FS-TOX: A FEW-SHOT BENCHMARK FOR TOXICITY PREDICTION USING CHEMICAL TRANSFORMERS

Seth Howes
School of Public Health
Imperial College London
London
ssh22@imperial.ac.uk

ABSTRACT

Datasets for drug toxicity assays typically contain only a few hundred records due to the high cost of experimental data collection. As such, drug toxicity prediction tasks are amenable to methods that learn quickly on small prediction tasks, termed few-shot learning methods. In the past year, chemical transformers have emerged as effective tools for few-shot molecular property prediction. However, it is unclear how well they perform at predicting small molecule toxicity. In this study, we introduce a few-shot benchmarking method for small molecule toxicity prediction, and assess both traditional and state-of-the-art chemical transformers. Furthermore, our benchmark allows for model evaluation in both single-task and multi-task approaches. We find that chemical transformers show improved learning with increasing training set sizes up to 128 samples, after which performance plateaus. Our examination of state-of-the-art techniques shows that chemical transformers slightly underperform when compared to logistic regression trained on ECFP4 fingerprints (median Δ-AUPRC difference -0.021). Notably, there is no significant performance difference between fine-tuned and non-fine-tuned chemical transformer models in predicting small molecule toxicity. Our results demonstrate that despite prominent successes of transformers for a wide variety of other NLP tasks, early chemical transformers currently do not outcompete more traditional approaches. We foresee that future iterations of larger chemical transformers, backed by more comprehensive pre-training, will soon surpass traditional prediction methods. Furthermore, refining fine-tuning strategies, such as model recycling will be particularly important at improving multi-task learning, especially when trained on out-of-distribution tasks.

1 Introduction

Computational methods are becoming increasingly important in the realm of preclinical drug development. This shift can be attributed in part to the "unreasonable effectiveness of deep learning" across an increasing wealth of domains [1]. The emergence of deep learning as a core tool in the drug discovery domain has been primarily realised by the unprecedented advancements in the power and availability of Graphics Processing Units (GPUs) [2]. This has facilitated GPU-enabled drug discovery algorithms for tasks such as molecular docking, protein folding prediction, and quantitative structure-activity relationship (QSAR) prediction [2–4].

Failure to predict drug toxicity in the preclinical stages is a significant reason for attrition through the drug discovery pipeline. Drug toxicity stands as the second greatest cause of attrition through the pipeline, with 30% of drugs lost due to excessive toxicity [5].

The need for better prediction methods for drug toxicity has garnered increasing attention towards computational approaches. A key moment in this shift took place in 2014 through the proposition of the Tox21 Data Challenge by the US National Institutes of Health (NIH) [6]. Researchers were tasked with generating predictions for 12 cellular toxicity assays for a subset of a 10,000 chemical library. The resultant best performance for this model took the form of a

vanilla fully-connected neural network. This winning model architecture, outperforming the more standard machine learning methods of the time, acted as a harbinger of the increasing importance of deep-learning methods for *in silico* drug toxicity prediction.

However, dataset sizes in the domain of drug discovery are typically much smaller than the close to 1 million record dataset used to train models for the Tox21 Data Challenge. Dataset sizes for drug toxicity assays usually fall in the range of tens to the low hundreds of observations. With the cost of bringing a single drug to market estimated to be between $1-2 billion dollars[7], there is an economic constraint on the number of drug candidates that can be screened using *in vitro* and *in vivo* toxicity assays. The small size of the resultant datasets means that they are less suitable for training traditional machine learning models, which are typically trained on datasets with hundreds-of-thousands to millions of observations.

Despite these limitations, a set of methods that can rapidly learn using very limited datasets, known as few-shot learning methods, have been developed over the past several years. For example, model-agnostic meta-learning (MAML)[8] is an algorithm that facilitates the fast adaptation of any model trained with gradient descent to new tasks using only a few training examples. An additional few-shot learning method, the prototypical network, has previously achieved state-of-the-art performance on multiple few-shot image classification benchmarks[9].

Over the past 12 months, transformer models have demonstrated remarkable capabilities on few-shot and zero-shot learning natural language tasks. These models are most commonly based upon the transformer architecture, which was pioneered by Google Brain in 2017[10]. This model uses a concept known as 'attention' to capture dependencies between input data regardless of their position in an input sequence. The most well known examples of the transformer architecture are the generative pre-trained transformer (GPT) class of models[11]. Models belonging to this class, such as Open AI's GPT-4, have demonstrated state-of-the-art performance across many benchmarks in the domain of natural language processing (NLP)[12].

Over the past year, transformer models have emerged that have been adapted specifically for tasks in the chemical domain. These models are typically trained using self-supervised learning on common string representations of small molecules, including SMILES[13], and SELFIES[14], as inputs. A prominent chemical language model is ChemGPT, which was pre-trained on a self-supervised causal language modelling task consisting of 10 million SMILES from the PubChem repository[15]. In this pre-training setup, a random token from the input sequence is masked. The model must then predict the identity of this token by conditioning on the remaining tokens in the input sequence. As part of the training setup, the model is also tasked with predicting a classification token, which is a vector embedding that represents the entirety of the small molecule in an abstract metric space[16]. This embedding is commonly used as the input to a final layer for classification tasks.

Despite the recent development of these models, it is unclear how well they perform on classification tasks that contain only a small number of samples. Furthermore, there is little evidence to indicate how their performance varies when fine-tuned on training sets of different sizes. Do these models learn quickly, if at all, when tasked with predicting molecular properties on only a few training examples?

Assessing the performance of few-shot models requires a specific training and evaluation procedure. Firstly tasks are split into a disjoint set of training and test tasks[17]. Samples making up each of these individual tasks are then split into a set for training, with the remaining samples being used for evaluation, termed support and query sets. An example of such a few-shot benchmark is FS-Mol. This was released in 2021 for evaluating the performance of models for QSAR classification. It contained just under 500,000 different observations grouped into different classification tasks which contained on average only a few hundred records each. However, it is unclear whether their findings translate to the toxicity prediction domain.

As is evidenced by the open-source nature of the aforementioned models and benchmarks, drug discovery is increasingly becoming an open endeavour. There has been a greater availability of open-source software tools for multiple stages of the drug discovery process. Examples of such tooling include DiffDock[3] (a library for molecular docking), and ColabFold[4] (a library for protein folding). There is also a growing repository of relevant publicly-accessible datasets, such as the Therapeutic Data Commons[18], and the EPA's ToxCast[19].

Benchmarks are an important tenet for validating and comparing the performance of computational drug discovery tools in a reproducible manner. As is common with much of science, drug discovery research is plagued with poor reproducibility[20]. Despite the necessity for adequate benchmarks for the toxicity prediction domain, existing prediction benchmarks are lacking. Current benchmarks are limited by their broad inclusion of different molecular prediction tasks, and concomitant insufficient coverage of toxicity prediction tasks. Toxcast, a component of the MoleculeNet benchmark[21], incorporates assays that do not directly assess cell viability, and therefore does not provide a thorough assessment of toxicity prediction performance. Similarly, another constituent dataset of MoleculeNet, ClinTox, consists of only two prediction targets. One of the outcome targets is whether or not the drug was FDA approved or not, and

therefore does not capture whether a drug is toxic. Another molecular property prediction benchmark we discussed previously, FS-Mol, is not explicitly designed for assessing toxicity prediction models; rather, it includes tasks relating to the prediction of different QSAR metrics. Finally, these datasets provide little coverage of drug toxicity prediction for human subjects.

We therefore seek to address the following three main research questions: 1) how do different training set sizes affect the performance of different toxicity prediction models, 2) do chemical transformers outperform traditional models, and 3) do multi-task approaches outperform single-task approaches for toxicity prediction?

In order to address these questions, we aim to create a toxicity prediction benchmark, which we term FS-Tox. This includes datasets spanning biological models with varying degrees of complexity: *in vitro* cell viability assays, *in vivo* toxicity assays, and a human toxicity dataset. Using this benchmark, we will then compare traditional machine learning prediction methods with state-of-the-art chemical transformers. We also aim to change the training set sizes for given tasks, to determine this affects predictive performance. Finally, we aim to test both traditional and state-of-the-art models using both single-task, and multi-task approaches.

2 Methods

2.1 Benchmark construction

2.1.1 Dataset inclusion

We began creation of our few-shot learning benchmark by sourcing publicly available datasets reporting results from real-world toxicity assays. We aimed to include datasets assessing toxicity across the hierarchy of biological models: *in vitro* cell-based assays, *in vivo* assays, and human case studies. We only include datasets containing at least one small molecule with an associated toxicity outcome measure. We included assays that defined toxicity in different ways, such as LC_{50}, and LD_{50}.

2.1.2 Dataset pre-processing

For each dataset, we carried out several pre-processing steps. This was to convert our datasets into a standardised format, consisting of columns with no missing data, standard string representations of small molecules, experimental descriptors, and toxicity assay outcome measures. The pre-processing steps are detailed as follows:

- Columns unrelated to drug identifier, toxicity outcome, or details of experimental setup were removed.
- Rows with missing data were removed.
- Different drug names were standardised to a canonical SMILES format using the PubChem API and the Python RDKit open-source chemoinformatics library.
- Rows for which we were unable to map the identifier to canonical SMILES were removed.
- Canonical SMILES were converted to SELFIES representations.

2.1.3 Assay derivation

Following dataset pre-processing and standardisation, we grouped records belonging to single experiments into a set of assays. Here, we define assays as a grouping of records with identical values for variables crucial to defining an individual assay setup. We grouped assays according to the following experimental variables: animal / cell model used, toxicity outcome measure, method of drug delivery, concentration / amount of drug delivered, experimental identifiers, and references to source material. We grouped records into assays using as many of the aforementioned variables as were present in each dataset.

In order to ensure that our assays were sufficiently large to both train and evaluate models on, we removed assays containing fewer than 32 samples.

2.1.4 Feature creation

Each assay was assigned to a meta-set depending upon the dataset that it was derived from. Datasets containing the results of assays of *in vitro* experiments were assigned to the *in vitro* meta-set ($\mathcal{D}_{\textit{in vitro}}$), with those assessing any live animal model assigned to the *in vivo* meta-set ($\mathcal{D}_{\textit{in vivo}}$), and those with human-level data to the human meta-set ($\mathcal{D}_{human}$).

2.1.5 Outcome standardisation

We standardised drug names to a single, common format: canonical SMILES. Our included datasets contained different chemical identifiers, such as International Chemical Identifiers (InChIs), common chemical names, non-canonical SMILES, and SELFIES. For this standardisation process, we used a combination of the PubChem API and the open-source chemoinformatics library, RDKit[22].

For each canonical SMILES, we generated ECFP4 fingerprints using the RDKit library. ECFP4 fingerprints are deterministic, binary string representations of small molecules that capture the different substructures belonging to a given chemical23. For each small molecule, we derived multiple ECFP4 lengths - 256, 512, 1024, and 2048.

We also generated embeddings using a chemical transformer model, ChemGPT, which belongs to the GPT family of model architectures. This model comes in three sizes, with parameter counts of 4.7 million, 19 million, and 1.2 billion, and generates embeddings with an output length of 128. Due to resource constraints, we generated embeddings using only the model with the smallest parameter count.

For each assay, we transformed the toxicity outcome to a binary outcome. We did this for two reasons. Firstly, within the range of datasets we sourced, there were a variety of methods employed to measure and evaluate toxicity. For example, some assays measured cellular viability using fluorescence intensity when exposed to a given concentration of different small molecules. Other assays measured LC_{50} values, which measure the concentration of a drug that will kill half of a given exposed population. These varied outcomes would have hindered the ability to carry out multi-task learning across different assays. Secondly, most toxicity outcomes we encountered were presented as continuous measures. Given that regression on these continuous toxicity outcomes is widely recognized as extremely challenging, we decided to incorporate only classification tasks within our benchmark[23].

We converted continuous to binary outcome values by firstly taking the negative log of the concentration if the toxicity outcome was reported as a concentration. This ensured more toxic compounds would have larger values. Furthermore, we excluded those assays for which the negative log range of concentrations was fewer than three orders of magnitude. This was to ensure that the range of toxicity values was sufficiently great to enable strong learning by our classification models. We then assigned each record a value of 1 if the toxicity value was above the median toxicity value for that assay, and 0 if it was below the median.

Additionally, to prevent heavily imbalanced tasks, we removed assays where the active ratio (number of toxic compounds / number of non-toxic compounds) was not in the range of 0.3 to 0.7. Assays falling outside of this active ratio range typically have a significant proportion of their samples as the median value, meaning all of these values are assigned to a single class. This therefore makes classification impractical.

2.1.6 Task derivation

We derived a set of classification tasks for each assay. Every sample for a given task was assigned to be part of either a support or query set. In few-shot learning setups, support sets are a subset of records from each task used to train few-shot models. We assigned the remaining records for each task to the query set, which is used to evaluate models trained on the support set.

In order to assess how different support set sizes impact predictive performance, we created multiple tasks with different support set sizes for each assay. We assessed the following support set sizes for each assay: 16, 32, 64, 128, 256. If there were not enough records available to create a specified support set size, or if the creation of a task resulted in a query set size of fewer than 16 records, we passed over the creation of this task. As a final step, we carried out stratified 3-fold random splitting for each support set size, in order to dampen the impact of invalid prediction accuracies resulting from lucky splits.

2.2 Toxicity prediction

2.2.1 Single-task approach

We applied four distinct single-task methods: logistic regression, random forest, gradient-boosted decision trees (XGBoost), and ChemGPT. Specifically for XGBoost, we incorporated hyperparameter tuning as a segment of our training process. The hyperparameter search was conducted on the first 10 tasks of a given run. We identified optimal hyperparameters using 4-fold cross-validation, alongside a randomised hyperparameter search. We carried out this search to identify modal *eta*, *gamma*, *max_depth*, and *gamma* parameters. The modal hyperparameter values resulting from these first 10 runs were used as the model hyperparameters for the remaining tasks.

Table 1: FS-Tox statistics

	in-vitro			in-vivo		human
Dataset	CancerRxGene	ToxCast	PRISM	Acute oral toxicity	ToxVal	MEIC
# tasks	967	125	482	1	38	3
# small molecules	227	3429	657	7342	1149	48
Median # compounds per task	227.0	113.0	604.5	7385.0	54.5	48.0
Source	Wellcome Sanger / Mass General	US EPA	Broad Institute	Zhu 2009	US EPA	Ekwall 1998
Raw values available?	Yes	No	Yes	Yes	Yes	Yes

2.2.2 Multi-task approach

We fine-tuned a single 4.7 million parameter ChemGPT model on up to 200 randomly-selected tasks belonging to $\mathcal{D}_{\textit{in vitro}}$, $\mathcal{D}_{\textit{in vivo}}$, which we refer to as auxiliary fine-tuning. The purpose of this approach was to mimic the drug discovery process, where one uses the results of testing on a simpler biological model to infer a drug's toxicity on a more complex, and more expensive model. We then combined the auxiliary fine-tuned models into a single model for each meta-set by taking the mean of each parameter value across the models.

We then carried out an additional fine-tuning step for each of our auxiliary fine-tuned models. Each auxiliary fine-tuned model corresponding to a meta-set was fine-tuned on tasks belonging to the meta-sets of higher complexity models. For example, the auxiliary fine-tuned model for $\mathcal{D}_{\textit{in vitro}}$ was fine-tuned on the support sets for tasks in $\mathcal{D}_{\textit{in vivo}}$, and $\mathcal{D}_{human}$. We refer to this process as target fine-tuning.

We evaluated the predictive performance of the target fine-tuned models on the associated query set for each task.

2.3 Outcome assessment

We used the Δ-AUPRC score to assess the predictive performance of our models across the previously detailed experimental setups. We chose Δ-AUPRC as it is a more valid outcome metric for cases where there is data imbalance, which was common across the tasks in our benchmark. AUPRC indicates the area under the precision-recall curve. For this metric, random classification by a model should equal the proportion of positive cases present in the dataset. As such the value of Δ-AUPRC shows the improvement in predictive performance of a model compared to random guessing.

3 Results

3.1 FS-Tox statistics

We included 6 datasets in our final benchmark. Three of these belonged to $\mathcal{D}_{\textit{in vitro}}$, two to $\mathcal{D}_{\textit{in vivo}}$ and the remaining to $\mathcal{D}_{human}$. All were sourced from publicly available repositories. The identity of these datasets as well as summary statistics are included in Table 1 .

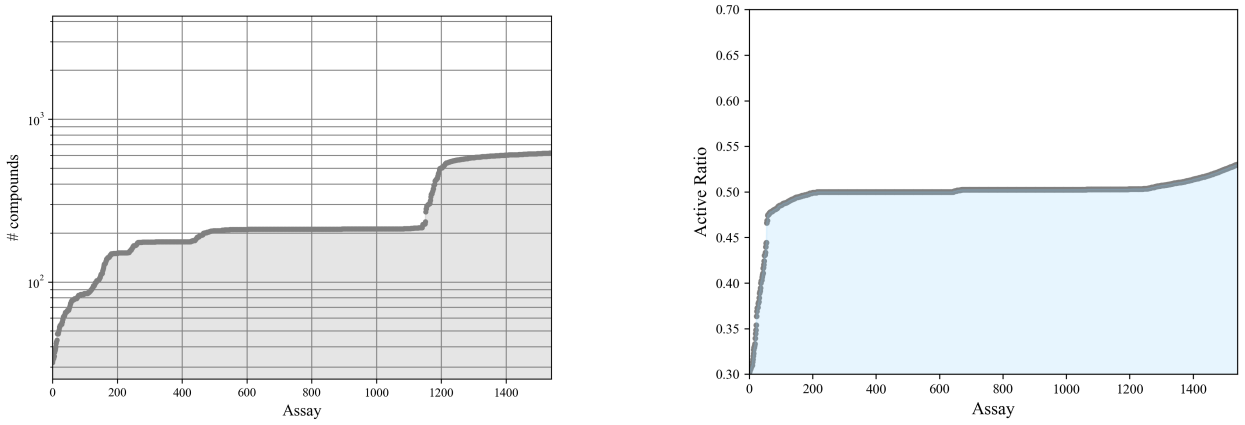
We considered 4 additional datasets in the construction of our benchmark that were ultimately excluded. These datasets did not pass our pre-processing criteria due to the active ratio of the assays being outside the range of 0.3-0.7. Three of these datasets were from the MoleculeNet, an existing drug toxicity benchmark (Tox21, BBBP, and ClinTox), and the final dataset from the National Cancer Institute (NCI60).

Across all datasets in our benchmark, the median number of compounds per assay was 113 (Figure 1a). Furthermore, the median value for active ratio across assays was 0.5 (Figure 1b).

There appears to be no clearly discernible difference between the small molecule structures of toxic and non-toxic compounds according to PCA plots (Appendix 1). This was true when modelling the structures as either ECFP4 fingerprints or as ChemGPT embeddings. Furthermore, there were no pairings of datasets for which small molecules were more structurally similar than other dataset pairings (Appendix 2).

3.2 Benchmarking single-task models on FS-Tox

There is a clear improvement in the predictive performance of single-task models when trained on larger support set sizes (Figure 2). The greatest improvement in predictive performance across all datasets occurred when moving from a support set size of 16, to a support set size of 32, with a mean Δ-AUPRC improvement across datasets of 0.031. The rate of Δ-AUPRC change diminishes beyond a support set size of 128, with a mean change of only 0.003 (Appendix 3).



(a) (b)

Figure 1: **FS-Tox statistics.** a) The median number of compounds per assay across each dataset. b) The active ratio of each assay across datasets.

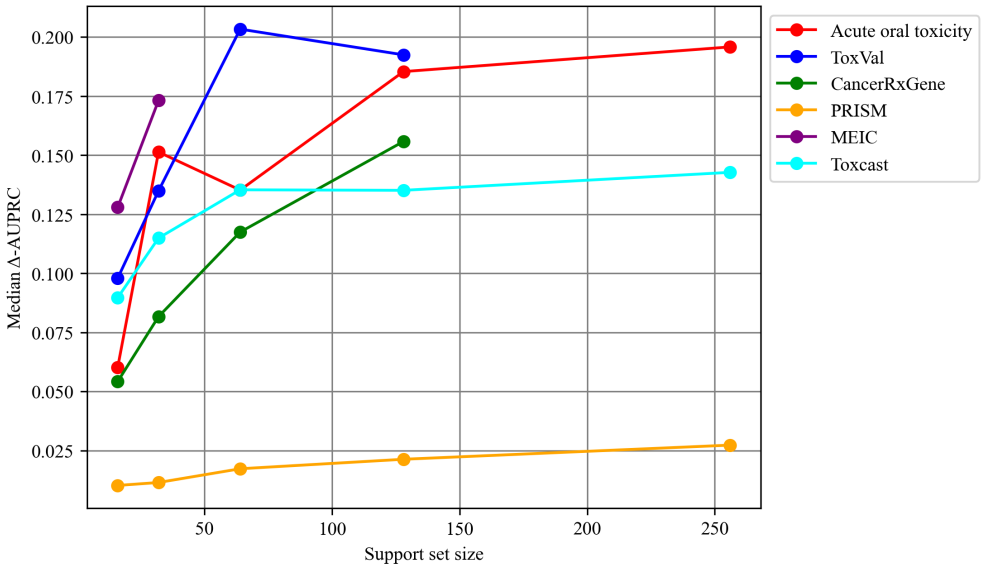


Figure 2: **Single-task learning on ECFP4 fingerprints**. Δ-AUPRC scores for models trained on different support set sizes. The experimental setup was logistic regression trained on ECFP4 fingerprints. Points are missing for datasets where the task size was too small to generate the relevant support set sizes.

However, the variation of Δ-AUPRC values for tasks in each constituent dataset was heavily variable. (Appendix 4). The PRISM dataset showed the smallest variation in scores across all tasks (S.D. 0.03), whereas ToxVal had the largest variation (S.D. 0.12).

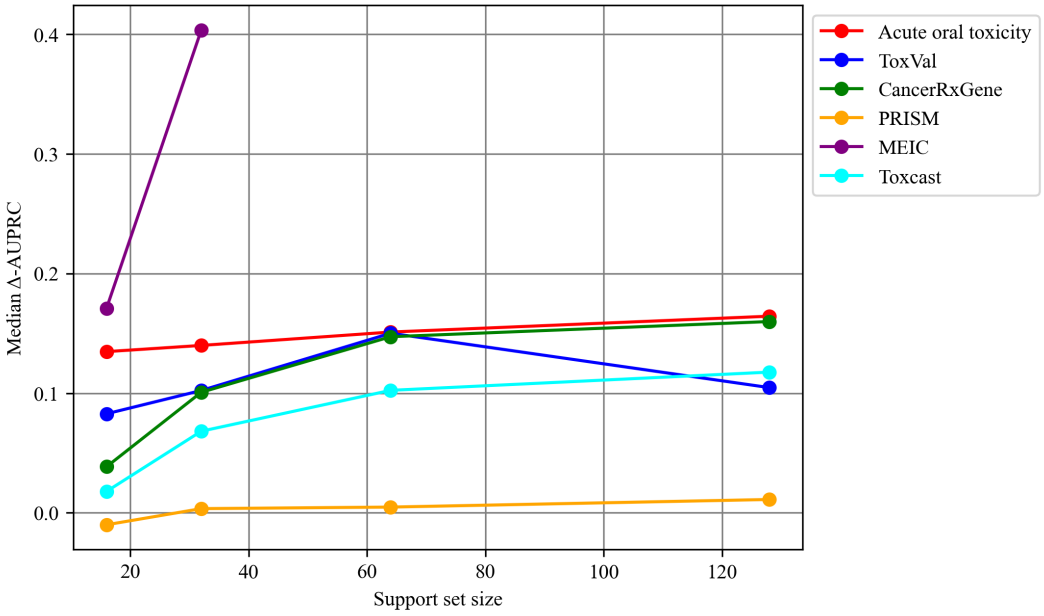
Logistic regression and random-forest were the best performing traditional single-task models, with XGBoost consistently having worse predictive performance than the other two models (Appendix 5).

Increasing the size of the ECFP4-fingerprints used to train the single-task models made little difference to the overall predictive performance of the trained models (Appendix 6). The median Δ-AUPRC score difference between ECFP4 sizes of 256 and 2048 was 0.011.

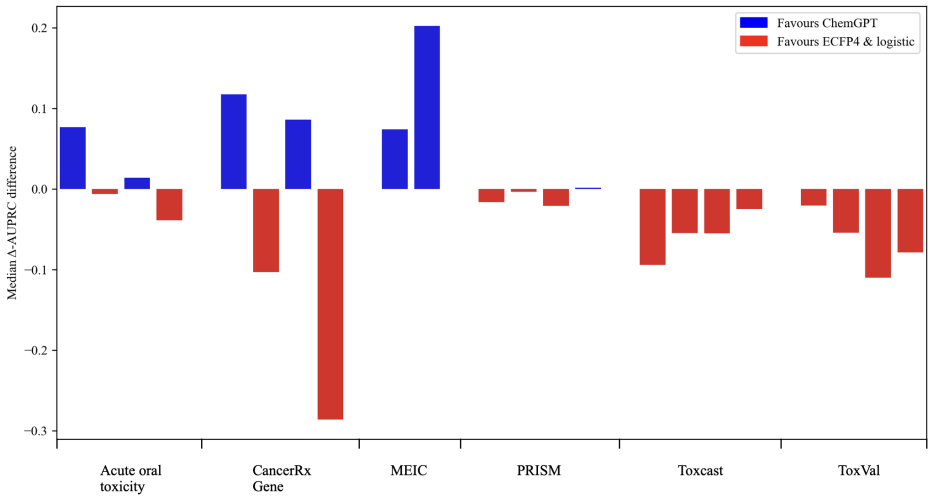
We found a slight indication that predictive performance correlated with how similar small molecules were between datasets (Appendix 7). The Tanimoto similarity between datasets, and the Δ-AUPRC scores showed a small negative correlation (-0.12).

3.3 Benchmarking fine-tuned models on FS-Tox

The predictive performance improved for the single-task fine-tuned ChemGPT model for increasing support set sizes (Figure 3a). This improvement was not consistent across datasets, with MEIC showing a Δ-AUPRC increase of 0.18 to 0.41 for support set sizes of 16 and 32 respectively. However, the acute oral toxicity dataset showed little improvement with increasing support set size. The PRISM dataset exhibited no learning regardless of the support set size used to fine-tune ChemGPT.



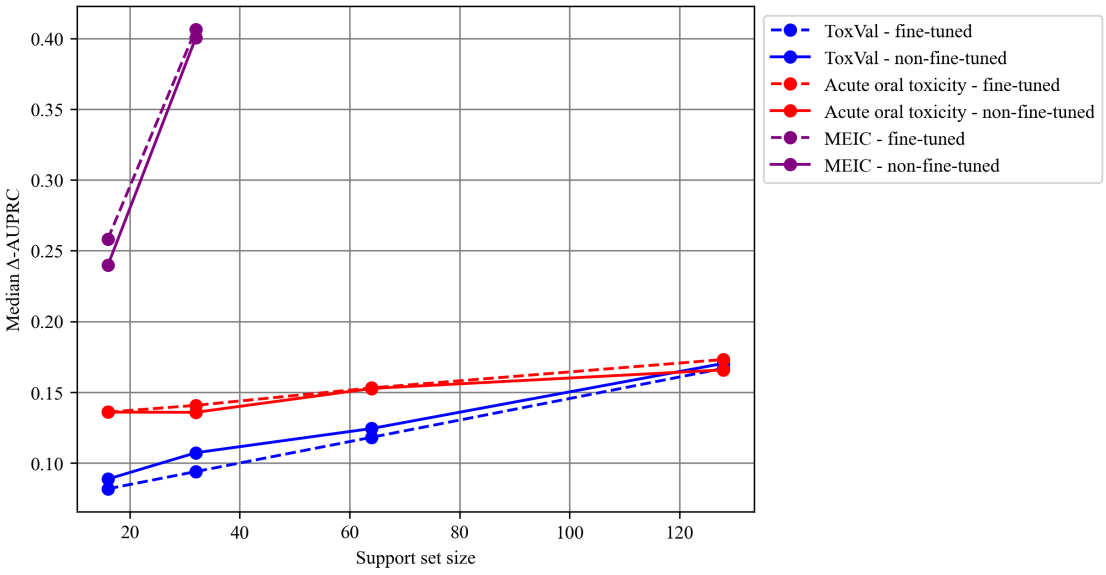
(a)



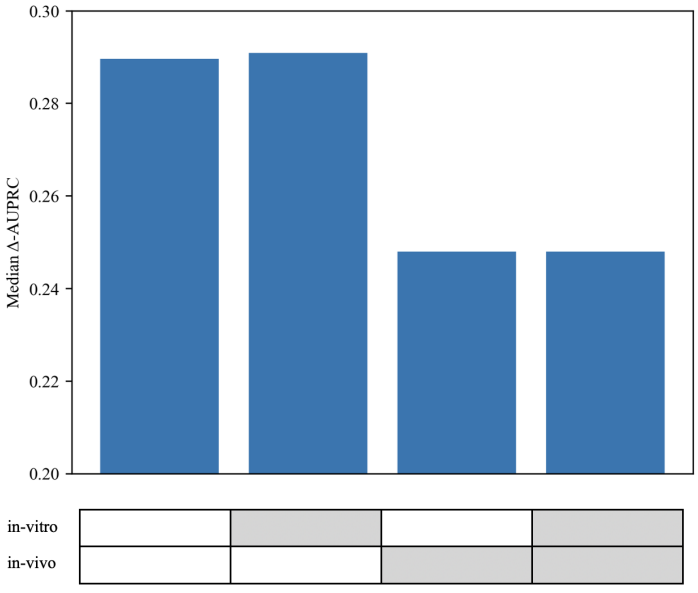
(b)

Figure 3: **ChemGPT single-task finetuning**. a) Δ-AUPRC scores for models evaluated on the query set for support set sizes of 16, 32, 64, and 128 samples. b) Mean difference in Δ-AUPRC scores for ChemGPT and logistic regression models trained on ECFP4 fingerprints of length 1024. Each bar represents a different support set size for a given dataset.

When directly comparing the predictive performance of a logistic regression trained on ECFP4 fingerprints, with ChemGPT models for the single-task approach, we found that logistic regression consistently outperformed ChemGPT (Figure 3b). This was the case for the majority of datasets, excluding the MEIC dataset, where ChemGPT had significantly better performance. Logistic regression outperformed ChemGPT across all support set sizes that we included in our benchmarking procedure.



(a)



(b)

Figure 4: **ChemGPT multi-task learning**. a) Predictive performance for models auxiliary fine-tuned on $\mathcal{D}_{in\ vitro}$ data vs. models with no auxiliary fine-tuning. b) Performance of models on predicting human toxicity with auxiliary fine-tuning on $\mathcal{D}_{in\ vitro}$ and $\mathcal{D}_{in\ vivo}$ meta-sets.

We found no difference in the predictive performance of ChemGPT models that underwent auxiliary fine-tuning on tasks from $\mathcal{D}_{in\ vitro}$, compared to those that did not undergo auxiliary fine-tuning (Δ-AUPRC 0.000). Performance

was no different between the auxiliary fine-tuned models compared to non-auxiliary fine-tuned models for all of the support set sizes that we assessed (Figure 4a).

We found that auxiliary fine-tuning on $\mathcal{D}_{\textit{in vitro}}$ tasks followed by target fine-tuning on $\mathcal{D}_{human}$ tasks did not improve predictive performance (Δ-AUPRC 0.005) compared to models with no auxiliary fine-tuning (Figure 4b). However, the model auxiliary fine-tuned on $\mathcal{D}_{\textit{in vitro}}$, the model fine-tuned on $\mathcal{D}_{\textit{in vivo}}$ and, the model fine-tuned on both $\mathcal{D}_{\textit{in vitro}}$ and $\mathcal{D}_{\textit{in vivo}}$ had a poorer performance relative to the model with no auxiliary fine-tuning (Figure 4b).

4 Discussion

We have created a benchmark for few-shot small molecule toxicity prediction to assess traditional prediction models, as well as state-of-the-art chemical transformers, in both single-task and multi-task approaches. We include datasets for assays of drug toxicity across a range of model complexities. We find that the predictive performance of single-task models increases most markedly when moving from support set sizes of 16 to 128, before performance plateaus. We show that chemical transformers do not outperform more traditional approaches to toxicity prediction. We also show that prediction performance of chemical transformer models has the same performance for all support set sizes for both single-task and multi-task approaches.

A major strength of our benchmarking procedure relative to existing benchmarks is the inclusion of toxicity prediction tasks that cover models of varying biological complexities. Importantly, we include tasks related to the assessment of toxicity in human subjects. For this, we included an underutilised resource in this domain, the MEIC dataset[24]. This dataset compiled the results from different case studies of the toxicity of 50 different common small molecules on humans. For each case study, they extracted small molecule blood concentrations from post-mortem reports for individuals who died from exposure to a given small molecule. Following this, they converted their toxicity measures into a standard outcome (LC_{50}). The inclusion of this dataset provides a ground-truth set of tasks that can be used to directly assess the transferability of learning from models trained on lower level biological models. Furthermore, we included datasets that assess toxicity of various *in vivo* models. The ToxVal dataset is an aggregation of various chemical toxicity studies submitted to the EPA by corporations that want to use a novel unlicensed chemical in a commercial setting. Acute oral toxicity is a dataset combining the results of different studies assessing toxicity in rats following administration of drugs via the oral route[25]. The inclusion of *in vivo* data is something that is lacking from other benchmarks, such as MoleculeNet. This consists of only one human-level toxicity assay, which is limited by its indirect assessment of toxicity in these subjects. Therefore, the inclusion of toxicity assays for a range of biological model complexities permits assessment of transfer learning to higher level models.

We show that learning for single-task models improves with increasing support set size, before performance plateaus at a support set size of 128. This is consistent with the findings from FS-Mol, a recently-proposed benchmark and testing procedure for assessing the performance of QSAR prediction on string representations of small molecules. They showed that model performance begins to plateau after a support set size exceeds 128 samples for the different models they assessed, including graph neural networks, prototypical networks, and random forests[23]. Additionally, we found that logistic regression and random forest models outperformed gradient boosting across all constituent datasets used in our benchmark. This finding is inconsistent with many current approaches to binary classification tasks. Gradient boosted decision trees are typically used as part of leading submissions for classification-based Kaggle competitions. It is unlikely that the poor performance of XGBoost relative to logistic regression and random forest models is due to overfitting, as the model still underperformed even with heavy regularisation. However, this finding is not without precedent. A systematic review by Christodoulou et al. 2019 showed consistent performance of logistic regression and ML methods for binary classification models in the clinical domain, with AUROC scores not differing between the two approaches[26]. As such, we find that simpler models may marginally outperform gradient-boosted methods for few-shot classification tasks.

We found that models typically experienced greater learning on tasks belonging to the *in vivo* and human datasets. We suspect this is because our *in vitro* assays were more commonly measured as part of a high-throughput setup. These setups typically cover compounds that fall within a narrow range of toxicity values. However, clinical assays typically include compounds covering a wider range of toxicity values, ranging from very well tolerated chemicals, to compounds with known extreme toxicity. This was evidenced in our benchmark, where we found the median toxicity values in the ToxVal dataset spanned 10 orders of magnitude, whereas the median toxicity range for the *in vitro* assays was 4. There are likely to be more distinctive molecular substructures with high toxicity present in small molecules from the ToxVal dataset, relative to the *in vivo* datasets. Therefore, predictive performance is likely to be better on the *in vivo* tasks as it is easier to distinguish between toxic and non-toxic compounds for these tasks relative to *in vitro* tasks.

There appeared to be little to no evidence of learning taking place on the PRISM dataset. PRISM assessed cell viability by measuring fluorescence intensity using a molecular barcoding method following exposure to different small

molecules[27]. This method was pioneered by a group at the Broad Institute, with PRISM containing the results of their initial testing of this novel method. There is a lack of benchmarking of this procedure relative to other, more standard assessments of toxicity. A paper for this method has yet to be published, and therefore we lack an understanding of the quality control methods used, and other methodological details. Furthermore, this study looked at previously licensed drugs to see whether they may be suitable to be repurposed as chemotherapeutic agents. As such, the toxicity range of these small molecules is likely to be very small, making our downstream classification tasks more challenging.

We found fine-tuning chemical transformers in a single-task setup results in strong learning. Existing transformer architectures, such as BERT and GPT, have shown state-of-the-art performance on text classification tasks[12,16]. There is also evidence that chemical transformers may be able to achieve strong performance in molecular prediction tasks. ChemBERTa-2 is a novel chemical transformer model that was released in late 2022 as a follow on to its predecessor ChemBERTa. It has been applied to small molecule property benchmarks and has outperformed prevailing approaches for molecular property prediction tasks, such as ChemProp, a graph convolutional neural network for learning graph-based representations of small molecules[28]. However, our study provides the first evidence that chemical transformers can predict small molecule toxicity. ChemBERTa-2 achieves state-of-the-art performance on ClinTox, one of the benchmark tasks from MoleculeNet[21]. However, this does not provide strong evidence that this model performs well at toxicity prediction, as we do not believe ClinTox to be an adequately valid indicator of drug toxicity. This dataset was excluded from our current benchmark as it had an active ratio outside of our inclusion range of 0.3 to 0.7. Additionally, it is unclear how well ChemBERTa-2 truly performed on this task, as the scoring metric used was the AUROC metric. This metric is not robust against heavily imbalanced datasets, such as ClinTox. Unfortunately, we were not able to assess the performance of this model ourselves, as it was not available on HuggingFace at the time we ran our experiments.

We found that training ChemGPT in a multi-task approach did not improve its predictive performance on *in vivo* and human toxicity prediction tasks. It is worth noting here that on the other prominent few shot property prediction dataset, FS-Mol, the only model that consistently outperformed the single-task random forest approach across all support set sizes was the prototypical network[23]. Other multi-task approaches, such as multi-task GNNs, did not outperform single-task random forests for molecular property prediction.

The lack of improvement in performance of models trained in a multi-task relative to a single-task setup might stem from our method of fusing models. After auxiliary fine-tuning on *in vitro* tasks, we merged each model by averaging the parameters to produce a single model for this meta-set[29]. Alternative ensemble techniques, such as model distillation[30], or a mixture of experts approach[31], have been shown to be more effective at integrating information from multiple models trained on different tasks. These approaches might therefore have offered better results.

Other fine-tuning methods might have enhanced our predictive performance. One such method is diverse weight averaging, wherein pre-trained models are fine-tuned on tasks from the target distribution prior to evaluation[32]. For us, this could have involved auxiliary fine-tuning on select *in vivo* tasks, merging the resulting models into a single model, which we would fine-tune on *in vivo* target tasks. Another method is model recycling, in which pre-trained models are first fine-tuned on out-of-distribution tasks, then on tasks from the target distribution. Finally, these models are then fused into a single model for target task prediction[33]. In our context, this might mean first fine-tuning on the *in vitro* dataset, then fine-tuning on *in vivo* tasks, and finally fusing these models into a single model for prediction on *in vivo* target tasks. Investigating varied model-fusing techniques and fine-tuning strategies presents a promising direction for future research.

We found that auxiliary fine-tuning on $\mathcal{D}_{\textit{in vivo}}$ tasks degrades predictive performance when these models are target fine-tuned and evaluated on $\mathcal{D}_{human}$ tasks. This is in contrast to auxiliary fine-tuning on $\mathcal{D}_{\textit{in vitro}}$ tasks, which generates models with the same predictive performance as target fine-tuning models without auxiliary fine-tuning. We suspect this is due to two main reasons. Firstly, $\mathcal{D}_{\textit{in vivo}}$ assays were of poorer quality than $\mathcal{D}_{\textit{in vitro}}$ assays. Here, 'poorer quality' means that samples from the assays are unlikely to have been generated from the same experimental setup. The ToxVal dataset is an amalgamation of many different datasets from the literature. Whilst we tried to ensure that each of the assays came from the same experimental procedure, it was not possible to ensure that this was definitely the case. The strongest indicator that two records were generated as part of the same experimental procedure was grouping by the name of the original source for each sample. However, it was not possible to tell whether those records with the same original source were definitely created as part of the same experimental setup, as some of the sources were themselves collections of data from different experiments. The second reason for poor performance may be that the animals used for the $\mathcal{D}_{\textit{in vivo}}$ tasks were poorer models of human toxicity than the $\mathcal{D}_{\textit{in vitro}}$ models. The majority of the species used in the ToxVal dataset, which contributes 99% of $\mathcal{D}_{\textit{in vivo}}$ tasks, were fish. However, all of the$\mathcal{D}_{\textit{in vitro}}$ tasks were based upon assays carried out using mammalian cells. Auxiliary fine-tuning on assays using species phylogenetically further from humans may result in poorer model performance. Kumar et al. (2022) showed that fine-tuning models on tasks with increasingly large distributional shifts from the target tasks increasingly degrades the quality of representational learning by language transformers[34]. As such, poorer quality assays, as well as a greater

phylogenetic difference between model species may have resulted in poorer transfer learning for models fine-tuned on $\mathcal{D}_{in\ vivo}$ tasks relative to $\mathcal{D}_{in\ vitro}$ tasks.

A limitation of the construction of our assays was the encoding of our target variable as a binary outcome. All of the constituent datasets making up our benchmarks encoded the toxicity outcome as a floating point number. However, treating QSAR outcomes as a regression target is known to be an extremely difficult problem. Reasons for this include measurement noise, a narrow measurement range, low or high values that exceed the range of the assay and are therefore encoded instead as a boundary constant[23]. We appreciate that the construction of our benchmark as a classification task carries its own set of problems; primarily, there is a loss of information and granularity by not including the full range of different toxicity values. Also, values surrounding the threshold between toxic and non-toxic values carry similar weight to those values far from this threshold. In light of these considerations, future research could explore methods that incorporate the full spectrum of toxicity values while accounting for the inherent challenges presented in regression-based approaches. An appropriate method may be quantile regression, which allows this to be framed as both a multi-class classification or regression based problem.

A compelling avenue for exploration is the comparison of emerging models within the rapidly evolving realm of chemical transformers for multi-task learning. In May 2023, Liu et al. released a mixed-modality chemical transformer that was pre-trained not only on SMILES, but also on the surrounding scientific text of PubMed articles from where the SMILES were sourced[35]. At the time of writing this article, the model currently has the best performance for most small molecule prediction tasks from the MoleculeNet benchmark.Before this breakthrough, Uni-Mol was the gold standard for molecular property prediction tasks, including those in the MoleculeNet benchmark[36]. This architecture of this model is an SE-(3) equivariant transformer, which can take in and represent the 3D confirmation of different small molecules. However, these models have not been tested specifically on toxicity prediction tasks. It is also unclear how the performance of these models varies with different training set sizes. We believe future research should aim to compare these models using a benchmark designed to test multi-task training, such as that detailed in this paper.

In conclusion, we found that chemical transformers show strong-learning when fine-tuned in a single-task setup. However, auxiliary fine-tuning these models prior to target fine-tuning does not improve the predictive performance of these models. In some cases, fine-tuning on poorer quality assays with a greater distributional shift relative to the target task may actually degrade predictive performance relative to models that do not undergo auxiliary fine-tuning. A clear next step to extend this research would be to benchmark the burgeoning set of chemical transformers, as well as test different fine-tuning strategies to see which results in better transfer learning.

5 Acknowledgements

We thank Niklas Rindtorff from LabDAO for offering support as to the construction of the single-task and multi-task analysis pipelines.

References

[1] Terrence J. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038, December 2020. doi: 10.1073/pnas.1907373117.

[2] Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C. Stern, and Artem Cherkasov. The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00463-x.

[3] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking, February 2023.

[4] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: Making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1.

[5] Helen Dowden and Jamie Munro. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*, 18(7):495–496, May 2019. doi: 10.1038/d41573-019-00074-z.

[6] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, 3, 2016. ISSN 2296-665X.

[7] Olivier J. Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853, March 2020. ISSN 1538-3598. doi: 10.1001/jama.2020.1166.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, July 2017.

[9] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning, June 2017.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.

[12] OpenAI. GPT-4 Technical Report, March 2023.

[13] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005.

[14] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, December 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/aba947.

[15] Nathan Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor Coley, and Vijay Gadepally. Neural Scaling of Deep Chemical Models, May 2022.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

[17] Victor Garcia and Joan Bruna. Few-Shot Learning with Graph Neural Networks, February 2018.

[18] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development.

[19] ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology | Chemical Research in Toxicology. https://pubs.acs.org/doi/10.1021/acs.chemrestox.6b00135.

[20] Nalini Schaduangrat, Samuel Lampa, Saw Simeon, Matthew Paul Gleeson, Ola Spjuth, and Chanin Nantasenamat. Towards reproducible computational drug discovery. *Journal of Cheminformatics*, 12(1):9, January 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-0408-x.

[21] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A Benchmark for Molecular Machine Learning, October 2018.

[22] RDKit: Open-source cheminformatics. https://www.rdkit.org.

[23] Megan Stanley, John F. Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-Mol: A Few-Shot Learning Dataset of Molecules. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, August 2021.

[24] L. Järkelid, P. Kjellstrand, E. Martinson, and A. Wieslander. Toxicity of 20 Chemicals from the MEIC Programme Determined by Growth Inhibition of L-929 Fibroblast-like Cells. *Alternatives to laboratory animals: ATLA*, 25 (1):55–59, 1997. ISSN 0261-1929.

[25] Hao Zhu, Todd M. Martin, Lin Ye, Alexander Sedykh, Douglas M. Young, and Alexander Tropsha. Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chemical Research in Toxicology*, 22(12):1913–1921, December 2009. ISSN 0893-228X. doi: 10.1021/tx900189p.

[26] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, June 2019. ISSN 0895-4356, 1878-5921. doi: 10.1016/j.jclinepi.2019.02.004.

[27] DepMap - Broad Institute. https://depmap.org/repurposing/.

[28] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, August 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00237.

[29] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pretraining. https://arxiv.org/abs/2204.03044v1, April 2022.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015.

[31] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards Understanding Mixture of Experts in Deep Learning, August 2022.

[32] Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse Weight Averaging for Out-of-Distribution Generalization, January 2023.

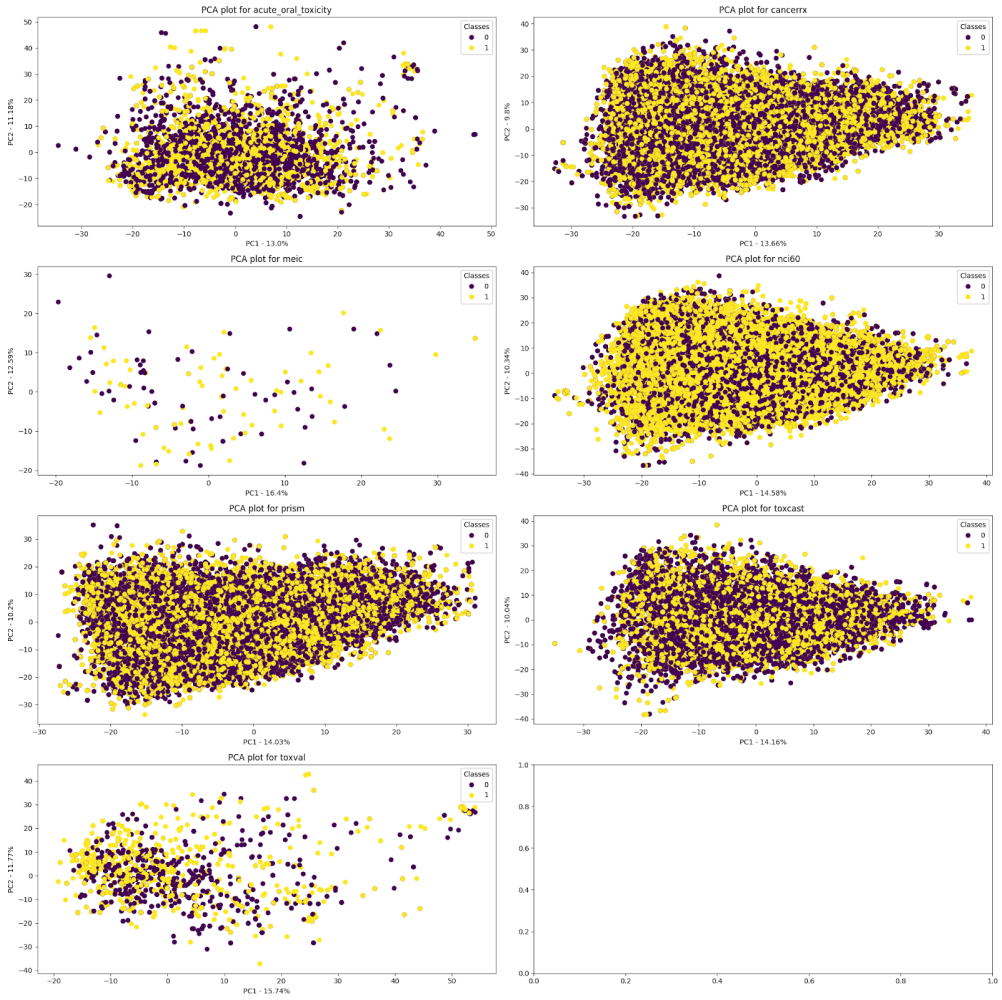
[33] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization, August 2023.

[34] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution, February 2022.

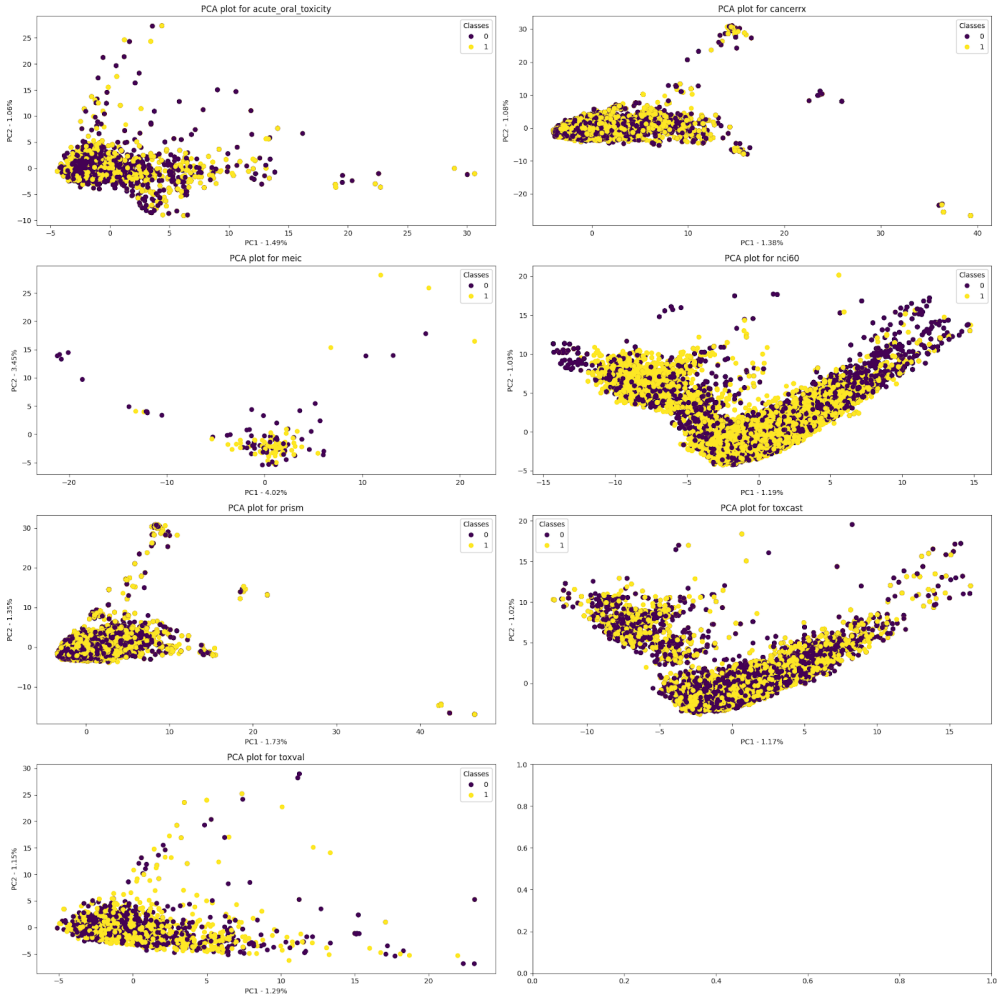
[35] Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. MolXPT: Wrapping Molecules with Text for Generative Pre-training, May 2023.

[36] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A Universal 3D Molecular Representation Learning Framework.

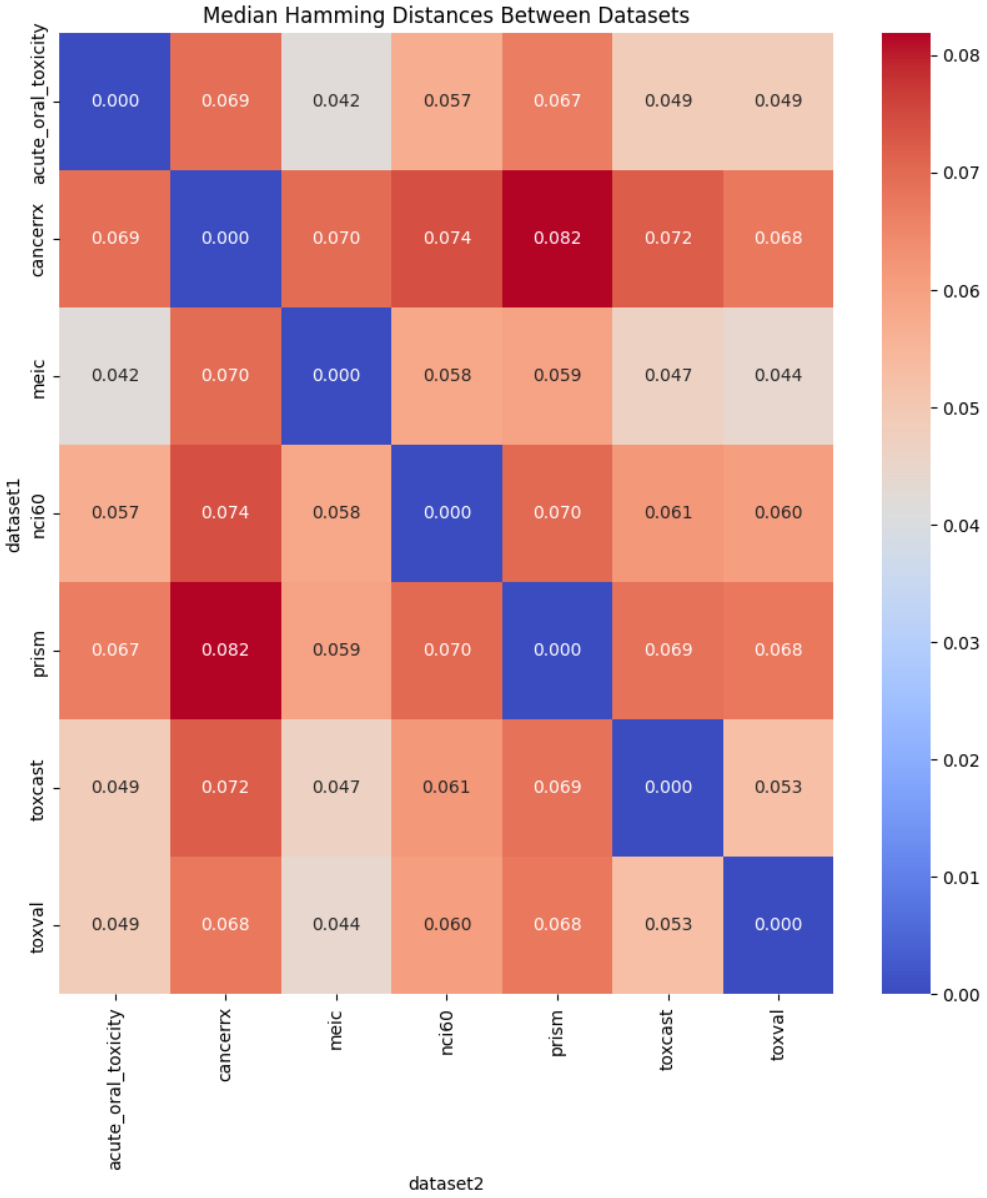
6 Appendix



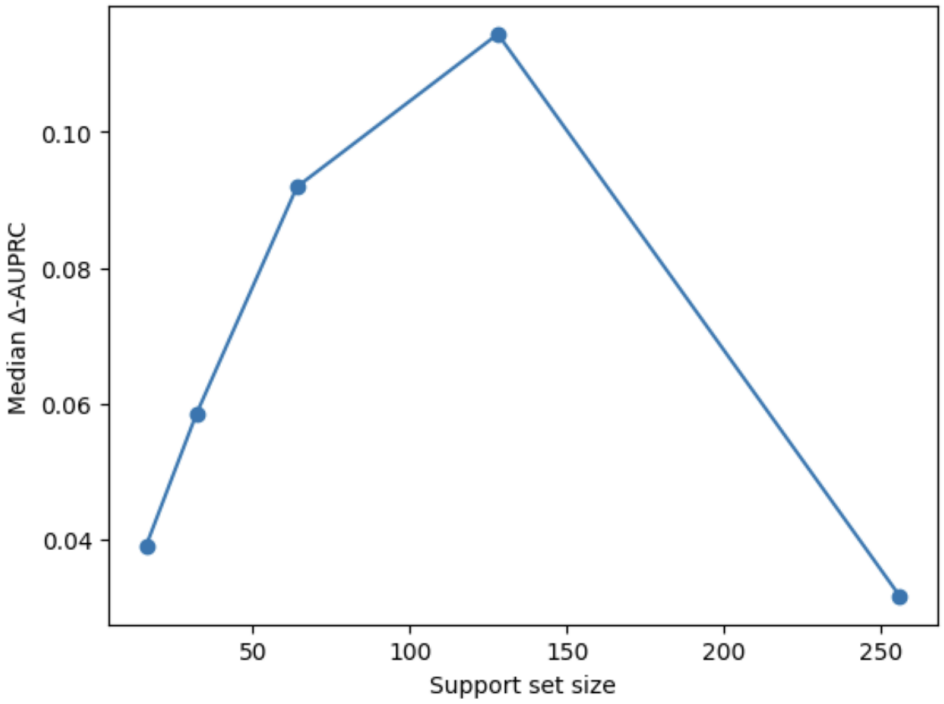
Appendix 1a. Principal components plot for ECFP4 fingerprints.



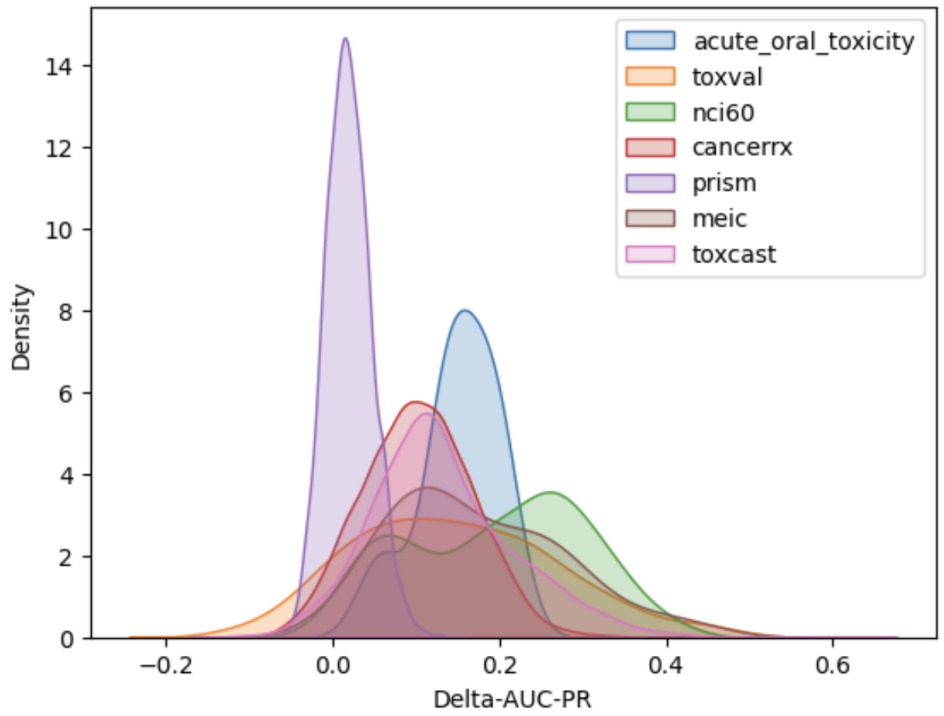
Appendix 1b. Principal components plot for ChemGPT embeddings.



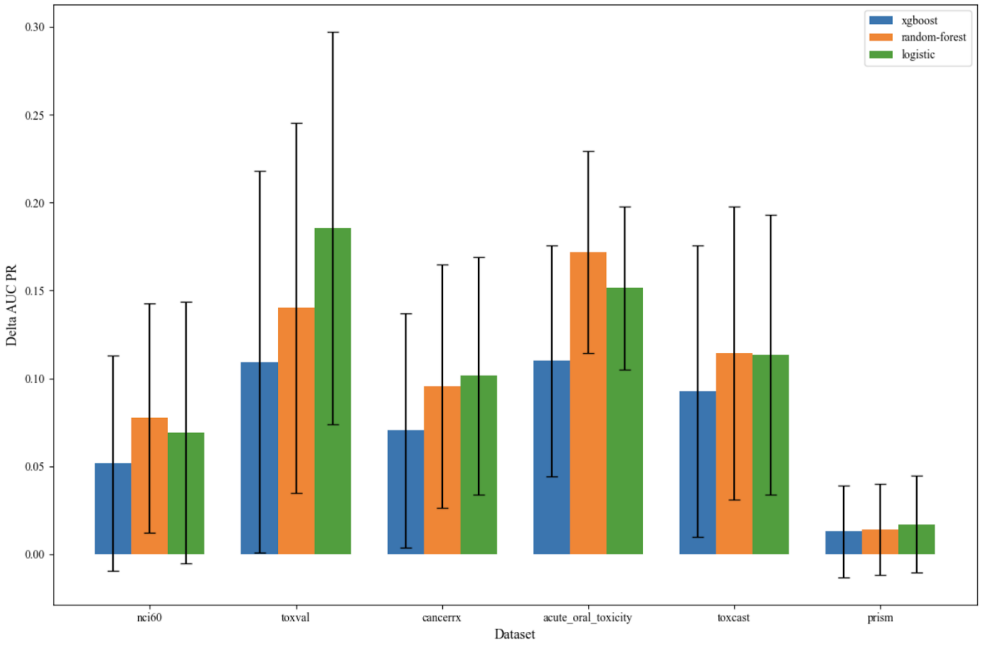
Appendix 2. Scatter plot of the difference in Δ-AUPRC scores between datasets, and their associated tanimoto similarity.



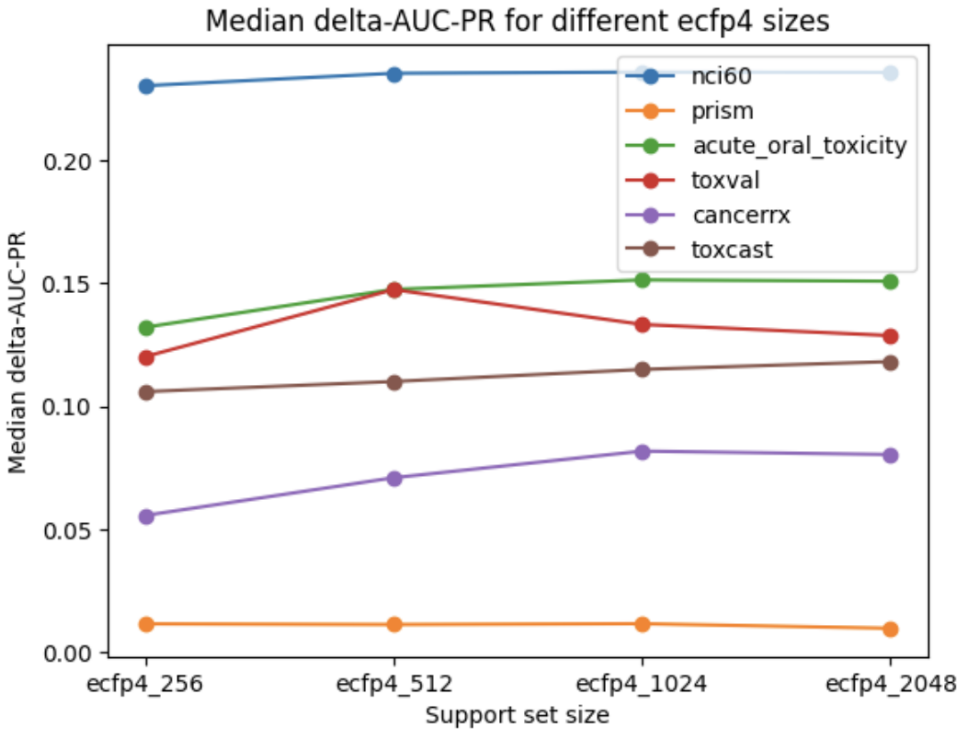
Appendix 3. Median Δ-AUPRC scores for different support set sizes across all datasets.



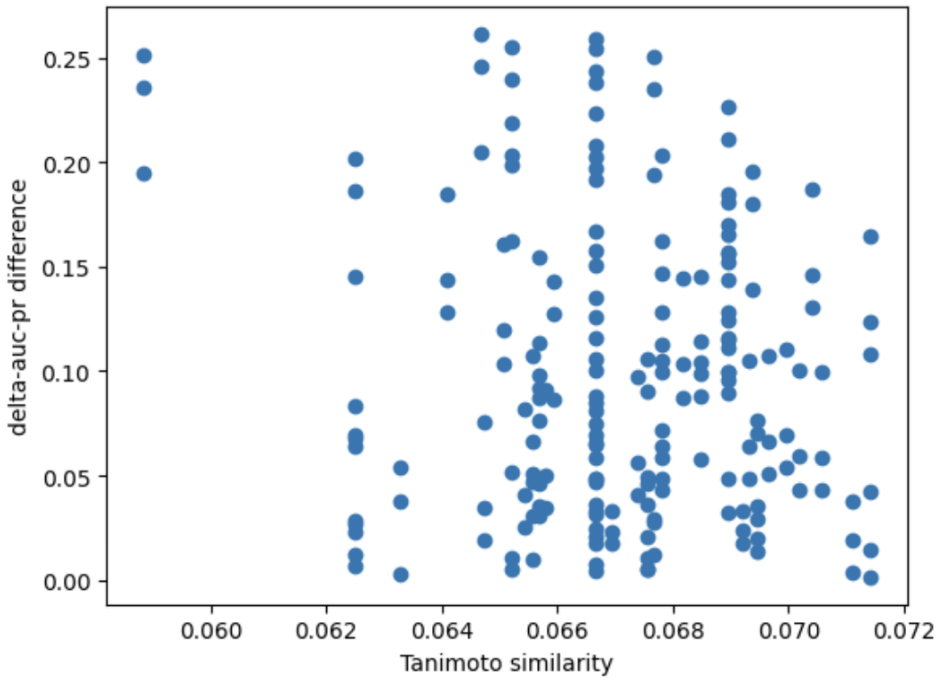
Appendix 4. Distribution of Δ-AUPRC for each constituent dataset.



Appendix 5. Median Δ-AUPRC scores comparing traditional machine learning prediction methods on ECFP4 fingerprints: logistic regression, random-forest, and XGBoost.



Appendix 6. Median Δ-AUPRC scores for different ECFP4 sizes.



Appendix 7. Scatter plot of the difference in Δ-AUPRC scores between datasets, and their associated tanimoto similarity.