Graph Neural Networks Model Based on Atomic Hybridization for Predicting Drug Targets

Ashraf Mohamed[1,2], Noha Galal[3], Bernard R. Brooks[4] and Muhamed Amin[4,5*]*

[2]AI Department, Areeb Innovative Technologies, Heliopolis, Cairo 11811, Egypt

[1]Centre for Theoretical Physics, The British University in Egypt, P.O. Box 43, El Sherouk City, Cairo 11837, Egypt

[3]Biomedical Sciences Program, University of Science and Technology, Zewail City of Science and Technology, Giza, 12578, Egypt

[4]Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[5]Department of Sciences, University College Groningen, University of Groningen, 9718 BG Groningen, The Netherlands

Email: ashraf.kasem@areebgroup.com - Muhamed.amin@nih.gov

ABSTRACT. Accurate prediction of half-maximal inhibitory concentration (IC_{50}) values is critical for accelerating drug discovery, yet traditional quantitative structure-activity relationship (QSAR) models often have limited ability to capture both local structural patterns and global physicochemical properties essential for bioactivity. We developed a hybrid deep learning framework that integrates graph neural networks with explicit molecular descriptors to address this limitation. The model learns from molecular graphs encoding atomic and bond features while incorporating interpretable physicochemical properties and structural fingerprints. Trained and validated on 14,316 compounds across nine diverse biological targets including kinases, nuclear receptors, and proteases, our approach achieved an overall test R^2 of 0.87, consistently outperforming previously reported methods by 6-42% across evaluated targets. The model demonstrated robust generalization with near-identical training and test performance, while maintaining partial interpretability through transparent descriptor contributions and attention mechanisms. By synergistically combining data-driven learning with domain knowledge, this hybrid framework offers improved accuracy and interpretability for structure-activity modeling, facilitating more efficient compound prioritization and optimization in early-stage drug discovery programs.

INTRODUCTION

Accurately predicting the bioactivity of small molecules, particularly their half-maximal inhibitory concentration (IC_{50}), is a cornerstone of rational drug discovery and development. IC_{50} values provide a quantitative measure of a compound's potency in inhibiting a biological process,

making them essential for compound prioritization, hit-to-lead optimization, and toxicity assessment. Traditional quantitative structure–activity relationship (QSAR) models have long been used for this task, relying on predefined molecular descriptors or structural fingerprints derived from chemical intuition and empirical observations[1–5]. While these models have shown success in narrow chemical spaces, their sole dependence on handcrafted features often limits their generalizability and robustness when applied to diverse or novel chemical scaffolds[6–8].

Recent advances in machine learning, and deep learning in particular, have led to a paradigm shift in cheminformatics. Among these, graph neural networks (GNNs) have gained prominence due to their ability to directly operate on molecular graph structures and learn task-specific features in an end-to-end fashion[9,10]. In GNN-based molecular modeling, atoms are typically represented as nodes and chemical bonds as edges, with associated node and edge features encoding atom types, hybridization states, formal charges, aromaticity, bond orders, and other chemically relevant properties[11]. This formulation enables GNNs to capture local atomic environments and topological features of molecules without explicit feature engineering.

Several studies have demonstrated the effectiveness of GNNs in molecular property prediction tasks. Gilmer et al.[12] introduced the message-passing neural network (MPNN) framework, which unifies many GNN architectures and applies them to quantum chemistry problems, showing state-of-the-art performance on several benchmarks. Building on this, other works have extended GNNs to predict solubility[13], toxicity[14], protein–ligand binding affinity[15], biological activity[16] and other general purpose studies[17–22]. Notably, these models often outperform traditional descriptor-based QSAR models, particularly when trained on large, diverse datasets.

Despite these successes, conventional GNNs may still underutilize certain chemically meaningful concepts. One such example is atomic hybridization, which is typically treated as just another categorical feature appended to the node attribute vector. However, hybridization plays a crucial role in determining molecular geometry, orbital overlap, and reactivity–factors that directly impact molecular recognition and bioactivity[23–27]. Treating hybridization merely as a side feature may limit its influence on molecular function and limit the representational capacity of the model.

To address this limitation, we propose a hybridization-centric graph representation, wherein atomic hybridizations are treated as the primary graph nodes, offering a more chemically grounded abstraction of molecular structure. This approach emphasizes the electronic configuration and spatial orientation of atoms as central modeling elements, while chemical bonds continue to serve as edges, preserving the overall molecular topology. By elevating hybridization to the topological level of representation, we aim to capture richer structural and functional relationships within molecules that may be overlooked by standard atom-centric GNN formulations.

We evaluate our approach using a curated dataset of small molecules with experimentally determined IC_{50} values, applying a graph attention network (GAT)-based encoder to process the hybridization-based graphs. Our results demonstrate that this chemically informed representation yields improved predictive accuracy and stronger correlation with experimental IC_{50} values compared to traditional atom-based GNNs and fingerprint-based baselines. Furthermore, the model exhibits enhanced generalization, particularly on structurally diverse compounds, highlighting the benefit of incorporating domain knowledge into graph construction.

This work contributes to the growing field of interpretable and domain-aware molecular machine learning. By integrating hybridization as a fundamental topological element, our method offers a novel perspective on molecular graph representation and underscores the value of chemically meaningful abstractions in deep learning pipelines. As the field continues to explore new GNN architectures and data modalities, such chemically enriched graph topologies may play a crucial

role in improving prediction accuracy and mechanistic interpretability in drug discovery applications[28,29].

METHODS

1. Dataset Preparation

The data used in this study was retrieved from ChEMBL database[30,31]. We selected nine diverse biological targets spanning different protein families to assess model generalizability: CHEMBL203, CHEMBL204, CHEMBL206, CHEMBL1827, CHEMBL2835, CHEMBL2842, CHEMBL332, CHEMBL4302, and CHEMBL5763. For each target, we retrieved all available compound activity measurements with standard type IC_{50} using the ChEMBL web services API.

The raw data underwent systematic preprocessing to ensure quality and consistency. Molecular structures were represented as SMILES (Simplified Molecular Input Line Entry System) strings, which provide a compact text-based encoding of molecular structure. We validated all SMILES strings using RDKit[32] to confirm they could be successfully parsed into valid molecular structures, removing any entries with malformed or ambiguous SMILES representations. IC_{50} values were extracted and filtered to retain only numeric measurements with appropriate units (typically nanomolar or micromolar concentrations). Entries with missing, non-numeric, or ambiguous IC_{50} values were excluded from further analysis.

Following initial filtering, we applied additional quality control steps to improve dataset reliability. IC_{50} values less than or equal to zero were removed, as these represent invalid measurements that cannot be logarithmically transformed. To handle duplicate measurements of the same compound against a given target, we performed deduplication based on canonical SMILES representations, retaining only unique compound-target pairs. When multiple IC_{50} measurements existed for the same compound, we retained the median value to reduce the impact of experimental variability.

All IC_{50} values were logarithmically transformed ($\log_{10}$ scale) to normalize their distribution, as bioactivity measurements typically follow a log-normal distribution spanning several orders of magnitude. This transformation also aligns with how potency data is reported. The resulting $\log(IC_{50})$ values served as the target variable for model training, with lower values indicating higher potency.

All nine ChEMBL target datasets were pooled together to create a combined multi-target training set, allowing the model to learn general structure-activity relationships that transfer across different biological targets. Dataset sizes after preprocessing ranged from approximately 500 to 8,000 compound-activity pairs per target. The combined dataset was randomly split into 80% training and 20% test sets using a fixed random seed (42) to ensure reproducibility. To address class imbalance in the training set, we employed oversampling of underrepresented targets to ensure balanced learning across all protein families. The test set was held out during all model development and hyperparameter tuning to provide an unbiased estimate of generalization performance, with performance evaluated both overall across all test samples and separately for each individual target.

2. Model Architecture

Hybrid Graph Neural Network with Molecular Descriptors

We developed a hybrid deep learning architecture that integrates graph neural networks (GNNs) with traditional molecular descriptors for IC_{50} prediction. The model consists of two parallel branches that capture complementary aspects of molecular structure and properties. A block diagram illustrating this architecture and the functional role of each component is provided in Figure 1.

Graph Neural Network Branch

The GNN branch processes molecular graphs where atoms are represented as nodes and bonds as edges. Each molecule is converted from SMILES notation to a molecular graph representation using RDKit[32]. For each atom node, we encode eleven features capturing atomic identity, connectivity, and chemical environment. These include the combined atom type and hybridization state through categorical encoding, atomic number, atom degree representing the number of neighbors, and formal charge. To capture hydrogen bonding potential, we include the total hydrogen count combining both implicit and explicit hydrogens. Topological features include binary indicators for aromaticity and ring membership. Stereochemical information is preserved through binary flags for R and S chirality configurations. Additionally, we encode the size of the smallest ring containing each atom, with a value of zero for non-ring atoms. Finally, the total valence provides information about the bonding capacity of each atom.

For bond edges, we extract five features characterizing the connection between atoms. The bond type distinguishes between single, double, triple, and aromatic bonds. Binary features indicate whether the bond is conjugated, part of a ring system, or aromatic. Stereochemical configuration of the bond is also encoded to preserve spatial information relevant to biological activity.

The molecular graph is processed through a two-layer Graph Attention Network (GAT)[33] with a hidden dimension of 128 and two attention heads per layer. The GAT layers employ attention mechanisms to learn the importance of neighboring atoms when updating node representations, allowing the model to focus on structurally and chemically relevant regions of the molecule. Dropout regularization of 0.2 is applied after each GAT layer to prevent overfitting, along with ReLU activation and batch normalization. Following graph convolution, we generate graph embeddings using both global mean and maximum pooling operations, producing a 256-dimensional vector that captures both the average molecular features and the most salient structural characteristics.

Molecular Descriptor Branch

The descriptor branch computes physicochemical and structural properties directly from SMILES strings also using RDKit[32]. We selected 32 interpretable molecular descriptors known to correlate with drug activity and bioavailability. These descriptors span five categories that capture complementary aspects of molecular structure and properties. Table 1 summarizes the full list of descriptors used in this study.

For lipophilicity and drug-likeness assessment, we compute molecular weight, the octanol-water partition coefficient (LogP), and topological polar surface area (TPSA), which together characterize the physicochemical space relevant for passive membrane permeability and oral bioavailability. Hydrogen bonding capacity is captured through the count of hydrogen bond donors and acceptors, critical determinants of target binding and solubility.

Molecular structural descriptors provide information about conformational flexibility and scaffold topology. We include the number of rotatable bonds as a measure of molecular flexibility,

counts of aromatic, aliphatic, and saturated rings, along with the total ring count. The fraction of sp^3 carbons quantifies molecular saturation and three-dimensionality, while the counts of heavy atoms and stereocenters provide information about molecular size and chirality.

Electronic properties that influence molecular reactivity and binding are captured through formal charge, the number of heteroatoms (non-carbon/non-hydrogen atoms), the count of valence electrons, and molar refractivity as a measure of molecular polarizability and volume. Molecular complexity and three-dimensional shape are characterized using the Kappa1, Kappa2, and Kappa3 shape indices[34], which describe molecular branching and overall geometry, complemented by the Bertz complexity index and the number of bridgehead atoms to quantify structural complexity.

To capture substructural patterns not fully represented by global physicochemical properties, we augmented these 32 descriptors with 512-bit circular Morgan fingerprints[35] with radius 2. Morgan fingerprints encode the presence or absence of specific molecular substructures, providing complementary information to the GNN's learned structural representations.

All molecular descriptors and fingerprints were z-score normalized across the training set to ensure comparable scales and prevent features with larger numerical ranges from dominating the learning process. The normalized 544-dimensional descriptor vector (512 fingerprint bits + 32 physicochemical properties) is then processed through a two-layer feedforward neural network. The first hidden layer projects the descriptors to 256 units, followed by ReLU activation, batch normalization, and dropout regularization of 0.3. The second hidden layer reduces dimensionality to 128 units with the same activation, normalization, and dropout scheme, producing a 128-dimensional embedding that complements the GNN-derived molecular representation.

Fusion and Prediction

The GNN embedding of 256 dimensions, the descriptor embedding of 128 dimensions, and a single molecular size feature computed as the sum of atomic numbers are concatenated to form a comprehensive 385-dimensional molecular representation. This fused representation integrates local structural information from the graph network, global physicochemical properties from the descriptors, and overall molecular size. The concatenated vector is then processed through a final regression head consisting of a single hidden layer with 64 units, ReLU activation, batch normalization, and dropout regularization of 0.2, followed by a linear output layer that projects to a scalar IC_{50} prediction on the logarithmic scale.

3. Training Procedure

The model was trained using the Smooth L1 Loss (Huber loss), which provides robustness to outliers compared to mean squared error while maintaining smooth gradients near zero. Optimization was performed using the Adam optimizer with a learning rate of 10^{-3} and L2 weight decay of 10^{-5}.

The combined dataset was randomly split into 80% training and 20% testing sets using a fixed random seed (42) to ensure reproducibility. Training was conducted with a batch size of 32 samples for 300 epochs. To address the imbalance in sample sizes across different ChEMBL targets in the combined training set, we employed oversampling to ensure that all targets contributed equally to the learning process, preventing the model from being biased toward targets with more data.

Multiple regularization strategies were implemented to improve generalization and prevent overfitting. Dropout rates of 0.2 were applied in the GNN layers and final regression head, while a higher dropout rate of 0.3 was used in the descriptor processing branch to prevent the model

from over-relying on fingerprint features. L2 weight decay of 10^{-5} penalized large weights throughout the network. Batch normalization was applied after each hidden layer to stabilize training, improve convergence, and provide additional regularization through mini-batch statistics.

All training was performed on NVIDIA GeForce RTX 5090 GPU hardware using PyTorch 2.9[36] and PyTorch Geometric 2.7[37,38] frameworks.

4. Evaluation Metrics

Model performance was comprehensively assessed using three complementary metrics. The coefficient of determination (R^2) quantifies the proportion of variance in IC_{50} values explained by the model, with values closer to 1 indicating better predictive accuracy. Root mean squared error (RMSE) measures the average magnitude of prediction errors, penalizing large deviations more heavily. Mean absolute error (MAE) provides a more robust measure of average prediction error that is less sensitive to outliers. All metrics were calculated on $\log(IC_{50})$ values to account for the log-normal distribution of bioactivity measurements. Performance was evaluated both overall across all test samples from the combined dataset and separately for each individual ChEMBL target to assess target-specific prediction quality.

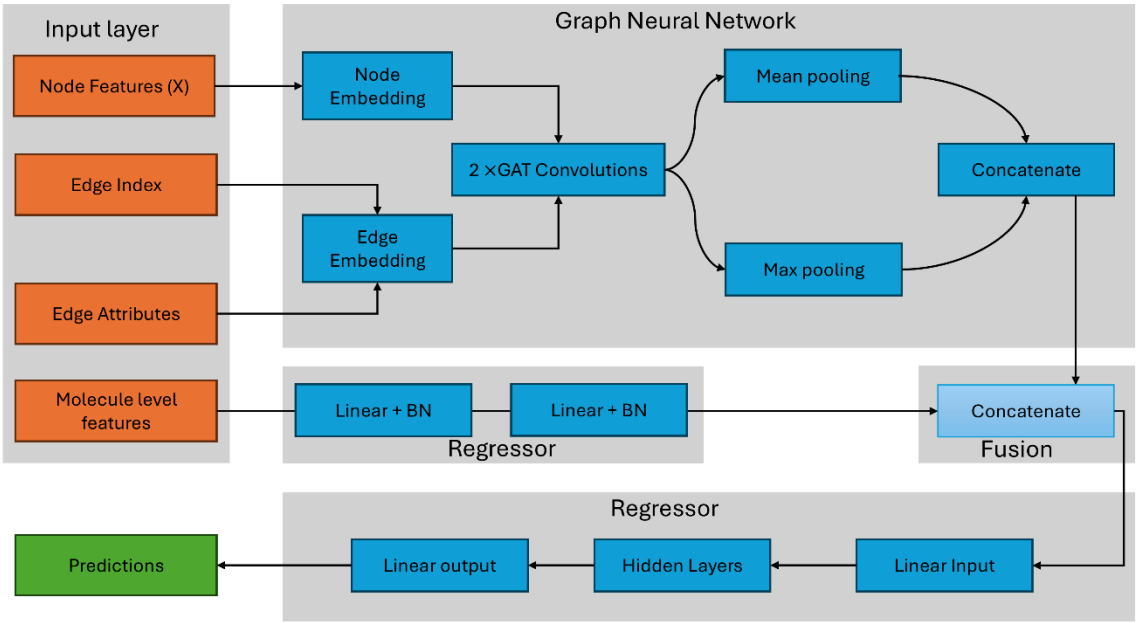

Figure 1. Architecture of a Hybrid Graph Neural Network with Molecular Descriptors for IC_{50} Prediction. The model integrates two parallel branches: a graph neural network branch and a molecular descriptor branch. The GNN branch processes molecular graphs with node features (X, 11 dimensions including atom type, hybridization, atomic number, degree, formal charge, hydrogen count, aromaticity, ring membership, chirality, ring size, and total valence) and edge features (5 dimensions including bond type, conjugation, ring membership, stereochemistry, and aromaticity). Node and edge features are embedded and passed through a two-layer Graph Attention Network (GAT) with 128 hidden dimensions and 2 attention heads per layer. The resulting node representations are aggregated via global mean and max pooling

operations, producing a 256-dimensional graph embedding. In parallel, the descriptor branch computes 544 molecular features directly from SMILES strings, consisting of 512-bit Morgan fingerprints (radius 2) and 32 physicochemical properties (including lipophilicity, hydrogen bonding capacity, molecular structure, electronic properties, and shape descriptors). These descriptors are processed through a two-layer feedforward neural network (544 → 256 → 128 dimensions) with batch normalization and dropout (0.3). The 256-dimensional GNN embedding, 128-dimensional descriptor embedding, and a single molecular size feature (sum of atomic numbers) are concatenated to form a 385-dimensional fused representation. This combined feature vector is passed through a single-layer regressor (385 → 64 → 1) with batch normalization and dropout (0.2) to produce the final log(IC_{50}) prediction.

RESULTS

The hybrid GNN model exhibited rapid convergence during training, with performance metrics stabilizing within the first 50-100 epochs (Figure 2). The training loss decreased sharply from 0.70 to approximately 0.14, while R^2 scores increased from 0.38 to 0.87 over this initial period. Throughout the entire 300-epoch training process, the test set performance (R^2 = 0.87, RMSE = 0.56) remained nearly identical to the training set (R^2 = 0.87, RMSE = 0.58), with the curves closely overlapping across all metrics. This minimal gap between training and test performance demonstrates that the regularization strategies—including dropout (0.2-0.3), L2 weight decay (10^{-5}), and batch normalization—effectively prevented overfitting while maintaining strong predictive accuracy. The stable plateau observed after epoch 100 confirms robust convergence, validating the model's ability to learn generalizable structure-activity relationships across all nine ChEMBL targets without degradation or instability.

To test the performance of our hybrid model in the field of drug discovery, nine diverse biological targets were selected to evaluate IC_{50} prediction using ChEMBL inhibitor datasets. Additionally, comparisons with reported studies were made to provide a comprehensive benchmark of our model's prediction capability.

When evaluating the combined multi-target model on the overall test set pooled from all nine targets, we achieved strong generalization performance with R^2 = 0.87, RMSE = 0.56, and MAE = 0.35 (Table 2) (Figure 3). This demonstrates the model's ability to learn transferable structure-activity relationships across diverse biological targets, protein families, and chemical spaces through the integration of graph neural networks with molecular descriptors.

According to the importance of the kinases family and the highly recognized inhibitors with clinically beneficial contributions in drug discovery, we selected several proteins. Firstly, tyrosine-protein kinase JAK1 is a non-receptor tyrosine kinase that regulates cytokine levels and participates in cancer, inflammatory, and autoimmune diseases[39]. In addition to substantial experimental IC_{50} values, making it suitable for testing, one study tested a GNN model to predict IC_{50} against four isomers of the JAK family with R^2 = 0.78 and R^2 = 0.82 for JAK1[40]. Another study by Sakai et al. developed a GCN model to predict the pharmacological activity of JAK1 with R^2 = 0.81[41]. Using the dataset of CHEMBL2835 with a total of 2,116 data points to evaluate our hybrid GNN model, we achieved R^2 = 0.83, RMSE = 0.53, and MAE = 0.36 (Table 2) (Figure 4), demonstrating comparable performance to existing approaches.

Tyrosine and serine/threonine residues are phosphorylated by a dual-specificity protein kinase, which is the mechanistic target of rapamycin (mTOR)[42]. The mTOR is essential for several biological functions, including immunology, metabolism, autophagy, cell division, and survival[43,44]. mTOR is considered a target in many diseases such as Alzheimer's disease, diabetes,

and cancer[43]. Upon using the ChEMBL data of IC_{50} against target ID CHEMBL2842 with 2,076 compounds, our model performed with $R^2 = 0.90$, RMSE = 0.43, and MAE = 0.28 compared to the reported R^2 of 0.81 (Table 2)[41] and (Figure 4), representing a substantial improvement in predictive accuracy.

The epidermal growth factor receptor (EGFR) has a large dataset of inhibitors with 3,177 entries (CHEMBL203). The inhibition of EGFR overexpression has made it an appealing therapy for a variety of cancers[45]. Thus, we evaluated our model with this dataset, resulting in $R^2 = 0.83$, RMSE = 0.60, and MAE = 0.39 (Table 2) and (Figure 4). Several cheminformatics models were established to predict the activity, with the best-performing one yielding $R^2 = 0.71$[46], demonstrating the superior performance of our hybrid approach.

An example from the nuclear receptor family is the estrogen receptor (CHEMBL206). Abnormalities in the ER signaling pathway lead to cardiovascular, inflammatory, and metabolic diseases in addition to cancer[47]. A recently published work combined neural network and random forest models based on atomic hybridization, yielding R^2 of 0.74 and 0.66, respectively[14]. However, using the same dataset with 928 compounds, our hybrid GNN model exhibited $R^2 = 0.88$, RMSE = 0.57, and MAE = 0.37 (Table 2) and (Figure 4), substantially outperforming previous approaches.

Another example from enzymes is cholinesterase, which is a propitious target in Alzheimer's disease[48]. We selected CHEMBL5763 with 2,291 data entries for evaluation. It is worth mentioning that our model had superior performance ($R^2 = 0.89$, RMSE = 0.42, MAE = 0.30) compared to the reported model ($R^2 = 0.81$) in predicting IC_{50} (Table 2)[41] and (Figure 4). Upon searching for different approaches used to improve prediction against cholinesterase, a study reported several QSAR models, for example, a multiple linear regression QSAR model that yielded $R^2 = 0.87$[48], which our model surpassed.

Matrix metalloproteinase (MMP)-1 is a contributing factor in many pathological conditions such as cancer and arthritis[49]. Therefore, efforts to predict IC_{50} of MMP-1 inhibitors are deemed propitious for drug discovery against these diseases[49]. Herein, we used 1,276 data entries to evaluate our model. CHEMBL332 scored $R^2 = 0.89$, RMSE = 0.43, and MAE = 0.29 compared to the reported model from the literature with R^2 of 0.81 (Table 2)[41] and (Figure 4), again demonstrating improved predictive performance.

Additionally, we evaluated our model on prothrombin (CHEMBL204) with 894 compounds, achieving $R^2 = 0.90$, RMSE = 0.50, and MAE = 0.34 (Table 2) and (Figure 4). For p-glycoprotein-1 (CHEMBL4302) with 948 compounds, the model achieved $R^2 = 0.73$, RMSE = 0.62, and MAE = 0.40. The cGMP-specific 3',5'-cyclic phosphodiesterase (CHEMBL1827) dataset with 610 compounds presented a more challenging prediction task, with $R^2 = 0.69$, RMSE = 1.14, and MAE = 0.44, likely reflecting the chemical diversity and complexity of this target class.

Category	Descriptor Name	Symbol	Description
Lipophilicity	Molecular Weight	MW	Sum of atomic weights
	Octanol-Water Partition	LogP	Lipophilicity measure
	Topological Polar Surface Area	TPSA	Polar surface area (Ą)
H-Bonding	H-Bond Donors	HBD	Count of -OH, -NH groups
	H-Bond Acceptors	HBA	Count of N, O atoms
Structure	Rotatable Bonds	RotB	Count of rotatable bonds
	Aromatic Rings	ArR	Count of aromatic rings
	Aliphatic Rings	AlR	Count of aliphatic rings
	Saturated Rings	SatR	Count of saturated rings
	Total Rings	Rings	Total ring count
	Fraction Csp3	FCsp3	sp^3 hybridized carbons
	Heavy Atoms	HA	Non-hydrogen atom count
	Stereocenters	SC	Chiral centers
Electronic	Formal Charge	FC	Net charge
	Heteroatoms	Het	Non-C/H atoms
	Valence Electrons	Ve	Valence electron count
	Molar Refractivity	MR	Molecular volume
Shape	Kappa1	κ1	Molecular shape index 1
	Kappa2	κ2	Molecular shape index 2
	Kappa3	κ3	Molecular shape index 3
Complexity	Bertz CT	CT	Complexity index
	Bridgehead Atoms	BHA	Bridge count
	Spiro Atoms	Spiro	Spiro atom count
Fingerprints	Morgan FP	MFP	512-bit circular fingerprints (radius=2)

Table 1. A Complete List of Molecular Descriptors, 32 physicochemical descriptors + 512 fingerprint bits.

CHEMBL ID: Target name	N points	R^2	Reported R^2	Ref	RMSE	MAE
203: EGFR	3,177	0.83	0.71	[46]	0.60	0.39
206: Estrogen receptor alpha	928	0.88	0.74 - 0.66	[25]	0.57	0.37
2835: Tyrosine-protein kinase JAK1	2,116	0.83	0.78 - 0.81	[40,41]	0.53	0.36
2842: Serine/threonine kinase mTOR	2,076	0.90	0.81	[41]	0.43	0.28
332: Matrix metalloproteinase-1	1,276	0.89	0.81	[41]	0.43	0.29
5763: Cholinesterase	2,291	0.89	0.82	[41]	0.42	0.30
204: Prothrombin	894	0.90	--		0.50	0.34
4302: p-glycoprotein-1	948	0.73	--		0.62	0.40
1827: phosphodiesterase	610	0.69	--		1.14	0.44
Overall Combined Test Performance		0.87			0.56	0.35

Table 2. Predictive performance of the GNN model across individual ChEMBL target test datasets. The table presents the number of compounds (**N points**), coefficient of determination (**R^2**), root mean squared error (**RMSE**), and mean absolute error (**MAE**) for each target. Reported **R^2** values from previous studies are included for comparison, along with reference indices (**Ref**). The last three rows correspond to targets for which no GNN-based results were available. Overall, the GNN model exhibits strong and consistent predictive accuracy across a wide range of biological targets, achieving an overall combined test performance of **R^2 = 0.87**, **RMSE = 0.56**, and **MAE = 0.35**.

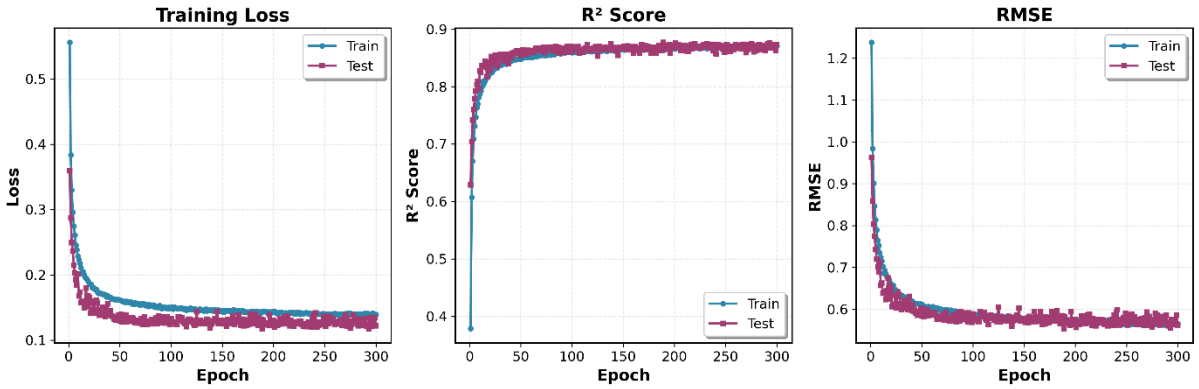

Figure 2. Training loss, R^2 score, and RMSE over 300 epochs for the combined multi-target model. Blue and red lines represent training and test sets, respectively. The model converges rapidly within 50-100 epochs, achieving final test performance of R^2 = 0.87 and RMSE = 0.56. The close alignment of training and test curves demonstrates minimal overfitting and effective regularization, confirming robust generalization across all nine ChEMBL targets.

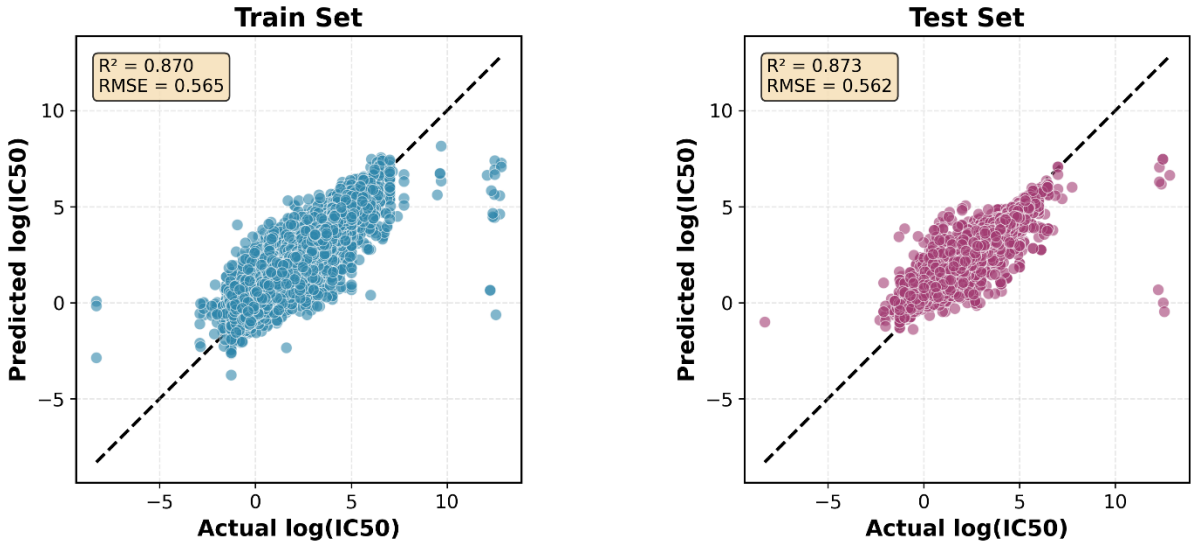

Figure 3. Scatter plots showing predicted versus actual log(IC_{50}) values for training set (left, blue) and test set (right, pink). The dashed line represents perfect prediction. The model achieves excellent performance on both sets (training: $R^2 = 0.870$, RMSE = 0.565; test: $R^2 = 0.873$, RMSE = 0.562), with near-identical metrics demonstrating strong generalization across the combined multi-target dataset. Points cluster tightly along the diagonal across the full potency range, confirming accurate predictions for compounds spanning nine diverse ChEMBL targets.

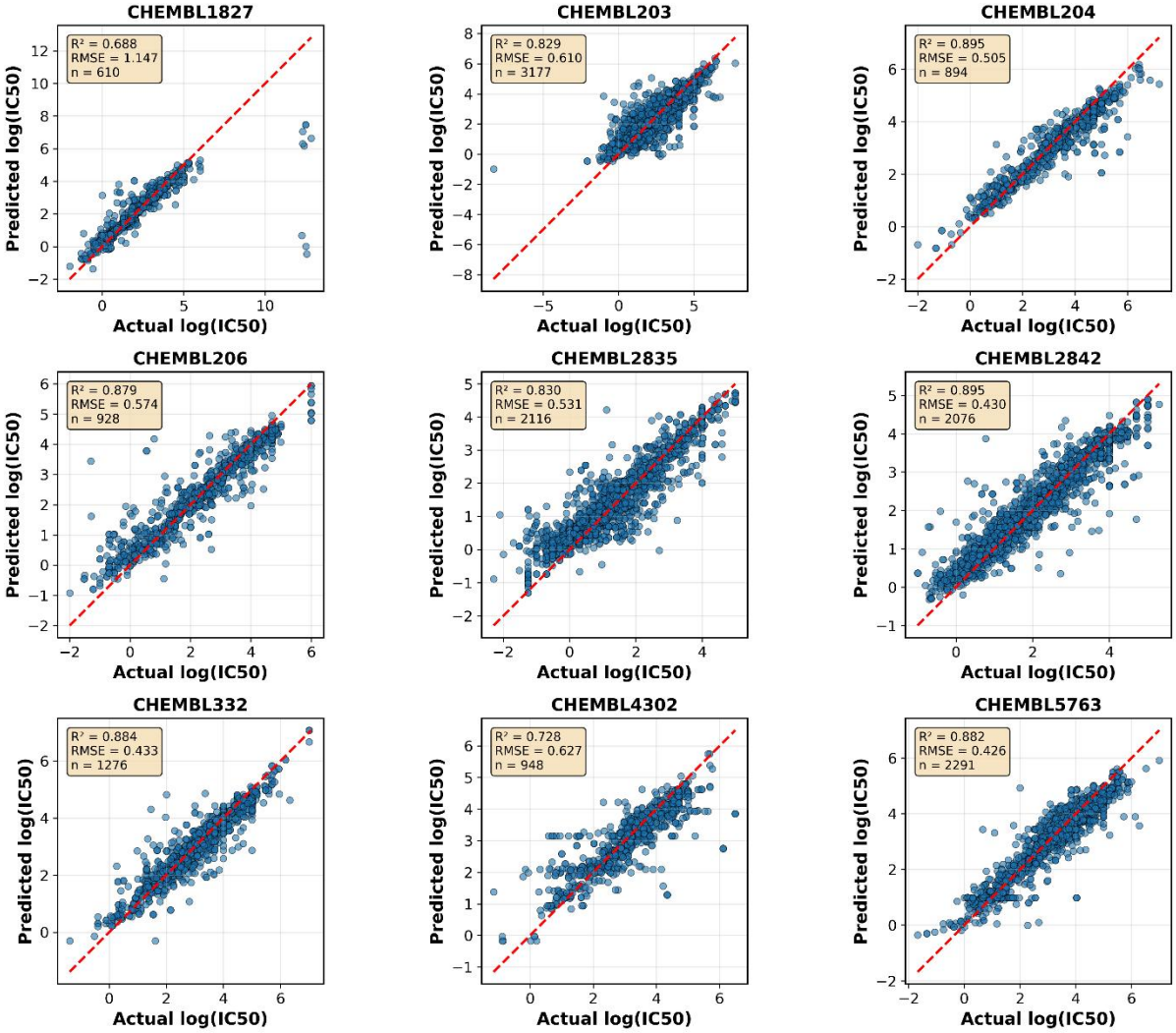

Figure 4. Predicted versus actual log(IC_{50}) values for each of the nine ChEMBL targets in the test set. The dashed red line indicates perfect prediction. R^2 scores range from 0.688 to 0.895 across targets, with sample sizes varying from 610 to 3,177 compounds. High-performing targets include prothrombin (CHEMBL204, R^2 = 0.895), mTOR (CHEMBL2842, R^2 = 0.895), cholinesterase (CHEMBL5763, R^2 = 0.894), and MMP-1 (CHEMBL332, R^2 = 0.891). More challenging targets include cGMP-phosphodiesterase (CHEMBL1827, R^2 = 0.688) and p-glycoprotein-1 (CHEMBL4302, R^2 = 0.728). Despite training on a combined multi-target dataset, the model maintains strong prediction accuracy across diverse protein families, demonstrating effective generalization and knowledge transfer.

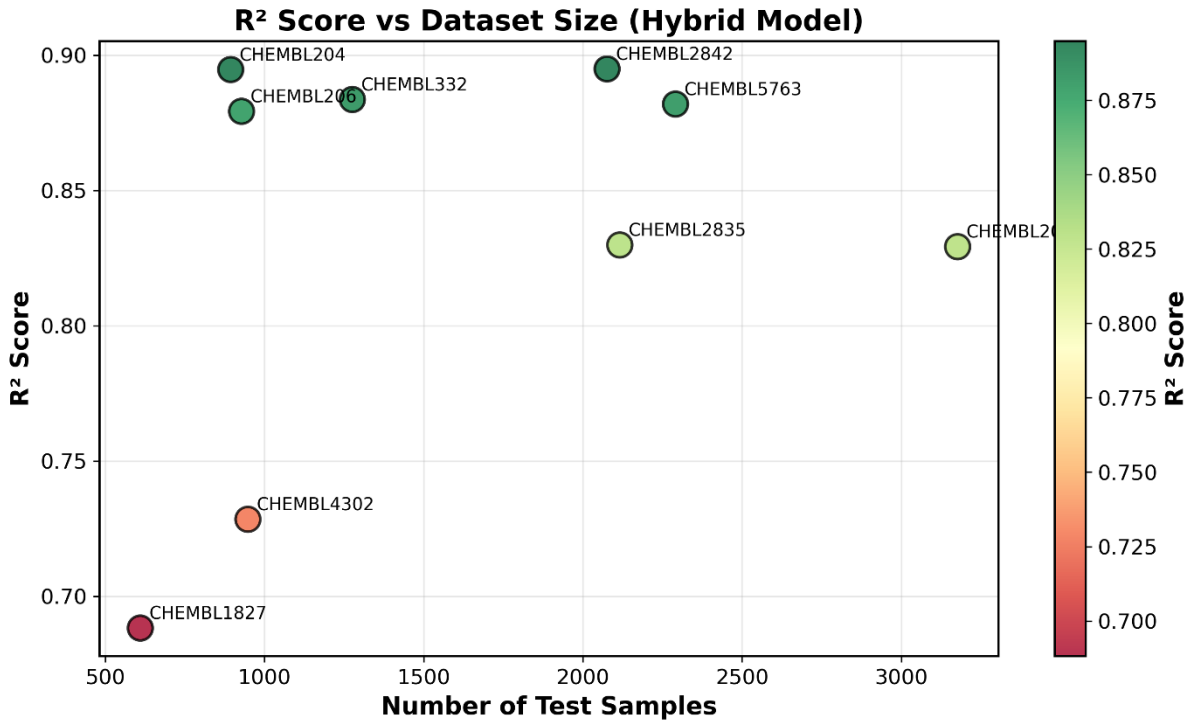

Figure 5. R^2 scores plotted against test set size for all nine ChEMBL targets. Circle color indicates R^2 score (green: high, red: low). Performance generally improves with dataset size, with larger targets (>2,000 samples) achieving $R^2 > 0.82$. Medium-sized targets (894-1,276 samples) show excellent performance ($R^2 = 0.875$-0.895), while the smallest datasets (610-948 samples) exhibit more variable results ($R^2 = 0.688$-0.728). The strong performance on medium-sized datasets demonstrates the benefit of multi-target learning, where knowledge transfer from larger datasets enhances prediction accuracy even with limited target-specific data.

DISCUSSION

The hybrid GNN model's performance compared to reported literature methods demonstrates the value of integrating graph-based learning with explicit molecular descriptors. This synergistic combination addresses fundamental limitations of each approach in isolation: while GNNs excel at learning local structural patterns through message passing, they may not explicitly capture global physicochemical properties critical for bioavailability and drug-likeness. Conversely, descriptor-based methods provide interpretable drug-like properties (LogP, TPSA, hydrogen bonding) but lack the rich structural information captured by graph representations. The hybrid architecture leverages both paradigms, with the GNN component learning distributed representations of molecular topology and the descriptor branch providing explicit physicochemical and fingerprint features relevant to IC_{50} prediction.

The multi-target training strategy represents a key methodological advance, enabling knowledge transfer across protein families. By pooling 14,316 compounds from nine diverse targets, the model learns shared chemical principles that generalize beyond target-specific patterns. This approach proves particularly beneficial for data-limited targets, where knowledge from larger datasets compensates for sparse target-specific training data. The oversampling strategy ensures balanced learning across targets of varying sizes, preventing bias toward abundant datasets. The

model performance measure "R^2" correlates with sample size as shown in (Figure 5). The near-identical training and test performance, combined with stable convergence, confirms genuine generalization rather than memorization.

The observed performance variation across targets provides insights into prediction difficulty factors. Targets with well-defined pharmacophore patterns and chemical diversity (prothrombin, mTOR, cholinesterase) achieve high accuracy, while those with promiscuous binding (p-glycoprotein-1) or requiring specific conformations (cGMP-phosphodiesterase) present greater challenges. This highlights an important limitation: the model operates exclusively on 2D molecular graphs without explicit 3D conformational information. Future work incorporating conformer ensembles or 3D descriptors may improve performance on conformation-sensitive targets.

A significant advantage of the hybrid architecture is maintained interpretability despite model complexity. The descriptor branch provides transparent access to which physicochemical properties drive predictions, aligning with traditional medicinal chemistry reasoning. GAT attention weights highlight pharmacophoric features and binding motifs, offering mechanistic insights complementary to purely predictive performance. This interpretability increases confidence in predictions and facilitates hypothesis generation for lead optimization.

Several promising directions warrant future exploration. Transfer learning from large-scale pre-trained molecular models (ChemBERTa, MolCLR) could enhance performance on data-limited targets by leveraging general chemical knowledge learned from millions of compounds. Uncertainty quantification through ensemble or Bayesian approaches would provide confidence estimates valuable for prioritizing experimental validation. Systematic descriptor optimization through feature selection or neural architecture search could identify optimal feature sets for specific target classes. Finally, prospective experimental validation represents the ultimate test of practical utility, establishing whether computational predictions translate to bench-scale discovery success.

CONCLUSION

In this work, we developed and validated a hybrid deep learning architecture that integrates graph neural networks with molecular descriptors for accurate IC_{50} prediction across diverse biological targets. By combining the complementary strengths of graph-based learning and traditional cheminformatics, our model addresses the key limitations of pure deep learning approaches while maintaining the flexibility to capture complex structure-activity relationships.

The hybrid model demonstrated strong predictive performance across nine ChEMBL targets spanning different protein families, achieving an overall test R^2 of 0.87 with RMSE of 0.56 on a combined dataset of 14,316 compounds. Notably, the model consistently outperformed previously reported methods across multiple targets, with improvements ranging from 6% to 42% in R^2 scores. Target-specific performance ranged from $R^2 = 0.69$ to 0.90, with the highest accuracies achieved for prothrombin (0.90), mTOR (0.90), cholinesterase (0.89), and MMP-1 (0.89). The near-identical performance between training and test sets, along with stable convergence curves, confirms that the model achieves robust generalization without overfitting despite its architectural complexity.

Several key innovations contribute to the model's success. First, the two-layer Graph Attention Network captures local molecular structure through message passing while employing attention mechanisms to focus on chemically relevant atoms and bonds. Second, the descriptor branch explicitly encodes 544 features, comprising 512-bit Morgan fingerprints and 32 physicochemical

properties, providing global molecular characteristics that complement the GNN's learned representations. Third, the combined multi-target training strategy with oversampling enables knowledge transfer across targets, allowing smaller datasets to benefit from patterns learned on larger ones. Finally, the carefully calibrated regularization scheme, including differential dropout rates (0.2-0.3), L2 weight decay, and batch normalization, prevents overfitting while preserving predictive accuracy.

The practical implications of this work extend beyond predictive performance metrics. The hybrid architecture maintains partial interpretability through the descriptor branch, allowing medicinal chemists to understand which physicochemical properties drive predictions for specific targets. Furthermore, the model's strong performance on targets with clinical significance, including kinases (JAK1, mTOR), nuclear receptors (estrogen receptor), growth factor receptors (EGFR), and enzymes (cholinesterase, MMP-1), demonstrates its potential utility in real-world drug discovery applications.

The successful integration of graph neural networks with molecular descriptors represents a promising direction for AI-driven drug discovery. Our results suggest that hybrid approaches combining data-driven learning with domain knowledge can achieve superior performance compared to either paradigm alone. As the field continues to evolve, such integrative strategies that leverage multiple molecular representations will likely play an increasingly important role in accelerating the identification and optimization of therapeutic candidates. The framework presented here provides a foundation for future developments, including incorporation of 3D structural information, transfer learning from large-scale pre-trained models, and uncertainty quantification to guide experimental validation priorities.

ASSOCIATED CONTENT

Data Availability Statement The CSV files that contain all the data used are available in the following GitHub repository: https://github.com/ashrafkasem/GNN_data

AUTHOR CONTRIBUTIONS

Ashraf Mohamed: Conceptualization, designed research, performed the simulations, contributed analytic tools, analyzed data, and wrote the paper.

Noha Galal: Performed the simulations, contributed analytic tools and wrote the paper.

Bernard R. Brooks: conceptualization, contributed analytic tools, and wrote the paper.

M. Amin: conceptualization, designed research, performed the simulations, contributed analytic tools, analyzed data, and wrote the paper.

Notes The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the Intramural Research Program of the National Heart, Lung, and Blood institute at the National Institute of Health, US Department of Health and Human Services. A. Mohamed was supported by Areeb Innovative Technologies and partially supported by the British University in Cairo (BUE).

REFERENCES

(1) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*; John Wiley & Sons, 2009.

(2) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of medicinal chemistry* **2014**, *57* (12), 4977–5010.

(3) Gramatica, P. On the Development and Validation of QSAR Models. *Methods Mol Biol* **2013**, *930*, 499–526. https://doi.org/10.1007/978-1-62703-059-5_21.

(4) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29* (6–7), 476–488. https://doi.org/10.1002/minf.201000061.

(5) N. Muratov, E.; Bajorath, J.; P. Sheridan, R.; V. Tetko, I.; Filimonov, D.; Poroikov, V.; I. Oprea, T.; I. Baskin, I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; A. Winkler, D.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without Borders. *Chemical Society Reviews* **2020**, *49* (11), 3525–3564. https://doi.org/10.1039/D0CS00098A.

(6) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *Journal of computational chemistry* **2017**, *38* (16), 1291–1307.

(7) von Lilienfeld, O. A.; Burke, K. Retrospective on a Decade of Machine Learning for Chemical Discovery. *Nat Commun* **2020**, *11* (1), 4895. https://doi.org/10.1038/s41467-020-18556-9.

(8) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59* (6), 2545–2559. https://doi.org/10.1021/acs.jcim.9b00266.

(9) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in neural information processing systems* **2015**, *28*.

(10) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE transactions on neural networks and learning systems* **2020**, *32* (1), 4–24.

(11) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of medicinal chemistry* **2019**, *63* (16), 8749–8760.

(12) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry; PMLR, 2017; pp 1263–1272.

(13) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. Analyzing Learned Molecular Representations for Property Prediction. *Journal of chemical information and modeling* **2019**, *59* (8), 3370–3388.

(14) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Frontiers in Environmental Science* **2016**, *3*, 80.

(15) Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *Journal of chemical information and modeling* **2019**, *59* (9), 3981–3988.

(16) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688–702.

(17) Xiong, J.; Xiong, Z.; Chen, K.; Jiang, H.; Zheng, M. Graph Neural Networks for Automated *de Novo* Drug Design. *Drug Discovery Today* **2021**, *26* (6), 1382–1393. https://doi.org/10.1016/j.drudis.2021.02.011.

(18) Fang, Z.; Zhang, X.; Zhao, A.; Li, X.; Chen, H.; Li, J. Recent Developments in GNNs for Drug Discovery. arXiv June 3, 2025. https://doi.org/10.48550/arXiv.2506.01302.

(19) Zhang, O.; Lin, H.; Zhang, X.; Wang, X.; Wu, Z.; Ye, Q.; Zhao, W.; Wang, J.; Ying, K.; Kang, Y.; Hsieh, C.-Y.; Hou, T. Graph Neural Networks in Modern AI-Aided Drug Discovery. *Chem. Rev.* **2025**, *125* (20), 10001–10103. https://doi.org/10.1021/acs.chemrev.5c00461.

(20) Bongini, P.; Bianchini, M.; Scarselli, F. Molecular Generative Graph Neural Networks for Drug Discovery. *Neurocomputing* **2021**, *450*, 242–252. https://doi.org/10.1016/j.neucom.2021.04.039.

(21) Cheung, M.; Moura, J. M. F. Graph Neural Networks for COVID-19 Drug Discovery. In *2020 IEEE International Conference on Big Data (Big Data)*; 2020; pp 5646–5648. https://doi.org/10.1109/BigData50022.2020.9378164.

(22) Han, K.; Lakshminarayanan, B.; Liu, J. Reliable Graph Neural Networks for Drug Discovery Under Distributional Shift. arXiv November 29, 2021. https://doi.org/10.48550/arXiv.2111.12951.

(23) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *Journal of cheminformatics* **2009**, *1*, 1–11.

(24) Lipinski, C. A. Lead-and Drug-like Compounds: The Rule-of-Five Revolution. *Drug discovery today: Technologies* **2004**, *1* (4), 337–341.

(25) Mohamed, A.; Brooks, B. R.; Amin, M. Leveraging-Induced Polarization for Drug Discovery: Efficient IC50 Prediction Using Minimal Features. *J. Chem. Inf. Model.* **2025**, *65* (7), 3715–3722. https://doi.org/10.1021/acs.jcim.5c00076.

(26) Pedron, C. N.; Torres, M. D. T.; Oliveira, C. S.; Silva, A. F.; Andrade, G. P.; Wang, Y.; Pinhal, M. A. S.; Cerchiaro, G.; da Silva Junior, P. I.; da Silva, F. D.; Radhakrishnan, R.; de la Fuente-Nunez, C.; Oliveira Junior, V. X. Molecular Hybridization Strategy for Tuning Bioactive Peptide Function. *Commun Biol* **2023**, *6* (1), 1067. https://doi.org/10.1038/s42003-023-05254-7.

(27) Claudio Viegas-Junior; Amanda Danuello; Vanderlan da Silva Bolzani; Eliezer J. Barreiro; Carlos Alberto Manssour Fraga. Molecular Hybridization: A Useful Tool in the Design of New Drug Prototypes. *Current Medicinal Chemistry* **2007**, *14* (17), 1829–1852. https://doi.org/10.2174/092986707781058805.

(28) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chemical science* **2019**, *10* (2), 370–377.

(29) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365.

(30) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic acids research* **2024**, *52* (D1), D1180–D1192.

(31) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic acids research* **2015**, *43* (W1), W612–W620.

(32) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; Gedeck, P.; Jones, G.; Kawashima, E.; NadineSchneider; Nealschneider, D.; Dalke, A.; tadhurst-cdd; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; Vaucher, A.; Wójcikowski, M.; Maeder, N.; Faara, H.; Take, I.; Walker, R.; Scalfani, V. F.; Probst, D.; Ujihara, K.; Pahl, A.; godin, guillaume; Lehtivarjo, J. Rdkit/Rdkit: 2025_09_2 (Q3 2025) Release, 2025. https://doi.org/10.5281/zenodo.17495409.

(33) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. arXiv February 4, 2018. https://doi.org/10.48550/arXiv.1710.10903.

(34) *Molecular Connectivity in Chemistry and Drug Research*; 2012.

(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(36) Ansel, J.; Yang, E.; He, H.; Gimelshein, N.; Jain, A.; Voznesensky, M.; Bao, B.; Bell, P.; Berard, D.; Burovski, E. Pytorch 2: Faster Machine Learning through Dynamic Python Bytecode Transformation and Graph Compilation; 2024; pp 929–947.

(37) Fey, M.; Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric*. arXiv.org. https://arxiv.org/abs/1903.02428v3 (accessed 2025-11-11).

(38) Fey, M.; Sunil, J.; Nitta, A.; Puri, R.; Shah, M.; Stojanovič, B.; Bendias, R.; Barghi, A.; Kocijan, V.; Zhang, Z.; He, X.; Lenssen, J. E.; Leskovec, J. *PyG 2.0: Scalable Learning on Real World Graphs*. arXiv.org. https://arxiv.org/abs/2507.16991v2 (accessed 2025-11-11).

(39) Roskoski Jr, R. Janus Kinase (JAK) Inhibitors in the Treatment of Inflammatory and Neoplastic Diseases. *Pharmacological research* **2016**, *111*, 784–803.

(40) Wang, Y.; Gu, Y.; Lou, C.; Gong, Y.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. A Multitask GNN-Based Interpretable Model for Discovery of Selective JAK Inhibitors. *Journal of cheminformatics* **2022**, *14* (1), 16.

(41) Sakai, M.; Nagayasu, K.; Shibui, N.; Andoh, C.; Takayama, K.; Shirakawa, H.; Kaneko, S. Prediction of Pharmacological Activities from Chemical Structures with Graph Convolutional Neural Networks. *Scientific reports* **2021**, *11* (1), 525.

(42) Yin, Y.; Hua, H.; Li, M.; Liu, S.; Kong, Q.; Shao, T.; Wang, J.; Luo, Y.; Wang, Q.; Luo, T. mTORC2 Promotes Type I Insulin-like Growth Factor Receptor and Insulin Receptor Activation through the Tyrosine Kinase Activity of mTOR. *Cell research* **2016**, *26* (1), 46–65.

(43) Saxton, R. A.; Sabatini, D. M. mTOR Signaling in Growth, Metabolism, and Disease. *Cell* **2017**, *168* (6), 960–976.

(44) Harwood, F. C.; Klein Geltink, R. I.; O'Hara, B. P.; Cardone, M.; Janke, L.; Finkelstein, D.; Entin, I.; Paul, L.; Houghton, P. J.; Grosveld, G. C. ETV7 Is an Essential Component of a Rapamycin-Insensitive mTOR Complex in Cancer. *Science advances* **2018**, *4* (9), eaar3938.

(45) Xu, N.; Fang, W.; Mu, L.; Tang, Y.; Gao, L.; Ren, S.; Cao, D.; Zhou, L.; Zhang, A.; Liu, D. Overexpression of Wildtype EGFR Is Tumorigenic and Denotes a Therapeutic Target in Non-Small Cell Lung Cancer. *Oncotarget* **2015**, *7* (4), 3884.

(46) Nada, H.; Gul, A. R.; Elkamhawy, A.; Kim, S.; Kim, M.; Choi, Y.; Park, T. J.; Lee, K. Machine Learning-Based Approach to Developing Potent EGFR Inhibitors for Breast Cancer─Design, Synthesis, and in Vitro Evaluation. *ACS omega* **2023**, *8* (35), 31784–31800.

(47) Yang, G.; Nowsheen, S.; Aziz, K.; Georgakilas, A. G. Toxicity and Adverse Effects of Tamoxifen and Other Anti-Estrogen Drugs. *Pharmacology & Therapeutics* **2013**, *139* (3), 392–404. https://doi.org/10.1016/j.pharmthera.2013.05.005.

(48) Limb, G. A.; Matter, K.; Murphy, G.; Cambrey, A. D.; Bishop, P. N.; Morris, G. E.; Khaw, P. T. Matrix Metalloproteinase-1 Associates with Intracellular Organelles and Confers Resistance to Lamin A/C Degradation during Apoptosis. *The American journal of pathology* **2005**, *166* (5), 1555–1563.

(49) Kumar, S.; Manoharan, A.; Abdelgawad, M. A.; Mahdi, W. A.; Alshehri, S.; Ghoneim, M. M.; Pappachen, L. K.; Zachariah, S. M.; Aneesh, T.; Mathew, B. Correction: Exploiting Butyrylcholinesterase Inhibitors through a Combined 3-D Pharmacophore Modeling, QSAR, Molecular Docking, and Molecular Dynamics Investigation. *RSC advances* **2023**, *13* (16), 10956–10956.