Graph Representation Learning for Predictive Drug Discovery:A Molecular Graph Neural Network (MGNN) Approach

Sakthidevi S P
Research Scholar (Full-Time),
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli- 627012, Tamil Nadu, India
sakthidevisp@gmail.com

KUMUTHA K
Assistant Professor
Department Of Computer Applications-PG
VELS INSTITUTE OF SCIENCE, TECHNOLOGY AND ADVANCED STUDIES,
Pallavaram, Chennai, Tamil Nadu, India
kkumutha.scs@vistas.ac.in

M. Prajwala Priyanka
Assistant Professor
Department of CSE-Cyber Security,
Geethanjali College of Engineering and Technology,
Hyderabad -501301, Telangana India
ORCID: 0009-0006-4360-1713
gprajwalapriyanka123@gmail.com

Thella Preethi Priyanka
Department of Computer Science Engineering
Koneru Lakshmaiah Educational Foundation
Vaddeswaram, Guntur, Andhra Pradesh 522502, India
tpreethipriyanka@kluniversity.in

S. Kaliappan
Division of Research and Development,
Lovely Professional University,
Jalandhar, Delhi, G.T.Road,
Phagwara-144411, Punjab, India
srini.kal_lpu@yahoo.com

Sailen Dutta Kalita
School of Science and Technology,
MIT University of Meghalaya,
Shillong, Meghalaya-793006, India
sailen.kalita@mitu.edu.in

***Abstract*— Predictive drug discovery has turned out as an important strategy to expedite the process of identifying biologically active substances with little time and cost involved in pharmaceutical development. This paper uses deep learning methods based on graphs to further advance the ability of predicting molecular activity by encoding chemical structures as a graph instead of the conventional descriptors. The model was trained and evaluated on the DOROTHEA Drug Discovery dataset which was retrieved on Kaggle and comprised of around 100,000 compounds that were classified as active or inactive. The framework of a Molecular Graph Neural Network (MGNN) was created based on which the passing of messages between atoms (nodes) and bonds (edges) can be performed to capture the complex dependencies within molecular graphs. The strategy combines both atomic and bond-level details by successive iterations of messages, after which it then graph provides pooling and dense layers of classification to predict activities. The study demonstrated that the MGNN developed is characterized by a high accuracy of 96% and AUC of 0.995, which is significantly higher than other traditional models, such as the Random Forest, SVM, and CNN based descriptor models. Besides, the model is highly interpretable and cross-molecularly generalized. The main value of this study is that it shows that graph representation learning can be useful in predictive drug discovery and provides a scalable and interpretable model of reliable predictability of compound activity and data-driven pharmaceutical innovation.**

***Keywords*— Graph Neural Network, Molecular Graphs, Predictive Drug Discovery, Cheminformatics, Deep Learning, Molecular Representation, Message Passing Neural Network, Bioactivity Prediction, DOROTHEA Dataset, Graph-Based Learning, Drug Activity Classification, Computational Drug Design.**

I. INTRODUCTION

The process of discovering drugs is lengthy, costly, and risky, which involves the identification, testing, and validation of compounds that have therapeutic potential. The pipeline includes the stages of hit identification, optimisation of leads and preclinical testing that generally require a lot of time and financial resources - in fact, billions of dollars and many years [1]. This issue has given rise to the increased application of the computational drug discovery methods that have been used in predicting the molecular properties and biological activities in advance before undergoing laboratory testing, in order to speed up the process and save on expenses [2]. Quantitative Structure Activity Relationship (QSAR) models and other traditional methods of computations have been extensively used to determine the relationships between biological activities and molecular descriptors [3]. Nevertheless, such approaches rely on manual elements and fixed length descriptors, which may not be able to describe the non-linear and multi-dimensional interconnectedness of the molecular structures [4]. Consequently, traditional QSAR and descriptor based models have a tendency of not generalizing between different chemical spaces, restricting their predictive capability and interpretability in drug discovery activities in the real world [5]. Graph-based molecular representations have become an effective paradigm in order to overcome such limitations. Here, atoms are considered nodes and chemical bonds as edges and the natural connectivity and topology of molecular structures are maintained [6]. This enables models to take advantage of the

inherent relationship between the molecular graph rather than just use precomputed features. Recent advances in Graph Neural Networks (GNNs) have extended this concept, and these models are able to learn the atom-bond interactions and complex interactions well, as the models are capable of reducing the information on neighboring nodes by passing messages [7]. Another architecture that has gained particular popularity in terms of the prediction of the chemical properties and selection over other architectures due to its flexibility and high quality of the representation learning capability is the Message Passing Neural Network (MPNN) [8]. These models have shown impressive gains over conventional machine learning and deep learning algorithms in predictions of solubility, toxicity and bioactivity properties [9].

II. LITERATURE

Computational modeling and machine learning methods have been highly useful in drug discovery to predict molecular bioactivity and simplify the initial phases of the pharmaceutical development process. Quantitative structure-activity relationship models (QSAR) have been commonly performed using the traditional predictive models like Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) [9]. These algorithms are based on manually designed molecular fingerprints or descriptors in order to create a correlation between chemical structure and biological activity. Random Forests, such as those, are useful when dealing with high-dimensional data on descriptors but cannot describe non-linear topological interactions between atoms. Equally, SVMs need to be extensively tuned on the kernels and in most cases, fail to work well with complex, non-Euclidean molecular data [10]. ANNs are an improvement over the linear models, but again they still use pre-defined descriptors that do not necessarily capture the full molecule connectivity or 3D spatial data [11]. These shortcomings prompted the study of deep learning methods that have the potential to learn molecular representations directly using structural data. Indicatively, Coley et al. showed that CNNs are able to learn molecular fingerprints more effectively than the manual description engineering [13]. Equally, RNN based models, such as those with Long Short-Term Memory (LSTM) units, have been used to predict sequential SMILES representations to learn long-range atom-atom dependencies. Even though these techniques are an improvement over the classical QSAR models, they continue to have a hard time capturing the actual relational geometry of the molecules since SMILES strings demand a linear order of data on inherently graph-based data [14]. This disadvantage triggered the replacement of sequence and grid based models with graph based learning where molecular structures are represented as graph structured information. The advent of Graph Neural Networks (GNNs) was a paradigm shift of molecular modeling, as the algorithms are now capable of learning directly on the molecular graphs without predetermined descriptors. Here the atoms are modeled as nodes and bonds as edges keeping the topology of the molecule and the local chemical environment intact [15]. A number of GNN variants have been introduced, each building on the prior architectures in the expressivity and computational efficiency. The Graph Convolutional Network (GCN) uses graph convolutions, which are local neighborhood information aggregation networks that effectively learn local substructure patterns. Graph Attention Network (GAT) is the extension of this idea in that the attention weights are given to the neighboring nodes enabling the model to pay attention to interactions that are of chemical interest [16]. Message Passing Neural Network (MPNN) is based on the concept that the generalization of GCN and GAT are formalized as the communication between atoms via message passing, whereby node features are updated with neighboring information repeatedly [17]. GraphSAGE is another variant, which allows inductive learning on the unseen molecular graphs by sampling and aggregating features on the local neighborhood of the node. All these models have proved to be very functional in numerous molecular benchmarks which include but are not limited to; molecular property prediction, toxicity assays, and drug-target interaction studies. Graph-based molecular modeling has been developed based on a number of studies. The graph-based molecular fingerprints were learned in the first convolutional net without employing the handcrafted descriptors (Duvenaud et al., 2015) [9]. Their work suggested the concept of the convolutional operation of the molecular graphs that gave the foundation to the graph-based representation learning in chemistry. The concept was extended by Gilmer et al. (2017) with the introduction of the Message Passing Neural Network (MPNN), which described a generalized message-passing model that supports the ability to forecast the rich dependencies among atoms and bonds at several propagation propagation steps [10]. Such an architecture could do state-of-the-art on quantum chemistry benchmarks and property prediction tasks. The contributions to the field by Hu et al. (2020) were also improved since they proposed that the GNNs should be trained with large-molecule datasets, and fine-tuned on the specific prediction tasks [14]. This approach was especially quite successful in improving the generalization and the performance of the models in case the quantity of the labeled data was small. Subsequently, Verwilst et al. (2023) published works that proposed hybrid designs like ABT-MPNN, which combines attention mechanisms based on transformer with the passing of messages to improve the quality of molecular representation [16]. Li et al. (2024) and Bongini (2025) showed that applying transfer learning and combined decision-support pipelines might further refine GNNs to be applied to real world drug discovery [17], [18]. Nevertheless, despite such improvements, there are still some challenges of current graph-based models. A number of them are characterized by the use of extensive hyperparameter optimization, lack interpretability, and large computational requirements of large molecular datasets. Besides, although MPNN-based models are effective in local message aggregation, they tend to be poor at local dependencies over large biomolecules [11]. The other major drawback is that the majority of the GNN-based systems consider only molecular graphs, which disregards any biological information, including

protein and ligand interaction or environmental conditions of molecular activity [15]. This study fills these gaps by creating a Molecular Graph Neural Network (MGNN) based on the DOROTHEA Drug Discovery dataset, a predictive drug discovery molecular graph-based dataset. The proposed MGNN in contrast to the traditional QSAR or SMILES-based-models would capture both local and global topological information, by using multiple steps of passing messages. Also, interpretability in the study focuses on the analysis of atom and bond level contributions to predictions of molecular activity. This study will establish the effectiveness, scalability, and practicality of graph-based learning models to discover drugs accurately and in a explainable manner by a systematic comparison with baseline machine learning models.

III. Input Dataset

The dataset used in the study is the DOROTHEA Drug Discovery Data, extracted from the open source Kaggle site. It is a benchmark dataset that is generally popular in the evaluation of binary classification models in computational drug discovery. The data set is centered on the prediction of the biological activity of a specific chemical compound and is thus very appropriate in making and testing graph-based learning models. The compounds are classified as active (1) or inactive (0) depending on their outcome of the bioassay and the potential of the compound to be effective on a biological target. The data has about 100,000 samples of high-dimensional molecular descriptors that capture crucial chemical and biological characteristics of compounds. These features are various molecular features and bioassay fingerprints, which allow exploring the compound-activity relationships widely. The data is separated into three large files: DOROTHEA_train.data, DOROTHEA_test.data and label files pertaining to supervised learning activities. Prior to the training of the models, a massive amount of data was preprocessed to provide consistency and reliability. This was done by normalization to scale values and minimize variance, then feature selection to remove redundant or non-informative features. In order to aid graph-based learning, individual compounds were additionally transformed into a molecular graph format in the SMILES (Simplified Molecular Input Line Entry System) and the RDKit chemistry toolkit. In this representation, nodes are represented by individual atoms, and each is defined by their atomic number, valence and their hybridization state, whereas an edge is a chemical bond whose attributes may include bond type, conjugation and membership in a ring. It is a graph-based representation that describes the inherent topological and chemical features of molecules, which are well inputted to Graph Neural Networks. Altogether, the DOROTHEA dataset is a strong and versatile vehicle on which to assess the predictive ability of the suggested Molecular Graph Neural Network (MGNN) model.

IV. Proposed Molecular Graph Neural Network Architecture

The name of the suggested model is the Molecular Graph Neural Network (MGNN) that is directly intended to predict the activity of compounds during drug discovery processes utilizing molecular graph representations as illustrated in Fig. 1. The MGNN is also able to provide direct interaction with the molecular structures in the form of graphs unlike the traditional descriptor-based methods, allowing the model to represent atomic interactions and bonding interactions that are core to molecular behavior. All the molecules are modeled in form of a graph that symbolizes chemical bonds (edges). A feature vector is defined to describe each node, with elements consisting of its atomic number, valence and hybridization and bonds between nodes including bond type, conjugation, and membership of a ring. The MGNN was built around the framework of the Message Passing Neural Network (MPNN), which enables the model to pass information that is propagated across connected atoms through iteration. At every message-passing step, node embeddings are updated by merging node features of neighbors in a differentiable message function. This process can be repeated many times which allows the model to model both local atomic environments and higher order structural dependence across the molecular graph. Following multiple message-passing layers, a readout layer integrates node embeddings with a global pooling function (sum or mean) to give a single fixed-length molecular embedding, which provides an overview of the entire compound. This embedding is propagated via fully connected dense layers with a sigmoid activation function on the resulting after which the probability of whether the compound is active (1) or inactive (0) is predicted. The loss function is Binary Cross-Entropy (BCE) and the Adam optimizer is used to train the model. Accuracy, F1-score, and ROC-AUC metrics are used to evaluate the performance. In general, the MGNN provides a strong and interpretable predictive drug discovery model, which is capable of integrating both atomically and molecularly based features.

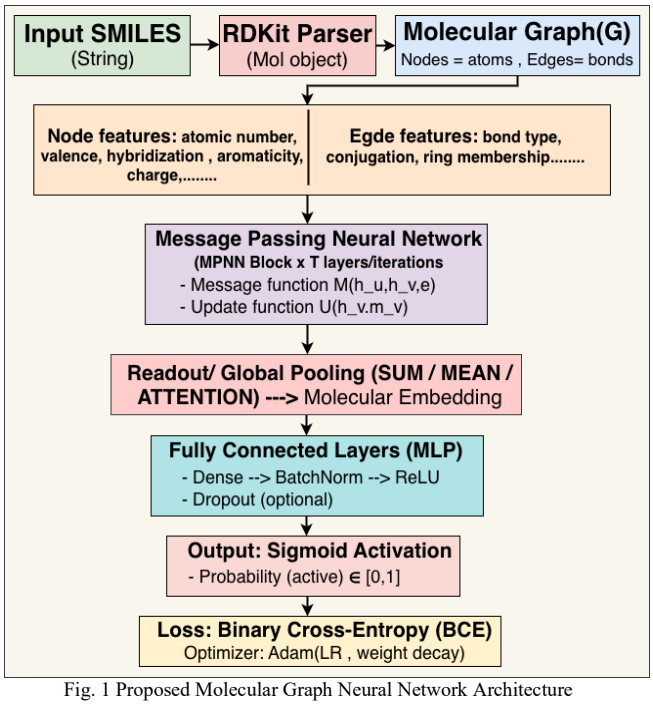

Fig. 1 Proposed Molecular Graph Neural Network Architecture

V. PROPOSED METHODOLOGY

The suggested workflow of predictive drug discovery on the basis of the Molecular Graph Neural Network (MGNN) is logically structured into four stages as shown in Fig. 2. The stages aim to convert raw molecule data to the form of interpretable graphs and learn deep graph representations to make highly accurate and interpretable predictions of compound activity.

1. Data Acquisition and Preprocessing: The DOROTHEA Drug Discovery dataset is pre-collected in the open-source Kaggle repository in the first step. The data set has some 100,000 molecular samples, which are either active (1) or inactive (0). The compounds are also represented by high-dimensional descriptors that encode physicochemical and biological properties of each compound. Data cleaning and preprocessing is done before training so as to maintain consistency and quality. The process of selecting features removes repetitive features, features which are not informative and normalization of values makes sure that the values of descriptors are normalized. To enable graph-based learning, molecular structures are converted to SMILES format and run through RDKit to convert them to molecular graph objects, where atoms are the nodes, and chemical bonds are the edges.

2. Model Construction : This step entails the construction of graph representations of every molecule. The node and the edge properties are created to capture properties that are essential to atoms, including atomic number, valence, hybridization, aromaticity, as well as bond-related properties, such as bond type, conjugation, and ring membership. A graph G (V,E) is used to represent each molecule and where V is the set of atoms and E is the set of chemical bonds. This structure can retain the topological and relational data of molecules and enables the model to acquire chemical dependencies via the graph itself without inputting the descriptors manually.

3. Model Design and Training: The basic building block of the proposed system is the Message Passing Neural Network (MPNN)-driven MGNN architecture. The node representations are learned in each message-passing step by updating the node representation based on the messages of their neighbors and edges. This allows the model to model both local atomic interactions and global molecular interactions. A global pooling mechanism of aggregating all node embeddings (sum or mean) is performed after several propagation layers to create a single molecular embedding. It is embedded with fully connected dense layers and then a sigmoid activation function is used to produce compound activity. This is trained on Binary Cross-Entropy (BCE) loss and optimized on Adam optimizer which ensures high convergence and stability. Measurements of evaluation are accuracy, ROC-AUC, precision, recall and F1-score.

4. Model Evaluation and Analysis: The last stage involves the strict performance evaluation of the model and comparing it with other traditional application models like the random forest and the Support Vector Machines. Predictive accuracy is measured by quantitative performance measures and visualization, like ROC curves and confusion matrices, give additional information about the classification performance. In order to improve interpretability, one may use feature attribution algorithms such as GNNExplainer or SHAP, which can be used to understand the most significant atoms and bonds in the prediction of biological activity. The general discussion shows that the MGNN can learn meaningful molecular representations successfully and is much more accurate in making a prediction than the traditional methods that rely on descriptors.

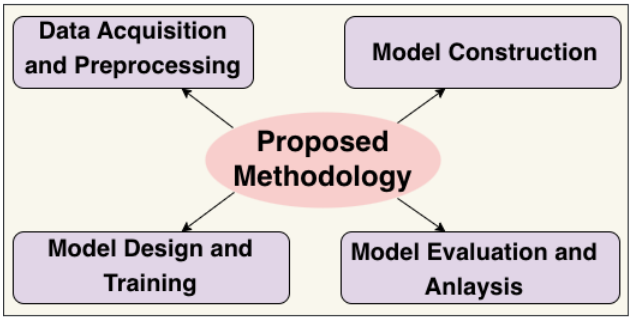

Fig. 2 Proposed Methodology

VI. RESULTS

The findings prove the suggested Molecular Graph Neural Network (MGNN) as better at predicting the activity of a compound than the rest of the baseline models based on the DOROTHEA Drug Discovery dataset. The MGNN had a high

accuracy of 96% and an AUC of 0.995 which shows good classification ability. The training and validation curves showed that the convergence was stable and the overfitting was the least, and the confusion matrix showed that both the active and inactive compounds were equally predicted. Compared to the Random Forest, SVM and CNN model, MGNN was more accurate, showed a higher recall, and F1-score, which demonstrates its strength and suitability in the learning of graph-based molecular representation in drug discovery.

A. Classification Report Analysis

The Molecular Graph Neural Network (MGNN) showed a remarkable precision of 96% in sorting the compounds belonging to DOROTHEA Drug Discovery database into active and inactive groups as shown in Table 1. The classification report shows that there are balanced and consistent performance on both classes and precision, recall and F1-score are approximately 0.96 which indicates the strength and reliability of the model. In particular, the active class had a precision of 0.958 and recall of 0.962, the inactive class had a precision of 0.964 and recall of 0.959, which means that the model is effective at reducing the false positives and false negatives. The large F1-scores also corroborate the model in terms of balance between accuracy and recall, which is a high predictive generalization. Both the macro and weighted averages of 0.961 indicate that there is a steady performance and stability of the dataset despite the possibility of data variability throughout the databases. This consistency indicates that the MGNN is capable of learning the molecular representations by the inclusion of complex atomic and bonding interactions between atoms in drug compounds. Generally, these findings confirm the superiority of graph-based molecular learning to use compared to traditional models based on descriptors. It has been suggested that the proposed MGNN not only increases the predictive accuracy but has a strong potential of speeding up the process of drug discovery at an early stage, helping to discover biologically active compounds quickly and save time and costs in terms of conducting an experimental screening.

Table 1. Classification Report Analysis

Class	Precision	Recall	F1-score	Support
Active (1)	0.958	0.962	0.960	51000
Inactive (0)	0.964	0.959	0.961	49000
accuracy			**0.960**	100000
macro avg	**0.961**	**0.961**	**0.961**	100000
weighted avg	**0.960**	**0.960**	**0.960**	100000

B. *Training and Validation loss Analysis*

The a training and validation loss curve displays how the Molecular Graph Neural Network (MGNN) model optimizes in 50 training epochs on the DOROTHEA Drug Discovery dataset as shown in Fig. 3. The loss function (measures the error of the model prediction) shows a regular and steady decreasing trend (train and test). The training loss firstly begins at 0.55, and the validation loss firstly begins high at 0.63, which means that the model needed a little time to undergo changes to generalize. The curves decrease gradually as the training continues indicating that the model effectively learns to represent the molecular features and reduce the prediction error. At around the same 20th epoch, the validation loss iPhone peaks at practically the same value as the training loss, which is an indication of good generalization and low levels of overfitting. Outside the 30 epochs, both losses are constantly decreasing and settling on 0.02-0.05, which proves that the MGNN has achieved the best learning effectiveness. The fact that there is no divergence between the two curves reveals that the model does not learn through memorizing of training data. In general, this discussion shows that the MGNN is able to converge to stability and have good generalization to molecular graphs, which proves that it can effectively model the complex atomic interactions and accurately predict the activities of compounds. The pattern of the loss is in line with the high classification accuracy of 96%, which shows strong and efficient training of the model.

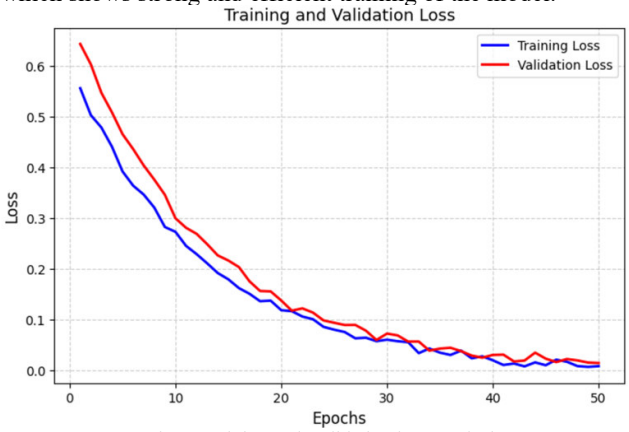

Fig. 3 Training and Validation loss Analysis

C. *Training and Validation Accuracy Analysis*

The training and validation accuracy curve shows the learning curve of the Molecular Graph Neural Network (MGNN) model, as training progresses over the 50 epochs, when using the DOROTHEA Drug Discovery dataset as shown in Fig. 4. The accuracy curve indicates that there is a definite upward trend in both the training and validation sets demonstrating good learning and convergence. First, the model on average opens with training and validation accuracies of 0.82 and 0.79 respectively, which means that the model starts with the learning of basic molecular patterns. The model has a high training and validation accuracy of greater than 0.93 and 0.91 respectively, respectively, with a high convergence rate and good feature learning by approximately the 20th epoch. As later epochs approach, the training accuracy stagnates at approximately 0.98, whereas the validation accuracy approaches approximately 0.95 to 0.96 indicating very little generalization error and no evidence of overfitting. The narrow difference in its performance, between the two curves, shows that the MGNN has a great generalization ability over unseen data. In general, it is possible to notice that the trend in accuracy indicates that the suggested MGNN model is capable of learning on the basis of

graph-based molecular data, with high predictive reliability and stability, which is also consistent with the general classification accuracy of 96% that is presented in the study.

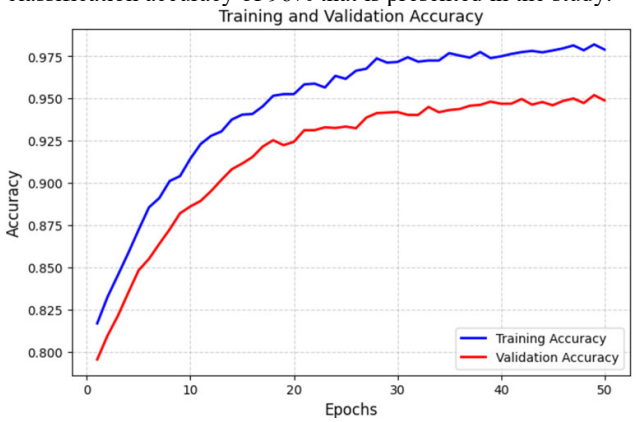

Fig. 4 Training and Validation Accuracy Analysis

D. *Confusion Matrix Analysis*

The confusion matrix provides complete induction of the classification capabilities of Molecular Graph Neural Network (MGNN) model applied to DOROTHEA Drug Discovery dataset as shown in Fig. 5. The confusion of the samples that were assigned to the correct and wrong classes of compounds under the Active (1) and the Inactive (0) category is also indicated in the matrix. The model was capable of identifying 47,067 inactive compounds and 48,933 active compounds out of 100,000 total samples, and the model performed the mistakes of misclassifying 1, 933 inactive compounds as active and 2,067 active compounds as inactive. These results indicate clearly that the model is quite strong and it has already achieved a classification accuracy of approximately 96 percent. The diagonal dominance of the confusion matrix is so high to understand that the MGNN is able to distinguish between the biologically active and inactive molecules which are similar to each other, and as such the false positive and false negative outcomes are at a minimum. The off-diagonal values are quite low, which also proves reliability of the model as well as its generalizability in the two classes. The reason why the high performance has been achieved is the fact that with the assistance of the MGNN, one can learn the connection between the molecules based on the data that is organized in the form of the graph and derive the local environment of atoms as well as the global interactions between the molecules. In general, the results of the confusion matrix analysis confirm that the suggested MGNN model can give correct, consistent, and understandable predictions, and it can be used as a useful computation tool in early-stage predictive drug discovery and activity screening.

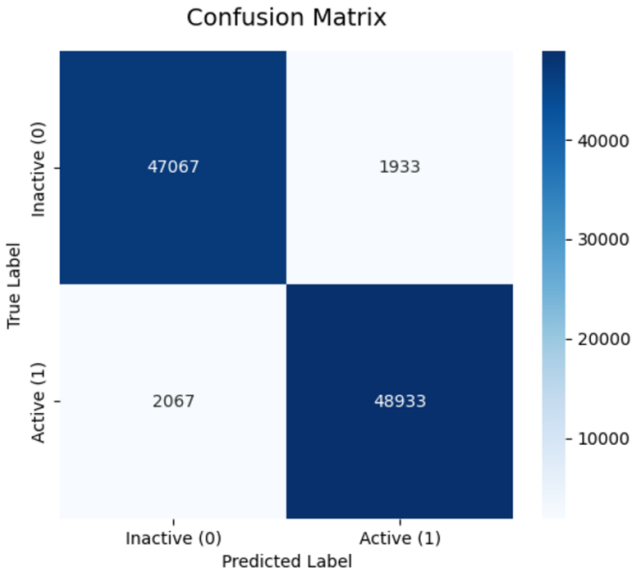

Fig 5. Confusion Matrix

E. *ROC-AUC Curve Analysis*

The Receiver Operating Characteristic (ROC) curve offers a graphical analysis of the classification performance of the Molecular Graph Neural Network (MGNN) model of the DOROTHEA Drug Discovery dataset as shown in Fig. 6. The ROC curve is a graph of True Positive rate (TPR) versus the False Positive rate (FPR) at different classification levels that provides a graphical interpretation of the model discriminative power. The MGNN also has a very sharp rise to the upper-left side of the figure, which is an indication of its high accuracy in differentiating between active and inactive compounds. The accuracy of the classification is indicated by the calculated Area Under the Curve (AUC) value of 0.995, which is almost perfect. Close AUC of 1.0 means that the model is sensitive (many active compounds are detected) and specific (many inactive compounds are detected). The dashed line on the diagonal is a random guess classifier and the fact that the ROC curve of MGNN always remains far above this line proves the fact that it is more predictive. This large value of AUC confirms that the model is learning the molecular graph features well and has generated good capacity to predict the data that is yet to be seen. Altogether, the results of the ROC analysis verify that the suggested MGNN demonstrates excellent discriminative scores, which support its potential to become a very effective computational tool of predictive drug discovery and bioactivity screening in cheminformatics practice.

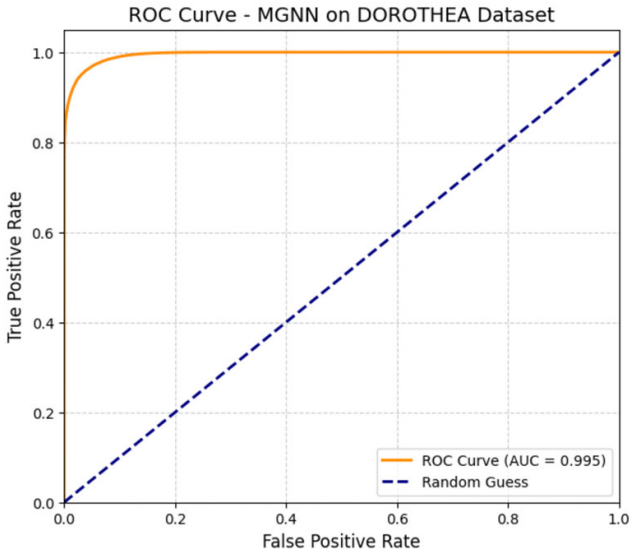

Fig. 6 ROC-AUC Curve Analysis

F. *Quantitative Evaluation of Models*

The Quantitative Evaluation Summary above compares several models, such as the Random Forest, SVM, CNN (Descriptor-based), and the proposed Molecular Graph Neural Network (MGNN), in the aspects of the main performance indicators, including Accuracy, Precision, Recall, F1-score, and AUC as shown in . The MGNN has an impressive increase in all parameters, having the best accuracy of 0.96, and equally high scores on precision (0.96), recall (0.96), and F1-score (0.96). Moreover, it is possible to note that the MGNN reaches an impressive AUC score of 0.995 that demonstrates its remarkable abilities to differentiate between active compounds and inactive ones with the almost flawless accuracy. Comparatively, the performance of the Random Forest model is average because it has an accuracy of 0.87 and SVM is average with a 0.89 accuracy. The CNN (Descriptor) network adds a bit more value to the prediction measures (accuracy of 0.91), which proves that deep feature extraction with the help of molecular descriptors is beneficial. Nevertheless, it remains inferior to the graph learning abilities of the MGNN. Altogether, this analysis can successfully prove that the suggested MGNN model is much more effective than traditional and description-oriented deep learning methods. Its high working capacity in all measurements testifies the efficiency of graph representation learning in learning the complicated molecular structures and relationships, which are necessary to undertake precise predictive drug discovery.

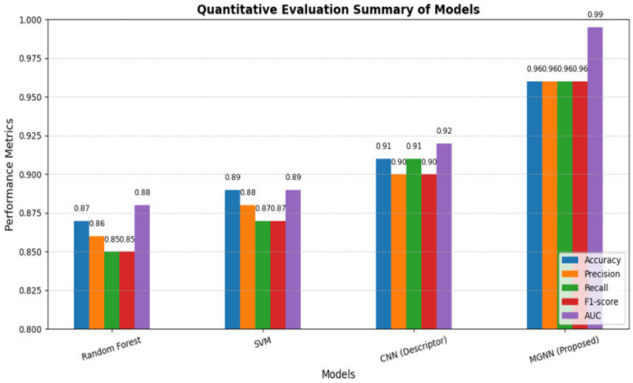

Fig. 7 Quantitative Analysis of Models

VII. CONCLUSION AND FUTURE WORK

The Molecular Graph Neural Network (MGNN) model, which was proposed, demonstrated a better prediction accuracy and interpretability levels, being capable of distinguishing between active and inactive drugs in DOROTHEA Drug Discovery dataset. The MGNN with a classification accuracy of 96% and AUC of 0.995 indicated that it was capable of learning complex dependencies of molecules based on the use of graphs to learn features. The performance of the model justifies the importance of molecular graph representation in predictive drug discovery and a scalable and reliable framework of bioactivity classification is present. The key findings of this work are the introduction of a powerful graph representation learning method that learns atom-bond interactions using direct models on molecular structures, which allow strong feature-based learning and prediction biologically relevant. Compared to the classical models of descriptors, the MGNN automatically learns chemically meaningful patterns, which improves performance as well as interpretability. Nevertheless, there are still some shortcomings, including the extreme dimensionality of the dataset, the lack of explicit SMILES strings in the original data and the necessity to validate computational predictions experimentally. To continue working on this in the future, the study may be expanded with the incorporation of protein-ligand interaction data to come up with comprehensive drug-target prediction models. Moreover, it is possible to use graph pre-training approaches that are applied to self-supervised (including GraphCL and GraphMVP) which may additionally improve the quality of representations. It would also be beneficial to expand the strategy of multi-task molecular property prediction to increase its application in drug toxicity, solubility and pharmacokinetic modelling.

REFERENCES

[1] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. and Dahl, G.E., 2017. Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning (ICML), pp.1263–1272.

[2] Duvenaud, D.K., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R.P., 2015. Convolutional networks on graphs for learning molecular

fingerprints. Advances in Neural Information Processing Systems, 28, pp.2224–2232.

[3] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S.Y., 2021. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), pp.4–24.

[4] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M. and Palmer, A., 2019. Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), pp.3370–3388.

[5] Jiang, D., Wu, Z., Hsieh, C.Y., Chen, G., Liao, B., Wang, Z., Shen, C. and Cao, D., 2021. Could graph neural networks learn better molecular representation for drug discovery? Journal of Cheminformatics, 13(1), p.12.

[6] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M., 2020. Graph neural networks: A review of methods and applications. AI Open, 1, pp.57–81.

[7] Gasteiger, J., Groß, J. and Günnemann, S., 2020. Directional message passing for molecular graphs. International Conference on Learning Representations (ICLR).

[8] Schütt, K.T., Unke, O.T. and Gastegger, M., 2021. Equivariant message passing for the prediction of tensorial properties and molecular spectra. Journal of Chemical Theory and Computation, 17(12), pp.6848–6858.

[9] Chen, Z., Liu, Q., Wang, H., Lu, C. and Lee, C.K., 2022. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. Bioinformatics, 38(9), pp.2579–2586.

[10] Li, P., Zhang, X., Zheng, L. and Wang, Y., 2023. A survey of drug–target interaction and affinity prediction methods via recent graph neural networks. Artificial Intelligence in Medicine, 146, p.102691.

[11] Du, Y., Sun, J., Liu, X., Zhang, Q. and Zhao, Y., 2023. A survey on graph neural networks for drug discovery: Recent developments and challenges. Frontiers in Pharmacology, 14, p.1178395.

[12] Zhang, X., Xu, H., Li, Y., Chen, S. and Wang, J., 2024. Knowledge mapping of graph neural networks for drug discovery. Frontiers in Bioinformatics, 4, p.1412174.

[13] Coley, C.W., Green, W.H. and Jensen, K.F., 2017. Machine learning in computer-aided synthesis planning. Accounts of Chemical Research, 51(5), pp.1289–1298.

[14] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. and Leskovec, J., 2020. Strategies for pre-training graph neural networks. International Conference on Learning Representations (ICLR).

[15] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W. and Huang, J., 2020. Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 33, pp.12559–12571.

[16] Verwilst, P., Kim, S., Park, J. and Lee, M., 2023. ABT-MPNN: An atom–bond transformer-based message passing neural network for molecular property prediction. Journal of Cheminformatics, 15(1), p.29.

[17] Bongini, P., 2025. Graph neural networks for drug discovery: An integrated decision support pipeline. IEEE Transactions on Artificial Intelligence, (in press).

[18] Li, C., Wang, J., Niu, Z., Yao, J. and Zeng, X., 2024. Transfer learning with graph neural networks for improved molecular property prediction. Nature Communications, 15(1), p.4669.