An Interpretable Framework for Drug-Target Interaction with Gated Cross Attention

Yeachan Kim   YEACHAN@DEARGEN.ME

Bonggun Shin   BONGGUN.SHIN@DEARGEN.ME

Deargen Inc., Seoul, South Korea

Abstract

In silico prediction of drug-target interactions (DTI) is significant for drug discovery because it can largely reduce timelines and costs in the drug development process. Specifically, deep learning-based DTI approaches have been shown promising results in terms of accuracy and low cost for the prediction. However, they pay little attention to the interpretability of their prediction results and feature-level interactions between a drug and a target. In this study, we propose a novel interpretable framework that can provide reasonable cues for the interaction sites. To this end, we elaborately design a *gated cross-attention* mechanism that crossly attends drug and target features by constructing explicit interactions between these features. The gating function in the method enables neural models to focus on salient regions over entire sequences of drugs and proteins, and the byproduct from the function, which is the attention map, could serve as interpretable factors. The experimental results show the efficacy of the proposed method in two DTI datasets. Additionally, we show that gated cross-attention can sensitively react to the mutation, and this result could provide insights into the identification of novel drugs targeting mutant proteins.

1. Introduction

Predicting the interactions between drugs and targets is a critical step in drug discovery. The importance is more evident in repurposing (or repositioning) already-approved drugs for fast-spreading and epidemic diseases (e.g., COVID-19) because the approval process of a new drug requires significant development timelines and overall costs. Recently, deep neural networks have shown promising results on this problem (Öztürk et al., 2018; Shin et al., 2019), demonstrating that they can work well in real-world scenarios (Beck et al., 2020; Choi et al., 2020).

Most recent studies for predicting the interaction focus on representation learning for drugs and targets. For example, Shin et al. (2019) used a transformer (Vaswani et al., 2017) to effectively capture salient features from molecules and further improved the generalization capability through a pre-training phase on a large-scale database (e.g., PubChem (Kim et al., 2019)). Nguyen et al. (2019) replaced the sequential representations of drugs with a molecular graph using graph convolutional networks (GCNs). Although advancements in representation learning have improved the performance, most studies pay little attention to *i)* the interpretability of their prediction results and *ii)* the interaction of drug and target features.

To further improve the current approaches, we propose a *gated cross-attention* (GCA), which is a novel interpretable and interaction framework that can provide the interpretabil-

ity of the prediction results through an explicit interaction between the drug and target features. Interpretability is served through the lens of attention to drugs and targets. In other words, the higher attention score at the specific parts of the drug and target, the more significant they are for prediction. To that end, we redesign the previous attention mechanism as a *gating function* so that the significant parts for the prediction easily affect to the prediction. Furthermore, unlike the previous attention mechanism that typically assigns all credits to all inputs, GCA sparsely attends a small portion of inputs to encourage the neural models to focus on salient parts over sequences.

To validate the efficacy of GCA, we apply the proposed method to existing neural models (e.g., convolutional neural networks (CNNs), GCNs) because the role of our method is to put current methods on a more solid footing by obtaining interpretability on their prediction. The comprehensive results clearly demonstrate that the proposed interaction networks not only provide the interpretable cues for the prediction by showing significant parts in drugs and targets, but also achieve performance improvement owing to the explicit interaction between features. We also conduct two case studies on mutations and possible interactions between drugs and targets and it can be the pharmaceutical use cases of the proposed method.

Clinical Relevance: The proposed method can help design new drugs by proposing a binding affinity with important regions of drugs and targets. It can also be used in drug design targeting specific mutant proteins by checking whether the attention in the mutation sites is activated differently.

Generalizable Insights about Machine Learning in the Context of Healthcare

Due to the large length of protein sequences, effectively handling protein representation is difficult. In this case, focusing salient parts over the large sequence is significant, and the sparse attention mechanism can alleviate the problem. There are several tasks taking two inputs as an input, such as drug-drug interaction and protein-protein interaction. Cross-attention mechanism over feature extraction layers could improve the performance in tasks while providing interaction interpretability.

2. Related Work

2.1. Drug Target Interaction

In silico prediction of DTI has been getting attention since it could provide useful guidance to drug discovery with reduced overall costs. The major approaches to DTI are three-folds which are molecular docking (Trott and Olson, 2010; Luo et al., 2016), similarity-based (Pahikkala et al., 2015; He et al., 2017) and deep learning-based approaches. In this paper, we mainly focus on deep learning-based methods since they could provide the results promptly and achieve reasonably accurate results compared to other approaches (Öztürk et al., 2018). In the beginning, several works utilized deep neural networks over character representations of drugs and targets. For drugs, SMILES (Simplified Molecular Input Line Entry System) format is utilized to represent chemical compound, and protein sequence (i.e., amino acids) is represented by FASTA (Lipman and Pearson, 1985). Based on such character representations, Öztürk et al. (2018) proposed DeepDTA that utilizes convolutional neural

networks (CNNs) to extract locally salient features from drug and target, and it achieved better performance than similarity-based approaches. Shin et al. (2019) used the same character representation with DeepDTA, but utilized the self-attention-based transformer (Vaswani et al., 2017) to consider a long-range relationship between atoms. Furthermore, they performed transfer learning to further generalize the molecule representations by pre-training the transformer on a large-scale database. Instead of character representations, (Öztürk et al., 2019) represented drug and target as words that correspond to substructure in compounds. Beyond such string representations, utilizing graph structures of drugs shows promising performance than others (Lim et al., 2019; Nguyen et al., 2019; Zhao et al., 2020).

The most related work to ours is Gao et al. (2018). It shares a similar strategy with ours in the aspect of attending to two inputs. However, our method adopts a multi-head attention mechanism in the transformer. It allows neural models to attend to different aspects of information that lie in input features. Thus, it serves more options to interpret the prediction results. For example, we found out that one of the heads from GCA is more committed to predicting the binding site than other heads (detailed in Section 5.2). Beyond the differences in architecture, the proposed model constraints the attention map to be sparse, which guides the model to more focus salient parts in inputs.

2.2. Attention Method in Neural Networks

Attention mechanism was invented to improve the quality of machine translation by aligning two different representations (e.g., source and target languages) (Bahdanau et al., 2015; Luong et al., 2015). There are two advantages in utilizing attention: First, it allows neural models to effectively consider a long-range relationship between features and it leads to the improved task performance (Yang et al., 2016; Anderson et al., 2018). Another advantage of the attention mechanism is that it could provide the interpretability of the model prediction (Seo et al., 2017; Annasamy and Sycara, 2019). In the interaction task of different domains, several works utilize the attention mechanism to provide interpretability. For example, Santos et al. (2016) utilized pooling operation over the interactive attention maps between question and answer for an answer selection task. Similarly, Gong et al. (2018) used densely-connected networks over attention maps to extract locally salient interactions between premise and hypothesis for a natural language inference task. For a molecule optimization task, Shin et al. (2021) utilized a self-attention-based transformer to optimize multiple properties of a given molecule.

For drug-target interaction, most works have shown the strengths of the attention mechanism at generating better representations. However, they pay little attention to the interpretability and interaction between two inputs. In this work, we focus on the above properties to improve the task performance and provide reasonable cues for the factors to the interaction binding, which may have broad pharmaceutical applications. Note that our method is parallel research lines with other works. Thus, we believe that existing methods could benefit from ours.

3. Gated Cross-Attention Networks

In this section, we elaborate the proposed method which is denoted as gated cross attention (GCA). Before delving into the method, we first describe the background of drug-target
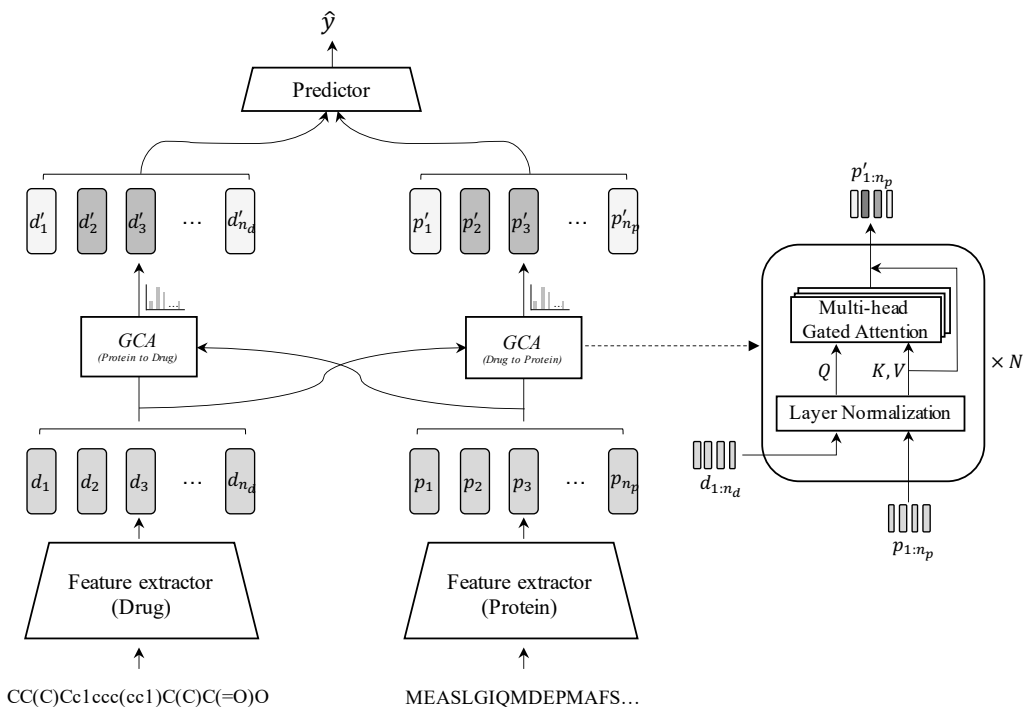
Figure 1: Overview of gated cross attention (GCA) networks and the detailed procedures when deriving the protein attention (right). GCA explicitly constructs the interaction between drug and target features, which are obtained from each feature extractor, using multi-head gated attention. The attention map, which is the by-product of the explicit interaction, serves as an interpretable factor for the prediction results.

interaction and formally define the task (Section 3.1). Then, we introduce the attention mechanism in the transformer with some negative observation at the interaction task (Section 3.2.1). Lastly, we detail gated cross attention (Section 3.2.2). The overall procedures are straightforward and illustrated in Figure 1. The main strategy of GCA is to generate drug and target features by considering counterpart information. The name **cross** in our method comes from the directionality of the attention in our method

3.1. Drug-Target Interaction

The task of this paper is to predict the affinity between a drug and a protein (also noted as target). The affinity can be measured as either scores or categories, the task of which is then regression or classification, respectively. Since predicting affinity scores is much beneficial in the process of drug discovery (Öztürk et al., 2018), we focus on regression-based modeling. Formally, the inputs are two kinds: a molecule represented by the SMILES (Weininger, 1988) sequence or the molecular graph, and a protein represented by the FASTA (Lipman and Pearson, 1985) sequence. A FASTA sequence is a series of various amino acids (also noted as residues).

4

The typical neural frameworks for drug-target interaction consist of two major components: **1)** two feature extraction layers for drugs and targets, and **2)** a single fully-connected layer to predict binding affinity scores based on the resultant features from **1)**. Let the drug and target features from each feature extractor be $d_{1:n_d}$ and $p_{1:n_p}$ where n_d, n_p are the sequential length of each feature and the subscript indicates the sequential range of features, we formally define the task as follows:

$$r_d = pooling(d_{1:n_d}), \ r_p = pooling(p_{1:n_p}), \tag{1}$$

$$\hat{y} = \phi(r_d, r_p; \theta_i). \tag{2}$$

where $pooling(\cdot)$ is the pooling methods such as global max pooling and average pooling, which depends on the architecture. $\hat{y}$ is the predicted binding affinity score by an interaction layer $\phi(\cdot)$ with its parameter θ_i. Most works used simple feed-forward networks for an interaction function. In this work, we enhance the interaction parts $\phi(\cdot)$ by explicitly making the interaction between drugs $d_{1:n_d}$ and targets $p_{1:n_p}$.

3.2. Cross Attention

To explicitly consider the interaction, we propose a *cross-attention* that considers protein representations to attend the drug representations and vice versa. We realize this by cross-attending each feature by its counterpart representations. We formalize this as follows:

$$d'_{1:n_d} = g_{p \to d}(d_{1:n_d}, p_{1:n_p}; \theta_{p \to d}), \tag{3}$$

$$p'_{1:n_p} = g_{d \to p}(p_{1:n_p}, d_{1:n_d}; \theta_{d \to p}). \tag{4}$$

where $g(\cdot)$ indicates an attention function that considers counterpart information. For example, $g_{p \to d}$ takes protein features to attend drug features with its learnable parameters $\theta_{p \to d}$ and returns the attended features of $d_{1:n_d}$. The attention weights, which are the byproducts of $g(\cdot)$, can serve as interpretable factors because the weights exhibit the most contributing sequences (i.e., amino acids and atoms) to predict affinity. In the following section, we will describe the functions are suitable for $g(\cdot)$.

3.2.1. MULTI-HEAD ATTENTION

As an attention function g, we follow multi-head attention in the transformer. There are two different types of attention: **1)** encoder attention and **2)** decoder attention since the transformer was constructed for a sequence-to-sequence task (e.g., machine translation). We only describe the process of calculating attention for drug features $d_{1:n_d}$ because the attention for a protein is symmetric with that of drugs.

Encoder Attention

The key idea of the encoder attention is to attend inputs by using the same inputs but differently transformed, which is also called *self-attention*. This attention is repeatedly performed with different heads and layers. For simplicity, we show the process of a single layer with a single head.

We first construct *query*, *key*, and *value* by transforming the same inputs $d_{1:n_d}$ by weight matrices W_Q, W_K, and W_V. These features are denoted as Q_d, K_d, and V_d, respectively and share the same dimensionality with the inputs (i.e., $Q, K, V \in \mathbb{R}^{n_d \times b}$). The attention weights $A \in \mathbb{R}^{n_d \times n_d}$ are computed based on the query and key:

$$A = \text{softmax}(\frac{Q_d K_d^T}{\sqrt{b}}), \tag{5}$$

where b is the dimensionality of input features and it serves as temperature to relax the distribution of the attention. Using the attention weights A and the values V_d, we compute attended inputs as follows:

$$V_d^{'} = A V_d. \tag{6}$$

Here, A can be utilized as interpretable factor to the prediction. The above encoder attention is widely utilized at feature extraction layers in various domains, such as image (Cornia et al., 2020), natural language processing (Devlin et al., 2019), and drug-discovery (Shin et al., 2019, 2021).

Decoder Attention

Decoder attention is slightly different from the self-attention. This attention is used at the decoder in the transformer, and the goal is to align decoder features with encoder features. To align two different features, it uses the counterpart features $p_{1:n_p}$ instead of its inputs $d_{1:n_d}$ to construct the key and value, which are denoted as K_p and V_p, and the attended features in Eq. 6 are obtained as follows:

$$V_d^{'} = \text{softmax}(\frac{Q_d K_p^T}{\sqrt{b}}) V_p, \tag{7}$$

The decoder attention can be used as a $g_{p \rightarrow d}(\cdot)$ in the cross attention. However, we empirically found that the decoder attention results in highly mixed representations between the two features. Figure 2(a) shows the pairwise feature similarity between drugs $(d, d^{'})$ and proteins $(p, p^{'})$ which are inputs and outputs from the decoder attention-based $g(\cdot)$ function. We observe that the protein features $p^{'}$ from the decoder attention are more similar to drug features d than its input features p over several training ranges (e.g., 3k-18k and near 30k) while the attended features for drugs are more similar to its input than protein features. In other words, a specific modality could dominate the other, and we found that such cases lead to poor performance in the task. These results are derived from the method of attention weighting in Eq. 7 because the attention maps A are applied to counterpart features rather than its input features. This is the intended design for many sequence-to-sequence tasks, where the goal is to effectively align the target and source. However, we argue that this should be redesigned for interaction tasks.

3.2.2. MULTI-HEAD GATED ATTENTION

For the interaction task, we redesign the decoder attention. It shares the same strategy of multi-head attention with the decoder. However, it is vastly different in that **1)** It uses context-level attention instead of token-level attention **2)** The attention is applied in the
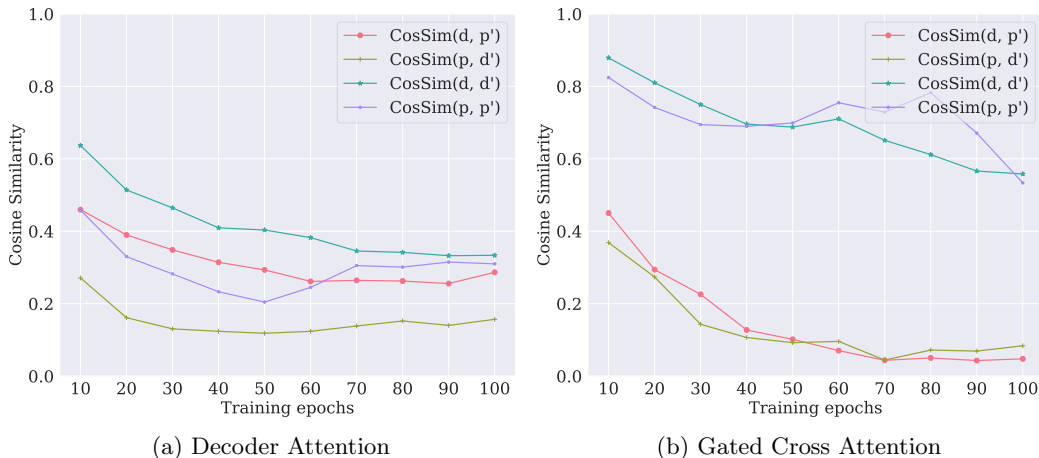
Figure 2: Pair-wise similarity between drugs (d, d') and protein features (p, p') over different training epochs when using different attention mechanism. It shows that utilizing decoder attention does not generate discriminate features between drug and target.

manner of *gating* mechanism in the variants of recurrent neural networks (Chung et al., 2014; Bradbury et al., 2017). The attention weights $a \in \mathbb{R}^{n_d}$ for drug features are calculated as follows:

$$a = \frac{1}{n_p} \sum_{i=1}^{n_p} \text{softmax}(\frac{Q_p K_d^T}{\sqrt{b}})_{:,i}, \tag{8}$$

In comparison to the decoder attention, the query and key inputs are reversed, and the attention map is aggregated by counterpart information, yielding a context-level attention vector rather than a token-wise attention matrix (Eq. 5). The attention vector a serves as a gating function in calculating the attended features with the values V_d:

$$V_d' = \text{softmax}(a) \odot V_d. \tag{9}$$

where $\odot$ is an element-wise multiplication operation[1]. After applying attention, we add a residual connection between attention layers, which creates good local minima (Yun et al., 2019), and follow pre-normalization (Xiong et al., 2020), as illustrated in Figure 1 (right).

The attended values V' are $d'_{1:n_d}$, and these features are pooled over the length n_d in Eq. 1. For the protein side, the attended features $p'_{1:n_p}$ are symmetrically computed using the above procedure. Lastly, the concatenated features of a drug and a target are fed to multi-layered feed-forward networks to predict the binding affinity. To train the overall networks for the drug-target interaction task, we minimize the mean squared error (MSE) between predicted affinity scores and ground-truth scores. Figure 2(b) shows that gated attention generates highly discriminative features while considering counterpart information.

---

1. If the dimensions do not match, the operation is performed in a broadcasting manner. We follow NumPy (Harris et al., 2020) convention (https://numpy.org/doc/stable/user/theory.broadcasting.html)

3.2.3. SPARSE ATTENTION

Our attention mechanism assigns probabilities (i.e., a in Eq 8 and 9) over all elements. However, most of them have a small but non-zero probability, which could weaken the attention given to the few truly significant elements (Shen et al., 2018; Zhao et al., 2019). In the context of drug-target interactions, it shares the same problem because a drug interacts with small portions of a protein. We constrain the attention map to be sparse to encourage the model to focus on a few salient parts. We achieve this by replacing the softmax function with the sparsemax (Martins and Astudillo, 2016) function, which considers the small weights to be unimportant and masks those values from the features when calculating the softmax-based attention.

4. Experiments

In this section, we evaluate the performance of our method in a drug-target interaction task on two different datasets.

4.1. Experimental Settings

4.1.1. COMPARISON METHODS

As previously stated, the GCA is designed to endow existing methods with interaction interpretability. Thus we compare the performance by applying our method to the following baselines:

- EmbDTA: Instead of a specific feature extractor, it only uses an embedding layer. We add this to see how much the proposed method improves the performance of the task using only the interaction. It uses SMILES and FASTA to represent the drugs and targets, respectively.

- DeepDTA (Öztürk et al., 2018): It uses CNNs for feature extraction layers to encode locally salient features from drugs and targets. It uses the same representations for drugs and targets as the EmbDTA.

- GraphDTA (Nguyen et al., 2019): It uses GCNs to exploit the properties of the molecular graph and utilizes CNNs for proteins as in DeepDTA. For an input representation, it uses atoms in drugs and connects them through graph information (i.e., adjacency matrix). To preserve sequential information, we use a sequence-wise convolution instead of feature-wise convolution, which is similar to Kim (2014) and we empirically found that this version produces nearly identical results with that of feature-wise convolution.

Note that the previous methods use multi-layered feed-forward networks for an interaction layer (i.e., $\phi(\cdot)$). We replace the networks with GCA without modifying the feature extractor on previous methods. For a fair comparison, we set the nearly same number of parameters used in the interaction layers.

In our model, we use a single cross-attention layer with two heads. The training setup is the same with the previous work Nguyen et al. (2019). We have implemented the proposed

Table 1: Statistics for KIBA and Davis dataset. Len. indicates the average string length of each format (SMILES, FASTA).

Dataset	# of Drugs	# of Targets	Len. Drugs (SMILES)	Len. Targets (FASTA)	# of Interactions	Affinity range
KIBA	2,111	229	56.6	656.4	118,254	[0.0, 17.2]
Davis	68	442	62.7	712.2	30,056	[5.0, 10.8]

method and all experiments with PyTorch (Paszke et al., 2017) and trained the models on a single NVIDIA Tesla V100 with 32GB of RAM.

4.1.2. DATASETS

We use two benchmark datasets to evaluate the performance in the drug-target interaction task , which are **KIBA** large-scale kinase inhibitors bioactivity dataset (Tang et al., 2014; He et al., 2017) and **Davis** kinase binding affinity dataset (Davis et al., 2011). The statistics of these datasets are listed in Table 1.

4.1.3. EVALUATION METRICS

We evaluate the methods by using mean squared error (MSE) and concordance-index (C-index) score following previous works (Öztürk et al., 2018; Nguyen et al., 2019). MSE is the average of regression loss in the test dataset, and C-index is the probability that randomly selected pairs are correctly ordered. It is computed as follows:

$$\text{C-index} = \frac{1}{N_{test}} \sum_{y_i > y_j} \mathbb{1}(\hat{y}_i > \hat{y}_j) \tag{10}$$

where N_{test} is the number of test pairs and, y_i, y_j are affinity scores of randomly selected pairs, $\hat{y}_i$, $\hat{y}_i$ are predicted affinity scores from the trained model. $\mathbb{1}(x)$ is a identity function that returns 1 if the statement x is true, otherwise 0.

4.2. Results

Table 2 shows the full comparison results in two datasets. Replacing the interaction networks with GCA consistently yields better results for most methods and datasets. Among them, EmbDTA, which has the least feature extraction parts, achieves greater performance improvement than others (roughly 9% improvement in terms of MSE). These results suggest that constructing explicit interaction is useful to improve the performance in the task. Furthermore, we see that GCA works well with the use of other input types (i.e., characters in EmbDTA and Deep-DTA, graphs in GraphDTA), and it shows that GCA can generally be applied to existing methods.

Table 3: Comparison with other interaction methods.

Method	MSE	C-index
EmbDTA	0.448	0.811
EmbDTA + DIN	0.523	0.776
EmbDTA + AP	0.439	0.824
EmbDTA + GCA	**0.393**	**0.832**

Table 2: Performance evaluation on KIBA and Davis. Best results are highlighted in boldface.

Method	KIBA		Davis	
	MSE	C-index	MSE	C-index
EmbDTA	0.342	0.761	0.438	0.811
EmbDTA + GCA	**0.311**	**0.784**	**0.393**	**0.832**
DeepDTA	0.186	0.865	0.246	0.881
DeepDTA + GCA	**0.170**	**0.881**	**0.238**	**0.888**
GraphDTA	**0.151**	**0.884**	0.251	0.883
GraphDTA + GCA	0.152	0.883	**0.242**	**0.891**

We also compare GCA with different interaction methods used in other domains. These methods are attentive pooling (AP, Santos et al. (2016)) and deep interaction networks (DIN, Gong et al. (2018)). Here, we experiment on the settings of EmbDTA in Davis to solely compare the interaction layers while minimizing the influence of the feature extractor. Table 3 presents the comparison results. Again, GCA works better than the others. In addition to achieving superior performance, we have several advantages. Compared to DIN, GCA explicitly enables the neural models to interact with a long-range relationships between drugs and target features owing to the attention mechanism. AP is the most similar method to GCA, as it also calculates attention maps and assigns weights to each feature. However, GCA adopts a multi-head and a multi-layer mechanism using the transformer, making the method more scalable by stacking multiple interaction layers and jointly attending to information from different representation subspaces at different positions (Vaswani et al., 2017). Indeed, we observe that the second head in GCA better captures the binding sites in proteins than the first head (detailed in Section 5.2). Such differences contribute to a better performance.

5. Analysis

5.1. Ablation Study

In this subsection, we analyze whether cross attention is effective compared to single direction attention and which attention is more significant for performance improvement. To this end, we measure the task performance when we remove the drug ($g_{p \to d}$) or target attention ($g_{d \to p}$), respectively. Here, we use EmbDTA trained on the KIBA dataset. Table 4 lists the ablation results. As can be seen from the table, the perfor-

Table 4: Ablation study for attention directionality

Method	MSE	C-index
GCA	0.311	0.787
GCA w/o Drug Attention	0.341	0.768
GCA w/o Target Attention	0.323	0.779

mance decreases as we remove each directionality from GCA. Specifically, removing drug

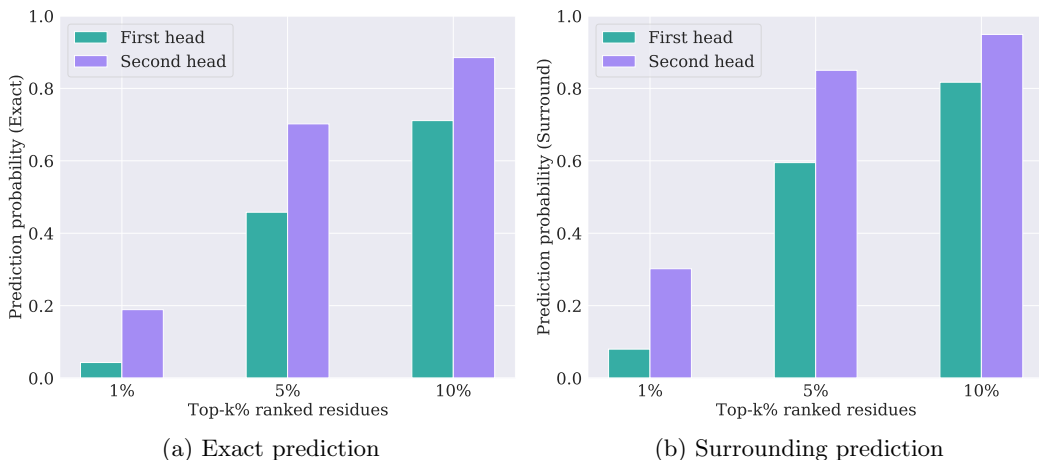(a) Exact prediction

(b) Surrounding prediction

Figure 3: Binding site prediction results. **Left** is the results of exact prediction. **Right** is the results of prediction with including nearest 1 position.

attention from targets results in the largest drop in performance. This indicates that drug attention is more significant than target attention. We also see that using both two types of attention leads to the best performance, demonstrating that cross-attention is indeed effective and that each attention complements usefully.

5.2. Binding site prediction

The proposed method assigns attention weights to a given protein such that the highlighted residues directly affect to the task performance. Therefore, residues receiving relatively high attention can be important local residues for prediction, such as binding sites. In this subsection, we analyze how the residues that are highly ranked by the attention correspond to actual binding sites. Here, we use GraphDTA with GCA, which is trained on KIBA, and use scPDB (Desaphy et al., 2015) for evaluation, which has atomic-level descriptions of proteins, ligands, and binding sites from complex structures. It contains interactions between proteins and pharmacological ligands for 7,179 entries of Vertebrata (the number of unique proteins is 809). We measure the probability of whether the top $k\%$ ranked residues include binding sites or not, and evaluate two predictions from two different heads in GCA. Figure 3 shows the prediction results. We first observe that the attention mostly finds the binding site when searching for the top 10% residues. Furthermore, the second head from our attention better captures binding sites than first head with a large margin. This shows that our method encode different aspects of the input features through multiple heads. These results suggest that our method can provide the insight in to the interpretability of binding sites by highlighting nearby residues.

Table 5: Rank of highlighted residues when the T790M mutation occurs

EGFR			EGFR with T790M		
Rank	Residue Index	Near Residues	Rank	Residue Index	Near Residues
1	655	GALLL	1	655	GALLL
...	...	...	...	...	...
12	792	**T**QLMP	5 (+7)	792	**M**QLMP
...	...	...	...	...	...
120	791	I**T**QLM	**11 (+109)**	791	I**M**QLM

5.3. Case Study

5.3.1. Interaction Interpretability

In our method, we provide two views on the prediction results: drug attention and the target attention. In this analysis, we confirm which atoms in drugs are receiving high attention from our method. To this end, we conduct a case study of the interaction between chemicals (PubChem ID: 117793281) and protein (MAP Kinase, UniProt ID: Q9HBH9).

Here, we use GraphDTA with GCA trained in KIBA and draw the top contributing atoms in Figure 4. As can be seen from the figure, the highlighted atoms in the drug attention include several nitrogen atoms that can form hydrogen bonds with amino acids at the binding site. As we use GCNs that interact with neighbors to represent drugs, several carbons near the nitrogen have received high attention, and they can also formulate potential solvent interaction since carbon has a relatively large solvent acces-

Figure 4: Highlighted atoms in the chemical (PubChem ID: 117793281)

sible surface. Overall, the results show that the highly attended parts of the drug attention can indicate potential interaction sites in drugs. Such interpretability can help researchers find leads by displaying the regions near the interaction sites with a binding affinity score.
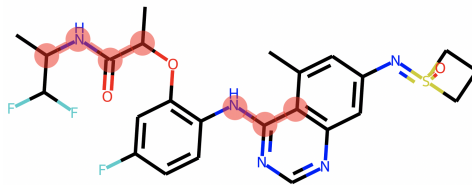
5.3.2. Mutation on epidermal growth factor receptor

Lastly, we analyze how the attention from ours is activated differently when a mutation occurs. We choose the mutation occurring at the epidermal growth factor receptor (EGFR) which is a well-known genes related to many cancer types. Specifically, we apply T790M mutation to EGFR, a well-known mutation found in most lung cancer patients. We use GraphDTA with GCA trained on KIBA. For a drug, we use Osimertinib which was developed as a third generation tyrosine kinase inhibitor targeting the T790M mutation in the exon of the EGFR gene (Soria et al., 2018). In this analysis, we compare the rank of residues in the target attention when the mutation occurs. Table 5 lists the ranked results before and after T790M mutation, in descending order of attention weights. We denote the ranks of the near residues where mutation occurs (i.e., 791, 792 residue)[2]. When the

---

2. As we use CNNs to extract features from proteins, the surrounding regions of attended parts are meaningful.

mutation occurs, the near residues of the mutation site are highly activated. While the increased ranking is not the most highlighted, such a drastic increase shows that GCA can react sensitively to the mutation and can be used as a guideline for identifying new drugs targeting specific mutant proteins.

6. Conclusion

In this paper, we have proposed a novel interpretable framework for drug-target interactions based on **_gated cross attention_**, denoted as GCA. GCA is elaborately designed to make the neural model explicitly interact with drugs and targets using the attention mechanism, and it assigns weights to each feature by considering counterpart information. Because the weights directly affect the prediction, the values in the attention can serve as an interpretable factor. The experimental results support the strengths of GCA in drug-target interaction and provide pharmaceutical insights gained from GCA for binding sites and potential interactions between drugs and targets. Furthermore, with the example of T790M EGFR, we have shown that GCA reacts sensitively to the mutant protein by mainly attending the surrounding mutation site. The comprehensive results suggest that our method can provide useful cues for the predictions of deep learning-based DTI methods, and we believe that existing methods can benefit from our work.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Raghuram Mandyam Annasamy and Katia Sycara. Towards better interpretability in deep q-networks. In *Proceedings of the Association for the Advancement of Artificial Intelligenc (AAAI)*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representation (ICLR)*, 2015.

Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal*, 2020.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-recurrent neural networks. *Proceedings of the International Conference on Learning Representation (ICLR)*, 2017.

Yoonjung Choi, Bonggun Shin, Keunsoo Kang, Sungsoo Park, and Bo Ram Beck. Target-centered drug repurposing predictions of human angiotensin-converting enzyme 2 (ace2) and transmembrane protease serine subtype 2 (tmprss2) interacting approved drugs for

coronavirus disease 2019 (covid-19) treatment through a drug-target interaction deep learning model. *Viruses*, 2020.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Proceedings of Neural Information Processing Systems (NeurIPS): Workshop*, 2014.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 2011.

Jérémy Desaphy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. sc-pdb: a 3d-database of ligandable binding sites—10 years on. *Nucleic acids research*, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Yichen Gong, Heng Luo, and Jian Zhang. Natural language inference over interaction space. *Proceedings of the International Conference on Learning Representation (ICLR)*, 2018.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 2020.

Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Cheminformatics*, 2017.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 2019.

Yoon Kim. Convolutional neural networks for sentence classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of chemical information and modeling*, 2019.

David J Lipman and William R Pearson. Rapid and sensitive protein similarity searches. *Science*, 1985.

Heng Luo, William Mattes, Donna L Mendrick, and Huixiao Hong. Molecular docking for identification of potential targets for drug repurposing. *Current topics in medicinal chemistry*, 2016.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Thin Nguyen, Hang Le, and Svetha Venkatesh. Graphdta: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv*, 2019.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 2018.

Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*, 2016.

Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of International Conference on recommender systems (RecSys)*, 2017.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Proceedings of Machine Learning for Healthcare (MLHC)*, pages 230–248. PMLR, 2019.

Bonggun Shin, Sungsoo Park, JinYeong Bak, and Joyce C Ho. Controlled molecule generator for optimizing multiple chemical properties. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 146–153, 2021.

Jean-Charles Soria, Yuichiro Ohe, Johan Vansteenkiste, Thanyanan Reungwetwattana, Busyamas Chewaskulyong, Ki Hyeong Lee, Arunee Dechaphunkul, Fumio Imamura, Naoyuki Nogami, Takayasu Kurata, et al. Osimertinib in untreated egfr-mutated advanced non–small-cell lung cancer. *New England journal of medicine*, 2018.

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 2014.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 1988.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Are deep resnets provably better than linear predictors? *Proceedings of Neural Information Processing Systems (NeurIPS)*, 32:15686–15695, 2019.

Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. Sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*, 2019.

Tianyi Zhao, Yang Hu, Linda R Valsdottir, Tianyi Zang, and Jiajie Peng. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in bioinformatics*, 2020.