Unlocking the Brain's Secrets: Trustworthy AI Models for Assessing Blood-Brain Barrier Permeability

Isakova A.M.[1], Shkil D.O.[2], Steshin I.S.[2] , Skorb E.V.[1], Fedorov M.V.[3] , Shityakov S.V.[1]

[1] ITMO University, [2] Syntelly LLC, [3] Kharkevich Institute for Information Transmission Problems of RAS

Abstract

The permeability of the blood-brain barrier (BBB) is crucial for evaluating drug candidates, traditionally measured through laborious in vitro and in vivo techniques. This study introduces a robust BBB classification ML model , achieving a peak **ROC-AUC$_{cv}$ of 0.963** and a superior regression model with **R^2 = 0.954, Q^2 = 0.728, RMSE$_{cv}$ = 0.321**. We have introduced an innovative idea **to incorporate classification labels as features in regression analysis**.

Keywords: blood-brain barrier, logBB, SMILES, machine learning, deep learning.

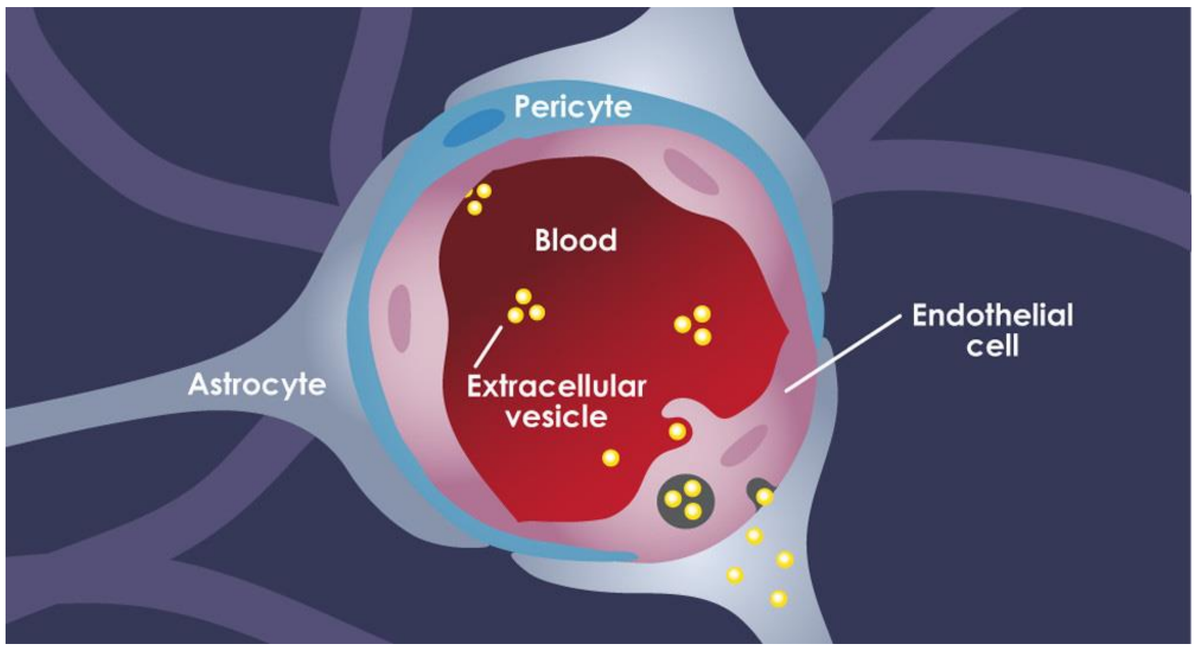

Figure 1. Human BBB representation

Introduction

The blood-brain barrier (BBB) serves as a selectively permeable interface that safeguards the central nervous system from detrimental substances (Figure 1). The capacity to traverse the BBB is a crucial factor in pharmacology, particularly concerning interventions for neurodegenerative disorders [1]. However, assessing permeability via experimental methodologies poses significant challenges, which has led to an increasing interest in the utilization of machine learning methods. At present, **all advanced machine learning models** developed for this purpose are either **not publicly accessible**, thus hindering the reproducibility of findings, or **have been trained on relatively small datasets**, subsequently compromising the reliability of these models [2].

Methods

We opted for **SMILES** notation to represent molecules due to its human-friendly nature and widespread use in such tasks. The training dataset was acquired from B3DB [3] and supplemented with molecular data from Tevosyan et al. [4], subsequently undergoing standard preprocessing procedures such as canonicalization, elimination of inorganic compounds, removal of duplicates and omissions. A variety of machine learning algorithms **(CatBoost, XGBoost, LGBM, RF, ANN)** and molecular fingerprints **(Avalon, MACCS keys, CATS2D, PubChem, ECFP6)**, in addition to **physico-chemical descriptors from RDKit**, were investigated, which included the formulation of bespoke fragment-based molecular descriptors **(Fragments) specifically designed for our task** and the integration of classification labels as supplementary features for regression analysis.

Results

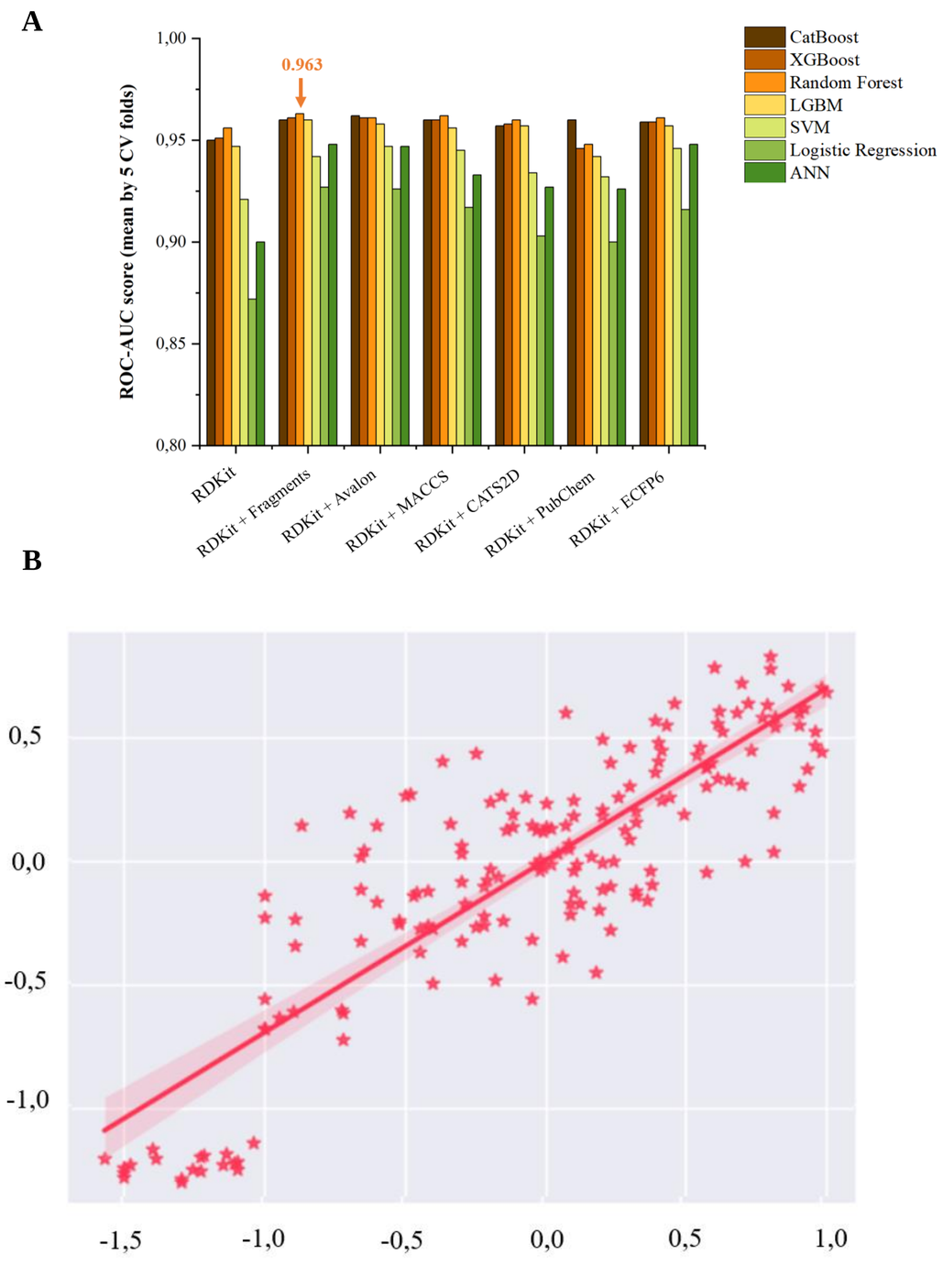

Figure 2. Average ROC-AUC of class labels predictions obtained by different classification algorithms and descriptor sets (**A**); relationship between the experimentally measured logBB and the predictions of the best regression model on a validation sample (**B**)

The **highest-performing** classification model attained a cross-validation **ROC-AUC score of 0.963** (Figure 2A), employing the Random Forest model with Fragments fingerprints as the input features. The regression model, which was developed on a dataset comprising **904 molecules** (this is **double the size of previously documented models**, another 226 molecules were left for the validation), demonstrated enhanced efficacy when utilizing the LGBM model with PubChem descriptors and class label as input features. The resultant **cross-validation RMSE** was recorded at **0.321** (Figure 2B). On a validation sample RMSE did not decrease dramatically: from 0.321 to 0.340.

Conclusion

Machine learning models that are openly available have been constructed for tasks related to classification and regression. It was found that fragments constituted the most effective input dataset for the entirety of regression algorithms and for a significant portion of classification algorithms. The inclusion of the classification label as a feature could substantially improve the precision of regression models. **All our models are available via the QR-code.**

References

1. Demystifying brain penetration in central nervous system drug discovery. In Journal of Medicinal Chemistry (Vol. 56, Issue 1). https://doi.org/10.1021/jm301297f
2. Ciura, K., Ulenberg, S., Kapica, H., Kawczak, P., Belka, M., & Bączek, T. (2020). Assessment of blood–brain barrier permeability using micellar electrokinetic chromatography and P_VSA-like descriptors. *Microchemical Journal*, *158*. https://doi.org/10.1016/j.microc.2020.105236
3. Meng, F., Xi, Y., Huang, J., & Ayers, P. W. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*, *8*(1). https://doi.org/10.1038/s41597-021-01069-5
4. Tevosyan, A., Khondkaryan, L., Khachatrian, H., Tadevosyan, G., Apresyan, L., Babayan, N., Stopper, H., & Navoyan, Z. (2022). Improving VAE based molecular representations for compound property prediction. Journal of Cheminformatics, 14(1). https://doi.org/10.1186/s13321-022-00648-x.