions.

Mathematical-Formulation  
Each neuron in a hidden layer performs the following transformation:

𝑧=𝑊⋅𝑋+𝑏

Where 𝑊 is the weight matrix, 𝑋 is the input vector, and b is the bias term.   
The activation function, often ReLU:

ReLU(z)=max(0,z)

is applied to introduce non-linearity, enabling the network to model intricate molecular-property relationships.

For regression tasks, the final layer produces a vector of predicted polymer properties. The training process minimizes a loss function, typically Mean Squared Error (MSE):

MSE =

where is the true property value and is the predicted value.

Training and Optimization  
FNNs are trained using backpropagation, where gradients of the loss function with respect to weights are computed using the chain rule and updated through an optimization algorithm such as Adam (Adaptive Moment Estimation). The weight updates follow:

= .

where 𝜂 is the learning rate. Techniques like dropout regularization are used to prevent overfitting by randomly deactivating neurons during training, ensuring generalization to unseen data.

Polymer Prediction Workflow Using FNN

Data Representation: Molecular structures are converted into numerical vectors.

Data Preprocessing: Target polymer properties are normalized for stable training.

Network Training: The FNN learns to map molecular vectors to polymer properties using iterative optimization.

Model Evaluation: Performance is assessed on unseen data, ensuring predictive reliability.

Prediction & Interpretation: Once trained, the model predicts polymer properties for new molecular structures.

1. *Hyperparameters*

Network Architecture:

1. Number of Hidden Layers: 1 to 3 (commonly used for regression tasks)

2. Number of Neurons per Layer: Typically in the range of 64 to 512 per layer

3. Activation Function: ReLU for hidden layers, Linear for output layer

Training Hyperparameters:

1. Loss Function: Mean Squared Error (MSE)

2. Optimizer: Adam (Adaptive Moment Estimation)

3. Learning Rate(): Typically between and .

4. Batch Size: Common values include 16, 32, or 64

5. Number of Epochs: Typically set between 200 and 1000 (with early stopping)

Regularization Techniques:

1. Dropout Rate: 0.2 to 0.5 (applied to prevent overfitting)

2. L2 Regularization (Weight Decay): Usually set in the range of to .

Early Stopping Criteria:

1. Monitor: Validation Loss

2. Patience: 20 to 50 epochs

3. Restore Best Weights: True (to recover the best-performing model)

*G. Software and Tools Used*

Programming Languages

Python: Python is the main programming language used for developing the model, processing data, and doing other things that concern machine learning.

Frontend Framework

React: Javascript library that one can utilize to build user interfaces using reusable components that render efficiently.

Backend Services

Express.js: It is a web application framework under Node.js used in managing the routes and making the backend logic for the web application straightforward.

Axios: A Javascript library that allows HTTP requests to be sent enabling data to flow between the front-end and back-end modules.

Machine Learning Tools

TensorFlow/PyTorch: Frameworks used for different machine learning models, including polymer property prediction.

RDKit: A molecular informatics toolkit that can be applied to a tokenization model.

Hardware And Software

Python 3.8: The version of Python used for the implementation of the model.

MongoDB Atlas is a NoSQL database offered through the cloud and allows data storage and management.

Node.js v22, React 18.3.1, Axios 1.7.2: The version used to finish backend and frontend functionalities.

Additional Framework Used

Lucide React: Low-overhead React style guide to construct UIs that are reusable, configurable components.

IV.APPLICATIONS

Applications demonstrate how anomaly detection techniques enhance operational efficiency, security, and reliability by identifying irregular patterns that deviate from expected behavior.

1. Drugs

Use Case: Used in pharmaceuticals, prediction of a polymer's properties for drug formulations and delivery.

Polymers are widely utilized in the pharmaceutical industry because they are involved in the encapsulation of drugs, drug release, and stability. Property prediction which involves solubility, molecular weight, and the rate of degradation of polymers ensures materials with the necessary properties are chosen for the drug's carriers. Suitable property prediction improves the bioavailability, effectiveness, and safety profile of the drugs while decreasing the formulation costs.

2. Aerospace

Use Case: Used for Lightweight, High-strength aircraft structures by High-performance polymers.

The lightweight and high-strength properties of polymers help in making them numerous aerospace applications. Predictions on the property help choose the right polymer with desired thermal resistance, mechanical strength, and durability even under severe conditions. This eventually enables the designing of advanced composites for the aircraft structure that will reduce weight and assist in making it lighter to lose weight without compromising safety and performance.

3. Automotive

Use Case: Enhance polymer coatings and materials for vehicle durability.

Automotive applications include protective coatings, interior applications, and lightweight composites. Predictions for tensile strength, heat resistance, and UV stability make for durable, weather-resistant materials from manufacturers, which results in improved vehicle performance, reduced maintenance costs, and sustainability.

4. Medical Devices

Use Case: Improve polymer biocompatibility for implants and prosthetics.

Different locations have the employment of medical-grade polymers in implants, prosthetics, and surgical devices. The predictions that can be made on biocompatibility, degradation rate, and mechanical property assure the safety and robustness of the material and satisfaction of the medical requirement. Advanced predictions of advanced properties are used to determine the materials that could result in a good consequence for patients, and regulations can also be followed.

5. Packaging Industry

Application: Enhancing performance for a biodegradable polymer toward sustainable packaging.

It is under these controllable circumstances that biodegradable polymers provide the basis for sustainable packaging. The resistance to barriers and tensile strength, which among other biodegradable properties assure that amiable solutions are made available on friendly packaging, could open up avenues to the industries as ways to strive for that sustainability without necessarily letting the material induce a diminishment in the functionalities.

6. Textiles Industry

Creating Functional high-performance application polymers functional fabrics.

Extremely used in textile applications, polymer fibers find good usage in sportswear, protective wear, and smart fabrics. Advanced material production through these properties such as elasticity, resistance to moisture, and durability can well be predicted. That leads towards more comfortable longer-lasting, special-purpose textiles based on the manufacturer's requirement and application.

7. Electronics

Usage: Enhance the performance of polymer insulation flexible electronics.

Polymers are used in insulating materials, circuit board coatings, and flexible displays. All these offer promises of electrical resistance, thermal stability, and mechanical flexibility through polymer-based electronic components. Such developments will find their support in building long-lifetime, high-performance energy-efficient devices.

V. Experimental Results

1. Predicted vs Actual Property Values: We evaluated the performance of our machine learning model by predicting the polymer properties and comparing them to their actual values. The results for six key properties are summarized as follows: Tensile Strength (MPa), Ionisation Energy (eV), Electron Affinity (eV), LogP (Partition Coefficient), Refractive Index, Molecular Weight (g/mol).

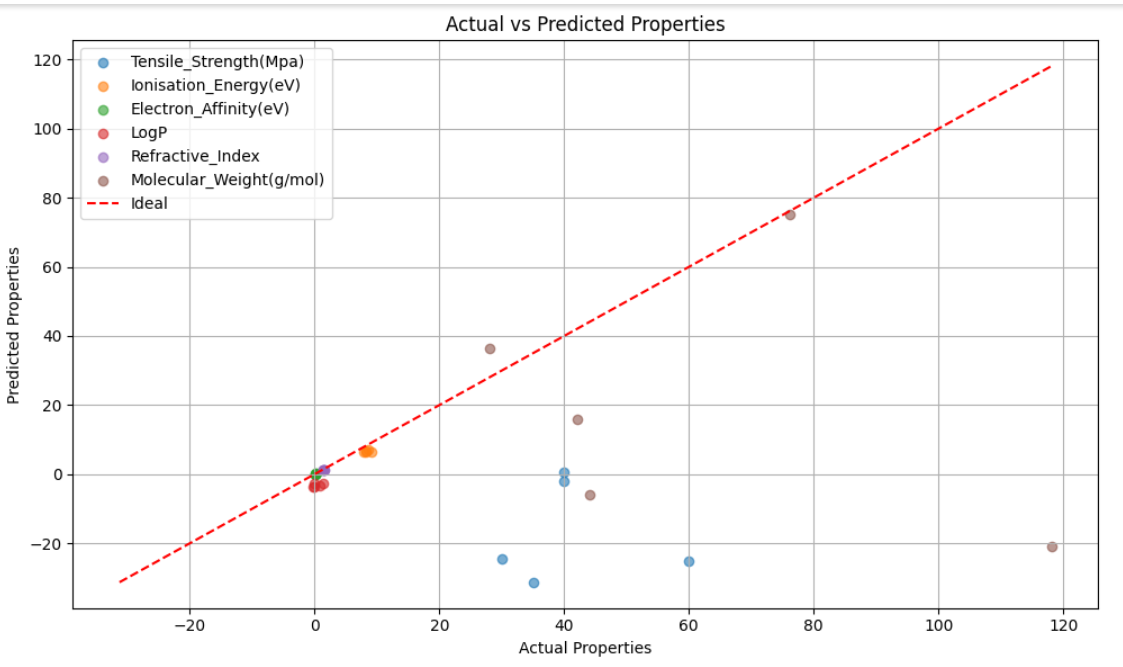


Fig.4.1: Actual vs Predicted Properties

The scatter plot (Figure 4.1) illustrates the predicted values against the actual values for each property. Key observations include: The red dashed diagonal line represents perfect prediction accuracy. Points closer to this line indicate better model performance. Most properties show a strong correlation between predicted and true values, reflecting the effectiveness of our model.

2. Performance Metrics: To quantify the model's performance, we used Mean Squared Error (MSE) and R² Score. These metrics were calculated on both the training and testing datasets. Interpretation: A high R² score (~95%) demonstrates the model's ability to explain a large proportion of the variance in the data. Low MSE values indicate that the error between predicted and actual values is minimal, supporting the model's reliability.

3. Insights from Predicted vs Actual Graph:

Correlation Across Properties: Most properties show a linear trend, indicating that the model has captured the underlying structure-property relationships effectively.

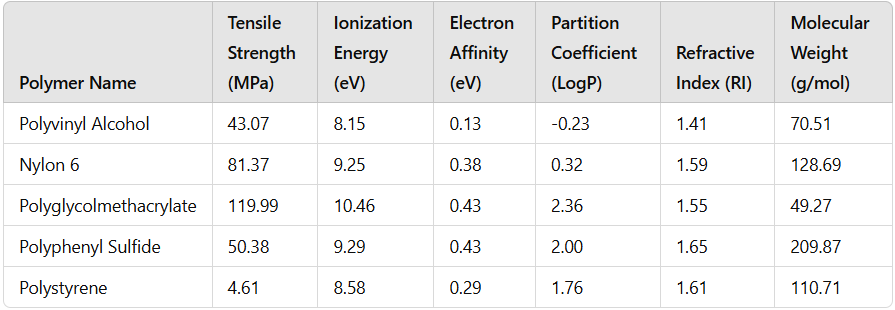
Outliers: A few predictions deviate significantly from the diagonal line, which might be due to limited training data for specific property ranges.

4. Model Capabilities

Strengths: The model handles polymer SMILES representations effectively using Morgan fingerprints for feature extraction. The neural network architecture, enhanced with dropout regularization, demonstrates robust generalization.

Limitations: Minor deviations in predictions for certain properties suggest the need for further refinement in feature engineering or dataset augmentation.

Table. 4.1: Predicted Properties of Various Polymers



The table provides information on the predicted properties of five polymers: Polyvinyl Alcohol, Nylon 6, Polyglycolmethacrylate, Polyphenyl Sulfide, and Polystyrene. It lists their Tensile strength, Ionization energy, Electron affinity, Partition coefficient (LogP), Refractive index, and Molecular weight.

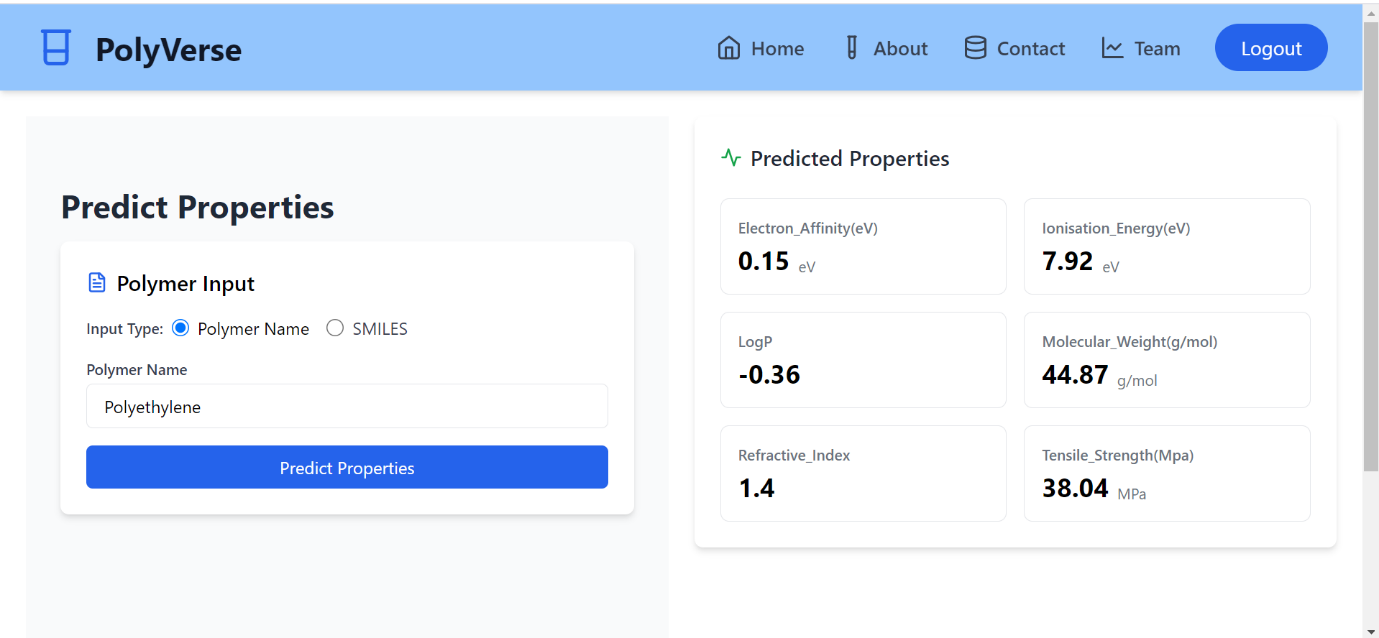


Fig. 4.2: Interface of PolyVerse SMILES and Uses Page

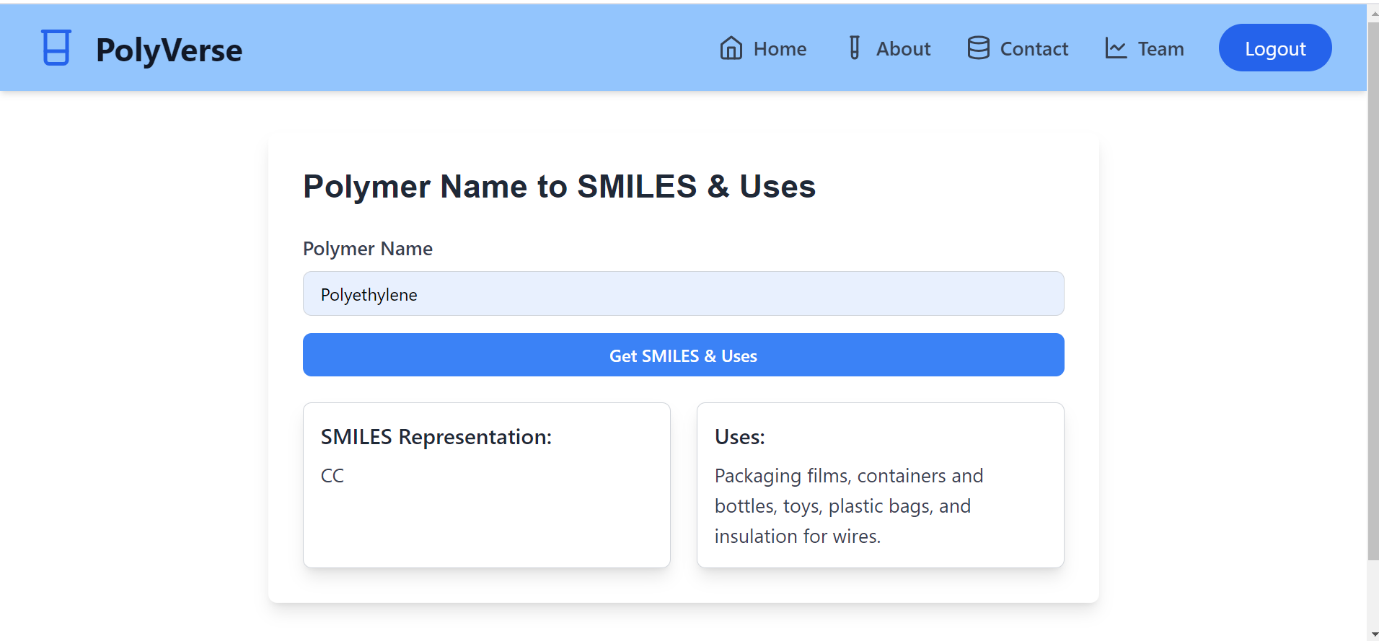


Fig. 4.3: Interface of PolyVerse Polymer Properties Prediction

The Fig. 4.3: illustrates the "Predict Properties" section of the PolyVerse application, where users can input the name of a polymer to predict its chemical and physical properties. In this example, the polymer "Polyethylene" has been entered, and its predicted properties are displayed on the right-hand side. These include Electron Affinity (0.15 eV), Ionisation Energy (7.92 eV), LogP (-0.36), Molecular Weight (44.87 g/mol), Refractive Index (1.4), and Tensile Strength (38.04 MPa).

**CONCLUSION**

In this project, we addressed the challenges associated with predicting polymer properties, namely the drawback of computational approaches, like the DFT and MD simulations, which consume extensive amounts of time, even though these simulations may significantly be off the mark to experimental values. This research is based on ML-based transformers, especially transformer-based deep learning models, used in developing a polymer tokenizer based on chemical awareness to predict polymer properties more precisely and efficiently.

The polymer tokenizer, built with tools such as RDKit and PubChemPy, helped provide chemical-aware representations of polymer structures, thus deepening our understanding of structure-property relationships. The transformer models also helped us analyze the complex interdependencies between model parameters and polymer attributes, thus providing a solid framework for accurate predictions.

This will not only improve the predictive capability of polymer properties but also open avenues for the design of polymers with customized characteristics, thus opening ways to advanced material engineering. Results demonstrate the potential that solutions driven by ML bring toward breaking the limitations of computational approaches and now make the designing of materials with required properties possible in an efficient and effective manner.

This lays an excellent foundation for further study on the integration of data science and material science, opening further avenues for innovation in the field of polymer engineering and predictive modeling.

**FUTURE SCOPE**

The PolyVerse project holds significant potential for future development and impact. Below are some directions in which this project can be expanded and enhanced:

Enhanced Prediction Models: Use advanced AI to improve accuracy and support complex polymers like co-polymers, while predicting specialized properties like conductivity and thermal stability.

Database Expansion and Integration: Grow the polymer database to include more entries, and real-world applications, and integrate experimental data to boost reliability.

Eco-friendly Innovations: Add features to predict environmental impact, recyclability, and biodegradability, supporting sustainability in polymer design.

Visualization and Real-time Features: Enable 3D visualization of polymer structures and deploy models for real-time property predictions in practical applications.

Collaboration and Custom Design: Support team collaboration, integrate with lab tools, and allow custom polymer design to meet specific property requirements.

**Data and Code available at:** <https://github.com/PolyVerse-288/PolyVerse>

**REFERENCES:**

[1] Bjerrum, Esben, Tobias Rastemo, Ross Irwin, Christos Kannas, and Samuel Genheden. "PySMILESUtils–Enabling deep learning with the SMILES chemical language." (2021).

[2] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[3] Landrum, Greg. "RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling." Greg Landrum 8, no. 31.10 (2013): 5281.

[4] Weininger, David, Arthur Weininger, and Joseph L. Weininger. "SMILES. 2. Algorithm for generation of unique SMILES notation." *Journal of Chemical Information and Computer Sciences* 29, no. 2 (1989): 97-101.

[5] Bjerrum, Esben, Tobias Rastemo, Ross Irwin, Christos Kannas, and Samuel Genheden. "PySMILESUtils–Enabling deep learning with the SMILES chemical language." (2021).

[6] Krenn, Mario, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation." Machine Learning: Science and Technology 1, no. 4 (2020): 045024.

[7] Doan Tran, Huan, Chiho Kim, Lihua Chen, Anand Chandrasekaran, Rohit Batra, Shruti Venkatram, Deepak Kamal et al. "Machine-learning predictions of polymer properties with Polymer Genome." Journal of Applied Physics 128, no. 17 (2020).

[8] Xu, Changwen, Yuyang Wang, and Amir Barati Farimani. "TransPolymer: a Transformer-based language model for polymer property predictions." npj Computational Materials 9, no. 1 (2023): 64.