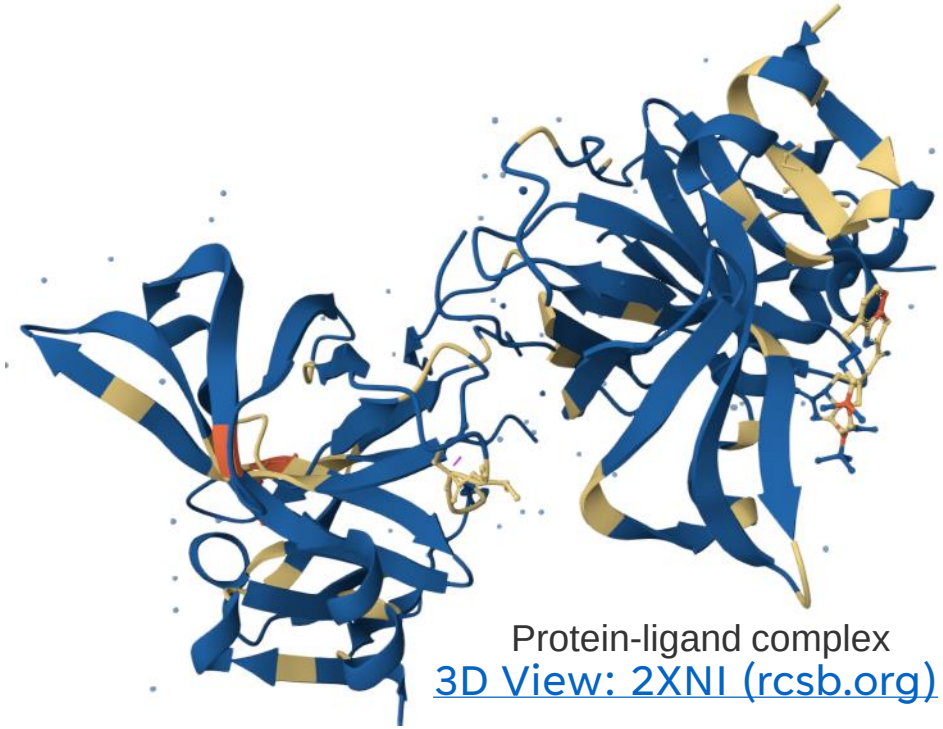
DMIS Data Mining & Information Systems Laboratory

KOREA UNIVERSITY 1905

# Perceiver CPI: a nested cross-attention network for compound–protein interaction prediction

Ngoc-Quang Nguyen, Gwanghoon Jang, Hajung Kim, Jaewoo Kang



Protein-ligand complex
3D View: 2XNI (rcsb.org)

Supervisor: Jaewoo Kang

First-author: Ngoc-Quang Nguyen

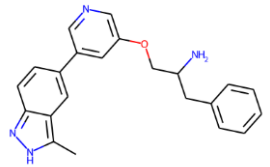
# CONTENTS


- Motivations
- Contributions
- Methods
  - Compound information encoding
  - Protein information encoding
  - Information integration
- Experiments and results
- Future works
- Q&A

# MOTIVATIONS

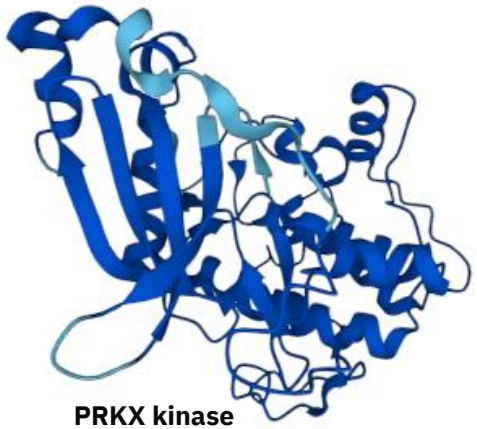
- **Biological motivations:**
  - Drug discovery is a high-cost low-efficient process.
  - Compound–protein interaction (CPI) plays an essential role in drug discovery.
  - Understanding drug–target binding affinity makes it possible to identify candidate drug.

- **Technical motivations:**
  - Extended-connectivity fingerprints (ECFPs) are informative, yet simple.
  - GNNs must always learn a meaningful chemical space embedding from scratch.
  - Integration of the compound network's and protein network's representation is often performed by a simple concatenation.



**A-674563**

"CC1=C2C=C(C=CC2=NN1)C3=CC(=CN=C3)OCC(CC4=CC=CC=C4)N"

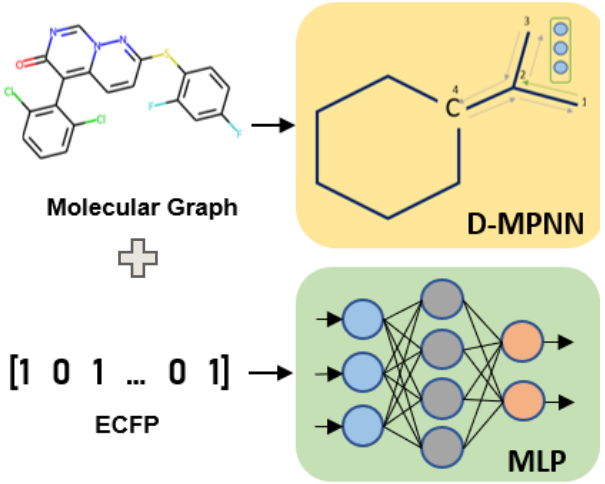


**PRKX kinase**

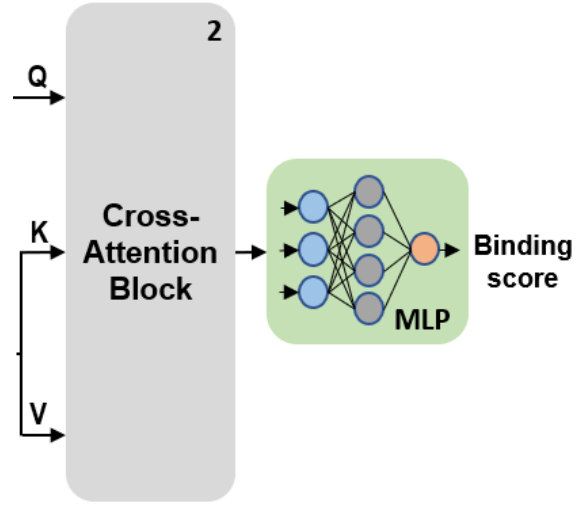
**Binding affinity (Bind? Not bind?)**

# CONTRIBUTIONS

- We propose novel method to enrich the representation of compounds by combining the information from both ECFPs and graph information.

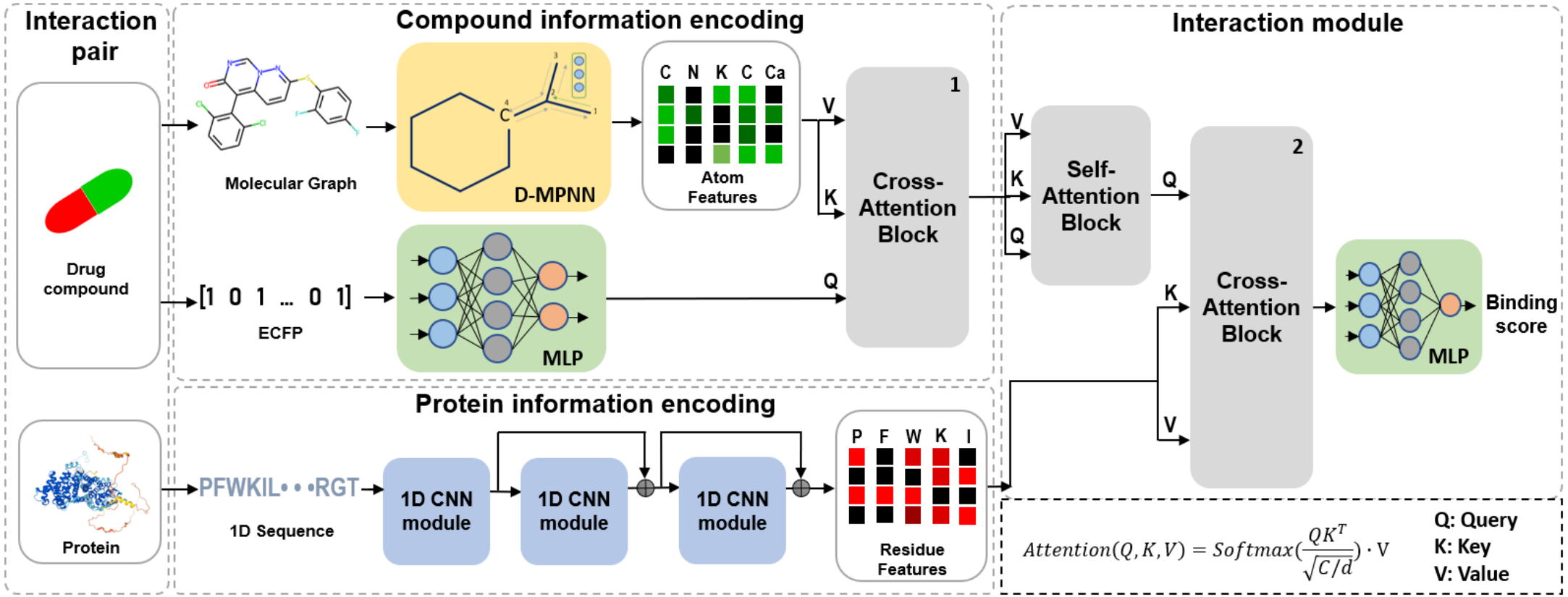



- The first approach to use nested cross-attention for capturing the relations between protein and molecule representations.



# METHODS

## *Perceiver CPI network*



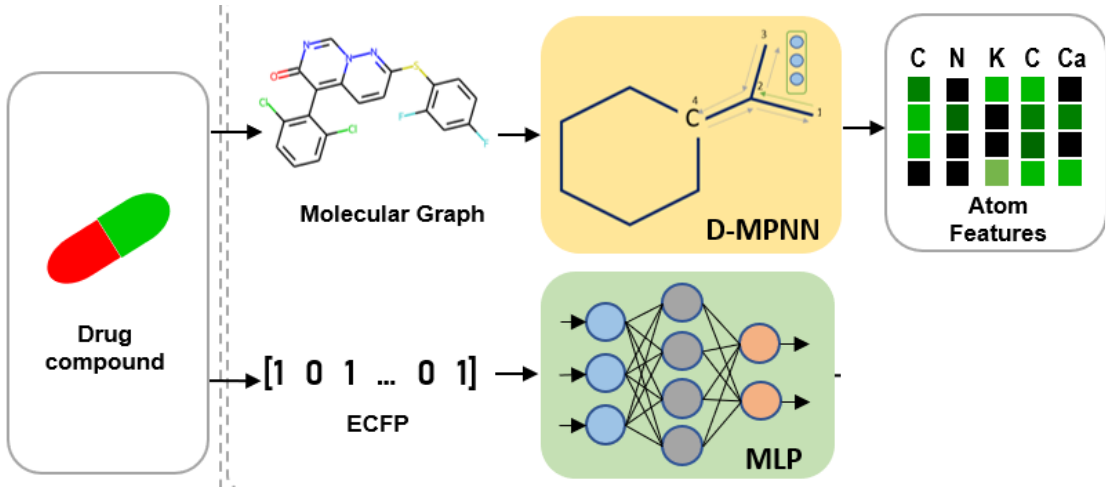
# METHODS

## *Compound information encoding*

- We represent a molecule s using two forms:

1. A molecular graph, which represents the interactions between a set of atoms by a set of bonds.

2. A Morgan/circular fingerprint vector as a binary vector.

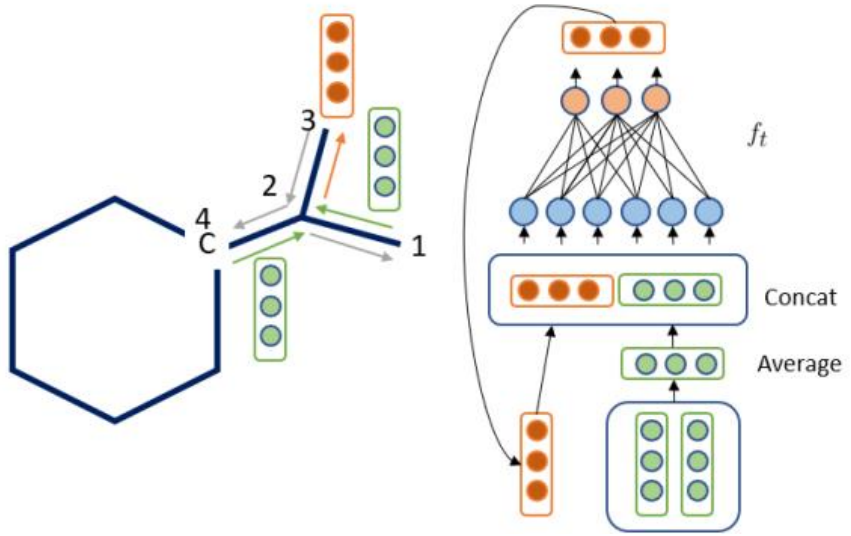



*a) Compound information encoding*

# METHODS

## *Compound information encoding Directed Message-Passing Neural Network (D-MPNN)*

- D-MPNN operates on hidden states and messages associated with directed edges (bonds) instead of messages associated with vertices (atoms).
- The main idea of the directed message-passing technique is to prevent the distortion of messages between atoms.




*a) Message passing mechanism*

*b) Hidden state update*

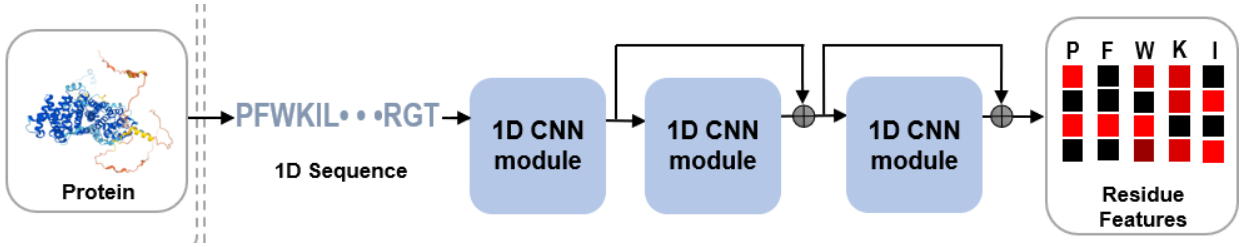
$$\begin{cases} m_{vw}^{t+1} = \sum_{k \in \{N(v) \setminus w\}} Average(x_v, x_k, h_{kv}^t) \\ h_{vw}^{t+1} = f_t(h_{vw}^t, m_{vw}^{t+1}) \end{cases}$$

*c) message-passing update equations*

# METHODS

## *Protein information encoding (1DCNN)*

- First, Tasks Assessing Protein Embeddings (TAPE) tokenizer was used to number protein characters.
- The input is zero-padded and then propagated into the blocks of 1DCNNs.
- The output of the 1D convolution block can be expressed as the **Alg. 1.**



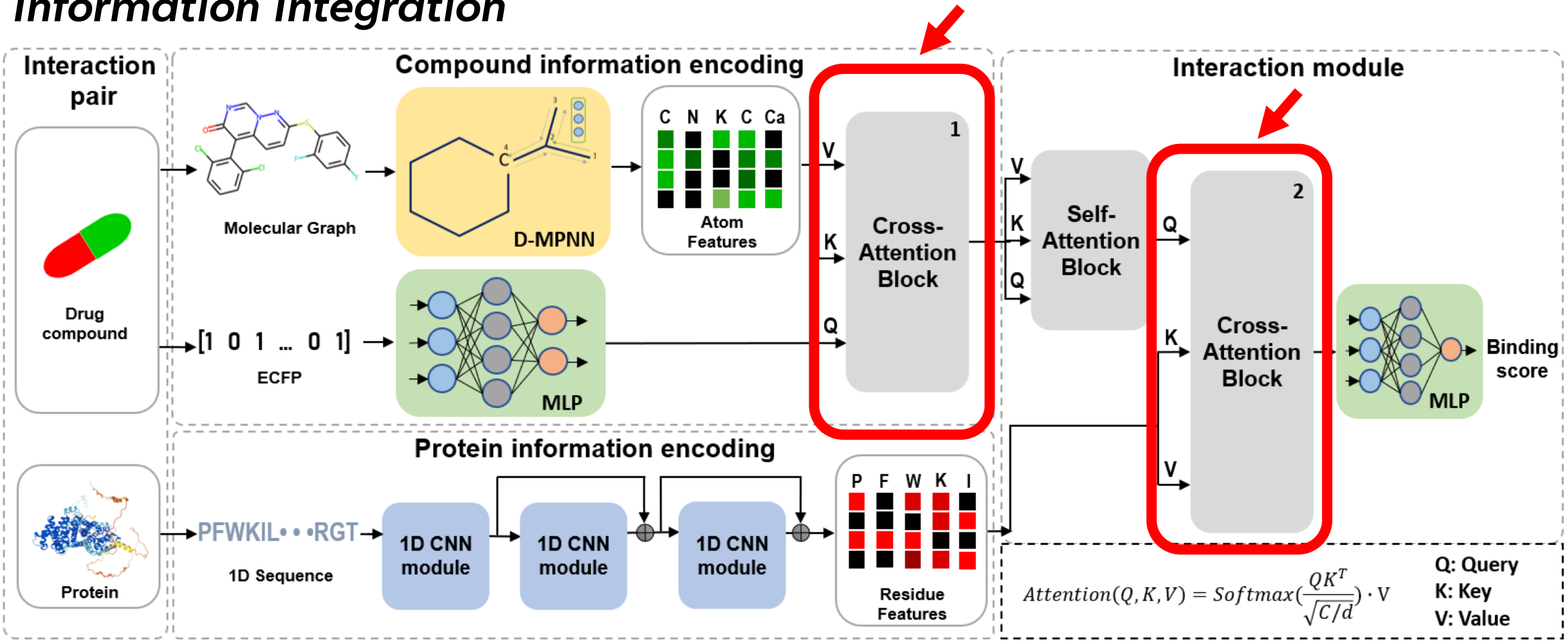

*a) Protein information encoding*

**Algorithm 1** An algorithm for residual block of 1DCNN

**Require:** $M$, $emb_{in} \leftarrow \text{Conv1D}(emb_0), \lambda$
**Result:** $Prot_t \leftarrow emb_{out}$
**for** $M$ **do**
  $emb_{out} \leftarrow \text{LN}\ (\text{Conv1D}(emb_{in}) + emb_{in} * \lambda)$;
  $emb_{in} \leftarrow \text{GLU}(emb_{out})$;
**end for**

# METHODS

## *Information Integration*

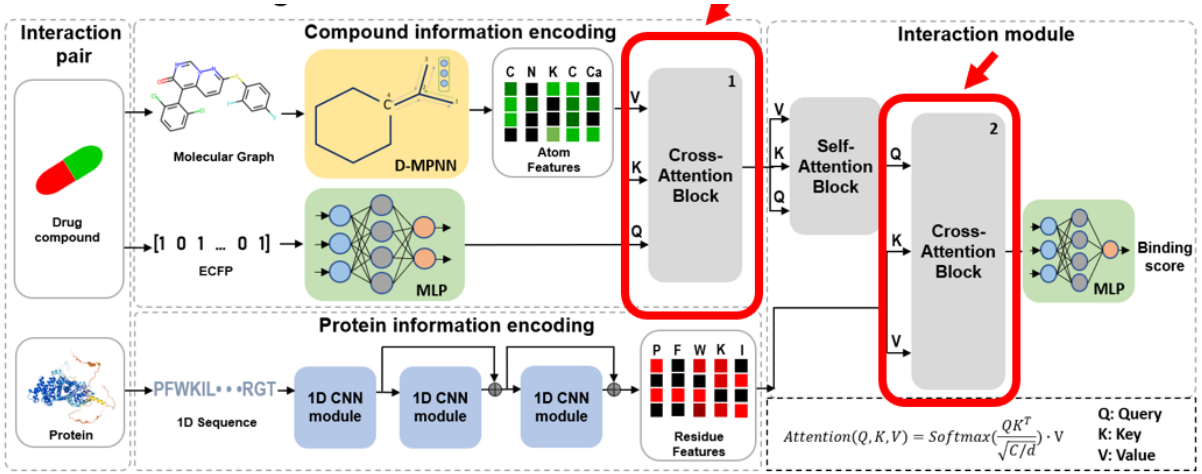


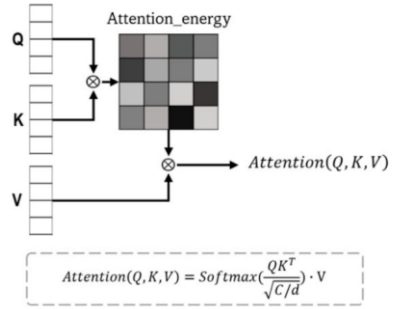
# METHODS

## *Information Integration*

- In the compound Information encoding module, we intend to take information from the EFCP to enrich the representation of compound network.

- In the interaction module, we try to capture the information which shows how the protein information affected by the compound information.




a) Cross-attention block

# EXPERIMENTS AND RESULTS

## *Experimental procedure*

- Novel pair (Davis, KIBA and Metz): There were no overlaps between the training and test datasets.
- Novel-hard pair (Davis): There were no overlaps between the training and test datasets.
- Novel compound (Davis): There were no intersections of compounds in the training set and compounds in the test set.
- Novel protein (Davis): There were no intersections of proteins in the training set and proteins in the test set.
- Cross-domain experiment (Davis and PDBbind): There were no overlaps between the training and test datasets.
- Enrichment factor analysis [GPCR, GPCR subset (DUD-E dataset), Diverse subset (DUD-E dataset)]: There were no overlaps between the training and test datasets (the duplicated target 'CXCR4' was removed from the Diverse subset).

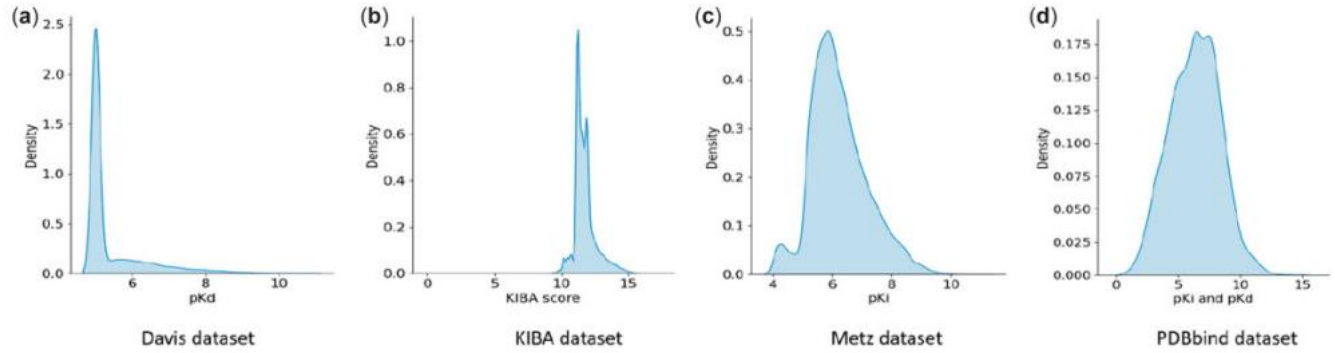



Fig. 2. Visualization of benchmark datasets with kernel density estimation

**Table 1.** Statistics of the benchmark datasets

| Dataset | Proteins | Drugs | Interactions | Density (%) |
|---|---|---|---|---|
| Davis (Davis *et al.*, 2011) | 442 | 68 | 30 056 | 100 |
| KIBA (Tang *et al.*, 2014) | 229 | 2068 | 117 657 | 24.84 |
| Metz (Metz *et al.*, 2011) | 170 | 1423 | 35 259 | 14.57 |
| PDBbind (Wang *et al.*, 2005) | 2079 | 5535 | 6989 | 0.06 |

**Table 2.** Statistic of GPCR dataset

| Proteins | Compounds | Positive pairs | Negative pairs | Density (%) |
|---|---|---|---|---|
| 356 | 5359 | 7989 | 7354 | 0.8 |

**Table 3.** Statistics of GPCR and diverse subsets from DUD-E database

| Subset | Number of target | Actives | Decoys |
|---|---|---|---|
| GPCR subset | 5 | 1480 | 99 856 |
| Diverse subset | 7 | 1759 | 107 591 |

# EXPERIMENT AND RESULTS

**Table 4.** Comparison of the proposed method with SOTA model in terms of three settings from the Davis dataset with 5-fold cross-validation

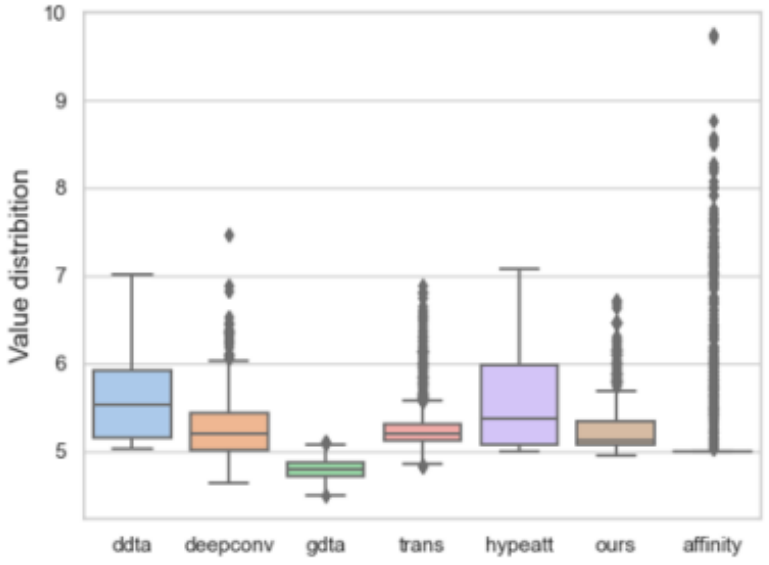
| Model | Novel pair | | Novel compound | | Novel protein | |
|---|---|---|---|---|---|---|
| | MSE | CI | MSE | CI | MSE | CI |
| DeepDTA (Öztürk *et al.*, 2018) | 0.631(±0.059) | 0.533(±0.027) | 0.482(±0.034) | 0.613(±0.029) | 0.701(±0.045) | **0.759(±0.015)** |
| DeepConvDTI (Lee *et al.*, 2019) | 0.598(±0.057) | 0.546(±0.043) | 0.512(±0.046) | 0.681(±0.012) | 0.789(±0.109) | 0.714(±0.034) |
| TransformerCPI (Chen *et al.*, 2020) | 0.549(±0.038) | 0.490(±0.032) | 0.522(±0.027) | 0.592(±0.026) | 0.708(±0.032) | 0.676(±0.005) |
| GraphDTA (GINs) (Nguyen *et al.*, 2021) | 0.846(±0.058) | 0.459(±0.032) | 0.452(±0.051) | 0.670(±0.018) | 0.970(±0.061) | 0.660(±0.016) |
| HyperattentionDTI (Zhao *et al.*, 2022) | 0.671(±0.045) | 0.517(±0.013) | 0.506(±0.015) | 0.578(±0.019) | 0.784(±0.063) | 0.674(±0.020) |
| Perceiver CPI (ours) | **0.463(±0.013)** | **0.638(±0.028)** | **0.378(±0.010)** | **0.726(±0.017)** | **0.667(±0.018)** | 0.758(±0.010) |

**Table 8.** Enrichment factor analysis results for subsets in the DUD-E database (UP: $EF_{1\%}$, DOWN: $BEDROC_{\alpha=80.5}$)

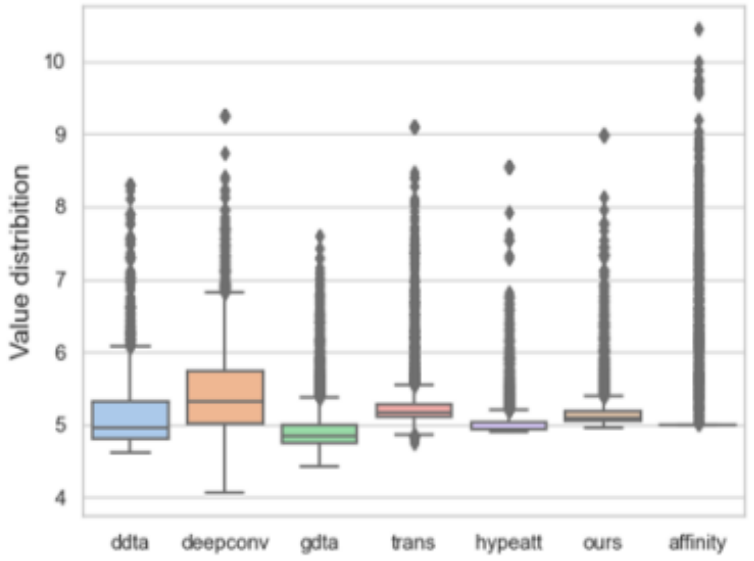
| Family | Deep ConvDTI | Transformer CPI | Hyperattention DTI | Perceiver CPI (ours) | Gold | Glide | Surflex | FlexX | Blaster |
|---|---|---|---|---|---|---|---|---|---|
| GPCR (DUD-E) | 9.728(±11.534) | 0.814(±1.178) | 3.982(±3.119) | **16.366(±15.921)** | N/a | N/a | N/a | N/a | 11.8(±8.136) |
| (DUD-E) | 0.152(±0.174) | 0.018(±0.040) | 0.071(±0.058) | 0.236(±0.177) | **0.282(±0.154)** | 0.198(±0.205) | 0.284(±0.098) | 0.156(±0.135) | N/a |
| Diverse | 0.292(±0.774) | 0.922(±0.819) | 1.075(0.876) | 1.88(±1.297) | N/a | N/a | N/a | N/a | **13.571(±12.908)** |
| (DUD-E) | 0.005(±0.015) | 0.021(0.016) | 0.023(±0.018) | 0.031(±0.022) | **0.295(±0.180)** | 0.258(±0.170) | 0.118(±0.093) | 0.104(±0.059) | N/a |

# EXPERIMENT AND RESULTS

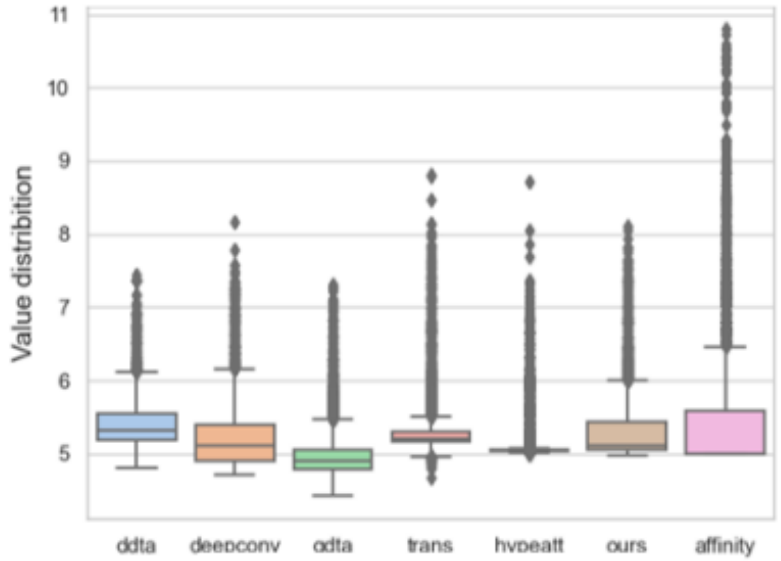
- The visualization of prediction for first folds of each setting from the Davis dataset.
- Here, we confirmed that the prediction distribution of Perceiver CPI is closely mimic to the distribution from the label.




Novel pair




Novel compound




Novel protein

# FUTURE WORKS

- Finding and extracting meaningful features from proteins remains a difficult but worthwhile task.
- Utilizing information from 3D structures produced from compounds.
- Adopting the transfer learning method for individual neural networks (compound and protein networks).
- The interpretability of Perceiver CPI might help addressing useful features which would form a valuable part of future work.

THANK YOU FOR YOUR ATTENTION

# Q&A