**AI-Powered Drug Discovery: A Novel Deep Learning**

**Approach for Molecular Bioactivity Prediction**

**Abstract**

Drug discovery is a time-consuming and expensive process, frequently limited by the requirement of significant laboratory experimentation. This work introduces a new deep learning method for predicting molecular bioactivity with SMILES (Simplified Molecular Input Line Entry System) sequences. Rather than depending on conventional chemical descriptors, the approach addresses SMILES as language-like data so that we can tokenize molecules into sequences and feed them into a 1D Convolutional Neural Network (CNN). Through a carefully curated dataset focusing on the EGFR protein, I trained the model to predict compounds as either bioactive or inactive. Accuracy/loss curves, confusion matrix, and ROC curves were utilized for assessment. The model attained validation accuracy of ~96.2% and test-time prediction confidence of 98.3%. This study illustrates the efficacy of NLP-motivated deep learning on molecular strings, providing a potential avenue in AI-based drug discovery.

**Introduction**

Artificial intelligence has over the past few years transformed drug discovery by making it possible to screen and evaluate molecular candidates at a fast pace. Conventional methods mainly rely on engineered features such as molecular fingerprints or physicochemical properties. But they frequently fail to recognize the rich sequential data available within SMILES representations of molecules. Through this project, I harnessed deep learning to identify insightful patterns directly out of tokenized SMILES strings. Drawing from natural language processing, I represented molecular bioactivity as a sequence classification problem. The main objective was to create a neural network that can predict with accuracy if a compound is bioactive, from raw molecular input, and to prove the performance of the model through state-of-the-art visual analytics.

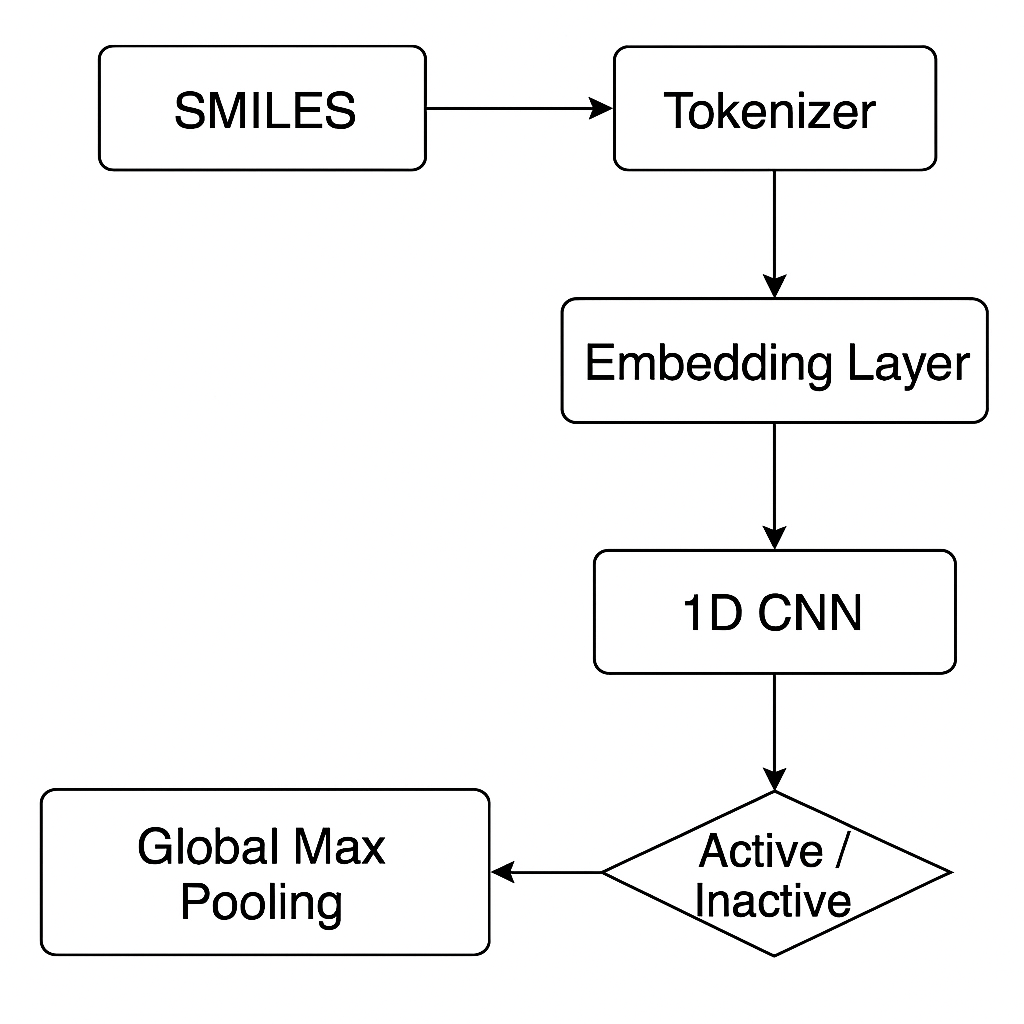
**Methodology**

I employed a publicly available dataset of molecules with 8,700+ SMILES entries against the EGFR protein. The dataset had molecular weight, LogP, TPSA, HBA, and HBD, but modeling was against the SMILES strings only. Every SMILES string was tokenized into a character sequence, and padded to ensure uniform input length. A 1D Convolutional Neural Network (CNN) with an embedding layer, convolutional filters, pooling layers, and fully connected output neurons was developed. The model was trained to classify molecules into two classes: active or inactive. Evaluation was performed with validation split (90/10), and performance was visualized with ROC curves, accuracy/loss plots, and confusion matrix.

**Novelty Introduced**

The novelty of the work is in interpreting molecular structures as language sequences instead of depending on engineered chemical descriptors. Conventional models rely heavily on feature extraction by domain knowledge (such as ECFP fingerprints), whereas this method directly applies deep learning over raw SMILES strings. Drawing on techniques in text classification, the work utilizes token-based CNNs to capture latent molecular patterns. This representation learning type avoids the need for manual features and takes advantage of convolutional networks' sequence modeling capability.

**Architectural Diagram**

****

**Results**

The model was trained for 10 epochs and tested with a 10% validation split. It had a validation accuracy of 96.2% and exhibited a prediction confidence of 98.3% on test input. Evaluation was done on:

Accuracy and Loss curves over training epochs

Confusion Matrix, which verified low misclassification

ROC Curve with good separation between classes

Also, a real-time prediction functionality was created that accepts a new SMILES string as input and returns predicted class ("Active"/"Inactive") with confidence score.

**Conclusion**

This work effectively illustrates that deep learning, specifically CNNs on SMILES sequences, can improve molecular activity prediction during drug discovery. By sidestepping descriptor-based input and employing an NLP-like sequence paradigm, I developed a more flexible, scalable pipeline. With validation accuracy over 96% and real-time inference possible, this model adds a new, data-driven approach to speed up early-stage drug screening.

**Reference Papers**

1. Zheng, S., Li, Y., Chen, S., Xu, J., Yang, Y., & Yang, Y. (2020). Predicting drug–protein interaction using a quasi-visual question answering system. *Nature Machine Intelligence, 2*(2), 134–142. https://doi.org/10.1038/s42256-019-0131-6
2. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell, 180*(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021
3. Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics, 34*(17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593
4. Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., ... & Luo, X. (2020). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry, 63*(16), 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959
5. Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence, 2*(10), 573–584. https://doi.org/10.1038/s42256-020-00236-4
6. Karpov, P., Godin, G., & Tetko, I. V. (2019). A transformer neural network for QSAR modeling. *Journal of Chemical Information and Modeling, 60*(6), 2652–2660. https://doi.org/10.1021/acs.jcim.0c00228
7. Gao, K., Nguyen, D. D., & Wei, G. W. (2020). Generative network complex (GNC) for drug discovery. *Communications in Information and Systems, 20*(2), 83–103. https://doi.org/10.4310/CIS.2020.v20.n2.a1
8. Zhang, Y., Jiang, D., Zhang, L., & Wang, S. (2020). Deep learning for drug–target interaction prediction. *Frontiers in Pharmacology, 11*, 587. https://doi.org/10.3389/fphar.2020.00587
9. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling, 57*(4), 942–957. https://doi.org/10.1021/acs.jcim.6b00740
10. Tang, B., He, F., Liu, D., Fang, M., Wu, Z., & Xu, D. (2019). AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2. *bioRxiv*. https://doi.org/10.1101/2020.03.03.972133