MoDTI: Modular Framework for Evaluating Inductive Biases in DTI Modeling

Roy Henha Eyono *
McGill University &
Mila-Quebec AI Institute
Montréal, QC, Canada

Cas Wognum
Valence Discovery
Montréal, QC, Canada

Emmanuel Noutahi
Valence Discovery
Montréal, QC, Canada

Prudencio Tossou
Laval University &
Valence Discovery
Montréal, QC, Canada

Abstract

Drug-Target Interaction (DTI) prediction is a critical problem in drug discovery, and machine learning (ML) has shown great potential in feature-based DTI prediction. However, selecting an appropriate ML architecture from the vast number of available biomolecular representations is challenging. To address this issue, we propose MoDTI, a modular framework that enables the exploration of three key inductive biases in DTI prediction: protein representation, multi-view learning, and modularity. We evaluate the impact of each inductive bias on DTI prediction performance and compare the performance of MoDTI against existing state-of-the-art models on multiple benchmarks. Our experiments with MoDTI provide valuable insights into the role of modularity, capacity, representation redundancy, and orthogonality in terms of generalization and interpretability. They also enable the provision of general guidelines for the rapid development of more accurate DTI models.

1 Introduction

Characterizing drug-target interactions (DTI) is a critical task in drug discovery with important implications for identifying therapeutic targets, optimizing drug efficacy and selectivity, and uncovering the mechanisms of action of drugs. To achieve this goal, computational methods for DTI prediction (Sachdev & Gupta, 2019; Xu et al., 2021) have become increasingly available, offering a fast, flexible, and scalable alternative to costly wet-lab experiments. In particular, feature-based DTI methods have gained popularity due to the success of machine learning techniques in drug discovery and the growing accessibility of pre-trained protein and molecular representations (Sachdev & Gupta, 2019).

Despite the advances in feature-based DTI prediction, there remains a significant challenge in selecting the appropriate inductive biases that ensure good modelling and generalization. Although several works have studied factors such as protein and molecular featurization, pre-training, and interaction bias, there is still a lack of systematic evaluation of the modeling biases required for generalization in DTI prediction.

In this paper, we explore the impact of several modeling biases, including protein representation, multi-view learning, and modularity, on DTI prediction performance through a modular framework. Our objective is to provide a comprehensive evaluation of the influence of these critical architectural decisions on the accuracy and robustness of DTI prediction. Our contributions include:

- the development of a modular framework (MoDTI) that enables the investigation of various modeling biases such as protein representation, multi-view learning, and modularity.

*Work done as intern at Valence Discovery. Correspondance at roy.eyono@mila.quebec

- an exploration of the effect of these biases on model generalization, with insights into the importance of both orthogonality and redundancy for good performance.

- the provision of guidelines for the development of new and improved DTI models.

2 Related Work

Feature-based DTI prediction: Feature-based DTI methods for predicting drug-target interactions (DTIs) utilize machine learning (ML) models, and differ primarily in their approach to representing bio-molecules (drugs and targets) and capturing their interactions. For instance, DeepConvDTI (Lee et al., 2019a) uses ECFP to represent molecules and a 1D CNN to process protein sequences, then it concatenates the features to predict the interaction using a fully connected neural network. Similarly, DeepDTI (Wen et al., 2017) uses a deep belief network with protein descriptors and molecular fingerprints, while DeepDTA (Öztürk et al., 2018) learns representations from SMILES and protein sequences using 1D CNNs before concatenating them for prediction. Other architectures were considered to process the SMILES and protein sequences in DeepAffinity (Karimi et al., 2019) and DrugVQA (Zheng et al., 2020).

In contrast, GNN–CPI (Tsubaki et al., 2019) and GraphDTA (Nguyen et al., 2021) use GNNs on molecular graphs to learn drug representations. Kim & Shin (2021) proposed a Gated Cross Attention that improves interpretability by constructing explicit interactions between the molecule's atoms and the target amino-acids.

In the aforementioned works, the generalization of DTI models remained limited despite architectural exploration due insufficient training data. In (Nguyen et al., 2022), this cold start problem was addressed through transfer learning from pretrained protein and molecular language models. Likewise, Atas Guvenilir & Doğan (2023) investigated the performance of conventional protein representations versus pretrained ones and found the latter to yield more accurate models. This finding was further confirmed by RapidDTI (Sledzieski et al., 2022) which showed that a low capacity fully-connected neural network significantly improves accuracy and generalization when using a pre-trained protein language models (PLMs). The demonstrated impact of PLMs in that work has inspired our focus on learned protein embeddings as a key modelling factor.

A few recent studies have explored the integration of diverse views of drugs and targets for DTI prediction(Zhang et al., 2017; Shang et al., 2022; Zhang et al., 2022; Agyemang et al., 2020). Among them, (Agyemang et al., 2020) stands out, as they use an approach combining the representations of multiple drugs and targets with a joint view attention mechanism. However, the representations are only learned from the dataset, which limits the ability to generalize across new protein and drug classes. To mitigate that, our architecture allows the consideration of pre-trained networks for the multi-view setting.

Modular architectures: Our proposed modular framework shares similarities with ensemble deep learning and Mixture-of-Experts (MoE), which have been widely used in drug discovery (Yu et al., 2022; Pittala & Bailey-Kellogg, 2019; Wu et al., 2022). Ensemble deep learning combines multiple individual models to achieve better generalization (Ganaie et al., 2022), while MoE models sparsely gather experts and leverage them through gating (Jordan & Jacobs, 1994). It is also known that ensembles better exploit orthogonality while MoEs exploits redundancy and by leveraging them both in MoDTI, we hope to study their importance for DTI modelling and generalization.

Previous studies have investigated the benefits of modularity and sparsity through the lenses of functional specialization in neural networks (Bakhtiari et al., 2021; Mittal et al., 2022), but only a few has been applied to DTI prediction. Herein, MoDTI seeks to leverage inductive biases arising from biomolecular representations and model design choices to improve DTI modeling generalization. It decomposes the problem into specialized sub-tasks and takes advantage of the diversity of biochemical information captured by different protein language representations. The proposed framework also offers interpretability, as the contribution of each module can be isolated and analyzed.

3 THE MODTI FRAMEWORK

To investigate the impact of various inductive biases in DTI, we propose the Modular DTI (MoDTI) framework. Within MoDTI, architectures particular to various inductive biases can be instantiated with different high-level parameters allowing to easily investigate those biases. In Section 3.1, we will first define the DTI problem, and then explain the MoDTI template architecture in Section 3.2.

3.1 DTI PREDICTION PROBLEM STATEMENT

DTI prediction consists of learning a function f that predicts the binding affinity from the drug-target pair. Let $\mathcal{D} = \{(x_i, y_i)\}^n$ be our training dataset of n experimentally tested drug-target pairs (x_i) and their associated binding affinities(y_i). In the pair $x_i = (d_i, t_i)$, d_i is the drug and t_i the target or *protein*). In the multi-view setting, both inputs will be transformed respectively by a set of molecular descriptors $L = \{l_1, ..., l_{|L|}\}$ and a set of protein featurizers $P = \{p_1, ..., p_{|P|}\}$. Then, combinations of these featurizers will be used to define a set of *views* $V = \{v_k\}^{|V|}$ to learn from. As such, the k-th view of a given drug-target pair x_i is given by $v_k(x_i) = (l_k(d_i), p_k(t_i))$, where $(l_k, p_k) \in L \times P$.

From the above, we can infer a single-view learning setting by considering $|V| = 1$. By changing the number of views $|V|$, the MoDTI framework can be used to investigate how this inductive bias and others contribute to the accuracy and generalization of f. In the following section, we detail the template architecture that forms the basis of our work.

3.2 TEMPLATE ARCHITECTURE

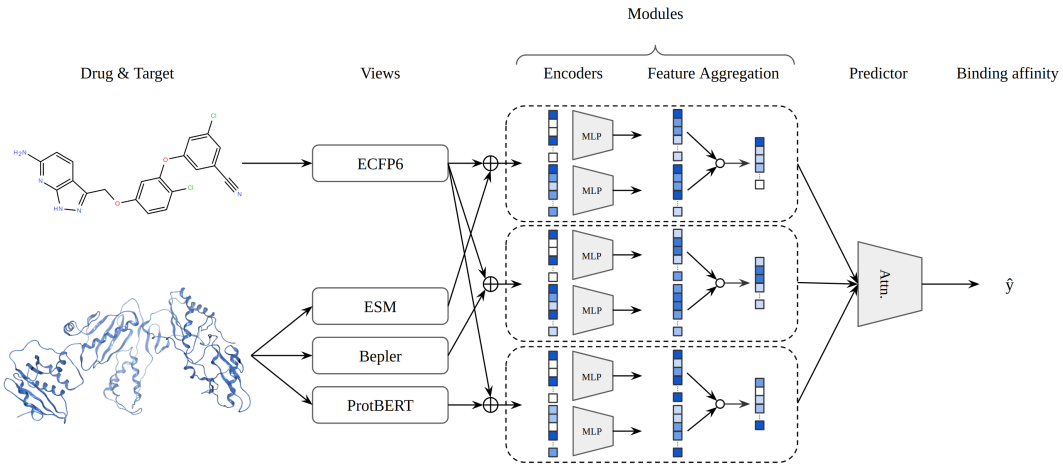


Figure 1: Illustration of the MoDTI template architecture. By varying the number of modules, the number of views and whether the modules are trained independently or jointly, this architecture template allows us to instantiate and investigate several model types.

At the center of this study is the MoDTI architecture template (see Figure 1) which consists of the following components:

- the input *views*, i.e. transformations $v_k \in V(\subseteq L \times P)$ that featurize the drug-target pairs. The considered views are discussed further in Section 4.1.
- the *modules* that map a drug-target *view* into a joint representation. Multiple modules can be used to process a single view and extract different representations. Each module comprises two subcomponents: the *encoders* and the *feature aggregator*. The *encoders* take in the featurized drug or target and independently transforms them into two latent-space embeddings. We use fully-connected neural networks as our encoders throughout our experiments. The *feature aggregator* combines the two latent-space representations into a single vector. Herein, we use a simple Hadamard Product as the feature aggregator.

- Finally, the *predictor* combines all latent representations from all modules and predicts the binding affinity. For this component, we use the attention-based Multi-Instance Learning (MIL) (Ilse et al., 2018).

Note that the flexibility of the MoDTI architecture allows for alternative choices of *encoders*, *feature aggregator*, and *predictor*. However, for the sake of simplicity and to focus our investigation on the inductive biases mentioned earlier, we have fixed these components based on what performed well in our early experiments.

Using this template, we can easily instantiate various models with the inductive biases we are interested in investigating. For instance, by denoting M the number of modules, we can retrieve :

- a Mixture-of-Experts (MoE) model when $M > 1$ and $|V| = 1$. We refer to this instance of MoDTI as **MoDTI-SV**. Likewise, the monolithic case where $M = 1$ and $|V| = 1$, will be referred as **MoDTI-SV-Mono**.
- a multi-view learning setup when $M > 1$ and $|V| > 1$. We will call this version of MoDTI: **MoDTI-MV**.
- an ensemble of several single-module **MoDTI-SV-Mono** allows us to retrieve an ensemble setup, which we will later refer to as **MoDTI-SV-Ens**.

4 RESULTS

4.1 EXPERIMENT SETUP

Benchmark Datasets: Throughout our experiments, we use three DTI classification benchmark datasets, namely BioSNAP, BindingDB (Liu et al., 2007), and DAVIS. Statistics for each dataset as well as the training, validation, and testing partitions are presented in Appendix A.1. Following previous studies (Sledzieski et al., 2022), we use both Area under the Precision-Recall curve (AUPR) and the Area Under the Receiver Operating Characteristic curve (AUROC) for performance evaluation.

Protein Featurization: We considered the following Protein Language Models (PLM): Bepler & Berger (Bepler & Berger, 2019), ESM (Rives et al., 2021), and ProtBERT (Elnaggar et al., 2021). They respectively transform the proteins into feature vectors of size 6165, 1280, and 1024. Baseline featurizations (Word2Vec and One-Hot-Encoding) were also used to better gauge PLMs contribution to performance in the modular setting. The three PLMs were selected to bring different perspectives on proteins representations based on their differing architectures, training objectives, and training datasets. Appendix A.2 provides more details on these differences.

Molecular featurization: Since we are mainly investigating the effect of protein language models, we use Extended-Connectivity Fingerprints with radius 3 (ECFP6) and 2048 bits (Morgan, 1965) for molecular representation.

Experimental design: We used the binary cross entropy (BCE) loss and optimized model weights with the Adam optimizer (lr $= 10^{-4}$) for a maximum of 50 epochs, with a batch size of 32.

4.2 BENCHMARKING AGAINST THE SOTA

Our experiments start by comparing MoDTI against multiple DTI baselines, namely RapidDTI (Sledzieski et al., 2022), MolTrans (Huang et al., 2021), GNN-CPI (Tsubaki et al., 2019), and DeepConv-DTI (Lee et al., 2019b). The performance of MoDTI was obtained from a grid search over combinations of protein representations, but no module parameter optimization was performed. Table 1 shows that, for all three datasets, both instances of MoDTI outperform the DTI baselines including the most recent SOTA model to our knowledge (Sledzieski et al., 2022).

It is worth observing that the Multi-View MoDTI (ModDTI-MV) slightly outperforms the Single-View MoDTI (MoDTI-SV) in two out of three datasets (DAVIS and BindingDB). The small difference in performance between MoDTI-SV and MoDTI-MV prompt many questions about the role of the architecture and the views in the overall performance. Therefore, in the following sections, we will investigate the contribution of views and modularity to the generalization performance. Our

aim is to unravel the modelling factors that lead to accurate DTI prediction and provide insights for future modelling decisions.

	DAVIS		BioSNAP		BindingDB	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
RapidDTI	0.481±.001	0.917±.003	0.895±.001	0.885±.001	0.623±.009	0.882±.002
MolTrans	0.335±.017	0.889 ±.007	0.885±.005	0.876±.007	0.598±.013	0.898±.009
GNN-CPI	0.269±.020	0.840±.012	0.890±.004	0.879±.007	0.578±.015	0.900±.004
DeepConv-DTI	0.299±.039	0.937±.004	0.889±.005	0.883±.002	0.611±.015	0.908±.004
MoDTI-SV	0.489±.006	0.936±.000	**0.923±.002**	**0.922±.001**	0.663±.007	0.924±.001
MoDTI-MV	**0.508±.008**	**0.940±.001**	0.920±.001	0.920±.001	**0.663±.003**	**0.926±.001**

Table 1: Benchmarking performance: Modular architectures outperform monolithic single view models over three random initializations. MoDTI-SV refers to the case where all module uses the same view/representation as input, and MoDTI-MV refers to the multi-view case which require at least two different views.

4.3 The Multi-View Inductive Bias

In this section, we investigate the inductive bias of multiple views for DTI modelling. More precisely, we examine how the number of views (quantity) and their information content (quality) affect performance.

4.3.1 Effect of views Redundancy vs Orthogonality

Taking the perspective of quality and information content, we measure the differences between redundant and orthogonal views towards DTI prediction. Recall that for MoDTI-SV models, there is redundancy (the views are repeated) while for MoDTI-MV models, many views are orthogonal. In Figure 2, we compare the best MoDTI-SV and MoDTI-MV models across the same number of modules. Figure 2 shows a clear trend in performance, where we observe that MoDTI-MV outperforms MoDTI-SV as we gradually increase the number of modules. This increase is up to 10% and 50% for AUROC and AUPR, respectively. These results suggest that *as we scale the number of views, the orthogonality of views becomes essential in improving performance*.

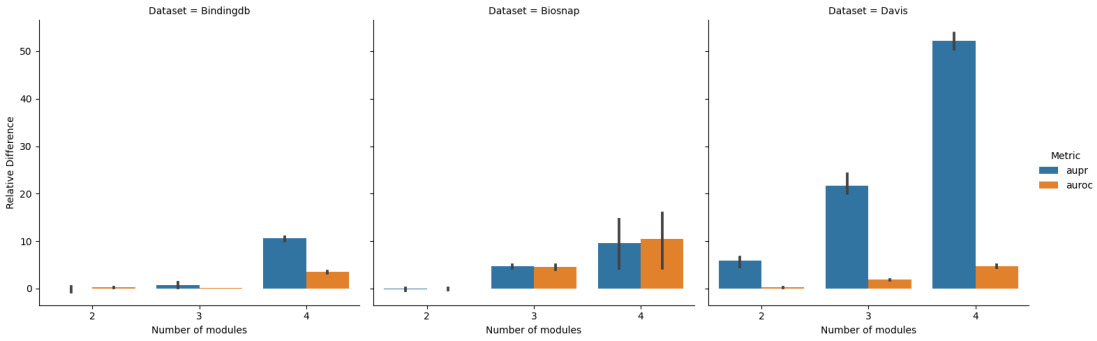


Figure 2: Difference in AUPR and AUROC between the best multi-view and single-view model. For all three datasets, we observe that the MoDTI model increasingly benefits from the orthogonal views compared to redundant views in MoDTI-SV. The error bars show the 95% confidence interval over three different seeds.

4.3.2 Scaling Orthogonal Views:

In this section, we investigate the impact of scaling the number of views while maintaining their orthogonality. Specifically, we aim to measure the performance gain obtained by incorporating a novel view into a given MoDTI model.

Figure 3 shows an upwards trend in AUPR as the number of orthogonal views increases, indicating that scaling the number of orthogonal views in our MoDTI framework will improve its performance.

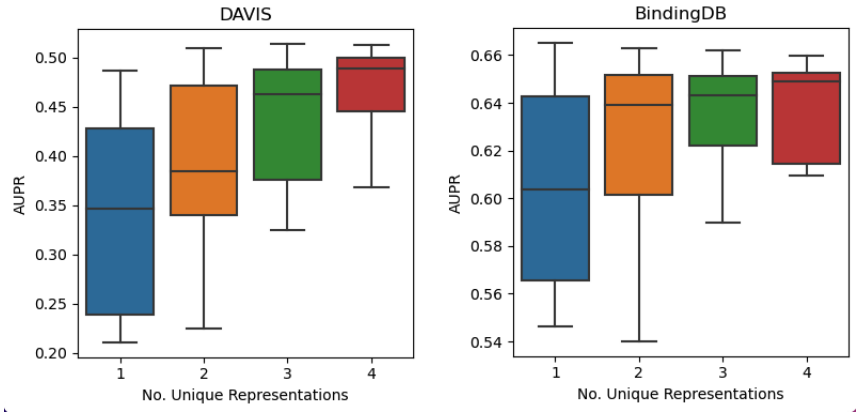


Figure 3: AUPR performance for best modules trained with unique and multiple-views for datasets, DAVIS and BindingDB. We observe a positive trend in the average AUPR across both datasets. In Appendix A.3, we report similar results for the BioSnap Dataset.

However, this also raises questions about the contribution of each PLM and which type of protein information is crucial for generalization.

To address this question, we examine the contribution of each view to the AUPR and AUROC metrics in Figure 5. We do this by fitting a linear regression model on the views (one-hot) and measuring their impact with respect to the held-out test. From this experiment, we observe that ESM and ProtBERT are significantly more impactful towards generalization than other views, with ESM contributing the most and in some instances twice as much as ProtBert. In contrast, Word2Vec and the One Hot embedding have slightly negative importance measures, which further validates the empirical benefits of protein language models over traditional word embedding methods. We expand on the impact of ESM and ProtBert on MoDTI in Figure 4 which shows the AUPR of different multi-view models that contain ESM, ProtBert, or neither. This analysis highlights the importance of the richness of information provided by the protein representations in DTI modelling and hints at it may be of greater importance than the number of views considered.

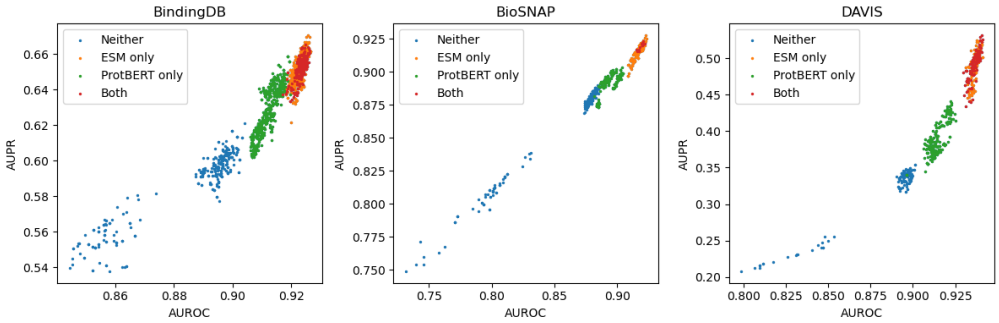


Figure 4: Impact of ESM and ProtBert on the AUPR and AUROC of multi-views multi-module models. Each dot represents the AUROC and AUPR of a different multi-view combination and the colors inform whether the combination include only ESM, only ProtBert, both or none. It confirms the dominance of both views in this study.

To summarize, our results demonstrate that both the quantity and quality of views are important for DTI prediction, with the latter being more crucial. We observe that increasing the number of orthogonal views leads to a performance improvement in MoDTI models, compared to increasing the number of redundant views. These findings suggest that the use of multiple orthogonal views provides an advantageous inductive bias for DTI prediction.”.

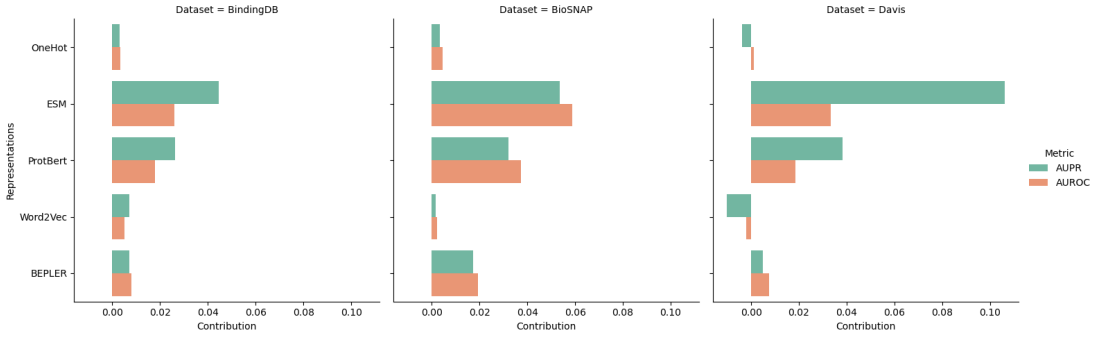


Figure 5: Contribution of each view to the test AUPR and AUROC. These values are the weights of each view in a linear regressor model learned with an intercept. This linear model is trained to predict a multi-module model performance from a count vector representing which views its received.

4.4 The Modularity Inductive Bias

One important distinction between ensemble methods and modularity is that modular networks are trained end-to-end. In contrast, ensemble methods are composed of base learners combined with an aggregator (e.g. mean). In this section, we compare the effect of modularity against ensemble learning to understand the inductive bias of modularity for DTI modelling.

4.4.1 Modularity versus Ensembling

To investigate the inductive bias of modularity, we build equivalent ensemble models using our MoDTI framework. The ensemble models (MoDTI-Ens) are composed of base learners (MoDTI-SV-Mono) which are trained individually on a specific view. All base-learners contribute equally in an ensemble, however in a MoDTI-MV, the predictions are determined by an attention mechanism.

	DAVIS		BioSNAP		BindingDB	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
MoDTI-SV-Ens (N=2)	**0.4936**	**0.9381**	**0.9273**	**0.9241**	0.6712	0.9262
MoDTI-MV-Ens (N=2)	0.4901	0.9359	0.9267	0.9223	**0.6727**	**0.9248**
MoDTI-SV-Ens (N=3)	**0.4984**	**0.9379**	**0.9287**	**0.9252**	**0.6726**	**0.9267**
MoDTI-MV-Ens (N=3)	0.4933	0.9359	0.9297	0.9209	0.6678	0.9211
MoDTI-SV-Ens (N=4)	**0.4956**	**0.9378**	**0.9301**	**0.9259**	**0.6745**	**0.9269**
MoDTI-MV-Ens (N=4)	0.4493	0.9312	0.9206	0.9082	0.6630	0.9142

Table 2: Ensemble models performance for different number of learners. It shows that SV ensemble models outperform MV ensemble models in most cases.

We present the performance of the ensemble models in Table 2. We observe that the single-view ensemble models outperform their multi-view counterparts with largely marginal, yet in some cases, significant differences.This contrasts with our findings on modularity with multiple views in Figure 2 where MoDTI-MV models are better than their counterpart MoDTI-SV. The contrast suggests that modular architectures might be better suited for orthogonal views compared to ensemble models and that might be a consequence of the differences in module contributions towards the predictions. MoDTI-MV might reduce the contributions of low-quality and thus noisy views.

Further comparison between Table 1 and Table 2 shows that MoDTI-Ens performs better than MoDTI-MV, except on the DAVIS dataset. The disparity in performance could be attributed to the end-to-end training of modules in MoDTI-MV, which is less parameter-efficient for training.

4.4.2 Modularity and Interpretability

Despite the disparity between ensemble models and modularity, we would like to investigate whether modularity lends itself naturally to interpretability. To study the interplay between modularity and

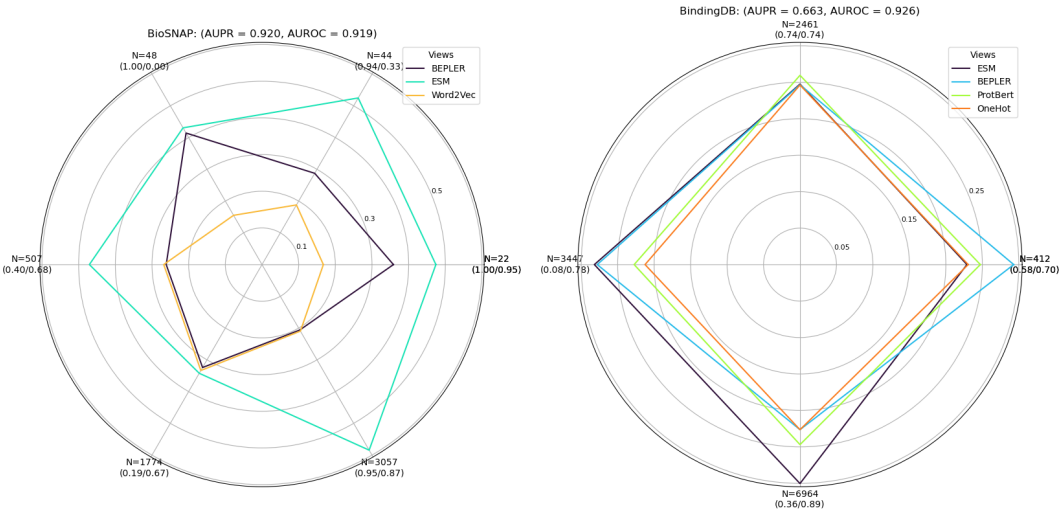


Figure 6: Cluster of the attention weights for models trained on BioSNAP and BindingDB. Each cluster is represented by a cardinal and the corresponding radii indicates the attention weight towards a particular module (i.e ESM). The metrics for each cluster are reported as "(AUPR/AUROC)".. The top caption on each plot is the overall performance of the model. (**Left**) We observe that instances from the largest cluster ($N = 3057$) in the BioSNAP model are highly attended to by the ESM module. (**Right**) Similarly, we observe for the BindingDB model, that its largest cluster ($N = 6964$) is also highly attended to by the ESM module.

interpretability within MoDTI (see Section 3.2), we measure the attention weights in MoDTI models. We cluster the attention weights using a density-based clustering algorithm and analyze the weights of each cluster to observe if specialization emerges and how this correlates with performance.

Our results (see Figure 6 & Appendix A.4), show that the clusters with high AUPR or high AUROC often have more weights on important features while views such as Word2Vec contribute negligibly to the predictions. These observations further demonstrate that MoDTI primarily leverages the most informative views for its predictions. They also suggest a weak specialization that emerges at the representation level. It is worth further investigating, in future work, if there are factors like protein and scaffold families that the network is specializing to.

5 Conclusion and Future Work

In this study, we examined three modelling biases that affect DTI model performance: protein featurization, multi-view learning, and modularity. Our MoDTI allowed to demonstrate that multi-view DTI learning is sensitive to architectural inductive biases. Specifically, we found that *Modular architectures are better at leveraging orthogonal views than ensemble models, which excel at exploiting redundant views.* Our findings suggests that modular architectures that exploit the orthogonality of different views to improve performance could be more advantageous for DTI. Improving the mechanism for leveraging multiple perspectives could also significantly enhance overall performance. In future work, we will focus on how to maintain the right balance of orthogonality and redundancy in multi-view DTI learning with end-to-end training of modular architectures. Overall, our study highlights the importance of considering the underlying biases of different DTI models and architectures, and provides a foundation for further research into improving the performance and interpretability of these models.

REFERENCES

Brighter Agyemang, Wei-Ping Wu, Michael Yelpengne Kpiebaareh, Zhihua Lei, Ebenezer Nanor, and Lei Chen. Multi-View Self-Attention for Interpretable Drug-Target Interaction Prediction. May 2020. doi: 10.1016/j.jbi.2020.103547. URL https://arxiv.org/abs/2005.00397v2.

Heval Atas Guvenilir and Tunca Doğan. How to approach machine learning-based prediction of drug/compound–target interactions. *Journal of Cheminformatics*, 15(1):1–36, 2023.

Shahab Bakhtiari, Patrick Mineault, Tim Lillicrap, Christopher C. Pack, and Blake A. Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning, October 2021. URL https://www.biorxiv.org/content/10.1101/2021.06.18.448989v3. Pages: 2021.06.18.448989 Section: New Results.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure, October 2019. URL http://arxiv.org/abs/1902.08661. arXiv:1902.08661 [cs, q-bio, stat].

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing, May 2021. URL http://arxiv.org/abs/2007.06225. arXiv:2007.06225 [cs, stat].

M. A. Ganaie, Minghui Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, October 2022. ISSN 09521976. doi: 10.1016/j.engappai.2022.105151. URL http://arxiv.org/abs/2104.02395. arXiv:2104.02395 [cs].

Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. *Bioinformatics (Oxford, England)*, 37(6):830–836, May 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa880.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. June 2018. doi: 10.48550/arXiv.1802.04712. URL http://arxiv.org/abs/1802.04712. arXiv:1802.04712 [cs, stat].

Michael I. Jordan and Robert A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2):181–214, March 1994. ISSN 0899-7667. doi: 10.1162/neco.1994.6.2.181. URL https://doi.org/10.1162/neco.1994.6.2.181.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Yeachan Kim and Bonggun Shin. An interpretable framework for drug-target interaction with gated cross attention. In *Machine Learning for Healthcare Conference*, pp. 337–353. PMLR, 2021.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129, 2019a.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology*, 15(6):e1007129, June 2019b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007129. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007129. Publisher: Public Library of Science.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(Database issue):D198–201, January 2007. ISSN 1362-4962. doi: 10.1093/nar/gkl999.

Sarthak Mittal, Yoshua Bengio, and Guillaume Lajoie. Is a Modular Architecture Enough?, June 2022. URL http://arxiv.org/abs/2206.02713. arXiv:2206.02713 [cs].

H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 0021-9576. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018. Publisher: American Chemical Society.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

Tri Minh Nguyen, Thin Nguyen, and Truyen Tran. Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring. *Briefings in Bioinformatics*, 23(4): bbac269, 2022.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Srivamshi Pittala and Chris Bailey-Kellogg. Mixture of experts for predicting antibody-antigen binding affinity from antigen sequence. *bioRxiv*, pp. 511360, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):e2016239118, April 2021. ISSN 1091-6490. doi: 10.1073/pnas.2016239118.

Kanica Sachdev and Manoj Kumar Gupta. A comprehensive review of feature based methods for drug target interaction prediction. *Journal of biomedical informatics*, 93:103159, 2019.

Yifan Shang, Xiucai Ye, Yasunori Futamura, Liang Yu, and Tetsuya Sakurai. Multiview network embedding for drug-target interactions prediction by consistent and complementary information preserving. *Briefings in Bioinformatics*, 23(3):bbac059, 2022.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. Adapting protein language models for rapid DTI prediction, November 2022. URL https://www.biorxiv.org/content/10.1101/2022.11.03.515084v1. Pages: 2022.11.03.515084 Section: New Results.

Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, January 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bty535. URL https://academic.oup.com/bioinformatics/article/35/2/309/5050020.

Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16(4): 1401–1409, 2017.

Jialu Wu, Junmei Wang, Zhenxing Wu, Shengyu Zhang, Yafeng Deng, Yu Kang, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Alipsol: An attention-driven mixture-of-experts model for lipophilicity and solubility prediction. *Journal of Chemical Information and Modeling*, 62(23): 5975–5987, 2022.

Lei Xu, Xiaoqing Ru, and Rong Song. Application of machine learning for drug–target interaction prediction. *Frontiers in Genetics*, 12:680117, 2021.

Tzu-Hui Yu, Bo-Han Su, Leo Chander Battalora, Sin Liu, and Yufeng Jane Tseng. Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying cns drugs with high prediction power. *Briefings in bioinformatics*, 23(1):bbab377, 2022.

Xin Zhang, Limin Li, Michael K. Ng, and Shuqin Zhang. Drug–target interaction prediction by integrating multiview network data. *Computational Biology and Chemistry*, 69:185–193, August 2017. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2017.03.011. URL https://www.sciencedirect.com/science/article/pii/S1476927117301950.

Yuanyuan Zhang, Mengjie Wu, Shudong Wang, and Wei Chen. Efmsdti: Drug-target interaction prediction based on an efficient fusion of multi-source data. *Frontiers in Pharmacology*, 13, 2022.

Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2): 134–140, 2020.

A Appendix

A.1 Dataset Statistics

Table 3 presents the statistics for each dataset as well as the training, validation, and testing partitions. Following (Sledzieski et al., 2022), it shows an effort to keep the training set balanced even if it results in unbalanced testing and validation sets.

Dataset	Unique Drugs	Unique Proteins	Training Pairs	Validation Pairs	Test Pairs
BioSNAP	4510	2181	9619/9619	1374/1374	2748/2748
BindingDB	10665	1413	6334/6334	927/5717	1905/11384
DAVIS	68	379	1043/1043	160/2846	303/5708

Table 3: Benchmark Datasets. Pairs are distinguished as (positive/negative) binding pairs.

A.2 Difference in information captured by the Protein Language Models

We assessed the difference in information captured by each PLMs by computing the correlation between pairwise similarity for any protein pair given each PLM representation (Figure 7). The low Pearson correlation (less than 0.37), is indicative that these different protein views encode orthogonal information.

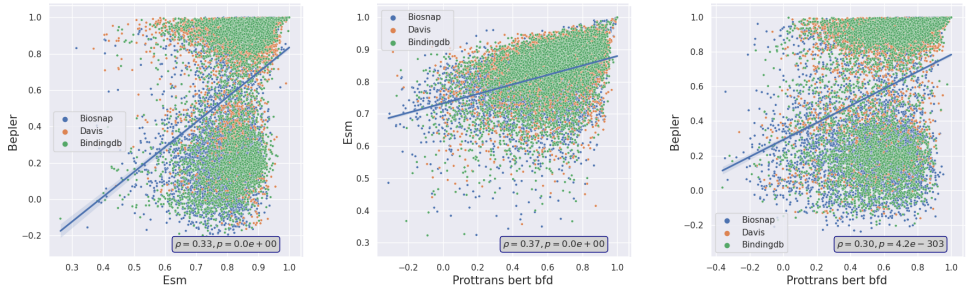


Figure 7: Similarity analysis of protein views. Each point represents a pair of proteins and the axes show the cosine distance between those proteins in different protein view spaces. If protein pairs are equally (dis)similar according to two different views, we consider the views to be similar. We argue that the different views encode orthogonal information about the protein.

A.3 Orthogonal Views:

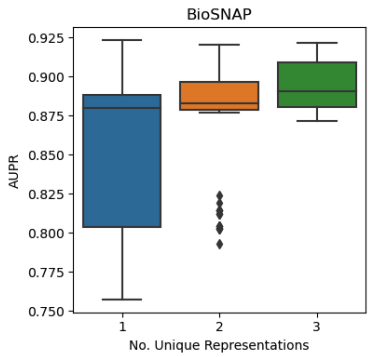
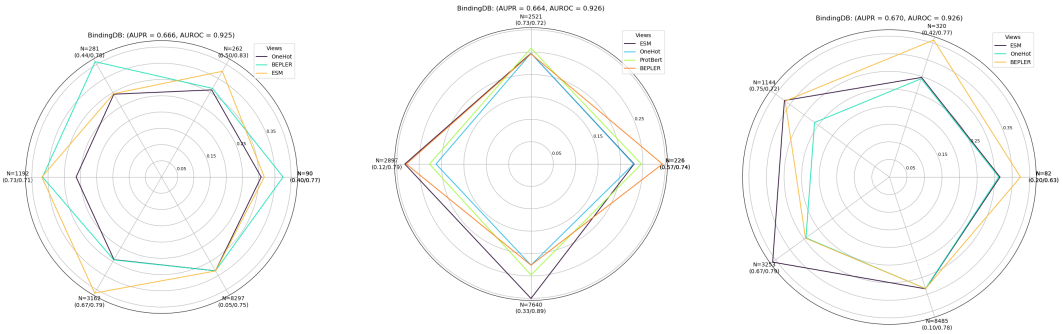
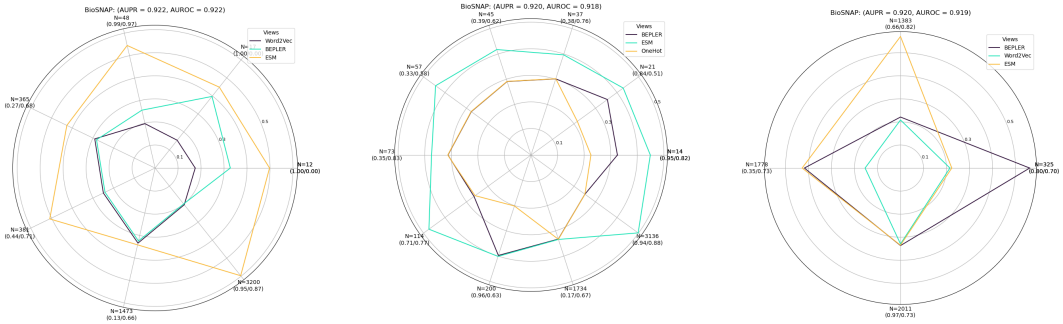


Figure 8: AUPR performance for best modules trained with unique and multiple-views for the BioSnap dataset. Here, we only run experiments for up to three unique representations.

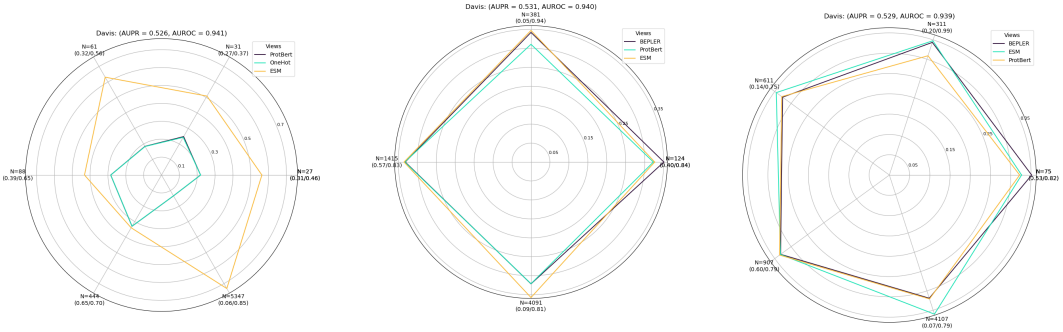
A.4 ATTENTION WEIGHTS



(a) BindingDB



(b) BioSNAP



(c) DAVIS

Figure 9: Average attention weights per cluster across three benchmark datasets.