Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BioGCN: Biologically Inspired Graph Convolutional Network to Predict Human Oral Bioavailability

Koyel Mandal
Indian Statistical Institute
Aritra Banerjee
Indian Statistical Institute
Anushree Bhattacharya
Indian Statistical Institute
Sanghamitra Bandyopadhyay
sanghami@isical.ac.in

Indian Statistical Institute

Article

Keywords:

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8668254/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BioGCN: Biologically Inspired Graph Convolutional Network to Predict Human Oral Bioavailability

Koyel Mandal[1,2], Aritra Banerjee[1], Anushree Bhattacharya[1], and Sanghamitra Bandyopadhyay[1,*]

[1]Indian Statistical Institute, Machine Intelligence Unit, Kolkata, 700108, West Bengal, India
[1]Departmnet of Computer Science and Engineering, SRM University-AP, Amaravati 522 240, Andhra Pradesh, India
[*]sanghami@isical.ac.in

ABSTRACT

Human Oral Bioavailability (HOB) is a critical issue in drug discovery. Accurate computational methods to predict HOB in the early stages of drug discovery significantly reduce costs and improve efficiency. The majority of existing methods rely on handcrafted molecular descriptors that require domain expertise and time-consuming feature selection. However, Graph Convolutional Network (GCN) can learn feature representation directly from molecular graphs. We have developed a method to predict HOB using a Biologically Inspired Graph Convolutional Network (BioGCN), incorporating the ATC code as domain knowledge to enhance prediction accuracy. The study uses a total of 1,480 drugs for training and 288 and 45 drugs for testing. The evaluation of BioGCN and traditional machine learning algorithms reveals significant performance improvements when ATC codes are incorporated. BioGCN achieves an average accuracy of 65.83%-76.5% on the training dataset and attains the best accuracy of 53.82%-77.78% on the test dataset across various scenarios. The result shows a consistent performance gain as compared to other machine learning algorithms as well as previously published HOB Models.

Introduction

Drug discovery is considered a complex, lengthy process spanning from the preclinical phase to final approval, typically taking 12-15 years and costing around 2.5 billion[1]. The statistic highlights that only 1 out of 5,000–10,000 discovered compounds reaches the U.S. Food and Drug Administration (FDA) review phase[2]. Regulatory agencies require rigorous testing phases to ensure the safety and effectiveness of new drug candidates before they are approved for clinical use[3]. The drug evaluation process contributes to the high attrition rate, in which 80-90% of projects are discontinued before entering human trials, and nearly 95% of drugs that enter human trials fail[4,5]. One of the major causes of the high attrition rate is poor ADMET properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity)[6]. Therefore, to reduce attrition rates in the later phases of drug research, it is important to understand a drug's pharmacokinetic properties, such as Human Oral Bioavailability (HOB).

HOB defines the fraction of an oral drug that reaches systemic circulation after ingestion[7]. Bioavailability (F) is measured by comparing the dose-corrected area under the curve (AUC) after oral administration to that after intravenous (IV) administration, as shown in Equation 1. Here, AUC_{iv} represents the systemic exposure measured after an IV dose D_{iv}, whereas AUC_{po} is the systemic exposure measured after an oral dose D_{po}.

$$F = \frac{AUC_{po}}{AUC_{iv}} \cdot \frac{D_{iv}}{D_{po}} \tag{1}$$

If administered intravenously, the drug can be delivered to the site where it exerts its pharmacological and therapeutic effects via the bloodstream. The HOB ranges from 0 to 100%. Many orally administered drugs undergo metabolism and elimination before they reach their target site in an active form. Therefore, it can be noted that a candidate drug with poor human oral bioavailability (around 50%) may fail to progress to Phase II or III clinical trials, even if it shows high efficacy in early in vitro or in vivo evaluations[8,9]. On the other hand, a higher HOB of the drug can reduce the required dosage to achieve the desired pharmacological effect, thereby minimizing side effects and toxicity risks. Hence, the level of HOB is the key factor in determining the success or failure of new drug development.

Laboratory experiments in new drug discovery can be both expensive and time-consuming. The total expenditure to bring a new drug to market can reach $2.6 billion USD, and this cost is rising exponentially over time. Hence, early prediction of the HOB of a drug candidate is crucial for making informed "Go/No Go" decisions in the drug discovery process[10]. Using computational methods to predict the $\%F$ of new drug candidates prior to synthesis or advancement into preclinical and clinical development can be significantly advantageous as it saves resources and offers a promising alternative to traditional

experimental approaches[8]. The accurate prediction of HOB is an extremely challenging task because HOB is influenced by a number of biological, chemical, and physical parameters, including the permeability of the intestinal membrane, the solubility of compounds in the gastrointestinal system, and first-pass metabolism[9]. Despite these complexities, there is still growing research interest in developing sophisticated computational methods, such as machine learning, to predict HOB based on molecular properties of compounds. These approaches simplify the drug discovery process, that reduces attrition rates, and accelerate the development of effective medications. This helps to more efficient use of resources, faster time-to-market for new drugs, and eventually, improved patient outcomes.

Numerous studies have been explored various approaches in order to predict HOB, starting from traditional Quantitative Structure-Activity Relationship (QSAR)[11,12] models to advanced ML techniques[1,13,14]. Yoshida and Topliss[15] developed a quantitative structure-bioavailability relationship model for 232 structurally diverse drugs, using physicochemical parameters such as lipophilicity, charge state, and structural features associated with metabolic reactivity. The model achieved 67% accuracy in leave-one-out cross-validation and 60% on a separate test set. Turner et al.[16] proposed a stepwise regression model using eight molecular descriptors based on a dataset of 167 compounds. Wang et al.[17] further expanded the scope by developing a correlation model for 577 compounds using a total of 50 molecular descriptors.

With the exponential power of the computational methods, machine learning (ML) algorithms dominate the field. Ma et al.[18] developed a model for human plasma protein binding rate and bioavailability using Support Vector Machines (SVM) in association with a genetic algorithm (GA). This hybrid approach resulted in 86% accuracy on the test set, demonstrating the potential of combining ML with optimization techniques. A notable advancement proposed by Kumar et al.[19] is an SVM model trained on 511 drugs. This model reported exceptional performance metrics, including a Matthews correlation coefficient, ROC AUC, and accuracy of 0.909, 0.943, and 96.1%, respectively. While these results significantly outperform state-of-the-art methods, they have also raised questions about potential overfitting or dataset biases.

HOB prediction has continued to evolve through various ensemble methods. In 2022, Wei et al.[9] introduced HOBPre, a consensus model using Random Forest (RF) for 1588 molecules. This larger dataset and ensemble approach aim to improve generalizability and robustness. Yang et al.[13] explored the impact of ADME properties on oral bioavailability and developed prediction models using various algorithms, including RF, XGBoost, CatBoost, and LightGBM, trained on Morgan fingerprints as molecular descriptors, where RF showed the best performance.

Recently, this work has been shifted towards deep learning and graph-based methods. In 2023, Ng and Lu[20] leveraged Graph Neural Networks (GNNs) for HOB prediction, eliminating the need for manual feature selection. This work leverages transfer learning from a pre-trained solubility model and achieves an average accuracy of 79.7%, demonstrating the potential of deep learning in this domain. Myung et al.[1] introduced Deep-PK, an advanced web-based deep learning platform that leverages GNNs and graph-based signatures to predict pharmacokinetics and toxicity. The study[21] introduced an improved deep forest model, known as bgmc-forest, to predict HOB by enhancing feature extraction and adopting balanced multi-grained scanning to address limitations in small-sample datasets. Several recent works have advanced general GNN theory[22–24], but no study has addressed HOB prediction.

Despite significant advancements in ML and deep learning (DL) in the field of HOB, it still faces significant challenges. Traditional approaches, likeQSAR, rely heavily on handcrafted molecular descriptors and fingerprints, requiring extensive domain expertise and time for feature selection. Although ML techniques such as SVM, RF, and ensemble methods have shown promising results and improved prediction accuracy, they still rely on predefined features. They also may struggle with the complexity of capturing the intricate relationships between chemical structures and pharmacokinetic properties. Recent approaches use Graph Neural Networks (GNNs), particularly Graph Convolutional Networks (GCNs), which provide and automatically learning feature representations from molecular graphs. Studies using GCNs have shown competitive performance without the need for manual feature engineering. However, existing GCN-based models often focus solely on structural information and fail to incorporate high-level domain knowledge, such as pharmacological or therapeutic context. This knowledge can provide additional insights to improve predictions. In particular, most GCN studies for HOB prediction have not fully explored integrating drug classification systems, such as the Anatomical Therapeutic Chemical (ATC) classification. It categorizes drugs by therapeutic use and mechanism of action. Moreover, ATC codes encapsulate valuable pharmacological information that could enhance the model's understanding of how structural and therapeutic characteristics jointly influence bioavailability.

The major contribution of this paper is the introduction of a biologically inspired graph convolutional network (BioGCN) that incorporates the ATC code as domain knowledge to enhance prediction accuracy.The Anatomical Therapeutic Chemical (ATC) classification system, developed by the World Health Organization, is a widely adopted scheme for classifying drugs and is commonly used in drug screening, repositioning, and similarity studies[25]. By understanding this classification, researchers can group drugs with similar ADMET profiles, thereby enhancing the HOB prediction model. The underlying assumption is that drugs within the same ATC code often exhibit similar pharmacokinetic properties, making ATC codes a valuable feature for bioavailability models. The evaluation analyzes the proposed GCN architecture and demonstrates significant performance

improvements when ATC codes are integrated into the graph learning framework compared to traditional approaches that rely solely on structural information.

Results

The experimental results of the prediction on HOB are performed in the Python programming language. First, the training dataset of 1480 drugs is assembled, along with the corresponding FM and CFM. Both matrices are separately preprocessed: missing values are imputed with each feature's average, and feature values are normalized using min-max scaling.

To compare the proposed method, BioGCN, ten different models are considered, including Random Forest (RF)[26], Support Vector Machine (SVM)[27], K-Nearest Neighbors (KNN)[28], Decision Tree classifier (DT), Logistic Regression (LR)[29], Gradient Boosting (GB)[30], Naive Bayes (NV)[31], Adapting Boosting (AdaBoost)[32], Extreme Gradient Boosting (EGB)[33], and Categorical Boosting (CatBoost)[34]. Scikit-learn[35], XGBoost, and CatBoost Python packages are used for different prediction models. Each SMILES notation is processed and calculates nodes and edge features using the RDKit package. For BioGCN, it is not possible to create a graph with a single atom, i.e., *Selenium*. Therefore, this drug is removed from the training process for the BioGCN model. The BioGCN model is constructed using PyTorch[36] and PyTorch Geometric[37].

Classification algorithms are sensitive to the choice of hyperparameters. Therefore, it is important to properly set the parameters of different algorithms. The parameter settings for all models using FM and CFM on the training dataset are optimized using five-fold cross-validation and grid search, with the optimal parameters chosen based on accuracy, as shown in Equation 8. The parameters searched and their explored ranges using FM and CFM are reported in the supplementary files, Tables S4 and S5, respectively. BioGCN uses a maximum of 100 epochs. The number of tunable hyperparameters has been limited to mitigate the risk of overfitting and underfitting. This simplifies the model and reduces the likelihood of learning noise or irrelevant patterns from the training dataset.

The performance of various models is evaluated using six different metrics, as mentioned in Section , using five-fold cross-validation, repeated 10 times. For BioGCN, embeddings are generated 10 times, each followed by 5-fold cross-validation, which is also repeated 10 times. The dataset has a severe class imbalance. Therefore, to overcome class imbalance, several sampling techniques are utilized, including undersampling, oversampling, and SMOTE. Table 1 summarizes the average metrics across cross-validation for all models.

Moreover, the effectiveness of various classification algorithms is compared using two separate test datasets comprising 288 and 45 drugs, respectively. For the testing dataset, BioGCN generates 10 embeddings for each feature matrix; each embedding predicts class levels 10 times. Other machine learning algorithms undergo the same procedure, which is repeated 10 times. The results are evaluated using six metrics and report the best results for CFM and FM in Tables 2, based on accuracy and balanced accuracy.

Comparison of FM and CFM

To compare the performance of two molecular feature sets, FM and CFM, cross-validation results on the training set are analyzed. For each evaluation metric, the prediction models are ranked by performance, with the highest-performing feature sets receiving the highest ranks. For example, if the accuracy of FM for a model is 0.7747 and that of CFM is 0.7845, CFM is ranked 1, and FM is ranked 2. These ranks are aggregated across all models and metrics. This ranking process is repeated for each sampling technique, and the final mean rank is calculated, as shown in Table 3. The table shows that CFM outperforms (similar performance) FM across almost all metrics, except recall, indicating that CFM is generally more effective than FM at HOB prediction. This study also uses testing data, which shows that the CFM performs better than FM. Further analysis is based on the CFM feature set for both training and testing data.

Comparison with different machine learning algorithms

BioGCN achieves average accuracies of 65.83%-76.5% on the training dataset and the best accuracies of 53.82%-77.78% on the test dataset across various experimental setups. This moderate increase in accuracy from training to test suggests that BioGCN generalizes effectively to unseen data and is relatively robust. This actually indicates that the model successfully captures relevant patterns for the HOB prediction without overfitting. For ease of comparison across algorithms, the same ranking methodology is followed. Specifically, the rank of each algorithm is determined by individual evaluation metrics. Subsequently, an aggregated rank for each model is computed across all metrics for every sampling technique. The average ranking results for training and testing are depicted in Figure 1.

Figure 1 illustrates that for the training data, CatBoost performs best when no sampling technique is applied. The results vary from one smapling technique to another. Random Forest achieved the best overall rank under undersampling, CatBoost consistently performed best under both the original and oversampled datasets, while Extra Gradient Boosting (EGB) attained the highest overall ranking with SMOTE.

	Algorithm	Accuracy		AUC		BAcc		Recall		Precision		F1 score	
		FM	CFM	FM	CFM	FM	CFM	FM	CFM	FM	CFM	FM	CFM
None	RF	0.7803	0.7816	0.7492	**0.7633**	0.607	0.5981	0.9481	0.9594	0.7984	0.7941	0.8666	0.8687
	SVM	0.7771	0.761	0.7179	0.7181	0.604	0.5971	0.9464	0.9203	0.7968	0.7954	0.8648	0.8529
	KNN	0.77	0.767	0.7048	0.6961	0.558	0.5515	0.9757	0.9755	0.7767	0.7742	0.8647	0.8631
	DT	0.7245	0.7413	0.6422	0.6146	0.6001	0.6041	0.8444	0.8752	0.8012	0.8007	0.8217	0.8358
	LR	0.7661	0.7751	0.7026	0.7347	0.5384	0.6178	**0.9866**	0.9281	0.7688	0.804	0.8639	0.8613
	GB	0.776	0.7789	0.7268	0.7254	0.6223	0.5988	0.9252	0.9538	0.8065	0.7945	0.8614	0.8666
	NB	0.7652	0.5293	0.6383	0.6008	0.5588	0.6008	0.9636	0.4602	0.7779	**0.8451**	0.8606	0.5953
	AdaBoost	0.7727	0.7736	0.695	0.7104	0.5721	0.5895	0.9675	0.9525	0.7826	0.7905	0.865	0.8637
	EGB	0.778	0.7796	0.751	0.7509	0.6307	**0.6372**	0.9208	0.9183	0.8107	0.8137	0.862	0.8625
	CatBoost	0.783	**0.7832**	0.7532	0.7486	0.6012	0.6134	0.9596	0.9476	0.7952	0.8012	**0.8694**	0.8681
	BioGCN	0.7725	0.765	0.7284	0.7212	0.5745	0.6307	0.9644	0.8959	0.7841	0.8123	0.8644	0.8516
Undersampling	RF	0.6879	**0.7062**	0.7416	**0.7543**	0.6778	**0.6873**	0.6967	0.7254	0.863	0.8637	0.7705	0.7877
	SVM	0.6791	0.6687	0.708	0.7221	0.6564	0.662	0.7008	0.675	0.8471	0.8557	0.7665	0.7538
	KNN	0.6774	0.6818	0.6992	0.6924	0.6412	0.6343	0.7126	0.729	0.8354	0.8301	0.7685	0.774
	DT	0.5925	0.5984	0.6395	0.6246	0.6016	0.5952	0.5828	0.6002	0.8255	0.8185	0.6821	0.6918
	LR	0.6863	0.665	0.6998	0.7231	0.6441	0.6624	0.7271	0.6676	0.8356	0.8569	0.7773	0.75
	GB	0.669	0.6744	0.7235	0.7232	0.6602	0.6642	0.6766	0.6836	0.8541	0.8553	0.7544	0.7592
	NB	0.7654	0.566	0.6342	0.6237	0.562	0.5839	**0.9617**	0.5489	0.7791	0.8263	**0.8605**	0.6311
	AdaBoost	0.6913	0.6729	0.6998	0.7044	0.6403	0.6437	0.7408	0.7014	0.8317	0.839	0.7829	0.7634
	EGB	0.6857	0.6775	0.7446	0.7317	0.6868	0.6731	0.685	0.6813	**0.8709**	0.8622	0.7662	0.7607
	CatBoost	0.6949	0.6846	0.7372	0.7371	0.6765	0.6748	0.7123	0.6931	0.8589	0.8613	0.7783	0.7676
	BioGCN	0.6625	0.6583	0.7122	0.7133	0.6524	0.658	0.6729	0.6585	0.8518	0.855	0.7447	0.7432
Oversampling	RF	0.7747	**0.7845**	0.7514	**0.7661**	0.6273	0.6349	0.918	**0.93**	0.8092	0.8118	0.8598	**0.8666**
	SVM	0.7194	0.748	0.7066	0.7283	0.651	0.6485	0.7856	0.8453	0.833	0.8252	0.8082	0.8347
	KNN	0.6725	0.6442	0.7025	0.6845	0.6481	0.6267	0.6977	0.6614	0.8418	0.8323	0.7622	0.7364
	DT	0.6624	0.6793	0.6189	0.6016	0.5938	0.5967	0.7295	0.7597	0.8052	0.8047	0.7647	0.7809
	LR	0.6926	0.7408	0.7055	0.7317	0.6526	0.6603	0.7316	0.8192	0.8403	0.8342	0.7818	0.8262
	GB	0.771	0.7401	0.7354	0.7303	0.6202	0.6564	0.9176	0.822	0.8058	0.8318	0.8578	0.8264
	NB	0.7634	0.5214	0.6449	0.5917	0.5577	0.5917	0.9621	0.4533	0.7772	0.8378	0.8596	0.5875
	AdaBoost	0.7085	0.7016	0.7043	0.7089	0.6445	0.6476	0.7708	0.7538	0.8305	0.835	0.7991	0.7918
	EGB	0.7774	0.7721	0.7543	0.7483	0.6522	0.6474	0.8996	0.8934	0.8223	0.8206	0.8589	0.8551
	CatBoost	0.7722	0.7714	0.7467	0.7538	0.6346	0.6631	0.906	0.8764	0.8135	0.8302	0.8569	0.8523
	BioGCN	0.6951	0.7326	0.7155	0.7091	**0.6628**	0.6456	0.7272	0.817	**0.8488**	0.8273	0.7796	0.8202
SMOTE	RF	0.7676	0.7731	0.7545	**0.7659**	0.6481	0.6516	0.8832	0.8915	0.8221	0.8229	0.8513	0.8553
	SVM	0.7143	0.6743	0.691	0.6943	0.6489	0.6393	0.7788	0.7094	0.8322	0.834	0.804	0.7661
	KNN	0.6155	0.5279	0.7002	0.6771	0.6444	0.6212	0.587	0.4377	0.8584	**0.8734**	0.6967	0.5822
	DT	0.6543	0.6786	0.6405	0.6206	0.6	0.6056	0.7073	0.7502	0.811	0.8101	0.7547	0.7783
	LR	0.6864	0.7129	0.704	0.7239	0.6512	0.6573	0.7213	0.767	0.8411	0.8388	0.7758	0.8007
	GB	0.7677	0.7466	0.7416	0.7337	0.6459	0.6499	0.8865	0.8409	0.8204	0.8262	0.8518	0.8332
	NB	0.7627	0.528	0.6519	0.5984	0.5575	0.5984	**0.9613**	0.459	0.7771	0.8434	**0.8591**	0.594
	AdaBoost	0.695	0.6995	0.6941	0.7058	0.6391	0.6431	0.749	0.7537	0.8301	0.8321	0.7871	0.7905
	EGB	0.7712	**0.7749**	0.7496	0.758	0.6598	**0.6675**	0.8796	0.8791	0.8282	0.832	0.8528	0.8546
	CatBoost	0.7641	0.7699	0.7492	0.756	0.6628	0.6548	0.8634	0.8825	0.831	0.8249	0.8464	0.8523
	BioGCN	0.6794	0.7305	0.7093	0.7063	0.6564	0.6402	0.7015	0.8177	0.8488	0.8238	0.7643	0.8202

Table 1. Prediction performance of training data (n=1480) using a five-fold cross-validation.

	Algorithm	Accuracy		AUC		BAcc		Recall		Precision		F1 score	
		Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2	Test1	Test2
None	RF	0.4653	0.6222	0.6737	0.6754	0.5966	0.6271	0.3302	0.6071	0.8765	0.7391	0.4797	0.6667
	SVM	0.7431	0.6667	0.7498	0.7542	0.6922	0.6628	0.7953	0.6786	0.8507	0.76	0.8221	0.717
	KNN	0.7535	0.6667	0.7392	0.6733	0.5589	0.5935	0.9535	**0.8929**	0.7707	0.6757	0.8524	**0.7692**
	DT	0.5347	0.4889	0.5068	0.4853	0.4893	0.4968	0.5814	0.4643	0.7396	0.619	0.651	0.5306
	LR	**0.7917**	**0.7556**	0.7817	0.792	0.7067	0.7111	0.8791	**0.8929**	0.8475	0.7576	**0.863**	0.8197
	GB	0.5625	0.5778	0.6863	0.6061	0.6617	0.6145	0.4605	0.4643	**0.9083**	0.7647	0.6111	0.5778
	NB	0.6389	0.6444	0.6858	0.7027	0.6858	0.7027	0.5907	0.4643	0.8881	**0.9286**	0.7095	0.619
	AdaBoost	0.4722	0.6	0.548	0.6922	0.547	0.5977	0.3953	0.6071	0.7944	0.7083	0.528	0.6538
	EGB	0.6424	0.6444	0.6181	0.5987	0.6021	0.5872	0.6837	0.8214	0.8077	0.6765	0.7406	0.7419
	CatBoost	0.4236	0.5556	0.6412	0.6681	0.5778	0.5504	0.2651	0.5714	0.8769	0.6667	0.4071	0.6154
	BioGCN	0.7118	0.7333	**0.7976**	**0.8088**	**0.721**	**0.7511**	0.7023	0.6786	0.8882	0.8636	0.7844	0.76
Undersampling	RF	0.5521	0.4	0.725	0.7447	0.6502	0.5179	0.4512	0.0357	0.8981	1	0.6006	0.069
	SVM	**0.7951**	**0.6444**	**0.802**	**0.7689**	**0.7135**	**0.6681**	**0.8791**	**0.5714**	0.8514	0.8	**0.865**	**0.6667**
	KNN	0.6875	0.6	0.736	0.7342	0.6867	0.6208	0.6884	0.5357	0.8655	0.75	0.7668	0.625
	DT	0.6667	0.5111	0.5983	0.5588	0.5641	0.584	0.7721	0.2857	0.7793	0.8	0.7757	0.4211
	LR	0.7396	0.6	0.7917	0.75	0.6899	0.6439	0.7907	0.4643	0.85	0.8125	0.8193	0.5909
	GB	0.5069	0.4444	0.7063	0.6334	0.6426	0.542	0.3674	0.1429	**0.9294**	0.8	0.5267	0.2424
	NB	0.6389	0.5333	0.6858	0.625	0.6858	0.625	0.5907	0.25	0.8881	1	0.7095	0.4
	AdaBoost	0.2743	0.4889	0.5682	0.6954	0.514	0.5893	0.0279	0.1786	1	1	0.0543	0.303
	EGB	0.6528	0.5556	0.6531	0.687	0.6408	0.6197	0.6651	0.3571	0.8363	**0.8333**	0.7409	0.5
	CatBoost	0.6424	0.4667	0.7302	0.6639	0.6836	0.5483	0.6	0.2143	0.8836	0.75	0.7147	0.3333
	BioGCN	0.5382	0.6222	0.7683	0.7647	0.6545	0.6502	0.4186	0.5357	0.9184	0.7895	0.5751	0.6383
Oversampling	RF	0.5521	0.6667	0.725	0.7069	0.6502	0.6513	0.4512	0.7143	0.8981	0.7407	0.6006	0.7273
	SVM	**0.7951**	0.7333	**0.802**	0.7983	**0.7135**	0.7164	**0.8791**	0.7857	0.8514	0.7857	**0.865**	0.7857
	KNN	0.6875	0.5556	0.736	0.6607	0.6867	0.5851	0.6884	0.4643	0.8655	0.7222	0.7668	0.5652
	DT	0.6667	0.6667	0.5983	0.6849	0.5641	0.6513	0.7721	0.7143	0.7793	0.7407	0.7757	0.7273
	LR	0.7396	0.7333	0.7917	0.7857	0.6899	0.7164	0.7907	0.7857	0.85	0.7857	0.8193	0.7857
	GB	0.5069	0.6444	0.7063	0.6975	0.6426	0.6796	0.3674	0.5357	**0.9294**	0.8333	0.5267	0.6522
	NB	0.6389	0.6444	0.6858	0.7027	0.6858	0.7027	0.5907	0.4643	0.8881	**0.9286**	0.7095	0.619
	AdaBoost	0.2743	0.5111	0.5682	0.6691	0.514	0.6071	0.0279	0.2143	1	1	0.0543	0.3529
	EGB	0.6528	0.5778	0.6531	0.6155	0.6408	0.5336	0.6651	0.7143	0.8363	0.6452	0.7409	0.678
	CatBoost	0.6424	0.6444	0.7302	0.6786	0.6836	0.6334	0.6	0.6786	0.8836	0.7308	0.7147	0.7037
	BioGCN	0.7535	**0.7778**	0.7805	**0.8067**	0.7127	**0.7405**	0.7953	**0.8929**	0.8636	0.7813	0.8281	**0.8333**
SMOTE	RF	0.5104	0.6889	0.727	0.7458	0.6359	0.7038	0.3814	0.6429	**0.9111**	0.8182	0.5377	0.72
	SVM	0.7361	0.6222	0.7494	0.6933	0.6152	0.6155	0.8605	0.6429	0.8009	0.72	0.8296	0.6792
	KNN	0.5868	0.4667	0.7026	0.6649	0.6554	0.5252	0.5163	0.2857	0.881	0.6667	0.651	0.4
	DT	0.4167	0.5778	0.5629	0.6901	0.5007	0.6145	0.3302	0.4643	0.7474	0.7647	0.4581	0.5778
	LR	0.7639	0.7111	0.7712	0.7626	**0.67**	0.7101	0.8605	0.7143	0.8296	0.8	0.8447	0.7547
	GB	0.4375	0.6667	0.641	0.6639	0.5916	0.6859	0.2791	0.6071	0.8955	0.8095	0.4255	0.6939
	NB	0.6389	0.6444	0.6858	0.7027	0.6858	0.7027	0.5907	0.4643	0.8881	**0.9286**	0.7095	0.619
	AdaBoost	0.2847	0.4889	0.5787	0.6387	0.4983	0.5777	0.0651	0.2143	0.7368	0.8571	0.1197	0.3429
	EGB	0.5799	0.5333	0.667	0.5336	0.6146	0.5441	0.5442	0.5	0.8357	0.6667	0.6592	0.5714
	CatBoost	0.5764	0.5333	0.7085	0.6218	0.6394	0.5788	0.5116	0.3929	0.8661	0.7333	0.6433	0.5116
	BioGCN	**0.7708**	**0.7778**	**0.7647**	**0.792**	0.661	**0.7521**	**0.8837**	**0.8571**	0.8225	0.8	**0.852**	**0.8276**

Table 2. Prediction performance on testing datasets.

	Accuracy		AUC		BAcc		Recall		Precision		F1 score	
	FM	CFM	FM	CFM	FM	CFM	FM	CFM	FM	CFM	FM	CFM
None	1.6364	1.3636	1.5455	1.4545	1.5455	1.4545	1.4545	1.5455	1.6364	1.3636	1.5455	1.4545
Undersampling	1.4545	1.5455	1.5455	1.4545	1.5455	1.4545	1.4545	1.5455	1.5455	1.4545	1.5455	1.4545
Oversampling	1.5455	1.4545	1.4545	1.5455	1.4545	1.5455	1.4545	1.5455	1.6364	1.3636	1.4545	1.5455
SMOTE	1.4545	1.5455	1.4545	1.5455	1.4545	1.5455	1.4545	1.5455	1.6364	1.3636	1.4545	1.5455
Average	1.5227	**1.4773**	1.5000	**1.5000**	1.5000	**1.5000**	**1.4545**	1.5455	1.6136	**1.3864**	1.5000	**1.5000**

Table 3. Comparison of six evaluation metric rankings of FM and CFM.

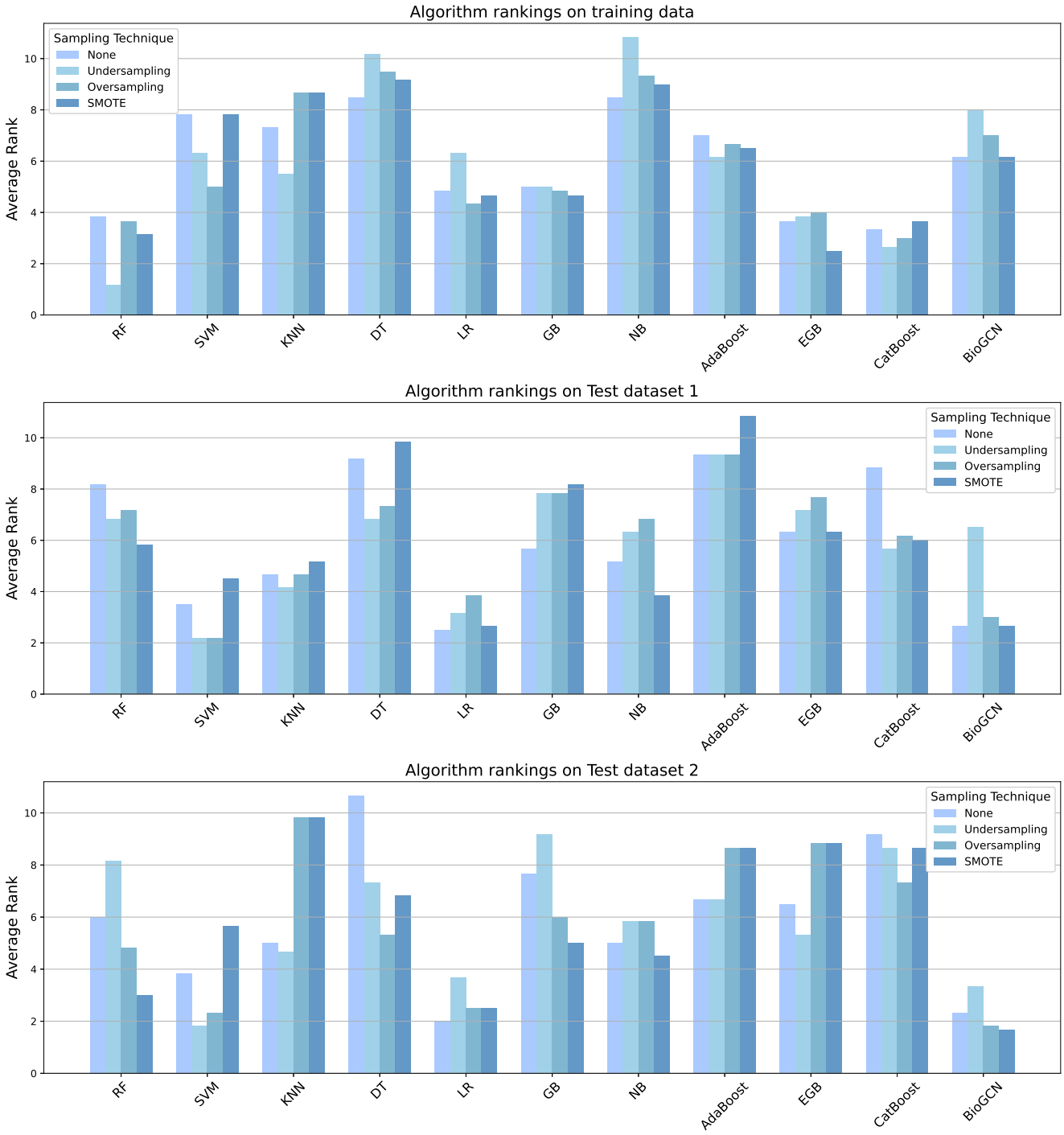


Figure 1. Comparison of average rankings of machine learning models across six evaluation metrics on both training and testing datasets.

The algorithms are arranged by performance for each sampling strategy as follows: CatBoost, RF, LR (without sampling); RF, CatBoost, EGB (undersampling); CatBoost, RF, EGB (oversampling); and EGB, RF, CatBoost (SMOTE). BioGCN exhibits moderate and relatively stable performance across different sampling strategies. It ranks sixth without sampling, eighth under undersampling and oversampling, and seventh with SMOTE according to the average rank across all evaluation metrics. In summary, these results suggest that no single classifier consistently outperformed all others across different sampling strategies.

Furthermore, Figure 1 presents the algorithm rankings based on the testing datasets 1 and 2. The figure indicates that BioGCN achieves the best overall performance with the SMOTE sampling strategy for both testing datasets and with the oversampling strategy for Testing Dataset 1. The LR and SVM algorithms are the best performers for no sampling and undersampling, respectively, across both datasets. It can be noted that CatBoost performs well on the training data but degrades on the test data. These results suggest that RF has overfit the training data, memorizing it rather than generalizing to test data.

Table 1 demonstrates a clearer overview of the performance of each model in four different scenarios. It is observed that the CatBoost model achieves an accuracy of 79.07%, followed closely by the EGB model with 77.96%, and the BioGCN model, which yields 76.5% without any sampling technique. Additionally, LR and NB exhibit the highest recall and precision, respectively, indicating their robustness in handling the HOB prediction task. Moreover, EGB can be considered as the best performer in terms of balanced accuracy.

This analysis help in understanding the impact of different sampling techniques in training data. Undersampling leads to lower performance than without sampling across most models and metrics, specifically in accuracy, AUC, recall, and F1 score. The decrease in both accuracy and recall suggests that undersampling is removing too much information from the majority class, resulting in poor performance and reducing the ability to identify positive classes. In contrast, oversampling generally improves performance compared to the result without any sampling, particularly in balanced accuracy and precision across all machine learning algorithms. Mixed results are observed in AUC: in some algorithms, it increases, and in others, it decreases. Oversampling performs better because it creates additional samples of minority classes. Similarly, applying SMOTE also improves performance, particularly in balanced accuracy and precision.

Table 2 demonstrates the insights into the generalization capability of our model. The BioGCN shows robust performance, achieving 77.78% accuracy with oversampling and SMOTE. More specifically, BioGCN also outperforms SMOTE in almost all evaluation metrics. In the case of undersampling, SVM achieves good accuracy, such as 79.51% and 64.44%, for test1 and test2, respectively. In contrast, LR and SVM exhibit superior performance without sampling and under the undersampling strategy, respectively. These findings suggest that the effectiveness of ML algorithms depends on the sampling technique used. In summary, we can say that BioGCN achieves better performance when averaging two test datasets and shows the best accuracy among all other methods.

Comparison with other HOB models

To further evaluate the performance of the proposed method, the accuracy of BioGCN on the test set 2 is compared with that reported in previously published HOB models (Table 4). Three recent studies—ADMETSAR 2.0[38], HobPre[9], and Deep-PK[1]—are included in this comparison due to the availability of their code or servers. The training sets for these state-of-the-art methods are sourced from the respective literature, which may use different datasets. For a fair comparison, a 20% cut-off is used. The comparison results, summarized in Table 4, show that BioGCN outperforms the other methods across accuracy, AUC, balanced accuracy (Bacc), recall, precision, and F1 scores. In this table, the bold numbers indicate the maximum value obtained for each corresponding evaluation metric.

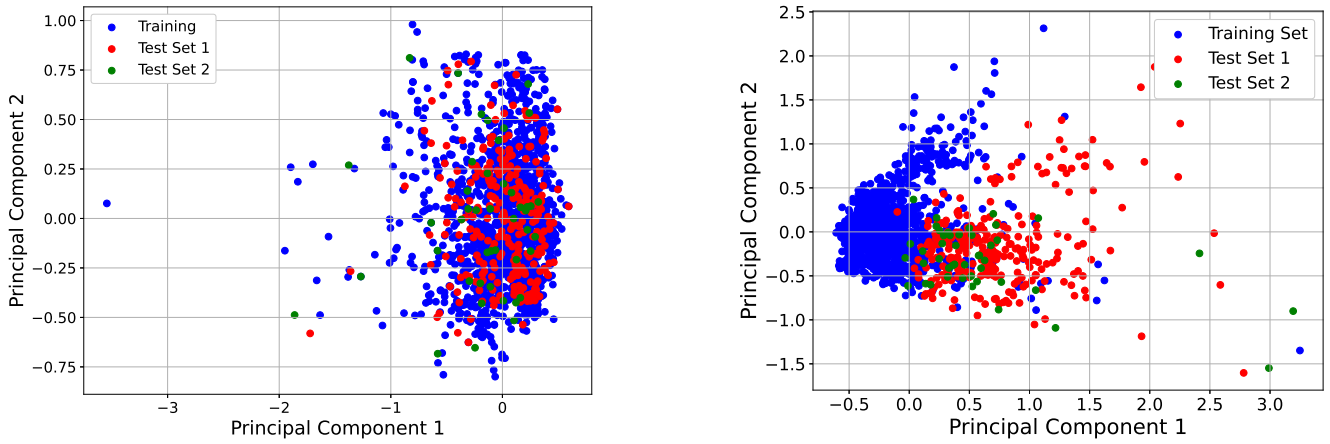
Method	Data set size	ACC	AUC	BAcc	Recall	Precision	F1 score
BioGCN	1480	**0.7778**	**0.7899**	**0.77**	**0.8929**	**0.8065**	**0.8475**
ADMETSAR 2.0 (RF)[38]	995	0.6818	0.6161	0.6161	0.8571	0.7059	0.7742
HobPre (RF)[9]	1558	0.7556	0.7689	0.7689	0.7143	0.7143	0.7843
Deep-PK (GNN)[1]	1005	0.6341	0.5762	0.5762	0.8400	0.6562	0.7368

Table 4. Prediction comparison with other state-of-the-art models on test data.

Diversity distribution of HOB dataset

Principal Component Analysis (PCA) is applied to the training and testing sets of drug molecules to examine their structural diversity, offering a potential means of evaluating the model's applicability domain. The two most important components are selected to construct a chemical space, using FM and CFM to characterize the distribution of the training and test sets, as depicted in Figure 2. The result suggests that the chemical space of the test set largely falls within that of the training set, supporting the use of the model trained on the training set to predict HOB values for the test set. PCA analysis is particularly

helpful for determining whether a model is suitable for predicting new molecules by evaluating their similarity to the training data. If a new molecule's projection falls within the training data's range on the PCA plot, it is considered "inside" the model's applicability domain, meaning its prediction is likely reliable.



(a) FM **(b)** CFM

Figure 2. The chemical space of the training and testing using PCA factorization.

Discussion

We propose a method for predicting HOB through a biologically inspired graph convolution network, namely BioGCN. The BioGCN is guided by biological knowledge, i.e., ATC, to improve HOB prediction. In this study, we have focused on two types of feature matrix: FM and CFM. First, a comparison of evaluation metrics using FM and CFM shows that CFM outperforms FM across all machine learning algorithms. CFM consistently outperforms FM for most models and metrics. This observation underscores the importance of integrating domain knowledge into feature engineering to enhance HOB prediction accuracy. The additional molecular descriptors included in CFM appear to capture important structural and chemical properties relevant to HOB prediction. The efficacy of BioGCN is also evaluated using a test dataset, where it outperforms other ML algorithms. Additionally, BioGCN achieves better predictive performance compared to previously published HOB models.

In conclusion, the comprehensive evaluation depicts the effectiveness of BioGCN in predicting HOB. BioGCN offers a unique advantage by capturing molecular features at the node and edge levels, facilitating the propagation of information across different layers. This model can provide valuable insights into HOB-related drug discovery and development processes. Future work will include refining feature engineering techniques, exploring additional deep learning architectures, and validating the HOB prediction models on larger, more diverse datasets.

Methods

Dataset Description

The human oral bioavailability (HOB) dataset, including bioavailability values, was collected from[9] and additional literature sources. Each drug in the dataset is associated with a corresponding SMILES (Simplified Molecular Input Line Entry System) representation. The training dataset consists of 1503 drugs. The performance of models predicting HOB is highly dependent on the criteria used to define bioavailability classes. Currently, there is no universally accepted threshold for bioavailability classification. In previous studies, four different thresholds have been used, namely 20%[9,18], 30%[39], 50%[40], and 80%[41]. This study defines the cutoff at 20%: drugs with HOB $< 20\%$ are low bioavailability, while those with HOB $\geq 20\%$ are high bioavailability. Some drugs from the training dataset cannot be classified as low- or high-bioavailability categories due to inaccurate experimental values, resulting in a final set of 1480 drugs. In the training dataset, there are 1116 molecules classified as high-bioavailability and 365 as low-bioavailability. The distribution of all training drugs with HOB values is displayed in Figure 3.

Additionally, there are three separate test datasets containing 288 (test 1), 140 (test 2), and 45 (test 3) drugs, respectively. Information about the datasets is summarized in Table 5.

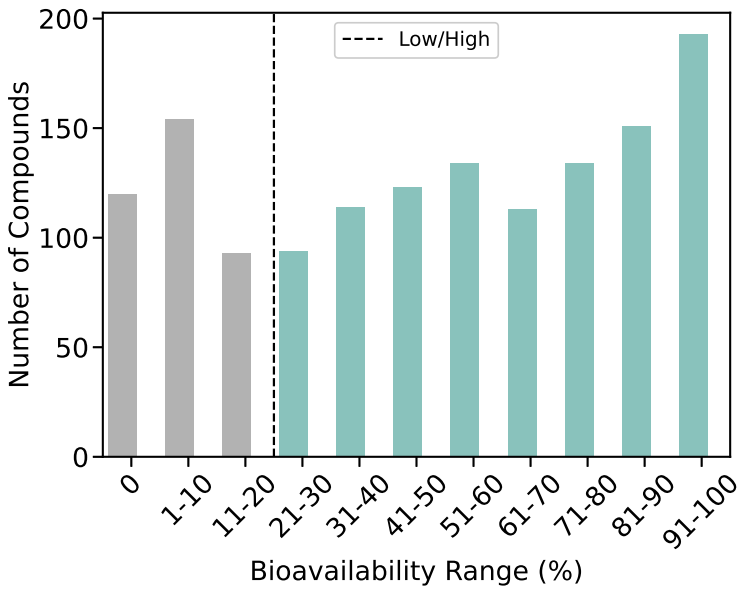


Figure 3. Distribution of compounds by different bioavailability ranges.

Datasets	Molecules	High	Low
Training	1480	1116	364
Test 1	288	215	73
Test 2	45	28	17

Table 5. Information of the training and testing datasets.

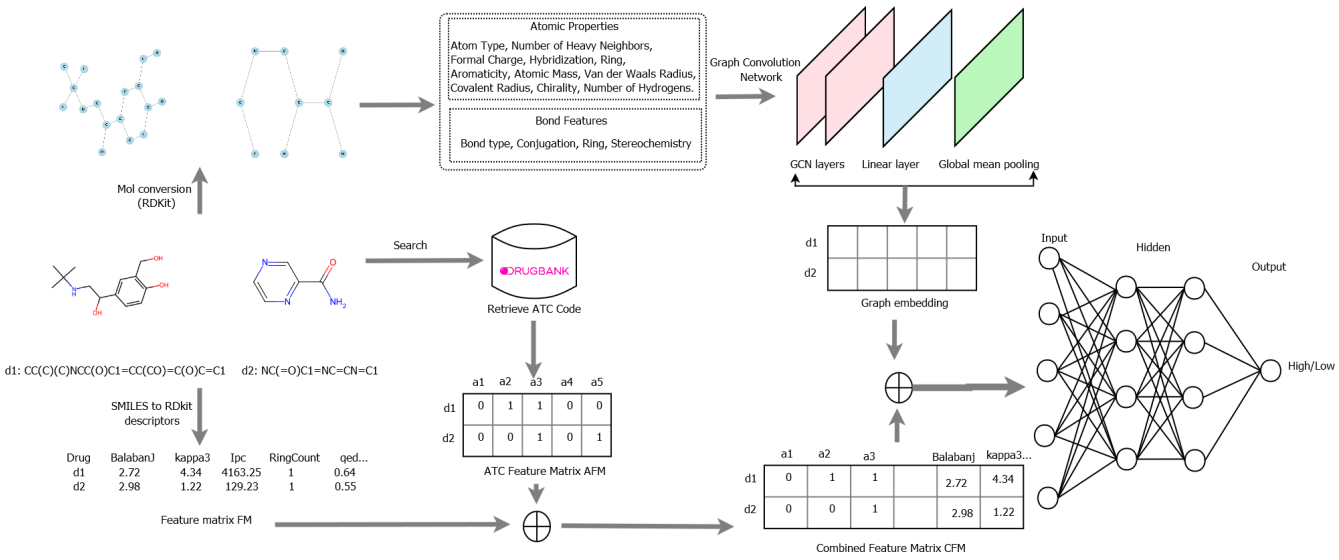


Figure 4. Proposed framework.

Descriptors

One of the key factors in HOB prediction is carefully selecting the molecular descriptors. All molecules in the training and test datasets are represented as 2D structures using their SMILES notations with the RDKit Python package[42]. A total of 200 molecular descriptors is calculated for each drug using this package. A detailed description of the descriptors is provided in the supplementary file Table S2. Descriptors with zero values across all molecules and those with more than 20% zero values are filtered out, resulting in a total of 75 molecular descriptors. Next, the Pearson correlation coefficient is computed for each pair of descriptors, as shown in the heatmap in the supplementary file Figure S1. To remove redundancy, descriptors with a similarity greater than 90% are excluded due to high correlation. This process further reduces the number of descriptors to 44. Hence, the feature matrix (FM) of size 1480×44 is obtained.

ATC code

The ATC codes are collected from the DrugBank database[43]. Some drugs are annotated with multiple ATC codes, and it is crucial to include ATC codes for drugs that are ingredients in other drugs as well. Let $D = \{d_1, d_2, \ldots, d_m\}$ be a set of drugs, and a drug d_i have a set of ATC codes $A = \{a_1, a_2, \ldots, a_n\}$. Each ATC code a_j is decomposed into multiple hierarchical levels to form its semantic set S. For instance, the drug *Propafenone* has an ATC code $C01BC03$. The semantic set of C01BC03 is $S = \{C, C01, C01B, C01BC, C01BC03\}$, where C (first level) represents the anatomical main group, $C01$ (second level) is the first therapeutic subgroup, $C01B$ (third level) is the pharmacological subgroup, $C01BC$ (fourth level) defines the chemical subgroup, and $C01BC03$ (fifth level) is the chemical substance. To construct the ATC feature matrix (AFM), the unique semantic sets across all drugs (training and testing) are considered. Each row of an AFM corresponds to a drug, and each column corresponds to a unique semantic set member. For each drug, the frequency of occurrences of each semantic set member is counted. If a drug is not a member of the semantic set, a value of 0 is assigned. This structured approach ensures that the hierarchical nature of ATC codes is incorporated into the model, enhancing its ability to predict bioavailability by leveraging detailed classification information. An example of the AFM is given in the supplementary file. The FM is combined with the AFM to create a Combined Feature Matrix (CFM).

BioGCN

The fundamental framework of BioGCN lies on the Graph Convolutional Network (GCN) introduced by Kipf and Welling[44]. It is essentially a deep neural network architecture for learning and processing molecular representations. The GCN operates on the graph data structure, allowing it to capture both local and global structural information. Figure 4 presents the workflow of the proposed framework. In this framework, each drug molecule is modeled as a graph. Formally, a graph $G = (V, E)$ is defined by a set of nodes V and a set of edges E that represent the relationships between the nodes. In the context of a drug molecular graph, each node $v_i \in V$ corresponds to an atom within the molecule, while each edge $e_j \in E$ represents a bond between two atoms. The molecular graph is typically considered an undirected graph because chemical bonds are bidirectional. The connections between atoms in the molecular graph are presented by an adjacency matrix $A \in \mathbb{R}^{M \times M}$, where $M = |V|$ is the

number of atoms. The atomic bonds are defined in $A = \{a_{mn}\}$ as shown in Equation 2.

$$a_{mn} = \begin{cases} 1, & \text{if atom } v_m \text{ and atom } v_n \text{ is connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Each node $v_i \in V$ is described by a feature vector, $f \in \mathbb{R}^{\mathbb{D}}$, considering 11 atom (node) features and four bond (edge) features. Each feature vector consists of 89 features: 79 atom features and 10 edge features. The 89 molecular features are summarized in Table 6. The final feature vector for each atom is defined by concatenating all features, whether one-hot encoded or represented as single values. All the feature vectors are organized in a node feature matrix, $H \in \mathbb{R}^{M \times N}$, with dimensions $M \times N$, where N refers to the number of features associated with each node. Each row f_i (for $i = 1, 2, \ldots, M$) of the matrix represents the feature vector of a different atom in the molecule. The SMILES notation of each drug is used to obtain the molecular graph.

Feature Category	Description	Size
Atom Features:		
Atom Type	C, N, O, S, F, Si, P, Cl, Br, Mg, Na, Ca, Fe, As, Al, I, B, V, K, Tl, Yb, Sb, Sn, Ag, Pd, Co, Se, Ti, Zn, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb, or unknown	43
Number of Heavy Neighbors	0, 1, 2, 3, 4, or more than four	6
Formal Charge	-3, -2, -1, 0, 1, 2, 3, or extreme	8
Hybridization	S, SP, SP2, SP3, SP3D, SP3D2, or other	7
Ring	If the atom is in a ring	1
Aromaticity	If the atom is aromatic	1
Atomic Mass	Scaled atomic mass	1
Van der Waals Radius	Scaled van der Waals radius	1
Covalent Radius	Scaled covalent radius	1
Chirality	Unspecified, clockwise, counter-clockwise, and Other	4
Number of Hydrogens	0, 1, 2, 3, 4, or more than four	6
Bond Features:		
Bond Type	Single, double, triple, or aromatic	4
Is Conjugated	If bond is conjugated	1
Is in Ring	If bond is part of a ring	1
Stereochemistry	E/Z, cis/trans	4

Table 6. Description of molecular features.

The BioGCN architecture is primarily divided into two parts: graph embeddings via GCN and a fully connected network for HOB prediction. The GCN component consists of two convolutional layers, a linear layer, and a pooling operation that aggregates node features to provide a global graph-level representation. The convolution layers propagate information across neighboring nodes, and the linear layer is applied to edge features to capture node interactions. The propagation rule in GCN is expressed in Equation 3.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3}$$

Here, $\tilde{A} = A + I_M$ is the adjacency matrix of the molecular graph with an added self-loop. I_M represents the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix, $W^{(l)}$ is the layer-specific trainable weight matrix, and σ defines the activation function, i.e., Rectified Linear Units (ReLU)[45]. $H^{(l)}$ is the node feature matrix at layer l, and $H^{(l+1)}$ is the output after the convolution at layer $l+1$. The edge features are processed using a linear layer as shown in Equation 4, where W_e and b_e are the weight matrix and bias, respectively.

$$\text{edge_features} = W_e \cdot \text{edge_attr} + b_e \tag{4}$$

Once the node and edge features are processed, a global mean pooling layer is applied to aggregate the node features across the entire graph as delineated in Equation 5, where x_i represents the node feature of node i, and M is the total number of nodes in

the graph. This produces a fixed-size embedding representing the molecular graph, regardless of the number of nodes.

$$\text{graph_embedding} = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{5}$$

In this study, a fully connected feedforward neural network is employed for a two-class problem. After that, the graph embeddings, along with FM and CFM, are subsequently fed to the neural network to predict HOB. The neural network architecture comprises three dense layers, interconnected through a non-linear activation function. The non-linear transformation is performed using the ReLU activation function. To mitigate overfitting, dropout is incorporated. The Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) is a loss function that combines a sigmoid layer with binary cross-entropy. The loss is computed as described in Equation 6, where P is the number of samples, y_i is the true label for the i^{th} sample, and Z_i is the logits for the i^{th} sample. Here, the $\sigma(.)$ is defined in Equation 7. Adam optimization is utilized to improve the learning efficiency.

$$\text{BCE_Loss} = -\frac{1}{P}\sum_{i=1}^{P}\left[y_i \log\left(\sigma(z_i)\right) + (1-y_i)\log\left(1-\sigma(z_i)\right)\right] \tag{6}$$

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}} \tag{7}$$

Evaluation metric

The performance of the different models is evaluated using six different statistical metrics, including accuracy (Equation 8), balanced accuracy (Equation 11), recall (Equation 12), precision (Equation 13), and F1 score (Equation 14). The equations of these measures are below,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity = \frac{TN}{FP+TN} \tag{10}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. Another important metric for evaluating binary classification performance is the Area Under the Receiver Operating Characteristic Curve (AUC). An AUC value near to 1 indicates stronger classification performance. Conversely, an AUC value of 0.5 indicates that the model's performance is no better than a random classifier.

References

1. Myung, Y., de Sá, A. G. & Ascher, D. B. Deep-pk: deep learning for small molecule pharmacokinetic and toxicity prediction. *Nucleic Acids Res.* gkae254 (2024).
2. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sinica B* **12**, 3049–3062 (2022).
3. Van Norman, G. A. Drugs and devices: comparison of european and us approval processes. *JACC: Basic to Transl. Sci.* **1**, 399–412 (2016).
4. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nat. reviews Clin. oncology* **8**, 189–190 (2011).
5. Fernandez-Moure, J. S. Lost in translation: the gap in scientific advancements and clinical application. *Front. Bioeng. Biotechnol.* **4**, 43 (2016).
6. Yan, A., Wang, Z. & Cai, Z. Prediction of human intestinal absorption by ga feature selection and support vector machine regression. *Int. journal molecular sciences* **9**, 1961–1976 (2008).
7. Wang, J. & Hou, T. Advances in computationally modeling human oral bioavailability. *Adv. drug delivery reviews* **86**, 11–16 (2015).
8. Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D. & Zhu, H. Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm. research* **31**, 1002–1014 (2014).
9. Wei, M. *et al.* Hobpre: Accurate prediction of human oral bioavailability for small molecules. *J. Cheminformatics* **14**, 1 (2022).
10. Cabrera-Pérez, M. Á. & Pham-The, H. Computational modeling of human oral bioavailability: what will be next? *Expert. opinion on drug discovery* **13**, 509–521 (2018).
11. Andrews, C. W., Bennett, L. & Yu, L. X. Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship. *Pharm. research* **17**, 639–644 (2000).
12. Moda, T. L., Montanari, C. A. & Andricopulo, A. D. Hologram qsar model for the prediction of human oral bioavailability. *Bioorganic & medicinal chemistry* **15**, 7738–7745 (2007).
13. Yang, Q. *et al.* Construction of an oral bioavailability prediction model based on machine learning for evaluating molecular modifications. *J. Pharm. Sci.* **113**, 1155–1167 (2024).
14. Sakai, M. *et al.* Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci. reports* **11**, 525 (2021).
15. Yoshida, F. & Topliss, J. G. Qsar model for drug human oral bioavailability. *J. medicinal chemistry* **43**, 2575–2585 (2000).
16. Turner, J. V., Glass, B. D. & Agatonovic-Kustrin, S. Prediction of drug bioavailability based on molecular structure. *Anal. Chimica Acta* **485**, 89–102 (2003).
17. Wang, J., Krudy, G., Xie, X.-Q., Wu, C. & Holland, G. Genetic algorithm-optimized qspr models for bioavailability, protein binding, and urinary excretion. *J. chemical information modeling* **46**, 2674–2683 (2006).
18. Ma, C.-Y. *et al.* Prediction models of human plasma protein binding rate and oral bioavailability derived by using ga–cg–svm method. *J. pharmaceutical biomedical analysis* **47**, 677–682 (2008).
19. Kumar, R., Sharma, A. & Varadwaj, P. K. A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine. *J. natural science, biology, medicine* **2**, 168 (2011).
20. Ng, S. S. & Lu, Y. Evaluating the use of graph neural networks and transfer learning for oral bioavailability prediction. *J. Chem. Inf. Model.* **63**, 5035–5044 (2023).
21. Ma, L. *et al.* Research on prediction of human oral bioavailability of drugs based on improved deep forest. *J. Mol. Graph. Model.* **133**, 108851 (2024).
22. Li, M. *et al.* Guest editorial: Deep neural networks for graphs: Theory, models, algorithms, and applications. *IEEE Transactions on Neural Networks Learn. Syst.* **35**, 4367–4372 (2024).
23. Bai, L. *et al.* Haqjsk: Hierarchical-aligned quantum jensen-shannon kernels for graph classification. *IEEE Transactions on Knowl. Data Eng.* **36**, 6370–6384 (2024).
24. Li, J., Zheng, R., Feng, H., Li, M. & Zhuang, X. Permutation equivariant graph framelets for heterophilous graph learning. *IEEE Transactions on neural networks learning systems* **35**, 11634–11648 (2024).

25. Yan, C., Suo, Z., Wang, J., Zhang, G. & Luo, H. Dacpgtn: drug atc code prediction method based on graph transformer network for drug discovery. *Front. Pharmacol.* **13**, 907676 (2022).

26. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

27. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. learning* **20**, 273–297 (1995).

28. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory* **13**, 21–27 (1967).

29. Cox, D. R. The regression analysis of binary sequences. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **20**, 215–232 (1958).

30. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals statistics* 1189–1232 (2001).

31. McCallum, A., Nigam, K. *et al.* A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, vol. 752, 41–48 (Madison, WI, 1998).

32. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. computer system sciences* **55**, 119–139 (1997).

33. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

34. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. *Adv. neural information processing systems* **31** (2018).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

36. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).

37. Wang, J. *et al.* Deepatomiccharge: a new graph convolutional network-based architecture for accurate prediction of atomic charges. *Briefings bioinformatics* **22**, bbaa183 (2021).

38. Yang, H. *et al.* admetsar 2.0: web-service for prediction and optimization of chemical admet properties. *Bioinformatics* **35**, 1067–1069 (2019).

39. Dong, J. *et al.* Admetlab: a platform for systematic admet evaluation based on a comprehensively collected admet database. *J. cheminformatics* **10**, 1–11 (2018).

40. Olivares-Morales, A. *et al.* The use of roc analysis for the qualitative prediction of human oral bioavailability from animal data. *Pharm. research* **31**, 720–730 (2014).

41. Ahmed, S. S. & Ramakrishnan, V. Systems biological approach of molecular descriptors connectivity: optimal descriptors for oral bioavailability prediction. *PLoS One* **7**, e40654 (2012).

42. Landrum, G. Rdkit documentation. *Release* **1**, 4 (2013).

43. Wishart, D. S. *et al.* Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research* **46**, D1074–D1082 (2018).

44. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

45. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814 (2010).

Acknowledgements

K.M. and S.B. would like to thank Anusandhan National Research Foundation (ANRF), Government of India (GoI), under the National Post Doctoral Fellowship (NPDF) scheme (File No. PDF/2023/002800) during the period March 2024 to March 2026 for providing all necessary support and facilities to carry out the research. S.B. acknowledges the JC Bose Fellowship grant No. JBR/2021/000036 from ANRF, GoI. The authors would like to extend their gratitude to Dr. Prabhat R. Mishra, Chief Scientist and Head of the Pharmaceutics and Pharmacokinetics Division at CSIR-Central Drug Research Institute, Lucknow, for providing the dataset used in this study.

Author contributions

K.M. conceived and designed the study, performed the experiments, analyzed the results, and drafted the manuscript. S.B. contributed to the initial conceptualization of the research topic, supervised the work, and provided critical revisions. All authors reviewed the manuscript, provided valuable feedback, and approved the final version of the manuscript.

Data availability

The code and the preprocessed data are available at https://github.com/Koyel-Mandal/BioGCN.git.

Competing interests

The authors declare no competing interests.

Supplementary

The additional information is summarized in the supplementary.pdf.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryrevised.pdf