Velodrome: Out-of-Distribution Generalization from Labeled and Unlabeled Gene Expression Data for Drug Response Prediction

Hossein Sharifi-Noghabi[1,2], Parsa Alamzadeh Harjandi[1], Olga Zolotareva[3,4], Colin C. Collins[2,5], Martin Ester[1,2,*]

[1]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada.
[2]Vancouver Prostate Center, Vancouver, British Columbia, Canada.
[3]Chair of Experimental Bioinformatics, School of Life Sciences, Technical University of Munich, Germany.
[4]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany.
[5]Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada.
*Corresponding author: ester@sfu.ca

Abstract

Data discrepancy between preclinical and clinical datasets poses a major challenge for accurate drug response prediction based on gene expression data. Different methods of transfer learning have been proposed to address this data discrepancy. These methods generally use cell lines as source domains and patients, patient-derived xenografts, or other cell lines as target domains. However, they assume that they have access to the target domain during training or fine-tuning and they can only take labeled source domains as input. The former is a strong assumption that is not satisfied during deployment of these models in the clinic. The latter means these methods rely on labeled source domains which are of limited size. To avoid this assumption, we formulate drug response prediction as an out-of-distribution generalization problem which does not assume that the target domain is accessible during training. Moreover, to exploit unlabeled source domain data, which tends to be much more plentiful than labeled data, we adopt a semi-supervised approach. We propose Velodrome, a semi-supervised method of out-of-distribution generalization that takes labeled and unlabeled data from different resources as input and makes generalizable predictions. Velodrome achieves this goal by introducing an objective function that combines a supervised loss for accurate prediction, an alignment loss for generalization, and a consistency loss to incorporate unlabeled samples. Our experimental results demonstrate that Velodrome outperforms state-of-the-art pharmacogenomics and transfer learning baselines on cell lines, patient-derived xenografts, and patients and therefore, may guide precision oncology more accurately.

Introduction

The goal of drug response prediction based on the genomic profile of a patient (also known as pharmacogenomics) -- a crucial task of precision oncology -- is to utilize the omics features of a patient to predict response to a given drug (Garraway, Verweij, and Ballman 2013; Cronin et al. 2018; Marquart, Chen, and Prasad 2018; Pal et al. 2019). Unfortunately, patient datasets with drug response are often small or not publicly available which motivated the creation of large-scale preclinical resources such as patient-derived xenografts (PDX) (Gao et al. 2015) or cancer cell lines (Garnett et al. 2012; Barretina et al. 2012; Basu et al. 2013; Seashore-Ludlow et al. 2015; Klijn et al. 2015; Iorio et al. 2016; Haverty et al. 2016) as proxies for patients.
Although preclinical datasets are viable proxies for patients, they differ in important ways from patients due to basic biological differences such as the lack of tumor microenvironment/the immune system (Mourragui et al. 2019; Sharifi-Noghabi et al. 2020) -- even two preclinical datasets may have discrepancies with each other (Haibe-Kains et al. 2013; Safikhani et al. 2016; Haverty et al. 2016; Mpindi et al. 2016; Geeleher et al. 2016).

Transfer learning has emerged as a machine learning paradigm for such scenarios (Pan and Yang 2010; Neyshabur, Sedghi, and Zhang 2020), where we have access to different datasets from multiple resources (known as source domains) and want to make predictions for a dataset of interest (known as target domain) and it has been employed in different problems (Raghu et al. 2019; Holmberg et al. 2020; Hu et al. 2020). Various methods of transfer learning have been proposed in the context of drug response prediction. These methods either address these discrepancies implicitly (Sharifi-Noghabi et al. 2019; Snow et al. 2020; Kuenzi et al. 2020), or explicitly which means they assume that the model has access to the desired labeled or unlabeled target domain during training (Sharifi-Noghabi et al. 2020; Mourragui et al. 2019, 2020; Ma et al. 2021; Zhu et al. 2020; Warren et al. 2020; Peres da Silva, Suphavilai, and Nagarajan 2021).

However, in the real-world we do not have access to the target domain(s) during training the model on the source domain, e.g., we do not know future patients that may walk into a clinic. Nevertheless, the trained model should generalize to the target domain and be able to make predictions for samples encountered during the deployment time. Since generating large high-quality labeled preclinical datasets is an expensive and time-consuming process and we do not know response to a given drug in the target domain (e.g., future patients), there is a need for a computational method that takes not only labeled but also unlabeled source domain data as input and learns a representation that generalizes to a future target domain. This problem is known as out-of-distribution generalization or domain generalization, where the target domain is not accessible during training (Gulrajani and

Lopez-Paz 2020; J. Wang et al. 2021; Zhou et al. 2021). Out-of-distribution generalization is particularly important for biomedical applications (Zhang et al. 2021).
There are two main approaches to out-of-distribution generalization: 1) generalizing via learning domain-invariant features (J. Wang et al. 2021), and 2) generalizing via learning hypothesis-invariant features (Zhao et al. 2020; Z. Wang, Loog, and van Gemert 2021). In the first approach, the goal is to map the input domains to a shared feature space in which the features of all domains are aligned, i.e. look similar to each other. However, forcing different domains to have very similar features is not always feasible because different domains may have unique characteristics, and completely aligning them ignores these unique characteristics. The second approach does not align the features but rather the predictions across domains. The idea is that if the extracted features of input domains are similar enough for an accurate predictor to make similar predictions, forcing the features to be more similar is not required anymore. We note that there is no existing method for out-of-distribution generalization, for either of the two approaches, that can exploit both labeled and unlabeled source domains.

In this paper, we propose Velodrome, a deep neural network method that combines the two above approaches and exploits both labeled and unlabeled samples. Velodrome takes gene expression from cell line (labeled) and patient (unlabeled) datasets as input domains and predicts the drug response (measured as area above dose-response curve, AAC) via a shared (between cell lines and patients) feature extractor and domain-specific predictors. The feature extractor and the predictors are trained using a novel loss function with three components: 1) a standard supervised loss to make the features predictive of drug response, 2) a consistency loss to exploit unlabeled samples in learning the feature representation, and 3) an alignment loss to make the features generalizable. We designed the loss function to balance between learning domain-invariant and hypothesis-invariant features. To the best of our knowledge, Velodrome is the first method for semi-supervised out-of-distribution generalization from labeled cell lines and unlabeled patients to different preclinical and clinical datasets.
We evaluated the performance of Velodrome and state-of-the-art methods of supervised out-of-distribution generalization, domain adaptation, and semi-supervised learning in terms of a diverse range of metrics including Pearson and Spearman correlation, the Area Under the Receiver Operating Characteristic curve (AUROC), and the Area Under the Precision-Recall curve (AUPR). We observed that Velodrome achieved substantially better performance across different clinical and preclinical pharmacogenomics datasets for multiple drugs, demonstrating the potential of semi-supervised out-of-distribution generalization for drug response prediction, a crucial task of precision oncology. Moreover, we showed that the responses predicted by Velodrome for TCGA patients (unlabeled, i.e. without drug response) with prostate and kidney cancers had statistically significant associations with the expression values of the target genes of the studied drugs. This shows

that Velodrome captures biological aspects of drug response. Finally, although Velodrome was trained only on solid tissue types, we showed that it made accurate predictions for cell lines originating from non-solid tissue types, showcasing the out-of-distribution capabilities of the Velodrome model.

Results

Datasets. We employed the following resources throughout this paper:

1. Patients without drug response: more than 3,000 samples obtained from TCGA (Cancer Genome Atlas Research Network et al. 2013) breast (TCGA-BRCA), lung (TCGA-LUAD), pancreatic (TCGA-PAAD), kidney (TCGA-KIRC), and prostate (TCGA-PRAD) cohorts with RNA-seq data.
2. Cell lines with drug response: The Cancer Therapeutics Response Portal (CTRPv2) (Basu et al. 2013; Seashore-Ludlow et al. 2015), The Genomics of Drug Sensitivity in Cancer (GDSCv2) (Garnett et al. 2012; Iorio et al. 2016), and The Genentech Cell Line Screening Initiative (gCSI) (Haverty et al. 2016; Klijn et al. 2015) pan-cancer datasets with a total of more than 2000 samples with RNA-seq data and AAC as the measure of the drug response across 11 drugs (in common for the three datasets). We focused on the following drugs for this paper: Erlotinib, Docetaxel, Paclitaxel, and Gemcitabine.
3. PDX samples with drug response: PDX Encyclopedia (PDXE) dataset (Gao et al. 2015) is a collection of more than 300 PDX samples with RNA-seq data screened with 34 drugs. We use the reported measure of response in RECIST (Schwartz et al. 2016) for Gemcitabine, Erlotinib, and Paclitaxel obtained from supplementary material of (Gao et al. 2015).
4. Patients with drug response: 2 cancer-specific datasets with microarray data and RECIST as the measure of drug response for Docetaxel (Hatzis et al. 2011), Paclitaxel (Hatzis et al. 2011), and Erlotinib (Byers et al. 2013). Plus, a pan-cancer dataset obtained from TCGA patients treated with Gemcitabine (Ding, Zu, and Gu 2016). We use clinical annotations of the drug response for some patients which were obtained from supplementary material of (Ding, Zu, and Gu 2016).

Table S1 presents characteristics of these datasets and indicates whether they were used as source domain for training or target domain for test.

Velodrome Overview. The proposed Velodrome method takes gene expression and AAC of cell line datasets (CTRPv2 and GDSCv2) as well as gene expression of patients without drug response (TCGA dataset) and learns a predictive and generalizable representation. To achieve this, Velodrome employs a shared feature extractor, which takes the gene expression of CTRPv2 and GDSCv2 samples and maps them to a shared feature space, and

domain-specific predictors (e.g. one for CTRPv2 and one for GDSCv2), which take the feature representation of the gene expression and predict the drug response.
The parameters are optimized using a novel objective function consisting of three loss components. 1) a standard supervised loss to make the representation predictive of drug response, 2) a consistency loss to exploit unlabeled samples in learning the representation, and 3) an alignment loss to make the representation generalizable.
The idea of the standard supervised loss is to make the representation predictive of the drug response via a mean squared loss.
To incorporate unlabeled patient samples, we add a consistency loss. The idea is to first extract features from patient samples using the feature extractor and then assign pseudo-labels to them by utilizing the predictors associated with CTRPv2 and GDSCv2. The consistency loss takes the pseudo-labels (i.e., predictions) from the predictors and regularizes the parameters of the feature extractor and the predictors by the $l_2$distance between the predictions of CTRPv2 predictor and those of the GDSCv2 predictor.
Finally, to make the feature representation generalizable, we add an alignment loss that regularizes the parameters of the feature extractor. This alignment loss takes the extracted features of any two input domains (eg., CTRPv2 and TCGA or CTRPv2 and GDSCv2) and minimizes the difference between the covariance matrices of those domains.
Figure 1 illustrates the schematic overview of the Velodrome method.

Evaluation. Drug response prediction using multiple labeled and unlabeled domains can be viewed in three approaches: 1) under the assumption that there is no data discrepancy, it can be viewed as a semi-supervised learning problem, 2) under the assumption that unlabeled patient samples are proxies to future patients, it can be viewed as an unsupervised domain adaptation problem, and 3) under the assumption that a generalizable representation can be obtained via only labeled domains, it can be viewed as a supervised domain generalization problem. It is important to note that the main contribution of the Velodrome method is that it is the first semi-supervised domain generalization method for drug response prediction.
To evaluate the performance of Velodrome, we compared it against the state-of-the-art methods of each approach. For the first approach, we compared Velodrome to Mean Teacher (Tarvainen and Valpola 2017) which is the state-of-the-art deep neural network for semi-supervised learning (Yang and Xu 2020). For The second approach, we compared Velodrome to PRECISE as a non-deep learning method based on subspace alignment and (Saito et al. 2018) as a deep learning method based on adversarial domain adaptation via disagreement between predictors. Finally, for the third approach, we compared Velodrome to Ridge-ERM (Ridge Regression) as a non-deep learning baseline and DeepAll-ERM as a deep learning baseline. Both of them are categorized as methods of Empirical Risk Minimization (ERM). ERM methods achieve state-of-the-art performance for

out-of-distribution generalization. They are trained in a supervised fashion by merging all available labeled input domains.

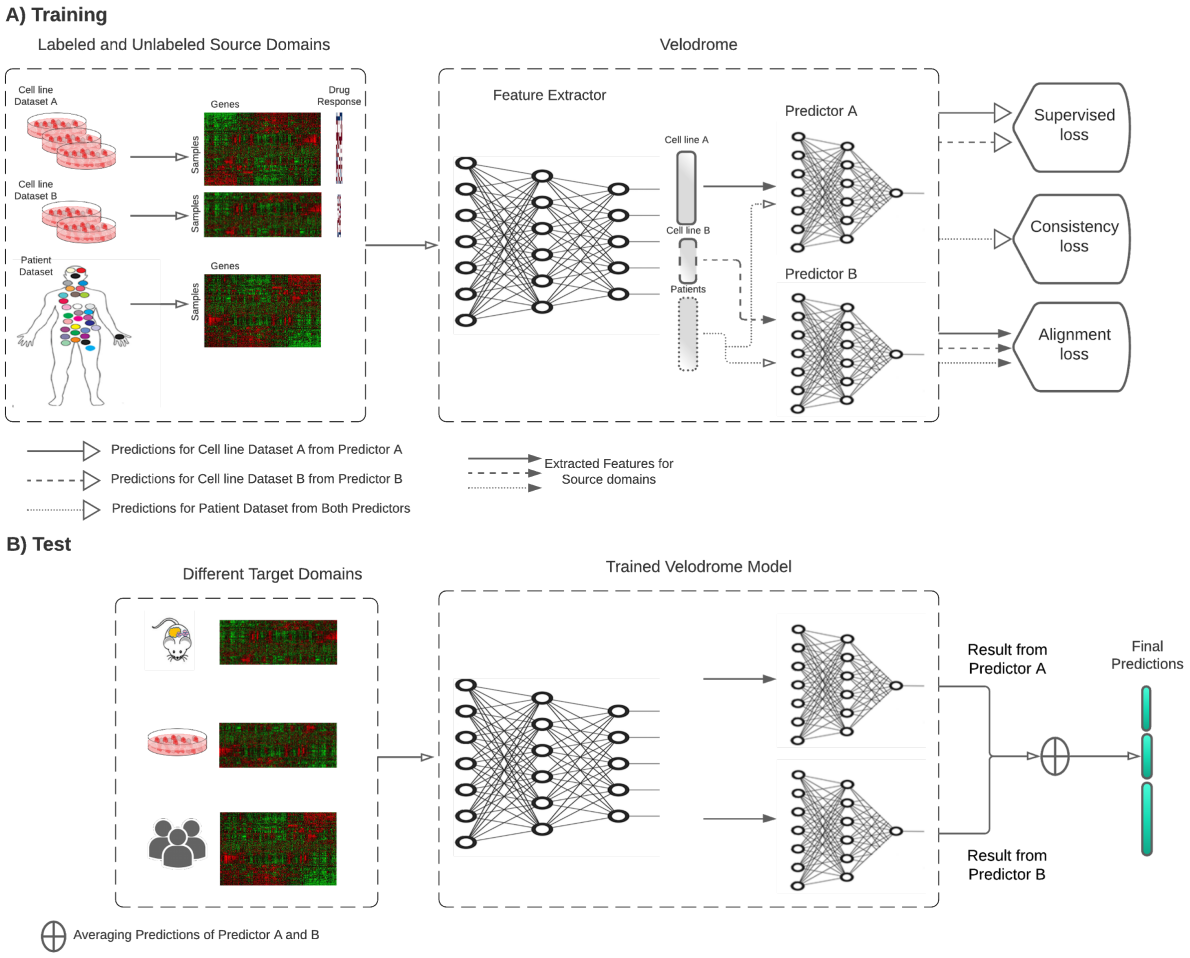

Figure 1 The schematic overview of the Velodrome method with three source domains (two labeled and one unlabeled). A) At training time, the feature extractor receives data from different source domains and extracts high-level abstract features. The extracted features of each labeled domain (cell line dataset) are input to the corresponding domain-specific predictor. Predictions are used to optimize the parameters of the predictors and the feature extractor via a standard supervised loss function. The extracted features of the unlabeled domain (patient dataset) are input to both predictors, and the predictions are used to optimize the parameters of predictors and the feature extractor via a consistency loss function. The extracted features of all source domains are used to optimize the parameters of the feature extractor via an alignment loss function.
B) At test time, the trained Velodrome model receives samples from different target domains, extracts features and makes predictions using the trained predictors. The predictions are then averaged to generate the final predictions for each sample.

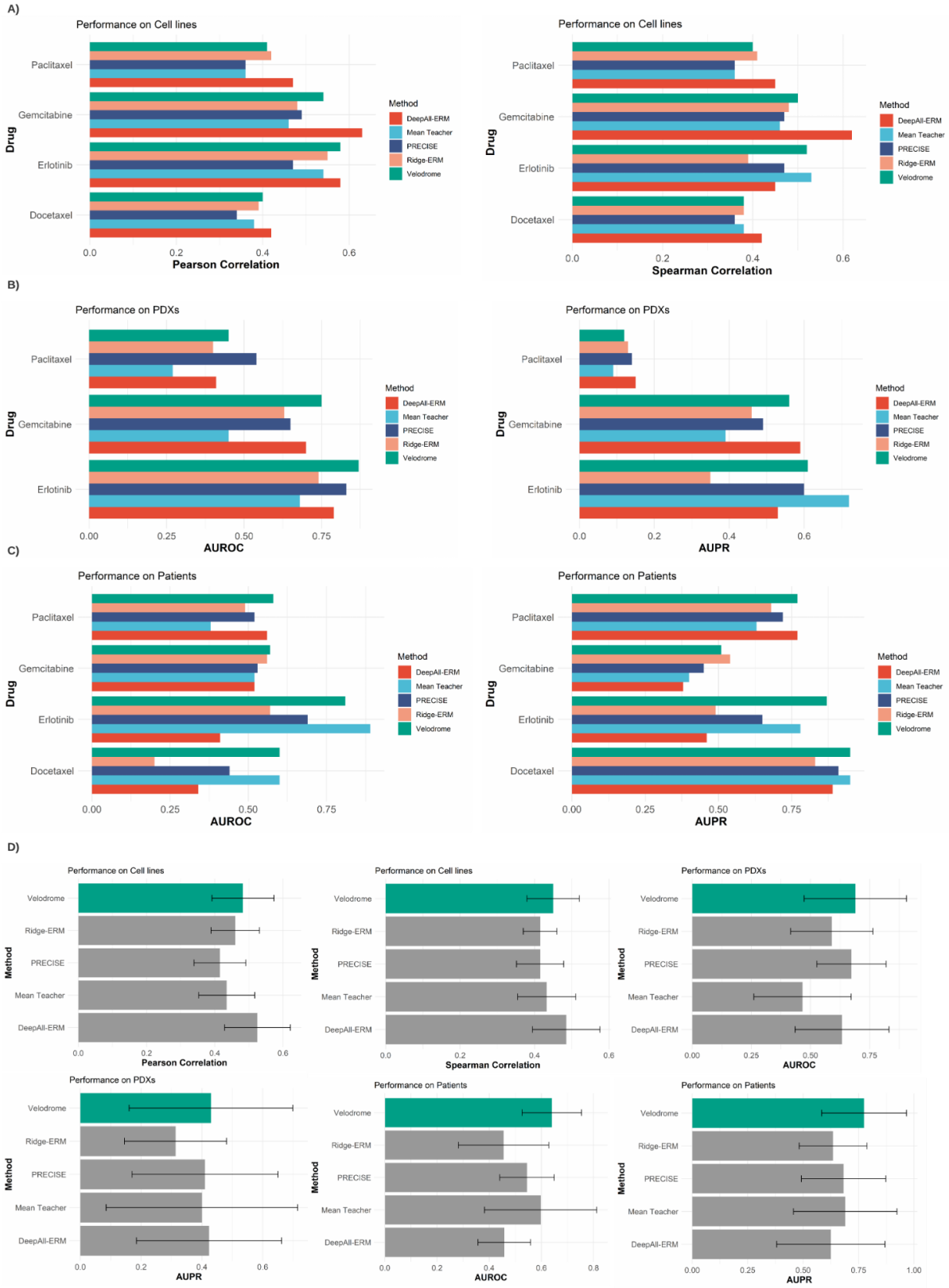
A)
Performance on Cell lines
Paclitaxel
Gemcitabine
Erlotinib
Docetaxel
Drug
Pearson Correlation
Spearman Correlation
Method
DeepAll-ERM
Mean Teacher
PRECISE
Ridge-ERM
Velodrome
B)
Performance on PDXs
AUROC
AUPR
C)
Performance on Patients
D)
Performance on Cell lines
Performance on PDXs
Performance on Patients
Pearson Correlation
Spearman Correlation
AUROC
AUPR
Method

Figure 2 Comparison of Velodrome and state-of-the-art of drug response prediction methods on cell lines in terms of Pearson and Spearman correlations (A), PDX models in terms of the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR) (B), and patients in terms of AUROC and AUPR (C). On average (D), Velodrome has the best or the second best performance on cell lines, PDX models, and patients compared to the baselines over the studied drugs.

Velodrome makes accurate predictions for cell lines. To investigate the generalization of Velodrome to other cell line datasets, we employed the gCSI dataset as the target domain and reported the performance of Velodrome and the baselines in terms of the Pearson and the Spearman correlation on this dataset. On average, DeepAll-ERM achieved the best performance (0.52 ± 0.09 for Pearson correlation coefficient and $0.48 \pm 0.09$for Spearman correlation coefficient - Figure 2D). Velodrome achieved the second best performance (0.48 ± 0.09 for Pearson correlation coefficient and $0.45 \pm 0.07$for Spearman correlation coefficient - Figure 2A and D). Ridge-ERM (0.46 ± 0.07- Figure 2A and D) and Mean Teacher (0.43 ± 0.07- Figure 2A and D) had the third best performance in terms of Pearson and Spearman correlation, respectively. These results indicate that although Velodrome is not the best performing model, it is fairly competitive on cell lines and generalizes well (Figure 2A).

Velodrome makes accurate predictions for PDXs samples. To investigate generalization of Velodrome to PDX samples, we employed the PDXE dataset as the target domain and reported the performance of Velodrome and the baselines discussed above in terms of the AUROC and the AUPR. On average, Velodrome achieved the best performance compared to the baselines (for 0.69 ± 0.21in AUROC and 0.43 ± 0.26in AUPR-Figure 2B and D). PRECISE obtained the second best performance in terms of AUROC (0.67 ± 0.14- Figure 2B and D) and DeepAll-ERM in terms of AUPR (0.42 ± 0.23- Figure 2B and D). Similarly, DeepAll had the third best performance in terms of AUROC (0.63 ± 0.19-Figure 2B and D) and PRECISE had the third best performance in terms of AUPR (0.41 ± 0.24-Figure 2B and D). These results indicate that utilizing both labeled and unlabeled samples from cell lines and patients improves drug response prediction on PDX samples.

Velodrome makes accurate predictions for patients. To investigate the generalization of Velodrome to patient samples, we employed the patient datasets obtained from clinical trials as target domains and reported the performance of Velodrome and the baselines discussed above in terms of AUROC and AUPR. On average, Velodrome achieved the best performance compared to the baselines and significantly outperformed them (0.64 ± 0.11in AUROC and 0.77 ± 0.19in AUPR-Figure 2C-D). Mean Teacher obtained the second best performance (0.59 ± 0.21in AUROC and 0.69 ± 0.23in AUPR-Figure 2C-D) and PRECISE had the third best performance (0.54 ± 0.1in AUROC and 0.68 ± 0.18in AUPR-Figure 2C-D). Interestingly, these three top-performing methods take for inputs both labelled and unlabeled samples, in contrast to other baselines considering only labeled

samples. These results indicate that incorporating unlabeled patient data along with labeled data significantly improves the generalization performance on patients. However, the results also demonstrate the advantage of learning features that are domain-invariant and hypothesis-invariant for out-of-distribution generalization, because the PRECISE method only ensures a domain-invariant representation.

Velodrome outperforms the baselines over multiple independent runs. To maximize the reproducibility, we utilized a fixed random seed for all methods (Velodrome and the baselines) and found the best settings for the hyper-parameters of each method with that seed. To investigate the performance of the best trained Velodrome model for each drug and those of the baselines, we re-trained all of the models from scratch using the same settings with 10 different random seeds and reported mean$\pm$std for each method (Figure 3A). Although we observed that the average performance (over the studied drugs) of all methods decreased, Velodrome still achieved the best performance on patients in terms of both AUROC and AUPR, and also the best performance in terms of both Pearson and Spearman correlation on cell lines. PRECISE and DeepAll-ERM obtained the best performance on PDX samples in terms of AUROC and AUPR, respectively (the performance of these two methods tied on AUPR). Velodrome had the third best performance in terms of AUROC and AUPR on PDX samples. Overall, these results indicate that Velodrome is more accurate and competitive compared to baselines particularly on patients and cell lines.

The complete version of Velodrome demonstrates the best performance. We performed an ablation study to investigate the impact of the different loss components of Velodrome separately. We studied three scenarios as follows: "*Velodrome w/o A*" represents a version of Velodrome without the alignment loss component, which means the neural network only uses supervised and semi-supervised losses. "*Velodrome w/o C*" represents a version of Velodrome without the consistency loss, which means the neural network only considers the supervised loss and the alignment loss. Finally, "*Velodrome w/o AC*" represents a version of Velodrome without both the alignment and the consistency loss, which means the neural network employed only has a standard supervised loss. Our results on patients demonstrate that on average (over 10 independent runs), the complete version of Velodrome outperforms its variants which indicates the added value of both alignment and consistency losses (Figure 3-B). Interestingly, removing the consistency loss from the objective function had the biggest impact on the Velodrome performance on patients. This may suggest that hypothesis alignment plays a more critical role than feature alignment for out-of-distribution generalization, which is compatible with recent observations in computer vision applications (Zhao et al. 2020).

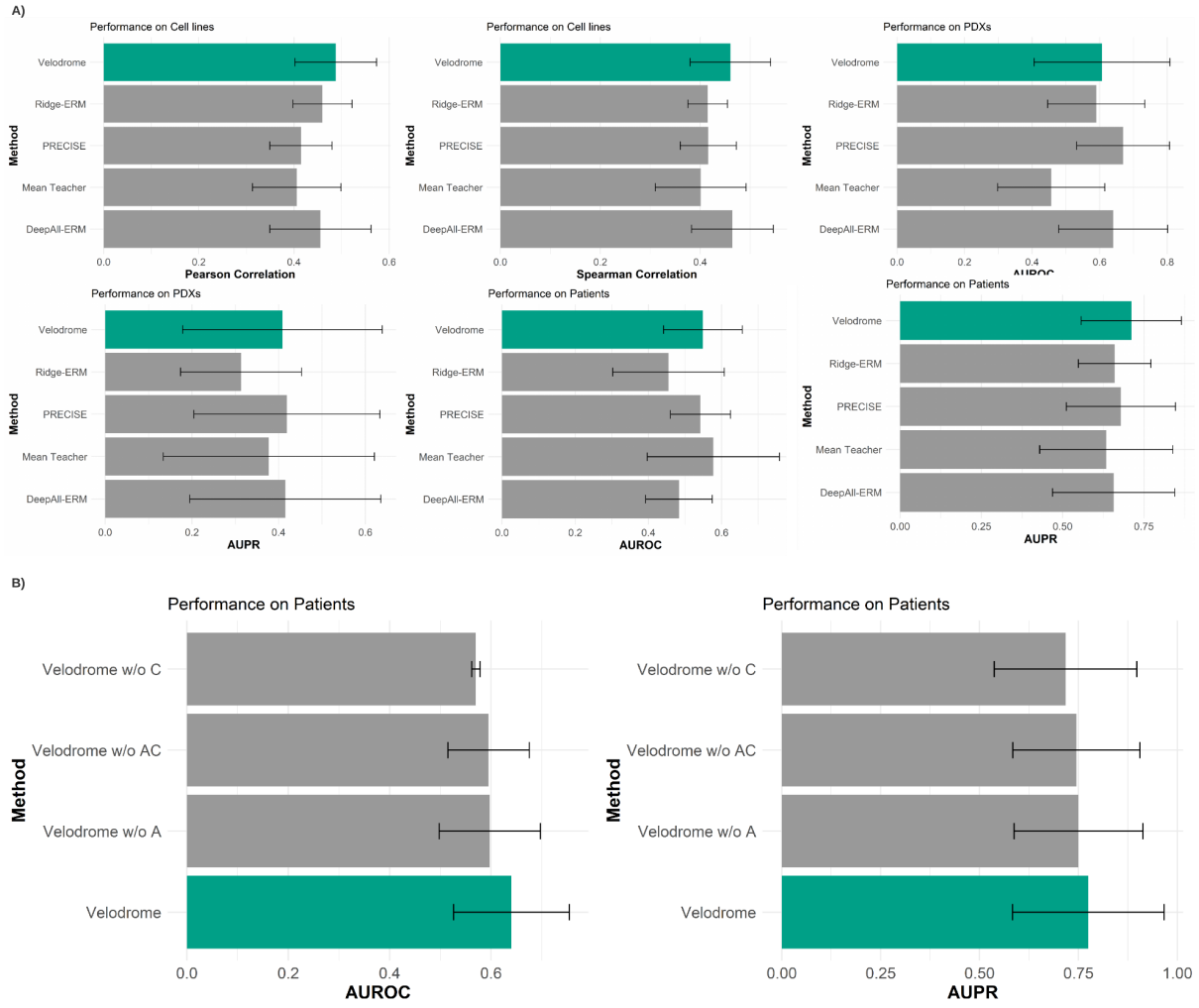

Figure 3 (A) Comparison of the average performance of the Velodrome and the state-of-the-art of drug response prediction methods over 10 independent runs . (B) An ablation study of the Velodrome performance on patients.

Velodrome generalizes to well-represented tissue types. To evaluate the performance of Velodrome on patients, we followed the experimental design of previous pharmacogenomics methods and designed an association study based on the known associated target genes for the investigated drugs (Geeleher, Cox, and Huang 2014; Mourragui et al. 2019; Sharifi-Noghabi et al. 2019, 2020). In this analysis, we employed the TCGA Kidney cancer cohort (TCGA-KIRC) as a tissue type well represented in our cell line datasets. In GDSCv2 and CTRPv2 combined, more than 3.3% of the samples originated from this tissue type (Figure S1).

We trained Velodrome models for each drug (Docetaxel, Erlotinib, Paclitaxel, and Gemcitabine) and applied them to the gene expression data of the patients of this cohort to predict their response. Then, we fit a linear regression model to the level of expression of

the known target genes of these drugs and the responses predicted by Velodrome. Based on the corrected p-values (two-tailed t-test) obtained from this multiple linear regression using the bonferroni correction method, there are a number of statistically significant associations between the target genes of the studied drugs and the responses predicted by Velodrome. For Docetaxel, MAP2 had a statistically significant association($P < 10^{-6}$). For Erlotinib, EGFR and ERBB2 had statistically significant associations (both $P < 10^{-6}$). For Paclitaxel BCL2 and MAP2 had significant associations (both $P < 10^{-6}$). Finally, for Gemcitabine, CMPK1 demonstrated a significant association($P < 10^{-6}$). These results suggest that the responses predicted by Velodrome are not random but capture biological aspects of the drug response.

Velodrome generalizes to under-represented tissue types. To further evaluate the performance of Velodrome, we performed a similar association study on the prostate cancer cohort in TCGA (TCGA-PRAD). We chose prostate because unlike kidney, prostate is a tissue type under-represented in our cell line datasets (only 0.3% of the samples originated from this tissue).
Similar to TCGA-KIRC, the Velodrome predictions for TCGA-PRAD patients demonstrated significant associations with known target genes of the studied drugs. For Docetaxel, MAP2 showed a statistically significant association ($P < 10^{-6}$). For Erlotinib, both EGFR and ERBB2 showed statistically significant associations (both $P < 10^{-6}$). For Paclitaxel, BCL2 ($P = 8 \times 10^{-6}$) and MAP2 ($P = 10^{-4}$) had significant associations. Finally, for Gemcitabine, CMPK1 demonstrated significant association ($P < 10^{-6}$). These results confirm again that the responses predicted by Velodrome are not random and they capture biological aspects of the drug response even for a tissue under-represented in the source domain.

Velodrome generalizes to new tissue types. Finally, we trained Velodrome and the baselines only on samples (cell lines and patients) that originated from solid tissue types because non-solid tissues such as haematopoietic and lymphoid have different molecular and pharmacological profiles compared to solid ones (Noghabi et al. 2021). Therefore, we wanted to examine the out-of-distribution capability of the Velodrome models on these tissue types that were completely absent during training. For that, we tested the trained Velodrome models for the studied drugs on samples originated from non-solid tissues in the gCSI cell line dataset and evaluated the performance in terms of Pearson correlation between the predictions and the actual AAC values and reported two-tailed p-value as well.
For Erlotinib and Gemcitabine, Velodrome demonstrated significant correlations of 0.4 ($P = 5 \times 10^{-3}$) and 0.39 ($P = 4 \times 10^{-3}$), respectively. For Docetaxel and Paclitaxel,

Velodrome did not make accurate predictions and had poor correlations of -0.07 and -0.02, respectively (both $P > 0.05$).

As a baseline to compare the Velodrome performance on non-solid tissues, we trained a Ridge Regression model on samples originated from non-solid tissues in CTRPv2 and GDSCv2 datasets and tested this predictor on non-solid samples of gCSI dataset. Therefore, we built a predictor specifically for non-solid samples and the performance of this model should act as an upper bound for the Velodrome. Similar to the Velodrome results, this predictor also achieved significant correlations of 0.34 ($P = 10^{-2}$) and 0.39 ($P = 5 \times 10^{-3}$) for Erlotinib and Gemcitabine and negative correlations of -0.11 ($P > 0.05$) and -0.4 ($P = 4 \times 10^{-3}$) for Docetaxel and Paclitaxel, respectively. These results suggest that Velodrome is as accurate (and even more accurate in the case of Erlotinib) as a non-solid predictor on these tissues even though it did not utilize them during training. The poor/negative correlation for Docetaxel and Paclitaxel may be dataset specific, particularly in the case of Paclitaxel where the non-solid predictor had a significant negative correlation, and requires further study.

Discussion

From the biological point of view, we found interesting connections between the known target genes of the studied drugs and the TCGA cohorts that we investigated (TCGA-PRAD and TCGA-KIRC). For example, BCL2 has known connections to prostate cancer progression (Chaudhary, Abel, and Lalani 1999; Catz and Johnson 2003) and survival (Renner et al. 2017). More importantly, the expression of BCL2 may have an antiapoptotic activity against androgen which is a key player in prostate cancer (Chaudhary, Abel, and Lalani 1999). Similarly, BCL2 can also act as an oncoprotein in kidney cancer (Paraf et al. 1995) and therapeutics roles (Adams and Cory 2007; Delbridge et al. 2016).

As another example, Microtubule-Associated Proteins including MAP2 have also been associated with different cancers including prostate (Bhat and Setaluri 2007) and kidney cancers (He et al. 2020). Moreover, Microtubule-targeting chemotherapy agents, Docetaxel and Paclitaxel, have been used in combination with anti-androgen therapeutics to increase the survival rate in prostate cancer patients (Martin, Kamelgarn, and Kyprianou 2014). Prostate cancer progression and lethal outcome have been associated with metabolic signaling pathways and CMPK1 (it mediates the mechanism of action for Gemcitabine) was shown to be highly expressed in prostate cancer patients (Kelly et al. 2016). A combination of Gemcitabine and other chemotherapy agents has shown to be effective for a subtype of kidney cancer (Numakura et al. 2014). Finally, EGFR and ERBB2 have been associated with different cancer types including prostate (El Sheikh et al. 2004; Pignon et al. 2009) and kidney (Reid et al. 2007) and they both showed therapeutic opportunities and increase in survival (Gordon et al. 2009).

From the computational point of view, it has been shown that methods of empirical risk minimization (ERM) are highly competitive for supervised domain generalization (Gulrajani and Lopez-Paz 2020). Therefore, it was also expected to see a competitive performance for a semi-supervised method (Mean Teacher) for the semi-supervised domain generalization setting. Moreover, Velodrome, PRECISE, and Mean Teacher were designed to take both labeled and unlabeled samples and, therefore, were expected to achieve better performance on patients than DeepAll-ERM and Ridge-ERM. On the other hand, these two methods achieved better performance on cell lines which makes sense since they were trained on cell lines.

We considered only TCGA-BRCA, TCGA-PAAD, and TCGA-LUAD for training, because these tissue types were well-represented in our cell line datasets (Figure S1) and because the four studied drugs were treatment options for these cell lines. This selection increases the relevancy of labeled (cell lines) and unlabeled (TCGA patients) data. Relevancy has been shown to improve semi-supervised learning performance even when both labeled and unlabeled datasets are imbalanced (Yang and Xu 2020), which is the case for drug response prediction.

Although methods of adversarial domain adaptation have shown great performance in different applications, especially computer vision (Ganin and Lempitsky 2015; Tzeng et al. 2017; Chen et al. 2017), we did not consider them as baselines because they were clearly outperformed by PRECISE (which we do use as baseline) in a recent study (Sharifi-Noghabi et al. 2020).

Although gene expression data has been shown many times to be the most effective genomic data type for drug response prediction (Iorio et al. 2016; Costello et al. 2014), in principle Velodrome can be extended to incorporate other omics data types. Especially promising are proteomics data (Ali et al. 2018) and germline variants (Menden et al. 2018), due to their predictive power. The advantage of proteomics is that it is closer to the phenotype and gene expression and protein abundance can be quite discordant. Velodrome can also be extended to incorporate additional information about the drug, such as the chemical representation, to improve the performance (Jiang et al. 2020). Finally, we did not discuss the explainability of the Velodrome model, but we note that the feature extractor of Velodrome can be replaced by a knowledge-based network (Snow et al. 2020) to offer explainability and transparency (Yu et al. 2018). A major limitation of our work is the output space discrepancy between cell lines, PDX samples, and patients, because on cell lines the drug response is measured based on the concentration of the drug but on PDX samples and patients the response is measured based on the change in the tumor volume after treatment. A recent method adjusts for this output space discrepancy and improves the prediction performance (Sharifi-Noghabi et al. 2020), but this method requires access to the target domain during training which violates the assumption of out-of-distribution generalization. In this work, we used AAC as the measure of drug response in cell line datasets and treated it as a score for making predictions for patients and PDX samples.

However, measuring AAC is dependent on the tested concentration range which generally differs between different pharmacogenomics studies. Recent efforts have demonstrated that adjusting concentration ranges across different datasets improves the prediction performance (Pozdeyev et al. 2016; Xia et al. 2021), however, we did not consider this adjustment because it reduces the sample size substantially.

Conclusion

In this paper, we proposed Velodrome, a transfer learning method for drug response prediction based on gene expression data. Velodrome is the first semi-supervised method of out-of-distribution generalization. We trained Velodrome on cell line datasets with drug response (measured in AAC) and patient datasets without drug response (i.e., unlabelled data) as source domains and successfully validated it on different target domains such as cell lines, PDX samples, and patient data across three chemotherapy agents and one targeted therapeutic. Our results suggest that Velodrome outperforms state-of-the-art methods of drug response prediction and transfer learning in terms of Pearson and Spearman correlations (on cell lines) and in terms of AUROC and AUPR (on PDX samples and patients). Moreover, we analyzed the biological significance of the predictions made by Velodrome and provided substantial evidence that these predictions have statistically significant associations with the expression level of numerous known target genes of the studied drugs in a tissue well-represented in our source domains, i.e. kidney cancer, and a tissue under-represented in our source domains, i.e. prostate cancer. Finally, we also demonstrated that Velodrome generalizes to new tissue types that were completely absent in the source domains. All these results demonstrate the superior out-of-distribution generalization capability of the Velodrome model and suggest that Velodrome may guide pharmacogenomics and precision oncology more accurately.

Methods

Data Preprocessing

We obtained all cell line datasets from the ORCESTRA platform (Mammoliti et al. 2020) which stores pharmacogenomics datasets in PharmacoSet (PSet) *R* objects. Samples with missing values were removed from both the gene expression and drug response data. The cell line datasets were generated via the same drug screening assay (CellTiter Glo) preprocessed using the PharmacoGx package (Smirnov et al. 2016) and are also comparable in terms of gene expression data which was preprocessed via Kallisto_0.46.1 (Bray et al. 2016). We also removed all the cell lines originating from non-solid tissue types from the cell line datasets.

We obtained the TCGA dataset via the Firehose (http://gdac.broadinstitute.org/) 28.01.2016. Expression values were converted to Transcripts Per Million (TPM) and log2-transformed. The PDX and clinical trial datasets were preprocessed similar to the approach described in (Sharifi-Noghabi et al. 2019).
For all of the employed datasets, all gene names were mapped to Entrez gene ids and the expression data were obtained before treatment and the response outcome after treatment. We reduced the number of genes to 2128 genes obtained from (Manica et al. 2019). After the preprocessing, all of the available datasets for each drug had the same number of genes (Table S1).

The Velodrome Method

We propose Velodrome, a method of drug response prediction using labeled and unlabeled source domains to build a predictive model that generalizes to unseen domains. To achieve this goal, Velodrome requires three different characteristics: 1) being predictive of drug response, 2) being generalizable to unseen domains, and 3) achieving these goals by taking both labeled and unlabeled data. We designed the objective function of Velodrome to meet these requirements by combining three loss functions: 1) a standard supervised loss based on labeled data to ensure that the model is predictive, 2) an alignment loss to ensure that the model has generalization capabilities, and 3) a consistency loss that exploits the unlabeled data.

Problem Definition: Following the notation of (Pan and Yang 2010), a domain D is defined by a raw input space $\mathbf{X}$, a probability distribution $p(X)$and a corresponding dataset $X = \{x_1, x_2,..., x_n\}$with $x_i \in \mathbf{X}$. A task $\mathrm{T} = \{Y, \mathcal{F}(.)\}$is associated with $D = \{X, p(X)\}$and is defined by a label space $Y \in \mathbf{Y}$ and a predictive function $\mathcal{F}(.)$ which is learned from training data $(X, Y) \in \mathbf{X} \times \mathbf{Y}$. In our case, $\mathbf{Y} \in [0, 1]$, which makes drug response prediction a regression problem.

Given multiple labeled and unlabeled source domains denoted by $D^L = \{D_i^l\}_{i=1}^{n_l}$ and $D^U = \{D_j^u\}_{j=1}^{n_u}$,the goal is to learn the predictive function $\mathcal{F}(.)$ which is implemented through a neural network.$\mathcal{F}(.)$consists of a shared (across all source domains) feature extractor $F_\theta(X)$ parameterized by θ, which maps X to latent features Z, and domain-specific predictors $G_{\phi_i}^i$ parametrized by ϕ_i, which takes Z_i (the extracted features ofD_i) as input and makes predictions (of the drug response) $\overline{Y}_i$for this source domain. θand$\{\phi_i\}_{i=1}^{n_l}$ are being optimized using an objective function

$J(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}) = l(D^L, \theta, \{\phi_i\}_{i=1}^{n_l}) + \Omega(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l})$,with a supervised loss$l(.)$ and some regularization terms $\Omega(.)$.

In drug response prediction, we have access to labeled source domains such as cell line datasets and unlabeled source domains such as cancer patients in TCGA . The goal is to learn a model that makes accurate predictions on patients, PDXes, or other cell lines as target domains that it may see during deployment. This is similar to out-of-distribution generalization (also known as domain generalization), where the goal is to optimize parameters of the model ($\theta$and $\{\phi_i\}_{i=1}^{n_l}$) in order to make the model generalizable and predictive of unseen domains. Out-of-distribution generalization assumes that there exists a d-dimensional latent feature space $Z \in R^d$that is invariant, predictive, and generalizable to seen and unseen domains on this given space.

Shared feature extractor: To map the raw input gene expression data to the latent space, Velodrome utilizes a feature extractor which is shared across all labeled and unlabeled source domains:

$$Z_i^j = F_\theta(X_i^j), j \in \{l, u\}, i \in D_i^j,$$

where, Z_i^jdenotes the features extracted by the feature extractor$F_\theta(.)$from X_i^j, the samples obtained from the i-th domain of type j (labeled or unlabeled). These extracted (latent) features will be provided as input to the domain-specific predictors.

Domain-specific predictors: To make predictions for the samples in the source domains, Velodrome utilizes n_l domain-specific predictors, meaning the number of domain-specific predictors that Velodrome utilizes is the same as the number of labeled source domains. These predictors are formulated as follows:

$$\overline{Y}_i^l = G_{\phi_i}^l(Z_i^l),$$

where, $\overline{Y}_i^l$denotes the predictions for the i-th labeled source domain obtained from predictor $G_{\phi_i}^l(.)$associated with the i-th labeled source domain and parameterized by ϕ_i. These predictions will be utilized to optimize the parameters of the feature extractor and the i-th predictor.

Supervised loss: To make the extracted latent features predictive of the drug response, Velodrome utilizes a standard supervised loss as follows:

$$l(D^l, \theta, \{\phi_i\}_{i=1}^{n_l}) = \frac{1}{n_l}\sum_{i=1}^{n_l} ||Y_i^l - \overline{Y}_i^l||_2^2,$$

where, $l(.)$denotes a standard supervised loss function in the form of the mean squared error (MSE). It is important to note that the parameters of the feature extractor are optimized by the total supervised loss but the parameters of the i-th predictor are optimized only by the supervised loss on predictions of the i-th predictor.

Alignment loss: Optimizing the parameters of the Velodrome model using only the supervised loss is likely to lead to overfitting to the labeled source domains. Therefore, we need an additional loss function to avoid overfitting to the source domains and to make the latent representation generalizable to unseen domains. To achieve this, Velodrome utilizes the CORAL loss function that regularizes the covariance matrices across input domains and has demonstrated state-of-the-art performance for learning invariant representations in computer vision applications (Sun and Saenko 2016; Gulrajani and Lopez-Paz 2020). The CORAL loss is defined as follows:

$$CORAL(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}) = \sum_{j=1}^{n_l}\sum_{i=1}^{n_u} ||C(Z_j^l) - C(Z_i^u)||_F^2,$$

where, C(.) is the covariance operator which receives the extracted features of a source domain and returns the covariance matrix of those features as follows:

$$C(Z) = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z}_i)(Z_i - \overline{Z}_i)^T,$$

where, n is the number of samples and $\overline{Z}$denotes the mean vector. Regularizing the covariance matrices across source domains ensures learning invariant feature vectors.
It is important to note that the objective function of Velodrome requires a combination of supervised and alignment loss because optimizing only the alignment loss is likely to lead to a trivial "zero" solution where all domains are mapped to the same point (Sun and Saenko 2016).

Consistency loss: Aligning the extracted features of the different domains imposes a strict constraint on learning an invariant latent representation because it disregards the unique domain-specific aspects of different source domains. To alleviate this, Velodrome utilizes a consistency loss to ensure that it learns a hypothesis invariant representation, i.e. predictions across source domains are similar when using different predictors. For example, if we have two predictors $G_{\phi_i}^l$ and $G_{\phi_j}^l$, we want them to generate similar predictions for the same unlabeled source domain. This consistency loss is defined as follows:

$$CON(D^u, \theta, \{\phi_i\}_{i=1}^{n_l}) = ||G^l_{\phi_i}(Z^u) - G^l_{\phi_j}(Z^u)||_2^2,$$

where, $Z^u$are extracted features for samples in a given unlabeled source domain and *MSE* is the mean squared error.

Objective function: Putting all of the loss functions together, the objective function of Velodrome is as follows:

$$J(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}) = l(D^L, \theta, \{\phi_i\}_{i=1}^{n_l}) + \lambda_1 CORAL(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}) + \lambda_2 CON(D^U, \theta, \{\phi_i\}_{i=1}^{n_l}).$$

where, $\lambda_1$and $\lambda_2 = 1 - \lambda_1$denote the regularization coefficients for the coral loss and consistency loss, respectively. Therefore, the function $\Omega(.)$that we defined in the problem definition is given by$\Omega(.) = \lambda_1 CORAL(.) + \lambda_2 CON(.)$. $\lambda_1$and $\lambda_2$control balancing between learning a domain-invariant representation and learning a hypothesis-invariant representation because the alignment loss ensures learning a domain-invariant representation, while the consistency loss ensures learning a hypothesis-invariant representation. The training steps of the Velodrome method are presented in Algorithm 1.

Algorithm 1: Velodrome

Input: Gene expression and AAC of multiple cell line datasets, gene expression of patients
Output: trained feature extractor and predictors of drug response

---

```
While the stopping condition is not reached:
    While the stopping condition is not reached:
        Sample a mini-batch from each cell line dataset
        Update feature extractor and predictors using supervised loss
    Sample a mini-batch from the patient dataset
    Sample a mini-batch from each cell line dataset
    Calculate the supervised loss
    Calculate the the covariance matrices and then  alignment loss
    Calculate the consistency loss
    Update feature extractor and predictors using all losses
```

Velodrome at test time: For a target sample x^t,Velodrome makes prediction as follows:

$$\overline{y^t} = \sum_i w_i G_{\phi_i}(F_\theta(x^t)),$$

where, w_i denotes the average supervised loss for the predictions of G_{ϕ_i}, normalized via a softmax function such that $\sum_i w_i = 1$. This means the final prediction will be a result of a weighted average of all predictors, and more accurate predictors will have higher weights.

Implementation Detail

Hyper-parameters: We considered a wide range of values for each hyper-parameter of the Velodrome model and optimized these values via a random search separately for each drug. The sets of values considered are as follows:

$Epoch = [10,\ 50,\ 100,\ 200]$, $Learning\ rate\ (LR) = [0.0001,\ 0.001,\ 0.01,\ 0.0005,\ 0.005,\ 0.05]$

$Dropout\ (DR) = [0.1,\ 0.3,\ 0.5,\ 0.8]$, $Weight\ Decay\ (WD) = [0.001,\ 0.0001,\ 0.01,\ 0.05,\ 0.005,\ 0.0$

$\lambda_1 = [1,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.01,\ 0.05,\ 0.001,\ 0.005,\ 0.0001,\ 0.0005]$

$Mini\ \ batch\ size\ (MB) = [17,\ 33,\ 65,\ 129]$

We considered separate learning rates and weight decays for the feature extractor and each predictor, but they all used the same sets of possible values.

We split the labeled cell line datasets (CTRPv2 and GDSCv2) into train and validation and considered 90% for train and 10% for validation. We merged the train splits into one training dataset and similarly, merged the validation splits into one validation set and used the merged validation set to optimize the values of these hyper-parameters.

Velodrome architecture: We followed previous works and designed predefined architectures (denoted by HD) for Velodrome (Noghabi et al. 2021; Sakellaropoulos et al. 2019). For the feature extractor, the first architecture has two hidden layers with the size 512×128, the second one has two layers with the size 256×256, the third one has three hidden layers with the size $128 \times 128 \times 128$ and the last architecture has four hidden layers with the size $64 \times 64 \times 64 \times 64$. We considered a batch normalization layer followed by an activation function (which we considered the Relu, the Tanh, and Sigmoid functions) as well as a dropout after the activation function for each hidden layer. The predictors have only one layer $HD \times 1$, where HD denotes the size of the last layer in the feature extractor. The final hyper-parameter and architecture of Velodrome for the studied drugs are as follows:

Drug: Epoch, MB, DR, WD1, WD2, WD3, HD, LR1, LR2, LR3, λ_1,λ_2

Docetaxel: 10, 65, 0.1, 0.05, 0.0005, 0.0001, 3, 0.001, 0.005, 0.0005, 0.2, 0.8

Gemcitabine: 10, 17, 0.1, 0.0001, 0.005, 0.01, 2, 0.01, 0.005, 0.05, 0.005, 0.99

Erlotinib: 50, 129, 0.1, 0.05, 0.005, 0.0005, 2, 0.001, 0.01, 0.001, 0.01, 0.99

Paclitaxel: 50, 129, 0.1, 0.005, 0.05, 0.005, 2, 0.05, 0.0005, 0.0001, 0.3, 0.7

WD1, WD2, and WD3 refers to the values we used for the feature extractor, predictor 1, and predictor 2, respectively (similar for LR1, LR2, and LR3).
For re-running and the ablation study of the trained models, we considered these random values for the random seed:
$Seed = [1, 21, 42, 84, 168, 336, 672, 1344, 2688, 5376]$.
We used 42 for the majority of the analyses in the paper (because it's the answer to life, the universe and everything!).
We used the same ranges for all of the baseline methods whenever using those values was applicable. For DeepAll-ERM and Ridge-ERM we used the existing implementations here: (https://github.com/bhklab/PGx_Guidelines), for PRECISE, we used the existing implementations here: (https://github.com/NKI-CCB/PRECISE). For Mean Teacher, we adopted an existing implementation for computer vision and modified it for this problem here: (https://github.com/CuriousAI/mean-teacher).
All of the deep neural network implementations were in the PyTorch framework and we employed the Adagrad optimizer to optimize the parameters of Velodrome as well as the baselines wherever applicable.

Performance evaluation: We employed the Scikit-learn and Scipy Python packages for the evaluation purposes. To be more specific, we utilized scikit-learn to calculate the AUROC and AUPR (for PDX samples and Patients) and we utilized the Scipy to calculate Pearson and Spearman correlations (for cell lines). For the association study, we utilized statsmodels.api Python package to fit the multiple linear regression and obtain the P-values and we obtained the list of known associated target genes for each drug by querying the PharmacoDB resource (Smirnov et al. 2018).

Data and Code Availability

All the final preprocessed data employed in this paper are publicly available here: https://zenodo.org/record/4793442#.YK1HVqhKiUk

All the codes, model objects, and supplementary material to run and reproduce our experimental results are publicly available here: https://github.com/hosseinshn/Velodrome

We also provided a conda environment to ensure version compatibility for future users.

Acknowledgement

We would like to thank Hossein Asghari (Ocean Genomics) and Shuman Peng (Simon Fraser University) for their support. We also would like to thank the Vancouver Prostate Centre

and Compute Canada (West Grid) for providing the computational resources for this research.

Funding

This work was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada (to M.E.), Canada Foundation for Innovation (33440 to C.C.C.), The Canadian Institutes of Health Research (PJT-153073 to C.C.C.), Terry Fox Foundation (201012TFF to C.C.C.), and The Terry Fox New Frontiers Program Project Grants (1062 to C.C.C.).

Authors' contributions

Study concept and design: H.S-N., M.E.
Deep learning design, implementations, and analysis: H.S-N.
Data preprocessing, analysis, and interpretation: H.S-N., O.Z.
Experiments: H.S-N., P.AH.
Analysis and interpretation of results: H.S-N., P.AH., O.Z.
Supervision: C.C.C., M.E.

Conflict of Interest

None declared.

References

Adams, J. M., and S. Cory. 2007. "The Bcl-2 Apoptotic Switch in Cancer Development and Therapy." *Oncogene* 26 (9): 1324–37.

Ali, Mehreen, Suleiman A. Khan, Krister Wennerberg, and Tero Aittokallio. 2018. "Global Proteomics Profiling Improves Drug Sensitivity Prediction: Results from a Multi-Omics, Pan-Cancer Modeling Approach." *Bioinformatics* 34 (8): 1353–62.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7.

Basu, Amrita, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Liu, Giannina I. Schaefer, Richard Y. Ebright, et al. 2013. "An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules." *Cell* 154 (5): 1151–61.

Bhat, Kumar M. R., and Vijayasaradhi Setaluri. 2007. "Microtubule-Associated Proteins as Targets in Cancer Chemotherapy." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 13 (10): 2849–54.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Erratum: Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (8): 888.

Byers, Lauren Averett, Lixia Diao, Jing Wang, Pierre Saintigny, Luc Girard, Michael Peyton, Li Shen, et al. 2013. "An Epithelial--Mesenchymal Transition Gene Signature Predicts Resistance to EGFR and PI3K Inhibitors and Identifies Axl as a Therapeutic Target for Overcoming EGFR Inhibitor Resistance." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19 (1): 279–90.

Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills

Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* 45 (10): 1113–20.

Catz, S. D., and J. L. Johnson. 2003. "BCL-2 in Prostate Cancer: A Minireview." *Apoptosis: An International Journal on Programmed Cell Death* 8 (1): 29–37.

Chaudhary, K. S., P. D. Abel, and E. N. Lalani. 1999. "Role of the Bcl-2 Gene Family in Prostate Cancer Progression and Its Implications for Therapeutic Intervention." *Environmental Health Perspectives* 107 Suppl 1 (February): 49–57.

Chen, Yi-Hsin, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. 2017. "No More Discrimination: Cross City Adaptation of Road Scene Segmenters." In *Proceedings of the IEEE International Conference on Computer Vision*, 1992–2001.

Costello, James C., NCI DREAM Community, Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, et al. 2014. "A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms." *Nature Biotechnology*. https://doi.org/10.1038/nbt.2877.

Cronin, Kathleen A., Andrew J. Lake, Susan Scott, Recinda L. Sherman, Anne-Michelle Noone, Nadia Howlader, S. Jane Henley, et al. 2018. "Annual Report to the Nation on the Status of Cancer, Part I: National Cancer Statistics." *Cancer* 124 (13): 2785–2800.

Delbridge, Alex R. D., Stephanie Grabow, Andreas Strasser, and David L. Vaux. 2016. "Thirty Years of BCL-2: Translating Cell Death Discoveries into Novel Cancer Therapies." *Nature Reviews. Cancer* 16 (2): 99–109.

Ding, Zijian, Songpeng Zu, and Jin Gu. 2016. "Evaluating the Molecule-Based Prediction of Clinical Drug Responses in Cancer." *Bioinformatics* 32 (19): 2891–95.

El Sheikh, Soha Salama, Jan Domin, Paul Abel, Gordon Stamp, and El-Nasir Lalani. 2004. "Phosphorylation of Both EGFR and ErbB2 Is a Reliable Predictor of Prostate Cancer Cell Proliferation in Response to EGF." *Neoplasia* 6 (6): 846–53.

Ganin, Yaroslav, and Victor Lempitsky. 2015. "Unsupervised Domain Adaptation by Backpropagation." In *Proceedings of the 32nd International Conference on Machine Learning*, edited by Francis Bach and David Blei, 37:1180–89. Proceedings of Machine Learning Research. Lille, France: PMLR.

Gao, Hui, Joshua M. Korn, Stéphane Ferretti, John E. Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, et al. 2015. "High-Throughput Screening Using Patient-Derived Tumor Xenografts to Predict Clinical Trial Drug Response." *Nature Medicine* 21 (11): 1318–25.

Garnett, Mathew J., Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, et al. 2012. "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells." *Nature* 483 (7391): 570–75.

Garraway, Levi A., Jaap Verweij, and Karla V. Ballman. 2013. "Precision Oncology: An Overview." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 31 (15): 1803–5.

Geeleher, Paul, Nancy J. Cox, and R. Stephanie Huang. 2014. "Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and in Vitro Drug Sensitivity in Cell Lines." *Genome Biology* 15 (3): R47.

Geeleher, Paul, Eric R. Gamazon, Cathal Seoighe, Nancy J. Cox, and R. Stephanie Huang. 2016. "Consistency in Large Pharmacogenomic Studies." *Nature* 540 (7631): E1–2.

Gordon, Michael S., Michael Hussey, Raymond B. Nagle, Primo N. Lara Jr, Philip C. Mack, Janice Dutcher, Wolfram Samlowski, et al. 2009. "Phase II Study of Erlotinib in Patients with Locally Advanced or Metastatic Papillary Histology Renal Cell Cancer: SWOG S0317." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (34): 5788–93.

Gulrajani, Ishaan, and David Lopez-Paz. 2020. "In Search of Lost Domain Generalization." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.01434.

Haibe-Kains, Benjamin, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C. Jin, Andrew H. Beck, Hugo J. W. L. Aerts, and John Quackenbush. 2013. "Inconsistency in Large Pharmacogenomic Studies." *Nature* 504 (7480): 389–93.

Hatzis, Christos, Lajos Pusztai, Vicente Valero, Daniel J. Booser, Laura Esserman, Ana Lluch, Tatiana Vidaurre, et al. 2011. "A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer." *JAMA: The Journal of the American Medical Association* 305 (18): 1873–81.

Haverty, Peter M., Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M. Neve, et al. 2016. "Reproducible Pharmacogenomic Profiling of Cancer Cell Line Panels." *Nature* 533 (7603): 333–37.

He, Zhaoyue, He Liu, Holger Moch, and Hans-Uwe Simon. 2020. "Machine Learning with Autophagy-Related Proteins for Discriminating Renal Cell Carcinoma Subtypes." *Scientific Reports* 10 (1): 720.

Holmberg, Olle G., Niklas D. Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, et al. 2020. "Self-Supervised Retinal Thickness Prediction Enables Deep Learning from Unlabelled Data to Boost Classification of Diabetic Retinopathy." *Nature Machine Intelligence* 2 (11): 719–26.

Hu, Jian, Xiangjie Li, Gang Hu, Yafei Lyu, Katalin Susztak, and Mingyao Li. 2020. "Iterative Transfer Learning with Neural Network for Clustering and Cell Type Classification in Single-Cell RNA-Seq Analysis." *Nature Machine Intelligence* 2 (10): 607–18.

Iorio, Francesco, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, et al. 2016. "A Landscape of Pharmacogenomic Interactions in Cancer." *Cell* 166 (3): 740–54.

Jiang, Yuepeng, Stefano Rensi, Sheng Wang, and Russ B. Altman. 2020. "DrugOrchestra: Jointly Predicting Drug Response, Targets, and Side Effects via Deep Multi-Task Learning."

Kelly, Rachel S., Jennifer A. Sinnott, Jennifer R. Rider, Ericka M. Ebot, Travis Gerke, Michaela Bowden, Andreas Pettersson, et al. 2016. "The Role of Tumor Metabolism as a Driver of Prostate Cancer Progression and Lethal Disease: Results from a Nested Case-Control Study." *Cancer & Metabolism* 4 (December): 22.

Klijn, Christiaan, Steffen Durinck, Eric W. Stawiski, Peter M. Haverty, Zhaoshi Jiang, Hanbin Liu, Jeremiah Degenhardt, et al. 2015. "A Comprehensive Transcriptional Portrait of Human Cancer Cell Lines." *Nature Biotechnology* 33 (3): 306–12.

Kuenzi, Brent M., Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. 2020. "Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells." *Cancer Cell* 38 (5): 672–84.e6.

Ma, Jianzhu, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F. A. Wessels, et al. 2021. "Few-Shot Learning Creates Predictive Models of Drug Response That Translate from High-Throughput Screens to Individual Patients." *Nature Cancer*, January. https://doi.org/10.1038/s43018-020-00169-2.

Mammoliti, Anthony, Petr Smirnov, Minoru Nakano, Zhaleh Safikhani, Chantal Ho, Gangesh Beri, and Benjamin Haibe-Kains. 2020. "ORCESTRA: A Platform for Orchestrating and Sharing High-Throughput Pharmacogenomic Analyses." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.09.18.303842.

Manica, Matteo, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and María Rodríguez Martínez. 2019. "Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders." *Molecular Pharmaceutics* 16 (12): 4797–4806.

Marquart, John, Emerson Y. Chen, and Vinay Prasad. 2018. "Estimation of the Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology." *JAMA Oncology* 4 (8): 1093–98.

Martin, Sarah K., Marisa Kamelgarn, and Natasha Kyprianou. 2014. "Cytoskeleton Targeting Value in Prostate Cancer Treatment." *American Journal of Clinical and Experimental Urology* 2 (1): 15–26.

Menden, Michael P., Francesco Paolo Casale, Johannes Stephan, Graham R. Bignell, Francesco Iorio, Ultan McDermott, Mathew J. Garnett, Julio Saez-Rodriguez, and Oliver Stegle. 2018. "The Germline Genetic Component of Drug Sensitivity in Cancer Cell Lines." *Nature Communications* 9 (1): 3385.

Mourragui, Soufiane, Marco Loog, Daniel J. Vis, Kat Moore, Anna Gonzalez Manjon, Mark A. van de Wiel, Marcel J. T. Reinders, and Lodewyk F. A. Wessels. 2020. "PRECISE+ Predicts Drug Response in Patients by Non-Linear Subspace-Based Transfer from Cell Lines and PDX Models." *bioRxiv*.

Mourragui, Soufiane, Marco Loog, Mark A. van de Wiel, Marcel J. T. Reinders, and Lodewyk F. A. Wessels. 2019. "PRECISE: A Domain Adaptation Approach to Transfer Predictors of Drug Response from Pre-Clinical Models to Tumors." *Bioinformatics* 35 (14): i510–19.

Mpindi, John Patrick, Bhagwan Yadav, Päivi Östling, Prson Gautam, Disha Malani, Astrid Murumägi, Akira Hirasawa, et al. 2016. "Consistency in Drug Response Profiling." *Nature* 540 (7631): E5–6.

Neyshabur, B., H. Sedghi, and C. Zhang. 2020. "What Is Being Transferred in Transfer Learning?" *Advances in Neural Information Processing Systems*. https://papers.nips.cc/paper/2020/file/0607f4c705595b911a4f3e7a127b44e0-Paper.pdf.

Noghabi, Hossein Sharifi, Soheil Jahangiri-Tazehkand, Casey Hon, Petr Smirnov, Anthony Mammoliti, Sisira Kadambat Nair, Arvind Singh Mer, Martin Ester, and Benjamin Haibe-Kains. 2021. "Drug Sensitivity Prediction From Cell Line-Based Pharmacogenomics Data: Guidelines for Developing Machine Learning Models." *bioRxiv*.

Numakura, Kazuyuki, Norihiko Tsuchiya, Susumu Akihama, Takamitsu Inoue, Shintaro Narita, Mingguo Huang, Shigeru Satoh, and Tomonori Habuchi. 2014. "Successful Mammalian Target of Rapamycin Inhibitor Maintenance Therapy Following Induction Chemotherapy with Gemcitabine and Doxorubicin for

Metastatic Sarcomatoid Renal Cell Carcinoma." *Oncology Letters* 8 (1): 464–66.
Pal, Sumanta K., Michael J. Miller, Neeraj Agarwal, Susan Marina Chang, Mariana Chavez-MacGregor, Ezra Cohen, Suzanne Cole, et al. 2019. "Clinical Cancer Advances 2019: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 37 (10): 834–49.
Pan, Sinno Jialin, and Qiang Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59.
Paraf, F., J. Gogusev, Y. Chrétien, and D. Droz. 1995. "Expression of Bcl-2 Oncoprotein in Renal Cell Tumours." *The Journal of Pathology* 177 (3): 247–52.
Peres da Silva, Rafael, Chayaporn Suphavilai, and Niranjan Nagarajan. 2021. "TUGDA: Task Uncertainty Guided Domain Adaptation for Robust Generalization of Cancer Drug Response Prediction from in Vitro to in Vivo Settings." *Bioinformatics* , May. https://doi.org/10.1093/bioinformatics/btab299.
Pignon, Jean-Christophe, Benjamin Koopmansch, Gregory Nolens, Laurence Delacroix, David Waltregny, and Rosita Winkler. 2009. "Androgen Receptor Controls EGFR and ERBB2 Gene Expression at Different Levels in Prostate Cancer Cell Lines." *Cancer Research* 69 (7): 2941–49.
Pozdeyev, Nikita, Minjae Yoo, Ryan Mackie, Rebecca E. Schweppe, Aik Choon Tan, and Bryan R. Haugen. 2016. "Integrating Heterogeneous Drug Sensitivity Data from Cancer Pharmacogenomic Studies." *Oncotarget* 7 (32): 51619–25.
Raghu, Maithra, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. "Transfusion: Understanding Transfer Learning for Medical Imaging." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, 32:3347–57. Curran Associates, Inc.
Reid, Alison, Laura Vidal, Heather Shaw, and Johann de Bono. 2007. "Dual Inhibition of ErbB1 (EGFR/HER1) and ErbB2 (HER2/neu)." *European Journal of Cancer* 43 (3): 481–89.
Renner, Wilfried, Uwe Langsenlehner, Sabine Krenn-Pilko, Petra Eder, and Tanja Langsenlehner. 2017. "BCL2 Genotypes and Prostate Cancer Survival." *Strahlentherapie Und Onkologie: Organ Der Deutschen Rontgengesellschaft ... [et Al]* 193 (6): 466–71.
Safikhani, Zhaleh, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, et al. 2016. "Revisiting Inconsistency in Large Pharmacogenomic Studies." *F1000Research* 5 (September): 2333.
Saito, Kuniaki, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2018.00392.
Sakellaropoulos, Theodore, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, Tyler J. Moss, et al. 2019. "A Deep Learning Framework for Predicting Response to Therapy in Cancer." *Cell Reports* 29 (11): 3367–73.e4.
Schwartz, Lawrence H., Saskia Litière, Elisabeth de Vries, Robert Ford, Stephen Gwyther, Sumithra Mandrekar, Lalitha Shankar, et al. 2016. "RECIST 1.1-Update and Clarification: From the RECIST Committee." *European Journal of Cancer* 62 (July): 132–37.
Seashore-Ludlow, Brinton, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, et al. 2015. "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset." *Cancer Discovery* 5 (11): 1210–23.
Sharifi-Noghabi, Hossein, Shuman Peng, Olga Zolotareva, Colin C. Collins, and Martin Ester. 2020. "AITL: Adversarial Inductive Transfer Learning with Input and Output Space Adaptation for Pharmacogenomics." *Bioinformatics* 36 (Supplement_1): i380–88.
Sharifi-Noghabi, Hossein, Olga Zolotareva, Colin C. Collins, and Martin Ester. 2019. "MOLI: Multi-Omics Late Integration with Deep Neural Networks for Drug Response Prediction." *Bioinformatics* 35 (14): i501–9.
Smirnov, Petr, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-Alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. 2018. "PharmacoDB: An Integrative Database for Mining in Vitro Anticancer Drug Screening Studies." *Nucleic Acids Research* 46 (D1): D994–1002.
Smirnov, Petr, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, et al. 2016. "PharmacoGx: An R Package for Analysis of Large Pharmacogenomic Datasets." *Bioinformatics* 32 (8): 1244–46.
Snow, Oliver, Hossein Sharifi-Noghabi, Jialin Lu, Olga Zolotareva, Mark Lee, and Martin Ester. 2020. "BDKANN -

Biological Domain Knowledge-Based Artificial Neural Network for Drug Response Prediction." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/840553.
Sun, Baochen, and Kate Saenko. 2016. "Deep CORAL: Correlation Alignment for Deep Domain Adaptation." In *Computer Vision – ECCV 2016 Workshops*, 443–50. Springer International Publishing.
Tarvainen, Antti, and Harri Valpola. 2017. "Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results." *arXiv [cs.NE]*. arXiv. http://arxiv.org/abs/1703.01780.
Tzeng, Eric, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. "Adversarial Discriminative Domain Adaptation." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2017.316.
Wang, Jindong, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. 2021. "Generalizing to Unseen Domains: A Survey on Domain Generalization." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2103.03097.
Wang, Ziqi, Marco Loog, and Jan van Gemert. 2021. "Respecting Domain Relations: Hypothesis Invariance for Domain Generalization." In *2020 25th International Conference on Pattern Recognition (ICPR)*, 9756–63.
Warren, Allison, Andrew Jones, Tsukasa Shibue, William C. Hahn, Jesse S. Boehm, Francisca Vazquez, Aviad Tsherniak, and James M. McFarland. 2020. "Global Computational Alignment of Tumor and Cell Line Transcriptional Profiles." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.03.25.008342.
Xia, Fangfang, Jonathan Allen, Prasanna Balaprakash, Thomas Brettin, Cristina Garcia-Cardona, Austin Clyde, Judith Cohn, et al. 2021. "A Cross-Study Analysis of Drug Response Prediction in Cancer Cell Lines." *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/2104.08961.
Yang, Yuzhe, and Zhi Xu. 2020. "Rethinking the Value of Labels for Improving Class-Imbalanced Learning." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2006.07529.
Yu, Michael K., Jianzhu Ma, Jasmin Fisher, Jason F. Kreisberg, Benjamin J. Raphael, and Trey Ideker. 2018. "Visible Machine Learning for Biomedicine." *Cell* 173 (7): 1562–65.
Zhang, Haoran, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. 2021. "An Empirical Framework for Domain Generalization in Clinical Settings." In *Proceedings of the Conference on Health, Inference, and Learning*. New York, NY, USA: ACM. https://doi.org/10.1145/3450439.3451878.
Zhao, Shanshan, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. 2020. "Domain Generalization via Entropy Regularization." *Advances in Neural Information Processing Systems* 33.
Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. "Domain Generalization: A Survey." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2103.02503.
Zhu, Yitan, Thomas Brettin, Yvonne A. Evrard, Alexander Partin, Fangfang Xia, Maulik Shukla, Hyunseung Yoo, James H. Doroshow, and Rick L. Stevens. 2020. "Ensemble Transfer Learning for the Prediction of Anti-Cancer Drug Response." *Scientific Reports* 10 (1): 18040.

Supplementary Material

Table S1 Characteristics of the employed datasets

Dataset	Drug	Type	Domain	Label	Tissue	#Samples	#Genes
CTRPv2	Docetaxel	Cell line	Source	AAC	Solid	292	1453
GDSCv2	Docetaxel	Cell line	Source	AAC	Solid	234	1453
TCGA-LUAD	Docetaxel	Patient	Source	Unlabeled	Solid	507	1453
TCGA-BRCA	Docetaxel	Patient	Source	Unlabeled	Solid	1051	1453
TCGA-PAAD	Docetaxel	Patient	Source	Unlabeled	Solid	131	1453
gCSI	Docetaxel	Cell line	Target	AAC	Solid	280	1453
GSE25065D	Docetaxel	Patient	Target	RECIST	Solid	51	1453
CTRPv2	Gemcitabine	Cell line	Source	AAC	Solid	514	2080
GDSCv2	Gemcitabine	Cell line	Source	AAC	Solid	226	2080
TCGA-LUAD	Gemcitabine	Patient	Source	Unlabeled	Solid	507	2080
TCGA-BRCA	Gemcitabine	Patient	Source	Unlabeled	Solid	1051	2080
TCGA-PAAD	Gemcitabine	Patient	Source	Unlabeled	Solid	131	2080
gCSI	Gemcitabine	Cell line	Target	AAC	Solid	277	2080
TCGA-Gem	Gemcitabine	Patient	Target	RECIST	Solid	66	2080
PDXE	Gemcitabine	PDX	Target	RECIST	Solid	25	2080
CTRPv2	Erlotinib	Cell	Source	AAC	Solid	607	2066

		line					
GDSCv2	Erlotinib	Cell line	Source	AAC	Solid	230	2066
TCGA-LUAD	Erlotinib	Patient	Source	Unlabeled	Solid	507	2066
TCGA-BRCA	Erlotinib	Patient	Source	Unlabeled	Solid	1051	2066
TCGA-PAAD	Erlotinib	Patient	Source	Unlabeled	Solid	131	2066
gCSI	Erlotinib	Cell line	Target	AAC	Solid	283	2066
GSE33072	Erlotinib	Patient	Target	RECIST	Solid	25	2066
PDXE	Erlotinib	PDX	Target	RECIST	Solid	21	2066
CTRPv2	Paclitaxel	Cell line	Source	AAC	Solid	445	1452
GDSCv2	Paclitaxel	Cell line	Source	AAC	Solid	230	1452
TCGA-LUAD	Paclitaxel	Patient	Source	Unlabeled	Solid	507	1452
TCGA-BRCA	Paclitaxel	Patient	Source	Unlabeled	Solid	1051	1452
TCGA-PAAD	Paclitaxel	Patient	Source	Unlabeled	Solid	131	1452
gCSI	Paclitaxel	Cell line	Target	AAC	Solid	284	1452
GSE25065P	Paclitaxel	Patient	Target	RECIST	Solid	84	1452
PDXE	Paclitaxel	PDX	Target	RECIST	Solid	43	1452
TCGA-PRAD	Docetaxel	Patient	Target	Unlabeled	Solid	498	1453
TCGA-KIRC	Docetaxel	Patient	Target	Unlabeled	Solid	534	1453
TCGA-PRAD	Paclitaxel	Patient	Target	Unlabeled	Solid	498	1452
TCGA-KIRC	Paclitaxel	Patient	Target	Unlabeled	Solid	534	1452

TCGA-PRAD	Gemcitabine	Patient	Target	Unlabeled	Solid	498	2080
TCGA-KIRC	Gemcitabine	Patient	Target	Unlabeled	Solid	534	2080
TCGA-PRAD	Erlotinib	Patient	Target	Unlabeled	Solid	498	2066
TCGA-KIRC	Erlotinib	Patient	Target	Unlabeled	Solid	534	2066
CTRPv2	Docetaxel	Cell line	Source	AAC	Non-solid	62	1453
GDSCv2	Docetaxel	Cell line	Source	AAC	Non-solid	69	1453
gCSI	Docetaxel	Cell line	Target	AAC	Non-solid	50	1453
CTRPv2	Paclitaxel	Cell line	Source	AAC	Non-solid	100	1452
GDSCv2	Paclitaxel	Cell line	Source	AAC	Non-solid	67	1452
gCSI	Paclitaxel	Cell line	Target	AAC	Non-solid	50	1452
CTRPv2	Gemcitabine	Cell line	Source	AAC	Non-solid	129	2080
GDSCv2	Gemcitabine	Cell line	Source	AAC	Non-solid	69	2080
gCSI	Gemcitabine	Cell line	Target	AAC	Non-solid	50	2080
CTRPv2	Erlotinib	Cell line	Source	AAC	Non-solid	135	2066
GDSCv2	Erlotinib	Cell line	Source	AAC	Non-solid	68	2066
gCSI	Erlotinib	Cell line	Target	AAC	Non-solid	49	2066

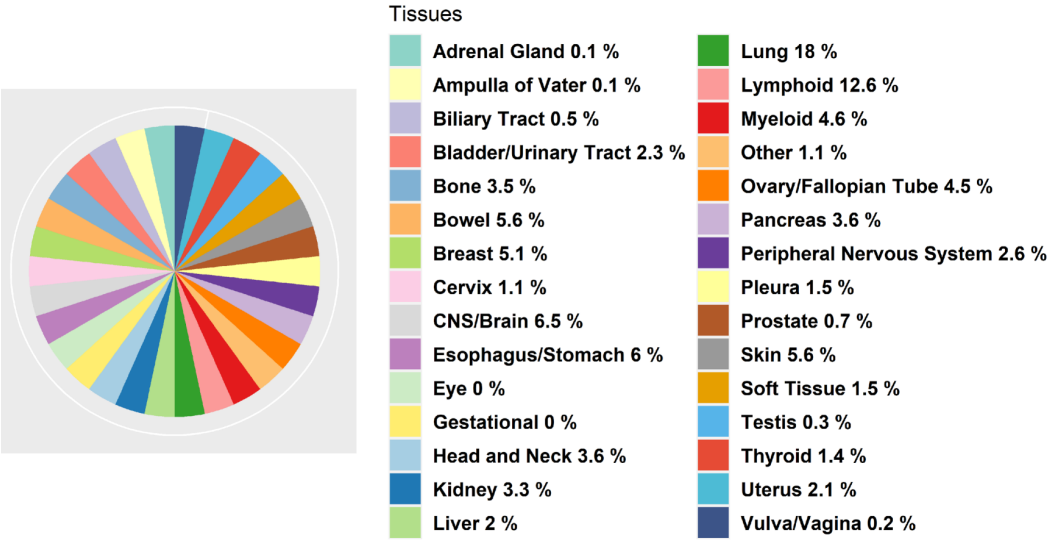

Figure S1 The percentage of tissue types in CTRPv2 and GDSCv2 cell line datasets combined.