BIO 523 - Chemoinformatics Assignment 1
Paper Summarized: Predicting blood–brain barrier permeability of molecules with a large language model and machine learning

Name: Anjaneya Sharma
Roll Number: 2021449

11th October 2024

Introduction

In their 2024 study, Huang et al. introduce a machine learning approach utilizing the large language model (LLM) MegaMolBART and the XGBoost algorithm to predict whether small-molecule compounds can cross the blood-brain barrier (BBB). The BBB is crucial for developing treatments for central nervous system (CNS) disorders, but its selective nature poses challenges for drug delivery. Prediction methods used uptil now relied on molecular properties such as molecular weight and lipophilicity which require significant computational resources and feature engineering.

The authors aim to use a model to directly predict permeability by encoding molecular structures in SMILES strings..

The authors compared the transformer-based MegaMolBART embeddings with the traditional Morgan fingerprints for representing molecular structures. After coupling the transformer model with XGBoost, they achieved better classification accuracy in predicting BBB permeability. Validation was done through *in silico* and *in vitro* experiments using 3D BBB spheroids to simulate the human BBB.

Research Focus

The primary objective of the research was to create AI models capable of predicting blood-brain barrier (BBB) permeability with greater accuracy than traditional methods. Models often rely on physicochemical properties thereby requiring extensive feature engineering and hence this process is quite time consuming and computationally intensive too, especially in case of large datasets.

Model and Data

In this study, the researchers employed a transformer-based model, specifically MegaMolBART, to encode molecular features for predicting BBB permeability. Key points include:

- **MegaMolBART**: A transformer architecture designed to process molecular structures represented as SMILES strings.
- **SMILES Representation**: SMILES strings encode the chemical structure of molecules in a linear format.
- **XGBoost Classifier**: Used to make final permeability predictions from the embeddings extracted by MegaMolBART.
- **Comparison with Morgan Fingerprints**: Morgan fingerprints were used as a baseline to compare against the transformer-based encoding.

Datasets

The models were trained and tested on datasets like LightBBB and B3DB, containing molecules classified as BBB-permeable or BBB-impermeable. Metrics such as AUC (Area Under the Curve) were used to evaluate the model performance.

In Vitro Validation

In addition to computational predictions, the researchers conducted *in vitro* experiments using 3D human BBB spheroids. The spheroids used actually consisted of cells such as brain endothelial cells, pericytes, and astrocytes thereby closely mimicking the human Blood Brain Barrier.

Methodology

The methodology for this research used the power of both transformer based models, with state-of-the-art gradient boosting techniques. The main steps used in the methodology were:

- **Molecular Encoding**: The SMILES strings representing the chemical structures were used as input to the MegaMolBART transformer model to generate embeddings that capture the relevant molecular features.
- **Training the Classifier**: The molecular embeddings generated by MegaMolBART were then fed into the XGBoost classifier, which was trained to predict whether a given molecule could cross the blood-brain barrier (BBB).
- **Comparison with Baseline**: In parallel, Morgan fingerprints, a traditional encoding method, were used with the same XGBoost classifier to provide a baseline for comparison with the transformer-based approach.
- **Evaluation Metrics**: The performance of the models was assessed using AUC, accuracy, and other relevant metrics, followed by experimental validation using *in vitro* techniques.

Results and Discussion

Model Performance

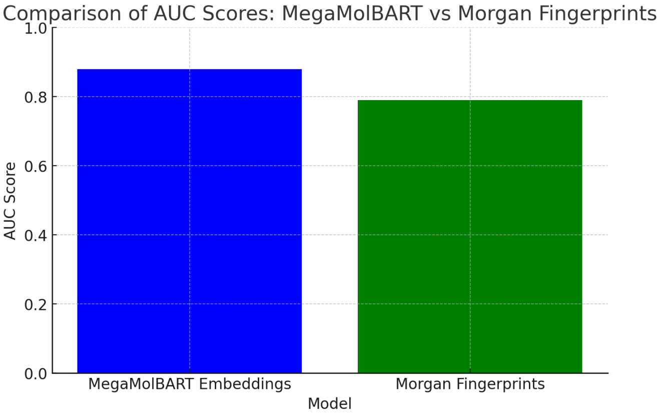

Figure 1: Comparison of AUC scores between MegaMolBART and Morgan fingerprints-based models.

Key results from the study:

- **AUC Score**:The MegaMolBART embeddings and the XGBoost classifier achieved an AUC score of 0.88 on the test dataset.
- **Improved Performance**: This performance was quite a significant improvement over traditional methods using Morgan fingerprints.
- **Complex Feature Extraction**: The transformer-based approach was able to capture complex molecular features better than conventional methods.

Experimental Validation

The results shown by the authors from *in vitro* experiments thus confirm that the model's predictions align quite well with the real world data. Blood Brain Barrier permeable compounds successfully penetrate the spheroids, while impermeable ones do not do so.[8]

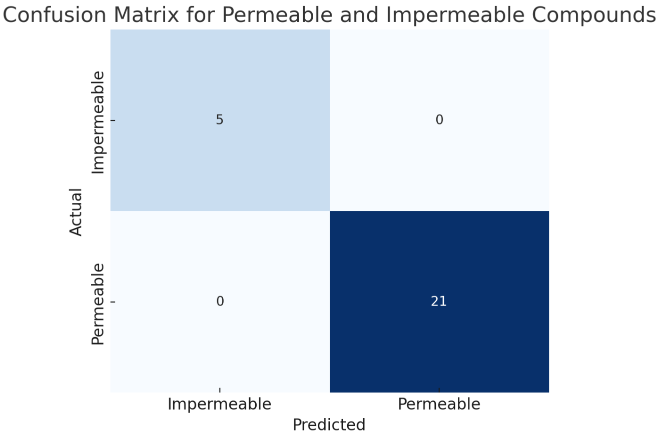

Figure 2: Confusion matrix for permeable and impermeable compounds.

Limitations

Even after showing promising results, the study faced a few limitations:

- **Overfitting**: Overfitting was present because of the relatively small size of the dataset.
- **Generalization**: The model finds it difficult to generalize to relatively new or unseen data.
- **Future Directions**: The authors propose using more diverse, varied and relatively larger datasets in future work to improve generalization.

Key Takeaways

Significance

This paper tries to show the advantages of using transformer-based models like MegaMolBART for predicting BBB permeability. The model's ability to automatically learn complex molecular patterns from SMILES representations makes it completely unnecessary for the need for extensive manual feature engineering, which is typically required in traditional methods.

Real-world Application

This approach taken by the authors is particularly more efficient to screen compounds for BBB permeability thereby making it particularly valuable in the CNS drug discovery. The model can predict the permeability early and hence can help pharma companies put focus on compounds with higher chances of clinical success.

Conclusion

The authors have shown that MegaMolBART and models based on transformers have shown great potential in future for drug discovery, especially in predicting BBB (Blood Brain Barrier) permeability. The validation that happens through *in vitro* experiments does highlight the real world utility it has thus providing as an efficient and accurate tool for screening CNS drug candidates.

Bibliography

[1] L. M. Ailioaie, C. Ailioaie, and G. Litscher. Photobiomodulation in alzheimer's disease: A complementary method to state-of-the-art pharmaceutical formulations and nanomedicine? *Pharmaceutics*, 15(3):916, 2023.

[2] H. Chen et al. In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms. *Journal of Molecular Graphics and Modeling*, 29(7):985–995, 2011.

[3] A. S. Guntner, T. Bogl, F. Mlynek, and W. Buchberger. Large-scale evaluation of collision cross sections to investigate blood-brain barrier permeation of drugs. *Pharmaceutics*, 13(12):2141, 2021.

[4] W. J. Harris et al. In vivo methods for imaging blood-brain barrier function and dysfunction. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(4):1051–1083, 2023.

[5] F. Meng, Y. Xi, J. Huang, and P. W. Ayers. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*, 8(1):289, 2021.

[6] F. Montanari and G. F. Ecker. Prediction of drug-abc-transporter interaction: Recent advances and future challenges. *Advanced Drug Delivery Reviews*, 86(1):17–26, 2015.

[7] D. Roy, V. K. Hinge, and A. Kovalenko. To pass or not to pass: Predicting the blood-brain barrier permeability with the 3d-rism-kh molecular solvation theory. *ACS Omega*, 4(7):16774–16780, 2019.

[8] Y. H. Zhao et al. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *Journal of Chemical Information and Modeling*, 47(1):170–175, 2007.