AnchorDrug: A system for drug-induced gene expression prediction in new contexts through active learning

Han Meng*†, Ruoqiao Chen*†, Bin Chen*§, Jiayu Zhou‡§

Abstract

Large pre-trained models have been extensively explored for numerous biomedical tasks. However, the diversity and complexity of biological systems often make zero-shot learning in a new context challenging. In many instances, the budget allows for the acquisition of a small number of labeled data through experiments for few-shot learning. Yet, the methodology for selecting the optimal set of samples for these experiments remains underexplored. In this work, we present an application focused on drug-induced gene expression prediction to demonstrate a data-driven approach for facilitating sample selection. We developed a system named AnchorDrug, which predicts drug-induced gene expression changes in new cell lines after fine-tuning with experimental data from a limited number of drugs. Initially, we built a pre-trained model with a large dataset of drug-induced gene expressions. We then adopted active learning to identify an optimal set of drugs (i.e. anchor drugs) for experiments, aiming to ensure that the experimental data used for subsequent fine-tuning would maximize model performance. Several acquisition functions are customized and incorporated into our pipeline. Compared with knowledge-based drug selection, our customized active learning methods proved more effective in selecting anchor drugs. A model trained using data from anchor drugs can even perform better than that trained using all available data in certain scenarios. We further provided insights into the reasons behind its superior performance. Our system is designed to mimic real-world scenarios, enabling its easy application to real biomedical research projects.

1 Introduction

Advancements in high-throughput technologies, coupling with the decreasing costs of biological experiments, have led to the generation of voluminous data, enabling the creation of large pre-trained models for various tasks [10], [17], [3], [14]. Example tasks include drug-target prediction, gene-regulatory network modeling, cell surface protein prediction, and drug-induced gene expression prediction. Given the diversity of domains (or contexts), it is common to fine-tune a pre-trained model for a specific context. In practice, Preparing data for fine-tuning a model to a new domain involves conducting experiments with a small sample set. This approach is both necessary and feasible. However, how to select an optimal set of samples for experiments within a specific budget is significant yet not fully explored.

Drug-induced gene expression profiles are extensively used to discover new drugs and understand drug mechanisms [24], [28], [23]. Initiatives like LINCS have invested heavily in generating such profiles for tens of thousands of drugs across various cell lines under diverse biological conditions (e.g., treatment duration, dosage) [24]. A single drug can exhibit a unique gene expression profile in a specific cell line under a particular condition. However, given the existence of thousands of cell lines and the wide range of biological conditions, it is impractical to experimentally profile each drug in each cell line. Even the largest database LINCS only covers a few dozen cell lines. Machine learning models have proven to be effective in predicting expression profiles for a cell line that has a substantial amount of data for training [27], [25], [18], but many cell lines have little to no data, imposing a challenge to generalize the model to various contexts. Typically, generating such profiles in a new context is feasible for a limited set of drugs. For instance, recent advancements in high-throughput transcriptomics allow profiling of 30 (three replicates per compound in a 96-well plate) or 100 (365-well plate) compounds at a reasonable cost (https://alitheagenomics.com), [1]. Several studies have demonstrated the feasibility of performing high-throughput transcriptomics for a small library of drugs in neuron cells [13], [20]. These profiles can be unitized to fine-tune the model so that the model can support the prediction of drug-induced gene expression in a new context, but currently, there is no established method for selecting the drugs from a large library to perform experiments. In this work, we present a system named AnchorDrug, designed to predict drug-induced gene ex-

*Michigan State University, {menghan1, chenruo4, chenbi12 }@msu.edu
†Equal contribution
‡University of Michigan, jiayuz@umich.edu
§Co-corresponding

pression in new cell lines using experimental data from a limited number of drugs. Initially, we adopted a model that inputs the embeddings of drugs, genes, and cell lines, predicting the categorical change in expression of a gene following treatment with a drug in a specific cell line. Unlike previous models that input embeddings of drugs and cell lines to predict the expression of all genes, our model significantly increases the sample size, thereby enhancing performance. To mimic real-world applications, we selected three cell lines for validation and pretrained the model with data from the remaining cell lines. To enable the pretrained model to predict expression change in these test cell lines, we fine-tuned it using expression data from a selected set of drugs, which we referred to as anchor drugs. To select anchor drugs, we leveraged active learning and investigated multiple active learning choices. We showed the superiority of the anchor drugs and provided mechanistic insights. We anticipated that this system could be applied not only to real research studies (e.g., predicting drug-gene expression in neurons) but also to different projects where the budget is limited to conducting a few experiments in a new context.

Our work made the following contributions:
1) We developed a model that enables the prediction of drug-induced gene expression changes in a cell line, using the drug's chemical structure, the gene name, and the gene expression profile of the untreated cell line.
2) We investigated various strategies for selecting an optimal set of drugs for experiments, ensuring that the experimental data used for fine-tuning would lead to optimal performance.
3) We integrated various active learning methods into our system for selecting anchor drugs, tailored them to our specific task for better performance and efficiency, and provided insights into their superior performance.
4) We designed our system to emulate real-world scenarios, ensuring it can be seamlessly integrated into real research applications.

2 Related Works

Drug-induced Gene Expression Prediction. The massive existing drug-induced gene expression profiles have enabled the development of advanced machine-learning models to infer gene expression based solely on chemical structure. [9] developed a computational framework that first arranged existing profiles into a three-dimensional array indexed by drugs, genes, and cell types and then used either local or global information to predict unmeasured profiles. By this means, the missing drug-gene-cell type profiles were predicted from known pairs in LINCS. However, only known drugs in LINCS could be predicted. Godwin et al. [27] developed DeepCOP, a deep learning-based approach to tackle this challenge. They used molecular fingerprints to represent compounds and Gene Ontology (GO) terms to embed genes, which could be likely extended to any compounds and genes of interest. However, around 2/3 of the compound profiles are of poor quality and failing to account for this variation might degrade performance. A subsequent study proposed a collaborative model to improve input data quality, further boosting the overall performance [25]. Similarly, [18] utilized a GNN and multi-head attention mechanism to model chemical substructure-gene and gene-gene associations. Using the predicted profiles, they proposed repurposed candidates for SARS-CoV-2. Zhu et al. trained a multi-task deep neural network model that used the latent embedding from a chemical autoencoder as the input to generate compound-induced gene expression changes and applied it to four drug screening cases based on the disease transcriptome reversal concept [29]. While all of these prior studies have demonstrated the feasibility of predicting gene expression based on chemical structures, none have addressed the specific scenario we consider, where the gene expression profiles of only a certain number of compounds in a given context are available for fine-tuning the model.

Batch Mode Active Learning. Label efficiency is of high significance in domains such as drug discovery, where the cost of labeling data can be prohibitively expensive. Many researchers have worked on predictive models to make use of limited labeled data to explore unlimited possibilities. However, training powerful models needs the support of massive labeled data, which again leads us to obtain expensive labels. Decades ago, active learning [22] was proposed to improve the labeling efficiency for model training. Active learning can be approximately divided into two categories. One type is based on uncertainty, and the other makes use of features extracted from the network. Entering the era of deep learning, many researchers have worked on batch-mode active learning. [2] proposed to use the gradients of the last layer's input as features, which could take both uncertainty and data distribution into consideration. [6] worked on Bayesian neural networks, which could provide reliable prediction uncertainty with empirical demonstration. And [5] proposed to use dropout to estimate uncertainty from any network. Another work [15] was inspired by adversarial sampling and estimated the distance between samples and the classification boundary using adversarial samples created by a generative adversarial network.

Active Learning in Drug Discovery. Several researchers have focused on employing active learning techniques to address various application challenges. [16] introduced

a benchmark study that conducted experiments on numerous genomics datasets utilizing different acquisition functions. [26] discussed selecting drugs based on the maximum margin hyperplane principle generated by Support Vector Machines (SVM). [19] carried out experiments with the ChEMBL dataset, employing two acquisition functions to capture two critical considerations in drug discovery: the exploitativeness of a drug, which concentrates on identifying target drugs, and the explorativeness of a drug, highlighting the novelty of the drugs. Other contributions, such as [8], and [7], have also made significant strides in applying active learning to efficiently explore the chemical space. Overall, the efficiency of labeling remains a key challenge for researchers in drug discovery. However, the area of research involving drug-induced gene expression data remains unexplored.

3 Method

3.1 Problem Definition. Our objective is to predict gene expression change following a drug treatment in a target cell line (denoted as c_T) under a specified biological condition, via utilizing data of only a few drugs in the target cell line for model training. This is considered a three-class classification problem with the constraints of labeling budgets B, where B represents the number of drugs (Since one drug corresponds to multiple genes' expression values in our data, we consider all data points corresponding to one drug collectively as a labeling budget). The input to our model, denoted as $X = X_D \bigoplus X_C \bigoplus X_G$, $X \in \mathbb{R}^p$ ($p = p_D + p_C + p_G$), is the concatenation of drug, gene, and cell line representations, where $X_D \in \mathbb{R}^{p_D}$, $X_G \in \mathbb{R}^{p_G}$ and $X_C \in \mathbb{R}^{p_C}$ denotes the drug, gene, and cell line representations, respectively (fig. 1(a)). The output is the drug-induced gene expression change, where 0, 1, and 2 represent down-regulation, no change, and up-regulation, respectively (fig. 1(a)).

3.2 AnchorDrug Pipeline AnchorDrug system comprised three stages: pre-training, fine-tuning, and prediction (fig. 1(b)). Our model was designed to leverage information on drug structure, cell line characteristics, and gene function together as input, enabling it to predict expression changes for any drug, gene, or cell line. During the pre-training stage, the model f_0 was pre-trained using a large number of samples from the LINCS database to effectively incorporate knowledge from the source cell lines (fig. 1(b)). During the fine-tuning stage, we optimized the fine-tuning data by employing active learning methods to select a small number of drugs (i.e., anchor drugs), obtaining their data in the target cell line from wet-lab experiments, and fine-tune the pre-trained model (fig. 1(b)-(c)). For anchor drug selection, we incorporated seven acquisition functions into our pipeline. Our budget constraint is tied to the number of drugs to profile rather than the quantity of data points. Thus we tailored each acquisition function to emphasize individual drugs rather than data points. These seven acquisition functions, which include state-of-the-art active learning methods, covered the three main principles of active learning: uncertainty, representativity, and diversity, thereby broadening our selection options. Finally, during the prediction stage, the fine-tuned model is applied to predict drug-induced gene expression change for all drugs in the target cell line.

3.3 Base Model The preparation of the base model involved extracting cell line embeddings using an autoencoder, followed by pretraining of the base model.

Cell Line Embeddings An autoencoder was used to generate a 128 bits cell line embeddings from the original cell line features $\hat{X_C} \in \mathbb{R}^{\hat{p_C}} (\hat{p_C} > 10,000)$. MSE loss is used for model training. The objective function is defined as:

$$argmin_{\phi\psi} \frac{1}{n} \sum_{o=1}^{n} (\hat{X_{Co}} - \psi(\phi(\hat{X_{Co}})))^2,$$

where $\phi : \hat{X_C} \to X_C$, $\psi : X_C \to \hat{X_C}$, $X_C \in \mathbb{R}^{128}$, and n is the number of cell lines.

The Universal Model We designed a universal model that enabled the prediction of drug-induced gene expression change for any provided drugs and genes in any cell lines. Gene representations (X_G, GO terms), drug representations (X_D, ECFP), and cell line representations (X_C, cell line embeddings) were concatenated together to serve as the input (fig. 1(a)). The model was a fully connected MLP that outputs the possibilities of the three categories. Softmax function was applied to the output. Cross-entropy loss was used for model training. The objective function was defined as:

$$argmin_{f_0} \frac{1}{m} \sum_{o=1}^{m} (- \sum_{c=1}^{M} \mathbb{I}_c \log(f_0(X_D \bigoplus X_C \bigoplus X_G)_{o,c}))$$

where M is the number of classes, m is the number of observations. $\mathbb{I}_c$ is the binary indicator if class label c is the correct label and f_0 maps $X_D \bigoplus X_C \bigoplus X_G$ to label.

3.4 Active Learning Active learning takes the model and drug pool as input and outputs the candidate drugs for experiments. As shown in the fig. 1(c), our active learning cycle was initialized by a pre-trained model, a data pool, and an empty training dataset. In every cycle, it works as follows:

1) Apply the model to inference on the data pool;
2) Select a drug set using the inference information according to the acquisition function;
3) Get labels of the selected drugs' data, add them to the training dataset, and delete them from the drug pool.
4) At the end of every cycle, fine-tune the pre-trained model to prepare for the next cycle if the budgets have not been used up. Otherwise, end the cycle and output

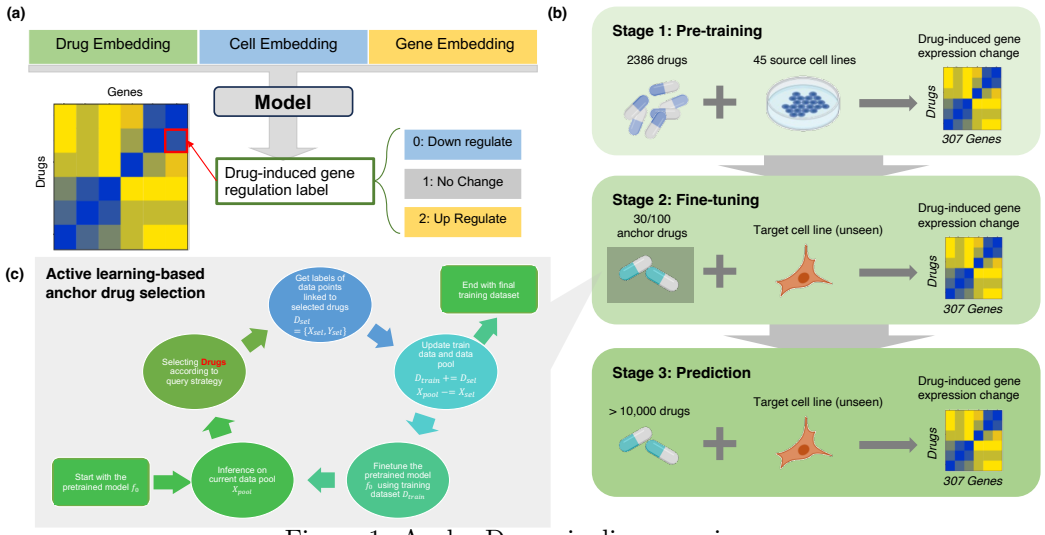

Figure 1: AnchorDrug pipeline overview.

the final training dataset.

We considered seven acquisition functions: AdversarialBIM [12], BALDDropout [6], BADGE [2], CoreSet [21], K-Means Sampling, Least-Confidence Sampling, and Margin Sampling. All of them were tailored to focus on drugs instead of individual data to achieve better performance or essential efficiency.

3.5 Model Fine-tuning The pre-trained model was fine-tuned in a cell line-specific manner, where data of the selected anchor drugs in the corresponding target cell line was used for fine-tuning, resulting in a model specifically for a target cell line (fig. 1(b)). The dominant label classes were down-sampled to mitigate the negative effect of data imbalance.

4 Experiments

4.1 Dataset Drug-induced gene expression profiles from LINCS L5 [24] data were collected, where only the labeled high-quality profiles under the biological condition of 10µM concentration and 24-hour treatment duration were used. The data were further filtered based on genes, where the profiles of 307 highly predictable genes [25] were filtered and used. We also filtered out cell lines that are not in OCTAD [28]. After filtering, the dataset covered 5886 unique drugs, 54 cell lines, and more than 20000 experiments. One experiment corresponded to one drug being tested in one cell line. Data of all 307 highly predictable genes were collected at the same time.

For validation use, the top three cell lines with the most number of profiles, A549, PC3, and MCF7 (containing 936, 1413, and 1337 drugs, respectively), were selected as the target cell lines, and their corresponding data were excluded from model pre-training (fig. 1(b)). Pretraining data covered 45 cell lines, 4872 unique drugs, and 17624 experiments. In each of the three target cell lines, random 10% of the drugs were held out, whose data served as the test set for the corresponding cell line. The rest of the data served as the drug pool for active learning for the corresponding cell line.

Gene expression profiles in 1210 untreated cell lines from OCTAD [28] were used for generating lower-dimensional cell line embeddings (128 bits). ECFP4 features (1024 bits) were used as drug representations. GO terms from [4] were used as gene representations.

The GitHub link for data and code of AnchorDrug system is: https://github.com/Bin-Chen-Lab/AnchorDrug_official.

4.2 Anchor Drug Selection Scenario We proposed two real-world scenarios of selecting anchor drugs for model fine-tuning:

S1 Design a common list of N anchor drugs which can be applied to any new target cell line, and

S2 For each new target cell line, design a specific list of N anchor drugs.

The considerations were that **S1** should be more economical than **S2**, whereas it would also be useful to know if cell line-specific anchor drug lists perform better than a common anchor drug list. In **S1**, we assumed that the common anchor drugs were representative enough for our model to learn contextual knowledge from various new cell lines.

To mimic these two scenarios with the three target cell lines, we first established an "initial pool", the pool which we would later select anchor drugs from, using the drugs in the intersection between the pre-trained LINCS data and the target cell lines' data. Drugs in the pre-trained LINCS data were used as a universal standard. In reality, what drugs have been measured for any new target cell line are unclear in advance. The pre-trained LINCS data consists of data that have already

Selected Drug Numbers		30				100			
Cell Line Name		A549	MCF7	PC3	Mean	A549	MCF7	PC3	Mean
Anchor Drug Pipeline	Margin	0.572	0.517	0.537	0.542	0.591	0.552	0.578	0.574
	LeastConfidence	0.575	0.523	0.540	0.546	0.591	0.552	0.576	0.573
	BALDDropout	**0.576**	**0.554**	0.555	**0.561**	**0.594**	0.566	**0.583**	**0.581**
	BADGE	0.569	0.543	0.552	0.555	0.569	0.557	0.571	0.565
	CoreSet	0.573	0.547	**0.559**	0.560	0.588	**0.568**	0.576	0.577
	K-Means	0.566	0.539	0.554	0.553	0.571	0.557	0.571	0.566
	AdversarialBIM	0.575	0.547	0.557	0.559	0.586	0.562	0.578	0.575
All Anchor Drug Options - Averaged		0.572	0.539	0.551	0.554	**0.584**	0.559	**0.576**	**0.573**
Baseline	Clustering	0.568	0.540	0.556	0.555	0.576	0.549	0.562	0.562
	MOA	0.563	0.533	0.554	0.550	0.569	0.554	0.563	0.562
	Random	0.564	0.535	0.556	0.552	0.571	0.553	0.565	0.563
All Baselines - Averaged		0.565	0.536	0.556	0.552	0.572	0.552	0.563	0.562
All Data Finetuned Model		**0.579**	**0.561**	**0.570**	**0.570**	0.579	**0.561**	0.570	0.570
Pre-trained Model		0.519	0.415	0.463	0.466	0.519	0.415	0.463	0.466

Table 1: Benchmarking evaluation of AnchorDrug's performance with a common list of anchor drugs shared by all target cell lines (averaged across three runs of random seeds for each method).

been measured, ensuring its popularity and general applicability to any newly given target cell lines to the greatest extent. Plus, we needed to ensure that the anchor drugs we selected had corresponding data in the target cell lines for the sake of validation. Thus, under Scenario 1 (**S1**), drugs shared by all the three target cell lines and the pre-trained LINCS data were used as our initial pool. For scenario 2 (**S2**), drugs shared by the target cell line and the pre-trained LINCS data were used as the initial pool for the corresponding target cell line. For both scenarios, the test set was held out in advance, as described in section 4.1.

4.3 Benchmarking Evaluation To evaluate AnchorDrug's performance, we compared AnchorDrug, including seven acquisition function options, with another three traditional drug section baselines:

1) Random selection: Random N drugs were selected from the initial pool.

2) Clustering-based selection: Drugs in the initial pool were grouped into N clusters by K-modes clustering [11] based on their drug representations (ECFP). From each cluster, a random drug was selected.

3) Mechanism of action (MOA)-based selection: Drugs in the initial pool were grouped by their MOAs, random N groups of MOAs were selected, where a random drug was selected from each group.

We also included the pre-trained base model and the model fine-tuned using all data of the initial pool in our comparison. Experiment settings except for fine-tuning data were consistent across experiments. Each drug selection method was run 3 times to select 30 or 100 drugs. The F1 score (macro) was averaged across three runs for each method.

*With a Common List of Anchor Drugs (**S1**)* Results in table 1 showed the averaged performance of AnchorDrug outperformed the other three baselines in all three target cell lines under both 30 and 100 anchor drug conditions, and was significantly better than the pre-trained model (t-test, $p < 0.05$), with an improvement of 18.89% and 22.95% in the mean F1 score across all cell lines under 30 and 100 anchor drug conditions, respectively. Surprisingly, when 100 anchor drugs were used for finetuning, the averaged performance of AnchorDrug even outperformed the model fine-tuned using all available data slightly.

In all three cell lines, BALDDropout achieved superior F1 score under both 30 and 100 anchor drug conditions, even showing a higher F1 score than all data fine-tuned models under the 100 anchor drug condition. Other acquisition functions such as CoreSet and AdversarialBIM also showed comparable performance.

Taken together, our results showed the superiority of active learning-based methods over random selection or domain knowledge-based methods in drug selection for training data (table 1). This indicated AnchorDrug's effectiveness in designing a common list of anchor drugs for different target cell lines.

*With Cell Line-specific Lists of Anchor Drugs (**S2**)* Results in table 2 showed that AnchorDrug outperformed the other three baselines in all three target cell lines under both 30 and 100 anchor drug conditions, and was significantly better than the pre-trained model baseline (t-test, $p < 0.05$), with an improvement of 17.60% and 22.10% in the mean F1 score across all cell lines under 30 and 100 anchor drug conditions, respectively. Similar to **S1**, in **S2**, the averaged performance of AnchorDrug outperformed the model fine-tuned using all available data slightly under 100 anchor drug conditions, with

Selected Drug Numbers		30				100			
Cell Line Name		A549	MCF7	PC3	Mean	A549	MCF7	PC3	Mean
Anchor Drug Pipeline	Margin	0.569	0.509	0.529	0.536	0.589	0.548	0.561	0.566
	LeastConfidence	0.571	0.512	0.534	0.539	0.588	0.555	0.562	0.568
	BALDDropout	0.569	0.539	**0.555**	0.554	**0.593**	0.566	**0.577**	**0.578**
	BADGE	0.568	0.542	0.554	0.555	0.571	0.557	0.567	0.565
	CoreSet	0.561	0.538	0.549	0.549	0.587	**0.569**	0.574	0.576
	K-Means	0.566	0.539	0.540	0.548	0.565	0.550	0.547	0.554
	AdversarialBIM	**0.580**	**0.542**	0.551	**0.558**	0.590	0.567	0.571	0.576
All Anchor Drug Options - Averaged		**0.569**	0.532	0.545	0.548	**0.583**	**0.559**	**0.566**	**0.569**
Baseline	Clustering	0.554	0.539	0.549	0.548	0.562	0.546	0.551	0.553
	MOA	0.541	0.527	0.535	0.534	0.550	0.535	0.547	0.544
	Random	0.542	0.532	0.540	0.538	0.547	0.539	0.545	0.544
All Baselines - Averaged		0.546	0.532	0.541	0.540	0.553	0.540	0.548	0.547
All Data Finetuned Model		0.563	**0.547**	**0.551**	**0.554**	0.563	0.547	0.551	0.554
Pre-trained Model		0.519	0.415	0.463	0.466	0.519	0.415	0.463	0.466

Table 2: Benchmarking evaluation of AnchorDrug's performance with cell line-specific lists of anchor drugs .

an improvement of 2.71%. Among all baselines including all data fine-tuned models and all acquisition functions, AdversarialBIM and BALDDropout achieved the highest mean F1 score across all cell lines under the 30 and 100 anchor drug conditions, respectively. Under the 100 anchor drug condition, six acquisition functions achieved better F1 scores in all cell lines than the model fine-tuned using all available data.

Interestingly, compared to the results in **S2**, the results in **S1** showed overall better performance (table 1, table 2). The model fine-tuned using all data in the **S1** initial pool also outperformed the model fine-tuned using all data in the **S2** initial pool. This might be due to the higher quality of the initial pool used for **S1** compared to **S2**. Another observation was that the performance gap between baselines and active learning-based methods was larger in **S2** than in **S1**. These findings together implied that in reality, when the drug pool is large and of low quality (useful information is sparse), the AnchorDrug system is highly promising to achieve a much better performance than other solutions.

5 Discussion

5.1 Anchor Drug Analysis To understand why the selected anchor drugs can improve model prediction performance, we used TSNE to visualize the distribution of gene profiles induced by the 30/100 selected anchor drugs in each target cell line (use the BALDDropout function in AnchorDrug as an example), and compared with the distribution of gene profiles induced by the other deselected, non-anchor drugs from the same initial drug pool. Results revealed that under both **S1** and **S2**, in each target cell line, gene profiles induced by the 30/100 anchor drugs show a sparse distribution across the whole space (fig. 2,). Also, as indicated by the black arrows (fig. 2), some anchor drugs were selected from outlier clusters in the distribution space. This suggested that AnchorDrug tends to select N anchor drugs which are more sparsely scattered across the whole space of the initial pool regarding their induced-gene expression profiles in target cell lines.

To further quantitatively evaluate the distribution space coverage of the anchor drugs' induced-gene expression profiles in target cell lines, we compared AnchorDrug (BALDDropout) to the other three drug selection method baselines (Random, Clustering, and MOA). In each target cell line, we calculated the Convex Hull for anchor drugs and all drugs from the initial pool, respectively, using their TSNE embeddings of induced gene expression profiles (fig. 2,), and then divided the two Convex Hull values to obtain a space coverage ratio. Results showed that in **S2**, under both 30/100 anchor drug conditions, selected anchor drugs showed a higher space coverage with their induced-gene expression profiles compared to drugs selected by the other three baselines . The same held true for **S1** under the 100 anchor drug conditions . Another metric of pairwise Euclidian distance variance also revealed consistent results . Moreover, a correlation test confirmed the positive relation between AnchorDrug performance and the distribution space coverage of selected anchor drugs' induced-gene expression profiles in target cell lines (fig. 3). Interestingly, anchor drugs selected in **S1** showed a higher space coverage compared to the anchor drugs selected in **S2** . This may account for the underlying reason why a common list of anchor drugs shared by all target cell lines performs better than the cell line-specific anchor drug lists (table 1, table 2).

Taken together, our results indicated that AnchorDrug tends to select a combination of N anchor drugs

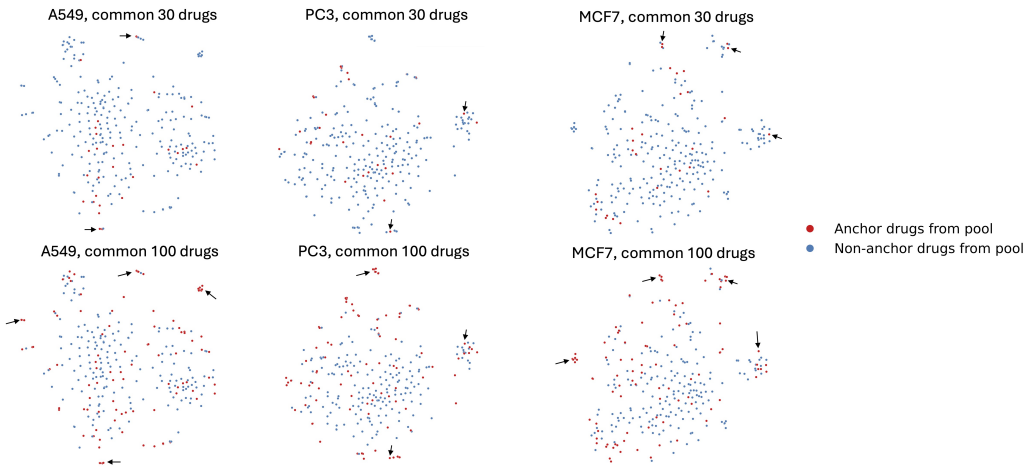

Figure 2: Distribution of the drug induced-gene expression profiles in the target cell lines from 30/100 anchor drugs selected by AnchorDrug (BALDDropout, one run using one random seed) compared to the whole initial pool (279 drugs) under **S1**.

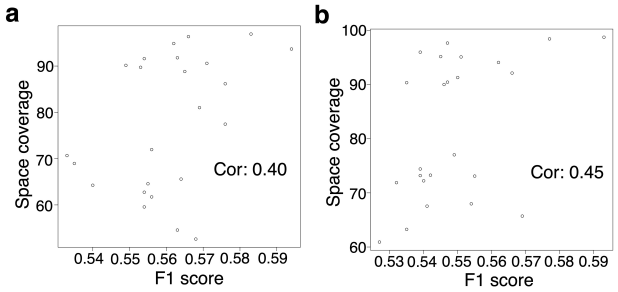

Figure 3: Correlation between the F1 score and the distribution space coverage of drug-induced gene expression profiles from selected drugs under **S1** (a) and **S2** (b). One dot represents one drug selection method with a certain number of selected drugs in one target cell line. For each scenario, three target cell lines, four drug selection methods (Random, Clustering, MOA, and AnchorDrug (BALDDropout)), and two numbers (30 and 100) of selected drugs are included.

whose induced-gene expression profiles in target cell lines optimally represent the distribution space of all drugs in the initial pool. This may result in optimal samples in the model fine-tuning stage, which boosted the effects of fine-tuning and contributed to AnchorDrug's good prediction performance.

5.2 Convergence of Performance with the Increase of Drug Number We first investigated the convergence of model performance with random addition of data. In all three cell lines, after 100 drugs were covered, the F1 score increased only slightly as the training data covered more drugs. When all data was included, the F1-score was still under 0.6, which is comparable to the low F1 scores reported in other papers [25]. This suggested that for the drug-induced gene expression prediction task, information scarcity of data remains a bottleneck for improving performance. The converge curve is zig-zag, which indicates that randomly added new data may have beneficial or detrimental effects on model performance, implying the importance of training data selection.

The model performance change during active learning is shown in fig. 4. Performance ranged from a pre-trained model (no fine-tuned data) to a model fine-tuned using data from 100 drugs. The curve of randomly adding training data under the same experiment setting was also included for a fair comparison. The overall trend showed the F1-score increased as the training sample size grew in three cell lines for every acquisition function. Across three cell lines, LeastConfidence stably achieved good F1-scores and Random was always among the poor-performance group.

As shown in fig. 4, there was a performance degradation from the pre-trained model to the model fine-tuned using data from 10 drugs. This may be due to the transition between different contexts. The pre-trained model was trained on other cell lines, while the fine-tuned model was fine-tuned on a new, unseen cell line, representing a new biological context. A very small amount of data can lead to a highly biased representation of the new context, resulting in a worse model than the pre-trained one. Among all the functions, LeastConfidence and Margin did not experience this performance degradation and achieved a relatively smooth transition from the pre-trained model to the fine-tuned model. This implied that for an extremely low labeling budget, these two acquisition functions may be a reasonable choice. Results in fig. 4 again demonstrated the superiority of active learning-based drug selection.

5.3 Adaptation of Active Learning to Drug-induced Gene Expression Data Because of the

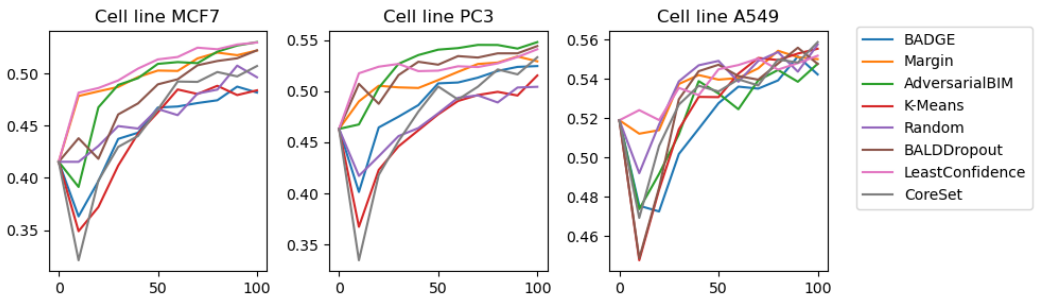

Figure 4: F1 Score v.s. Number of Drugs in Finetunning Samples during active learning. The experiments are conducted under scenario 2. For every step, we increase the data of 10 drugs into training data. (averaged across three runs of random seeds for each method)

nature of wet lab experiments, testing one drug in one cell line simultaneously generates expression data for multiple genes. Available active learning acquisition functions select data points. Applying them to our real-world scenarios means that when one data point is selected and labeled, all other data points of the same drug are also labeled regardless of whether they are selected or not. Intuitively, this would cause efficiency problems and lead to a poorly designed training dataset. Therefore, we modified several widely used acquisition functions to focus on drugs instead of individual data points to better suit our needs. These customized query strategies fall into three categories based on our adaptations.

LeastConfidence, BALDDropout, and MarginSampling are uncertainty-based acquisition functions. The original algorithms select the most uncertain data and its corresponding drug, then add all associated data to the training set. Our adapted version treated the average uncertainty of all associated data as the uncertainty of the corresponding drug. Comparing results in table 3 and table 2, for LeastConfidence, MarginSampling, and BALDDropout, our adapted version consistently achieved higher F1 scores in all three cell lines, which validated the superiority of our modification.

method	cell	A549	MCF7	PC3	mean
Margin		0.573	0.532	0.546	0.550
LeastConfidence		0.569	0.531	0.548	0.549
BALDDropout		0.580	0.557	0.570	0.569

Table 3: Performance of the original MarginSampling, LeastConfidence, and BALDDropout methods under the 100 anchor drug condition in scenario 2 .

KMeansSampling, KCenterGreedy, and BadgeSampling all utilize model embeddings. The original methods generate embeddings for individual data points and then apply clustering, which involves extensive computations like pairwise distance calculations among all embeddings, making it impractical to implement for our task. Our adapted version generated embeddings for drugs and avoided calculations between data points of the same drug, whose difference solely came from gene variance. (Cost analysis can be found in the github). Thus, our adapted methods are more efficient, practical, and biologically reasonable than the original methods.

The third category, AdversarialBIM, updates the input according to gradient descent until the output differs from the original one. The difference between the original and the updated input is regarded as the distance between the data and the decision boundary. It then picks data closest to the decision boundary. The original one calculates gradients and updates each data individually, making it impractical for our task. In contrast, our revised version updates all data of one drug simultaneously, significantly improving efficiency. Plus, the original one alters the whole input including drug, cell line, and gene representations, while our modified version only updates drug representation, providing a more accurate distance between drugs and the decision boundary. To sum up, our modified version is more reasonable and practical compared to the original counterparts.

6 Conclusion

In this work, we proposed the AnchorDrug system, which incorporated pertaining, fine-tuning, and active learning, providing a promising framework to explore new contexts and train a powerful predictive model of drug-induced gene expression. We further proposed two real-world scenarios and demonstrated that under both scenarios, the AnchorDrug system is superior in selecting an optimal combination of anchor drugs for model fine-tuning.

7 Acknowledgment

This research was supported by the National Institute of General Medical Sciences R01GM134307 and R01GM145700, the National Institute of Aging 1RF1AG072449, the National Science Foundation IIS-2212174, and the MSU Foundation Strategic Partnership Grant (SPG).

References

[1] Daniel Alpern, Vincent Gardeux, Julie Russeil, Bastien Mangeat, Antonio CA Meireles-Filho, Romane Breysse, David Hacker, and Bart Deplancke. Brb-seq: ultra-affordable high-throughput transcriptomics enabled by bulk rna barcoding and sequencing. *Genome biology*, 20(1):1–15, 2019.

[2] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds, 2020.

[3] Bin Chen, Lana Garmire, Diego F Calvisi, Mei-Sze Chua, Robin K Kelley, and Xin Chen. Harnessing big 'omics' data and ai for drug discovery in hepatocellular carcinoma. *Nature Reviews Gastroenterology & Hepatology*, 17(4):238–251, 2020.

[4] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):1–7, 2009.

[5] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[6] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017.

[7] David E Graff, Eugene I Shakhnovich, and Connor W Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical science*, 12(22):7866–7881, 2021.

[8] Filipp Gusev, Evgeny Gutkin, Maria G Kurnikova, and Olexandr Isayev. Active learning guided drug design lead optimization based on relative binding free energy modeling. *Journal of Chemical Information and Modeling*, 63(2):583–594, 2023.

[9] Rachel Hodos, Ping Zhang, Hao-Chih Lee, Qiaonan Duan, Zichen Wang, Neil R Clark, Avi Ma'ayan, Fei Wang, Brian Kidd, Jianying Hu, et al. Cell-specific prediction and application of drug-induced gene expression profiles. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 32–43. World Scientific, 2018.

[10] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, 18(10):1033–1036, October 2022.

[11] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. *Dmkd*, 3(8):34–39, 1997.

[12] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[13] Jingyao Li, Daniel J Ho, Martin Henault, Chian Yang, Marilisa Neri, Robin Ge, Steffen Renner, Leandra Mansur, Alicia Lindeman, Brian Kelly, et al. Drug-seq provides unbiased biological activity readouts for neuroscience drug discovery. *ACS Chemical Biology*, 17(6):1401–1414, 2022.

[14] Jianzhu Ma, Samson H Fong, Yunan Luo, Christopher J Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk FA Wessels, Marc Hafner, Roded Sharan, Jian Peng, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021.

[15] Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3071–3079, 2020.

[16] Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. GeneDisco: A Benchmark for Experimental Design in Drug Discovery. In *International Conference on Learning Representations (ICLR)*, 2022.

[17] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[18] Thai-Hoang Pham, Yue Qiu, Jucheng Zeng, Lei Xie, and Ping Zhang. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. *Nature machine intelligence*, 3(3):247–257, 2021.

[19] Daniel Reker and Gisbert Schneider. Active-learning strategies in computer-assisted drug discovery. *Drug discovery today*, 20(4):458–465, 2015.

[20] Steve Rodriguez, Clemens Hug, Petar Todorov, Nienke Moret, Sarah A Boswell, Kyle Evans, George Zhou, Nathan T Johnson, Bradley T Hyman, Peter K Sorger, et al. Machine learning identifies candidates for drug repurposing in alzheimer's disease. *Nature communications*, 12(1):1033, 2021.

[21] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[22] Burr Settles. Active learning literature survey. 2009.

[23] Marina Sirota, Joel T Dudley, Jeewon Kim, Annie P Chiang, Alex A Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96):96ra77–96ra77, 2011.

[24] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles.

Cell, 171(6):1437–1452, 2017.

[25] Mengying Sun, Jing Xing, Bin Chen, and Jiayu Zhou. Robust collaborative learning with noisy labels. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1274–1279. IEEE, 2020.

[26] Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

[27] Godwin Woo, Michael Fernandez, Michael Hsing, Nathan A Lack, Ayse Derya Cavga, and Artem Cherkasov. Deepcop: deep learning-based approach to predict gene regulating effects of small molecules. *Bioinformatics*, 36(3):813–818, 2020.

[28] Billy Zeng, Benjamin S Glicksberg, Patrick Newbury, Evgeny Chekalin, Jing Xing, Ke Liu, Anita Wen, Caven Chow, and Bin Chen. Octad: an open workspace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features. *Nature protocols*, 16(2):728–753, 2021.

[29] Jie Zhu, Jingxiang Wang, Xin Wang, Mingjing Gao, Bingbing Guo, Miaomiao Gao, Jiarui Liu, Yanqiu Yu, Liang Wang, Weikaixin Kong, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature biotechnology*, 39(11):1444–1452, 2021.