Artificial Intelligence in Molecular Physics

[Vol-01, Issue-02, July 2026]
DOI: XX
ISSN: XXXX-XXXX

Artificial Intelligence in Molecular Property Prediction: Advances in Machine Learning, Molecular Representations and Explainable Modelling

Alec Lamens

Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany

Received: 06/06/2026 **Revised:** 12/07/2026

Accepted: 20/07/2026 **Published:** 27/07/2026

Abstract— *Artificial intelligence has proven to be a significant way to predict molecular properties in drug discovery, materials science, toxicology, chemical engineering, and environmental research. This review covers the recent developments of machine learning techniques, molecular representations and explainable modelling for molecular property prediction. Structured data and interpretable analysis continue to benefit from the use of traditional algorithms such as linear regression, random forest, support vector machines, decision trees, and gradient boosting. But deep learning architectures, graph neural networks, graph convolutional networks, message-passing models, graph attention networks, transformers, and foundation models have greatly enhanced the learning of complex structure-property relationships. Major molecular representation approaches, such as descriptors, fingerprints, molecular graphs, 3-D structures and learned multimodal embeddings, are also analysed in the review. Explainable AI from Integrated Gradients, attention mechanisms, and saliency mapping are explained in detail, creating transparency and chemically meaningful interpretation. Also, public benchmark datasets and evaluation metrics for regression and classification are discussed. Although significant advances have been made, data quality and model generalization, data interpretability, computational burden, uncertainty, and lack of reproducibility are still limitations. Future developments are anticipated to be focused on physics-informed models, transferable foundation architectures, standardized benchmarking, and better integration with experimental validation, while multimodal learning is expected to be a key focus.*

Keywords— ***Artificial intelligence; Molecular property prediction; Machine learning; Molecular representations; Graph neural networks; Explainable artificial intelligence***

1. INTRODUCTION

The prediction of molecular properties is now a fundamental and indispensable part of many of the most current fields of chemistry, materials science, pharmaceutical research, and chemical engineering, where the molecules' structural, physical, chemical and biological properties can be quickly estimated before experimental testing. Key properties of interest, such as solubility, toxicity, bioactivity, binding affinity, electronic structure and thermodynamic stability, are crucial for the identification of promising compounds and reduce the amount of time and money spent in the lab required for their synthesis and testing. While traditional computational methods, such as quantum mechanical calculations and molecular simulations, have been very useful in understanding how molecules behave, they can be extremely computationally intensive and difficult to apply to the large chemical space of billions of potential molecular structures (Ananikov, 2024). As a result, there has been a marked need for scalable computational methods that can efficiently compute molecular properties in a timely fashion. In recent years, the field of molecular property prediction has seen a revolution with the emergence and advancement of artificial intelligence (AI) and machine learning (ML) technologies. The development and evolution of artificial intelligence (AI) and machine learning (ML) technologies have revolutionized the way molecular properties are predicted, enhancing prediction accuracy, computational efficiency, and generalizability by learning complex relationships directly from vast molecular datasets (Damewood et al., 2023; Ding et al., 2026).

Application of artificial intelligence in chemistry has shifted from basic statistical modelling and quantitative structure–activity relationship (QSAR) techniques to more advanced deep learning techniques that can learn complex molecular representations. Early machine learning techniques used handcrafted molecular descriptors and fingerprints, which necessitated a large amount of feature engineering to describe the molecular structure. These methods were successful but were often limited in their predictive power because of the quality and completeness of the features they relied on, which were manually designed. However, the advent of deep learning has revolutionized this approach, facilitating automated extraction of molecular features from raw data and eliminating the need for manual feature definitions, which enhances model robustness and reduces reliance on expert knowledge (Corso et al., 2024). Recently, molecular modelling has been further enhanced by the development of graph neural networks (GNNs), transformer architectures, geometric deep learning and multimodal learning techniques, which represent molecules as interconnected graph structures that preserve atomic and bonding relationships, and which incorporate multiple types of chemical information. Alongside this, AI has become an essential tool in modern chemical research, with the constant improvement of AI-powered chemistry, leading to the emergence of increasingly sophisticated computational workflows that combine predictive modelling with automated molecular design, optimization and discovery (Sanchez-Lengeling & Aspuru-Guzik, 2018).

AI has been a gamechanger in the field of molecular discovery, making the process much faster and more efficient by reducing the time-consuming iterative molecular design and screening, optimization and validation process. AI models can quickly process thousands to millions of candidate molecules, rank those that are most likely to have a desirable function and uncover unexpected connections between structure and properties. Rather than relying only on costly laboratory experiments or computationally intensive simulations, AI models can rapidly analyze thousands to millions of candidate molecules, prioritize molecules with the greatest potential for a desired function, and identify unexpected relationships between molecular structure and functional properties. In the field of drug discovery, AI's ability to aid in virtual screening and molecular optimization has been especially useful, helping to identify new drug candidates in significantly reduced experimental timelines compared to traditional methods. One notable example of this capability is the recent use of deep learning to discover new antibiotics that are resistant to antimicrobial resistance, which has demonstrated the relevance of AI to biomedical problems in the real world (Stokes et al., 2020). In addition to pharmaceutical research, molecular property prediction via AI is now being used in materials discovery, catalysis, environmental chemistry, polymer science, and energy storage, where predictions of molecular properties are essential for creating high-performing materials and sustainable chemical processes. The integration of structural, physicochemical, and contextual information in a single framework, as achieved in recent advances of molecular representation learning, graph-based neural architectures, and multimodal fusion, has further boosted prediction accuracy. Due to this, AI has been progressing from a prediction tool to an intelligent decision support system that can inform experimental design, decrease discovery expenses, and enhance scientific productivity in various areas of molecular science (Wang et al., 2026; Wieder et al., 2020).

The objective of this review is to give an overview of the latest developments in the field of artificial intelligence for molecular property prediction, focusing on the advances in machine learning methodologies, more recent molecular representation methods and explainable modelling methods. The review covers the following key areas of improvement traditional machine learning algorithms, deep learning architectures, graph neural networks, transformer-based models, and emerging multimodal frameworks, and systematically discusses them. It also explores the development of molecular representation from simple descriptors and fingerprints to graph-based, 3D and learned embedding methods, which allow for better characterization of complex structures. Moreover, the review investigates the increasing significance of explainable AI (xAI) for improving model transparency, interpretability, and trustworthiness, which promotes the successful implementation

of AI models in scientific studies and industrial use.

2. FUNDAMENTALS OF MOLECULAR PROPERTY PREDICTION

2.1 Definition of Molecular Properties

Molecular properties are measurable characteristics related to the structural, physicochemical, electrical and biological behavior of molecules. These properties are crucial in the interaction with its surroundings and explain why molecules are suitable for use in drug discovery, materials science, catalysis and environmental chemistry. Predicting molecular properties accurately can help researchers to assess candidate compounds before experimental testing, which decreases development time and cost. With the advances of machine learning methods, the prediction of these properties has been further improved by learning complex structure–property relationships from a large molecular dataset (Wu et al., 2018; Yang et al., 2019).

2.2 Categories of Molecular Properties

The properties of molecules are often put into groups according to the information that they represent. These are physical properties such as molecular weight, density, boiling and melting point, solubility, lipophilicity, and refractive index, which affect molecular stability and processability. Chemical properties include intermolecular interactions, stability, polarity, basicity or acidity (pKa values) and molecular reactivity and govern chemical transformations. Electronic properties such as dipole moment, molecular orbital energies, charge distribution, electron affinity, ionization potential and the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are important in the comprehension of molecular properties in chemical and materials applications. Biological properties include bioactivity, toxicity, pharmacokinetics, ADMET properties, protein binding affinity and therapeutic efficacy, among others, all of which are of paramount significance in pharmaceutical research. Quantum mechanical properties encompass the total energy, formation energy, atomization energy, vibrational frequencies and conformational energies of molecules that give fundamental insight into molecular structure and stability (Brown et al., 2019; Damewood et al., 2023).

2.3 Conventional Computational Approaches

In the pre-Artificial Intelligence era, the methods used for molecular property prediction were limited to experimental measurements, quantum mechanical calculations, molecular mechanics, molecular dynamics simulations, density functional theory (DFT) and quantitative structure–activity or property relationship (QSAR/QSPR) models (Wieder et al., 2020). While significant progress has been made in computational chemistry with these methods, some of them are extremely expensive to compute or rely heavily on handcrafted molecular descriptors. They are not as good at analyzing large chemical libraries or highly diverse molecular structures, which makes them less scalable for use in modern high-throughput discovery (Wu et al., 2023; Yang et al., 2019).

2.4 Transition from Classical Methods to Artificial Intelligence

The recent evolution of molecular databases, computational power, and algorithms based on data has driven the shift from traditional computational approaches to AI-driven ones. Descriptive, graph-based, SMILES-based, and multimodal molecular representations can be learned directly from them using machine learning and deep learning models, which eliminates the need for intensive manual feature engineering. In response, AI-driven solutions have shown promise in achieving high prediction accuracy, scalability, and generalizability across a wide range of molecular prediction tasks, which could be vital for the advancement of computational chemistry, drug discovery, and materials design in the modern era (Thameem et al., 2026; Wang et al., 2026).

3. MACHINE LEARNING APPROACHES FOR MOLECULAR PROPERTY PREDICTION

3.1 Traditional Machine Learning Algorithms

Traditional machine learning (ML) algorithms laid the groundwork for molecular property prediction by learning between molecular descriptors and molecular properties. Linear Regression, Random Forest, Support Vector Machine (SVM), Decision Trees, and Gradient Boosting have been widely used to predict physicochemical, biological and pharmacological properties. They are efficient, easy to interpret, and effective on well-structured data sets, but they are usually highly dependent on manually engineered molecular features and are often poor at representing complex non-linear relationships within large, diverse chemical spaces (Brown et al., 2019; Yang et al., 2019). The characteristics, advantages, limitations and common applications of traditional machine learning algorithms for molecular property prediction are summarized in Table 1.

Table 1. Comparison of Major Machine Learning Algorithms for Molecular Property Prediction

Algorithm	Principle	Advantages	Limitations	Common Applications	Reference(s)
Linear Regression	Linear relationship modelling	Simple and interpretable	Poor for nonlinear data	QSAR, regression	Yang et al. (2019)
Random Forest	Ensemble decision trees	High accuracy, robust	Less interpretable	Toxicity, ADMET	Yang et al. (2019)

Support Vector Machine	Maximum-margin classification/regression	Effective for small datasets	Kernel selection required	Bioactivity prediction	Yang et al. (2019)
Decision Tree	Rule-based partitioning	Easy interpretation	Overfitting	Property classification	Brown et al. (2019)
Gradient Boosting	Sequential ensemble learning	High predictive power	Computationally intensive	Molecular property prediction	Brown et al. (2019)

3.2 Deep Learning Architectures

Recently, deep learning has been used to greatly enhance the ability to predict molecular properties by learning hierarchical molecular representations directly from raw data. ANNs are used to model complex nonlinear relationships, and CNNs are used well to extract structural features of the image at a local level. Recurrent Neural Networks (RNNs) are generally employed for molecular representations that follow a sequence, like SMILES strings, while Autoencoders are used to learn a compact latent representation for learning features and generating molecules from scratch. These architectures have shown high predictive accuracy in many molecular prediction problems in comparison to conventional ML methods (Stokes et al., 2020; Wang et al., 2026).

3.3 Graph-Based Deep Learning

Due to the natural graph representation of molecules, with atoms as nodes and chemical bonds as edges, graph-based deep learning has become the mainstream approach in molecular modelling. Graph Neural Networks (GNNs) learn expressive molecular embeddings, which are the aggregation of the neighborhood information, and Graph Convolutional Networks (GCNs) learn local structural patterns as graph convolution operations (Corso et al., 2024; Cremer et al., 2023). MPNNs propagate information between atoms, which is iterated to model molecular interactions, and Graph Attention Networks (GATs) use attention mechanisms to give more weight to the atoms which are influential. They can capture the molecular topology well and yield state-of-the-art performance for property prediction, toxicity assessment, and even molecular representation learning (Fang et al., 2022; Xia et al., 2022).

3.4 Transformer and Foundation Models for Molecular Prediction

Recent progress has enabled architectures and foundation models to be developed that are able to learn rich molecular representations from large-scale chemical datasets. Transformer-based models use self-attention to capture long-range dependencies within molecules, while foundation models benefit from extensive pre-training to enhance transfer learning and generalization for various molecular prediction tasks (Maziarka et al., 2024). These methods have yielded spectacular results in various molecular representation learning, molecular optimization, and generative molecular design tasks, marking the state-of-the-art in AI-driven molecular modelling (Jiang et al., 2024; Rampášek et al., 2022).

3.5 Comparative Strengths and Limitations of AI Algorithms

The benefits of each AI algorithm vary from one prediction task to another and are based on the available data. Manual features engineering of descriptors limits the applicability of traditional machine learning approaches to the specific tasks for which they are designed and can be hard to interpret (Muneer et al., 2026). Deep learning models automatically learn complex nonlinear features and, in general, possess higher predictive accuracy, but need greater number of data sets and computationally intensive processing. The graph-based neural networks offer highly expressive molecular representations by maintaining the structure of molecules and the transformer and foundation models have been shown to be highly scalable and transferable but require extensive pre-training and computational infrastructure. Thus, the choice of the algorithm should match the amount of data, model interpretability, computational expense, and complexity of the targeted molecular prediction problem. Artificial Intelligence models for molecular properties prediction are shown in Figure 1, from the very early days of machine learning algorithms to deep learning, graph neural networks, transformer architectures and foundation models.

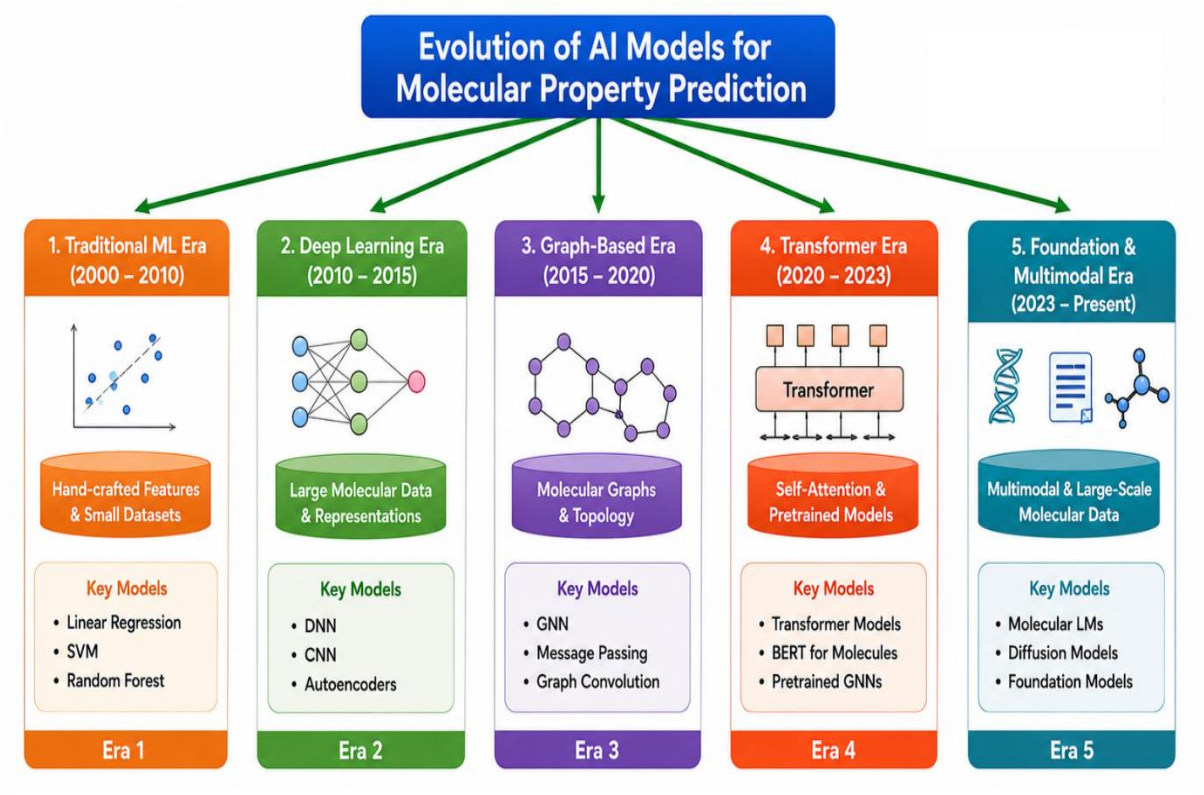

Figure 1. Evolution of AI models for molecular property prediction

4. Molecular Representations in Artificial Intelligence

4.1 Molecular Descriptors

The molecular descriptors are numerical features that quantitatively describe the structural, physicochemical, and topological properties of molecules. They feed machine learning models with informative variables that describe the molecular composition, size, polarity and connectivity, which form the basis for traditional molecular property prediction methods (Pocha et al., 2021; Wang et al., 2019).

4.2 Molecular Fingerprints

Molecular fingerprints are a representation of molecular structure, stored in fixed-length binary or numeric vectors that are used for molecular comparison and prediction. The most popular fingerprinting approaches are MACCS keys, which correspond to defined structural fragments; Morgan fingerprints, which reflect circular atomic neighbourhoods; and Extended Connectivity Fingerprints (ECFPs), which can capture local structural environments effectively. They are easily computed and are extensively used in similarity search, virtual screening, and QSAR modelling (Pocha et al., 2021; Wang et al., 2019).

4.3 String-Based Representations

String-based molecular representation: Chemical structures are represented as string sequences. It is essential to note that SMILES (Simplified Molecular Input Line Entry System) is the most widely used representation due to its simplicity and compatibility with deep learning models. SELFIES (Self-Referencing Embedded Strings) is an extended version of SMILES with the additional capability to check for syntactic validity, especially useful for molecular generation and transformer-based learning systems. Recently, large-scale pre-training (Ross et al., 2024; Xiong et al., 2026) has improved the effectiveness of string-based representations in chemical language models.

4.4 Graph-Based Molecular Representations

Graph-based representations represent molecules as graphs, atoms as nodes and chemical bonds as edges. This representation naturally maintains the molecular topology and allows graph neural networks to learn useful structural relationships directly from molecular graphs. The representation of graphs is the method that has received significant improvements in order to better represent complex molecular properties, acquiring both local and global structural

information (Fang et al., 2022; Xia et al., 2022).

4.5 Three-Dimensional Molecular Representations

Three-dimensional (3D) molecular representations include atomic coordinates, bond geometry, and 3D conformation of the molecule, which can be hidden in 2D representations. These models, which are geometry-aware, can explicitly account for the shape of molecules and their spatial relationships, thus enhancing the prediction of molecular interactions, binding affinity, and quantum mechanical properties (Townshend, 2021; Stärk et al., 2022).

4.6 Learned Molecular Embeddings

Learned Molecular Embedding: Vector representation obtained by deep learning model through self-supervised or multimodal learning. These embeddings are designed to embed structural, sequential, and geometric information into a compact feature space that enables accurate prediction for a variety of molecular tasks. They are very important for modern molecular property prediction due to their adaptability and transferability (Han et al., 2026; Xiong et al., 2026).

4.7 Comparative Analysis of Molecular Representation Techniques

The different molecular representations used have specific strengths for particular prediction tasks. Molecular descriptors and fingerprints are computationally efficient, suitable for conventional machine learning models, but require features to be manually designed. While string-based representations work well with language models and generative AI, graph-based representations maintain molecular topology and tend to be more predictive in general. Three-dimensional embedding and learned embedding approaches offer more structure and better generalization, but are more computationally intensive and need larger training sets. Therefore, the selection of molecular representation needs to be determined by the application, availability of data, and modelling goals. Table 2 shows a comparison of commonly used molecular representation techniques.

Table 2. Molecular Representation Techniques Used in Artificial Intelligence

Representation	Information Captured	Advantages	Limitations	Typical AI Models	Reference(s)
Molecular Descriptors	Physicochemical properties	Simple	Manual feature engineering	ML	Yang et al. (2019)
MACCS	Structural fragments	Fast similarity search	Limited information	ML	Wang et al. (2019)
Morgan/ECFP	Circular atom environments	Highly informative	Ignores 3D geometry	ML, DL	Wang et al. (2019)
SMILES	Sequential representation	Compatible with Transformers	Non-unique syntax	RNN, Transformer	Ross et al. (2024)
SELFIES	Robust molecular strings	Always valid	Longer sequences	Language Models	Ross et al. (2024)
Molecular Graph	Topology	Preserves connectivity	Higher computational cost	GNN	Fang et al. (2022)
3D Representation	Spatial geometry	Better structural accuracy	Requires conformers	Geometric DL	Stärk et al. (2022)
Learned Embeddings	Latent molecular features	Excellent generalization	Data intensive	Foundation Models	Han et al. (2026)

5. Explainable Artificial Intelligence for Molecular Property Prediction

5.1 Importance of Explainability in Molecular AI

The ability to explain in molecular AI is a key part, since predictions need to be very good without them knowing how! Chemists, pharmaceutical companies developing drugs and biologists working in the field of toxicology must be aware of the features of the molecule that affect the prediction before using the model for experimental or regulatory purposes. Thereby, explainable AI (XAI) contributes to the scientific interpretability of AI, aids in generating hypotheses and contributes to the identification of spurious correlations or unreliable model behaviour (Proietti et al., 2024; Rao et al., 2022).

5.2 Global and Local Explainability

Global explainability refers to the overall decision-making

behaviour of the model over the entire dataset, while local explainability explains why a specific molecule is predicted. Global methods show general structure–property relationships, and local methods show the atoms, bonds, fragments or descriptors that affect an individual outcome. Both levels are crucial to assess the molecular model, as well as to assess if the model is chemically meaningful (Corso et al., 2024; Wang et al., 2023).

5.3 Feature Attribution Methods

The feature attribution methods estimate the contribution of individual molecular features to model predictions. SHAP can give both global and local explanations and is based on cooperative game theory. LIME, which creates interpretable local approximations around each prediction, and Integrated Gradients, which quantifies the effect of a given input feature on the output of a neural network based on a baseline. They can identify key descriptors, atomic environments, or molecular fragments, though the interpretations can differ across model architectures and feature representations (Proietti et al., 2024; Rao et al., 2022). The major explainable AI techniques and their use in molecular property prediction are summarized in Table 3.

Table 3. Explainable Artificial Intelligence Methods for Molecular Property Prediction

XAI Method	Explanation Type	Strength	Limitation	Typical Application	Reference(s)
SHAP	Global & Local	High interpretability	Computational cost	Feature importance	Proietti et al. (2024)
LIME	Local	Easy visualization	Approximation error	Local explanations	Rao et al. (2022)
Integrated Gradients	Local	Suitable for DL	Baseline dependent	Neural networks	Rao et al. (2022)
GNNExplainer	Graph	Graph interpretation	Computationally intensive	Graph neural networks	Corso et al. (2024)
Attention Maps	Graph	Visual importance	Attention ≠ explanation	GAT models	Wang et al. (2023)
Saliency Maps	Graph	Atom-level visualization	Sensitive to noise	Molecular graphs	Cremer et al. (2023)

5.4 Explainability in Graph Neural Networks

Graph neural networks have to be interpreted with care, as their predictions involve intricate interactions of atoms and bonds. GNNExplainer identifies the influential subgraphs and node features, and attention-based methods give importance weights to individual atoms or interactions in a molecule. Saliency maps represent the sensitivity of prediction to variations in the structure of the graph. These methods provide insight into structural motifs linked to toxicity and bioactivity, as well as physicochemical properties, but not all attention scores are fully reliable causal explanations (Corso et al., 2024; Cremer et al., 2023).

5.5 Model Transparency and Trustworthiness

Explanations in transparent molecular models should be stable, repeatable, chemically plausible and consistent with known biology and chemistry. Uncertainty estimation, robustness testing, bias detection and validation in chemical space are also essential components of trustworthiness (Wodrich et al., 2026). Confident predictions can be backed by unstable explanations, and these are considered separate from the predictive/explanatory uncertainties. For the reasons mentioned above, the need for strong interpretability frameworks becomes even greater in the case of scale and architecture complexity of emerging foundation and multimodal models (Segura-Alabart & Serraosa, 2026; Song et al., 2026).

5.6 Case Studies of Explainable Molecular Prediction

The applications of XAI in molecular prediction have proven helpful in the field of toxicity prediction, drug discovery, structure–property relationship, etc. There are methods of connecting predictions to chemically meaningful concepts (self-interpretable graph models); and methods of identifying molecular fragments that are associated with toxic or biologically active behaviour (attribution methods and graph-explanation methods). The case studies demonstrate that explainability can enhance model validation, drive molecular modification, and facilitate interactions between computational

scientists and experimental researchers. However, explanations must be backed up by chemical knowledge and observations of experiments and should not be used as proof of cause (Proietti et al., 2024; Wodrich et al., 2026). The full EAI loop for molecular property prediction is summarized in Figure 2 with input molecules, molecular representation learning, prediction of the molecular property, generation of explanations, expert interpretation of the explanations, and iterative model refinement.

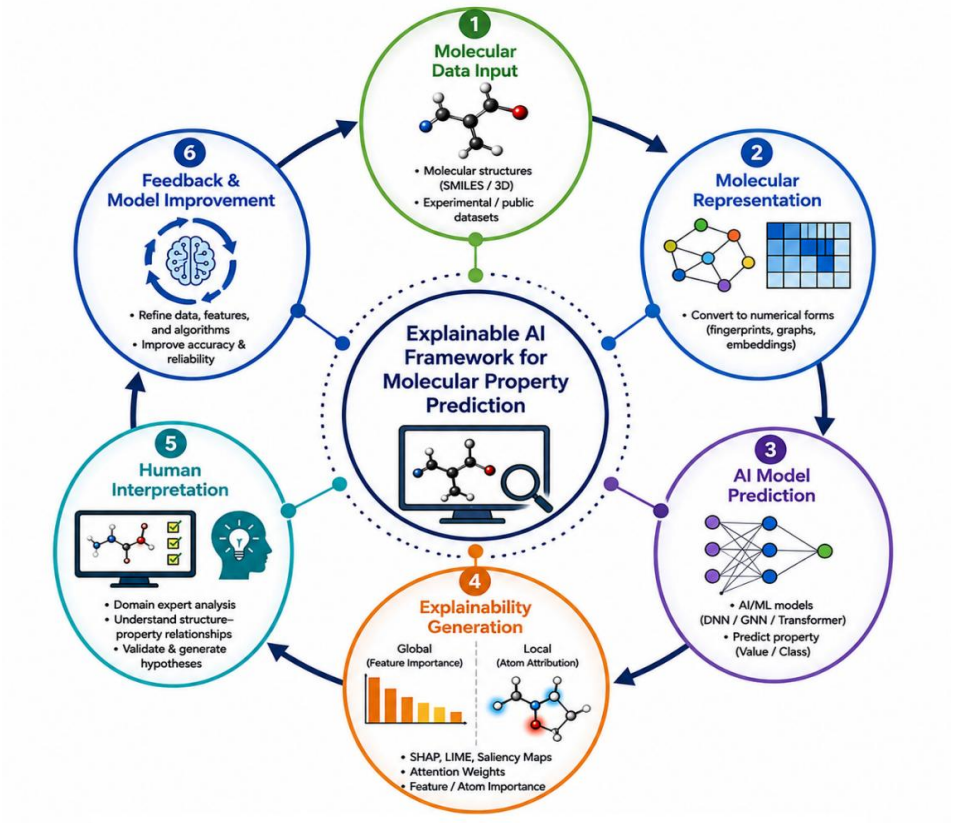

Figure 2. Explainable AI framework for molecular property prediction

6. Applications of AI-Based Molecular Property Prediction

6.1 Drug Discovery and Development

The use of AI for molecular property prediction is now a key component in contemporary drug discovery, playing a vital role in virtual screening, lead identification, target prioritization, and candidate optimization. Machine learning models can quickly screen vast chemical libraries to find molecules with desired biological activity, binding potential and physicochemical properties. In addition, deep learning has also proven to be useful in medicine, specifically in the discovery and optimization of promising therapeutic compounds, as shown in various studies, such as in the field of antibiotic discovery (Brown et al., 2019) and chemical language modelling (Ross et al., 2024; Stokes et al., 2020).

6.2 Materials Discovery

AI models can predict materials properties like thermal stability, conductivity, mechanical strength, band gaps and catalytic activity in material science. These properties, in turn, lower reliance on costly experimental screening and aid in quicker identification of battery, semiconductor, polymer and energy materials. The ability to represent material structure and composition is key to capturing structural and compositional relationships between a wide variety of material classes (Damewood et al., 2023).

6.3 Toxicity Prediction

AI-based toxicity prediction helps identify potentially harmful molecules early in the drug development process before lab or clinical trial testing. Using graph-based and equivariant neural networks to learn structural patterns linked to toxic responses can enhance safety evaluation and minimize the use of animal experiments. This type of model is used more and more to predict the toxicity of an organ, the mutagenicity, the carcinogenicity and the environmental hazards of a substance (Cremer et al., 2023).

(Costamagna et al., 2025; Guo et al., 2026).

6.4 ADMET Property Prediction

In order to assess the pharmacokinetic suitability of a candidate molecule, it is necessary to predict the absorption, distribution, metabolism, excretion and toxicity properties of the molecule. The molecular structure can be used to estimate solubility, permeability, metabolic stability, bioavailability and clearance with AI models. Molecular graphs and descriptors, along with information obtained from SMILES, are used in conjunction with each other to further enhance the prediction of ADMET properties in integrated pipelines and multimodal approaches (Han et al., 2026; & McGuire, 2025).

6.5 Chemical Reaction Prediction

Reaction outcome prediction, retrosynthetic planning, reagent selection, and reaction-condition optimization are all emerging new applications of AI systems. Chemical language models are able to understand reactions as structured sequences, and to learn transformations from vast databases of reactions. These approaches facilitate quicker synthesis planning and enhance the efficiency of experimental chemistry by selecting viable pathways and potential products, thereby reducing time and effort (Ross et al., 2024).

6.6 Molecular Design and Optimization

Generative models and optimization algorithms allow the design of molecules that have defined properties such as potency, stability, solubility and selectivity. While the molecular similarity methods are useful for identifying compounds with similar structures, domain-specialised language models can inform the optimisation of candidate molecules from a smaller set of examples. These methods enable traversing chemical space and the design of molecules which meet multiple design features simultaneously

6.7 Environmental and Green Chemistry Applications

Finally, AI is also critical for environmental and green chemistry, with the ability to predict biodegradability, persistence, ecotoxicity and environmental mobility of properties. These predictions are used to find hazardous compounds, to design safer alternatives and to save on resource-intensive experimental testing. AI can aid sustainable chemical design, pollution prevention, and the development of environmentally responsible materials by leveraging multimodal molecular representations and efficient predictive pipelines (Han et al., 2026; Marimuthu & McGuire, 2025).

7. Benchmark Datasets and Performance Evaluation

7.1 Public Molecular Datasets

Benchmark datasets are crucial for the creation, comparison, and validation of AI models that predict molecular properties in a public context. QM7, QM8 and QM9 also contain quantum mechanical properties and have been used extensively for regression problems involving molecular energies, electronic structure, and other physicochemical properties. MoleculeNet is a standardized library of datasets in the areas of quantum chemistry, physical chemistry, biophysics, and physiology. In this context, ESOL, FreeSolv and Lipophilicity predict solubility, hydration free energy and lipophilicity, respectively, while Tox21, ClinTox and BBBP are frequently used to predict toxicity, clinical safety and blood–brain barrier permeability, respectively. In addition to these resources, GuacaMol offers benchmarks for molecular generation and goal-directed design (Brown et al., 2019; Wu et al., 2018). Some common standard sets of data that are used to benchmark molecular property prediction are summarized in Table 4.

Table 4. Benchmark Datasets Used for Molecular Property Prediction

Dataset	Primary Task	Property Type	Molecules	Common AI Models	Reference(s)
QM7	Quantum prediction	Atomization energy	~7,000	GNN	Wu et al. (2018)
QM8	Electronic properties	Quantum chemistry	~22,000	GNN	Wu et al. (2018)
QM9	Quantum prediction	Multiple quantum properties	~134,000	GNN, Transformer	Wu et al. (2018)
ESOL	Solubility	Regression	1,128	ML, DL	Wu et al. (2018)

FreeSolv	Hydration free energy	Regression	642	GNN	Wu et al. (2018)
Lipophilicity	LogD	Regression	4,200+	Transformer	Wu et al. (2018)
BBBP	Blood-brain barrier	Classification	2,039	GNN	Wu et al. (2018)
ClinTox	Clinical toxicity	Classification	1,478	DL	Wu et al. (2018)
Tox21	Toxicity	Multi-label classification	~8,000	DL	Wu et al. (2018)

7.2 Data Preprocessing and Feature Engineering

The quality of the model performance is highly sensitive to the quality of data and feature preprocessing. Major steps in this include molecular standardization, duplicate removal, treatment of missing data, class balancing, calculation of descriptors, generation of fingerprints, graph construction, and splitting of the dataset. Scaffold-based splitting is preferred to random splitting, as it yields more realistic values of the model generalization to a chemical structure that was not used during training. In the case of deep-learning models, molecular graphs, SMILES sequences and learned embeddings can be used to decrease the reliance on manually engineered features while retaining the chemically relevant information, (Ding et al., 2026; Yang et al., 2019).

7.3 Evaluation Metrics

The choice of evaluation metrics depends on the type of prediction task: continuous vs categorical. The mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) are commonly used to quantify the accuracy of predictions and the amount of variance explained. The ranking measures used for classification, such as ROC-AUC and PR-AUC, and the classification quality measures used, such as accuracy, precision, recall, and F1-score, provide a means for comparing ranking performance and classification quality. In imbalanced datasets, PR-AUC is especially informative since it places a focus on the performance of the minority class. Multiple metrics allow for a more comprehensive and accurate evaluation of model behaviour (Pocha et al., 2021).

7.4 Comparative Benchmarking of AI Models

Models should be compared on the same datasets, splits, preprocessing methods, and hyperparameters to create a basis for comparative benchmarking. Traditional machine learning models can serve as baseline models, graph neural networks, graph transformers, self-supervised models, and foundation models generally have better representation-learning power (Rampášek et al., 2022). It is important to consider, however, gains in predictive performance in the context of computational cost, robustness, transferability, and reproducibility. The recent work demonstrates that the graph architecture, the representation of atoms, the pre-training strategy and the size of datasets can significantly affect the performance and can make the results from different methods difficult to compare (Song et al., 2026; Wang et al., 2023). Reporting of benchmarking results should, therefore, be presented clearly and in a way that allows for comparison with existing research, to gauge if the improvements are methodologically significant or dataset-specific (Ding et al., 2026; Yang et al., 2019). The entire workflow to create, test, and use molecular property prediction models powered by artificial intelligence is described in Figure 3.

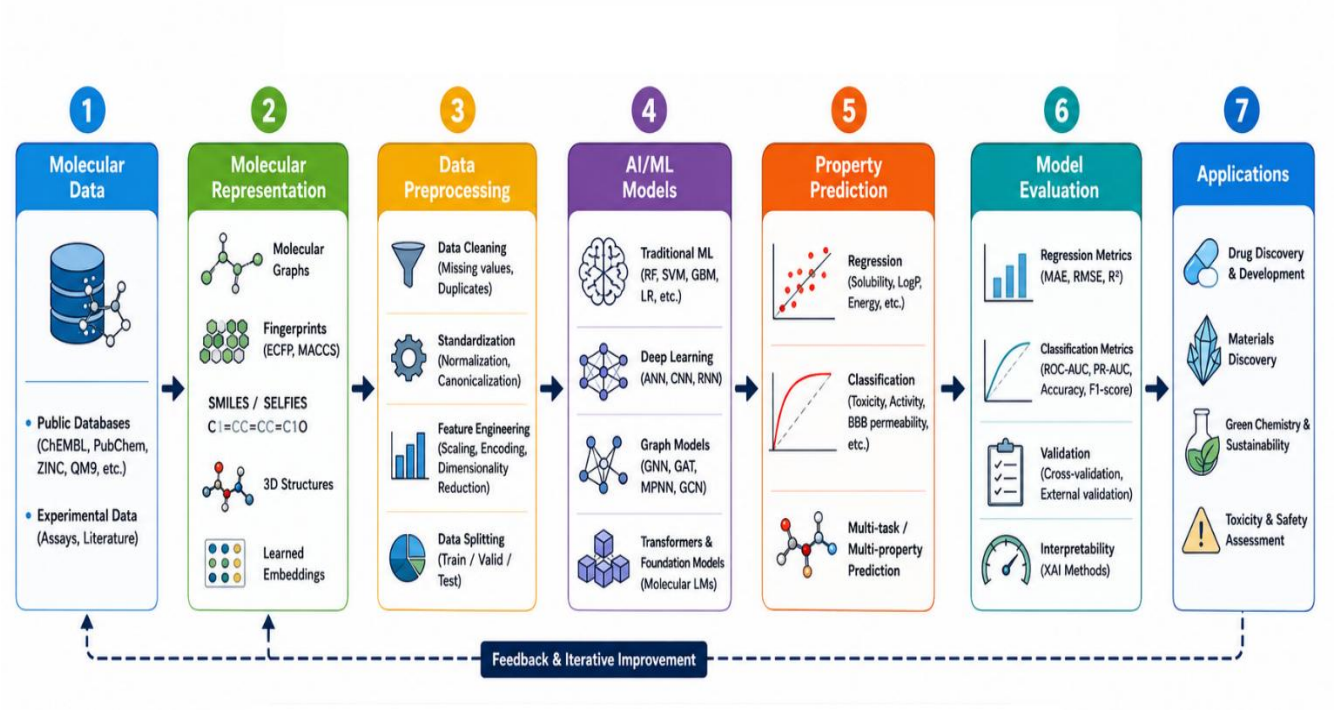

Figure 3. Workflow of AI-based molecular property prediction

8. Challenges and Future Perspectives

While significant strides have been made, AI-driven molecular property prediction still has several critical challenges to address: data quality and data difficulty, generalization of the model, interpretability, computational complexity, and reproducibility. The availability of many molecular datasets is small, skewed, inaccurate, or focused on a particular space of molecules, making the reliability of models for structurally new compounds limited. However, there are difficulties when comparing studies due to different preprocessing pipelines, data splitting and evaluation methods. While deep learning, graph neural networks, transformers, and foundation models have led to better predictive performance, their black-box nature often hinders scientific understanding and regulatory acceptance. Explainable AI methods are able to find influential atoms, bonds or fragments of a molecule, but explanations are not always stable, chemically meaningful or causally valid. These results should thus focus future research on transparent modelling, uncertainty quantification, external validation and standardization of benchmarking. Additionally, more focus should be placed on the three-dimensional shape of molecules, stereochemistry, conformational changes and dynamics, and the nature of molecular interactions, which are sometimes underrepresented in existing models. Multi-modal learning combining molecular graphics, SMILES, 3D structures, spectra, experimental measurement and biological data could lead to more comprehensive and transferable representations. Physics-informed and hybrid models, which integrate data-driven learning and quantum mechanics, may provide for greater accuracy and fewer reliance on an extremely large training set. Few-shot learning, transfer learning and multiproperty prediction in expanded chemical space are anticipated to be enabled by foundation and domain-specific molecular language models. They come with high environmental and other costs, however, and need to be considered, along with being accessible and reproducible.

9. Conclusion

AI has made a significant impact on molecular property prediction, making it more rapid, accurate, and scalable across various fields such as chemistry, drug discovery, materials science, toxicology, and environmental research. The application of traditional machine learning techniques works well with structured datasets and interpretable predictions, while deep learning, graph neural networks, transformers, and foundation models have increased the capacity to learn complex structure property relationships from molecular data. Molecular representation, from classical descriptors and fingerprints to SMILES strings, graphs, 3D structures and learned embeddings, has a strong impact on the quality of prediction. There are advantages and disadvantages in each representation, and their appropriateness for any given property

depends on the data available, as well as the computing needs. Explainable AI is also crucial for enhancing transparency and distinguishing important molecular features, to boost scientific trust in model predictions. There are public benchmark datasets and standardized evaluation metrics that can be used to compare models, but there are still some differences in preprocessing, splitting strategies and reporting that lead to reproduction challenges. The field is moving towards more multimodal, geometry-aware, physics-informed, and transferable AI systems for broader chemical spaces overall. Moving forward, it will be critical to enhance data quality, uncertainty estimates, interpretability, external validation, and experimental integration.

REFERENCES

1. Ananikov, V. P. (2024). Artificial intelligence chemistry. Artificial Intelligence, 2, 100075.
2. Brown, N., Fiscato, M., Segler, M. H., & Vaucher, A. C. (2019). GuacaMol: benchmarking models for de novo molecular design. Journal of chemical information and modeling, 59(3), 1096-1108.
3. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. Nature Reviews Methods Primers, 4(1), 17.
4. Costamagna, M., Foscato, M., Grellscheid, D., & Jensen, V. R. (2025). Hypershape recognition: a general framework for moment-based molecular similarity. Journal of Chemical Information and Modeling, 65(12), 5960-5972.
5. Cremer, J., Medrano Sandonas, L., Tkatchenko, A., Clevert, D. A., & De Fabritiis, G. (2023). Equivariant graph neural networks for toxicity prediction. Chemical Research in Toxicology, 36(10), 1561-1573.
6. Damewood, J., Karaguesian, J., Lunger, J. R., Tan, A. R., Xie, M., Peng, J., & Gómez-Bombarelli, R. (2023). Representations of materials for machine learning. Annual Review of Materials Research, 53(1), 399-426.
7. Ding, H., Hua, P., & Huang, Z. (2026). Survey on recent progress of AI for chemistry: methods, applications, and opportunities. Frontiers of Computer Science, 20(11), 2011358.
8. 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.
9. Guo, Y., Luo, M., Zhang, W., Liu, P., Liu, J., Huang, S., ... & Liu, X. (2026). Few-shot molecular property optimization via a domain-specialized large language model. Chemical Science, 17(10), 4928-4941.
10. Han, Q., Zhou, C., Wang, R., & Ma, Z. (2026). Multimodal Molecular Representation Learning with Graph Neural Networks, Deep & Cross Networks, and SMILES Embeddings. arXiv preprint arXiv:2607.05736.
11. Jiang, J., Ke, L., Chen, L., Dou, B., Zhu, Y., Liu, J., ... & Wei, G. W. (2024). Transformer technology in molecular science. Wiley Interdisciplinary Reviews: Computational Molecular Science, 14(4), e1725.
12. 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.
13. Maziarka, Ł., Majchrowski, D., Danel, T., Gaiński, P., Tabor, J., Podolak, I., ... & Jastrzębski, S. (2024). Relative molecule self-attention transformer. Journal of Cheminformatics, 16(1), 3.
14. Muneer, A., Waqas, M., Saad, M. B., Showkatian, E., Bandyopadhyay, R., Xu, H., ... & Wu, J. (2026). From classical machine learning to emerging foundation models: review on multimodal data integration for cancer research. Artificial Intelligence Review, 59(4), 119.
15. Pocha, A., Danel, T., Podlewska, S., Tabor, J., & Maziarka, Ł. (2021, July). Comparison of atom representations in graph neural networks for molecular property prediction. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
16. Proietti, M., Ragno, A., Rosa, B. L., Ragno, R., & Capobianco, R. (2024). Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. Machine Learning, 113(4), 2013-2044.
17. Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., & Beaini, D. (2022). Recipe for a general, powerful, scalable graph transformer. Advances in Neural Information Processing Systems, 35, 14501-14515.
18. Rao, J., Zheng, S., Lu, Y., & Yang, Y. (2022).

Quantitative evaluation of explainable graph neural networks for molecular property prediction. Patterns, 3(12).

19. Ross, J., Hoffman, S., Belgodere, B., Chenthamarakshan, V., Mroueh, Y., & Das, P. (2024, December). Learning to Optimize Molecules with a Chemical Language Model. In Annual Conference on Neural Information Processing Systems.
20. Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400), 360-365.
21. Segura-Alabart, N., & Serratosa, F. (2026). A survey of models composed of graph neural networks and large language models for molecular science. Bioinformatics, 42(7), btag387.
22. Song, B., Zhang, J., Liu, Y., Liu, Y., Jiang, J., Yuan, S., ... & Liu, Y. (2026). A systematic review of molecular representation learning foundation models. Briefings in Bioinformatics, 27(1), bbaf703.
23. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., & Jaakkola, T. (2022, June). Equibind: Geometric deep learning for drug binding structure prediction. In International conference on machine learning (pp. 20503-20521). PMLR.
24. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.
25. Thameem, M., AlHmoudi, O., Al Salloum, A., Al Darmaki, N., Elkamel, A., & AlHammadi, A. A. (2026). Molecular Property Prediction: Input Types and Information Processing in Machine Learning Models. Results in Engineering, 109241.
26. Townshend, R. J. L. (2021). Geometric Learning of Biomolecular Structure. Stanford University.
27. Wang, C., Huang, L., Li, P., Shan, D., Han, Q., & Mao, J. (2026). Deep learning for molecular property prediction: From molecular representations to multi-modal fusion. Frontiers in Bioinformatics, 6, 1863358.
28. Wang, H., Kaddour, J., Liu, S., Tang, J., Lasenby, J., & Liu, Q. (2023). Evaluating self-supervised learning for molecular graph embeddings. Advances in Neural Information Processing Systems, 36, 68028-68060.
29. Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2019, September). Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics (pp. 429-436).
30. Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. Drug Discovery Today: Technologies, 37, 1-12.
31. Wodrich, M., Feragen, A., & Schmidt, M. N. (2026, January). Predictive and Explanatory Uncertainties in Graph Neural Networks: A Case Study in Molecular Property Prediction. In Northern Lights Deep Learning Conference 2026.
32. Wu, X., Wang, H., Gong, Y., Fan, D., Ding, P., Li, Q., & Qian, Q. (2023). Graph neural networks for molecular and materials representation. Journal of Materials Informatics, 3(2), N-A.
33. 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.
34. Xia, J., Zhu, Y., Du, Y., & Li, S. Z. (2022). Pre-training graph neural networks for molecular representations: retrospect and prospect. In ICML 2022 2nd AI for Science Workshop.
35. Xiong, Z., Wang, Z., Huang, F., Qiu, M., Fang, S., Yang, L., ... & Zhang, W. (2026). Multi-to-uni modal knowledge transfer pre-training for molecular representation learning. Nature Communications, 17(1), 3797.
36. 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