ARTICLES
https://doi.org/10.1038/s42256-020-0152-y

nature machine intelligence



# Predicting drug–protein interaction using quasi-visual question answering system

Shuangjia Zheng[1,2,3], Yongjian Li[2,3], Sheng Chen[2], Jun Xu[1]✉ and Yuedong Yang[2]✉

**Identifying novel drug–protein interactions is crucial for drug discovery. For this purpose, many machine learning-based methods have been developed based on drug descriptors and one-dimensional protein sequences. However, protein sequences cannot accurately reflect the interactions in three-dimensional space. However, direct input of three-dimensional structure is of low efficiency due to the sparse three-dimensional matrix, and is also prevented by the limited number of co-crystal structures available for training. Here we propose an end-to-end deep learning framework to predict the interactions by representing proteins with a two-dimensional distance map from monomer structures (Image) and drugs with molecular linear notation (String), following the visual question answering mode. For efficient training of the system, we introduce a dynamic attentive convolutional neural network to learn fixed-size representations from the variable-length distance maps and a self-attentional sequential model to automatically extract semantic features from the linear notations. Extensive experiments demonstrate that our model obtains competitive performance against state-of-the-art baselines on the directory of useful decoys, enhanced (DUD-E), human and BindingDB benchmark datasets. Further attention visualization provides biological interpretation to depict highlighted regions of both protein and drug molecules.**

Prediction of drug–protein interactions (DPIs) is of crucial importance for drug design and development. Although experimental assays remain the most reliable approach for determining DPIs, experimental characterization of every possible drug–protein pair is daunting due to the vast costs and labour involved in experiments.

Computational prediction of DPIs has therefore made rapid progress recently. In general, it falls roughly into two categories: physics-based and machine learning methods. Physics-based methods, such as molecular docking, apply physics-inspired predetermined energy functions to assess drug–protein interactions at the atomic level[1,2]. However, these methods are usually of limited accuracy due to difficulties in evaluating the conformational entropy and solvent contributions. Furthermore, these atom-level methods are sensitive to structural fluctuations and can't process protein flexibility well.

With the recent increase in protein structural data and protein–ligand interaction datasets, there has been a rapid progress in machine learning-based methods [3–5]. Usually, the prediction is treated as a task of binary classification by integrating information on ligands, proteins and their interactions in a unified framework.

Drug molecules can be well represented by their linear notations as most drugs contain less than 100 heavy atoms, and thus have a relatively small structural space. Recent studies have proven that current deep learning techniques can accurately predict structural properties from their linear representation[6–8]. In contrast, protein molecules are much bigger, typically containing more than 1,000 heavy atoms. Prediction of the process from a one-dimensional (1D) sequence to a 3D structure, called protein folding, is a well-known challenging problem. Therefore, traditional representation by a 1D protein sequence is insufficient to capture the structural features in 3D space that decide the prediction of DPIs. Although the direct input of 3D structure has been attempted in recent studies[3,9,10], they achieved relatively low accuracy for a few reasons. First, the irregular protein 3D structure needs a large-scale 3D matrix to contain the whole structure. The high-dimension, sparse matrix caused a vast number of tedious input variables. Second, these studies suffered from a small number of high-quality 3D structures because they require co-crystal structures of protein–ligand pairs, which are difficult to determine by experiments.

As a balance, proteins can be alternatively represented by a 2D pairwise distance map. A distance map is a compact representation of the 3D structure of a protein via the pairwise contacts between the amino acids constituting the protein. Previous studies have indicated that distance maps can be used for generating and comparing protein 3D structures[11–13].

Inspired by these studies, here we utilize a 2D distance map to represent proteins, and thus the DPI task can be converted into a classical visual question answering (VQA) problem[14–17]. Given an image and its corresponding question, the VQA system is required to provide a correct answer. In early studies, the VQA tasks[18] involved answering 'yes' or 'no' or a single word, and were treated as a classification task that is similar to DPI prediction. Here, images are the distance maps for proteins, questions are the simplified molecular input line entry system (SMILES)[19] for drugs, and answers are whether they will interact. This framework enables training on protein monomer structures without needing co-crystal structures with their binding ligands, which greatly expands usable datasets for training.

However, differences exist between VQA and DPI prediction. First, in many VQA scenarios, image size can be resized to a fixed value, but each pixel of the pairwise distance map represents the relation between one pair of amino acids and it results in an information loss if the map is downsampled. Second, the syntax of SMILES is different from natural language, which forces us to utilize a customized tokenization process and suitable model for obtaining the semantic feature of molecular linear notations. Third, our training set is still much smaller than other applications, which requires us to carefully design the networks.

[1]School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou, China. [2]National Supercomputer Center in Guangzhou, Sun Yat-sen University, Guangzhou, China. [3]These authors contributed equally: Shuangjia Zheng and Yongjian Li. ✉e-mail: junxu@biochemomes.com; yangyd25@mail.sysu.edu.cn

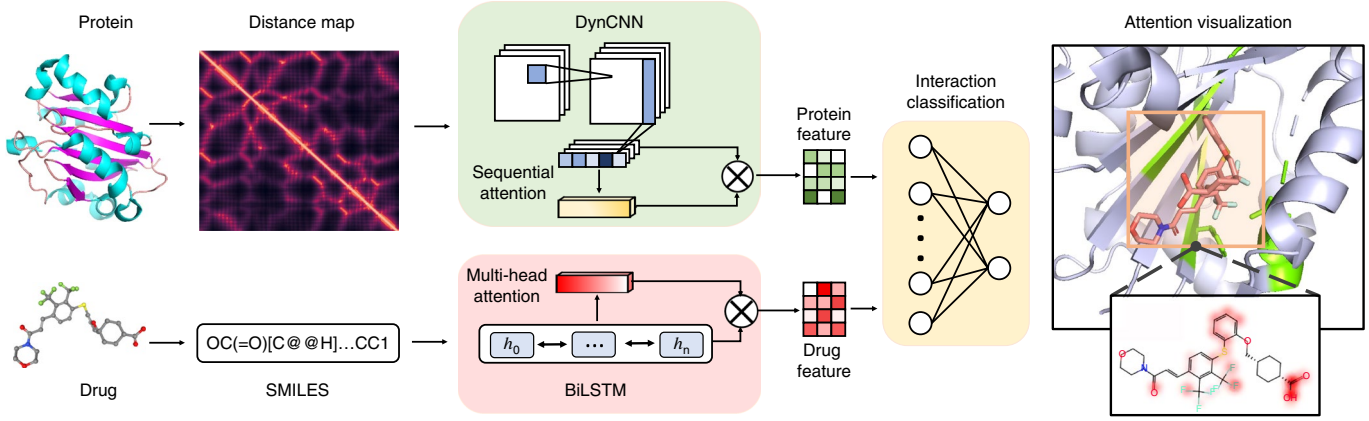


**Fig. 1 | The framework of the proposed DrugVQA model.** The model consists of two main components: a dynamic CNN with sequential attention and a BiLSTM with multi-head self-attention. The learned attention weights enable the visualization of individual contributions of each protein residue and ligand component to the classification decision.

To address the above problems, we present a VQA-inspired interpretable model that predicts DPIs directly from a protein distance map and molecular SMILES. The distance maps and SMILES are encoded by a dynamic convolutional neural network (DynCNN) and a bidirectional long short-term memory (BiLSTM)[20] with attention mechanism, respectively, and the outputs are concatenated to fully connected layers to make predictions. The proposed model is shown to outperform state-of-the-art approaches over three public DPI datasets. In addition, the learned attention enables visualization of the individual contributions between binding regions on proteins and ligands, which is important for ligand refinement.

## Related works

In this section, we introduce the related works in two areas. First, we briefly review the current models for VQA task. Second, we focus on summarizing current approaches fro DPIs prediction.

**Visual question answering.** VQA is the task of answering a related question based on a given image. Early VQA systems tended to infer the answer by directly learning a joint representation of the image and text features extracted by convolutional neural network (CNN) and recurrent neural networks (RNNs)[18,21]. Later on, attention-based models[22,23], which learn to attend the visual components that provide informative evidence to the question, showed great successes. Recently, Schwartz et al.[24] proposed multimodal attention, which focuses on not only the objects of images but also questions. Here we employ the multimodal attention framework to capture the key components of both the drug molecules and proteins.

**DPI prediction.** Docking-based methods, such as refs. [1,2], are widely used to predict the binding mode and affinity given the 3D structure inputs of a drug compound and a protein. These methods apply predefined force fields to estimate the binding score to assess DPIs at the atomic level.

Machine learning-based methods have also been investigated to predict DPIs. For example, Bleakley and Yamanishi[25] proposed the bipartite local model by training local support vector machine classifiers from chemical structure similarity and protein sequence similarity. Ballester and Mitchell[26] used a random forest algorithm to capture binding effects during molecular docking process. Durrant and McCammon[27] presented a scoring function based on fully connected neural networks to characterize the binding affinities of protein–ligand complexes. Tabei and Yamanishi[28] further improved DPI prediction by using a hashing algorithm with more compact fingerprints of compound–protein pairs.

Recently, deep learning techniques have been introduced to predict DPIs by direct use of 3D protein–compound complexes[3,9,10]. As the input features were based on a 3D matrix defined around pocket–ligand complexes, these methods generated a large number of input variables, and suffered from a limited number of training sets. In addition, though a few graph neural network-based representation learning studies predicted DPIs on the basis of molecular topological structure and protein sequence or their functional annotations[4,5,7], their accuracies were limited owing to a lack of protein structural information.

## Deep learning framework

In this section, we will introduce the framework of our VQA-based DPIs prediction model (DrugVQA).

**Problem formulation.** Our task is to predict the interaction between a drug compound and a target protein. Concretely, drug compounds are represented in SMILES format, a text string for the topological information based on chemical bonding rules. For example, the benzene ring can be encoded as 'c1ccccc1'. Each lowercase 'c' represents an aromatic carbon atom and '1' indicates the start and closing of a cycle. All hydrogen atoms aren't shown because they can be inferred via simple rules. To preserve important chemical features, we tokenized drug molecules using the following regular expression inspired by the work of ref. [29]:

$$\text{token}_{\text{regex}} = \text{//}(\backslash[\lceil\backslash\rceil]\{1,6\}\backslash])\text{//} \quad (1)$$

In addition, we replaced the multicharacter symbols using the following rules: 'Br':R, 'Cl':L, 'Si':A, 'Se':Z.

Suppose we have a drug molecular linear notation containing $n$ tokens, the molecule can be represented in a sequence of molecular embeddings as $M=(t_1, \ldots, t_n)$, where $t_i$ is a vector of $d$-dimensional token embedding for the $i$th token. Thus, $M \in \mathbb{R}^{n\times d}$ is a representation of a 2D matrix by concatenating all the token embeddings together.

Similarly, a protein can be simply described as a linear sequence that consists of a list of amino acid residues $P=(r_1, \ldots, r_l)$, where $r_i$ is a one-hot representative vector with a length of 20 for the amino acid type at position $i$, and $l$ is the sequence length. To capture structural information, we instead describe a protein as a 2D pairwise distance map calculated by the following formula:

$$\hat{s}(r_i, r_j) = \frac{1}{1 + d(r_i, r_j)/d_0}, r_i, r_j \in P \quad (2)$$

**Table 1 | Ablation results on the human, BindingDB and DUD-E datasets**

| Ablation tests | Human | BindingDB | DUD-E |
|---|---|---|---|
| LSTM+CNN | 0.940 ± 0.005 | 0.913 ± 0.003 | 0.937 ± 0.003 |
| Att-LSTM+CNN | 0.958 ± 0.004 | 0.928 ± 0.002 | 0.951 ± 0.005 |
| LSTM+Att-CNN | 0.951 ± 0.004 | 0.921 ± 0.003 | 0.946 ± 0.004 |
| VQA-seq | 0.964 ± 0.005 | 0.897 ± 0.004 | 0.948 ± 0.003 |
| DrugVQA | **0.979 ± 0.003** | **0.936 ± 0.002** | **0.972 ± 0.003** |

The first three models show the effectiveness of sequential attention and self-attention modules. The VQA-seq denotes a version by replacing the protein distance map with protein sequence. We reported the average AUC and standard deviations of DrugVQA from three runs with different random seeds. Att, attention modules.

**Table 2 | Comparison results of proposed models and baselines on human dataset**

| Method | AUC | Recall | Precision |
|---|---|---|---|
| k-NN | 0.860 | 0.927 | 0.798 |
| RF | 0.940 | 0.897 | 0.861 |
| L2 | 0.911 | 0.913 | 0.861 |
| GNN | 0.970 | 0.918 | 0.923 |
| DrugVQA | **0.979 ± 0.003** | **0.961 ± 0.002** | **0.954 ± 0.003** |

where $d(r_i, r_j)$ is the distance between Cα atoms of residues $i$ and $j$, and $d_0$ is set to 3.8 Å, which is the distance between neighbouring Cα atoms. Let $\hat{s}_i \in [\hat{s}(r_i, r_1), \ldots, \hat{s}(r_i, r_l)]$ be the $l$-dimensional distance vector with all the residues in $P$ of $r_i$. The protein can be represented as a distance matrix:

$$P = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_l]_{l\times l} \tag{3}$$

The goal of DPI prediction is to learn a system that takes a pair ($M$;$P$) as input and outputs label $y \in \{0, 1\}$, where $y = 1$ means an interaction between $M$ and $P$.

As shown in the Fig. 1, our DrugVQA model consists of two main components: a dynamic CNN with sequential attention ('Dynamic attentive CNN') and a BiLSTM with multi-head self-attention ('Self-attentive BiLSTM').

**Dynamic attentive CNN.** *Dynamic process.* In our model, an adapted CNN is employed to code protein distance maps to fixed-size vector representations. The CNN module consists of stacked residual blocks and a sequential self-attention block. The residual block we used is borrowed from Resnet[30] and each residual unit is stacked by a 5×5 convolutional layer and a 3×3 convolutional layer. We utilize the exponential linear unit (ELU)[31] instead of the traditionally used rectified linear unit (ReLU)[32]. Different from VQA tasks that often preprocess images to the same size, the real-world proteins are of different lengths of amino acids and cannot be scaled. Therefore, we want to design a dynamic neural network that could (1) handle inputs of variable lengths and (2) predict the importance of each amino acid. For this purpose, we first take off the pooling layers between the residual blocks and use zero padding to two sides of input to ensure that the results of residual blocks have the same size as the input. Concretely, given a protein distance map $P \in \mathbb{R}^{l\times l}$, the output of the last residual block remains the dimension of $l\times l\times N_f$, where $N_f$ is the number of filters of the last convolution layer. Afterwards, we use average pooling to compress the information-enriched output of residual blocks for the downstream processing.

*Sequential attention.* Through average pooling, we obtain a convolutional protein feature map $P_c \in \mathbb{R}^{l\times N_f}$. Practically, $P_c$ can be viewed as protein sequential representation where $l$ is the number of amino acids (sites) in the protein, and $N_f$ represents the spatial feature of each site. As most sites are not directly related to the binding with drugs, recognizing the small portion of binding sites is critical for the accurate prediction of DPIs. To handle the varying size of feature maps from the convolution layer and to emphasize the important binding sites, we adopt a sequential self-attention mechanism[33] to fully use these features for classification. Concretely, the attention mechanism takes $P_c$ as input, and outputs a vector of weights $a^p$ (the attention matrix for the protein):

$$\begin{aligned} a^p &= \mathrm{softmax}(w_{p2}\tanh(W_{p1}P_c^T)) \\ &\textstyle\sum_{i=1}^{I}(a_i^p) = 1, \forall i, 1\le i\le l \end{aligned} \tag{4}$$

where $W_{p1} \in \mathbb{R}^{d_p\times N_f}$, and $w_{p2}$ is a row vector of parameters with size $d_p$, where $d_p$ is an adjustable hyperparameter. This vector representation usually focuses on a set of consecutive sites of protein sequence. Since a protein binding pocket is composed of multiple consecutive sites neighboured in space, we further extend the $w_{p2}$ into a $r_p$-by-$d_p$ matrix, noted as $W_{p2}$, to capture the overall structural information of the binding pocket. Thus, $a^p$ is converted to a multi-head attention weight $A^p \in \mathbb{R}^{r_p\times l}$ as

$$A^p = \mathrm{softmax}(W_{p2}\tanh(W_{p1}P_c^T)) \tag{5}$$

Practically, equation (5) can be deemed as a two-layer multilayer perceptron (MLP) without bias, whose hidden unit numbers is $d_p$, and parameters are $\{W_{p1}, W_{p2}\}$. We compute the $r_p$ weighted sums by multiplying the annotation matrix $A^p$ and feature map $P_c$:

$$P_a = A^pP_c \tag{6}$$

where $P_a$ is an attentional feature map containing the latent relationship between contribution of sites on the interaction. The size of $P_a$ is $r_p$-by-$N_f$, where $r_p$ is an adjustable hyperparameter representing the number of attention vectors.

**Self-attentive BiLSTM.** Each drug molecular SMILES string is encoded to a 2D embedding matrix $M \in \mathbb{R}^{n\times d}$. Token vectors in the molecular matrix $M$ are independent of each other. To gain some dependency between adjacent tokens within a molecule, a BiLSTM is used to process a molecule:

$$\overrightarrow{h_i} = \overrightarrow{\mathrm{LSTM}}(t_i, \overrightarrow{h_{i-1}}) \tag{7}$$

$$\overleftarrow{h_i} = \overleftarrow{\mathrm{LSTM}}(t_i, \overleftarrow{h_{i+1}}) \tag{8}$$

$\overrightarrow{h_i}$ is concatenated with $\overleftarrow{h_i}$, and a hidden state $h_i$ is obtained to replace token embedding $t_i$, and thus $h_t$ becomes a more information-enriched vector that gains some dependency between adjacent tokens in a molecule. For simplicity, we note all $h_i$ in every time step $i$ as $H$.

$$h_i = (\overrightarrow{h_i}, \overleftarrow{h_i}) \tag{9}$$

$$H = (h_0, h_i \cdots h_n) \tag{10}$$

If the hidden unit number for each unidirectional LSTM is set as $u$, the shape of $H$ would be $n$-by-$2u$.

The next goal is to know which part of the molecule contributed most to the interaction prediction. In other words, we want to

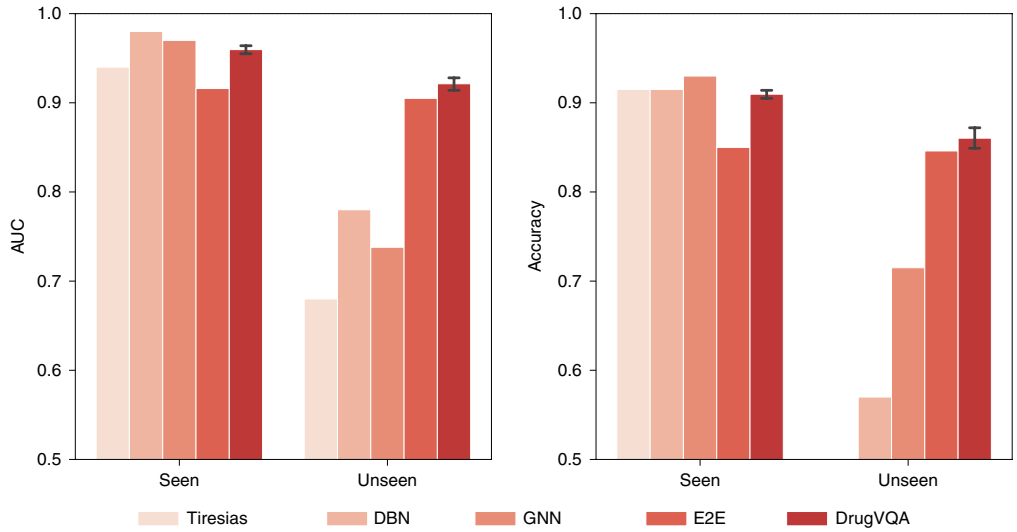


**Fig. 2 | Performance comparisons of our proposed method and baselines on seen and unseen protein targets from the BindingDB dataset.** Error bars indicate the standard deviations.

**Table 3 | Comparison of the proposed DrugVQA model to different types of baselines across targets in the DUD-E dataset**

| Category | Model | AUC | 0.5% RE | 1.0% RE | 2.0% RE | 5.0% RE |
|---|---|---|---|---|---|---|
| ML scoring function | NNscore | 0.584 | 4.166 | 2.980 | 2.460 | 1.891 |
| | RF-score | 0.622 | 5.628 | 4.274 | 3.499 | 2.678 |
| Docking based | Vina | 0.716 | 9.139 | 7.321 | 5.811 | 4.444 |
| | Smina | 0.696 | - | - | - | - |
| Deep learning based | 3D-CNN | 0.868 | 42.559 | 26.655 | 19.363 | 10.710 |
| | AtomNet | 0.895 | - | - | - | - |
| | PocketGCN | 0.886 | 44.406 | 29.748 | 19.408 | 10.735 |
| | GNN | 0.940 | - | - | - | - |
| Proposed | DrugVQA | **0.972 ± 0.003** | **88.17 ± 4.88** | **58.71 ± 2.74** | **35.06 ± 1.91** | **17.39 ± 0.94** |

The results of DrugVQA are the averages and standard deviations of AUC and RE from the threefold cross-validation.

identify the relationship between tokens and interaction, which can be used for a chemist to design or improve chemical compounds. Similarly, we achieve this by introducing multi-head self-attention mechanism into BiLSTM. The attention mechanism takes the whole LSTM hidden states $H$ as input, and outputs a vector of weights $A^{\mathrm{m}}$ (attention matrix for the molecule) as

$$A^{\mathrm{m}} = \mathrm{softmax}(\mathrm{MLP}(H^T)) \tag{11}$$

where hidden unit numbers of MLP is $d_{\mathrm{m}}$, and parameters are $\{W_{\mathrm{m1}}, W_{\mathrm{m2}}\}$. We compute the weighted sums by multiplying the annotation matrix $A^{\mathrm{m}}$ and LSTM hidden states $H$, the resulting matrix is the self-attentive molecular embedding:

$$M_{\mathrm{a}} = A^{\mathrm{m}}H \tag{12}$$

where $M_{\mathrm{a}}$ is a self-attentive drug molecular feature map that contains the latent relationship between tokens contribution of interaction. The size of $M_a$ is $r_{\mathrm{m}}$-by-$2u$.

**Classifier.** For $P_{\mathrm{a}}$ and $M_{\mathrm{a}}$, we summed up over all the attention vectors, and then normalized the resulting weight vector to sum up to 1. This process enables us to obtain two information-enriched 1D vectors $\widehat{P}_{\mathrm{a}}$ and $\widehat{M}_{\mathrm{a}}$, which will be fed into the classification layer. We concatenate $\widehat{P}_{\mathrm{a}}$ and $\widehat{M}_{\mathrm{a}}$, that is, $[\widehat{P}_{\mathrm{a}};\widehat{M}_{\mathrm{a}}]$, and obtain an output $o \in \mathbb{R}$:

$$o = W_o[\widehat{P}_{\mathrm{a}};\widehat{M}_{\mathrm{a}}] + b_o \tag{13}$$

where $W_o \in \mathbb{R}^{N_f+2u}$ is the weight matrix and $b_o \in \mathbb{R}$ is the bias. Given a dataset $D=\{(m_i, p_i, y_i)\}$, the training objective is to minimize the cross entropy as follows:

$$\mathcal{L}_{\mathrm{CE}}(\Theta) = -\sum_{i=1}^{N}(y_i\log(\sigma(o_i)) + (1-y_i)\log(1-\sigma(o_i))) + \frac{\lambda}{2}\|\Theta\|_2^2 \tag{14}$$

where $\Theta$ is the set of all weight matrices and bias vectors in our system, $N$ is the total number of drug–protein pairs in the training dataset and $\sigma$ denotes the sigmoid function. $\lambda$ is an L2 regularization hyperparameter.

## Experiments

Please add some text here.

**Datasets.** To enable head-to-head comparisons of DrugVQA to existing machine learning-based methods and docking programs,

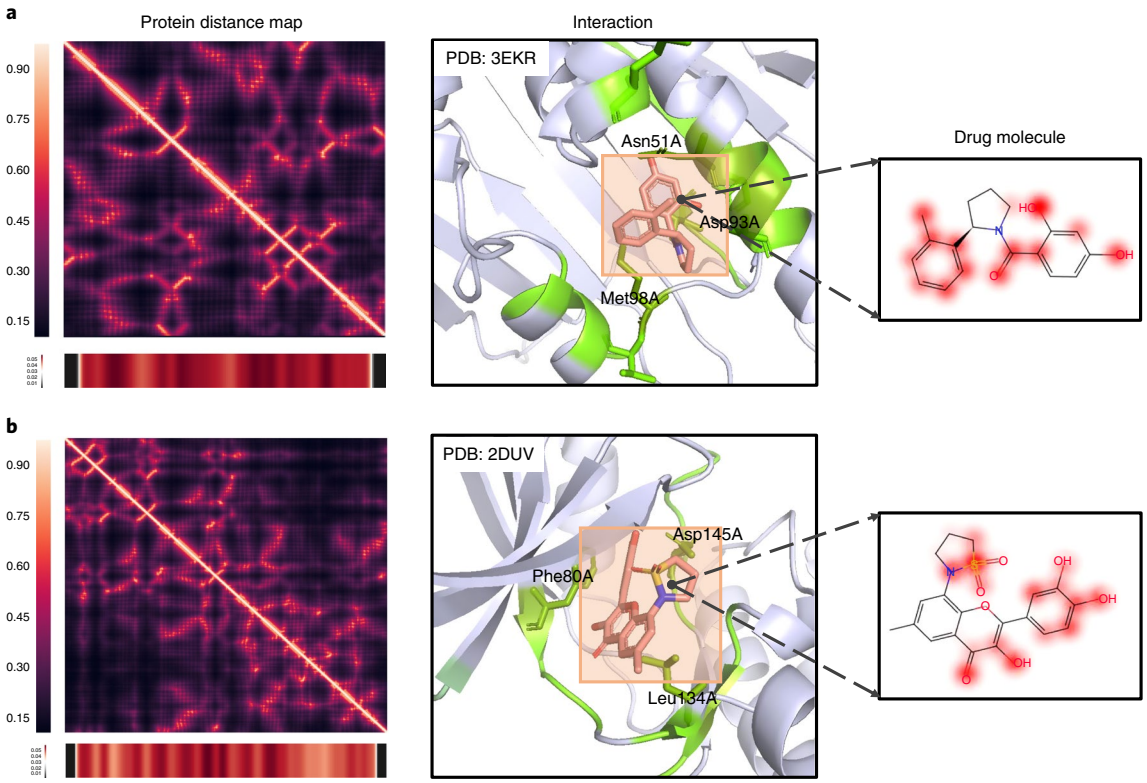


**Fig. 3 | Importance visualization of pocket and ligand pairs. a**, 3EKR. **b**, 2DUV. The corresponding sequential attention bars are shown below the pairwise maps. The green highlights the sites with high attentions within or surrounding the binding pocket, and the red cloud indicates drug atoms with attentions. Darker colours indicate higher attention coefficients.

we evaluated our proposed model on three public DPI datasets: the directory of useful decoys, enhanced (DUD-E) dataset, the human dataset and the BindingDB dataset.

*DUD-E.* The DUD-E dataset is a well-known benchmark consisting of 102 targets across 8 protein families[34]. Each target has 224 actives and over 10,000 decoys, on average. The decoys are chosen such that they are physically similar but topologically dissimilar to the actives. The final dataset contains 22,645 positive examples and 1,407,145 negative examples. We adopt a threefold cross-validation strategy to train and evaluate our model on the DUD-E dataset following ref. [3]. The folds were split between targets, where all ligands of the same target belong to the same fold. To avoid the impact of homologous proteins, targets belonging to the same protein families were strictly kept in the same fold. For fast training of models, we used a balanced set (all positives and randomly chosen equivalent negatives for each target) for training, but kept using the whole set (unbalanced ones) for evaluation.

*Human.* Created by ref. [35], this dataset includes highly credible negative samples of compound–protein pairs obtained by using a systematic screening framework. Following ref. [4], we used a balanced dataset, where the ratio of positive and negative samples was 1:1. Finally, the human dataset contains 5,423 interactions and 1,803 unique proteins. Also following ref. [4], we use an 80%/10%/10% training/validation/testing random split.

*BindingDB.* We further choose the BindingDB dataset[36] as the real-world dataset to evaluate our model. BindingDB is a public database of experimentally measured binding affinities, focusing chiefly on the interactions of small molecules and proteins. In our experiments, we used the customized BindingDB dataset constructed by ref. [5] for head-to-head comparisons. The dataset contains 39,747 positive examples and 31,218 negative examples from BindingDB. The dataset was divided into three splits: train (50,155 interactions), validation (5,607 interactions) and test (5,508 interactions) sets. To further validate the generalization of our model towards the unseen proteins, we then split the testing interactions into two parts, the proteins are observed in the training set and not.

**Implementation and evaluation strategy.** *Implementation details.* We implemented the proposed model with Pytorch 0.4.0[37]. The training process lasts at most 50 epochs on all the datasets using the Adam optimizer[38] with a learning rate of 0.001 and batch size of 1. Considering the limitation of memory of the used GPU (GTX1080Ti 12GB), we employed 32 residual blocks with 16 and 32 filters, respectively. The hidden state of the BiLSTM was set to 64 (the $u$ in 'Self-attentive BiLSTM') and 0.2 dropout was applied on the BiLSTM. In addition, attention MLPs both in CNN and BiLSTM had a hidden layer with 100 units (the $d_p$ and $d_m$), and we chose the matrix embedding to have 10 rows (the $r_p$) for protein and 20 rows (the $r_m$) for drug. The coefficient of L2 regularization was 0.001. We explored hyperparameters in a wide range and found the above set of hyperparameters yields the highest performance on the human dataset. Hyperparameter space and learning curves for various hyperparameters on the human validation set are shown in Supplementary Table 1 and Supplementary Fig. 1. We note that the rest of the hyperparameters on the other two datasets were not tuned, as we found the model performance was not sensitive to reasonable settings. All experiments were repeated three times, each with a different random seed.

*Evaluation metrics.* The performance was evaluated by the area under the receiver operating characteristic curve (AUC). In addition,

for the human dataset, we report the precision and recall value following ref. [4]. For the DUD-E dataset, we report the ROC enrichment metric (RE) following the work of ref. [3]. Specifically, the RE score is defined as the ratio of the true positive rate to the false positive rate (FPR) at a given FPR threshold. Here, we report the RE scores at 0.5%, 1%, 2% and 5% FPR thresholds. For the BindingDB dataset, we also report the accuracy following ref. [5].

**Ablation study.** We conducted ablation studies on the three benchmarks to investigate factors that influence the performance of the proposed DrugVQA framework. The results are shown in Table 1. We first investigated the influence of the attention modules. As shown in the first three lines, removal of the attention modules decrease the AUC by 2.3– 3.9%. This demonstrates the effectiveness of the two attention modules. To validate the influence of the protein distance map, we also included a version of our model by replacing the protein distance map with protein sequence (denoted as VQA-seq). The protein sequences were processed as drug SMILES through self-attentive BiLSTM. As shown in the last two lines, using distance maps leads to consistent gains on all three datasets. This agrees with our expectation that the distance maps contain more information than sequences for the prediction of DPI.

**Comparisons on the human dataset.** *Compared models.* In this section, we compare our DrugVQA with the state-of-the-art DPI approaches on the human dataset. We compare it with k-nearest neighbour (k-NN), random forest (RF), L2-logistic (L2) (results obtained from ref. [35]) and graph neural network (GNN)[4] (we retrained the model with the same parameter settings as in the original papers). We reported the average AUC, recall, precision and their standard deviations by DrugVQA from three runs with random seeds.

*Results.* As shown in Table 2, our DrugVQA system outperforms the current state-of-the-art GNN model with an increase of 0.9%, 4.3% and 3.1% for AUC, recall and precision, respectively. This phenomenon is in line with our expectation. Our system contains the protein structure information, which is helpful to learn a more informative representation of protein. Other descriptor-based machine learning techniques have low performance with AUC ranging from 0.86 to 0.94, indicating that the end-to-end learned representations can learn important information from proteins and drugs for DPI prediction.

**Comparisons on the BindingDB dataset.** *Compared models.* We further assess our model on the BindingDB dataset. We compare our model with four baselines: (1) similarity-based method Tiresias[39]; (2) deep belief networks (DBN)[40], a deep learning method using middle-level features from predefined molecular fingerprints and protein descriptors; (3) end-to-end neural network model E2E[5] using GCN and LSTM to process drug molecules and protein high-level information (Gene Ontology annotations), respectively; and (4) GNN[4].

*Results.* The experimental results for the BindingDB dataset are shown in Fig. 2. Our approach consistently performs well across the test sets and all metrics. When the tested proteins were observed in the training data (seen), DrugVQA achieved an AUC of $0.955 \pm 0.005$ and an accuracy of $0.916 \pm 0.005$. The AUC and accuracy were $0.922 \pm 0.007$ and $0.861 \pm 0.009$ when the tested proteins were not observed (unseen). Three baselines (Tiresias, DBN and GNN) perform well on seen proteins, but have much worse performance on unseen proteins. This indicates there is over-fitting over proteins used for training. However, E2E gives a consistent performance for seen and unseen proteins, but it is consistently lower than DrugVQA by 2.6% and 2.3% for AUC and accuracy in average.

**Comparisons on the DUD-E dataset.** *Compared models.* We compare our DrugVQA with the state-of-the-art DPI approaches on DUD-E dataset, which can be divided into three categories: (1) conventional docking approaches Vina[2] and Smina[1]; (2) machine learning scoring functions NNScore[27] and RF-Score[26]; and (3) deep learning-based methods 3D-CNN[3], AtomNet[9], PocketGCN[41] and GNN[4].

*Results.* As listed in Table 3, DrugVQA achieved an order-of-magnitude improvement over baselines at a level of accuracy useful for drug discovery. Note that we calculated the results on a per-target basis and then we reported the average results (based on the three-fold cross-validation) on a total of 102 targets. The experiments were repeated three times to obtain the standard deviations. On the full DUD-E dataset, DrugVQA outperforms the state-of-the-art GNN model with an average AUC of 0.972 versus 0.94. Although 3D-CNN employs 3D structure for training, it has the lowest performance among the three deep learning methods. This is probably due to the sparse data in 3D space, whereas the 2D pairwise distance map provides a good balance. As a result, DrugVQA outperforms 3D-CNN for 96% of the DUD-E targets on a per-target basis. We also show the cross-validation performance as well as variations of DrugVQA model on the DUD-E benchmark compared with the Vina scoring function and 3D-CNN in Supplementary Fig. 1.

**Attention visualization.** Another advantage of our model is its interpretability. To exemplify this, we selected two top predicted interactions in the DUD-E dataset. By input of the protein distance map and SMILE of compounds, the model produced multi-head attentions for proteins and compounds. We coloured the top-15 weighted residues of the example proteins with green and compared them to their diagram and interactions (annotated by experts) retrieved from the Protein Data Bank (PDB)[42]. We found that the highest-weighted amino acids (green) and compound atoms (red) overlap substantially with the real interaction sites. For protein Hsp90 (Fig. 3a), the attention bar highlights residues Asn51A, Asp93A and Met98A, which highly overlap with the key pocket residues observed in the co-crystal complex (PDB: 3EKR). For protein CDK2 (Fig. 3b), the highlighted residues (Phe80A, Asp145A, Leu134A) and ligand functional groups in the importance maps show high similarity to observed interactions in 2DUV. Thus, our model gives reasonable clues on the factors for the binding, which may have broad biomedical applications. A few more examples are shown in Supplementary Figs. 3–5.

## Conclusion

Here we have presented a novel end-to-end deep learning framework-like VQA task to predict DPIs. We have used self-attentive convolutional and recurrent structures to extract features simultaneously from a protein 2D distance map and molecular language in a DPI study. Experimental evaluations show that our model consistently has the best performance on three public datasets. Furthermore, the model is shown to be able to provide biological insights for understanding the nature of molecular interactions. The substantial improvement over the DPI task and the modular system suggests a new strategy of combining VQA with biological questions, including all 3D structure-based predictions, such as protein interaction, protein function and protein design.

## Data availability

All data used in this paper are publicly available and can be accessed at http://dude.docking.org for the DUD-E dataset, https://github.com/IBMInterpretableDTIP for the BindingDB-IBM dataset, https://github.com/masashitsubaki/CPI_prediction/tree/master/dataset for human dataset and https://www.rcsb.org for the protein crystal structure.

## Code availability

Demo, instructions and code for DrugVQA are available at https://github.com/prokia/drugVQA.

Received: 19 September 2019; Accepted: 13 January 2020;
Published online: 14 February 2020

## References

1. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
2. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
3. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
4. Tsubaki, M., Tomii, K. & Sese, J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**, 309–318 (2018).
5. Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S. & Zhang, P. Interpretable drug target prediction using deep neural representation. In *Int. Joint Conf. on Artificial Intelligence* 3371–3377 (IJCAI, 2018).
6. Zheng, S., Yan, X., Yang, Y. & Xu, J. Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *J. Chem. Inf. Model.* **59**, 914–923 (2018).
7. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
8. Jastrzebski, S., Leśniak, D. & Czarnecki, W. M. Learning to SMILE(S). Preprint at https://arxiv.org/abs/1602.06289 (2016).
9. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Preprint at https://arxiv.org/abs/1510.02855 (2015).
10. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).
11. Skolnick, J., Kolinski, A. & Ortiz, A. R. MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217–241 (1997).
12. Namrata, A. & Possu, H. Generative modeling for protein structures. *Adv. Neural Inf. Process. Syst.* **31**, 7494–7505 (2018).
13. Bepler, T. & Berger, B. Learning protein sequence embeddings using information from structure. Preprint at https://arxiv.org/abs/1902.08661 (2019).
14. Yang, Z., He, X., Gao, J., Deng, L. & Smola, A. Stacked attention networks for image question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 21–29 (2016).
15. Xu, K. et al. Show, attend and tell: neural image caption generation with visual attention. In *Int. Conf. on Machine Learning* **37**, 2048–2057 (PMLR, 2015).
16. Noh, H., Seo, P. H. & Han, B. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 30–38 (IEEE, 2016).
17. Agrawal, A. et al. VQA: visual question answering. *Int. J. Comput. Vis.* **123**, 4–31 (2017).
18. Antol, S. et al. VQA: Visual Question Answering. In *Proc. IEEE International Conference on Computer Vision* 2425–2433 (IEEE, 2015).
19. Weininger, D. et al. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
20. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
21. Ma, L., Lu, Z. & Li, H. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence* (AAAI, 2016).
22. Shih, K. J., Singh, S. & Hoiem, D. Where to look: focus regions for visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4613–4621 (IEEE, 2016).
23. Xu, H. & Saenko, K. Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision* (Springer, 2016).
24. Schwartz, I., Schwing, A. & Hazan, T. High-order attention models for visual question answering. *Adv. Neural Inf. Process. Syst.* 3664–3674 (2017).
25. Bleakley, K. & Yamanishi, Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* **25**, 2397–2403 (2009).
26. Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).
27. Durrant, J. D. & McCammon, J. A. NNScore 2.0: a neural-network receptor–ligand scoring function. *J. Chem. Inf. Model.* **51**, 2897–2903 (2011).
28. Tabei, Y. & Yamanishi, Y. Scalable prediction of compound–protein interactions using minwise hashing. *BMC Syst. Biol.* **7**, S3 (2013).
29. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
30. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision* 630–645 (Springer, 2016).
31. D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). Preprint at https://arxiv.org/abs/1511.07289 (2015).
32. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th International Conference on Machine Learning* 807–814 (ICML, 2010).
33. Lin, Z. et al. A structured self-attentive sentence embedding. Preprint at https://arxiv.org/abs/1703.03130 (2017).
34. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
35. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **31**, i221–i229 (2015).
36. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2015).
37. Paszke, A. et al. Automatic differentiation in PyTorch. In *Neural Information Processing Systems Workshop Autodiff* (NeurIPS, 2017).
38. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
39. Fokoue, A., Sadoghi, M., Hassanzadeh, O. & Zhang, P. Predicting drug–drug interactions through large-scale similarity-based link prediction. In *European Semantic Web Conference* 774–789 (Springer, 2016).
40. Wen, M. et al. Deep-learning-based drug–target interaction prediction. *J. Proteome Res.* **16**, 1401–1409 (2017).
41. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model.* **59**, 4131–4149 (2019).
42. Burley, S. K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2018).

## Acknowledgements

The work was supported in part by the National Key R&D Program of China (2018YFC0910500), GD Frontier and Key Tech Innovation Program (2018B0101090 06,2019B020228001), the National Natural Science Foundation of China (61772566, U1611261 and 81801132, 81903540) and the programme for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

## Author contributions

S.Z., Y.L. and Y.Y. contributed concept and implementation. S.Z. and Y.L. co-designed experiments. S.Z. and Y.L. were responsible for programming. All authors contributed to the interpretation of results. S.Z. and Y.Y. wrote the manuscript. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s42256-020-0152-y.

**Correspondence and requests for materials** should be addressed to J.X. or Y.Y.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.