Interdisciplinary Sciences: Computational Life Sciences
https://doi.org/10.1007/s12539-022-00523-1

ORIGINAL RESEARCH ARTICLE

Improving the Prediction of Potential Kinase Inhibitors with Feature Learning on Multisource Knowledge

Yichen Zhong[1,2] · Cong Shen[3] · Huanhuan Wu[1] · Tao Xu[1] · Lingyun Luo[1,2]

Received: 2 December 2021 / Revised: 14 April 2022 / Accepted: 15 April 2022

Abstract

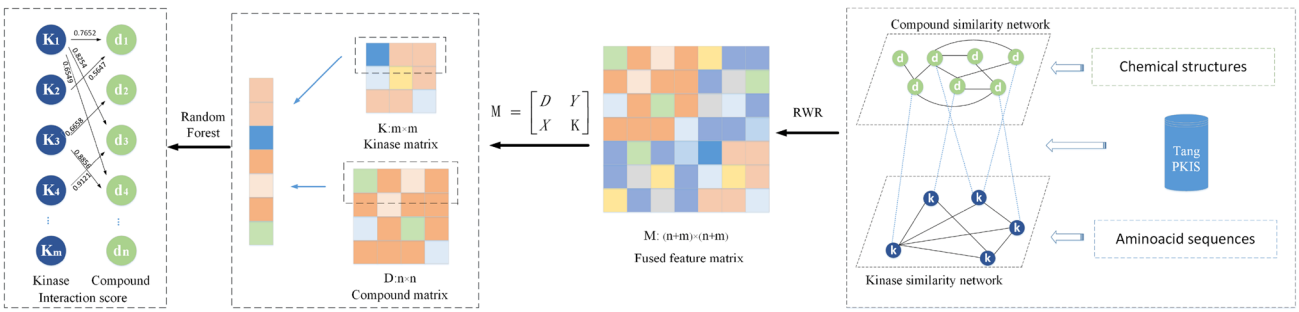
Purpose The identification of potential kinase inhibitors plays a key role in drug discovery for treating human diseases. Currently, most existing computational methods only extract limited features such as sequence information from kinases and inhibitors. To further enhance the identification of kinase inhibitors, more features need to be leveraged. Hence, it is appealing to develop effective methods to aggregate feature information from multisource knowledge for predicting potential kinase inhibitors. In this paper, we propose a novel computational framework called FLMTS to improve the performance of kinase inhibitor prediction by aggregating multisource knowledge.

Method FLMTS uses a random walk with restart (RWR) to combine multiscale information in a heterogeneous network. We used the combined information as features of compounds and kinases and input them into random forest (RF) to predict unknown compound–kinase interactions.

Results Experimental results reveal that FLMTS obtains significant improvement over existing state-of-the-art methods. Case studies demonstrated the reliability of FLMTS, and pathway enrichment analysis demonstrated that FLMTS could also accurately predict signaling pathways in disease treatment.

Conclusion In conclusion, our computational framework of FLMTS for improving the prediction of potential kinase inhibitors successfully aggregates feature information from multisource knowledge, yielding better prediction performance than existing state-of-the-art methods.

Graphical Abstract



Keywords Kinase inhibitor · Multisource knowledge · Feature learning · Heterogeneous network

✉ Lingyun Luo
luoly@usc.edu.cn

1 School of Computer Science, University of South China, Hengyang 421001, China

2 Hunan Provincial Base for Scientific and Technological Innovation Cooperation, Hengyang 421001, China

3 College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China

1 Introduction

For the past 30 years, kinases have been intensively investigated as drug targets. There are 518 kinases in the human kinome, constituting approximately 1.7% of all human genes [1]. Deregulation of kinase function has been proven to be

Published online: 10 May 2022

important in many human diseases, including cancer and immunological, inflammatory, and infectious diseases [2, 3]. Despite the importance of kinases, the functions of more than 100 kinases (approximately 25%) are still completely unknown, and kinases that are basically uncharacterized account for approximately 50% [4]. As of December 2020, 62 kinase inhibitors have been approved by the US Food and Drug Administration (US FDA) for the treatment of cancer and other diseases [5]; however, these drugs have only proven successful in targeting a small number of human kinases (approximately 80). Among them, many kinases are targeted by several kinase inhibitors [6]. The above facts indicate that kinase inhibitor field development is still in its infancy [7, 8]. As a result, finding new kinase inhibitors remains a promising, but difficult task.

Traditional approaches of studying kinase inhibitors often utilize biological experiments, but it is costly with low efficiency even for large-scale pharmaceutical companies [9, 10]. As a result, computation-based models have been proposed to predict kinase ligands in many studies [11–13]. Compared with traditional drug design approaches, computational approaches are generally more flexible and faster. Thus, it is necessary to develop computation-based models to compensate for the shortcomings of traditional approaches for the discovery of kinase inhibitors.

Machine learning-based technologies have significantly improved the effect of forecasting the biological activities of massive kinase inhibitors in recent years. The models used include random forest (RF) [10, 14], support vector machine (SVM) [15], naive Bayesian (NB) [13], K-nearest neighbors (KNN) [16], and deep neural network (DNN) [17, 18]. These models made predictions based on the features of kinases and kinase inhibitors. For instance, Avram et al. [14] developed a PFPECFP model to compute molecular encoding (2D structures of inhibitors) and used the random forest (RF) to compute classification probabilities of kinases. However, the lack of 3D structures of kinases (proteins) and inhibitors (compounds) limits the forecast precision and generalization of the model.

In addition to the features of kinases and inhibitors, the interaction network information between inhibitors and kinases is also useful for predicting kinase inhibitors. Existing studies indicate that network-based methods are critical in drug discovery as well as other tasks [19–21], including drug–target interactions (DTI) [22, 23], drug repositioning [24, 25] and drug combinations [20, 26]. For instance, Chen et al. [27] utilized the KATZ measure and constructed a heterogeneous network for predicting human microbe-disease associations. Li et al. and lv et al. [28, 29] built heterogeneous networks and utilized the random walk with restart (RWR) to find multiple entity relationships. Shen et al. [30] proposed an information fusion model based on a heterogeneous network (IDDkin) to improve the prediction performance of kinase inhibitors and achieved excellent performance. In addition, a method of nonnegative matrix factorization has shown excellent performance on target prediction [31, 32]. It can utilize the similarity information of entities to reconstruct an interaction matrix and predict unknown interactions. Therefore, it is necessary to combine the advantages of feature-based methods and network-based methods to improve the prediction performance of kinase inhibitors.

In this paper, we presented feature learning on a multisource knowledge (FLMTS) model to predict inhibitor (compound)-kinase (protein) interactions. First, we used the features of kinases (proteins) and inhibitors (compounds) to construct two similarity networks. Together with the compound–-kinase interaction network, we constructed a heterogeneous network. Second, we used the RWR to fuse the global network topology information of the heterogeneous network. Third, we used the features of nodes that have fused network information as input of the RF model to output the classification probabilities. The major contributions of our work are summarized as follows:

- To the best of our knowledge, this is the first attempt to extract features by performing RWR on a heterogeneous network for predicting potential kinase inhibitors. RWR can fuse multisource knowledge and extract features without supervision.
- We presented a novel FLMTS model based on RWR and RF. FLMTS could present better performance based on limited a priori knowledge, since it may be less inclined to overfitting.
- Diverse experiments on two public datasets demonstrate that FLMTS not only outperforms the state-of-the-art baselines but can also accurately target disease related signaling pathways.

2 Materials and Methods

2.1 Framework Overview

By fusing the topological information of the heterogeneous network and the features of the nodes (compounds and kinases) in the network, FLMTS employs the RF algorithm to predict the potential kinase inhibitors. The overview of FLMTS is shown in Fig. 1 First, a kinase similarity matrix and a compound similarity matrix were constructed by using the amino acid sequences of the kinases and the MACCS fingerprints of the compounds, respectively. Then we utilized known compound–kinase interactions to construct an adjacency matrix. Second, we used the three matrices to construct three networks, namely the compound similarity network, the kinase similarity network, and the

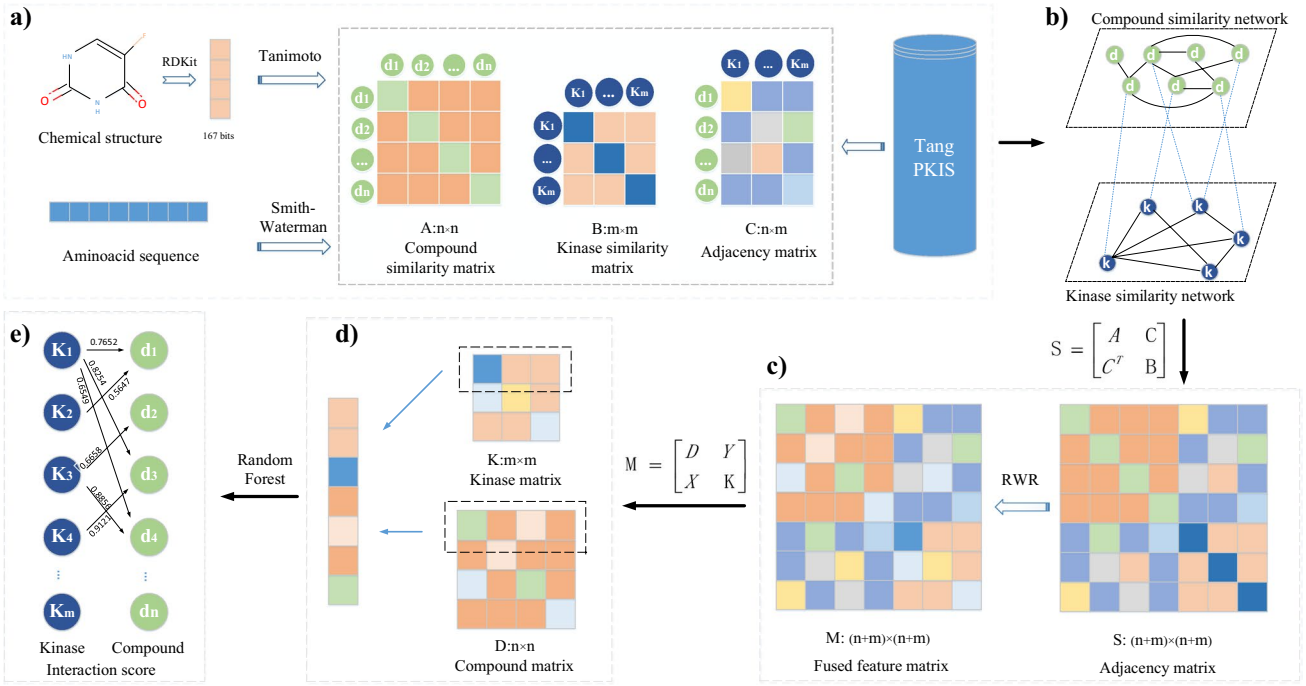


Fig. 1 Overview of the FLMTS architecture. **a** Construct an adjacency matrix C by known compound–kinase interactions, and construct a compound similarity matrix A and a kinase similarity matrix B based on chemical structures and acid sequences separately. **b** Construct a heterogeneous network by combining the three matrices A, B and C above. **c** The global network topology information of nodes is fused by using RWR in the heterogeneous network. **d** Extract the kinase matrix K and the compound matrix D into RF as input. **e** FLMTS provides interaction scores between the compounds and the kinases

compound–kinase interaction network, and employed the three networks to structure a heterogeneous network. Third, the RWR was applied to fuse the global network topology information of nodes in the heterogeneous network. Finally, we used the fusion information of the nodes as features and input them into the RF classifier to train the model.

2.2 Heterogeneous Network Construction

In this study, we used the Tang and PKIS [30] datasets to train the compound–kinase interaction prediction model. In the following, we took an example, the PKIS dataset, to demonstrate the process of network construction. In our experiment, the same operation was also performed on the Tang dataset.

2.2.1 Data Source and Adjacency Matrix Construction

In the Tang and PKIS [30] datasets, Davis' kinase profiling datasets were used to structure the Tang dataset [33–35], and the PKIS dataset stands for the published kinase inhibitor set [36, 37]. In the two datasets, each compound–kinase pair was classified into the positive (active) or the neutral (inactive) class by setting the threshold of pIC_{50} as 6.3 (equivalent to 500 nM). Hence, there are not real negative samples, but only positive samples and neutral samples. In this paper, all neutral samples are regarded as negative samples. There were only 15,660 and 2414 kinase-inhibitor (compound) pairs (positive samples) in the Tang and PKIS datasets, respectively. Notably, there are far fewer known kinase–inhibitor pairs than the unknown ones in real-world scenes. There is no literature or database reporting negative kinase inhibitor pairs. It is worth noting that the PKIS is smaller and more sparsely distributed than the Tang dataset. In this study, the number of negatives is approximately 28 times and 15 times that of the positives in the PKIS and Tang datasets, respectively. Detailed information is shown in Table 1. Let $\mathrm{K} = \{\mathrm{k}_1, \mathrm{k}_2, ..., \mathrm{k}_m\}$ and $\mathrm{P} = \{\mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_n\}$ denote the set of kinases and the set of compounds, respectively. The compound–kinase interaction information was applied to build an adjacency matrix $C \in \mathrm{R}^{m \times n}$. For each element C_{ij} in C, $C_{ij} = 1$ (positive sample) if compound p_i

Table 1 Detailed descriptions of the two datasets

Datasets	Kinases	Compounds	Negative	Positive	Total
Tang	188	1351	238,328	15,660	253,988
PKIS	195	366	68,956	2414	71,370

and kinase k_j are kinase-inhibitor pairs; otherwise, $C_{ij} = 0$ (negative sample).

2.2.2 Similarity Calculation

For compounds, we used the MACCS fingerprints to calculate the similarity score by using the Tanimoto coefficient (T) [38]. Notably, each MACCS fingerprint is a 167-dimensional string composed of values 0 and 1 and was assembled by RDKit (http://www.rdkit.org/). The Tanimoto coefficient T of two compounds is defined as:

$$T = \frac{s}{e + h - s},$$

where s is the number of same dimensionals with the same values between the fingerprints of the two compounds, and e and h denote the length of the MACCS fingerprints of the two compounds. The results were represented by a compound similarity matrix $A \in R^{n \times n}$. In matrix A, the values of entries (i, j) and (j, i) are calculated by using compound p_i and compound p_j. Additionally, we assembled the names of kinases from the PKIS set to obtain the UniProt ID of kinases from the KinHub database (http://kinhub.org/). Then, we downloaded the amino acid sequences of all the kinases from the UniProt database (http://www.uniprot.org/) in the PKIS set and the similarity of the kinases was calculated based on the amino acid sequences using the Smith-Waterman algorithm [39]. The Smith Waterman algorithm determines the similarity region between the two amino acid sequences of the kinase by comparing all possible length fragments and optimizing the similarity measure, to perform local sequence alignment [20]. The pair of segments with maximum similarity could be found by first locating the maximum element of score matrix [39], and the similarity score of two kinase amino acid sequences is determined by the maximum score in the matrix. The results of all kinase similarity scores were represented as a kinase similarity matrix $B \in R^{m \times m}$. Finally, we normalized matrix B by row.

2.2.3 The Heterogeneous Network

In this paper, the heterogeneous network is structured by using a compound similarity network, a kinase similarity network, and a compound–kinase interaction network. First, we used the adjacency matrix C to construct the compound–kinase interaction network, adding an edge between the kinase k_j and the compound p_i if $C_{ij} = 1$; otherwise, no edge was added between k_j and p_i. The compound–kinase interaction network includes 195 kinase nodes, 366 compound nodes, and 2414 edges from the PKIS dataset.

Second, we constructed a compound similarity network using the compound similarity matrix A. The network contains 366 nodes (compounds) and 133,956 edges, and the edge weight was set as the similarity score between the two nodes. Similarly, a kinase similarity network was also constructed, including 195 nodes (kinases) and 38,025 edges. Finally, we constructed a heterogeneous network based on the PKIS dataset, which contains two types of nodes (195 kinases and 366 compounds) and three types of edges (133,956 compound-compound similarities, 38,025 kinase-kinase similarities, and 2414 compound–-kinase interactions). Similarly, a heterogeneous network was also constructed based on the Tang dataset, containing two types of nodes (188 kinases and 1351 compounds) and three types of edges (35,344 kinase-kinase similarities, 1,825,201 compound–-compound similarities, and 15,660 compound–kinase interactions) (Table 2).

2.3 Heterogeneous Information Fusion

In the FLMTS model, we used the RWR to fuse node information in the heterogeneous network. The random walker starts at the seed node and spreads information throughout the network by either (1) randomly moving to its connected neighbors at every step according to the probability transition matrix [29, 40] or (2) restarting from the seed node according to the restarting probability α. Ultimately, the representation of each node in the heterogeneous network was able to fuse the global information of the network. Specifically, let q_0 be the initial vector in which the seed node u is set to 1 and the other nodes are equal to 0, and let q_n be the vector for the n-th iteration, where the i-th value represents the probability of the random walker taking n steps at the seed node to node i. The vector of seed node u at step $n + 1$ can be represented by the equation below:

$$q_{n+1} = (1 - \alpha)\tilde{S}q_n + \alpha q_0,$$

where $\tilde{S}$ is the normalized adjacency matrix $S = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$. The iteration will stop, if the difference between q_{n+1} and q_n (measured by the L_2 norm) falls below 10^{-16} or the number of iterations is over 100. When it stops at step t, we obtain the representation of u as q_t. By stacking the representations of all the nodes, we obtained a feature matrix.

Table 2 Detailed information of two heterogeneous networks

	Tang		PKIS	
	Nodes	Edges	Nodes	Edges
Compound similarity network	1351	1,825,201	366	133,956
Kinase similarity network	188	35,344	195	38,025
Interaction network	1539	15,660	561	2414

2.4 Prediction with Random Forest

The feature matrix constructed above is represented as $M = \begin{bmatrix} D & Y \\ X & K \end{bmatrix}$, where $D \in R^{n\times n}$, $Y \in R^{n\times m}$, $X \in R^{m\times n}$, and $K \in R^{m\times m}$. The matrix D and the matrix K represent the compound feature matrix and the kinase feature matrix, respectively, where the j-th row of K (the i-th row of D) expresses the feature of the j-th kinase (the i-th compound). We connected the feature of the i-th compound and the j-th kinase, and input it into the random forest to identify their relationship score. Similarly, we also used the matrix $(D + Y)$ and the matrix $(X + K)$ as the compound feature and kinase feature, respectively, to predict their relationship score. The performance was evaluated with the baselines.

2.5 Implementation Details and Evaluation Metrics

We applied the same setting of FLMTS on two kinase profiling datasets and implemented FLMTS with NumPy = 1.21.2 and Scikit-Learn = 1.0. We trained a random forest classifier using the function sklearn.ensemble.RandomForestClassifier with parameters (n_estimators = 1000, min_samples_split = 2, random_state = 0) in Python = 3.7.6. The parameter α of the restarting probability is from the set {0.1, 0.2, …, 0.9}. Two metrics were applied to evaluate the performance of FLMTS, including ROC-AUC and AUPR score. In addition, for a more comprehensive evaluation of the model, we also applied precision, recall, balance-accuracy (BA), and F1-score (F1) to measure its performance:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall},$$

$$BA = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right),$$

where TP, FP, TN and FN represent the numbers of positive samples identified correctly, positive samples identified incorrectly, negative samples identified correctly and negative samples identified incorrectly, respectively. We compared the performance of FLMTS and baselines performing fivefold cross-validation: we randomly split the compound–-kinase interactions into five different sets with equal size, among which four sets were selected for training, and the last one was used to test the model. By choosing different test sets each time, this procedure was repeated five times. In total, the process of cross validation was repeated 10 times, and we calculated the average score as the result.

2.6 Baselines

We introduced several state-of-the-art approaches, which are classical methods based on heterogeneous networks.

- IDDkin [30] is a new information fusion model based on heterogeneous network. It combines compound–-kinase interaction information and compound similarity to construct heterogeneous networks. It improves the prediction performance by fusing the structure information and the topology information of heterogeneous networks.
- RWR [29] based on the topology information of the heterogeneous network and first applies network retrieval methods to integrated biological interactions. We adopted this approach on the heterogeneous network to predict unknown interactions.
- Nonnegative Matrix Factorization (NMF) [31]. The multisource information is effectively integrated through matrix decomposition, and then the possible interactive information is predicted by matrix synthesis. We utilized known interactions, compound similarity, and kinase similarity to predict undiscovered interactions.
- Katz [27] is also a network-based method to obtain network topology information. This method combines KATZ measurement [41], a similarity metric, and Gaussian interaction profile kernel similarity to predict the interaction relationship.
- FLMTS + . Variant of our proposed method FLMTS, where we added the interaction information of each compound (kinase) with each kinase (compound) and concatenated it to the original compound (kinase) feature. It is worth noting that both the interaction information and the original feature fused the heterogeneous information by RWR.

3 Results

We illustrate our experimental results in detail, and the superiority of FLMTS was demonstrated by comparing our results with baselines.

3.1 Influence of the Hyperparameter

In FLMTS, the restarting probability (α) is the only hyperparameter. We chose the value of α according to the best performance of FLMTS judged by the AUC and the AUPR. The results demonstrate that FLMTS achieves optimal performance at $\alpha = 0.7$ (Fig. 2). We found that α has a slight

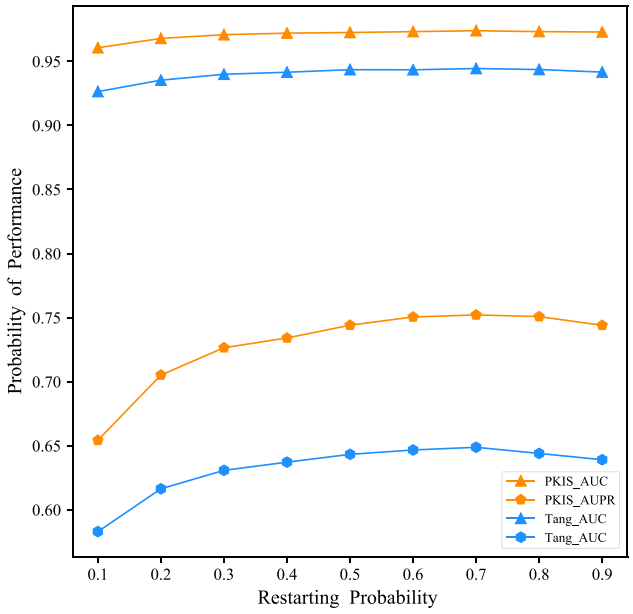


Fig. 2 The AUC and the AUPR of FLMTS with different restarting probabilities on the Tang and PKIS datasets

effect on the AUC but an obvious effect on the AUPR, and the performance based on the PKIS dataset outperformed Tang in both AUC and AUPR metrics. Notably, the PKIS dataset is smaller and sparser than the Tang dataset. Figure 3 shows that our method is more suitable for datasets with less data volume and fewer positive samples. In the field of kinase inhibitors, the amount of data tends to be smaller and there are even fewer experimentally validated known inhibitors; therefore, our model has excellent practical value for the prediction of potential kinase inhibitors.

3.2 Comparison with Other Methods

The results of evaluating the performance of FLMTS are presented in Fig. 3. Both FLMTS and its variant FLMTS + outperformed the other models by a large margin, and FLMTS obtained the best performance. On the PKIS dataset, FLMTS achieves the highest AUC value of 0.9732, which is 3.44% higher than the best NMF model in the state-of-the-art (Fig. 3a). For the other evaluation metric AUPR on the PKIS dataset, FLMTS reached a score of 0.7463, which is 17.16% higher than the best IDDkin model in the state of the art (Fig. 3b). On the Tang dataset, the AUC and AUPR of the model are 0.69 and 5.59% higher than those of the NMF model, respectively (Fig. 3c, d). The performance of FLMTS is better than that of FLMTS + , which shows that the similarity networks of compounds and kinases contain enough information for predicting kinase inhibitors, while the interaction matrices Y and K may contain noisy data. In summary, FLMTS performs best among all the listed methods on the two datasets, indicating its excellent prediction capability.

Furthermore, both Tang and PKIS are very sparse datasets, and the number of negative samples is approximately 28 times and 15 times that of the positive samples in the PKIS and Tang datasets, respectively. When the dataset is too sparse, it is essential to correctly predict positive samples [42]. Therefore, four additional evaluation metrics, balanced-accuracy (BA), F1-score, recall, and precision, were used to comprehensively assess the performance of the model. As shown in Table 2, FLMTS achieved better performances than the other models. For instance, compared with the best IDDkin model in the state-of-the-art on the PKIS dataset, FLMTS has the smallest increase of 0.0267 and the largest increase of 0.0786 among the four metrics. Based on the Tang dataset, FLMTS also achieves the best results compared to the state-of-the-art methods. Moreover, we found that the results of FLMTS on PKIS were improved more significantly than the results of FLMTS on Tang, and the ratio of positive data over negative data on PKIS is much smaller than that of Tang. This shows that FLMTS has better practical application prospects; because the number of positive samples is extremely small in real-world scenarios. In general, Table 3 demonstrates the advantages of FLMTS over baselines on the two datasets.

3.3 Case Studies: Sorafenib, Vandetanib, Sunitinib

We selected three anticancer drugs approved by the FDA as the objects of the case studies, namely, sorafenib [43], vandetanib [44], and sunitinib [45]. They are not included in our datasets. Sorafenib, a small molecule B-RAF and VEGFR inhibitor, is a drug for the therapy of nephron-cell carcinoma and unresectable hepatocellular carcinoma. Vandetanib is the first systemic therapy drug for treating symptomatic or progressive advanced medullary thyroid cancer. Sunitinib, an inhibitor targeting PDGFR β and VEGFR2, has been approved by the FDA for the medical diagnosis and treatment of neuroendocrine tumors of the pancreas. For each drug (kinase inhibitor), Table 4 shows the predicted top ten kinases in the PKIS dataset. For sorafenib, seven of the ten top kinases were supported by the literature. The numbers for vandetanib and sunitinib were 5 and 4, respectively. For instance, the literature [46] indicates that the target kinases of sorafenib were PDGFR α < DDR2 < RET < HIPK4 < FLT 4 < FLT1 < KDR < PDGFR β < RAF1 < FLT3 in rank of IC_{50} values. Moreover, most of the predicted kinases focus on one kinase group. For example, the top ten predicted kinases of sorafenib focused on the tyrosine kinase (TK) group, which demonstrated the accuracy of FLMTS in the prediction of kinase inhibitors.

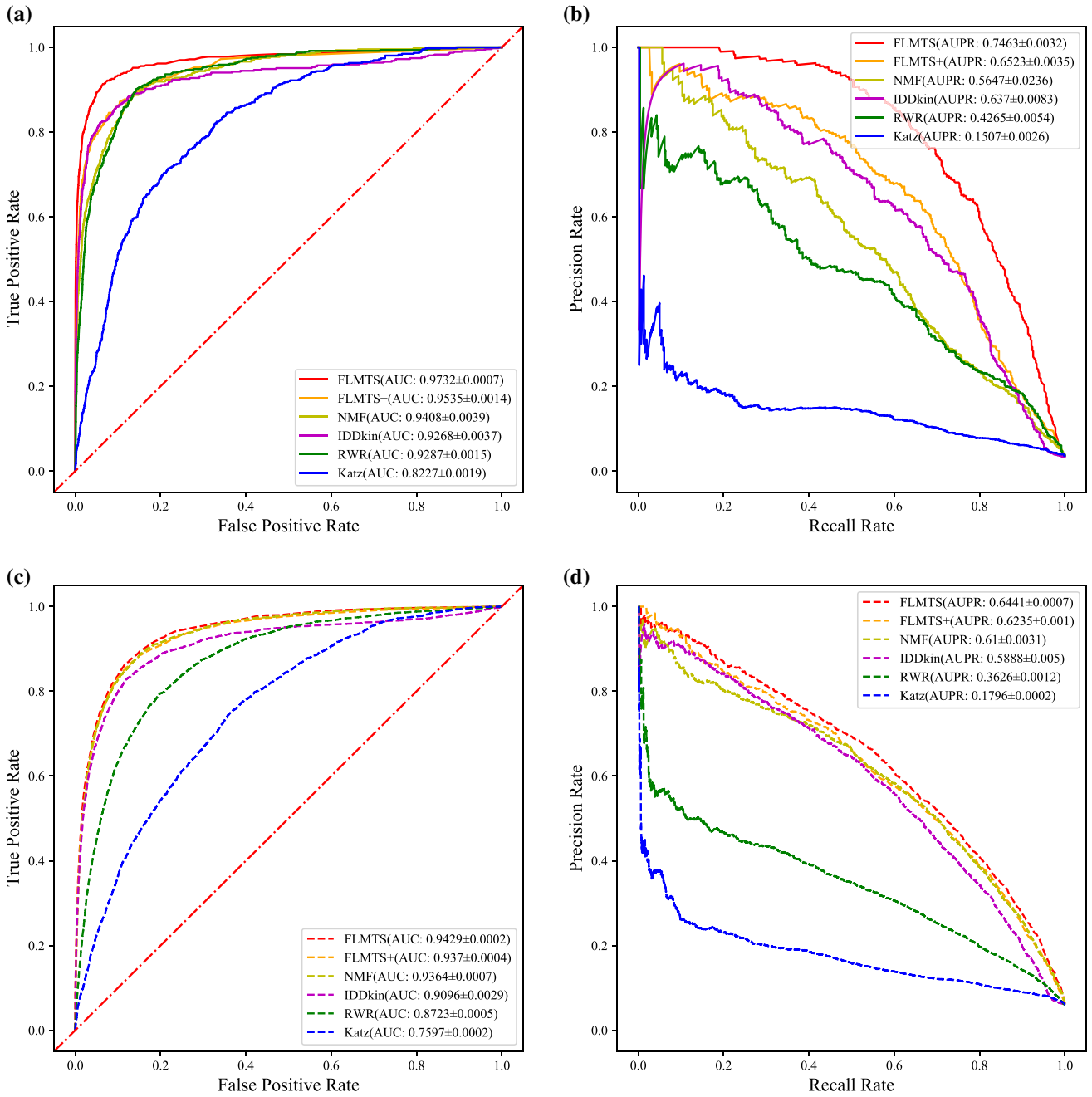


Fig. 3 Performance of different models. **a** ROC curves of FLMTS and comparison models on the PKIS dataset. **b** The AUPR of FLMTS and the comparison models on the PKIS dataset. **c** ROC curves of FLMTS and comparison models on the Tang dataset. **d** The AUPR of FLMTS and the comparison models on the Tang dataset

3.4 Pathway Enrichment Analysis

In cancer therapy, tyrosine kinase inhibitors can accurately target receptor tyrosine kinase signaling pathways, which can inhibit cell signal transduction, thereby inhibiting the growth and proliferation of cancer cells and promoting apoptosis [47]. Thus, it is particularly important to accurately target the signaling pathways in disease treatment. To further validate the practicability of FLMTS in treating disease, we carried out gene pathway enrichment analysis using the first 20 predicted genes (kinases) of sorafenib [43], which has been used in the treatment of unresectable hepatocellular carcinoma (HCC), advanced renal cell carcinoma (RCC), and differentiated thyroid carcinoma (DTC) from the DailyMed database (https://www.dailymed.nlm.nih.gov/). The pathway enrichment results are shown in Fig. 4.

Table 3 The experimental results of FLMTS with baselines on the two datasets

	Recall	Precision	F1-score	BA
PKIS				
FLMTS	**0.9187**	**0.3108**	**0.4643**	**0.9237**
FLMTS +	0.8469	0.2864	0.4279	0.8865
IDDkin	0.8401	0.2841	0.4245	0.8830
NMF	0.8011	0.2713	0.4052	0.8629
RWR	0.7620	0.2578	0.3851	0.8426
Katz	0.4249	0.1437	0.2148	0.6682
Tang				
FLMTS	**0.7523**	**0.4639**	**0.5738**	**0.8476**
FLMTS +	0.7362	0.4540	0.5616	0.8390
NMF	0.7361	0.4539	0.5615	0.8389
IDDkin	0.7071	0.4360	0.5393	0.8235
RWR	0.5489	0.3385	0.4187	0.7392
Katz	0.3173	0.1957	0.2420	0.6158

Bold values indicate the best results

Table 4 The first ten candidate kinases of the three selected compounds predicted by FLMTS

Compounds	Kinase	Group	Score	Evidence
Sorafenib	MUSK	TK	0.318	25,965,825
	KDR	TK	0.304	23,279,183
	PYK2	TK	0.229	23,153,798
	CSK	TK	0.227	–
	JAK3	TK	0.201	22,368,270
	FLT1	TK	0.201	23,279,183
	DDR2	TK	0.168	23,279,183
	PDGFRA	TK	0.155	19,212,337
	EphA3	TK	0.129	–
	EphA4	TK	0.127	–
Vandetanib	DDR2	TK	0.369	25,806,311
	ErbB2	Tk	0.347	17,136,225
	MSSK1	CMGC	0.287	–
	P38a	CMGC	0.198	–
	EphB4	TK	0.097	28,332,573
	CaMK1d	CAMK	0.057	–
	PDK1	AGC	0.042	17,136,225
	AurC	Other	0.041	–
	TSSK2	CAMK	0.038	–
	MAP2K1	STE	0.035	23,822,199
Sunitinib	MSSK1	CMGC	0.102	–
	DYRK2	CMGC	0.083	–
	GSK3A	CMGC	0.072	–
	P38a	CMGC	0.061	26,815,723
	PYK2	TK	0.041	–
	FLT4	TK	0.038	27,149,458
	GCK	STE	0.035	–
	FLT1	TK	0.034	27,149,458
	TSSK2	CAMK	0.030	–
	KDR	TK	0.030	27,149,458

Genes (kinases) were enriched in 12 pathways, and most of them were enriched in the first six pathways, including PI3K-Akt, focal adhesion, Rap1, Ras, calcium, and MAPK. This shows that sorafenib may target the 12 significant pathways and have a more significant effect on the first six pathways. In addition, we selected genes associated with diseases of HCC, RCC, and DTC with a threshold of $\text{Score}_{\text{gda}} = 0.5$ in the DisGeNET database (http://www.disgenet.org) and used these genes for pathway enrichment analysis. These genes were enriched in 96 pathways, suggesting that these pathways hold a critical relationship in the treatment of HCC, RCC, and DTC. The results are shown in the Supporting Materials. We found that 9 pathways could be found on both the 12 pathways of sorafenib and the 96 pathways that performed on the selected genes. Among the first 6 significant pathways of 12 pathways, 5 pathways (PI3K-Akt, focal adhesion, Rap1, Ras, and MAPK) were found among the 96 pathways. This shows that FLMTS could also accurately predict the signaling pathways in disease treatment and further explains the internal reason for the predicted these kinase inhibitors. In summary, the pathway enrichment analysis showed that the FLMTS model has excellent practicability and high accuracy as well as application value in the screening of potential kinase inhibitors.

4 Conclusion and Discussion

In this study, a new computational framework (FLMTS) was developed for predicting potential kinase inhibitors. This framework incorporated multisource information including gold standard compound–kinase interaction network information, the similarity information of compounds and kinases, and the structural information of compounds and kinases. The FLMTS includes four steps: (1) we calculated the structure similarity of compounds and kinases and obtained the interaction information of compound–kinase from datasets; (2) the similarity of compounds and kinases and the interaction information of compound–kinase were applied to structure a heterogeneous network; (3) FLMTS uses the RWR algorithm to obtain the topology information of nodes on the heterogeneous network; (4) the topology information as features of compounds and kinases were input to the random forest classifier for predicting the interaction score of compound–-kinase.

In conclusion, we presented a novel framework, named FLMTS, to enhance performance. First, to fuse multisource knowledge as features, we used the RWR to combine the topology information and sequence characteristics of nodes

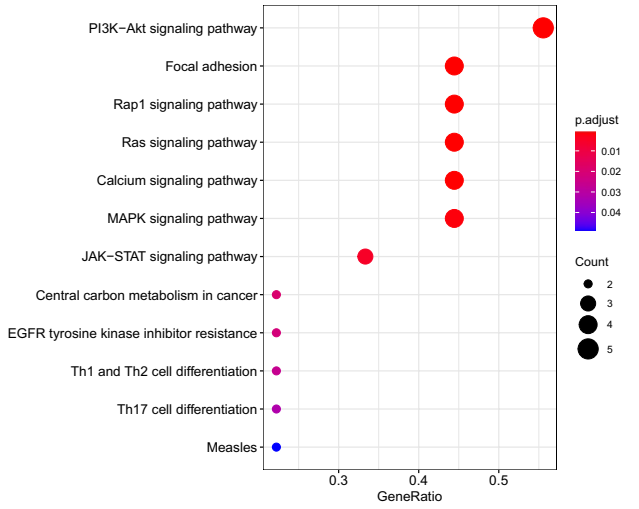


Fig. 4 Pathway enrichment analysis of the top 20 predicted genes (kinases) for sorafenib in the PKIS dataset

as the features. Second, to prevent overfitting based on limited a priori knowledge, RF was applied to predict the interaction of compounds and kinases. Compared with the state-of-the-art methods, the FLMTS has better performance. FLMTS makes the most of the heterogeneous network information, the structural information of the compounds and the kinases, and two similarity network knowledge, but the conventional models for predicting kinase inhibitors (i.e., MTDNN [17], PEPECFP [14]) do not apply the structural information of kinases and the two similar network knowledge. The models (i.e., NMF [31, 32], RWR [28, 29]) do not incorporate similarity network knowledge, and it is not surprising that their prediction performance is worse than that of FLMTS. Compared with some complex models (i.e., IDDkin [30]), the FLMTS could present better performance based on limited priori knowledge (e.g., 2414 positive samples in PKIS), since it may be less inclined to overfitting. In addition, compared with the variant model FLMTS+, FLMTS has less information but better performance. It may be that the interaction network is too sparse and the features of the interaction network contain too much noise, which reduces the performance of FLMTS+. In the end, case studies demonstrate that FLMTS is an effective method for improving the prediction of potential kinase inhibitors, and pathway enrichment analysis demonstrates that FLMTS could also accurately predict the signaling pathways in disease treatment and further explains the internal reason for the predicted result. In brief, diverse experiments demonstrate that FLMTS achieves excellent performance and has practical value for kinase inhibitor prediction.

In our future work, we will consolidate more useful interaction information such as inhibitor-disease interactions and relevant feature knowledge such as RNA sequences from other databases and literature and not only obtain the topology information in a heterogeneous network but also obtain semantic information. It may improve the model's prediction performance. Moreover, we will further study the binding mode of kinases and inhibitors and apply new methods to predict whether there is irreversible inhibition between potential kinase inhibitors and corresponding kinases. It is worth noting that irreversible kinase inhibitors offer many latent advantages over reversible kinase inhibitors [47]. Hence, it is worth studying whether the inhibitors are irreversible. As far as we know, there are some reversible inhibitors in the datasets (e.g., infigratinib phosphate [48] in Tang and GSK180736A [49] in PKIS) and few irreversible inhibitors in the datasets, but there are no divided into reversible or irreversible inhibitors in the Tang and PKIS datasets. Thus, we will conduct this research in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12539-022-00523-1.

Acknowledgements This work was supported in part by the Hunan Provincial Natural Science Foundation of China (No. 2019JJ50520), Hunan Provincial Innovation Foundation For Postgraduate (No. CX20210937).

Data Availability Source code and the dataset supporting the conclusions of this article are available from the corresponding author upon reasonable request.

Declaration

Conflict of Interest The authors declare that they have no competing interests.

References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298(5600):1912–1934. https://doi.org/10.1126/science.1075762
2. Levitzki A (2003) Protein kinase inhibitors as a therapeutic modality. Acc Chem Res 36(6):462–469. https://doi.org/10.1021/ar0201207
3. Muller S, Chaikuad A, Gray NS, Knapp S (2015) The ins and outs of selective kinase inhibitor development. Nat Chem Biol 11(11):818–821. https://doi.org/10.1038/nchembio.1938
4. Bhullar KS et al (2018) Kinase-targeted cancer therapies: progress, challenges and future directions. Mol Cancer 17(1):48. https://doi.org/10.1186/s12943-018-0804-2
5. Roskoski R Jr (2021) Properties of FDA-approved small molecule protein kinase inhibitors: a 2021 update. Pharm Res 165:105463. https://doi.org/10.1016/j.phrs.2021.105463
6. Fabbro D, Cowan-Jacob SW, Moebitz H (2015) Ten things you should know about protein kinases: IUPHAR review 14. Br J Pharm 172(11):2675–2700. https://doi.org/10.1111/bph.13096
7. Fedorov O, Muller S, Knapp S (2010) The (un)targeted cancer kinome. Nat Chem Biol 6(3):166–169. https://doi.org/10.1038/nchembio.297

8. Botta M (2014) New frontiers in kinases: special issue. ACS Med Chem Lett 5:270. https://doi.org/10.1021/ml500071m
9. Dickson M, Gagnon JP (2004) Key factors in the rising cost of new drug discovery and development. Nat Rev Drug Discov 3(5):417–429. https://doi.org/10.1038/nrd1382
10. Merget B, Turk S, Eid S, Rippmann F, Fulle S (2017) Profiling prediction of kinase inhibitors: toward the virtual assay. J Med Chem 60(1):474–485. https://doi.org/10.1021/acs.jmedchem.6b01611
11. Bora A, Avram S, Ciucanu I, Raica M, Avram S (2016) Predictive models for fast and effective profiling of kinase inhibitors. J Chem Inf Model 56(5):895–905. https://doi.org/10.1021/acs.jcim.5b00646
12. Cao D-S et al (2013) Large-scale prediction of human kinase–inhibitor interactions using protein sequences and molecular topological structures. Anal Chim Acta 792:10–18. https://doi.org/10.1016/j.aca.2013.07.003
13. Niijima S, Shiraishi A, Okuno Y (2012) Dissecting kinase profiling data to predict activity and understand cross-reactivity of kinase inhibitors. J Chem Inf Model 52(4):901–912. https://doi.org/10.1021/ci200607f
14. Avram S, Bora A, Halip L, Curpan R (2018) Modeling kinase inhibition using highly confident data sets. J Chem Inf Model 58(5):957–967. https://doi.org/10.1021/acs.jcim.7b00729
15. Yabuuchi H et al (2011) Analysis of multiple compound–protein interactions reveals novel bioactive molecules. Mol Syst Biol 7(1):472. https://doi.org/10.1093/bioinformatics/btaa577
16. Schürer SC, Muskal SM (2013) Kinome-wide activity modeling from diverse public high-quality data sets. J Chem Inf Model 53(1):27–38. https://doi.org/10.1021/ci300403k
17. Li X et al (2020) Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. J Med Chem 63(16):8723–8737. https://doi.org/10.1021/acs.jmedchem.9b00855
18. Manallack DT et al (2002) Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. J Chem Inf Comput Sci 42(5):1256–1262. https://doi.org/10.1021/ci020267c
19. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113. https://doi.org/10.1038/nrg1272
20. Cheng F, Kovács IA, Barabási A-L (2019) Network-based prediction of drug combinations. Nat Commun 10(1):1–11. https://doi.org/10.1038/s41467-019-09186-x
21. Ding P, Ouyang W, Luo J, Kwoh CK (2020) Heterogeneous information network and its application to human health and disease. Brief Bioinform 21(4):1327–1346. https://doi.org/10.1093/bib/bbz091
22. Luo Y et al (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8(1):573. https://doi.org/10.1038/s41467-017-00680-8
23. Shen C, Luo J, Ouyang W, Ding P, Wu H (2020) Identification of small molecule–miRNA associations with graph regularization techniques in heterogeneous networks. J Chem Inf Model 60(12):6709–6721. https://doi.org/10.1021/acs.jcim.0c00975
24. Cheng F et al (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun 9(1):2691. https://doi.org/10.1038/s41467-018-05116-5
25. Xuan P et al (2019) Drug repositioning through integration of prior knowledge and projections of drugs and diseases. Bioinformatics 35(20):4108–4119. https://doi.org/10.1093/bioinformatics/btz182
26. Ding P, Yin R, Luo J, Kwoh CK (2019) Ensemble prediction of synergistic drug combinations incorporating biological, chemical, pharmacological, and network knowledge. IEEE J Biomed Health Inform 23(3):1336–1345. https://doi.org/10.1109/JBHI.2018.2852274
27. Chen X, Huang YA, You ZH, Yan GY, Wang XS (2017) A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics 33(5):733–739. https://doi.org/10.1093/bioinformatics/btw715
28. Li Y, Patra JC (2010) Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics 26(9):1219–1224. https://doi.org/10.1093/bioinformatics/btq108
29. Lv YL et al (2015) Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. Bioinformatics 31(22):3638–3644. https://doi.org/10.1093/bioinformatics/btv417
30. Shen C, Luo J, Ouyang W, Ding P, Chen X (2020) IDDkin: Network-based influence deep diffusion model for enhancing prediction of kinase inhibitors. Bioinformatics 36(22–23):5481–5491. https://doi.org/10.1093/bioinformatics/btaa1058
31. Xiao Q, Luo J, Liang C, Cai J, Ding P (2018) A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. Bioinformatics 34(2):239–248. https://doi.org/10.1093/bioinformatics/btx545
32. Jamali AA, Kusalik A, Wu F-X (2020) MDIPA: a microRNA–drug interaction prediction approach based on non-negative matrix factorization. Bioinformatics 36(20):5061–5067. https://doi.org/10.1093/bioinformatics/btaa577
33. Davis MI et al (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051. https://doi.org/10.1038/nbt.1990
34. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol 29(11):1039–1045. https://doi.org/10.1038/nbt.2017
35. Metz JT et al (2011) Navigating the kinome. Nat Chem Biol 7(4):200–202. https://doi.org/10.1038/nchembio.530
36. Elkins JM et al (2016) Comprehensive characterization of the published Kinase Inhibitor Set. Nat Biotechnol 34(1):95–103. https://doi.org/10.1038/nbt.3374
37. Knapp S et al (2013) A public-private partnership to unlock the untargeted kinome. Nat Chem Biol 9(1):3–6. https://doi.org/10.1038/nchembio.1113
38. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11(23–24):1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005
39. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197. https://doi.org/10.1016/0022-2836(81)90087-5
40. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4):949–958. https://doi.org/10.1016/j.ajhg.2008.02.013
41. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43. https://doi.org/10.1007/BF02289026
42. Zeng X et al (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics 35(24):5191–5198. https://doi.org/10.1093/bioinformatics/btz418
43. Wilhelm S et al (2006) Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. Nat Rev Drug Discov 5(10):835–844. https://doi.org/10.1038/nrd2130
44. Wells SA Jr et al (2010) Vandetanib for the treatment of patients with locally advanced or metastatic hereditary medullary thyroid cancer. J Clin Oncol 28(5):767. https://doi.org/10.1200/JCO.2009.23.6604

45. Motzer RJ et al (2007) Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. N Engl J Med 356(2):115–124. https://doi.org/10.1056/NEJMoa065044
46. Kitagawa D et al (2013) Activity-based kinase profiling of approved tyrosine kinase inhibitors. Genes Cells 18(2):110–122. https://doi.org/10.1111/gtc.12022
47. Ferguson FM, Gray NS (2018) Kinase inhibitors: the road ahead. Nat Rev Drug Discov 17(5):353–377. https://doi.org/10.1038/nrd.2018.21
48. Tang LWT et al (2021) Infigratinib is a reversible inhibitor and mechanism-based inactivator of cytochrome P450 3A4. Drug Metab Dispos 49(9):856–868. https://doi.org/10.1124/dmd.121.000508
49. Wang H et al (2022) Decreased CXCR2 expression on circulating monocytes of colorectal cancer impairs recruitment and induces Re-education of tumor-associated macrophages. Cancer Lett 529:112–125. https://doi.org/10.1016/j.canlet.2022.01.004