ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

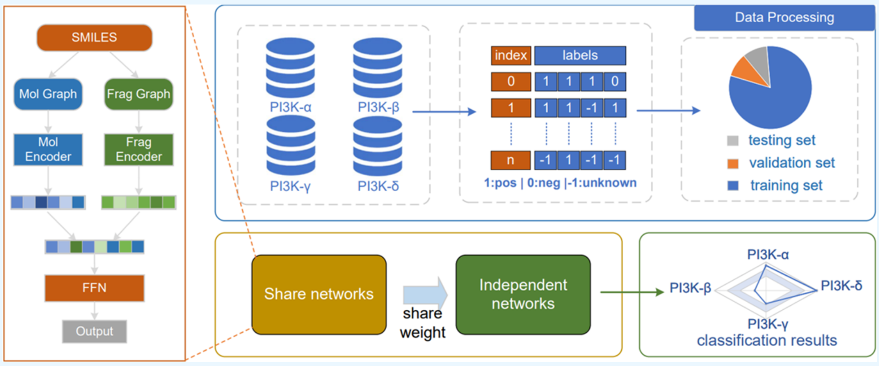
MVGNet: Prediction of PI3K Inhibitors Using Multitask Learning and Multiview Frameworks

Yanlei Kang, Qiwei Xia, Yunliang Jiang, and Zhong Li*

Cite This: *ACS Omega* 2024, 9, 45159−45168

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: PI3K (phosphatidylinositol 3-kinase) is an intracellular phosphatidylinositol kinase composed of a regulatory subunit, p85, and a catalytic subunit, p110. Based on the different structures of the p110 catalytic subunit, PI3K can be divided into four isoforms: PI3Kα, PI3Kβ, PI3Kγ, and PI3Kδ. As molecularly targeted drugs, PI3K inhibitors have demonstrated antiproliferative effects on tumor cells and can also induce cancer cell death. In this study, a multiview deep learning framework (MVGNet) is proposed, which integrates fragment-based pharmacophore information and utilizes multitask learning to capture correlation information between subtasks. This framework predicts the inhibitory activity of molecules against the four PI3K isoforms (PI3Kα, PI3Kβ, PI3Kγ, and PI3Kδ). Compared to baseline prediction models based on three traditional machine learning methods (RF, SVM, and XGBoost) and four deep learning algorithms (GAT, D-MPNN, CMPNN, and KANO), our model demonstrates superior performance. The evaluation results show that our model achieves the highest average AUC-ROC and AUC-PR values on the test set, which are 0.927 ± 0.006 and 0.980 ± 0.002, respectively. This study provides a reference for exploring the structure−activity relationship of PI3K inhibitors.

1. INTRODUCTION

PI3K, also known as phosphatidylinositol 3-kinase, catalyzes the phosphorylation of phosphatidylinositol molecules. It is divided into three categories: Class I, Class II, and Class III. For drug discovery and development, the primary focus lies on class I PI3K.[1−3] Within Class I PI3K, its active subunit p110 has four isoforms: α, β, γ, and δ. Among these, p110α and p110β are widely distributed in tissues, while p110γ and p110δ are highly enriched in leukocytes. PI3K inhibitors are promising small molecule drugs with antiproliferative effects on tumor cells and beneficial effects on tumor cells and the immune system leading to cancer cell death.[4−6] PI3Kα inhibitors exhibit promising results for breast cancer indications; PI3Kβ inhibitors are considered potentially effective for targeting PTEN-deficient tumors; PI3Kγ inhibitors show potential for controlling immune disorders; and PI3Kδ inhibitors are effective against hematological tumors. Based on their selectivity, currently identified PI3K inhibitors can be classified into three main groups: pan-PI3K inhibitors, PI3K isoform-selective inhibitors, and dual inhibitors.[7] Pan-PI3K inhibitors, such as copanlisib, act simultaneously on all four class I subtypes.[8] However, the lack of selectivity of Pan-PI3K inhibitors may lead to nonspecific inhibition of the entire pathway, leading to various side effects. To address this issue, PI3K subtype-selective inhibitors targeting specific subtype isoforms have been extensively studied. In addition, dual PI3K-mTOR inhibitors capable of inhibiting two isoforms, such as Pictilisib (GDC-0941), have been proposed.[9] However, despite the potential of PI3K inhibitors, there are very few PI3K inhibitor drugs currently on the market, mainly due to

Received: July 5, 2024
Revised: October 9, 2024
Accepted: October 15, 2024
Published: October 30, 2024

© 2024 The Authors. Published by
American Chemical Society

https://doi.org/10.1021/acsomega.4c06224

the high cost of translating drugs from basic science to early stage clinical trials. To develop new PI3K inhibitors, experts must screen a large number of candidate compounds for factors such as activity and toxicity. However, relying solely on chemical methods for this process results in high time and economic costs, extensive experimentation, and a high failure rate. Although PI3K inhibitors are widely recognized and used in clinical practice, their molecular properties, including toxicity, selectivity, and resistance, limit their clinical utility.[10−12]

Computational methods have been increasingly used to explore structure−activity relationships of PI3K inhibitors. For example, Kumar et al. characterized salvianolic acid A as a dual inhibitor of PI3K and mTOR using a structure-based computational approach.[13] Similarly, Das et al. identified PI3Kα inhibitors for the treatment of hepatocellular carcinoma through virtual screening and watermap analysis.[14] Zhu et al. developed a hybrid virtual screening (VS) approach combining ligand pharmacophore modeling and molecular docking with multiple PI3Kδ inhibitor complexes to target PI3Kδ proteins.[15] In addition, perturbation theory machine learning (PTML) models have been applied to dual-target/multitarget drug discovery. They can be categorized into multitarget QSAR, multiconditional QSAR, and multitasking models for quantitative structure-biological effect relationships (mtk-QSBER). For example, Speck-Planche et al. developed two different multitarget models based on quantitative conformational relationships (mt-QSAR) to predict multitarget BET bromodomain inhibitors.[16] Speck-Planche et al. also used multiconditional QSAR models to virtually design and predict molecules with dual pan-antiviral and anti-CS profiles.[17] In this study, a new model was developed for the design of molecules with multitarget activity. Kleandrova et al. combined perturbation theory and machine learning to construct a multilayer perceptron network, PTML-MLP, which was applied to the design of multiprotein and multicellular inhibitors in pancreatic cancer research.[18] These computational models and schemes have been instrumental in the discovery of new PI3K inhibitors.

However, identifying highly selective molecules for specific PI3K isoforms using structure-based VS methods or QSAR modeling is challenging. This difficulty arises because the binding active sites across the PI3K family exhibit high sequence homology and structural similarity.[19,20] To address this, multitask models are often used to predict compounds with high sequence homology and structural similarity. For example, Nguyen-Vo et al. developed the iCYP-MFE framework using multitask learning and molecular fingerprint embedding encoding to predict inhibitory activity for five CYP isoforms (1A2, 2C9, 2C19, 2D6, and 3A4).[21] In 2022, Ai et al. introduced a multitask FP-GNN framework that accurately predicted the inhibitory activity of molecules against four PARP isoforms (PARP-1, PARP-2, PARP-5A, and PARP-5B).[22] Recent advancements in machine learning have significantly impacted the field of chemistry and materials, particularly in drug property prediction. Due to the unique properties of 2D molecular graphs, these graphs have been widely used for drug property prediction. For example, Gilmer et al. designed various message-passing neural networks (MPNN) for molecular property prediction and achieved high accuracy.[23] Song et al. improved MPNNs by proposing a directed graph-based communicating message-passing neural network (CMPNN) that improved molecular graph embedding through interactive updates of edge and node embeddings, which greatly improved molecular property prediction.[24] The combination of molecular graphs with other molecular features is also popular in research. For example, Cai et al. combined molecular fingerprints and molecular graphs and proposed the FP-GNN model for molecular property prediction, which has a good performance in terms of noise immunity.[25] Zhu et al. combined molecular fragments and molecular graphs and designed a plug-and-play feature-level attention block to propose a hierarchical infographics neural network framework (HiGNN), which is capable of recognizing the key components of a molecule.[26] Li et al. proposed the FG-BERT model by combining functional groups and molecular graphs, which has a good performance in terms of noise immunity. Li et al. proposed the FG-BERT model by combining functional groups and molecular graphs. This is a self-supervised pretrained deep learning model by masking functional groups to learn more useful molecular representations in pretraining.[27] In addition, with the development of knowledge graphs, Fang et al. introduced KANO, a molecular property prediction method based on the knowledge graph of chemical elements and functional group hints. This method uses element-oriented knowledge graphs as a priori, designs element-guided graph expansions, and learns functional hints during fine-tuning to evoke relevant knowledge for downstream tasks.[28] Therefore, applying machine learning and deep learning methods to predict the bioactivity of candidate PI3K inhibitors can efficiently screen out compounds with substandard bioactivity. This approach effectively shortens the trial period, saves research and development funds, and improves the clinical translation rate.

In this study, a multiview PI3K inhibitor activity prediction model called MVGNet was constructed. This model is based on chemical reaction information and CMPNN, where the molecule is cut into multiple fragments using BRICS.[29] These fragments are considered as a whole, with the bonds between fragments serving as edges to form a new fragment view. The multiview design enables the model to efficiently extract pharmacophore and reaction information from molecular fragments based on chemical reactions. In addition, multitask learning is used to simultaneously predict four isoforms of PI3K inhibitors (PI3K-α, PI3K-β, PI3K-γ, and PI3K-δ). Meanwhile, in order to train and validate the performance of the model, a new PI3K inhibitor prediction dataset was generated based on data from online public databases, and data cleaning and preprocessing steps were applied.

2. MATERIALS AND METHODS

2.1. Dataset Collection and Preparation. The modeling dataset of PI3K inhibitors was mainly obtained from the CHEMBL database (version 33),[30] which is a large drug discovery database containing therapeutic targets and indications of clinical trial drugs and approved drugs. We collected four subtypes of inhibitors of PI3K p110 (including PI3K-α, PI3K-β, PI3K-γ, and PI3K-δ) acting on humans on the CHEMBL database, and then cleaned and annotated them, which were processed as follows: (1) Compounds with a clear bioassay value (assay type = B), such as IC_{50}, EC_{50}, K_d, or K_i, were retained. Compounds with null bioactivity data were discarded; (2) all units of bioactivity data (e.g., g/mL, M, nM, etc.) were converted to standard units of μM; (3) if a molecule has more than one bioactivity data, the average of them was taken as the final value; (4) duplicate molecules were removed; (5) the bioactivity values (e.g., pIC_{50},pEC_{50},pK_d and pK_i) $\leq$ 1 μM were labeled as active molecules, and vice versa as inactive molecules;

After the above steps, the final dataset contained 17622 unique compounds with 21701 bioactive data points involving four isoforms (i.e., PI3K-α, PI3K-β, PI3K-γ, PI3K-δ). In addition, we divided the data into two categories: shared data and distinct data. Shared data are compounds that are present in all four subtype data sets. All other relevant compounds except those that are shared data are referred to as distinct data. Among them, the number of shared samples (compounds present in all four subtype data sets) was 1894, all of which were used as the training set. For the unique sample data sets, each dataset is randomly divided into: training set, validation set and test set according to 8:1:1 respectively. Table 1 lists the total number of compounds and the number of active and inactive compounds in each PI3K isoform dataset. Figure 1 illustrates the percentage of active and inactive compounds in each PI3K isoform species, which is relatively balanced except for the PI3K-δ dataset with a data imbalance problem (which is difficult to solve in real-world drug discovery).

Table 1. PI3K Isoform Dataset

dataset	molecules	actives	inactive
PI3K-α	7055	4926	2129
PI3K-β	2732	1607	1125
PI3K-γ	3792	2292	1500
PI3K-δ	4043	3146	896

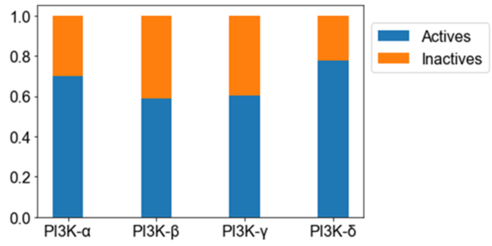


Figure 1. Distribution of PI3K isoform samples.

2.2. Methods. CMPNN is a directed graph-based communication message-passing neural network that improves molecular graph embedding by interactively updating edge and node embeddings. Although CMPNN has been greatly improved in the task of molecular property prediction, in the task of PI3K inhibitor activity prediction, considering that the binding active sites in the PI3K family exhibit a high degree of sequence homology and their structures also show similarity. The use of CMPNN model alone is not enough to learn more association information between substructures, while the use of multitask learning strategy can effectively solve this problem. In addition, CMPNN pays more attention to the connection between nodes and edges, and does not take into account the influence of substructures such as functional groups on molecular properties; in fact, the molecular properties of drugs are often highly correlated with substructures, and the inclusion of a priori knowledge can often help the model to learn more effective knowledge. Therefore, in this study, the molecular fragment view was introduced to partition the molecule into different substructures, and the pharmacophore information and reaction information on the substructures were extracted to serve as the initial features.

For these considerations, a multiview PI3K inhibitor activity prediction model (Figure 3) was constructed based on chemical reaction information and CMPNN using multitask learning as a way to predict the biological activity of four subtypes of PI3K inhibitors. First, CMPNN was used as a molecular map feature extractor for the model. BRICS is an algorithm for segmenting molecules based on chemical reaction templates that follow chemically sound rules.[29] Jiang et al. used BRICS to segment molecules and construct heterogeneous molecular maps to predict molecular properties.[31] BRICS is used to cut molecules into fragments while preserving reaction information on the edges. Next, a fragment view is constructed by treating each fragment as a different point on the graph and the bonds connecting the fragments as edges. Then, the features of the original molecular view and the fragment view obtained after the BRICS cut are extracted separately using CMPNN, and finally, the two features (molecular graph features and fragment graph features) are stitched together using a fully connected layer, and then the prediction results of compound activities are output.

It should be noted that for the fragment view, RDKIT's feature factory was used to extract fragment features for the pharmacophore of each fragment. Taking Figure 2 as an example, the molecular fragment view uses Molecular Access System (MACCS) molecular fingerprints (selected atomic features include hydrogen bond acceptor, hydrogen bond, and whether it has aromatic atoms or not, etc.), and extracts the pharmacophore information on the substructures and splices them with the MACCS molecular fingerprints as the initial feature vectors of the fragments:

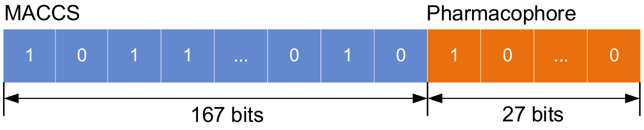


Figure 2. Fragment feature representation, splicing the 167-bit MACCS fingerprint with the 27-bit pharmacophore feature.

$$FP_{\text{Fragment}} = FP_{\text{MACCS}} || FP_{\text{Pharmacophore}} \tag{1}$$

The initial features for the bonds connecting the fragments were derived from the residual information obtained from the BRICS cut. The inclusion of the fragment view allowed us to efficiently obtain more compound-related a priori knowledge and thus more key features.

In addition, since the binding active sites in the PI3K family have a high degree of sequence homology and structural similarity, the PI3K inhibitor dataset contains important connections between subtasks. Using a single-task model during training would result in the loss of correlation information between these subtasks. This could lead to misclassification, especially when there are two structurally similar compounds. Therefore, multitask learning is critical for predicting the activity of PI3K inhibitors. By sharing parameters, the model can first learn the association information between subtasks using multitarget samples, and then fine-tune the model using single-target samples. This approach improves the efficiency of sample usage and increases the predictive accuracy of the model.

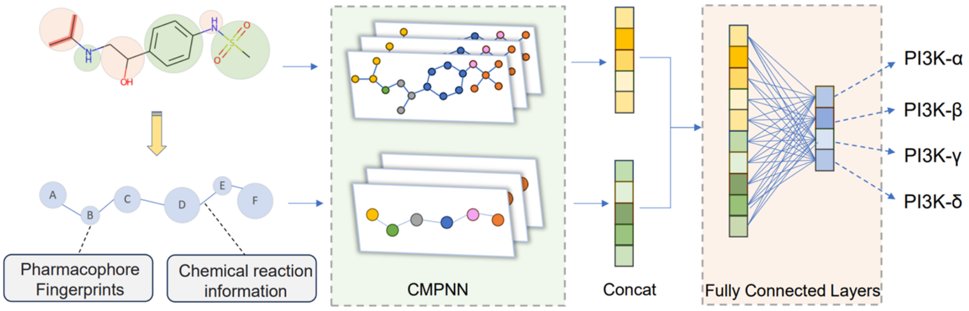


Figure 3. MVGNet basic framework diagram.

Table 2. Comprehensive Performance of the PI3K Inhibitor Dataset on Different Models[a]

model	AUC-ROC	AUC-PR	ACC	MCC	RE	SP	F1
RF	0.806 ± 0.022	0.852 ± 0.020	0.734 ± 0.025	0.430 ± 0.057	0.790 ± 0.042	0.622 ± 0.069	0.768 ± 0.025
XGBoost	0.809 ± 0.022	0.859 ± 0.020	0.736 ± 0.023	0.441 ± 0.049	0.779 ± 0.037	0.654 ± 0.049	0.768 ± 0.024
SVM	0.701 ± 0.028	0.761 ± 0.030	0.657 ± 0.021	0.257 ± 0.047	0.776 ± 0.026	0.459 ± 0.060	0.716 ± 0.020
GAT	0.746 ± 0.022	0.837 ± 0.027	0.719 ± 0.026	0.301 ± 0.043	0.850 ± 0.078	0.404 ± 0.117	0.794 ± 0.028
DMPNN	0.886 ± 0.016	0.929 ± 0.013	0.818 ± 0.017	0.576 ± 0.043	0.887 ± 0.036	0.664 ± 0.066	0.862 ± 0.015
CMPNN	0.894 ± 0.020	0.921 ± 0.019	0.810 ± 0.015	0.603 ± 0.030	0.864 ± 0.049	0.719 ± 0.058	0.834 ± 0.021
KANO	0.891 ± 0.020	0.915 ± 0.020	0.806 ± 0.022	0.597 ± 0.048	0.850 ± 0.050	**0.726 ± 0.076**	0.826 ± 0.023
MVGNet	**0.913 ± 0.011**	**0.949 ± 0.01**	**0.856 ± 0.012**	**0.659 ± 0.023**	**0.890 ± 0.027**	0.725 ± 0.048	**0.887 ± 0.013**

[a]The best-performing results are marked in bold.

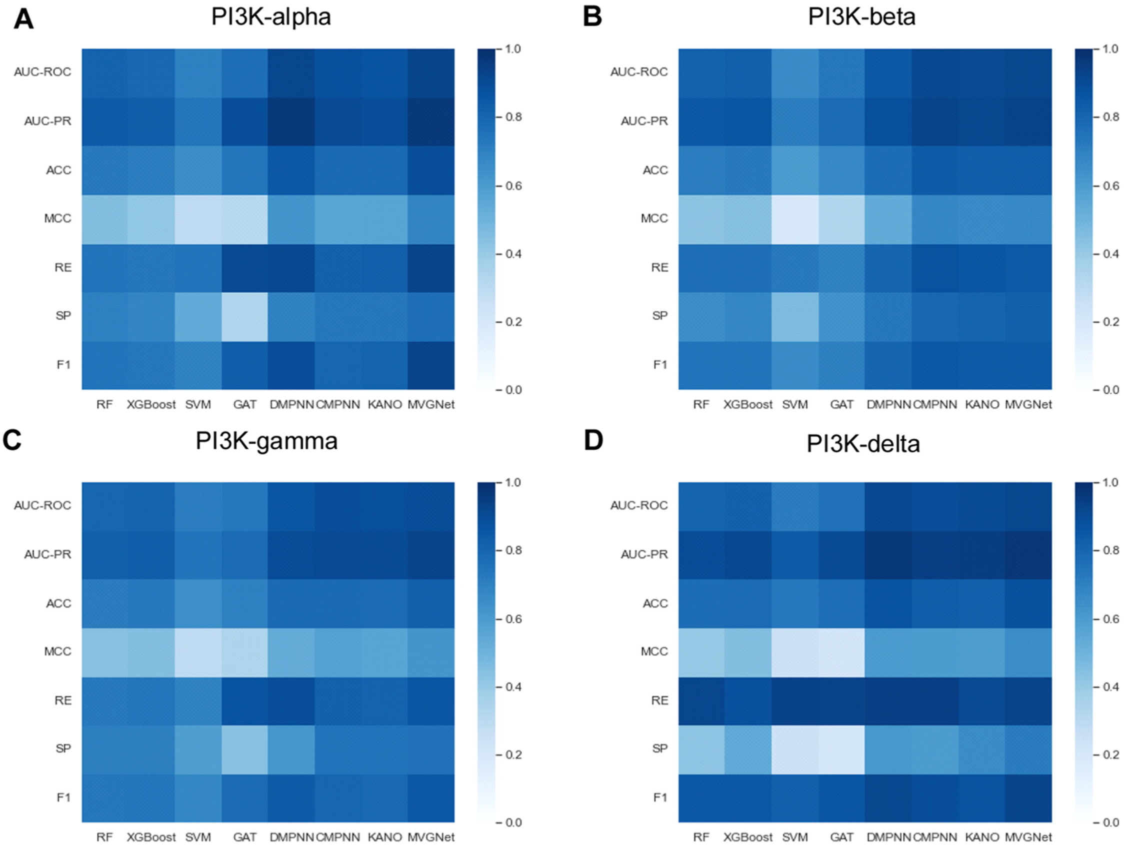


Figure 4. Performance of all models on PI3K-*α* (A), PI3K-*β* (B), PI3K-*γ* (C) and PARP-delta (D) test sets.

2.3. Baseline Machine Learning and Deep Learning Algorithms. To better illustrate the advantages of the model in this study for the PI3K inhibitor prediction task, MVGNet is compared to three machine learning models and four deep learning models.

Machine Learning Baselines: Support Vector Machine (SVM) is a binary classification model that achieves good classification by finding a hyperplane that maximizes the distance between two classes.[32] Random Forest (RF) is a classifier that uses multiple trees to train and predict samples

Table 3. Comprehensive Performance of the PI3K Inhibitor Dataset on the Multitask Model[a]

model	AUC-ROC	AUC-PR	ACC	MCC	RE	SP	F1
GAT	0.726 ± 0.028	0.871 ± 0.022	0.755 ± 0.021	0.285 ± 0.051	0.887 ± 0.061	0.350 ± 0.116	0.841 ± 0.019
DMPNN	0.893 ± 0.025	0.954 ± 0.013	0.861 ± 0.018	0.600 ± 0.052	0.931 ± 0.029	0.631 ± 0.096	0.909 ± 0.012
CMPNN	0.912 ± 0.016	0.965 ± 0.012	0.874 ± 0.015	0.636 ± 0.045	0.928 ± 0.020	0.690 ± 0.067	0.917 ± 0.012
KANO	0.908 ± 0.016	0.960 ± 0.008	0.870 ± 0.014	0.636 ± 0.046	0.933 ± 0.019	0.673 ± 0.062	0.914 ± 0.009
MVGNet	0.927 ± 0.006	0.980 ± 0.002	0.889 ± 0.008	0.643 ± 0.028	0.940 ± 0.014	0.730 ± 0.045	0.930 ± 0.005

[a]The best-performing results are marked in bold.

by randomly selecting the number of features, randomly selecting the training data, and taking the prediction label with the most occurrences for the same prediction as the final prediction label.[33] Extreme Gradient Boosting (XGBoost) belongs to the Gradient Boosting Tree, which is a variant of the Gradient Boosting Tree and is suitable for classification and regression problems due to its efficient performance, automatic handling of missing values, feature importance assessment, regularization, etc;[34]

Deep learning baselines: Graph Attention Networks (GAT) introduce the attention mechanism to spatial domain-based graph neural networks, which can be used to learn different weight depths for different neighbors.[35,36] Directed Message Passing Neural Networks (DMPNN) passes information no longer between atoms but between individual bonds.[37] CMPNN (2020) is an interactive update edge and node-embedded communication message-passing neural network.[24] KANO is a molecular property prediction method based on chemical element knowledge mapping with functional group hints.[28]

2.4. Performance Evaluation of Models. The following metrics were used to evaluate the performance of all models in this study, including specificity (SP), recall (RE), accuracy (ACC), F1 score (F1), Matthews correlation coefficient (MCC), the area under the receiver operating characteristic (ROC) curve (AUC-ROC), and the area under the precision recall (PR) curve (AUC-PR), in comparison with the results of the multibaseline model comparison experiments. The results of the comparison experiments with several baseline models. The area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) are the two determinants used for model evaluation. These evaluation metrics are defined below:

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{PR} \times \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{PN})}} \tag{6}$$

where TP is true positive; FP is false positive; TN is true negative; FN is false negative.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation of Models. To fairly compare model performance, the single-task MVGNet model was compared to all baseline models, and then the multitask MVGNet model was compared to four deep learning baseline models with multitask learning. All models were cross-validated with 5-fold cross-validation, with three independent experiments for each task. All models were trained on 30 epochs with a learning rate of 1 × 10^{-4}. The best-performing model on the validation set was selected as the final model.

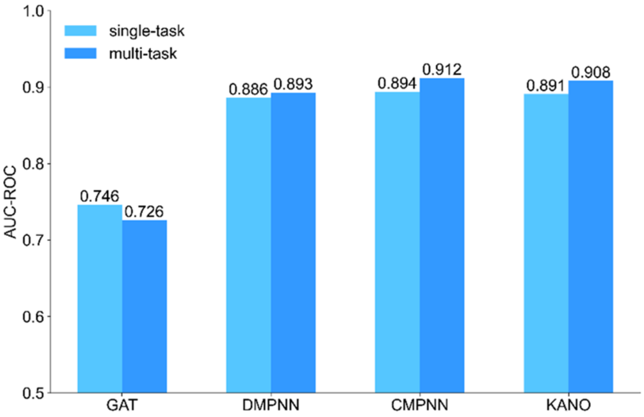


Figure 5. Comparison of AUC-ROC for single-task and multitask models.

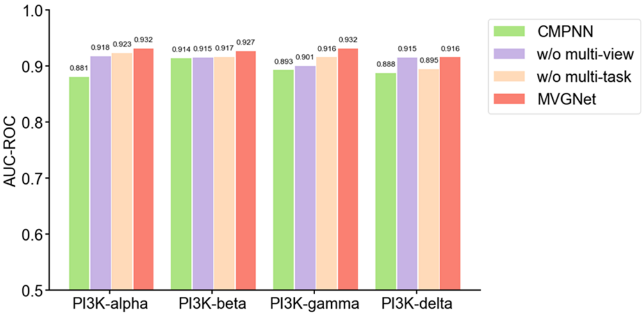


Figure 6. Results of ablation experiments on four PI3K subtype inhibitor data sets, green represents the baseline model CMPNN used in this study, purple represents the results obtained without multiview and only with multitask learning, light orange represents the results obtained without multitask learning and only with multiview, and light red is the model MVGNet used in this study that combines multiview and multitask learning.

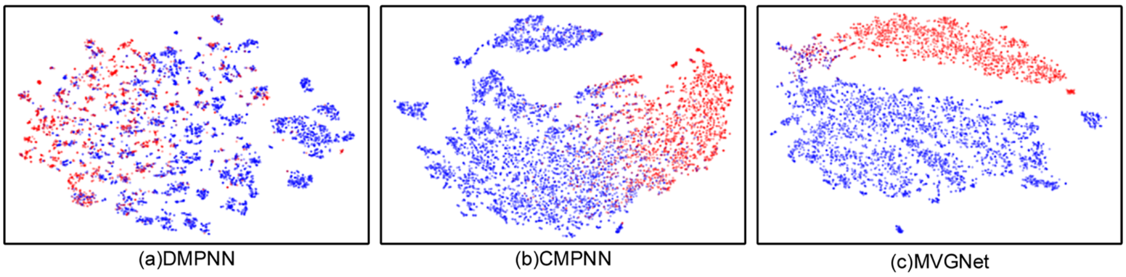
Three machine learning methods (RF, XGBoost, SVM) and four deep learning methods (GAT, DMPNN, CMPNN, KANO) were applied to the four subtyped data sets of PI3K. The mean performance across these methods was used as the final overall performance measure to compare all model

https://doi.org/10.1021/acsomega.4c06224
ACS Omega 2024, 9, 45159−45168

Table 4. Overall Prediction Performance of CMPNN, the Model Using Only Multitask Learning, the Model Using Only Multiviews, and MVGNet[a]

model	AUC-ROC	AUC-PR	ACC	MCC	RE	SP	F1
CMPNN	0.894 ± 0.020	0.921 ± 0.019	0.810 ± 0.015	0.603 ± 0.030	0.864 ± 0.049	0.719 ± 0.058	0.834 ± 0.021
w/o multiview	0.909 ± 0.005	0.964 ± 0.004	0.862 ± 0.009	0.633 ± 0.026	0.912 ± 0.018	0.704 ± 0.049	0.906 ± 0.007
w/o multitask	0.913 ± 0.011	0.949 ± 0.010	0.856 ± 0.012	0.659 ± 0.023	0.890 ± 0.027	0.725 ± 0.048	0.887 ± 0.013
MVGNet	0.927 ± 0.006	0.980 ± 0.002	0.889 ± 0.008	0.643 ± 0.028	0.940 ± 0.014	0.730 ± 0.045	0.930 ± 0.005

[a] The best-performing results are marked in bold.


(a)DMPNN

(b)CMPNN

(c)MVGNet

Figure 7. Visualization of molecular features. Molecular features of PI3K-α from (a) DMPNN, (b) CMPNN and (c) MVGNet were visualized using t-SNE. Any molecule with a label of 1 is an active compound, while any molecule with a label of 0 is an inactive compound, where active compounds are colored blue and inactive ones are colored red.

Table 5. Information on the Structural Features of the Pharmacophore in Duvelisib and Tenalisib

pharmacophore	Duvelisib	Tenalisib
SingleAtomDonor	1	1
SingleAtomAcceptor	1	1
imidazole	1	1
ZnBinder5	1	1
ZnBinder6	1	1
Arom6	1	1
Arom7	1	1
ThreeWayAttach	1	1
ChainTwoWayAttach	0	1
RH6_6	1	1

performances. The area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) were the two crucial metrics used for model evaluation. The experimental results (Table 2) show that MVGNet performs better among all the models. Compared with other message-passing graph neural network models, MVGNet achieves a relatively large improvement in both AUC-ROC and AUC-PR metrics.

In addition, a visual heat map was created to illustrate the performance of the four subtypes of PI3K data sets across all models. It used AUC-ROC as the primary evaluation metric, where darker colors indicated higher metric values, implying better model performance. As shown in Figure 4, MVGNet performed well on all four different test sets, with AUC-ROC and AUC-PR metrics leading the other models. The remaining three message-passing neural network-based models (DMPNM, CMPNN, and KANO) also showed good performance compared to the machine learning model and the classical GAT model, demonstrating the superiority of messaging neural networks in predicting PI3K inhibitor activity.

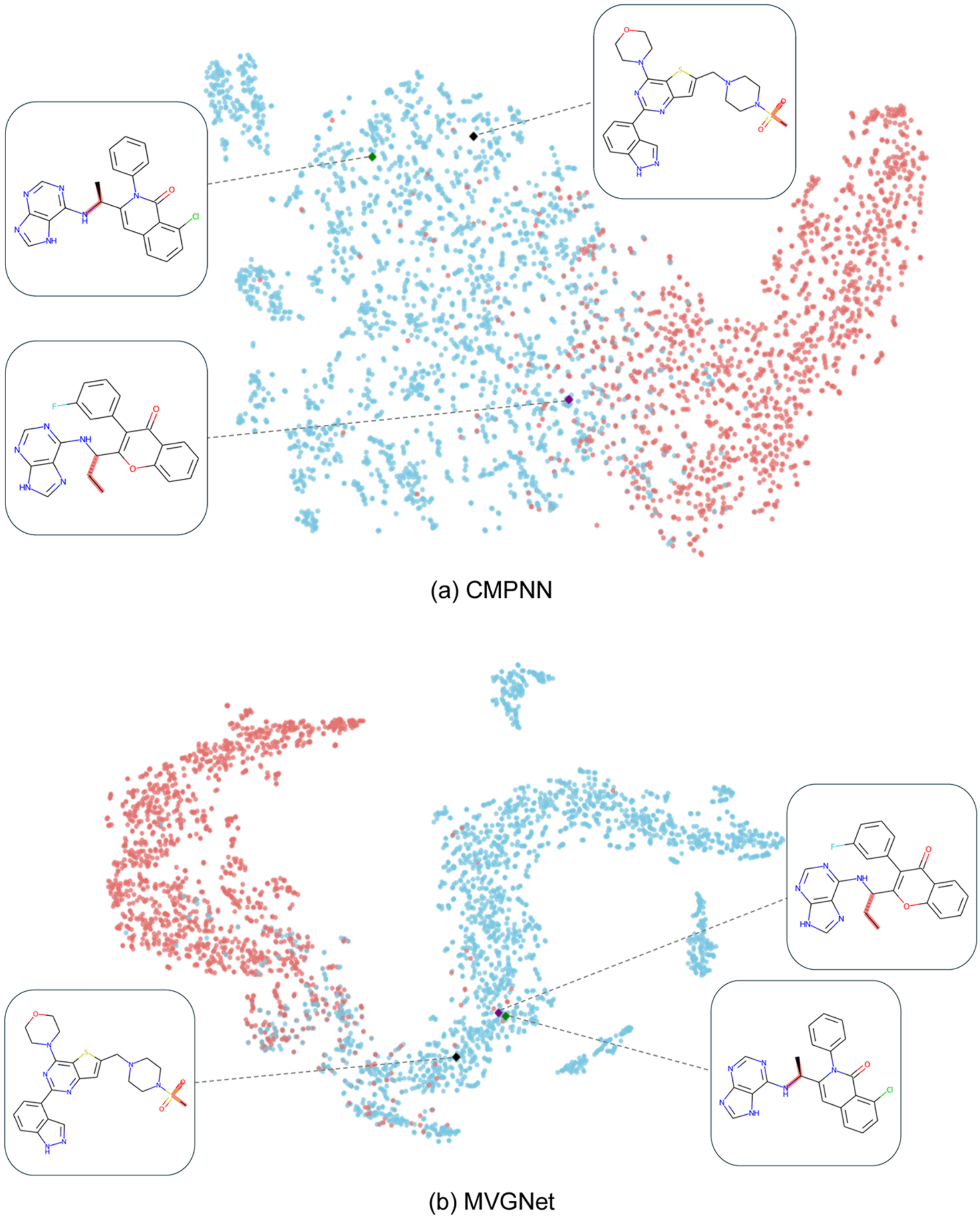
Table 3 shows the combined performance of the PI3K inhibitor dataset on a multitask MVGNet with four deep learning baseline models after increasing multitask learning. The results in Table 3 and Figure 5 show that most models have a relatively significant improvement in their performance in predicting the biological activity of PI3K inhibitors with the addition of multitask learning. This also demonstrates the positive role of multitask learning in predicting the biological activity of PI3K inhibitors.

3.2. Ablation Experiment.

To demonstrate the effectiveness of multiview and multitask learning, ablation experiments were conducted. Figure 6 visually compares the baseline model CMPNN, the model utilizing only multitask learning, the model employing only multiview learning, and our proposed model MVGNet. AUC-ROC was chosen as the primary evaluation metric across four PI3K subtype inhibitor data sets to evaluate these models. As shown in the figure, both multiview and multitask learning show a relatively large improvement compared to the original results of the baseline model CMPNN. The best experimental results are obtained with MVGNet after combining multiview and multitask learning. In addition, the experimental results of different models for predicting the activity of the four PI3K subtype inhibitors were averaged. Table 4 displays the final results of the four models across seven evaluation metrics on the test set of PI3K inhibitors. The results show that either the model using only multitask learning or the model using only multiview worsens the prediction performance, with the CMPNN-only model having the worst results. And after adding multiview and multitask learning, all the metrics have a relatively large improvement.

3.3. Feature Visualization.

To better demonstrate the robust representation learning capabilities of our model, t-distributed stochastic neighborhood embedding (t-SNE) was used to visualize the default hyperparameters of molecular representations within the PI3K-α dataset. Active compounds were labeled 1 and inactive compounds were labeled 0. Given that feature spaces with similar activities tend to have greater similarity, t-SNE was used to visualize their embeddings and assess whether molecular representations were effectively learned by the model based on the clarity of the boundaries between active and inactive molecules. As shown in Figure 7,


(a) CMPNN

(b) MVGNet

Figure 8. Case study. A case study of molecular characterization on the PI3K-γ dataset based on t-SNE visualization of CMPNN and MVGNet, where the active molecules of the PI3K-γ inhibitors were labeled as blue, the inactive molecules were labeled as light red, and the three special cases of active molecules were labeled as a diamond pattern and differentiated by different colors. Three of these molecules, which were shown to be active in clinical trials, were selected for case studies.

the molecular representation generated by MVGNet clearly separates active from inactive molecules in PI3K-α, followed by CMPNN, while DMPNN performs poorly. This further confirms that MVGNet has achieved enhanced representation learning capabilities by integrating reaction information and fragment pharmacophore information from multiview, coupled with multitask weight sharing.

3.4. Case Study. Dual inhibitors are increasingly used in the current treatment of breast cancer and tumors, etc. Among these new-generation drugs, orally active Pictilisib (GDC0941), a PI3K α/δ inhibitor, has shown a favorable safety profile in combination with other anticancer drugs for the treatment of breast cancer, advanced solid tumors, and nonsquamous nonsmall-cell lung cancers.[9] It is also moderately selective for p110β and p110γ, with an IC_{50} value of 75 nM for PI3Kγ. Duvelisib (IPI-145, marketed as Copiktra) is a novel selective PI3K δ/γ inhibitor, and clinically, Duvelisib is approved for the treatment of hematological malignancies.[38] Tenalisib (RP6530) is a new generation of PI3K δ/γ inhibitors, and preclinical studies have demonstrated the ability to induce

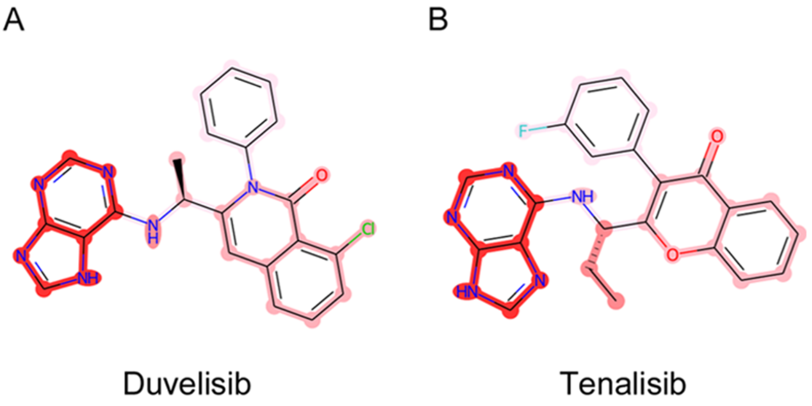


Figure 9. Molecule-fragment similarity visualization. Minimum-maximum normalization of cosine similarity scores was used to measure the importance of the Duvelisib (A) and Tenalisib (B) fragments, with darker colors representing higher cosine similarity scores.

apoptosis and antiproliferative activity, thereby reducing angiogenesis.[39]

Pharmacophores are the core structural units of molecules that interact with biological targets, and the ability to identify and add useful structural information about pharmacophores can help strengthen molecular characterization. To demonstrate that pharmacophore structure information was successfully learned by MVGNet, the molecules in the PI3Kγ dataset were visualized, focusing on the three molecules known to be active in PI3Kγ clinical trials. The PI3Kγ inhibitor active molecules were labeled in blue, the inactive molecules were labeled in light red, and the three special molecules were labeled in a diamond pattern if they were active and were distinguished using different colors to differentiate them, where black represents Pictilisib, green represents Duvelisib, and purple represents Tenalisib. As PI3Kδ/γ inhibitors, both Duvelisib and Tenalisib are structurally characterized by the presence of the 6-(methylamino)-purine substructure and have many of the same pharmacophores. As shown in Table 5, the leftmost column of the table lists all the pharmacophore structural information in Duvelisib and Tenalisib, and if there is relevant pharmacophore structural information, it is marked as 1, otherwise it is marked as 0. It is obvious that they both have the same pharmacophore structural features such as SingleAtomDonor, SingleAtomAcceptor, and so on. Therefore, if the model has the ability to capture the pharmacophore information, then these two molecules will be clustered together. Figure 8a shows that in CMPNN, Duvelisib and Tenalisib with similar pharmacophore structure information do not cluster well after visualization. In contrast, Figure 8b shows that MVGNet can cluster Duvelisib and Tenalisib with similar pharmacophore structure information together and distinguish them from inactive molecules. This indicates that the embedded representation learned by MVGNet can effectively capture pharmacophore structure information.

To investigate exactly which fragment contributes more to the prediction of PI3K inhibitor biological activity, molecule-fragment similarity assessment was introduced. Here, molecule-fragment cosine similarity was calculated using the following equation:

$$\cos(\theta_{\mathrm{F}}) = \frac{h_{\mathrm{G}} \cdot h_{\mathrm{F}}}{\|h_{\mathrm{G}}\| \cdot \|h_{\mathrm{F}}\|} \qquad (7)$$

where h_{G} and h_{F} represent the feature vectors of the molecular graph and fragments, respectively, and · denotes the vector dot product.

Figure 9 visualizes the importance of each fragment in Duvelisib and Tenalisib, and it can be seen that both Duvelisib, and Tenalisib are more focused on purines, and the cosine similarity scores are much higher than the other fragments. Purines are also one of the common backbones of PI3K inhibitors and play an important role in the inhibition of PI3Kγ by Duvelisib and Tenalisib. This also indicates that MVGNet can accurately identify the key backbones.

4. CONCLUSIONS

In this study, we propose a multitask MVGNet model. The model obtains the fragment view of the molecule through the BRICS algorithm, extracts the pharmacophore information on the fragment and the reaction information between the fragments, and obtains more molecular feature information through multiple views. In addition, multitask learning allows models to learn more about the associations between subtasks by sharing parameters. The experimental results show that MVGNet achieves state-of-the-art performance in predicting PI3K inhibitor activity. The results of ablation experiments show that the combination of molecular fragment pharmacophore information and multitask learning can effectively obtain the correlation information between subtasks. Finally, three molecules that exhibited activity in clinical trials were selected for case studies, and MVGNet successfully learned similar substructural information and distinguished them from inactive molecules.

■ ASSOCIATED CONTENT

Data Availability Statement

All data sets and codes used in this study are available at GitHub: https://github.com/xqwq123/MVGNet.

■ AUTHOR INFORMATION

Corresponding Author

Zhong Li − *Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources, School of Information Engineering, Huzhou University, Huzhou 313000 Zhejiang Province, China*; Email: lizhong@zjhu.edu.cn

Authors

Yanlei Kang − *Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources, School of Information Engineering, Huzhou University, Huzhou 313000 Zhejiang Province, China*

Qiwei Xia − *Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources, School of Information Engineering, Huzhou University, Huzhou 313000 Zhejiang Province, China;* orcid.org/0009-0009-9182-8746

Yunliang Jiang − *Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources, School of Information Engineering, Huzhou University, Huzhou 313000 Zhejiang Province, China; School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004 Zhejiang Province, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c06224

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (nos. 62302165 and 12171434) and Science and Technology Plan Project of Huzhou City, China (nos. 2022YZ03 and 2022GZ51).

■ REFERENCES

(1) Vanhaesebroeck, B.; Perry, M. W. D.; Brown, J. R.; André, F.; Okkenhaug, K. PI3K inhibitors are finally coming of age. *Nat. Rev. Drug Discovery* **2021**, *20* (10), 741−769.

(2) Hawkins, P. T.; Stephens, L. R. PI3K signalling in inflammation. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2015**, *1851* (6), 882−897.

(3) Bilanges, B.; Posor, Y.; Vanhaesebroeck, B. PI3K isoforms in cell signalling and vesicle trafficking. *Nat. Rev. Mol. Cell Biol.* **2019**, *20* (9), 515−534.

(4) Costa, R. L. B.; Han, H. S.; Gradishar, W. J. Targeting the PI3K/AKT/mTOR pathway in triple-negative breast cancer: a review. *Breast Cancer Res. Treat.* **2018**, *169* (3), 397−406.

(5) Sun, E. J.; Wankell, M.; Palamuthusingam, P.; McFarlane, C.; Hebbard, L. Targeting the PI3K/Akt/mTOR Pathway in Hepatocellular Carcinoma. *Biomedicines* **2021**, *9* (11), No. 1639.

(6) Wullenkord, R.; Friedrichs, B.; Erdmann, T.; Lenz, G. Therapeutic potential of PI3K signaling in distinct entities of B-cell lymphoma. *Expert Rev. Hematol.* **2019**, *12* (12), 1053−1062.

(7) Dienstmann, R.; Rodon, J.; Serra, V.; Tabernero, J. Picking the Point of Inhibition: A Comparative Review of PI3K/AKT/mTOR Pathway Inhibitors. *Mol. Cancer Ther.* **2014**, *13* (5), 1021−1031.

(8) Liu, N.; Rowley, B. R.; Bull, C. O.; Schneider, C.; Haegebarth, A.; Schatz, C. A.; Fracasso, P. R.; Wilkie, D. P.; Hentemann, M.; Wilhelm, S. M.; et al. BAY 80−6946 is a highly selective intravenous PI3K inhibitor with potent p110α and p110δ activities in tumor cell lines and xenograft models. *Mol. Cancer Ther.* **2013**, *12* (11), 2319−2330. From NLM.

(9) Sarker, D.; Ang, J. E.; Baird, R.; Kristeleit, R.; Shah, K.; Moreno, V.; Clarke, P. A.; Raynaud, F. I.; Levy, G.; Ware, J. A.; et al. First-in-Human Phase I Study of Pictilisib (GDC-0941), a Potent Pan-Class I Phosphatidylinositol-3-Kinase (PI3K) Inhibitor, in Patients with Advanced Solid Tumors. *Clin. Cancer Res.* **2015**, *21* (1), 77−86.

(10) Fruman, D. A.; Rommel, C. PI3K and cancer: lessons, challenges and opportunities. *Nat. Rev. Drug Discovery* **2014**, *13* (2), 140−156.

(11) Janku, F.; Yap, T. A.; Meric-Bernstam, F. Targeting the PI3K pathway in cancer: are we making headway? *Nat. Rev. Clin. Oncol.* **2018**, *15* (5), 273−291.

(12) Shan, K. S.; Bonano-Rios, A.; Theik, N. W. Y.; Hussein, A.; Blaya, M. Molecular Targeting of the Phosphoinositide-3-Protein Kinase (PI3K) Pathway across Various Cancers. *Int. J. Mol. Sci.* **2024**, *25* (4), No. 1973.

(13) Kumar, B. H.; Manandhar, S.; Choudhary, S. S.; Priya, K.; Gujaran, T. V.; Mehta, C. H.; Nayak, U. Y.; Pai, K. S. R. Identification of phytochemical as a dual inhibitor of PI3K and mTOR: a structure-based computational approach. *Mol. Diversity* **2023**, *27* (5), 2015−2036.

(14) Das, S.; Halder, D.; Jeyaprakash, R. S. Computational-guided approach for identification of PI3K alpha inhibitor in the treatment of hepatocellular carcinoma by virtual screening and water map analysis. *J. Biomol. Struct. Dyn.* **2024**, 1−23.

(15) Zhu, J.; Meng, H.; Li, X.; Jia, L.; Xu, L.; Cai, Y.; Chen, Y.; Jin, J.; Yu, L. Optimization of virtual screening against phosphoinositide 3-kinase delta: Integration of common feature pharmacophore and multicomplex-based molecular docking. *Comput. Biol. Chem.* **2024**, *109*, No. 108011. Journal Article

(16) Speck-Planche, A.; Scotti, M. T. BET bromodomain inhibitors: fragment-based in silico design using multi-target QSAR models. *Mol. Diversity* **2019**, *23* (3), 555−572.

(17) Speck-Planche, A.; Kleandrova, V. V. Multi-Condition QSAR Model for the Virtual Design of Chemicals with Dual Pan-Antiviral and Anti-Cytokine Storm Profiles. *ACS Omega* **2022**, *7* (36), 32119−32130.

(18) Kleandrova, V. V.; Speck-Planche, A. PTML Modeling for Pancreatic Cancer Research: In Silico Design of Simultaneous Multi-Protein and Multi-Cell Inhibitors. *Biomedicines* **2022**, *10* (2), 491.

(19) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A Practical Overview of Quantitative Structure-activity Relationship. *Excli J.* **2009**, *8*, 74−88.

(20) Fang, X.; Gao, Y.; Wang, C.; Chen, H.; Zhu, T. Exploring the selective mechanism of inhibitors towards different subtypes of class I PI3K. *Chem. Phys. Lett.* **2022**, *786*, No. 139174.

(21) Nguyen-Vo, T.-H.; Trinh, Q. H.; Nguyen, L.; Nguyen-Hoang, P.-U.; Nguyen, T.-N.; Nguyen, D. T.; Nguyen, B. P.; Le, L. iCYP-MFE: Identifying Human Cytochrome P450 Inhibitors Using Multitask Learning and Molecular Fingerprint-Embedded Encoding. *J. Chem. Inf. Model.* **2022**, *62* (21), 5059−5068.

(22) Ai, D.; Wu, J.; Cai, H.; Zhao, D.; Chen, Y.; Wei, J.; Xu, J.; Zhang, J.; Wang, L. A multi-task FP-GNN framework enables accurate prediction of selective PARP inhibitors. *Front. Pharmacol.* **2022**, *13*, No. 971369, DOI: 10.3389/fphar.2022.971369.

(23) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning*, 2017; Vol. *70*.

(24) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. Communicative Representation Learning on Attributed Molecular Graphs. *IJCAI* **2020**, *2020*, 2831−2838.

(25) Cai, H.; Zhang, H.; Zhao, D.; Wu, J.; Wang, L. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. *Brief Bioinform* **2022**, *23* (6), No. bbac408, DOI: 10.1093/bib/bbac408.

(26) Zhu, W.; Zhang, Y.; Zhao, D.; Xu, J.; Wang, L. HiGNN: A Hierarchical Informative Graph Neural Network for Molecular Property Prediction Equipped with Feature-Wise Attention. *J. Chem. Inf. Model.* **2023**, *63* (1), 43−55.

(27) Li, B.; Lin, M.; Chen, T.; Wang, L. FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings Bioinf.* **2023**, *24* (6), No. bbad398, DOI: 10.1093/bib/bbad398.

(28) Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat. Mach. Intell.* **2023**, *5* (5), 542−553.

(29) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503−1507.

(30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, María P.; Mosquera, Juan F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47* (D1), D930−D940. (acccessed 3/15/2024).

(31) Jiang, Y.; Jin, S.; Jin, X.; Xiao, X.; Wu, W.; Liu, X.; Zhang, Q.; Zeng, X.; Yang, G.; Niu, Z. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Commun. Chem.* **2023**, *6* (1), No. 60.

(32) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20* (3), 273−297.

(33) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(34) Chen, T.; Guestrin, C.et al. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016.

(35) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. arXiv:1710.10903. arXiv.org e-Print archive. https://arxiv.org/abs/1710.10903 (accessed 2017).

(36) Jiang, S. L.; Balaprakash, P.Graph Neural Network Architecture Search for Molecular Property Prediction *8th IEEE International Conference on Big Data (Big Data)*, (Electr Network, Dec 10−13), 2020; Vol. *2020*, pp 1346−1353 .

(37) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370−3388.

(38) Flinn, I. W.; Hillmen, P.; Montillo, M.; Nagy, Z.; Illés, A.; Etienne, G.; Delgado, J.; Kuss, B. J.; Tam, C. S.; Gasztonyi, Z.; et al. The phase 3 DUO trial: duvelisib vs ofatumumab in relapsed and refractory CLL/SLL. *Blood* **2018**, *132* (23), 2446−2455.

(39) Huen, A.; Haverkos, B. M.; Zain, J.; Radhakrishnan, R.; Lechowicz, M. J.; Devata, S.; Korman, N. J.; Pinter-Brown, L.; Oki, Y.; Barde, P. J.; et al. Phase I/Ib Study of Tenalisib (RP6530), a Dual PI3K δ/γ Inhibitor in Patients with Relapsed/Refractory T-Cell Lymphoma. *Cancers* **2020**, *12* (8), No. 2293.