UMC Utrecht

Utrecht University

Graph Neural Networks in Molecular Property Prediction

Writing Assignment – Literature review

Master in Bioinformatics and Biocomplexity

Author: Markel Benito
Examiner: Jeroen De Ridder
Daily supervisor: Dieter Stoker

LAYMAN'S SUMMARY

For effective drug design, it is essential to understand how molecules behave. Scientists use tools to predict important properties such as toxicity, how harmful a molecule can be, and solubility, how well it dissolves in water. These properties are crucial in determining whether a drug can act safely and effectively. Traditionally, these predictions have been made using methods based on molecular descriptors, which are numerical representations of molecules that summarize their chemical and physical properties, such as size, weight, or polarity. Initially, these descriptors were manually designed, but over time, the process was automated through computational tools (1). Although useful, these methods have limitations when working with complex molecules, as they often fail because of the bias problems (2).

In recent years, Graph Neural Networks (GNNs) have emerged as a solution by treating molecules as graphs. In this approach, atoms are represented as points (nodes) and bonds as connections (edges). This representation allows GNNs to directly analyze the structure of a molecule and identify patterns that traditional methods might overlook (3). For example, GNNs are particularly effective at capturing how atoms interact within a molecule, which is essential for predicting molecular properties (4). This graph-based network employs a process called *message passing*, where atoms exchange information with their neighbors, enabling the network to build a detailed representation of the molecule. Another key stage is the *readout phase*, where the collected information is combined to generate a global representation of the molecule, which can then be used in tasks such as classification (e.g. is the molecule toxic or not?) or regression (e.g. solubility) (5).

In addition to their advantages, GNNs face significant challenges. Scalability is one of the main issues, as working with very large molecules can be computationally expensive. Another limitation is the accurate representation of stereochemistry, where small differences in the 3D arrangement of atoms can lead to vastly different molecular behaviors. To address these limitations, advanced versions of GNNs have been developed, such as Hierarchical Informative GNNs (HiGNNs), which simplify the analysis of large molecules by grouping similar nodes, or Equivariant GNNs (EGNNs), which ensure consistent predictions regardless of the molecule's three-dimensional orientation (6,7). Ongoing research is also tackling other limitations through further advancements in GNN architectures.

Given the potential of these graph-based networks, this work consists of a literature review that analyzes the current state of GNNs in molecular property prediction. It describes their functioning, main applications, and recent advancements in architecture that overcome traditional limitations. Unlike other complex and fragmented resources, this review provides a clear and accessible introduction to GNNs applied to molecular property prediction. By consolidating foundational knowledge and recent advancements, it aims to serve as a practical guide for researchers and practitioners, whether new to the field or experienced, exploring this innovative methodology (8).

ABSTRACT

Molecular property prediction is a key component in drug development, enabling early and accurate assessments of the chemical and biological characteristics of potential compounds. This step optimizes the selection process in vast molecular libraries and helps eliminate compounds with adverse profiles early, minimizing risks before advancing to later stages (1). Historically, these predictions have been made manually using molecular descriptor design, which required extensive expertise and has been prone to bias. To address these limitations, computational tools like SMILES and molecular fingerprints were introduced, automating the process and enhancing scalability. However, these approaches still retained biases, prompting the development of more advanced methods (2).

Recently, Graph Neural Networks (GNNs) have emerged as an innovative solution to address these challenges. GNNs model molecules directly as graphs, where atoms are nodes and bonds are edges (3). Through processes such as *message passing*, GNNs allow neighboring nodes to exchange information, enabling the construction of enriched representations that capture complex structural and spatial relationships. In the *readout phase*, this information is aggregated into a global vector called embedding, which is then used for tasks such as classification or regression (Fig.1) (5).

Despite their advantages, GNNs face challenges such as stereochemistry, and interpretability. Recent advancements, including Equivariant GNNs (EGNNs) and Graph Attention Networks (GATs), have made significant progress in mitigating these limitations and enhancing the overall performance of these networks (7,9,10). However, further research is required to more effectively tackle these and other persistent challenges.

Given the potential of Graph Neural Networks (GNNs) in molecular property prediction, this review provides a comprehensive analysis of their current state, highlighting their capabilities, challenges, and recent advancements. By consolidating foundational knowledge and the latest developments, this review serves as a clear and accessible resource for researchers and practitioners seeking to understand and leverage GNNs in molecular property prediction.

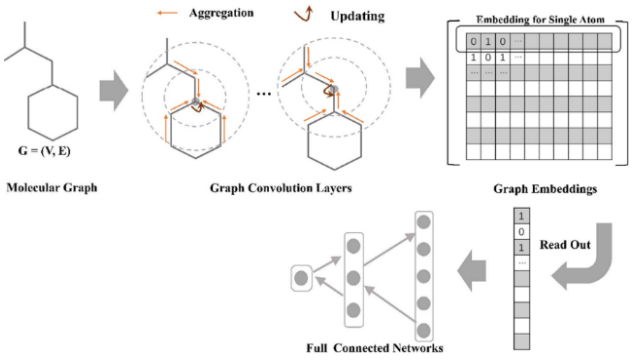


Figure 1: Overview of Graph Neural Networks (GNNs) for molecular property prediction. Molecules are represented as graphs where atoms are nodes and bonds are edges. Through message passing and aggregation, GNNs create enriched atom embeddings, which are combined in the readout phase into global molecular embedding. This embedding is used for property prediction tasks like classification or regression. Taken from (11).

TABLE OF CONTENTS

ABBREVIATIONS

QSPR	Quantitative structure-property relationship
QSPR	Quantitative structure-activity relationship
2D	2-dimensional
3D	3-dimensional
ML	Machine learning
SMILE	Simplified Molecular Input Line Entry System
SVM	Support-Vector-Machine
RF	Random Forest
NN	Neural Network
GNN	Graph Neural Network
HiGNN	Hierarchical Informative Graph Neural Network
GCN	Graph Convolutional Network
GAT	Graph Attention Network
RMSE	Root Mean Square Error
GraphSAINT	Graph Sampling Based Inductive Learning Method
EGNN	Equivariant Graph Neural Network
MAE	Mean Absolute Error

1. Introduction

1.1 Molecular property prediction

Molecular property prediction plays a critical role in drug discovery, as it enables early and accurate assessment of the chemical and biological characteristics of potential compounds. This step not only optimizes the screening of vast molecular libraries but also allows researchers to identify and exclude compounds with likely adverse profiles early in the process, minimizing risks in later stages of development (12).

Historically, molecular property prediction has relied on molecular descriptors, numerical features that summarize the physicochemical (e.g., dipole moment), topological (e.g., connectivity), and compositional characteristics (e.g., molecular weight) of molecules (13). By translating molecular structures into a format that can be used in mathematical models, these descriptors have been foundational in approaches such as QSAR (Quantitative Structure-Activity Relationship) and QSPR (Quantitative Structure-Property Relationship) (14). QSAR focuses on predicting biological activities, such as toxicity, while QSPR models physicochemical focuses on properties like solubility and boiling point (13). These methods, which range from simple linear regression to more advanced machine learning techniques, establish relationships between molecular descriptors and target properties. Their evolution and applications will be discussed in greater detail in the following sections.

1.2 Descriptor-based classical methods

Initially, experts in chemistry used to manually design these molecular descriptors by selecting features they believed were relevant to specific properties. For example, a simple descriptor is molecular weight, which is calculated by adding up the atomic masses of all the atoms in a molecule. This measure provides basic information about the size of a molecule and can be linked to properties like solubility (1). In early QSAR and QSPR models, performance evaluation was often conducted using statistical methods such as linear regression and multiple regression to establish relationships between descriptors and target properties. These approaches provided a straightforward way to predict molecular properties based on structural features (13).

However, while these handcrafted descriptors have been widely used, the design of these descriptors require expertise from chemistry professionals, which introduces potential biases and restricts the generalizability of the models (15). These shortcomings led to the development of more advanced computational methods that can address these challenges (2).

1.3 Descriptor based ML-based methods

The limitations of handcrafted molecular descriptors, which were traditionally combined with simple statistical methods like linear and multiple regression, led to the development of computational models based on machine learning (ML). This shift introduced automated molecular representations such as the Simplified Molecular Input Line Entry System (SMILES) and molecular fingerprints, which systematically encode molecular structures without requiring manual feature design (1). SMILES encodes molecules as text strings that describe their composition and connectivity (Fig.2), while fingerprints transform molecular structures into bond matrices or triangular matrices that reflect the presence or absence of specific chemical substructures (Fig.3). These representations allowed ML models to process molecular data more systematically and efficiently, reducing their dependency on handcrafted descriptors (15).

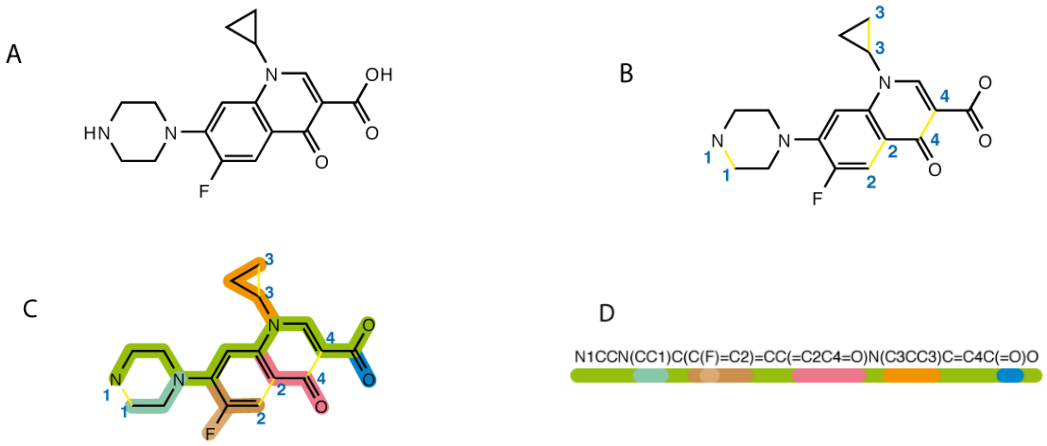
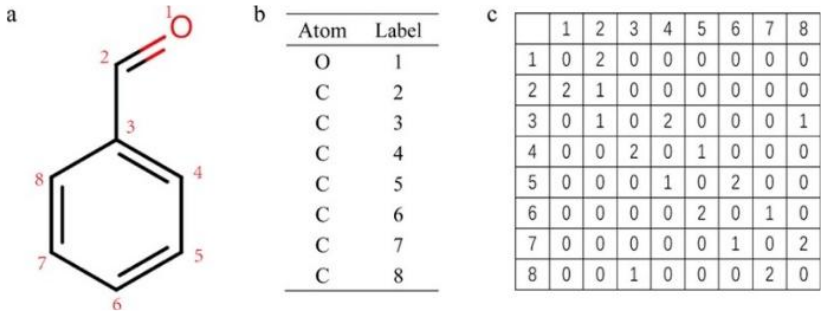


Figure 2: SMILES representation of a chemical molecule, showing: (A) the chemical structure, (B) atom numbering, (C) fragments highlighted by groups, and (D) the resulting SMILES string. The groups highlighted in (C) are automatically identified by software that analyzes the molecule, which follows specific rules to define the main carbon chain and side chains. Taken from *Wikimedia Commons, file "SMILES.png" (CC BY-SA 4.0)*.



Atom	Label
O	1
C	2
C	3
C	4
C	5
C	6
C	7
C	8

	1	2	3	4	5	6	7	8
1	0	2	0	0	0	0	0	0
2	2	1	0	0	0	0	0	0
3	0	1	0	2	0	0	0	1
4	0	0	2	0	1	0	0	0
5	0	0	0	1	0	2	0	0
6	0	0	0	0	2	0	1	0
7	0	0	0	0	0	1	0	2
8	0	0	1	0	0	0	2	0

Figure 3: 2D connection table of benzaldehyde as a molecular fingerprint. Non-hydrogen atoms are labeled (a), with their corresponding atom types listed (b). The table (c) represents the bond types between atoms numerically: 1 for single bonds, 2 for double bonds, and 0 for no direct connection between atoms. This encoding captures the connectivity of the molecule in a format useful for machine learning applications. Taken from (16).

Among the most widely used ML models are support vector machines (SVMs), random forests (RFs), and neural networks (NNs). Unlike earlier statistical methods, RF and SVM have already simplified molecular property prediction by leveraging automated decision rules and nonlinear feature mappings, reducing the need for handcrafted models. However, both still rely heavily on manual feature engineering, where descriptors must be carefully designed to capture relevant molecular properties. In contrast, NNs automatically learn hierarchical feature representations directly from raw data, eliminating the need for predefined descriptors. This is particularly advantageous for complex molecular datasets, as NNs can uncover intricate patterns that may not be captured by predefined features. In a typical NN, the input layer processes raw data, hidden layers extract progressively abstract features, and the output layer generates predictions. The application of NNs combined with SMILES and fingerprint representations has further simplified feature extraction, ensuring that consistent patterns are learned across all data while minimizing potential biases (1).

However, biases can still arise from the datasets used for training and assumptions within the models themselves (17). To address these shortcomings, recent advances have shifted towards graph-based representations and deep learning. Deep learning typically refers to neural networks with more than 8 layers, which allow for deeper and more sophisticated hierarchical feature extraction. These concepts will be explained further in the next sections.

1.4 Graph Neural Networks (GNNs)

Graph neural networks (GNNs) have emerged as an innovative solution to address the current challenges in molecular property prediction. These networks excel in modeling molecular structures directly as graphs, where atoms are represented as nodes and bonds as edges. Each node in the graph contains information in the form of a vector about the atom it represents, such as its atomic number, hybridization state, or partial charge. Similarly, edges store information about the type of chemical bond, such as single, double, or aromatic bonds. This rich representation allows GNNs to encode relevant information of molecular structures in ways that traditional descriptor-based methods cannot (1). One of the core concepts of GNNs is the *message passing* phase. In this step, each node exchanges information with its neighbors to update its own features. This iterative exchange progressively enriches the nodes with information about their chemical environment, resulting in graph representations that capture complex relationships between atoms and their connections (17). The other core concept is the *readout phase*, where the enriched node representations are pooled, meaning they are aggregated using methods such as summation, averaging, or maximization, to generate a global graph-level representation. This step is crucial as it condenses the information from the entire graph into a fixed-size vector, making it suitable for downstream tasks such as regression or classification. These concepts will be explained in more detail in the following sections.

1.5 Objectives of the review

The primary goal of this review is to provide a clear and comprehensive analysis of the role of Graph Neural Networks (GNNs) in molecular property prediction. While the field of graph-based methods is vast, encompassing a wide range of applications and specialized methodologies, this review will focus specifically on their application to predicting molecular properties. This review will begin by exploring the foundations of GNNs, explaining how these networks are capable of modeling complex molecular structures to predict molecular properties. To maintain accessibility for readers new to the field, the review will minimize the use of overly technical mathematical explanations, emphasizing core concepts and practical applications instead. It will then address the limitations of GNNs and how researchers have developed specialized variants and strategies to overcome these obstacles. These include techniques for improved scalability, enhanced modeling of stereochemical and three-dimensional molecular interactions, and better interpretability. Because GNNs are a relatively new approach in this field, the lack of a clear and accessible review makes it challenging for researchers and practitioners to fully understand their applications. This work aims to fill that gap by providing a foundational guide that highlights their strengths, limitations, and recent advancements, offering insights into their potential impact on molecular property prediction for drug discovery and molecular research.

2. GNNs in molecular property prediction

2.1 Molecular graph

When talking about Graph Neural Networks (GNNs) for molecular property prediction, one of the main questions is how to effectively represent molecules as graphs, where atoms are represented as nodes and chemical bonds as edges (Fig.4) (5). Each node contains information about atomic characteristics, such as type, hybridization state, and charge (Table 1). Similarly, edges encode bond properties, including bond order (single, double, triple, or aromatic), conjugation, and ring membership (Table 2). These features are represented as one-hot vectors, with dimensions tailored to the type of information they encode (1,18). A one-hot vector is a binary representation where only one element is active (set to 1), while all others remain inactive (set to 0). For example, an atom type feature could be represented as a one-hot vector, where carbon might be encoded as [1,0,0,0], oxygen as [0,1,0,0], and nitrogen as [0,0,1,0]. This method ensures that categorical information is efficiently captured in a numerical format, making it suitable for processing within GNN architectures (5).

While this representation eliminates the need for traditional molecular descriptors, it is important to note that the initial features of nodes and edges are often predefined, e.g. obtained using tools like RDKit, and serve as the starting point for the GNN's learning process. During training, the GNN refines these initial representations, adapting them to better capture patterns and relationships relevant to the specific prediction task. The *message-passing* process, described in detail in Section 2.2, plays a crucial role in this refinement. At each step, nodes aggregate information from their neighbors and edges, allowing the model to capture both local interactions, such as bonding patterns, and global molecular

structures. Additionally, this representation does not depend on the order of the nodes, meaning the graph remains the same no matter how the input is arranged. This is different from traditional methods, which usually assign features to nodes in a fixed order, so if the order changes the entire representation can change. GNNs focus on the connections between nodes, not their order, making molecular modeling more stable and reliable (19).

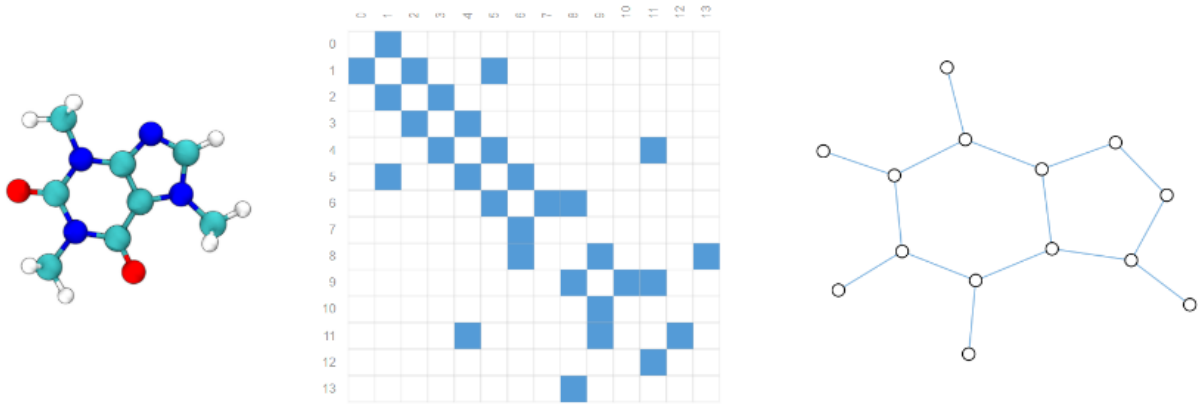


Figure 4: (Left) 3D structure of the caffeine molecule. (Center) Adjacency matrix of the bonds in the molecule. (Right) Graph-based representation of the molecule. Taken from (18).

Feature	Description/example values
Atom type	Type of atom (C, O, S, F)
Hybridization	sp, sp2, sp3, sp3d, or sp3d2
Charge	Formal charge of the atom
Bonds	Number of bonds the atom is involved in

Table 1: Each atom is represented by features like atom type, hybridization state, formal charge, and number of bonds. For example, hybridization state is encoded as a one-hot vector, where sp = [1, 0, 0, 0, 0], sp2 = [0, 1, 0, 0, 0], and sp3 = [0, 0, 1, 0, 0]. Taken from (1).

Feature	Description/example values
Bond order	Single, double, triple, or aromatic
Conjugated	Whether the bond is conjugated
Ring membership	Whether the bond is part of a ring

Table 2: Bond features for the initial edge feature vector. Each bond is represented by features like bond order, conjugation, and ring membership. For example, bond order is encoded as a one-hot vector, where single = [1, 0, 0, 0], double = [0, 1, 0, 0], triple = [0, 0, 1, 0], and aromatic = [0, 0, 0, 1]. Taken from (1).

2.2 Message passing phase

After constructing the molecular graph, the *message-passing* process is a key step in predicting molecular properties (Fig.5). This process enables the graph to integrate information from neighboring nodes and edges, progressively updating the internal representation of each target node to better reflect its molecular environment. At each layer of the network, *message passing* refines node representations by aggregating structural and chemical information from their neighbors. Common aggregation functions include *summation*, *mean*, and *maximum* aggregation, which determine how node features are combined. The aggregated vector is then transformed using a learnable weight matrix, which helps the model assign different levels of importance to various features. This transformation allows the network to emphasize the most relevant atomic and bonding properties for the prediction task. Finally, a bias vector and a non-linear activation function, such as ReLU, are applied to enable the model to learn more complex molecular interactions and relationships (5,20). During training, the network adjusts its weights and biases through backpropagation, updating them at each layer to improve molecular feature representations. This process allows the GNN to identify and prioritize the most relevant features, enhancing its accuracy in predicting properties such as solubility or toxicity (1,5). The learning mechanisms of these networks will be explored further in Section 2.4. Unlike traditional approaches like molecular fingerprints or SMILES, which rely on predefined, static descriptors, GNNs dynamically learn representations directly from the raw structural data of molecular graphs. This adaptive learning process enables GNNs to capture intricate chemical relationships, such as non-covalent interactions and electronic resonance effects, which are often overlooked by conventional methods. As a result, GNNs provide a flexible and highly accurate framework for molecular property prediction (1,9).

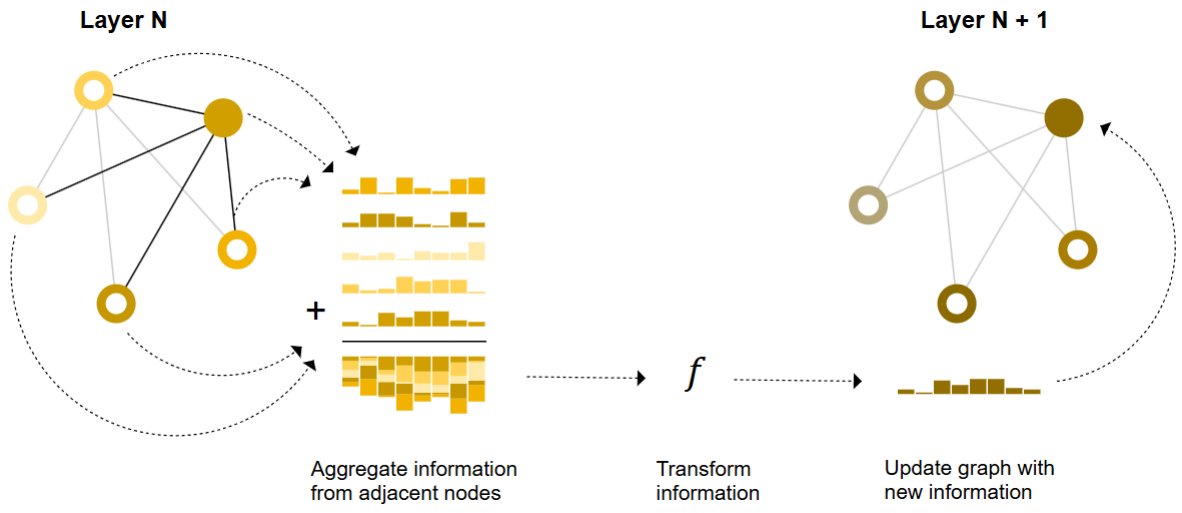


Figure 5: Schematic representation of the *message-passing* process in a GNN. In Layer N, information from neighboring nodes is aggregated and transformed through a learnable weight matrix and a non-linear activation function, updating the node representations in Layer N+1. The updated graph incorporates the newly computed node embeddings, facilitating the modeling of relationships and patterns within the graph. This can be repeated for however many layers there are in the graph neural network. Taken from (18).

2.3 Readout phase

After the *message-passing*, where the graph's node and edge vectors are progressively updated, the *readout phase* merges this information into a unique graph-level vector representing the entire graph (Fig.6). This step is essential as it transforms the detailed embeddings generated during *message passing* into a representation suitable for predictions of molecule-level properties (18).

In this step, pooling layers play a central role by summarizing node-level information (and in some cases, edge-level) into a simpler graph representation. Traditional pooling techniques include *summation*, *averaging*, or selecting the *maximum* value, which are computationally efficient and invariant to the order of nodes in the graph. For example, *summation* pooling aggregates the embeddings of all nodes, producing a comprehensive representation that reflects the contributions of each atom in the molecular graph. Beyond these basic techniques, more advanced pooling strategies, such as attention-based pooling, dynamically assign weights to nodes based on their relevance to the task, improving the model's ability to focus on the most informative parts of the molecular graph. Finally, the resulting vector serves as the input for predicting molecular properties (21).

In molecular property prediction, this *readout phase* typically produces a *global* prediction, meaning that the entire graph is condensed into a single representation for tasks such as regression or classification. However, in other contexts, GNNs can also perform *node-level* predictions, where each node's embedding is used to make individual predictions, such as identifying specific functional groups within a molecule. Additionally, *edge-level* predictions can be made, where relationships between nodes, such as bond properties or molecular interactions, are analyzed independently. This flexibility allows GNNs to be applied in a wide range of tasks beyond molecular property prediction, such as protein structure analysis or social network modeling (12,16).

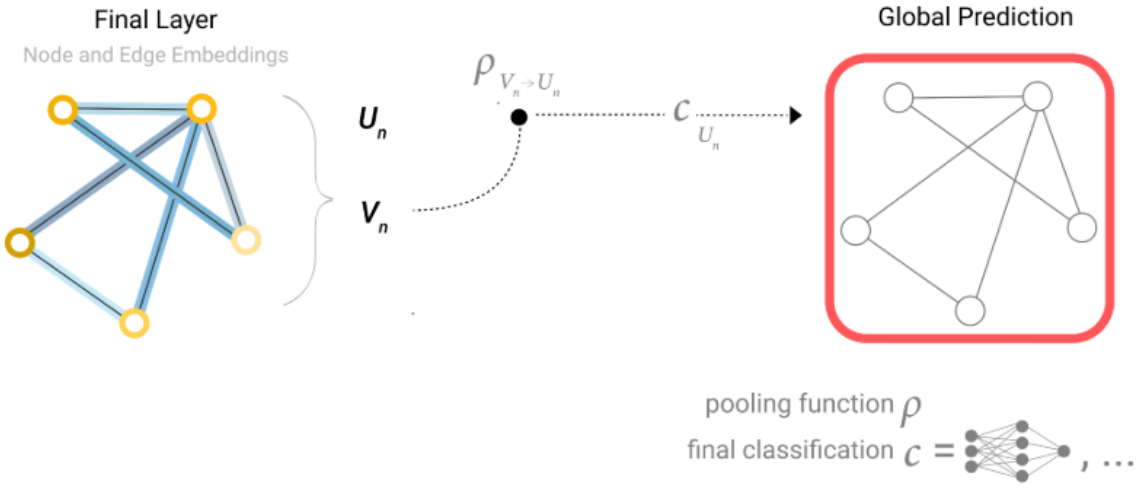


Figure 6: Diagram of the pooling process in GNNs, where node and edge embeddings are combined into a global vector using a pooling function, which is then employed for molecular property predictions in a classification or regression task. Taken from (18).

2.4 End-to-end learning

Graph Neural Networks (GNNs) are a clear example of end-to-end learning, as they integrate all stages of the process into a unified model optimized by backpropagation. This process ensures that all parameters, including those governing *message passing* and *readout*, are continuously adjusted to minimize the error between predictions and actual labels (Fig.7) (4,15).

During training, backpropagation propagates the error from the output layer backward through the network, updating the model's parameters at each step. In the *message-passing*, this process refines the weights used to aggregate and transform information from neighboring nodes and edges, enabling the model to emphasize the most relevant local and global features. Likewise, in the *readout phase*, using classic pooling techniques such as *summation*, *mean*, or *max* pooling, the pooling operation itself does not involve learnable parameters. However, backpropagation optimizes the weights in the final transformation layers, such as a fully connected neural network that processes the global graph representation (1,5).

This integrated learning approach removes the need for manually defined molecular descriptors, allowing GNNs to learn directly from data and uncover complex patterns in a more dynamic and adaptive way. By jointly optimizing all components, these models enhance their ability to predict molecular properties with high accuracy and efficiency (13).

However, while this end-to-end framework offers significant advantages, it also presents challenges. One issue is *over-smoothing*, where many layers cause node representations to become too similar, reducing the model's ability to distinguish unique molecular structures. Additionally, the scalability of GNNs remains a concern, as large and intricate molecular graphs require significant computational resources (15). In the next section, we will delve deeper into these and other limitations, exploring ongoing research aimed at addressing these issues.

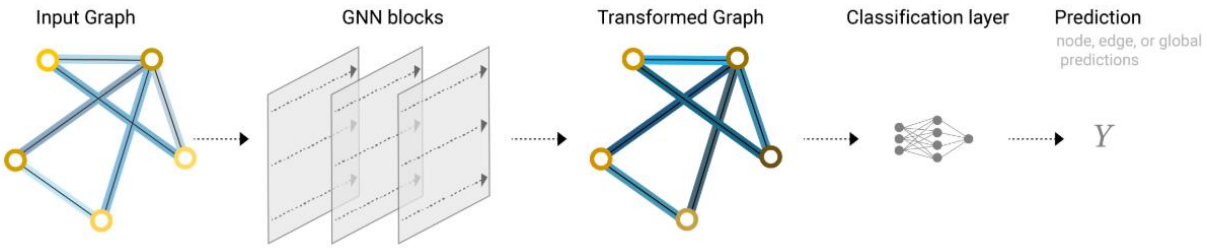


Figure 7: Schematic of end-to-end learning in GNNs. Starting from the input graph, GNN blocks transform the representations of nodes and links. This transformed graph goes through a classification layer that generates the final prediction. For molecular property predictions, this global representation is typically used. Taken from (18).

3. Limitations and GNN variations for solving them

3.1 Scalability

As mentioned at the end of the previous section, one of the main challenges of Graph Neural Networks (GNNs) is their scalability. When molecular graphs increase in size and complexity, the computational cost required to process them grows significantly. This is particularly problematic in drug discovery, where datasets often contain thousands of nodes (atoms) and edges (connections between atoms). For example, large molecules like paclitaxel ($C_{47}H_{51}NO_{14}$), a chemotherapy drug, have a molecular graph with 113 non-hydrogen atoms and approximately 122 bonds, including complex ring structures and branches. As graphs like these expand, the number of messages exchanged during the message-passing phase in GNNs increases significantly, leading to computational bottlenecks that hinder model efficiency (22–24).

A key component of HiGNN is its *Feature-Wise Attention* mechanism, which enhances predictive accuracy by assigning greater importance to the most relevant atomic features after the *message-passing* phase. Instead of treating all features equally, this mechanism prioritizes those with the most impact on the model's final prediction, allowing HiGNN to focus on key molecular characteristics. Another innovation is the BRICS algorithm (Breaking of Retrosynthetically Interesting Chemical Substructures), which decomposes molecules into chemically meaningful fragments by breaking bonds according to predefined chemical rules. This reduces graph complexity while preserving essential chemical information, improving the model's ability to predict molecular properties and enhancing interpretability by highlighting structurally relevant substructures.

Extensive experiments demonstrate that HiGNN achieves state-of-the-art results on datasets like ESOL (RMSE 0.532), outperforming both traditional GNNs (GCN, RMSE 0.708) and non-GNN models like Random Forest (RF, RMSE ~0.582) and Support Vector Machines (SVM, RMSE ~0.600). These results confirm HiGNN's ability to efficiently process complex molecular datasets while maintaining high interpretability (6).

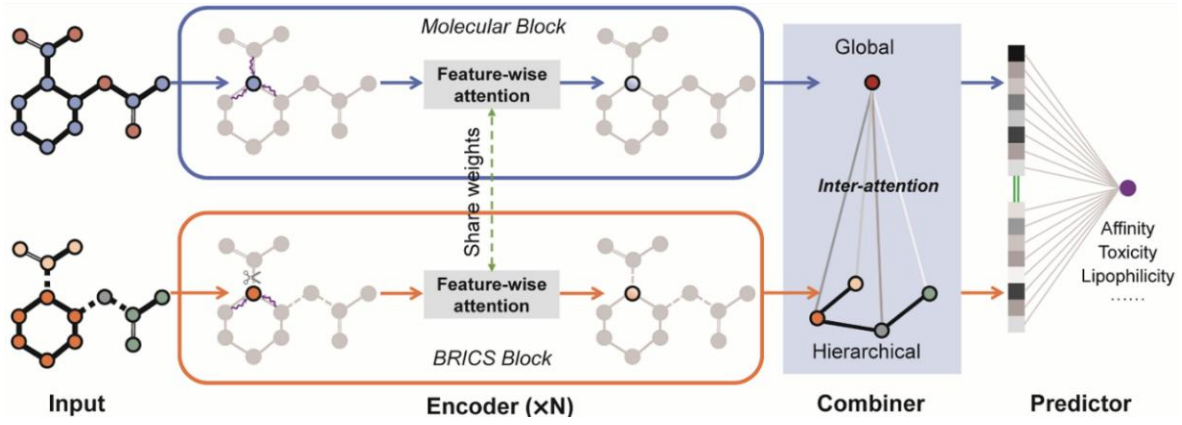


Figure 8. Architecture of the HiGNN model. The *Molecular Block* (top) processes the entire molecular graph using a *Feature-Wise Attention* mechanism to highlight important atomic features. The *BRICS Block* (bottom)

decomposes the molecule into chemically meaningful fragments using its algorithm (indicated by the scissor symbol), which cuts specific bonds based on chemical rules. Weights are shared between the two blocks, ensuring consistency between global and fragment-level representations. These outputs are combined through an Inter-Attention mechanism, allowing the model to predict properties like toxicity with high accuracy. Taken from (6).

Another innovative strategy in graph learning is adaptive sampling, as implemented in GraphSAINT (Graph Sampling-Based Inductive Learning Method). This method reduces computational load by dividing large graphs into smaller, manageable subgraphs through node, edge, or subgraph sampling (Fig. 9). During this process, GraphSAINT prioritizes influential nodes to preserve essential relationships and applies normalization techniques to mitigate sampling biases.

GraphSAINT is primarily used for node classification tasks, such as predicting product categories in the Amazon dataset, which models co-purchase patterns. On this dataset, GraphSAINT achieved an F1 score of 0.815, significantly outperforming GCN, which scored 0.281. The F1 score balances precision and recall, evaluating how well the model distinguishes true relationships from false positives. Additionally, GraphSAINT enhanced training efficiency, reducing computational costs by over 40% compared to models like Graph Attention Networks (GATs) (23).

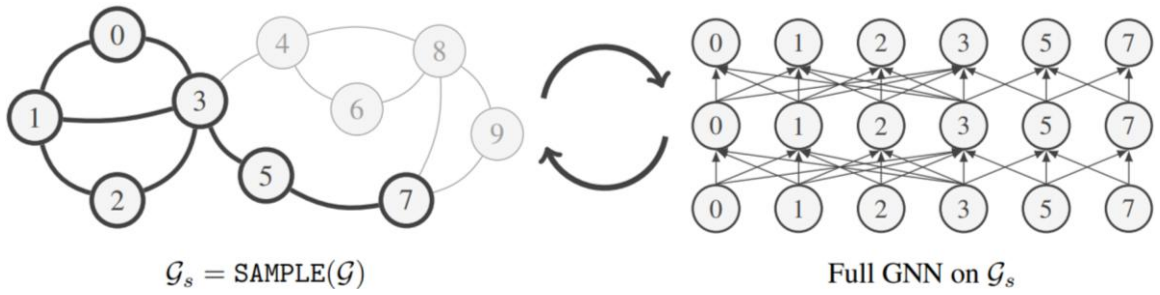


Figure 9: Illustration of the GraphSAINT sampling method. The figure demonstrates how a large graph is sampled to create a subgraph, which is then used for training a full GNN. This sampling approach reduces the computational burden by focusing on smaller subgraphs while preserving the key structural and relational information of the original graph. The method ensures efficient and scalable training for large-scale graph neural networks. Taken from (23).

3.2 Over-smoothing

Over-smoothing is another fundamental challenge in Graph Neural Networks (GNNs), which occurs when node representations become too similar to each other after several layers of *message passing*, causing the loss of distinguishing features. As a result, nodes that should retain unique characteristics based on their connections and attributes end up sharing overly homogeneous representations, reducing the model's ability to capture relevant relationships in the graph and make accurate predictions of drug's solubility, toxicity, or binding affinity (18).

To mitigate these effects, several strategies have been proposed. A common solution is limiting the depth of GNNs to reduce excessive message propagation between nodes. However, while this approach can help reduce *over-smoothing*, it also restricts the model's ability to capture long-range

dependencies, which are crucial in tasks such as molecular property prediction (18). A more effective alternative is DropEdge, a method that randomly removes certain connections between nodes during training (Fig.10). By reducing excessive message propagation, DropEdge allows models to maintain greater differentiation between nodes without losing generalization capacity. In node classification tasks conducted on the Cora dataset, which contain nodes representing scientific articles and edges representing citations between them, DropEdge increased the accuracy of an 8-layer GCN from 78.7% to 85.8% and a 32-layer GCN from 71.6% to 74.6% (25). These results demonstrate how DropEdge can preserve structural information in deeper GNNs, enhancing their ability to differentiate nodes and make more accurate predictions (25).

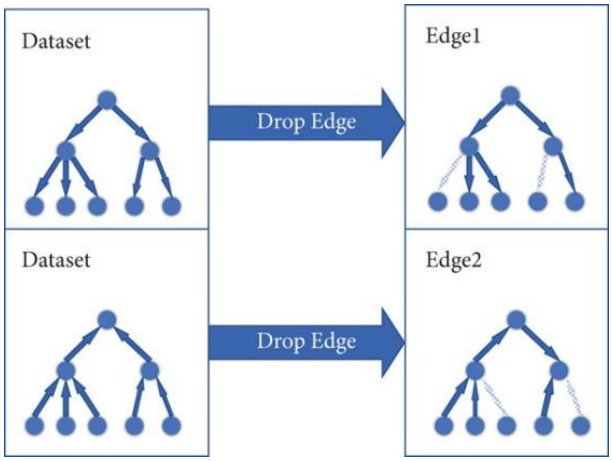


Figure 10: Schematic illustration of the DropEdge technique. The diagram shows how DropEdge operates by randomly removing edges from the input graph during training. This process reduces the extent of message propagation between nodes, thereby mitigating the over-smoothing phenomenon. By selectively dropping edges (Edge1 and Edge2), the technique ensures that node representations retain diversity, which improves the model's generalization capabilities. Taken from (26).

3.3 Stereochemistry

Another challenge for Graph Neural Networks (GNNs) is their difficulty in accurately capturing stereochemical information, such as chirality, which is essential for the three-dimensional characterization of molecules. Chirality, defined as the property of a molecule being non-superimposable on its mirror image, directly influences the biological activity and physicochemical properties of chemical compounds. This limitation affects the predictive capacity of GNNs in key tasks such as ligand binding affinity, toxicity, and solubility, where spatial orientation and atomic interactions are crucial. For example, a slight change in a molecule's three-dimensional structure can significantly alter its interaction with a biological receptor, leading to either therapeutic or adverse effects. These limitations directly impact the real-world applicability of GNNs (27). To address these issues, variants of GNNs have been developed to integrate geometric information into their models.

One of the most notable solutions is the Equivariant Graph Neural Network (EGNN). Unlike traditional GNNs, which represent molecular structures solely as undirected two-dimensional graphs, EGNNs incorporate three-dimensional spatial coordinates as additional node features within the message-passing mechanism (Fig.11). These coordinates are processed in a way that ensures equivariance rather than invariance. This means the network accounts for transformations like rotations and translations by adjusting how messages are passed and aggregated, rather than requiring multiple representations of the same molecule. Importantly, the spatial coordinates are treated as relative distances to neighboring atoms, preserving the local geometry of the molecule without being affected by its absolute position or orientation in space. This property enables EGNNs to more accurately capture three-dimensional molecular interactions that influence reactivity and chemical properties. In a study conducted on the QM9 dataset, widely used in quantum chemistry, EGNNs were evaluated for their ability to predict various molecular properties, such as dipole moment, energy levels, heat capacity, and enthalpy. Compared to standard GNNs such as Graph Convolutional Networks (NMP, SchNet, L1Net, LieConv, and TFN) and Graph Attention Networks (Cormorant and SE(3)-Tr), EGNNs achieved superior performance in 9 out of 12 property prediction tasks, with significantly lower mean absolute errors (MAE) (Table 3). In the remaining three tasks, EGNNs achieved the second-best performance, further highlighting their ability to capture key three-dimensional molecular information.

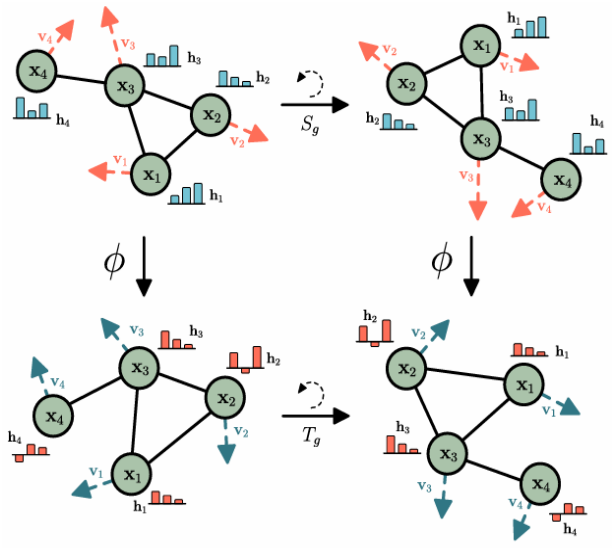


Figure 11: Illustration of rotation equivariance in a graph neural network. The diagram shows how the network maintains rotation equivariance by recalculating relative distances and directions between nodes under spatial transformations. The top row depicts the original graph Sg and its rotated version, with nodes (x1, x2, x3, x4) and their embeddings (h1, h2, h3, h4) adapting to the new orientation. The bottom row shows the corresponding transformations ϕ applied to node embeddings and edge vectors to preserve geometric relationships. This ensures that learned features remain consistent, enabling accurate modeling of three-dimensional molecular structures. Taken from (7).

Task Units	α bohr^3	$\Delta\varepsilon$ meV	$\varepsilon_{\text{HOMO}}$ meV	$\varepsilon_{\text{LUMO}}$ meV	μ D	C_ν cal/mol K	G meV	H meV	R^2 bohr^3	U meV	U_0 meV	ZPVE meV
NMP	.092	69	43	38	.030	.040	19	17	.180	20	20	**1.500**
Schnet	.235	63	41	34	.033	.033	14	14	**.073**	19	14	1.700
Cormorant	.085	61	34	38	.038	**.026**	20	21	.961	21	22	2.027
L1Net	.088	68	46	35	.043	.031	14	14	.354	14	13	1.561
LieConv	.084	49	30	**25**	.032	.038	22	24	.800	19	19	2.280
TFN	.223	58	40	38	.064	.101	-	-	-	-	-	-
SE(3)-Tr.	.142	53	35	33	.051	.054	-	-	-	-	-	-
EGNN	**.071**	**48**	**29**	**25**	**.029**	.031	**12**	**12**	.106	**12**	**11**	1.554

Table 3: Mean Absolute Error (MAE) for predicting molecular properties using the QM9 dataset. Each column represents a specific property: polarizability (α), electronic energy gap (Δε), HOMO energy (εHOMO), LUMO energy (εLUMO), dipole moment (μ), heat capacity (C_v), free energy (G), enthalpy (H), spatial extent (R^2), internal energy (U, U_0), and zero-point vibrational energy (ZPVE). EGNN achieved state-of-the-art performance in 9 out of 12 tasks, including α and Δε, with the lowest MAE values. For the remaining three tasks, it ranked second-best. Some models, such as TFN and SE(3)-Transformer, were not run for certain properties like G or H due to computational constraints or task-specific limitations. This highlights EGNN's efficiency in capturing molecular geometry while handling a wide range of molecular properties. Taken from (7).

3.4 Interpretability

Despite the effectiveness of Graph Neural Networks (GNNs) in tasks such as molecular property prediction, determining which specific features or relationships most influence a given prediction remains a challenge. This lack of interpretability complicates the validation of results. For instance, if a model predicts that a molecule is toxic or has low solubility, researchers must understand which structural aspects or chemical interactions led to this prediction. Without interpretability, it is difficult to trust these models and refine them for critical applications such as drug discovery (28).

To address this limitation, several interpretability techniques have been explored. Among them, Graph Attention Networks (GATs) introduce attention mechanisms that assign different weights to node connections, prioritizing the most relevant relationships. This enhances both model performance and interpretability by identifying which atomic interactions contribute most to predictions (10). For example, in solubility prediction tasks, GATs were trained on approximately 10,000 organic molecules, and by analyzing their attention weights, it was possible to determine that these type of models could highlight key molecular features, such as hydrophilic functional groups, which play a crucial role in solubility, demonstrating it capacity to provide insights into the molecular substructures that drive model decisions (9).

Another promising approach is GNNExplainer, which provides post-hoc explanations for GNN predictions without modifying the model's architecture (Fig.12). Unlike GATs, which offer global interpretability through attention scores, GNNExplainer generates instance-specific explanations by identifying the smallest subgraph and subset of node features that significantly influence a particular prediction. To achieve this, GNNExplainer optimizes a differentiable mask function that assigns weights to nodes and edges, learning which components should be retained for explanation purposes. In experiments on molecular datasets like MUTAG, GNNExplainer effectively identified functional groups such as NO_2 and NH_2, which are strong indicators of mutagenicity. The base model used in these experiments was a Graph Convolutional Network (GCN) trained for molecular classification. The masking function was optimized by iteratively removing nodes and edges while observing if the model's prediction remained stable, thus learning to highlight the most relevant substructures (28).

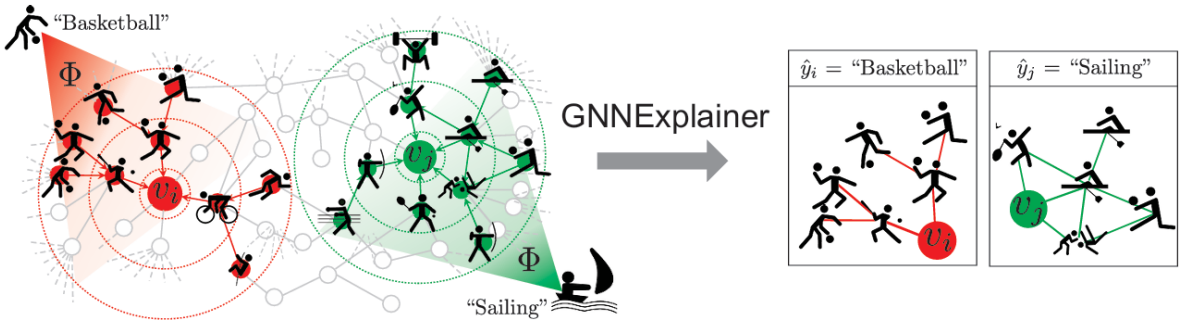


Figure 12: Illustration of the GNNExplainer interpretability approach. The image shows how GNNExplainer provides interpretable explanations for predictions made by GNN models. In this example, a GNN trained on a social interaction graph predicts future sport activities. For node vi ("Basketball"), GNNExplainer identifies a relevant subgraph and key features (e.g., links to other sports involving balls) that influenced the prediction. Similarly, for node vj ("Sailing"), the explanation highlights substructures and relationships associated with water and beach sports. This approach enhances the transparency and interpretability of GNN predictions, making them more comprehensible for complex graph-based tasks. Taken from (28).

4. Discussion and conclusion

Graph Neural Networks (GNNs) have revolutionized molecular property prediction by modeling molecules as graph structures, which allows capturing structural and spatial relationships more accurately than traditional descriptor-based methods (29). This review has examined the fundamental principles of GNNs in property prediction, their key limitations, and the adaptations developed to address them. However, several open questions persist.

Have Previous Problems Been Solved?

Despite advancements in scalability, such as hierarchical representations, computational demands remain a significant challenge when managing vast chemical libraries in industrial applications (6). Additionally, adaptive sampling methods like GraphSAINT may introduce biases if high-importance nodes, those with many connections or critical roles, are not appropriately included (23). Similarly, while DropEdge has been introduced to mitigate *oversmoothing*, its effects on overfitting and memory consumption in deeper networks require further exploration (25). Stereochemistry remains another significant hurdle, particularly in cases where molecular flexibility and conformational changes critically influence bioactivity. Even models incorporating equivariant mechanisms still struggle to accurately represent dynamic molecular structures, limiting their generalization across diverse chemical environments (7). Finally, interpretability continues to be a crucial challenge, especially in regulatory settings where transparency is essential for decision-making. Although methods like GNNExplainer and Graph Attention Networks (GATs) have improved interpretability, they remain insufficient for fully explaining how molecular features contribute to predictions, as they primarily focus on node and edge relevance but often fail to capture higher-order dependencies within molecular substructures (9,10,28).

Are Graph-Based Methods Better Than Descriptor-Based Methods?

Graph Neural Networks (GNNs) provide a flexible framework for modeling molecular structures, capturing connectivity patterns that influence bioactivity and reactivity. Unlike traditional methods, they represent molecules as graphs, where atoms act as nodes and bonds as edges. This allows them to automatically learn molecular features without predefined descriptors. However, most GNNs operate on 2D molecular graphs, meaning they focus on atomic connectivity but do not explicitly model 3D spatial arrangements. This limitation makes it difficult to account for stereochemistry and conformational flexibility, both crucial in many molecular tasks. More advanced architectures, such as Equivariant GNNs (EGNNs), incorporate 3D information but at a significantly higher computational cost. By contrast, descriptor-based methods remain a strong alternative, particularly for tasks where molecular properties can be effectively encoded using predefined features. Machine learning models like Support Vector Machines (SVMs) or Random Forest (RF) leverage molecular descriptors, which often provide sufficient information for predicting physicochemical properties. These models are typically more computationally efficient than GNNs and have demonstrated higher accuracy in many regression and classification tasks. Additionally, descriptor-based models offer better interpretability, as techniques like SHAP (Shapley Additive Explanations) allow researchers to identify which molecular features drive prediction. Despite their differences, both approaches have their strengths. Therefore, the choice between GNNs and descriptor-based models depends on the complexity of the problem. When molecular connectivity plays a crucial role and predefined descriptors are insufficient, GNNs may offer an advantage. However, for many applications, descriptor-based models remain a good option due to their higher efficiency, accuracy, and interpretability (14).

What Are the Future Directions?

Beyond addressing the limitations discussed in this review, research in GNNs is evolving in multiple directions. One major challenge is improving generalization to novel molecules, as many GNNs struggle with compounds outside their training distribution, leading to performance drops when encountering new chemical scaffolds. To address this, researchers are exploring contrastive learning and self-supervised learning, which enhances model robustness by learning invariant molecular representations across diverse datasets. Contrastive learning trains a model to distinguish between similar and dissimilar molecules by maximizing the similarity between different representations of the same molecule while ensuring distinct molecules remain separate in the learned space. Self-supervised learning, on the other hand, eliminates the need for large, labeled datasets by employing auxiliary prediction tasks, such as estimating electronic properties or molecular forces, allowing the model to extract meaningful molecular features before fine-tuning on downstream prediction tasks (29). Another key focus is reducing computational costs, as GNNs remain resource intensive. Strategies such as subgraph sampling techniques like GraphSAINT or hierarchical pooling like HiGNN mitigate scalability issues by reducing redundant computations while preserving structural information (23). Additionally, hybrid approaches that integrate GNNs with descriptor-based methods or physics-informed models are gaining attention. These models leverage the efficiency and interpretability of traditional descriptors while benefiting from the structural learning capabilities of deep learning models, leading to improvements in molecular property prediction tasks (8).

Conclusion: Take-away Idea.

Graph Neural Networks (GNNs) have redefined molecular property prediction, providing an unprecedented level of accuracy in modeling chemical structures. However, despite their potential, challenges such as scalability, interpretability, and computational efficiency still hinder their widespread adoption. Overcoming these limitations requires continuous research and innovation. The key takeaway is that while GNNs are not yet a universal solution, they are rapidly evolving into a fundamental tool for computational chemistry and drug discovery. With the right advancements, they have the potential to become the standard approach for molecular property prediction and beyond, driving significant progress in the pharmaceutical and chemical industries.

5. Acknowledgments

I would like to thank my supervisor, Dieter Stoker, for his guidance, support, and advice throughout the preparation of this literature review. His help has been essential, and I am grateful for the opportunity to learn under his supervision. I also want to acknowledge the use of AI-based tools, specifically ChatGPT, which played a key role in supporting this work. These tools helped translate text from Spanish to English, refine the clarity and style of the writing, and suggest improvements to the structure and flow of the content (Fig.13). Additionally, they assisted in identifying relevant references for specific topics I was exploring. The outputs were carefully reviewed, edited, and cross-checked against peer-reviewed references and original research to ensure academic standards were met. Using these tools allowed me to focus more on developing my ideas and addressing challenges efficiently, while maintaining the originality and integrity of the work.

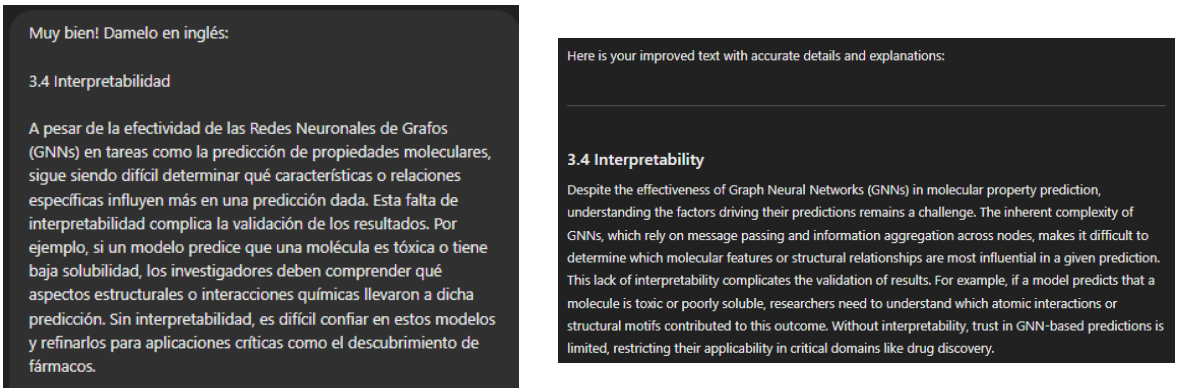


Figure 13: Screenshot of a conversation where I used ChatGPT to translate a text I wrote into English.

6. References

1. Rittig JG, Gao Q, Dahmen M, Mitsos A, Schweidtmann AM. Graph neural networks for the prediction of molecular structure-property relationships. 2022 Jul 25; Available from: http://arxiv.org/abs/2208.04852

2. Cremer J, Medrano Sandonas L, Tkatchenko A, Clevert DA, De Fabritiis G. Equivariant Graph Neural Networks for Toxicity Prediction. Chem Res Toxicol. 2023;

3. Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2021 Apr 27; Available from: http://arxiv.org/abs/2104.13478

4. Wang Y, Li Z, Barati Farimani A. Graph Neural Networks for Molecules. In: Challenges and Advances in Computational Chemistry and Physics. Springer Science and Business Media B.V.; 2023. p. 21–66.

5. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. 2017 Apr 4; Available from: http://arxiv.org/abs/1704.01212

6. Zhu W, Zhang Y, Zhao D, Xu J, Wang L. HiGNN: A Hierarchical Informative Graph Neural Network for Molecular Property Prediction Equipped with Feature-Wise Attention. J Chem Inf Model. 2023 Jan 9;63(1):43–55.

7. Satorras VG, Hoogeboom E, Welling M. E(n) Equivariant Graph Neural Networks. 2021 Feb 19; Available from: http://arxiv.org/abs/2102.09844

8. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nat Biomed Eng. 2022 Dec 1;6(12):1353–69.

9. Ahmad W, Tayara H, Chong KT. Attention-Based Graph Neural Network for Molecular Solubility Prediction. ACS Omega. 2023 Jan 24;8(3):3236–44.

10. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. 2017 Oct 30; Available from: http://arxiv.org/abs/1710.10903

11. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021 Dec 1;13(1).

12. Wu Y, Ni X, Wang Z, Feng W. Enhancing drug property prediction with dual-channel transfer learning based on molecular fragment. BMC Bioinformatics. 2023 Dec 1;24(1).

13. Todeschini & Consonni. QSPR/QSAR Analysis Using SMILES and Quasi-SMILES [Internet]. Toropova AP, Toropov AA, editors. Cham: Springer International Publishing; 2023. (Challenges and Advances in Computational Chemistry and Physics; vol. 33). Available from: https://link.springer.com/10.1007/978-3-031-28401-4

14. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021 Dec 1;13(1).

15. Wang Z, Liu M, Luo Y, Xu Z, Xie Y, Wang L, et al. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. Bioinformatics. 2022 May 1;38(9):2579–86.

16. Yang J, Cai Y, Zhao K, Xie H, Chen X. Concepts and applications of chemical fingerprint for hit and lead screening. Vol. 27, Drug Discovery Today. Elsevier Ltd; 2022.

17. Wang H, Zhang A, Zhong Y, Tang J, Zhang K, Li P. Chain-aware graph neural networks for molecular property prediction. Wren J, editor. Bioinformatics [Internet]. 2024 Oct 1;40(10). Available from: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btae574/7818417

18. Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko A. A Gentle Introduction to Graph Neural Networks. Distill. 2021 Aug 17;6(8).

19. Corso G, Cavalleri L, Beaini D, Liò P, Veličković P. Principal Neighbourhood Aggregation for Graph Nets. 2020 Apr 12; Available from: http://arxiv.org/abs/2004.05718

20. Lu Y, Chen C, Huang K, Zitnik M, Xu M, Wang Q. GNN 101.

21. Filippo Maria Bianchi. An introduction to pooling in GNNs. 2024 [cited 2025 Jan 8]; Available from: https://gnn-pooling.notion.site/

22. Zhong Z, Li CT, Pang J. Hierarchical Message-Passing Graph Neural Networks. 2020 Sep 8; Available from: http://arxiv.org/abs/2009.03717

23. Zeng H, Zhou H, Srivastava A, Kannan R, Prasanna V. GraphSAINT: Graph Sampling Based Inductive Learning Method. 2019 Jul 10; Available from: http://arxiv.org/abs/1907.04931

24. National Center for Biotechnology Information (NCBI). Paclitaxel - PubChem [Internet]. 2025 [cited 2025 Feb 8]. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Paclitaxel

25. Rong Y, Huang W, Xu T, Huang J. DropEdge: Towards Deep Graph Convolutional Networks on Node Classification. 2019 Jul 25; Available from: http://arxiv.org/abs/1907.10903

26. Xu D, Liu Q, Zhu L, Tan Z, Gao F, Zhao J. GCNRDM: A Social Network Rumor Detection Method Based on Graph Convolutional Network in Mobile Computing. Wirel Commun Mob Comput. 2021;2021.

27. Pattanaik L, Ganea OE, Coley I, Jensen KF, Green WH, Coley CW. Message Passing Networks for Molecules with Tetrahedral Chirality. 2020 Nov 23; Available from: http://arxiv.org/abs/2012.00094

28. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating Explanations for Graph Neural Networks. 2019 Mar 9; Available from: http://arxiv.org/abs/1903.03894

29. Pang C, Tong HHY, Wei L. Advanced deep learning methods for molecular property prediction. Vol. 11, Quantitative Biology. John Wiley and Sons Inc; 2023. p. 395–404.