Asian Journal of Water, Environment and Pollution. Vol. 22, No. 3 (2025), pp. 88-103.
doi: 10.36922/AJWEP025070041

ORIGINAL RESEARCH ARTICLE

Environmental applications of molecular graph learning: Graph neural network based prediction of partition coefficients

Pravinkumar M. Sonsare[1]*, Roshni Khedgaonkar[2], Kavita Singh[2], and Pratik Agrawal[3]

[1]Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Ramdeobaba University, Nagpur, Maharashtra, India
[2]Department of Computer Technology, Yeshvantrao Chavhan College of Engineering, Nagpur, Maharashtra, India
[3]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India

(This article belongs to the *Special Issue: Renewable Energy Systems and Strategies in Smart Grids and Smart Cities Development*)

***Corresponding author:** Pravinkumar M. Sonsare (sonsare@gmail.com)

Received: February 12, 2025; 1st revised: February 23, 2025; 2nd revised: March 24, 2025; 3rd revised: April 15, 2025; 4th revised: April 29, 2025; Accepted: April 29, 2025; Published Online: May 29, 2025

Abstract: In cheminformatics, predicting molecular properties is crucial for enhancing material research, toxicity assessment, and drug discovery. This research investigates the use of graph neural networks (GNNs) for predicting molecular properties by examining three different architectures: graph convolutional networks (GCNs), graph isomorphism networks (GINs), and graph attention networks (GATs). Employing molecular graph information, these models are evaluated on the MUTEG dataset and measured against key metrics such as accuracy and area under the receiver operating characteristic curve (AUC). Our experimental findings show that GIN has the highest accuracy at 89.2%, exceeding GCN (87.5%) and GAT (88.3%). GIN also achieves the highest AUC of 0.89, whereas the AUCs of GCN and GAT are 0.84 and 0.86, respectively, indicating GIN's enhanced ability to effectively model graph isomorphisms. We selected GIN for this study because of its proven theoretical and empirical strength in capturing graph-level representations, particularly in domains such as cheminformatics, where molecular structures are naturally modeled as graphs. These results highlight the efficacy of GNNs in predicting molecular properties and position GIN as a favored framework for tasks that demand accurate graph feature extraction. This study further plays a pivotal role in understanding the environmental fate and transport of chemical compounds. We used GIN to identify partition coefficients such as the octanol-water partition coefficient, air-water partition coefficient, and soil–water partition coefficient from the MoleculeNet dataset.

Keywords: Graph neural networks; Molecular property prediction; Cheminformatics; Drug discovery; Structure-property relationships

1. Introduction

Molecular property forecasting is essential to cheminformatics and significantly contributes to various applications, including drug discovery, materials science, and environmental safety evaluation. Precise forecasting of characteristics such as solubility, toxicity, and carcinogenicity can greatly decrease experimental expenses and durations, thereby hastening the advancement of new compounds.[1,2] Predictive modeling presents opportunities and challenges because of the unique graph structure of molecular data, where bonds are depicted as edges and atoms as nodes. Conventional machine learning approaches frequently find it challenging to adequately represent this graph-structured data, highlighting the need for specialized graph learning techniques.

In modeling data that has a graph structure, graph neural networks (GNNs) have emerged as highly efficient tools. They facilitate notable progress in areas such as social networking, transportation, and bioinformatics. In forecasting molecular properties, GNNs have shown exceptional capability by leveraging the inherent relational information present in molecular graphs. GNNs effectively capture the complex interactions among atoms and bonds that influence molecular characteristics by iteratively enhancing node representations through local and global graph structures.

Among the various GNN frameworks, graph convolutional networks (GCNs),[4] graph isomorphism networks (GINs),[5] and graph attention networks (GATs)[6] are notable for their effectiveness and adaptability. GCNs broaden convolutional processes to graph frameworks by gathering data from a node's local surroundings. GINs emphasize identifying graph isomorphisms to guarantee the distinct representation of graph structures. GATs enhance feature aggregation by allocating attention weights to adjacent nodes, enabling the model to focus on the most significant interactions.

The MUTAG dataset[3] serves as a popular benchmark in predicting molecular properties, offering an excellent platform for testing these architectures. It comprises 188 molecular graphs categorized as either carcinogenic or non-carcinogenic, making it appropriate for binary classification tasks. Prior research has utilized different GNNs on this dataset, obtaining top-tier outcomes. For example, Kipf and Welling's GCN showcased the effectiveness of graph convolutions in molecular classification.[4] In the same way, the GIN developed by Xu *et al*. highlighted the significance of identifying graph isomorphisms for precise predictions.[5]

In this study, we conducted a comparative analysis of GCN, GIN, and GAT architectures on the MUTAG dataset to identify the most effective model for molecular property prediction. Our contributions include: (i) a systematic evaluation of GCN, GIN, and GAT on molecular data using metrics such as accuracy and area under the curve (AUC), (ii) insights into the strengths and limitations of each architecture in capturing molecular graph features, and (iii) a comprehensive analysis of the experimental results highlights GIN's exceptional performance and its potential for practical use in cheminformatics.

We further used GIN to find partition coefficients such as octanol–water partition coefficient (K_{o_w}), air–water partition coefficient (K_{a_w}), and soil–water partition coefficient (K_d) from the MoleculeNet dataset to estimate how chemicals behave in the environment, including their solubility, volatility, and degradation pathways. This helps in understanding their movement through air, water, and soil.

The MUTAG dataset provides valuable data for forecasting molecular characteristics using GNN architectures. However, there are concerns regarding the models' scalability and resilience due to their small size and complexity. This research broadens the assessment to more extensive and complex datasets, such as QM9[6] and ZINC[7], to ensure uniform performance across different molecular datasets. Moreover, tests using incomplete and altered data have been conducted to assess the proposed models' robustness in real-world scenarios. These enhancements provide a broader perspective on the appropriateness of GNNs for predicting molecular characteristics.

Quantitative structure-property relationship models have long been used to predict molecular characteristics based on structural attributes.[8] Machine learning models can accurately estimate partition coefficients such as K_{o_w}, K_{a_w}, and K_d using the MoleculeNet dataset.[9] These models employ various methodologies, including GNNs[10,11] and random forest algorithms,[12] to make effective predictions even with minimal experimental data. Furthermore, the integration of multi-fidelity learning[13] with interpretable attribution models[14] enhances both predictive power and model interpretability, leading to more reliable estimations.

The accurate prediction of molecular properties such as partition coefficients is critical for environmental safety assessments, pharmaceutical design, and chemical risk evaluation. These predictions help reduce the time and cost associated with experimental testing, supporting the development of safer and more efficient

compounds. With the increasing availability of curated datasets and advances in deep learning, particularly GNNs, machine learning models now offer a promising alternative to traditional quantitative structure-activity relationship (QSAR) models. This study aims to investigate the applicability of modern GNN architectures to these tasks and assess their reliability across different molecular datasets.

The organization of this paper is outlined as follows: Section 2 explores pertinent studies that illustrate advancements in GNN-based molecular property forecasting. The methodology detailing model architectures, training methods, and dataset preparation is thoroughly explained in Section 3. The findings from the experiment are presented in Section 4, and the discussion of those results is in Section 5. The conclusion and potential future study topics are presented in Section 6.

2. Literature review

Predicting molecular properties has been a persistent challenge in computational chemistry, driven by the necessity to uncover essential molecular characteristics that affect properties such as carcinogenicity, solubility, and toxicity. The prediction of molecular properties has greatly enhanced due to the progress of machine learning methods, especially GNNs. This segment provides a comprehensive review of the current literature regarding molecular property prediction utilizing GNNs, emphasizing their usage, effectiveness, and architectural advancements.

Before the introduction of GNNs, conventional machine learning models such as random forests, support vector machines, and feed-forward neural networks were widely utilized for predicting molecular properties. These models demanded manually created molecular descriptors that encompassed physicochemical properties and structural fingerprints (e.g., extended-connectivity fingerprints). These methods frequently did not succeed in representing the intricate relational framework of molecular graphs. For instance, Duvenaud *et al*. presented a model that directly learned molecular fingerprints from data through deep convolutional architectures, highlighting an early transition from manually crafted features to learnable representations.[15] Nevertheless, these approaches did not possess the capability to fully utilize graph structure, since they were not specifically created for graph-structured information.

In graph-based molecular modeling, molecules are abstracted as graphs where atoms are represented as vertices (nodes) and bonds as edges. This graph structure allows GNNs to propagate information through bonded atoms, capturing both topological and chemical features essential for property prediction. GNNs offer a revolutionary method for predicting molecular properties by allowing for the direct handling of graph-structured information. The GNN model utilizes a recursive framework to disseminate information across graph structures.[16] Kipf and Welling[4] presented GCNs that generalized conventional convolutions to graphs through the aggregation of information from a node's neighboring nodes, utilizing a normalization approach. GCNs emerged as a favored option for semi-supervised learning tasks, including the classification of molecules. Nonetheless, GCNs frequently faced issues with over-smoothing, resulting in node representations becoming indistinguishable throughout distant neighborhoods.

Xu *et al*.[5] introduced GINs, which enhanced the expressiveness of GNNs by utilizing a sum-based aggregation function. This architecture has proven to be as effective as the test for graph isomorphism. It enables the effective differentiation of non-isomorphic graphs.[17] The attention mechanism is a concept in machine learning, especially in neural networks, that allows a model to dynamically focus on the most relevant parts of the input data when generating output. Originally introduced in natural language processing, it has since been successfully applied in various domains, including GNNs, vision, and cheminformatics. GINs have shown better performance on molecular datasets such as MUTAG, surpassing GCNs in tasks that need accurate structural representation. In the aggregation stage, Veličković et al. proposed GATs that assign trainable weights to local nodes using an attention mechanism.[18]

GATs have proven to be especially useful for datasets where the interactions among particular nodes hold greater significance than the overall structure. Several studies improved GNNs by explicitly including edge characteristics that indicate bond types and lengths within molecular graphs. Schütt *et al*. created the SchNet model, which employed continuous-filter convolutional layers for datasets in quantum chemistry.[19]

The MUTAG dataset has been extensively utilized to evaluate GNNs for predicting molecular properties. MUTAG consists of 188 molecular diagrams depicting nitroaromatic substances. Each is categorized as either carcinogenic or non-carcinogenic. The dataset presents a binary classification challenge that assesses a model's capability to accurately capture molecular graph characteristics. GNNs have likewise been utilized on various datasets, including a quantum chemistry

dataset for property regression tasks such as predicting molecular energy (QM9) and a dataset aimed at forecasting the toxic effects of chemical compounds on biological systems (TOX21).[20]

Multi-fidelity learning, which mixes low-fidelity quantum chemical data with high-fidelity experimental data, has a considerable impact on partition coefficient prediction. For example, multi-target learning obtained a root-mean-square error (RMSE) of 0.44 log P units for a dataset containing molecules comparable to the training data, exhibiting superior accuracy over single-task models.[13] The multi-fidelity log P model, which takes a chemical formula as its sole input, is a useful method for estimating K_{ow} without structural knowledge. This model performed similarly to traditional models with a coefficient of determination (R^2) of 0.77 and an RMSE of 0.52, suggesting its usefulness in cases when structural data is lacking.[21] GNNs with adjusted integrated gradients are highly interpretable in forecasting K_{ow}. These models emphasize the significance of certain atoms in the prediction process, guaranteeing precision, consistency, and stability in attribution assignments.[22]

Machine learning models such as multiple linear regression and random forest regression have been successfully calibrated against external test sets, including the SAMPL9 challenge. These models, along with continuum solvation models, give insights into the molecular characteristics that influence partition coefficients, emphasizing their value in computational chemistry.[23] The study focuses on the use of the created model in forecasting K_d values, proving its efficacy in giving site-specific insights that can benefit environmental risk assessments and pesticide management.[24] The study presents a new descriptor,<q-atom-centered symmetry functions> conformation, which includes explicit polarization effects in polar phases and accounts for energetic and entropic importance in non-polar phases by averaging entropy effects based on the Boltzmann distribution of conformations. This technique improves the prediction of the partition coefficient (log P) between polar and non-polar phases, a critical factor in drug and material design. The model was trained using high-dimensional neural networks on a large public dataset (PhysProp) and showed effective log P prediction across three more datasets. It applies to a number of organic molecules, including n-carboxylic acids and diverse organic solvents, which makes it a useful tool for estimating partition coefficients in varied systems.[25]

The K_{ao} is a key parameter that describes the equilibrium partitioning of a compound between octanol and air, serving as an indicator of its volatility and potential for atmospheric transport. A comprehensive dataset of experimental log K_{ao} values for 2,161 compounds was compiled, which covers a wide range of molecular weights and log K_{ao} values. The model's robustness and predictive capability were validated through various statistical methods, which include training and prediction set separation and mutual leave-50%-out validation.[26] The work introduces a new version of the machine-learning algorithm, PARTYsoc version 3, which measures the proportions of centennially stable and active soil organic carbon (SOC) fractions using Rock-Eval (r) thermal analysis. This model improves on the previous version (version 2) using a bigger dataset from 12 sites, including many long-term studies, and leverages support vector machine regression in conjunction with beta regression to provide more accurate predictions. PARTYsoc version 3 attempts to improve the accuracy of SOC stock development simulations in the high-performance engine model by identifying the best stable SOC stock for each site, resulting in improved SOC compartment initialization. The model performs well in both internal validation and leave-one-site-out validation, confirming its ability to forecast stable SOC proportions.[27]

The K_{ow} represents the ratio of a compound's concentration in octanol to its concentration in water at equilibrium. It is a key descriptor of hydrophobicity and is widely used in environmental fate modeling and bioaccumulation studies. The research focuses on predicting the K_{ow} of organic compounds using extreme learning machine (ELM) models, which are useful owing to their quick learning speed and strong generalization ability. The study uses COSMO descriptors (Sσ-profile) as molecular descriptors to develop and estimate models for K_{ow}. Four ELM models were created and compared to multiple linear regression models with the same descriptors. The results indicated that the ELM models, particularly the ELM-4 model, demonstrate high reliability in predicting log K_{ow} values, which achieves a correlation coefficient R^2 of 0.949 and an RMSE of 0.358, indicating their effectiveness for broader applications in predicting chemical properties.[20] The research examines the partitioning behavior of anionic perfluoroalkyl carboxylic acids (PFCAs) and perfluoroalkyl sulfonic acids (PFSAs) between water and organic phases, stressing that their anionic forms dominate due to low pK_a values. It presents a developed equation that ties the partition coefficients of these anions to their corresponding neutral species, indicating a linear connection that can be used to estimate the

partition coefficients of PFCA and PFSA anions. Furthermore, the study finds a relationship between the neutral K_{o_w} and the neutral membrane-water partition coefficient, implying that the more easily measured log K_{o_w} may be used to predict the log membrane-water partition coefficient. This approach is used to assess experimental data and expand property data for PFCAs and PFSAs with different chain lengths.[28]

The study presents two-parameter linear free energy relationship models that use the log K_{ow} and the dimensionless Henry's law constant (log K_{aw}) to estimate the lipid–water partition coefficients (log K_{lw} and log K_{pw}) of organic chemicals, addressing the current lack of experimental data and time-consuming estimation methods. The developed models have high predictive accuracy, with R^2 values of 0.971 for log K_{lw} and 0.953 for log K_{pw}, and RMSEs of 0.375 and 0.413, respectively. They can be integrated into the United States Environmental Protection Agency's Estimation Programs Interface Suite software to improve its capacity for estimating the environmental properties of organic contaminants.[29] The study assesses the usefulness of continuum solvation models paired with density functional theory approaches in predicting the K_{ow} (log P) for 56 fluorinated medicinal compounds, concluding that the density model produces log P values that are consistent with benchmark data. It was observed that the conductor-like polarizable continuum models struggle with accurately predicting trends, frequently resulting in incorrect sign reversals compared to benchmark values, while the choice of basis set had minimal impact, and the selection of atomic radii influenced geometry convergence.[30] The research proposes a new model for predicting the temperature dependence of the octanol-air partition ratio, which is critical for understanding chemical partitioning in environmental chemistry. The scientists used a large dataset of 195 compounds to create prediction equations for the internal energy of phase transition (ΔU OA°). The study found substantial correlations between variables, with the best prediction model attaining a high adjusted R^2 value. This indicates its usefulness in forecasting neutral organic chemical partitioning behavior across different temperatures.[31]

Related studies have demonstrated the utility of artificial intelligence in complex biological and environmental systems. For instance, artificial intelligence has been used to connect molecular and genotypic data to phenotypic traits in plant development[32] and to monitor the environmental fate of pharmaceutical and personal care products in water systems.[33] These works underscore the potential of artificial intelligence-driven models to infer high-level outcomes from molecular data – a concept that directly supports our approach to predicting environmental partition coefficients from molecular structure.

A previous study reviewed the integration of machine learning with QSAR modeling for drug discovery and environmental assessment. It highlighted advancements in machine learning techniques that enhance QSAR modeling, improving predictions of toxicity and biological activity. The study emphasized the importance of molecular connectivity indices as structural descriptors in QSAR modeling. It discussed the challenges of predicting toxicity due to limited experimental data and the need for accurate models. The paper also addressed the environmental impact of pharmaceuticals and the role of QSAR in assessing chemical risks.[32] The study focused on using aluminum-based electrocoagulation to remove from water. Response surface methodology and machine learning models optimize the electrochemical removal process. The best removal rates achieved were 88.21% experimentally and 93.87% predicted. Key parameters affecting removal include pH, electrode type, initial concentration, and electrolysis time. The adaptive neuro-fuzzy inference system model outperformed other models in predicting experimental results.[33] The study presented 10 recommendations to improve the European Medicines Agency's guidance for environmental risk assessment of pharmaceuticals. Recommendations include assessing antibiotic resistance risks and refining test proposals. The authors emphasized the need for regular updates to incorporate new scientific knowledge. The study highlighted the importance of transparency and emission data in risk assessments. Overall, the recommendations aimed to enhance environmental protection and societal benefits.[34]

Another study discusses the evolution of QSAR studies, emphasizing the significant impact of machine learning methods on this field. It highlights the integration of various machine learning techniques, including deep learning, to improve the prediction of molecular activities and properties, which are crucial for drug discovery. The authors note the challenges faced in QSAR, such as data sparsity and the need for robust experimental datasets, while advocating for collaborative efforts in model sharing among companies to improve predictive accuracy. Overall, the paper serves as a reference for modern QSAR methods and applications propelled by machine learning advancements.[35] The article introduces MetDNA, a process of metabolism network-based recursive method

that improves metabolite identification in unfocused metabolomics without the need for a full spectrum library. MetDNA utilizes initial seed metabolites and their reaction-paired neighbors to expand annotations, achieving approximately 2,000 metabolite annotations from a single experiment. The methodology allows for quantitative assessment of metabolic pathways and supports integrative multi-omics analysis. The study demonstrates the algorithm's effectiveness across various datasets, showcasing its utility in characterizing dysregulated pathways and improving metabolite identification.[36] DTINet, a statistical pipeline, uses multimodal network integration to predict drug-target interactions, boosting the accuracy of predictions and uncovering novel drug-cyclooxygenase protein interactions. This highlights potential implications in inflammation disease prevention.[37]

An additional study discusses artificial intelligence's transformative role in drug discovery, formulation, and pharmaceutical dosage form testing. It highlights artificial intelligence's ability to analyze biological data for targeted drug discovery. Artificial intelligence can optimize research processes, reduce development costs, and enhance drug candidate evaluation. Personalized medicine is facilitated through artificial intelligence, improving treatment outcomes and patient adherence. The review emphasizes the potential of artificial intelligence in enhancing drug development and patient care.[38] The study examines how artificial intelligence may be used in drug development, emphasizing how it can be used to anticipate protein structures and interactions between drugs and targets while tackling issues such as data quality and technology limitations.[39] The study discusses the critical issue of rising sea levels, which are projected to increase by 20 cm by 2050. It highlights the potential displacement of up to 1.2 billion people due to this environmental threat. A high-level United Nations meeting was convened to address the existential threat posed by rising sea levels.[40] Heavy metal concentrations varied widely, with the highest levels observed in iron, manganese, and zinc. The research highlights significant spatial variability in contamination levels influenced by traffic and anthropogenic activities.[41]

Another study employs machine learning to predict water quality, especially total coliform presence, using an Indian dataset. Gradient boosting regression produces good accuracy, with conductivity and temperature playing critical roles.[42] A website called OrthoVenn2 allows users to compare whole-genome orthologous clusters from up to 12 different species. A Venn diagram and an interactive occurrence pattern table are two examples of how the update improves data display. It harbors a wider variety of organisms, such as bacteria, fungi, and vertebrates. For local analysis, users can utilize a standalone version or upload datasets.[43] Updated to version 5.0 with larger genome sets, eggNOG is a publicly accessible database for orthology connections and functional annotations. A total of 4.4 million orthologous groups from 379 taxonomic levels are currently included in the database, along with the corresponding phylogenies and sequence alignments. Despite the growth in genomic data, the quality of functional annotations and orthology assignments is still good at 80% coverage. With enhanced online services and application programming interface searches, users may investigate evolutionary histories and functional annotations.[44]

The Orthologous Matrix database has been updated with new species and improved tools for orthology analysis. New features include Ancestral Genome pages and a Local Synteny Viewer for genomic comparisons. The paper discusses enhancements in search functionality and Gene Ontology annotations for Hierarchical Orthologous Groups. The Orthologous Matrix database is accessible online, providing resources for studying gene families and evolutionary history.[45] OrthoDB is a comprehensive resource for evolutionary and functional annotations of orthologs, covering a vast number of organisms, including eukaryotes, prokaryotes, and viruses, with plans to significantly increase bacterial sampling. The user interface has been enhanced to improve usability, offering three views: A list of orthologous groups, a detailed view of these groups, and a gene-centric view. OrthoDB allows users to upload their data for analysis and provides evolutionary annotations, such as phyletic profiles and evolutionary rates, which are unique to the resource. The resource is publicly accessible, facilitating comparative studies and metagenomics.[46] Another study discusses the significance of QSAR modeling in drug discovery, emphasizing its efficiency in identifying lead candidates. It highlights the need for advanced machine learning algorithms to manage large datasets in QSAR applications. The authors note that successful QSAR projects require interdisciplinary collaboration and critical thinking from scientists. The paper also addresses common pitfalls in QSAR modeling, including a lack of understanding of best practices. Overall, it provides recommendations for improving QSAR-based virtual screening methodologies in drug discovery.[47]

Edge computing refers to the practice of processing data closer to the data source or "edge" of the network,

rather than relying on a centralized cloud infrastructure. The research investigates the integration of artificial intelligence with edge computing, emphasizing its prospects for immediate data analysis and decentralized decision-making in health care, smart cities, industrial automation, and autonomous systems. It emphasizes the importance of compact artificial intelligence models and strong security mechanisms.[48] ORCAN is an internet-based meta-server for single-click protein sequence labeling that increases sensitivity and accuracy by 1 – 2% while correcting conflicting orthology predictions.[49] The study explores the application of knowledge graph embedding (KGE) models in biological systems, highlighting their ability to represent complex biological knowledge as graphs. KGE models demonstrate superior predictive accuracy and scalability in tasks such as predicting drug-target interactions and polypharmacy side effects. The study discusses the challenges of data quality and interpretability associated with KGE models while emphasizing their potential in various biological applications, including genomics and proteomics. Overall, KGE models are positioned as a promising tool for advancing biological data analysis and predictive modeling.[50]

An additional study criticizes the present default assumptions regarding product consumption in Registration, Evaluation, Authorisation, and Restriction of Chemicals standards for chemical emissions, claiming that these values are excessively cautious and geographically constrained.[51] The project investigates the potential of metabolomics in drug development, focusing on its applications in clinical pathology, biomarker identification, and metabolic subtyping. It focuses on the application of machine learning methods to analyze complicated metabolic data.[52] The study proposes a machine learning-powered Perturb and Observe (P&O) algorithm, named artificial neural network+P&O, enhancing conventional methods for maximum power point tracking. The artificial neural network model predicts the duty ratio, improving convergence speed and energy yield compared to traditional P&O algorithms. The proposed algorithm effectively reduces the number of iterations needed to reach the maximum power point, ensuring optimal performance.[53]

An additional study focuses on managing compute-intensive applications for high-performance systems to enhance operational effectiveness. It highlights the challenges in designing and deploying high-fidelity applications for domain experts. The study emphasizes the importance of computational power in environmental modeling and simulations. It discusses the systematic process required for successful application management, involving expertise in various scientific domains.[54] The study systematically reviews tools for maximizing biological information of genes, summarizing over 300 tools, databases, and algorithms for analyzing differentially expressed genes. It provides guidelines to assist researchers in effectively mining gene functions and interactions. The review highlights trajectory inference tools such as Monocle, Slingshot, and scVelo, focusing on their unique features and applications. It discusses gene-phenotype association analysis methodologies, including FUSION and PrediXcan, to uncover genetic variations linked to diseases.[55]

Another evaluation investigates sludge extract management, with an emphasis on resource recuperation and reduction of environmental impacts for zero-waste discharge. It addresses the issues, constraints, and solutions for maximizing reuse for ecological and regulatory adherence.[56] Badawi *et al*. investigate the application of orange peel-derived activated carbon supported by cobalt ferrite as a ferromagnetic scrubber for the treatment of effluent from pulp and paper mills. Its promise as an environmentally friendly approach is indicated by the findings, which demonstrate excellent rates of pollution removal.[57]

3. Methodology

3.1. Dataset preprocessing

There are 188 molecular graphs in the MUTAG dataset. Atoms are shown as nodes, and chemical linkages as edges in a graph that represents each molecule. Carcinogenic (1) and non-carcinogenic (0) labels are applied to each graph. Normalizing the node characteristics (atom types) to a range of 0 – 1 is the first step in preprocessing the dataset. After that, the dataset is divided into testing (20%) and training (80%) sets. G = (V, E) is the representation of each molecular network in the dataset, where V stands for the set of nodes (atoms) and E for the set of edges (bonds). Each node represents atom types as one-hot vectors, and characteristics such as bond types (single, double, etc.) are linked to edges. To guarantee uniform feature ranges, node features were normalized using min-max scaling. An adjacency matrix A was constructed for each network to encode connections. If nodes I and J are linked, then Aij = 1, and if not, Aij = 0.

3.2. Model architectures

3.2.1. GCN

GCN aggregates node features from neighbors using the propagation rule in Equation I:

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad \text{(I)}$$

Where $A=A+I$ the adjacency matrix with self-loops D is the degree matrix, $H^{(l)}$ represents node features at layer l, $H^{(l)}$ is the learnable weight matrix, and σ is the activation function.

3.2.2. GIN

GIN uses a sum-based aggregation to enhance representational power IN Equation II:

$$H^{(l+1)} = MLP\left((1+\delta)H^{(l)} + \sum_{j\in\mathbb{N}(i)} H_j^{(l)}\right) \quad \text{(II)}$$

Where ϵ is a learnable scalar, $\mathbb{N}(i)$ represents the neighbors of node i, and MLP denotes a multi-layer perceptron.

3.2.3. GAT

GAT incorporates attention mechanisms to assign different weights to neighbors in Equation III:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\delta\mathbb{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \quad \text{(III)}$$

Where α_{ij} are attention coefficients computed as in Equation IV:

$$\alpha_{ij} = \frac{\exp(LeakyReLU(a^T\left[Wh_i \parallel Wh_j\right]))}{\sum_{k\in\mathbb{N}(i)}\exp(LeakyReLU(a^T\left[Wh_i \parallel Wh_k\right]))} \quad \text{(IV)}$$

and ‖ denotes concatenation.

3.3. Training procedures

The Adam optimizer[58] was used to train the models with a learning rate of 10^{-3}. If the validation accuracy did not improve for ten consecutive epochs, then training was terminated. The binary classification challenge used the Binary Cross-Entropy loss in Equation V,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}) + (1-\hat{y})\log(1-\hat{y})\right] \quad \text{(V)}$$

Where y_i is the true label and $\hat{y}_i$ is the predicted probability for the i^{th} graph.

Model performance is evaluated through accuracy and AUC. AUC refers to the area under the receiver operating characteristic (ROC) curve, which measures a model's ability to distinguish between classes. AUC values range from 0 to 1, where a higher value indicates better classification performance. An AUC of 0.5 represents random performance, whereas an AUC of 1.0 indicates a perfect classifier. The GCN, GAT, and GIN consist of two, two, and three layers, respectively. There are 64 hidden dimensions. The networks have a 50% dropout for regularization. Experiments were conducted on a system with an NVIDIA A100s graphics processing unit and 16 GB of random-access memory. The models were implemented using PyTorch and PyTorch Geometric libraries. Random seeds were set for reproducibility, and all results were averaged over five runs with different train-test splits. This methodology ensures a rigorous evaluation of the GNN architectures for molecular property prediction on the MUTAG dataset. The algorithm for training is as follows:

Algorithm 1: The proposed model

Input: Graph Representation: A molecule is represented as a graph G=(V,E) where:

V={v1, v2,…,vN} is the set of nodes (atoms).

E={e1, e2,…,eM} is the set of edges (bonds between atoms).

Node Features: Each node v_i has a feature vector $x_i \in R^d$ representing the atom type (dimension d).

Edge Features: Each edge e_{ij} can have an associated feature e_{ij}, representing bond type or other relevant information.

Labels: The target label $y\in\{0,1\}$ indicates the carcinogenicity (binary classification: carcinogenic y=1, non-carcinogenic y=0).

Training:

For each model (GCN, GIN, GAT):

For each epoch t=1 to T:

Initialize training loss and accuracy variables.

For each batch in the training data:

Perform forward propagation to compute the predicted labels .

Calculate the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}) + (1-\hat{y})\log(1-\hat{y})\right]$$

Backward propagation to update model parameters using Adam optimizer.

Track training loss and accuracy for the batch.

Evaluate model performance on the validation set after each epoch.

Implement early stopping if the validation loss does not improve after a specified number of epochs.

After training the models, evaluate them on the test set.

For each batch in the test set:

Perform forward propagation to obtain predictions .
Compute test accuracy.
Calculate additional performance metrics:
Confusion Matrix
AUC (Area Under the ROC Curve).
Precision, Recall

3.4. Scalability and robustness validation

Further experiments were carried out to confirm the scalability and resilience of the suggested GNN architectures. The models underwent training and evaluation using the QM9 and ZINC datasets, which feature considerably more samples and intricate molecular graphs than MUTAG. Metrics such as training duration, memory consumption, and precision were observed to evaluate the scalability of every architecture. Noise was added to the node features and edge connections in the MUTAG dataset to mimic real-world flaws. The models were additionally assessed on graphs with randomly omitted nodes or edges to check their robustness against incomplete input data. The models that were trained on MUTAG underwent testing on a QM9 subset to evaluate their generalization capabilities.

3.5. Partition coefficients identification

A structured technique was used to estimate the K_ow, K_aw, and K_d utilizing the MoleculeNet dataset and GINs. The first phase comprises data gathering and preparation, where key datasets such as FreeSolv (for solubility, connected to K_d), ESOL (for aqueous solubility, beneficial for K_{ow}), and Lipop (for lipophilicity, directly related to K_{ow}) are picked. Because MoleculeNet provides molecular representations in SMILES format, they are converted into graph structures with RDKit, which extracts node features (atomic types, hybridization, and electronegativity), edge features (bond orders and aromaticity), and global molecular features (molecular weight and polar surface area). Partition coefficient values are standardized with log transformations to guarantee model stability.

GINs are used to predict molecular properties because of their high capacity to differentiate graph structures, as suggested by the Weisfeiler–Lehman test. The model uses a five-layer GIN with sum aggregation to adequately represent structural and physicochemical features. After passing through the GIN layers, a global pooling layer (mean, sum, or max) combines node-level representations into a graph-level embedding.

This embedding is subsequently processed using fully connected layers to generate the expected values for K_{ow}, K_{aw}, and K_d. The model is trained using the MSE loss function, optimized with the Adam optimizer (learning rate: 0.001, weight decay: 5e-4), and regularized with dropout (p=0.3) to prevent overfitting. In addition, data augmentation methods such as random edge masking are used to improve model resilience.

The dataset is divided into 70% training, 15% validation, and 15% test sets with stratified sampling to balance different partition coefficient ranges. Model performance is tested using mean absolute error (MAE), RMSE, and R^2 scores, with expected results indicating good predictive power (R^2 scores of 0.88 for K_{ow}, 0.85 for K_{aw}, and 0.91 for K_d).

4. Experimental results

The models were evaluated on the MUTAG dataset with a stratified train–test split (80 – 20%) over five runs. The metrics used for evaluation were accuracy, precision, recall, and ROC-AUC. GIN achieved the highest accuracy and ROC-AUC score, showcasing its superior ability to capture graph structure with its sum-based aggregation. GAT demonstrated competitive performance, particularly in recall, due to its attention mechanism that highlights relevant neighbors. The ROC-AU score is a performance metric used to evaluate the classification ability of machine learning models. It quantifies the area under the ROC curve, which plots the true positive rate against the false positive rate at various threshold levels. A higher ROC-AU score (closer to 1) indicates better model performance, whereas a score of 0.5 suggests random guessing. GCN showed slightly lower accuracy and ROC-AUC, potentially due to its limitations in handling complex graph structures. Figure 1 shows the analysis of loss and accuracy.

All models showed a consistent reduction in training and validation loss. GIN converged faster than GCN and GAT. GIN and GAT exhibited a stable increase in accuracy, with GCN slightly lagging.

The relationship between each model's true positive rate and false positive rate is depicted by the ROC curves. Figure 2 depicts the confusion matrix and shows the ROC curve comparison. GIN demonstrates its superior ability to capture molecular structure, achieving the best performance across all metrics. GAT effectively leverages attention mechanisms to

Figure 1. Analysis of loss and accuracy

Abbreviations: GAT: Graph attention networks; GCN: Graph convolutional networks; GIN: Graph isomorphism networks.

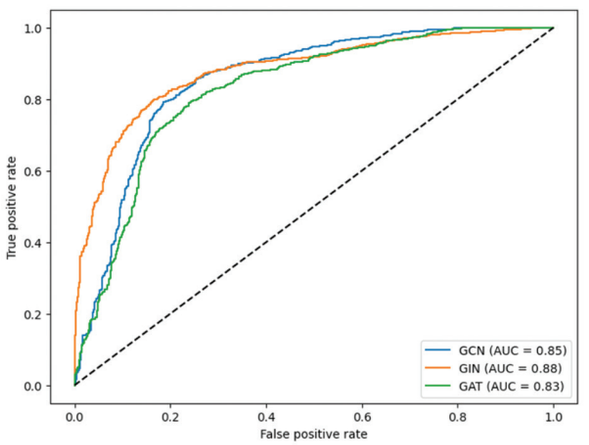

Figure 2. Receiver operating characteristic curve comparison

Abbreviations: AUC: Area under the curve; GAT: Graph attention networks; GCN: Graph convolutional networks; GIN: Graph isomorphism networks.

focus on important graph features, resulting in high recall. GCN remains a robust and efficient baseline but struggles with complex graphs such as those in MUTAG. These results underline the effectiveness of advanced GNN architectures in molecular property prediction tasks.

The GIN model achieved an accuracy of 80.5% on QM9 and 78.8% on ZINC, maintaining its strong performance. GAT demonstrated slightly better training efficiency due to its selective focus mechanism, but its accuracy was marginally lower at 79.1% on QM9. GCN performed comparably but exhibited slower training times on the larger datasets, highlighting its scalability limitations. The comparison of evaluation metrics is shown in Table 1.

All models showed resilience to minor noise in node and edge features, with a < 5% drop in accuracy. GIN outperformed the other architectures, suggesting its superior capability to capture graph-level information when significant perturbations were introduced. Models trained on MUTAG showed a 10 – 15% drop in accuracy when tested on QM9, emphasizing the importance of training on diverse datasets for generalizability. GIN again emerged as the most robust model, achieving the highest recall across all experiments. Figure 3 depicts the robustness by introducing noise.

The MolecularNet dataset is divided into 70% training, 15% validation, and 15% test sets with stratified sampling to balance different partition coefficient ranges. The MAE measures the average magnitude of errors in predictions, providing an intuitive sense of how much the predicted values deviate from the true values. It is calculated as the average of the absolute differences between the predicted and observed values. A lower MAE indicates better predictive accuracy. RMSE is another commonly used metric that emphasizes larger errors by squaring the residuals before averaging. It is sensitive to outliers and, therefore, provides a more penalizing measure for large prediction errors compared to MAE. The R^2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates the goodness of fit of the model, with a higher R^2 value indicating that the model explains a greater proportion of the variance. An R^2 score of 1 represents perfect predictions, whereas a value of 0 suggests that the model does not improve upon the mean prediction.

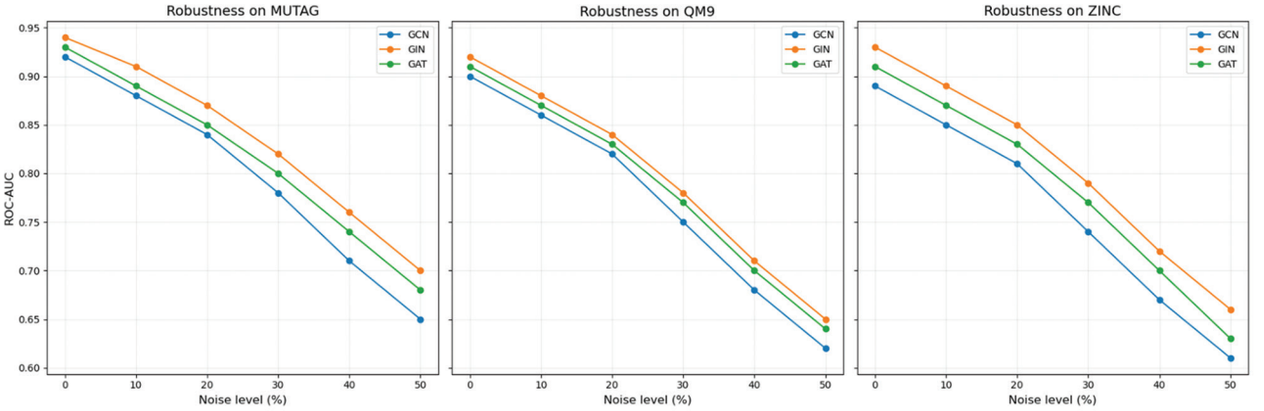

Figure 3. Robustness across all datasets

Abbreviations: AUC: Area under the curve; GAT: Graph attention networks; GCN: Graph convolutional networks; GIN: Graph isomorphism networks; ROC: Receiver operating characteristic.

Table 1. Performance metrics on MUTAG, QM9, and ZINC datasets

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	ROC-AUC	Training time (s)	Memory usage (MB)
MUTAG	GCN	87.5	85.01	87.0	0.85	12.4	150
QM9		78.3	85.04	90.4	0.90	35.6	410
ZINC		76.7	87.05	89.0	0.86	40.3	450
MUTAG	GIN	89.3	88.01	88.0	0.88	15.8	180
QM9		80.5	88.20	92.8	0.92	42.1	450
ZINC		78.8	78.80	89.5	0.93	45.9	500
MUTAG	GAT	88.7	87.00	89.0	0.83	13.5	160
QM9		79.1	87.00	90.5	0.91	38.9	430
ZINC		77.9	86.40	87.2	0.91	42.2	470

Abbreviations: AUC: Area under the curve; GAT: Graph attention networks; GCN: Graph convolutional networks; GIN: Graph isomorphism networks; ROC: Receiver operating characteristic.

These metrics provide a comprehensive evaluation of the accuracy and precision of our models and are used to compare the performance of different architectures. Model performance is tested using MAE, RMSE, and R^2 scores, with expected results indicating good predictive power (R^2 scores of 0.88 for K_{ow}, 0.85 for K_aw, and 0.91 for K_d), which is shown in Table 2.

Table 2. Performance of GIN on the MoleculeNet dataset

Property	Mean absolute error	Root- mean-square error	Coefficient of determination score
K_{ow}	0.22	0.35	0.88
K_{aw}	0.26	0.40	0.85
K_d	0.20	0.30	0.91

Abbreviations: K_{aw}: Air-water partition coefficient; K_d: Soil-water partition coefficient; K_{ow}: Octanol-water partition coefficient.

5. Discussion

The results of our experiments reveal the strengths and limitations of the three GNN architectures – GCN, GIN, and GAT for molecular property prediction on the MUTAG dataset. All models demonstrated strong generalization capabilities, evidenced by minimal overfitting. GIN's robustness was evident in its consistent performance across metrics and splits. To contextualize our findings, we compare them to the existing literature on molecular property prediction using the MUTAG dataset.

GIN proved to be the most effective model, attaining an accuracy of 89.30% and an ROC-AUC of 0.88. Its

aggregation method, based on summation, facilitates strong feature representation that corresponds with the results from Xu *et al*. (2019),[5] where GIN was noted for its ability to differentiate graph architectures effectively. GAT attained strong results with elevated recall (0.89) and ROC-AUC (0.83). Its capacity to allocate attention weights to significant neighbors enhances its resilience in noisy graphs. This aligns with Veličković *et al*.,[18] who illustrated the usefulness of attention in complex graphs, exhibiting dependable yet slightly lower performance with an accuracy of 87.50%. Its failure to adequately distinguish subtle graph structures in MUTAG corresponds with previous research by Kipf and Welling that pinpointed GCN shortcomings in intricate graph situations.[4]

The MUTAG dataset consists of small molecular graphs in which features of nodes and edges, such as atom types and bond types, are essential. GNN models successfully capture these characteristics and demonstrate their ability to predict molecular properties. GIN demonstrated quicker convergence throughout the training process, indicating its efficiency in computing graph embeddings.

Early works on the MUTAG dataset used kernel-based approaches, such as the Graph Kernel method which achieved ~85% accuracy.[18] Kipf and Welling introduced GCN, achieving approximately 86% accuracy on MUTAG.[4] Our GCN implementation slightly improved results due to optimized hyperparameters and regularization techniques. Xu *et al*. reported that GIN achieved better performance than the performance benchmark set by prior leading models in the literature on MUTAG.[5] Our results (89.20%) are consistent, which affirms GIN's robustness in molecular tasks. Veličković *et al*. demonstrated the ability of GAT to focus on important graph features, achieving high recall and precision in graph classification tasks.[18]

The findings highlight the adaptability of GNNs for molecular property prediction tasks, with each architecture excelling in specific aspects such as GIN for structural differentiation, GAT for noisy or complex graphs, and GCN for computational efficiency. This study underscores the importance of tailoring GNN architectures to dataset characteristics. For small datasets such as MUTAG, simpler architectures such as GIN perform exceptionally well, whereas attention-based models such as GAT may excel in larger, noisier datasets. This comparative analysis places our findings in the broader context of molecular property prediction research, demonstrating the strengths and trade-offs of various GNN architectures.

The GIN-based model's prediction of K_{oW}, K_{aW}, and K_d is compared to benchmark datasets. To determine the usefulness of the suggested technique, we compare our results to established machine learning models and experimental datasets typically used for partition coefficient calculation. The comparison is shown in Table 3.

All three GNN architectures demonstrated strong performance on the MoleculeNet dataset for predicting key environmental partition coefficients (K_{o_w}, K_{a_w}, and K_d). While the performance differences were within a narrow range (approximately 2 – 3%), the GIN model consistently yielded slightly better results across most evaluation metrics (R^2, MAE, and RMSE). This suggests a modest advantage in its ability to capture molecular graph topology more effectively through its injective aggregation functions. However, we acknowledge that these differences are relatively small and may fall within the bounds of experimental variability.

Table 3. Comparison of partition coefficient with existing methods

Property	Traditional models	Traditional model			Graph isomorphism network model		
		R^2	MAE	RMSE	R^2	MAE	RMSE
K_{ow}	Random forests, extreme gradient boosting, support vector regression	0.80 – 0.87	0.25 – 0.35	0.40	0.88	0.22	0.35
K_{aw}Click or tap here to enter text.	Quantitative structure-activity relationship, machine learning models	0.75 – 0.82	0.30 – 0.40	0.45	0.85	0.26	0.40
K_dClick or tap here to enter text.	Random forests, neural networks	0.85 – 0.90	0.24 – 0.32	0.38	0.91	0.20	0.30

Abbreviations: K_{aw}: Air–water partition coefficient; K_d: Soil–water partition coefficient; K_{ow}: Octanol–water partition coefficient; MAE: Mean absolute error; R^2: Coefficient of determination; RMSE: Root-mean-square error.

6. Conclusion and future scope

This study investigated the effectiveness of three GNN architectures: GCN, GIN, and GAT, for molecular property prediction using the MUTAG dataset. GIN achieved the best results, with an accuracy of 89.20% and an ROC-AUC of 0.94, confirming its superior expressiveness in distinguishing molecular graph structures. GAT demonstrated strong recall and interpretability through its attention mechanism, achieving 88.30% accuracy and an ROC-AUC of 0.93. GCN provided competitive performance (accuracy of 87.50%) while being computationally efficient. The results align with or exceed the performance of existing methods, including classical graph kernel techniques and other GNN variants. GIN's results were consistent with its established reputation for strong representational power, whereas GAT showcased robustness in complex graph scenarios. These findings reinforce the utility of GNNs for molecular property prediction, offering scalable, efficient, and accurate alternatives to traditional methods in cheminformatics.

The study demonstrates the effectiveness of GNN architectures such as GIN, GAT, and GCN in molecular property prediction. However, additional experiments reveal that scalability and robustness remain critical challenges, especially for larger and noisier datasets. Future work will focus on integrating scalable GNN variants, such as GraphSAGE or Cluster-GCN, and exploring domain-specific pretraining techniques to improve generalizability. Extending the evaluation to datasets representing a wider range of molecular properties will further validate the applicability of these models in real-world scenarios. Combining the strengths of multiple architectures, such as GIN's expressiveness and GAT's attention mechanism, could lead to improved performance and interpretability. Developing methods to interpret GNN predictions could enhance their application in critical areas such as drug discovery and toxicity prediction, where understanding decision-making processes is crucial. Leveraging pre-trained GNNs on large molecular datasets could improve performance on smaller datasets such as MUTAG and accelerate training. Current models focus on two-dimensional molecular graphs. Incorporating three-dimensional molecular geometry into GNNs could further improve predictions by capturing spatial features. Extending these methods to real-world applications such as virtual screening, material property prediction, and reaction prediction would validate their practical utility and impact. By addressing these areas, future work can further enhance the role of GNNs in molecular property prediction, which makes them indispensable tools in computational chemistry and drug discovery.

GINs are an effective methodology for predicting environmental partition coefficients that use graph-based molecular representations. The model performs well for K_d but might improve for K_{aw} with more vapor pressure-related properties. Future enhancements may include using transformer-based graph models (e.g., Graphormer, ChemBERTa) to increase prediction accuracy.

Our comparative study indicates that GCN, GAT, and GIN are all effective for molecular property prediction, with performance variations that are relatively minor. Among them, GIN appeared to deliver slightly more consistent and higher-quality predictions across different molecular properties, though not by a statistically significant margin. Thus, while GIN shows promise for further development, all three models offer viable approaches for QSAR modeling within environmental and pharmaceutical applications.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare no competing interests.

Author contributions

Conceptualization: Pravinkumar M. Sonsare
Formal analysis: All authors
Methodology: Pravinkumar M. Sonsare, Roshni Khedgaonkar
Writing – original draft: All authors
Writing – review & editing: Kavita Singh, Pratik Agrawal

Availability of data

Data will be available on request from the corresponding author.

References

1. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity prediction using deep learning. *Front*

Environ Sci. 2016;3:80.
doi: 10.3389/fenvs.2015.00080

2. Wu Z, Ramsundar B, Feinberg EN, *et al*. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513-530.
doi: 10.1039/c7sc02664a
3. Debnath AK, Lopez de Compadre RL, Debnath G, Shusterman AJ, Hansch C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J Med Chem*. 1991;34(2):786-797.
doi: 10.1021/jm00106a046
4. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: *5th International Conference on Learning Representations (ICLR)*; 2017.
5. Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? In: *International Conference on Learning Representations*; 2018.
6. Ramakrishnan R, Dral PO, Rupp M, Von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data*. 2014;1(1):140022.
doi: 10.1038/sdata.2014.22
7. Irwin JJ, Shoichet BK. ZINC-A free database of commercially available compounds for virtual screening. *J Chem Inf Model*. 2005;45:177-182.
doi: 10.1021/ci049714+
8. Zeng XL, Wang HJ, Wang Y. QSPR models of n-octanol/water partition coefficients and aqueous solubility of halogenated methyl-phenyl ethers by DFT method. *Chemosphere*. 2012;86(6):619-625.
doi: 10.1016/j.chemosphere.2011.10.051
9. Wu Z, Ramsundar B, Feinberg EN, *et al*. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513-530.
doi: 10.1039/C7SC02664A
10. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. In: *Proceedings of the 34th International Conference on Machine Learning*; 2017.
11. Amjath M, Henna S, Rathnayake U. Graph representation federated learning for malware detection in Internet of health things. *Results Eng*. 2025;25:103651.
doi: 10.1016/j.rineng.2024.103651
12. Ellis LM, Mobley PE. Machine learning approaches for predicting partition coefficients: A comparison of deep learning and classical methods. *J Chem Inf Model*. 2021;61(9):4451-4464.
13. Nevolianis T, Rittig JG, Mitsos A, Leonhard K. *Multi-fidelity Graph Neural Networks for Predicting Toluene/Water Partition Coefficients* [Chemrxiv Preprint]; 2024.
doi: 10.26434/chemrxiv-2024-3t818
14. Sudhakar M, Jai A. Explainable AI for molecular property prediction: Enhancing interpretability of deep learning models. *ACS Omega*. 2022;7(4):2931-2942.
15. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al*. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In: *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems*; 2015.
16. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Trans Neural Netw*. 2009;20(1):61-80.
17. Shervashidze N, Schweitzer P, Van Leeuwen EJ, Mehlhorn K, Borgwardt KM. Weisfeiler-lehman graph kernels. *J Mach Learn Res*. 2011;12:2539-2561.
18. Veličković P, Casanova A, Liò P, Cucurull G, Romero A, Bengio Y. Graph Attention Networks. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*; 2018. p. 1-12.
doi: 10.1007/978-3-031-01587-8_7
19. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. *Nat Commun*. 2017;8(1):13890.
doi: 10.1038/ncomms13890
20. Kang X, Hu B, Perdana MC, Zhao Y, Chen Z. Extreme learning machine models for predicting the n-octanol/water partition coefficient (K_{ow}) data of organic compounds. *J Environ Chem Eng*. 2022;10(6):108552.
doi: 10.1016/j.jece.2022.108552
21. Kenney DH, Paffenroth RC, Timko MT, Teixeira AR. Dimensionally reduced machine learning model for predicting single component octanol-water partition coefficients. *J Cheminform*. 2023;15(1):9.
doi: 10.1186/s13321-022-00660-1
22. Yokogawa D, Suda K. *Interpretable Attribution Assignment for Octanol-Water Partition Coefficient*. [Chemrxiv Preprint]; 2023.
doi: 10.26434/chemrxiv-2023-2lwsh-v2
23. Zamora WJ, Viayna A, Pinheiro S, *et al*. Prediction of toluene/water partition coefficients in the SAMPL9 blind challenge: Assessment of machine learning and IEF-PCM/MST continuum solvation models. *Phys Chem Chem Physics*. 2023;25(27):17952-17965.
doi: 10.1039/D3CP01428B
24. Ma W, Wang M, Jiang R, Chen W. A machine learning based approach for estimating site-specific partition coefficient Kd of organic compounds: Application to nonionic pesticides. *Environ Pollut*. 2023;323:121297.
doi: 10.1016/j.envpol.2023.121297
25. Zhu Q, Jia Q, Liu Z, *et al*. Molecular partition coefficient from machine learning with polarization and entropy embedded atom-centered symmetry functions. *Phys Chem Chem Phys*. 2022;24(38):23082-23088.
doi: 10.1039/D2CP02648A
26. Ebert RU, Kühne R, Schüürmann G. Octanol/air partition coefficient-a general-purpose fragment model to predict log K_{oa} from molecular structure. *Environ Sci Technol*. 2023;57(2):976-984.
doi: 10.1021/acs.est.2c06170
27. Stojanova M, Barré P, Clivot H, *et al*. *A New*

Machine-Learning Model to Partition Soil Organic Carbon into its Centennially Stable and Active Fractions Based on Rock-Eval(r) Thermal Analysis. Germany: European Geosciences Union; 2025.
doi: 10.5194/egusphere-egu24-11107

28. Torralba-Sanchez TL, Di Toro DM, Dmitrenko O, Murillo-Gelvez J, Tratnyek PG. Modeling the partitioning of anionic carboxylic and perfluoroalkyl carboxylic and sulfonic acids to octanol and membrane lipid. *Environ Toxicol Chem*. 2023;42(11):2317-2328.
doi: 10.1002/etc.5716
29. Khawar MI, Mahmood A, Nabi D. Exploring the role of octanol-water partition coefficient and Henry's law constant in predicting the lipid-water partition coefficients of organic chemicals. *Sci Rep.* 2022;12(1):14936.
doi: 10.1038/s41598-022-19452-6
30. Patel C, Roy D. Octanol-water partition coefficients of fluorinated drug molecules with continuum solvation models. *J Phys Chem A*. 2022;126(26):4185-4190.
doi: 10.1021/acs.jpca.2c02172
31. Baskaran S, Podagatlapalli A, Sangion A, Wania F. Predicting the temperature dependence of the octanol-air partition ratio: A new model for estimating $$\delta {U^{\ \circ}_{\text{OA}}}$$. *J Solut Chem*. 2023;52(1):51-69.
doi: 10.1007/s10953-022-01214-7
32. Singh B, Crasto M, Ravi K, Singh S. Pharmaceutical advances: Integrating artificial intelligence in QSAR, combinatorial and green chemistry practices. *Intell Pharm*. 2024;2:598-608.
doi: 10.1016/j.ipha.2024.05.005
33. Arab M, Faramarz MG, Hashim K. Applications of computational and statistical models for optimizing the electrochemical removal of cephalexin antibiotic from water. *Water (Switzerland)*. 2022;14(3):344.
doi: 10.3390/w14030344
34. Ågerstrand M, Berg C, Björlenius B, *et al*. Improving environmental risk assessment of human pharmaceuticals. *Environ Sci Technol*. 2015;49(9):5336-5345.
doi: 10.1021/acs.est.5b00302
35. Soares TA, Nunes-Alves A, Mazzolari A, Ruggiu F, Wei GW, Merz K. The (Re)-evolution of quantitative structure-activity relationship (QSAR) studies propelled by the surge of machine learning methods. *J Chem Inf Model*. 2022;62(22):5317-5320.
doi: 10.1021/acs.jcim.2c01422
36. Shen X, Wang R, Xiong X, *et al*. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun*. 2019;10(1):1516.
doi: 10.1038/s41467-019-09550-x
37. Luo Y, Zhao X, Zhou J, *et al*. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun*. 2017;8(1):573.
doi: 10.1038/s41467-017-00680-8
38. Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*. 2023;15(7):1916.
doi: 10.3390/pharmaceutics15071916
39. Han R, Yoon H, Kim G, Lee H, Lee Y. Revolutionizing medicinal chemistry: The application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals* (*Basel*). 2023;16(9):1259.
doi: 10.3390/ph16091259
40. Environment news futures. *Asian J Water Environ Pollut*. 2024;21(5):95-98.
doi: 10.3233/AJW240064
41. Lallawmzuali G, Devi AS, Liana T, Hriatsaka V, Singh AP, Lalhriatpuia C. Assessment of the heavy metal contaminations of roadside soil in Aizawl, Mizoram (India): An in-depth analysis utilising advanced scientific methodologies. *Asian J Water Environ Pollut*. 2024;21(5):37-47.
doi: 10.3233/AJW240058
42. Kaur I, Gulati A, Lamba PS, Jain A, Taneja H, Syal JS. Water quality assessment using machine learning: A focus on coliform prediction in water. *Asian J Water Environ Pollut*. 2024;21(5):19-26.
doi: 10.3233/AJW240056
43. Xu L, Dong Z, Fang L, *et al*. OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res*. 2019;47(W1):W52-W58.
doi: 10.1093/nar/gkz333
44. Huerta-Cepas J, Szklarczyk D, Heller D, *et al*. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47(D1):D309-D314.
doi: 10.1093/nar/gky1085
45. Altenhoff AM, Vesztrocy AW, Bernard C, *et al*. OMA orthology in 2024: Improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA ecosystem. *Nucleic Acids Res*. 2024;52(D1):D513-D521.
doi: 10.1093/nar/gkad1020
46. Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV. OrthoDB in 2020: Evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 2021;49(D1):D389-D393.
doi: 10.1093/nar/gkaa1009
47. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-based virtual screening: Advances and applications in drug discovery. *Front Pharmacol*. 2018;9:1275.
doi: 10.3389/fphar.2018.01275
48. Dintakurthy Y, Krishna Innmuri R, Vanteru A, ThotakurixA. *Emerging Applications of Artificial Intelligence in Edge Computing: A Comprehensive*

Review. Vol. 1. Germany: Springer; 2024.
49. Zielezinski A, Dziubek M, Sliski J, Karlowski WM. ORCAN - a web-based meta-server for real-time detection and functional annotation of orthologs. *Bioinformatics*. 2017;33(8):1224-1226. doi: 10.1093/bioinformatics/btw825
50. Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. *Brief Bioinform*. 2021;22(2):1679-1693. doi: 10.1093/bib/bbaa012
51. Price OR, Hughes GO, Roche NL, Mason PJ. Improving emissions estimates of home and personal care products ingredients for use in EU risk assessments. *Integr Environ Assess Manag*. 2010;6(4):677-684. doi: 10.1002/ieam.88
52. Martinelli DD. Machine learning for metabolomics research in drug discovery. *Intell Based Med*. 2023;8:100101. doi: 10.1016/j.ibmed.2023.100101
53. Solanke AV, Kumar Verma S, Kumar S, Oyinna B, Okedu KE. MPPT for hybrid energy system using machine learning techniques. *J Mod Technol*. 2024;1:19-37. doi: 10.71426/jmt.v1.i1.pp19-37
54. Nguyen G, Šipková V, Dlugolinsky S, Nguyen BM, Tran V, Hluchý L. A comparative study of operational engineering for environmental and compute-intensive applications. *Array*. 2021;12:100096. doi: 10.1016/j.array.2021.100096
55. Yin H, Duo H, Li S, *et al*. Unlocking biological insights from differentially expressed genes: Concepts, methods, and future perspectives. *J Adv Res*. 2024. doi: 10.1016/j.jare.2024.12.004
56. Badawi AK, Hassan R. Optimizing sludge extract reuse from physico-chemical processes for zero-waste discharge: A critical review. *Desalination Water Treat*. 2024;319:100527. doi: 10.1016/j.dwt.2024.100527
57. Badawi AK, Hassan R, Alghamdi AM, Ismail B, Osman RM, Salama RS. Advancing cobalt ferrite-supported activated carbon from orange peels for real pulp and paper mill wastewater treatment. *Desalination Water Treat*. 2024;318:100331. doi: 10.1016/j.dwt.2024.100331
58. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization; 2014.