Accelerating Drug Discovery with Graph Neural Networks and Molecular Property Prediction

Article · June 2025

8 authors, including:

Rapheal Alamu
Obafemi Awolowo University
74 PUBLICATIONS **52** CITATIONS

SEE PROFILE

Accelerating Drug Discovery with Graph Neural Networks and Molecular Property Prediction

Authors: Raphael Alamu, Sudarshana Karkala, Sazzad Hossain, Mahendra Krishnapatnam, Ankur Aggarwal, Zarif Zahir, Harshad Vijay Pandhare, Varun Shah

Date: 12 June 2025

Abstract: ***This research article provides a comprehensive review of the transformative role of Graph Neural Networks (GNNs) in accelerating drug discovery, with a particular focus on molecular property prediction. Traditional drug discovery faces significant challenges, including high costs, lengthy timelines, and low success rates, largely due to the inherent complexity of biological systems and the limitations of conventional computational methods. GNNs offer a powerful paradigm shift by treating molecules as graphs, enabling the direct capture of complex structural and spatial relationships. The fundamental principles of GNNs, their architectural variants, and their profound impact on predicting crucial molecular properties such as solubility, toxicity, and binding affinity are explored. Beyond property prediction, the broader applications of GNNs across the drug discovery pipeline, including de novo molecule generation, drug-target interaction prediction, and virtual screening, are discussed. Despite their immense potential, GNNs in drug discovery face challenges related to data scarcity, interpretability, and scalability. Recent advancements addressing these limitations and outlining future research directions are examined, emphasizing the critical need for robust validation and seamless integration with experimental workflows to realize the full promise of AI-driven pharmaceutical innovation.***

Keywords: ***Graph Neural Networks, Drug Discovery, Molecular Property Prediction, Artificial Intelligence, Cheminformatics, Deep Learning, Computational Chemistry***

---

1. Introduction

The pursuit of new therapeutic agents is a cornerstone of modern medicine, yet the process of drug discovery remains one of the most challenging and resource-intensive endeavors in scientific research. Historically, this arduous journey has been characterized by protracted timelines, exorbitant costs, and a high probability of failure. The advent of artificial intelligence (AI) and machine learning (ML) has heralded a new era, offering unprecedented opportunities to revolutionize pharmaceutical research. Among the diverse AI methodologies, Graph Neural Networks (GNNs) have emerged as a particularly promising approach, uniquely suited to the intricate nature of molecular data. This article provides a comprehensive overview of how GNNs are

fundamentally transforming drug discovery, with a specific emphasis on their role in accelerating molecular property prediction.

1.1 The Traditional Drug Discovery Paradigm: Challenges and Limitations

Drug development is a notoriously long, costly, and high-risk endeavor. The average time required to bring a new drug to market typically ranges from 10 to 15 years, with an average cost that can exceed $2 billion.[1] This extensive and meticulous pipeline is plagued by increasing costs and significant challenges that contribute to its inefficiencies.[2]

A major hurdle in this process is the alarmingly high attrition rate. Only approximately 4% to 10% of drug development programs ultimately result in licensed drugs.[1] A primary cause of these failures is often attributed to a lack of demonstrable efficacy, stemming from early missteps during target identification or validation.[1] This is profoundly complicated by the immense biological complexity of human systems, where a drug target that appears promising in

in vitro experiments may exhibit unforeseen interactions or no observable efficacy within a whole organism.[3] The traditional "one gene, one drug, one disease" paradigm frequently falls short in capturing the polypharmacology often observed in clinical settings, where drugs can act on multiple targets.[3]

Furthermore, traditional methods for molecular property prediction, such as manual molecular descriptor design, were inherently prone to bias and demanded extensive expert knowledge, severely limiting their scalability for the vast chemical space.[5] Even the introduction of computational tools like SMILES (Simplified Molecular Input Line Entry System) and molecular fingerprints, while automating aspects of the process, retained certain biases and, critically, lacked the ability to explicitly represent atom-to-atom interactions, leading to suboptimal predictive performances.[5] The challenges are further compounded by a poor understanding of the underlying biological mechanisms of many diseases and the inaccuracies inherent in current disease modeling technologies, which often lead to unsuccessful translation of preclinical model results to human patients.[7]

The cumulative effect of these challenges creates an urgent imperative for more

efficient and accurate methods in drug discovery. The high financial risk associated with traditional approaches often deters investment in truly groundbreaking new therapies, instead incentivizing the development of "me-too" drugs—incremental improvements over existing treatments.[3] This dynamic not only stifles radical innovation but also directly impacts healthcare systems and patient access, contributing to increased prescription drug prices and a disproportionate focus on diseases with large patient populations rather than those with high unmet medical needs.[1] Consequently, the acceleration of drug discovery through advanced computational methods carries significant societal and economic implications, promising more affordable and accessible treatments.

1.2 The Dawn of AI/ML in Pharmaceutical Research

In response to the formidable challenges faced by traditional drug discovery, the integration of deep learning techniques has ushered in a period of remarkable success in molecular property prediction.[6] Artificial intelligence and machine learning offer a transformative potential to address these persistent challenges, providing a promising pathway toward increased efficiency and success rates in drug development by correcting existing limitations and opening entirely new avenues for research.[4]

Machine learning and deep learning methods have been extensively investigated within the domain of molecular structures, recognizing their central role in drug discovery and prediction.[8] These technologies empower researchers to uncover complex patterns, predict outcomes, and significantly accelerate drug development processes that previously demanded months or even years of labor-intensive research.[9] AI-driven methods are fundamentally transforming the pharmaceutical industry, making the entire process faster and more efficient.[10] Deep learning, a sophisticated subset of machine learning, is particularly adept at this transformation. It employs multi-layered artificial neural networks designed to simulate the neural networks of the human brain, enabling the learning of complex data representations. This architecture makes deep learning exceptionally powerful and flexible in handling the complex and high-dimensional data characteristic of molecular biology.[4]

This represents a profound shift in the approach to drug discovery, moving beyond mere incremental improvements to a qualitatively different way of processing

information and driving scientific breakthroughs. The capacity of AI to synthesize vast and complex datasets and generate predictive insights [9] directly addresses the limitations of traditional methods that struggled with the sheer volume and intricacy of molecular information.[11] This paradigm shift suggests a future where computational methods are not merely supplementary tools but become central to the entire drug discovery pipeline, influencing decisions from the initial target identification to the final lead optimization stages. Such a fundamental change also highlights the increasing demand for interdisciplinary expertise, particularly in AI/ML, within pharmaceutical companies.

1.3 Graph Neural Networks: A Transformative Approach for Molecular Data

Among the various AI and ML methodologies, Graph Neural Networks (GNNs) have emerged as an innovative and exceptionally suitable solution for addressing the challenges inherent in molecular property prediction.[5] GNNs operate on a fundamental principle that aligns perfectly with the inherent structure of chemical compounds: they directly model molecules as graphs, where individual atoms are represented as nodes and the chemical bonds connecting them are represented as edges.[5] This graph-based representation allows GNNs to intrinsically capture the complex structural and spatial relationships within a molecule, a capability that traditional computational methods often overlook.[5]

The ability of GNNs to directly analyze the graph structure of molecules enables them to identify intricate patterns, such as how atoms interact within a molecule, which is absolutely essential for accurately predicting molecular properties.[5] This approach is particularly powerful in the biomedical domain [8], where GNN models have consistently demonstrated superior performance compared to traditional methods in the critical task of molecular property prediction.[12] Their capacity to comprehend and process intricate molecular patterns is a key distinguishing feature that underscores their utility and potential.[15]

The success of GNNs stems from their ability to act as the "native language" for molecular data. Molecules naturally lend themselves to graph representations because they are fundamentally composed of atoms interconnected by bonds.[17] Unlike one-dimensional string representations, such as SMILES, which inherently lack explicit

information about atom-to-atom interactions and thus lead to suboptimal performances [6], GNNs directly model these irregular structural features.[6] This direct mapping of chemical topology and geometry is both intuitive and highly expressive [18], allowing GNNs to capture intricate structural and spatial relationships.[5] This inherent suitability means that GNNs are not merely applied to molecular data; they are designed to process it in a way that mirrors its intrinsic organization. This capability allows GNNs to learn directly from the chemical structure itself, rather than relying on pre-computed, and potentially biased, molecular descriptors. The profound consequence is that advancements in GNN theory and architecture are likely to directly translate into significant breakthroughs in molecular science, establishing GNNs as a dominant deep learning architecture for chemical information processing.

1.4 Scope and Organization of This Article

This article provides a comprehensive analysis of the current state of Graph Neural Networks in molecular property prediction, highlighting their capabilities, the persistent challenges they face, and recent advancements aimed at overcoming these limitations.[5] The review consolidates foundational knowledge with the latest developments in the field, serving as a clear and accessible resource for researchers and practitioners seeking to understand and leverage GNNs in drug discovery.[5] Beyond molecular property prediction, the paper also covers the broader applications of GNNs in computational drug development, including

de novo molecule generation and drug-drug interaction prediction.[15] The subsequent sections will delve into the fundamentals of GNNs, their specific applications in accelerating molecular property prediction, their wider impact across the drug discovery pipeline, and finally, current challenges and future research directions.

2. Fundamentals of Graph Neural Networks for Molecular Data

To appreciate the transformative power of Graph Neural Networks in drug discovery, it is essential to understand how these models represent and process molecular

information. Unlike traditional deep learning architectures designed for grid-like data such as images or sequences, GNNs are specifically engineered to handle the irregular, non-Euclidean structure of graphs.

2.1 Representing Molecules as Graphs: Nodes, Edges, and Features

In the GNN paradigm, molecules are explicitly treated as graphs, a representation that naturally aligns with their chemical structure. Atoms within a molecule are represented as individual points, or **nodes**, while the chemical bonds connecting these atoms are represented as **edges**.[5] This approach contrasts sharply with conventional molecular representations that have historically dominated cheminformatics.

Traditional methods, such as SMILES strings and molecular fingerprints, have significant limitations when applied to complex machine learning tasks. SMILES, a compact one-dimensional string representation [6], fails to explicitly encode crucial information about atom-to-atom interactions and complex structural features like rings and functional groups.[6] While efficient for storage and database searching, molecular fingerprints, which are typically fixed-size binary vectors representing substructures or topological patterns, cannot easily encode global molecular features such as overall size and three-dimensional shape.[19]

In contrast, the graph representation allows for a rich encoding of molecular information. Each node (atom) can be associated with a vector of **node features** that describe its chemical properties, such as its element type (often one-hot encoded), formal charge, partial charges (e.g., Gasteiger-Marsili, MMFF94), and stereochemical information like chirality tags.[17] Similarly, each edge (bond) can carry

edge features that describe the bond's properties, such as bond order (single, double, triple, aromatic) or the distance between bonded atoms.[29] Molecules are inherently undirected graphs, meaning chemical bonds have no directional concept, leading to symmetric adjacency matrices.[12] Specialized cheminformatics tools, such as RDKit, are commonly used to convert SMILES strings or other input formats into these detailed graph representations, making them amenable to GNN processing.[17]

The inherent information richness of graph representations is a primary driver of GNNs' success. Molecules are naturally structured as graphs, with their atoms and

interconnecting bonds forming an intuitive network.[17] This direct encoding of chemical topology and geometry is intuitive and expressive [18], providing a more faithful representation of molecular structure than linear strings or pre-computed descriptors. Unlike SMILES, which omits explicit atom-to-atom interaction information, leading to suboptimal performance [6], graph representations directly capture the irregular structural features of atomic interactions [6] and even long-range dependencies within a molecule.[33] This fundamental advantage allows GNNs to learn directly from the intrinsic chemical structure, without relying on potentially biased or incomplete hand-crafted features. The ability to seamlessly incorporate three-dimensional (3D) information, including atomic coordinates, bond angles, and torsion angles, further enriches these representations, making them particularly effective for tasks requiring a deep understanding of molecular geometry.[17] This shift from traditional, often limited, descriptors to direct feature learning from graph structures is a cornerstone of GNNs' superior performance in molecular property prediction.

Representation Type	Description	Strengths for AI/ML	Limitations for AI/ML
SMILES	Linear string notation for chemical structures.	Compact, easy to store and transmit.	Lacks explicit atom-to-atom interaction information; difficult to capture complex 2D/3D relationships; prone to producing invalid molecules during generation.[6]
Molecular Fingerprints	Fixed-size binary vectors representing substructures or topological patterns.	Computationally efficient; good for similarity searches.	Expert-dependent feature engineering; may retain biases; cannot easily encode global features like size/shape; loses fine-grained structural information.[5]
Graph Representation	Atoms as nodes, bonds as edges, with associated features.	Directly captures complex topological and spatial relationships; handles irregular structures;	Requires specialized GNN architectures; can be computationally intensive for very

		naturally incorporates atom-to-atom interactions; can include 3D information.[5]	large graphs; interpretability can be challenging.[14]

Table 1: Evolution of Molecular Representations in Drug Discovery

2.2 Core GNN Architectures and the Message Passing Framework

Graph Neural Networks are a class of deep learning models that extend the principles of neural networks to non-Euclidean domains, enabling them to process data structured as graphs.[17] The primary objective of GNNs is to learn effective, low-dimensional representations (embeddings) of nodes and entire graphs that capture their structural and feature information.[45]

The foundational mechanism underpinning most GNN architectures is the **message passing framework**. This iterative process allows nodes within the graph to exchange and aggregate information with their immediate neighbors, effectively building a comprehensive understanding of the local chemical environment.[5] During each message passing iteration, a node's hidden state is updated based on messages received from its neighbors and its own previous state. This enables the network to construct a detailed and enriched representation of the entire molecule by progressively integrating information from increasingly distant parts of the graph.[5]

Following the message passing phase, a **readout phase** aggregates the updated node embeddings into a single, global molecular embedding.[5] This global vector then serves as the input for downstream tasks, such as molecular property classification (e.g., predicting toxicity) or regression (e.g., predicting solubility).[5]

The message passing framework embodies a powerful inductive bias for molecular data. Chemical interactions are inherently local, with the properties of an atom being heavily influenced by its immediate neighbors and their connections. The iterative message passing process naturally mirrors this local interaction mechanism, allowing GNNs to learn complex relationships between atoms and bonds without the need for

explicit feature engineering.[5] This local aggregation strategy enables GNNs to effectively capture unique structural patterns in molecules, such as rings and functional groups, which are crucial for accurate property prediction.[6] Furthermore, the message passing paradigm allows GNNs to be robust to varying molecular sizes and structures, as the learning mechanism adapts to the local connectivity rather than relying on fixed-size input vectors. This adaptability is a significant advantage in drug discovery, where the chemical space is vast and molecules exhibit immense structural diversity.

2.3 Advanced GNN Variants for Enhanced Molecular Understanding

The foundational message passing framework has given rise to a diverse array of GNN architectures, each designed with specific mechanisms to enhance molecular understanding and predictive power for various chemical tasks. The choice of GNN architecture often depends on the specific molecular property to be predicted and the type of information (e.g., 2D graph, 3D coordinates) available.

- **Graph Convolutional Networks (GCNs):** GCNs were among the pioneering GNNs that applied spectral graph convolutions to effectively learn node representations.[14] They simplify the complex spectral convolution operation into a localized first-order approximation, making them scalable for large graphs.[56] GCNs aggregate information from direct neighbors, enabling them to capture local structural patterns.
- **Message Passing Neural Networks (MPNNs):** Introduced as a unifying framework, MPNNs abstract the commonalities among many existing GNN models.[34] They consist of a message function that defines how information is passed between nodes and an update function that determines how a node's state is updated. MPNNs have demonstrated state-of-the-art results in predicting various quantum chemical properties of molecules, highlighting their effectiveness in learning representations from graph data.[34]
- **Graph Attention Networks (GATs):** GATs address a limitation of earlier GNNs by introducing masked self-attentional layers.[5] Instead of assigning uniform weights or pre-defined filters, GATs allow each node to implicitly specify different weights to its neighbors based on their features, enabling the model to focus on the most relevant parts of the neighborhood for aggregation.[58] This attention mechanism

enhances the model's ability to capture complex relationships and improves its applicability to inductive problems where the graph structure is not known upfront.[57]

- **Equivariant GNNs (EGNNs):** For tasks involving 3D molecular structures, standard GNNs that operate on 2D graphs may lose crucial geometric information. EGNNs are specifically designed to preserve geometric properties, such as invariance to rotations and translations of the molecule in 3D space.[5] These networks incorporate geometric vector features and spherical harmonic convolutions, making them highly effective for tasks like predicting properties that depend on precise 3D atomic arrangements, such as bond angles and torsion angles.[14]
- **Other Notable Architectures:**
  - **GraphSAGE:** This inductive framework learns a function that generates node embeddings by sampling and aggregating features from a node's local neighborhood, making it particularly useful for large, evolving graphs where not all nodes are present during training.[17]
  - **SchNet:** This architecture employs continuous-filter convolutional layers to model quantum interactions in molecules. It is designed to respect fundamental quantum-chemical principles, providing rotationally invariant energy predictions and achieving state-of-the-art performance on benchmarks for quantum chemical properties.[34]

The architectural diversity of GNNs is a testament to the ongoing effort to tailor these models for specific chemical challenges. Different GNN architectures are designed to address the nuances of various molecular properties or structural patterns. For example, the need to explicitly account for 3D geometry led to the development of Equivariant GNNs, while the desire to capture long-range dependencies more effectively spurred the creation of Graph Attention Networks.[5] This continuous evolution and specialization of GNNs underscore their adaptability and increasing sophistication in tackling complex problems in molecular science.

Architecture	Core Mechanism	Key Advantages for Molecular Data
GCN (Graph Convolutional Network)	Spectral graph convolutions, localized first-order approximation.	Efficiently learns node representations; captures local structural patterns; scalable for large graphs.[14]

MPNN (Message Passing Neural Network)	General framework for message passing and node updating.	Unifies many GNN models; highly effective for quantum chemical property prediction; learns from graph topology.[34]
GAT (Graph Attention Network)	Masked self-attentional layers to assign dynamic weights to neighbors.	Learns importance of neighboring nodes; robust to varying node degrees; applicable to inductive problems.[5]
EGNN (Equivariant GNN)	Preserves geometric properties (rotation, translation) in 3D molecular graphs.	Crucial for tasks sensitive to 3D geometry; incorporates spatial information directly.[5]
GraphSAGE	Learns aggregation functions by sampling and aggregating features from local neighborhoods.	Inductive capability for unseen nodes; efficient for large, evolving graphs.[60]
SchNet	Continuous-filter convolutional layers for modeling quantum interactions.	Respects quantum-chemical principles; rotationally invariant energy predictions; state-of-the-art for quantum properties.[62]

Table 2: Key GNN Architectures for Molecular Property Prediction

3. Accelerating Molecular Property Prediction with GNNs

Molecular property prediction (MPP) is a cornerstone of drug discovery, serving as a critical bottleneck in the traditional pipeline. Graph Neural Networks have emerged as powerful tools to accelerate this process, offering unprecedented accuracy and efficiency in predicting a wide array of chemical and biological characteristics.

3.1 The Critical Role of Molecular Property Prediction in Drug Development

Molecular property prediction is a key component in the early stages of drug development, enabling accurate and early assessments of the chemical and biological characteristics of potential drug candidates.[5] This crucial step optimizes the selection process from vast molecular libraries and facilitates the early elimination of compounds with undesirable or adverse profiles, thereby minimizing risks before advancing to later, more costly stages.[5]

The properties predicted are diverse and encompass a wide range of physicochemical and biological attributes. These include, but are not limited to, solubility (a key pharmacokinetic property influencing drug absorption and distribution), toxicity (predicting potential adverse effects), binding affinity (how strongly a molecule binds to a target protein), molecular stability, and reactivity.[8] The ability to accurately predict these properties

in silico allows researchers to prioritize promising compounds, design more effective experiments, and significantly reduce the time and resources expended on synthesizing and testing molecules that are unlikely to succeed.

3.2 Key Applications and Predictive Successes

GNNs have demonstrated remarkable success in accelerating molecular property prediction across various domains. By learning representations of r-radius subgraphs, often referred to as molecular fingerprints, GNNs can capture the essence of molecular fragments that are critical for specific properties.[67] This end-to-end learning approach does not require pre-defined input features or descriptors, a significant departure from traditional chemoinformatics methods.[67]

Specific applications and predictive successes include:

- **Drug Efficacy and Photovoltaic Efficiency:** GNNs have been implemented to predict a range of molecular properties, including drug efficacy and the efficiency of photovoltaic materials.[67]
- **Cardiotoxicity Detection:** GNNs have been successfully applied to molecule

classification tasks, such as the detection of cardiotoxicity, demonstrating their reliability in complex biological problems.[10]

- **HIV Inhibitor Prediction:** GNNs have been utilized to predict and identify potential HIV inhibitor molecules by analyzing their graph-based representations, offering a promising direction for antiviral drug development.[8]
- **Quantum Chemical Property Prediction:** GNNs have achieved state-of-the-art accuracy in predicting various quantum chemical properties of molecules and materials, such as HOMO-LUMO gap and energy bandgaps.[34]
- **Solubility and Lipophilicity:** Advanced GNN models have shown strong performance in predicting crucial properties like lipophilicity and solubility, vital for drug pharmacokinetics.[14]
- **3D Molecular Property Prediction:** The ability to leverage 3D molecular data allows GNNs to predict properties that are highly dependent on spatial arrangement, such as bond angles and torsion angles, further enhancing predictive accuracy.[14]

The success of GNNs in these applications signifies a fundamental shift from traditional feature engineering to feature learning. Historically, molecular property prediction relied on hand-crafted features based on predetermined fingerprint extraction rules, a process that was time-consuming, expert-dependent, and did not always yield optimal results.[5] GNNs, through their message passing and aggregation mechanisms, automatically learn relevant features directly from the molecular graph structure.[6] This capability allows them to capture complex non-linear relationships between atomic structure and molecular properties with higher accuracy and efficiency. This transition to data-driven, automated feature learning marks a significant advancement, moving towards more robust and less expert-dependent predictive models in drug discovery.

3.3 Benchmark Datasets and Evaluation Methodologies

The robust evaluation and advancement of GNNs for molecular property prediction rely heavily on high-quality, standardized benchmark datasets and rigorous evaluation methodologies. These resources allow researchers to compare models fairly and track progress in the field.[20]

Key benchmark datasets widely used in GNN-based MPP include:

- **QM9:** This dataset comprises small organic molecules with computed quantum chemical properties, serving as a standard for evaluating models on fundamental molecular characteristics.[8]
- **MoleculeNet:** A large-scale benchmark collection for molecular machine learning, MoleculeNet curates multiple public datasets and provides standardized metrics for evaluation across various molecular tasks.[8]
- **HIV:** A classification dataset used to predict anti-HIV activity, demonstrating GNNs' utility in specific therapeutic areas.[49]
- **Tox21, ClinTox, SIDER, BACE, FreeSolv, Lipophilicity:** These datasets cover a wide range of molecular properties, including toxicity, clinical toxicity, side effects, BACE inhibition, aqueous solubility, and lipophilicity, providing diverse challenges for GNN models.[49]

Evaluation methodologies typically employ metrics tailored to the prediction task. For classification tasks, common metrics include the Area Under the Receiver Operating Characteristic curve (AUROC).[10] For regression tasks, Root Mean Squared Error (RMSE) is frequently used.[10] Beyond predictive accuracy, other metrics like Expected Calibration Error (ECE), Brier score, and Negative Log Likelihood (NLL) are used to assess model robustness and confidence in predictions, particularly under distributional shifts.[10]

Validation types are also crucial for assessing a model's generalizability. While random splits of data are common, more challenging evaluations involve **dissimilar-molecules splits**.[71] In this approach, the test set contains molecules structurally dissimilar from those in the training data, better mimicking real-world scenarios where models encounter novel compounds. This rigorous validation helps ensure that GNNs can generalize effectively to new chemical spaces, a critical requirement for practical drug discovery.

Dataset Name	Description	Typical Tasks	Key Characteristics
QM9	~134k small organic molecules with quantum chemical properties.	Regression (e.g., HOMO-LUMO gap, energies)	Small, diverse molecules; high-quality computed properties; often used for fundamental quantum chemistry predictions.[8]

MoleculeNet	Curated collection of various public molecular datasets.	Classification (e.g., toxicity, bioactivity), Regression (e.g., solubility)	Large-scale benchmark; diverse tasks and molecule types; facilitates comparative studies.[8]
HIV	Molecules tested for anti-HIV activity.	Binary Classification (active/inactive)	Specific biological activity; often used to evaluate GNNs for antiviral drug development.[49]
Tox21, ToxCast, ClinTox, SIDER	Toxicity and adverse effect data.	Multi-label Classification (e.g., binding to toxicity targets, side effects)	Focus on safety and adverse profiles; critical for early compound elimination.[49]
FreeSolv, Lipophilicity, ESOL	Aqueous solubility and lipophilicity data.	Regression (e.g., logP, solubility in water)	Physicochemical properties; important for pharmacokinetics and drug formulation.[34]
BACE	Beta-secretase 1 (BACE1) inhibition data.	Binary Classification (inhibitor/non-inhibitor)	Enzyme inhibition; relevant for neurodegenerative diseases like Alzheimer's.[49]

Table 3: Benchmark Datasets for GNN-based Molecular Property Prediction

4. Broader Impact of GNNs Across the Drug Discovery Pipeline

Beyond their significant contributions to molecular property prediction, Graph Neural Networks are exerting a profound influence across multiple stages of the drug discovery pipeline. Their ability to effectively learn from and represent complex molecular structures makes them invaluable tools for tasks ranging from the *de novo*

design of new chemical entities to the prediction of intricate biological interactions.

4.1 De Novo Molecule Generation and Optimization

One of the most exciting applications of GNNs is in *de novo* molecule generation and optimization. This involves creating novel chemical entities with desired biological and physicochemical properties from scratch, rather than merely screening existing compound libraries.[15] GNNs are uniquely positioned for this task because they can directly operate on molecular graphs, enabling the generation of chemically valid and diverse structures that meet specific design constraints.

Various GNN-based methods have been developed for this purpose:

- **Iterative Sampling Approaches:** Models like GraphINVENT use iterative sampling techniques, where atoms or bonds are progressively added to a growing molecular graph, guided by GNN-learned action probability distributions.[20] This ensures chemical validity at each step.
- **Ligand-Protein Based Generation:** More advanced approaches focus on generating molecules tailored for specific protein binding sites, often incorporating 3D information. Models such as AR, GraphBP, Pocket

References

1. Raman, Siddhant. "The Rise Of AI-Powered Applications: Large Language Models In Modern Business." (2023).
2. [2] Vishwakarma, S., Hernandez-Hernandez, S., & Ballester, P. J. (2024). Graph neural networks are promising for phenotypic virtual screening on cancer cell lines.
3. *Biology Methods and Protocols*, 9(1), bpae065. https://doi.org/10.1093/biomethods/bpae065
4. [3] Lifebit. (n.d.).
5. *What are the current challenges of drug discovery?* Retrieved from https://www.lifebit.ai/blog/current-challenges-of-drug-discovery/
6. [4] ACS Omega. (2025). AI-Driven Drug Discovery: A Comprehensive Review.
7. https://doi.org/10.1021/acsomega.5c00549

8. [5] Fang, Z., Zhang, X., Zhao, A., Li, X., Chen, H., & Li, J. (2025). Recent Developments in GNNs for Drug Discovery. arXiv preprint arXiv:2506.01302.
9. https://doi.org/10.48550/arXiv.2506.01302
10. [6] Wang, H., Zhang, A., Zhong, Y., Tang, J., Zhang, K., & Li, P. (2024). Chain-aware graph neural networks for molecular property prediction.
11. *Bioinformatics*, 40(10), btae574. https://doi.org/10.1093/bioinformatics/btae574
12. [7] CAS. (n.d.).
13. *Dealing with the challenges of drug discovery*. Retrieved from https://www.cas.org/resources/cas-insights/dealing-challenges-drug-discovery
14. [8] Nguyen, Q. (n.d.).
15. *Graph Neural Network for Molecular Structure: Application in HIV Inhibitor Molecule Prediction*. OSF. Retrieved from https://osf.io/c3mqy/download
16. [9] Chapman University. (2025).
17. *Raman, S. (2024). DECIPHERING DESTINY: The Rise of Large Language Models in Decision-Making Mastery*.
18. [10] Drug Discovery Today Technology. (2020). A compact review of molecular property prediction with graph neural networks.
19. *Drug Discovery Today Technology*, 37, 1-12. https://doi.org/10.1016/j.ddtec.2020.11.009
20. [11] Molecules. (2024). Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges.
21. https://doi.org/10.3390/molecules29040903
22. [12] Zou, Z., Wang, D., & Tiwary, P. (2025). A graph neural network-state predictive information bottleneck (GNN-SPIB) approach for learning molecular thermodynamics and kinetics.
23. *Digital Discovery*, 4(1).(https://doi.org/10.1039/D4DD00315B)
24. [13] Zhang, Z., et al. (2020). Deep learning on graphs for drug discovery.
25. *Journal of Chemical Information and Modeling*, 60(10), 4800-4812. https://doi.org/10.1021/acs.jcim.0c00321
26. [14] Stärk, H., et al. (2021). Equivariant Graph Neural Networks for 3D Macromolecular Structure. arXiv preprint arXiv:2106.03843.
27. https://doi.org/10.48550/arXiv.2106.03843
28. [15] Bongini, P. (2023). Graph Neural Networks for Drug Discovery: An Integrated Decision Support Pipeline. In
29. *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering*

(MetroXRAINE).(https://doi.org/10.1109/MetroXRAINE58569.2023.10405789)
30. [16] Zhang, O., et al. (2025). Graph Neural Networks in Modern AI-aided Drug Discovery. arXiv preprint arXiv:2506.06915.
31. https://doi.org/10.48550/arXiv.2506.06915
32. [17] Tsubaki, M., Tomii, K., & Sese, J. (2018). Compound-protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences.
33. *Bioinformatics*, 35(2), 309-318. https://doi.org/10.1093/bioinformatics/bty535
34. [18] Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. In
35. *Proceedings of the 35th International Conference on Machine Learning* (PMLR, Vol. 80, pp. 2323-2332). https://proceedings.mlr.press/v80/jin18a.html
36. [19] Krenn, M., et al. (2020). Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation.
37. *Raman, S. (2023). The Rise Of AI-Powered Applications: Large Language Models In Modern Business.*
38. [20] Landrum, G. A. (2014). RDKit: Open-source cheminformatics. Release 2014.03.1.(https://doi.org/10.5281/ZENODO.10398)
39. [21] Duvenaud, D. K., et al. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. In
40. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. https://papers.nips.cc/paper/5954-convolutional-networks-on-graphs-for-learning-molecular-fingerprints
41. [22] Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In
42. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*. https://arxiv.org/abs/1609.02907
43. [23] Gilmer, J., et al. (2017). Neural Message Passing for Quantum Chemistry. In
44. *Proceedings of the 34th International Conference on Machine Learning* (PMLR, Vol. 70, pp. 1263-1272). https://proceedings.mlr.press/v70/gilmer17a.html
45. [24] Velickovic, P., et al. (2018). Graph Attention Networks. In
46. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. https://arxiv.org/abs/1710.10903
47. [25] Schütt, K. T., et al. (2017). SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In
48. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://papers.nips.cc/paper/6700-schnet-a-continuous-filter-convolutional-

neural-network-for-modeling-quantum-interactions
49. [26] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In
50. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://papers.neurips.cc/paper/6703-inductive-representation-learning-on-large-graphs
51. [27] Wu, Z., et al. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning.
52. *Chemical Science*, 9(2), 513-530.(https://doi.org/10.1039/C7SC02664A)
53. [28] Wang, C., Kumar, G. A., & Rajapakse, J. C. (2025). Drug discovery and mechanism prediction with explainable graph neural networks.
54. *Nature Communications*, 16, 1-15. https://doi.org/10.1038/s41598-024-83090-3
55. [29] You, J., et al. (2020). Graph Contrastive Learning with Augmentations. In
56. *Raman, Siddhant. "DECIPHERING DESTINY: The Rise of Large Language Models in Decision-Making Mastery." (2024).*
57. [30] Hu, W., et al. (2020). Strategies for Pre-training Graph Neural Networks. In
58. *International Conference on Learning Representations (ICLR 2020)*. https://cs.stanford.edu/people/jure/pubs/pretrain-iclr20.pdf
59. [31] Loesch, J., Yang, Y., Ekmekci, P., Dumontier, M., & Celebi, R. (2024). Explaining Graph Neural Network Predictions for Drug Repurposing. In
60. *CEUR Workshop Proceedings*. Retrieved from https://ceur-ws.org/Vol-3890/paper-5.pdf
61. [32] Ioannidis, V., et al. (2019). Multi-relation graph neural networks for protein-protein interaction prediction.
62. *Bioinformatics*, 35(2), 309-318. https://doi.org/10.1093/bioinformatics/btz500
63. [33] Gligorijevic, V., et al. (2021). Deep learning for protein function prediction.
64. *Nature Communications*, 12(1), 1-14. https://doi.org/10.1038/s41467-021-24811-y
65. [34] J. Chem. Theory Comput. (2024). Graph Neural Network-Based Molecular Property Prediction with Patch Aggregation.
66. *Journal of Chemical Theory and Computation*, 20(20), 8886-8896. https://doi.org/10.1021/acs.jctc.4c00798
67. [35] Ingraham, J., Garg, V., Barzilay, R., & Jaakkola, T. (2019). Generative Models for Graph-Based Protein Design. In
68. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. https://papers.nips.cc/paper/9711-generative-models-for-graph-based-protein-design

69. Chowdhury, R. H. (2024). Blockchain and AI: Driving the future of data security and business intelligence. *World Journal of Advanced Research and Reviews*, *23*(1), 2559-2570.
70. Chowdhury, R. H. (2025). *Digital Leadership and Organizational Learning: Technologies for Business Transformation and Operational Excellence*. Deep Science Publishing.
71. Sharmin, S., & Chowdhury, R. H. (2025). Digital transformation in governance: The impact of e-governance on public administration and transparency. *Journal of Computer Science and Technology Studies*, *7*(1), 362-379.
72. Sharmin, S. (2025). Refugee Resettlement & AI-Powered Resource Allocation Optimizing social services for displaced populations. *Journal of Public Administration Research*, *2*(1), 13-36.
73. Chowdhury, N. R. H., & Chowdhury, N. M. S. H. (2024). Impact of social crises on brand perception and consumer trust. *International Journal of Science and Research Archive*, *13*(2), 527–536. https://doi.org/10.30574/ijsra.2024.13.2.2162
74. Okika, N., Nwatuzie, G. A., Olarinoye, H. S., Nwaka, A. A., Igba, E., & Dunee, R. (2025). Assessing the vulnerability of traditional and post-quantum cryptographic systems through penetration testing and strengthening cyber defenses with zero trust security in the era of quantum computing. *International Journal of Innovative Science and Research Technology*, *10*(2).
75. Nwatuzie, G. A., Enyejo, L. A., & Umeaku, C. (2025). Enhancing Cloud Data Security Using a Hybrid Encryption Framework Integrating AES, DES, and RC6 with File Splitting and Steganographic Key Management. *International Journal of Innovative Science and Research Technology*, *10*(1).
76. Nwatuzie, G., & Peyravi, H. (2025). DRNMS: An Enhanced Deep Learning-Based System for Data Recovery and Anomaly Detection in Network Monitoring. In *International Conference on Computational Science and Computational Intelligence* (pp. 87-103). Springer, Cham.