Computers in Biology and Medicine 179 (2024) 108729

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# A graph neural network approach for predicting drug susceptibility in the human microbiome

Maryam [a], Mobeen Ur Rehman [b], Irfan Hussain [b], Hilal Tayara [c,*], Kil To Chong [a,d,**]

[a] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea
[b] Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University, United Arab Emirates
[c] School of International Engineering and Science, Jeonbuk National University, Jeonju, 54896, South Korea
[d] Advances Electronics and Information Research Centre, Jeonbuk National University, Jeonju, 54896, South Korea

ARTICLE INFO

*Keywords:*
Microbiome
Graph neural network
Molecular docking
Bioinformatics

ABSTRACT

Recent studies have illuminated the critical role of the human microbiome in maintaining health and influencing the pharmacological responses of drugs. Clinical trials, encompassing approximately 150 drugs, have unveiled interactions with the gastrointestinal microbiome, resulting in the conversion of these drugs into inactive metabolites. It is imperative to explore the field of pharmacomicrobiomics during the early stages of drug discovery, prior to clinical trials. To achieve this, the utilization of machine learning and deep learning models is highly desirable. In this study, we have proposed graph-based neural network models, namely GCN, GAT, and GINCOV models, utilizing the SMILES dataset of drug microbiome. Our primary objective was to classify the susceptibility of drugs to depletion by gut microbiota. Our results indicate that the GINCOV surpassed the other models, achieving impressive performance metrics, with an accuracy of 93% on the test dataset. This proposed Graph Neural Network (GNN) model offers a rapid and efficient method for screening drugs susceptible to gut microbiota depletion and also encourages the improvement of patient-specific dosage responses and formulations.

## 1. Introduction

Microorganisms such as bacterial species, viruses, and fungi are highly diverse and exhibit dynamic behavior. They can colonize human cells and play a substantial role in human cells, specifically in the gut, intestines, and skin [1]. The microbial role of the microbiome in the human body is to protect against pathogens and enhance metabolic and immunity capabilities [2]. Microbes possess the ability to resist the invasion of external pathogens [3] and contribute to the synthesis of essential vitamins and sugar metabolism to boost T-cell responses [4]. Therefore, the abnormal growth of microorganisms in human cells influences health and can lead to diseases such as obesity [5], inflammatory bowel disease [6], and even cancer [7].

Similarly, the microbiome also exerts a significant effect on drugs. However, several studies have shown that microbiomes can also have pharmacological effects on drugs used to treat various diseases [1,8–13]. Javdan et al. proposed an experimental procedure known as Microbiome-Derived Metabolism (MDM)-screen to investigate the ability of gut microbiota to metabolize drugs. Their findings elucidated the diversity of individual human microbiota and underscored the importance of the microbiome in drug development [14]. In another study, Maier et al. compared 1079 marketed drugs in representative intestinal symbiotic microorganisms and found that 24% of the drugs had inhibitory effects on microorganisms [15]. Concetta et al. explained that the association between anticancer drugs and the microbiome affects drug efficacy and can lead to toxic effects [16].

Microbial drug metabolism relies heavily on the production of specific enzymes by particular individuals. However, this expression varies from patient to patient [17–19]. For example, Tacrolimus is a medication whose microbial metabolism has been associated with *F. prausnitzii* [20]. Lee et al. highlight the abundance of *F. prausnitzii*, which has a direct correlation with the dosage of tacrolimus. This association is the result of tacrolimus being converted by *F. prausnitzii* into a metabolite M1, as shown in Fig. 1, that has 15-fold less immunosuppressant activity than tacrolimus itself [21]. Another drug, digoxin, which involves the production of cardiac glycoside reductase enzyme (CGR), is deactivated by the *E. lenta* bacterial strain [22]. Studies have revealed a substantial correlation between patients' ex vivo metabolism of digoxin and the abundance of the CGR gene compared to *E. lenta*

* Corresponding author.
** Corresponding author at: Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea.
*E-mail addresses:* hilaltayara@jbnu.ac.kr (H. Tayara), kitchong@jbnu.ac.kr (K.T. Chong).

https://doi.org/10.1016/j.compbiomed.2024.108729
Received 22 April 2024; Received in revised form 4 June 2024; Accepted 8 June 2024
Available online 1 July 2024

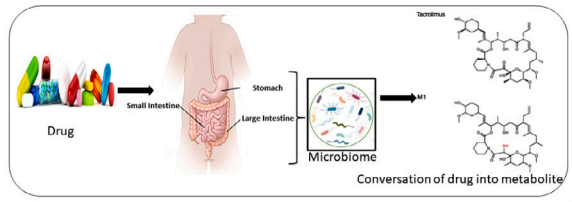


**Fig. 1.** Illustration of drug metabolism by microbiome.

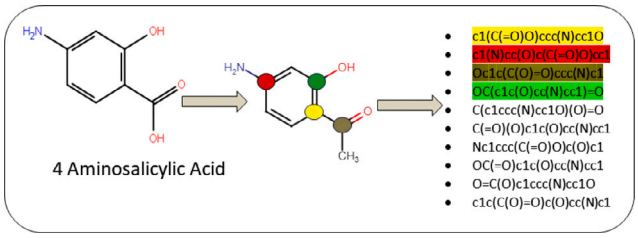


**Fig. 2.** 10X SMILES augmentation representation of compound. The first four SMILES highlighted the labeling initialization of the SMILE string.

concentration in their feces, demonstrating that individuals with CGR-encoding bacteria have greater capacity for in vivo metabolism of digoxin [20].

Individual pharmacokinetic diversity resulting from microbial drug depletion can contribute to toxicity and treatment failure in certain patients [23,24]. Moreover, these challenges can impede the progression of novel drugs to the market. However, the metabolic toxicity of drugs from the microbiome is not commonly assessed during clinical development [25]. In many rare cases, drug stability is primarily evaluated in the colonic environment rather than exploring pharmacokinetic variability [26]. To date, no accepted method has been proposed to quantify the toxicity of drugs due to microbiome activity [19]. Nevertheless, testing the association between drugs and the microbiome involves intestinal fluids, microbial cultures, and incubation of specific drugs in human fecal samples [9,14]. These experimental methods are time-consuming and resource-intensive. Therefore, in silico methods play a crucial role in predicting drug-microbial associations and drug depletion [8,27–29].

Recently, researchers have employed machine learning [30,31] and deep learning methods [32–36] to predict microbial drug toxicity. For instance, Elmassrt et al. utilized the common substructural algorithm to predict drug-microbial associations [37]. Zimmerman et al. identified the essential functional groups of drugs that form bonds with microbes, leading to microbial depletion [9]. Similarly, in 2017, Sharma et al. developed a machine learning model called “Drug Bug”, employing a random forest classifier and achieving predictive performance exceeding 90%, focusing exclusively on bacterial reactions with drugs [38].

The primary objective of this study was to identify the most optimal neural network for predicting microbial drug interactions. This study unfolds the performance of graph neural networks, such as Graph Convolutional Networks (GCN) [39], Graph Isomorphic Networks (GINCOV) [40], and Graph Attention Networks (GAT) [41] for predicting microbial drug interactions. Unlike previous studies that focused solely on machine learning models, this study conducted a comparison of graph neural networks (GNN) models. Furthermore, comparing with previous studies posed challenges due to differences in evaluation metrics. Our proposed models demonstrated improved performance compared to previously employed machine learning models. Among all the models tested, the Graph Isomorphic Neural Network outperformed the others and was selected as the final model for predicting microbial drug interactions.

## 2. Methods and material

### 2.1. Dataset preparation

Datasets related to microbial drug interactions were gathered from various literature sources, with a primary focus on the works of Zimmermann et al. [9] and Javdan et al. [14]. In Zimmerman et al.’s study, 271 drugs were incubated with 76 gut bacterial strains for 24 h under anaerobic conditions. The results suggested that a drug is considered depleted if its starting concentration is reduced by at least 20% by at least one bacterial strain. Similarly, in Javdan et al.’s [14] research, 438 drugs were incubated in the presence of gut microbiome strains obtained from a single human donor. A drug was labeled as metabolized if it was entirely consumed or transformed into other metabolites within 24 h in at least 2 or 3 independent experiments [26,42–44].

Furthermore, studies addressing drug metabolism in animals were excluded from our analysis since the microbiome composition in animals significantly differs from that in humans. We considered only studies based on gut bacterial isolates or human fecal/intestinal fluid samples [45].

Based on the studies mentioned above, drugs interacting with the microbiome (active) were labeled as (1), while those not interacting (inactive) were labeled as (0) as mentioned by McCoubrey et al. [27]. Subsequently, data preprocessing steps were carried out, including the removal of invalid SMILE strings, correction of molecular SMILEs valency, and removal of duplicate SMILEs from the dataset. As a result, we finalized a dataset comprising 429 compounds, which were used for training and testing the performance of the graph-based neural network model shown in Table 1.

### 2.2. Smile-based augmentation for training data

To attain strong performance and extract valuable insights from deep learning models, a substantial amount of data is essential. Therefore, in our comprehensive analysis of the model’s performance, we emphasize the use of multiple SMILES representations. Our proposed approach involves augmenting the available data by generating diverse representations of molecular structures using SMILES strings. This technique aids deep learning models in better generalization from a wide range of datasets, mitigating issues of overfitting and underfitting. Furthermore, it offers a practical solution to enhance prediction performance in molecular property prediction tasks, even when labeled data is limited [46].

The “Simplified Molecular Input Line Entry System” (SMILES) format for all active and inactive drugs was obtained from the PubChem database. SMILES notation is a widely accepted method for representing small molecules in a concise line notation format. To create the multiple SMILES representations, we implemented a mapping function from the RDKit library. In this process, a random SMILES sequence is first generated, and then atom numbering is performed using RDKit’s “MolToSmiles” method, with the canonical setting set to “False” [47].

As depicted in Fig. 2, we generated ten different SMILES representations for each molecules in the training dataset, each displaying distinct atom numbering in randomly generated SMILES sequences. Enumerate small molecules were labeled as positive and negative data based on the represented SMILE molecule before training the deep learning model.

### 2.3. Molecular feature extraction

Molecular features from the SMILES notation were computed using the RDKit library, which includes nine types of atomic features and four types of bond features. These features were encoded using one-hot encoding methods. For instance, feature hybridization was encoded with a 7-bit one-hot vector. Chirality was represented as a 2-bit one-hot vector, as illustrated in Table 2 [48].

**Table 1**
Detail of microbial drugs interaction dataset used for training and testing the GNN models.

| Dataset | Training dataset | Testing dataset | Total dataset |
|---|---|---|---|
| Experimental data | 349 | 80 | 429 |
| SMILES Enumeration Ratio 10X | 3775 | 80 | 3855 |

**Table 2**
Atomic and bond level features used for one hot encoding of molecular SMILES.

| Type of attribute | Description | Encoding dimension |
|---|---|---|
| Atom attribute | | |
| Atom symbol (Node) | 'Ag', 'Al', 'As', 'Au', 'B', 'Ba', 'Be', 'Bi', 'Br', 'C', 'Ca', 'Cd', 'Cl', 'Co', 'Cr', 'Cs', 'Cu', 'F', 'Fe', 'Ga', 'Ge', 'H', 'Hf', 'Hg', 'I', 'In', 'Ir', 'K', 'La', 'Li', 'Lu', 'Mg', 'Mn', 'Mo', 'N', 'Na', 'Nb', 'Ni', 'O', 'Os', 'P', 'Pb', 'Pd', 'Pt', 'Rb', 'Re', 'Rh', 'Ru', 'S', 'Sb', 'Sc', 'Se', 'Si', 'Sn', 'Sr', 'Ta', 'Te', 'Ti', 'Tl', 'U', 'V', 'W', 'Y', 'Zn', 'Zr', 'Unknown' | 66 |
| Degree | Number of covalent bonds (0,1,2,3,4,5) | 8 |
| Atomic charges | Electric charges | 1 |
| Hybridization | s, sp1, sp2, sp3, sp3d, Sp3d2 | 7 |
| Chirality | Atomic chirality | 1 |
| Type of chirality | Rectus and Sinister | 2 |
| Aromaticity | Atom in aromatic part (true/false) | 1 |
| Radical electrons | Total number of radical electrons | 1 |
| Hydrogens | Number of connected hydrogens | 5 |
| Bond attribute | | |
| Type of bond | Single, Double, Triple, Aromatic | 4 |
| Conjugation | Bond conjugation | 1 |
| Stereo | StereoE, StereoZ, StereoAny, StereoN | 4 |
| Ring | Number of bonds in ring parts | 1 |

### 2.4. Graph neural network models

Graph neural networks (GNNs) have the ability to learn molecular representations by directly applying convolutional operations to encoded molecular graphs. In this context, atoms and bonds are represented as nodes and edges within the graph, denoted as G = (V, E). Here, V (nodes) encompass atomic features, including atom symbol, atom degree, atomic charges, radical electrons, aromaticity, hybridization, hydrogens, and chirality type. The edges (E) in the graph capture bond features such as bond types, ring relationships, stereochemistry, and conjugation. All of these features are one-hot encoded based on molecular SMILES representations and serve as inputs to the graph-based models [49].

#### 2.4.1. Graph convolutional neural network

GCNs adopt the concept of convolutional operation from a convolutional neural network (CNN) and perform an aggregation over the neighborhood of each node which is connected directly to its edge. The GCNs convolutional operation work as follow: 1. Collect data from neighbor nodes. 2. Perform aggregation function. 3. Update the node information based on aggregation information. 4. Compute the Classification or regression task. The GCN uses a normalization-based aggregation function as the initial formulation. The graph's adjacency matrix is represented as X, where $X(ab) = 1$ if node a and b are connected, otherwise 0. Likewise, $A^I$ represents the node matrix at layer I in a neural network. The updated rule for each node is as follows.

$$A^{(i+1)} = \sigma\left(D^{-1/2} \cdot X \cdot D^{-1/2} \cdot A^{(i)} \cdot W^{(i)}\right)$$

In this equation: $\sigma$ = ReLU activation function, D = Degree matrix, W = weight matrix for layer i in the network. Additionally, if the node has more than one neighbor node then the average of nodes is taken to update the node.

#### 2.4.2. Graph attention network

The Graph Attention Network (GAT), introduced by Veličković et al. [50], is a novel neural network architecture that employs attention mechanisms to assess the importance of neighboring nodes when collecting feature information for a given node. In contrast to other graph neural networks that assign equal importance to all neighboring nodes, GAT calculates attention coefficients using learnable parameters. This innovative approach enhances the quality of node representations by prioritizing them based on their significance while filtering out noisy and less informative connections [51].

The GAT neural network involves the following key steps: node embedding, graph attention layer, and aggregation and refinement. During these steps, feature representations of neighboring nodes are combined using attention coefficients to derive the updated representation of a node $x$. The mathematical representation of this process is given by:

$$x_i' = \sigma\left(\sum_{j \in N_i} \sigma_{ij} \cdot W \cdot h_j\right) \tag{1}$$

where:

$\alpha$ represents the activation function. $x_i'$ denotes the updated representation of node $x_i$ after attention and aggregation. This equation succinctly illustrates how GAT calculates the updated representation of a node by taking into account the attention coefficients assigned to its neighboring nodes.

#### 2.4.3. Graph isomorphic neural network

Graph isomorphic neural network was recently proposed by Xu et al. [48] based on the Weisfeiler–Lehman isomorphic test. The architecture of GINCOV model is composed of three components including a message-passing layer, a read-out layer, and linear layers [52].

*Message passing layer* In the GINCOV model, the message-passing layer employs an enhanced neighboring aggregation method, where each node performs a linear transformation and aggregates the feature vectors of its neighboring nodes to create a new feature vector. For a given node $v$, the node features $x_v^{(i+1)}$ are updated as follows:

$$x_v^{(i+1)} = \mathrm{ReLU}\left(W_g^{(i)} \times \left(x_v^{(i)} + \sum_{u \in N(v)} x_u^{(i)}\right) + b_g^{(i)}\right) \tag{2}$$

where:

$x_v^{(i+1)}$ represents the node feature of node $v$ after increment $i + 1$. $N(v)$ is the set of neighboring nodes of node $v$. $W_g^{(i)}$ denotes the node weights. $b_g^{(i)}$ stands for the node bias. ReLU is the activation function.

*Read-out layer* After the message-passing layer, a permutation-invariant function is employed as a readout function to aggregate all the node features from the last iteration into a graph embedding for the complete molecular graph. The summation function used for aggregating all-node features into the graph embedding is as follows:

$$x_g = \sum_{v \in g} x_v^{(i)} \tag{3}$$

Here, $g$ represents the entire molecular graph.

*Linear layer* Subsequently, the entire graph embedding $x_g$ is passed through fully connected layers and undergoes a nonlinear transformation as follows:

$$x_{i+1} = \mathrm{ReLU}\left(x_i^{W_i} + b_i\right) \tag{4}$$

where:

$x_{(i+1)}$ represents the output of the $(i + 1)$th layer. $x_i$ is the input to the $i$th layer. $W_i$ represents the weight matrix of the $i$th layer. $b_i$ is the bias term of the $i$th layer. As depicted in Fig. 3, the GINCOV model architecture model consists of graph isomorphic layers that capture node-level features from neighboring nodes and update the target node. The output of this layer is then passed to the next linear layer as input, and the process repeats. The result of the linear readout layer serves as the output for the graph classification problem.

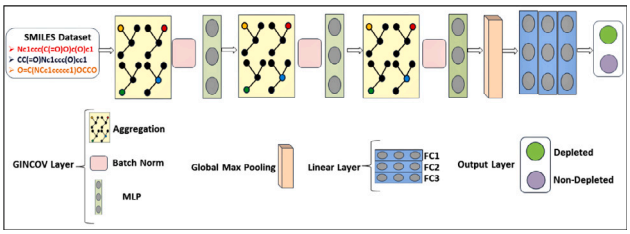


**Fig. 3.** GINCOV model architecture used in this study.

### 2.5. Model implementation protocol

Graph-based neural network models, including GINCOV, GCN, and GAT, were trained using the PyTorch library [https://pytorch.org/] on a Linux OS, as illustrated in Fig. 4. The initial SMILES dataset was converted into graph data using RDKit [https://www.rdkit.org/] and the PyTorch graph library PyG https://pyg.org/. Subsequently, the data was split into training, validation, and testing sets using the Sklearn library [https://scikit-learn.org/].

Training of the graph neural network models utilized an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of RAM. The models were optimized using SGD and Adam optimizers for GINCOV, GCN, and GAT, respectively, as detailed in Table 1 in the supplementary material.

The cross-entropy loss function [53] was employed to calculate the loss, measuring the disparity between the actual and predicted labels of drug molecules. The cross-entropy loss function is defined as follows:

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

where:

$\mathcal{L}(\hat{y}, y)$ represents the cross-entropy loss. $\hat{y}$ denotes the model's predicted label vector. $y$ is the true label vector. $N$ is the number of classes. To mitigate overfitting and reduce training time, early stopping [54] was implemented. A maximum epoch limit of 200 was set for the training process. If the performance metric did not improve after 20 iterations on both the training and validation sets, the training process was terminated early. It is worth noting that the specific criteria for early stopping may vary based on empirical data and dataset size. Early stopping is a practical approach to find the optimal set of hyperparameters for the model.

### 2.6. Model evaluation metrics

In this study, the proposed model was computed using Accuracy, Sensitivity, Specificity, AUROC, MCC, and F1 score metrics [55,56], to comprehensively evaluate model's performance in predicting the drug susceptibility. The evaluation metrics better align with our study objective to develop a reliable model for drug susceptibility prediction. Accuracy(ACC), F1 score, AUROC metrics explain the model's overall performance whereas, the ability to accurately identify both susceptible and non-susceptible cases was analyzed using Sensitivity, Specificity evaluation method. Furthermore, Matthews Correlation Coefficient (MCC) particularly used to evaluate the data imbalances. The other metrics Balanced Accuracy, Weighted Recall, Weighted Precision were computed to compare with the previous proposed models [27] performances.

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\mathbf{F}_1 = \frac{2 \times \mathrm{TP} \times \mathrm{TN}}{\mathrm{TP} + \mathrm{TN}} \tag{9}$$

$$\textbf{BalancedAccuracy} = \frac{1}{2} \times \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{10}$$

$$\textbf{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + TN) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{11}$$

where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

### 2.7. In silico drug bank screening

The final graph-based model (GINCOVnet) was further employed to evaluate the potential interaction of small molecules against beta-glucuronidase with in drug bank database, that comprising over 12 685 drugs. This step serves as an external validation of our suggested model by predicting the compounds from drug bank, that may exhibit comparable conformation to our active dataset and be involve in drug-microbiome interactions. Several steps were performed to find the probability of hits to bind with microbiome target protein. Firstly, drugs were filters to remove the redundancy, salts and metal ions. After prepossessing, compounds were screened through GINCOVnet model to predict the probability scores of active and inactive. Compounds having probability score of more than 0.9 is considered as active hits and were further processed for molecular docking.

### 2.8. Molecular docking simulation

Molecular docking is a computational modeling approach used to predict the optimized binding conformation between drugs and biological targets, such as proteins or DNA. This method relies on a scoring function to compute interacting parameters, which are then employed to predict the energy profiling, binding stability, and strength of the complex formed by these molecules, including drugs and proteins [57–59]. In this study, molecular docking was employed to determine the binding strength of the drugs that were screened by proposed deep learning models.

#### 2.8.1. Protein and ligands preparations

The crystal structure of *E. coli* beta-glucuronidase (PDB: 3K4D) was retrieved from RCSB Protein Data Bank https://www.rcsb.org/ shown in Fig. 5 [60]. Specifically, we targeted the enzyme E. coli beta-glucuronidase found in the gut microbiome. Research has demonstrated the pivotal role of gut microbiota beta-glucuronidase in reactivating drugs, processing xenobiotics, and modulating dietary metabolites. Furthermore, it plays a significant role in regulating active and potentially toxic metabolites produced through enterohepatic recirculation in the gastrointestinal tract (GI). As a result, beta-glucuronidase is considered a crucial therapeutic target in conditions such as Crohn's disease, colon cancer, and drug-induced gastrointestinal toxicity [60,61]. The crystal structure of $\beta$- glucuronidase was employed for protein preparation wizard embedded in Maestro 9.3 (Schrodinger 2019 suites). Protein preparation steps include the addition of hydrogen bonds, assigning bond orders, atomic charge fixing, and creating disulfide bonds were incorporated.

For ligand preparation, the Ligand preparation wizard named as Ligprep module from Maestro (Schrodinger 2023-2 suites) was incorporated. Ligand preparation includes the addition of hydrogen bonds in small molecules, correction of bond angle, bond degree, atom charges, ring conformation, and low energy structures. Additionally, forcefield OPLS 2005 was used for optimization and 32 conformations of each ligand were generated.

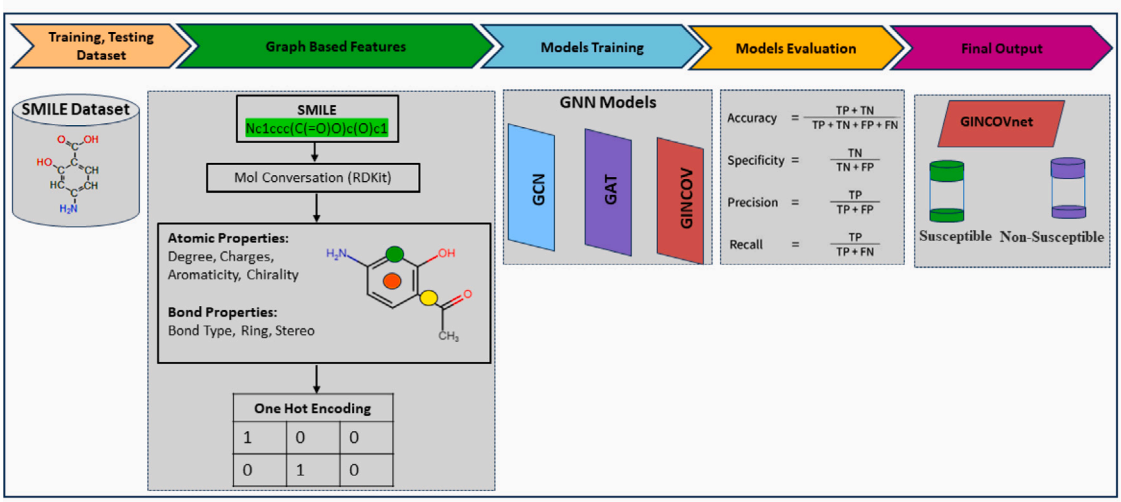


**Fig. 4.** Mapping out the workflow: From data collection to model development, validation, and evaluation, uncovering the intricate process of predicting drug susceptibility to microbiome interactions using graph neural network.

**Fig. 5.** Cartoon representation of 3D crystal structure of $\beta$- glucuronidase receptor used for molecular docking.

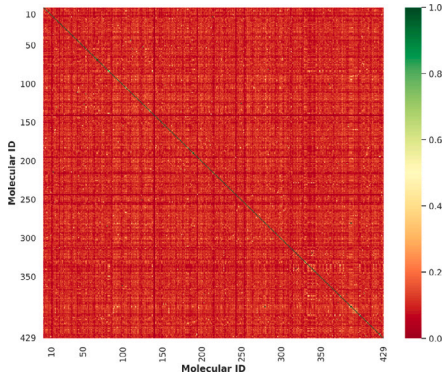


**Fig. 6.** Heatmap plot illustrated the Tanimoto similarity of the molecules using Morgan Fingerprints. x-and y-axis represent the number of compounds used in this study.

#### 2.8.2. Grid generation and GLIDE molecular docking

The minimized crystal structure of $\beta$- glucuronidase and optimized drugs were further employed for molecular docking using the GLIDE docking protocol. GLIDE is a rapid and accurate molecular docking technique embedded in maestro based on empirical scoring function. Moreover, the parameter used to evaluate the GLIDE docking output was GSCORE. GSCORE is a combination of different energies and interactions such as hydrogen bonds, hydrophobic and hydrophilic interactions, and pi-pi stacking [62].

Before molecular docking, a Grid box was generated to specify the active site of the protein [63,64]. A 3D box was generated to specify the active site of $\beta$- glucuronidase in the X, Y, and Z axis with the size of 40 Å. Further, Standard precision (SP) docking of active compounds was performed and ranked the compounds based on docking score and molecular interaction at the active site of target protein $\beta$- glucuronidase.

In the next step, the structural based molecular clustering was performed to select the best compound with high Glide GSCORE from each cluster, these selected compounds were further docked with extra precision maintaining rigid grid box, and utilizing extra precision (XP) glide docking. A total four compounds (one from each cluster) were selected and compare with training active compounds to analyze the binding conformations.

## 3. Results

### 3.1. Data analysis and visualization

Fig. 6 shows the chemical diversity of microbial interacting drugs dataset used in Graph-based neural network models. For this purpose, the Tanimoto similarity index calculation method was applied to the Morgan Fingerprints that were computing from SMILES data with a radius of 2. The heatmap plot demonstrated that most of the compounds within the training and testing datasets exhibited a similarity index below 0.3 with a mean value of 0.11, suggesting that the chemical compounds used in this study were diverse. After evaluating the chemical diversity of chemical compounds, the dataset was split into 80% for training and 20% for testing using the Sklearn library (https://scikit-learn.org/). Subsequently, the SMILES enumeration technique was applied to the training dataset using a tool developed by Bjerrum and each SMILES was enumerated with a 10X enumeration ratio as shown in Table 2.

### 3.2. Model training and evaluation

In this study, we have implemented the graph neural network models GINCOV, GCN, and GAT. For this purpose, initially, SMILES data was transformed into molecular structure (MOL) format using RDKit in Python. Subsequently, the one hot encoding has been used for encoding

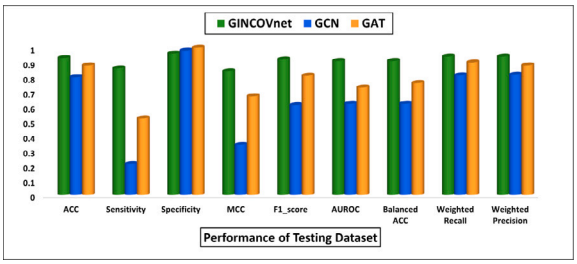


**Fig. 7.** Bar plot shows the performance of GNN models (GCN, GAT, GINCOV) on testing dataset in terms of different evaluation metrics.

**Table 3**
The performance comparison with previously proposed models.

| Comparison of GINCOVnet with existing models | | | |
|---|---|---|---|
| Model | Balanced accuracy | Weighted recall | Weighted precision |
| Extra trees algorithm (McCoubrey et al.) | 69 | 79.2 | 80.2 |
| GINCOVnet | 91 | 94 | 94 |

the atomic and bond properties of each molecule and then converted into graph representation using Pytorch geometric library (PyG) https://pyg.org/. The graph data of molecules and their interacting labels applied for training the GNN models such as GCN, GAT, and GINCOVnet using batch size 32. To evaluate the performance of graph-based neural networks, accuracy, sensitivity, and specificity matrix were evaluated. As our dataset is imbalanced, sensitivity and specificity better represent model performance. All three models were evaluated as shown in Fig. 7 shows that GINCOVnet performed well compared to the other two GNN models, with accuracy of 91% and 93% across and testing datasets respectively. The sensitivity and specificity score of GINCOVnet model is high, with value of 0.86 and 0.96 respectively. The values of sensitivity and specificity show that GINCOVnet also has better performance for test datasets. The AUROC plot, accuracy, sensitivity, and specificity for the test dataset are shown in the supplementary material.

### 3.3. Comparison with previous models

We have also compared the performance of our proposed model with previous machine learning models. The previous machine learning models achieved a balanced accuracy of 69%, whereas our proposed GINCOV model outperformed them with a balanced accuracy of 91%. Our proposed algorithm yields better predictions due to its high capacity to learn both local and non-local features from chemical structures, as demonstrated in Table 3. Furthermore, deep learning models exhibit superior performance, as they possess greater capability to learn complex hierarchical features, interactions, and nonlinear relationships in data when compared to traditional machine learning models.

### 3.4. GNN-based virtual screening and molecular docking analysis

Our proposed model was efficient discerning the depleted and non depleted drugs, even when dealing with the diverse set of molecules of the drug bank database. Drugbank database was screened by best model, resulting in the set of compounds that were predicted as positives by the GINCov model. The positive compounds were docked with the microbial receptor $\beta$-glucuronidase to assess the binding strength of the proposed depleted drugs. The results of molecular docking revealed that the top positive compounds exhibited GSCORE values ranging from −8 kcal/mol to −6 kcal/mol, as shown in Supplementary File 2. The ligand interaction diagrams of the top 5 compounds, presented in the supplementary material, illustrate that GLU413 and TYR472 of $\beta$-glucuronidase interacted with most of the ligands, forming pi-pi stacking and hydrogen bonds. GLU413 established hydrogen bonds

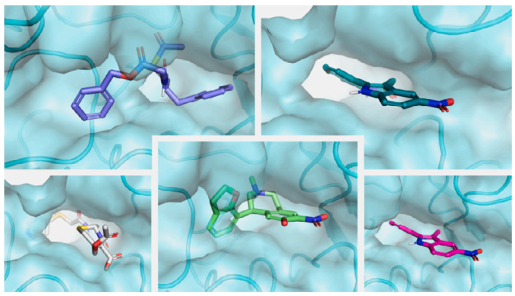

**Fig. 8.** Docking confirmation of top five screened compounds in the active site of $\beta$-glucuronidase.

with the amino group (NH) of the ligands, while TYR472 interacted with the ligands' oxygen atoms, as depicted in Fig. 8. These types of bonds are of utmost significance in the binding between small molecules and target proteins.

## 4. Conclusion

Recent studies have revealed the intricate and bidirectional relationship between the human microbiome and pharmaceutical compounds. Importantly, the transformation of drugs into inactive metabolites by the intestinal microbiome makes them toxic and also affects their pharmacokinetic and pharmacodynamic properties. In this study, graph neural network models have been developed to predict the drugs' depletion by intestinal microbiome. For this purpose, a dataset of 429 compounds was extracted from the literature and split into training and testing datasets, later the SMILE enumeration technique was applied to the training dataset to get the optimal performance of models. Models were trained into graph-based features of the compounds and performance scores were evaluated. The performance comparison of three GNN models with different learning parameters was compared and a GINCOV model was selected based on its high performance. The best model has achieved good performance with an accuracy: 91%, sensitivity: 82%, and specificity: 94% on the training dataset. Subsequently, the performance of the model on the testing dataset is also better than previously proposed models with accuracy: 93%, sensitivity: 86%, specificity: 96% showed that our proposed model have good generalization and stability. The key findings reveal that the rich representation of molecular interaction and ability to model complex relationship are crucial for accurate predictions. The proposed model accurately classifies the drug to be depleted or non-depleted by intestinal microbiota and could be used in the early stage of drug development for screening the drugs' susceptibility to the microbiome. For instance, the performance of the model to predict drug microbial depletion is likely to be more refined with an increased number of data samples in future research. Overall, this study underscores the potential of GNNs in biomedical research, setting a precedent for future studies in this field.

The datasets, the source code, and the pre-trained models are available on GitHub at https://github.com/MaryamRasoolSatti/drug_microbiome_interaction_prediction.

## Funding

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C2005612).

## CRediT authorship contribution statement

**Maryam:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization. **Mobeen Ur Rehman:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Irfan Hussain:** Writing – review & editing, Validation, Conceptualization. **Hilal Tayara:** Writing – review & editing, Visualization, Supervision, Methodology, Formal analysis, Conceptualization. **Kil To Chong:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108729.

## References


[1] Structure, function and diversity of the healthy human microbiome, Nature 486 (7402) (2012) 207–214.

[2] M. Ventura, S. O'flaherty, M.J. Claesson, F. Turroni, T.R. Klaenhammer, D. Van Sinderen, P.W. O'toole, Genome-scale analyses of health-promoting bacteria: probiogenomics, Nat. Rev. Microbiol. 7 (1) (2009) 61–71.

[3] F. Sommer, F. Bäckhed, The gut microbiota—masters of host development and physiology, Nat. Rev. Microbiol. 11 (4) (2013) 227–238.

[4] A.L. Kau, P.P. Ahern, N.W. Griffin, A.L. Goodman, J.I. Gordon, Human nutrition, the gut microbiome and the immune system, Nature 474 (7351) (2011) 327–336.

[5] R.E. Ley, P.J. Turnbaugh, S. Klein, J.I. Gordon, Human gut microbes associated with obesity, Nature 444 (7122) (2006) 1022–1023.

[6] J. Durack, S.V. Lynch, The gut microbiome: Relationships with disease and opportunities for therapy, J. Exp. Med. 216 (1) (2019) 20–40.

[7] R.F. Schwabe, C. Jobin, The microbiome and cancer, Nat. Rev. Cancer 13 (11) (2013) 800–812.

[8] L.E. McCoubrey, S. Gaisford, M. Orlu, A.W. Basit, Predicting drug-microbiome interactions with machine learning, Biotechnol. Adv. 54 (2022) 107797.

[9] M. Zimmermann, M. Zimmermann-Kogadeeva, R. Wegmann, A.L. Goodman, Mapping human microbiome drug metabolism by gut bacteria and their genes, Nature 570 (7762) (2019) 462–467.

[10] Y. Luo, P. Wang, M. Mou, H. Zheng, J. Hong, L. Tao, F. Zhu, A novel strategy for designing the magic shotguns for distantly related target pairs, Brief. Bioinform. 24 (1) (2023) bbac621.

[11] W. Xue, T. Fu, S. Deng, F. Yang, J. Yang, F. Zhu, Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram, ACS Chem. Neurosci. 13 (3) (2022) 340–351.

[12] W. Xue, F. Yang, P. Wang, G. Zheng, Y. Chen, X. Yao, F. Zhu, What contributes to serotonin–norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation, ACS Chem. Neurosci. 9 (5) (2018) 1128–1140.

[13] J. Yin, H. Zhang, X. Sun, N. You, M. Mou, M. Lu, Z. Pan, F. Li, H. Li, S. Zeng, et al., Decoding drug response with structurized gridding map-based cell representation, IEEE J. Biomed. Health Inf. (2023).

[14] B. Javdan, J.G. Lopez, P. Chankhamjon, Y.-C.J. Lee, R. Hull, Q. Wu, X. Wang, S. Chatterjee, M.S. Donia, Personalized mapping of drug metabolism by the human gut microbiome, Cell 181 (7) (2020) 1661–1679.

[15] L. Maier, M. Pruteanu, M. Kuhn, G. Zeller, A. Telzerow, E.E. Anderson, A.R. Brochado, K.C. Fernandez, H. Dose, H. Mori, et al., Extensive impact of non-antibiotic drugs on human gut bacteria, Nature 555 (7698) (2018) 623–628.

[16] C. Panebianco, A. Andriulli, V. Pazienza, Pharmacomicrobiomics: Exploiting the drug-microbiota interactions in anticancer therapies, Microbiome 6 (1) (2018) 1–13.

[17] I.D. Wilson, J.K. Nicholson, Gut microbiome interactions with drug metabolism, efficacy, and toxicity, Transl. Res. 179 (2017) 204–222.

[18] S.A. Flowers, S. Bhat, J.C. Lee, Potential implications of gut microbiota in drug pharmacokinetics and bioavailability, Pharmacother.: J. Hum. Pharmacol. Drug Ther. 40 (7) (2020) 704–712.

[19] M. Klünemann, S. Andrejev, S. Blasche, A. Mateus, P. Phapale, S. Devendran, J. Vappiani, B. Simon, T.A. Scott, E. Kafkia, et al., Bioaccumulation of therapeutic drugs by human gut bacteria, Nature 597 (7877) (2021) 533–538.

[20] Y. Guo, C.M. Crnkovic, K.-J. Won, X. Yang, J.R. Lee, J. Orjala, H. Lee, H. Jeong, Commensal gut bacteria convert the immunosuppressant tacrolimus to less potent metabolites, Drug Metab. Dispos. 47 (3) (2019) 194–202.

[21] J.R. Lee, T. Muthukumar, D. Dadhania, Y. Taur, R.R. Jenq, N.C. Toussaint, L. Ling, E. Pamer, M. Suthanthiran, Gut microbiota and tacrolimus dosing in kidney transplantation, PLoS One 10 (3) (2015) e0122399.

[22] H.J. Haiser, K.L. Seim, E.P. Balskus, P.J. Turnbaugh, Mechanistic insight into digoxin inactivation by Eggerthella lenta augments our understanding of its pharmacokinetics, Gut Microbes 5 (2) (2014) 233–238.

[23] R. Hitchings, L. Kelly, Predicting and understanding the human microbiome's impact on pharmacology, Trends Pharmacol. Sci. 40 (7) (2019) 495–505.

[24] L.F. Mager, R. Burkhard, N. Pett, N.C. Cooke, K. Brown, H. Ramay, S. Paik, J. Stagg, R.A. Groves, M. Gallo, et al., Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy, Science 369 (6510) (2020) 1481–1489.

[25] P. Timmerman, S. Blech, S. White, M. Green, C. Delatour, S. McDougall, G. Mannens, J. Smeraglia, S. Williams, G. Young, Best practices for metabolite quantification in drug development: updated recommendation from the European bioanalysis forum, Bioanalysis 8 (12) (2016) 1297–1305.

[26] V. Yadav, S. Gaisford, H.A. Merchant, A.W. Basit, Colonic bacterial metabolism of corticosteroids, Int. J. Pharmaceut. 457 (1) (2013) 268–274.

[27] L.E. McCoubrey, M. Elbadawi, M. Orlu, S. Gaisford, A.W. Basit, Machine learning uncovers adverse drug effects on intestinal bacteria, Pharmaceutics 13 (7) (2021) 1026.

[28] N. Koppel, V. Maini Rekdal, E.P. Balskus, Chemical transformation of xenobiotics by the human gut microbiota, Science 356 (6344) (2017) eaag2770.

[29] W. Wang, Z. Ye, H. Gao, D. Ouyang, Computational pharmaceutics-A new paradigm of drug delivery, J. Control. Release 338 (2021) 119–136.

[30] M. Ahmadi, M.F. Nia, S. Asgarian, K. Danesh, E. Irankhah, A.G. Lonbar, A. Sharifi, Comparative analysis of segment anything model and U-Net for breast tumor detection in ultrasound and mammography images, 2023, arXiv preprint arXiv:2306.12510.

[31] M. Ahmadi, D. Javaheri, M. Khajavi, K. Danesh, J. Hur, A deeply supervised adaptable neural network for diagnosis and classification of Alzheimer's severity using multitask feature extraction, PLoS One 19 (3) (2024) e0297996.

[32] L. Zheng, S. Shi, M. Lu, P. Fang, Z. Pan, H. Zhang, Z. Zhou, H. Zhang, M. Mou, S. Huang, et al., AnnoPRO: A strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding, Genome Biol. 25 (1) (2024) 41.

[33] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research 6 (2023) 0240.

[34] Y. Wang, Z. Pan, M. Mou, W. Xia, H. Zhang, H. Zhang, J. Liu, L. Zheng, Y. Luo, H. Zheng, et al., A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder, Nucleic Acids Res. 51 (21) (2023) e110.

[35] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Brief. Bioinform. 21 (4) (2020) 1437–1447.

[36] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Brief. Bioinform. 21 (5) (2020) 1825–1836.

[37] M.M. Elmassry, S. Kim, B. Busby, Predicting drug-metagenome interactions: Variation in the microbial $\beta$-glucuronidase level in the human gut metagenomes, PLoS One 16 (1) (2021) e0244876.

[38] A.K. Sharma, S.K. Jaiswal, N. Chaudhary, V.K. Sharma, A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota.

[39] S. Zhang, H. Tong, J. Xu, R. Maciejewski, Graph convolutional networks: A comprehensive review, Comput. Soc. Netw. 6 (1) (2019) 1–23.

[40] L. Meng, J. Zhang, Isonn: Isomorphic neural network for graph representation learning and classification, 2019, arXiv preprint arXiv:1907.09495.

[41] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, Stat 1050 (20) (2017) 10–48550.

[42] Z. Coombes, V. Yadav, L.E. McCoubrey, C. Freire, A.W. Basit, R.S. Conlan, D. Gonzalez, Progestogens are metabolized by the gut microbiota: implications for colonic drug delivery, Pharmaceutics 12 (8) (2020) 760.

[43] V. Yadav, Y. Mai, L.E. McCoubrey, Y. Wada, M. Tomioka, S. Kawata, S. Charde, A.W. Basit, 5-aminolevulinic acid as a novel therapeutic for inflammatory bowel disease, Biomedicines 9 (5) (2021) 578.

[44] T. Sousa, V. Yadav, V. Zann, A. Borde, B. Abrahamsson, A.W. Basit, On the colonic bacterial metabolism of azo-bonded prodrugsof 5-aminosalicylic acid, J. Pharm. Sci. 103 (10) (2014) 3171–3175.

[45] G.B. Hatton, V. Yadav, A.W. Basit, H.A. Merchant, Animal farm: Considerations in animal gastrointestinal physiology and relevance to drug delivery in humans, J. Pharm. Sci. 104 (9) (2015) 2747–2776.

[46] C.-K. Wu, X.-C. Zhang, Z.-J. Yang, A.-P. Lu, T.-J. Hou, D.-S. Cao, Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules, Brief. Bioinform. 22 (6) (2021) bbab327.

[47] C. Li, J. Feng, S. Liu, J. Yao, et al., A novel molecular representation learning for molecular property prediction with a multiple SMILES-based augmentation, Comput. Intell. Neurosci. 2022 (2022).

[48] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, et al., Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism, J. Med. Chem. 63 (16) (2019) 8749–8760.

[49] L. Bao, Z. Wang, Z. Wu, H. Luo, J. Yu, Y. Kang, D. Cao, T. Hou, Kinome-wide polypharmacology profiling of small molecules by multi-task graph isomorphism network approach, Acta Pharm. Sin. B 13 (1) (2023) 54–67.

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[52] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? 2018, arXiv preprint arXiv:1810.00826.

[53] A. Mao, M. Mohri, Y. Zhong, Cross-entropy loss functions: Theoretical analysis and applications, 2023, arXiv preprint arXiv:2304.07288.

[54] Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constr. Approx. 26 (2007) 289–315.

[55] A. Baratloo, M. Hosseini, A. Negida, G. El Ashal, Part 1: simple definition and calculation of accuracy, sensitivity and specificity, 2015.

[56] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: An overview, 2020, arXiv preprint arXiv:2008.05756.

[57] S. Agarwal, R. Mehrotra, An overview of molecular docking, JSM Chem. 4 (2) (2016) 1024–1028.

[58] L.B. Silva, E.F. Ferreira, Maryam, J.M. Espejo-Román, G.V. Costa, J.V. Cruz, N.M. Kimani, J.S. Costa, J.A. Bittencourt, J.N. Cruz, et al., Galantamine based novel acetylcholinesterase enzyme inhibitors: A molecular modeling design approach, Molecules 28 (3) (2023) 1035.

[59] R.S. Bastos, L.R. de Lima, M.F. Neto, Maryam, N. Yousaf, J.N. Cruz, J.M. Campos, N.M. Kimani, R.S. Ramos, C.B. Santos, Design and identification of inhibitors for the spike-ACE2 target of SARS-CoV-2, Int. J. Mol. Sci. 24 (10) (2023) 8814.

[60] X.-G. Tian, J.-K. Yan, C.-P. Sun, J.-X. Li, J. Ning, C. Wang, X.-K. Huo, W.-Y. Zhao, Z.-L. Yu, L. Feng, et al., Amentoflavone from selaginella tamariscina as a potent inhibitor of gut bacterial $\beta$-glucuronidase: Inhibition kinetics and molecular dynamics stimulation, Chem. Biol. Interact. 340 (2021) 109453.

[61] S. Gori, A. Inno, L. Belluomini, P. Bocus, Z. Bisoffi, A. Russo, G. Arcaro, Gut microbiota and cancer: How gut microbiota modulates activity, efficacy and toxicity of antitumoral therapy, Crit. Rev. Oncol. / Hematol. 143 (2019) 139–147.

[62] P. Singh, F. Bast, In silico molecular docking study of natural compounds on wild and mutated epidermal growth factor receptor, Med. Chem. Res. 23 (12) (2014) 5074–5085.

[63] N. Kausar, W.T. Shier, M. Ahmed, N.A. Albekairi, A. Alshammari, M. Saleem, M. Imran, M. Muddassar, et al., Investigation of the insecticidal potential of curcumin derivatives that target the helicoverpa armigera sterol carrier protein-2, Heliyon (2024).

[64] N. Yousaf, R.D. Alharthy, I. Kamal, M. Saleem, M. Muddassar, et al., Identification of human phosphoglycerate mutase 1 (PGAM1) inhibitors using hybrid virtual screening approaches, PeerJ 11 (2023) e14936.