Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

Early Drug Discovery: Target Identification, Screening and Optimization using GCN

Bheshajaa P, Nikhil K Nagavar, Pavan Dev L, Shashank B Poojary, Prof. Shobha Y
Department of Artificial Intelligence and Machine Learning
Bangalore Institute of Technology
Bangalore, India

***Abstract*— The proposed AI-driven framework will revolutionize drug discovery by leveraging innovative technologies and overcoming inefficiencies in the traditional approaches. PPI networks along with Graph Convolutional Networks are there to identify the target in question; 3D- Convolutional Neural Networks and Generative Adversarial Networks are then used for hit generation and compound screening, which is then done using reinforcement learning algorithms to come up with a robust lead as it identifies some of the best candidates. The ADMET predictions further support the pharmacokinetic and toxicological properties, hence avoiding late-stage failures. Further acceleration in the process of drug discovery comes with a data-driven process by greatly reducing the costs and increasing the success rate of clinical trials associated with reduced attrition and better decision-making.**

***Keywords*— AI-based drug discovery, Protein-Protein Interaction networks, Graph Convolutional Networks, 3D Convolutional Neural Networks, Generative Adversarial Networks, Reinforcement Learning, ADMET prediction, drug development.**

I. INTRODUCTION

Early drug discovery is a very complex, very time and resource-consuming process involving the identification of therapeutic targets, design of lead compounds, and optimization towards efficacy and safety. Current traditional methods are often inefficient, not scalable, and costly, which makes the process slow and unsustainable. This project proposes an AI- driven framework that is said to transform the drug discovery pipeline by transcending the barriers. The framework is based on some computational methods including Protein-Protein Interaction (PPI) networks, Graph Convolutional Networks (GCNs), and Generative Adversarial Networks (GANs) for the identification of targets and compound generation. It checks 3D-CNN binding potential through virtual screening while applying Reinforcement Learning (RL) to optimize lead compounds for toxicity and efficacy. The ADMET prediction further ensures pharmacokinetics and toxicological viability; hence, the risk of late-stage failures is highly minimized. This integrated approach accelerates timelines in the discovery phase and conserves a lot of cost while maximizing success rates in clinical situations.

II. RELATED WORK

[1] This paper proposes an AI-driven framework that leverages advanced computational methods to address challenges in drug discovery. Previous studies have demonstrated the effectiveness of AI and in which machine learning techniques in this domain. A tensor factorization model combined with knowledge graph embeddings has been used to predict those potential drug targets for diseases, significantly improving prediction accuracy and outperforming different machine learning methods in identifying disease-target interactions.

[2] Machine learning approaches such as SVM and CNN are used for drug identification based on large character recognition from medicine labels. Fragment-link methods have been employed to divide and recognize drug names, providing audio output for identified drugs.

[3] Neural networks have been benchmarked for drug candidate selection, highlighting machine learning's role in reducing costs and improving drug development efficiency.

[4] Generative Adversarial Networks (GANs) combined with Reinforcement Learning (RL) have been proposed to automate drug discovery pipelines, accelerating the identification of drug candidates and improving pre-clinical productivity.

[5] Various feature selection techniques are compared to address high dimensionality in drug datasets. Machine learning models have been evaluated using K-fold cross-validation to enhance predictive accuracy.

[6] Deep learning and natural language processing (NLP) are employed for drug repurposing by mining biomedical literature. Automated text mining reveals new drug-disease relationships, offering a cost-effective and accelerated approach to drug development.

[7] AI advancements have also been applied in drug discovery to predict drug properties and interactions, utilizing key data resources and molecular representation methods.

[8] Explores in silico toxicity prediction for workplace chemicals. Highlights computational methods for hazard assessment. Aims to improve chemical safety and risk management. Supports regulatory decision-making in toxicology.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

[9] Reviews drug discovery with a focus on ADMET properties. Discusses absorption, distribution, metabolism, excretion, and toxicity. Highlights computational tools for ADMET evaluation. Enhances drug screening and safety assessment.
[10] Introduces DeepAffinity for compound-protein interaction prediction. Uses recurrent and convolutional neural networks for modeling. Improves drug binding affinity predictions. Enhances AI-driven drug discovery strategies.
[11] Analyses adverse drug reactions from medication switches. Examines patient safety and pharmacovigilance data. Identifies risks associated with switching medications. Supports improved drug monitoring systems.
[12] Compares cryo-electron microscopy and X-ray crystallography. Explores structural biology applications in drug discovery. Highlights their complementary strengths in molecular analysis. Enhances protein structure determination for pharmaceuticals.
[13] Review machine learning in drug discovery applications. Discuss various ML techniques for drug candidate identification. Improve efficiency and accuracy in pharmaceutical research. Highlight AI's role in accelerating drug development.
[14] Machine learning in drug discovery applications. Drug candidate identification by different ML techniques. Efficiency and accuracy improvement in pharmaceutical research. AI in accelerating drug development.

III. METHODOLOGY

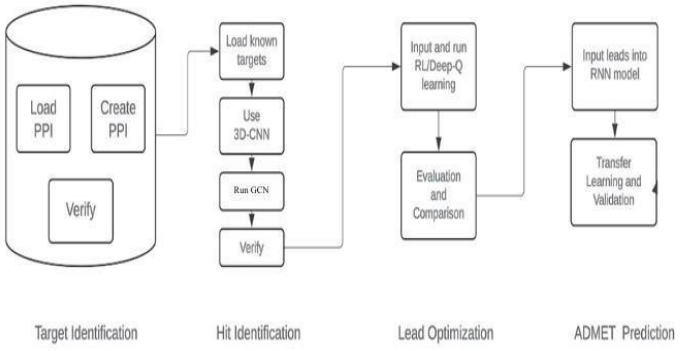

Fig. 1. System Architecture

A. Target Identification:

The PPI networks have dominated the identification of critical proteins that have been linked up with disease mechanisms. Analyzing all of these networks which gives the researcher an opportunity to focus on hub proteins as central nodes in disease pathways. Advanced computational methods, such as clustering and each network analysis that we do for the target and, are available in prioritizing those proteins. Additionally, structural and functional annotations of these proteins provide insights into their biological roles, aiding in the selection of viable drug targets. This systematic approach ensures efficient utilization of resources in the early stages of drug discovery and reduces the time required for downstream processes.

B. Hit Identification:

Three-dimensional convolutional neural networks (3D-CNNs) are utilized to predict the binding potential of small molecules to target proteins. These models assess the molecular shape, flexibility, and electrostatic interactions using structural data obtained during target identification.

They offer high-resolution views of ligand-receptor interactions, allowing the identification of promising compounds with high binding affinity. Moreover, cheminformatics techniques, such as molecular docking and virtual screening, are integrated with 3D-CNN predictions to enhance hit identification accuracy.

A pipeline of lead candidate validation in follow-up experiments becomes the basis of down-stream optimization and development in a drug discovery.

C. Lead Optimization:

Chemical structures under DQL optimization can be made by iterative improvements through reinforcement learning. In it, several criteria like potency, safety, solubility, and bioavailability are balanced, while advanced simulation tools of the molecular type combined with DQL explore space to generate a new molecule showing desired properties.

Another target for optimization is the pharmacokinetic and pharmacodynamic profiles. This process will ensure effective ness alongside safety of any candidates. It is at this point that c omputational predictions are bridged toward experimental testing, offering much reduced attrition rates in later stages of development.

D. ADMET Prediction :

RNNs, which feed on sequential data, assess ADMET profile s of drug candidates for absorption, distribution, metabolism, excretion, and toxicity. These models analyze molecular descriptors and time-series data to predict pharmacokinetic behavior and toxicological risks. By incorporating experimental datasets and results of high-throughput screening, RNNs improve the accuracy of the ADMET predictions we have got from above. Later from this, by integrating various forms of the various bioinformatics of which the tools, and any of the results are made species-specific so that the candidates not only work in humans but also strictly adhere to the regulatory standards. This reduces the risk of any late failures and hastens the way toward moving to the clinical trials stage.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

IV. MODEL DEVELOPMENT

A. Model Architecture

The proposed model for drug the early discovery drug connects both drug features and protein features to predict drug-protein interactions by using advanced neural network frameworks:

- Input Layer:
  Protein features are extracted using this DL2vec and extracted from the phenotypic features, which will be processed using a Graph Convolutional Neural Network.
  Similarly, the drug features were obtained using SMILES Transformer and DeepGO Plus for molecular embeddings, then processed using a Deep Neural Network.

- Graph Convolutional Neural Network (GCN):
  Update the PPI graph nodes embedding that learns the protein features. Include input, hidden, and output layers for drugs to extract relevant features during drug-protein interaction.

- Deep Neural Network (DNN):
  Process drug embeddings through input, hidden, and output layers to extract drug- related features.

- Output Layer:
  Calculate the similarity between drug and protein features in order to predict strength of interaction.

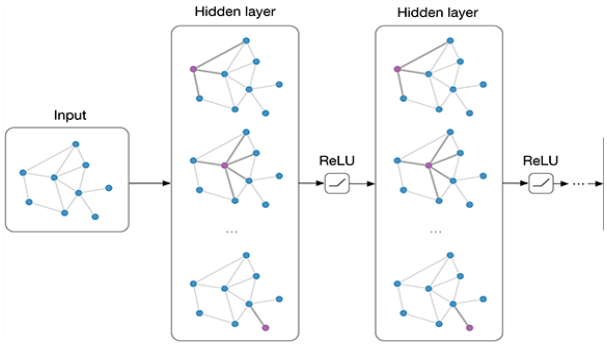

Fig. 2. GCN working

B. Model Compilation

Model was built using the following configurations:

- Optimizer: Adam optimizer to update weights effectively and ensure faster convergence.
- Loss Function: In this the Binary Cross-Entropy loss to ensure high precision of interaction prediction.
- Metric for Evaluation: Accuracy to check on how well the model learns and generalizes when training.

C. Model Training

The training process involved the following steps to ensure the development of robust predictive models:

- Dataset: Preprocessed datasets were utilized, which included:
  - Protein Interaction Graphs: Representing the complex network of protein-protein interactions, providing structural and functional insights for target identification.
  - Drug Molecular Representations: Encoded as SMILES strings, fingerprints, or graph-based molecular structures, these representations enabled precise modeling of drug-target interactions. Encoded as SMILES strings, fingerprints, or graph-based molecular structures, these representations enabled precise modeling of drug-target interactions.

  - Parameters:

1. Batch Size: A batch size of 20 was used, striking a balance between computational efficiency and model convergence stability.
2. Learning Rate: Fine-tuned through experimentation to optimize gradient descent and improve model performance.

D. Model Deployment

The trained model was deployed with the following features:

- Integration with FastAPI:
  - Enables users to upload drug SMILES strings and protein data for real-time predictions.
- Model Loading:
  - Dynamically loads the saved model for inference.
- Input Preprocessing:
  - Transforms uploaded drug and protein data into graph embeddings and molecular embeddings as required by the model.
- Prediction Pipeline:
  - Outputs the similarity score to determine whether the drug will effectively interact with the target protein.

E. Model Evaluation

The model's performance was assessed using test datasets, with the following results:

- Accuracy: Achieved high accuracy in predicting interactions.
- Sensitivity and Specificity: Demonstrated robust predictive power for identifying both active and inactive targets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

V. RESULTS

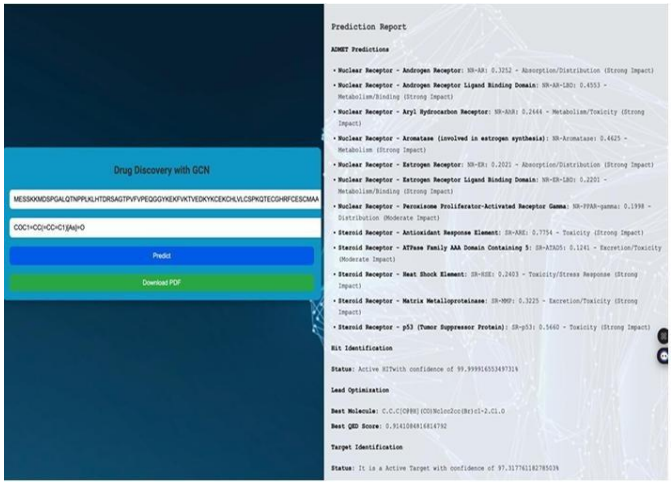

Fig. 3 Active Target

The image presents an active target prediction where the system predicts that the target is active at a 97.31% confidence level. In other words, the considered molecule is most likely to be potent in the interaction with the target, using the input sequence of protein and the SMILES string. Receptor interactions revealed through ADMET predictions give deeper insight into the behavior and effectiveness of a molecule in the development of drugs.

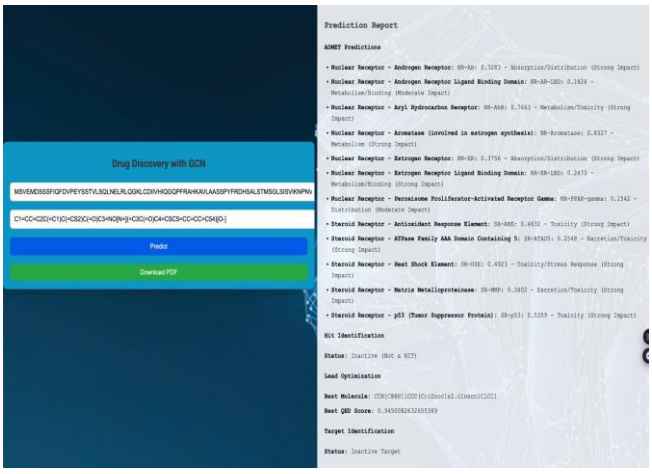

Fig. 4 Inactive Target

This image would show an inactive target, with a low confidence with the score that the protein target is unlikely to interact with the compound in an effective manner. The predictions for receptor interactions, metabolism, and toxicity would suggest that the molecule does not have a strong biological effect or favorable interactions with the target, hence not viable for therapeutic purposes. The inactive target would have much lower.

VI. CONCLUSION

This system of this drug discovery developed in the project offers a highly advanced, it is automated, and highly efficient approach to dealing with the early stages of drug development. Conventional of drug discovery process is lengthy process, and for this resource-consuming, and very complicated, involving massive laboratory work, trial, and error.

In contrast, the system here brings together crucial modules like the Target Identification, the Hit and the Identification, the Lead Optimization, and everything at the point that is the last with all this what we have got ww are going to do the ADMET Prediction with the updates we have got is that we are going to do the streamlining and speeding up the entire process to ultimately make drug discovery efficient and accessible.

The Target Identification module enables the system to make rapid and accurate predictions about whether a given protein is an active drug target. This will be the first instance of early indication of potential therapeutic interventions.

With the use of this SMILES string and the representations, the Hit Identification module will enable fast screening of molecular candidates and candidates that may possibly bind effectively to the target. Lead Optimization is the iterative cycle of improving those candidates into even better lead compounds by enhancing the binding affinity as well as their pharmacological properties.

Finally, ADMET prediction evaluates key drug properties such as absorption, distribution, metabolism, excretion, and toxicity for researchers to understand whether drugs can be safe for testing in the clinical setting. The system reduces the extensive time and material required for conducting drug discovery; it does so by incorporating a cutting-edge system of computational models, machine- learning algorithms, as well as advanced predictive analytics in the process of drug discovery. With its comprehensive capabilities, this system can become a game-changer in pharmaceutical research, being a really valuable tool for accelerating the discovery and development of new, safe, and effective drugs. Moreover, it's adaptable to be scaled and customized for various drug development projects for any researcher to advance therapeutic innovations and eventually improve their overall drug discovery outcomes.

REFERENCES

[1] Y. Cheng, R. Swiers, S. Bonner, and I. Barrett, "A Knowledge Graph Enhanced Tensor Factorisation Model for Discovering Drug Targets," Bioinformatics, vol. 38, no. 3, pp. 594–601, 2022, doi: 10.1093/bioinformatics/btab908.

[2] A. Abdelkrim, A. Bouramoul, and I. Zenbout, "Identification of Drug Discovery for Patients Using Machine Learning," Journal of Biomedical Informatics, vol. 113, p. 103635, 2021, doi: 10.1016/j.jbi.2020.103635.

[3] L. J. A. Marcilin, I. R. Sheeba, M. Sugadev, B. Velan, and P. Chitra, "Neural Drug Discovery," Computational Biology and Chemistry, vol. 91, p. 107511, 2021, doi: 10.1016/j.compbiolchem.2020.107511.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

[4] C. H. Chang, C. L. Hung, and C. Y. Tang, "Drug Discovery using Generative Adversarial Network with Reinforcement Learning," Journal of Chemical Information and Modeling, vol. 59, no. 8, pp. 3247–3258, 2019, doi: 10.1021/acs.jcim.9b00220.
[5] L. Moumné, A. C. Marie, and N. Crouvezier, "Oligonucleotide Therapeutics: From Discovery and Development to Patentability," Pharmaceutics, vol. 14, no. 1, p. 41, 2022, doi: 10.3390/pharmaceutics14010041
[6] S Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Transforming Drug Discovery: Leveraging Deep Learning and NLP for Accelerated Drug Repurposing. Frontiers 10.3389/fphar.2020.574618.
[7] W. Chen, X. Liu, S. Zhang, and S. Chen, "Artificial intelligence for drug discovery: Resources, methods, and applications," Molecular Therapy - Nucleic Acids, vol. 28, pp. 19–35, 2023, doi: 10.1016/j.omtn.2023.01.005.
[8] Rim, K. T. (2020). In silico prediction of toxicity and its applications for chemicals at work. Toxicology and Environmental Health Sciences, 12(2), 205-213. DOI: 10.1007/s13530-020-00450-0.
[9] Patil, P. (2016). Drug Discovery and ADMET process: A Review. Figshare. DOI: 10.6084/m9.figshare.3440570.
[10] Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics, 35(18), 3558-3565. DOI: 10.1093/bioinformatics/btz205.
[11] Glerum, P. J., Maliepaard, M., de Valk, V., Scholl, J. H. G., van Hunsel, F. P. A. M., van Puijenbroek, E. P., Burger, D. M., & Neef, K. (2020). Quantification of Adverse Drug Reactions Related to Drug Switches in The Netherlands. Clinical and Translational Science, 13(3), 495-501. DOI: 10.1111/cts.12723.
[12] Vénien-Bryan, C., Li, Z., Vuillard, L., & Boutin, J. A. (2017). Cryo-electron microscopy and X-ray crystallography: Complementary approaches to structural biology and drug discovery. Acta Crystallographica Section F: Structural Biology Communications, 73(4), 296-304. DOI: 10.1107/S2053230X17025771.
[13] Dara, S., Dharmercherula, S., Jadav, S. S., Babu, C. H. M., & Ahsan, J. (2021). Machine Learning in Drug Discovery: A Review. Molecular Diversity, 25(2), 553-577. DOI: 10.1007/s11030-020-10175-y.
[14] Dadhwal, A., & Gupta, M. (2021). Machine Learning Methods in Drug Discovery. Molecules, 26(16), 4853. DOI: 10.3390/molecules2616485