Utilizing a GNN machine learning model to predict anti-tuberculosis compounds

Antimicrobial Chemotherapy Virtual Conference 2026

Sohail Mahmood[1] & Zidane Virani[2], [1] Living In Silico Inc. [2] Western University

INTRODUCTION

- **WHO statistics:** 1.25 million TB deaths worldwide in 2023
- Problem statement: Multidrug-resistant (MDR) and Extensively drug-resistant (XDR) tuberculosis remain major global health threats, necessitating novel therapeutics
- Need for AI/ML: Traditional drug discovery is slow and costly; Artificial Intelligence and Machine Learning methods offer accelerated screening and lead identification

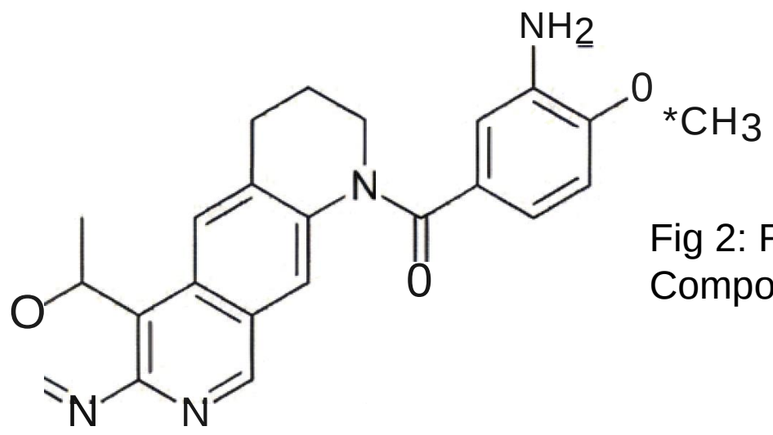


Fig 2: Representative Tuberculosis Compound (e.g., Bedaquiline)

METHODS

- Dataset: PubChem AID 588726 (359,483 compounds), including active and inactive compounds
- **Molecular representation:** Atoms as nodes with 7 features (atomic number, degree, charge, hybridization, aromaticity, hydrogens, ring membership)
- GCN Architecture: 3 convolutional layers (128 hidden channels) → Global mean pooling → 3-layer classifier (45,185 parameters)
- Training details: 50 eposhs, Adam optimizer, learning rate 0.001, 70-10-20 data split (train-validation-test)

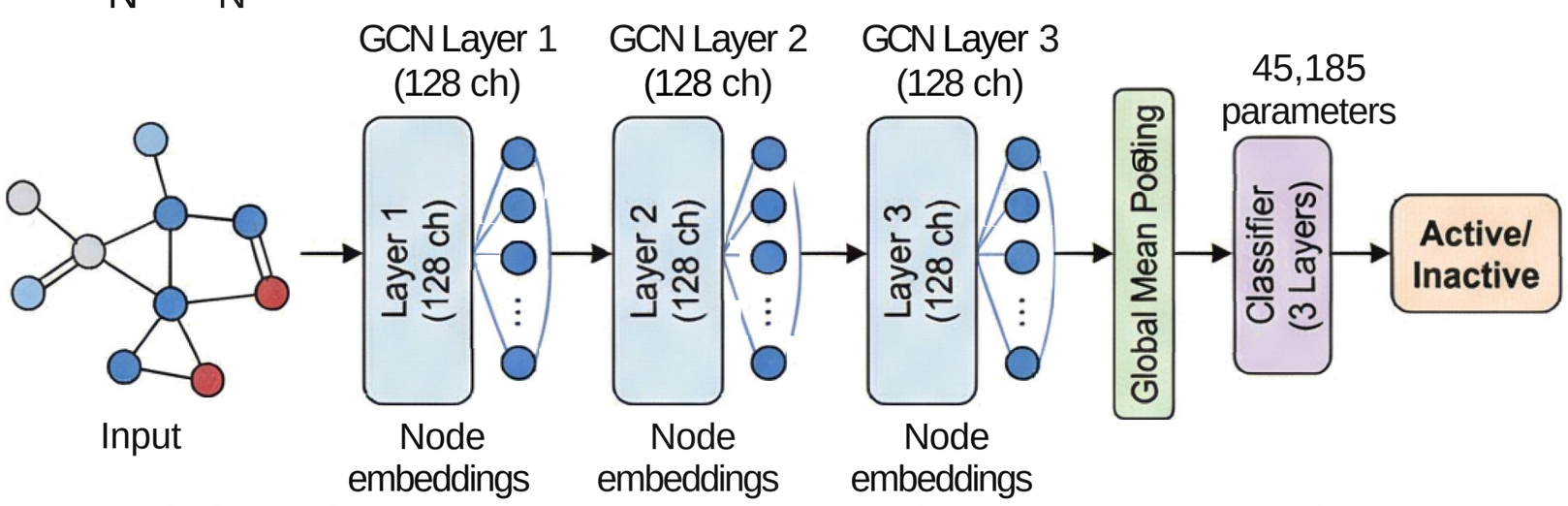


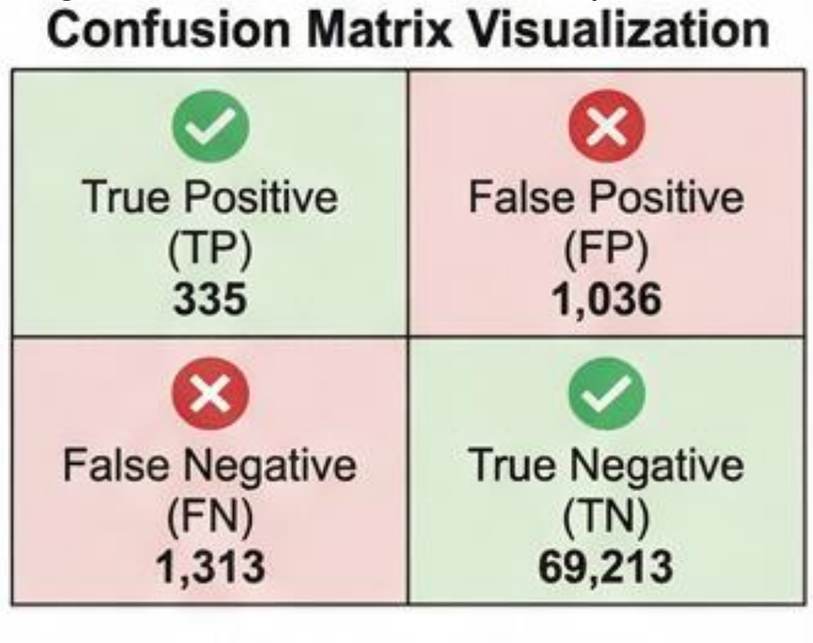
Fig 3: Graph Convolutional Network architecture for TB compound classification

RESULTS

Table 1: Model performance metrics

Metric	Value
AUC-ROC	0.74
Accuracy	96.73%
Specificity	98.53%
Recall	20.33%
Precision	24.43%

Fig. 1: Confusion matrix of GCN predictions



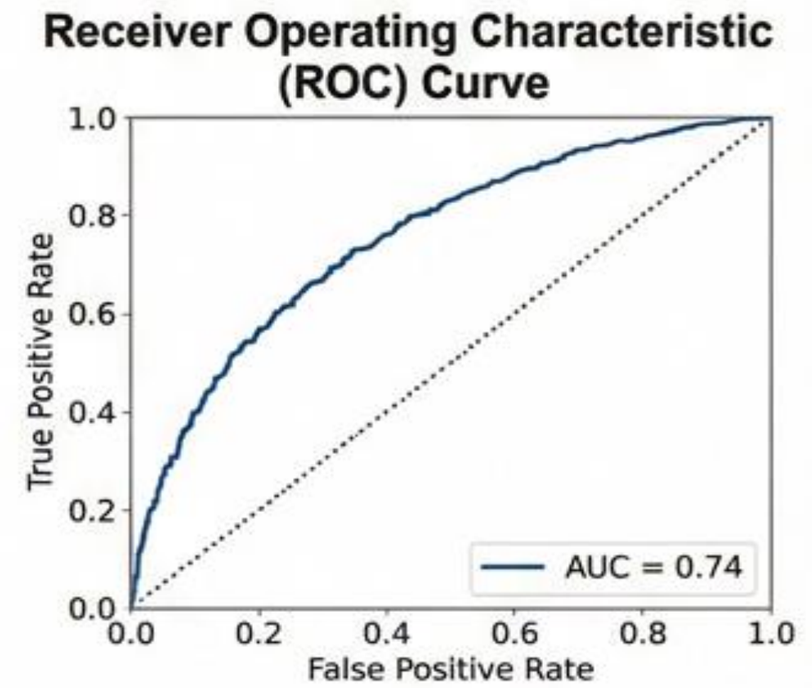


Fig. 4: ROC curve showing model performance

CONCLUSION

- Moderate discrimination ability (AUC=0.74) demonstrates potential for TB drug discovery
- High specificity (98.53%) useful for pre-screening large chemical libraries, reducing experimental costs
- Low recall (20.33%) limits standalone application; many active compounds are missed
- Class imbalance (1:41 ratio of actives to inactives) contributed to model bias
- Future work: Explore weighted loss functions, advanced sampling techniques, and incorporate enzymatic pathway data