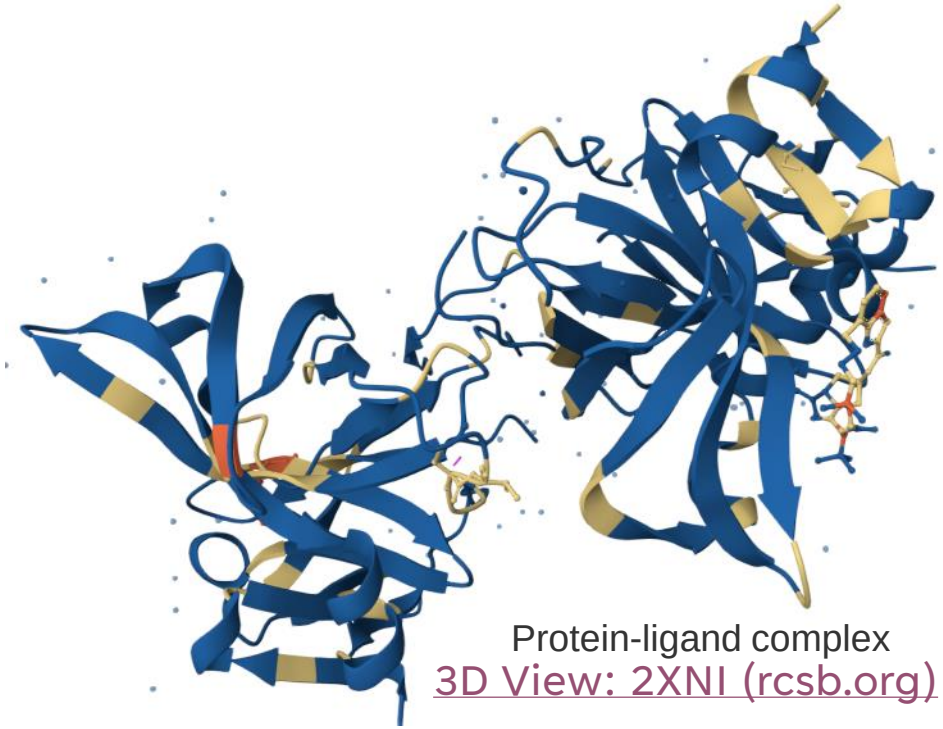
DMIS Data Mining & Information Systems Laboratory

KOREA UNIVERSITY 1905

MulinforCPI: enhancing precision of compound–protein interaction prediction through novel perspectives on multi-level information integration

Ngoc-Quang Nguyen, Sejeong Park, Mogan Gim and Jaewoo Kang



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

Supervisor: Jaewoo Kang

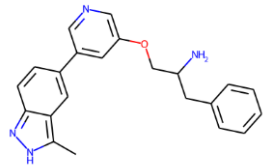
First-author: Ngoc-Quang Nguyen

CONTENTS

MOTIVATIONS

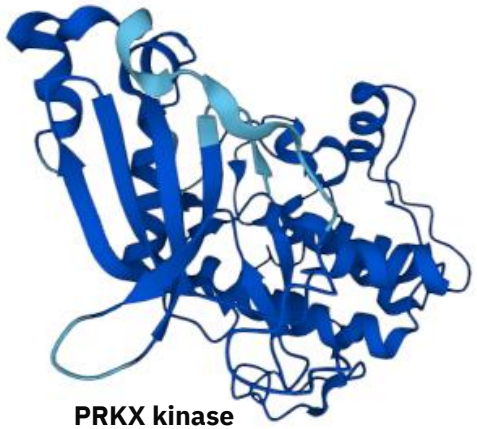
- **Biological motivations:**
  - Drug discovery is a high-cost low-efficient process.
  - Compound–protein interaction (CPI) plays an essential role in drug discovery.
  - Understanding drug–target binding affinity makes it possible to identify candidate drug.

- **Technical motivations:**
  - The constraints of previous studies reside in the utilization of plaintext to portray protein sequences.
  - Most of prior works use the information regardless to the 3D information from compounds.
  - Prior approaches have typically relied on preexisting datasets to tackle the task at hand.
  - K-folds splitting method impedes the model's capacity when confronted with substantially disparate test sets.



A-674563

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



PRKX kinase

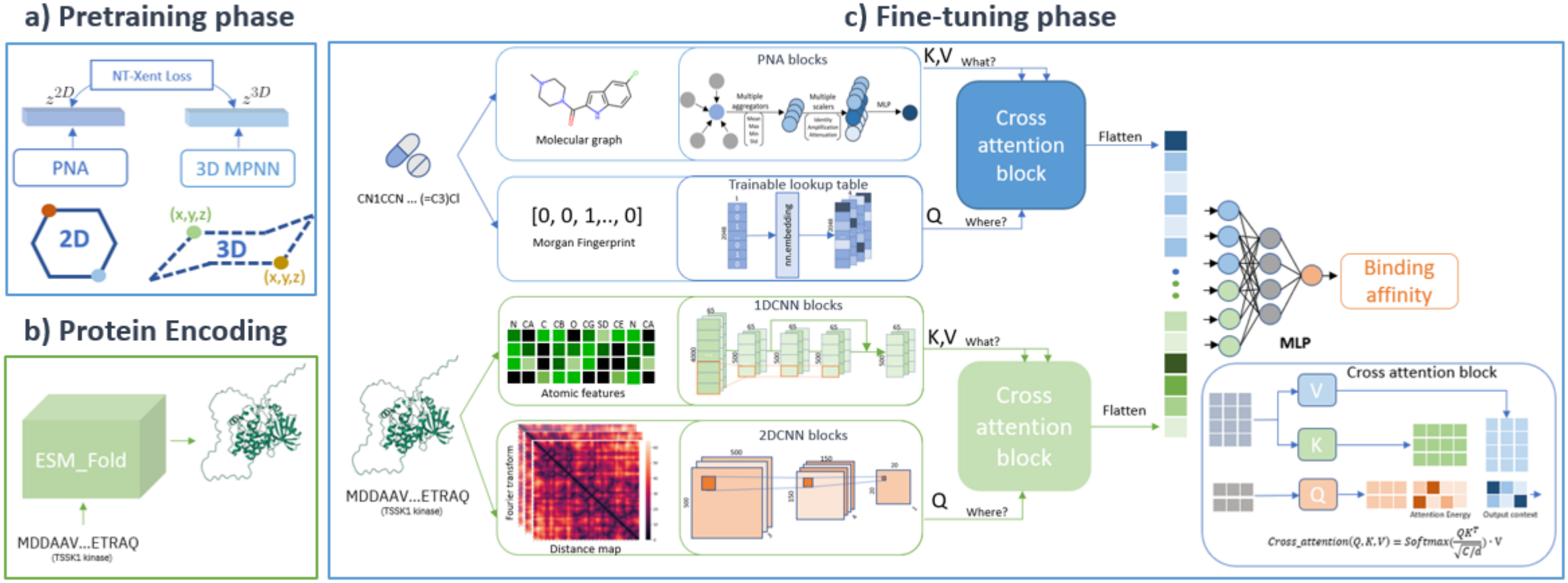
Binding affinity (Bind? Not bind?)

CONTRIBUTIONS

- We propose that the MulinforCPI DL model, which utilizes multi-level information from compounds and proteins, can address significant challenges in CPI prediction tasks.
- In contrast to prior research where most end-to-end models used sequences of amino acid characters to conduct protein representations, our approach involved leveraging both atomic-level attributes and 3D information extracted from proteins to augment the model's capacity.
- The developed transfer learning technique leverages the extensive Quantum-Mechanical Properties of Drug-like Molecules (QMugs) dataset and employs it for fine-tuning of CPI datasets.
- Our separation strategy enables the model to closely approximate the actual problem when faced with unfamiliar test sets.
- Our research reveals the gap between first-principle methods and data-driven approaches. We believe these findings open up opportunities for future research on CPI prediction tasks.

METHODS

Mulinfor CPI network



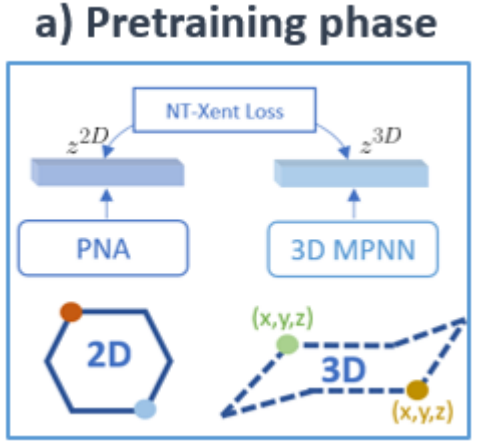
METHODS

Pretraining phase

- It is unachievable to procure 3D configurations in all practical scenarios.
- To tackle this issue, we follow the suggestion of training strategy proposed by 3Dinfomax.
  - 3D molecular dataset QMugs (Quantum Mechanical Properties of Drug-like Molecules) [REF] is used for the pre-training purpose, resulting in GNN 3D geometry information aware.
  - By minimizing the contrastive learning loss LNT-Xent (normalized temperature-scaled cross entropy loss)

Table 1 Descriptive statistics of QMugs dataset.

Dataset	Unique compounds	Total conformations	Heavy atoms max (mean)
QMugs	665,911	1,992,984	100 (30.6)



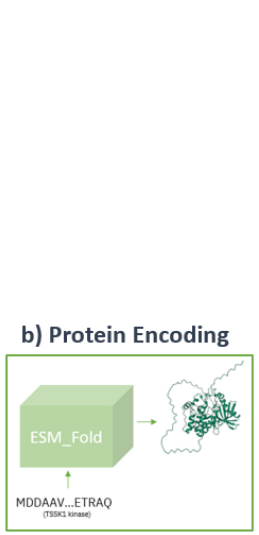
$$\mathbf{L}_{NT-Xent} = \frac{1}{N}\sum_{i=0}^{N}\left[log\frac{\sum_{j=1}^{c} e^{sim(z_i^{2D}, z_{i,j}^{3D})/\tau}}{\sum_{\substack{k=1 \\ k\neq i}}^{N}\sum_{j=1}^{c} e^{sim(z_i^{2D}, z_{k,j}^{3D})/\tau}}\right]$$

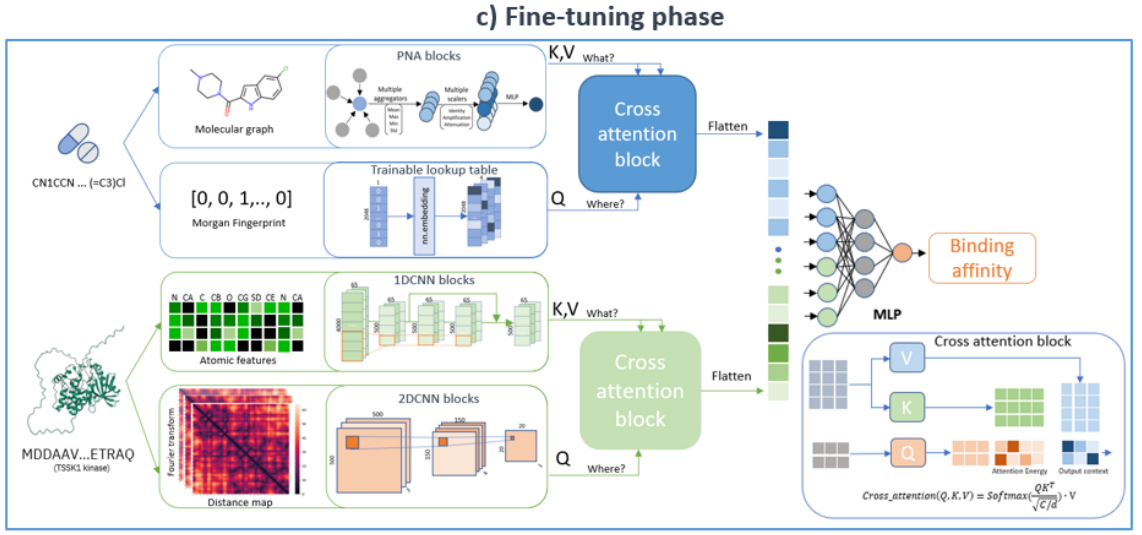
[REF] Isert, Clemens, et al. "QMugs, quantum mechanical properties of drug-like molecules." *Scientific Data* 9.1 (2022): 273.

METHODS

Fine-tuning phase

- Compound work
  - First, we have access to GNNs that can effectively tackle the question of “what” to learn from the molecular graphical structure.
  - Second, to augment the model’s capacity to incorporate the global information of the molecular structure, we employ Morgan fingerprints (MFs).
- Protein work
  - First, we extract residue-residue Euclidean distance information obtained from the interatomic alpha carbon (α-carbon or Cα) coordinates.
  - Second, we utilize the information that discloses atomic properties, including the specific type of atom in a given residue).





We measure a distance by applying:

$$d_{i,j} = \sqrt[2]{\sum_{c=1}^{3}(x_c - y_c)^2}$$

Using Fourier feature mapping function:

$$\gamma(d_{i,j}) = [d_{i,j}, \frac{sin(d_{i,j})}{2^0}, \frac{cos(d_{i,j})}{2^0}, ..., \frac{cos(d_{i,j})}{2^{F-1}}]$$

[REF] Stärk, Hannes, et al. "3d infomax improves gnns for molecular property prediction." *International Conference on Machine Learning*. PMLR, 2022.

DATASETS

Our study involves 6 benchmark datasets.

For 2 tasks:

Regression :

Davis, KIBA, Metz, CASF2016

Classification:

BindingDB, DUD_E Diverse subset

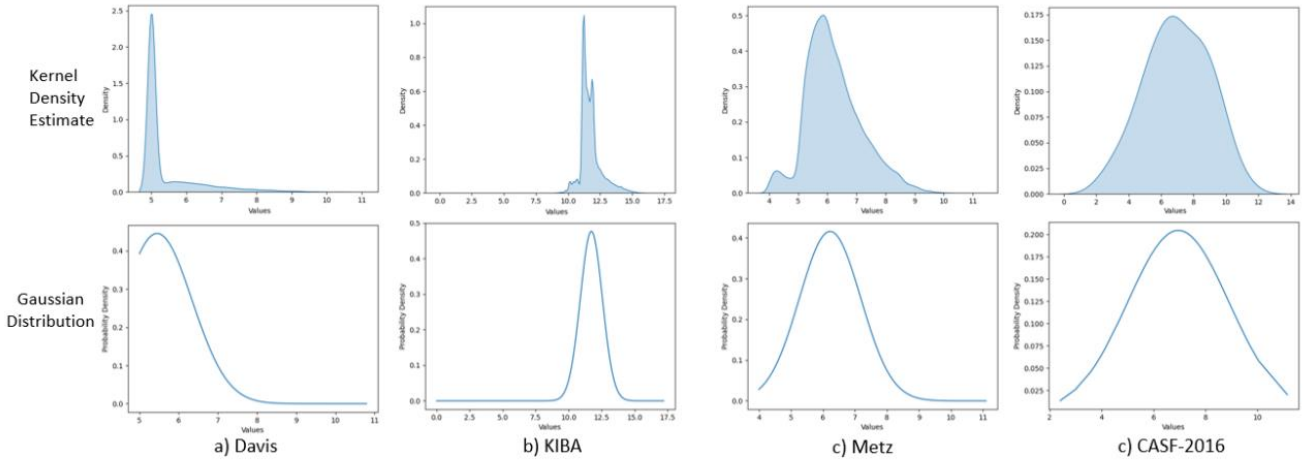


Fig. 1 The label distribution of four regression datasets. a) KIBA dataset, b) Davis dataset, c) Metz dataset and d) CASF-2016 dataset.

Table 2 Statistics of the benchmark datasets.

Dataset	Task	Proteins	Drugs	Interactions		Density (%)
				Negatives	Positives	
Davis	Regression	442	68	30,056		100
KIBA	Regression	229	2,068	117,657		24,84
Metz	Regression	170	1,423	35,259		14,57
CASF2016	Regression	15	57	57		6.6
DUD_E Diverse	Classification	7	108,212	107,590	1,759	14,43
BindingDB	Classification	813	49,752	27,493	33,777	0,15

EXPERIMENTAL SETTINGS

- Novel pair (Davis): There were no overlaps between the training and test datasets. Neither the training compound nor the training protein appeared in the test set.
- 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.
- Novel hard pair (Metz, KIBA): We removed interactions from the training dataset if either the protein sequence or the compound had a similarity score exceeding the threshold 0.3
- Cross-domain (Metz, CASF-2016): We removed interactions involving 56 proteins and 105 compounds with similarities higher than 0.3 from the Metz dataset.
- Enrichment factor analysis (BingdingDB, DUD_E diverse): we removed interactions for two proteins and compounds that appeared in both datasets (GCR HUMAN (P04150) and AKT1 HUMAN (P31749) and 102 compounds) from training set.

EXPERIMENT AND RESULTS

All of these models exhibited notably low Spearman correlation values:

- A low Spearman correlation value suggests that these models fail to capture features from training dataset and may not offer a reliable fit.
- The 3DinforCPI model shows robustness in its ability to learn from training datasets, consistently achieving the highest R-squared values across the majority of experiments.

Table 3: Restult for novel-pair in Davis dataset (MSE ↓ better, CI ↑ better, Spearman Correlation ↑ better, mean and standard deviation values were computed from 5-fold results' averages).

Models	MSE	CI	Spearman Correlation
DeepDTA	0.719(±0.312)	0.456(±0.107)	−0.054(±0.162)
DeepConvDTI	0.602(±0.221)	0.580(±0.065)	0.141(±0.105)
TransformerCPI	0.565(±0.252)	0.552(±0.024)	0.087(±0.037)
GraphDTA (GINs)	1.078(±0.564)	0.499(±0.100)	0.011(±0.139)
HyperattentionDTI	0.633(±0.249)	0.529(±0.046)	0.049(±0.078)
PerceiverCPI	0.668(±0.357)	0.547(±0.071)	0.062(±0.124)
MulinforCPI (ours)	**0.547(±0.256)**	**0.646(±0.05)**	**0.237(±0.061)**
MulinforCPI (ours) Freeze 95%	0.580(±0.258)	0.528(±0.073	0.055(±0.093)

Table 4: Result for novel-comp in Davis dataset (MSE ↓ better, CI ↑ better, Spearman Correlation ↑ better, mean and standard deviation values were computed from 5-fold results' averages).

Models	MSE	CI	Spearman correlation
DeepDTA	0.873(±0.274)	0.549(±0.036)	0.086(±0.068)
DeepConvDTI	0.750(±0.275)	0.674(±0.048)	0.312(±0.075)
TransformerCPI	0.831(±0.244)	0.615(±0.039)	0.205(±0.051)
GraphDTA (GINs)	0.750(±0.283)	0.688(±0.05)	**0.333(±0.062)**
HyperattentionDTI	0.757(±0.269)	0.589(±0.057)	0.157(±0.104)
PerceiverCPI	0.746(±0.245)	0.669(±0.036)	0.303(±0.054)
MulinforCPI (ours)	0.690(±0.275)	0.679(±0.072)	0.317(±0.113)
MulinforCPI (ours) Freeze 95%	**0.679(±0.219)**	**0.688(±0.028)**	0.290(±0.084)

Table 5: Result for novel-prot in Davis dataset (MSE ↓ better, CI ↑ better, Spearman Correlation ↑ better, mean and standard deviation values were computed from 5-fold results' averages).

Models	MSE	CI	Spearman correlation
DeepDTA	0.529(±0.130)	0.729(±0.014)	0.396(±0.031)
DeepConvDTI	**0.465(±0.151)**	0.755(±0.062)	0.433(±0.094)
TransformerCPI	0.487(±0.172)	0.660(±0.040)	0.278(±0.066)
GraphDTA (GINs)	1.122(±0.887)	0.694(±0.051)	0.333(±0.088)
HyperattentionDTI	0.542(±0.219)	0.707(±0.040)	0.352(±0.044)
PerceiverCPI	0.513(±0.213)	0.748(±0.022)	0.427(±0.033)
MulinforCPI (ours)	0.488(±0.138)	**0.756(±0.017)**	**0.439(±0.022)**
MulinforCPI (ours) Freeze 95%	0.478(±0.140)	0.753(±0.020)	0.435(±0.027)

EXPERIMENT AND RESULTS

- Comparing with others, MulinforCPI shows the better trend in the scatter plots from Davis experiment.
- MulinforCPI explicitly outperforms in terms of metric measurement

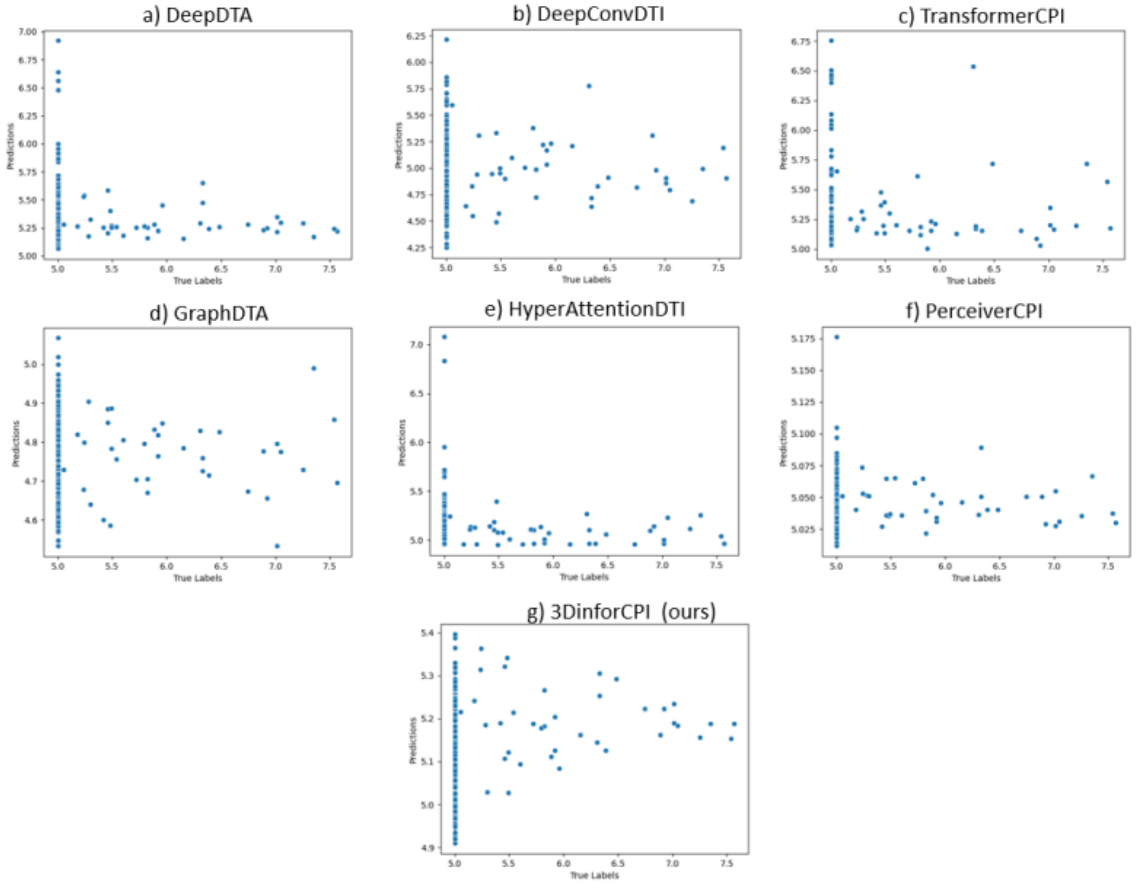


Figure 1: The scatter plot for the third fold in novel pair setting.

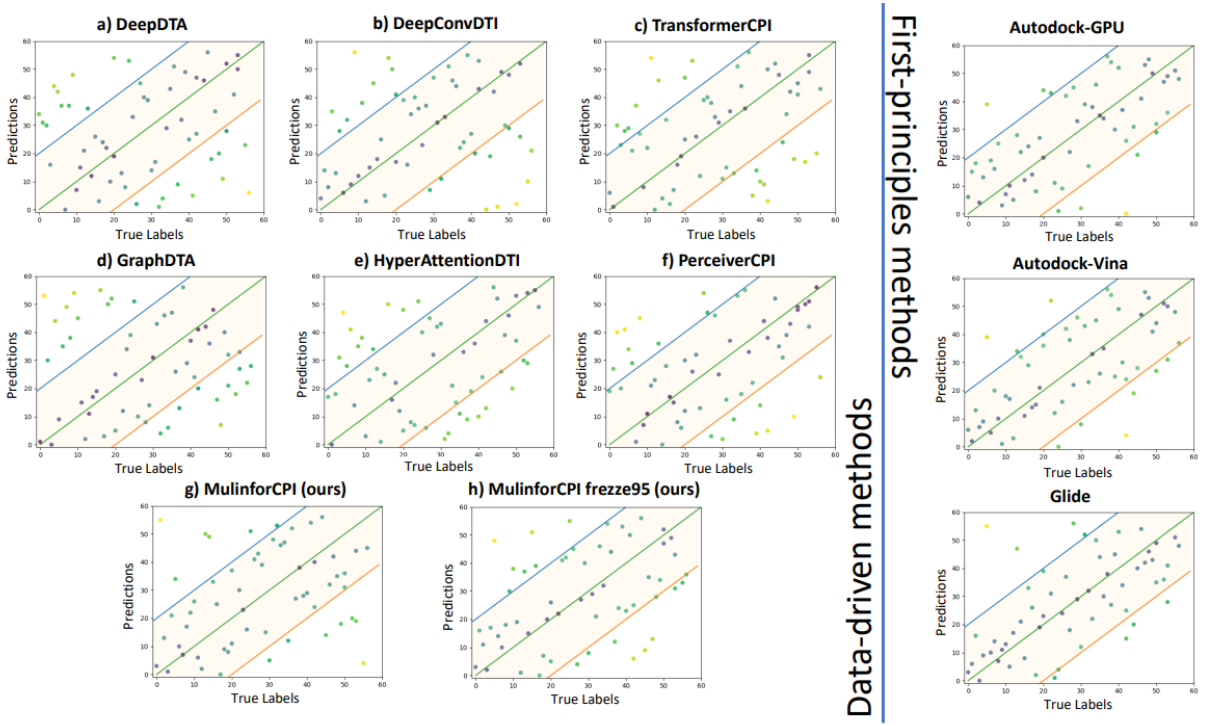
EXPERIMENT AND RESULTS

Result for Cross-domain

- This experiment:
  - Train with Metz dataset.
  - Test with 57 interactions from CAFS-2016.
  - (Interactions involving 56 proteins and 105 compounds with similarities higher than 0.3 were removed from training dataset.)
- The performance of end-to-end models is increased when we low down the hard of threshold.

Table 6: The results cross-domain experiments when similarity threshold = 0.3 (MSE ↓ better, CI ↑ better, Spearman Correlation ↑ better).

Model	MSE	CI	Spearman correlation
DeepDTA	6.193	0.542	0.135
DeepConvDTI	6.611	0.562	0.176
TransformerCPI	4.999	0.6	0.298
GraphDTA (GINs)	6.676	0.512	0.02
HyperattentionDTI	5.484	0.606	0.314
PerceiverCPI	5.279	0.615	0.342
MulinforCPI (ours)	4.698	0.602	0.297
MulinforCPI (ours) Freeze 95%	**4.391**	0.642	0.395
Autodock-GPU	N/A	0.717	**0.620**
Autodock-Vina	N/A	0.711	0.608
Glide	N/A	**0.722**	0.614



EXPERIMENT AND RESULTS

The performance of data-driven method has not reached to the level of strong docking simulations such as Glide and Gold.

Table 7: The enrichment factor analysis results on a Diverse subset from the DUD-E database ($EF_{1\%}\uparrow$ better, $BEDROC_{\alpha=80.5}\uparrow$ better, mean and standard deviation values were computed from per protein results' averages).

Models	$EF_{1\%}$ (±std)	$BEDROC_{\alpha=80.5}$ (±std)
DeepConvDTI	6.357(±6.173)	0.118(±0.109)
TransformerCPI	7.039(±12.496)	0.117(±0.192)
HyperattentionDTI	1.753(±2.551)	0.038(±0.051)
PerceiverCPI	4.649(±3.136)	0.094(±0.067)
MulinforCPI (ours)	7.886(±10.642)	0.137(±0.167)
MulinforCPI (ours) Freeze 95%	4.248(±5.787)	0.078(±0.095)
Random Guessing	0.940(±0.844)	0.022(±0.010)
Gold	N/A	0.253(±0.182)
Glide	N/A	**0.259(±0.171)**
Surflex	N/A	0.119(±0.093)
FlexX	N/A	0.104(±0.060)
Blaster	**13.571(±12.908)**	N/A

FUTURE WORKS

- Based on the data obtained from ESMFold, MulinforCPI requires a substantial amount of memory for preprocessing before proceeding to GPU training. Enhancing the input while maintaining optimal performance can accelerate the training process.
- The interpretability of our DL network is constrained by the dimensionality reduction of the CNNs and the MLP layers. Addressing these significant characteristics will form an integral part of future endeavors.
- Leveraging equivariant networks, such as E(n) Equivariant GNNs [41] and Euclidean Neural Networks [42], to incorporate positional information (rotation, translation, inversion) has the potential to enhance the model's capacity to capture more informative patterns.

THANK YOU FOR YOUR ATTENTION

Q&A