Deep Learning Models for Activity Prediction Against the Low-data COVID-19 Targets

LIANG Shuran[1], HUO Wenbo[1], SHEN Wanxiang[2], CHEN Yuzong[3], JIANG Yuyang[1*], TAN Ying[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: OBJECTIVE In response to Corona Virus Disease 2019(COVID-19), reusable drugs and new drugs against the low-data COVID-19 targets (with <300 known inhibitors) need to be discovered. **METHODS** Employing MolMapNet, a deep learning architecture that outperformed the state-of-the-art deep learning models on pharmaceutical benchmark datasets, new deep learning models were developed for predicting pharmaceutical properties with broadly-learned knowledge-based molecular representations. Predicted activities against 6 low-data COVID-19 targets with 34, 51, 81, 155, 161, 241 known inhibitors respectively. Compared with machine learning and deep learning models(with 5 478–10 000 known inhibitors) trained with targets in higher datasets. **RESULTS** Tested under the 10-fold cross-validation, our models predicted the activity values of the test-set inhibitors of these 6 targets with RMSE 0.442–0.917, MAE 0.358–0.749, and R^2 0.436–0.761. **CONCLUSION** The screening of approved drugs for potential drug repurposing agents against COVID-19 identified 3 drugs that are consistent with the literature-reported experimental findings. These indicate the potential of our deep learning method for the low-data targets against COVID-19 and other diseases.

KEYWORDS: COVID-19; drug discovery; drug repurposing; deep learning; activity prediction; low-data targets

针对低数据量 COVID-19 靶点的活性预测深度学习模型

梁书然[1]，霍文博[1]，申万祥[2]，陈宇综[3]，蒋宇扬[1*]，谭英[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：目的 为应对新型冠状病毒肺炎(Corona Virus Disease 2019，COVID-19)，发现针对低数据量 COVID-19 靶点(已知抑制剂<300 种)的可再利用药物和新药。**方法** 使用一种性能优于药物基准数据集上最先进的深度学习模型的深度学习架构 MolMapNet，开发新的深度学习模型，用于预测基于知识的分子表示方式的药物特性。针对 6 个低数据量 COVID-19 靶点进行活性预测，这些靶点分别有 34，51，81，155，161，241 种已知抑制剂。并与使用更高数据集靶点训练的机器学习和深度学习模型(具有 5 478~10 000 种已知抑制剂)进行比较。**结果** 在 10 倍交叉验证下进行模型测试，并使用测试集预测了这 6 个靶点的抑制剂的活性值。RMSE 为 0.442~0.917，MAE 为 0.358~0.749，R^2为 0.436~0.761。**结论** 在已批准药物中筛选针对 COVID-19 的潜在药物，确定了 3 种与文献报道的实验结果一致的可再利用药物。这些表明了该深度学习模型在针对 COVID-19 和其他疾病的低数据量靶点活性预测方面的潜力。

关键词：新型冠状病毒肺炎；药物发现；药物再利用；深度学习；活性预测；低数据量靶点

中图分类号：R914.2　　**文献标志码：**A　　**文章编号：**1007-7693(2022)21-2872-07

DOI: 10.13748/j.cnki.issn1007-7693.2022.21.025

引用本文：梁书然，霍文博，申万祥，等. 针对低数据量 COVID-19 靶点的活性预测深度学习模型[J]. 中国现代应用药学，2022, 39(21): 2872-2878.

Apart from vaccine development, extensive efforts have been directed at drug repurposing and new drug discovery for the treatment of Corona Virus Disease 2019(COVID-19)[1-2]. New targets, particularly the host targets of high therapeutic potential have been discovered by such investigations as the virus-host interactions[3] and infection-induced host proteomic changes[4]. The exploration of the host targets for the treatment of COVID-19 is advantageous over the viral targets, because these host targets are more abundant for drug targeting and less prone to drug resistances[1], which has become the focus of drug repurposing[2] and new drug discovery efforts[5]. Moreover, computational methods, particularly machine learning(ML)[6-9] and deep learning(DL) methods, have been explored for finding drug repurposing candidates and predicting new inhibitors against the COVID-19 targets.

Some of the ML and DL models against the COVID-19 targets have been developed for activity prediction with good performances(RMSE 0.29–1.002, R^2 0.047–0.838, MAE 0.115–1.244)[6-7,10-11], which facilitate the search of highly potent inhibitors against COVID-19 targets. These models have been trained by ~5 478 to 10 000 inhibitors(Tab. 1).

基金项目：国家重点研究计划合成生物学专项(2019YFA0905901)

作者简介：梁书然，女，硕士　E-mail: lsr18@tsinghua.org.cn　*通信作者：谭英，女，博士，副教授　E-mail: tan.ying@sz.tsinghua.edu.cn　蒋宇扬，男，博士，教授　E-mail: jiangyy@sz.tsinghua.edu.cn

Tab. 1 Training data size and activity prediction performance of the published machine learning(ML) and deep learning(DL) models

表 1 已发布的机器学习和深度学习模型的训练数据大小及活性预测表现

Target name	Training data size	Models	RMSE	MAE	R^2
		Mall[10]			
papain-like proteinase; 3C-like proteinase; Spike glycoprotein	60 195 drug-target interactions	ML: Random Forest	0.925	0.630	0.546
		ML: Support Vector Machines	0.883	0.596	0.588
		ML: XGBoost	0.868	0.567	0.599
		DL: CNN	0.909	0.587	0.575
		DL: LSTM	0.899	0.597	0.571
		CNN-LSTM	1.002	0.646	0.490
		GAT-CNN	0.872	0.576	0.597
		Beck[11]			
3C-like proteinase; RNA-dependent RNA polymerase; helicase; 3'-to-5' exonuclease; endoRNAse; 2'-*O*-ribose methyltransfe rase	drug-target interaction datasets of 97 092 853 interactions	DL: MT-DTI			0.81
		Batra[6]			
S-protein	8 120 molecules against target	ML: Random Forest	0.29	0.21	0.81
S-protein: ACE2 interface	5 478 molecules against target		0.84	0.57	0.70
		Kowalewski[7]			
ABCC1	Bio-assay data from ChEMBL 25	ML: Support Vector Machines		0.446–0.520	0.312–0.329
BRD2				0.341–0.405	0.435–0.529
BRD4				0.419–0.431	0.685–0.727
CSNK2A2				0.847–0.902	0.151–0.243
CSNK2B				0.414–0.449	0.670–0.750
DCTPP1				0.275–0.283	0.424–0.573
DNMT1				0.190–0.197	0.047–0.129
GFER				0.547–0.553	0.076–0.105
HDAC2				0.504–0.534	0.385–0.517
IMPDH2				0.423–0.463	0.352–0.493
ITGB1				0.565–0.638	0.616–0.699
MARK2				0.576–0.587	0.058–0.102
MARK3				0.450–0.473	0.134–0.177
NSD2				0.266–0.294	0.075–0.128
PABPC1				0.115–0.151	0.094–0.323
PLAT				0.522–0.613	0.283–0.460
PRKACA				0.502–0.517	0.483–0.522
PSEN2				0.582–0.603	0.502–0.542
PTGES2				0.421–0.490	0.644–0.716
RIPK1				1.099–1.244	0.252–0.352
SIGMAR1				0.504–0.555	0.572–0.639
TBK1				0.405–0.433	0.436–0.497
VCP				0.355–0.417	0.536–0.639
ACE2				0.530–0.858	0.748–0.838

However, many of the promising COVID-19 targets are of low-data targets with <300 known inhibitors in ChEMBL database[12-13] (some of which are in Tab. 2). Because DL typically requires larger training data, it is difficult to develop DL models for the low-data targets. As a result, low-data targets are inadequately explored by means of DL methods, particularly the low-data COVID-19 targets. To further explore DL capability for drug repurposing and new drug discovery against the low-data COVID-19 targets, it is desirable to explore new DL algorithms for activity prediction against these targets.

Significant progress has recently been made in the exploration of the broadly-learned knowledge-based molecular representations MolMap[14] and the graph-based de-novo learning of molecular representations[1-4] for DL of pharmaceutical (bio-activity, toxicological and pharmacokinetic) properties, which outperformed the previous state-of-the-art(SOTA) DL models[1-4,14]. The enhanced learning capability of these methods may be extended for the low-data targets, which has not yet been explored. In this work, we employed our recently developed MolMap representations and the DL architecture MolMapNet[14] to develop single-task regression DL models for activity prediction against 6 low-data COVID-19 targets (Tab. 2).

Tab. 2 Low-data COVID-19 targets and inhibitors selected in this study

表 2 本研究选择的低数据量新型冠状病毒肺炎靶点及其已知抑制剂数

Target group	Target (Gene name)	Number of known inhibitors
Kinase	Casein kinase II alpha prime (CSNK2A2)[15]	161
	Janus kinase 2 (JAK-2)][16]	34
	Cyclin G-associated kinase (GAK)[17]	241
Immune cell receptor	Toll-like receptor 2 (TLR2)[18]	81
	Toll-Toll-like receptor 9 (TLR9) (ClinicalTrials.gov Identifier: NCT04312997)	51
Protease	Dibasic-processing enzyme (Furin) (ClinicalTrials.gov Identifier: NCT04334460)	155

Based on the broad profiling of 1 456 molecular descriptors and 12 108 fingerprints against 8 206 960 unique molecules, our MolMap feature-generation algorithm maps molecular structures into correlational-arranged 2D feature maps of molecular descriptors and fingerprint features, and the corresponding MolMapNet architecture enables robust out-of-the-box(OOTB) DL of diverse pharmaceutical properties, including activity

prediction of various pharmaceutical properties[14]. Such OOTB DL models are constructed with the same set of default parameters for all learning tasks, which aim at taking human out of the DL processes, allowing more people to use them for different DL tasks[19]. The capability of the MolMapNet architecture was further tested on the activity prediction tasks against the 6 low-data COVID-19 targets(with 34–241 known inhibitors), and the corresponding performance was evaluated against the published performances of the DL and ML models for the higher-data COVID-19 targets with 5 478 to 10 000 inhibitors.

1 Methods

1.1 Data collection and processing

The COVID-19 targets were searched from the special COVID-19 sections of several databases[20], such as the therapeutic target database TTD[12] and DrugBank[21] with a specific focus on the targets in human host. The identified COVID-19 targets have been found to be the key target proteins or regulators of COVID-19 viral infection[12], which have been tested in COVID-19 clinical trials[22], or discovered by the COVID-19 omics investigations[15,23], or investigated for COVID-19 drug repurposing and new drug discovery[24]. For the identified targets, their known inhibitors were searched from the ChEMBL database[13]. We identified 6 low-data COVID-19 targets with 34, 51, 81, 155, 161, 241 known inhibitors respectively. For exploring drug repurposing opportunities against these targets, the approved drugs were collected from TTD[12] and DrugBank[21]. The SMILES codes of the inhibitors of the identified targets together with their activity values against the respective target were downloaded from the ChEMBL database[13]. For unified representation, a standard pChEMBL, which allows multiple roughly comparable measures to be compared on a negative logarithmic scale, were used to represent the activity values of inhibitors. pChEMBL is defined as: −Log (molar IC_{50}, XC_{50}, EC_{50}, AC_{50}, K_i, K_d or Potency)[13].

1.2 MolMap molecular representations

SMILES codes of the inhibitors were converted to canonical SMILES codes by RDKit[25-26]. The molecular descriptors and fingerprints of these inhibitors were computed from their canonical SMILES codes by using MolMap package[14]. We used MolMap to further convert these molecular descriptors and fingerprints into a MolMap 2D molecular feature map[14], which embeds the broadly-learned correlation relationships of the molecular descriptors and fingerprints in the 2D feature space by means of UMAP[27].

1.3 MolMapNet DL architecture

MolMapNet has a dual-path CNN architecture, one path for molecular descriptors and another for fingerprint features to enable simultaneous learning (Fig. 1)[14]. In this work, we choose 13 classes of molecular descriptors and 3 sets of fingerprints (MACCSFP, PharmacoErGFP, and PubChemFP) for representing the inhibitors. Deeper feature extraction processes were conducted through the CNN layers. The first convolution layer consists of a larger number of kernels(48) for increased data dimension and a larger kernel size(13×13/1) for more expressive capability and more extensive perception[27]. The maximum parameter of our model is <0.83 million in the general tasks, but with relatively complex topology and depth.

The left path is for learning molecular descriptors with multi-channel input layer of up to 13 descriptor classes. The right path is for learning fingerprints with multi-channel input layer of up to 3 fingerprint sets. Trainable parameters: left path: ~ 0.40 million, right path: ~0.32 million, dual path: ~ 0.80 million.

1.4 MolMapNet hyperparameters and training

The activation function ReLU was used for all tasks, with a small learning rate(0.000 1) and batch size(128). Other regularization options such as dropout and weight decay(L2 regularization) were not used, because MolMapNet models have relatively fewer parameters and are easily trained to convergence. In the regression tasks, the loss function was set to mean squared error. In model training, the early-stopping strategy was used in MolMapNet for alleviating over-fitting and computational cost[1-4, 28-30]. We performed 10-fold cross-validation for each model by splitting the full data set randomly into train set, test set in proportions(0.9∶0.1) for training and validation purposes respectively as mentioned in the comparative works. All models were developed by TensorFlow 2.0.0 on GeForce RTX 2080 Ti (12 GB memory in each card).

1.5 Performance evaluation and metrics

Three metrics were used for evaluating our developed models: RMSE, MAE and R^2. For RMSE and MAE, each of these metrics are estimated using the predicted pChEMBL values *vs* the ground truth pChEMBL values for inhibitor-target interactions. R^2 is squared Pearson correlation coefficient between predicted and observed values. For metrics, MAE and RMSE, the lower the value and closer to 0, the better the predictive performance of the model, whereas for metrics, R^2, the higher and closer the value to 1, the better the efficiency of the predictive model.

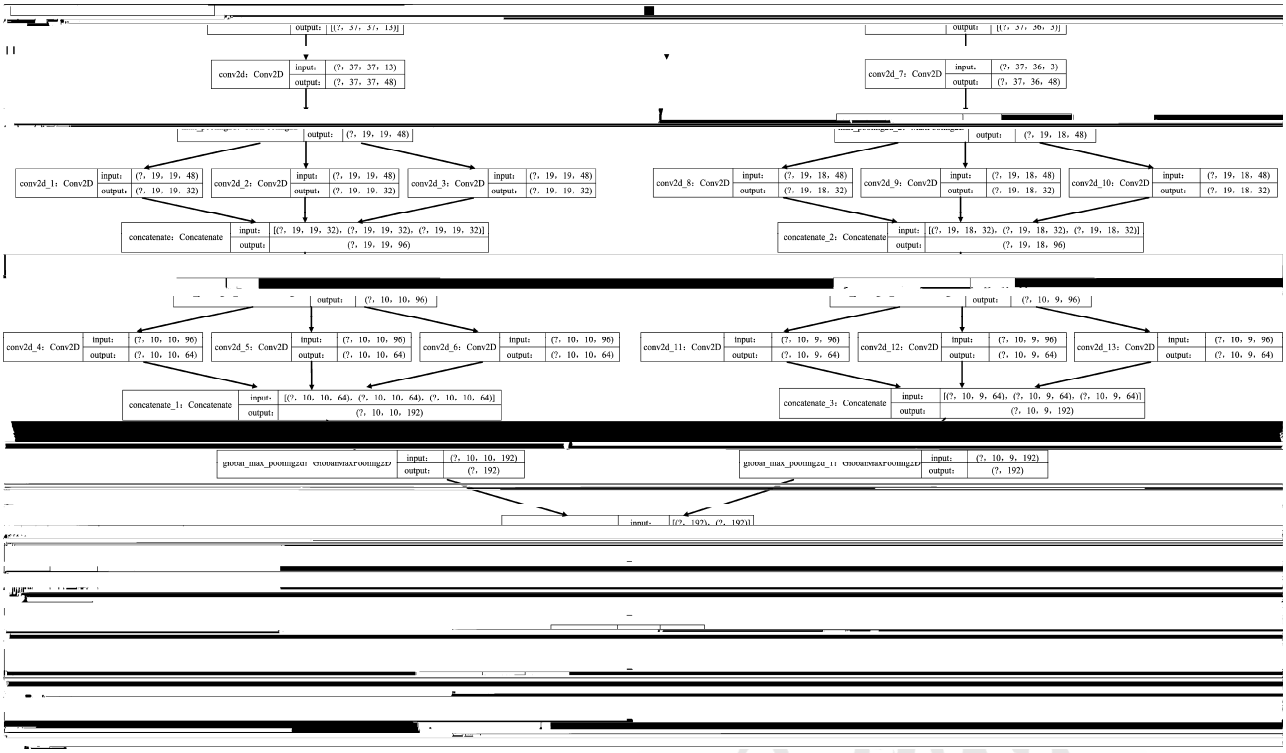

Fig. 1 MolMapNet deep learning architecture with three components(multichannel input feature mapping, dual-path CNN feature learning, nonlinear task learning with fully-connected layers)

图 1 三组件 MolMapNet 深度学习架构(多通道输入特征映射、双路径 CNN 特征学习、具有全连接层的非线性任务学习)

2 Results and Discussion

MolMapNet regression models were trained and tested on each of the 6 low-data COVID-19 targets using three metrics(RMSE, MAE and R^2)(Tab. 3), which were compared with the reported performance of the published ML and DL models of higher-data targets(with 5 478–10 000 known inhibitors) measured by the same metrics(RMSE, MAE and R^2) (Tab. 1). The performance of MolMapNet was evaluated by 10-fold cross-validation as described in the Methods section. Significantly, our low-data models showed comparable or better performance with respect to those of the published ML and DL models of higher-data targets. The best RMSE of MolMapNet(0.442) outperformed that of 9 in 10 published ML and DL models. While the best MAE of MolMapNet(0.358) was better than those of most published ML and DL models. The best R^2 of MolMapNet(0.761) was better than those of 32 in 34 published ML and DL models. Both the MolMapNet model and the published ML model have training on target casein kinase II alpha prime(Gene name: CSNK2A2), the MolMapNet outperformed the published ML model on both MAE and R^2. In the case of much less training data, good expectations have been reached, and it might be exceeded if training data increased. These results showed that the MolMapNet method has certain competitiveness in predicting with small data sizes.

Tab. 3 Activity prediction performances of MolMapNet for the low-data COVID-19 targets

表 3 MolMapNet 对低数据新型冠状病毒肺炎靶点的活性预测表现

COVID-19 targets	Number of known inhibitors	RMSE	MAE	R^2
Kinase group				
Casein kinase II alpha prime (CSNK2A2)	161	0.575	0.466	0.595
Janus kinase 2 (JAK-2)	34	0.442	0.358	0.631
Cyclin G-associated kinase (GAK)	241	0.599	0.474	0.761
Immune cell receptor group				
Toll-like receptor 2 (TLR2)	81	0.917	0.749	0.436
Toll-like receptor 9 (TLR9)	51	0.530	0.437	0.468
Protease group				
Dibasic-processing enzyme (Furin)	155	0.578	0.440	0.484

We obtained 2 895 approved drugs from DrugBank[21] and TTD[12] , which were screened by our developed DL models of the COVID-19 targets for finding potential drug repurposing agents. We specifically screened against two targets with relatively higher number of known inhibitors (161 to 241) and with a relatively good prediction performance(R^2>0.59). The top 10 ranked drugs of each target, based on higher predicted pChEMBL values, are shown in Tab. 4. Among the identified potential drug repurposing candidates, two drugs Atorvastatin and Telavancin were predicted to be

Tab. 4 Screening results of top 10 ranked approved drugs against COVID-19 targets Casein kinase II alpha prime(CSNK2A2) and Cyclin G-associated kinase(GAK)

表 4 针对新型冠状病毒肺炎靶点 CSNK2A2 和 GAK 的已批准药物筛选结果前 10 名

Approved drug names	DrugBank ID	Original target	Predict value
COVID-19 target : Casein kinase II alpha prime (CSNK2A2)			
Celecoxib	DB00482	Cc1ccc(-c2cc(C(F)(F)F)nn2-c2ccc(S(N)(=O)=O)cc2)cc1	6.993
Indapamide	DB00808	CC1Cc2ccccc2N1NC(=O)c1ccc(Cl)c(S(N)(=O)=O)c1	6.991
Metolazone	DB00524	Cc1ccccc1N1C(=O)c2cc(S(N)(=O)=O)c(Cl)cc2NC1C	6.912
Cyclothiazide	DB00606	NS(=O)(=O)c1cc2c(cc1Cl)NC(C1CC3C=CC1C3)NS2(=O)=O	6.890
Tenapanor	DB11761	CN1Cc2c(Cl)cc(Cl)cc2[C@H](c2cccc(S(=O)(=O)NCCOCCOCCNC(=O)NCCCCNC(=O)NCCOCCO CCNS(=O)(=O)c3cccc([C@@H]4CN(C)Cc5c(Cl)cc(Cl)cc54)c3)c2)C1	6.867
Chlorthalidone	DB00310	NS(=O)(=O)c1cc(C2(O)NC(=O)c3ccccc32)ccc1Cl	6.783
Benzthiazide	DB00562	NS(=O)(=O)c1cc2c(cc1Cl)N=C(CSCc1ccccc1)NS2(=O)=O	6.771
Zotepine	DB09225	CN(C)CCOC1=Cc2ccccc2Sc2ccc(Cl)cc21	6.763
Oritavancin	DB04911	CN[C@H](CC(C)C)C(=O)N[C@H]1C(=O)N[C@@H](CC(N)=O)C(=O)N[C@H]2C(=O)N[C@H]3C (=O)N[C@H](C(=O)N[C@H](C(=O)O)c4cc(O)cc(O)c4-c4cc3ccc4O)[C@H](O[C@H]3C[C@](C) (N)[C@@H](O)[C@H](C)O3)c3ccc(c(Cl)c3)Oc3cc2cc(c3O[C@@H]2O[C@H](CO)[C@@H](O) [C@H](O)[C@H]2O[C@H]2C[C@](C)(NCc3ccc(-c4ccc(Cl)cc4)cc3)[C@@H](O)[C@H](C)O2) Oc2ccc(cc2Cl)[C@H]1O	6.763
Lumefantrine	DB06708	CCCCN(CCCC)CC(O)c1cc(Cl)cc2c1-c1ccc(Cl)cc1/C2=C/c1ccc(Cl)cc1	6.760
COVID-19 target : Cyclin G-associated kinase(GAK)			
Crizotinib	DB08865	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl	7.829
Isavuconazonium	DB06636	CNCC(=O)OCc1cccnc1N(C)C(=O)OC(C)[n+]1cnn(C[C@](O)(c2cc(F)ccc2F)[C@@H](C)c2nc(-c3ccc (C#N)cc3)cs2)c1	7.778
Atorvastatin	DB01076	CC(C)c1c(C(=O)Nc2ccccc2)c(-c2ccccc2)c(-c2ccc(F)cc2)n1CC[C@@H](O)C[C@@H](O)CC(=O)O	7.622
Dabrafenib	DB08912	CC(C)(C)c1nc(-c2cccc(NS(=O)(=O)c3c(F)cccc3F)c2F)c(-c2ccnc(N)n2)s1	7.600
Ceritinib	DB09063	Cc1cc(Nc2ncc(Cl)c(Nc3ccccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1	7.462
Telavancin	DB06402	CCCCCCCCCCCNCCN[C@@]1(C)C[C@H](O[C@H]2[C@H](Oc3c4cc5cc3Oc3ccc(cc3Cl)[C@@H] (O)[C@@H](NC(=O)[C@@H](CC(C)C)NC)C(=O)N[C@@H](CC(N)=O)C(=O)N[C@H]5C(=O)N [C@H]3C(=O)N[C@H](C(=O)N[C@H](C(=O)O)c5cc(O)c(CNCP(=O)(O)O)c(O)c5c5cc3ccc5O) [C@H](O)c3ccc(c(Cl)c3)O4)O[C@H](CO)[C@@H](O)[C@@H]2O)O[C@@H](C)[C@H]1O	7.301
Fosnetupitant	DB14019	Cc1ccccc1-c1cc(N2CC[N+](C)(COP(=O)([O-])O)CC2)ncc1N(C)C(=O)C(C)(C)c1cc(C(F)(F)F)cc(C(F) (F)F)c1	7.281
Avatrombopag	DB11995	O=C(Nc1nc(-c2cc(Cl)cs2)c(N2CCN(C3CCCCC3)CC2)s1)c1cnc(N2CCC(C(=O)O)CC2)c(Cl)c1	7.222
Lusutrombopag	DB13125	CCCCCCO[C@@H](C)c1cccc(-c2csc(NC(=O)c3cc(Cl)c(/C=C(\\C)C(=O)O)c(Cl)c3)n2)c1OC	7.151
Letermovir	DB12070	COc1cccc(N2CCN(C3=Nc4c(F)cccc4[C@H](CC(=O)O)N3c3cc(C(F)(F)F)ccc3OC)CC2)c1	7.133

Note: Predict value refers to predicted pChEMBL value.

注：预测值是指预测的 pChEMBL 值。

active against the COVID-19 target Cyclin G-associated kinase(GAK) (UniProt ID: O14976). An on-going randomized controlled Atorvastatin adjuvant trial is underway to study the effects of Atorvastatin on COVID-19 disease progression within 30 d (ClinicalTrials.gov Identifier: NCT04380402). A recent study has shown that Atorvastatin can reduce the expression of IL10, to achieve anti-inflammatory effects[31]. GAK has been found to interact with IL12 receptor[32]. IL12 and IL10 signaling is interdependent in response to certain infections[33]. Therefore, the clinical effect of Atorvastatin against COVID-19 may be attributable to its interference of GAK-IL12-IL10 signaling pathways. A recent study has shown that Teicoplanin is able to block MERS and SARS envelope pseudoviruses, it inhibits the activities of the host cell's cathepsin L and cathepsin B specifically, thereby, blocks the receptor-binding domains exposure of the core genome and subsequent release into the cytoplasm of the host cell[34]. Telavancin, as a derivative of Teicoplanin, is a novel inhibitor of histopsin L-dependent virus[35]. Our prediction of the targeting of Atorvastatin and Telavancin against GAK is consistent with these experimental findings.

A Cox2 inhibitor drug Celecoxib was predicted to be active against the COVID-19 target casein kinase(CK2) (UniProt ID: P19784)[12]. A randomized trial of hospitalized COVID-19 patients has shown that Celecoxib prevented clinical deterioration, and is associated with rapid pulmonary CT-chest improvement[36]. Another clinical trial has demonstrated that treatment with an anti-CK2 synthetic peptide improves clinical response in COVID-19 patients with pneumonia[37]. An investigation has found that CK2 is a regulator of

TBK1 and IRF3, and blockade of CK2 activity by a small molecule inhibitor leads to TBK1 activation, whereby eliciting effective host defense mechanisms against such viral infection as hepatitis C infection[38]. Another study has revealed that COVID-19 proteins NSP6 and NSP13 bind and block TBK1 phosphorylation, which suppress IRF3 phosphorylation and nuclear translocation for the evasion of COVID-19 against human host's type-1 interferon responses[39]. Hence, the clinical effect of Celecoxib may be due to its regulation of the CK2-TBK1-IRF3-interferon signaling in addition to its Cox2 inhibitory anti-inflammatory effects[40]. Our predicted targeting of Celecoxib against CK2 is consistent with these reports. For the remaining 17 identified drugs, we have not found literature reports for judging the validity of the predicted COVID-19 targets of these drugs. Nonetheless, the consistency of the prediction results of 3 drugs with the literature-reported experimental findings indicates the usefulness of our DL models for searching potential COVID-19 drug repurposing agents.

3 Conclusion

Accurate learning and prediction of activates against a target is a challenging task[41], particularly for low-data targets[42] and novel prediction tasks[43]. Appropriate molecular representations are critical for enhanced DL capabilities[1, 44-48]. Our developed DL model MolMapNet is based on a CNN architecture with broadly-learned knowledge-based molecular representations, which has shown good prediction performances for various activity, toxicity and pharmacokinetic properties[14]. This work further demonstrated the capability of MolMapNet for activity prediction against the low-data targets. In particular, the prediction performances for the 5 low-data COVID-19 targets are comparable or even better than the published ML and DL models of higher-data targets. In the screening of approved drugs for potential drug repurposing agents against 2 low-data COVID-19 targets, the prediction results of 3 identified drugs are consistent with the literature- reported experimental findings. Taken together, our studies suggested the usefulness of DL methods with broadly-learned knowledge-based molecular representations for activity prediction against low-data targets, particularly for drug discovery and drug repurposing against the low-data COVID-19 targets.

REFERENCES

[1] WU Z Q, RAMSUNDAR B, FEINBERG E N, et al. MoleculeNet: a benchmark for molecular machine learning[J]. Chem Sci, 2017, 9(2): 513-530.

[2] KRENN M, HÄSE F, NIGAM A, et al. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation[J]. Mach Learn: Sci Technol, 2020, 1(4): 045024.

[3] 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.

[4] 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.

[5] 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.

[6] BATRA R, CHAN H, KAMATH G, et al. Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking studies[J]. J Phys Chem Lett, 2020, 11(17): 7058-7065.

[7] KOWALEWSKI J, RAY A. Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space[J]. Heliyon, 2020, 6(8): e04639.

[8] IVANOV J, POLSHAKOV D, KATO-WEINSTEIN J, et al. Quantitative structure-activity relationship machine learning models and their applications for identifying viral 3CLpro- and RdRp-targeting compounds as potential therapeutics for COVID-19 and related viral infections[J]. ACS Omega, 2020, 5(42): 27344-27358.

[9] MOHAPATRA S, NATH P, CHATTERJEE M, et al. Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking[J]. PLoS One, 2020, 15(11): e0241543.

[10] MALL R, ELBASIR A, MEER H A, et al. Data-driven drug repurposing for COVID-19[J/OL]. 2020-7-16 [2021-12-15]. https://doi.org/10.26434/chemrxiv.12661103.v1.

[11] BECK B R, SHIN B, CHOI Y, et al. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model[J]. Comput Struct Biotechnol J, 2020(18): 784-790.

[12] WANG Y X, ZHANG S, LI F C, et al. Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics[J]. Nucleic Acids Res, 2020, 48(D1): D1031-D1041.

[13] MENDEZ D, GAULTON A, BENTO A P, et al. ChEMBL: towards direct deposition of bioassay data[J]. Nucleic Acids Res, 2019, 47(D1): D930-D940.

[14] SHEN W X, LIU Y, CHEN Y, et al. AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks[J]. Nucleic Acids Res, 2022, 50(8): e45.

[15] GORDON D E, JANG G M, BOUHADDOU M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing[J]. Nature, 2020, 583(7816): 459-468.

[16] ZHANG W, ZHAO Y, ZHANG F C, et al. The use of anti-inflammatory drugs in the treatment of people with severe coronavirus disease 2019 (COVID-19): The Perspectives of clinical immunologists from China[J]. Clin Immunol, 2020(214): 108393.

[17] RICHARDSON P, GRIFFIN I, TUCKER C, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease[J]. Lancet, 2020, 395(10223): e30-e31.

[18] PATTERSON B K, SEETHAMRAJU H, DHODY K, et al.

Disruption of the CCL5/RANTES-CCR5 pathway restores immune homeostasis and reduces plasma viral load in critical COVID-19[J]. medRxiv, 2020, Doi: 10.1101/2020.05.02.20084673.
[19] YAO Q, WANG M, CHEN Y, et al. Taking human out of learning applications: A survey on automated machine learning[J]. arXiv preprint arXiv:1810.13306, 2018.
[20] WANG Y, LI F, ZHANG Y, et al. Databases for the targeted COVID-19 therapeutics[J]. Br J Pharmacol, 2020, 177(21): 4999-5001.
[21] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Res, 2017, 46(D1): D1074-D1082.
[22] LYTHGOE M P, MIDDLETON P. Ongoing clinical trials for the management of the COVID-19 pandemic[J]. Trends Pharmacol Sci, 2020, 41(6): 363-382.
[23] BOJKOVA D, KLANN K, KOCH B, et al. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets[J]. Nature, 2020, 583(7816): 469-472.
[24] LI G D, DE CLERCQ E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV)[J]. Nat Rev Drug Discov, 2020, 19(3): 149-150.
[25] BECHT E, MCINNES L, HEALY J, et al. Dimensionality reduction for visualizing single-cell data using UMAP[J]. Nat Biotechnol, 2018. Doi: 10.1038/nbt.4314.
[26] MCINNES L, HEALY J, SAUL N, et al. UMAP: uniform manifold approximation and projection[J]. J Open Source Softw, 2018, 3(29): 861.
[27] PENG C, ZHANG X Y, YU G, et al. Large kernel matters—improve semantic segmentation by global convolutional network[J]. 2017 IEEE Conf Comput Vis Pattern Recognit CVPR, 2017: 1743-1751.
[28] LI X, XU Y J, LAI L H, et al. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network[J]. Mol Pharm, 2018, 15(10): 4336-4345.
[29] 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.
[30] 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): 41.
[31] BIFULCO M, GAZZERRO P. Statin therapy in COVID-19 infection: Much more than a single pathway[J]. Eur Heart J Cardiovasc Pharmacother, 2020, 6(6): 410-411.
[32] LIN Y, TANG Y J, ZONG H L, et al. Cyclin G associated kinase interacts with interleukin 12 receptor beta2 and suppresses interleukin 12 induced IFN-gamma production[J]. FEBS Lett, 2007, 581(26): 5151-5157.
[33] RETINI C, KOZEL T R, PIETRELLA D, et al. Interdependency of interleukin-10 and interleukin-12 in regulation of T-cell differentiation and effector function of monocytes in response to stimulation with *Cryptococcus neoformans*[J]. Infect Immun, 2001, 69(10): 6064-6073.
[34] WANG Y Z, CUI R, LI G M, et al. Teicoplanin inhibits Ebola *Pseudovirus* infection in cell culture[J]. Antiviral Res, 2016(125): 1-7.
[35] ZHOU N, PAN T, ZHANG J S, et al. Glycopeptide antibiotics potently inhibit cathepsin L in the late endosome/lysosome and block the entry of Ebola virus, middle east respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus (SARS-CoV)[J]. J Biol Chem, 2016, 291(17): 9218-9232.
[36] TOMERA K, MALONE R, KITTAH J. Hospitalized COVID-19 patients treated with celecoxib and high dose famotidine adjuvant therapy show significant clinical responses[J/OL]. 2020-7-8 [2021-12-15]. https://ssrn.com/abstract=3646583 or http://dx.doi.org/10.2139/ssrn.3646583.
[37] CRUZ L R, BALADRÓN I, RITTOLES A, et al. Treatment with an anti-CK2 synthetic peptide improves clinical response in COVID-19 patients with pneumonia. A randomized and controlled clinical trial[J]. ACS Pharmacol Transl Sci, 2020, 4(1): 206-212.
[38] DU M, LIU J H, CHEN X, et al. Casein kinase II controls TBK_1/IRF_3 activation in IFN response against viral infection[J]. J Immunol, 2015, 194(9): 4477-4488.
[39] XIA H, CAO Z, XIE X, et al. Evasion of type I interferon by SARS-CoV-2[J]. Cell reports, 2020, 33(1): 108234.
[40] HONG W X, CHEN Y, YOU K, et al. Celebrex adjuvant therapy on coronavirus disease 2019: An experimental study[J]. Front Pharmacol, 2020(11): 561674.
[41] VAN DE WATERBEEMD H, GIFFORD E. ADMET in silico modelling: Towards prediction paradise?[J]. Nat Rev Drug Discov, 2003, 2(3): 192-204.
[42] ALTAE-TRAN H, RAMSUNDAR B, PAPPU A S, et al. Low data drug discovery with one-shot learning[J]. ACS Cent Sci, 2017, 3(4): 283-293.
[43] GLAVATSKIKH M, LEGUY J, HUNAULT G, et al. Dataset's chemical diversity limits the generalizability of machine learning predictions[J]. J Cheminform, 2019(11): 69.
[44] PAOLINI G V, SHAPLAND R H, VAN HOORN W P, et al. Global mapping of pharmacological space[J]. Nat Biotechnol, 2006, 24(7): 805-815.
[45] WILLETT P. Similarity-based virtual screening using 2D fingerprints[J]. Drug Discov Today, 2006, 11(23/24): 1046-1053.
[46] WETZEL S, KLEIN K, RENNER S, et al. Interactive exploration of chemical space with Scaffold Hunter[J]. Nat Chem Biol, 2009, 5(8): 581-583.
[47] ZHAVORONKOV A, IVANENKOV Y A, ALIPER A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors[J]. Nat Biotechnol, 2019, 37(9): 1038-1040.
[48] WINTER R, MONTANARI F, NOÉ F, et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations[J]. Chem Sci, 2018, 10(6): 1692-1701.

收稿日期：2022-09-07
(本文责编：沈倩)