Artificial Intelligence in the Life Sciences 3 (2023) 100065

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

An improved 3D quantitative structure-activity relationships (QSAR) of molecules with CNN-based partial least squares model

Xuxiang Huo[a,b], Jun Xu[b,c,*], Mingyuan Xu[a,*], Hongming Chen[a,*]

[a] Guangzhou Laboratory, Guangzhou 510005, China
[b] School of Biotechnology and Health Sciences, Wuyi University, Jiangmen 529020, China
[c] School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China

A B S T R A C T

Ligand-based virtual screening plays an important role for cases in which protein structures are not available. Among ligand-based methods, accurate and fast prediction of protein-ligand binding affinity is crucial for reducing computational cost and exploring the chemical search space efficiently. Here we proposed a CNN-based method, termed as L3D-PLS for building the quantitative structure-activity relationships without target structures. In L3D-PLS, a CNN module was designed for extracting the key interaction features from the grids around aligned ligands, and a partial least square (PLS) model fits the binding affinity with the extracted features of the pre-trained CNN module. In 30 publicly available pre-aligned molecular datasets, L3D-PLS outperformed the traditional CoMFA method. This results highlight that L3D-PLS can be useful for lead optimization based on small datasets which is often true in drug discovery compaign.

Introduction

Ligand-based virtual screening, one of the commonly used computer-aided drug discovery approaches [1], has become routine process for lead identification, lead optimization and scaffold hopping. Among ligand based methods, quantitative structure-active relationships (QSAR) analysis aims to find the statistically significant correlations between a series of molecular structures and their associated biological properties [2], guide the optimization of lead series.

Conventional QSAR models are built usually using molecular physicochemical properties/descriptors [3], composition of specific substructures or force field based interaction energies etc. Many statistic methods including multiple-linear regression (MLR) [4], principal component analysis (PCA) [5], principal component regression (PCR) [6], partial least square (PLS) [7] are employed for fitting those descriptors against the biological activities or DMPK properties [8]. The common workflow of QSAR model can be described in Fig. 1.

There are usually two scenarios in predicting ligand-protein binding affinity, one is the type of method that employs physics based simulation using protein-ligand complex structures, like FEP [9], thermodynamic integration [10] or umbrella sampling methods [11]. Derived from statistical mechanics, these methods could provide accurate estimation of protein-ligand binding affinity with expensive complex-structure sampling computations. For example, FEP+ shows considerable correlations between calculated and experimental relative binding free energies, and average errors in the range of only 1 kcal/mol [12]. Although available protein structures have increased dramatically in recent decades, for some targets such as ion channels, getting target structure information is still challenge and it is still very common that for some drug discovery compaigns target structure information is missing. In the other type of scenario, one has to rely on ligand based method to estimate ligand binding affinitywhich also called quantitative structure-activity relationships modeling (QSAR), Although historically some 2D based QSAR methods have been proposed [13], such as Free-Wilson [14], Hansh-Fujita methods [15] and fragment based QSAR models propsed by Klopman,.3D based QSAR models using molecular interaction field (MIF) [16] can provide chemical insights regarding the relationship between 3D based molecular properties and biological activities. However, the biological properties of a compound are the function of its three dimensional structure, therefore these methods still suffer from limitations due to the lack of 3D ligand structural information [17], which otherwise can provide explicitly guidelines for medicinal chemists to elucidate ligand binding mode, and modify chemical structures for improving biological activity. To address this problem, comparative molecular field analysis (CoMFA) method was the first and commonly used attempt to unitize the 3D spatial information in QSAR modeling [18].

In CoMFA, a series of probe atoms are introduced to describe 3D features of a ligand by calculating the electrostatic and steric interaction energies between probe atom and ligand on the 3D lattice surrounding molecules. The PLS method is further used to extract the relationships between molecular interaction field (MIF) and biological activity. Following CoMFA method, other 3D QSAR methods like CoMSIA [19], EVA [20], WHIM [21] also proposed in bridging the molecular structure and bio-activity. These 3D QSAR models require selection of a template conformation which mimicks the bioactive conformation of ligand and pre-alignment of all template conformations. One difficulty that these

* Corresponding authors.
E-mail addresses: junxu@biocheemomes.com (J. Xu), mingyuan.xu.sci@gmail.com (M. Xu), chen_hongming@gzlab.ac.cn (H. Chen).

https://doi.org/10.1016/j.ailsci.2023.100065
Received 7 November 2022; Received in revised form 21 January 2023; Accepted 21 February 2023
Available online 24 February 2023

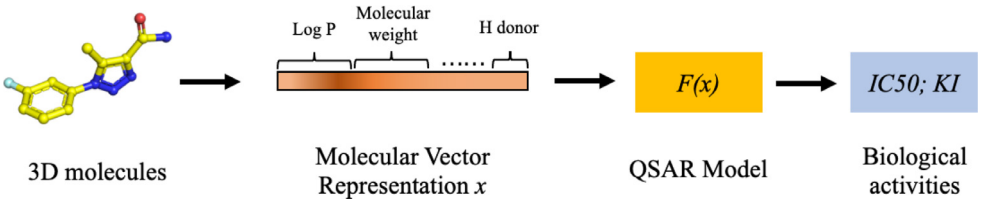

Fig. 1. The workflow of QSAR modeling.

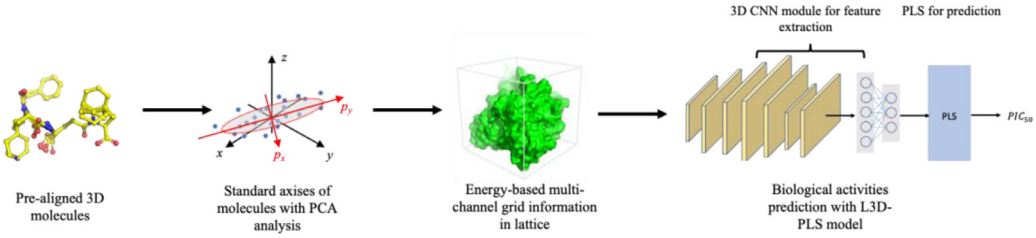

Fig. 2. The workflow of L3D-PLS for binding affinity predictions.

methods suffer from is the variable selection among thousands of interaction features on the 3D grid.

Recent decades, it has been seen that machine learning technologies have brought new directions into the field of QSAR in the form of supervised, unsupervised, semi-supervised learning, especially on accurate prediction of binding affinity [22]. Early attempts include introduction of models like linear regression [23], kernel ridge regression [24], support vector machine [25] and random forest [26] to improve the accuracy of predictions. For example, a random-forest regression model, RF-score [27], was proposed to predict protein-ligand binding affinity. With recent development of deep learning (DL) methods, examples like RosENet [28], GAT-Score [29] tried to introduce CNN (convolutional neural network) model or graph neural network [30] to extract structural features of protein-ligand complex structures and greatly improve the prediction accuracy of protein-ligand binding affinity [31]. One unique feature of DL based method is that they can handle large input features and can carry out automatic feature extraction. Although DL based methods have been applied for study of protein-ligand interaction, there still lack of research on applying DL method on ligand based 3D QSAR model, this maybe due to the fact that the size of ligand based QSAR datasets is usually small.

Here, we propose a new grid-based 3D QSAR method, L3D-PLS for modeling protein-ligand binding affinity based on ligand information only. In L3D-PLS, a CNN module was designed for deriving ligand electrostatic and steric potential related features from the multi-channel 3D grid information of pre-aligned ligands and a partial least squares (PLS) model fits the binding affinity with the output of the pre-trained CNN module. In a series of 30 publicly available pre-aligned molecular datasets, L3D-PLS outperformed the traditional CoMFA method

Methods

A. L3D-PLS workflow

The whole work-flow for L3D-PLS model is shown in Fig. 2. Firstly, the aligned molecules go through a preparation procedure to standardize molecular positions to make sure they are rotation and translation invariance. The origin of the coordinate system are moved to the geometric center of the aligned conformations and the XYZ axes are also rotated to align with the principal axes of the conformation set. Secondly, a grid box is created to encircle the whole aligned conformation set with certain step size, then a set of probe atoms walked through the grids for calculating the interaction energies between the probe atom and every molecule on each grid point. Once the energy based features are collected for the molecules, multi-layer CNN model is used for feature extraction from the multi-channel grid information and a PLS module are involved to correlate the CNN derived features with bioactivity data.

B. Dataset and the standard orientation

Altogether 30 publicly available datasets were used for model evaluation and were downloaded from PyCoMFA website [32]. These datasets were used by PyCoMFA program which is a python version of CoMFA model and in each dataset, all molecules were pre-aligned. To make sure models are not affected by the initial rotation and translation, all the molecular structures in each dataset went through a standardization process. The origin of the coordinate system were first moved to the geometric center of the whole conformation set and then rotated to align with the principal axes of the conformation set. The orientation vector $\{\alpha, \beta, \gamma\}$ of principal axises against original axis is obtained by doing PCA analysis of the whole set of molecular coordinates $\boldsymbol{C}$ as shown in Eq (1).

$$\{\alpha, \beta, \gamma\} = F_{PCA}(\boldsymbol{C}) \tag{1}$$

Then the rotated coordinates C_i' of ith molecules can be obtained with Eq (2).

$$\begin{aligned} C_i' &= (C_i^0 - C_{mass}) R_X(\alpha) R_Y(\beta) R_z(\gamma) \\ &= (C_i^0 - C_{mass}) \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\alpha & sin\alpha \\ 0 & -sin\alpha & cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & sin\beta \\ 0 & 1 & 0 \\ -sin\beta & 0 & \cos\beta \end{bmatrix} \\ &\times \begin{bmatrix} \cos\gamma & -sin\gamma & 0 \\ sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{2}$$

Where C_i^0 represents the original coordinates of ith molecule, R_{mass} is the geometric center of molecule sets.

C. Generation of molecular interaction field grids

Generation of molecular interaction field (MIF) grids of pre-aligned molecules are same as CoMFA methods [33], derived from semi-empirical free capacity field embedding in Autogrid [4]. Eight types of probe atom, including $H^*, HD^*, C^*, A^*, N^*, NA^*, OA^*$ and SA^*, are used by default, as illustrated in Table 1. The interactions between probe atom i and molecule atom j are as shown in Eq (3)~Eq (5).

$$V_{ij}(r_{ij}) = \frac{C_i}{r_{ij}^n} - -\frac{C_j}{r_{ij}^m} \tag{3}$$

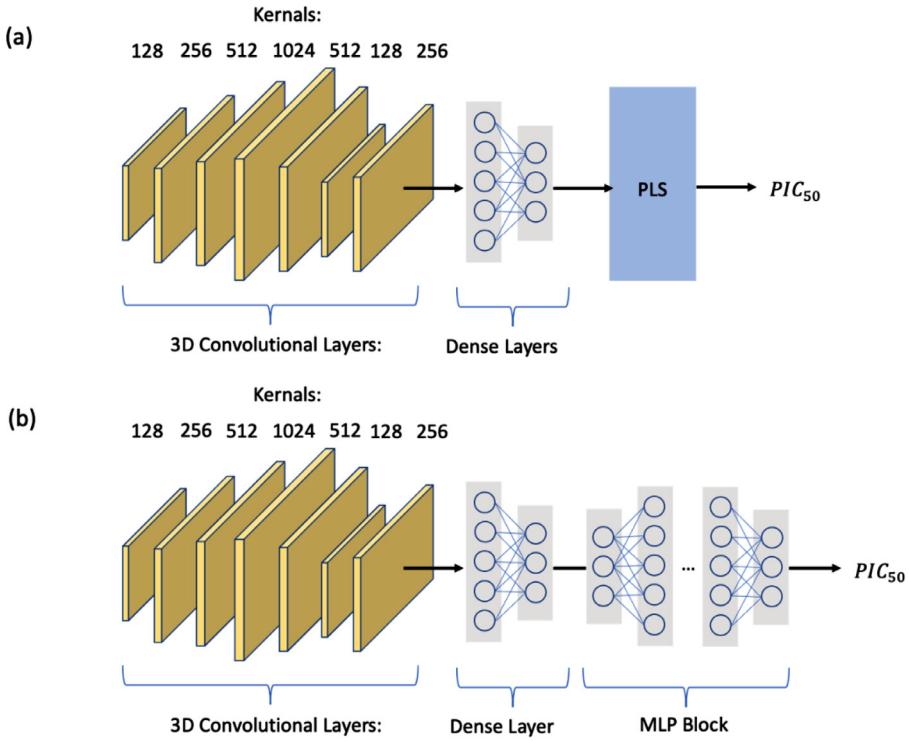

Fig. 3. The model structure of (a) L3D-PLS and (b) L3D-MLP.

Table 1
Eight default probe atoms types in Autogrid 4 MIF calculations.

symbol	Probe atoms
H*	Non H-bonding Hydrogen
HD*	Donor 1 H-bond Hydrogen
C*	Non H-bonding Aliphatic Carbon
A*	Non H-bonding Aromatic Carbon
N*	Non H-bonding Nitrogen
NA*	Acceptor 1 H-bond Nitrogen
OA*	Acceptor 2 H-bonds Oxygen
SA*	Acceptor 2 H-bonds Sulphur

$$C_i = \frac{m}{n-m} * \varepsilon * r_{eq}^{n} \tag{4}$$

$$C_j = \frac{n}{n-m} * \varepsilon * r_{eq}^{m} \tag{5}$$

Where the parameters are: r_{eq}, ε, n, m, and the two atom types, where r_{eq} is the pre-defined equilibribium distance for the bottom of the interaction energetic well between atom i and j, ε is the depth of the well, n and m are the coefficients in Autogrid [4]. The MIF grid data are calculated in a lattice box with size of 30 Å and interval of 0.375 Å.

D. L3D-PLS model construction

In current study, besides the L3D-PLS model, another model L3D-MLP was also created for comparison. The model architectures are depicted in Fig. 3.

As seen in Fig. 3, altogether seven convolutional layers and two dense layers are used for processing grid data to generate embedding features. The size of convolutional kernels of seven 3D convolutional layers are 128, 256, 512, 1024, 512, 128 and 256 respectively. The model output is pIC_{50} value of compound. For the L3D-MLP model, six extra dense layers (i.e. the MLP block) were added to correlate with bioactivity data, while in the L3D-PLS model, a PLS model replaced the MLP block to predict bioactivity. For both models, the mean standard error (MSE) of pIC_{50} was used as the loss function as following:

$$Loss = MSE\left(pIC_{50}^{predict}, pIC_{50}^{reference}\right) \tag{6}$$

In this work, Pytorch framework was applied to build the CNN model, the PLS code of this workflow is from Scikit-learn python package.

E. model evaluation

The composition of benchmark datasets is listed in Table 2. The performance of L3D-PLS on 30 different publicly available datasets is evaluated with correlation coefficient R^2 and cross-validated Q^2. For a set of predicts y^{pred} and references y^{ref}, R^2 and Q^2 are both defined as shown in Eq (7):

$$R^2 = 1 - \sum (y^{ref} - y^{pred})^2 \,/\, \sum (y^{ref} - \overline{y^{ref}})^2 \tag{7}$$

The R^2 was calculated for compounds in test set, however, for small benchmark datasets like AT2, a 5-fold cross-validation was performed, and Q^2 on cross-validation was calculated to evaluate the model performance.

For training the L3D-PLS model, the input of last dense layer of the L3D-MLP model was taken as the input descriptor and the PLS model was then trained with the descriptors, in which the MSE Loss in Eq (6) was used. Both the L3D-PLS and L3D-MLP were trained with the piece-wise constant attenuation of learning rate decay scheme and Adam optimizer. ReLU activation function was used in the model.

Results and discussion

A. The influence of translation, rotation and scaling of the molecular lattice on bio-activity predictions

Current study aims for predicting the binding affinity of small molecules against proteins using deep learning methods based on 3D-QSAR method. Here we introduce a method which combines the idea

Table 2
Composition of 30 public QSAR benchmark datasests.

Index of dataset	Dataset	Numbers of molecules			Activity range	
		Total	Training	Test	Training	Test
1	ACE	114	76	38	2.14~9.88	2.7~9.94
2	ACHE	111	74	37	4.34~9.52	4.27~9.22
3	BZR	147	98	49	6.34~8.92	5.52~8.85
4	GPB	66	44	22	1.3~4.8	1.4~6.8
5	COX2	282	188	94	4.03~8.77	4.03~8.7
6	DHFR	361	237	124	3.3~9.81	3.57~9.4
7	THERM	76	51	25	0.52~8.82	0.52~10.17
8	THR-1	88	59	29	4.57~8.48	4.36~8.38
9	ATA	94	72	22	2.64~7.53	2.64~6.72
10	AT2	28	28	N/A	4.27~8.2	N/A
11	CCR5	75	63	12	5.14~8.07	5.29~8.64
12	YOPH	39	35	4	2.26~6.628	2.85~6.178
13	KOA	39	31	8	7.66~9.602	6.921~9.553
14	MX	29	29	N/A	−2.41~8.65	N/A
15	DAT	42	36	6	3.76~7.81	6.04~7.16
16	TP2A	25	25	N/A	2.89~6.43	N/A
17	CBRA	32	32	N/A	4.3~7	N/A
18	AI	78	78	N/A	−1.65~2.85	N/A
19	HIVPR	113	93	20	5.24~10.96	6.8~10.7
20	GSK3B	42	34	8	6~8.15	6.89~8.4
21	STEROIDS	21	21	N/A	5.322~9.74	N/A
22	GHS	31	31	N/A	5.05~8.52	N/A
23	D2R	38	32	6	5.66~10.3	5.65~8.55
24	D4R	38	32	6	6.28~10.3	7.22~9.37
25	DIAZEPAM DI/DS	42	42	N/A	−3.42~0.67	N/A
26	DIAZEPAM DI	42	42	N/A	5.55~8.77	N/A
27	DIAZEPAM DS	42	42	N/A	6.41~10	N/A
28	THR-2	88	72	16	4.357~8.377	4.745~8.481
29	TRY	88	72	16	4.796~7.699	4.337~7.638
30	FXA	88	72	16	3.745~6.046	4.284~5.509

*N/A means data is not available.

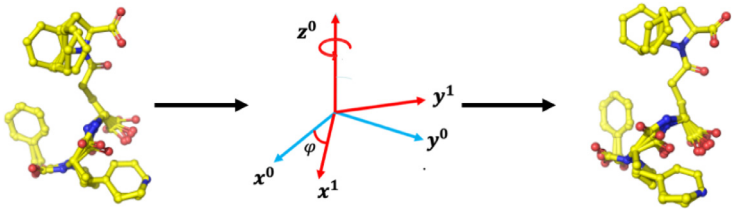

Fig. 4. The rotation of molecule among z-axis with the rotation matrix.

of calculating molecular field on 3D grid which was adopted in CoMFA method and the deep neural network approach. The 3D based features of small molecules structures can be derived by applying convolutional neural network on the 3D molecular field data and they are fit with the binding affinities. But the interaction data derived from 3D grid and molecular conformations is rotational, and translational variant. So we first examine how the translation and rotation operation can affect the model quality.

In order to verify the effect of rotation on the binding affinity predictions, we built L3D-PLS and L3D-MLP models on BZR dataset as an example case when the aligned conformations were rotated around Z-axis from 0° to 360° with step interval of 20°, while the grids are fixed. The model performance is shown in Table 3. When a molecule is rotated round Z-axis with torsion of ϕ, the rotation matrix is represented as shown in Eq (8):(Fig. 4)

$$R_z(\varphi) = \begin{bmatrix} \cos\varphi & -sin\varphi & 0 \\ sin\varphi & cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

So the new molecular coordinates *C'* after rotation can be computed as Eq (9).

$$C' = C_0 \times R_z(\varphi) \tag{9}$$

where C_0 is the original coordinate.

Table 3
The R^2 on BZR test set under different rotations.

Dataset	Rotations (°)	L3D-MLP	L3D-PLS
BZR	0	0.127	0.026
	20	−0.011	−0.154
	40	−0.041	−0.411
	60	−0.136	−0.314
	80	0.073	−0.075
	100	−0.111	−0.377
	120	−0.247	−0.485
	140	−0.051	−0.066
	160	−0.066	−0.331
	180	−0.276	−0.519
	200	−0.111	−0.38
	220	−0.301	−1.061
	240	−0.181	−0.579
	260	−0.177	−0.511
	280	−0.197	−0.552
	300	−0.118	−0.254
	320	−0.206	−0.622
	340	−0.106	−0.242

As it can be seen in Table 3, due to the fact that 3D-CNN is not rotation invariance, the rotation of small molecules in the grid space has clearly affected the accuracy of model quality. In addition, we also studied the effect of translation of molecules in the grid box on the result

Table 4
The R^2 on BZR and TP2A dataset with different translation.

Dataset	Model	Original result	The result after translation
BZR	L3D-MLP	0.127	0.0921
	L3D-PLS	0.025	0.022
TP2A	L3D-MLP	0.428	0.439
	L3D-PLS	0.917	0.865

by simply increase all conformations 1 Å along the X axis in lattice, the result is as shown in Table 4. There, slightly decrease of prediction accuracy is observed, indicating the 3D-CNN based feature extraction may not be that sensitive on translation.

The influence of the size of grid box on accuracy of model prediction was also examined. The size of grid box was changed and models were built on three different datasets and the results can be seen in Table 5. It seems that when the size increased from 51 Å to 61 Å, its accuracy slightly decreases in general, which means too large box is not good for CNN-based feature extraction.

B. The performance of L3D-PLS on 30 datasets

To address the rotation and translation variance mentioned above, we designed a standard preparation step to make the aligned molecules rotational and translational invariance as discribed in the method part. We compared the performance of three QSAR models including PyCoMFA, L3D-MLP and L3D-PLS. The R^2 between predictions and references on 20 test sets are shown in Table 6.

L3D-PLS performed best in 10 datasets, PyCoMFA and L3D-MLP had best performance on 7 and 3 datasets, respectively. Comparing with PyCoMFA model alone on these 20 benchmarks, L3D-MLP and L3D-PLS show superior performance on 10 and 12 datasets.

For other benchmark datasets without available test sets, the performance of three QSAR models were evaluated through Q^2 in 5-fold cross validation and the results are listed in Table 7. Similarly, L3D-PLS models performed best among 6 of 10 models, while PyCoMFA and L3D-MLP models give best result on three and one datasets, respectively. The poor performance of a L3D-MLP model reflects the limitation of deep-learning methods on small datasets, while the combination of 3D-CNN and PLS method out-performed the traditional CoMFA method, indicating the advantages of 3D-CNN method in catching the key interaction features from high-dimensional data.

The performance of L3D-PLS model on the 9 datasets in terms of mean absolute error (MAE) and root mean square error also outperformed L3D-MLP and PyCoMFA mdoel, in which L3D-PLS model achieved better performance among 8 of 9 benchmark datasets (as listed in Table S1). The average of RMSE and MAE of L3D-PLS is 0.366 and 0.283, both outperforms than PyCoMFA and L3D-MLP. We also performed the Wilcoxon signed-rank test to verify the statistically advantages of L3D-PLS. In the test between RMSE results of PyCoMFA and L3D-MLP, the p-value is 0. 017, while it is 0.028 for MAE results. and it is 0.05 and 0.07 in the test between L3D-PLS and L3D-MLP for RMSE and MAE results, respectively. It indicates a statistically advantages of L3D-PLS. The correlation between experimental pIC_{50} values and L3D-PLS predictions, PyCoMFA, L3D-MLP on these random selected datasets are depicted in Fig. 5, S2 and S3.

Table 5
The influence of grid size on model performance.

Dataset	Model	The size of grid			
		51 × 51 × 51	53 × 53 × 53	55 × 55 × 55	61 × 61 × 61
GPB	L3D-MLP	0.284	0.276	0.262	0.251
	L3D-PLS	0.258	0.248	0.251	0.241
D2R	L3D-MLP	0.034	0.030	0.027	0.028
	L3D-PLS	0.512	0.501	0.460	0.432
THR	L3D-MLP	0.421	0.417	0.419	0.417
	L3D-PLS	0.440	0.446	0.432	0.434

Table 6
R^2 and RMSE between predictions of 3 models and reference on the 20 datasets.

Datasets	Number of molecules	R^2			Root mean square error (RMSE)		
		PyCoMFA	L3D-PLS	L3D-MLP	PyCoMFA	L3D-PLS	L3D-MLP
ACE	114	0.523	0.336	0.374	N/A	0.332	0.320
Ache	111	0.525	0.332	0.264	N/A	0.298	0.379
BZR	147	0.046	0.025	0.127	N/A	0.451	0.472
GPB	66	0.246	0.257	0.251	0.466	0.272	0.700
COX2	282	0.072	0.322	0.368	N/A	0.364	0.328
DHFR	361	0.569	0.261	0.328	N/A	0.349	0.413
THERM	76	0.565	0.521	0.466	N/A	0.275	0.295
THR	88	0.662	0.433	0.417	N/A	0.273	0.265
ATA	94	−1.402	−1.070	−0.253	N/A	1.136	0.903
CCR5	75	−0.302	0.221	−0.112	0.214	0.214	1.007
YOPH	39	0.933	0.767	0.703	0.141	0.361	0.428
KOA	39	0.660	0.695	0.311	0.203	0.144	0.460
DAT	42	−4.323	0.151	−0.446	0.175	0.139	0.166
HIVPR	113	0.497	0.502	0.401	N/A	0.253	0.377
GSK3B	42	0.266	0.366	0.282	N/A	0.278	0.520
D2R	38	0.420	0.432	0.028	0.208	0.186	0.399
D4R	38	−0.134	0.365	−0.106	0.455	0.368	0.389
THR-2	88	0.416	0.433	0.417	N/A	0.273	0.265
TRY	88	0.655	0.564	0.142	N/A	0.272	0.427
FXA	88	−0.160	0.183	−0.042	N/A	0.378	0.556

*N/A means data is not available.

Table 7
Q^2 and RMSE between predictions of 3 models and reference on the 20 datasets.

Datasets	Number of molecules	Q^2			Root mean square error (RMSE)		
		PyCoMFA	L3D-PLS	L3D-MLP	PyCoMFA	L3D-PLS	L3D-MLP
AT2	28	0.190	0.320	0.370	N/A	0.375	0.341
MX	29	0.770	0.750	0.110	0.639	0.688	1.329
TP2A	25	0.620	0.920	0.430	N/A	0.129	0.321
CBRA	32	0.620	0.690	0.510	0.531	0.156	0.837
AI	78	0.500	0.220	0.190	N/A	0.403	0.413
DIAZEPAM_DS	42	0.420	0.730	0.060	0.498	0.326	0.726
STEROIDS	21	0.700	0.720	0.050	0.742	0.699	3.376
GHS	31	0.320	0.520	0.380	0.554	0.469	0.530
DIAZEPAM_DI	42	0.420	0.730	0.060	N/A	0.206	0.500

*N/A means data is not available.

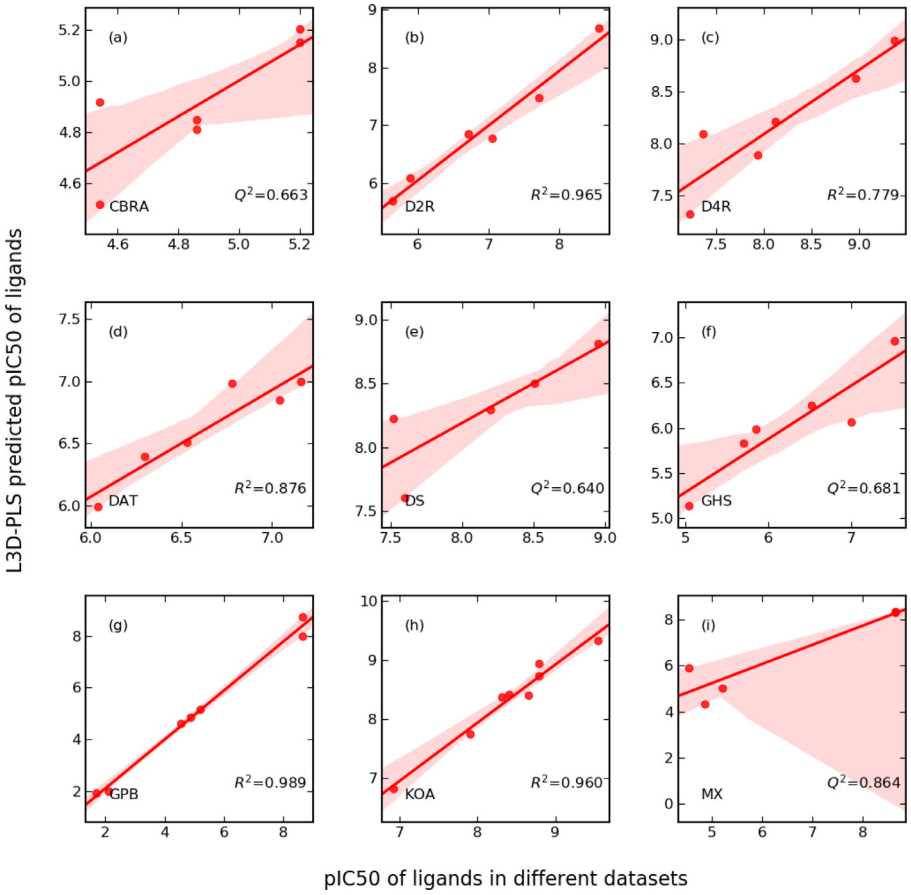

Fig. 5. The correlation between reference pIC_{50} values and L3D-PLS predictions on 9 random selected datasets including (a) CBRA, (b) D2R, (c) D4R, (d) DAT, (e) DS, (f) GHS, (g) GPB, (h) KOA, (i) MX.

Conclusions

In current study, 3D-CNN model is proposed to derive spatial and electrostatic features for pre-aligned small molecules through multi-channel grids, these features are then combined with PLS algorithm to fit bioactivity data. This methodology is useful for carrying out 3D QSAR study when target information is not available, and it has been applied on 30 publicly available molecular datasets. Two variants L3D-MLP, L3D-PLS models were built and compared with the traditional 3D QSAR method, CoMFA. Our results show that L3D-PLS model perform best, while L3D-MLP model perform worse than CoMFA method. The poor performance of the L3D-MLP may be due to the fact that all the 30 datasets are not big dataset, so that deep learning method doesn't demonstrate it advantage. Anyhow, the L3D-PLS model, which combines 3D-CNN based feature extraction and PLS based linear correlation technology, demonstrated improved results. The interpretability of L3D-PLS is limited by its convolution mechanism. Although the contributions of a chemical group on pIC50 can be estimated from the difference between original ligands and analogs without this chemical group, it's still difficult to privide pharmocophore informations from single predictions. This method is still useful for building 3D QSAR model on small datasets when only ligand information is avaliable. And in next steps, graph attention mechanism will be introduced to improve the interpretability.

Code availability

The source code of L3D-MLP and L3D-PLS are available from https://github.com/huoxuxinag/L3D-PLS.git.

Supporting information available

1. The evolution of loss, learning rate and r^2 in L3D-PLS training process is as shown in Figure S1.
2. The correlation between reference pIC_{50} values, Pycomfa and L3D-MLP predictions are as shown in Figure S2~S3.

3. The MAE and RMSE of PyCoMFA, L3D-MLP and L3D-PLS predictions are as shown in Table S1. The R^2 of L3D-PLS trained on three datasets with randomized MIFs and original MIFs are shown in Table S2.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100065.

References

[1] Geppert H, Vogt M, Bajorath J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 2010;50:205–16.
[2] Papa E, Dearden J, Gramatica P. Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. Chemosphere 2007;67:351–8.
[3] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 1996;96:1027–44.
[4] Viswanadhan VN, Mueller GA, Basak SC, Weinstein JN. Comparison of a neural net-based QSAR algorithm (PCANN) with hologram-and multiple linear regression-based QSAR approaches: application to 1, 4-dihydropyridine-based calcium channel antagonists. J Chem Inf Comput Sci 2001;41:505–11.
[5] Eriksson L, Andersson PL, Johansson E, Tysklind M. Megavariate analysis of environmental QSAR data. Part I–A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Mol Divers 2006;10:169–86.
[6] Hemmateenejad B, Miri R, Jafarpour M, Tabarzad M, Foroumadi A. Multiple linear regression and principal component analysis-based prediction of the anti-tuberculosis activity of some 2-aryl-1, 3, 4-thiadiazole derivatives. QSAR Comb Sci 2006;25:56–66.
[7] Loader R, Singh N, O'malley P, Popelier P. The cytotoxicity of ortho alkyl substituted 4-X-phenols: a QSAR based on theoretical bond lengths and electron densities. Bioorg Med Chem Lett 2006;16:1249–54.
[8] Lindström A, Pettersson F, Almqvist F, Berglund A, Kihlberg J, Linusson A. Hierarchical PLS modeling for predicting the binding of a comprehensive set of structurally diverse protein– ligand complexes. J Chem Inf Model 2006;46:1154–67.
[9] Cole DJ, Tirado-Rives J, Jorgensen WL. Molecular dynamics and Monte Carlo simulations for protein–ligand binding and inhibitor design. Biochim Biophys Acta 2015;1850:966–71.
[10] Mitchell MJ, McCammon JA. Free energy difference calculations by thermodynamic integration: difficulties in obtaining a precise value. J Comput Chem 1991;12:271–5.
[11] Kästner J. Umbrella sampling. Wires Comput Mol Sci 2011;1:932–42.
[12] Abel R, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing drug discovery through enhanced free energy calculations. Acc Chem Res 2017;50:1625–1632.
[13] Roy K, Das NR. A review on principles, theory and practices of 2D-QSAR. Curr Drug Metab 2014;15:346–79.
[14] Kubinyi H. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. Quant Struct-Activity Relation 1988;7:121–33.
[15] Srikanth, K.; Kumar, C.; Goswami, D.; De, A.; Jha, T., Quantitative structure activity relationship (QSAR) studies of some substituted benzenesulphonyl glutamines as tumour suppressors. 2001.
[16] Chakravarti SK, Saiakhov RD, Klopman G. Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. J Chem Inf Model 2012;52:2609–18.
[17] Ajmani S, Jadhav K, Kulkarni SA. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. J Chem Inf Model 2006;46:24–31.
[18] Mittal RR, Harris L, McKinnon RA, Sorich MJ. Partial charge calculation method affects CoMFA QSAR prediction accuracy. J Chem Inf Model 2009;49:704–709.
[19] Silverman B, Platt DE. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem 1996;39:2129–40.
[20] Turner DB, Willett P, Ferguson AM, Heritage TW. Development and validation of the eva descriptor for QSAR studies. ACS Publications; 1997.
[21] Todeschini R, Vighi M, Provenzani R, Finizio A, Gramatica P. Modeling and prediction by using WHIM descriptors in QSAR studies: toxicity of heterogeneous chemicals on Daphnia magna. Chemosphere 1996;32:1527–45.
[22] Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, Sun S, Yang J, Ramsundar B, Pande V. PotentialNet for molecular property prediction. ACS Cent Sci 2018;4:1520–30.
[23] Papa, E.; Dearden, J.; Gramatica, P.J.C., Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. 2007, 67, 351–8.
[24] Li BYS, Yeung LF, Ko KT. Indefinite kernel ridge regression and its application on QSAR modelling. Neurocomputing 2015;158:127–33.
[25] Mei H, Zhou Y, Liang G, Li Z. Support vector machine applied in QSAR modelling. Chin Sci Bull 2005;50:2291–6.
[26] Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE. Application of random forest approach to QSAR prediction of aquatic toxicity. J Chem Inf Model 2009;49:2481–8.
[27] Zilian D, Sotriffer CA. Sfcscore rf: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. J Chem Inf Model 2013;53:1923–33.
[28] Hassan-Harrirou H, Zhang C, Lemmin T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. J Chem Inf Model 2020;60:2791–802.
[29] Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R., Gated graph sequence neural networks. *arXiv preprint* 2015.
[30] Liu K, Sun X, Jia L, Ma J, Xing H, Wu J, Gao H, Sun Y, Boulnois F, Fan J. Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. Int J Mol Sci 2019;20:3389.
[31] Gomes, J.; Ramsundar, B.; Feinberg, E.N.; Pande, V.S., Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint* 2017.
[32] Ragno R. www.3d-qsar.com: a web portal that brings 3-D QSAR to all electronic devices—the Py-CoMFA web application as tool to build models from pre-aligned datasets. J Comput Aided Mol Des 2019;33:855–64.
[33] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110:5959–67.