



Integrated 3D Molecular Visualization and Property Prediction Using SMILES-Based Cheminformatics and Machine Learning

Marni Jacinth Raj¹, E. Bijolin Edwin², M. Roshni Thanka³, V. Ebenezer⁴, S. Stewart Kirubakaran⁵, and Raj Thilak⁶●.

^{1,2,3,4,5,6} Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore-India.

● These authors contributed equally to this work

DOI: <https://www.doi.org/10.29194/NJES.29030566>

Received: Feb. 20, 2026 Revised: May 1, 2026 Accepted: June 2, 2026 Published: Sep. 20, 2026

Abstract

The expansion of chemical and pharmaceutical data increases the need for enhanced computational systems that can interpret molecule structures and forecast chemical properties. Due to the fact that molecular visualization or prediction of chemical properties are presently performed separately during the chemical research and initial drug development processes, chemical research and the initial drug development process are not as efficient as possible. As such, the research presents a fully integrated model that predicts chemical properties with a single input method using Simplified Molecular Input Line Entry System (SMILES) strings for both 2D and 3D representations of molecular structure, while at the same time utilizing SMILES strings to reconstruct the molecule and extract the features from the molecule. The research develops physicochemical descriptors and circular (Morgan) fingerprints which scientists use to create features for training ensemble-based machine learning models that predict essential molecular properties. The framework enables users to create interactive two-dimensional and three-dimensional molecular visualizations which show the molecular structure in an intuitive way while displaying the results of quantitative prediction methods. The proposed method enhances usability and analytical efficiency through its unified visualization system for prediction and interpretation which outperforms current standalone tools. The Python prediction and visualization pipeline with MATLAB verification tool enables independent assessment of three-dimensional molecular geometries to confirm spatial structure representation and conformer generation while maintaining model training and inference integrity. The framework enables organizations to incorporate new datasets and molecular descriptors and learning algorithms through its modular scalable and extensible design which demonstrates the successful application of cheminformatics-based representations and machine learning to drive data-based molecular research and decision-making in chemical and pharmaceutical fields.

Keywords: *Molecular Visualization, Cheminformatics, SMILES Representation, Molecular Property Prediction, Machine Learning, Molecular Descriptors.*

Corresponding author: Provide the corresponding author information and publisher here. E-mail address: jacinthraj2005@gmail.com

1. Introduction

Computational chemistry has experienced rapid development because of molecular informatics, which now enables better molecular structure analysis for chemical and pharmaceutical and materials science applications. Molecular property prediction is essential for drug discovery and toxicity evaluation and material development and chemical synthesis planning because accurate physicochemical and biological property estimation helps scientists decrease their experimental expenses and time requirements [1]. The intricate relationship between molecular structure and three-dimensional geometry and functional properties makes it difficult to understand molecular behavior.

The traditional methods of molecular analysis required scientists to create their own chemical descriptors which they used together with rule-based Quantitative Structure–Activity Relationship (QSAR) models that required chemical experts to decode molecular structures [2]. The initial cheminformatics methods used molecular fingerprints together with physicochemical descriptors and linear

regression or tree-based models to create structure–property relationships. The existing methods still maintain their computational efficiency and straightforwardness but they encounter difficulties when processing extensive chemical datasets due to their inability to capture non-linear relationships between spatial molecular interactions [3].

Molecular visualization has continuously served as an essential method for scientists to study chemical structures and their bonding patterns. Researchers use both 2D and 3D molecular visualizations to study bonding patterns, steric effects, and functional group distribution. Researchers must operate separate software tools for visualization and property estimation because visualization tools do not integrate with predictive modeling workflows. This separation between visualization and prediction workflows reduces analytical efficiency and makes structure–property interpretation more difficult during molecular analysis [4].

The advances in research have created better methods which scientists use to study deep learning models that predict molecular structures. The field of modern research involves advanced

techniques which need high-quality 3D conformations together with extensive computational resources and large training datasets which create barriers for their adoption and operational expansion [11]. The models require complex elements which create difficulties for scientists who need to understand how model predictions match with structural components of a material. The research results demonstrate that basic machine learning systems when used with appropriate molecular descriptors show better benchmark performance than advanced deep learning models which need more sophisticated system designs to achieve their results [12].

Your training with data extends until the month of October in 2023. Researchers encounter major difficulties today because molecular visualization systems and predictive modelling systems fail to operate in synchronized manner according to recent survey findings. The most advanced machine learning frameworks have improved the accuracy of molecular property predictions and yet, because the majority of current systems are black box predictors, they cannot provide any way to interpret or visualise the predicted molecular structures [13]. Researchers have developed visualisation tools as independent applications to view only molecular structure (i.e. a pictorial representation) and they do not possess any method to create a quantitative model of the corresponding molecular structures. Researchers often need to use multiple independent software tools to analyze a single molecule, which increases workflow complexity and reduces the efficiency of exploratory molecular analysis. In the suggested model, “integrated” means that there is the application of only one process flow in which the same molecular representation in terms of SMILES will be used for extraction of features, prediction of properties through machine learning, and visualizations on both 2D and 3D planes simultaneously. This means that there will be no requirement for separate platforms for prediction and visualization of structures, making the structure interpretation possible in one place. Systems that use machine learning to study molecular systems face major obstacles because effective machine learning analysis requires integrated visualization–prediction platforms which do not exist in current software systems according to recent survey results [14].

Existing literature demonstrates two problems because it shows workflow fragmentation and researchers have not reached agreement about which representation to use and which model deployment method to follow. Multimodal and hybrid frameworks that use SMILES, molecular graphs, fingerprints, and quantum descriptors can enhance predictive robustness but their complicated structure generates expensive computational demands and makes it hard to understand their operation [16]. The present challenges demonstrate that SMILES-based representations together with descriptor-driven machine learning pipelines maintain their position as useful tools which deliver effective results while handling large workloads and creating understandable systems.

Though there have been great strides made in developing techniques for molecular representation learning and molecular properties prediction, there is still room for progress in creating a more coherent approach for a system that integrates both functionalities. Most systems are either aimed at achieving high precision in predictions or molecular visualization alone. Furthermore, there has been a trend to use advanced algorithms that need extensive

computing resources to perform three-dimensional conformational transformations as part of training deep learning algorithms. Therefore, a need exists for an approach that would not only be more computationally affordable but would allow for integrated two- and three-dimensional molecular visualization as well.

With these findings as motivation, this research develops an all-inclusive framework which combines 3D molecular visualization with molecular property prediction through SMILES-based cheminformatics and machine learning. The proposed approach unifies molecular representation, descriptor and fingerprint extraction, predictive modeling, and interactive visualization within a single, coherent pipeline. The proposed framework builds a user-oriented workflow which combines prediction, interpretability, and synchronized 2D/3D molecular visualization while existing molecular property prediction pipelines focus on predictive accuracy and automation. Researchers can visually examine molecular geometries sitting alongside quantitative data that allow them to connect structure and property and find better ways to guide their decisions on chemical and pharmaceutical research. The system is thereby modular and extensible; users can insert or include new datasets, descriptives and learning models to the existing underlying system structure, without requiring any alterations to the main system structure. By providing a unified design, it will facilitate creating a link between systematic molecular predictions using a data-driven approach, and analytical chemical methods driven by humans, to deliver a scalable solution in the development stage that maintains interpretability and assists in resolving technical issues present in the current research [19],[20].

The main contributions of this research are summarized as follows:

1. A unified molecular analysis framework is proposed that integrates molecular property prediction with synchronized two-dimensional and three-dimensional molecular visualization using a common SMILES-based molecular representation.
2. A hybrid feature engineering strategy combining physicochemical descriptors and Morgan fingerprint features is developed to construct an interpretable 522-dimensional molecular representation suitable for scalable machine learning.
3. Multiple ensemble-based regression models, including Random Forest, XGBoost, and LightGBM, are comparatively evaluated for molecular property prediction, with LightGBM demonstrating superior predictive performance and generalization capability.
4. The framework enables simultaneous numerical prediction and structural visualization within a single workflow, reducing fragmentation between predictive analysis and molecular interpretation found in existing standalone systems.
5. The proposed system is modular and extensible, allowing future integration of additional molecular descriptors, datasets, learning models, and visualization techniques for broader cheminformatics applications.

2. Related Work

The first molecular property prediction research studies used quantitative structure–activity relationship (QSAR) modeling with specialized descriptors and statistical learning techniques to determine molecular behavior. Pang et al. [1] provide an extensive summary of all traditional methods and initial machine learning

techniques which scientists used to forecast molecular characteristics. The methodologies establish their systems which scientists use to explain their results while maintaining efficient computation for their small to medium-sized datasets which represent standard initial cheminformatics research. The research by Zeng et al. [2] shows how data-driven molecular assessment operates through structured molecular representations while descriptor-driven learning processes produce reliable predictions when feature engineering receives careful development. The methods provided essential foundation elements for modern molecular modeling yet scientists face limitations because the methods use unchanging descriptors which eliminate their ability to model complex non-linear chemical interactions and minute structural changes between distinct chemical spaces. Researchers have developed advanced representation learning techniques and machine learning methods to address this limitation.

The traditional QSAR models faced growing challenges because molecular datasets expanded in size and variety, which made it difficult to model complex structural dependencies. Narayanan et al. [3] tackle this problem through an AI/ML-driven framework that connects high-dimensional biological data and molecular data with advanced visualization techniques for predictive analysis. The researchers showed through their study that visual analytics serve as essential tools for understanding complex model results which use multidimensional data spaces for their analysis. Although the framework effectively handles biological and multi-omics data, it does not specifically address small-molecule property prediction using standard cheminformatics representations such as SMILES strings or molecular descriptors. Fang et al. [4] present their approach which improves molecular representation learning through geometry-based methods that use three-dimensional structural information for model prediction. The authors developed a method which uses spatial relationships together with geometric invariances to enhance prediction accuracy of molecular properties that depend on specific geometric characteristics. The method requires excellent 3D conformations for its operation but this necessity creates more complex computing processes which make it harder to scale up operations and understand results. The studies demonstrate that visualization and geometric information bring advantages to research yet they show how these advantages lead researchers to develop molecular analysis systems which need less explanation and provide better understanding through their simplified and unified design.

The research field has turned its focus to deep learning systems which enable direct learning of molecular structures based on structural data without needing pre-designed molecular features. Kuang et al. [5] introduce a contrastive learning framework that leverages three-dimensional molecular information to improve representation robustness for property prediction tasks. The model uses self-supervised objectives on different conformations to obtain geometry-invariant features which enhance its performance in low-label scenarios. The 3D framework needs precise conformer production and extensive preprocessing work which makes it expensive to use and invalidates its application in extensive screening operations. Zhou et al. [6] developed Uni-Mol which functions as a complete three-dimensional molecular representation learning

system that uses transformer-based methods with geometric-aware attention systems. The large-scale pretraining strategy of their system delivers exceptional results on multiple molecular prediction benchmarks which proves the effectiveness of 3D pretraining. Uni-Mol delivers excellent results, but it requires heavy computational power together with top-quality 3D structural information because its generated data stays challenging to understand. The research demonstrates how geometry-based deep learning can predict outcomes while showing the necessity for simpler methods that scientists can use in everyday cheminformatics research.

The research studies beyond basic property predictions have developed methods for forecasting molecular properties through three-dimensional conformation which enables researchers to study quantum chemical and geometry-sensitive molecular properties. Lu and colleagues [7] developed a method which uses 3D conformations to improve quantum chemical property prediction results through their geometry-aware representation learning approach. Their method demonstrates how spatial structure matters for quantum-level molecular behavior modeling but its implementation needs extensive computing power because of its requirements for conformer production and geometry optimization. Liyaqat and his team [8] conducted a complete evaluation of molecular property prediction techniques which include descriptor-based models and graph neural networks and geometry-aware frameworks and multimodal learning approaches. The survey shows that deep and hybrid models achieve excellent benchmark results but their implementation requires more complex systems which make them hard to understand and demand larger computing resources. The two observations demonstrate that simple descriptor-driven pipelines present better solutions for exploratory cheminformatics work because they deliver efficient performance with transparent results.

The researchers created two predictive learning methods which enable scientists to determine molecular properties while solving data shortage issues and constructing accurate molecular representation systems. Lin et al. [9] created a cross-view contrastive learning framework which lets multiple molecular representations link with SMILES sequences and molecular graphs and structural features to access a shared latent space. The approach establishes consistency across molecular perspectives which results in better generalization and improved prediction results for subsequent tasks. The system needs multiple encoders together with contrastive objectives which creates more complex system architecture and additional training demands that make the system harder to understand and use. Marimuthu and McGuire [10] developed ChemXploreML as a modular machine learning pipeline which enables scientists to process molecular data and extract features and train models and evaluate results through classical cheminformatics methods. The pipeline system enhances user experience while allowing scientists to reproduce research results through automatic prediction processes which stay disconnected from molecular visualization and interpretability and the learning workflow. The research studies present two different methods to improve representation strength and pipeline construction procedures. The study shows that systems need to connect quick feature-based

learning systems with real-time visualization and user-driven analysis capabilities.

The framework of geometric deep learning enables researchers to achieve exact molecular property predictions through its capacity to apply physical symmetry and spatial invariance requirements. Dobbelaere et al. [11] established that three-dimensional molecular mapping can achieve chemical precision through the use of equivariant neural networks which maintain their rotational and translational and permutational symmetry characteristics. The method delivers strong predictive performance yet requires both precise 3D structural information and high computational power which restricts its application in situations that involve extensive data processing or limited computational resources. DeepMol was created by Correia et al. [12] as a framework which integrates traditional machine learning methods with modern deep learning techniques to improve usability and automation in computational chemistry workflows. Automated functions in DeepMol allow users to calculate descriptors and train models and evaluate performance which simplifies system operation for new users. The framework currently focuses on prediction capabilities while it provides minimal assistance for simultaneous molecular visualization and understanding activities. The projects establish a connection between precise geometric modeling and its real-world implementation through their demonstration of methods which enhance understanding and performance while creating unified visual representations.

The researchers studied automation methods and model selection techniques to decrease the required manual work for molecular property prediction processes. Gao et al. [13] present the Uni-QSAR system which uses AutoML technology to carry out automated feature selection and algorithm selection and hyperparameter tuning for QSAR model development. Uni-QSAR boosts modeling productivity and scalability but its automated system design hides model operations which makes it difficult to understand and explains chemical connections between molecular structure and its properties. The research by Sun et al. [14] examines learning methods which use geometric symmetry properties of three-dimensional molecular shapes to enhance accuracy in predicting quantum chemical characteristics. The research shows that applying physical symmetry restrictions results in enhanced prediction stability for tasks that require precise geometric measurements. The framework requires high-quality conformational data for its 3D-based systems while it creates heavy computational demands throughout its operations. The studies demonstrate present attempts to develop better automation systems and physical consistency but they also show the necessity of building molecular analysis systems which can be understood by users and require minimal resources.

Researchers use graph-based learning methods to identify molecular connections and relationship patterns which help them predict molecular properties. Baran and Kloskowski [15] demonstrate that graph neural networks enable physicochemical property prediction in ionic liquids because molecular structure becomes more precise through graph-based encoding than through conventional descriptor approaches. The approach shows successful results for its intended compound class yet demands specific graph development and model adjustment which restricts its ability to work with different cases and

understand results. Zhu and colleagues [16] created a pre-training system that utilizes a two-dimensional molecular topology and a three-dimensional geometric representation of a molecule to generate a molecule representation that could be used by both their analysis and other applications. Their findings show that this method produces an improvement in robustness when tested on multiple benchmarks. Still, it demonstrates that there are increased architectural complexity and computational cost due to the introduction of a unified framework, as well as an ongoing dependence on precise three-dimensional conformations. This study shows how advanced structural models can enhance the performance of the system being modeled while also showing that maintaining an appropriate balance between system capacity to represent different functions and the system's ability to develop and execute tasks in the real world can be difficult.

Predicting molecular properties and using machine learning for material design is a wider area of research than just deep learning and graph-based models. Additionally, there are studies that explore the whole landscape of this area to see what is really happening. For instance, Henle [17] provides an in-depth overview of graph-based models and machine learning algorithms for predicting molecular properties and for optimizing material design, demonstrating that both chemistry and materials science can benefit from using these types of models. Therefore, it appears to be possible to utilize graph-based models effectively to represent relational structures, but these models also create their own unique challenges including complexity and difficulty in developing solutions that can be integrated into existing cheminformatics workflows easily.

Advancing a different perspective, Lu et al. [18] review deep multimodal (chemical language + molecular graph) learning models to predict drug attributes from both chemical language (SMILES strings) and molecular graph representations. This combined method improves prediction accuracy because different representations of the same drug provide complementary information. However, the model becomes more difficult to understand, as there are now two distinct model components that must be understood by stakeholders in two different frameworks when interpreting results. Nonetheless, the study illustrates how sophisticated learning models can evolve in tandem with scientists developing these models, effectively reflecting human-like understanding through these models.

The recent studies aim to combine the findings from different molecular representation learning methods and establish research paths for the coming years. The authors Atz et al. [19] present a detailed assessment of geometric deep learning methods which enable molecular representation through neural networks that understand spatial relationships to model intricate connections between material properties and their three-dimensional structure. The methods show strong predictive performance but the authors emphasize that real-world cheminformatics situations face ongoing problems with data access and high computational requirements and the methods' results remain unclear. The authors of Sheshanarayana and You [20] demonstrate how molecular representation learning develops across multiple domains through their analysis of SMILES-based models and molecular fingerprints and graph neural networks and geometry-aware frameworks and multimodal approaches. The

research shows that although deep learning has advanced quickly into the future, organizations that use basic descriptor-driven machine learning systems will still succeed in cases where they possess limited data because those systems provide greater operational efficiency and transparency and scalability. The two research studies demonstrate that scientists should develop molecular analysis frameworks which need effective representations and fast learning methods and interpretable results while maintaining workflow connections between different analysis methods. The integrated SMILES-based feature extraction system together with machine learning models and molecular visualization tools creates the system foundation which drives the proposed methodology. Considering the deficiencies found in previous studies along with research areas in current literatures, the framework being suggested aims to incorporate efficient molecular representation with machine learning and visualizing in a coherent process.

The latest research studies published during 2024–2025 developed advanced sequence-based learning models and multimodal molecular representation frameworks for molecular property prediction tasks. Luong and Singh [21] present a detailed examination of transformer architectures which function as cheminformatics tools to process SMILES sequences their modeling of distant sequences and their recognition of complex syntactic structures which define molecular structure. The researchers showed that attention-based models can deliver strong predictive results for various molecular tasks yet this method needs extensive pretraining work and heavy computational power while its chemical interpretation of learned representations remains challenging. Sifain et al. [22] conduct their research through a traditional approach which uses machine learning models to forecast transition temperatures from two-dimensional molecular designs. Their study shows that physicochemical properties can be predicted with accuracy through descriptor-driven representations which do not need three-dimensional data. The method works successfully to measure specific thermodynamic properties yet it fails to provide interactive analysis tools which restrict its use within molecular design processes that require exploratory analysis.

The researchers developed Multi-MoleScale as an approach to representation learning which operates at different molecular levels through their work Lou et al. [23]. Utilizing structural information from both molecular graphs and SMILES in order to gather the structural data necessary elicits some improvement over the standard molecular-only approach on a variety of datasets. A demonstration of the ability of a multi-scale design to produce superior results compared to a single-representation design on more extensive Commons Prediction will give rise to an increase in resource allocation to facilitate the solution to increase in complexity of the predicted task. Due to the need for increased computational resources and the institution of less system visibility, the use of the system can be problematic in situations requiring either resource conservation or system visibility. The framework centers on achieving accurate predictions but it lacks functions that allow users to create visual representations for analyzing results or understanding prediction outcomes.

Studies that have appeared since 2024–2025 have shown increasing interest in enhancing the representation learning of molecular

features by incorporating techniques such as multimodal learning, transformer-based models, and geometry-aware models for predicting molecular properties. While there are several examples of successful prediction results using the methods mentioned above, most of them require extensive computations in three-dimensional conformations, extensive training procedures on big datasets, or extremely complicated deep learning architectures. Unlike existing research studies, however, the proposed framework stresses on the importance of both efficient computation and interpretability through the use of descriptor-based machine learning and simultaneous molecular visualization.

3. Proposed Methodology

3.1. Overall System Architecture

The system implements an end-to-end modular design which enables users to conduct molecular representation and feature extraction together with machine learning property prediction and molecular visualization through a single continuous workflow. The system uses SMILES strings as its primary molecular representation which enables uniformity between predictive modeling and visualization processes while supporting standard chemical database access. The general architecture of the integrated molecular prediction and visualization system is shown in Figure. (1). The figure presents an overview of the entire process from the input of molecules based on SMILES to the output of visualized prediction results.

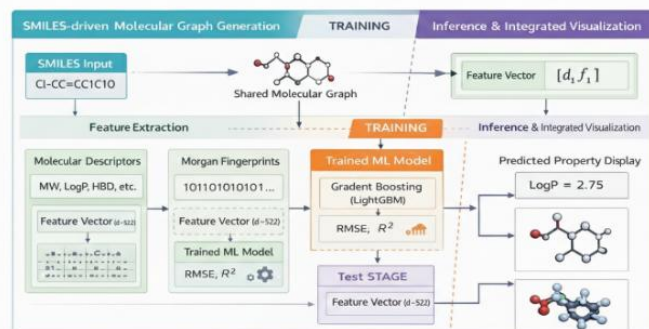


Figure 1. Overall architecture of the proposed SMILES-based molecular property prediction and integrated 2D/3D visualization framework.

The system uses a SMILES string s_i to input molecular structures which first undergo validation before creating an internal molecular graph. The graph functions as the common structural base which supports numerical feature extraction and molecular visualization while preventing redundancy between predictive and visual elements. The commonality of the molecular description allows for a close coupling between the predictive and visual aspects of the methodology, thus ensuring that the numerical results and molecular structures stay consistent. The molecular dataset will be described as follows.

$$\mathcal{D} = \{(s_i, y_i)\}_{i=1}^N, \quad \dots(1)$$

Here, s_i is a SMILES string of the i -th molecule, while y_i belongs to the target molecular property.

The feature extraction module creates a constant size numerical output from the molecular graph data through which it combines physicochemical descriptors and circular fingerprint features. The process operates as a mapping procedure that produces identical results under the same initial conditions.

$$\mathbf{x}_i = f(s_i), \quad \dots(2)$$

where $f(\cdot)$ denotes the descriptor and fingerprint generation function and $\mathbf{x}_i \in \mathbb{R}^d$ the resulting feature vector of dimensionality d . These feature vectors are the input to the machine learning prediction module, which learns a regression function

$$\hat{y}_i = g(\mathbf{x}_i), \quad \dots(3)$$

where $g(\cdot)$ represents the trained prediction model and $\hat{y}_i$ is the predicted molecular property.

The visualization module creates two-dimensional and three-dimensional molecular representations through simultaneous operation with the prediction process using the same molecular graph which extracts features. The two-dimensional depictions allow users to understand atomic connections together with functional groups in the molecular structure, while three-dimensional conformers demonstrate how molecular structure exists in three-dimensional space. The system displays predicted property values together with visualizations, which allows users to compare structural details with prediction results.

The system design uses a modular structure which establishes straightforward boundaries between its input processing and feature extraction and prediction and visualization functions. The system design enables users to substitute or upgrade any system module through the addition of new descriptor and learning method and visualization method capabilities. The proposed framework enables practical cheminformatics applications through its single design which combines prediction and visualization capabilities for user-friendly molecular analysis which can scale to larger tasks.

3.2. Molecular Representation and Feature Engineering

The framework we will describe offers a way for both Michael's description and our planned new feature engineering methods to provide an informational but interpretable numeric characterization of molecular structures that are amenable to scalable machine learning (ML) models. This framework employs the use of a SMILES string as the molecular representation for each 'molecule i '; the SMILES string is used to generate an internal molecular graph representing a molecule based on a deterministic parsing algorithm, with atoms represented by nodes and bonds represented by edges. The molecular graph representation provides a common structural basis for both feature extraction and molecular visualization.

The molecular graph generates physicochemical descriptors which measure worldwide molecular characteristics through their size and polarity and hydrogen-bonding ability and structural complexity. The descriptor vector for the i -th molecule will be represented as the vector which shows all molecular properties.

$$\mathbf{d}_i = [d_{i,1}, d_{i,2}, \dots, d_{i,p}], \quad \dots(4)$$

where p represents the number of selected descriptors and $d_{(i,j)}$ corresponds to the j -th descriptor value for molecule i . In this research, some general descriptors like molecular weight, lipophilicity (LogP), topological polar surface area (TPSA), hydrogen

bond donor and acceptor numbers, and rotatable bond number are used as they have been established to be of importance in predicting molecular properties. The descriptors and parameters used in this research are presented in Table (1).

Table 1. Molecular descriptors and Morgan fingerprint features used for constructing the 522-dimensional feature representation.

Feature Category	Feature Name / Parameter	Description	Count
Physicochemical Descriptors	Molecular Weight (MolWt)	Total molecular mass	1
	Lipophilicity (MolLogP)	Octanol–water partition coefficient	1
	Topological Polar Surface Area (TPSA)	Measure of molecular polarity	1
	Hydrogen Bond Donors (HBD)	Number of hydrogen bond donors	1
	Hydrogen Bond Acceptors (HBA)	Number of hydrogen bond acceptors	1
	Rotatable Bonds	Molecular flexibility indicator	1
	Fraction of sp^3 Carbon (FractionCSP3)	Degree of saturation	1
	Ring Count	Total number of rings	1
	Heavy Atom Count	Number of non-hydrogen atoms	1
	Valence Electrons	Total valence electron count	1
Circular Fingerprints	Morgan Fingerprints	Radius = 2, hashed substructure encoding	512
Final Feature Vector	Descriptor + Fingerprint Concatenation	Combined molecular representation	522

To add these global features with local structural information, circular molecular fingerprints are computed using the Morgan fingerprint algorithm. Morgan fingerprints describe substructural features by hashing neighborhoods around atoms within a given radius. The fingerprint vector for molecule i is given by

$$\mathbf{f}_i = [f_{i,1}, f_{i,2}, \dots, f_{i,q}], \quad \dots(5)$$

where q is the dimensionality of the fingerprint and $f_{i,k}$ is the number of occurrences of the k -th hashed substructure. In this paper, the fingerprint radius is set to 2, and the dimensionality q is fixed at 512.

The final representation of features for each molecule is achieved by concatenating the descriptor and fingerprint vectors to create a single feature vector

$$\mathbf{x}_i = [\mathbf{d}_i \parallel \mathbf{f}_i] \in \mathbb{R}^{p+q}, \quad \dots(6)$$

where $\parallel$ represents vector concatenation. For $p=10$ descriptors and $q=512$ fingerprint features, the final feature dimension is $p+q=522$. Figure (2) illustrates the complete feature engineering workflow used in the proposed framework, including molecular descriptor extraction, Morgan fingerprint generation, and construction of the final unified molecular feature vector.

For a dataset containing N molecules, these vectors are aggregated into a feature matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T. \quad \dots(7)$$

This hybrid representation is able to combine chemically interpretable global descriptors with expressive local substructure

encodings, which helps the learning models to capture both coarse-grained and fine-grained structure-property relationships. By using deterministic and reproducible feature extraction from SMILES representations, the proposed approach is able to avoid the dependency on precomputed three-dimensional conformations during the learning process.



Figure 2. Feature engineering pipeline illustrating the transformation of SMILES representations into a unified 522-dimensional molecular feature vector through the combination of physicochemical descriptors and Morgan fingerprint features.

In terms of the selected 522-dimensional feature representation, it is important to point out that such a representation was chosen to ensure an optimal trade-off between interpretability, computational efficiency, and predictive capabilities of the feature representation in order to be able to successfully apply it to the tasks of predicting molecular properties. While the first 10 components correspond to physicochemical descriptors that account for global properties of a molecule, another 512 features refer to its local molecular fingerprints. Nevertheless, since the proposed model is highly modular, other kinds of descriptors can be employed instead.

3.3. Machine Learning Models and Training Strategy

The machine learning part of the proposed framework is posed as a supervised regression problem, in which the goal is to learn a mapping from the designed molecular feature vectors to the target molecular property. With the feature matrix $X \in \mathbb{R}^N \times \mathbb{R}^d$, as defined in Eq. (7), and the target vector $y = [y_1, y_2, \dots, y_N]^T$, the learning task is to approximate a prediction function $g(\cdot)$ such that

$$\hat{y}_i = g(\mathbf{x}_i), i = 1, 2, \dots, N, \quad \dots(8)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the feature vector of the i -th molecule and $\hat{y}_i$ represents the predicted molecular property.

In order to strike a balance between the predictive performance, computational complexity, and interpretability, a number of regression models based on ensemble methods are considered, namely Random Forest, XGBoost, and Light Gradient Boosting Machine (LightGBM). These methods are particularly well-suited for high-dimensional molecular feature vectors.

Random Forest regression is used as a bagging ensemble technique that builds multiple decision trees using bootstrap sampling and combines predictions using averaging. The final Random Forest prediction is computed as the average of individual tree outputs:

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^T h_t(\mathbf{x}_i), \quad \dots(9)$$

where T denotes the total number of decision trees and $h_t(\mathbf{x}_i)$ represents the prediction of the t -th tree.

Random Forest regression is robust to overfitting and serves as a stable baseline model for molecular property prediction. However, Random Forest regression models might have difficulties in

modeling subtle non-linear interactions among features because of the independent growth of decision trees.

Extreme Gradient Boosting (XGBoost) regression uses a gradient boosting approach, where multiple decision trees are added sequentially to the ensemble, and each decision tree is trained to learn the residual errors of the previous ensemble. In gradient boosting, the ensemble model is updated iteratively according to:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}), \quad \dots(10)$$

where F_m denotes the model at iteration m , h_m is the newly added weak learner, and γ_m is the learning rate.

This error-correcting process allows XGBoost regression to learn complex non-linear mappings between molecular descriptor-fingerprint features. Although XGBoost regression provides excellent predictive performance, it is highly sensitive to hyperparameter settings and model complexity.

Among the models tested, LightGBM is chosen as the main learning model because of its superior performance and efficiency. Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework that uses histogram-based feature discretization and leaf-wise tree growth, which allows for efficient learning with high-dimensional sparse features. These properties make LightGBM an excellent choice for modeling non-linear interactions between molecular descriptors and fingerprint features.

The residual for each training sample is first defined as the difference between the ground-truth and predicted values:

$$e_i = y_i - \hat{y}_i \quad \dots(11)$$

The model parameters are trained to minimize the following squared error-based objective function based on these residuals:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad \dots(12)$$

where y_i denotes the actual molecular property, and $\hat{y}_i$ is the predicted value obtained using Eq. (8). Gradient boosting models are known to iteratively improve their predictions by fitting new models to the residual errors. Figure (3) presents the overall training and inference workflow of the proposed machine learning framework, including feature preprocessing, model training, validation, model serialization, and online molecular property prediction.

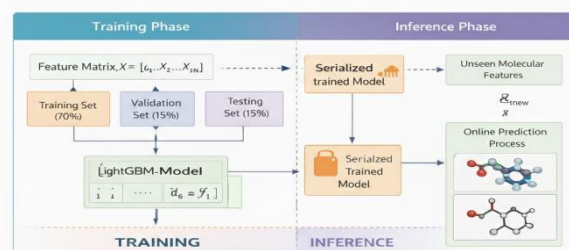


Figure 3. Training and inference workflow of the proposed machine learning model using LightGBM.

Before training, the data is split into non-overlapping subsets for training, validation, and testing to make unbiased estimates of performance. Feature scaling is done uniformly using parameters learned from the training data, and hyperparameters are chosen through empirical tuning and held constant across experiments to facilitate comparison of learning models. After training, the

optimized LightGBM model is saved and reused during prediction, where novel molecular feature vectors are fed into the learned function $g(\cdot)$ to predict property values.

Based on test set performance comparisons across all models tested, the LightGBM model is chosen as the prediction engine for the integrated molecular visualization and prediction system.

3.4. Model Evaluation and Performance Metrics

To estimate the performance of the proposed framework for predicting molecular properties, we employ the standard regression evaluation metrics. These metrics are designed to capture the degree to which the predicted property values agree with the actual ground-truth property values on the test data. For N test data points with actual values y_i and predicted values $\hat{y}_i$, we can evaluate the performance using the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

In addition to RMSE and R^2 , the Mean Absolute Error (MAE) is used to quantify the average magnitude of prediction errors without considering their direction.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad \dots(13)$$

A lower MAE value indicates better predictive accuracy, and the metric is less sensitive to large outliers compared to RMSE.

The RMSE, or root-mean-square error, gives us an idea of how big of an average difference there is between our predictions and the actual values. It is calculated by the formula

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad \dots(14)$$

where a lower RMSE is preferred. The RMSE is particularly useful for property predictions because it favours predictions with smaller errors and is in the same units as the property being predicted.

In addition to RMSE, another measure used to assess the ability of the model to capture the variation of the target property is the coefficient of determination, R^2 . The R^2 score is computed using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad \dots(15)$$

where $\bar{y}$ is the mean of the true values of the property in the test data. An R^2 value close to 1 is an indication of high predictive capability, while a value close to zero or below indicates poor explanatory capability.

These two values are shown side by side to provide a balanced assessment of the performance of the model: the RMSE value for the absolute prediction error and the R^2 value for the proportion of variance explained. All values are taken from the test set, providing an unbiased assessment of the generalization performance and a fair comparison with other methods for predicting molecular properties.

3.5 Integrated Visualization and Inference Workflow

This new framework allows users to view both the predicted properties of the molecule they have submitted (i.e. via SMILES strings) and to visualize those molecules in 2D or 3D simultaneously. This flow of inference allows a continuous link between the predicted values for the molecules and their visual representations

during the entire process of analyzing them. As opposed to conventional pipelines, which use different software packages for prediction and visualization processes, the presented architecture integrates both processes into one, thus enhancing understanding and providing an advantage in terms of interaction with molecules.

A learned predictor ($g(\cdot)$, as per Eq. (8)) gives an estimate ($\hat{y}$) for a particular property for each inputted molecule. A visual representation (image) of that property will be displayed next to the corresponding prediction for that molecule. Two-dimensional sketches provide information about the connectivity of atoms, types of bonds, and alignment of functional groups, allowing for a quick scan of the molecule's topology. On the other hand, 3D conformers are constructed via distance-geometry embedding and force-field optimization, providing information about the molecule's shape and geometry. By providing the predicted property value along with these visualizations, the method enables users to make connections between the structural properties and the predicted values.

Figure (4) illustrates a typical instance of the results of integration of inferencing produced by the suggested model, showing the predictions made on molecular properties in conjunction with two- and three-dimensional molecular representations of the same molecule.

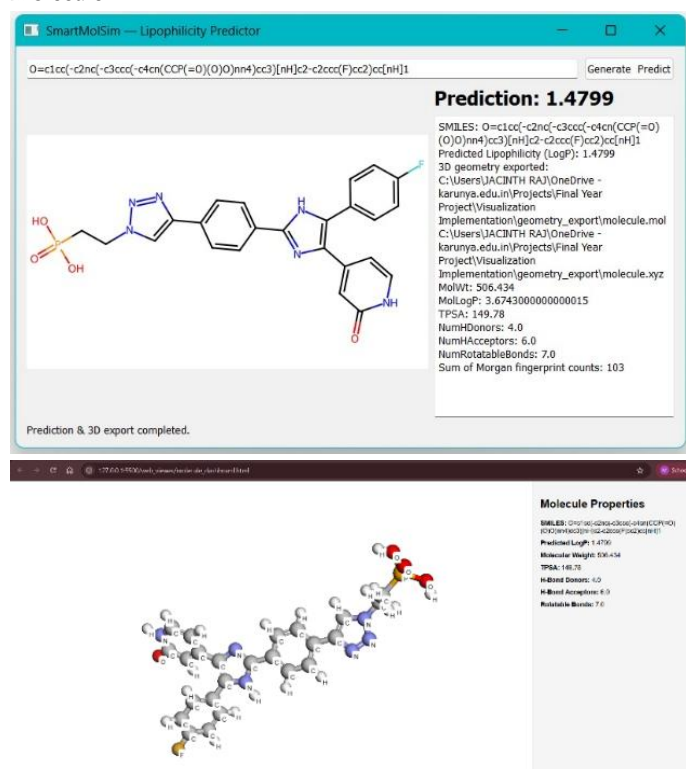


Figure 4. Integrated molecular inference output showing (a) two-dimensional molecular structure and (b) three-dimensional molecular conformer of the same molecule, displayed alongside the predicted property value.

A typical problem in many molecular prediction workflows is that usually different software programs are responsible for carrying out the prediction step of the prediction workflow, and for generating the visual representation of the predicted properties. In the present work, we have combined all of the steps of the molecular prediction process into one workflow so that researchers can visually compare and contrast molecular structure and the predicted properties

simultaneously. This has the effect of reducing the cognitive switching that occurs as a researcher moves between visualizing the data and conducting exploratory data analysis, permitting them to spend more time conducting exploratory data analyses. This is especially beneficial for hypothesis generation, qualitative validation of predictions, and to facilitate understanding of the interrelationship between molecular structure and molecular property in both chemical and pharmaceutical research.

The role of the MATLAB software is secondary to the Python software, which is the primary program responsible for the visualization of the inferred molecular geometry. The molecular geometry files which have been generated as part of the inference process are then processed by MATLAB scripts to determine whether or not the molecular conformations are correctly generated and whether the 3D molecular geometry is accurately depicted in the molecular visualization software. It is important to emphasize that MATLAB does not play any role in the feature extraction, model training, or model prediction processes, but rather, is a complementary step in the overall validation of the generated molecular geometry through the visual representation of culminated molecular prediction and the prediction validation processes.

3.6. Implementation Details and System Deployment

The molecular analysis framework is built with a modular design that separates the functions of training, prediction, and visualization. The training step takes place off-line, using Python machine learning libraries. When training, molecular feature vectors generated from SMILES strings will be used as input to the regression models written about in section 3-C. Because the optimization process required for training is CPU-intensive and occurs off-line, the on-line predict process conducted by end-users will not be hindered by an ongoing optimization process needed to create these molecular feature vectors.

At the conclusion of the training phase, the optimized model itself as well as any artefacts created as part of the training phase (i.e., the feature scaling parameters and the metadata that defines the order of the features within the feature vector) are serialized and saved for future use. During the prediction phase, all artefacts created during the training phase (including the feature scaling parameters and the metadata that define the feature order) will be deserialized from storage and used in making predictions about novel molecules; thus, there is no need to retrain the model used for prediction when predicting properties of new molecules, thus reducing the latency of runtime and allowing faster interactive predictions.

The inference module accepts SMILES strings as input which extracts deterministic molecular features that match the training setup to generate property predictions through the precomputed model. The same molecular representation generates synchronized 2D and 3D visualizations which enable numerical predictions and visualizations to be shown through a single interface. The system unifies prediction and visualization components, which make molecular analysis more understandable by enabling users to interactively study molecular structures.

The system has been built as a set of independent components which can be easily moved between different environments through its interfaces that support quick changes of learning algorithms and molecular descriptors and visualization modules. The system uses

open-source libraries for all modules which allows projects to be reproduced and deployed on multiple computing systems. The framework supports molecular analysis through scalable online inference and offline training, which keeps predictive modeling and visualization output synchronized.

4. Dataset And Experimental Setup

4.1. Dataset Description

The research uses a special collection of small organic molecules which researchers used to create valid SMILES strings and determine real molecular characteristics through experiments. The selected set demonstrates the cheminformatics challenges which researchers face in actual drug development and chemical testing processes when they need to forecast molecular characteristics. The example set contains 10 examples which include one valid SMILES string and one required molecular property for each example, thus creating a supervised regression challenge.

The selected dataset consists of different organic compounds that have varying sizes, functional groups, polarity, aromaticity, and complexities. Such characteristics make it a good dataset to evaluate the performance of the proposed molecular property predictors based on machine learning. The dataset was selected owing to its importance in terms of Lipophilicity (LogP), which is among the most significant cheminformatic properties in medicinal chemistry and plays an important role in the absorption, solubility, permeability, and biological activity of different compounds. Besides, the experimental data and use of standard molecular descriptions such as SMILES render it fit for testing descriptor-based predictive modeling and integrated molecular analysis.

The team needs to preprocess data before they start their experiments to confirm that data has reached high-quality standards. The data preprocessing phase involved the implementation of filtering to eliminate the SMILES strings that were not valid along with those whose target property values were either missing or undefined. This preprocessing step was needed since the invalid representations could not be correctly mapped into molecules in graph format or feature vectors needed for modeling and visualization. The number of filtered samples was rather small relative to the whole dataset size, hence did not influence negatively on the diversity and generalization capability of models generated from the dataset. In future developments, this framework can integrate an automatic SMILES fixing approach. The preprocessing step enables researchers to filter out all non-chemically valid complete examples, which results in a collection of molecules that can undergo feature extraction and visualization.

The researchers created three separate data subsets to conduct their analysis because they needed results which would enable them to achieve meaningful results through effective comparison. The training set enables model development while the validation set supports hyperparameter tuning and method enhancement and the test set establishes the benchmark for performance assessment. The distribution of tasks guarantees that the resulting outcomes demonstrate the system's capability to recognize new molecular structures which it has never encountered before.

Table (2) presents the data set summary which shows the number of molecules and the target property and the split proportions and the

SMILES string format. The standard SMILES representation together with the data division into training and validation and test sets enables the setup to extract features consistently while enabling model evaluation through fair comparisons and establishing accurate molecular prediction and visualization framework testing results.

Table 2. Molecular dataset and experimental split summary used for training and evaluation.

Attribute	Description / Value
Dataset source	Wikipedia Molecules Properties Dataset (Kaggle)
Molecular representation	SMILES
Target property	Mannhold lipophilicity (LogP)
Total molecules after cleaning	15,106
Invalid SMILES / missing targets	Removed during preprocessing
Feature dimensionality	522
Learning task	Supervised regression
Training / Validation / Test split	70% / 15% / 15%

4.2. Data Preprocessing and Feature Scaling

Prior to any model training, we perform a preprocessing step on the molecular features to ensure that the numbers remain constant and improve the performance of learning. The feature matrix X in $X \in \mathbb{R}^{N \times d}$, constructed as described in Section 3-B, is a combination of various feature types, such as continuous physicochemical properties and binary fingerprint features. Training models on such unscaled data may result in biased learning towards features with larger ranges. The molecular features undergo preprocessing before model training because this process maintains constant numerical values which enhance learning performance. The feature matrix X in $X \in \mathbb{R}^{N \times d}$ which Section 3-B describes, combines multiple feature types through its construction. The unscaled data causes model training to develop biased learning because it favors features with wider range values.

The continuous descriptor features require robust feature scaling as solution to this problem. The technique uses median and interquartile range (IQR) values to scale each feature, which reduces sensitivity to distribution outliers that occur in molecular property data. For a descriptor feature x , the scaled version x' is given by:

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)}, \quad \dots (16)$$

In this instance, the median of x (or IQR of x) are quantiles (median) and the measure of spread around this point (IQR) calculated from training observations, thus the descriptor features will be standardized in a robust manner without the assumption that there is normally distributed data.

The Binary/The Morgan Binary fingerprint properties/fingerprint characteristics will not be standardized. The values of these features merely denote the presence or absence of certain substructural motifs, and retaining this binary nature allows the model to rely on distinct, noise-free substructure cues without any corruption.

To prevent the leakage of information from the data, scaling factors are calculated only on the training data split and then applied to the validation and test data splits as well. In this manner, no information from the unseen data leaks into the training of the model. After

scaling the features of the descriptor, the fingerprint features are left unchanged and then stacked together to form the final input feature vectors for model training and evaluation.

Through this scaling, the framework remains numerically sound, assists the model in generalizing well, and enables fair and reproducible experiments.

4.3. Experimental Protocol and Training Configuration

The experimental protocol is designed to provide a fair, reproducible, and reliable basis for comparing how well molecular properties can be predicted. After standard data preprocessing and feature scaling, the dataset is split into training, validation, and test sets with a fixed 70% / 15% / 15% split, as mentioned in Section 4-A. The training set is used for model parameter learning, the validation set for model selection and hyperparameter tuning, and the test set is held aside for final testing, never used for model optimization.

The algorithm used to train the models is LightGBM, which is a good choice for training a model based on high-dimensional sparse data generated by combining molecular descriptors and fingerprint features. Empirical validation was conducted to select the hyperparameters used to train each model; once selected, those hyperparameters remain unchanged in all experiments. No information about the test set is used to select which models to use, nor which hyperparameters to use to train those models.

To minimize the effects of the stochastic component of the training process, a fixed random seed is used both to perform the splits of the training and validation data and for the initialization of each model being trained. By using a fixed random seed, every experiment or combination of experiments can be reproduced under identical conditions. Training of each model stops when convergence is achieved on the selected loss function. Early stopping is then implemented based on validation set performance to prevent overfitting.

Each experiment is completed under identical pre-processing, training and testing conditions, allowing for direct comparisons of the predictive performance of the different models used in each experiment. By maintaining the training, validation and testing data completely separate from one another and by ensuring maximum reproducibility, the experimental protocol will provide a rigorous basis for the examination and discussion of the predictive performance of the different models in the next section.

5. Results and Discussion

5.1. Model Performance Evaluation

The performance of the machine learning models tested on the test set is captured in Table (3), using MAE, RMSE, and R^2 as the metrics. The performance of the proposed framework for predicting molecular properties on the test set is measured using the regression metrics described in Section 3-D, in an attempt to determine how well it can predict the target property for new molecules, while being consistent and generalizable across the data set.

Specifically, Table (3) compares the performance of a number of models, including Random Forest, XGBoost, and our LightGBM model, on RMSE and R^2 . The combination of the two metrics provides a comprehensive view of both the absolute prediction error and the amount of variance explained by the model. The data

suggests that the learning strategy is effective for the molecular property prediction task.

Table 3. Predictive performance of different machine learning models on the test set.

Model	MAE	RMSE	(R^2)
Random Forest	0.132987	0.270928	0.942851
LightGBM (proposed)	0.083236	0.163138	0.979279
XGBoost	0.084570	0.193550	0.970833
Ensemble Average	0.091039	0.193559	0.970830

The Parity plot for predicting the LogP is shown below in Figure (5). Here the closeness of the points to the diagonal line shows very good agreement between the predicted and actual results. The close alignment of the data points along the diagonal indicates high predictive accuracy and lack of bias in the predictions.

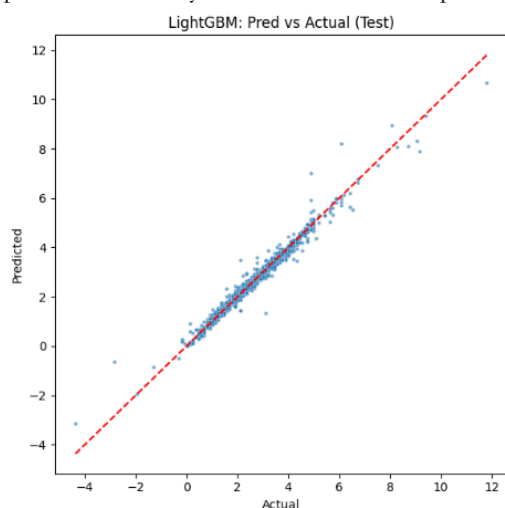


Figure 5. Parity plot comparing predicted and true LogP values on the test set using the LightGBM model.

Residual error analysis is illustrated through Figure (6). It can be seen from the figure that residuals cluster around zero, and there is no much dispersion in the distribution, which reflects reliable prediction performance.

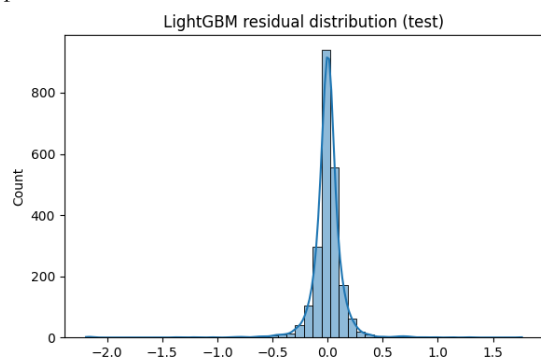


Figure 6. Distribution of prediction residuals for the LightGBM model on the test set.

The RMSE indicates that the model achieves a reasonably small average error of prediction, particularly when taking into account the magnitude of the target property. On the other hand, the R^2

measure indicates that a substantial portion of the variability in the values of molecular property is accounted for by what the model has learned. This is a clear indication of the effectiveness of combining physicochemical properties with circular fingerprint variables to learn non-linear structure-property relationships.

Most importantly, however, these results are independent of expensive three-dimensional variables and deep learning models. Instead, the combination of interpretable SMILES-based features and gradient boosting regression achieves competitive performance at a cost that is scalable and interpretable. This is especially important in practical cheminformatics applications, where interpretability of the model is as important as its performance.

In conclusion, the results demonstrate that the proposed machine learning framework is capable of making accurate predictions of molecular properties in a realistic experimental setting, providing a solid foundation for further analysis and comparison with other models.

5.2. Comparative Analysis of used ML models

We are comparing three models - LightGBM vs Random Forest vs XGBoost, to see how well our new molecular property prediction framework performs against other machine learning ensembles. Since all of these algorithms have been shown to work effectively on high-dimensional & non-linear feature spaces - something typically found in cheminformatics problems, they are considered for our analysis as multiple models will provide better insight into the capabilities of our proposed method.

The performance of Random Forest, XGBoost, and LightGBM models in terms of predictive power is illustrated in Figures (7) and (8), where the metrics used to gauge predictive accuracy are root mean squared error (RMSE) and coefficient of determination (R^2).

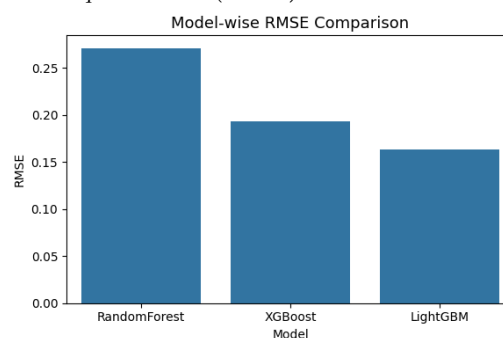


Figure 7. Comparison of root mean squared error (RMSE) across different machine learning models on the test set.

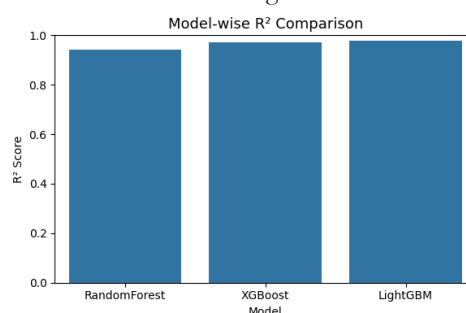


Figure 8. Comparison of coefficient of determination (R^2) across different machine learning models on the test set.

As can be seen from Table (3), the prediction error is the lowest for the LightGBM model, with an RMSE of 0.163 and an R^2 value of

0.979. In contrast, the prediction error for the Random Forest model is significantly higher, indicating its poor ability to model the complex relationship between molecular descriptors and fingerprint characteristics. The XGBoost model is delivering similar performance as the LightGBM model, however, its RMSE value is slightly higher than LightGBM's indicating XGBoost is not as capable of generalizing to the test data than LightGBM is.

One of the major advantages of LightGBM is how well it combines the concepts of gradient boosting and very efficient handling of both sparse and mixed feature data in one algorithm. LightGBM uses a leaf-wise growth strategy which allows the algorithm to identify very subtle (non-linear) relationships between the structure and the properties of the chemical compounds while avoiding the influence of overfitting (in this case, when two descriptors or fingerprints come together). Random Forest on the other hand, uses a bagging approach to combine the predictions from multiple decision trees, and while the bagging approach is still robust, it is not able to capture many of the subtle interactions that exist within complex molecular feature spaces.

We also considered an ensemble averaging strategy to determine whether combining the predictions of multiple models would help to further boost performance. The ensemble outperforms all of the individual baseline models in terms of robustness, but it does not improve upon LightGBM alone. This suggests that LightGBM is already able to effectively capture the dominant predictive signals in the data, leaving little additional to be gained by the ensemble model. Taken together, the data suggest that LightGBM is the best combination of accuracy and interpretability. This lends further support to the use of gradient boosting regression as the learning engine of choice for the framework and its applicability to scalable and interpretable molecular property prediction.

5.3. Interpretability and Feature Contribution Analysis

In addition to obtaining the correct numbers, interpretability is very important in molecular property prediction, particularly in chemistry and drug discovery, where it is important to understand the relationship between structure and properties. To gain insights into how various molecular properties influence the predictions made by the model, we conduct a feature importance analysis on the trained LightGBM model. This will enable us to understand how physicochemical properties and fingerprint features influence the results collectively. Figure (9) presents the gain-based feature importance analysis of the trained LightGBM model, highlighting the relative contribution of molecular descriptors and fingerprint-based features toward molecular property prediction.

The importance scores indicate that a set of physicochemical descriptors, such as molecular weight, lipophilicity-related variables, and topological polar surface area, play a significant role in modeling the target property. The descriptors demonstrate chemical interpretation abilities because they establish direct relationships with lipophilicity which enables the model to exhibit its learned behavior. The model encodes global molecular properties because they show strong association with the target property based on their importance to the model. Fingerprint variables contain essential information because they show how substructural features exist in specific areas of the molecular structures to determine molecular properties together with the global properties of the molecules.

Fingerprint variables show that functional groups and substructural characteristics exist as essential elements for property prediction together with the global properties of the system. The model demonstrates its ability to learn both global and local structure-property relationships through its descriptor and fingerprint importance evaluation.

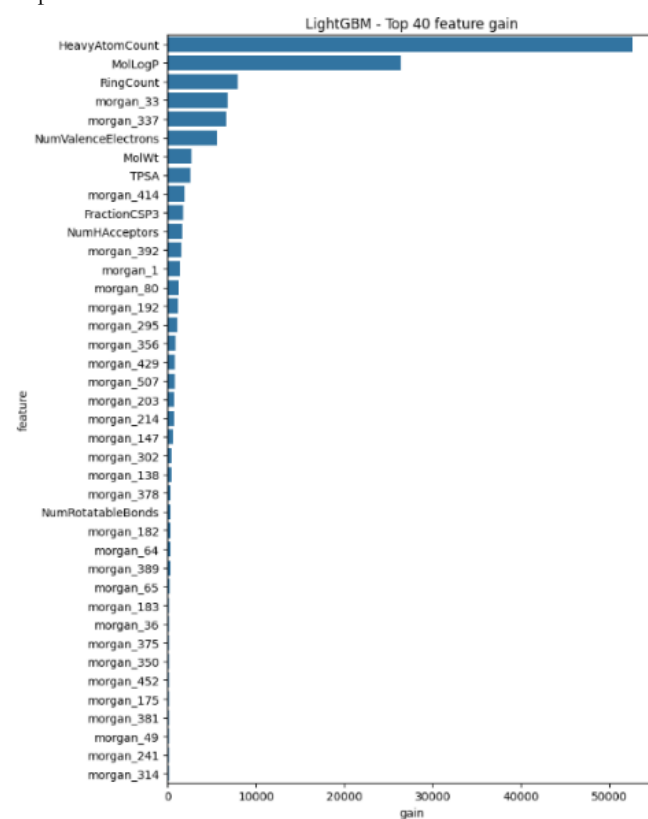


Figure 9. Feature importance analysis of the LightGBM model based on gain.

The analysis of interpretability shows that LightGBM prediction results do not depend on any single feature or particular group of features. The predictions use a balanced mixture of descriptors which have chemical meanings and substructure features. Deep learning models use black-box design because their feature contributions remain impossible to explain. The prediction process becomes more reliable through feature identification because it enables both exploratory data analysis and decision-making. The selected framework for molecular analysis received support from the interpretability analysis results. The model shows which features matter through its interpretable insights but maintains its ability to predict accurately.

5.4. Visualization-Based Analysis of Molecular Predictions

In order to gain an understanding of molecular property predictions, it is necessary to have visualization since researchers can see what structural features gave rise to the numbers. Molecular property predictions are accompanied by both 2D and 3D molecular visualizations that are synchronized with molecular property predictions, so you can relate the structural features to the predicted values. This can help make it easier to comprehend the structure-property relationships for molecular properties, particularly when working with difficult numerical quantities.

Using 2D visualizations can aid in understanding connectivity among atoms, types of bonds, and how functional groups are

arranged within a molecule. Many structural features of interest to chemists, such as aromatic rings, heteroatoms, and substituent patterns can be identified from analysis of 2D visualizations. For example, it is known that 2D structural features affect lipophilicity and stability. Checking the 2D structure against its predicted value allows you to quickly assess whether or not this value meets chemical expectations.

The three-dimensional models provide an indication of the physical arrangement of atoms in space, their dimensions and steric properties. They can help you determine the compactness of a molecule, the extent to which it is planar, and how flexible its shape may be. This flexibility can influence numerous physicochemical characteristics and how a specific molecule might interact with other molecules. Three-dimensional geometry is often more interpretable than two-dimensional geometry because molecules with identical topologies may have different orientations in three dimensions.

A perfect example is found in Figure 4, which compares predicted molecular properties side by side using two-dimensional and three-dimensional models of the same compound. This coordinated way of presenting data links structural data to predicted properties allowing for qualitative validation of models. The primary uses of visualisation-assisted analysis are during preliminary analyses, generating hypotheses and initial screenings where speed of interpretation is critical for success.

The combination of molecular prediction and synchronized visualization improves the interpretability of molecular analysis compared with numerical prediction results alone. Thus, by successfully linking quantitative predictions with visual chemical reasoning the overall system becomes more interpretable and, therefore, useful for cheminformatics and drug discovery research.

5.5. Discussion of Findings and Practical Implications

Overall, the results from these experiments provide confirmation of SMILES-centred frameworks for predicting and visualising the behaviour of molecules using highly accurate predictions and without loss of interpretability or usability. The results demonstrate that the use of LightGBM provides an effective method for modelling the properties of molecules by combining their physicochemical features with their circular fingerprint descriptors, since it provides low error and a good explanation of the variance when predicting molecular properties from test data.

In relation to other ensemble learning methods used, such as Random Forest and XGBoost, LightGBM appears to be highly competitive. Both Random Forest and XGBoost produce beneficial baselines for predicting properties of molecules, but overall LightGBM has outperformed both of these methods; thus, demonstrating the effectiveness of LightGBM for understanding the complex and nonlinear patterns created within the space of high-dimensional molecular descriptors. Most importantly, this is achieved without requiring the use of heavy deep learning models or 3D descriptors during training.

The molecular visualization component significantly improves the interpretability of the predicted molecular properties. Synchronized structural visualization supports qualitative interpretation of the predicted molecular properties. If you look at the variables that're most important for the molecular predictions you will see that molecular weight and lipophilicity scores are very important. These

are things that chemists know about. Measures of complexity are also important for the predictions. This makes sense because molecular weight and lipophilicity scores are important, in chemistry. Molecular visualization is a help when you are looking at molecular predictions. Displaying molecular structures alongside their predicted properties enables more intuitive interpretation of structure–property relationships. This way we can understand what the predictions mean and see how the structure of something is connected to its properties. Molecular structures and their properties are easier to understand when we can compare them side by side.

The system demonstrates its practical value by serving three specific functions which include exploratory cheminformatics research and initial molecular screening and educational purposes where both precise results and understandable interpretations are essential. The system's modular design enables users to add new property specifications and data sets and machine learning methods which results in a system that can be customized for multiple chemical and pharmaceutical research needs. The combination of interpretable machine learning and molecular visualization creates an effective molecular analysis method which works in real-world situations.

6. Conclusion And Future Work

An approach to molecular property prediction was described, which includes a combination of a SMILES-based representation, descriptor fingerprint feature extraction, interpretable machine learning, and molecular visualization synchronized together into a single pipeline. The combination of physicochemical descriptors, Morgan fingerprint features, and the LightGBM regression model enables accurate molecular property prediction while maintaining computational efficiency and model interpretability. From experimental performance testing of the framework on a large-scale molecular dataset, the results indicate that the framework models complex structure-property relationships while not relying upon expensive 3D computing-based descriptors or on deep learning architectures when constructing the model.

Moreover, another major contribution of the study is the development of a tightly coupled prediction and visualisation framework that enables the analysis of molecular properties concurrently with synchronised 2D and 3D representations of the molecular structure. This coupling aids in enhancing interpretability of predictions by providing a means for the user to visually relate molecular substructures and structural geometrical characteristics to corresponding predicted values. Additionally, conducting feature contribution analysis on the dataset supports the notion that the model based on LightGBM consistently predicts to chemically meaningful descriptors/substructures thus strengthening confidence in the modelled relationships as the basis for predictive modelling of molecular structures. Compared with other ensemble learning approaches evaluated in this study, the LightGBM-based framework demonstrated superior predictive performance and generalization capability, supporting its suitability for scalable cheminformatics applications.

The proposed framework provides multiple paths to future development despite showing successful performance. Future research may consider the use of other molecular representations, larger sets of descriptors, adaptive fingerprints, graph embeddings,

and quantum chemical descriptors that would increase the prediction capabilities of geometry-dependent molecular properties, as well as more cheminformatics uses. The framework needs to be expanded in two ways to support multi-task learning and to include extra physicochemical and biological endpoints which should be studied as future research paths. The system will benefit from advanced interpretability techniques which will help users understand the system better while interactive visualization functions will support their decision-making process. These future extensions are expected to further improve predictive robustness, scalability, and interpretability while preserving the computational efficiency and practical usability of the proposed framework.

7. References

- [1] Pang, C., Tong, H. H., & Wei, L. (2023). Advanced deep learning methods for molecular property prediction. *Quantitative Biology*, 11(4), 395-404. <https://doi.org/10.1002/qub.2.23>
- [2] Zeng, X., Xiang, H., Yu, L., Wang, J., Li, K., Nussinov, R., & Cheng, F. (2022). Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11), 1004-1016. <https://doi.org/10.1038/s42256-022-00557-6>
- [3] Narayanan, R., Peker, E., DeGroat, W., Mendhe, D., Zeeshan, S., & Ahmed, Z. (2025). 3D IntelliGenes: AI/ML application using multi-omics data for biomarker discovery and disease prediction with multi-dimensional visualization. *BMC medical research methodology*, 25(1), 193. <https://doi.org/10.1186/s12874-025-02649-4>
- [4] Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., ... & Wang, H. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2), 127-134. <https://doi.org/10.1038/s42256-021-00438-4>
- [5] Kuang, T., Ren, Y., & Ren, Z. (2024). 3D-Mol: a novel contrastive learning framework for molecular property prediction with 3D information. *Pattern Analysis and Applications*, 27(3), 71. <https://doi.org/10.1007/s10044-024-01287-8>
- [6] Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., ... & Ke, G. (2023, May). Uni-mol: A Universal 3d Molecular Representation Learning Framework. In *The eleventh international conference on learning representations*. <https://doi.org/10.26434/chemrxiv-2022-ijm0j-v4>
- [7] Lu, S., Gao, Z., He, D., Zhang, L., & Ke, G. (2024). Data-driven quantum chemical property prediction leveraging 3d conformations with uni-mol+. *Nature communications*, 15(1), 7104. <https://doi.org/10.1038/s41467-024-51321-w>
- [8] Liyaqat, T., Ahmad, T., & Saxena, C. (2025). Advancements in Molecular Property Prediction: A Survey of Single and Multimodal Approaches: T. Liyaqat et al. *Archives of Computational Methods in Engineering*, 1-31. <https://doi.org/10.1007/s11831-025-10317-5>
- [9] Lin, J., Zheng, Y., Chen, X., Ren, Y., Pu, X., & He, J. (2024, October). Cross-view Contrastive Unification Guides Generative Pretraining for Molecular Property Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 2108-2116). <https://doi.org/10.1145/3664647.3681193>
- [10] Marimuthu, A. N., & McGuire, B. A. (2025). Machine Learning Pipeline for Molecular Property Prediction Using ChemXploreML. *Journal of Chemical Information and Modeling*, 65(11), 5424-5437. <https://doi.org/10.1021/acs.jcim.5c00516>
- [11] Dobbelaere, M. R., Lengyel, I., Stevens, C. V., & Van Geem, K. M. (2024). Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *Journal of Cheminformatics*, 16(1), 99. <https://doi.org/10.1186/s13321-024-00895-0>
- [12] Correia, J., Capela, J., & Rocha, M. (2024). Deepmol: an automated machine and deep learning framework for computational chemistry. *Journal of Cheminformatics*, 16(1), 136. <https://doi.org/10.1186/s13321-024-00937-7>
- [13] Gao, Z., Ji, X., Zhao, G., Wang, H., Zheng, H., Ke, G., & Zhang, L. (2023). Uni-qsar: an auto-ml tool for molecular property prediction. *arXiv preprint arXiv:2304.12239*.
- [14] Sun, J., Cao, Y., Hu, H., & Qi, B. (2025). Equivariant learning leveraging geometric invariances in 3D molecular conformers for accurate prediction of quantum chemical properties. *Scientific Reports*, 15(1), 26969. <https://doi.org/10.1038/s41598-025-09842-x>
- [15] Baran, K., & Kloskowski, A. (2023). Graph neural networks and structural information on ionic liquids: a cheminformatics study on molecular physicochemical property prediction. *The Journal of Physical Chemistry B*, 127(49), 10542-10555. <https://doi.org/10.1021/acs.jpcc.3c05521>
- [16] Zhu, J., Xia, Y., Wu, L., Xie, S., Qin, T., Zhou, W., ... & Liu, T. Y. (2022, August). Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 2626-2636). <https://doi.org/10.1145/3534678.3539368>
- [17] Henle, E. A. (2024). Graph-Based Algorithms and Machine Learning Approaches for Molecular Property Prediction and Material Design.
- [18] Lu, X., Xie, L., Xu, L., Mao, R., Chang, S., & Xu, X. (2023). Integrating Chemical Language and Molecular Graph in Multimodal Fused Deep Learning for Drug Property Prediction. *arXiv preprint arXiv:2312.17495*.
- [19] Atz, K., Grisoni, F., & Schneider, G. (2021). Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12), 1023-1032. <https://doi.org/10.1038/s42256-021-00418-8>
- [20] Sheshanarayana, R., & You, F. (2025). Molecular representation learning: cross-domain foundations and Future Frontiers. *Digital Discovery*, 4(9), 2298-2335. <https://doi.org/10.1039/D5DD000170F>
- [21] Luong, K. D., & Singh, A. (2024). Application of transformers in cheminformatics. *Journal of Chemical Information and Modeling*, 64(11), 4392-4409. <https://doi.org/10.1021/acs.jcim.3c02070>
- [22] Sifain, A. E., Rice, B. M., Yalkowsky, S. H., & Barnes, B. C. (2021). Machine learning transition temperatures from 2D

- structure. *Journal of Molecular Graphics and Modelling*, 105, 107848. <https://doi.org/10.1016/j.jmglm.2021.107848>
- [23] Lou, X., Cai, J., & Siu, S. W. (2025). Multi-MoleScale: a multi-scale approach for molecular property prediction with graph contrastive and sequence learning. *Journal of Cheminformatics*, 18, 5. <https://doi.org/10.1186/s13321-025-01126-w>
- [24] Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., ... & Pande, V. S. (2018). PotentialNet for molecular property prediction. *ACS central science*, 4(11), 1520-1530. <https://doi.org/10.1021/acscentsci.8b00507>
- [25] Le, H. V., Ren, W., Kim, J., Yun, Y., Park, Y. B., Kim, Y. J., ... & Choi, J. M. (2025). CaliciBoost: Performance-driven evaluation of molecular representations for caco-2 permeability prediction. *Journal of Cheminformatics*. <https://doi.org/10.1186/s13321-025-01137-7>
- [26] Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., ... & Chen, A. F. (2015). ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of cheminformatics*, 7(1), 60. <https://doi.org/10.1186/s13321-015-0109-z>
- [27] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8), 3370-3388. <https://doi.org/10.1021/acs.jcim.9b00237>
- [28] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2), 513-530. <https://doi.org/10.1039/C7SC02664A>
- [29] W. Jin, C. Coley, R. Barzilay, and T. Jaakkola, "Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network", <https://doi.org/10.48550/arXiv.1709.04555>
- [30] "Hierarchical graph representation learning for molecular property prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2126-2137, 2021.
- [31] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "QSAR modeling using machine learning methods: A review," *Engineering Applications of Artificial Intelligence*, vol. 98, 104122, 2021.