



This is a repository copy of *Interpretable bilinear attention network with domain adaptation improves drug-target prediction*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/195230/>

Version: Accepted Version

Article:

Bai, P., Miljković, F., John, B. et al. (1 more author) (2023) Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence*, 5 (2). pp. 126-136. ISSN 2522-5839

<https://doi.org/10.1038/s42256-022-00605-1>

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <http://dx.doi.org/10.1038/s42256-022-00605-1>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Interpretable bilinear attention network with domain adaptation improves drug-target prediction

Peizhen Bai¹, Filip Miljković², Bino John³, and Haiping Lu^{1*}

¹Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

²Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

³Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, USA

*corresponding author: Haiping Lu (h.lu@sheffield.ac.uk)

ABSTRACT

Predicting drug-target interaction is key for drug discovery. Recent deep learning-based methods show promising performance but two challenges remain: (i) how to explicitly model and learn local interactions between drugs and targets for better prediction and interpretation; (ii) how to generalize prediction performance on novel drug-target pairs from different distribution. In this work, we propose DrugBAN, a deep bilinear attention network (BAN) framework with domain adaptation to explicitly learn pair-wise local interactions between drugs and targets, and adapt on out-of-distribution data. DrugBAN works on drug molecular graphs and target protein sequences to perform prediction, with conditional domain adversarial learning to align learned interaction representations across different distributions for better generalization on novel drug-target pairs. Experiments on three benchmark datasets under both in-domain and cross-domain settings show that DrugBAN achieves the best overall performance against five state-of-the-art baselines. Moreover, visualizing the learned bilinear attention map provides interpretable insights from prediction results.

Drug-target interaction (DTI) prediction serves as an important step in the process of drug discovery¹⁻³. Traditional biomedical measuring from *in vitro* experiments is reliable but has notably high cost and time-consuming development cycle, preventing its application on large-scale data⁴. In contrast, identifying high-confidence DTI pairs by *in silico* approaches can greatly narrow down the search scope of compound candidates, and provide insights into the causes of potential side effects in drug combinations. Therefore, *in silico* approaches have gained increasing attention and made much progress in the last few years^{5,6}.

For *in silico* approaches, traditional structure-based and ligand-based virtual screening (VS) methods have been studied widely for their decent performance⁷. However, structure-based VS requires molecular docking simulation, which is not applicable if the target protein's three-dimensional (3D) structure is unknown. On the other hand, ligand-based VS predicts new active molecules based on the known actives to the same protein, but the performance is poor when the number of known actives is insufficient⁸.

More recently, deep learning (DL)-based approaches have rapidly progressed for computational DTI prediction due to their successes in other areas, enabling large-scale validation in a relatively short time⁹. Many of them are constructed from a chemogenomics perspective^{3,10}, which integrates the chemical space, genomic space, and interaction information into a unified end-to-end framework. Since the number of biological targets that have available 3D structures is limited, many DL-based models take linear or two-dimensional (2D) structural information of drugs and proteins as inputs. They treat DTI prediction as a binary classification task, and make predictions by feeding the inputs into different deep encoding and decoding modules such as deep neural network (DNN)^{11,12}, graph neural network (GNN)^{9,13-15} or transformer architectures^{16,17}. With the advances of deep learning techniques, such models can automatically learn data-driven representations of drugs and proteins from large-scale DTI data instead of only using pre-defined descriptors.

Despite these promising developments, two challenges remain in existing DL-based methods. The first challenge is explicit learning of interactions between local structures of drug and protein. DTI is essentially decided by mutual effects between important molecular substructures in the drug compound and binding sites in the protein sequence¹⁸. However, many previous studies learn global representations in their separate encoders, without explicitly learning local interactions^{2,11,13,19,20}. Consequently, drug and protein representations are learned for the whole structures first and mutual information is only implicitly learned in the black-box decoding module. Interactions between drug and target are particularly related to their crucial substructures, thus separate global representation learning tends to limit the modeling capacity and prediction performance. Moreover, without explicit learning of local interactions, the prediction result is hard to interpret, even if the prediction is accurate.

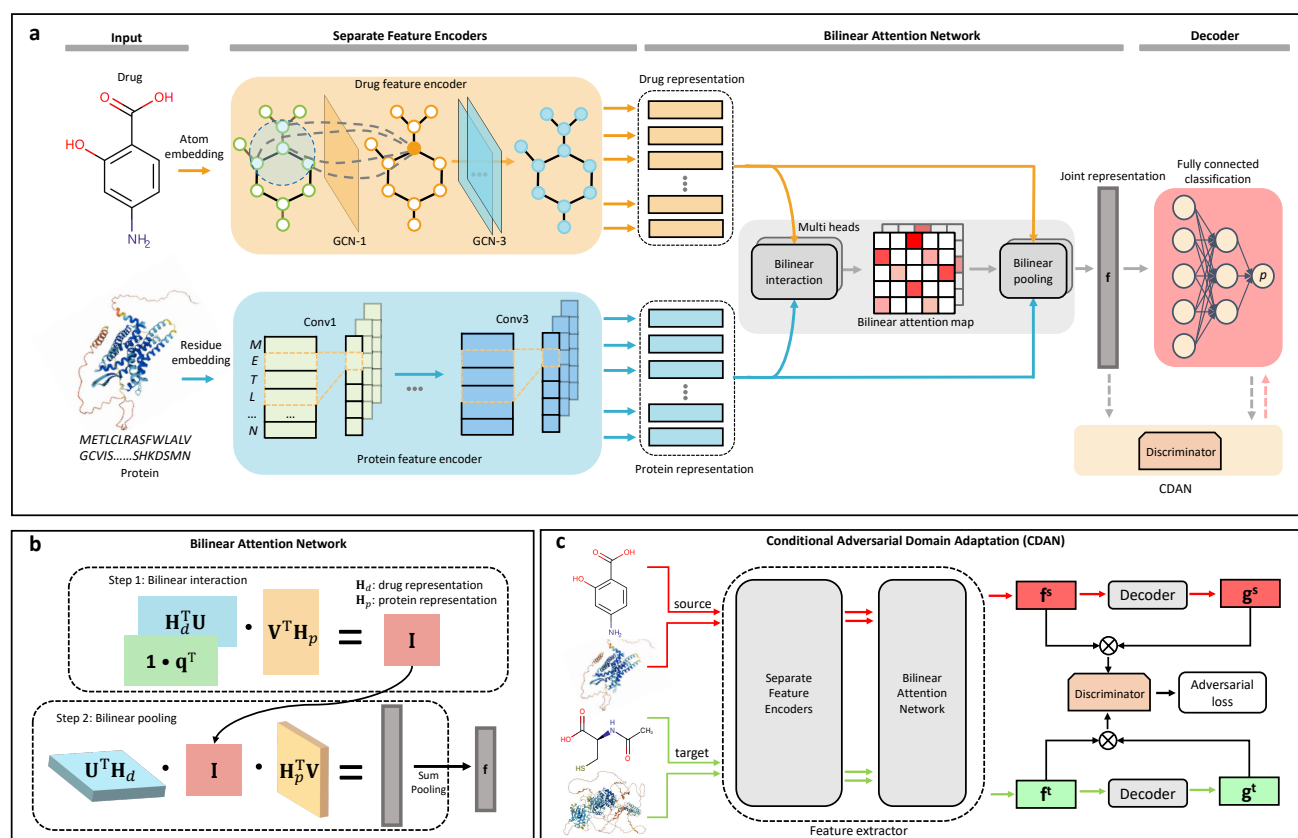


Figure 1. Overview of the DrugBAN framework. (a) The input drug molecule and protein sequence are separately encoded by graph convolutional networks and 1D-convolutional neural networks. Each row of the encoded drug representation is an aggregated representation of adjacent atoms in the drug molecule, and each row of the encoded protein representation is a subsequence representation in the protein sequence. The drug and protein representations are fed into a bilinear attention network to learn their pairwise local interactions. The joint representation \mathbf{f} is decoded by a fully connected decoder module to predict the DTI probability p . If the prediction task is cross-domain, the conditional domain adversarial network²¹ (CDAN) module is employed to align learned representations in the source and target domains. (b) The bilinear attention network architecture. \mathbf{H}_d and \mathbf{H}_p are encoded drug and protein representations. In Step 1, the bilinear attention map matrix \mathbf{I} is obtained by a low-rank bilinear interaction modeling via transformation matrices \mathbf{U} and \mathbf{V} to measure the substructure-level interaction intensity²². Then \mathbf{I} is utilized to produce the joint representation \mathbf{f} in Step 2 by bilinear pooling via the shared transformation matrices \mathbf{U} and \mathbf{V} . (c) CDAN is a domain adaptation technique to reduce the domain shift between different distributions of data. We use CDAN to embed joint representation \mathbf{f} and softmax logits \mathbf{g} for source and target domains into a joint conditional representation via the discriminator, a two-layer fully connected network that minimizes the domain classification error to distinguish the target domain from the source domain.

39 The second challenge is generalizing prediction performance across domains, i.e. out of learned distribution. Due to the
 40 vast regions of chemical and genomic space, drug-target pairs that need to be predicted in real-world applications are often
 41 unseen and dissimilar to any pairs in the training data. They have different distributions and thus need cross-domain modeling.
 42 A robust model should be able to transfer learned knowledge to a new domain that only has unlabeled data. In this case, we
 43 need to align distributions and improve cross-domain generalization performance by learning transferable representations, e.g.
 44 from "source" to "target". To the best of our knowledge, this is an underexplored direction in drug discovery.

45 To address these challenges, we propose an interpretable bilinear attention network-based model (DrugBAN) for DTI
 46 prediction, as shown in Figure 1a. DrugBAN is a deep learning framework with explicit learning of local interactions between
 47 drug and target, and conditional domain adaptation for learning transferable representations across domains. Specifically, we
 48 first use graph convolutional network²³ (GCN) and convolutional neural network (CNN) to encode local structures in 2D
 49 drug molecular graph and 1D protein sequence, respectively. Then the encoded local representations are fed into a pairwise

50 interaction module that consists of a bilinear attention network^{24,25} to learn local interaction representations, as depicted in
51 Figure 1b. The local joint interaction representations are decoded by a fully connected layer to make a DTI prediction. In this
52 way, we can utilize the pairwise bilinear attention map to visualize the contribution of each substructure to the final predictive
53 result, improving the interpretability. For cross-domain prediction, we apply conditional domain adversarial network²¹ (CDAN)
54 to transfer learned knowledge from source domain to target domain to enhance cross-domain generalization, as illustrated in
55 Figure 1c. We conduct a comprehensive performance comparison against five state-of-the-art DTI prediction methods on both
56 in-domain and cross-domain settings of drug discovery. The results show that our method achieves the best overall performance
57 compared to state-of-the-art methods, while providing interpretable insights for the prediction results.

58 To summarize, DrugBAN differs from previous works by (i) capturing pairwise local interactions between drugs and targets
59 via a bilinear attention mechanism, (ii) enhancing cross-domain generalization with an adversarial domain adaptation approach;
60 and (iii) giving an interpretable prediction via bilinear attention weights instead of black-box results.

61 Results

62 Problem formulation

63 In DTI prediction, the task is to determine whether a pair of a drug compound and a target protein will interact. For target
64 protein, denoting each protein sequence as $\mathcal{P} = (a_1, \dots, a_n)$, where each token a_i represents one of the 23 amino acids. For
65 drug compound, most existing deep learning-based methods represent the input by the Simplified Molecular Input Line Entry
66 System (SMILES)²⁶, which is a 1D sequence describing chemical atom and bond token information in the drug molecule.
67 The SMILES format allows encoding drug information with many classic deep learning architectures. However, since the 1D
68 sequence is not a natural representation for molecules, some important structural information of drugs could be lost, degrading
69 model prediction performance. Our model converts input SMILES into its corresponding 2D molecular graph. Specifically, a
70 drug molecule graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of vertices (atoms) and \mathcal{E} is the set of edges (chemical bonds).

71 Given a protein sequence \mathcal{P} and a drug molecular graph \mathcal{G} , DTI prediction aims to learn a model \mathcal{M} to map the joint feature
72 representation space $\mathcal{P} \times \mathcal{G}$ to an interaction probability score $p \in [0, 1]$. Supplementary Table 3 provides the commonly used
73 notations in this paper.

74 DrugBAN framework

75 Figure 1a shows the proposed DrugBAN framework. Given an input drug-target pair, firstly, we employ separate graph
76 convolutional network (GCN) and 1D-convolutional neural network (1D-CNN) blocks to encode molecular graph and protein
77 sequence information, respectively. Then we use a bilinear attention network module to learn local interactions between encoded
78 drug and protein representations. The bilinear attention network consists of a bilinear attention step and a bilinear pooling step
79 to generate a joint representation, as illustrated in Figure 1b. Next, a fully connected classification layer learns a predictive
80 score indicating the probability of interaction. For improving model generalization performance on cross-domain drug-target
81 pairs, we further embed CDAN into the framework to adapt representations for better aligning source and target distributions,
82 as depicted in Figure 1c.

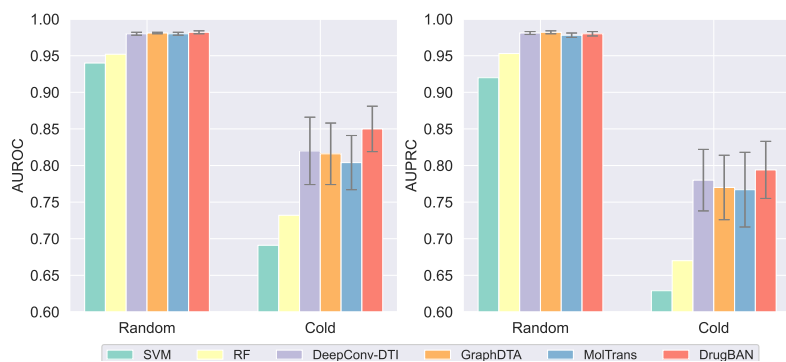
83 Evaluation strategies and metrics

84 We study classification performance on three public datasets separately: BindingDB²⁷, BioSNAP²⁸ and Human^{16,29}, with test
85 sets holding out as ‘unknown’ for evaluation. We use two different split strategies for in-domain and cross-domain settings. For
86 in-domain evaluation, each experimental dataset is randomly divided into training, validation, and test sets with a 7:1:2 ratio.
87 For cross-domain evaluation, we propose a clustering-based pair split strategy to construct cross-domain scenario. We conduct
88 cross-domain evaluation on the large-scale BindingDB and BioSNAP datasets. For each dataset, we firstly use the single-linkage
89 algorithm to cluster drugs and proteins by ECFP4 (extended connectivity fingerprint, up to four bonds)³⁰ fingerprint and pseudo
90 amino acid composition (PSC)³¹, respectively. After that, we randomly select 60% drug clusters and 60% protein clusters
91 from the clustering result, and consider all drug-target pairs between the selected drugs and proteins as source domain data.
92 All the pairs between drugs and proteins in the remaining clusters are considered to be target domain data. The clustering
93 implementation details are provided in Supplementary Section 1. Under the clustering-based pair split strategy, the source and
94 target domains are non-overlapping with different distributions. Following the general setting of domain adaptation, we use all
95 labeled source domain data and 80% unlabeled target domain data as the training set, and the remaining 20% labeled target
96 domain data as the test set. The cross-domain evaluation is more challenging than in-domain random split but provides a better
97 measure of model generalization ability in real-world drug discovery. For a more comprehensive study, we report additional
98 experiments across different protein families, on unseen drugs/targets, and with high fraction of missing data in Supplementary
99 Sections 4-6, respectively.

100 The AUROC (area under the receiver operating characteristic curve) and AUPRC (area under the precision-recall curve) are
101 used as the major metrics to evaluate model classification performance. In addition, we also report the accuracy, sensitivity, and

Table 1. In-domain performance comparison on the BindingDB and BioSNAP datasets with random split (**Best**, **Second Best**).

Method	AUROC	AUPRC	Accuracy	Sensitivity	Specificity
BindingDB					
SVM ³²	0.939±0.001	0.928±0.002	0.825±0.004	0.781±0.014	0.886±0.012
RF ³³	0.942±0.011	0.921±0.016	0.880±0.012	0.875±0.023	0.892±0.020
DeepConv-DTI ¹¹	0.945±0.002	0.925±0.005	0.882±0.007	0.873±0.018	0.894±0.009
GraphDTA ¹³	0.951±0.002	0.934±0.002	0.888±0.005	0.882±0.012	0.897±0.008
MolTrans ¹⁷	0.952±0.002	0.936±0.001	0.887±0.006	0.877±0.016	0.902±0.009
DrugBAN	0.960±0.001	0.948±0.002	0.904±0.004	0.900±0.008	0.908±0.004
BioSNAP					
SVM ³²	0.862±0.007	0.864±0.004	0.777±0.011	0.711±0.042	0.841±0.028
RF ³³	0.860±0.005	0.886±0.005	0.804±0.005	0.823±0.032	0.786±0.025
DeepConv-DTI ¹¹	0.886±0.006	0.890±0.006	0.805±0.009	0.760±0.029	0.851±0.013
GraphDTA ¹³	0.887±0.008	0.890±0.007	0.800±0.007	0.745±0.032	0.854±0.025
MolTrans ¹⁷	0.895±0.004	0.897±0.005	0.825±0.010	0.818±0.031	0.831±0.013
DrugBAN	0.903±0.005	0.902±0.004	0.834±0.008	0.820±0.021	0.847±0.010

**Figure 2.** In-domain performance comparison on the Human dataset with random split and cold pair split. Left: AUROC scores. Right: AUPRC scores. The grey lines are error bars indicating the standard deviation.

102 specificity at the threshold of the best F1 score. We conduct five independent runs with different random seeds for each dataset
 103 split. The best performing model is selected to be the one with the best AUROC on the validation set. The selected model is
 104 then evaluated on the test set to report the performance metrics.

105 In-domain performance comparison

106 Here we compare DrugBAN with five baselines under the random split setting: support vector machine³² (SVM), random
 107 forest³³ (RF), DeepConv-DTI¹¹, GraphDTA¹³, and MolTrans¹⁷. This is the in-domain scenario so we use vanilla DrugBAN
 108 without embedding the CDAN module. Table 1 shows the comparison on the BindingDB and BioSNAP datasets. DrugBAN
 109 has consistently outperformed baselines in AUROC, AUPRC, and accuracy, while the performance in sensitivity and specificity
 110 is also competitive. The results indicate that data-driven representation learning can capture more important information than
 111 pre-defined descriptor features in in-domain DTI prediction. Moreover, DrugBAN can capture interaction patterns via its
 112 pairwise interaction module, further improving prediction performance.

113 Figure 2 shows the in-domain results on the Human dataset. Under the random split, the deep learning-based models all
 114 achieve similar and promising performance (AUROC > 0.98). However, Chen et al. (2020)¹⁶ pointed out that the Human
 115 dataset had some hidden ligand bias, resulting in the correct predictions being made only based on the drug features rather than
 116 interaction patterns. The high accuracy could be due to bias and overfitting, not indicating a model’s real-world performance on
 117 prospective prediction. Therefore, we further use a cold pair split strategy to evaluate models to mitigate the overoptimism
 118 of performance estimation under random split due to the data bias. This cold pair split strategy guarantees that all test drugs
 119 and proteins are not observed during training so that prediction on test data cannot rely only on the features of known drugs
 120 or proteins. We randomly assign 5% and 10% DTI pairs into the validation and test sets respectively, and remove all their
 121 associated drugs and proteins from the training set. Figure 2 indicates that all models have a significant performance drop

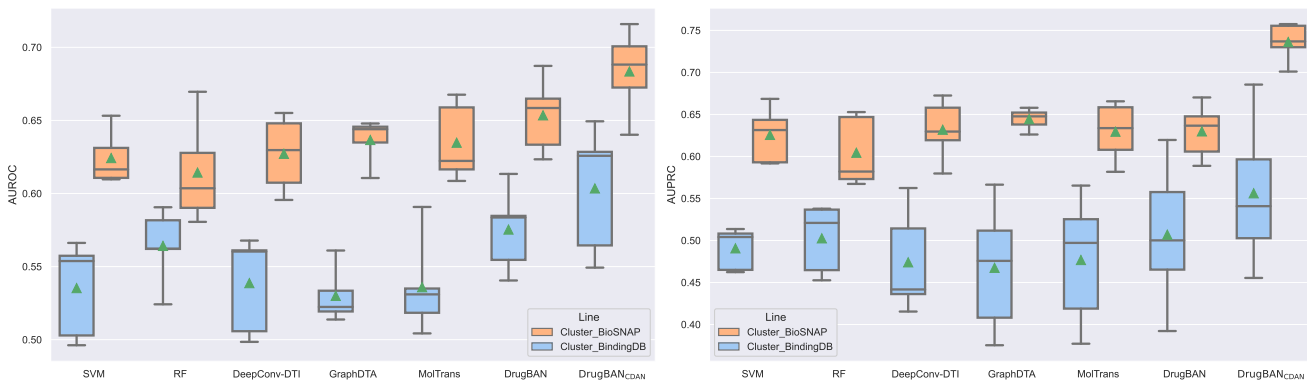


Figure 3. Cross-domain performance comparison on the BindingDB and BioSNAP datasets with clustering-based pair split. Left: AUROC scores. Right: AUPRC scores. The box plots show the median as the center lines, and the mean as the green triangles.

Table 2. Ablation study in AUROC on the BindingDB and BioSNAP datasets with random and clustering-based split strategies (averaged over five random runs). The first four models show the effectiveness of our bilinear attention module, and the last three models show the strength of DrugBAN_{CDAN} on cross-domain prediction (**Best**, **Second Best**).

Ablation tests	BindingDB _{random}	BioSNAP _{random}	BindingDB _{cluster}	BioSNAP _{cluster}
Linear concatenation ^{2, 11, 13}	0.949±0.002	0.887±0.007	-	-
One-side target attention ¹⁴	0.950±0.002	0.890±0.005	-	-
One-side drug attention ¹⁴	0.953±0.002	0.892±0.004	-	-
DrugBAN	0.960±0.001	0.903±0.005	0.575±0.025	0.654±0.023
MolTrans _{CDAN}	-	-	0.575±0.038	0.656±0.028
DrugBAN _{DANN}	-	-	0.592±0.042	0.667±0.030
DrugBAN _{CDAN}	-	-	0.604±0.039	0.684±0.026

122 from random split to cold pair split, especially for SVM and RF. However, we can see that DrugBAN still achieves the best
 123 performance against other state-of-the-art deep learning baselines.

124 Cross-domain performance comparison

125 In-domain classification under random split is an easier task and of less practical importance. Therefore, next, we study more
 126 realistic and challenging cross-domain DTI prediction, where training data and test data have different distributions. To imitate
 127 this scenario, the original data is divided into source and target domains by the clustering-based pair split. We turn on the
 128 CDAN module of DrugBAN to get DrugBAN_{CDAN} for studying knowledge transferability in cross-domain prediction.

129 Figure 3 presents the performance evaluation on the BindingDB and BioSNAP datasets with clustering-based pair split.
 130 Compared to the previous in-domain prediction results, the performance of all DTI models drops significantly due to much less
 131 information overlap between training and test data. In this scenario, vanilla DrugBAN still outperforms other state-of-the-art
 132 models on the whole. Specifically, it outperforms MolTrans by 2.9% and 7.4% in AUROC on the BioSNAP and BindingDB
 133 datasets, respectively. The results show that DrugBAN is a robust method under both in-domain and cross-domain settings.
 134 Interestingly, RF achieves good performance and even consistently outperforms other deep learning baselines (DeepConv,
 135 GraphDTA and MolTrans) on the BindingDB dataset. The results indicate that deep learning methods are not always superior
 136 to shallow machine learning methods under the cross-domain setting.

137 Recently, domain adaptation techniques have received increasing attention due to the ability of transferring knowledge
 138 across domains, but they are mainly applied to computer vision and natural language processing problems. We combine
 139 vanilla DrugBAN with CDAN to tackle cross-domain DTI prediction. As shown in Figure 3, DrugBAN_{CDAN} has significant
 140 performance improvements with the introduction of a domain adaptation module. On the BioSNAP dataset, it outperforms
 141 vanilla DrugBAN by 4.6% and 16.9% in AUROC and AUPRC, respectively. By minimizing the distribution discrepancy across
 142 domains, CDAN can effectively enhance DrugBAN generalization ability and provide more reliable results.

143 These results demonstrate the strength of DrugBAN in generalizing prediction performance across domains.

144 Ablation study

145 Here we conduct an ablation study to investigate the influences of bilinear attention and domain adaptation modules on DrugBAN.
146 The results are shown in Table 2. To validate the effectiveness of bilinear attention, we study three variants of DrugBAN that
147 differ in the joint representation computation between drug and protein: one-side drug attention, one-side protein attention,
148 and linear concatenation. The one-side attention is equivalent to the neural attention mechanism introduced by Tsubaki et al.
149 (2019)¹⁴, which is used to capture the joint representation between a drug vector representation and a protein subsequence
150 matrix representation. We replace the bilinear attention in DrugBAN with one-side attention to generate the two variants.
151 Linear concatenation is a simple vector concatenation of drug and protein vector representations after a max-pooling layer.
152 As shown in the first four rows of Table 2, the results demonstrate that bilinear attention is the most effective method to
153 capture interaction information for DTI prediction. To examine the effect of CDAN, we study two variants: DrugBAN with
154 domain-adversarial neural network (DANN)³⁴ (i.e. DrugBAN_{DANN}) and MolTrans with CDAN (i.e. MolTrans_{CDAN}). DANN
155 is another adversarial domain adaptation technique without considering classification distribution. The last four rows of Table 2
156 indicate that DrugBAN_{CDAN} still achieves the best performance improvement in cross-domain prediction.

157 Interpretability with bilinear attention visualization

158 A further strength of DrugBAN is to enable molecular level insights and interpretation critical for drug design efforts, utilizing
159 the components of the bilinear attention map to visualize the contribution of each substructure to the final predictive result.
160 Here, we examine the top three predictions (PDB ID: 6QL2, 5W8L and 4N6H) of co-crystallized ligands from Protein Data Bank
161 (PDB)³⁷. Only X-ray structures with resolution greater than 2.5 Å that corresponded to human protein targets were proceeded
162 for selection. In addition, co-crystallized ligands were required to have pIC₅₀ ≤ 100 nM and not to be part of the training set. The
163 visualization results are shown in Figure 4a alongside the ligand-protein interaction maps originating from the corresponding
164 X-ray structures. For each molecule, we colored its top 20% weighted atoms in bilinear attention map with orange.

165 For PDB structure 6QL2 (ethoxzolamide complexed with human carbonic anhydrase 2), our model correctly interpreted
166 sulfonamide region as essential for ligand-protein binding (in 6QL2: sulfonamide oxygen as a hydrogen bond acceptor to the
167 backbone of Leu198 and Thr199, and amino group as a hydrogen bond donor to the side chains of His94 and Thr199). On
168 another hand, ethoxy group of ethoxzolamide was incorrectly predicted to form specific interactions with the protein, although
169 its exposure to the solvent may promote further binding (blue highlight). In addition, benzothiazole scaffold, which forms an
170 arene-H interaction with Leu198, is only partly highlighted by our interpretability model. It is worth mentioning that though
171 top 20% of interacting atoms of ethoxzolamide only corresponded to three highlighted atoms, all of them indicated different
172 ligand-protein interaction sites corroborated by the X-ray structure.

173 In 5W8L structure (9YA ligand bound to human L-lactate dehydrogenase A), the interpretability feature once more
174 highlighted important interaction patterns for ligand-protein binding. For example, sulfonamide group was once more indicated
175 to form specific interactions with the protein (in 5W8L: amino group as a hydrogen bond donor to the side chains of Asp140
176 and Glu191, and sulfonamide oxygen as a hydrogen bond acceptor to the backbone of Asp140 and Ile141). Similarly, we noted
177 that carboxylic acid group was also partly highlighted (in 5W8L: carboxylic acid oxygens act as hydrogen bond acceptors to the
178 side chains of Arg168, His192, and Thr247). Moreover, biphenyl rings were correctly predicted to participate in ligand-protein
179 binding (in 5W8L: arene-H interaction with Arg105 and Asn137). Although 9YA (bound to 5W8L) was much larger and complex
180 than ethoxzolamide (bound to 6QL2), the model showed good interpretability potential for the majority of the experimentally
181 confirmed interactions.

182 In the third example, 4N6H X-ray complex of human delta-type opioid receptor with EJ4 ligand, main interacting functional
183 groups of EJ4 were once more highlighted correctly. Here, a hydroxyl group of the aliphatic ring complex and a neighboring
184 tertiary amine (in 4N6H: both as hydrogen bond donors to the side chain of Asp128) were correctly interpreted to form specific
185 interactions. On the other hand, phenol group was wrongly predicted to participate in protein binding.

186 As for the more challenging protein sequence interpretability, the results were overall weaker than those for the ligand
187 interpretability. Although many amino acid residues that were predicted to potentially participate in ligand binding were in fact
188 distantly located to the respective compounds, a number of amino acid residues forming the binding sites were yet correctly
189 predicted, which is shown in Figure 4b. For example, in 6QL2 complex the following residues were highlighted: His94, His96,
190 Thr200, Pro201, Pro202, Leu203, Val207, Trp209. Among these, only His94 forms specific interaction with ethoxzolamide.
191 In 5W8L, none of the residues that constitute the ligand-protein binding site were highlighted. However, in 4N6H structure,
192 there were several correctly predicted residues within the binding site: Lys214, Val217, Leu300, Cys303, Ile304, Gly307, and
193 Tyr308. Unfortunately, none of the residues participated in the specific interactions with the ligand. Given these results, it is
194 expected that protein sequence interpretability would be less confident because the one-dimensional protein sequence (used
195 as protein information input in our model) does not necessarily imply the three-dimensional configuration and locality of the
196 binding pocket. However, the results from the primary protein sequence are encouraging enough to safely assume that the
197 further incorporation of three-dimensional protein information into the modeling framework would eventually improve the

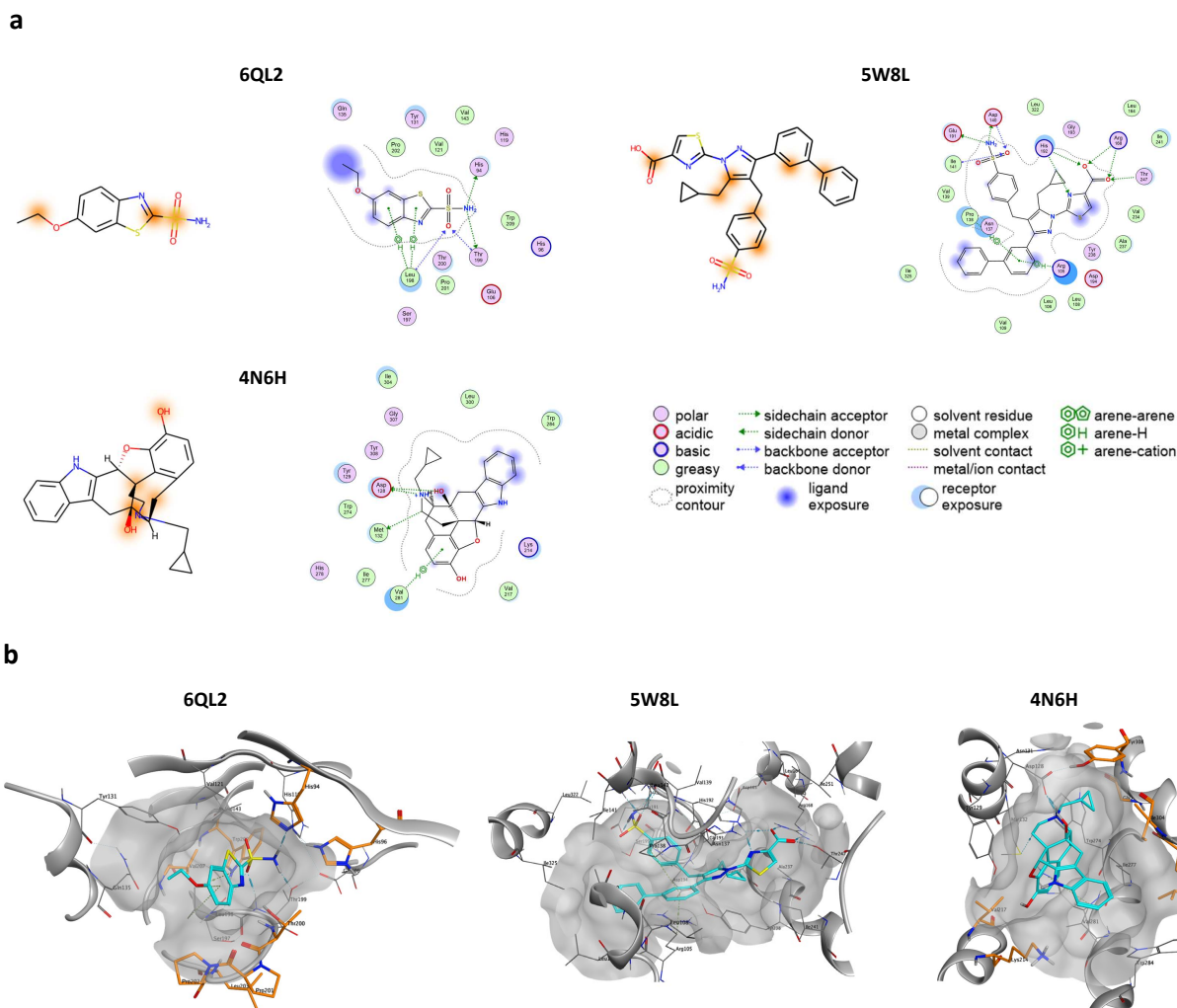


Figure 4. Importance visualization of ligands and binding pockets. (a) Interpretability of co-crystallized ligands. The left-hand side of each panel shows the two-dimensional structures of ligands with highlighted atoms (orange) that were predicted to contribute to protein binding. All structures were visualized using RDKit³⁵. In addition, ligand-protein interaction maps (right-hand side of each panel) from the corresponding crystal structures of these ligands are provided. At the right bottom, the legend panel for the ligand-protein interaction maps is displayed. (b) Interpretability of binding pocket structures. The three-dimensional representations of ligand-protein binding pockets are provided highlighting the correctly predicted amino acid residues (orange) that surround the corresponding ligands (cyan). Remaining amino acid residues, secondary structure elements, and surface maps are colored in grey. All ligand-protein interaction maps and three-dimensional representations of X-ray structures were visualized using the Molecular Operating Environment (MOE) software³⁶.

198 model interpretability of drug-target interaction networks.

199 In addition, as the interpretability provided by DrugBAN is adaptively learned from DTI data itself, such interpretation
 200 has potential to find some hidden knowledge of local interactions that has not been explored, and could help drug hunters to
 201 improve binding properties of a given scaffold, or to reduce the off-target liabilities of a compound.

202 Conclusion

203 In this work, we present DrugBAN, an end-to-end bilinear attention deep learning framework for DTI prediction. We have
 204 integrated CDAN, an adversarial domain adaptation network, into the modeling process to enhance cross-domain generalization
 205 ability. Compared with other state-of-the-art DTI models and conventional machine learning models, the experimental results
 206 show that DrugBAN consistently achieves improved DTI prediction performance in both in-domain and cross-domain settings.
 207 Furthermore, by mapping attention weights to protein subsequences and drug compound atoms, our model can provide biological

208 insights for interpreting the nature of interactions. The proposed ideas are general in nature and can be extended to other
209 interaction prediction problems, such as the prediction of drug-drug interaction and protein-protein interaction.

210 This work focuses on chemogenomics-based DTI using 1D protein sequence and 2D molecular graph as input. Since
211 the number of highly accurate 3D structured proteins only accounts for a small fraction of the known protein sequences, this
212 work did not consider the modeling with such structural information. Nevertheless, DeepMind’s AlphaFold³⁸ is making great
213 progress in protein 3D structure prediction, recently generating 2 billion protein 3D structure predictions from 1 million species.
214 Such progress opens doors for utilizing 3D structural information in chemogenomics-based DTI prediction. Following the
215 idea of pairwise local interaction learning and domain adaptation, we believe that extending our ideas further on complex 3D
216 structures can lead to even better performance and interpretability in future work. Finally, this work studies different datasets
217 separately, combining dataset integration with DrugBAN will be another interesting future direction to explore.

218 Methods

219 Bilinear attention network

220 This is an attention-based model and was first proposed to solve the problem of visual question answering (VQA)²⁵. Given an
221 image and relevant natural language question, VQA systems aim to provide a text-image matching answer. Therefore, VQA can
222 be viewed as a multimodal learning task, similar to DTI prediction. Bilinear attention network (BAN) uses a bilinear attention
223 map to gracefully extend unitary attention networks for adapting multimodal learning, which considers every pair of multimodal
224 input channels, i.e., the pairs of image regions and question words to learn an interaction representation. Compared to using a
225 unitary attention mechanism directly on multimodal data, BAN can provide richer joint information but keep the computational
226 cost at the same scale. Due to the problem similarity between VQA and DTI, we design a BAN-inspired pairwise interaction
227 module for DTI prediction.

228 Domain adaptation

229 These approaches learn a model that reduces domain distribution shift between the source domain and target domain, which is
230 mainly developed and studied in computer vision³⁹. Early domain adaptation methods tended to reweight sample importance or
231 learn invariant feature representations in shallow feature space, using labeled data in the source domain and unlabeled data in
232 the target domain. More recently, deep domain adaptation methods embed the adaptation module in various deep architectures
233 to learn transferable representations^{40,41}. In particular, Long et al. (2018)²¹ proposed a novel deep domain adaptation method,
234 CDAN, that combines adversarial networks with multilinear conditioning for transferable representation learning. By introducing
235 classifier prediction information into adversarial learning, CDAN can effectively align data distributions in different domains.
236 We embed CDAN as an adaptation module in DrugBAN to enhance model performance for cross-domain DTI prediction.

237 DrugBAN architecture

238 **CNN for protein sequence.** The protein feature encoder consists of three consecutive 1D-convolutional layers, which transforms
239 an input protein sequence to a matrix representation in the latent feature space. Each row of the matrix denotes a subsequence
240 representation in the protein. Drawing on the concept of word embedding, we first initialize all amino acids into a learnable
241 embedding matrix $\mathbf{E}_p \in \mathbb{R}^{23 \times D_p}$, where 23 is the number of amino acid types and D_p is the latent space dimensionality. By
242 looking up \mathbf{E}_p , each protein sequence \mathcal{P} can be initialized to corresponding feature matrix $\mathbf{X}_p \in \mathbb{R}^{\Theta_p \times D_p}$. Here Θ_p is the
243 maximum allowed length of a protein sequence, which is set to align different protein lengths and make batch training. Following
244 previous works^{2,14,17}, protein sequences with maximum allowed length are cut, and those with smaller length are padded with
245 zeros.

246 The CNN-block protein encoder extracts local residue patterns from the protein feature matrix \mathbf{X}_p . Here a protein sequence
247 is considered as an overlapping 3-mer amino acids such as “METLCL...DSMN” \rightarrow “MET”, “ETL”, “TLC”, ..., “DSM”, “DLK”.
248 The first convolutional layer is utilized to capture the 3-mer residue-level features with kernel size = 3. Then the next two
249 layers continue to enlarge the receptive field and learn more abstract features of local protein fragments. The protein encoder is
250 described as follows:

$$\mathbf{H}_p^{(l+1)} = \sigma(\text{CNN}(\mathbf{W}_c^{(l)}, \mathbf{b}_c^{(l)}, \mathbf{H}_p^{(l)})), \quad (1)$$

251 where $\mathbf{W}_c^{(l)}$ and $\mathbf{b}_c^{(l)}$ are the learnable weight matrices (filters) and bias vector in the l -th CNN layer. $\mathbf{H}_p^{(l)}$ is the l -th hidden
252 protein representation and $\mathbf{H}_p^{(0)} = \mathbf{X}_p$. $\sigma(\cdot)$ denotes a non-linear activation function, with ReLU(\cdot) used in our experiments.

253 **GCN for molecular graph.** For drug compound, we convert each SMILES string to its 2D molecular graph \mathcal{G} . To represent
254 node information in \mathcal{G} , we first initialize each atom node by its chemical properties, as implemented in the DGL-LifeSci⁴²

255 package. Each atom is represented as a 74-dimensional integer vector describing eight pieces of information: the atom type, the
 256 atom degree, the number of implicit Hs, formal charge, the number of radical electrons, the atom hybridization, the number of
 257 total Hs and whether the atom is aromatic. Similar to the maximum allowed length setting in a protein sequence above, we
 258 set a maximum allowed number of nodes Θ_d . Molecules with less nodes will contain virtual nodes with zero padded. As a
 259 result, each graph’s node feature matrix is denoted as $\mathbf{M}_d \in \mathbb{R}^{\Theta_d \times 74}$. Moreover, we use a simple linear transformation to define
 260 $\mathbf{X}_d = \mathbf{W}_0 \mathbf{M}_d^T$, leading to a real-valued dense matrix $\mathbf{X}_d \in \mathbb{R}^{\Theta_d \times D_d}$ as the input feature.

261 We employed a three-layer GCN-block to effectively learn the graph representation on drug compounds. GCN generalizes
 262 the convolutional operator to an irregular domain. Specifically, we update the atom feature vectors by aggregating their
 263 corresponding sets of neighborhood atoms, connected by chemical bonds. This propagation mechanism automatically captures
 264 substructure information of a molecule. We keep the node-level drug representation for subsequent explicit learning of local
 265 interactions with protein fragments. The drug encoder is written as:

$$\mathbf{H}_d^{(l+1)} = \sigma(\text{GCN}(\tilde{\mathbf{A}}, \mathbf{W}_g^{(l)}, \mathbf{b}_g^{(l)}, \mathbf{H}_p^{(l)})), \quad (2)$$

266 where $\mathbf{W}_g^{(l)}$ and $\mathbf{b}_g^{(l)}$ are the GCN’s layer-specific learnable weight matrix and bias vector, $\tilde{\mathbf{A}}$ is the adjacency matrix with added
 267 self-connections in molecular graph \mathcal{G} , and $\mathbf{H}_d^{(l)}$ is the l -th hidden node representation with $\mathbf{H}_d^{(0)} = \mathbf{X}_d$.

268 **Pairwise interaction learning.** We apply a bilinear attention network module to capture pairwise local interactions between
 269 drug and protein. It consists of two layers: (i) A bilinear interaction map to capture pairwise attention weights and (ii) a bilinear
 270 pooling layer over the interaction map to extract joint drug-target representation.

271 Given the third layer’s hidden protein and drug representations $\mathbf{H}_p^{(3)} = \{\mathbf{h}_p^1, \mathbf{h}_p^2, \dots, \mathbf{h}_p^M\}$ and $\mathbf{H}_d^{(3)} = \{\mathbf{h}_d^1, \mathbf{h}_d^2, \dots, \mathbf{h}_d^N\}$ after
 272 separate CNN and GCN encoders, where M and N denote the number of encoded substructures in a protein and atoms in a
 273 drug. The bilinear interaction map can obtain a single head pairwise interaction $\mathbf{I} \in \mathbb{R}^{N \times M}$:

$$\mathbf{I} = ((\mathbf{I} \cdot \mathbf{q}^\top) \circ \sigma((\mathbf{H}_d^{(3)})^\top \mathbf{U})) \cdot \sigma(\mathbf{V}^\top \mathbf{H}_p^{(3)}), \quad (3)$$

274 where $\mathbf{U} \in \mathbb{R}^{D_d \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_p \times K}$ are learnable weight matrices for drug and protein representations, $\mathbf{q} \in \mathbb{R}^K$ is a learnable
 275 weight vector, $\mathbf{I} \in \mathbb{R}^N$ is a fixed all-ones vector, and \circ denotes Hadamard (element-wise) product. The elements in \mathbf{I} indicate
 276 the interaction intensity of respective drug-target sub-structural pairs, with mapping to potential binding sites and molecular
 277 substructures. To intuitively understand bilinear interaction, an element $\mathbf{I}_{i,j}$ in Equation (3) can also be written as:

$$\mathbf{I}_{i,j} = \mathbf{q}^\top (\sigma(\mathbf{U}^\top \mathbf{h}_d^i) \circ \sigma(\mathbf{V}^\top \mathbf{h}_p^j)), \quad (4)$$

278 where \mathbf{h}_d^i is the i -th column of $\mathbf{H}_d^{(3)}$ and \mathbf{h}_p^j is the j -th column of $\mathbf{H}_p^{(3)}$, respectively denoting the i -th and j -th sub-structural
 279 representations of drug and protein. Therefore, we can see a bilinear interaction as first mapping representations \mathbf{h}_d^i and \mathbf{h}_p^j to a
 280 common feature space with weight matrices \mathbf{U} and \mathbf{V} , then learn an interaction on Hadamard product and the weight of vector
 281 \mathbf{q} . In this way, pairwise interactions provide interpretability on the contribution of sub-structural pairs to the predicted result.

282 To obtain the joint representation $\mathbf{f}' \in \mathbb{R}^K$, we introduce a bilinear pooling layer over the interaction map \mathbf{I} . Specifically,
 283 the k -th element of \mathbf{f}' is computed as:

$$\begin{aligned} \mathbf{f}'_k &= \sigma((\mathbf{H}_d^{(3)})^\top \mathbf{U})_k^\top \cdot \mathbf{I} \cdot \sigma((\mathbf{H}_p^{(3)})^\top \mathbf{V})_k \\ &= \sum_{i=1}^N \sum_{j=1}^M \mathbf{I}_{i,j} (\mathbf{h}_d^i)^\top (\mathbf{U}_k \mathbf{V}_k^\top) \mathbf{h}_p^j, \end{aligned} \quad (5)$$

284 where \mathbf{U}_k and \mathbf{V}_k denote the k -th column of weight matrices \mathbf{U} and \mathbf{V} . Notably, there are no new learnable parameters at this
 285 layer. The weight matrices \mathbf{U} and \mathbf{V} are shared with the previous interaction map layer to decrease the number of parameters
 286 and alleviate over-fitting. Moreover, we add a sum pooling on the joint representation vector to obtain a compact feature map:

$$\mathbf{f} = \text{SumPool}(\mathbf{f}', s), \quad (6)$$

287 where the $\text{SumPool}(\cdot)$ function is a one-dimensional and non-overlapped sum pooling operation with stride s . It reduces the
 288 dimensionality of $\mathbf{f}' \in \mathbb{R}^K$ to $\mathbf{f} \in \mathbb{R}^{K/s}$. Furthermore, we can extend the single pairwise interaction to a multi-head form by

289 calculating multiple bilinear interaction maps. The final joint representation vector is a sum of individual heads. As the weight
 290 matrices \mathbf{U} and \mathbf{V} are shared, each additional head only adds one new weight vector \mathbf{q} , which is parameter-efficient. In our
 291 experiments, the multi-head interaction has a better performance than a single one.

292 Thus, using the novel bilinear attention mechanism, the model can explicitly learn pairwise local interactions between drug
 293 and protein. This interaction module is inspired by and adapted from Kim et al. (2018)²⁵ and Yu et al. (2018)²⁴, where two
 294 bilinear models are designed for the VQA problem. To compute the interaction probability, we feed the joint representation \mathbf{f}
 295 into the decoder, which is one fully connected classification layer followed by a sigmoid function:

$$p = \text{Sigmoid}(\mathbf{W}_o \mathbf{f} + \mathbf{b}_o), \quad (7)$$

296 where \mathbf{W}_o and \mathbf{b}_o are learnable weight matrix and bias vector.

297 Finally, we jointly optimize all learnable parameters by backpropagation. The training objective is to minimize the
 298 cross-entropy loss as follows:

$$\mathcal{L} = - \sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) + \frac{\lambda}{2} \|\Theta\|_2^2, \quad (8)$$

299 where Θ is the set of all learnable weight matrices and bias vectors above, y_i is the ground-truth label of the i -th drug-target
 300 pair, p_i is its output probability by the model, and λ is a hyperparameter for L2 regularization.

301 **Cross-domain adaptation for better generalization.** Machine learning models tend to perform well on similar data from
 302 the same distribution (i.e. in-domain), but poorer on dissimilar data with different distribution (i.e. cross-domain). It is a key
 303 challenge to improve model performance on cross-domain DTI prediction. In our framework, we embed conditional adversarial
 304 domain adaptation (CDAN) to enhance generalization from a source domain with sufficient labeled data to a target domain
 305 where only unlabeled data is available.

306 Given a source domain $\mathcal{S}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ of N_s labeled drug-target pairs and a target domain $\mathcal{S}_t = \{x_j^t\}_{j=1}^{N_t}$ of N_t unlabeled
 307 drug-target pairs, we leverage CDAN to align their distributions and improve prediction performance across domains. Figure 1c
 308 shows the CDAN workflow in our framework, including three key components: the feature extractor $F(\cdot)$, the decoder $G(\cdot)$, and
 309 the domain discriminator $D(\cdot)$. We use $F(\cdot)$ to denote the separate feature encoders and bilinear attention network together
 310 to generate joint representations of input domain data, i.e., $\mathbf{f}_i^s = F(x_i^s)$ and $\mathbf{f}_j^t = F(x_j^t)$. Next, we use the fully connected
 311 classification layer mentioned above followed by a softmax function as $G(\cdot)$ to get a classifier prediction $\mathbf{g}_i^s = G(\mathbf{f}_i^s) \in \mathbb{R}^2$ and
 312 $\mathbf{g}_j^t = G(\mathbf{f}_j^t) \in \mathbb{R}^2$. Furthermore, we apply a multilinear map to embed joint representation \mathbf{f} and classifier prediction \mathbf{g} into a
 313 joint conditional representation $\mathbf{h} \in \mathbb{R}^{2K/s}$, which is defined as the flattening of the outer product of the two vectors:

$$\mathbf{h} = \text{FLATTEN}(\mathbf{f} \otimes \mathbf{g}), \quad (9)$$

314 where \otimes is the outer product.

315 The multilinear map captures multiplicative interactions between two independent distributions^{43,44}. Following the CDAN
 316 mechanism, we simultaneously align the joint representation and predicted classification distributions of source and target
 317 domains by conditioning the domain discriminator $D(\cdot)$ on the \mathbf{h} . The domain discriminator $D(\cdot)$, consisting of a three-layer
 318 fully connected networks, learns to distinguish whether a joint conditional representation \mathbf{h} is derived from the source domain
 319 or the target domain. On the other hand, the feature extractor $F(\cdot)$ and decoder $G(\cdot)$ are trained to minimize the source domain
 320 cross-entropy loss \mathcal{L} with source label information, and simultaneously generate indistinguishable representation \mathbf{h} to confuse
 321 the discriminator $D(\cdot)$. As a result, we can formulate the two losses in the cross-domain modeling:

$$\mathcal{L}_s(F, G) = \mathbb{E}_{(x_i^s, y_i^s) \sim \mathcal{S}_s} \mathcal{L}(G(F(x_i^s)), y_i^s), \quad (10)$$

$$\mathcal{L}_{adv}(F, G, D) = \mathbb{E}_{x_i^t \sim \mathcal{S}_t} \log(1 - D(\mathbf{f}_i^t, \mathbf{g}_i^t)) + \mathbb{E}_{x_j^s \sim \mathcal{S}_s} \log(D(\mathbf{f}_j^s, \mathbf{g}_j^s)), \quad (11)$$

$$(12)$$

322 where \mathcal{L}_s is the cross-entropy loss on the labeled source domain and \mathcal{L}_{adv} is the adversarial loss for domain discrimination. The
 323 optimization problem is written as a minimax paradigm:

$$\max_D \min_{F, G} \mathcal{L}_s(F, G) - \omega \mathcal{L}_{adv}(F, G, D), \quad (13)$$

where $\omega > 0$ is a hyperparameter to weight \mathcal{L}_{adv} . By introducing the adversarial training on \mathcal{L}_{adv} , our framework can reduce the data distribution shift between source and target domains, leading to the improved generalization on cross-domain prediction.

Experimental setting

Datasets. We evaluate DrugBAN and five state-of-the-art baselines on three public DTI datasets: BindingDB, BioSNAP and Human. The BindingDB dataset is a web-accessible database⁴⁵ of experimentally validated binding affinities, focusing primarily on the interactions of small drug-like molecules and proteins. We use a low-bias version of the BindingDB dataset constructed in our earlier work Bai et al. (2021)⁴⁶, with the bias-reducing preprocessing steps described in Supplementary Section 2. The BioSNAP dataset is created from the DrugBank database⁴⁷ by Huang et al. (2021)¹⁷ and Marinka et al. (2018)²⁸, consisting of 4,510 drugs and 2,181 proteins. It is a balanced dataset with validated positive interactions and an equal number of negative samples randomly obtained from unseen pairs. The Human dataset is constructed by Liu et al. (2015)²⁹, including highly credible negative samples via an *in silico* screening method. Following previous studies^{14,16,20}, we also use the balanced version of Human dataset containing the same number of positive and negative samples. To mitigate the influence of the hidden data bias¹⁶, we use additional cold pair split for performance evaluation on the Human dataset. Supplementary Table 2 shows statistics of the three datasets.

Implementation. DrugBAN is implemented in PyTorch 1.7.1⁴⁸. The batch size is set to be 64 and the Adam optimizer is used with a learning rate of $5e-5$. We allow the model to run for at most 100 epochs for all datasets. The best performing model is selected at the epoch giving the best AUROC score on the validation set, which is then used to evaluate the final performance on the test set. The protein feature encoder consists of three 1D-CNN layers with the number of filters [128, 128, 128] and kernel sizes [3, 6, 9]. The drug feature encoder consists of three GCN layers with hidden dimensions [128, 128, 128]. The maximum allowed sequence length for protein is set to be 1200, and the maximum allowed number of atoms for drug molecule is 290. In the bilinear attention module, we only employ two attention heads to provide better interpretability. The latent embedding size k is set to be 768 and the sum pooling window size s is 3. The number of hidden neurons in the fully connected decoder is 512. Our model performance is not sensitive to hyperparameter settings. The configuration details and sensitivity analysis are provided in Supplementary Section 3. We also present a scalability study in Supplementary Section 7.

Baselines. We compare DrugBAN with the following five models on DTI prediction: (1) Two shallow machine learning methods, support vector machine (SVM) and random forest (RF) applied on the concatenated fingerprint ECFP4 and PSC features; (2) DeepConv-DTI¹¹ that uses CNN and one global max-pooling layer to extract local patterns in protein sequence and a fully connected network to encode drug fingerprint ECFP4; (3) GraphDTA¹³ that models DTI using graph neural networks to encode drug molecular graph and CNN to encode protein sequence. The learned drug and protein representation vectors are combined with a simple concatenation. To adapt GraphDTA from the original regression task to a binary classification task, we follow the steps in earlier literature^{16,17} to add a Sigmoid function in its last fully connected layer, and then optimize its parameters with a cross-entropy loss. (4) MolTrans¹⁷, a deep learning model adapting transformer architecture to encode drug and protein information, and a CNN-based interactive module to learn sub-structural interaction. For the above deep DTI models, we follow the recommended model hyper-parameter settings described in their original papers.

Data availability

The experimental data with each split strategy is available at <https://github.com/pz-white/DrugBAN/tree/main/datasets>. All data used in this work are from public resource. The BindingDB source is at <https://www.bindingdb.org/bind/index.jsp>; The BioSNAP source is at <https://github.com/kexinhuang12345/MolTrans> and the Human source is at <https://github.com/lifanchen-simm/transformerCPI>.

Code availability

The source code and implementation details of DrugBAN are freely available at GitHub repository (<https://github.com/pz-white/DrugBAN>) and archived on Zenodo (<https://doi.org/10.5281/zenodo.7231658>).

References

1. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8** (2017).
2. Öztürk, H., Olmez, E. O. & Özgür, A. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821 – i829 (2018).
3. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232 – i240 (2008).

- 373 4. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, Practice, and Opportunities. *Inf.*
374 *Fusion* **50**, 71–91 (2019).
- 375 5. Bagherian, M. *et al.* Machine learning approaches and databases for prediction of drug–target interaction: a survey paper.
376 *Briefings Bioinforma.* **22**, 247 – 269 (2021).
- 377 6. Wen, M. *et al.* Deep-learning-based drug-target interaction prediction. *J. Proteome Res.* **16** **4**, 1401–1409 (2017).
- 378 7. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: Evaluating chemical data for machine learning in structure-
379 based virtual screening. *J. Chem. Inf. Model.* **59** **3**, 947–961 (2019).
- 380 8. Lim, S. *et al.* A review on compound-protein interaction prediction methods: Data, format, representation and model.
381 *Comput. Struct. Biotechnol. J.* **19**, 1541 – 1556 (2021).
- 382 9. Gao, K. Y. *et al.* Interpretable drug target prediction using deep neural representation. In *IJCAI*, 3371–3377 (2018).
- 383 10. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5**,
384 262–275 (2004).
- 385 11. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on
386 protein sequences. *PLoS Comput. Biol.* **15** (2019).
- 387 12. Hinnerichs, T. & Hoehndorf, R. DTI-Voodoo: machine learning over interaction networks and ontology-based background
388 knowledge predicts drug–target interactions. *Bioinformatics* **37**, 4835 – 4843 (2021).
- 389 13. Nguyen, T. *et al.* GraphDTA: Predicting drug-target binding affinity with graph neural networks. *Bioinformatics* **37**,
390 1140–1147 (2021).
- 391 14. Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with end-to-end learning of neural networks
392 for graphs and sequences. *Bioinformatics* **35**, 309–318 (2019).
- 393 15. Feng, Q., Dueva, E., Cherkasov, A. & Ester, M. PADME: A deep learning-based framework for drug-target interaction
394 prediction. *arXiv preprint arXiv:1807.09741* (2018).
- 395 16. Chen, L. *et al.* TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with
396 self-attention mechanism and label reversal experiments. *Bioinformatics* (2020).
- 397 17. Huang, K., Xiao, C., Glass, L. & Sun, J. MolTrans: Molecular interaction transformer for drug–target interaction prediction.
398 *Bioinformatics* **37**, 830 – 836 (2021).
- 399 18. Schenone, M., Dancík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical
400 biology and drug discovery. *Nat. Chem. Biol.* **9** **4**, 232–40 (2013).
- 401 19. Öztürk, H., Ozkirimli, E. & Özgür, A. WideDTA: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*
402 (2019).
- 403 20. Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering
404 system. *Nat. Mach. Intell.* **2**, 134–140 (2020).
- 405 21. Long, M., Cao, Z., Wang, J. & Jordan, M. I. Conditional Adversarial Domain Adaptation. In *NeurIPS* (2018).
- 406 22. Kim, J.-H. *et al.* Hadamard Product for Low-rank Bilinear Pooling. In *ICLR* (2017).
- 407 23. Kipf, T. & Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR* (2017).
- 408 24. Yu, Z., Yu, J., Xiang, C., Fan, J. & Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for
409 visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 5947–5959 (2018).
- 410 25. Kim, J.-H., Jun, J. & Zhang, B.-T. Bilinear Attention Networks. In *NeurIPS* (2018).
- 411 26. Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules.
412 *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
- 413 27. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. Bindingdb: a web-accessible database of experimentally
414 determined protein–ligand binding affinities. *Nucleic Acids Res.* **35**, D198 – D201 (2007).
- 415 28. Marinka Zitnik, S. M., Rok Sosič & Leskovec, J. BioSNAP Datasets: Stanford biomedical network dataset collection
416 (2018).
- 417 29. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound–protein interaction prediction by building up highly
418 credible negative samples. *Bioinformatics* **31**, i221 – i229 (2015).
- 419 30. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50** **5**, 742–54 (2010).

- 420 **31.** Cao, D., Xu, Q. & Liang, Y. Propy: a tool to generate various modes of chou's pseaac. *Bioinformatics* **29** 7, 960–2 (2013).
- 421 **32.** Cortes, C. & Vapnik, V. Support-vector networks. *Mach. learning* **20**, 273–297 (1995).
- 422 **33.** Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*,
423 vol. 1, 278–282 (1995).
- 424 **34.** Ganin, Y. *et al.* Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.* (2016).
- 425 **35.** Greg Landrum *et al.* RDKit: Open-source cheminformatics (2006).
- 426 **36.** Molecular Operating Environment (MOE). 2020.09 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite
427 910, Montreal, QC, Canada, H3A 2R7 (2022).
- 428 **37.** Burley, S. K. *et al.* Rcsb protein data bank: biological macromolecular structures enabling research and education in
429 fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464 – D474 (2019).
- 430 **38.** Jumper, J. M. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583 – 589 (2021).
- 431 **39.** Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
- 432 **40.** Gong, B., Grauman, K. & Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features
433 for unsupervised domain adaptation. In *ICML* (2013).
- 434 **41.** Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. & Schölkopf, B. Correcting sample selection bias by unlabeled data.
435 In *NIPS* (2006).
- 436 **42.** Li, M. *et al.* DGL-LifeSci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega* (2021).
- 437 **43.** Song, L., Huang, J., Smola, A. & Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to
438 dynamical systems. In *ICML* (2009).
- 439 **44.** Song, L. & Dai, B. Robust low rank kernel embeddings of multivariate distributions. In *NIPS* (2013).
- 440 **45.** Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems
441 pharmacology. *Nucleic acids research* **44**, D1045–D1053 (2016).
- 442 **46.** Bai, P. *et al.* Hierarchical clustering split for low-bias evaluation of drug-target interaction prediction. *2021 IEEE Int. Conf.*
443 *on Bioinforma. Biomed. (BIBM)* 641–644 (2021).
- 444 **47.** Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901 –
445 D906 (2008).
- 446 **48.** Paszke, A. *et al.* Automatic differentiation in PyTorch. (2017).

447 **Acknowledgements**

448 We thank S. Zhou, X. Liu and L. Schobs for their helpful suggestions and discussion on the work. We thank anonymous reviewers
449 for their insightful comments to improve the work. P.B. is supported by a University of Sheffield Faculty of Engineering
450 Research Scholarship.

451 **Author contributions**

452 P.B., F.M., B.J., and H.L. conceived and designed the work presented here. P.B. developed the models and performed the
453 experiments under the guidance of B.J. and H.L. F.M. and P.B. analyzed the data and conducted method comparisons. F.M.
454 contributed to materials/analysis tool. All authors contributed to write the paper.

455 **Additional information**

456 **Competing interests:** the authors declare no competing interests.

Supplementary Information

Interpretable bilinear attention network with domain adaptation improves drug-target prediction

Peizhen Bai¹, Filip Miljković², Bino John³, and Haiping Lu^{1*}

¹Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

²Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

³Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, USA

*corresponding author: Haiping Lu (h.lu@sheffield.ac.uk)

1. Clustering-based pair split strategy

As mentioned in the main text, we separately cluster drug compounds and target proteins of the BindingDB and BioSNAP datasets for cross-domain performance evaluation. Specifically, we choose the single-linkage clustering, a bottom-up hierarchical clustering to ensure that the distances between samples in different clusters are always larger than a pre-defined distance, i.e., minimum distance threshold γ . This property can prevent clusters from being too close to help to generate the cross-domain scenario.

We use binarized ECFP4 feature to represent drug compounds, and integral PSC feature to represent target proteins. For accurately measuring the pairwise distance, we use the Jaccard distance and cosine distance on ECFP4 and PSC, respectively. We choose $\gamma = 0.5$ in both drug and protein clusterings since this choice can prevent over-large clusters and be ensure separate dissimilar samples. We obtained 2,780 clusters of drugs and 1,693 clusters of proteins for the BindingDB dataset, and 2,387 clusters of drugs and 1,978 clusters of proteins for the BioSNAP dataset. Table 1 shows the number of samples in the ten largest clusters of the clustering results. It shows that BindingDB has a more balanced cluster distribution than BioSNAP in drug clustering. In addition, the protein clustering result tends to generate many small clusters with only a few proteins in both datasets, indicating that the average similarity between proteins is lower than that between drugs. We randomly select 60% drug clusters and 60% protein clusters from clustering result, and regard all associated drug-target pairs with them as source domain data. The associated pairs in the remaining clusters are considered to be source domain data. We conduct five independent clustering-based pair splits with different random seeds for downstream model training and evaluation. Clustering-based pair split allows quantitatively constructing cross-domain tasks by considering the similarity between drugs or proteins.

Table 1. Size of the ten largest clusters in the BindingDB and BioSNAP datasets generated by the clustering-based pair split.

Dataset	Object	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
BindingDB	Drug	598	460	304	290	253	250	203	202	198	158
BioSNAP	Drug	294	267	75	68	36	35	28	26	24	24
BindingDB	Protein	17	15	15	12	10	10	10	9	9	8
BioSNAP	Protein	8	8	8	6	5	4	4	4	4	4

2. Dataset statistics, notations, and preprocessing steps

Table 2 shows the statistics of experimental datasets and Table 3 lists the notations used in this paper with descriptions. The BioSNAP and Human datasets were created by Huang et al. (2021)¹ and Liu et al. (2015)², respectively. For the BindingDB dataset, we created a low-bias version from the BindingDB database source³ following the bias-reducing preprocessing steps in our earlier work⁴: i) We considered a drug-target pair to be positive only if its IC50 is less than 100 nM, and negative only if its IC50 was greater than 10,000 nM, giving a 100-fold difference to reduce class label noise. These IC50 thresholds were selected following earlier works^{5,6}. ii) We removed all DTI pairs where the drugs only had one type of pairs (positive or negative) to improve drug-wise pair class balance and reduce hidden ligand bias that can lead to the correct predictions based only on drug features.

3. Hyperparameter setting and sensitivity analysis

Table 4 shows a list of model hyperparameters and their values used in experiment. As our model performance is not sensitive to hyperparameter setting, we use the same hyperparameters on all experimental datasets (BindingDB, BioSNAP and Human).

Table 2. Experimental dataset statistics

Dataset	# Drugs	# Proteins	# Interactions
BindingDB ⁴	14,643	2,623	49,199
BioSNAP ¹	4,510	2,181	27,464
Human ²	2,726	2,001	6,728

Table 3. Notations and descriptions

Notations	Description
$\mathbf{E}_p \in \mathbb{R}^{23 \times D_p}$	protein amino acid embedding matrix
$\mathbf{f} \in \mathbb{R}^{K/s}$	drug-target joint representation
$F(\cdot), G(\cdot), D(\cdot)$	feature extractor, decoder and domain discriminator in CDAN
$\mathbf{g} \in \mathbb{R}^2$	output interaction probability by softmax function
$\mathbf{H}_p^{(l)}, \mathbf{H}_d^{(l)}$	hidden representation for protein (drug) in l -th CNN (GCN) layer
$\mathbf{I} \in \mathbb{R}^{N \times M}$	pair-wise interaction matrix between drug and protein substructures
$\mathbf{M}_d \in \mathbb{R}^{\Theta_d \times 74}$	drug node feature matrix by its chemical properties
$p \in \mathbb{R}^1$	output interaction probability by Sigmoid function
\mathcal{P}, \mathcal{G}	protein amino acid sequence, drug 2D molecular graph
$\mathbf{q} \in \mathbb{R}^K$	weight vector for bilinear transformation
$\mathbf{U} \in \mathbb{R}^{D_d \times K}$	the weight matrix for encoded drug representation
$\mathbf{V} \in \mathbb{R}^{D_p \times K}$	the weight matrix for encoded protein representation
$\mathbf{W}_c, \mathbf{b}_c$	the weight matrix and bias for protein CNN encoder
$\mathbf{W}_g, \mathbf{b}_g$	the weight matrix and bias for drug GCN encoder
$\mathbf{W}_o, \mathbf{b}_o$	the weight matrix and bias for decoder
$\mathbf{X}_p \in \mathbb{R}^{\Theta_p \times D_p}$	latent protein matrix representation
$\mathbf{X}_d \in \mathbb{R}^{\Theta_d \times D_d}$	latent drug matrix representation

39 Figure 1 illustrates the learning curves with the different choices of hyperparameters on the BindingDB validation set, including
 40 bilinear embedding size, learning rate and heads of attention. It shows that the performance differences are not large and
 41 typically converges between 30 and 40 epochs.

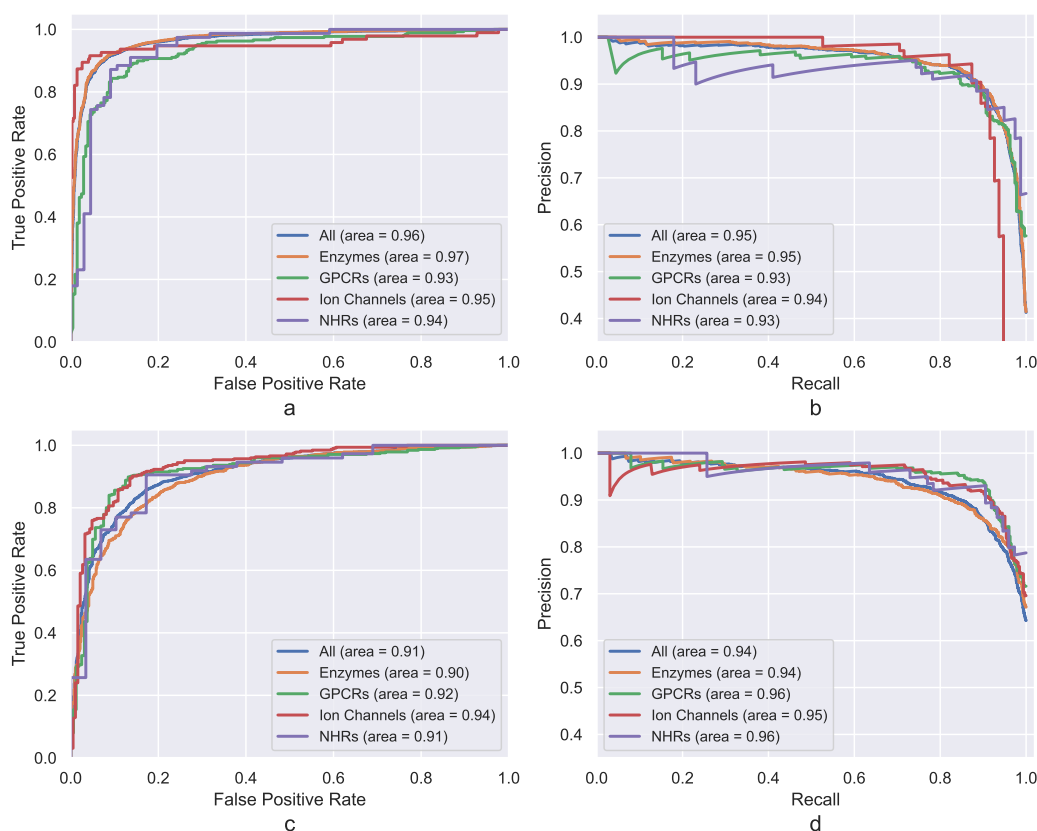
**Figure 1.** Learning curves with the different choices of hyperparameters on the BindingDB validation set.

Table 4. DrugBAN hyperparameter configuration

Module	Hyperparameter	Value
Optimizer	Learning rate	5e-5
Mini-batch	Batch size	64
Three-layer CNN protein encoder	Initial amino acid embedding	128
	Number of filters	[128, 128, 128]
	Kernel size	[3, 6, 9]
Three-layer GCN drug encoder	Initial atom embedding	128
	Hidden node dimensions	[128, 128, 128]
Bilinear interaction attention	Heads of bilinear attention	2
	Bilinear embedding size	768
	Sum pooling window size	3
Fully connected decoder	Number of hidden neurons	512
Discriminator	Number of hidden neurons	256

Table 5. Number of interactions for major protein families in the test sets.

Dataset	# Enzymes	# GPCRs	# Ion channels	# NHRs
BindingDB	5,277	472	440	144
BioSNAP	1,956	536	510	103

**Figure 2.** DrugBAN performance on different protein families. (a) AUROC curves on the BindingDB dataset. (b) AUPRC curves on the BindingDB dataset. (c) AUROC curves on the BioSNAP dataset. (d) AUPRC curves on the BioSNAP dataset.

4. Performance comparison across different protein families

We conduct experiments to study the performance of DrugBAN on different protein families. Following the previous studies^{1,7}, we select four major protein families: enzymes, G protein-coupled receptors (GPCRs), ion channels and nuclear hormone receptors (NHRs). We randomly retrieve one in-domain test set of BindingDB and BioSNAP respectively, and map their proteins to the four protein families using GtoPdb database (<https://www.guidetopharmacology.org/targets.jsp>). Table 5 presents the number of interactions for each protein family in the test sets. Figure 2 shows the performance (AUROC and AUPRC) varying only slightly given different protein families.

5. Performance comparison on unseen drugs/targets

Table 6. Performance (average AUROC over five random runs) comparison on the BindingDB and BioSNAP datasets with random split, unseen drug, and unseen target settings (**Best**, Second Best).

Setting	DeepConv-DTI ⁸	GraphDTA ⁹	MolTrans ¹	DrugBAN
	BindingDB			
Random Split	0.945±0.002	0.951±0.002	<u>0.952±0.002</u>	0.960±0.001
Unseen Drug	0.943±0.004	<u>0.950±0.004</u>	0.945±0.004	0.959±0.002
Unseen Target	0.627±0.070	<u>0.670±0.023</u>	0.661±0.037	0.692±0.038
	BioSNAP			
Random Split	0.886±0.006	0.887±0.008	<u>0.895±0.004</u>	0.903±0.005
Unseen Drug	0.856±0.005	<u>0.858±0.007</u>	0.856±0.008	0.886±0.005
Unseen Target	0.692±0.017	<u>0.704±0.010</u>	0.714±0.014	<u>0.710±0.016</u>

To study how DrugBAN and other deep learning baselines perform on unseen drugs/targets, we conduct additional experiments on BindingDB and BioSNAP. For each dataset, we randomly select 20% drugs/target proteins. Then we evaluate predictive performance on all DTI pairs associated with these drugs/target proteins (70% as test set for evaluation and 30% as validation set for determining early stopping), and the rest pairs as training set for model optimization. Each unseen setting has five independent runs. Table 6 presents the AUROC results on the test sets, including the results on the usual random split for comparison. DrugBAN achieves the best performance in five of the six settings, while its performance in the unseen target setting of BioSNAP is also very competitive.

We need to point out that the model performance under the unseen drug setting only dropped slightly compared to that under the random split for all methods on BindingDB. This is because there are many highly similar molecules in the DTI datasets, and naive unseen drug setting does not distinguish them. A better strategy is the clustering-based split strategy in our previous study to alleviate this issue, leading to a more challenging cross-domain task.

6. Performance comparison with high fraction of missing data

Table 7. Performance comparison (average AUROC over five random runs) on the BindingDB and BioSNAP datasets with high fraction of missing data (**Best**, Second Best)

Missing (%)	DeepConv-DTI ⁸	GraphDTA ⁹	MolTrans ¹	DrugBAN
	BindingDB			
95	0.773±0.005	0.831±0.002	<u>0.846±0.004</u>	0.856±0.003
90	0.840±0.002	0.867±0.002	<u>0.874±0.003</u>	0.887±0.004
80	0.877±0.002	0.897±0.003	<u>0.905±0.001</u>	0.920±0.003
70	0.890±0.005	0.916±0.002	<u>0.923±0.001</u>	0.934±0.001
	BioSNAP			
95	0.710±0.005	<u>0.768±0.005</u>	0.767±0.006	0.770±0.008
90	0.781±0.003	0.798±0.003	<u>0.800±0.004</u>	0.802±0.003
80	0.816±0.003	0.829±0.003	<u>0.835±0.001</u>	0.836±0.002
70	0.839±0.002	0.851±0.002	<u>0.853±0.002</u>	0.860±0.003

We conduct experiments to clarify how the proposed model performs with high fraction of missing data on BindingDB and BioSNAP. Following the missing data setting in MolTrans¹, we train DrugBAN and deep learning baselines with only 5%, 10%, 20% and 30% of each dataset, and evaluate predictive performance on the rest of data (90% as test set and 10% as validation

65 set for determining early stopping). Table 7 presents the obtained results, showing DrugBAN has the best performance in all
66 settings. In particular, the improvement is larger on the bigger dataset (BindingDB).

67 7. Scalability

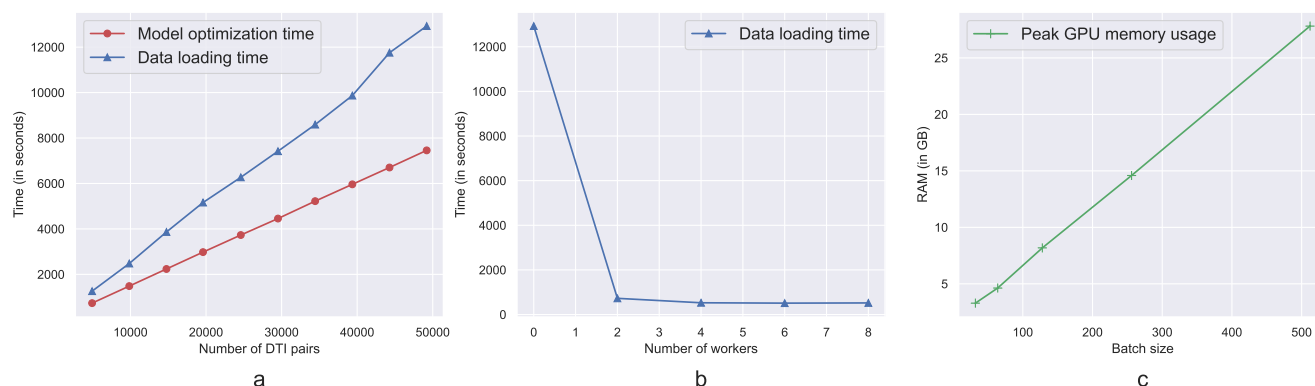


Figure 3. Scalability of DrugBAN on the BindingDB dataset (a) Model optimization and data loading time increase almost linearly with the number of DTI pairs. (b) Data loading time significantly reduces with the increasing number of workers. (c) Peak GPU memory usage increases linearly with the batch size.

68 We study the scalability of DrugBAN from three different perspectives: model optimization time, data loading time and
69 GPU memory usage. We use the default hyperparameter configuration in Table 4, and a single Nvidia V100 GPU to train the
70 model in 100 epochs. Figure 3a illustrates the model optimization time and data loading time against the number of DTI pairs
71 for 4,919 (10%) - 49,199 (100%) from the BindingDB dataset. We empirically observe that the optimization time (red line) of
72 DrugBAN increases almost linearly with the number of DTI pairs. It takes about two hours for 49,199 DTI pairs to complete
73 the optimization. The data loading process (blue line) takes more time than model optimization. Nevertheless, since the data
74 loading can be done on CPU, we can accelerate the process with multiple loading workers (subprocesses) in parallel. Figure 3b
75 shows the data loading time changes with respect to the number of workers, and it reduces significantly with only two additional
76 workers added. Figure 3c shows the peak GPU memory usage against the batch size. We find that DrugBAN only takes up 4.63
77 GB RAM with the default batch size 64, which is highly efficient. Similar to the optimization time, the memory usage also
78 increases linearly with the batch size. This study demonstrates the scalability of DrugBAN.

79 References

- 80 1. Huang, K., Xiao, C., Glass, L. & Sun, J. MolTrans: Molecular interaction transformer for drug–target interaction prediction.
81 *Bioinformatics* **37**, 830 – 836 (2021).
- 82 2. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound–protein interaction prediction by building up highly
83 credible negative samples. *Bioinformatics* **31**, i221 – i229 (2015).
- 84 3. Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems
85 pharmacology. *Nucleic acids research* **44**, D1045–D1053 (2016).
- 86 4. Bai, P. *et al.* Hierarchical clustering split for low-bias evaluation of drug–target interaction prediction. *2021 IEEE Int. Conf.*
87 *on Bioinforma. Biomed. (BIBM)* 641–644 (2021).
- 88 5. Gao, K. Y. *et al.* Interpretable drug target prediction using deep neural representation. In *IJCAI*, 3371–3377 (2018).
- 89 6. Wang, Z., Liang, L., Yin, Z. & Lin, J. Improving chemical similarity ensemble approach in target prediction. *J. cheminform-*
90 *atics* **8**, 1–10 (2016).
- 91 7. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug–target interaction networks from
92 the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232 – i240 (2008).
- 93 8. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug–target interactions via deep learning with convolution on
94 protein sequences. *PLoS Comput. Biol.* **15** (2019).
- 95 9. Nguyen, T. *et al.* GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**,
96 1140–1147 (2021).