Low Sample Kinase Inhibitory Activity Prediction Capability of Multi-task Deep Convolutional Neural Networks Under Knowledge-based Molecular Representations

HUO Wenbo[1], TAN Ying[1*], SHEN Wanxiang[2], JIANG Yuyang[1], CHEN Yuzong[3], CHEN Yan[1](*1.Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China; 2.National University of Singapore, Singapore 117543, Singapore; 3.Shenzhen Bay Laboratory, Shenzhen 518000, China*)

ABSTRACT: To facilitate drug discovery, deep learning models have been developed for the prediction of inhibitors of various targets including kinases, achieving high prediction performances. Nonetheless the ability of deep learning on low-sample targets (<100 known active molecules) has not been adequately tested. Leveraging the good activity prediction capability of a recently emerged deep convolutional neural network MolMapNet method under knowledge-based molecular representations, this study developed multi-task MolMapNet models for inhibitory activity prediction of 19 low-sample kinases and 43 higher-sample kinases of 6 kinase subfamilies. The developed multi-task MolMapNet models for all low-sample and higher-sample kinases significantly enhanced the activity prediction performance over the single-task models. The activity prediction indicators such as R^2 values were in the good performance ranges of 0.651 3–0.749 8 for most kinases. These suggest the usefulness of the multi-task transfer learning strategy in activity prediction of low-sample targets.

KEYWORDS: deep learning; kinase inhibitors; activity prediction; multi-task learning; regression method; drug discovery

基于先验知识的分子表征多任务深度卷积神经网络的小样本激酶抑制剂活性预测

霍文博[1]，谭英[1*]，申万祥[2]，蒋宇扬[1]，陈宇综[3]，陈妍[1](1.清华大学深圳国际研究生院，广东 深圳 518055；2.新加坡国立大学，新加坡 117543；3.深圳湾实验室，广东 深圳 518000)

摘要：为了推动药物研发，深度学习模型被开发用于预测包括激酶在内的各种靶点的抑制剂，且能够达到较好的预测性能。尽管如此，深度学习在小样本靶点(<100 个已知的活性化合物)上的预测性能还没有得到充分的测试。本研究利用最近出现的深度卷积神经网络 MolMapNet 方法在基于先验知识的分子表示下具有的良好的活性预测能力，开发出了多任务 MolMapNet 模型，用于预测 6 个激酶亚家族的 19 个小样本激酶和 43 个大样本激酶的抑制剂活性。开发的用于所有小样本和大样本激酶的多任务 MolMapNet 模型明显比单任务模型具有更好的活性预测性能。R^2 值等活性预测指标在 0.651 3~0.749 8 内表现较好。这证明了多任务迁移学习在小样本靶标活性预测中的鲁棒性。

关键词：深度学习；激酶抑制剂；活性预测；多任务学习；回归策略；药物研发

中图分类号：R914.2　　**文献标志码：**A　　**文章编号：**1007-7693(2022)21-2819-09

DOI: 10.13748/j.cnki.issn1007-7693.2022.21.018

引用本文：霍文博, 谭英, 申万祥, 等. 基于先验知识的分子表征多任务深度卷积神经网络的小样本激酶抑制剂活性预测[J]. 中国现代应用药学, 2022, 39(21): 2819-2827.

Improvement of the efficiency of new drug development remains a key issue and challenge to the pharmaceutical communities. Machine learning and deep learning(DL) methods have been extensively explored in biomedicine and drug discovery, achieving promising performances with artificial intelligence assisted drugs(e.g. ISM00055[1]) entering clinical trials. Pharmaceutical DL can be based on different various representations of molecular properties, which include molecular graph-based feature representations[2-5], molecular string-based representations[6-8], image-based representations[9-10], and knowledge-based molecular representations[11-12]. These DL methods achieved promising performances in the prediction of various pharmaceutical properties such as activities, pharmacokinetic properties and toxicological properties. Nonetheless, the ability of these DL methods on low-sample targets(with <100 known active molecules) has not been adequately tested. Low-sample targets are important in drug discovery because they are mostly in the earlier-stages of discovery processes, typically in need for finding more active molecules as potential drug candidates. Therefore, the development and testing of DL methods for low-sample targets are useful for facilitating earlier-stage drug discovery.

作者简介：霍文博，男，硕士　E-mail: 1029730566@qq.com　***通信作者：**谭英，女，博士，副教授　E-mail: tan.ying@sz.tsinghua.edu.cn

In this work, we probed this question by testing the ability of a newly developed MolMapNet deep convolutional neural network method under the knowledge-based molecular representations[12] in the prediction of inhibitors of low-sample kinase targets. MolMapNet was selected because of its good performances on multiple benchmark datasets, including multi-task datasets[12]. We specifically developed multi-task MoMapNet models for 6 kinase subfamilies, each containing both low-sample kinases(<100 known inhibitors) and higher-sample kinases(>250 known inhibitors). The performance of these models on the low-sample kinases as well as higher-sample kinases were evaluated. MolMapNet explores broadly leaned knowledge-based molecular representations for DL of pharmaceutical properties. The molecular features were derived from the MolMap algorithms trained from 8 million compounds from the PubChem database using a hierarchical method[12]. This enables efficient learning of pharmaceutically relevant molecular features by a CNN architecture. In this work, MolMapNet regression prediction models were developed for predicting the activity of kinase inhibitors[13].

Kinases constitute a large family of therapeutic targets, and many kinases are the targets of approved and clinical trial drugs[14]. Nonetheless, at least 30% of kinases are still underdeveloped[15]. For instance, LATS2 of the NDR subfamily is a key target for malignant peritoneal mesothelioma[16], but there are only 27 known inhibitors in the ChEMBL database, and there is no approved drug for this target. It is very difficult to develop DL models for such low-sample target, and special strategies are needed[17]. The core principle of low-sample DL is to achieve the effect of higher-sample DL based on the existing data, albeit being far less than the typical data-sizes of conventional DL[18]. One useful strategy is the transfer learning[19]. In particular, the multi-task strategy allows the collective learning of low-sample and higher-sample targets of the same subfamily under the same DL architecture, wherein the features learned from higher-sample targets can be transferred for the learning of low-sample targets. Therefore, we aimed to establish a multi-task regression DL model for the activity prediction of low-sample kinase targets based on the MolMapNet multi-task architecture[12].

1 Material and methods

1.1 Data collection and processing

Human kinase targets were searched from the ChEMBL database, and the inhibitor activity data of the identified targets were downloaded through the ChEMBL_websource_client on the ChEMBL database[15]. In order to eliminate the dimensional gap and the variation degree of kinase inhibitor activity data, the standard pChEMBL is used to represent the activity value of inhibitors, allowing comparison of multiple roughly comparable measurements on a negative log scale, which is defined as: −Log(molar IC_{50}, EC_{50}, K_i)+9. The activity values based on these three indicators are combined by a Python code. For the inhibitors with multiple activity values against the same, kinase, the median activity value was tentatively chosen as the activity values. We selected three H-type kinase subfamilies dominated by higher-sample(>200 known inhibitors) kinases and three L-type kinase subfamilies dominated by low-sample(<100 known inhibitors) kinases(Tab. 1). The list of kinases and the number of inhibitors of these subfamilies are provided in Supplementary Tab. 1. The inhibitors of all kinases of each subfamily forms a dataset for developing multi-task DL models. These datasets were normalised based on established normalization algorithm[20].

Tab. 1 List of the H-type kinase subfamilies dominated by higher-sample(>200 inhibitors) kinases and L-type kinase subfamilies dominated by low-sample kinases(<100 inhibitors)

表 1 H-型激酶亚家族(靶点抑制剂数目>200)和 L-型激酶亚家族(靶点抑制剂数目<100)信息汇总

Type	Subfamily	HSKs≥200	LSKs≤100	HSK *vs* LSK ratio
H-type subfamily	Tyrosine protein kinase EGFR subfamily	4	1	4
	AGC protein kinase AKT subfamily	3	1	3
	Atypical protein kinase PIKK subfamily	4	2	2
L-type subfamily	TKL protein kinase RAF subfamily	2	3	0.67
	CMGC protein kinase CDK subfamily	12	27	0.44

Note: Each subfamily was used for developing multi-task deep learning models to predict the inhibitors of each kinase in the subfamily. HSKs indicates number of higher-sample kinases, and LSKs indicates number of low-sample kinases.

注：每个亚家族被用于构建多任务深度学习模型，以预测亚家族中每个激酶的抑制剂活性。HSKs 表示高样本激酶的数量，LSKs 表示低样本激酶的数量。

1.2 MolMap molecular representations

The SMILES codes of kinase inhibitors were converted to standard SMILES codes by RDKit. The MolMap software package can be used to convert the standard SMILES code of inhibitors into advanced two-dimensional features such as molecular descriptors and fingerprints, and further build a two-dimensional molecular feature map, embedding the feature relationship into the two-dimensional space by means of UMAP. In order to ensure the efficiency of model training, it is

essential to remove the repetitive inhibitor SMILES sequence and the ligand activity files with no activity value before transformation[21].

1.3 MolMapNet architecture and hyperparameters

MolMapNet architecture is in Fig. 1. Each model has comparatively few parameters, which uses an early stop strategy to reduce overfitting and computational costs. We performed stratiKFOLD 10-fold cross validation for each model. After the activity value data of small molecule inhibitors against kinase targets were divided into 0, 1, 2, 3, 4, 5, 6 and 7 categories according to (0, 10], (10, 100], (100, 1 000], (1 000, 10 000], (10 000, 100 000], (100 000, 1 000 000], (1 000 000, 10 000 000], (10 000 000, infinity), data was stratified by means of stratiKFOLD. The training of the model is carried out by inputting the divided training set, and then the test set tests the training effect of the model, makes a prediction, and then obtains the evaluation result. The verification set is mainly used to evaluate the results and determine how to adjust the super parameters of the model[22-23].

1.4 Model evaluation

Three metrics were used for evaluating our developed models: R^2, MAE, RMSE. R^2 reflects the linear correlation between the predicted value and the real value of the model, ranging from −1 to 1. The greater the absolute value of R^2, the stronger the linear correlation between the predicted value and the real value. RMSE represents the degree of deviation between the predicted value and the real value. MAE is the average value of the difference between the predicted value and the real value after taking the absolute value. For R^2, the higher and closer the value to 1, the better the efficiency of the predictive model, whereas for MAE and RMSE, the lower the value and closer to 0, the better the predictive performance of the model[24].

1.5 Model optimization

We utilize dual-input channel to input data converted into descriptors and fingerprints into MolMapNet model for operation. When the number of kinase target inhibitors is <50, the number of compounds in part of the fold may be <2, so that the Nan value needs to be returned to avoid the abnormal interruption of the program caused by the inability of R^2 to calculate. If the test set data is <1, the Nan value needs to be returned to avoid program exceptions caused by RMSE calculation interruption[25]. In the regression task, the loss function is generally set as mean squared error. After many updating iterations, it has proved that the fitting efficiency of the primary loss function is not high. We introduced MASK and Pos_weight[i] parameter to improve the fitting efficiency of the model. The loss function is multiplied by MASK. The core function of MASK is to eliminate the influence of the vacancy value in the data files on the loss function, replacing the active null value in targets and distinguishing it from the non null value samples[26]. In the regression, the value of MASK should be greater than the maximum active value, otherwise the convergence anomaly of the model loss function will occur. Pos_weight[i] is defined as total row number of the data column divided by the effective row number. Pos_weight[i] is able to weight the data and reduce the error caused by the target with small data volume when optimizing the parameters of the loss function during model training and prediction[27].

2 Results

2.1 Inhibitory activity prediction performance of multi-task MolMapNet on the H-type kinase subfamilies dominated by higher-sample kinases

We first evaluated the performance of multi-task MolMapNet models on three H-type kinase subfamilies dominated by higher-sample kinases. The three subfamilies are ERBB, AKT, PIKK with 4 and 1, 3 and 1, 4 and 2 higher-sample and low-sample kinases respectively(Tab. 1). ERBB subfamily of kinases are important for non small cell lung cancer and breast cancer treatment with multiple approved drugs[15]. AKT and PIKK kinases have been explored for anticancer therapeutics[28-29]. With sufficient number of inhibitors for vast majority of kinases, these three subfamilies represent the relatively-easier DL tasks. The multi-task MolMapNet models achieved good overall performances for the low-sample and higher-sample kinases in the ERBB, AKT and PIKK subfamilies(Tab. 2). For the low-sample kinases in these three subfamilies, the R^2 value are 0.749 8, 0.668 3, and 0.559 4 respectively, the RMSE values are 0.791 9, 0.410 9 and 0.991 6, and the MAE values are 0.673 2, 0.351 0 and 0.738 9. For the higher-sample kinases in these three subfamilies, the R^2 value are in the range of 0.337 9–0.749 8, 0.668 3–0.841 1, 0.180 8–0.274 3 respectively, the RMSE values are in the range of 0.527 9–0.861 1, 0.410 9–0.692 8, 0.695 0–1.065 6 respectively, and the MAE values are in the range of 0.361 8–0.692 1, 0.347 4–0.526 2, 0.568 7–0.907 5 respectively. In the evaluation of regression models such as QSAR models, the threshold of R^2 values is 0.6[30]. Therefore, the multi-task MolMapNet models are of good performances in four CDK kinase inhibitors have been approved for the inhibitory activities of low-sample and higher-sample kinases in the H-type kinase subfamilies.

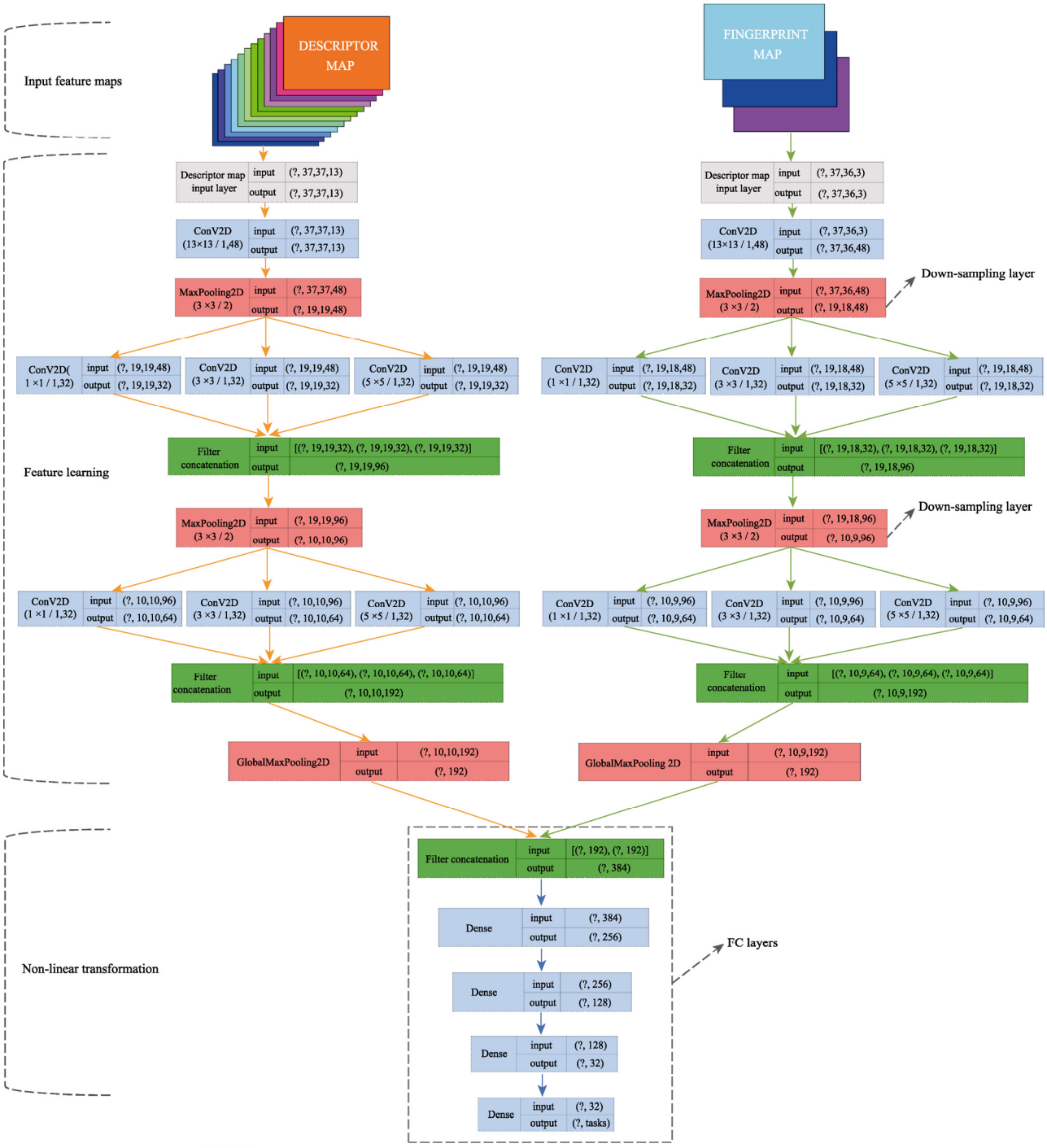

Fig. 1 Workflow of MolMapNet architecture[11]

Left input: descriptor map; right input: fingerprint map. Trainable parameters: left single path is 0.40 million; right single path is 0.32 million; both path is 0.80 million.

图 1 MolMapNet 架构图示

左输入：描述符；右输入：分子指纹。可训练参数：左单路径为 40 万；右单路径为 32 万；双路径为 80 万。

2.2 Inhibitory activity prediction performance of multi-task MolMapNet on the L-type kinase subfamilies dominated by low-sample kinases

We then evaluated the performance of multi-task MolMapNet models on three L-type kinase subfamilies dominated by low-sample kinases. The three subfamilies are RAF, CDK, and MAPK with 2 and 2, 12 and 8, and 10 and 6 higher-sample and low-sample kinases respectively(Tab. 1). The BRAF V600E/K mutant of RAF subfamilies is the target of approved drugs for the treatment of melanoma, and intensified efforts have been directed at the development of RAF inhibitors as potential anticancer therapeutics[15,31]. So far, four CDK kinase inhibitors have been approved for anticancer therapeutics and on-going efforts are being directed

Tab. 2 Inhibitory activity prediction performance of multi-task MolMapNet on each of the four H-type kinase subfamilies EGFR, AKT, PIKK and RAF

表 2 多任务 MolMapNet 对 4 个 H 型激酶亚家族 EGFR、AKT、PIKK 和 RAF 的抑制剂活性预测结果

Type	Kinase	NOI	R^2	RMSE	MAE
EGFR	EGFR, ERB3, ERBB4	254	0.703 7	0.702 3	0.550 5
subfamily	ERBB4	901	0.736 9	0.527 7	0.357 2
	ERBB2	3 127	0.676 6	0.635 3	0.459 4
	EGFR	9 425	0.681 2	0.821 7	0.621 1
AKT	AKT3	1 082	0.782 1	0.532 5	0.350 8
subfamily	AKT2	1 699	0.836 0	0.523 7	0.379 6
	AKT1	4 074	0.780 4	0.677 4	0.513 6
RAF	RAF1	1 597	0.772 5	0.688 1	0.510 0
subfamily	BRAF	4 685	0.733 1	0.744 9	0.582 6

Note: Average R^2, RMSE and MAE of 10-fold cross-validation results are shown. NOI indicates the number of kinase inhibitors.

注：结果展示了 10 倍交叉验证结果的平均 R^2、RMSE 和 MAE。NOI 表示激酶抑制剂的数量。

at the development of multi-target drugs and target selective drugs that avoid the non-ideal CDK isoforms[15,32]. Members of MAPK subfamilies are part of key components in cellular signaling networks, which are being explored as potential targets for CNS diseases[33] and cancers[34]. The multi-task MolMapNet models produced good performances for the low-sample and higher-sample kinases in the RAF, CDK, and MAPK subfamilies (Tab. 3–5, Fig. 2–4). For the low-sample kinases in these three subfamilies, the R^2 value are 0.696 3–0.746 0, 0.327 7–0.745 5, and 0.431 1–0.696 2 respectively, the RMSE values are 0.532 9, 0.350 5–0.800 0, and 0.713 6–1.171 6. The MAE values are 0.380 4, 0.278 9–0.706 1 and 0.530 2–0.923 9. For the higher-sample kinases in these three subfamilies, the R^2 value are in the range of 0.671 4–0.766 8, 0.337 2–0.665 4, and 0.678 9–0.802 2 respectively, the RMSE values are in the range of 0.715 3–1.675 5, 0.678 5–0.884 0 and 0.252 0–0.904 0 respectively, and the MAE values are in the range of 0.540 3–1.481 1, 0.502 5–0.669 3 and 0.170 9– 0.923 9 respectively. Our studies showed the good ability of the multi-task MolMapNet models in inhibitory activity prediction for both low-sample and higher-sample kinases in the L-type subfamilies.

Tab. 3 Inhibitory activity prediction performance of multi-task MolMapNet on the L-type RAF subfamily

表 3 多任务 MolMapNet 对 L 型 RAF 亚家族的抑制剂活性预测结果

Kinase	NOI	R^2	RMSE	MAE
ARAF	94	0.746 0	0.532 9	0.380 4
RAF1, MP2K1	137	0.696 3	0.687 1	0.519 6
RAF1	1597	0.766 8	0.715 3	0.540 3
BRAF	4685	0.671 4	1.675 5	1.481 1

Note: Average R^2, RMSE and MAE of 10-fold cross-validation results are shown. NOI indicates the number of kinase inhibitors.

注：结果展示了 10 倍交叉验证结果的平均 R^2、RMSE 和 MAE。NOI 表示激酶抑制剂的数量。

Tab. 4 Inhibitory activity prediction performance of multi-task MolMapNet on the L-type CDK subfamily

表 4 多任务 MolMapNet 对 L 型 CDK 亚家族的抑制剂活性预测结果

Kinase	NOI	R^2	RMSE	MAE
CDK14	24	0.669 7	0.524 1	0.425 9
CDK1,CCND3 dual kinase	30	0.745 5	0.350 5	0.278 9
CDK8,CDK19 dual kinase	55	0.327 7	0.757 3	0.571 3
CCNC	62	0.528 4	0.705 7	0.531 2
CCNA1	68	0.645 3	0.941 6	0.706 1
CDK1,CCNA2 dual kinase	77	0.383 9	0.763 4	0.554 3
CCNK	92	0.665 5	0.772 1	0.551 5
CCNY	96	0.447 5	0.800 0	0.642 0
CDK19	205	0.337 2	0.803 7	0.560 2
CCNC	285	0.384 2	0.884 0	0.669 3
CCNB2,CCNB3 dual kinase	586	0.579 3	0.846 7	0.631 2
CDK7	589	0.442 2	0.714 2	0.536 8
CDK8	705	0.569 9	0.846 2	0.641 3
CDK1,CCNB1 dual kinase	939	0.617 2	0.795 3	0.584 2
CDK2,CCNA2 dual kinase	930	0.583 3	0.826 7	0.628 3
CDK9	1170	0.631 4	0.678 5	0.502 5
CDK5	1305	0.510 1	0.684 3	0.504 7
CCNA1	1303	0.665 4	0.799 9	0.611 7
CDK1	2223	0.625 9	0.746 6	0.560 1
CDK2	3285	0.651 3	0.788 4	0.600 8

Note: Average R^2, RMSE and MAE of 10-fold cross-validation results are shown. NOI indicates the number of kinase inhibitors.

注：结果展示了 10 倍交叉验证结果的平均 R^2、RMSE 和 MAE。NOI 表示激酶抑制剂的数量。

Tab. 5 Inhibitory activity prediction performance of multi-task MolMapNet on the L-type MAPK subfamily

表 5 多任务 MolMapNet 对 L 型 MAPK 亚家族的抑制剂活性预测结果

Kinase	NOI	R^2	RMSE	MAE
MARK1	24	0.646 5	0.894 0	0.744 3
AAPK2	26	0.696 2	0.843 3	0.600 9
SIK2	51	0.522 3	1.065 1	0.825 1
KCC4	53	0.582 4	0.713 6	0.530 2
MARK4	72	0.431 1	0.841 3	0.577 3
NUAK1	75	0.618 6	1.171 6	0.923 9
AAPK2, AAKG1, AAKB1 multi-kinase	103	0.798 8	0.680 8	0.539 7
MARK1	245	0.709 1	0.894 0	0.744 3
KCC1D	375	0.604 7	0.515 3	0.387 5
AAPK2, AAKB2, AAKG1 multi-kinase	396	0.802 2	0.670 9	0.485 4
BRSK1	623	0.204 8	0.514 6	0.359 3
MARK2	662	0.266 2	0.556 8	0.401 1
KCC1A	711	0.238 9	0.395 9	0.257 9
MARK3	842	0.491 0	0.613 0	0.411 2
PASK	855	0.538 4	0.252 0	0.170 9
AAPK1	946	0.209 0	0.629 6	0.461 7
MELK	1 419	0.792 7	0.639 6	0.480 3
CHK1	3 202	0.631 3	0.904 0	0.699 4

Note: Average R^2, RMSE and MAE of 10-fold cross-validation results are shown. NOI indicates the number of kinase inhibitors.

注：结果展示了 10 倍交叉验证结果的平均 R^2、RMSE 和 MAE。NOI 表示激酶抑制剂的数量。

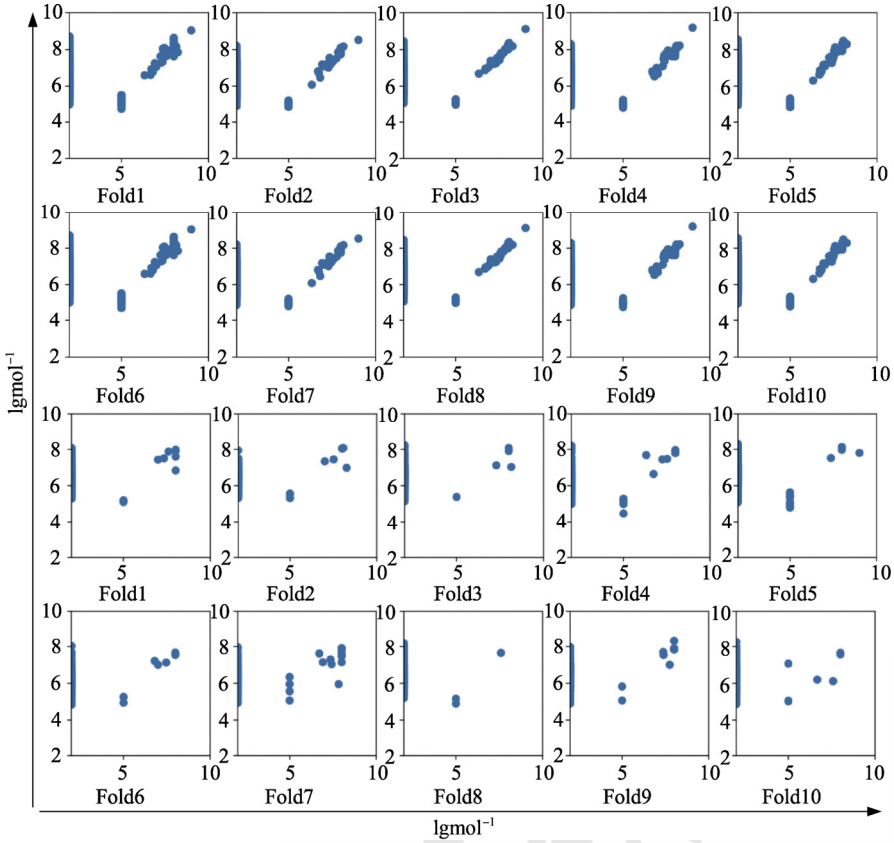

Fig. 2 RAF target (CHEMBL1169596) multi-task regression training and testing ten-fold scatter plot

图 2 RAF 靶点(CHEMBL1169596)多任务回归训练和测试十折散点图

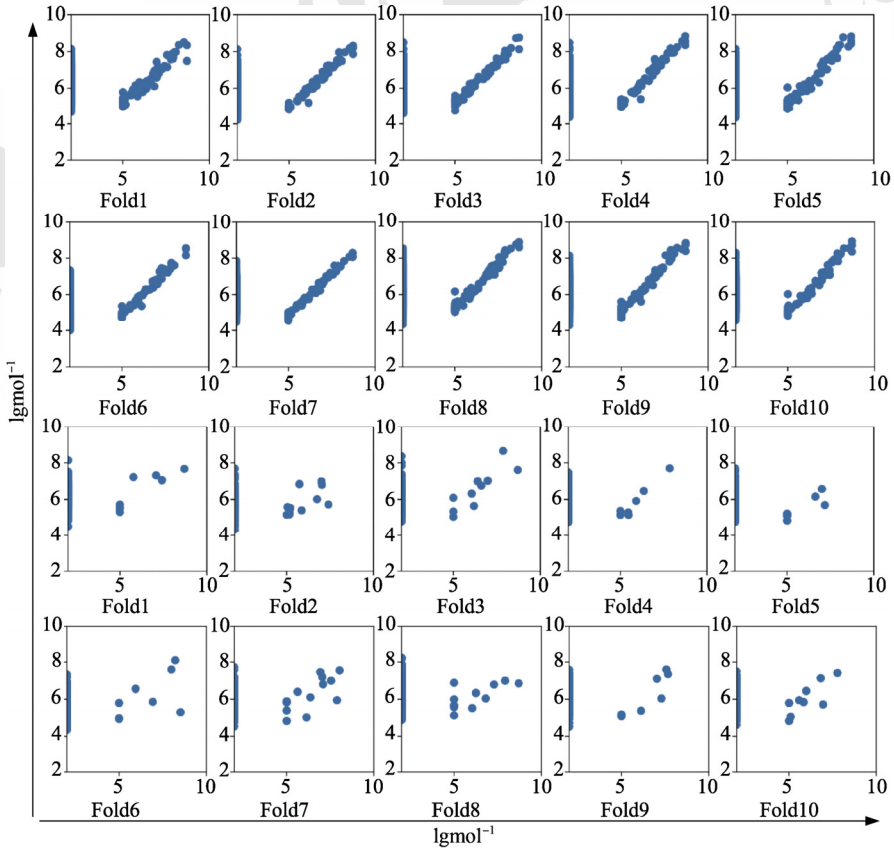

Fig. 3 CNKK target (CHEMBL3038475) multi-task regression training and testing ten-fold scatter plot

图 3 CNKK 靶点(CHEMBL3038475)多任务回归训练和测试十折散点图

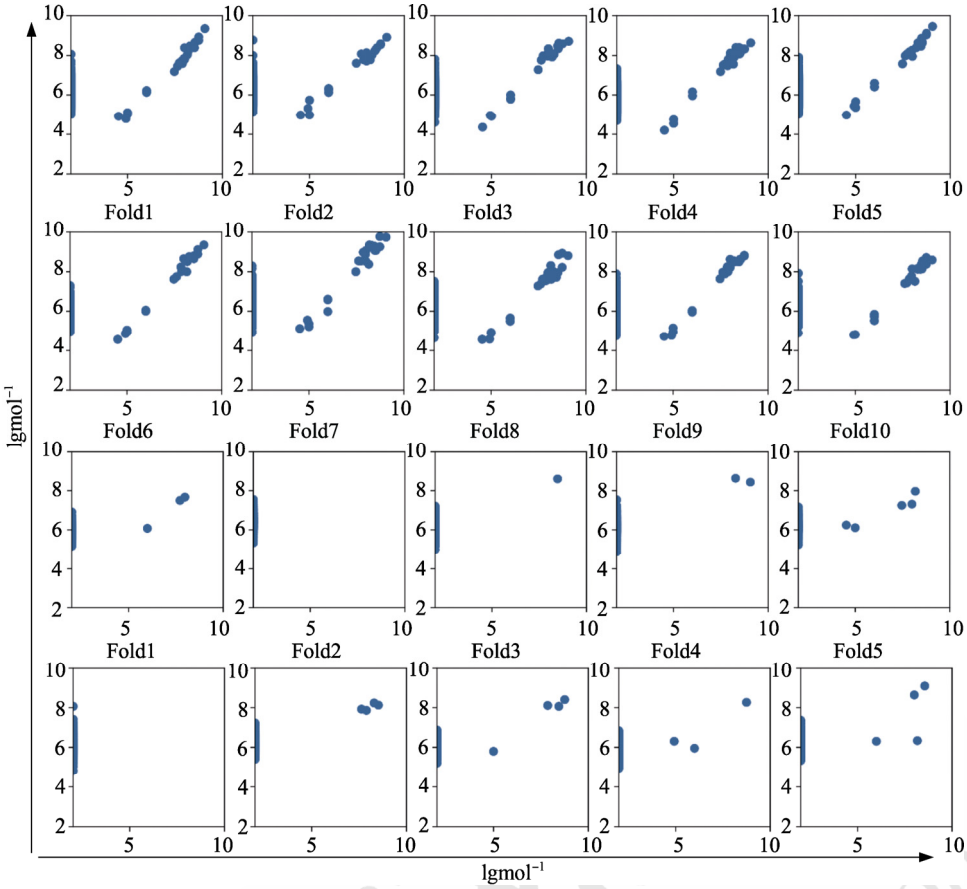

Fig. 4 AAPK2 target(CHEMBL2116) multi-task regression training and testing ten-fold scatter plot

图 4 AAPK2 靶点(CHEMBL2116)多任务回归训练和测试十折散点图

2.3 Comparison of the inhibitory activity prediction performance of multi-task MolMapNet and single-task MolMapNet on low-sample kinases

Our study indicated that multi-task DL strategies are capable of enhanced inhibitory activity prediction for low-sample targets. To further evaluate this capability, we conducted an additional study to compare the inhibitory activity prediction performance of multi-task and single-task MolMapNet models on 6 low-sample kinases of the CDK, RAF, and CAMK1 subfamilies. Tab. 6 provides the average R^2, RMSE and MAE of 10-fold cross-validation results for the multi-task and single-task MolMapNet models of each kinase

Tab. 6 Comparison of the inhibitory activity prediction performance of multi-task and single-task MolMapNet models on 2, 2 and 3 low-sample kinases of the CDK, RAF and MAPK subfamilies

表 6 多任务和单任务 MolMapNet 模型分别对 CDK、RAF 和 MAPK 亚家族中 2，2 和 3 个低样本激酶的抑制剂活性预测性能比较

Family Name	Target Name	CHEMBL_ID	NOI	R^2_m	R^2_s	R^2_(m-s)/s/%	RMSE_m	RMSE_s	RMSE_(m-s)/s/%	MAE_m	MAE_s	MAE_(m-s)/s/%
CDK	CCNK	CHEMBL3038475	92	0.605 7	0.616 2	**−1.71**	0.859 6	0.707 7	**21.47**	0.645 3	0.581 6	**10.95**
CDK	CCNY	CHEMBL4296115	96	0.403 4	0.268 5	**50.24**	0.803 7	0.763 7	**5.24**	0.615 1	0.638 3	**−3.63**
CAMK1	MARK4	CHEMBL5754	72	0.524 8	0.247 5	**112.03**	0.756 6	0.785 0	**−3.61**	0.543 6	0.626 9	**−13.29**
CAMK1	NUAK1	CHEMBL5784	75	0.512 2	0.325 6	**57.31**	1.196 4	1.365 0	**−12.35**	0.990 0	1.151 1	**−13.99**
CAMK1	AAPK2, AAKG1, AAKB1	CHEMBL3038455	103	0.814 0	0.867 1	**−6.12**	0.610 6	0.543 3	**12.39**	0.466 3	0.418 5	**11.42**
RAF	ARAF	CHEMBL1169596	94	0.850 0	0.652 5	**30.26**	0.531 8	0.730 8	**−27.24**	0.377 8	0.619 6	**−39.02**
RAF	RAF1, MP2K1	CHEMBL2111351	137	0.640 3	0.609 8	**4.99**	0.715 5	0.695 4	**2.89**	0.547 4	0.556 2	**−1.59**

Note: R^2_m, and R^2_s, RMSE_m and RMSE_s, and MAE_m and MAE_s are the average R^2, RMSE and MAE of 10-fold cross-validation results for the multi-task and single-task MolMapNet models respectively. The (m-s)/s value represents the relative difference of the R^2, RMSE, and MAE values of the multi-task and single-task models. The bold characters indicates the positive improvement of the multi-task model over single-task model(increased R^2 or decreased RMSE or MAE).

注：R^2_m 和 R^2_s、RMSE_m 和 RMSE_s、MAE_m 和 MAE_s 分别为多任务和单任务 MolMapNet 模型 10 倍交叉验证结果的平均值 R^2、RMSE 和 MAE。(m-s)/s 值表示多任务模型和单任务模型的 R^2、RMSE 和 MAE 值的相对差值。加粗字体表示多任务模型较单任务模型有积极的改进(增加 R^2 或降低 RMSE 或 MAE)。

respectively. For the CDK subfamily, there are 50% of the low-sample kinases with the relative R^2 values increased by 50.24%, though both of the low-sample kinases with the relative RMSE increased by <0.05, the CCNY low-sample kinases with the relative MAE decreased by <0.62. For the RAF subfamily, there are 50% of the low-sample kinases with the relative R^2 values increased by >0.2, 50% of the low-sample kinases with the relative RMSE decreased by <0.2, 50% of the low-sample kinases with the relative MAE decreased by <0.23 respectively. For the CAMK1 subfamily, there are 66.7% of the low-sample kinases with the relative R^2 values increased by >50%, though one of the low-sample kinases with the relative RMSE increased by comparative average of 0.06, and both of the low-sample kinases with the relative MAE decreased by comparative average of 0.08. These results further showed the significantly enhanced capability of the multi-task transfer learning approach, particularly the multi-task MolMapNet method, in the prediction of the inhibitory activity values of the low-sample kinases.

3 Concluding Remarks

Substantial number of kinases have not yet been fully explored as therapeutic targets in terms of drug approval[35]. There is big room for the development of drugs targeting these kinases for the treatment of cancers and other diseases. The rapid development and successful applications of artificial intelligence in other fields has allowed it to be actively explored in drug discovery, with the expectation to shorten the cycle and cost of traditional drug development with the help of DL technology. The bioactivity data on the public database has the problem of lack of data magnitude and quality dimensions, and the number of active inhibitors corresponding to some kinase targets and newly discovered kinase targets is insufficient. How to solve such low sample data is also the current core baffle. In our research, we introduced MolMapNet, a high-efficiency model for map reinforcement learning based on CNN. Its built-in MolMap has collected and mapped > 8 000 000 compounds in databases such as PubChem in advance, and has established rich structures and physicochemical properties. and help to improve the generalization performance of the model, based on this basis, we construct single-task regression and multi-task regression models of the kinase family[36].

The results of our studies clearly demonstrated that MolMapNet has good generalization capability, and is significantly better than single-task regression in multi-task regression modeling, and can establish a good transfer learning effect between different functional kinase family targets[37]. The framework of our established multi-task regression activity prediction model for kinase targets can also be transferred to other types of targets such as G Protein-Coupled Receptors and partial ion channels, to predict the activity value of the active compound of the corresponding target, that is the model has the characteristics of being portable and general. Our work still has some limitations, such as part of the target activity prediction accuracy of inhibitors is insufficient, and the overfitting phenomenon in low sample training and prediction, but the model structure and parameters of loss function optimized in greater depth can better solve the shortcomings existing in current models. There is still room for further optimization of the model, and it is expected to achieve improved activity prediction performances. Further development of the multi-task DL methods can offer useful tools for drug discovery against low-sample targets.

REFERENCES

[1] SAMORODNITSKY D. AI Widens Search Spaces and Promises More Hits in Drug Discovery: AI platforms are enhancing discovery efforts across modalities—mall-molecule drugs, RNA-based therapeutics, and protein-based therapeutics[J]. Genet Eng Biotechnol News, 2022, 42(4): 34-36, 38.

[2] XIONG Z, WANG D, LIU X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism[J]. J Med Chem, 2020, 63(16): 8749-8760.

[3] YANG K, SWANSON K, JIN W G, et al. Analyzing learned molecular representations for property prediction[J]. J Chem Inf Model, 2019, 59(8): 3370-3388.

[4] DUVENAUD D K, MACLAURIN D, IPARRAGUIRRE J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. NIPS, 2015.

[5] SUN M Y, ZHAO S D, GILVARY C, et al. Graph convolutional networks for computational drug development and discovery[J]. Brief Bioinform, 2020, 21(3): 919-935.

[6] POPOVA M, ISAYEV O, TROPSHA A. Deep reinforcement learning for de novo drug design[J]. Sci Adv, 2018, 4(7): eaap7885.

[7] GOH G B, HODAS N O, SIEGEL C, et al. SMILES2Vec: an interpretable general-purpose deep neural network for predicting chemical properties[J]. arXiv preprint arXiv:1712.02034, 2017.

[8] KARPOV P, GODIN G, TETKO I V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation[J]. J Cheminform, 2020, 12(1): 1-12.

[9] GOH G B, SIEGEL C, VISHNU A, et al. ChemNet: A transferable and generalizable deep neural network for small-molecule property prediction[R]. No. PNNL-SA-129942. Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2017, 12(8).

[10] CORTÉS-CIRIANO I, BENDER A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images[J]. J Cheminform, 2019, 11(1): 1-16.
[11] WENZEL J, MATTER H, SCHMIDT F. Predictive multitask deep neural network models for ADME-tox properties: Learning from large data sets[J]. J Chem Inf Model, 2019, 59(3): 1253-1268.
[12] SHEN W X, ZENG X, ZHU F, et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations[J]. Nat Mach Intell, 2021, 3(4): 334-343.
[13] JONES D, KIM H, ZHANG X H, et al. Improved protein-ligand binding affinity prediction with structure-based deep fusion inference[J]. J Chem Inf Modeling, 2021, 61(4): 1583-1592.
[14] KANNAIYAN R, MAHADEVAN D. A comprehensive review of protein kinase inhibitors for cancer therapy[J]. Expert Rev Anticancer Ther, 2018, 18(12): 1249-1270.
[15] ATTWOOD M M, FABBRO D, SOKOLOV A V, et al. Trends in kinase drug discovery: Targets, indications and inhibitor design[J]. Nat Rev Drug Discov, 2021, 20(11): 839-861.
[16] LI H Y, LIANG Z F, ZHANG C Y, et al. SuperDTI: Ultrafast DTI and fiber tractography with deep learning[J]. Magn Reson Med, 2021, 86(6): 3334-3347.
[17] LAVECCHIA A. Deep learning in drug discovery: Opportunities, challenges and future prospects[J]. Drug Discov Today, 2019, 24(10): 2017-2032.
[18] LIAO J Y, WAY G, MADAHAR V. Target Virus or Target Ourselves for COVID-19 Drugs Discovery? -Lessons learned from anti-influenza virus therapies[J]. Med Drug Discov, 2020(5): 100037.
[19] LIU G N, SINGHA M, PU L M, et al. GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data[J]. J Cheminform, 2021, 13(1): 58.
[20] XU Y Q, YAO H Q, LIN K J. An overview of neural networks for drug discovery and the inputs used[J]. Expert Opin Drug Discov, 2018, 13(12): 1091-1102.
[21] SONI A, BHAT R, JAYARAM B. Improving the binding affinity estimations of protein-ligand complexes using machine-learning facilitated force field method[J]. J Comput Aided Mol Des, 2020, 34(8): 817-830.
[22] RODRIGUES T, BERNARDES G J L. Machine learning for target discovery in drug development[J]. Curr Opin Chem Biol, 2020(56): 16-22.
[23] RAMSUNDAR B, KEARNES S, RILEY P, et al. Massively multitask networks for drug discovery[J]. arXiv preprint arXiv:1502.02072, 2015.
[24] RIFAIOGLU A S, NALBAT E, ATALAY V, et al. DEEPScreen: High performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations[J]. Chem Sci, 2020, 11(9): 2531-2557.
[25] RIFAIOGLU A S, ATAS H, MARTIN M J, et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases[J]. Brief Bioinform, 2019, 20(5): 1878-1912.
[26] TRABELSI A, CHAABANE M, BEN-HUR A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities[J]. Bioinformatics, 2019, 35(14): i269-i277.
[27] LIU B, LIU F L, FANG L Y, et al. repDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects[J]. Bioinformatics, 2015, 31(8): 1307-1309.
[28] MUNDI P S, SACHDEV J, MCCOURT C, et al. AKT in cancer: New molecular insights and advances in drug development[J]. Br J Clin Pharmacol, 2016, 82(4): 943-956.
[29] SHAIK A, KIRUBAKARAN S. Evolution of PIKK family kinase inhibitors: A new age cancer therapeutics[J]. Front Biosci (Landmark Ed), 2020, 25(8): 1510-1537.
[30] ALEXANDER D L J, TROPSHA A, WINKLER D A. Beware of R(2): Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models[J]. J Chem Inf Model, 2015, 55(7): 1316-1322.
[31] KAROULIA Z, GAVATHIOTIS E, POULIKAKOS P I. New perspectives for targeting RAF kinase in human cancer[J]. Nat Rev Cancer, 2017, 17(11): 676-691.
[32] XIE Z L, HOU S Z, YANG X X, et al. Lessons learned from past cyclin-dependent kinase drug discovery efforts[J]. J Med Chem, 2022, 65(9): 6356-6389.
[33] ASIH P R, PRIKAS E, STEFANOSKA K, et al. Functions of p38 MAP kinases in the central nervous system[J]. Front Mol Neurosci, 2020(13): 570586.
[34] SMORODINSKY-ATIAS K, SOUDAH N, ENGELBERG D. Mutations that confer drug-resistance, oncogenicity and intrinsic activity on the ERK MAP kinases-current state of the art[J]. Cells, 2020, 9(1): 129.
[35] ANUSUYA S, KESHERWANI M, PRIYA K V, et al. Drug-target interactions: Prediction methods and applications[J]. Curr Protein Pept Sci, 2018, 19(6): 537-561.
[36] CHEN M, HAO Y X, HWANG K, et al. Disease prediction by machine learning over big data from healthcare communities[J]. IEEE Access, 2017(5): 8869-8879.
[37] CICHOŃSKA A, RAVIKUMAR B, ALLAWAY R J, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors[J]. Nat Commun, 2021, 12(1): 3307.

收稿日期：2022-09-06
(本文责编：沈倩)