DrugOOD: Out-of-Distribution Dataset Curator and Benchmark for AI-Aided Drug Discovery – a Focus on Affinity Prediction Problems with Noise Annotations

Yuanfeng Ji[1,3*], **Lu Zhang**[1,2*], **Jiaxiang Wu**[1], **Bingzhe Wu**[1], **Lanqing Li**[1], **Long-Kai Huang**[1], **Tingyang Xu**[1], **Yu Rong**[1], **Jie Ren**[1], **Ding Xue**[1], **Houtim Lai**[1], **Wei Liu**[1], **Junzhou Huang**[1], **Shuigeng Zhou**[2], **Ping Luo**[3], **Peilin Zhao**[1], **Yatao Bian**[1†]

[1] Tencent AI Lab, China
[2] Fudan University, China
[3] The University of Hong Kong, China
u3008013@connect.hku.hk

Abstract

AI-aided drug discovery (AIDD) is gaining popularity due to its potential to make the search for new pharmaceuticals faster, less expensive, and more effective. Despite its extensive use in numerous fields (e.g., ADMET prediction, virtual screening), little research has been conducted on the out-of-distribution (OOD) learning problem with noise. We present `DrugOOD`, a systematic OOD dataset curator and benchmark for AIDD. Particularly, we focus on the drug-target binding affinity prediction problem, which involves both macromolecule (protein target) and small-molecule (drug compound). `DrugOOD` offers an automated dataset curator with user-friendly customization scripts, rich domain annotations aligned with biochemistry knowledge, realistic noise level annotations, and rigorous benchmarking of SOTA OOD algorithms, as opposed to only providing *fixed* datasets. Since the molecular data is often modeled as irregular graphs using graph neural network (GNN) backbones, `DrugOOD` also serves as a valuable testbed for graph OOD learning problems. Extensive empirical studies have revealed a significant performance gap between in-distribution and out-of-distribution experiments, emphasizing the need for the development of more effective schemes that permit OOD generalization under noise for AIDD.

Introduction

Traditional drug discovery procedures are lengthy and expensive. To accelerate the drug development process, drugmakers and investors are turning to artificial intelligence techniques (Muratov et al. 2020) for drug discovery (e.g., ADMET prediction (Rong et al. 2020), target identification, protein structure prediction(Shen et al. 2021)), which aim to rapidly identify new compounds and model complex mechanisms to automate previously manual processes (Schneider 2018). In this paper, we focus on one of the most challenging applications, called drug-target binding affinity prediction, which aims to identify a subset of compounds with high binding affinity for a given protein target among many candidate compounds.

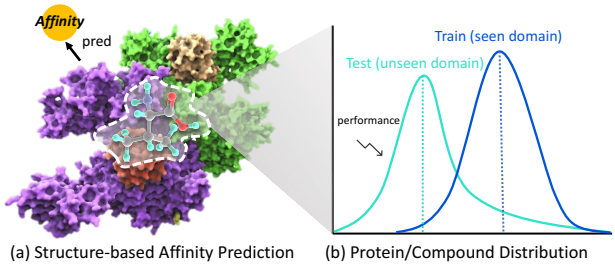

Figure 1: (a) Structure-based affinity prediction aims to predict binding affinity values between a pair of target (protein) and compound (molecule), and (b) model performance is often severely degraded when the data distributions are shifted.

Distribution shift is a ubiquitous problem in the field of AIDD, where the training distribution differs from the test distribution. Typically, the prediction model is trained on known target proteins when conducting virtual screening for hit findings. A "black swan" event, such as COVID-19, may nevertheless occur, resulting in a new target with unseen data distributions. Hence, the performance of the unseen target will decline drastically. To address the performance degradation (Koh et al. 2021) caused by distribution shift, it is necessary to develop robust and generalizable algorithms for this challenging issue in AIDD. Despite its importance in real-world problems, the community still lacks curated OOD datasets and benchmarks for inspiring relevant research. Besides, *label noise* is another critical issue. Generally, AI models are trained using publicly deposited datasets, such as ChEMBL, whereas the bioassay data are typically noisy (Kramer et al. 2012; Cortés-Ciriano and Bender 2016). For instance, the activity data from ChEMBL is manually extracted from the full texts of seven Medicinal Chemistry journals (Mendez et al. 2019). Various factors, including but not limited to different confidence levels for activities measured through experiments, unit-translation errors, repeated citations of single measurements, and different "cut-off" noise[1],

[*]Equal contribution. Order was determined by tossing a coin.
[†]Corresponding author: Yatao Bian.

[1]E.g., measurements could be recorded with $<$, $\leq$, $\approx$, $>$, $\geq$,

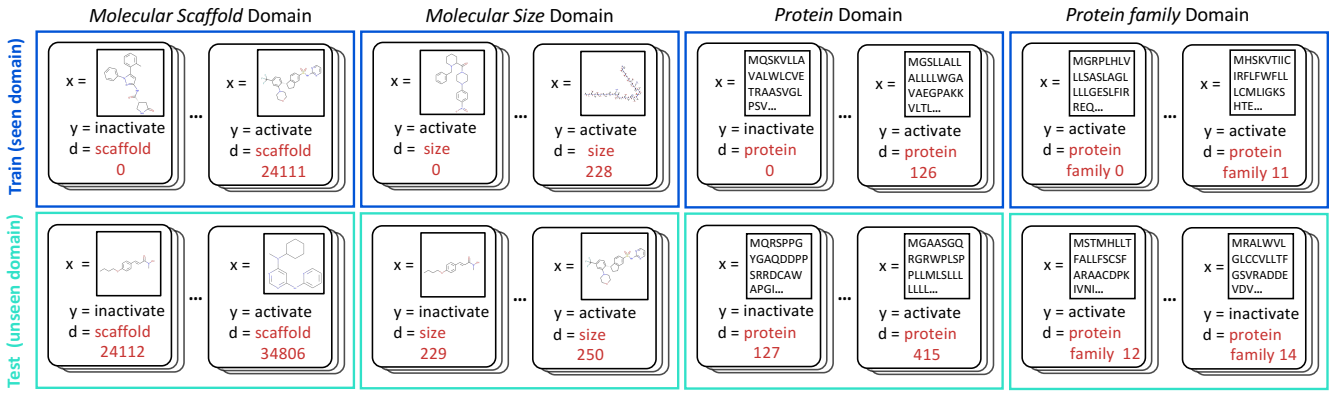

Figure 2: Exemplar curated datasets spanning different domain shifts from `DrugOOD`. Each data sample (x, y, d) in dataset is associated with a domain annotation d, which corresponds to a distribution P over data points which are similar in some way, e.g., molecules with the same scaffold. Specifically, `DrugOOD` focuses on the problem of domain generalization, in which we train (seen domain) and test (unseen domain) the model on disjoint domains, e.g., molecules with a new scaffold. Additionally, `DrugOOD` identifies and annotates three noise levels (core, refined, general), whose level increases with data volume and noise sources.

can cause noise in these data. Figure 2 shows examples with different noisy levels. Meanwhile, real-world data with noise level annotations is lacking for learning tasks under noise labels (Angluin and Laird 1988; Han et al. 2020).

To help accelerate research by focusing community attention and simplifying systematic comparisons between data collection and implementation method, we present `DrugOOD`, a systematic OOD dataset curator and benchmark for AI-assisted drug discovery that includes an open-source Python package which fully automates the data curation and OOD benchmarking processes. We focus on the most challenging OOD setting: domain generalization (Zhou et al. 2021) problem in AI-aided drug discovery, though `DrugOOD` can be easily adapted to other OOD settings, such as subpopulation shift (Koh et al. 2021) and domain adaptation (Zhuang et al. 2020). Our dataset is also the first AIDD dataset curator with realistic noise level annotations that can serve as an important testbed for the setting of learning under noise.

In contrast to only providing fixed datasets, we present an automated dataset curator based on the large-scale bioassay deposition website ChEMBL (Mendez et al. 2019). Figure 3 provides a summary of the automated dataset curator. Using this dataset curator, researchers/practitioners can generate new OOD datasets based on their needs by simply modifying the configuration files in the Python package. Specifically, we also realize this dataset curator by generating 45 OOD datasets spanning various domains, noise levels, and measurement types. This mechanism offers two benefits: i) It ensures accurate reproduction of our datasets and benchmarks, ii) It provides flexibility for future use, since it is often difficult, even for domain experts, to agree on a specific configuration. For example, agreeing on a threshold for partitioning IC50 meansurements to get active/inactive pairs for all domain experts might be challenging.

In summary, our contributions are fourfold:

- **Automated dataset curator**: We provide a fully customizable pipeline for curating OOD datasets for AI-aided drug discovery from the large-scale bioassay deposition website ChEMBL.
- **Rich domain annotations**: We present various approaches to generate specific domains that are aligned with the domain knowledge of biochemistry.
- **Realistic noise level annotations**: We aggregate real-world noise according to the measurement confidence score, "cut-off" noise, etc., offering a valuable testbed for learning under real-world noise.
- **Rigorous OOD benchmarking**: We benchmark six SOTA OOD algorithms with various backbones for the 45 realized dataset instances and gain insight into OOD learning under noise for AI-aided drug discovery.

Related Work

In this section, we review related literature from the perspectives of binding affinity prediction, out-of-distribution generalization, and commonly used drug discovery benchmarks. We refer the readers to Appendix A for more detailed discussions.

Binding Affinity Prediction. The aim of virtual screening is to identify the most promising molecules based upon both target-independent and target-dependent properties. The former evaluates the likelihood that one molecule itself qualifies as a drug candidate (e.g., toxicity and hydrophobicity), while the latter characterizes the tendency of its potential interaction with the target (and other unrelated proteins), which often heavily depends on the joint formulation of the candidate molecule and target (Hu, Chan, and You 2016; Karimi et al. 2019; Lim et al. 2019). In this paper, we focus on the binding affinity between the molecule and target protein, which falls into the domain of predicting target-dependent properties. In this circumstance, the out-of-distribution issue may result in severe performance degradation (e.g., when the distribution

etc, which would introduce the "cut-off" noise when translated into supervision labels.

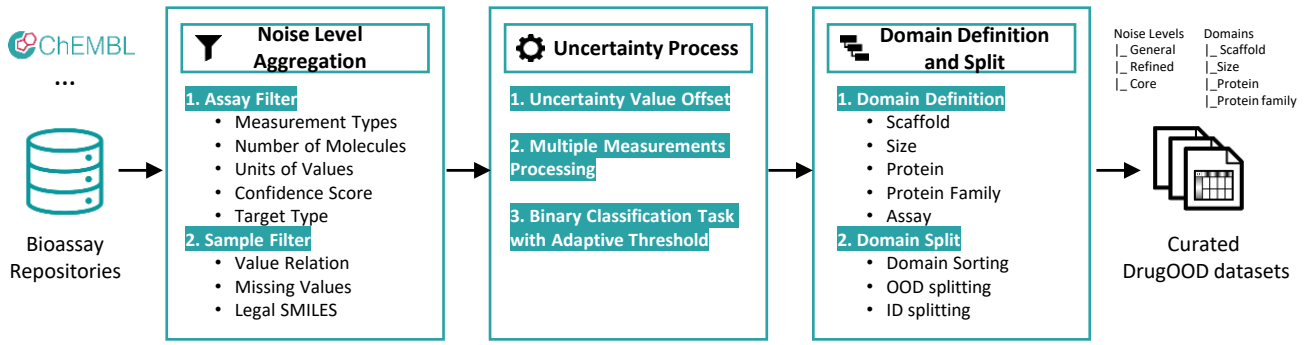

Figure 3: Workflow of the proposed dataset curator. We mainly implement three steps based on the ChEMBL data source, including noise level aggregation, uncertainty processing, as well as domain splitting.

of target proteins dramatically differs between model training and inference), leading to this work's major motivation.

Drug Discovery Benchmarks. There is a myriad of databases with biomedical and chemical information. ChEMBL (Mendez et al. 2019) is a large-scale, open-access drug discovery database that aims to capture medicinal chemistry data and knowledge across the pharmaceutical R&D process. BindingDB (Gilson et al. 2016) provides binding affinity data from scientific articles and US patents primarily; PDBbind (Liu et al. 2014) collects biomolecular complexes from the PDB database (Burley et al. 2020), along with experimental binding affinity data. Recently, several benchmarks for AIDD's development have been established. MoleculeNet (Wu et al. 2018) provides a collection of molecular property prediction tasks on which various featurization and models are benchmarked. TDC (Huang et al. 2021) provides 66 datasets across 22 tasks, covering the full pipeline of drug discovery. Recently, FS-Mol (Stanley et al. 2021) proposed a comprehensive benchmark for few-shot molecular learning. Nonetheless, most databases and benchmarks do not draw special attention to the challenge of out-of-distribution generalization. The training and evaluation subsets are often randomly partitioned, which could lead to an over-optimistic model evaluation. In contrast, `DrugOOD` covers comprehensive sources of out-of-distribution in structure-based affinity prediction and provides a dataset curator for highly customizable generation of OOD datasets. Additionally, noise level annotations are aggregated and considered so that algorithms can be evaluated in a more realistic setting, which further bridges the gap between research and pharmaceutical applications.

Out-of-Distribution Generalization. Some work focuses on aligning feature representations across different domains to improve the generalization ability over out-of-distribution test samples. The minimization of feature discrepancy can be conducted over various distance metrics, including second-order statistics (Sun and Saenko 2016a), the maximum means discrepancy (Tzeng et al. 2014), Wasserstein distance (Zhou et al. 2021), or measured by adversarial networks (Ganin et al. 2016a). Others apply data augmentation to generate new samples or domains to promote the consistency of feature representations, such as Mixup across existing domains (Xu et al. 2020; Yan et al. 2020), or in an adversarial manner (Zhao et al. 2020; Qiao, Zhao, and Peng 2020). With the label distribution further taken into consideration, recent work aims at enhancing the correlation between domain-invariant representations and labels. Invariant risk minimization (Arjovsky et al. 2019) seeks a data representation on which the optimal classifier is trained to match all training distributions. Additional regularization terms are proposed to align gradients across domains (Koyama and Yamaguchi 2021), reduce the variance of risks of all domains (Krueger et al. 2021), or smooth inter-domain interpolation paths (Chuang and Mroueh 2021). In `DrugOOD`, we provide rigorous benchmark tests over state-of-the-art OOD algorithms with a unified standard and offer performance analysis as well as suggestions for future research.

DrugOOD

In this work, we present an automated dataset curator and benchmark, named `DrugOOD`, based on the large-scale bioassay deposition website ChEMBL (Mendez et al. 2019), to facilitate OOD research for AI-aided drug discovery.

Automated Data Curator

Overview. We curate all datasets on the basis of the large-scale, open-access bioactivity website: ChEMBL[2], and consider the settings of OOD and of different noise levels. Specifically, we focus on one of the most critical AIDD tasks (Sliwoski et al. 2013): structure-based affinity prediction (SBAP), which involve both the target and compound information for predicting binding activity. The dataset curation pipeline is illustrated in Figure 3. It consists of three steps: 1) noise level aggregation with different filters; 2) processing uncertainty, averaging multi-measurements, and generating binary classification tasks with an adaptive threshold; and 3) splitting the domains by assays, scaffolds, sizes, proteins, or protein families. We provide 45 built-in configuration files to generate exemplar datasets by the configuration of three noise

[2]We use its latest release (released on Feb. 2022): ChEMBL30, downloaded from http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_30/chembl_30_sqlite.tar.gz

levels, three measurement types (Ki, IC50, EC50) and five domains. Besides, with our `DrugOOD` dataset curator, practitioners can easily generate their desired datasets through customizing the configuration files. This is especially useful when they want to use a new biochemistry setup.

Noise Level Aggregation. There are numerous heterogeneous noise sources in bioactivity data, such as the target type, value relation, confidence score, assay quality etc. In order to facilitate development of learning under noise, inspired by the practice of PDBbind (Gilson et al. 2016) which considers noise levels of protein-ligand complex data, we summarize the various types of noise and delineate different subsets based on noise severity. In particular, we aggregated chaotic data with three noise levels (*core, refined, general*) through the proposed various filter configurations. We refer the readers for more details in Appendix C.1.

Processing Uncertainty and Multiple Measurements. As mentioned above, many of the activity values recorded by ChEMBL are inherently uncertain. And the reported activity values may be higher or lower than the highest or lowest concentration tested. Here, we follow the practice in pQSAR 2.0 (Martin et al. 2017) to offset the activity values by 10-fold. Meanwhile, the protein-ligand pair reported in multiple sources may result in multiple assays, resulting in the "multiple measurements" problem. Following the common practice (Hu et al. 2020), we average the activity values of the same protein-ligand pair, which has been shown to be effective in previous work (Kalliokoski et al. 2013) and does not introduce excessive data bias.

Binary Classification Task. While ChEMBL records raw activity values as floating-point numbers, it is difficult to benchmark OOD tasks as regression tasks due to various noises, uncertainty measurements, and that low/high values are often only recorded as a boundary constant. In this work, we benchmark OOD tasks as a binary classification problem. In practice, however, the threshold for binary classification is contingent on the particulars of the drug development project. Here, we employ an adaptive thresholding method that is capable of adapting to a wider variety of situations. Details are deferred to Appendix C.2.

Domain Definition and Split. As previously stated, the distributional shift is a common occurrence in the drug development process. When predicting the bioactivities of paired protein-compound in the deployment scenario, the molecular scaffolds (Koh et al. 2021), sizes, the protein, protein family, etc. may differ significantly from the training data. To meet this purpose of covering a wide range of shifts that naturally occur in the area of AIDD, we cautiously consider five domains, including assay, molecular size, molecular scaffold, protein, and protein family.

- **Molecular Scaffold:** Molecular scaffold plays a critical role in driving molecules to show different properties. Following the strategies in (Koh et al. 2021; Hu et al. 2020), the molecules with the same molecular scaffold are grouped into the same domain, and the model needs to be able to generalize to unseen domains with novel scaffolds.

Data subset	#Domain	#Sample
sbap-core-ic50-assay	1,503	123,028
sbap-refined-ic50-assay	6,635	348,248
sbap-general-ic50-assay	22,376	552,347
sbap-core-ic50-protein	693	123,028
sbap-refined-ic50-protein	1,515	348,248
sbap-general-ic50-protein	2,608	552,347
sbap-core-ic50-protein-family	13	123,028
sbap-refined-ic50-protein-family	15	348,248
sbap-general-ic50-protein-family	15	552,347

Table 1: Statistical information of some realized datasets. #Domain and #Sample represents the number of domains and data points respectively.

- **Molecular Size:** Similarly, the size of the molecule is often related to the biomedical properties (Bevilacqua, Zhou, and Ribeiro 2021; Yehudai et al. 2021). Thus, the molecules with same size (i.e., the number of atoms) are categorized into one domain, and models need to infer on unseen domains with varied molecular sizes.
- **Protein:** In this case, paired protein-compound data with the same target protein are grouped into the same domain. Model performance is estimated on data samples with a never-seen-before protein target.
- **Protein Family:** In a similar vein, data with targets from the same protein family are categorized as one domain. Compared with protein domains, there are much fewer domains albeit with greater differences from each other.
- **Assay:** The data samples generated from the same binding assay are classified into one domain. Due to the differences in different binding assay environments, the activity values measured by different assays will naturally have a distribution shift. Under this setting, the model needs to test on data of unseen bioassay environments.

For the next step, we need to split the data with domain annotations into training, OOD validation, and OOD testing sets and ensure sufficient domain shifts amongst them. This raises the question of how domain differences can be measured and sorted into subsets. Here, we design a general pipeline, that is, firstly generate *domain descriptor* for each domain, and then sort the domains with descriptors. Then the sorted domains are sequentially divided into the training set, OOD validation, and testing set (Figure 6 (a) in Appendix). Meanwhile, the number of domains in different splits is controlled by the number of total samples in each splits, and the proportion of sample numbers is kept at 6:2:2 for training, validation, and testing. For additional details on *domain descriptor* and data split scheme, please refer to Appendix C.3.

Statistics of Exemplar Curated Datasets. In order to realize the above curation process, `DrugOOD` provides a total of 45 built-in datasets, with different noise levels, different domain definitions, as well as three affinity measurements (i.e., IC50, EC50, Ki)[3]. The statistics of resulting domains and samples are partially displayed in Table 1. It can be

[3]We were also considering the "Potency" measurement type. Since it was verified to have much higher noise than these three types, we decided to not put it in the exemplar datasets.

Domain	Val (ID)		Val (OOD)		Test (ID)		Test (OOD)	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Assay	87.90 (0.48)	89.93 (2.49)	91.31 (0.38)	68.60 (0.20)	87.67 (0.53)	89.73 (2.52)	81.97 (0.23)	70.86 (0.38)
Molecular Scaffold	94.56 (0.26)	92.20 (1.13)	85.44 (0.08)	78.80 (1.20)	89.62 (0.11)	84.24 (1.83)	77.06 (0.05)	68.74 (1.07)
Molecular Size	93.10 (0.05)	92.33 (0.11)	84.09 (0.06)	81.90 (0.50)	93.10 (0.08)	92.75 (0.24)	72.86 (0.18)	66.37 (0.35)
Protein	88.92 (0.24)	89.62 (1.49)	85.74 (0.24)	72.26 (0.85)	89.23 (0.30)	90.32 (1.49)	83.04 (0.25)	68.62 (0.45)
Protein-family	88.30 (0.38)	86.97 (2.56)	92.25 (0.13)	66.38 (0.60)	87.91 (0.47)	86.79 (2.85)	79.80 (0.23)	71.84 (1.01)

Table 2: Results of ERM on datasets with different domain shift under the core noise level: sbap-core-ic50-assay, sbap-core-ic50-scaffold, sbap-core-ic50-size, sbap-core-ic50-protein, and sbap-core-ic50-protein-family. Parentheses show standard deviation across 3 replicates.

seen that as the number of samples increases, the noise level also increases. Meanwhile, for the same noise level, there are huge differences in the number of domains generated by different domain split methods, which will challenge the applicability of OOD algorithms for different domain numbers. In order to show the comparison of data volume under different measurement types, we count the samples of different measurement types under different noise levels, as shown in Figure 7 of the appendix. As we can see, the number of samples varies greatly under different measurement types in ChEMBL. Meanwhile, different measurement types may also bring different noise levels. Besides, our curation process can generate specific measurement types of datasets according to the needs of specific drug development scenarios. More statistical information on the `DrugOOD` datasets under different settings are summarized in Table 7 of the Appendix.

Extension to LBAP Task. We have also extended `DrugOOD` for the LBAP (ligand based affinity prediction) task, which aims to predict affinity for a molecular input. Further details are in the full version on the project page.

Benchmarking State-of-the-art OOD Algorithms

To comprehensively evaluate different learning algorithms on the proposed `DrugOOD` datasets, we implement and evaluate algorithms that aim to address the distribution shift problem from two perspectives, i.e., architecture design and domain generalization algorithms. To our best knowledge, this is the first work to evaluate a large set of approaches in different settings of the Drug OOD problem.

Architecture Design. It is well known that the power of a model depends significantly on its network architecture. Popular research topics of OOD data problems include how to design network architecture for improved ability to fit the target function and noise resilience. Based on the DGL-LifeSci package (Li et al. 2021), we benchmark and evaluate nine graph-based and two transformer-based backbones, including GIN (Xu et al. 2018), GAT (**?**), SchNet (Schütt et al. 2017), GCN (Kipf and Welling 2016), Weave (Kearnes et al. 2016), MGCN (Lu et al. 2019), ATi-FPGNN (Xiong et al. 2019), NF (Duvenaud et al. 2015), GTransformer (Rong et al. 2020), Bert (Devlin et al. 2018) and ProteinBert (**?**). For the main experiments in this paper, we use a standard model structure for each type of data: GIN (Kipf and Welling 2016) for molecular graphs and Bert (Devlin et al. 2018) for protein amino acid sequences, a readout function, and an MLP layer is further extended for the classification task.

Domain Generalization Algorithms. In the traditional machine learning area, models are usually optimized by the empirical risk minimization (ERM) algorithm, which trains the model to minimize the average training loss across all the training domains. Since ERM is proposed assuming that the training and testing data share the same distribution, it can be sensitive to the distribution shift between training and test data and may not handle such a situation well. To address such a problem, many methods try to improve the model's robustness from various perspectives. In this work, in addition to the ERM baseline, we implement and evaluate several representative OOD methods, including IRM (Arjovsky et al. 2019), DeepCoral (Sun and Saenko 2016b), DANN (Ganin et al. 2016b), Mixup (Zhang et al. 2017), and GroupDro (Sagawa et al. 2019). Detailed information about them is deferred to Appendix D.2.

Empirical Studies

In this section, we perform experimental validation on the realized datasets for the SBAP task to investigate the rationality of `DrugOOD`. Further details on the LBAP (ligand based affinity prediction) tasks can be found at the project homepage (https://drugood.github.io). First, we introduce the experimental settings, including the problem definition and implementation details. Then, experiments are carefully designed to report the performance and findings from multiple perspectives.

Implementations

SBAP Problem Definition. Precisely predicting the affinity of paired protein-compound will greatly boost the process of drug discovery by reducing the need for costly laboratory experiments. In this paper, we study a domain generalization problem where the model needs to be generalized to the paired protein-compound from different domain splits. As an illustration, we treat the SBAP problem as a binary classification problem, where the input x is the paired input of a small molecule and target protein, label y is the ground truth (active or inactive) of binary affinity classification, and the d represents domain identifier for one specific domain splits.

Data Info. As mentioned before, we run the designed data curator and generate 45 exemplar datasets with varying noise levels, measurements types, and domain definitions. Each

Algos	Assay		Scaffold		Size		Protein		Protein family	
	Test (ID)	Test (OOD)	Test (ID)	Test (OOD)	Test (ID)	Test (OOD)	Test (ID)	Test (OOD)	Test (ID)	Test (OOD)
ERM	89.73	70.86	84.24	68.74	92.75	66.37	90.32	68.62	86.79	71.84
IRM	83.55	68.72	83.87	67.74	74.92	56.62	91.29	67.66	85.63	70.44
DeepCoral	84.91	68.68	80.21	67.83	79.34	59.41	90.33	67.26	_	_
DANN	75.98	65.16	75.86	64.18	90.60	66.05	78.12	62.58	86.00	70.28
Mixup	88.07	70.85	86.05	68.61	91.98	66.21	91.11	68.25	86.14	73.10
GroupDro	84.31	68.49	81.11	67.79	82.71	59.92	89.36	67.62	87.72	72.76

Table 3: Baseline results of six OOD algorithms on dataset with different domain shift: sbap-core-ic50-assay, sbap-core-ic50-scaffold, sbap-core-ic50-size, sbap-core-ic50-protein, and sbap-core-ic50-protein-family. The results are reported in AUC score. - means unavailable experiments due to limited numbers of domain.

small molecule in each dataset is represented as a graph, where the nodes are atoms and edges are chemical bonds. While the protein is represented as amino acid sequences, it can be easily extended by incorporating 3D structure information of protein targets by referring to protein structure deposition databases, such as PDB (Berman et al. 2000) and UniProt (Consortium 2014). This will be left as an important future work. Following the pre-processing strategy of (Xiong et al. 2019), we preprocess the molecules via the RDKit package (**?**). Input node features are 39-dimensional vectors including atomic symbol, hybridization, hydrogens, etc. Input edge features are 10-dimensional vectors including bond type, conjugation, ring and bond stereo chemistry.

Model Details. We used a standard two-tower model for the SBAP task, in which two networks extract molecular and protein features independently, the generated features are then concatenated and fed into a fully connected layer to predict the interaction probabilities. We train the model on each dataset from scratch with a learning rate at 1e-4, a batch size of 256 samples, and without L2-regularization. For molecular inputs, we used the GIN backbone (Xu et al. 2018) to extract 256-dimensional features; for protein sequences, we used the pre-trained Bert (Devlin et al. 2018) 'bert-base-uncase' to extract the 768-dimensional protein representations. To avoid performance degradation caused by inappropriate hyper-parameters, following the strategy in WILDS (Koh et al. 2021), we conducted a grid search strategy over learning rates of $\{0.00003, 0.0001, 0.0005, 0.001, 0.01\}$, batch size of $\{64, 128, 256, 512, 1024\}$. We report averaged results aggregated over 3 random seeds.

Evaluation Metric. We evaluate models' performance by the area under the receiver operating characteristic (AUROC), which indicates the ability of a classifier to distinguish between classes (e.g., inactive or active). The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Meanwhile, we also provide the results of accuracy metric.

Experimental Results

ERM Results and Performance Drops. As shown in Table 2, model performance dropped significantly going from the in-distribution (ID) setting to the official out-of-distribution (OOD) setting. For the assay domain, ERM achieves an average AUC score of 87.90% on the ID validation set and 87.67% on the ID test set, but only 68.60% on the OOD val set, 70.86% on the OOD test set. Similarly, for the scaffold and size domain, ERM obtains 84.24%, 92.75% AUC score on the ID test set but 68.74%, 66.37% on the OOD test set, respectively. The test performance of ERM drops by 18.87%, 15.50%, 26.38%, 21.7%, 14.95% points AUC score when the assay, scaffold, size, protein, and protein family split are used, respectively, suggesting that these splits are indeed harder than conventional random split, and can be used to estimate the realistic ID-OOD gap in the task of structure based affinity prediction.

Results of Other Baselines. Table 3 shows the performance of other representative domain generalization (DG) algorithms. For a fair comparison, all algorithms adopt the same backbone network. Besides, we also make additional grid searches on algorithms' specific hyper-parameters separately: IRM's penalty weight in $\{1, 10, 100, 1000\}$ and penalty anneal iteration in $\{100, 500, 1000\}$. DeepCoral's penalty weight in $\{0.1, 1, 10\}$, GroupDro's step size in $\{0.001, 0.01, 0.1\}$, DANN's inverse factor between $\{0.0001, 1\}$ and Mixup's probability and interpolate strength between $\{0.0001, 1\}$. As shown in Table 3, ERM almost always performs better than DeepCoral, IRM, and Group DRO for the five domain splits, indicating these existing domain generalization methods can not well solve the defined OOD problems. Moreover, similar to the findings of the WILDS benchmark, current DG methods render the model hard to fit the training data, For instance, under the size domain split, DeepCoral, IRM achieves 80.21%, 83.87% AUC score in the ID test set, respectively, while the AUC score of ERM baseline is 92.75%. Also, these methods are primarily designed for the case when each group contains a decent number of samples, which is not the common case for the drug development scenario. Finally, the SOTA OOD algorithms do not work well in the `DrugOOD` setting, suggesting that better methods need to be developed to solve the OOD problem for graph data.

Performance Drops of Different Domains Figure 4 shows the performance degradation for different domain partitions, the gap values are computed on the test set and averaged over all measurement types. From the results, we can conclude the following. 1) Among the domain splits, the size domain often brings the largest performance degradation, which is consistent with daily experimental findings,

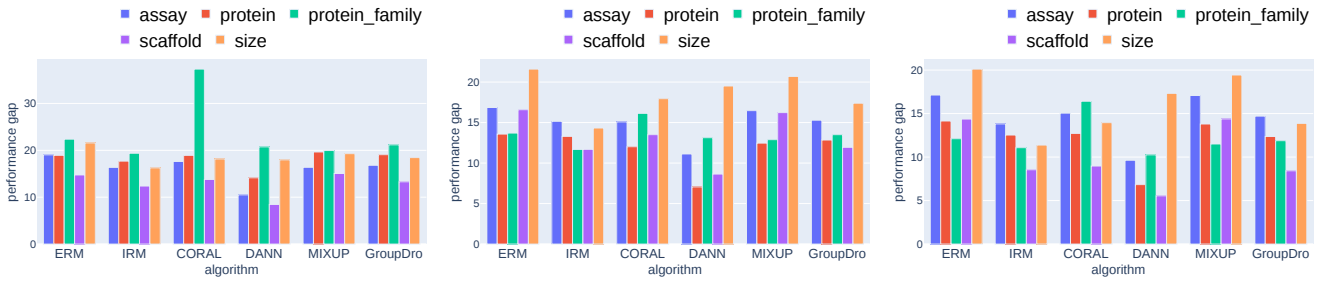

Figure 4: Performance gap on AUC of different domain splits with different noise levels and OOD algorithms on the DrugOOD-sbap datasets. The gap values are calculated on the OOD test set and averaged over measurement types. Left to right: core noise level, refined noise level, general noise level. Color indicates domain splits for SBAP tasks.

Noise Level	Val (ID)	Val (OOD)	Test (ID)	Test (OOD)
Core	89.62	72.26	90.32	68.62
Refined	82.87	73.53	82.92	68.00
General	78.97	71.63	78.94	68.06

Table 4: AUC score of ERM on datasets with assay shift under different noise levels: sbap-core-ic50-assay, sbap-refined-ic50-assay, sbap-general-ic50-assay.

where molecules of different sizes often have very different properties. 2) As the noise level of the data set increases, the performance degradation of different domains is somewhat mitigated, and the increased data to some extent increase the generalization ability of the model. In addition, the degradation of the model performance is not further mitigated as the data continues to increase, indicating that the increase in the amount of data brings limited improvement and that a truly effective method needs to be developed to address the noise issue.

Performance Drops of Different Noise Levels We also investigated the ID and OOD performance at different noise levels. As shown in Table 4, we summarize the ERM baseline's ID and ODD performance under three noise levels. Limited by the page left, the reader can refer to Table 6 of the Appendix for other algorithms' detailed results. One can observe that: 1) In the presence of more noise, the introduced noise produces pollution, which progressively affects the model's performance. 2) By increasing the noise level, more data is collected, providing more information about the dataset. There is a narrowing of the gap between ID and OOD performance from core to refined levels. There is, however, no significant improvement from the refined level to the general level as the improvement reaches a bottleneck. 3) By combining the above two points, we can see that the introduction of large amounts of data with noise affects the learning of the model to a certain degree. However, the noisy data, in turn, provides additional information that can be used to improve the generalization capability of the model.

Studies for Different Measurement Types. The automated dataset curator supports variant measurement types, e.g., EC50 and IC50. As different measurement types will generate datasets with different distributions and noise, we analyze the performance of different measurement types by varying the noise levels and algorithms. As shown in Figure 9 of the appendix, one can see that: 1) For different measurement types, ID and OOD performance can differ. This might be because of varying amounts of data, and data collection procedures of different measurement types. 2) For almost all types of measurements, the benchmarked algorithms have acceptable accuracy.

Discussions and Future Work

In this work, we have presented an automated dataset curator and benchmark based on the large-scale bioassay deposition website ChEMBL, in order to facilitate OOD research for AIDD. It is very worthwhile to explore more in the following respects. As observed in current benchmark results, existing general OOD methods do not significantly outperform the baseline ERM method. Most of these OOD methods are designed and validated with visual and/or textual data, which may fail in capturing critical information for the affinity prediction problem. This implies that to further improve the performance under various OOD scenarios, it is essential to develop more advanced OOD methods, particularly with drug-related domain knowledge integrated. Another key characteristic of DrugOOD database is that the majority of data falls into the highest noise level ("general"). Simply discarding such noisy labels and only referring to high-quality ones may severely limit the model performance due to insufficient training data. It would be worthwhile investigating whether large-scale unsupervised pre-training methods can be utilized to construct better representations for molecules and target proteins, which are critical to accurate affinity predictions. Additionally, learning with noisy labels has been extensively studied in the general context, but it may be crucial to take the generation process of noisy affinity annotations into consideration. This includes different experimental precision, measurement types, activity relation annotation types, etc. It is possible that data quality can be further improved with carefully-designed denoising techniques, so that more accurate affinity prediction models can be trained. For *societal impact*, we foresee the following positive impacts: a potentially positive impact on the ability to withstand unexpected epidemic diseases, a potentially positive impact on reducing the time and cost required to bring drugs to market, and the potential to leverage large-scale computation-based, data-driven approaches to develop more effective, targeted therapies.

References

Angluin, D.; and Laird, P. 1988. Learning from noisy examples. *Machine Learning*, 2(4): 343–370.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research*, 28(1): 235–242.

Bevilacqua, B.; Zhou, Y.; and Ribeiro, B. 2021. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*, 837–851. PMLR.

Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ganesan, S.; Goodsell, D. S.; Ghosh, S.; Green, R. K.; Guranović, V.; Guzenko, D.; Hudson, B. P.; Lawson, C.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Persikova, I.; Randle, C.; Rose, A.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Tao, Y.-P.; Voigt, M.; Westbrook, J.; Young, J. Y.; Zardecki, C.; and Zhuravleva, M. 2020. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1): D437–D451.

Chuang, C.-Y.; and Mroueh, Y. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*.

Consortium, T. U. 2014. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1): D204–D212.

Cortés-Ciriano, I.; and Bender, A. 2016. How consistent are publicly reported cytotoxicity data? Large-scale statistical analysis of the concordance of public independent cytotoxicity measurements. *ChemMedChem*, 11(1): 57–71.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2224–2232.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016a. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016b. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; and Chong, J. 2016. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1): D1045–D1053.

Han, B.; Yao, Q.; Liu, T.; Niu, G.; Tsang, I. W.; Kwok, J. T.; and Sugiyama, M. 2020. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*.

Hu, P.-W.; Chan, K. C.; and You, Z.-H. 2016. Large-scale prediction of drug-target interactions from deep representations. In *2016 international joint conference on neural networks (IJCNN)*, 1236–1243. IEEE.

Hu, R.; Xu, H.; Jia, P.; and Zhao, Z. 2020. KinaseMD: kinase mutations and drug response database. *Nucleic Acids Research*, 49(D1): D552–D561.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv e-prints*, arXiv:2005.00687.

Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; and Zitnik, M. 2021. Therapeutics data Commons: machine learning datasets and tasks for therapeutics. *arXiv preprint arXiv:2102.09548*.

Kalliokoski, T.; Kramer, C.; Vulpetti, A.; and Gedeck, P. 2013. Comparability of mixed IC50 data–a statistical analysis. *PloS one*, 8(4): e61007.

Karimi, M.; Wu, D.; Wang, Z.; and Shen, Y. 2019. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18): 3329–3338.

Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; and Riley, P. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B.; Haque, I.; Beery, S. M.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5637–5664. PMLR.

Koyama, M.; and Yamaguchi, S. 2021. When is invariance useful in an Out-of-Distribution Generalization problem ? arXiv:2008.01883.

Kramer, C.; Kalliokoski, T.; Gedeck, P.; and Vulpetti, A. 2012. The experimental uncertainty of heterogeneous public K i data. *Journal of medicinal chemistry*, 55(11): 5165–5173.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 5815–5826. PMLR.

Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; and Karypis, G. 2021. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS omega*, 6(41): 27233–27238.

Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; and Kim, W. Y. 2019. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of chemical information and modeling*, 59(9): 3981–3988.

Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; and Wang, R. 2014. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3): 405–412.

Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; and He, L. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1052–1060.

Martin, E. J.; Polyakov, V. R.; Tian, L.; and Perez, R. C. 2017. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *Journal of Chemical Information and Modeling*, 57(8): 2077–2088. PMID: 28651433.

Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1): D930–D940.

Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. 2020. QSAR without borders. *Chemical Society Reviews*, 49(11): 3525–3564.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to Learn Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *NeurIPS*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Schneider, G. 2018. Automating drug discovery. *Nature reviews drug discovery*, 17(2): 97–113.

Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566*.

Shen, T.; Wu, J.; Lan, H.; Zheng, L.; Pei, J.; Wang, S.; Liu, W.; and Huang, J. 2021. When homologous sequences meet structural decoys: Accurate contact prediction by tFold in CASP14—(tFold for CASP14 contact prediction). *Proteins: Structure, Function, and Bioinformatics*, 89(12): 1901–1910.

Sliwoski, G.; Kothiwale, S.; Meiler, J.; and Lowe Jr., E. W. 2013. Computational methods in drug discovery. *Pharmacological reviews*, 66(1): 334–395. 24381236[pmid].

Stanley, M.; Bronskill, J. F.; Maziarz, K.; Misztela, H.; Lanini, J.; Segler, M.; Schneider, N.; and Brockschmidt, M. 2021. FS-Mol: A Few-Shot Learning Dataset of Molecules. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Sun, B.; and Saenko, K. 2016a. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, 443–450. Springer.

Sun, B.; and Saenko, K. 2016b. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, 443–450. Springer.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. arXiv:1412.3474.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.

Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16): 8749–8760.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*.

Xu, M.; Zhang, J.; Ni, B.; Li, T.; Wang, C.; Tian, Q.; and Zhang, W. 2020. Adversarial Domain Adaptation with Domain Mixup. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 6502–6509. AAAI Press.

Yan, S.; Song, H.; Li, N.; Zou, L.; and Ren, L. 2020. Improve Unsupervised Domain Adaptation with Mixup Training. arXiv:2001.00677.

Yehudai, G.; Fetaya, E.; Meirom, E.; Chechik, G.; and Maron, H. 2021. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, 11975–11986. PMLR.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. 2020. Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 14435–14447. Curran Associates, Inc.

Zhou, F.; Jiang, Z.; Shui, C.; Wang, B.; and Chaib-draa, B. 2021. Domain Generalization via Optimal Transport with Metric Similarity Learning. *Neurocomputing*, 456(C): 469–480.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.