DEBIASEDDTA: MODEL DEBIASING TO BOOST DRUG-TARGET AFFINITY PREDICTION

Rıza Özçelik, Alperen Bağ,* Berk Atıl* & Arzucan Özgür †
Department of Computer Engineering
Boğaziçi University, İstanbul, Turkey
{riza.ozcelik,alperen.bag,berk.atil,arzucan.ozgur}@boun.edu.tr

Elif Ozkirimli†
Data and Analytics Chapter, Pharma International Informatics
F. Hoffmann-La Roche AG, Switzerland
elif.ozkirimli@roche.com

ABSTRACT

Computational models that accurately identify high-affinity protein-chemical pairs can accelerate drug discovery pipelines. These models, trained on available protein-chemical interaction datasets, can be used to predict the binding affinity of an input protein-chemical pair. However, the training datasets may contain surface patterns, or dataset biases, such that the models memorize dataset-specific biomolecule properties instead of learning affinity prediction rules. As a result, the prediction performance of models drops for unseen biomolecules. Here, we present DebiasedDTA, a novel drug-target affinity (DTA) prediction model training framework that addresses dataset biases to improve affinity prediction for novel biomolecules. DebiasedDTA uses ensemble learning and sample weight adaptation to identify and avoid biases and is applicable to most DTA prediction models. The results show that DebiasedDTA can boost models while predicting the interactions between unseen biomolecules. In addition, prediction performance for seen biomolecules also improves when the surface patterns are debiased. The experiments also show that DebiasedDTA can avoid biases of different sources and augment DTA prediction models of different input and model structures. An open-source python package, pydta, is published to facilitate the adoption of DebiasedDTA by future DTA prediction studies. Out-of-the-box, pydta allows debiasing custom DTA prediction models with only two lines of code and eliminates two sources of bias. pydta is designed to be the go-to library for model debiasing in the field of computational drug discovery.

1 INTRODUCTION

The first step toward drug discovery is to identify high affinity protein-chemical pairs. However, the number of possible protein-chemical combinations makes this task a "needle in the haystack" problem (∼560K proteins in UniProt (Apweiler et al., 2004) and ∼2.1M chemicals in ChEMBL (Davies et al., 2015)). This is where drug-target affinity (DTA) prediction models come into play; they can rapidly identify high-affinity protein-chemical pairs in the combination space after learning generalizable affinity prediction rules from large interaction datasets.

The interaction datasets report affinity measurements for millions of protein-chemical pairs and stand as invaluable resources to learn rules of affinity prediction. However, they also contain spurious patterns that can misguide the learning (Chaput et al., 2016; Wallach & Heifets, 2018; Sieg et al., 2019; Yang et al., 2020; Scantlebury et al., 2020). For instance, a single atom may be separating actives and inactives of a target (Bietz et al., 2015) and the prediction models can learn to predict interaction strength through that atom exclusively, instead of learning generalizable affinity prediction

*These authors contributed equally.
†Corresponding author.

rules. Consequently, models struggle to estimate the binding affinity between unseen biomolecules, for which the learned shortcuts are unavailable (Chen et al., 2019; Tran-Nguyen et al., 2020; Yang et al., 2020; Boyles et al., 2020; Özçelik et al., 2021). These dataset shortcuts are the dataset biases and form a major problem to discover drugs for rare diseases or to identify novel chemical moieties to which proteins have not yet acquired resistance.

To the best of our knowledge, there is no study with a focus on boosting drug-target affinity prediction on novel biomolecules. Recent works studied the generalizability problem in a similar task, drug-target interaction prediction. They focused on the datasets and designed train-test splits with dissimilar biomolecules so that the training set biases are less rewarding on the test set (Wallach & Heifets, 2018; Tran-Nguyen et al., 2020). However, counter to the aim, these "dataset-oriented" approaches introduced the risk of degrading model generalizability and inaccurate estimation of dissimilar test set performance (Sundar & Colwell, 2019). Furthermore, their use in the affinity prediction task would require non-trivial adaptations, as they exploit the two-class structure (active or inactive) in the drug-target interaction task.

An alternative perspective to cope with biases and improve model generalizability is to focus on the prediction models instead of datasets. This "model-oriented" perspective is free from the limitations of the task and has been recently successfully used in natural language processing (Clark et al., 2020; Sanh et al., 2021; Utama et al., 2020), computer vision (Bissoto et al., 2020; Majumdar et al., 2021), as well as for structure-based virtual screening (Scantlebury et al., 2020). Unfortunately, the impact of the model-oriented perspective on computational drug discovery was limited by the number of available 3D structures (Scantlebury et al., 2020).

In this paper, we propose DebiasedDTA, a novel model training framework to address dataset biases and boost the generalizability of DTA prediction models. DebiasedDTA adopts the model-oriented perspective and, unlike the dataset-oriented approaches proposed for drug-target interaction prediction, it is applicable to datasets with continuous and discrete labels without requiring modifications. In addition, DebiasedDTA can be used to debias DTA prediction models with any biomolecule representation and finds a wider application range than 3D-structure based approaches.

DebiasedDTA ensembles a "guide" and a "predictor" to train debiased DTA prediction models. The guide quantifies a particular type of training set bias and prepares a debiasing roadmap for the predictor. The predictor utilizes the roadmap in order to adapt the sample weights during training to avoid biases and to achieve higher generalizability on novel biomolecules.

We test DebiasedDTA with two guides on different bias sources and with three predictors to evaluate across biomolecule representations. Experiments on two datasets and ten test sets show that the proposed approach is robust to different bias sources and can boost prediction performance of DTA models with different drug-target representations. Noteworthy, the improvement is not only observed for novel biomolecules but also for the seen ones.

DebiasedDTA is a novel approach that boosts the generalizability of DTA prediction models. Using a model-oriented perspective and a biomolecule representation independent sample weight adaptation strategy, DebiasedDTA can be adopted to enhance the prediction performance of any DTA prediction model that allows sample weighting.

2 DEBIASEDDTA

DebiasedDTA is a model debiasing framework to boost drug-target affinity prediction on novel biomolecules and consists of two DTA prediction models, the guide and the predictor. The guide aims to identify dataset biases only, and thus uses biomolecule representations that target a specific bias source in the dataset. When the guide is trained on the training set, it prepares a training roadmap for the predictor and the predictor follows the roadmap to drive its training away from dataset biases; towards generalizable information. The debiased model is used standalone to predict the affinity between target protein-chemical pairs. Figure 1 illustrates the DebiasedDTA training framework.

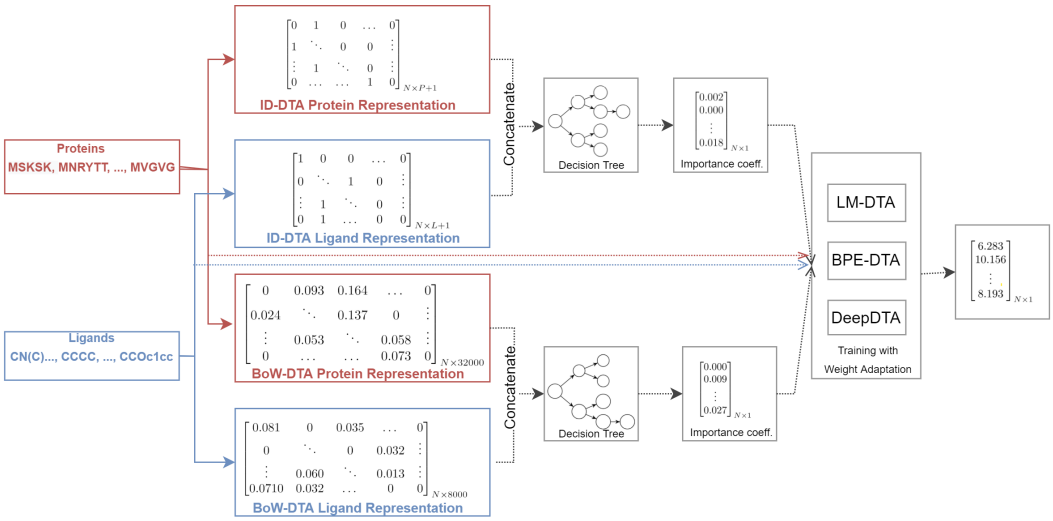

Figure 1: DebiasedDTA. DebiasedDTA training framework boosts the generalizability of DTA prediction models on novel biomolecules by driving them away from spurious dataset patterns called dataset biases. In DebiasedDTA, the guide models adopt biomolecule representation schemes to identify the training samples that contain targeted biasing patterns. Here, we target bimolecular word and identity-driven biases with BoW-DTA and ID-DTA models. The guides are trained on the training protein-chemical pairs with 10-fold cross validation and median squared error of each training sample is calculated. The predictors use these errors to prioritize training samples and attribute more importance to the instances that challenge the guides. The biomolecule representation of the predictors can take any form and we experiment with three models, DeepDTA, BPE-DTA, and LM-DTA. The experiments show that all models leverage DebiasedDTA training framework to boost their performance on unseen biomolecules.

2.1 THE GUIDE

The guides in DebiasedDTA are designed to learn merely dataset biases and should have limited learning capacity. So, we design two weak learners with simple biomolecule representations to identify different bias sources: an identifier-based model (ID-DTA) and a biomolecule word-based model (BoW-DTA). ID-DTA is motivated by the fact that mere use of random biomolecule identifiers can produce high-achieving models for similar test sets (Özçelik et al., 2021), and thus, can be a strong bias source. ID-DTA featurizes the interactions by concatenating the one-hot encoded vectors of chemicals and proteins. BoW-DTA, on the other hand, bases on natural language inference studies in which the use of certain words in a sentence produces a strong bias with its semantic label (Gururangan et al., 2018; Poliak et al., 2018). Here, we investigate a similar bias in biomolecular sequences and create BoW-DTA. BoW-DTA represents the proteins and chemicals with bag-of-words vectors and concatenates their vectors to represent the interaction.

BoW-DTA segments the biomolecule sequences into their words via BPE vocabularies and this might create an inconsistency between BoW-DTA and the predictor, if the predictor uses different vocabularies. So, we fork BoW-DTA and create BoW-LM-DTA to use with LM-DTA, a predictor introduced in Section 2.2, which has different vocabularies. BoW-LM-DTA adopts the same word segmentation strategy as LM-DTA and same vectorization method as BoW-DTA. ID-DTA, BoW-DTA, BoW-LM-DTA use decision tree regression for prediction, as decision trees have limited learning capacity and yet are effective to learn spurious patterns.

We adopt 5-fold cross-validation to quantify dataset biases with the guides. First, we randomly divide the training set into five folds and construct five different mini-training and mini-validation sets. We train the guide on each mini-training set and compute the squared errors of its predictions on the corresponding mini-validation set. One run of cross-validation yields one squared-error measurement per protein-chemical pair as each pair is placed in the mini-validation set exactly once. In order to better estimate the performance on each sample, we run the 5-fold cross-validation 10

times and obtain 10 error measurements per sample. We compute the median of the 10 squared errors and name it as the "importance coefficient" of a protein-chemical pair. If the affinity of pair is easily predictable via exploiting dataset biases, i.e. the guide has a low prediction error, then the pair might contain biasing patterns for DTA prediction models and has a low importance coefficient. Otherwise, the pair is more likely to contain generalizable information about binding affinity and has a high importance coefficient. The importance coefficients guide the training of the predictor.

2.2 THE PREDICTOR

In DebiasedDTA training framework, the predictor is the model to debias and use on the target protein-chemical pairs. The predictor is free to adopt any biomolecule representation, but have to be able to weight the training samples during training to comply with the weight adaptation strategy proposed in DebiasedDTA.

The proposed strategy initializes the training sample weights to 1 and updates them at each epoch such that the weight of each training sample converges to its importance coefficient at the last epoch. When trained with this strategy, the predictor attributes more importance to samples with less biasing patterns as the learning continues, that is the bias in the model decays over time. Our weight adaptation strategy is formulated in Equation 1.

$$\vec{w}_e = (1 - \frac{e}{E}) + \vec{i} \times \frac{e}{E} \tag{1}$$

where w_e is the vector of training sample weights at epoch e, E is the number of training epochs, and $\vec{i}$ is the importance coefficients vector. Here, e/E increases as the training continues, and so does the impact of $\vec{i}$ on the sample weights. This ensures that the importance of samples with less biasing patterns is increased towards the end of training.

We implement three drug-target affinity prediction models to observe the performance of DebiasedDTA training framework with different predictors. The first one is DeepDTA (Öztürk et al., 2018), an influential affinity prediction model that uses SMILES strings of chemicals and amino-acid sequences of proteins to represent biomolecules. DeepDTA applies three layers of character-level convolutions over input sequences and uses a three-layered fully-connected neural network for prediction. Here, we slightly modify DeepDTA and treat chemical groups in the SMILES strings ([OH], [COH], [COOH] etc.) as a single token, while the original DeepDTA processes these groups as character-by-character, too.

In the second model, we alter DeepDTA to use biomolecular word-level convolutions, where the words are identified via BPE algorithm and name the resulting model BPE-DTA. We experiment with BPE vocabulary sizes of 8K, 16K, and 32K for SMILES and protein sequences and pick the combination of 8K-32K as it yields the highest scores on datasets of our previous studies (Özçelik et al., 2021). We report the results for all vocabulary combinations in our GitHub repository for completeness.

Third, we utilize ChemBERTa (Chithrananda et al., 2020) and ProtBERT (Elnaggar et al., 2020) to create another drug-target affinity prediction model, LM-DTA. LM-DTA vectorizes SMILES and amino-acid sequences via the language models and concatenates their vectors to represent the interaction. Finally, LM-DTA uses a two-layered fully connected neural network for prediction.

3 EXPERIMENTAL SETUP

3.1 DATASETS

We test DebiasedDTA on BDB (Özçelik et al., 2021) and KIBA (Tang et al., 2014) datasets. KIBA contains 118K affinity measurements of 229 kinase family proteins and 2111 chemicals, such that the affinities are reported in terms of KIBA score. KIBA score combines different measurement sources such as K_d, IC_{50}, and K_i, and ranges from 1.3 to 17.2 in the dataset, the latter denoting a higher binding affinity.

BDB is a dataset filtered from BindingDB database (Liu et al., 2007) and comprises 31K binding affinity measurements of 490 proteins and 924 chemicals. The binding affinities are recorded in terms of pK_d (Özçelik et al., 2021), which correlates positively with the binding strength and changes between 1.6 and 13.3 in the dataset. Protein diversity is higher in BDB than KIBA as it contains fewer interactions, but more proteins from different families.

3.2 Experimental Settings

We create five distinct train-test setups per dataset to evaluate the models. To create different setups, we cluster the proteins and chemicals in the datasets and randomly divide the clusters into two as "warm" and "cold". We interpret the warm clusters as already known biomolecules and the cold clusters as novel biomolecules. The dissimilarity of known and novel biomolecules is enforced by the clustering-based split.

To produce training and test sets from warm and cold biomolecule clusters, we first filter the interactions between proteins and chemicals in the warm clusters. We use these interactions mainly as the training set, but also separate small subsets as "validation" and "warm test" sets. The validation fold is used to tune model hyper-parameters, whereas the warm test set is to evaluate models on the interactions between known biomolecules.

We create two more test sets called "cold chemical" and "cold protein", where the cold chemical test set consists of the interactions between chemicals in the cold cluster and proteins in the warm cluster. This test set is used to measure model performance in the scenarios in which new drugs are searched to target existing proteins. The cold protein test set is created similarly and used to evaluate models in the scenarios where existing drugs are searched to target a novel protein.

Last, we create a "cold both" test set, that is the set of interactions between the proteins and chemicals in the cold clusters. This is the most challenging test set of every setup, as both the proteins and the chemicals are unavailable in the training set.

To tune the hyper-parameters, we train models on the training set of each setup and measure the performance on the corresponding validation set. We pick the hyper-parameter combination that scores the lowest validation average mean squared error to predict the test set interactions.

3.3 Evaluation Metrics

We evaluate DebiasedDTA models with two metrics, concordance index (CI) (Gönen & Heller, 2005) and R^2. We use CI in order to evaluate the consistency of the predicted binding affinity ranking of protein-chemical pairs with the expected one. Evaluating a ranking, CI is independent of the output range and allows comparisons across datasets. CI is expected to be around 0.5 for random predictions and reaches 1 when two rankings match exactly.

We also calculate R^2, a regression metric that measures how much of the variance in the expected labels is explained by the predictions. R^2 is 0 when all predictions are equal to mean of labels and equals to 1 when labels and predictions are the same. We use its `scikit-learn` (Pedregosa et al., 2011) implementation.

4 Results

We debias three DTA prediction models, namely DeepDTA, BPE-DTA, and LM-DTA with two debiasing approaches, BoW-DTA and ID-DTA, on BDB and KIBA datasets in DebiasedDTA training framework and report CI and R^2 on the test sets in Table 1.

The Overall Gain of Debiasing We first examine the performance boost due to DebiasedDTA and compare the best DebiasedDTA score on each setup with no debiasing score. Table 2 reports the percent increase in CI and absolute increase in R^2 thanks to debiasing.

Table 2 demonstrates that in 17 of 24 ($\sim$71%) evaluation setups, at least one model trained in DebiasedDTA outperforms the non-debiased counterpart, highlighting the strength of the proposed training framework to boost DTA prediction performance. To show that the performance increase

Table 1: The effect of debiasing on the model performance.

			Warm		Cold Chemical		Cold Protein		Cold Both	
		The Guide	CI	R^2	CI	R^2	CI	R^2	CI	R^2
BDB	DeepDTA	None	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
		BoW-DTA	**0.899 (0.004)**	0.799 (0.013)	**0.698 (0.037)**	**0.043 (0.108)**	**0.777 (0.014)**	**0.351 (0.090)**	0.568 (0.044)	**-0.092 (0.132)**
		ID-DTA	0.898 (0.005)	**0.804 (0.011)**	0.693 (0.058)	0.026 (0.109)	0.771 (0.007)	0.339 (0.067)	**0.585 (0.040)**	-0.128 (0.056)
	BPE-DTA	None	0.883 (0.006)	0.774 (0.013)	0.657 (0.083)	-0.143 (0.202)	0.653 (0.060)	**-0.256 (0.411)**	0.522 (0.054)	-0.442 (0.349)
		BoW-DTA	0.888 (0.008)	**0.781 (0.016)**	0.687 (0.082)	-0.091 (0.302)	**0.664 (0.067)**	-0.386 (0.593)	**0.568 (0.084)**	**-0.334 (0.347)**
		ID-DTA	**0.891 (0.005)**	0.777 (0.019)	**0.692 (0.065)**	**-0.045 (0.252)**	0.650 (0.039)	-0.689 (0.476)	0.565 (0.090)	-0.426 (0.231)
	LM-DTA	None	0.876 (0.005)	0.745 (0.011)	**0.688 (0.046)**	-0.027 (0.175)	0.780 (0.016)	0.384 (0.083)	0.572 (0.028)	-0.226 (0.205)
		BoW-DTA	0.882 (0.006)	**0.762 (0.003)**	**0.688 (0.069)**	**-0.005 (0.169)**	0.781 (0.017)	0.386 (0.081)	0.563 (0.032)	**-0.182 (0.136)**
		ID-DTA	0.883 (0.006)	0.758 (0.003)	0.683 (0.067)	-0.016 (0.270)	0.782 (0.017)	0.387 (0.080)	**0.581 (0.017)**	-0.198 (0.174)
		BoW-LM-DTA	**0.884 (0.009)**	0.761 (0.008)	0.662 (0.074)	-0.096 (0.227)	**0.784 (0.016)**	**0.395 (0.078)**	0.548 (0.033)	-0.244 (0.137)
KIBA	DeepDTA	None	0.873 (0.005)	0.756 (0.021)	0.753 (0.018)	0.337 (0.081)	0.719 (0.029)	0.330 (0.109)	0.654 (0.019)	**0.087 (0.099)**
		BoW-DTA	**0.888 (0.005)**	**0.775 (0.019)**	**0.761 (0.004)**	0.349 (0.046)	0.713 (0.036)	0.308 (0.115)	0.639 (0.028)	0.045 (0.147)
		ID-DTA	0.887 (0.006)	**0.775 (0.018)**	**0.761 (0.020)**	**0.350 (0.101)**	**0.725 (0.038)**	**0.333 (0.124)**	**0.660 (0.034)**	0.084 (0.195)
	BPE-DTA	None	0.881 (0.005)	0.760 (0.016)	0.735 (0.025)	**0.274 (0.105)**	0.680 (0.020)	**0.185 (0.077)**	**0.605 (0.033)**	**-0.006 (0.117)**
		BoW-DTA	0.891 (0.003)	0.774 (0.016)	**0.736 (0.018)**	0.231 (0.093)	0.679 (0.030)	0.174 (0.103)	0.604 (0.017)	-0.046 (0.082)
		ID-DTA	**0.893 (0.003)**	**0.776 (0.012)**	**0.736 (0.021)**	0.229 (0.099)	**0.684 (0.023)**	0.179 (0.060)	0.590 (0.014)	-0.037 (0.079)
	LM-DTA	None	0.858 (0.005)	0.756 (0.012)	0.749 (0.012)	0.409 (0.067)	0.713 (0.049)	0.366 (0.137)	0.650 (0.041)	0.107 (0.122)
		BoW-DTA	**0.865 (0.005)**	**0.769 (0.013)**	0.756 (0.013)	0.435 (0.064)	0.717 (0.051)	0.382 (0.139)	**0.653 (0.028)**	**0.159 (0.121)**
		ID-DTA	0.864 (0.006)	0.767 (0.014)	**0.759 (0.011)**	0.436 (0.056)	0.718 (0.053)	**0.385 (0.143)**	0.652 (0.036)	0.151 (0.126)
		BoW-LM-DTA	0.864 (0.005)	0.768 (0.012)	0.758 (0.010)	**0.441 (0.055)**	**0.719 (0.054)**	0.382 (0.145)	0.646 (0.032)	0.139 (0.115)

due to DebiasedDTA is statistically significant, we conduct a one-sided one-sample t-tests with the null hypotheses that mean CI and R^2 gains are 0. The statistical tests result in the rejection of the null hypothesis with p-value < 0.01, suggesting that DebiasedDTA boosts prediction performance in general, with 99% significance.

The improvement in performance due to debiasing is more evident in the cold test sets of BDB, because BDB is a more diverse dataset than KIBA. Since the BDB biomolecules are more diverse, the training biases are less applicable to the unknown test biomolecules and their elimination boosts the DTA prediction performance more than KIBA.

Table 2 also highlights that, training models in DebiasedDTA improves the performance on every warm test set, though it is mainly designed to boost DTA prediction on novel biomolecules. This shows that eliminating the training set biases helps models to better represent the known biomolecules, too.

Finally, Table 2 shows that debiasing improves the performance of all affinity prediction models in the study on at least one test setup. This emphasizes that DTA prediction models are susceptible to dataset biases irrespective of their input representation and the proposed training framework is powerful and abstract enough to eliminate biases in different biomolecule representation settings.

Effect of The Guides We investigate the effect of the guide selection in DebiasedDTA on the affinity prediction performance by comparing BoW-DTA models with ID-DTA. For BDB, models debiased with BoW-DTA yield higher scores in both metrics in 5 cases and ID-DTA based models outperform BoW-DTA 2 times in terms of CI and R^2. 5 out of 12 times, no guide can outscore the other in both metrics.

On KIBA, ID-DTA achieves higher CI and R^2 than BoW-DTA in 7 cases whereas BoW-DTA outperforms ID-DTA 4 times in terms of both metrics. The higher performance of ID-DTA on KIBA compared to BDB (7 wins vs. 2 wins) suggests that biomolecule identities cause more bias in this dataset. We relate this with the fact that KIBA contains more interactions per biomolecule and thus the models can infer more information about the biomolecule identities from the training set. In total, both guides outperform the other 9 times, indicating that the performance of ID-DTA and BoW-DTA is comparable to each other and both chemical word based and identity based biases are prevalent in the datasets.

Last, we examine the effect of using the same biomolecule vocabularies in the guide and predictor by comparing BoW-DTA and BoW-LM-DTA. BoW-LM-DTA, which uses the same vocabulary as LM-

Table 2: The gain of debiasing. The percentile improvement in CI and increase in R^2 are displayed for each model on every test set. The statistics are computed by comparing the best DebiasedDTA score with the non-debiased one. Negative statistics are reported if the non-debiased model outperforms every debiasing configuration.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
BDB	DeepDTA	1.239%	0.023	1.601%	0.004	2.372%	0.036	5.596%	0.062
	BPE-DTA	0.906%	0.007	5.327%	0.098	1.685%	-0.141	8.812%	0.108
	LM-DTA	0.913%	0.017	0.000%	0.022	0.513%	0.011	1.573%	0.044
KIBA	DeepDTA	1.718%	0.019	1.062%	0.013	0.834%	0.003	0.917%	-0.003
	BPE-DTA	1.362%	0.017	0.136%	-0.045	0.588%	-0.006	-1.157%	-0.031
	LM-DTA	0.816%	0.013	1.335%	0.032	0.842%	0.019	0.462%	0.052

DTA, outperforms BoW-DTA based models on only 2 of 8 setups in Table 1, whereas BoW-DTA outscores BoW-LM-DTA on 4 setups. This shows that the guide and the predictor architectures do not have to be similar for a cohesive learning. Because, once the guides quantify the dataset biases, the predictors acquire all the information they need for weight adaptation – they become indifferent to the underlying computation.

5 PYDTA: DEBIASING DRUG-TARGET AFFINITY PREDICTION IN PYTHON

DebiasedDTA boosts every predictor designed in this study with its low-cost guides and widely-adoptable weight adaptation strategy. Its effectiveness, low training overhead, and compatibility with most DTA prediction models promise that it can be a standard approach to train DTA prediction models. We present a pip-installable python package, pydta, that can be used to train DTA models within the DebiasedDTA framework with only a couple of lines of code.

pydta adopts object oriented programming to offer a simple and intuitive interface to use DebiasedDTA. DebiasedDTA is a class in pydta and requires a guide definition, a predictor definition, and a dictionary of predictor construction parameters for initialization. BoW-DTA and ID-DTA are already available as guide models in pydta and custom DTA prediction models can easily use these guides to improve their performance.

To use DebiasedDTA, pydta enforces the custom prediction model to be implemented as a class which has an n_epochs attribute and a train method with arguments training_chemicals, training_proteins, training_labels, and sample_weights_by_epoch. DebiasedDTA imposes no restriction on the inner-workings of the train function and the content of the arguments. Code 1 displays the template to debias a custom DTA prediction model with ID-DTA and more examples are available in the pydta repository.

6 CONCLUSION

Dataset bias is a major hurdle on the path to develop robust and generalizable ML models and one approach is to obtain a sampling from all knowledge space. However, protein-chemical interaction space is not sampled evenly, either because some protein targets are privileged due to their association with certain disease states, or because some chemicals or chemical moieties are privileged due to their relatively easier synthesis, or because the study of some interactions is experimentally infeasible. As some proteins or chemicals are over-represented, machine learning models tend to overfit and memorize these patterns and perform well when the training and test sets are similar to each other. However, it is difficult to learn generalizable patterns about protein-chemical interactions and machine learning methodologies fail when they are tasked with predictions about unseen biomolecules. In this work, we propose DebiasedDTA, a novel training framework that boosts the performance of DTA prediction methods both on known and unknown biomolecules. The performance improvement is observed for similar and distant test sets and underlines the value of DebiasedDTA.

```python
from pydta.models import IDDTA, DebiasedDTA

class CustomDTAModel:
    # The constructor can have other arguments and/or the class
        have other attributes.
    def __init__(self, n_epochs):
        self.n_epochs = n_epochs

    # The last argument will be filled by DebiasedDTA.
    def train(self, train_chemicals, train_proteins, train_labels,
        sample_weights_by_epoch):
        pass

train_chemicals, train_proteins, train_labels = [...], [...],
    [...]
debiaseddta = DebiasedDTA(IDDTA, CustomDTAModel, predictor_params
    ={'n_epochs': 100})
debiaseddta.train(train_chemicals, train_proteins, train_labels)
```

Code 1: Debiasing a custom model in pydta.

DebiasedDTA owes the performance boost to the guides that are designed to identify specific types of bias sources. Here, we experiment with biomolecule word and identity driven biases and find that elimination of either of the two can improve prediction performance. We also find that DebiasedDTA does not require a similarity in biomolecule representations of guides and predictors and can improve predictors of diverse architectures. We publish DebiasedDTA as a python package to promote its use to debias assorted predictors.

The predictors weight training samples for debiasing, which tunes the contribution of input features to the predictions. We show that elimination of biomolecule word biases pushes the models to learn more from the proteins and can reduce the effect of pharmacologically unimportant substructures to the predictions.

Here, we present DebiasedDTA as a pioneering work toward overcoming dataset bias from the model's perspective and creating more generalizable DTA prediction models. Prioritization of informative training samples, proposed in DebiasedDTA can also find applications in debiasing natural language processing and computer vision models, where out-of-distribution generalization is also an essential problem.

REFERENCES

Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. Uniprot: The universal protein knowledgebase. *Nucleic Acids Research*, 32(suppl_1):D115–D119, 2004.

Stefan Bietz, Karen T Schomburg, Matthias Hilbig, and Matthias Rarey. Discriminative chemical patterns: Automatic and interactive design. *Journal of Chemical Information and Modeling*, 55 (8):1535–1546, 2015.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 740–741, 2020.

Fergus Boyles, Charlotte M Deane, and Garrett M Morris. Learning from the ligand: Using ligand-based features to improve binding affinity prediction. *Bioinformatics*, 36(3):758–764, 2020.

Ludovic Chaput, Juan Martinez-Sanz, Nicolas Saettel, and Liliane Mouawad. Benchmark of four popular virtual screening programs: Construction of the active/decoy dataset remains a major determinant of measured performance. *Journal of Cheminformatics*, 8(1):1–17, 2016.

Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J Dickson, Jose S Duca, Viktor Hornak, David R Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS One*, 14(8):e0220113, 2019.

Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 3031–3045, 2020.

Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. Chembl web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1):W612–W620, 2015.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, 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. *bioRxiv*, 2020.

Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018.

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, 2007.

Puspita Majumdar, Richa Singh, and Mayank Vatsa. Attention aware debiasing for unbiased model prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4133–4141, 2021.

Rıza Özçelik, Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Chemboost: A chemical language based approach for protein–ligand binding affinity prediction. *Molecular Informatics*, 40 (5):2000212, 2021.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*, 2021.

Jack Scantlebury, Nathan Brown, Frank Von Delft, and Charlotte M Deane. Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes and highlight important binding interactions. *Journal of Chemical Information and Modeling*, 60(8): 3722–3730, 2020.

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, 2019.

Vikram Sundar and Lucy Colwell. The effect of debiasing protein–ligand binding data on generalization. *Journal of Chemical Information and Modeling*, 60(1):56–62, 2019.

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: An unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling*, 2020.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7597–7610, 2020.

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, 2018.

Jincai Yang, Cheng Shen, and Niu Huang. Predicting or pretending: Artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Frontiers in Pharmacology*, 11:69, 2020.