# 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

Date: 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 up till 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. They 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, which is time-consuming and computationally intensive, especially in the 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 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 and state-of-the-art gradient boosting techniques. The main steps 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: Morgan fingerprints, a traditional encoding method, were used with the same XGBoost classifier to provide a baseline for comparison.

• Evaluation Metrics: The performance of the models was assessed using AUC, accuracy, and other metrics, followed by experimental validation using in vitro techniques.

## Results and Discussion

### Model Performance

• 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 a significant improvement over traditional methods using Morgan fingerprints.  
• Complex Feature Extraction: The transformer-based approach captured complex molecular features better than conventional methods.

### Experimental Validation

The in vitro experiments confirmed that the model's predictions align well with real-world data. Blood Brain Barrier permeable compounds successfully penetrated the spheroids, while impermeable ones did not.

### Limitations

Despite promising results, the study had a few limitations:

• Overfitting: Overfitting was present due to the relatively small dataset size.

• Generalization: The model had difficulty generalizing to new or unseen data.

• Future Directions: The authors propose using more diverse, varied, and larger datasets in future work to improve generalization.

## Key Takeaways

### Significance

This paper demonstrates 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 eliminates the need for extensive manual feature engineering.

### Real-world Application

The approach taken by the authors is more efficient for screening compounds for BBB permeability, making it valuable in CNS drug discovery. The model can predict permeability early, helping pharmaceutical companies focus on compounds with higher chances of clinical success.

## Conclusion

The authors have shown that MegaMolBART and transformer-based models have great potential for drug discovery, especially in predicting BBB permeability. The in vitro experiments validate the real-world utility, providing an efficient and accurate tool for screening CNS drug candidates.