Bias in the Benchmark: Systematic experimental errors in bioactivity databases confound multi-task and meta-learning algorithms

Leo Klarner [*1] **Michael Reutlinger** [2] **Torsten Schindler** [2] **Charlotte Deane** [1] **Garrett Morris** [1]

Abstract

There is considerable interest in employing deep learning algorithms to predict pharmaceutically relevant properties of small molecules. To overcome the issues inherent in this low-data regime, researchers are increasingly exploring multi-task and meta-learning algorithms that leverage sets of related biochemical and toxicological assays to learn robust and generalisable representations. However, we show that the data from which commonly used multi-task benchmarks are derived often exhibits systematic experimental errors that lead to confounding statistical dependencies across tasks. Representation learning models that aim to acquire an inductive bias in this domain risk compounding these biases and may overfit to patterns that are counterproductive to many downstream applications of interest. We investigate to what extent these issues are reflected in the molecular embeddings learned by multi-task graph neural networks and discuss methods to address this pathology.

1. Introduction

The impressive performance of modern deep learning algorithms in traditional machine learning domains such as computer vision and natural language processing has led to a resurgence of interest in applying them to various problems throughout the drug discovery pipeline. While many parts of the pre-clinical drug discovery process stand to benefit from robust computational models, the tasks of predicting the bioactivity or toxicity of a molecule from its structure have emerged as one of the most popular benchmarks for algorithmic progress in this area (Wu et al., 2018; Mayr et al., 2018).

One of the fundamental practical issues of adapting existing and developing novel approaches to this problem is that the size of available high-quality labelled datasets is orders of magnitude below that of the well-known benchmarks to which much of the recent progress in e.g. computer vision has been ascribed (Sun et al., 2017). This limitation stems from the inherently experimental nature of medicinal chemistry, and while high-throughput screening (HTS) technologies are steadily improving, it is unlikely that this bottleneck will be overcome in the near future.

To train models that are viable in this low-data regime, a considerable amount of recent work has investigated the ability of multi-task and meta-learning algorithms to leverage a collection of related datasets (e.g. Ramsundar et al. (2015); Lenselink et al. (2017); Mayr et al. (2018); Nguyen et al. (2020); Stanley et al. (2021)). The rationale behind these approaches is that requiring a model to perform well on a set of related but distinct tasks leads to more powerful and generalisable representations, mediated by both the increased quantity of training data and the implicit regularisation of optimising multiple predictive performances. This objective of attending to underlying causal factors that are shared across tasks, while de-emphasising task-specific noise, is referred to as learning a domain-specific inductive bias (Caruana, 1997).

However, this approach is clearly counter-productive if the set of tasks from which an inductive bias is derived exhibits statistical dependencies that are adversarial to a particular downstream application of interest. For example, one of the main limitations of high-throughput bioactivity and toxicity screens is their tendency to produce large numbers of reproducible false positives. These well-known artefacts often stem from compound-dependent interference with the assay system and can be responsible for up to 95% of ostensibly active compounds (Thorne et al., 2010). As many of the mechanisms that are known to produce these false positives act orthogonally to any specific biochemical interaction of interest, it is plausible to assume that they persist across different assays, which is indeed what is observed in practice (Baell & Holloway, 2010; M Nissink & Blackburn, 2014; Schorpp et al., 2014).

[*]Equal contribution [1]Department of Statistics, University of Oxford [2]Pharma Research and Early Development, Roche Innovation Center Basel. Correspondence to: Leo Klarner <leo.klarner@stats.ox.ac.uk>.

2nd AI4Science Workshop at the 39^{th} International Conference on Machine Learning (ICML), 2022. Copyright 2022 by the author(s).

The prevalence of these falsely active compounds and their recurrence across biologically unrelated systems presents a challenge for many multi-task and meta-learning algorithms. If a substantial proportion of positive labels can be consistently attributed to a small set of assay-interfering substructures, this may present a more attractive inductive bias than learning complex biochemical interactions, especially if the representational capacity of a model is restricted by further regularisation, as is common in low-data domains.

In the following, we will review how bioactivity data is measured and highlight different forms of assay interference (Section 2). We then demonstrate that the resulting biases are prevalent many popular bioactivity benchmarks and influence the representations learned by multi-task graph neural networks (Sections 3 and 4). Finally, we propose approaches to address this issue (Section 5).

2. Measuring and Mismeasuring Bioactivity

Identifying molecules that have a strong effect on a given target is challenging, as usually only very few compounds elicit a desired response. This is compounded by the fact that chemical space is vast, with estimates for the number of unique drug-like molecules ranging from 10^{20} to 10^{60} (Bohacek et al., 1996; Ertl, 2003; Polishchuk et al., 2013).

To increase the coverage with which molecules can be investigated, drug discovery research makes use of high-throughput screens (HTS) to identify promising lead candidates for extensive follow-up studies. These usually consist of a large molecular library being passed in an automated fashion through a miniaturised assay system that translates the physiological effect of a molecule into an easily measurable readout.

As in any chemical system, the magnitude of the effect that a compound exerts is dependent on its concentration, and the objective of most screening campaigns is to find molecules that show a strong biological effect at a minimal dose. A common approach to identifying such compounds is to first assay the entire compound library at a single concentration, and then select the molecules with the strongest response to be re-measured in a dilution series. A logistic model is then fit to this dose-response data and used to derive quantitative measures of the binding dynamics, such as the widely-used half-maximal active concentration[1].

One common issue with HTS assay systems is that the biological target of interest is not the only or even the main component that molecules can interact with. Different classes of substructural motifs can elicit a wide variety of behaviours that are orthogonal to the physiological interaction being investigated, but nevertheless result in a positive readout. Examples include compound aggregation (McGovern et al., 2002; Ryan et al., 2003; Feng et al., 2007), reactivity towards the protein target (Rishton, 1997), and interference with the readout method (Baell & Holloway, 2010; Baell & Nissink, 2018).

As these confounding interactions are not experimental noise, but reproducible, concentration-dependent, and often platform-independent positive readouts, it is generally difficult to differentiate them from truly active compounds without extensive validation experiments (termed counter-screens). Researchers have extensively examined historic HTS data and collated lists of molecular substructures that are enriched in compounds with high frequencies of hits, exemplified by the well-known library of Pan-Assay INterference compoundS, or PAINS (Baell & Holloway, 2010).

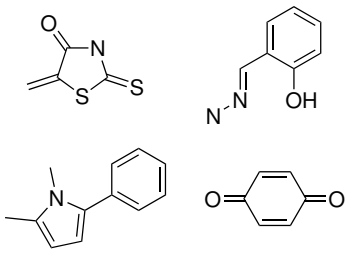

Figure 1. Representative structural submotifs that are correlated with hit frequencies, adapted from (Baell & Holloway, 2010).

And while medicinal chemists have grown cognisant of seemingly active compounds containing known assay-interfering substructures, they remain prevalent in popular bioactivity databases and published research (Dahlin & Walters, 2016). Moreover, many unidentified or less extensively characterised substructural motifs are bound to exist in high-throughput screening datasets, further complicating analysis.

As outlined in Section 1, representations learned by multi-task and meta-learning models trained on datasets that are populated with these problematic compound sets risk favouring this more accessible inductive bias over the challenging and complex biochemical interactions that are of practical relevance.

3. Bias in Bioactivity Benchmarks

To quantify the prevalence of this issue in multi-task bioactivity prediction datasets, we investigated the extent to which the hit frequency of a compound across screens can predict binding in any individual screen with a very simple model.

[1] i.e. the concentration at which a compound exerts half of its maximal physiological effect; depending on the context often also referred to as the half-maximal inhibitory/effective/lethal concentration/dose (IC_{50}/EC_{50}/LD_{50})

Consider the label matrix $\mathbf{Y} \in \{0, 1, -\}^{|M|\times|A|}$ of a multi-task binary classification problem, denoting whether a molecule $m_i \in M = \{m_1, m_2, \ldots, m_L\}$ was inactive, active, or not measured in an assay $a_j \in A = \{a_1, a_2, \ldots, a_K\}$. For all molecules that were measured in a given assay a_j

$$M_j = \{m_i \in M | \mathbf{Y}_{ij} \in \{0, 1\}\}$$

we calculate the frequency of hits across all other assays they were measured in

$$h_{ij} = \frac{|\{a_k \in A \setminus j | \mathbf{Y}_{ik} = 1\}|}{|\{a_k \in A \setminus j | \mathbf{Y}_{ik} \in \{0, 1\}\}|}$$

and use these hit frequencies to calculate the areas under the receiver operator characteristic (AUC-ROC) and the precision recall curve (AUC-PRC) with respect to the labels in a_j. Molecules that are only measured in a single assay are assigned the global hit frequency across $\mathbf{Y}$[2]. The summary of this characterisation is shown in Table 1, where the metrics are averaged across all assays and compared against the expected performance of a random classifier using the label ratio of each a_j.

If the assays in A investigate reasonably distinct biological systems, as is the case in public repositories such as PubChem (Kim et al., 2021) and ChEMBL (Gaulton et al., 2017), this approach is not expected to perform significantly better than a random classifier. These databases cover hundreds of different targets spanning many physiological functionalities and scales, meaning that any consistent activity across screens is much more likely to stem from target-independent interference mechanisms than selective biochemical interactions.

To characterise this behaviour, five multi-task benchmarks were retrieved from their respective sources and the MoleculeNet repository (Wu et al., 2018), including:
CHEMBL - a preprocessed subset of the ChEMBL database (Mayr et al., 2018), designed for multi-task bioactivity prediction;
MUV - an extensively preprocessed subset of the PubChem database (Rohrer & Baumann, 2009), originally designed for virtual screening;
PCBA - a large, minimally preprocessed subset of the PubChem database (Ramsundar et al., 2015), designed for multi-task bioactivity prediction; and
TOX21/TOXCAST - a collection of assays measuring binding to targets involved in known adverse physiological reactions (Dix et al., 2007).

The degree to which the issue of target-independent assay interference is addressed varies between benchmarks.

[2]Alternative approaches of assigning the hit rate of a_j or ignoring these compounds entirely only negligibly changes the results.

Table 1. The AUC-ROC and AUC-PRC classification metrics of the frequency of hits model discussed in Section 3, with the expected performance of a random classifier in parentheses.

DATASET	AUC-ROC	AUC-PRC	$\|A\|$	$\|M\|$
CHEMBL	0.76 (0.50)	0.69 (0.43)	1310	456 331
PCBA	0.80 (0.50)	0.12 (0.02)	128	439 863
MUV	0.57 (0.50)	0.01 (0.00)	17	93 127
TOX21	0.81 (0.50)	0.32 (0.08)	12	8 014
TOXCAST	0.79 (0.50)	0.44 (0.20)	617	8 615

While the authors of the PCBA dataset acknowledge the possibility of multi-task models focussing on interfering substructures, they argue that the predictive performance of networks trained on cleaner data alleviates this concern. The authors of the MUV dataset go further and apply extensive preprocessing steps, including the removal of all compounds with a hit rate over 26%. The curators of TOXCAST pursue an even more rigorous approach and perform extensive experimental counter-screens for each assay. However, in the form that this dataset is currently distributed and used these counter-screens are simply treated as additional predictive tasks, strongly amplifying the adverse incentive presented by assay-interfering compounds.

The results of our experiment in Table 1 closely mirror these considerations. The datasets that have not been preprocessed to remove potential artefacts (CHEMBL, PCBA, TOX21, and TOXCAST) show a substantial correlation between the readouts of different assays. The only dataset on which the performance is close to random is MUV, where many potentially interfering compounds are filtered out.

4. Impact on Learned Representations

To investigate to what extent these statistical dependencies are reflected in the representations that models learn, we examined the molecular embeddings that a multi-head graph neural network derives from the CHEMBL dataset. Specifically, we used the graph isomorphism network (Xu et al., 2018) architecture introduced in Hu et al. (2019), consisting of a representation-learning network $\psi : \mathcal{G} \to \mathbb{R}^d$ that maps a molecular graph $\mathcal{G}$ to a continuous embedding in $\mathbb{R}^d$, followed by a multi-head linear layer $\phi : \mathbb{R}^d \to \mathbb{R}^{|A|}$ to model bioactivity labels. In the following experiments, we compare the molecular representations generated by a random initialisation[3] of ψ to those that were optimised to perform well in the supervised multi-task setting by (Hu et al., 2019).

[3]In contrast to other deep learning models, it is well-established that graph neural networks can already extract highly useful representations in their randomly initialised state (Kipf & Welling, 2016; Hamilton et al., 2017; Velickovic et al., 2019).

Our hypothesis is that the strong predictive performance of a frequency of hits regression model demonstrated above leads to representations that encode molecular promiscuity and propensity to target-independent assay interference, rather than modelling the biochemical concepts underlying target-specific interactions.

To quantify this, we replaced the last layer of the multi-task network with a single-head regression output $\phi' : \mathbb{R}^d \to \mathbb{R}$, keeping ψ fixed, and trained this new predictive layer on hit frequencies across the CHEMBL dataset. As expected, we observe that a linear model fit on top of the pretrained embeddings is able to predict molecular hit frequencies much better than one that is trained on the naive embeddings, summarised by Pearson correlations of 0.43 and 0.28 respectively.

A complimentary approach to testing this hypothesis is to examine the topology of the learned representation space. Rohrer & Baumann (2009) outline the method of refined nearest neighbour analysis, which can be used to describe how evenly two classes of molecules are distributed in chemical space. We adapt this method to quantify how clearly compounds containing well-known assay-interfering molecular substructures are separated from compounds that do not and use the resulting metric to compare the pretrained and naïve embedding spaces.

First, we annotate all compounds from the CHEMBL benchmark containing one or more of the substructures enumerated in the PAINS library (Baell & Holloway, 2010). As a sanity check, we compare the distribution of hit frequencies between these and the remaining compounds and find that flagged molecules have a significantly[4] higher hit rate than molecules without PAINS substructures. Next, we define $G(d_i)$ as the proportion of flagged compounds for which the distance to the nearest other flagged compound is less than d_i and the function $F(d_i)$ as the proportion of PAINS-free compounds for which the distance to the nearest flagged compound is less than d_i. The resulting distributional metric

$$S = \frac{1}{N_d} \sum_{i=1}^{N_d} (F(d_i) - G(d_i))$$

captures the cumulative difference between F and G over a set of distance thresholds $\{d_i\}_{i=1}^{N_d}$. Intuitively, if $F(d_i)$ is consistently lower than $G(d_i)$ over a range of different distance thresholds, i.e. when $S < 0$, molecules that contain PAINS substructures are clustered in the embedding space and are thus more easily distinguishable from molecules that do not. We follow Rohrer & Baumann (2009) and generate the distance threshold sets by creating $N_d = 500$ evenly spaced splits in $[0, 3d_{\text{m}}]$, where d_{m} is the median nearest neighbour distance.

[4] one-sided two-sample K-S test statistic of 0.22, $p < 10^{-300}$

Consistent with the findings from the previous experiment, the representation space of the trained network exhibits a significantly greater separability of PAINS-containing compounds ($S = -0.39$) than the representation space derived from a random initialisation ($S = -0.29$).

While a more thorough investigation of different models in different settings is necessary, these results indicate that multi-task datasets on which significant discriminative performance can be achieved by recognising assay-interfering substructures are able to induce representations that reflect this shortcut. And while these representations may perform well on the benchmark in question, they are inadequate for many downstream applications, as for example a generative model built on them would preferentially generate assay-interfering compounds and not molecules with actual physiological activity.

5. Conclusions and Future Work

Systematic experimental errors that cause assay-independent false positive readouts in biochemical and toxicological high-throughput screens are present in many popular multi-task benchmarks. The resulting statistical dependencies lead to a high correlation between tasks and can compel multi-task and meta-learning algorithms to focus on predicting hit frequencies, leading to representations that are counterproductive to many impactful downstream applications of interest.

The most robust way to address this issue is to make use of experimental information from counter-screens and remove active compounds that are likely assay-interfering or non-selective. Several instances of high-throughput screens with associated validation assays exist (see e.g. Butkiewicz et al. (2013)) and could be compiled into a collection of high-quality data that is used as a clean test set.

However, the vast majority of screening data in repositories such as ChEMBL or PubChem is not linked to assay-specific counter-screens that could be used to filter out potential false positives. To still recognise problematic compounds in these screens, it may be beneficial to examine the associated measurement meta-data, as the MUV dataset shows that even highly empirical rules-of-thumb can already significantly improve the quality of a dataset. Going even further, the large quantities of meta-data in repositories such as PubChem could present an opportunity to systematically develop more sophisticated de-biasing and de-confounding techniques.

Conclusions from these techniques could then in turn be used to process new single- and multi-task datasets and increase the efficacy with which molecular property prediction models can learn from publicly available bioactivity data.

References

Baell, J. B. and Holloway, G. A. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010.

Baell, J. B. and Nissink, J. W. M. Seven year itch: Pan-assay interference compounds (pains) in 2017-utility and limitations. *ACS Chemical Biology*, 13(1):36–44, 2018.

Bohacek, R. S., McMartin, C., and Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal Research Reviews*, 16 (1):3–50, 1996.

Butkiewicz, M., Lowe, E. W., Mueller, R., Mendenhall, J. L., Teixeira, P. L., Weaver, C. D., and Meiler, J. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules*, 18(1):735–756, 2013.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Dahlin, J. L. and Walters, M. A. How to triage pains-full research. *Assay and Drug Development Technologies*, 14 (3):168–174, 2016.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J. The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, 95(1):5–12, 2007.

Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *Journal of Chemical Information and Computer Sciences*, 43(2):374–380, 2003.

Feng, B. Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B. K., and Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *Journal of Medicinal Chemistry*, 50(10):2385–2390, 2007.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. The chembl database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2017.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 2021.

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

Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., IJzerman, A. P., and Van Westen, G. J. Beyond the hype: deep neural networks outperform established methods using a chembl bioactivity benchmark set. *Journal of Cheminformatics*, 9(1):1–14, 2017.

M Nissink, J. W. and Blackburn, S. Quantification of frequent-hitter behavior based on historical high-throughput screening data. *Future Medicinal Chemistry*, 6(10):1113–1126, 2014.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science*, 9(24):5441–5451, 2018.

McGovern, S. L., Caselli, E., Grigorieff, N., and Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *Journal of Medicinal Chemistry*, 45(8):1712–1722, 2002.

Nguyen, C. Q., Kreatsoulas, C., and Branson, K. M. Meta-learning gnn initializations for low-resource molecular property prediction. *arXiv preprint arXiv:2003.05996*, 2020.

Polishchuk, P. G., Madzhidov, T. I., and Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of Computer-Aided Molecular Design*, 27(8):675–679, 2013.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Rishton, G. M. Reactive compounds and in vitro false positives in hts. *Drug Discovery Today*, 2(9):382–384, 1997.

Rohrer, S. G. and Baumann, K. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of Chemical Information and Modeling*, 49(2):169–184, 2009.

Ryan, A. J., Gray, N. M., Lowe, P. N., and Chung, C.-w. Effect of detergent on "promiscuous" inhibitors. *Journal of Medicinal Chemistry*, 46(16):3448–3451, 2003.

Schorpp, K., Rothenaigner, I., Salmina, E., Reinshagen, J., Low, T., Brenke, J. K., Gopalakrishnan, J., Tetko, I. V., Gul, S., and Hadian, K. Identification of small-molecule frequent hitters from alphascreen high-throughput screens. *Journal of Biomolecular Screening*, 19(5):715–726, 2014.

Stanley, M., Bronskill, J. F., Maziarz, K., Misztela, H., Lanini, J., Segler, M., Schneider, N., and Brockschmidt, M. Fs-mol: A few-shot learning dataset of molecules. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 843–852, 2017.

Thorne, N., Auld, D. S., and Inglese, J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Current Opinion in Chemical Biology*, 14(3):315–324, 2010.

Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. *ICLR (Poster)*, 2(3):4, 2019.

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

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.