# Deep learning on the fusion of chemical sequences and molecular grids for ligand-based virtual screening

*Rongji* Ke[1,*] and *Debby D.* Wang[1,**]

[1]School of Science and Technology, Hong Kong Metropolitan University, 81 Chung Hau St, Ho Man Tin, Hong Kong

**Abstract.** Computational techniques have been widely applied in modern drug discovery to reduce cost and time. As a crucial component of computational drug discovery, predicting compound-protein interactions is becoming increasingly prevalent. Virtual screening serves as a cost-effective tool for predicting such interactions. However, current models often require input from both compounds and proteins, which is unfriendly for scenarios where the protein information is not valid. In this work, we propose a ligand-based prediction approach that leverages only the SMILES sequences and 3D grids of compounds in target-specific tasks. By learning these features through a cross-attention mechanism, our model can capture high-level structural features of the compounds in the prediction tasks. Experimental validation on the DUD-E dataset demonstrates that our model achieves competitive performance in both accuracy and efficiency. Particularly, it performs decently when large proteins are involved.

## 1 INTRODUCTION

Developing a new drug often requires a considerable amount of time, money, and effort [1]. Before clinical trials, processes such as hit identification and lead optimization play an indispensable role in early stage drug discovery [2]. These processes aim to identify compounds that show preferred pharmacological activity against a disease-associated protein, from a large library (e.g. in-house or synthesized). High-throughput screening (HTS) [3] is a well-known technique for experimentally screening the compounds in a large library, but with high costs and much time. Virtually screening large compound libraries using computational approaches has therefore become an efficient alternative to HTS in recent decades [4].

The core of a virtual screening (VS) task is to predict the interactions between a compound and its target protein. Such interactions can be either binary values (e.g. binder or non-binder for a target) or continuous values (e.g. binding constant regarding a target). In earlier VS tasks, traditional machine learning models were extensively applied, often with the molecules represented as descriptors or fingerprints. For example, Rawan *et al.* [5] employed a random forest model to extract various graph-based descriptors from heterogeneous graphs for prediction. Aghakhani *et al.* [6] proposed an approach, which integrates the k-means clustering algorithm with descriptors from social network analysis techniques, to predict drug-target interactions. With many successful applications in different areas, deep learning has

*e-mail: rke@hkmu.edu.hk
**e-mail: dwang@hkmu.edu.hk

entered the VS field and has garnered increasing attention nowadays. Mostly, convolutional neural networks (CNNs) and graph neural networks (GNNs) are adopted in these prediction works. The *CPI* model [7] integrates a GNN and a CNN to learn compound-protein interactions from compound graphs and protein sequences. *TransformerCPI* [8] involves a transformer neutral network to learn compound-protein interactions from protein sequences and compound chemical sequences.

Although these predictions have shown good potential for deep learning techniques in the VS field, they mostly need input from both proteins (e.g. amino acid sequences or crystal structures) and compounds (e.g. chemical strings or 2D graphs) as shown in Figure 1A. Proteins are large molecular systems, and therefore involving them often increases the com-

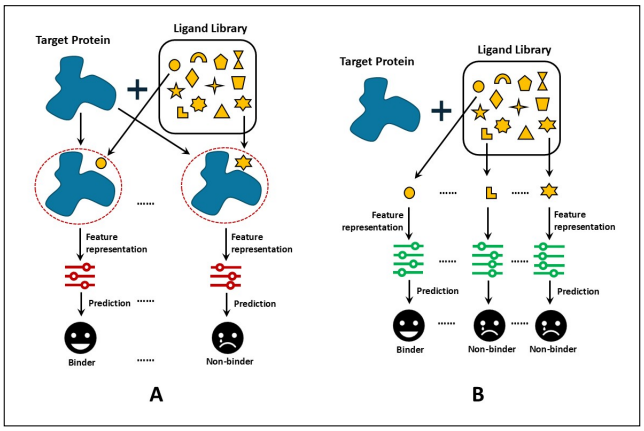


**Fig. 1.** The target-specific drug-protein interaction prediction.

plexity of feature representation and lowers the VS efficiency. In this regard, handling VS tasks in a target-specific manner and based only on compound information is more efficient (Figure 1B). In this study we propose a ligand-based, deep learning model that only relies on compound information and predicts the compound-protein interactions properly. The highlights of this work are listed below.

- In a target-specific context, we represents a compound by both its SMILES string and a 3D grid. It captures the topological and spatial information of a compound.
- The features extracted from the SMILES strings and those from the 3D grids are fused through a cross-attention mechanism. This conserves the important, high-level structural features of each compound.
- Experimental validation on the DUD-E database confirms the prediction accuracy and computational efficiency of our model in the prediction of compound-protein interaction. Particularly, it performs decently in cases involving large target proteins.

## 2 MATERIALS AND METHODS

The network architecture of our model is shown in Figure 2. First, a hybrid molecular representation that describes both the chemical sequence (SMILES string) and the spatial information (3D grids) of compounds is constructed. This presentation is only based on the information of compounds, without any input from the target proteins. In the next phase, a

fusion of these two types of features is accomplished by a cross-attention mechanism. Finally, the fused features are fed into the prediction head to determine if a molecule is a binder or non-binder to a specific target protein (target-specific tasks).

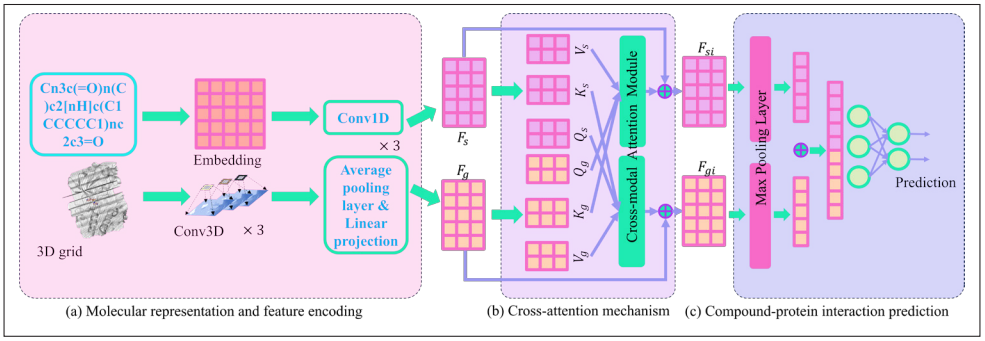


**Fig. 2.** The network architecture of our DL-FSG model.

## 2.1 Molecular representation and feature encoding

When representing a small-molecule compound, we considered both the *sequential* and *spatial* information.

*Sequential information.* A SMILES (Simplified Molecular Input Line Entry System) string is a structured, chemical sequence that describes a molecule using plain text. Encoding such a chemical string can promisingly capture the sequential and topological information of a molecule. It often involves a tokenization process, with tokens referring to specific substrings that cover atoms, bonds, branches, rings and other information. Below shows some commonly used tokens in this context.

$$\begin{aligned} ''Cl'' &- \text{Chlorine atom} \\ '' ='' &- \text{Double bond} \\ ''()'' &- \text{Branch sign} \\ ''C1CCCCC1'' &- \text{Cyclohexane} \end{aligned} \tag{1}$$

To enhance the efficiency of pattern matching, we mainly focused on atoms, bonds, and branches. The tokens $\{t_n\}_{n=1}^{N}$ in the SMILES string of a given molecule can be mapped into an embedding space as follows.

$$\mathbf{e}_n^s = f_{SMILES}(t_n) \tag{2}$$

Stacking the embeddings of all the tokens leads to a matrix $\mathbf{E}_s \in \mathbb{R}^{N \times d_s}$, where $N$ is the number of tokens and $d_s$ is the size of embedding for those tokens. $\mathbf{E}_s$ is further processed by three consecutive 1D-CNN layers, which can efficiently extract sequence semantic information [9].

$$\mathbf{F}_s = f_{CNN}^3(f_{CNN}^2(f_{CNN}^1(\mathbf{E}_s))) \tag{3}$$

$\mathbf{F}_s \in \mathbb{R}^{N \times f}$ indicates the latent features, where $f$ is the number of channels of the last 1D-CNN layer.

*Spatial information.* The encoded SMILES string of a molecule can convey important sequential and topological information, but it can hardly describe the spatial details of the

molecule. Capturing the spatial information is quite necessary because the molecular binding process mostly concerns the 3D structures of molecules. In this regard, representing a molecule as a 3D grid is a rational way. Given a molecule, it can be confined to a box with its original center, and the box can be further partitioned into small cells by a specific interval. Suppose the box has a size of $L\mathring{A} \times W\mathring{A} \times H\mathring{A}$ and the interval is $1\mathring{A}$, then there are $L \times W \times H$ cells for this molecule. Each atom $\{\mathbf{a}_k\}_{k=1}^{K}$ in this molecule is assigned to a cell $\{l, w, h\}$ according to the nearest distance strategy, and this cell can be further characterized by the atomic properties (Equation 4).

$$\mathbf{TS}_g(l, w, h) = (p_1^k, p_2^k, \ldots, p_{d_g}^k) \tag{4}$$

Here $(l, w, h)$ indicates the center position of a cell and $p_i^k$ is the $i$-th property of the $k$-th atom. If multiple atoms are assigned to the same cell, then the atomic properties will be fused as follows.

$$\mathbf{TS}_g(l, w, h) = (\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_{d_g}) \tag{5}$$

Here $\bar{p}_i$ is the average property value for a group of atoms assigned to cell $(l, w, h)$. It leads to the generation of a tensor $\mathbf{TS}_g$ that stores the spatial and structural information of the input molecule. This tensor is then processed by three consecutive 3D-CNN layers, which effectively extract the physical features from 3D molecular grids [10].

$$\mathbf{E}_g = f_{CNN3}^3(f_{CNN3}^2(f_{CNN3}^1(\mathbf{TS}_g))) \tag{6}$$

A subsequent average pooling layer and a linear projection layer yield the latent features $\mathbf{F}_g \in \mathbb{R}^{N \times f}$ as below.

$$\mathbf{F}_g = f_{linear}(f_{avgpool}(\mathbf{E}_g)) \tag{7}$$

Consequently, each molecule can be represented by $(\mathbf{F}_s, \mathbf{F}_g)$, which conveys both sequential and spatial messages.

### 2.2 Cross-attention mechanism

A cross-attention mechanism [11] is used to fuse $\mathbf{F}_s$ and $\mathbf{F}_g$. First, each type of features $\mathbf{F}_j$ $(j \in \{s, g\})$ are transformed into query, key and value vectors ($\mathbf{Q}_j$, $\mathbf{K}_j$ and $\mathbf{V}_j$) as below.

$$\begin{cases} \mathbf{Q}_j = f_{CNN}^{j,1}(\mathbf{F}_j) \\ \mathbf{K}_j = f_{CNN}^{j,2}(\mathbf{F}_j) \\ \mathbf{V}_j = f_{CNN}^{j,3}(\mathbf{F}_j) \end{cases} \tag{8}$$

Then, the fused features $\mathbf{A}_s \in \mathbb{R}^{N \times f}$ for the SMILES embedding ($\mathbf{V}_s$) can be derived by referring to $\mathbf{Q}_g$ and $\mathbf{K}_g$. Similarly, the fused features $\mathbf{A}_g \in \mathbb{R}^{N \times f}$ for the 3D grid embedding ($\mathbf{V}_g$) can be derived by referring to the query and key vectors of the SMILES embedding ($\mathbf{Q}_s$ and $\mathbf{K}_s$). The formulas are as below.

$$\mathbf{A}_s\left(\mathbf{Q}_g, \mathbf{K}_g, \mathbf{V}_s\right) = Softmax\left(\frac{\mathbf{Q}_g \mathbf{K}_g{}^T}{\sqrt{f}}\right)\mathbf{V}_s \tag{9}$$

$$\mathbf{A}_g\left(\mathbf{Q}_s, \mathbf{K}_s, \mathbf{V}_g\right) = Softmax\left(\frac{\mathbf{Q}_s \mathbf{K}_s{}^T}{\sqrt{f}}\right)\mathbf{V}_g \tag{10}$$

Finally, the features $\mathbf{F}_j$ $(j \in \{s, g\})$ are updated to $\mathbf{F}'_j$ by a simple average weighting method as follows.

$$\mathbf{F}'_j = 0.5 \cdot \mathbf{F}_j + 0.5 \cdot \mathbf{A}_j \tag{11}$$

### 2.3 Compound-protein interaction prediction

The prediction module comprises a max pooling layer, a concatenation layer and a multi-layer fully connected neural network (FCNN).

$$\hat{y} = f_{FCNN}(f_{maxpool}(\mathbf{F}'_s) \| f_{maxpool}(\mathbf{F}'_g)) \tag{12}$$

Here $\|$ indicates concatenation, and the FCNN employs the Leaky ReLU activation function with a negative slope of 0.01 [12]. To mitigate overfitting, each FCNN layer is followed by a Dropout layer. The module outputs $\hat{y}$ represent the probability of interaction likelihood. For this binary classification task, the binary cross-entropy loss is used as the loss function. This prediction module is displayed in Figure 2(c).

$$L = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right] \tag{13}$$

where $y$ is the ground truth label.

## 3 EXPERIMENT AND RESULT

### 3.1 Dataset

The DUD-E database, an enhanced and rebuilt version of The Directory of Useful Decoys (DUD) [13], has been widely used as the benchmark data for VS tasks and molecular docking programs. Decoys, which serve as *negative samples*, are chosen to possess similar physicochemical properties but dissimilar 2D topologies compared to the corresponding active molecules. DUD-E contains 22,886 active compounds against 102 diverse targets, with an average of 50 decoys per active compound. It corresponds to 102 target-specific VS tasks in our study.

### 3.2 Experimental setup and evaluation metrics

The experiment was implemented using PyTorch. The SMILES strings were tokenized, and each token was encoded as a 64-dimensional dense vector. The CNN block for processing SMILES embedding has output channels of 40, 80, and 160, with window sizes of 4, 6, and 8, respectively. The CNN block for processing 3D grids has output channels of 64, 64, and 128, with a uniform window size of 3. The average pooling layer has a window size of 3.

The prediction module consists of four fully connected layers, with 1,024, 1,024, 512, and 2 neurons, respectively. A dropout rate of 0.1 was applied, and early stopping (no loss improvement within five epochs) was used to prevent overfitting. The model was trained with a default learning rate of $1 \times 10^{-4}$ and a weight decay coefficient of $1 \times 10^{-4}$. The Adam optimizer was used for parameter optimization, and CyclicLR was employed for learning rate adjustment. The batch size was set to 8.

To evaluate model performance, we employed accuracy, precision, recall, the area under the receiver operating characteristic curve (AUC), and the area under the precision-recall curve (AUPR) as key metrics. For each target-specific task, the dataset was partitioned into training, validation, and testing sets by a ratio of 8:1:1. To better reflect real-world conditions, we ensure that the number of negative samples exceeds the number of positive samples. Accordingly, we set the decoy-to-active ratio to 9:1 in the training, validation, and testing sets. A 5-fold cross-validation was adopted for the final model evaluation.

### 3.3 Prediction performance and comparison to other models

We used a series of state-of-the-art VS models as our performance baselines. These comparison models include *CPI* (2018) [7], *BACPI* (2018) [14], *TransformerCPI* (2020) [8], *DrugVQA* (2020) [15], *MCANet* (2023) [11], DrugLAMP (2024) [16], and MMDG-DTI (2025) [17]. As our model applies Deep Learning on the Fusion of chemical Sequences and molecular Grids for predicting compound-protein interactions, we abbreviate this model as *DL-FSG*. The feature representation methods for these involved models are listed in Table 1, and the model performance is presented in Table 2.

**Table 1.** Feature representation methods for the involved models.

| Model | Protein | Compound |
|---|---|---|
| **DL-FSG (Ours)** | **-** | **SMILES + 3D grid** |
| CPI (18') | Sequence | Graph |
| BACPI (18') | Sequence | Graph |
| TransformerCPI (20') | Sequence | Graph |
| drugVQA (20') | Distance Map | SMILES |
| MCANet (23') | Sequence | SMILES |
| DrugLAMP (24') | Sequence | Graph |
| MMDG-DTI (25') | Sequence | Graph |

**Table 2.** Performance comparison of our method with the baseline methods.

| Model | AUC | Accuracy | AUPR | Recall | Precision |
|---|---|---|---|---|---|
| **DL-FSG (Ours)** | **0.9910** | **0.9894** | **0.9725** | **0.9285** | **0.9559** |
| CPI (18') | 0.9708 | 0.9639 | 0.8477 | 0.8325 | 0.8464 |
| BACPI (18') | 0.9789 | 0.9656 | 0.8398 | 0.6659 | 0.8900 |
| TransformerCPI (20') | 0.9219 | 0.9382 | 0.6965 | 0.5366 | 0.6794 |
| drugVQA (20') | 0.9718 | 0.9873 | 0.8665 | 0.7598 | 0.9207 |
| MCANet (23') | 0.9887 | 0.9835 | 0.9637 | 0.8812 | 0.9501 |
| DrugLAMP (24') | 0.9616 | 0.9690 | 0.9311 | 0.8811 | 0.9252 |
| MMDG-DTI (25') | 0.9663 | 0.9783 | 0.9413 | 0.8960 | 0.9346 |

According to the two tables, our model was built in a target-specific context and achieved state-of-the-art performance on average. Since a real-world VS task usually faces an imbalance problem, AUC and AUPR serve as more comprehensive metrics to evaluate model performance. These metrics are illustrated in Fig. 3, demonstrating that our model outperforms others with respect to many protein targets. It verifies the effectiveness of our ligand-based deep learning techniques in VS tasks. Incorporating 3D grid features with general sequence features enables the model to capture high-level structural features of each compound, leading to a more comprehensive understanding of its characteristics.

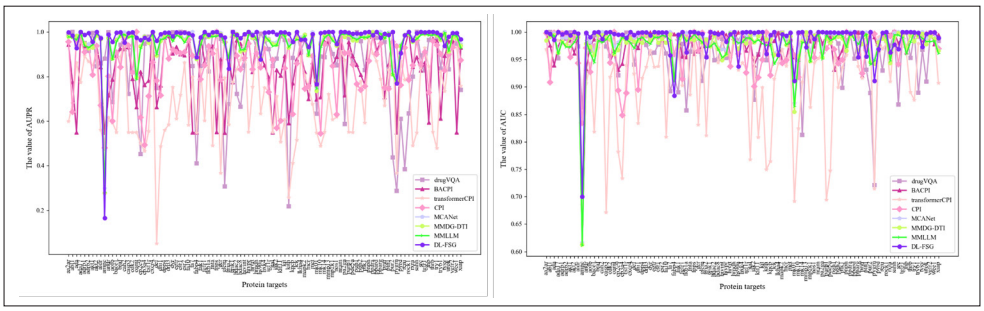


**Fig. 3.** Model performance evaluation with respect to AUC and AUPR on all the 102 target proteins in DUD-E. Each protein indicates a target-specific VS task.

**Table 3.** Comparison of the computational efficiency of our method and the baseline methods.

| Model | Times(s) | FLOPs(G) | Params (MB) |
|---|---|---|---|
| **DL-FSG** | **0.1375** | **1.1929** | **113.4753** |
| CPI (18') | 0.2751 | 0.0078 | 0.0033 |
| BACPI (18') | 0.4501 | 0.7563 | 0.1952 |
| TransformerCPI (20') | 0.3587 | 0.0811 | 0.3486 |
| drugVQA (20') | 0.3537 | 15.2092 | 0.4275 |
| MCANet (23') | 0.3382 | 1.5943 | 2.2259 |
| DrugLAMP (24') | 0.1486 | 683.39 | 166.68 |
| MMDG-DTI (25') | 0.6581 | 502.4307 | 149.95 |

Meanwhile, we evaluated the runtime, FLOPs, and parameters of each model (Table 3). Although our model has a higher parameter amount, its FLOPs remain at a moderate level compared to the baselines. Since computation time directly reflects efficiency in practical applications, we evaluated the model's computational efficiency by measuring the processing time for a batch of 8 samples. The results indicate that our model achieves the shortest runtime, demonstrating good computational efficiency.

## 3.4 Ablation

### *3.4.1 Attention*

In this section, we examine the impact of different attention mechanisms on model performance. As shown in Table 4, we compare three variants, namely (a) 'Cross', representing the method proposed in this paper; (b) 'Self', which replaces cross-attention with a self-attention mechanism; and (c) 'None', where cross-attention is replaced by a three-layer convolutional module. The results show that the 'Cross' attention mechanism outperforms the others across all evaluation metrics. This improvement stems from its ability to enable bidirectional information exchange. It allows the model to achieve a more comprehensive understanding of the characteristics of drug molecules.

**Table 4.** Performance evaluation on different attention mechanisms.

| Method | AUC | Accuracy | AUPR | Recall | Precision |
|---|---|---|---|---|---|
| **Cross (Ours)** | **0.9910** | **0.9894** | **0.9725** | **0.9285** | **0.9559** |
| Self | 0.9855 | 0.9817 | 0.9665 | 0.8984 | 0.9587 |
| None | 0.9798 | 0.9618 | 0.8448 | 0.8325 | 0.8480 |

### *3.4.2 Feature representation*

In this section, we investigate the impact of different feature representation methods on model performance. Specifically, we compare the proposed method, which integrates both SMILES and 3D grid representations (denoted as 'Both'), with strategies that use SMILES only ('SMILES') or 3D grid only ('Grid'). In the 'SMILES' and 'Grid' settings, we replace the original cross-attention mechanism with self-attentions. As shown in Table 5, the proposed method outperforms the others across all evaluation metrics. This improvement is attributed to the combination of SMILES and 3D grid information, which enables a more comprehensive understanding of the compounds.

### *3.4.3 Target-involved tasks*

In this section, we investigate the impact of the participation of protein information on model performance in target-specific tasks. As shown in Table 6, we compare two mod-

**Table 5.** Performance evaluation on different feature representation methods.

| Method | AUC | Accuracy | AUPR | Recall | Precision |
|---|---|---|---|---|---|
| **Both (Ours)** | **0.9910** | **0.9894** | **0.9725** | **0.9285** | **0.9559** |
| SMILES | 0.9683 | 0.9535 | 0.8323 | 0.8143 | 0.8338 |
| Grid | 0.9646 | 0.9436 | 0.8266 | 0.8046 | 0.8258 |

els, namely (a) the proposed method ('GS') and (b) a model variant involving protein information ('PGS'). The 'PGS' model introduces a protein-processing branch, analogous to the SMILES branch, consisting of a tokenization layer, embedding layers, and three successive 1D-CNN layers. The resulting protein features are then concatenated with the cross-attention features derived from the compound representations, and passed through a prediction head. It demonstrates that our proposed 'GS' model outperforms 'PGS'.

**Table 6.** Performance evaluation on different input strategies (involving or removing protein information).

| Method | AUC | Accuracy | AUPR | Recall | Precision |
|---|---|---|---|---|---|
| **GS (Ours)** | **0.9910** | **0.9894** | **0.9725** | **0.9285** | **0.9559** |
| PGS | 0.9783 | 0.9678 | 0.8445 | 0.8382 | 0.8458 |

### 3.5 Case studies

Given that earlier deep learning models often struggle with long protein sequences in VS tasks, we investigated the performance of our model in tasks involving proteins of different lengths, including (a) Epidermal Growth Factor Receptor erbB1 (*egfr*) with a sequence length of 315, (b) Cytochrome P450 2C9 (*cp2c9*) with a sequence length of 477, and (c) Monoamine Oxidase B (*aofb*) with a sequence length of 520 (Figure 4).

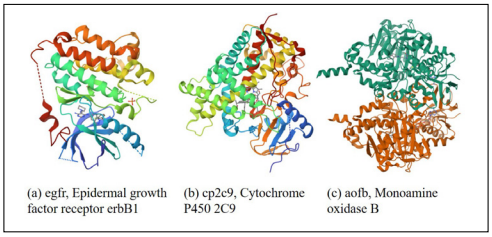


**Fig. 4.** Structures of several representative proteins.[1]

These proteins are important targets in various disease-related studies. For example, *egfr* is crucial in regulating the progression of non-small-cell lung cancer, and active binders to this protein may serve as promising candidates for cancer therapies [18–20]. Table 7 summarizes the performance of our model and the baselines in these target-specific tasks. Compared to baselines, our approach demonstrates significant improvements in recall and precision, particularly in the cases involving large proteins (*aofb*). Our ligand-based deep learning approach does not require input from the proteins, reducing the difficulties of parsing proteins and simplifying the training of models.

[1]The protein images are obtained from RCSB PDB: (a) egfr: https://doi.org/10.2210/pdb2RGP/pdb, (b) cp2c9: https://doi.org/10.2210/pdb1R9O/pdb, (c) aofb: https://doi.org/10.2210/pdb1S3B/pdb.

**Table 7.** Model comparisons with respect to the performance on targets *egfr*, *cp2c9*, and *aofb*.

| Target | Metric | DL-FSG (Ours) | CPI | BACPI | TransformerCPI | drugVQA | MCANet | DrugLAMP | MMDG-DTI |
|---|---|---|---|---|---|---|---|---|---|
| *egfr* | AUC | **0.9998** | 0.9919 | 0.9949 | 0.9361 | 0.9979 | 0.9989 | 0.9816 | 0.9886 |
| | Accuracy | **0.9917** | 0.9806 | 0.9852 | 0.9244 | 0.9889 | 0.9863 | 0.9761 | 0.9800 |
| | AUPR | **0.9986** | 0.9088 | 0.9333 | 0.6096 | 0.9941 | 0.9892 | 0.9798 | 0.9847 |
| | Recall | **0.9159** | 0.8333 | 0.8519 | 0.2870 | 0.8920 | 0.9026 | 0.9518 | 0.9496 |
| | Precision | **1.0000** | 0.9677 | 1.0000 | 0.8611 | 1.0000 | 1.0000 | 0.9786 | 0.9821 |
| *cp2c9* | AUC | **0.9954** | 0.8935 | 0.9311 | 0.7814 | 0.9210 | 0.9812 | 0.9640 | 0.9695 |
| | Accuracy | **0.9833** | 0.9083 | 0.9625 | 0.9125 | 0.9780 | 0.9750 | 0.9576 | 0.9566 |
| | AUPR | **0.9644** | 0.6167 | 0.8206 | 0.4994 | 0.4514 | 0.9382 | 0.9264 | 0.9188 |
| | Recall | **0.9167** | 0.6667 | 0.6667 | 0.2917 | 0.1000 | 0.9167 | 0.8988 | 0.9019 |
| | Precision | **0.9467** | 0.5333 | 0.9112 | 0.6364 | 0.9345 | 0.8462 | 0.8303 | 0.8292 |
| *aofb* | AUC | **0.9909** | 0.9275 | 0.9762 | 0.9426 | 0.9680 | 0.9821 | 0.9626 | 0.9717 |
| | Accuracy | **0.9833** | 0.8566 | 0.9547 | 0.8975 | 0.9710 | 0.9500 | 0.9336 | 0.9350 |
| | AUPR | **0.9549** | 0.6000 | 0.7861 | 0.5922 | 0.6853 | 0.8960 | 0.8781 | 0.8876 |
| | Recall | **0.9167** | 0.7917 | 0.7917 | 0.6667 | 0.6500 | 0.7917 | 0.7706 | 0.7823 |
| | Precision | **0.9167** | 0.3878 | 0.7600 | 0.4848 | 0.7220 | 0.7308 | 0.7164 | 0.7234 |

[1] *egfr*: Epidermal Growth Factor Receptor; *cp2c9*: Cytochrome P450 2C9; *aofb*: Monoamine Oxidase B.

## 4 CONCLUSIONS

Computational drug discovery plays a critical role in reducing costs and accelerating timelines in new drug development. Predicting the interaction between a small-molecule compound and its target protein is a fundamental problem in this field, making the virtual screening of binders to a specific protein a key task. While recent deep learning models often extract features from both compounds and proteins, it is limited to scenarios where protein information is missing or the protein has a long sequence. To overcome these challenges, we propose a ligand-based prediction method that integrates SMILES sequences and 3D grids of compounds. Our approach captures essential features at both one-dimensional and three-dimensional levels. Experimental validation on the DUD-E dataset demonstrates the competitive performance of our model, providing an effective framework for predicting the compound-protein interaction.

## ACKNOWLEDGMENT

This work is supported by Hong Kong Metropolitan University (Project PFDS/2024/01) and Hong Kong Research Grants Council (Project UGC/FDS16/E16/23).

## References

[1] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, Future Medicinal Chemistry **12**, 939 (2020).

[2] R.A. Goodnow Jr, Hit and lead identification: Integrated technology-based approaches, Drug Discovery Today: Technologies **3**, 367 (2006).

[3] L.M. Mayr, D. Bojanic, Novel trends in high-throughput screening, Current Opinion In Pharmacology **9**, 580 (2009).

[4] J. Bajorath, Integration of virtual and high-throughput screening, Nature Reviews Drug Discovery **1**, 882 (2002).

[5] R.S. Olayan, H. Ashoor, V.B. Bajic, DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches, Bioinformatics **34**, 1164 (2018).

[6] S. Aghakhani, A. Qabaja, R. Alhajj, Integration of k-means clustering algorithm with network analysis for drug-target interactions network prediction, International Journal of Data Mining and Bioinformatics **20**, 185 (2018).

[7] M. Tsubaki, K. Tomii, J. Sese, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics (2018).

[8] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, M. Zheng, Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, Bioinformatics **36**, 4406 (2020).

[9] Y. Kim, Convolutional Neural Networks for Sentence Classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, edited by A. Moschitti, B. Pang, W. Daelemans (2014), pp. 1746–1751

[10] J. Jiménez, M. Skalic, G. Martinez-Rosell, G. De Fabritiis, K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks, Journal of Chemical Information and Modeling **58**, 287 (2018).

[11] J. Bian, X. Zhang, X. Zhang, D. Xu, G. Wang, Mcanet: shared-weight-based multihead-crossattention network for drug–target interaction prediction, Briefings in Bioinformatics **24**, bbad082 (2023).

[12] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1026–1034

[13] M.M. Mysinger, M. Carchia, J.J. Irwin, B.K. Shoichet, Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking, Journal of Medicinal Chemistry **55**, 6582 (2012).

[14] M. Li, Z. Lu, Y. Wu, Y. Li, Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction, Bioinformatics **38**, 1995 (2022).

[15] S. Zheng, Y. Li, S. Chen, J. Xu, Y. Yang, Predicting drug–protein interaction using quasi-visual question answering system, Nature Machine Intelligence **2**, 134 (2020).

[16] Z. Luo, W. Wu, Q. Sun, J. Wang, Accurate and transferable drug–target interaction prediction with druglamp, Bioinformatics **40**, btae693 (2024).

[17] Y. Hua, Z. Feng, X. Song, X. Wu, J. Kittler, MMDG-DTI: drug-target interaction prediction via multimodal feature fusion and domain generalization, Pattern Recognition **157**, 110887 (2025).

[18] L. Ma, D.D. Wang, B. Zou, H. Yan, An eigen-binding site based method for the analysis of anti-egfr drug resistance in lung cancer treatment, IEEE/ACM transactions on computational biology and bioinformatics **14**, 1187 (2016).

[19] M. Zhu, D.D. Wang, H. Yan, Genotype-determined egfr-rtk heterodimerization and its effects on drug resistance in lung cancer treatment revealed by molecular dynamics simulations, BMC molecular and cell biology **22**, 34 (2021).

[20] D.D. Wang, Y. Huang, Scoring protein-ligand binding structures through learning atomic graphs with inter-molecular adjacency, PLOS Computational Biology **21**, e1013074 (2025).