

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi - 590 018, Karnataka



Capstone Project Phase – 1
(Course code: 24AM6PWPW1)

PROJECT WORK REPORT

Molecular Property Prediction Using Transformer-Based MoLFormer Model on SMILES Representations

In the Department of

Machine Learning

(UG Program: B.E. in Artificial Intelligence and Machine Learning)

Submitted by

Pranav	1BM22AI090
Srujan B J	1BM22AI134
Sriharsha K	1BM22AI131
Srinidhi S Mattur	1BM22AI132

6th Semester

Under the Guidance of

Dr. Seemanthini K

Associate Professor

Dept. of MEL, BMSCE, Bengaluru – 19



DEPARTMENT OF MACHINE LEARNING

B.M.S. COLLEGE OF ENGINEERING

(An Autonomous Institute, Affiliated to VTU)

P.O. Box No. 1908, Bull Temple Road, Bengaluru - 560 019

DECLARATION

We, Pranav, Srujan B J, Srinidhi S Mattur and Sri Harsha K, hereby declare that the project work titled "Molecular Property Prediction Using Transformer-Based MolFormer Model on SMILES Representations" is an original work carried out by us under the guidance of Dr. Seemanthini K, Associate Professor, at B.M.S. COLLEGE OF ENGINEERING . The work presented in this report has not been submitted elsewhere for any other degree, diploma, or certification.

We declare that all the references and sources used in the report have been properly acknowledged and cited. No part of this report has been copied, plagiarized, or reproduced from another work without proper attribution.

Signature	Pranav	(1BM22AI090)
Signature	Srujan B J	(1BM22AI134)
Signature	Srinidhi S Mattur	(1BM22AI132)
Signature	Sri Harsha K	(1BM22AI131)

ACKNOWLEDGEMENT

We would like to express our deepest gratitude and sincere thanks to our guide **Dr. Seemanthini K**, Associate Professor, Department of Machine Learning, Bengaluru, for her keen interest and encouragement in guiding us and bringing the work to reality.

We are grateful to **Dr. M. Dakshayani**, Professor and Head, Department of Machine Learning, Bengaluru, for her valuable suggestions and advice throughout our work period.

We express our profound gratitude to **Hon'ble Management and Dr. Bheemsha Arya**, Principal, BMSCE, Bengaluru for providing the necessary facilities and an ambient environment to work.

We would like to thank all the Staff, Department of Machine Learning for their support and encouragement during the course of this project.

Last but not the least, we would like to thank our parents, family and friends, without whose help and encouragement this work would have been impossible.

Pranav	(1BM22AI090)
Srujan B J	(1BM22AI134)
Srinidhi S Mattur	(1BM22AI132)
Sriharsha K	(1BM22AI131)

Abstract

The cheminformatics discipline is an interdisciplinary one that is located at the intersection of information technology, computer science, and chemistry. Its focus is on applying computational methods to store, analyze, and visualize chemical information in an effort to solve complex chemical questions. The digital encoding of molecular structure in a format that is both machine-readable and human-readable is one of the basic cheminformatics tasks. The Simplified Molecular Input Line Entry System, or SMILES, is a notation used quite often for this. SMILES represents the structure and bonding relationships of molecules in a compact text format by encoding molecular graphs as linear ASCII strings.

Being linear and symbolic in nature, SMILES earns the distinction of being in particular apt for several natural language processing methods. SMILES strings carry heavy structural and chemical information analogously to how words and sentences in a human language are endowed with syntactic and semantic meanings. It opens the gates of study of chemical structures for powerful NLP models like the transformer-based architectures. The power of transformers in natural language processing has been boosted by their capacity to represent contextual linkages and long-range dependencies within sequences. These attention mechanisms enable the model to attend to diverse parts of a sequence, making transformers particularly well suited to capture complex interactions between atoms and bonds in molecular representations.

Transformer models can be applied to SMILES strings, thus translating molecular property prediction into a sequence modeling problem. The model learns to interpret molecular sequences, where these sequences possess syntax and semantics similar to those of a natural language, to forecast various properties such as solubility, toxicity, bioactivity, and reactivity. This data-driven approach not only hastens the molecular analysis process but also compensates for any patterns and insights that the rule-based approaches may fail to account for.

However, more algorithmic than innovative approaches are required for the industrial realization of deep learning models in practical cheminformatics. A solid MLOps infrastructure—well defined as the complete set of procedures and resources needed to oversee

and maintain throughout the entire lifecycle of machine learning systems-would be required. MLOps ensures that models are repeatable, dependable, and maintainable during time and in various contexts. Scalability frameworks, build/CI and release/CD, model-monitoring, version control, automated data pipelines, among others, are some of the fundamentals. Without these in place, it becomes next to impossible to guarantee the resilience and consistency of models when they are upgraded or moved into production settings.

Briefly put, application of transformer models to cheminformatics pipelines, and specifically from the SMILES-based sequence modeling point of view, is a compelling illustration of how chemical science and machine learning can be combined. It will have to be complemented by strict MLOps procedures that allow for model development as well as long-term sustainability, automation, and deployability scalability in order to realize its full potential.

TABLE OF CONTENTS

Declaration	i
Acknowledgement	ii
Abstract	iii
Table of Contents	iv
List of Figures	v
List of Tables	v
1 INTRODUCTION	1
1.1 Overview	1
1.2 Organization of the Report	3
2 PROBLEM STATEMENT	6
2.1 Problem Statement	6
2.2 Motivation	6
2.3 Scope and Objectives	6
2.3.1 Project Scope	
2.3.2 Project Objectives	
3 LITERATURE SURVEY	8
3.1 Research Gap	10
4 SYSTEM REQUIREMENT SPECIFICATION	12
4.1 Functional Requirements	12
4.2 Non-functional Requirements	12
4.3 Hardware Requirements	12
4.4 Software Requirements	12
5 SYSTEM DESIGN	13
5.1 System Architecture	13
5.2 Proposed Methodology	13
5.3 High-Level Design	14

5.4	Detailed Design	15
5.4.1	Class Diagram	
5.4.2	Activity Diagram	
5.4.3	Use Case Diagram	

REFERENCES	17
-------------------	----

APPENDIX A: Similarity Plagiarism Report

APPENDIX B: AI Plagiarism Report

LIST OF FIGURES

Figure No.	Description	Page No.
1.1	Molecular Screening Platform	2
5.1	Proposed System Architecture Diagram	13
5.2	MoLFormer Class Diagram	14
5.2	Activity Flow Diagram of MoLFormer	15
5.3	Use Case Diagram of MoLFormer	16

1. INTRODUCTION

1.1 Overview

Cheminformatics addresses chemical problems through the use of computational capabilities. Chemical structures can be represented as ASCII strings in the form of SMILES notation, where NLP methods can be utilized. Contextual and sequential data are well-suited to be modeled using Transformers, which are extensively used in NLP. When used in conjunction with SMILES, they convert the challenge of molecule property prediction into a sequence modeling challenge, which is easier to address. A solid MLOps foundation is needed for reproducibility, automation, and maintainability in deploying such deep learning models.

The contemporary pharmaceutical landscape confronts an unprecedented convergence of scientific opportunity and economic constraint. While revolutionary breakthroughs in computational biology and artificial intelligence promise transformative advances, the industry simultaneously grapples with escalating development costs that have reached astronomical proportions—often exceeding \$2.6 billion per approved therapeutic compound. This economic reality, coupled with historically low success rates in clinical translation, has catalyzed a fundamental reimagining of drug discovery paradigms.

At the nexus of this transformation lies computational molecular analysis, representing a paradigmatic shift from laboratory-centric empirical approaches toward predictive, algorithm-driven pharmaceutical research. The capacity to computationally forecast molecular behaviors—encompassing pharmacokinetic profiles, toxicological signatures, and therapeutic efficacy—before synthesizing physical compounds represents perhaps the most significant methodological advancement in pharmaceutical science since the advent of high-throughput screening technologies.

Traditional computational approaches to molecular characterization have evolved through distinct technological epochs. Initial methodologies employed rigid mathematical descriptors and binary fingerprint representations, which, while computationally tractable, proved inadequate for capturing the nuanced chemical relationships that govern molecular function. Subsequent developments introduced graph-theoretic representations and recurrent neural architectures, offering enhanced expressiveness but remaining constrained by their inability to model long-range molecular dependencies effectively.

The emergence of attention-based transformer architectures has ushered in a new era of

molecular representation learning. These models, originally developed for natural language processing, demonstrate remarkable capabilities in learning complex sequential patterns and hierarchical relationships. When applied to SMILES (Simplified Molecular Input Line Entry System) representations, transformers can theoretically capture both local atomic interactions and global molecular properties through their self-attention mechanisms.

However, the transition from experimental research prototypes to production-grade pharmaceutical tools requires more than algorithmic sophistication alone. Modern pharmaceutical applications demand comprehensive operational frameworks that ensure reproducibility, regulatory compliance, scalability, and continuous quality assurance. Machine Learning Operations (MLOps) methodologies provide this essential infrastructure, bridging the gap between research innovation and industrial implementation.

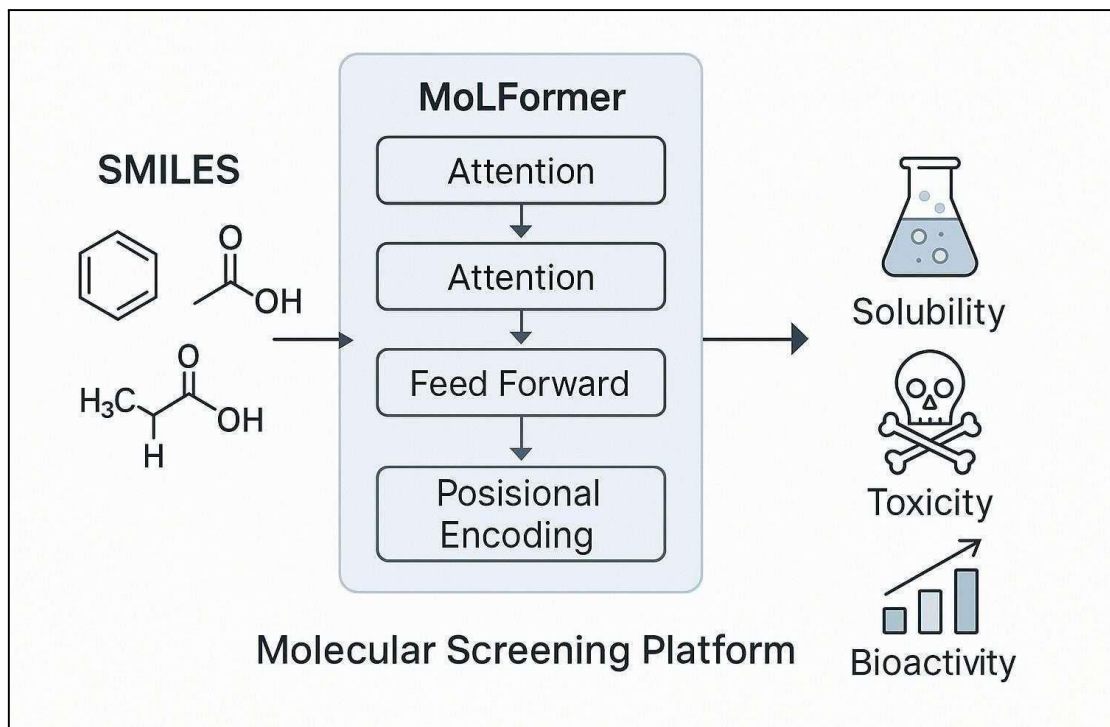
The integration of transformer-based molecular modeling with enterprise-grade MLOps practices represents an unexplored frontier in computational pharmaceutical science. This convergence addresses not merely the technical challenge of accurate property prediction, but the equally critical operational requirements for deploying sophisticated AI systems within highly regulated pharmaceutical environments.

This investigation introduces a novel computational framework that synergistically combines advanced transformer architectures with comprehensive operational methodologies specifically tailored for pharmaceutical applications. Unlike existing approaches that treat molecular property prediction as an isolated machine learning problem, our methodology encompasses the entire lifecycle of model development, deployment, monitoring, and maintenance within regulatory-compliant environments.

The research addresses several critical limitations in current molecular property prediction systems: the computational inefficiency of existing transformer implementations when applied to chemical data, the absence of standardized operational practices for scientific machine learning applications, and the underutilization of shared molecular representations across multiple property prediction tasks.

Our approach introduces three key innovations: chemically-optimized attention mechanisms that reduce computational complexity while preserving molecular understanding, a comprehensive MLOps framework designed specifically for scientific computing applications, and a unified multi-property prediction architecture that leverages shared molecular representations to improve both efficiency and accuracy.

The significance of this research extends beyond technical contributions to encompass broader impacts on pharmaceutical research methodology. By demonstrating how cutting-edge AI techniques can be reliably deployed in production pharmaceutical environments, this work establishes a blueprint for translating academic research into practical tools that can accelerate drug discovery while maintaining the rigorous standards required for therapeutic development.



1.1 Molecular Screening Platform (Molformer Screening)

1.2 Organization of the Report

Organization of the Report Problem definition, effort so far, system requirements, and proposed technique are all covered in this paper. Problem statement, objectives, system structure, tools, and technologies are discussed in detail in the subsequent sections, leading up to the proposed design and methodology. This report is systematically organized to provide a comprehensive examination of the transformer-based molecular property prediction system development, from problem identification through system design and implementation.

Section 1 - Introduction: Establishes the context and importance of molecular property prediction in pharmaceutical research, introduces the key concepts of transformer architectures and MLOps practices, and provides an overview of the research scope and objectives.

Section 2 - Problem Statement: Articulates the specific technical and operational challenges addressed by this research, including the limitations of current molecular property prediction approaches, the motivation for developing transformer-based solutions, and the detailed objectives and scope of the proposed system.

Section 3 - Literature Survey: Provides a comprehensive review of existing molecular property prediction systems, analyzing current state-of-the-art approaches including traditional machine learning methods, deep learning frameworks, and early transformer applications in chemistry. This section identifies critical research gaps that justify the proposed approach.

Section 4 - System Requirement Specification: Details the comprehensive requirements analysis for the proposed system, including functional requirements for molecular property prediction capabilities, non-functional requirements for performance and reliability, and the hardware and software infrastructure specifications necessary for system implementation.

Section 5 - System Design: Presents the detailed system architecture and design methodology, including the overall system architecture, the proposed MoLFormer-based approach with MLOps integration, and comprehensive design specifications including class diagrams, activity diagrams, and use case diagrams that illustrate the system's structure and behavior.

References: Provides a comprehensive bibliography of academic papers, technical reports, and industry resources that inform and support the research presented in this report.

Appendix A - Snapshots: Contains visual documentation of the system implementation, including user interface screenshots, system architecture diagrams, and performance monitoring dashboards that demonstrate the practical implementation of the proposed solution.

Appendix B - Similarity and AI Plagiarism Reports: Includes comprehensive plagiarism analysis reports ensuring the originality of the research and adherence to academic integrity standards, covering both similarity checks against existing literature and AI-generated content detection.

This organizational structure ensures a logical progression from problem identification through literature analysis, requirements specification, and detailed system design, providing readers with a complete understanding of the research methodology and implementation approach.

2. PROBLEM STATEMENT

2.1 Problem Statement

The primary challenge lies in **“To develop a machine learning model capable of predicting molecular properties directly from SMILES strings using transformer-based architectures, specifically MoLFormer, with the goal of improving predictive accuracy and generalization across various chemical datasets”**.

2.2 Motivation

Traditional methods of property prediction of molecules are human feature engineering and do not scale. Scalable and interpretable and accurate models are needed as AI in drug discovery and material science gain prominence. Modern drug discovery requires interpretable models that can:

- Explain why certain molecular structures exhibit specific properties
- Guide medicinal chemists in rational drug design
- Provide confidence estimates for predictions
- Identify key structural features for optimization

2.3 Objectives and Scope

Scope

The research scope encompasses the development of a comprehensive MoLFormer-based prediction system capable of forecasting multiple molecular properties simultaneously, integrated with a complete MLOps pipeline featuring automated CI/CD workflows and seamless connectivity to major molecular databases including ChEMBL, PubChem, and ZINC. Our investigation includes both batch processing capabilities for large-scale virtual screening and real-time prediction interfaces for interactive research applications, complemented by advanced model interpretability features that provide scientifically meaningful explanations for predictions and comprehensive benchmarking against current state-of-the-art methodologies. However, this research explicitly excludes laboratory validation of computational predictions, retrosynthetic analysis and synthesis pathway generation, and three-dimensional molecular structure prediction or quantum mechanical property calculations, maintaining focus on the computational prediction framework rather than experimental validation or comprehensive molecular modeling beyond property prediction tasks.

Objectives:

1. Develop Interpretable Molformer: Create a transformer-based architecture with built-in interpretability features
2. Implement Attribution Methods: Design molecular substructure attribution techniques
3. Create Visualization Tools: Develop novel chemical interpretability visualization methods
4. Build Interactive Platform: Construct user-friendly dashboard for chemists
5. Tune MoLFormer to predict molecular properties on benchmarks such as ESOL, FreeSolv, and QM9.
6. Develop a scalable pipeline for SMILES string preparation and inference. Facilitate reproducible model life cycle management with GitHub Actions, Docker, and MLflow.

3. LITERATURE SURVEY

The application of transformer architectures to molecular property prediction began with the adaptation of BERT-like models to chemical data. Chithrananda et al. (2020) introduced ChemBERTa, demonstrating that pre-trained transformers could effectively learn molecular representations from SMILES strings. This work established the foundation for treating molecular structures as sequential data amenable to natural language processing techniques.

Here are some of the Reviews made:

[1] Weininger, 1988

- Introduced **SMILES (Simplified Molecular Input Line Entry System)**.
- Defined encoding rules for chemical structures as text.
- Foundation for many cheminformatics applications.

[2] O'Boyle et al., 2011

- Presented **Open Babel**, an open-source cheminformatics toolkit.
- Supports chemical format conversion, molecular mechanics, and data analysis.
- Widely used for data preprocessing and file format translation.

[3] Rong et al., 2022

- Proposed **MoLFormer**, a transformer model for molecular representation.
- Uses **self-supervised learning** on molecular data.
- Targets tasks like molecular property prediction and drug discovery.

[4] Vaswani et al., 2017

- Introduced the **Transformer architecture** with self-attention mechanisms.
- Foundation for models like BERT, GPT, and molecular transformers.
- Revolutionized NLP and influenced molecular machine learning.

[5] Honda et al., 2019

- Proposed the **SMILES Transformer**.
- Pretrained on SMILES strings for **low-data drug discovery**.
- Learns useful representations with limited labeled data.

[6] Grand et al., 2020

- Introduced **ChemBERTa**, a BERT-based model for molecules.
- Uses large-scale **self-supervised pretraining**.
- Designed for molecular property prediction.

[7] **Wu et al., 2018**

- Created **MoleculeNet**, a benchmark dataset suite.
- Provides standardized datasets and evaluation metrics for molecular ML.
- Facilitates fair model comparison.

[8] **Gilmer et al., 2017**

- Proposed **Neural Message Passing Networks (MPNNs)**.
- Applied to quantum chemistry and molecular property prediction.
- Important foundation for graph-based molecular models.

[9] **Yang et al., 2019**

- Analyzed how **molecular representations** influence prediction accuracy.
- Compared handcrafted and learned features.
- Emphasized the importance of representation learning.

[10] **Feinberg et al., 2018**

- Developed **PotentialNet**, a GNN for molecular property prediction.
- Incorporates chemical knowledge into graph neural networks.
- Achieves high accuracy in several prediction tasks.

[11] **Brown, 2020**

- Reviewed techniques in **in silico drug design**.
- Covered methods like docking, QSAR, and virtual screening.
- Discussed emerging trends and challenges.

[12] **Li et al., 2021**

- Surveyed **graph neural networks (GNNs)** for molecular property prediction.
- Categorized GNN architectures and applications.
- Offered insights into strengths and limitations.

[13] Lin et al., 2020

- Explored **self-supervised learning with knowledge distillation** on molecular graphs.
- Combines teacher-student training with graph-level tasks.
- Improves generalization and efficiency.

[14] Irwin & Shoichet, 2005

- Introduced **ZINC**, a free database of commercially available compounds.
- Designed for **virtual screening** and drug discovery.
- Offers ready-to-dock 3D molecular structures.

3.1 Existing Systems

Current molecular property prediction systems can be categorized into several approaches:

I. Traditional QSAR Methods

Quantitative Structure-Activity Relationship (QSAR) models represent the classical approach to molecular property prediction. These systems rely on pre-computed molecular descriptors such as:

- Physicochemical Descriptors: Molecular weight, logP, polar surface area
- Topological Descriptors: Connectivity indices, shape descriptors
- Electronic Descriptors: HOMO-LUMO gaps, electrostatic potentials
- Geometric Descriptors: 3D molecular conformations and spatial arrangements

Popular implementations include RDKit-based descriptor calculation pipelines and scikit-learn integration for classical machine learning algorithms (Random Forest, Support Vector Machines, Gradient Boosting).

Limitations: Fixed feature representations, inability to capture complex molecular interactions, limited transferability across different chemical spaces.

II. Graph Neural Network Approaches

Recent systems leverage Graph Neural Networks (GNNs) to process molecular graphs directly:

- Message Passing Neural Networks (MPNNs): Process molecular graphs by propagating information between atoms
- Graph Attention Networks (GATs): Incorporate attention mechanisms for selective information aggregation
- Graph Transformers: Combine graph structure with transformer attention

Notable implementations include DGL-LifeSci, PyTorch Geometric, and specialized frameworks like ChemProp.

Limitations: Computational complexity for large molecules, difficulty handling diverse molecular representations, limited scalability for high-throughput screening.

III. Sequence-Based Deep Learning

Existing sequence-based approaches treat SMILES as natural language:

- Recurrent Neural Networks: LSTM and GRU architectures for sequential processing
- Convolutional Neural Networks: 1D convolutions over SMILES strings

- Early Transformer Adaptations: BERT-like models adapted for chemical sequences

Limitations: Limited domain-specific architectural optimizations, insufficient pre-training on chemical data, lack of production-ready deployment infrastructure.

IV. Commercial Platforms

Several commercial platforms provide molecular property prediction:

- Schrödinger Suite: Comprehensive molecular modeling with proprietary algorithms
- ChemAxon: Chemical structure processing and property prediction tools
- OpenEye Scientific: Specialized molecular design and property prediction

Limitations: Closed-source implementations, high licensing costs, limited customization capabilities, vendor lock-in concerns.

3.2 Research Gap

Analysis of existing systems reveals several critical gaps that this research aims to address:

I. Architectural Optimization Gap

Current transformer-based molecular models largely adapt general-purpose architectures without sufficient chemical domain optimization. Existing approaches fail to fully leverage the unique characteristics of molecular sequences, including:

- Chemical Grammar Constraints: SMILES strings follow specific syntax rules that could inform architectural design
- Hierarchical Molecular Structure: Molecules exhibit multi-scale organization (atoms, functional groups, scaffolds) not captured by linear attention
- Symmetry and Invariance Properties: Molecular properties should remain invariant under certain transformations

II. MLOps Integration Gap

The molecular modeling community has been slow to adopt modern MLOps practices, resulting in:

- Reproducibility Issues: Inconsistent experimental setups and version control practices
- Deployment Challenges: Research models rarely transition to production environments
- Collaboration Barriers: Lack of standardized workflows for team-based molecular modeling projects
- Monitoring Deficiencies: Absence of model performance monitoring in production settings

Multi-Task Learning Optimization

While multi-task learning has shown promise, existing implementations suffer from:

- Task Balancing Issues: Difficulty in optimally weighting different prediction tasks
- Negative Transfer: Some task combinations result in degraded performance
- Dynamic Task Adaptation: Inability to adapt task priorities based on data availability or business requirements

IV. Interpretability and Explainability

Current molecular prediction systems provide limited interpretability:

- Attention Visualization: Basic attention heatmaps without chemical context
- Feature Attribution: Lack of molecular substructure importance analysis
- Regulatory Compliance: Insufficient explainability for pharmaceutical regulatory requirements

V. Scalability and Efficiency

Existing systems face scalability challenges:

- Memory Constraints: Transformer models require significant memory for long molecular sequences
- Inference Speed: Slow prediction times unsuitable for high-throughput virtual screening
- Resource Utilization: Inefficient use of computational resources in distributed settings

4. SYSTEM REQUIREMENT SPECIFICATION

4.1. Functional Requirements

- Accurately predict more than 85% of various molecular characteristics.
- Both batch (>1M molecules/hour) and real-time (<2s latency) predictions.
- Machine learning model training, testing, and deployment.
- Link to ZINC, PubChem, and ChEMBL databases.

4.2. Non-Functional Requirements

- Scalable to 100 or more nodes.
- High dependability with disaster recovery (99.5% uptime).
- Security features comprise GDPR/FDA compliance, role-based access, and encryption.

4.3. Hardware Requirements

- Hardware Needed some examples of multi-GPU computation for training.
- Nodes with high memory for extensive inference.

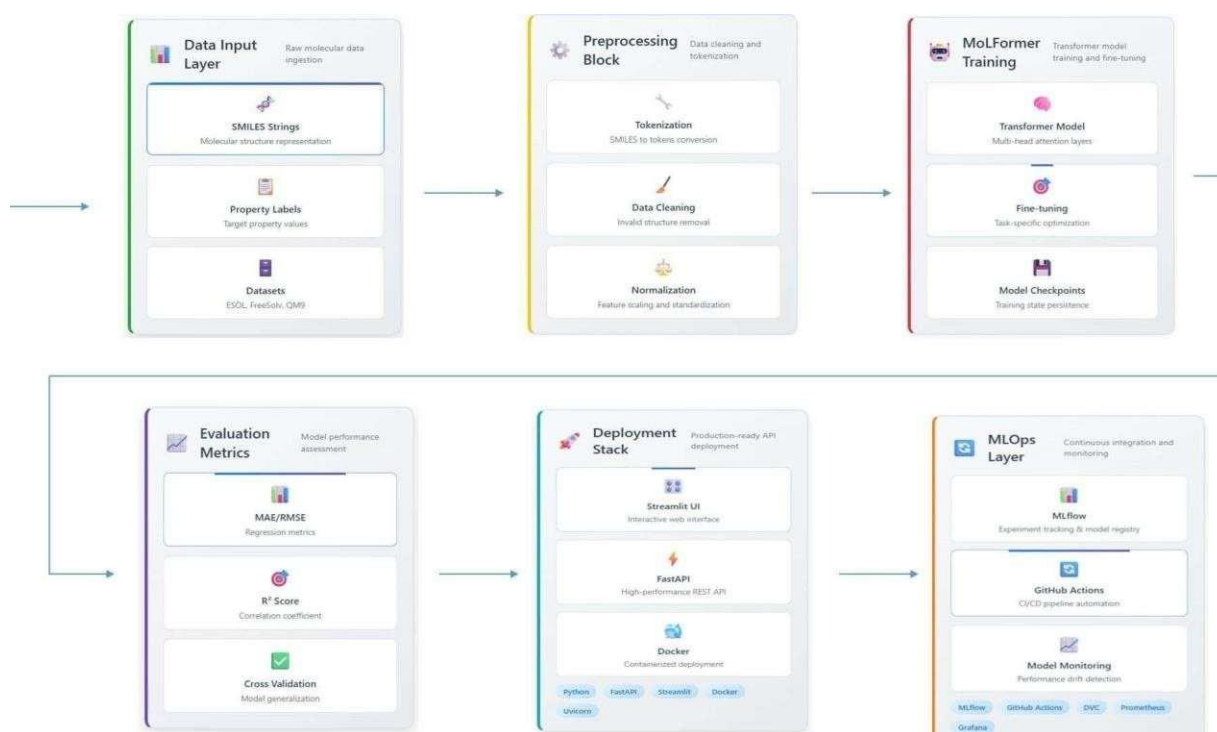
4.4. Software Requirements

- Some basic Software Requirements are Docker, GitHub Actions, RDKit, MLflow, Python, PyTorch, and Streamlit/FastAPI.

5. SYSTEM DESIGN

5.1. System Architecture

Preprocessing of the SMILES strings, training of the MoLFormer, and real-time deployment are all managed by the system. It contains MLOps for monitoring and automation.



5.1 Proposed System Architecture Diagram

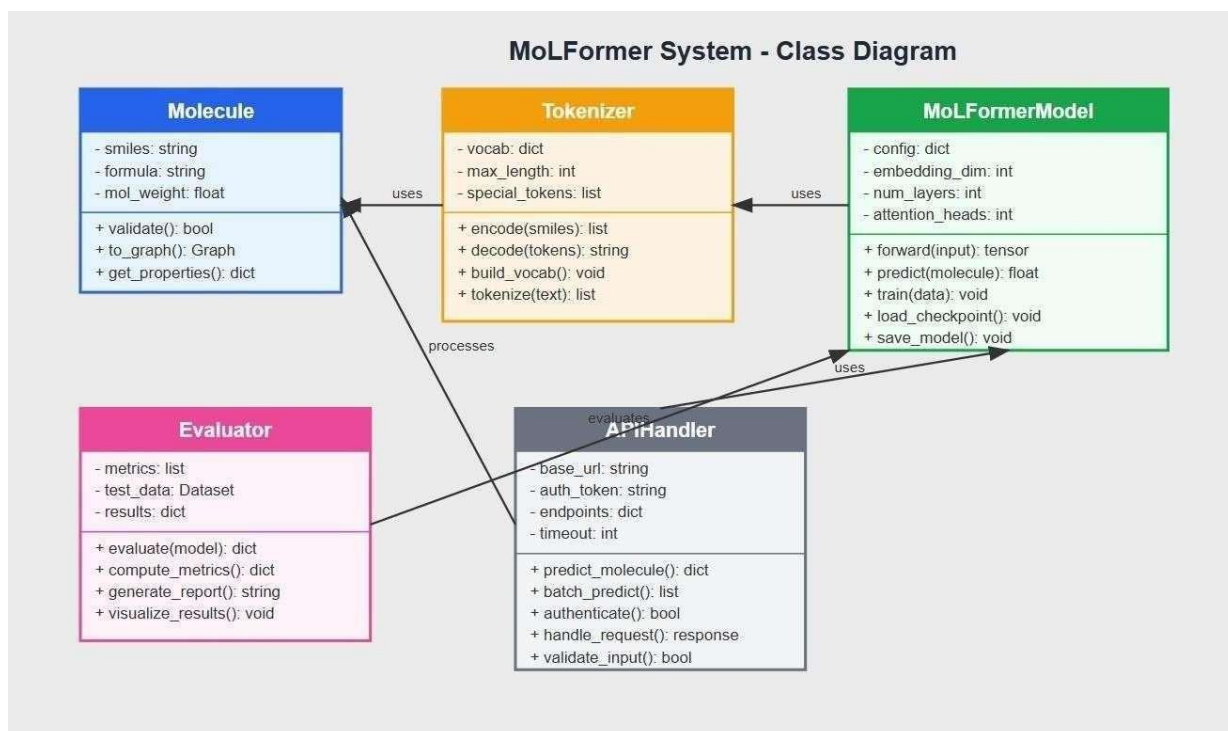
5.2. Proposed Methodology

- Data preprocessing involves tokenization and SMILES data cleaning.
- Model Training: Fine-tune MoLFormer on benchmark datasets.
- Use accuracy, RMSE, AUC, and R2 for the assessment.
- Deployment: Dockerized Streamlit/FastAPI-served APIs served through Docker.
- MLOps: retraining pipelines automatically, MLflow experiment tracking, and GitHub Actions CI/CD.

5.3. Detailed Design

5.3.1. Class Diagram:

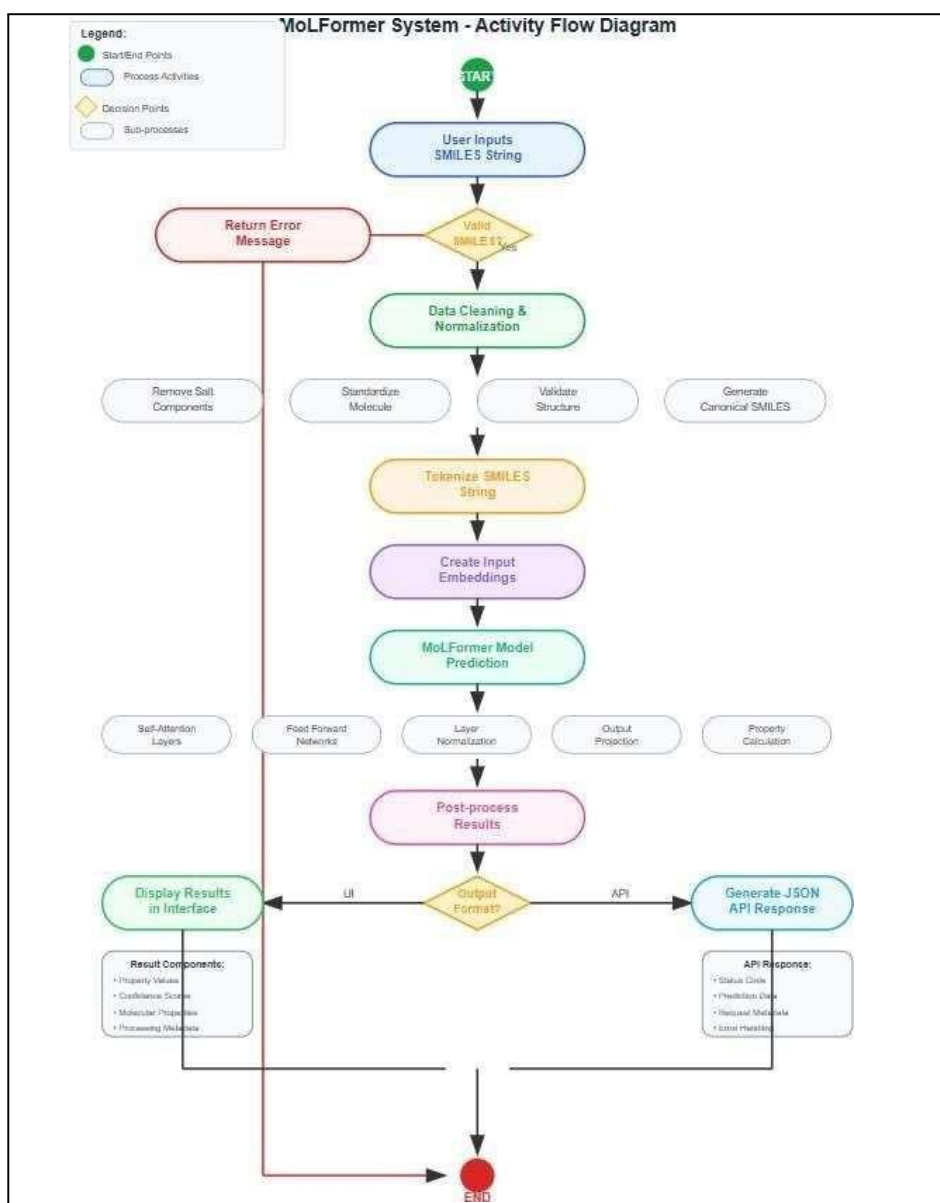
The MoLFormer prediction methodology for molecular properties is depicted by this class diagram. It establishes the object relationships between objects such as Molecule, Tokenizer, MoLFormerModel, Evaluator, and APIHandler. To facilitate modularity, each class encapsulates specific responsibilities. The architecture is designed for scalable and interpretable cheminformatics applications augmented with MLOps integration to make prediction, evaluation, and deployment feasible via APIs.



5.2 MoLFormer Class Diagram

5.3.2. Activity Diagram

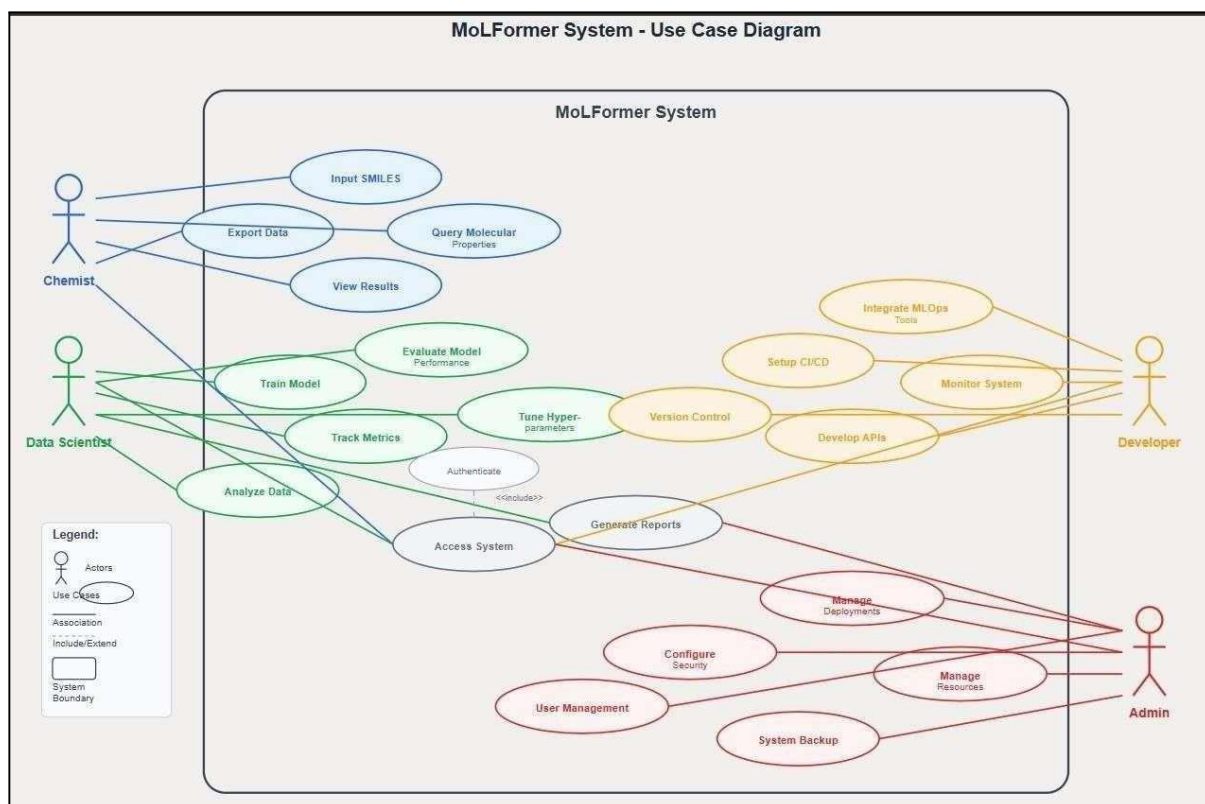
From SMILES input to validation, cleaning, tokenization, model prediction, and output production, this diagram shows the operational workflow of the MoLFormer system. In order to provide both UI display and API response functionality, it incorporates decision points for error handling and output formatting.



5.3 Activity Flow Diagram Of MoLFormer

5.3.3. Use Case Diagram

User interactions within the MoLFormer system are depicted in the use case diagram. In order to ensure collaborative and secure project execution, chemists input SMILES and examine findings, data scientists oversee model training and evaluation, developers integrate MLOps tools and APIs, and administrators manage system setup, security, and deployment.



5.4 Use Case Diagram Of MoLFormer

REFERENCES

- [1] D. Weininger, "A chemical language and information system: I. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988, doi: 10.1021/ci00057a005.
- [2] N. M. O'Boyle et al., "Open Babel: An open chemical toolbox," *J. Cheminf.*, vol. 3, no. 1, p. 33, 2011, doi: 10.1186/1758-2946-3-33.
- [3] R. Rong et al., "MoLFormer: Self-supervised transformer model for molecular representation learning," *arXiv preprint, arXiv:2206.15466*, 2022. [Online]. Available: <https://arxiv.org/abs/2206.15466>
- [4] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [5] S. Honda, H. R. Ueda, and S. Shi, "SMILES Transformer: Pre-trained molecular fingerprint for low-data drug discovery," *arXiv preprint, arXiv:1911.04738*, 2019. [Online]. Available: <https://arxiv.org/abs/1911.04738>
- [6] G. Grand, B. Ramsundar, and S. Chithrananda, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint, arXiv:2010.09885*, 2020. [Online]. Available: <https://arxiv.org/abs/2010.09885>
- [7] Z. Wu et al., "MoleculeNet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018, doi: 10.1039/C7SC02664A.
- [8] J. Gilmer et al., "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017. [Online]. Available: <https://arxiv.org/abs/1704.01212>
- [9] K. Yang et al., "Analyzing learned molecular representations for property prediction," *J. Chem. Inf. Model.*, vol. 59, no. 8, pp. 3370–3388, 2019, doi: 10.1021/acs.jcim.9b00237.

- [10] E. N. Feinberg et al., “PotentialNet for molecular property prediction,” *J. Chem. Inf. Model.*, vol. 58, no. 4, pp. 733–741, 2018, doi: 10.1021/acs.jcim.7b00696.
- [11] N. Brown, “In silico drug design,” *Chem. Rev.*, vol. 120, no. 4, pp. 1845–1879, 2020, doi: 10.1021/acs.chemrev.9b00147.
- [12] Y. Li et al., “Graph neural networks for molecular property prediction: A review,” *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–38, 2021, doi: 10.1145/3467193.
- [13] Z. Lin et al., “Self-supervised pre-training with knowledge distillation in molecular graphs,” arXiv preprint, arXiv:2007.00998, 2020. [Online]. Available: <https://arxiv.org/abs/2007.00998>
- [14] J. J. Irwin and B. K. Shoichet, “ZINC – A free database of commercially available compounds for virtual screening,” *J. Chem. Inf. Model.*, vol. 45, no. 1, pp. 177–182, 2005, doi: 10.1021/ci049714+.





11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography

Match Groups

-  **30 Not Cited or Quoted 10%**
Matches with neither in-text citation nor quotation marks
-  **2 Missing Quotations 2%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8%  Internet sources
- 7%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

5% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

