Acta Pharmaceutica Sinica B 2023;13(1):54–67

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

Kinome-wide polypharmacology profiling of small molecules by multi-task graph isomorphism network approach

Lingjie Bao[a,†], Zhe Wang[a,†], Zhenxing Wu[a], Hao Luo[a], Jiahui Yu[a], Yu Kang[a,*], Dongsheng Cao[c,*], Tingjun Hou[a,b,*]

[a]*Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China*
[b]*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China*
[c]*Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, China*

Received 21 March 2022; received in revised form 15 April 2022; accepted 30 April 2022

KEY WORDS

Kinome-wide polypharmacology;
Machine learning;
Kinases;
Graph neural networks;
Artificial intelligence

Abstract Prediction of the interactions between small molecules and their targets play important roles in various applications of drug development, such as lead discovery, drug repurposing and elucidation of potential drug side effects. Therefore, a variety of machine learning-based models have been developed to predict these interactions. In this study, a model called auxiliary multi-task graph isomorphism network with uncertainty weighting (AMGU) was developed to predict the inhibitory activities of small molecules against 204 different kinases based on the multi-task Graph Isomorphism Network (MT-GIN) with the auxiliary learning and uncertainty weighting strategy. The calculation results illustrate that the AMGU model outperformed the descriptor-based models and state-of-the-art graph neural networks (GNN) models on the internal test set. Furthermore, it also exhibited much better performance on two external test sets, suggesting that the AMGU model has enhanced generalizability due to its great transfer learning capacity. Then, a naïve model-agnostic interpretable method for GNN called edges masking was devised to explain the underlying predictive mechanisms, and the consistency of the interpretability results for 5 typical epidermal growth factor receptor (EGFR) inhibitors with their structure–activity relationships could be observed. Finally, a free online web server called KIP was developed to predict the kinome-wide polypharmacology effects of small molecules (http://cadd.zju.edu.cn/kip).

*Corresponding authors. Tel./fax: +86 571 88208412.
E-mail addresses: tingjunhou@zju.edu.cn (Yu Kang), oriental-cds@163.com (Dongsheng Cao), yukang@zju.edu.cn (Tingjun Hou).
[†]These authors made equal contributions to this work.
Peer review under responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2022.05.004
2211-3835 © 2023 Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

© 2023 Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The human kinome contains around 500 kinases, accounting for roughly 1.7% of the whole human genome[1]. Protein kinases can catalyze the transfer of the terminal phosphate group of adenosine triphosphate (ATP) to substrate proteins, which plays a pivotal role in signal transductions and regulation of a wide range of cellular processes[2]. Moreover, aberrant kinase signaling has been linked to a variety of diseases, such as cancer, autoimmune disorders, diabetes, and neurological disorders[3]. The US Food and Drug Administration (FDA) has approved 71 small molecule kinase inhibitors by May 2021[4]. Despite the significant progress made in recent years, some challenges still need to be overcome in the field of kinase drug discovery. On the one hand, previous studies were focusing on only a small subset of human kinases, while most others were overlooked[5,6]. Thus, new techniques to reveal the activities of these understudied kinases and discovery of new small-molecule inhibitors for the treatment of associated complicated indications are urgently required[6,7]. On the other hand, most kinase inhibitors bind to the highly conserved ATP binding pockets in a competitive manner, which may lead to undesirable off-target effects[6,8]. Certainly, inhibiting numerous kinases at the same time may improve the efficacy of a kinase inhibitor and its ability to treat several types of cancers and many incurable diseases[5,6]. Hence, detecting the interactions between small molecules and kinase targets is quite critical to elucidate potential off-target effects, facilitate drug repurposing and discover new kinase inhibitors[9]

With the rapid accumulation of experimental bioactivity data for small molecules against kinases, many machine learning (ML)-based ligand-centric models have been developed to predict the kinome-wide polypharmacology of small molecules[10−16]. For example, in 2017, Merget et al.[17] developed a series of single-task ligand-based kinase inhibition classification models based on the connectivity-based and feature-based Morgan fingerprints for over 280 kinases using random forest (RF), naïve Bayes (NB), K-nearest neighbor (KNN), and deep neural network (DNN). The single-task RF models achieved the best performance with an average Area Under the Receiver Operating Characteristic curve (AUROC) of 0.76. In 2018, Sorin et al.[18] created the single-task RF models for 104 kinases based on the extended connectivity fingerprints (ECFPs) and pharmacophoric fingerprints (PFPs), which were then used to predict the inhibitory activities of small molecules against 104 kinases retrieved from ChEMBL. The models yielded good predictions with median sensitivity and specificity of higher than 0.8 for 90 kinase tasks and a median AUROC higher than 0.9 for 96 kinase tasks. Later, Janssen et al.[19] proposed the Drug Discovery Maps (DDM) method, which used the t-distributed stochastic neighbor embedding (t-SNE) algorithm to generate a visualization map of chemical similarity based on molecular fingerprints and biological similarity. DDM was also employed to find a novel inhibitor toward FMS-like tyrosine kinase 3 (FLT3), which was confirmed by biochemical assays. Recently, multi-task learning has attracted extensive attention and it is an inductive transfer approach that improves generalization by utilizing domain information contained in the training data of multiple related learning tasks as an inductive bias[20]. In 2019, Rodríguez-Pérez et al.[21] developed the single-task support vector machines (SVM), single-task RF, and multi-task DNN (MT-DNN) models based on the ECFPs and Molecular ACCess System (MACCS) fingerprints to predict the inhibitory activities of small molecules against 103 kinases. The MT-DNN models achieved the best prediction performance with a median Balanced Accuracy (BA) exceeding 0.8 and a median Matthews correlation coefficient (MCC) exceeding 0.75. Following that, Li et al.[22] developed the MT-DNN model based on the ECFPs to predict the inhibitory effects of small molecules against 391 kinases using the large-scale bioactivity data. In the internal test set, the model outperformed the standard single-task RF models with an AUROC of 0.90, especially for these kinases with limited activity data.

Despite the advances made in this field, there are still flaws and issues that need to be addressed. First, almost all models mentioned above were developed based on expert-crafted descriptors as molecular representation, which may not fully exploit the data's characteristics. As a novel form of deep learning (DL) algorithm, GNN can produce task-specific representation for molecules from data in an adaptable manner. In many molecular property prediction tasks, GNN models show better performances than descriptor-based models[23−27]. However, the GNN algorithms have never been utilized to predict protein kinase inhibition profiles. Therefore, it is quite valuable to explore the application of these state-of-the-art methods in predicting kinase inhibitory activities. Secondly, while multi-task learning has been effectively implemented in this field, previous researches have failed to account for task conflicts during the training process, which may result in inferior model performance[28,29]. To address these issues in multi-task learning, dynamic weighting strategies were suggested, but it is unknown whether these dynamic weighting algorithms are useful in multi-task learning for drug discovery. Third, the reported studies have never provided any interpretability to explain the underlying predictive mechanisms for the prediction models.

In this study, we proposed auxiliary multi-task graph isomorphism network with uncertainty weighting (AMGU), a new multi-task GNN model that can predict the inhibition profiles for small molecules against 204 kinases. As a comparison, four different descriptor-based models and five GNN-based models were also built. In both the internal and external testing, the AMGU model beat the other 9 models, highlighting its superiority in the prediction of kinase inhibition profiles. Compared with single-task models, the AMGU model could effectively enhance the performances of the separate tasks from related tasks, and the advantages were more noticeable for the tasks with fewer data. In addition, AMGU has the potential to uncover the relevance between the inhibition data of different kinases, which could aid in the discovery of "group-selective" kinase inhibitors. Moreover, we

proposed a naïve model-agnostic explanation method named edges masking to interpret the underlying predictive mechanisms behind the AMGU model. Finally, a web server for the kinome-wide polypharmacology profiling of small molecules was developed and freely accessible at http://cadd.zju.edu.cn/kip.

2. Material and methods

2.1. Dataset collection and preparation

The "Human and mouse protein kinases: classification and index" file was downloaded from the UniProt database (https://www.uniprot.org/docs/pkinfam) to extract a list of the specified UniProt identifiers of human kinases (organism: "*Homo sapiens*")[30]. The ChEMBL database (Release 27) was searched for the experimental bioactivity data (*i.e.*, IC_{50}, K_d and K_i) by querying the UniProt identifiers of human kinases on the ChEMBL website[15]. Five independent kinase assays were additionally collected, including the Davis dataset[12], Anastassiadis dataset[11], Metz dataset[14], Published Kinase Inhibitor Set (PKIS)[16], and Published Kinase Inhibitor Set 2 (PKIS 2)[13]. The inhibitory activity data points in the Anastassiadis dataset were transformed to IC_{50} values *via* the equation previously defined[31]. The dataset was then processed through the following steps to assure data quality:

(1) The organic molecules without biological activity records or clear chemical structures (SMILES string) and the inorganic compounds were removed. For each molecule, additional salts and solvents in the structure were removed using the Python script from Merget et al.[17]
(2) High-confidence biochemical assays were kept (confidence level $\geq$8, ensuring that there was a reported direct interaction between the ligand and its protein target), while the assays for mutated kinase targets were removed[32].
(3) For the classification tasks, a reasonable threshold of 1 μmol/L was defined to distinguish active and inactive compounds according to the previous study[22]. For the PKIS and PKIS 2 datasets, an inhibition rate over 50% at 1 μmol/L was defined as the active threshold[22]. The Davis dataset, Metz dataset, Anastassiadis dataset, and ChEMBL dataset were integrated as the training set. The compound-kinase pairs with both positive and negative labels in all datasets were eliminated due to data conflict. Moreover, the tasks with less than 40 positive and 40 negative samples were excluded from the training dataset to ensure data quality. The PKIS and PKIS 2 datasets were served as the external test sets for further evaluation. The external test datasets were also stripped of the compound-kinase pairs found in the training set. Furthermore, only the tasks on the external test sets with at least one positive and one negative sample were evaluated.
(4) For the regression tasks, the bioactivity data points with exact IC_{50} values were further extracted from the ChEMBL dataset and Anastassiadis dataset. The geometric mean of all the potency values was determined as the final potency annotation for any inhibitor that has multiple IC_{50} values for the same kinase[33]. At last, the negative logarithm values of IC_{50} of the bioactivity data points were recorded in the datasets. These bioactivity data for the regression tasks were also used in the auxiliary learning.

A summary for all the datasets can be seen in Table 1. The detailed information for each task in the dataset can be found in Supporting Information Table S1. All datasets can be accessed at http://cadd.zju.edu.cn/kip.

2.2. Molecule graph representation

In GNN, a molecule can be seen as a topological molecular graph with hydrogen-depleted nodes and edges, where nodes represent atoms and edges represent bonds. In this study, molecules were converted to molecular graphs and utilized as the inputs for the GNN-based models. As indicated in Tables 2 and 3, eight types of atom features and four types of bond features were used as the initial features of atoms and bonds. All atomic information was represented by the one-hot form except that the formal charge was in the integer form. All bond information was encoded in the one-hot form. According to the study reported by Kip et al.[34], self-connected undirected edges were added to the atoms for the GNN-based models except for the Directed Message Passing Neural Network (DMPNN) model. The DGL-LifeSci (version 0.2.5)[35] was used to transform molecules into bi-directed molecular graphs.

2.3. AMGU

AMGU is a multi-task GNN with the uncertainty weighting (UN) and the auxiliary learning strategy. The graph neural network layers employed in AMGU are Graph Isomorphism Network (GIN). The extra regression tasks functioned as the auxiliary learning components were incorporated and embedded in the last layer of the model to improve the generalization ability of the classification tasks. Then the uncertainty weighting strategy was utilized to resolve the potential conflicts between tasks by dynamically adjusting the weight for each task. The AMGU model is schematically depicted in Fig. 1.

Graph Isomorphism Network. Graph Isomorphism Network (GIN) is a simple graph neural network proposed by Xu et al[36]. The Xu's study illustrates that its discriminative/representational ability is equal to the power of the Weisfeiler-Lehman test[36]. As shown in Fig. 1A, the architecture of GIN can be divided into three different parts: (1) message-passing layer, (2) read-out layer, and (3) fully connected layers[36,37].

In the message-passing layer, the GIN model follows a recursive neighborhood aggregation scheme, where each node aggregates the feature vectors of its neighbors to form the new feature vector through nonlinear transformation. Specifically, for every node v, the node features h_v^l are updated as Eq. (1):

$$h_v^{l+1} = \mathrm{ReLU}\left(W_g^l \times \left(h_v^l + \sum_{u \in N(v)} h_u^l\right) + b_g^l\right) \quad (1)$$

Table 1 Summary of the datasets.

Dataset	Classification				Regression	
	Total	Positive	Negative	Task	Total	Task
Training set	260,183	116,515	143,668	204	122,676	202
PKIS	46,179	2758	43,421	131	/	/
PKIS 2	116,052	10,509	105,543	186	/	/

Table 2 The atom initial features used in GNN.

Atom feature	Size	Description
Atom symbol	43	[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, H, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb] (one-hot)
Atom degree	11	Number of covalent bonds [0,1,2,3,4,5,6,7,8,9,10] (one-hot)
Implicit valence	6	The implicit valence of an atom [1,2,3,4,5,6] (one-hot)
Hydrogens	7	The number of implicit Hs on the atom [0,1,2,3,4,5,6] (one-hot)
Atom Hybridization	5	[SP, SP2, SP3, SP3D, SP3D2] (one-hot)
Aromaticity	1	Whether this atom is part of an aromatic system [0/1] (one-hot)
Formal charge	1	−2−2 (integer)
Chirality	4	The chirality type of an atom [unspecified, tetrahedral CW, CCW, or other] (one-hot)

where h_v^{l+1} is the node features of node ν after $l+1$ iterations, $N(v)$ denotes the set of the neighbors of node ν, W_g^l and b_g^l are the weight and bias, respectively, and ReLU denotes the rectified linear activation function. After k iterations, the node features vector h_v^k captures the structural information within the node's k-hop network neighborhoods. Followed by the message-passing layer, a permutation invariant function as the readout function is designed to aggregate all features h_v^k in the final iteration into the graph embedding h_G of the entire graph G. Here, a summation function was used to directly aggregate the node features to gain the graph embedding h_G as Eq. (2):

$$h_G = \sum_{v \in G} h_v^l \tag{2}$$

where G denotes the whole molecular graph.

Subsequently, the graph embedding h_G is fed to the fully connected layers and undertakes the nonlinear transformation as follows:

$$h^{l+1} = \mathrm{ReLU}\left(h^l W^l + b^l\right) \tag{3}$$

where W^l is the weights of l^{th} layer in neural network, b^l is the bias of l^{th} layer in neural network, h^l denotes the input layers, and h^{l+1} denotes the output layer. Then the output of this layer is fed to the subsequent unit in the next layer and performs the same operation. The result of the final output layer is used as the solution for the problem[38]. For the regression task, the output is calculated as the same as Eq. (3) but without the ReLU activation function. While for the binary classification task, the sigmoid activation function is applied in the last layer to ensure getting the probability output in the range [0,1] for a particular task as Eq. (4):

$$h^{l+1} = \sigma\left(h^l W^l + b^l\right) \tag{4}$$

Table 3 The initial edge features used in GNN.

Edge feature	Size	Description
Bond type	4	[single, double, triple, aromatic] (one-hot)
Conjugation	1	Whether the bond is conjugated [0/1] (one-hot)
Ring	1	Whether the bond is part of a ring [0/1] (one-hot)
Stereo	6	[none, any, *E*/*Z* or *cis*/*trans*] (one-hot)

where σ denotes the sigmoid activation function. Finally, a loss function is used to compute the losses between the neural network outputs and true labels, which is used to guide the updating of neural network parameters. The loss function for each task is the weighted binary cross entropy, which imposes a higher penalty for the misclassification of the minority class and aids in the development of a discriminative model to handle the imbalanced distribution of the active and inactive points in our datasets[39]. The expression is as Eq. (5):

$$L^j = \frac{1}{N^j}\sum_i -\frac{N_{\mathrm{neg}}^j}{N^j} y^{i,j}\log(\hat{y}^{i,j}) - \frac{N_{\mathrm{pos}}^j}{N^j}(1-y^{i,j})\log(1-\hat{y}^{i,j}) \tag{5}$$

where $\hat{y}^{i,j}$ is the output of the model for sample i in task j, $y^{i,j}$ is the ground-truth value for sample i in task j, N^j denotes the number of the samples in task j, N_{neg}^j is the number of the negative samples in task j and N_{pos}^j is the number of the positive samples in task j.

For single-task Graph Isomorphism Network (ST-GIN), the weighted binary cross entropy was used as the loss function and ST-GIN only outputted one prediction value. While for multi-task Graph Isomorphism Network (MT-GIN), the loss function and network structure were slightly different from that for ST-GIN. In the MT-GIN, the hard-parameter sharing architecture was used (Fig. 1A)[40]. The shared chunk, which was the layer preceding the last layer in the network, shared parameters between tasks. The last layer in MT-GIN, on the other hand, had its own set of task-specific parameters and could predict several targets at once. The loss function of every task was set as mentioned above Eq. (5) and the final loss of all tasks was commonly set to be an average of the single tasks' losses L^j. The loss function of multi-task is as Eq. (6):

$$L_{\mathrm{all}} = \frac{1}{C}\sum_{j=0}^{C} w^j \cdot L^j \tag{6}$$

where C is the number of total tasks, and w^j is the task-specific weight for task j which is set to be uniform in common multi-task learning models.

In addition to the previously mentioned primary components in the architecture of neural networks, the batch norm (BN) and dropout sections were added and shown in Fig. 1B. The batch norm component was used to speed up and improve the stability of neural networks[41]. The dropout portion was employed to keep the neural networks from overfitting[42].

Uncertainty weighting (UN). Uncertainty weighting is a dynamic weighting strategy that can adaptively change task-specific weights $w^j(t)$ during the training process to balance these potential conflicts and it may achieve high performance. As shown in Eq.

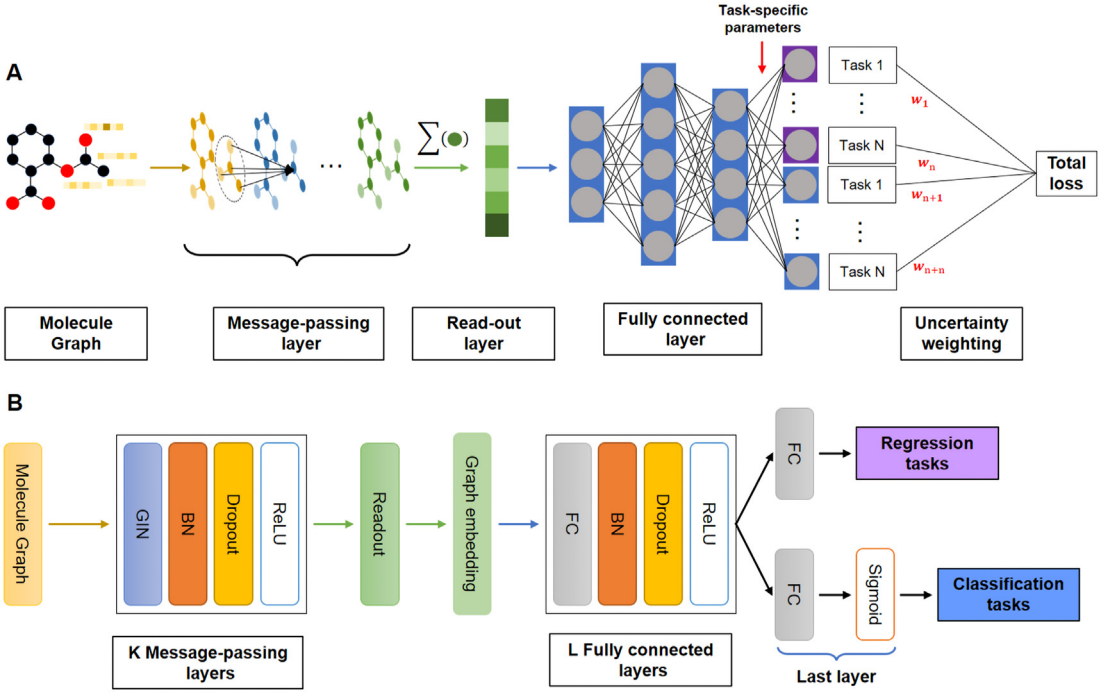

Figure 1 (A) The schematic overview of the AMGU model. The molecule is first represented by its initial atom and edge features and then fed into the AMGU model. The atom and edge features are fed into the message-passing layer to transform their features from the previous adjacency layer. The outputs from the final message-passing layer are reduced to vectors by the readout function (summation here), which is then used for predicting the inhibitory activities of molecules towards different kinases *via* the stacked fully connected layers. The extra regression tasks (in purple) are served as the auxiliary learning component to improve the generalization of the classification tasks. The total loss of AMGU is the linear weighted average of the tasks with the task weights updated by the uncertainty weighting strategy. The task-specific parameters from the last layer of the model are retrieved to capture the relevance between tasks. (B) The details of the AMGU model. The GIN module is solely used to aggregate the data from the nearby nodes and perform linear transformations (without nonlinear activation function), while the ReLU module serves as the activation function in the network.

(6), the final loss function is often assumed to be a linear weighted average of the single tasks' losses L^j in multi-task learning. When using stochastic gradient descent to minimize this loss function, the network parameters in the shared layers W_{sh} are updated as Eq. (7):

$$W_{\text{sh}} = W_{\text{sh}} - \eta \frac{1}{C} \sum_{j=0}^{C} w^j \cdot \frac{\partial L^j}{\partial W_{\text{sh}}} \tag{7}$$

where W_{sh} denotes shared parameters and η denotes learning rate in the neural network. According to Eq. (5), we can draw the following conclusions that the network parameters updated may be suboptimal when the task gradients conflict, or dominated by one task when its gradient magnitude is much higher than those for the other tasks[29,43]. To deal with this difficulty, some approaches by setting the task-specific weights w^j in the loss[44]. However, searching for appropriate task-specific weights by hand is a difficult and expensive process. Hence, uncertainty weighting was proposed as a dynamic weighting strategy by Kendall et al.[44]. During the training process, the homoscedastic uncertainty was exploited to balance the task-specific weights optimally. The relative confidence between tasks can be captured using homoscedastic uncertainty as task-dependent uncertainty. Uncertainty weighting strategy would automatically assign high task-specific weights for the tasks with low homoscedastic uncertainty. In their study, the loss function of multi-task classification tasks can be written in Eq. (8):

$$L_{\text{all}} = \frac{1}{C} \sum_{j=0}^{C} \left(\frac{1}{\sigma_j^2} L^j + \log \sigma_j^2 \right) \tag{8}$$

where σ_j is j^{th} task's noise parameters, which is relevant to the task's homoscedastic uncertainty in task j. Here, σ_j, as a learnable parameter for task j, is updated through standard backpropagation in every batch and can essentially balance the task-specific losses during training process. Large value σ_j denotes high task's homoscedastic uncertainty hence reducing the task weight for task j. In all tasks, $1/\sigma_j^2$ was initialized initialize to 1, which meant that each task weight was uniform at first and then updated according to the homoscedastic uncertainty of tasks.

Auxiliary learning with regression tasks. Auxiliary learning is similar to multi-task learning but aims to improve the performance on some primary tasks[40]. The auxiliary module is optimized in tandem with the multi-task learning network during training, and it serves as additional regularization by applying an inductive bias to the shared layers. In the testing phase, only the original multi-task learning network was retained[45]. With the goal of improving the generalization of these classification tasks, the regression tasks were used as the auxiliary tasks, and a new model termed Aux-MT-GIN was constructed. The mean squared error loss (MSE) was used as the loss function and as shown in Eq. (9):

$$L^r = \frac{1}{N^r} \sum_i (y^{i,r} - \hat{y}^{i,r})^2 \tag{9}$$

where L^r denotes the loss of regression task r, N^r denotes the number of the samples in regression task r, $y^{i,r}$ is the true label and $\hat{y}^{i,r}$ is the predicted value of the model. The total loss of auxiliary learning can be calculated as Eq. (10):

$$L_{\text{all}} = \frac{1}{C}\sum_{j=1}^{C} L^j + \frac{1}{R}\sum_{r=1}^{R} L^r \tag{10}$$

where R denotes the number of the regression tasks.

Furthermore, to handle potential conflicts between the classification tasks and regression tasks, the uncertainty weighting strategy was introduced and a new model named AMGU was developed, and the total loss was recast as Eq. (11):

$$L_{\text{all}} = \frac{1}{C}\sum_{j=1}^{C}\left(\frac{1}{\sigma_j^2}L^j + \log\sigma_j^2\right) + \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{2\sigma_r^2}L^r + \log\sigma_r^2\right) \tag{11}$$

where σ_r denotes the homoscedastic uncertainty in regression task. All $\frac{1}{\sigma_j^2}$ and $\frac{1}{\sigma_j^2}$ are initialized as 1 at the beginning.

2.4. *Model construction and evaluation protocols*

For the classification task, the training data were split into the training set, validation set, and internal test set with the ratio of 8:1:1 by using the "random stratified shuffle split" strategy to ensure that the percentage of the samples for each class was approximately preserved in each subset. In each task, the data from the training set, validation set, and internal test set were blended to be used for the multi-task learning. For the regression tasks in the auxiliary learning, the regression labels were provided if the same compound-kinase pair can be found in the classification tasks otherwise removed from the training set, validation set and internal test set. The training set was used to train the model, the validation set was used to search for the best combination of hyperparameters, and the internal test set and two external test sets were further used to evaluate the performance of each model.

The AMGU model was developed by using PyTorch (version 1.5.0)[46] and Deep Graph Library package (version 0.6.0)[47]. The total loss of AMGU was the linear weighted average of the weighted cross entropy for every task with task weight updated according to the uncertainty weighting strategy. The Adam algorithm was used to optimize the parameters in the model and the task weights[48]. To avoid overfitting, an early stop was utilized with the patience of 20 based on the average AUROC of all tasks in the internal validation set. If there is no improvement on the average AUROC in 5 consecutive epochs, the learning rate halves. The maximum number of the training epochs was set to 500. The detailed hyperparameters of different models can refer to Supporting Information. The final results were given as the mean and standard deviation for all models, which were performed with the seeds ranging from 0 to 9. For all models, the open-source library scikit-optimize was used to search for hyperparameters based on the average AUROC for all tasks in the internal validation sets[49]. To achieve acceptable performance, the model was subjected to 100 trials of hyperparameters search.

Several binary classification evaluation metrics were used to evaluate the performance of the classification models, including accuracy, precision (P), recall (R), F1-measure (F1), Matthews correlation coefficient (MCC), balanced accuracy (BA), the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC)[50]. These metrics are defined as Eqs. (12)−(17):

$$\text{Accuracy} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{TP}+\text{FP}+\text{FN}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{14}$$

$$F1 = \frac{2\times\text{Recall}\times\text{Precision}}{(\text{Recall}+\text{Precision})} \tag{15}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{TN})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}} \tag{16}$$

$$\text{Balanced accuracy} = \frac{1}{2}\times\left(\frac{\text{TP}}{\text{TP}+\text{FN}}+\frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{17}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. The AUROC was usually used to illustrate the model's ability to discriminate between positive samples and negative samples. When the dataset is imbalanced, especially when there are very few positive samples, AUROC may provide an optimistic view towards the model and AUPRC may be a better choice than AUROC.

For the regression tasks, coefficient of determination (R^2) was used to evaluate the performance of these tasks and defined as follows for every task as Eq. (18):

$$R^2 = 1 - \frac{\sum(y^{i,r}-\hat{y}^{i,r})^2}{\sum(y^{i,r}-\overline{y^{i,r}})^2} \tag{18}$$

where $\overline{y^{i,r}}$ is the average of all $y^{i,r}$.

Pearson correlation coefficient (PCC) was used to evaluate the relevance between the task-specific parameters in the classification models, and it is defined as Eq. (19):

$$\text{PCC}_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{19}$$

where $\text{cov}(X,Y)$ is the covariance between X and Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Other four descriptor-based models and five GNN-based classification models were established and evaluated in the same way. The uncertainty weighting strategy was not applied in these models and the total loss was the average cross entropy for all tasks. The descriptor-based models include single-task random forest (ST-RF)[51], single-task extreme gradient boosting (ST-XGBoost)[52], single-task Deep Neural Networks (ST-DNN) and multi-task Deep Neural Networks (MT-DNN)[38], and the GNN-based models include single-task Graph Isomorphism Network (ST-GIN)[36], multi-task Graph Attention network (MT-GAT)[53], multi-task Attentive FP (MT-Attentive FP)[54], multi-task Directed Message Passing Neural Network (MT-DMPNN)[55] and multi-task Graph Isomorphism Network (MT-GIN). The detailed information of these ML models can be seen in Supporting Information.

2.5. *Interpretation for models*

Although DL has posed a considerable impact in chemistry, one important issue that cannot be overlooked is the lack of

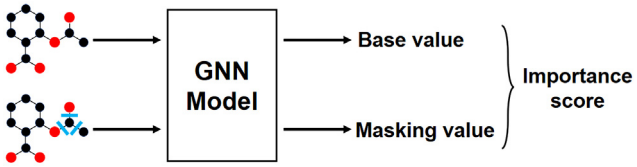

Figure 2 The computation method of the importance score for a specified atom.

interpretability in this field. Interpretability is essential because it guarantees our trust and transparency in the decision process of ML models[24]. Hence, we designed a strategy called edges masking to understand the results of the models in order to ensure that the conclusion derived from the model is rational. The edges masking was designed to offer the interpretability results for the molecular property prediction tasks in the GNN model, and it was inspired by the idea from this article[56].

As seen in Fig. 2, the concept behind this strategy is simple and straightforward. The importance score for each atom in the molecule was calculated as follows. First, an original molecular graph is fed into a GNN model to get a base prediction value. Then, by masking out some chemically significant substructures in the molecule, a modified molecular graph is created. Briefly, the edges around a specified atom are masked out to construct the modified molecule graph, which is roughly equivalent to masking out the related functional groups at the molecular level with some rationality in chemistry. The modified molecule graph is fed into the GNN model and it gets a masking prediction value. The Important score of this specified atom is the difference between the base prediction value and the masking prediction value as Eq. (20):

$$\text{Importance score}_{a_i} = \left(F(G) - F\left(G_i^{1-m_i}\right)\right) \tag{20}$$

where G represents the original molecule graph, a_i represents the i^{th} atom in the molecule, $G_i^{1-m_i}$ represents the modified molecule graph by masking out the edges around a_i, and F represents any GNN model. The importance score of each atom in a molecule is calculated. Then, by dividing by the largest absolute importance score value, these importance scores of all atoms in the molecule are normalized. The more important the contribution of the chemical environment around an atom to the model's output, the higher its importance score.

3. Result and discussion

3.1. *The overall performance of AMGU on 204 kinases*

The average AUROC and average AUPRC are used to assess the performances of AMGU and the other ML models. The performances of all the tested models are summarized in Table 4, and the detailed information of each model performance is presented in Supporting Information Table S2. We can observe that the

Table 4 The performances of all the models on the internal test set and external test sets[a].

Dataset	Model	AUROC	AUPRC[b]
Internal test set	ST-RF	0.8777 ± 0.0659	0.7974 ± 0.1621
	ST-XGB	0.8666 ± 0.0747	0.7831 ± 0.1719
	ST-DNN	0.8677 ± 0.0750	0.7891 ± 0.1649
	ST-GIN	0.8643 ± 0.0762	0.7723 ± 0.1766
	MT-DNN	0.9403 ± 0.0321	0.8849 ± 0.1032
	MT-Attentive_FP	0.9345 ± 0.0345	0.8778 ± 0.1059
	MT-DMPNN	0.9380 ± 0.0330	0.8831 ± 0.1026
	MT-GAT	0.9347 ± 0.0353	0.8788 ± 0.1033
	MT-GIN	0.9415 ± 0.0332	0.8879 ± 0.1005
	AMGU	**0.9425 ± 0.0321**	**0.8907 ± 0.1003**
PKIS	ST-RF	0.7699 ± 0.1150	0.2601 ± 0.2153
	ST-XGB	0.7611 ± 0.1131	0.2504 ± 0.1990
	ST-DNN	0.7842 ± 0.1110	0.2709 ± 0.1923
	ST-GIN	0.7820 ± 0.1176	0.2554 ± 0.1913
	MT-DNN	0.8323 ± 0.0900	0.3266 ± 0.1922
	MT-Attentive_FP	0.8431 ± 0.0833	0.3136 ± 0.1999
	MT-DMPNN	0.8526 ± 0.0820	0.3291 ± 0.2038
	MT-GAT	0.8387 ± 0.0779	0.3208 ± 0.1948
	MT-GIN	0.8455 ± 0.0769	0.3319 ± 0.2045
	AMGU	**0.8708 ± 0.0773**	**0.3773 ± 0.2097**
PKIS 2	ST-RF	0.7071 ± 0.0744	0.2352 ± 0.1290
	ST-XGB	0.6921 ± 0.0741	0.2273 ± 0.1240
	ST-DNN	0.6951 ± 0.0786	0.2379 ± 0.1300
	ST-GIN	0.7029 ± 0.0787	0.2273 ± 0.1270
	MT-DNN	0.7482 ± 0.0663	0.3015 ± 0.1463
	MT-Attentive_FP	0.7607 ± 0.0722	0.3039 ± 0.1440
	MT-DMPNN	0.7594 ± 0.0701	0.3090 ± 0.1503
	MT-GAT	0.7573 ± 0.0692	0.3040 ± 0.1442
	MT-GIN	0.7619 ± 0.0704	0.3188 ± 0.1539
	AMGU	**0.7796 ± 0.0741**	**0.3436 ± 0.1607**

[a]The maximum values of AUROC and AUPRC for different datasets are bold.
[b]The low AUPRC values in two external test sets due to low positive rate in these datasets.

Table 5 The performances of different methods on the different dataset in terms of main classification metrics[a].

Dataset	Metric	ST-GIN	MT-GIN	AMGU
Internal test set	AUROC	0.8643 ± 0.0762	0.9415 ± 0.0332	**0.9425 ± 0.0321**
	AUPRC	0.7723 ± 0.1766	0.8879 ± 0.1005	**0.8907 ± 0.1003**
	P	0.6795 ± 0.1878	**0.7888 ± 0.1348**	0.7849 ± 0.1401
	R	0.7613 ± 0.1293	0.8564 ± 0.0712	**0.8571 ± 0.0724**
	F1	0.7077 ± 0.1639	**0.8140 ± 0.1074**	**0.8122 ± 0.1109**
	Accuracy	0.8163 ± 0.0650	**0.8854 ± 0.0330**	0.8846 ± 0.0330
	BA	0.7930 ± 0.0785	0.8720 ± 0.0427	**0.8716 ± 0.0433**
	MCC	0.5653 ± 0.1653	**0.7222 ± 0.1002**	**0.7204 ± 0.1030**
PKIS	AUROC	0.7820 ± 0.1176	0.8455 ± 0.0769	**0.8708 ± 0.0773**
	AUPRC	0.2554 ± 0.1913	0.3319 ± 0.2045	**0.3773 ± 0.2097**
	P	0.1560 ± 0.1038	0.1732 ± 0.1137	**0.2020 ± 0.1249**
	R	0.6181 ± 0.2147	**0.7438 ± 0.1700**	0.7390 ± 0.1929
	F1	0.2303 ± 0.1243	0.2610 ± 0.1350	**0.2938 ± 0.1427**
	Accuracy	0.7874 ± 0.0942	0.7853 ± 0.0832	**0.8157 ± 0.0866**
	BA	0.7077 ± 0.1072	0.7659 ± 0.0832	**0.7797 ± 0.0926**
	MCC	0.2138 ± 0.1141	0.2673 ± 0.1050	**0.3007 ± 0.1185**
PKIS 2	AUROC	0.7029 ± 0.0787	0.7619 ± 0.0704	**0.7796 ± 0.0741**
	AUPRC	0.2273 ± 0.1270	0.3188 ± 0.1539	**0.3436 ± 0.1607**
	P	0.1974 ± 0.1125	0.2340 ± 0.1265	**0.2631 ± 0.1409**
	R	0.4745 ± 0.1555	**0.5822 ± 0.1519**	0.5693 ± 0.1682
	F1	0.2575 ± 0.1127	0.3133 ± 0.1278	**0.3346 ± 0.1329**
	Accuracy	0.7744 ± 0.0893	0.7904 ± 0.0730	**0.8139 ± 0.0757**
	BA	0.6381 ± 0.0658	0.6954 ± 0.0660	**0.7027 ± 0.0720**
	MCC	0.1859 ± 0.0970	0.2601 ± 0.1041	**0.2844 ± 0.1145**

[a]The maximum values of metrics for different datasets are bold.

Table 6 The comparison of the performances between our model and the Li's model.

Dataset	AMGU	Li's Model	*P* value[a]
PKIS	0.8724 ± 0.0777	0.8203 ± 0.1017	0.000
PKIS 2	0.7740 ± 0.0700	0.7073 ±0.0756	0.000

[a]The significance differences for the Mann−Whitney U test.

multi-task learning models significantly outperform the single-task learning models across all the three sets, with a 6.48%−10.97% absolute gap in the average AUROC and a 9.33%−12.69% absolute gap in the average AUPRC between the multi-task learning models and single-task learning models across the three test sets, demonstrating that the multi-task learning models are more superior for kinase inhibition prediction. In addition, the AMGU model outperforms the other multi-task learning models, with an average AUROC of 0.9425 and an average AUPRC of 0.8907 for the internal test set, an average AUROC of 0.8708 and an average AUPRC of 0.3773 for the PKIS dataset, and an average AUROC of 0.7796 and an average AUPRC of 0.3436 for the PKIS2 dataset. The AMGU model outperforms MT-GIN on both the internal and external test sets, demonstrating that the auxiliary learning and uncertainty weighting strategy have favorable contributions the multi-task learning in kinase inhibition prediction.

Table 7 The performances of the Auxiliary leaning.

Dataset	Model	AUROC	AUPRC
Internal test set	MT-GIN	0.9415 ± 0.0306	0.8877 ± 0.0967
	MT-GIN-UN	0.9384 ± 0.0320	0.8824 ± 0.1007
	Aux-MT-GIN	0.9379 ± 0.0329	0.8826 ± 0.1045
	AMGU	0.9425 ± 0.0321	0.8907 ± 0.1003
PKIS	MT-GIN	0.8455 ± 0.0769	0.3319 ± 0.2045
	MT-GIN-UN	0.8546 ± 0.0756	0.3436 ± 0.2043
	Aux-MT-GIN	0.8303 ± 0.0872	0.3182 ± 0.1966
	AMGU	0.8708 ± 0.0773	0.3773 ± 0.2097
PKIS 2	MT-GIN	0.7619 ± 0.0704	0.3188 ± 0.1539
	MT-GIN-UN	0.7661 ± 0.0701	0.3187 ± 0.1550
	Aux-MT-GIN	0.7522 ± 0.0686	0.2919 ± 0.1478
	AMGU	0.7796 ± 0.0741	0.3436 ± 0.1607

Some other important binary classification evaluation metrics were also utilized to evaluate the models, with the ST-GIN model and MT-GIN as the baseline to further verify the advantages of the AMGU model. The performances of AMGU, MT-GIN and ST-GIN on various datasets are listed in Table 5. When compared with the ST-GIN model, the performance of AMGU is always better than that of ST-GIN on the different datasets under the main classification metrics. When compared with the MT-GIN model, the AMGU model shows close or even slightly higher performance than the MT-GIN model in the internal test set. However, the main advantages of the AMGU model over the MT-GIN model are shown in two external datasets collected from the result of high throughput screening. It is noteworthy that all metrics except Recall of the AGMU model have been boosted. Compared with the MT-GIN model, more stringent criteria are taken in the AMGU model to distinguish positive samples from negative samples, which results in the decline of Recall and the boost of

Table 8 The task weights and training loss of all tasks in AMGU[a].

Task	Task weights	Training loss
Regression tasks	1.1299 ± 0.0266	0.0956 ± 0.0478
Classification tasks	2.3144 ± 0.0209	0.0616 ± 0.0212

[a]The task weights and training loss of all tasks are presented mean ± standard deviation.

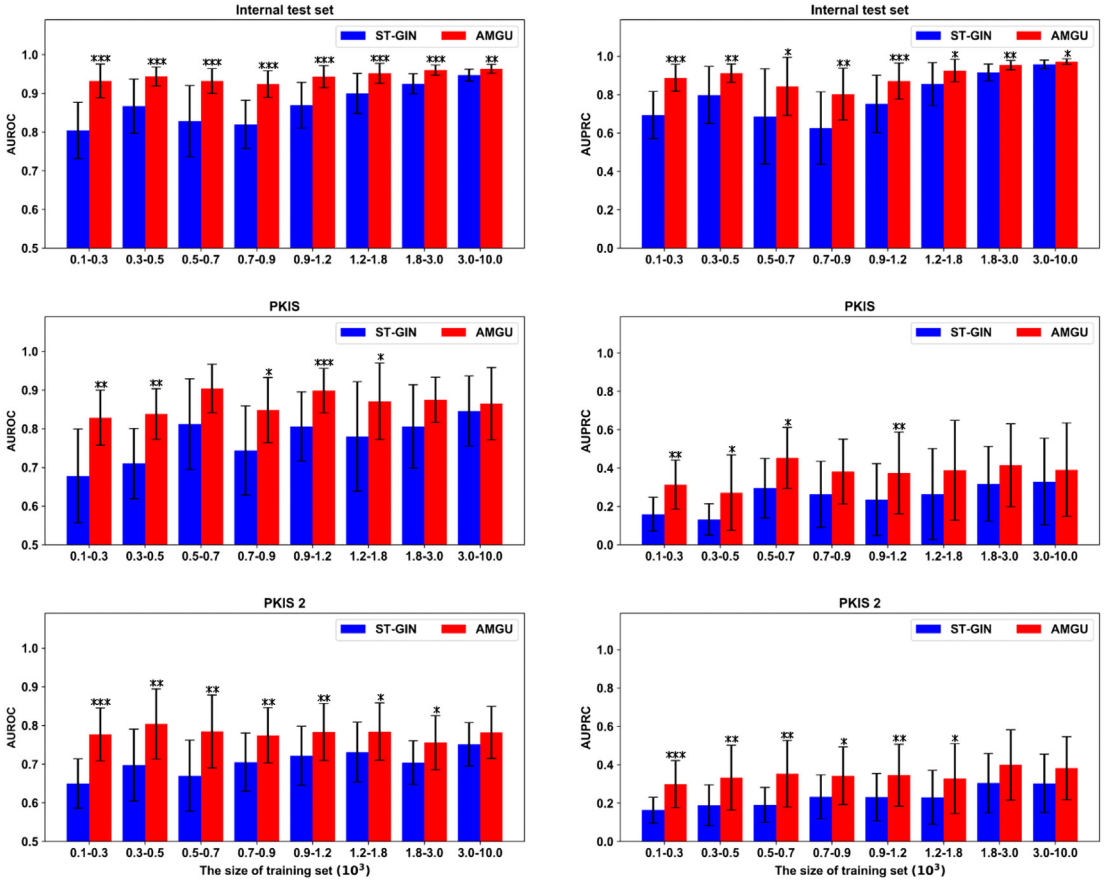

Figure 3 The performances of ST-GIN and AMGU on the tasks with different data volumes. A bar indicates the average AUROC or AUPRC of tasks with the number of bioactivity data points within the underlying range. The significance differences are under Mann−Whitney U test and shown in the figure according to the following standards: *$0.01 \leq P < 0.05$; **$0.001 \leq P < 0.01$; ***$P < 0.001$.

Precision. But taking overall improvement under the main classification metrics into consideration, it is acceptable for the lower value of Recall in the AMGU model than the MT-GIN model (but still higher than the ST-GIN model). Hence, the AMGU model is more suitable than the MT-GIN model for applications in real-world scenarios to some extent. These results further illustrate the advantages of the AMGU model.

Furthermore, we compared the performances of our model and the Li's model[22] for the same tasks on the PKIS and PKIS 2 datasets in terms of AUROC because only the metric of AUROC was reported by Li et al. There are 121 same targets in the PKIS dataset and 170 same targets in the PKIS 2 dataset. The average AUROC values of the same tasks for the AMGU and Li's models were calculated and compared. As shown in Table 6, on both PKIS and PKIS 2 datasets, our model performs better in those tasks. Furthermore, the AUROC values for the two separate external test sets show substantial difference between our model and Li's model, highlighting the superiority of our model. More details of the comparison between our model and Li's model can be found in Supporting Information Table S3.

3.2. The contribution of auxiliary learning and uncertainty weighting strategy

For kinase inhibition prediction, we also investigated the role of the auxiliary learning and uncertainty weighting strategies in the multi-task learning. For conducting the ablation experiments, the MT-GIN model with the auxiliary learning (Aux-MT-GIN) and the MT-GIN model with the uncertainty weighting strategy (MT-GIN-UN) were developed. The results of the ablation experiments are shown in Table 7. The overall performances of the different models on the two external test sets rank as follows: AMGU > MT-GIN-UN > MT-GIN > Aux-MT-GIN. When the auxiliary learning strategy was solely used with the regression tasks in MT-GIN, it (Aux-MT-GIN) degrades the performances of the classification tasks when compared with the original MT-GIN. This might be attributed to the fact that the regression tasks have larger-scale losses than the classification tasks, making the Aux-MT-GIN model pay more attention to regression tasks so that the auxiliary learning part in the model can't really work even downgrade the performance of original tasks. When using only the uncertainty weighting in MT-GIN, it (MT-GIN-UN) can achieve a minor improvement in the two independent external test sets, demonstrating the effectiveness of the uncertainty weighting in this dataset to some extent. When these two techniques were combined in MT-GIN to generate the AMGU model, it yields a minor improvement in all the test sets, with a maximum absolute improvement of 3.6% in AUPRC and 2.16% in AUROC when compared with the original MT-GIN model. Therefore, the improved performance of the AMGU model might be attributed to the combination of these two tactics. The losses and task weights in the classification and regression tasks are

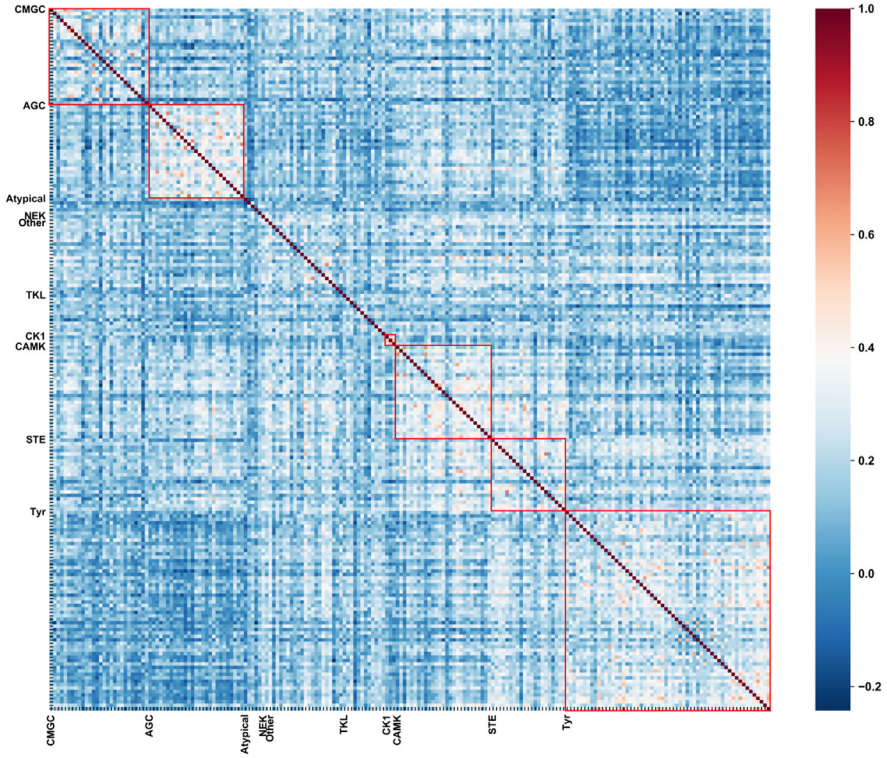

Figure 4 The Pearson correlation coefficient between different classification tasks. All tasks are organized by the kinase groups, which are based on the UniProt database's classification. The tasks against the same group begin with the name of the group and end with the name of the other group on the axis.

taken into account for further analysis, and their outcome is presented in Table 8. Due to the low homoscedastic uncertainty (training loss) in the classification tasks, the uncertainty weighting strategy might automatically balance any conflicts between the regression and classification tasks and provide larger weights to the classification tasks than the regression tasks. The classification task weights are nearly twice as high as the regression task weights, causing our model to pay more attention to the classification tasks and improving its performance on the classification tasks.

3.3. Tasks with small data volumes benefit more from AMGU model

As mentioned above, the AMGU model outperforms the ST-GIN model on nearly 90% of the tasks. Our results show that the tasks with small data quantities benefit more than those with large data volumes. Fig. 3 depicts the association between the training data size and the performances of the AMGU and ST-GIN models. Obviously, the AMGU model consistently outperforms the ST-GIN model due to the significant transfer learning effect of

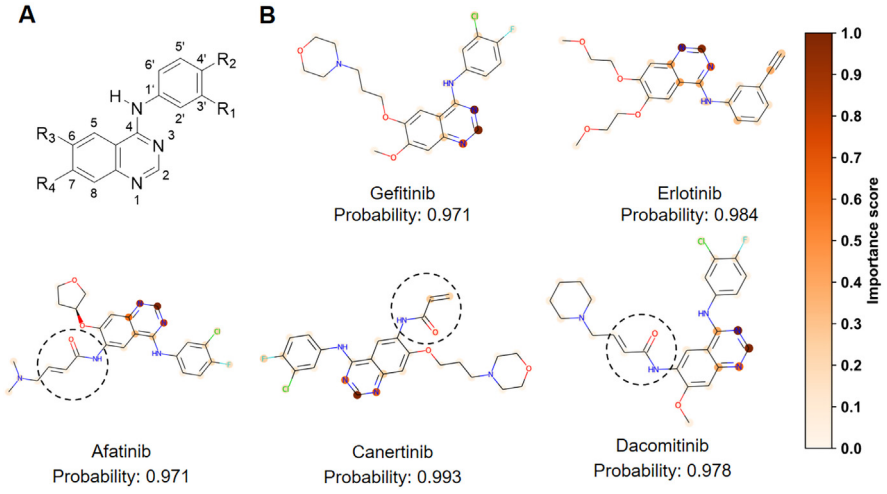

Figure 5 (A) The scaffold structure of 4-anilinoquinazolines derivatives. (B) The interpretability results of representative EGFR inhibitors. The chemical name and the model's anticipated value are annotated below the molecule structure. The first-generation and second-generation EGFR inhibitors are shown in the first row and second row, respectively, and the acrylamide moiety is highlighted with a black dotted circle.

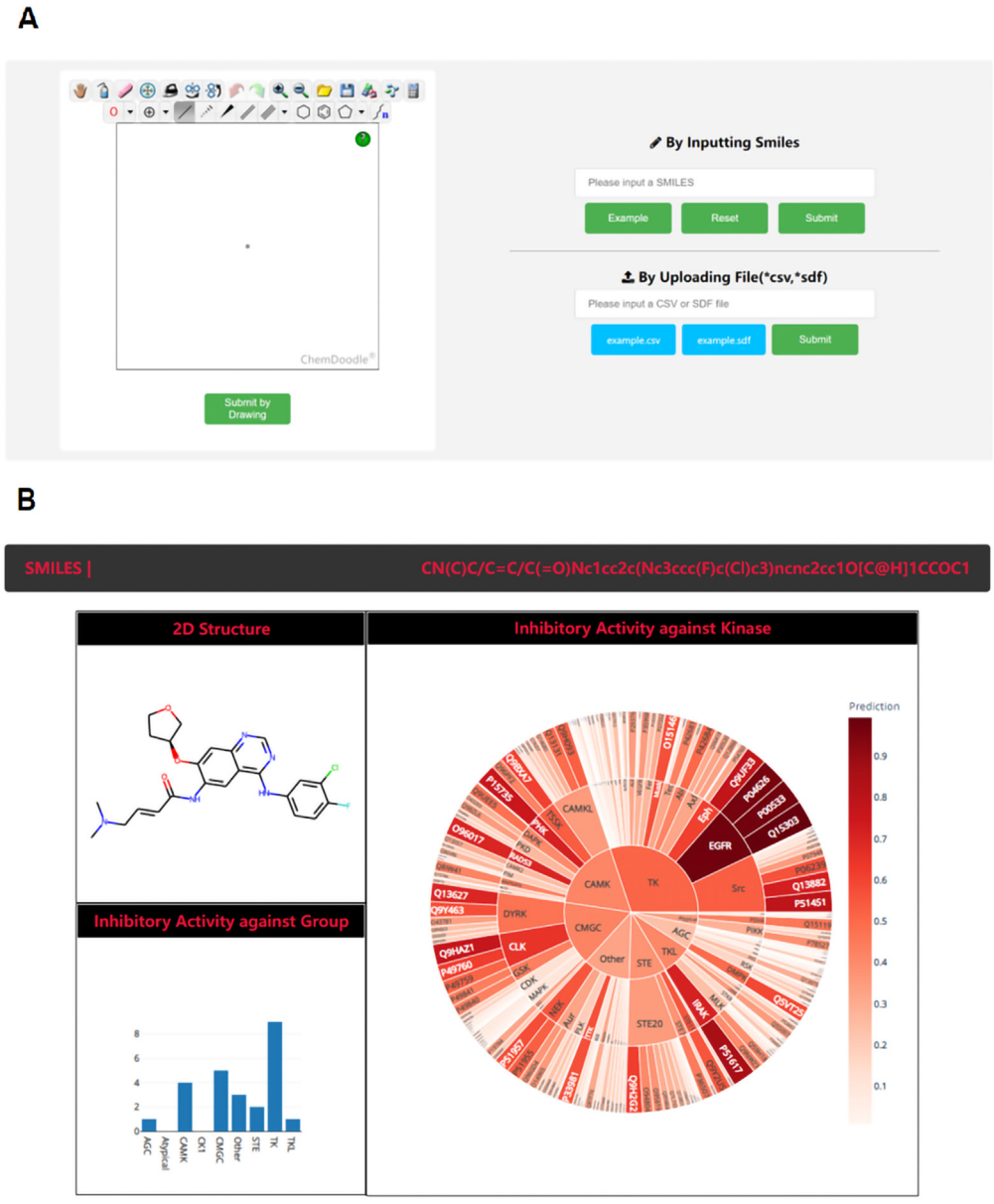

Figure 6 The illustrations of the KIP webservice. (A) Input page of the website; (B) Result page of the website.

multi-task learning[22,57], suggesting that related tasks can benefit from the AMGU model. In addition, as demonstrated in Fig. 3, the tasks with small data volumes benefit more from the AMGU model than those with large data volumes. It is quite challenging to build highly discriminative classifiers using single-task learning methods because some kinase targets have limited activity data. The AMGU model can exploit the domain information from related tasks and boost the performances of these tasks with limited data points, which may hopefully aid the discovery of new inhibitors for these understudied kinases.

3.4. AMGU can reveal the correlation between different kinases

Another advantage of our multi-task learning model is that it can recognize data characteristics and learn the underlying correlations between different kinases automatically. The parameters of the task-specific parameters of the classification tasks in the final fully-connected layer of the AMGU model (Fig. 1) were extracted independently to analyze the relevance between these tasks. Every task was represented by a vector with 1001 dimensions. The Pearson correlation coefficient for each vector pair was calculated. As illustrated in Fig. 4, the inhibitors within the same group tend to be clustered more closely than those outside the group, particularly for the CMGC, AGC, CK1, CAMK, and Tyr kinase groups (in red box). This result is consistent with the fact that there are some kinase inhibitors referred to as "group-selective inhibitors" which are broadly active against a single group of kinases but selective outside that group[12]. As a result, we can deduce that multi-task learning can conceptually capture the inherent characteristics of data.

3.5. *Interpretation for EGRF inhibitors*

In addition to the model's accuracy, it is quite important to interpret the underlying predictive mechanisms behind the AMGU model. Here, 4-anilinoquinazolines derivatives, one of the most important classes of EGFR inhibitors, were selected to interpret the prediction results of the EGFR inhibition prediction task for the AMGU model because their quantitative structure−activity relationships (QSAR) have been well-established[58,59]. The scaffold structure of 4-anilinoquinazolines is shown in Fig. 5A. As proven previously, the N_1 and N_3 atoms in 4-anilinoquinazolines play the most critical roles in suppressing the activity of EGFR[59].

The edges masking strategy was used to interpret the prediction results of the AMGU model. Two first-generation inhibitors (*i.e.*, gefitinib and erlotinib) and three second-generation inhibitors (*i.e.*, afatinib, canertinib, and dacomitinib) of EGFR were analyzed as the representative examples. Fig. 5B depicts the results of the interpretability test. For all these inhibitors, the edges masking interpretability approach highlights the atoms around N_1 and N_3, indicating the relevance of the chemical environment around these atoms contributing to the model's output. Moreover, the acrylamide moiety is the key component of the second-generation EGFR inhibitors, and it serves as an electrophilic warhead that conducts Michael addition with the conserved C797 residue in the EGFR active region[60]. The removal of this substructure from the molecule alters this type of inhibitors from covalent to non-covalent, but does not significantly change its inhibitory activity. Our model also identified these small variations in molecules as well, as seen in Fig. 5B, but did not give this component a high importance score.

Overall, the consistency of the interpretability results with the QSAR of 4-anilinoquinazolines demonstrates that our model has learned several critical molecular structures to some level, which raises our confidence to the model.

3.6. *Web server for the identification of kinase inhibitors*

To share our models with other chemists and pharmacologists, we developed a web server called Kinase Inhibition Prediction (KIP) (http://cadd.zju.edu.cn/kip) to profile the kinome-wide polypharmacology effects of small molecules (Fig. 6). The KIP, which was developed based on the Django framework, is freely available to non-commercial users. The web server is able to predict the biological activities towards kinases for small molecules, and it can also provide an interpretable explanation from the model to the positive outcome. Given the website's user-friendliness, the RDKit (Release 2019.09.1)[61] and Plotly (https://plotly.com/python/) were employed for the depiction of molecules and the visualization of results, respectively. The datasets utilized in this study and the trained models are also available on this website.

The workflow of KIP is as follows: (i) The client-side (browser) submits a query molecule for bioactivity prediction by sketching its structure from the ChemDoodle panel[62] or inputting its SMILES. Multiple-molecule Comma-Separated Values (CSV) or Structure Data File (SDF) files are also acceptable for submission. (ii) The kinome-wide inhibitory activities against 204 kinases are predicted at the server-side based on the AMGU model. (iii) The model's outputs are saved in a CSV file that can be downloaded. The visualization of the result and the explanation of the result are also available online for inspection.

4. Conclusions

In this study, the AMGU model based on the MT-GIN method combined with the auxiliary learning and uncertainty weighting strategy was proposed and used to simultaneously predict the inhibition profiles of small molecules against 204 kinases. The calculation results illustrate that the overall performance of the AMGU model is better than those of the other descriptor-based and GNN-based models, including ST-XGB, ST-RF, ST-DNN, MT-DNN, ST-GIN, MT-Attentive FP, MT-DMPNN, MT-GAT and MT-GIN on the test datasets. The ablation studies were carried out to further verify the effectiveness of the AMGU model. When both the auxiliary learning and uncertainty weighting strategy were integrated with the MT-GIN method, the corresponding models yielded better performance, suggesting that the combination of these two strategies jointly improved the performance of the model. The AMGU model can automatically assign higher task weights for the classification tasks due to low homoscedastic uncertainty in the classification tasks when analyzing the task weights and training losses of the classification and regression tasks. The AMGU mode has two advantages. One is that, due to the strong transfer learning ability, it can improve the generalizability for nearly all tasks, especially those with limited data. Another advantage is that our multi-task learning model can comprehend the data's characteristics and learn the potential correlation between tasks automatically to a certain extent. The model's correlation is based on the kinase group, which corresponds to the real-world situation where some "group-selective inhibitors" exist that are active against a single kinase group but selective outside of that group. Then, a simple model-agnostic interpretable strategy for GNN called edges masking was designed to understand the model's potential decision-making process. The consistency between the interpretability results of five representative EGFR inhibitors and their QSAR showed that our model could learn some key molecular structures. Finally, a freely accessible webserver named KIP was developed for the implementation of the well-trained models. Overall, our multi-task learning model makes large-scale kinome-wide virtual profiling of small molecular simple, which could help explain unwanted side effects, repurpose drugs, and discover new hit compounds.

Acknowledgments

This work was financially supported by National Key Research and Development Program of China (2021YFF1201400), National Natural Science Foundation of China (21575128, 81773632, 22173118), and Natural Science Foundation of Zhejiang Province (LZ19H300001, China), and Science and Technology Innovation Program of Hunan Province (2021RC4011, China).

Author contributions

Lingjie Bao and Zhe Wang designed and performed experiments, analyzed data, and wrote the paper. Zhenxing Wu, Hao Luo and Jiahui Yu performed some experiments. Yu Kang, Dongsheng Cao and Tingjun Hou initiated the study, organized, designed, and

wrote the paper. All authors read and approved the final manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Appendix A. Supporting information

Supporting data to this article can be found online at https://doi.org/10.1016/j.apsb.2022.05.004.

References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 2002;**298**:1912−34.
2. Cohen P. The origins of protein phosphorylation. Nat. *Cell Biol* 2002;**4**:E127−30.
3. Eglen RM, Reisine T. The current status of drug discovery against the human kinome. *Assay Drug Dev Technol* 2009;**7**:22−43.
4. Roskoski Jr R. Properties of FDA-approved small molecule protein kinase inhibitors: a 2021 update. *Pharmacol Res* 2021;**165**:105463.
5. Attwood MM, Fabbro D, Sokolov AV, Knapp S, Schioth HB. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nat Rev Drug Discov* 2021;**20**:839−61.
6. Wu P, Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol Sci* 2015;**36**:422−39.
7. Essegian D, Khurana R, Stathias V, Schürer SC. The clinical kinase index: a method to prioritize understudied kinases as drug targets for the treatment of cancer. *Cell Rep Med* 2020;**1**:100128.
8. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 2009;**9**:28−39.
9. Tucker JA, Martin MP. Recent advances in kinase drug discovery part I: the editors' take. *Int J Mol Sci* 2021;**22**:7560.
10. Ferrè F, Palmeri A, Helmer-Citterich M. Computational methods for analysis and inference of kinase/inhibitor relationships. *Front Genet* 2014;**5**:196.
11. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1039−45.
12. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046−51.
13. Drewry DH, Wells CI, Andrews DM, Angell R, Al-Ali H, Axtman AD, et al. Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PLoS One* 2017;**12**:e0181585.
14. Metz JT, Johnson EF, Soni NB, Merta PJ, Kifle L, Hajduk PJ. Navigating the kinome. *Nat Chem Biol* 2011;**7**:200−2.
15. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 2014;**42**:D1083−90.
16. Elkins JM, Fedele V, Szklarz M, Azeez KRA, Salah E, Mikolajczyk J, et al. Comprehensive characterization of the published kinase inhibitor set. *Nat Biotechnol* 2016;**34**:95−103.
17. Merget B, Turk S, Eid S, Rippmann F, Fulle S. Profiling prediction of kinase inhibitors: toward the virtual assay. *J Med Chem* 2017;**60**:474−85.
18. Avram S, Bora A, Halip L, Curpan R. Modeling kinase inhibition using highly confident data sets. *J Chem Inf Model* 2018;**58**:957−67.
19. Janssen APA, Grimm SH, Wijdeven RHM, Lenselink EB, Neefjes J, van Boeckel CAA, et al. Drug discovery maps, a machine learning model that visualizes and predicts kinome-ilnhibitor interaction landscapes. *J Chem Inf Model* 2019;**59**:1221−9.
20. Caruana R. Multitask learning. *Mach Learn* 1997;**28**:41−75.
21. Rodriguez-Perez R, Bajorath J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. *ACS Omega* 2019;**4**:4367−75.
22. Li X, Li Z, Wu X, Xiong Z, Yang T, Fu Z, et al. Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. *J Med Chem* 2020;**63**:8723−37.
23. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**:513−30.
24. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. *Adv Neural Inf Process Syst* 2019;**32**:9240−51.
25. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Cheminf* 2021;**13**:12.
26. Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining toxicity information from large amounts of toxicity data. *J Med Chem* 2021;**64**:6924−36.
27. Zhang XC, Wu CK, Yang ZJ, Wu ZX, Yi JC, Hsieh CY, et al. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings Bioinf* 2021;**22**:bbab152.
28. Vandenhende S, Georgoulis S, Van Gansbeke W, Proesmans M, Dai D, Van Gool L. Multi-task learning for dense prediction tasks: a survey. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:3614−33.
29. Sener O, Koltun V. Multi-task learning as multi-objective optimization. *arXiv* 2019:1810.04650.
30. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158−69.
31. Sutherland JJ, Gao C, Cahya S, Vieth M. What general conclusions can we draw from kinase profiling data sets?. *Biochim Biophys Acta* 2013;**1834**:1425−33.
32. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model* 2014;**54**:735−43.
33. Hu Y, Bajorath J. Exploring the scaffold universe of kinase inhibitors. *J Med Chem* 2015;**58**:315−32.
34. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv* 2016:02907.1609.
35. Li M, Zhou J, Hu J, Fan W, Zhang Y, Gu Y, et al. Dgl-lifesci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega* 2021;**6**:27233−8.
36. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks?. *arXiv* 2018:00826.1810.
37. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. *arXiv* 2017:01212.1704.
38. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019;**7**:53040−65.
39. Brownlee J. Imbalanced classification with python: choose better metrics, balance skewed classes, cost-sensitive learning. Available from: https://download.csdn.net/download/DomicZhong/19844813.
40. Ruder S. An overview of multi-task learning in deep neural networks. *arXiv* 2017:1706.05098.
41. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv* 2015:1502.03167.
42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929−58.
43. Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. Gradient normalization for adaptive loss balancing in deep multitask networks. *arXiv* 2018:1711.02257.
44. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv* 2018:07115v2.1705.
45. Liu Y, Zhuang B, Shen C, Chen H, Yin W. Auxiliary learning for deep multi-task learning. *arXiv* 2019:02214.1909.

46. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. *arXiv* 2019:01703.1912.
47. Wang M, Zheng D, Ye Z, Gan Q, Li M, Song X, et al. Deep graph library: a graph-centric, highly-performant package for graph neural networks. *arXiv* 2019:01315.1909.
48. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv* 2014:1412.6980.
49. Head T, Kumar M, Nahrstaedt H, Louppe G, Shcherbatyi I. Scikit-optimize (0.9.7). Zenodo. Available from: https://doi.org/10.5281/zenodo.6451894.
50. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett* 2006;**27**:861−74.
51. Breiman. Random forests. *Mach Learn* 2001;**45**:5−32.
52. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *arXiv* 2016:1603.02754v3.
53. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. *Stat* 2017;**1050**:20.
54. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**:8749−60.
55. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 2019;**59**:3370−88.
56. Riniker S, Landrum GA. Similarity maps—a visualization strategy for molecular fingerprints and machine-learning methods. *J Cheminf* 2013;**5**:43.
57. de la Vega de León A, Chen B, Gillet VJ. Effect of missing data on multitask prediction methods. *J Cheminf* 2018;**10**:1−12.
58. Kamath S, Buolamwini JK. Targeting EGFR and HER-2 receptor tyrosine kinases for cancer drug discovery and development. *Med Res Rev* 2006;**26**:569−94.
59. Ismail RSM, Ismail NSM, Abuserii S, Abou El Ella DA. Recent advances in 4-aminoquinazoline based scaffold derivatives targeting EGFR kinases as anticancer agents. *Future J Pharm Sci* 2016;**2**:9−19.
60. Ghosh AK, Samanta I, Mondal A, Liu WR. Covalent inhibition in drug discovery. *ChemMedChem* 2019;**14**:889−906.
61. Landrum G. RDKit: open-source cheminformatics from machine learning to chemical registration. Release 1 Mar 2014. Available from: https://doi.org/10.5281/zenodo.10398.
62. Burger MC. ChemDoodle Web Components: HTML5 toolkit for chemical graphics, interfaces, and informatics. *J Cheminf* 2015;**7**:1−7.