Binding Activity Prediction of the Low-data G-protein Coupled Receptors Targets by Deep Learning of Knowledge-based Molecular Representations

LI Yin, TAN Ying*(*Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China*)

ABSTRACT: OBJECTIVE To construct new deep learning(DL) models for binding activity prediction against each of 23 low-data G-protein coupled receptors(GPCRs)(known binders <250) using MolMapNet, assisting in the novel drug discovery of GPCRs. **METHODS** Binding activity datasets of low-data GPCRs were collected from multiple databases and preprocessed, and DL models were constructed by MolMapNet; the established models were compared with published DL models and ML models; Neuropeptide S receptor proprietary compounds to evaluate the constructed model. **RESULTS** Under 10-fold cross-validation tests, MolMapNet DL models predicted the binding activity values of the test-set compounds for each GPCR with RMSE 0.373 6–1.199 8(20 among which RMSE<1), MAE 0.299 4–1.008 3(21 among which MAE<1), and R^2 0.136 9–0.810 7(15 among which R^2 >0.5, 9 among which R^2 >0.6). Our low-sample models showed comparable performances to those of the published DL models trained with higher-data GPCRs(>250 known binders). Our models also performed well in activity prediction of patented GPCR binders. **CONCLUSION** The 23 models constructed here can predict the biological activity of a compound against a specific target with good performance, have the potential to screen drugs with novel structures, and MolMapNet architecture is useful for activity prediction against the low-sample GPCR targets.
KEYWORDS: binding activity; deep learning; G-protein coupled receptors; low-data

基于分子理化性质特征的小样本G蛋白偶联受体靶点结合活性预测的深度学习模型

李因，谭英*(清华大学深圳国际研究生院，广东 深圳 518055)

摘要：目的 使用 MolMapNet 构建深度学习(deep learning，DL)模型，预测化合物对 23 个小样本(已知活性数据<250)G 蛋白偶联受体(G-protein coupled receptors，GPCRs)的结合活性，辅助发现 GPCRs 的新型药物。**方法** 从多个数据库搜集小样本 GPCRs 的活性数据集并进行预处理，使用 MolMapNet 构建 DL 模型；将建立的模型与已公布 DL 模型和 ML 模型进行比较；用神经肽 S 受体专利化合物对构建的模型进行评估。**结果** 构建了 23 个小样本 GPCRs 靶点的单回归模型，在 10 折交叉验证测试下，构建的模型在测试集上的均方根误差为 0.373 6~1.199 8(其中 20 个<1)，平均绝对误差为 0.299 4~1.008 3(其中 21 个<1)，R^2 为 0.136 9~0.810 7(其中 15 个>0.5，9 个>0.6)；与已发表的大样本 GPCRs(已知活性数据>250) DL 模型和小样本靶点的 ML 模型相比，显示出相当的性能；使用构建的模型对专利中化合物进行活性预测，模型表现良好。**结论** 构建的 23 个回归模型能够预测化合物对特定靶点的生物活性，具有筛选结构新颖的药物的潜力，MolMapNet 可用于小样本 GPCRs 的活性预测。

关键词：结合活性；深度学习；GPCR；小样本
中图分类号：R914.2 **文献标志码：**A **文章编号：**1007-7693(2022)21-2842-08
DOI: 10.13748/j.cnki.issn1007-7693.2022.21.021
引用本文：李因，谭英. 基于分子理化性质特征的小样本 G 蛋白偶联受体靶点结合活性预测的深度学习模型[J]. 中国现代应用药学, 2022, 39(21): 2842-2849.

G-protein coupled receptors(GPCRs) constitute one of the largest drug target families. They are targeted by 475(~34%) FDA approved drugs and 321 clinical trial agents[1]. Despite successful exploration of GPCRs, 20% of the 66 novel GPCRs in clinical trials are without an approved drug, and there are additional 224(56%) non-olfactory GPCRs with broad untapped therapeutic potential and yet to enter clinical investigations[1]. Therefore, methods that facilitate the efficient discovery of binders of novel GPCRs are highly desired. Computational methods such as molecular docking and machine learning(ML) have been explored for facilitating the discovery of binders of novel GPCRs[2-7]. The rapid advance of the exploration of deep learning(DL) methods for drug discovery has raised great interest in applying

基金项目：国家重点研究计划合成生物学专项(2019YFA0905901)
作者简介：李因，女，硕士 E-mail: naturele@126.com *__通信作者__：谭英，女，博士，副教授 E-mail: tan.ying@sz.tsinghua.edu.cn

DL for discovering binders of novel GPCRs[8-15].

Many novel GPCRs are low-data targets with < 250 known binders in ChEMBL database(Tab. 1)[16-17]. DL models have been typically trained by larger data. The recently-developed graph convolutional neural network(GCN) models for higher-data GPCRs(>250 known binders) binder prediction have been develped by using 2 135−11 632 binders[18]. In other study, ML models of low-sample non GPCRs targets have been developed by using 61−170 binders[3-6]. Efficient DL methods for low-data targets are needed for extended coverage of low-data GPCRs and other targets. Moreover, GPCR targeted drugs are mostly high potent binders. There is a need for the prediction of binding activity level against GPCRs in order to find potent binders. Although DL regression models have been developed for the prediction of binding activity values of various targets[10-15], to the best of our knowledge, rarely ML or DL regression model has been developed for low-data GPCRs of <250 known binders. The enhanced capability of the emerging DL algorithms may be explored for prediction of the binding activity levels against the low-data GPCRs.

Tab. 1 Low-sample GPCRs and known binders evaluated in this study

表 1 本文研究的小样本 GPCRs 及其已知配体数量

GPCR group	GPCRs	No of Known Binders
Short peptide receptor	Neuropeptide S receptor	104
	Apelin receptor	149
	Prolactin-releasing peptide receptor	164
	Neuropeptide FF receptor 1	172
	Neuromedin B receptor	243
Lipid-like ligand receptor	Oxoeicosanoid receptor 1	81
	Uracil nucleotide/cysteinyl leukotriene receptor	90
	Lysophosphatidic acid receptor 5	105
	C-X-C chemokine receptor type 5	108
Chemokine receptor	C-C chemokine receptor type 10	50
	C-X-C chemokine receptor type 5	108
	C-C chemokine receptor type 8	211
Carboxylic acid receptor	Succinate receptor 1	48
	HM74 nicotinic acid GPCR	143
Nucleotide-like receptor	Purinergic receptor P2Y11	65
G-protein coupled receptor B family	Vasoactive intestinal polypeptide receptor 1	52
	Corticotropin releasing factor receptor 2	58
	Glucagon-like peptide 2 receptor	111
	Vasoactive intestinal polypeptide receptor 2	116
	Pituitary adenylate cyclase-activating polypeptide type I receptor	144
	Gastric inhibitory polypeptide receptor	203
G-protein coupled receptor C family	GABA-B receptor 1	60
	Metabotropic glutamate receptor 7	133

Advanced DL algorithms have recently emerged for the prediction of pharmaceutical properties based on broadly-learned knowledge-based molecular representations MolMap and the graph-based de-novo learning of molecular representations[10-14]. The DL models with these algorithms outperformed the previous state-of-the-art (SOTA) DL models in the prediction of activity values, pharmacokinetic and toxicological properties[10-14]. These algorithms may be applied for the improved prediction of binding activity against low-data GPCRs. In this work, we employed our recently developed MolMap representations and the DL architecture MolMapNet for developing single-task regression DL models for binding activity prediction against 23 low-data GPCRs(Tab. 1)[14].

MolMap algorithm converts unordered molecular descriptors and fingerprints of compounds into correlationally-arranged 2D feature maps, based on which highly-efficient convolutional neural networks MolMapNet architecture can be constructed for robust out-of-the-box(OOTB) DL of diverse pharmaceutical properties, including activity prediction of various pharmaceutical properties[14]. The robustness of MolMap and MolMapNet models is supported by broad profiling of 1 456 molecular descriptors and 12 108 fingerprints against 8 206 960 unique molecules. The OOTB DL models are with fixed set of default parameters for all learning tasks, which takes human out of the DL processes and thus allows more people to develop DL models[19]. The prediction performance of our developed MolMapNet models was evaluated with respect to published DL models of higher-sample GPCRs and ML models of low-data protein targets. The first set of models are GCN models of DL for the prediction of binders of 33 GPCRs[18]. The second set of models are ML models for the prediction of binders of diverse low-sample targets[3-6]. While the 23 low-sample GPCRs of our models are different from the 33 higher-data GPCRs of the GCN models and other low-data targets of ML models, the range of the performance metric RMSE, MAE and R^2 values of these models can nonetheless be tentatively compared, which provide some indication about whether our low-data GPCR models can reach the comparable level of performance of the higher-sample GPCR models and low-data protein targets ML models[15].

1 Methods

1.1 Data collection, processing and molecular representation

The GPCR family of proteins were obtained from the Uniprot database[20]. Based on their respective Uniprot ID, the binders of each GPCR were obtained from the ChEMBL database[17], PubChem database and BindingDB database using web crawler, with the IC_{50}, EC_{50} or K_i value of each binder recorded. We identified 23 low-data GPCRs with 48–243 known binders. For unified representation of binding potencies, a standard pChEMBL value, defined as –Log(molar IC_{50}, EC_{50}, K_i)[17], was used for measuring the binding activity value of each binder against its respective GPCR target. For each binder, its SMILES code was converted to canonical SMILES code by RDKit[21]. Using the SMILES code, the molecular descriptors and fingerprints of each binder were computed by means of MolMap package[14]. We used MolMap to further convert these molecular descriptors and fingerprints into a MolMap 2D feature map[14], which embeds the broadly-learned correlation relationships of the molecular descriptors and fingerprints in the 2D feature space based UMAP manifold learning[22-23].

1.2 MolMapNet DL architecture

MolMapNet adopts a dual-path CNN architecture, one path is for learning molecular descriptors, and the second path is for simultaneous learning of fingerprint features(Fig. 1)[14]. Following our previous work[14], we choosed 13 classes of molecular descriptors and 3 sets of fingerprints (MACCSFP, PharmacoErGFP, and PubChemFP) for representing the GPCR binders. The first convolution layer of MolMapNet contained 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[24]. Deeper feature extraction processes were conducted through the CNN layers of MolMapNet. The maximum number of parameter of a MolMapNet model was <0.83 million for general tasks, while the robustness of these models was facilitated by the relatively complex topology and depth.

1.3 MolMapNet hyperparameters, training and performance metrics

For all learning tasks, the activation function ReLU was used in MolMapNet along 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 in MolMapNet models because these models were with relatively fewer parameters and was easily trained to convergence. In the regression tasks, mean squared error was used as the loss function. In model training, the early-stopping strategy was used in MolMapNet for avoiding over-fitting problem and for reduced computational cost[10,12-13,25-26]. We performed 10-fold cross-validation(10-FCV) for each model by splitting the full dataset randomly into train set, test set in proportions(0.9 ∶ 0.1) for training and validation purposes respectively. All models were developed by TensorFlow 2.0.0 on DGX-1(32 GB memory in each card). In this work, three popular performance metrics for regression tasks were used for evaluating our developed MolMapNet models, namely, RMSE, MAE and R^2. For RMSE and MAE, each of these metrics were estimated using the predicted pChEMBL values *vs* the ground truth pChEMBL values for agonist/inhibitor-target interactions. R^2 is squared Pearson correlation coefficient between predicted and observed values.

2 Results and Discussions

2.1 Low-sample GPCRs

Most of the 23 low-sample GPCRs in Tab. 1 are being actively explored for drug discovery against a variety of diseases. For instance, efforts have been directed at the discovery of novel agonists against apelin receptor for the treatment of several diseases, due to the involvement of this GPCR in cardiovascular diseases, liver fibrosis, obesity, diabetes and neuroprotection[27]. Antagonists of Neuropeptide S receptor(NPSR) have been developed for the treatment of various CNS disorders, because modulation of this GPCR and the neuropeptide S system are closely associated with CNS disorders such as panic disorder, anxiety, sleeping disorder, asthma, obesity, and substance abuse[28]. Agonists of Oxoeicosanoid receptor 1 have been designed as antiinflammatory and anticancer agents, because this GPCR is involved in inflammatory processes and oncogenesis[29]. Antagonists of Lysophosphatidic acid receptor 5 have been discovered as potential analgesic agents, because this GPCR is highly expressed in spinal cord and dorsal root ganglion associated with pain[30]. Novel antagonists of Succinate receptor 1 have been developed for the treatment of such illnesses as rheumatoid arthritis, liver fibrosis and obesity, because this GPCR senses the citric cycle intermediate succinate and is implicated in these illnesses[31]. Possibly because of the low-data nature of these 23 GPCRs[17], rarely DL and ML model has been published for these GPCRs.

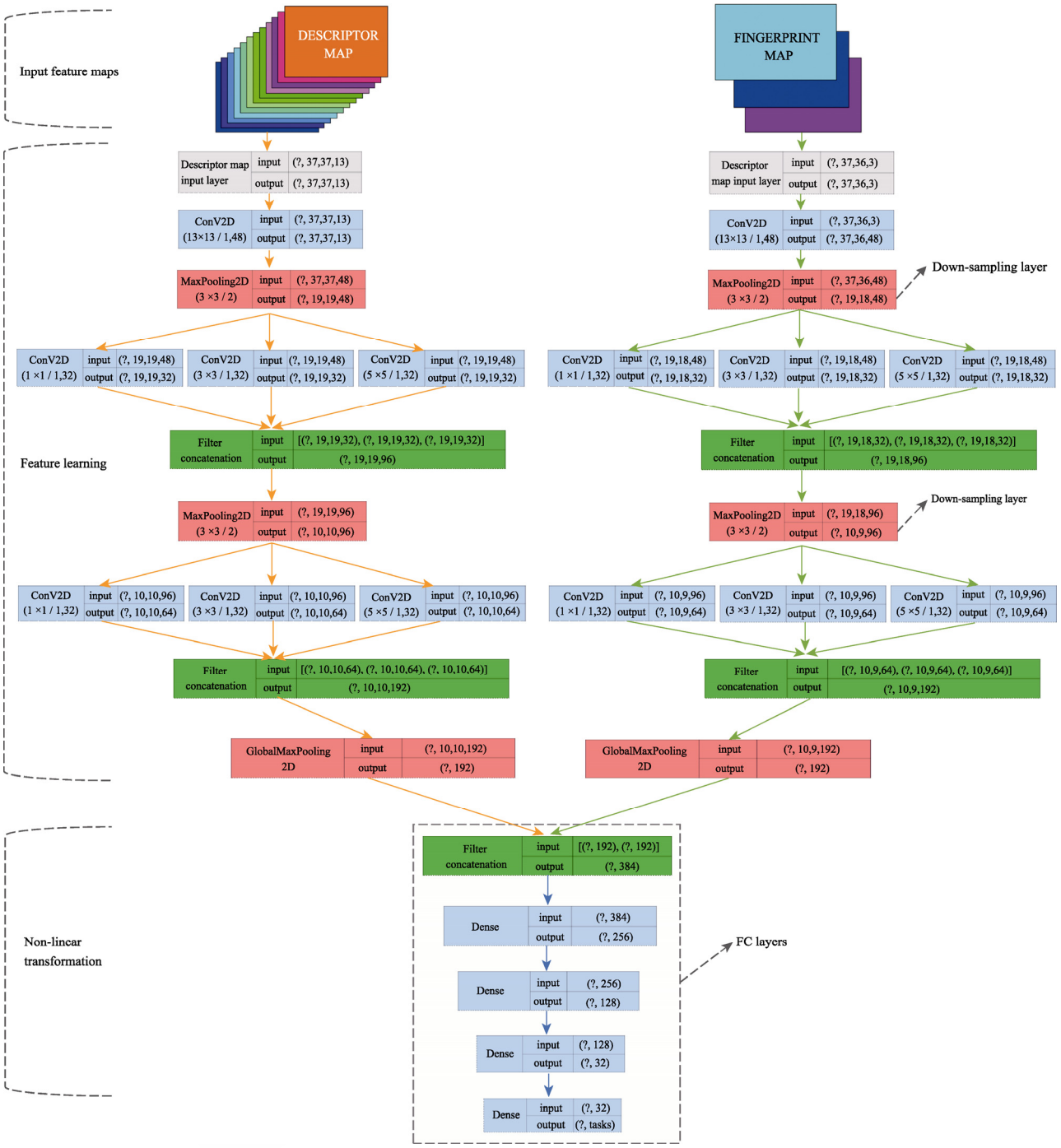

Fig. 1 MolMapNet deep learning architecture[14]

From top to bottom: Multichannel input feature mapping, dual-path CNN feature learning, nonlinear task learning with fully-connected layers.

图 1 深度学习网络 MolMapNet 框架[14]

从上到下依次为：多通道输入特征映射层、双路径 CNN 特征学习层、全连通非线性任务学习层。

2.2 Binding activity prediction performance of MolMapNet models of the low-sample GPCRs

The single-task MolMapNet regression models were trained and tested on each of the 23 low-data GPCRs using three metrics RMSE, MAE and R^2. For each of the 23 low-data GPCRs, the performance of our MolMapNet model was evaluated by 10-FCV as described in the Methods section. The average RMSE, MAE and R^2 value of the 10-FCV results for each GPCR is recorded in Tab. 2. The MAE values of these 23 low-sample GPCR models were tentatively compared with the MAE values of the GCN models of 33 higher-sample GPCRs[18], and the RMSE values of these 23 low-data GPCR models were compared with the RMSE values of the ML models of low-data targets[3-6]. In developing GCN

models, 33 models have been built only on features that are automatically extracted from compound structures by using ensemble learning for higher-data GPCR, and the performance of each model has been evaluated by MAE values on test set[18]. The 33 higher-sample models and 23 low-sample models were mixed and then ranked according to their MAE values regardless of the sample sizes. The models with lower MAE values were ranked higher. The top-10 models with lower MAE values are in Tab. 3. Significantly, 3 of the top-5 and 6 of the top-10 models are MolMapNet models, suggesting that the low-sample MolMapNet models are close in performance to the higher-sample models. The R^2 values of the 23 low-sample models were compared with those of the 33 higher-sample GCR models. Overall, 15 of the 23 low-sample models are with R^2 values higher than the lowest R^2 value(0.51) of the higher-sample models, indicating that the low-sample MolMapNet models are close to the higher-sample models.

Tab. 2 Binding activity prediction performances of models built by using MolMapNet for the binders of 23 low-sample GPCRs

表 2 使用 MolMapNet 在 23 个小样本 GPCRs 配体集上构建的活性预测模型的表现

UniProt ID	No of Known Binders	RMSE	MAE	R^2
Q9BXA5	48	0.480 4	0.428 7	0.508 6
P46092	50	0.842 8	0.728 7	0.500 3
P32241	52	1.199 8	1.008 3	0.454 7
Q13324	58	0.486 5	0.405 0	0.810 7
Q9UBS5	60	0.966 0	0.808 1	0.530 6
Q96G91	65	0.631 7	0.533 8	0.361 6
Q8TDS5	81	0.929 9	0.756 8	0.416 2
Q13304	90	0.770 3	0.650 5	0.639 8
Q6W5P4	104	0.833 1	0.686 8	0.278 3
Q9H1C0	105	0.723 6	0.590 2	0.367 3
P32302	108	0.703 4	0.547 7	0.594 9
O95838	111	0.637 4	0.512 8	0.388 9
P41587	116	0.974 5	0.773 5	0.671 3
Q14831	133	0.373 6	0.299 4	0.756 3
P49019	143	0.649 3	0.524 0	0.534 5
P41586	144	0.981 3	0.804 8	0.138 1
P35414	149	0.760 3	0.600 6	0.719 9
P49683	164	0.664 6	0.526 7	0.136 9
Q9GZQ6	172	0.581 0	0.453 2	0.651 4
P48546	203	0.465 1	0.364 0	0.774 7
Q9UBY5	204	0.767 7	0.603 1	0.584 1
P51685	211	0.850 5	0.679 4	0.637 0
P28336	243	0.625 2	0.439 5	0.721 9

Note: Average RMSE, MAE and R^2 of 10-fold cross-validation results were shown.

注：表中 RMSE、MAE 和 R^2 是 10 折交叉验证结果的均值。

Tab. 3 Comparison of the proposed model with the published DL regression model on MAE

表 3 本文模型和已公布的 DL 回归模型在 MAE 上的比较

GPCR	No of Known Binders	MAE
Metabotropic glutamate receptor 7	133	0.299 4
Orexin receptor 1	2 852	0.36
Gastric inhibitory polypeptide receptor	203	0.364
Corticotropin releasing factor receptor 2	58	0.405
Serotonin 7 (5-HT7) receptor	2 395	0.42
Succinate receptor 1	48	0.428 7
Neuromedin B receptor	243	0.439 5
Orexin receptor 2	3 079	0.45
Neuropeptide FF receptor 1	172	0.453 2
Cannabinoid CB1 receptor	6 966	0.46

Apart from GPCRs, there have been literature-reported ML regression models of 11 low-sample non-GPCR targets(92–170 binders) with computed RMSE and R^2 values[3-6]. These ML regression models explore random forest(RF), support vector regression(SVR), decision trees(DT). These 11 low-sample ML models and the 23 low-sample MolMapNet models were mixed then ranked according to their RMSE values regardless the sample sizes(Tab. 4). The models with lower RMSE values were ranked higher. Overall, 5 of the top-5 and 8 of the top-10 models are the MolMapNet models. Comparison of the R^2 values of the ML models and those of the MolMapNet models showed that 9 of the 23 MolMapNet models are with R^2 values higher than the lowest R^2 values of the 11 ML models. The MolMapNet model with the best R^2 value(0.810 7) is better than the R^2 values of 72% of the ML models.

Tab. 4 Comparison of the proposed model with published ML models on RMSE

表 4 本文模型和已公布的 ML 模型在 RMSE 上的比较

Targets	No of Known Binders	RMSE
Metabotropic glutamate receptor 7	133	0.373 6
Gastric inhibitory polypeptide receptor	203	0.465 1
Succinate receptor 1	48	0.480 4
Corticotropin releasing factor receptor 2	58	0.486 5
Neuropeptide FF receptor 1	172	0.581 0
Trypsin	110	0.610 0
Human serum albumin	95	0.620 0
Neuromedin B receptor	243	0.625 2
Purinergic receptor P2Y11	65	0.631 7
Glucagon-like peptide 2 receptor	111	0.637 4

The performance of MolMapNet measured by MAE and R^2 values also indicates fairly good capability in predicting these metrics for low-data GPCRs. Overall, 6(26%) and 20(87%) of the 23 GPCRs are with average RMSE value <0.6 and <1 respectively, 6(26%) and 21(92%) of the 23 GPCRs are with average MAE value <0.5 and <1 respectively, and 9(39%) and 65(87%) of the 23 GPCRs are with average R^2 value >0.6 and >0.5 respectively. These results indicated that the MolMapNet is a useful tool for low-data learning tasks, and it may be employed for the development of DL models for the prediction of potential binders of low-data GPCRs.

2.3 Influence of data size on the performance of MolMapNet models

The average R^2 values of our MolMapNet models showed some sensitivity to the size of training data. The 23 low-data GPCRs(Tab. 1) can be divided into three groups. The first is the extremely low data group of 40–70 known binders, which inlcudes 6 GPCRs. The second is the intermediately low data group of 71–150 known binders, which contains 11 GPCRs. The third is the fairly low data groups of 151–250 known binders, which consists of 6 GPCRs. In general, a QSAR model for bioactivity prediction is acceptable. when it has an R^2 value > 0.6[32]. Using R^2 value > 0.6 as a tentative criterion for acceptable low-data regression models, there are acceptable MolMapNet models for 1(17%), 4(36%) and 4(67%) of the GPCRs in the extremely, intermediately and fairly low data group respectively (Fig. 2). Therefore, data size has a significantly negative impact on the performance of MolMapNet models in the extremely low data groups, while it has lower impact on the performance of MolMapNet models of the intermediately and fairly low data groups. MolMapNet models were able to score acceptable. R^2 values for more than 36% of the GPCRs in the intermediately and fairly low data groups, indicating its usefulness in low-data learning tasks for GPCRs with >70 known binders.

2.4 Performance of MolMapNet models on novel compounds

To further test our developed MolMapNet models on novel compounds, we searched the PubMed database for the patented compounds against the 23 low-data GPCRs not in the ChEMBL, PubChem and BindingDB database. Our search resulted in the molecular structures of 7 patented agents of NPSR[33]. The experimental activity values (pIC_{50}) of NPSR of these patented agents were obtained from the literature[33]. These patented agents were used for testing the MolMapNet models of NPSR. The structures and activity values of our searched patented agents are provided in Fig. 3 and Tab. 5. The average RMSE, MAE and R^2 values of the 10-FCV MolMapNet models of NPSR on the 7 ligands are 1.504 8, 1.370 3 and 0.674 9 respectively. The MAE between actual binding value and predicted value of compound 7 is only 0.17, while the max extended connectivity fingerprints similarity between compound 7 and 104 known binders of NPSR is 0.38(Fig. 4). These results suggested that MolMapNet models for low-data GPCRs have some capability in predicting the binding activity of some novel compounds, and have the potential to screen drugs with novel structures, thus they can be used to screen potential drugs corresponding to the GPCR target in the early stage of drug development.

3 Conclusion

It is a challenging task to accurately learn and predict the binding activities against a target, particularly for low-data targets and novel molecular structures[34-36]. A key factor for enhanced learning and prediction capability is the appropriate representations of the compounds[10,37-41]. Our

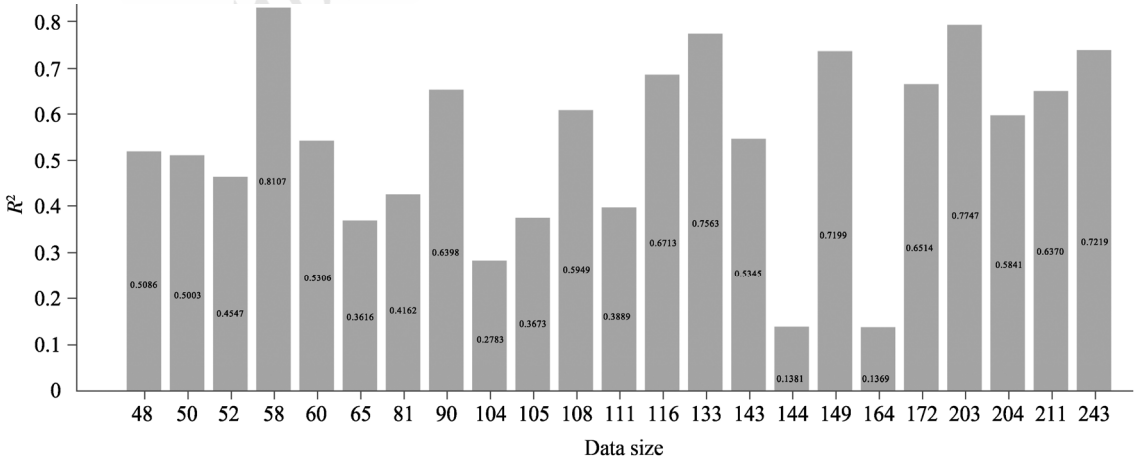

Fig. 2 Relationship between model performance and data volume

图 2 模型表现与数据量的关系图

Tab. 5 Selected patented neuropeptide S receptor binders, experimental activities, MolMapNet predicted activities, and the structural similarity to the closest compound of the training dataset of known

表 5 专利中神经肽 S 受体配体的实验活性，MolMapNet 预测活性，以及与已知训练数据集最接近的配体的结构相似性

ID	SMILES of Patented Compound	Experimental pIC_{50}	Predicted value	Structural Similarity to the known binders in trying dataset
3b	Cc1ccccc1CN1CCN(C2(c3ccccc3)C(=O)c3ccccc3C2=O)CC1	8.52	6.793 6	>0.74
3c	COc1ccccc1CN1CCN(C2(c3ccccc3)C(=O)c3ccccc3C2=O)CC1	8.22	6.870 0	>0.74
3t	O=C1c2ccccc2CC1(c1ccccc1)N1CCN(Cc2ccccc2)CC1	7.77	6.774 5	0.74
3u	Cc1ccccc1CN1CCN(C2(c3ccccc3)Cc3ccccc3C2=O)CC1	8.7	6.616 1	0.60

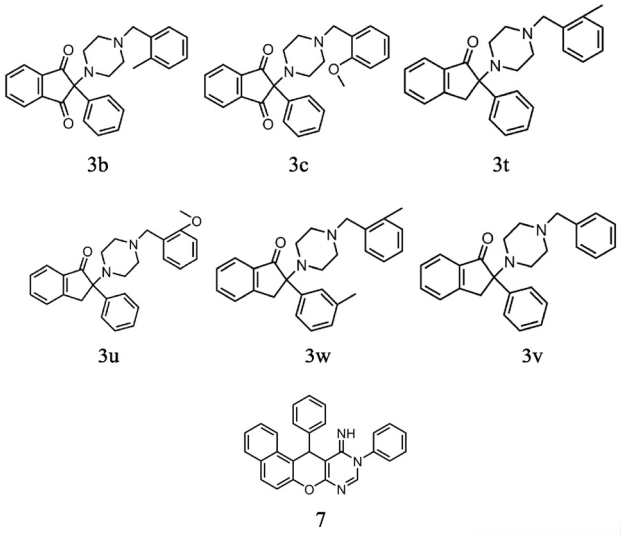

Fig. 3 Molecular structures of the selected patented neuropeptide S receptor binders

图 3 专利中神经肽 S 受体化合物分子结构图

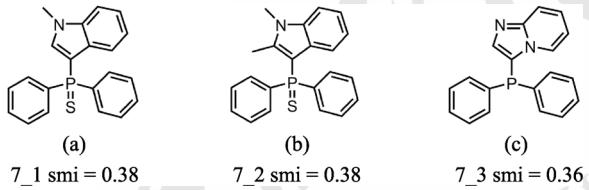

Fig. 4 Molecular structure and corresponding similarity value of the NPSR ligand with the highest similarity of ECFP to patent compound 7

图 4 与 7 号专利化合物 ECFP 相似性最高的 NPSR 配体的分子结构及对应相似值

MolMapNet models is based on broad learning of knowledge-based molecular representations, which enable the restructuring of unordered molecular descriptors and fingerprints into ordered 2D feature maps for subsequent DL with CNN architecture[14]. MolMapNet and other advanced DL methods has shown good prediction performances for various activity, toxicity and pharmacokinetic properties[14]. Our study in this work further demonstrated the capability of MolMapNet for binding activity prediction against the low-data GPCRs. DL algorithms that explore wider variety of feature representations, e.g. the graph-based DNN fingerprints[42], have continuously progressed. The collective exploration of these and other established strategies enable more enhanced DL and prediction of molecular binding activities and other pharmaceutical properties.

REFERENCES

[1] HAUSER A S, ATTWOOD M M, RASK-ANDERSEN M, et al. Trends in GPCR drug discovery: New agents, targets and indications[J]. Nat Rev Drug Discov, 2017, 16(12): 829-842.

[2] WEISS D R, KARPIAK J, HUANG X P, et al. Selectivity challenges in docking screens for GPCR targets and antitargets[J]. J Med Chem, 2018, 61(15): 6830-6845.

[3] KUNDU I, PAUL G, BANERJEE R. A machine learning approach towards the prediction of protein-ligand binding affinity based on fundamental molecular properties[J]. RSC Adv, 2018, 8(22): 12127-12137.

[4] XUE C X, ZHANG R S, LIU H X, et al. QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine[J]. J Chem Inf Comput Sci, 2004, 44(5): 1693-1700.

[5] DENG W, BRENEMAN C, EMBRECHTS M J. Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods[J]. J Chem Inf Comput Sci, 2004, 44(2): 699-703.

[6] WANG Y, GUO Y Z, KUANG Q F, et al. A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach[J]. J Comput Aided Mol Des, 2015, 29(4): 349-360.

[7] BUSHDID C, DE MARCH C A, FIORUCCI S, et al. Agonists of G-protein-coupled odorant receptors are predicted from chemical features[J]. J Phys Chem Lett, 2018, 9(9): 2235-2240.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.
[13] 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.
[14] 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.
[15] WU J S, ZHANG Q M, WU W J, et al. WDL-RF: Predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest[J]. Bioinformatics, 2018, 34(13): 2271-2282.
[16] WHITE J. PubMed 2.0[J]. Med Ref Serv Q, 2020, 39(4): 382-387.
[17] 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.
[18] SAKAI M, NAGAYASU K, SHIBUI N, et al. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks[J]. Sci Rep, 2021, 11(1): 525.
[19] YAO Q, WANG M, CHEN Y, et al. Taking human out of learning applications: A survey on automated machine learning[J]. arXiv e-prints, 2018, 1810.13306:1-20.
[20] CONSORTIUM U. UniProt: the universal protein knowledgebase in 2021[J]. Nucleic Acids Res, 2021, 49(D1): D480-D489.
[21] LANDRUM G. Rdkit documentation[J]. Release, 2013, 1(1-79): 4.
[22] 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.
[23] MCINNES L, HEALY J, MELVILLE J. UMAP: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.
[24] PENG C, ZHANG X Y, YU G, et al. Large kernel matters—improve semantic segmentation by global convolutional network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4353-4361.
[25] 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.
[26] 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.
[27] NARAYANAN S, MAITRA R, DESCHAMPS J R, et al. Discovery of a novel small molecule agonist scaffold for the APJ receptor[J]. Bioorg Med Chem, 2016, 24(16): 3758-3770.
[28] HASSLER C, ZHANG Y N, GILMOUR B, et al. Identification of neuropeptide S antagonists: Structure-activity relationship studies, X-ray crystallography, and *in vivo* evaluation[J]. ACS Chem Neurosci, 2014, 5(8): 731-744.
[29] STEPNIEWSKI T M, TORRENS-FONTANALS M, RODRÍGUEZ-ESPIGARES I, et al. Synthesis, molecular modelling studies and biological evaluation of new oxoeicosanoid receptor 1 agonists[J]. Bioorg Med Chem, 2018, 26(12): 3580-3587.
[30] KAWAMOTO Y, SEO R, MURAI N, et al. Identification of potent lysophosphatidic acid receptor 5 (LPA5) antagonists as potential analgesic agents[J]. Bioorg Med Chem, 2018, 26(1): 257-265.
[31] VELCICKY J, WILCKEN R, COTESTA S, et al. Discovery and optimization of novel SUCNR1 inhibitors: Design of zwitterionic derivatives with a salt bridge for the improvement of oral exposure[J]. J Med Chem, 2020, 63(17): 9856-9875.
[32] FRIMAYANTI N, YAM M L, LEE H B, et al. Validation of quantitative structure-activity relationship (QSAR) model for photosensitizer activity prediction[J]. Int J Mol Sci, 2011, 12(12): 8626-8644.
[33] RUZZA C, CALÒ G, DI MARO S, et al. Neuropeptide S receptor ligands: A patent review (2005-2016)[J]. Expert Opin Ther Pat, 2017, 27(3): 347-362.
[34] VAN DE WATERBEEMD H, GIFFORD E. ADMET in silico modelling: Towards prediction paradise? [J]. Nat Rev Drug Discov, 2003, 2(3): 192-204.
[35] 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.
[36] 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.
[37] LIPINSKI C A, LOMBARDO F, DOMINY B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[J]. Adv Drug Deliv Rev, 2001, 46(1/2/3): 3-26.
[38] PAOLINI G V, SHAPLAND R H B, VAN HOORN W P, et al. Global mapping of pharmacological space[J]. Nat Biotechnol, 2006, 24(7): 805-815.
[39] WILLETT P. Similarity-based virtual screening using 2D fingerprints[J]. Drug Discov Today, 2006, 11(23/24): 1046-1053.
[40] 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.
[41] 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.
[42] DUVENAUD D K, MACLAURIN D, IPARRAGUIRRE J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. NeurIPS Proceedings , 2015, 28.

收稿日期：2022-09-07
(本文责编：沈倩)