IMPROVING THE ASSESSMENT OF DEEP LEARNING MODELS IN THE CONTEXT OF DRUG-TARGET INTERACTION PREDICTION

Mirko Torrisi, Antonio de la Vega de León, Guillermo Climent, Remco Loos, Alejandro Panjkovich
Center for Innovation and Translational Research Europe (CITRE)
Bristol Myers Squibb
Seville, Spain
{mirko.torrisi,alejandro.panjkovich}@bms.com

ABSTRACT

Machine Learning techniques have been widely adopted to predict drug-target interactions, a central area of research in early drug discovery. These techniques have shown promising results on various benchmarks although they tend to suffer from poor generalization. This is typically related to very sparse and nonuniform datasets available, which limits the applicability domain of machine learning techniques. Moreover, widespread approaches to split datasets (into training and test sets) treat a drug-target interaction as an independent entities, when in reality the drug and target involved may take part in other interactions, breaking apart the assumption of independence. We observe that this leads to overly optimistic test results and poor generalization of out-of-distribution samples for various state-of-the-art sequence-based machine learning models for drug-target prediction. We show that previous approaches to reduce bias in binding datasets focus on drug or target information only and, thus, lead to similar pitfalls. Finally, we propose a minimum viable solution to evaluate the generalization capability of a machine learning model based on the systematic separation of test samples with respect to drugs and targets in the training set, thus discerning the three out-of-distribution scenarios seen at test time: (1) drug or (2) target present in the training set, or (3) neither.

1 INTRODUCTION

Recently developed deep learning methods to predict drug-target interactions (DTI) hold great potential for drug discovery (Bagherian et al., 2021). In spite of the interest sparked both in academia and industry towards these methods, the lack of standard pipelines or benchmarking criteria discourages adoption and further developments. Particularly, as we show here, it is important to assess deep learning models beyond usual predictive measures and increase scrutiny on what the models are learning.

Data quantity and quality play a fundamental role in developing and applying deep learning models successfully. In the case of DTI, multiple databases are publicly available and focus on different characteristics of the data, such as: quality of three-dimensional structure (Binding MOAD (Smith et al., 2019)), manual curation of annotations (PDBbind (Liu et al., 2015)), or sparsity, i.e. limiting the number of drugs and targets (ExcapeDB (Sun et al., 2017)). Here, we perform an independent benchmark of DTI predictors based on BindingDB (Liu et al., 2007), a database collecting interactions from scientific articles and patents, which has already been used to create DTI benchmark datasets by others (Yingkai Gao et al., 2018; Karimi et al., 2019).

Beyond the specifics of each dataset, a crucial aspect for developing and assessing a DTI predictor is how to divide or split available data into testing and training sets, while preserving a balanced representation of the interaction space. This has been demonstrated to be particularly challenging when using chemical data for machine learning, given the amplitude of chemical space and the inherent biases present in the sparse and nonuniform DTI benchmark datasets available (Sieg et al., 2019).

Moreover, deep learning models are capable of easily fitting nuances observed in the training set, i.e. from technical noise to random labeling (Zhang et al., 2021). Although some approaches have been deployed for evaluating the redundancy between training and test sets of drug-based (Wallach & Heifets, 2018), and target-based (Urban et al., 2020) models, none of these approaches is directly applicable for DTI models. Thus, state-of-the-art DTI predictors may learn biases observed in DTI benchmark datasets instead of chemical and physical features characterizing potential interactions.

In this work, we benchmark three recently developed DTI methods and observe competitive results on two previously defined benchmark datasets. However, the models show relatively poor generalization on four Out Of Distribution (OOD) scenarios. Intuitively, a DTI deep learning model will generalize to OOD scenarios by learning fundamental chemical and physical features that govern the interaction between a drug and a target at the molecular level. To further understand the generalization challenge, we investigated if considering interactions as independent entities, as done previously when splitting the training and test samples in these defined benchmark datasets, may constitute a potential source of information leakage which, in turn, may hinder the learning and evaluation of DTI predictors. With a baseline analysis, we show that the association of DTI in these benchmark datasets by drug or target alone contains information and predictive power and, thus, constitutes a case of information leakage.

To improve the assessment of DTI predictors, we propose a minimum viable solution for measuring their generalization capability, i.e. discerning the three OOD scenarios seen at test time: (1) drug or (2) target present in the training set, or neither (3). Our solution strengthens the assessment of DTI predictors evaluating empirically their ability to generalize to OOD scenarios, without altering the training set, while facilitating the identification of potential sources of bias in benchmark datasets.

To summarize, we show that considering DTI as independent entities introduces potential sampling biases which may hinder the prediction and generalization capability of a DTI model. Correspondingly, we propose a simple approach to gauge the generalization capability of a DTI model emphasizing the evaluation on OOD scenarios.

2 METHODS

2.1 DATASETS

All the datasets in this work are derived from BindingDB (Liu et al., 2007): a publicly accessible and regularly updated collection of binding affinity values between proteins considered to be drug-targets, and drug-like molecules. In particular, we adopt two benchmark datasets derived from BindingDB, one released by Yingkai Gao et al. (2018) and the other as defined by Karimi et al. (2019), which have been used for benchmarking recent DTI predictors (Chen et al., 2020; Born et al., 2022). Both benchmark datasets are outlined in Table 1.

The first version, which we name *BindingDB*^{S}, contains 59,136 interactions (involving 48,084 drugs and 798 targets), distributed in training (*train*^{S}), validation (*val*^{S}), test (*test*^{S}) set. It was assembled with the intent of simulating practical scenarios, i.e. "given a pair of drug and target at testing time, the drug, the target, or both of them may have not been observed at training time" (Yingkai Gao et al., 2018). We remove from the original sets any drug and target exceeding the length constraints of DeepAffinity (Karimi et al., 2019) (see Section 2.2).

The second version, *"BindingDB*^{L}*"*, contains 472,925 interactions (involving 321,950 drugs and 3,350 targets), distributed in training (*train*^{L}), validation (*val*^{L}), test (*test*^{L}), and 4 OOD sets, i.e. ion Channels (*test*^{Ch}), nuclear Estrogen Receptors (*test*^{ER}), G-Protein-Coupled Receptors (*test*^{GPCR}), and Receptor Tyrosine Kinases (*test*^{RTK}). This collection of datasets was released for assessing various deep learning techniques and their generalization capability, constraining four protein families only in the respective OOD sets (Karimi et al., 2019), i.e. leaving ion Channels, nuclear Estrogen Receptors (*test*^{ER}), G-Protein-Coupled Receptors, and Receptor Tyrosine Kinases out of the training set. Although none of the method in this work needs a validation set, we follow the approach used by the authors of DeepAffinity (Karimi et al., 2019), and randomly select 10% of the samples in the training set to split the validation set.

Table 1: Number of unique drugs, targets, and interactions observed in the benchmark datasets, i.e. *BindingDB*^{S} and *BindingDB*^{L}, as well as sparsity and ratio of Positive/Negative samples. Values within parenthesis are the overlap with the training set, omitted when equal to 0%, bolded when greater than 50%, and underlined in cases of data leakage.

Dataset	Drugs	Targets	Interactions	Sparsity	P/N
train^S	41,708	741	48,512	0.16%	1.26
val^S	4,882 (30%)	456 (**90%**)	5,376	0.24%	1.02
test^S	4,798 (30%)	455 (**90%**)	5,248	0.24%	0.97
Total (unique entities)	48,084	798	59,136	0.15%	1.21
train^L	180,736	2,632	232,044	0.05%	1.30
val^L	25,215 (39%)	1,778 (**98%**)	26,272 (<u>4%</u>)	0.06%	1.33
test^L	97,358 (37%)	2,370 (**95%**)	111,906 (<u>4%</u>)	0.05%	1.31
test^{Ch}	12,795 (18%)	125 (1%)	14,107	0.88%	1.02
test^{ER}	2,115 (13%)	6	3,228	25%	1.5
test^{GPCR}	48,712 (13%)	313	57,957	0.38%	2.74
test^{RTK}	24,608 (25%)	127 (1%)	33,189	1.06%	3.31
Total (unique entities)	321,950	3,350	472,925	0.05%	1.48

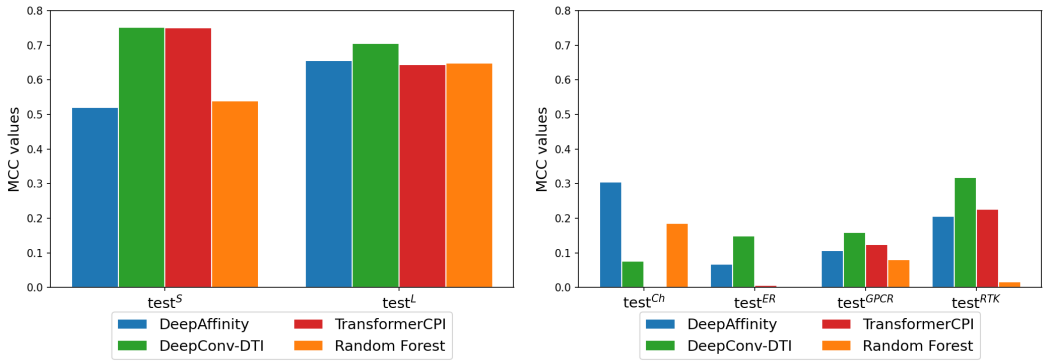
2.2 DTI PREDICTORS

In this work, we assess the predictive performance of three recent deep learning methods, together with a machine learning baseline: DeepAffinity (Karimi et al., 2019), DeepConv-DTI (Lee et al., 2019), TransformerCPI (Chen et al., 2020), and Random Forest (Ho, 1995). All deep learning methods are published along with the code for training and assessing them. We do not perform any hyperparameter optimization and aim to reproduce the methods as described in the respective manuscripts.

DeepAffinity (Karimi et al., 2019) is the only regressor in this work, which makes comparison to the other methods challenging. However, for completeness, we include comparison with this method. To do this, we impose a classification threshold of 0.5 for *BindingDB*^{S} (which was released already binarized), and of 6 for *BindingDB*^{L} (as done previously on *BindingDB*^{S} by Yingkai Gao et al. (2018)). DeepAffinity requires the PDB (Berman et al., 2000), Pfam (Mistry et al., 2021) and UniRef (Suzek et al., 2015) for building the vocabulary of 72 four-mers describing the targets. Drugs are described using an alphabet of 68 letters derived from the corresponding SMILES. Moreover, UniRef50 (Suzek et al., 2015) and Stitch (Szklarczyk et al., 2016) are used for pre-training the embeddings of drugs and targets, respectively. DeepAffinity accepts targets of up to 1,500 amino acids, and SMILES of up to 100 symbols.

DeepConv-DTI (Lee et al., 2019) is the only method not relaying on external databases, nor k-mers, nor pre-training. Although it relays on an embedding for representing protein sequences, such embedding is trained from scratch on the training set. Drugs are represented by Morgan fingerprints of radius 2 (Rogers & Hahn, 2010). It accepts targets of up to 2,500 amino acids.

TransformerCPI (Chen et al., 2020) is inspired from a Transformer architecture (Vaswani et al., 2017), although the drugs are passed in input only to the decoder. Protein sequences are split into overlapping three-mers of amino acids, and then passed to a word2vec embedding (Mikolov et al., 2013) (pre-trained on human proteins from UniProt (The UniProt Consortium, 2017)). Each atom of the drugs is represented via 34 chemical and physical properties calculated with RDKit (RDKit), which are then passed to a graph.



(a) MCC on the main Test Sets

(b) MCC on the OOD Sets

Figure 1: MCC observed on (a) $test^S$ and $test^L$, and on (b) the four OOD sets. All methods provide good predictive performance on the test sets, and lower performance on the OOD sets.

Random Forest (Ho, 1995) is implemented using the code released along with DeepAffinity (Karimi et al., 2019). In particular, protein sequences are represented by a vector of size 2,500 (padded with zeros for shorter sequences), where each amino acid is encoded via a label encoder, i.e. mapped to a number from one to twenty-five. Drugs are represented using Morgan fingerprints of radius 2 (Rogers & Hahn, 2010) calculated with RDKit (RDKit).

2.3 Metrics

We use Matthews Correlation Coefficient (MCC) to measure predictive performance. MCC summarizes the confusion matrix in a single value, and is considered to be more informative than other metrics on imbalanced datasets (Brown, 2018). Specifically, the MCC can be calculated from the confusion matrix as follow:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

with TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives). The MCC can assume values from -1 to 1, where the extremes indicate perfect agreement or disagreement between ground truth and model predictions, and 0 indicates no relationship.

For completeness, we include Accuracy and F1 metrics for all results in this work in the AppendixA.

3 Results

3.1 DTI benchmark

We benchmark DeepAffinity, DeepConv-DTI, TransformerCPI, and a Random Forest on $BindingDB^S$ and $BindingDB^L$. To do so, we train all methods twice to obtain a model for each benchmark dataset.

We observe good performance across the board on $test^S$ and $test^L$, as shown in Figure 1a. DeepConv-DTI achieves a MCC over 0.7, slightly outperforming other methods. Notably, the Random Forest is never the worst method in this setting, and appears to be on par with the deep learning methods.

Lower performance is observed regarding generalization, as measured on the four OOD sets, i.e. $test^{Ch}$, $test^{ER}$, $test^{GPCR}$, and $test^{RTK}$ (see Figure 1b). All methods achieve lower MCC values than on $test^S$ and $test^L$, and exceed a MCC value of 0.3 only in two cases, i.e. DeepAffinity on $test^{Ch}$, and DeepConv-DTI on $test^{RTK}$. DeepConv-DTI achieves the highest MCC values in most

cases, although no method outperforms the competitors on all four OOD sets, and the Random Forest is generally the worst by MCC value.

The results of our benchmark outline that no method outperforms all the other methods assessed, in all benchmarking scenarios. Importantly, all methods achieve good MCC on *test*^{S} and *test*^{L}, but much lower performance on OOD scenarios, pointing towards poor generalization. Regarding DeepAffinity, we observe particularly good predictive performance on *test*^{L} and *test*^{Ch}, although it is challenging to compare the only regressor with the other methods.

3.2 Drugs and Targets distribution

To identify possible sources of overoptimistic performance on *test*^{S} and *test*^{L}, which may justify the poor generalization on *test*^{Ch}, *test*^{ER}, *test*^{GPCR}, and *test*^{RTK}, we look deeper in the composition of the benchmark datasets. In Section 3.2.1, we explore the overlap between training and testing sets, first at the level of interactions and then at the level of their constitutive drugs and targets. In Section 3.2.2, we performed a baseline analysis to measure the potential information leakage because of the overlap of drugs and targets across different interactions between training and testing sets. Our results show that treating DTI as independent entities causes information leakage in the context of train-test splits.

3.2.1 Overlaps across sets

We find overlapping interactions in *BindingDB*^{L}, a simple case of data leakage. Specifically, we observe that 4% of the interactions in both *test*^{L} and *val*^{L} are present in *train*^{L} (see Table 1). We also observe DTI with multiple labels, i.e. the same drug-target pair associated to multiple binding observations. This case of data leakage may be due to the lack of stereochemistry in the SMILES of *BindingDB*^{L}, leading different molecules to be treated as if they were the same.

Beyond direct overlap of interactions between datasets, we also look at interactions not as independent entities, but in terms of their constitutive drugs and targets (see Table 1). In this regard, we find overlap of drugs and targets across all datasets with respect to their training set, including *test*^{Ch}, *test*^{ER}, *test*^{GPCR}, and *test*^{RTK}.

We find that 30% of drugs and 90% of targets in *test*^{S} are also in *train*^{S}, as shown in Table 1. We find an even greater overlap between *test*^{L} and *train*^{L}, i.e. 37% of drugs and 95% of targets. Similar overlaps exist also between the validation sets and the respective training sets for both *BindingDB*^{S} and *BindingDB*^{L}, as shown in Figure 2. In the case of the four OOD sets, we find that 13-25% of the drugs in these sets, as well as one target in *test*^{Ch} and a second one in *test*^{RTK}, are present in *train*^{L}.

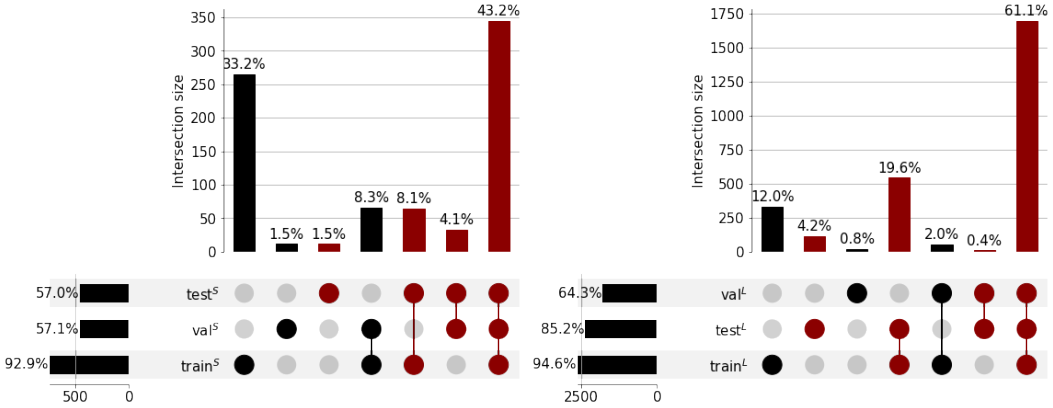
We further investigate the large overlaps of targets across training, validation, and test set seen in *BindingDB*^{S} and *BindingDB*^{L}, looking at the most represented targets. In particular, we pick the top 10% targets by number of interactions in each dataset, i.e. considering the number of interactions in each dataset individually. We observe that nearly all top targets in *test*^{S} and *test*^{L} are among the top targets in the respective training and validation set, as shown in Figure 3. Thus, the overlap of targets across training, validation, and test set seen in *BindingDB*^{S} and *BindingDB*^{L} is strongly present even when looking at the top targets.

3.2.2 Identity matching

The overlaps of drugs and targets we observe across the benchmark datasets may represent a problematic case of information leakage, which may hinder the learning of DTI model.

We investigate this hypothesis performing a baseline analysis by extracting the average activity value for each drug and target in the training set. Then, for each interaction in the test set, if it contains a drug or target that is present in the training set, a drug-based prediction (p_drug) and/or a target-based prediction (p_target) is provided, corresponding to their average activity in the training set.

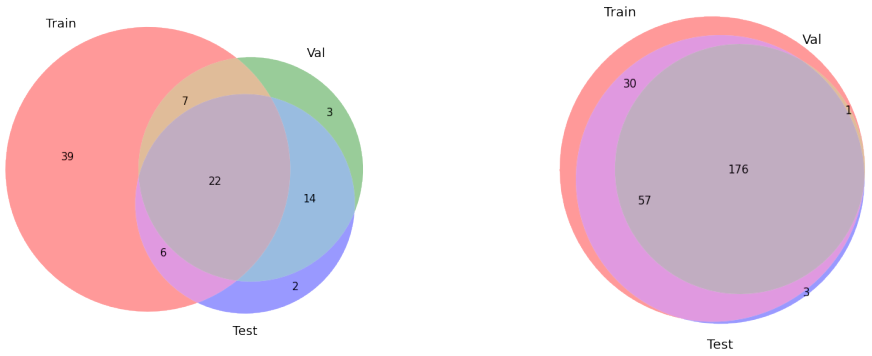
This baseline analysis matches identities for drugs and targets across training and test sets; it performs no machine learning, does not access any protein sequence or compound chemical information, nor does it use any similarity or clustering calculation. Thus, this baseline cannot analyze interactions involving a drug and a target when both are not present in the training set (see Table 6



(a) Targets in *BindingDB*^{S}

(b) Targets in *BindingDB*^{L}

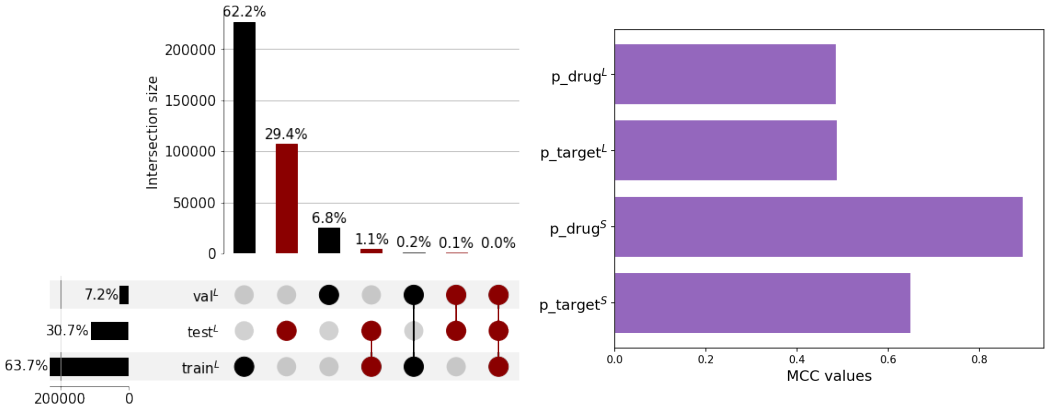
Figure 2: Target overlaps in (a) *BindingDB*^{S} and in (b) *BindingDB*^{L} with respect to the entire benchmark datasets. In both cases, a large portion of targets are present in training, validation, and test set at the same time.



(a) Most represented targets in *BindingDB*^{S}

(b) Most represented targets in *BindingDB*^{L}

Figure 3: The overlap of targets across training, validation, and test set is strongly present even when looking at the subset of the most represented targets in (a) *BindingDB*^{S} and in (b) *BindingDB*^{L}. The overlap between training and test set is shown in magenta, and in grey when overlapping also the validation set.



(a) Interactions in *BindingDB*^{L}

(b) MCC of identity matching baseline

Figure 4: (a) Overlaps of interactions in *BindingDB*^{L} with respect to the entire benchmark dataset; (b) baseline analysis matching identities for drugs and targets across training and test sets: for each interaction in the test set, a drug-based prediction (p_drug) and/or a target-based prediction (p_target) is provided, corresponding to their average activity in the training set.

for the number of predictions for each dataset). Given the lack of chemical and protein sequence information in this approach, one would expect a performance close to random (MCC close to 0.0). Nonetheless, this approach was able to produce surprisingly high MCC values on both *BindingDB*^{S} and *BindingDB*^{L}, as shown in Figure 4b (see full results in Table 6).

Therefore, this analysis provides evidence of the information leakage present in the dataset due to the non-independent nature of interactions with respect to their constitutive drugs and targets.

3.3 Evaluation of Generalization Capability

Motivated by the information leakage we find in *BindingDB*^{S} and *BindingDB*^{L}, we investigate whether filtering the test samples according to the overlap with the training set provides a more informative evaluation of the generalization capability of a method. Therefore, we propose to disaggregate OOD scenarios to improve the benchmarking of DTI predictors.

In particular, we derive 3 subsets for each test set collecting any interaction where:

1. the drug is not in the training set (D);
2. the target is not in the training set (T);
3. neither the target nor the drug is in the training set (I).

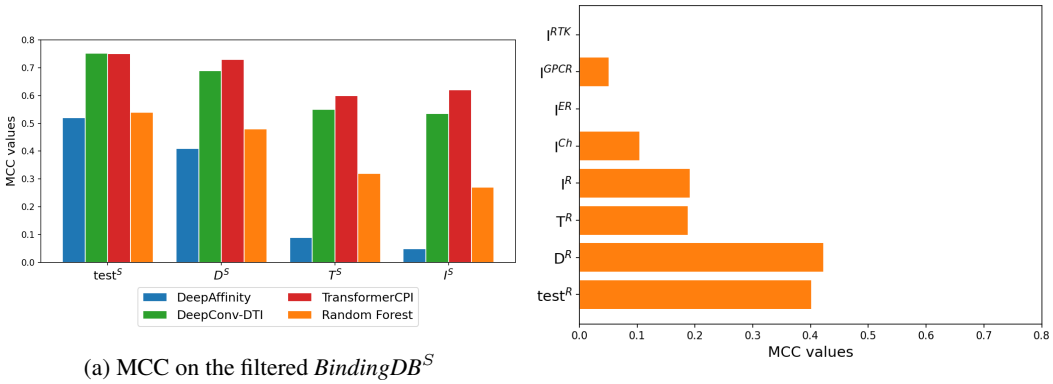
In the Sections below, we use this approach to shed additional light on the information leakage that partially causes the overoptimistic predictive performance observed on *test*^{S} and *test*^{L}.

3.3.1 BindingDB^S

The filtering approach we propose results in three subsets of *test*^{S}, i.e. D^S, T^S, and I^S (see Table 2). For all methods, we observe good predictive performance on *test*^{S} and D^S, as shown in Figure 5a. Lower performance is shown on T^S and I^S, meaning that filtering by target has a similar effect to filtering by both drug and target in this dataset. The stronger relevance of filtering by target matches the larger overlap seen across targets for D^S (as show in Table 2). In part, this can be expected due to the composition of the data (one order of magnitude fewer targets than drugs), but it also points towards the information leakage present in *test*^{S}, due to the non-independent nature of drugs and targets within interactions.

Table 2: Number of unique drugs, targets, and interactions observed in the modified benchmark datasets, as well as sparsity and ratio of Positive/Negative samples. Values within parenthesis are the overlap with the training set, omitted when equal to 0%, and bolded when greater than 50%.

Dataset	Drugs	Targets	Interactions	Sparsity	P/N
D^S	3,354	356 (**90%**)	3,623	0.3%	1.33
T^S	2,210 (15%)	45	2,357	2.36%	1.03
I^S	1,890	35	2,001	3.03%	1.36
Total (unique entities)	3,674	366	3,979	2.41%	1.13
$train^R$	195,479	2,440	265,510	0.06%	1.3
val^R	28,170 (42%)	1,678 (**97%**)	29,501	0.06%	1.29
$test^R$	57,250 (22%)	1,996 (**85%**)	69,433	0.06%	1.21
D^R	35,268	1,945 (**87%**)	54,662	0.34%	1.14
T^R	43,594 (48%)	305	36,694	0.07%	1.26
I^R	21,612	254	21,923	0.4%	1.22
I^{Ch}	9,917	116	11,016	0.96%	1.13
IER	1,779	6	2,773	25.98%	1.29
I^{GPCR}	40,606	272	46,746	0.42%	1.8
I^{RTK}	12,254	93	21,255	1.33%	3.99
Total (unique entities)	321,891	3,268	446,234	0.04%	1.46



(a) MCC on the filtered *BindingDB*^{S}

(b) MCC on *BindingDB*^{R}

Figure 5: MCC observed on (a) the test set and subsets of *BindingDB*^{S}, and on (b) the sets of *BindingDB*^{R}. Figure (b) shows the MCC of a Random Forest.

3.3.2 BINDINGDB^L

The large overlap between $train^L$ and $test^L$ leaves only 64 interactions in I^L. Therefore, we create a new benchnmark dataset, called *BindingDB*^{R}, by resplitting the training, validation, and test set of *BindingDB*^{L} (see Table 2). To do so, we randomly select 10% of targets and 10% of drugs, and allocate any interaction involving those to $test^R$. We allocate 90% of any other interaction to $train^R$, and the remaining 10% to val^R. Thus, all samples in $test^R$ are in D^R, T^R, or I^R.

BindingDB^R contains a similar ratio of samples as in *BindingDB*^{L}, while reducing the overlap of both drugs and targets between training and test set, and remediating the data leakage in $test^L$ (see

Figure 4a). We also filter the four OOD sets removing any interaction involving a drug or target in $train^R$, i.e. I^{Ch}, IER, I^{GPCR}, and I^{RTK}.

We train a Random Forest on $train^R$ and we expect analogous results for the deep learning methods in this study. The Random Forest performs similarly well on $test^R$ and D^R, and worse on T^R and I^R, as shown in Figure 5b. This matches what we observe for $BindingDB^S$ in Section 3.3.1, and provides additional evidence of the non-independence of drugs and targets within interactions.

The predictive performance on T^R and I^R, and to same extend those on T^S and I^S, resemble the results on $test^{Ch}$, $test^{ER}$, $test^{GPCR}$, and $test^{RTK}$. Thus, we see our approach as an alternative solution to evaluate OOD predictive performance, without depriving the training set of entire protein families (as done for the four OOD sets).

4 Conclusion

In this work, we perform an independent benchmark of recent DTI predictors based on deep learning. We find information leakage in previously used DTI benchmark datasets due to the non-independent nature of drugs and targets within interactions. We examine if such information leakage is related to overoptimistic predictive performance and relatively poor generalization observed. To support this idea, we show the high predictive performance obtained with a baseline analysis based solely on drugs and targets identity. To remove this source of bias and improve the assessment of DTI predictors, we propose a novel approach for assessing DTI predictors with respect to the content of their training set. Our novel approach results in a more comprehensive assessment of DTI predictors, discerning the OOD scenarios seen at test time, and highlighting potential source of bias without altering the training set. For future work, we aim to consider similarity of drugs and targets, instead of identity, to extend our minimal solution for evaluating the generalization capability of DTI predictors.

Acknowledgments

The authors are grateful to Brian Kidd and Wilbert Copeland from BMS, and the anonymous reviewers for useful comments and suggestions.

References

Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in Bioinformatics*, 22(1):247–269, jan 2021. ISSN 1467-5463. doi: 10.1093/bib/bbz157.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, jan 2000. ISSN 03051048. doi: 10.1093/nar/28.1.235.

Jannis Born, Tien Huynh, Astrid Stroobants, Wendy D. Cornell, and Matteo Manica. Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction and Inhibitor Generation: 3D Effects in a 1D Model. *Journal of Chemical Information and Modeling*, 62(2):240–257, jan 2022. ISSN 15205142. doi: 10.1021/acs.jcim.1c00889.

J. B. Brown. Classifiers and their Metrics Quantified. *Molecular Informatics*, 37(1-2):1700127, jan 2018. ISSN 1868-1751. doi: 10.1002/MINF.201700127.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 05 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa524.

Tin Kam Ho. Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1:278–282, 1995. ISSN 15205363. doi: 10.1109/ICDAR.1995.598994.

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, 02 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz111.

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, jun 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007129.

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(SUPPL. 1):D198–D201, jan 2007. ISSN 03051048. doi: 10.1093/nar/gkl999.

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31 (3):405–412, feb 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu626.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, jan 2013.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L.L. Sonnhammer, Silvio C.E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, jan 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA913.

RDKit. Open-source cheminformatics. URL `http://www.rdkit.org/`.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, may 2010. ISSN 15499596. doi: 10.1021/CI100050T.

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 59(3):947–961, mar 2019. ISSN 15205142. doi: 10.1021/acs.jcim.8b00712.

Richard D. Smith, Jordan J. Clark, Aqeel Ahmed, Zachary J. Orban, James B. Dunbar, and Heather A. Carlson. Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *Journal of Molecular Biology*, 431(13):2423–2433, jun 2019. ISSN 10898638. doi: 10.1016/j.jmb.2019.05.024.

Jiangming Sun, Nina Jeliazkova, Vladimir Chupakin, Jose Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of Cheminformatics*, 9(1):1–9, mar 2017. ISSN 17582946. doi: 10.1186/s13321-017-0203-5.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, mar 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btu739.

Damian Szklarczyk, Alberto Santos, Christian Von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1):D380–D384, jan 2016. ISSN 13624962. doi: 10.1093/nar/gkv1277.

The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, nov 2017. ISSN 13624962. doi: 10.1093/nar/gkw1099.

Gregor Urban, Mirko Torrisi, Christophe N Magnan, Gianluca Pollastri, and Pierre Baldi. Protein profiles : Biases and protocols. *Computational and Structural Biotechnology Journal*, 18:2281–2289, 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.08.015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, jun 2017. ISSN 10495258. doi: 10.48550/arxiv.1706.03762.

Izhar Wallach and Abraham Heifets. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *Journal of Chemical Information and Modeling*, 58(5): 916–932, may 2018. ISSN 15205142. doi: 10.1021/acs.jcim.7b00403.

Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. *IJCAI International Joint Conference on Artificial Intelligence*, 2018-July:3371–3377, 2018. ISSN 10450823. doi: 10.24963/IJCAI.2018/468.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, nov 2021. ISSN 15577317. doi: 10.1145/3446776.

A APPENDIX

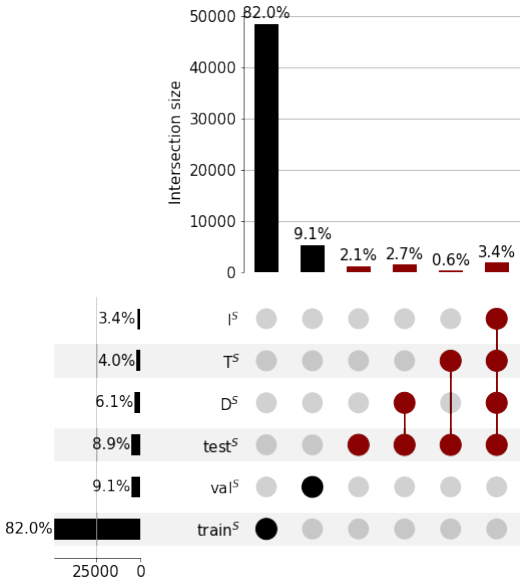
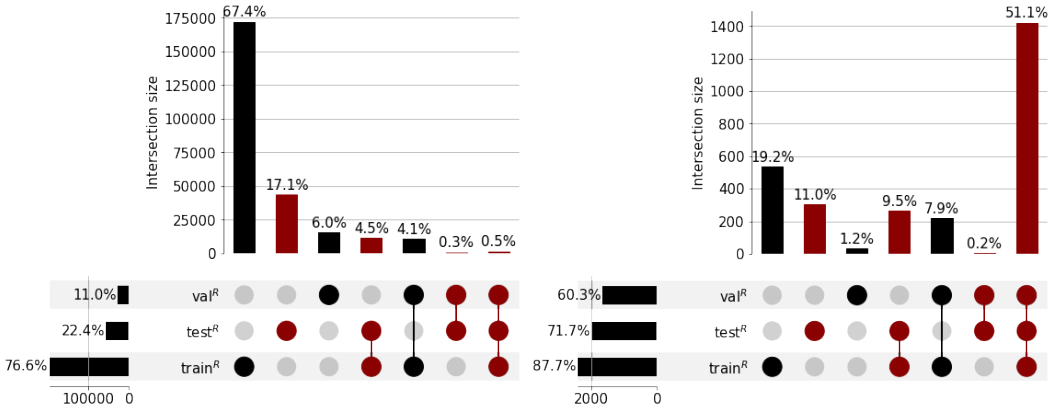


Figure 6: Overlaps of interactions in *BindingDB*^{S} with respect to the entire benchmark dataset. The figure shows that there is no simple case of data leakage with the training set, in net contrast with what we observe on *BindingDB*^{L} (see Figure 4a).



(a) Overlap of drugs in *BindingDB*^{R}

(b) Overlap of targets in *BindingDB*^{R}

Figure 7: Overlaps of (a) drugs and (b) targets in *BindingDB*^{R} with respect to the entire benchmark dataset. In both cases, the overlaps with the training set is reduced with respect to *BindingDB*^{L} (see Table 1 and Figure 2b).

Table 3: Results of DTI benchmark on the test set and subsets of *BindingDB*^{S}.

	Predictor	**Accuracy**	**F1**	**MCC**
test^S	*DeepAffinity*	0.7510	0.7450	0.5197
	DeepConv-DTI	0.8752	0.8749	0.7517
	TransformerCPI	0.8748	0.8748	0.7496
	Random Forest	0.7681	0.7669	0.5386
D^S	*DeepAffinity*	0.6729	0.6713	0.4058
	DeepConv-DTI	0.8424	0.8414	0.6893
	TransformerCPI	0.8653	0.8631	0.7266
	Random Forest	0.7345	0.7340	0.4823
T^S	*DeepAffinity*	0.5286	0.4758	0.0880
	DeepConv-DTI	0.7692	0.7676	0.5500
	TransformerCPI	0.8014	0.8014	0.6033
	Random Forest	0.6474	0.6380	0.3186
I^S	*DeepAffinity*	0.4713	0.4380	0.0537
	DeepConv-DTI	0.7556	0.7555	0.5359
	TransformerCPI	0.8091	0.8066	0.6159
	Random Forest	0.6127	0.6119	0.2741

Table 4: Results of DTI benchmark on the test set and OOD sets of *BindingDB*^{L}.

	Predictor	**Accuracy**	**F1**	**MCC**
test^L	*DeepAffinity*	0.8317	0.8279	0.6560
	DeepConv-DTI	0.8559	0.8525	0.7055
	TransformerCPI	0.8259	0.8215	0.6439
	Random Forest	0.8279	0.8219	0.6480
test^{Ch}	*DeepAffinity*	0.6468	0.6435	0.3052
	DeepConv-DTI	0.5372	0.5367	0.0767
	TransformerCPI	0.4961	0.4953	-0.0091
	Random Forest	0.5420	0.4509	0.1858
test^{ER}	*DeepAffinity*	0.4899	0.4879	0.0671
	DeepConv-DTI	0.5101	0.5029	0.1482
	TransformerCPI	0.4988	0.4942	0.0063
	Random Forest	0.3856	0.2783	0.0000
test^{GPCR}	*DeepAffinity*	0.6245	0.5501	0.1060
	DeepConv-DTI	0.6022	0.5615	0.1596
	TransformerCPI	0.6261	0.5577	0.1240
	Random Forest	0.3884	0.3860	0.0808
test^{RTK}	*DeepAffinity*	0.5516	0.5304	0.2060
	DeepConv-DTI	0.7150	0.6475	0.3173
	TransformerCPI	0.6853	0.6051	0.2263
	Random Forest	0.2397	0.2070	0.0159

Table 5: Results of DTI benchmark on the test set, subsets, and OOD sets of *BindingDB*^{R}.

	Accuracy	**F1**	**MCC**
test^R	0.6894	0.6890	0.4023
D^R	0.7040	0.7040	0.4227
T^R	0.5631	0.5389	0.1883
I^R	0.5729	0.5441	0.1920
I^{Ch}	0.4971	0.3952	0.1047
IER	0.3567	0.2629	0.0000
I^{GPCR}	0.3636	0.3606	0.0520
I^{RTK}	0.2031	0.1719	-0.0173

Table 6: Results of sequence identity baseline, and number of interactions analyzed.

Pred	**#preds**	**Accuracy**	**F1**	**MCC**
p_target^S	2,891	0.8201	0.8199	0.6500
p_drug^S	1,625	0.9551	0.9482	0.8964
p_target^L	112,995	0.7511	0.7986	0.4883
p_drug^L	45,745	0.7425	0.7500	0.4860