ARTICLES
https://doi.org/10.1038/s42256-021-00408-w

nature machine intelligence

Check for updates

Out-of-distribution generalization from labelled and unlabelled 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] and Martin Ester[1,2] ✉

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 such data discrepancy in drug response prediction for different cancers. These methods generally use cell lines as source domains, and patients, patient-derived xenografts or other cell lines as target domains; however, it is assumed that the methods have access to the target domain during training or fine-tuning, and they can only take labelled source domains as input. The former is a strong assumption that is not satisfied during deployment of these models in the clinic, whereas the latter means these methods rely on labelled source domains that are of limited size. To avoid these assumptions, we formulate drug response prediction in cancer as an out-of-distribution generalization problem, which does not assume that the target domain is accessible during training. Moreover, to exploit unlabelled source domain data—which tends to be much more plentiful than labelled data—we adopt a semi-supervised approach. We propose Velodrome, a semi-supervised method of out-of-distribution generalization that takes labelled and unlabelled 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 unlabelled 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. Finally, we showed that Velodrome models generalize to different tissue types that were well-represented, under-represented or completely absent in the training data. Overall, our results suggest that Velodrome may guide precision oncology more accurately.

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[1,2]. Unfortunately, patient datasets with drug response are often small or not publicly available, motivating the creation of large-scale preclinical resources such as patient-derived xenografts (PDX)[3] or cancer cell lines[4–10] 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 tumour microenvironment/the immune system[11,12]. This has been a source of discussions in the community showing discrepancy between preclinical resources[13] and providing evidence regarding consistency between them[10,14,15].

Transfer learning has emerged as a machine learning paradigm for such scenarios[16,17], in which 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 domains). It has been employed for different problems[18,19]. Various methods of transfer learning have been proposed in the context of drug response prediction. These methods either address these discrepancies implicitly[20–22] or explicitly, which means they assume that the model has access to the desired labelled or unlabelled target domain during training[11,12,23–29].

However, in the real-world we do not have access to the target domain(s) while training the model on the source domain, for example, 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. As generating large high-quality labelled preclinical datasets is an expensive and time-consuming process, and as we do not know the response to a given drug in the target domain (for example, future patients), there is a need for a computational method that takes not only labelled but also unlabelled 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[30–32]. Out-of-distribution generalization is particularly important for biomedical applications[33].

There are two main approaches to out-of-distribution generalization: (1) via learning domain-invariant features[31] and (2) via learning hypothesis-invariant features[34,35]. 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, that is, look similar to each other; however, forcing different domains to have very similar features is not always feasible as different domains may have unique characteristics, and completely aligning them ignores these unique characteristics. The second approach does not align the features

[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, Munich, Germany. [4]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany. [5]Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada. ✉e-mail: hsharifi@sfu.ca

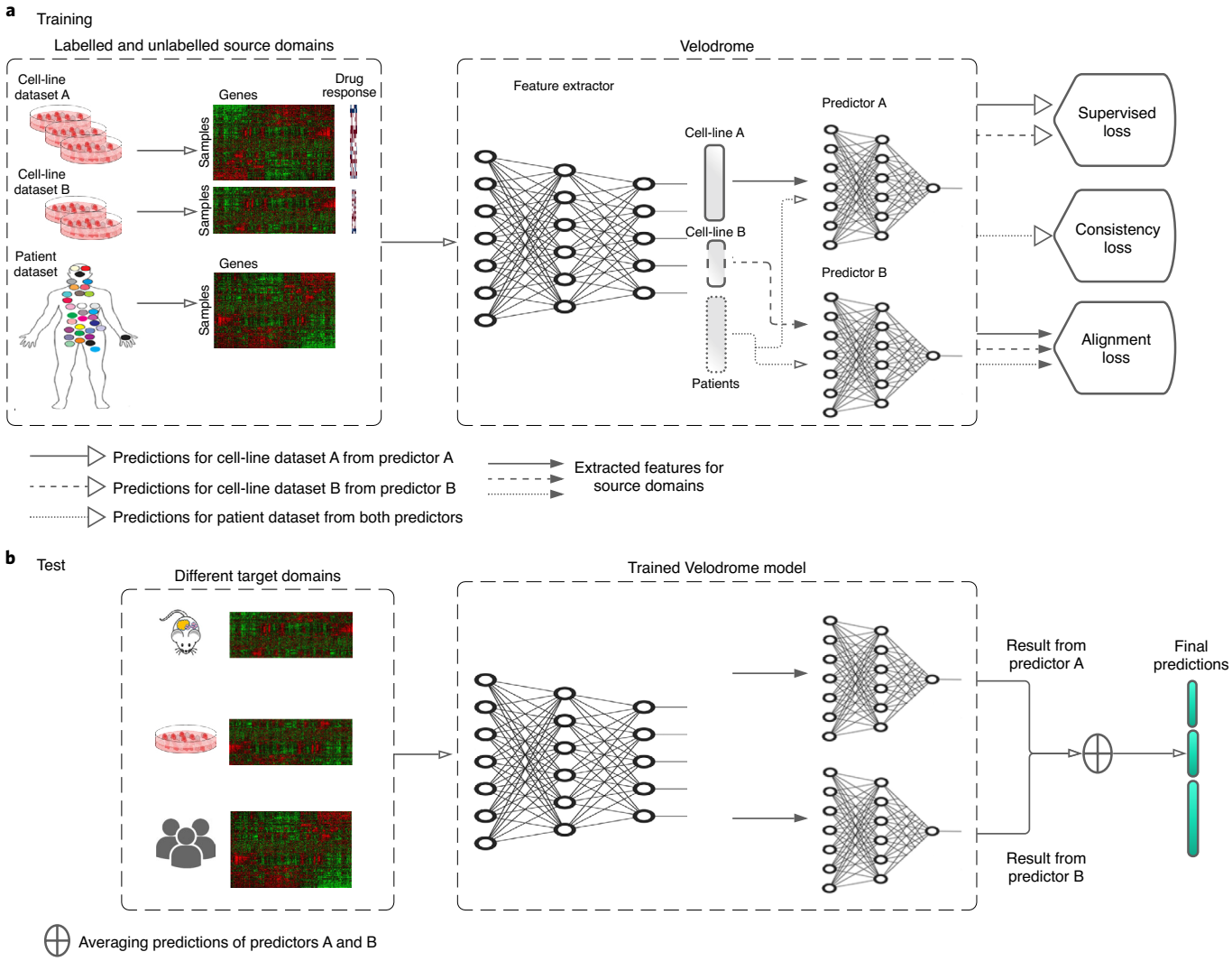


Fig. 1 | Schematic of the Velodrome method with three source domains (two labelled and one unlabelled). a, At training time, the feature extractor receives data from different source domains and extracts high-level abstract features. The extracted features of each labelled domain (cell-line dataset) are input into 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 unlabelled domain (patient dataset) are input into 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.

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.

In this paper we propose Velodrome, a deep neural network method that combines the two above approaches and exploits both labelled and unlabelled samples. Velodrome takes gene expression from cell-line (labelled) and patient (unlabelled) datasets as input domains and predicts the drug response (measured as the area above the 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 unlabelled 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 labelled cell lines and unlabelled 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 correlations, the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPR). We observed that Velodrome exhibited 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 The Cancer Genome Atlas (TCGA) patients (unlabelled,

Fig. 2 | Comparisons between Velodrome and state-of-the-art drug response prediction methods. a, Cell-line comparisons in terms of Pearson and Spearman correlations. **b**, PDX model comparisons in terms of the AUROC and the AUPR. **c**, Patient comparisons in terms of AUROC and AUPR. **d**, The average ± s.d. over the studied drugs for each method; Velodrome has the best or the second best performance on cell lines, PDX models and patients compared with the baseline methods. The error bars indicate s.d. of performance across studied drugs.

that is, 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.

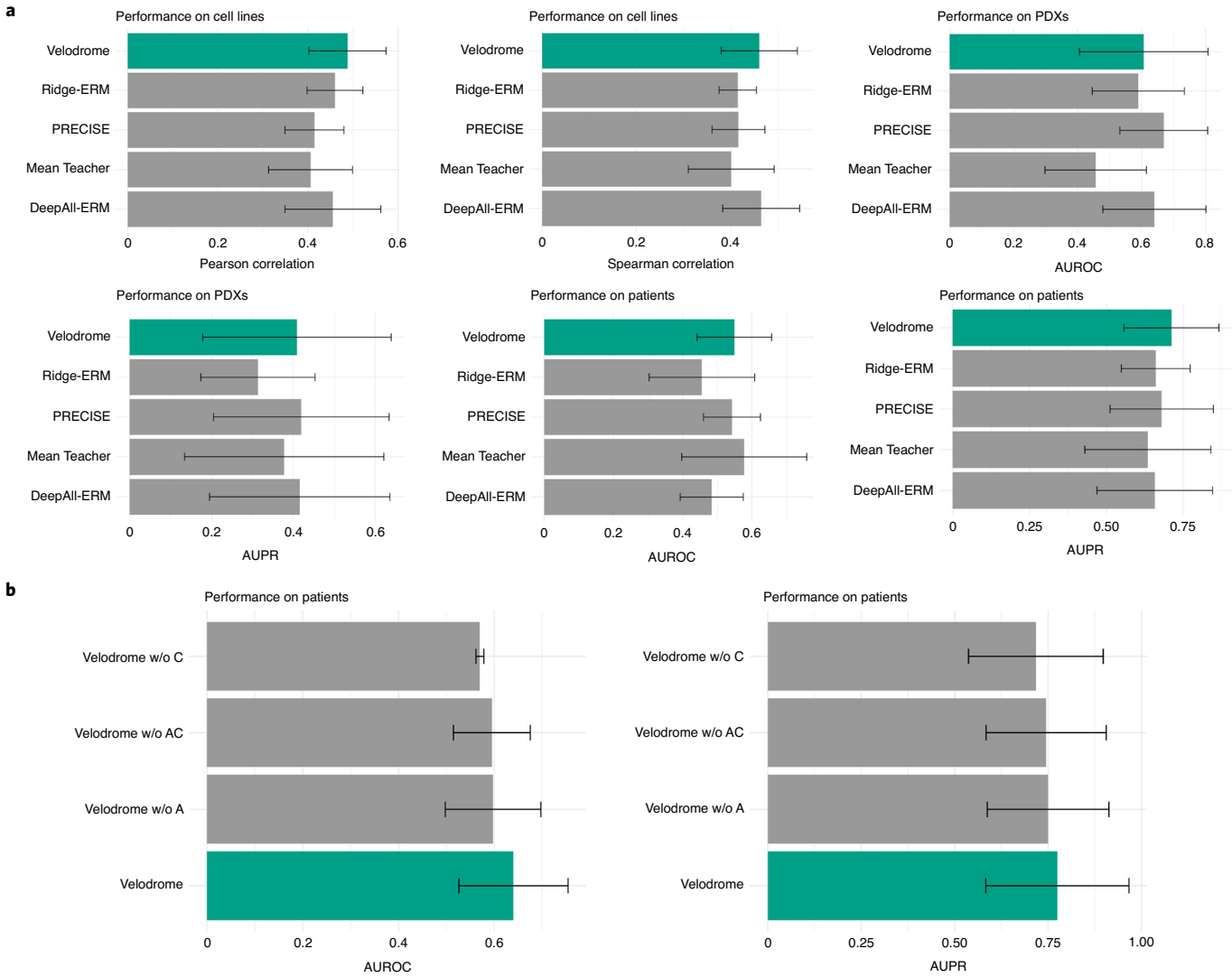


Fig. 3 | a, Comparisons of the average ± s.d. performance of the Velodrome and the state-of-the-art of drug response prediction methods over ten independent runs for the studied drugs in terms of Pearson and Spearman correlations as well as the AUROC and the AUPR. **b**, Average ± s.d. of an ablation study to examine different components of the Velodrome on patients in terms of AUROC and AUPR. The error bar indicates standard deviation of performance across studied drugs.

Results

Datasets. We employed the following resources throughout this paper:

1. Patients without drug response: more than 3,000 samples obtained from TCGA[36] breast (TCGA-BRCA), lung (TCGA-LUAD), pancreatic (TCGA-PAAD), kidney (TCGA-KIRC), prostate (TCGA-PRAD), myeloid (TCGA-LAML) and lymphoma (TCGA-DLBC) cohorts with RNA-seq data.
2. Cell lines with drug response: The Cancer Therapeutics Response Portal (CTRPv2)[6,7], The Genomics of Drug Sensitivity in Cancer (GDSCv2)[4,9] and The Genentech Cell Line Screening Initiative (gCSI)[8,10] pan-cancer datasets with a total of more than 2,000 samples with RNA-seq data and AAC as the measure of the drug response across eleven drugs (in common across the three datasets). We focused on the following drugs for this paper: Erlotinib, Docetaxel, Paclitaxel and Gemcitabine.
3. PDX samples with drug response: the PDX Encyclopaedia (PDXE) dataset[3] 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[37] for Gemcitabine, Erlotinib and Paclitaxel obtained from the supplementary material of ref. [3].
4. Patients with drug response: two cancer-specific datasets with microarray data and RECIST as the measure of drug response for Docetaxel[38], Paclitaxel[38] and Erlotinib. Furthermore, pan-cancer dataset obtained from TCGA patients treated with Gemcitabine[39]. We use clinical annotations of the drug response for some patients which were obtained from supplementary material of ref. [39].

Supplementary Tables 1, 2 and 3 present characteristics of these datasets and indicate whether they were used as source domain for training or target domain for test.

Velodrome overview. Velodrome takes gene expression and AAC of cell-line datasets (CTRPv2 and GDSCv2)—as well as gene expression of patients without a drug response (TCGA dataset)—and learns a predictive and generalizable representation. To achieve this, Velodrome employs a shared feature extractor, which

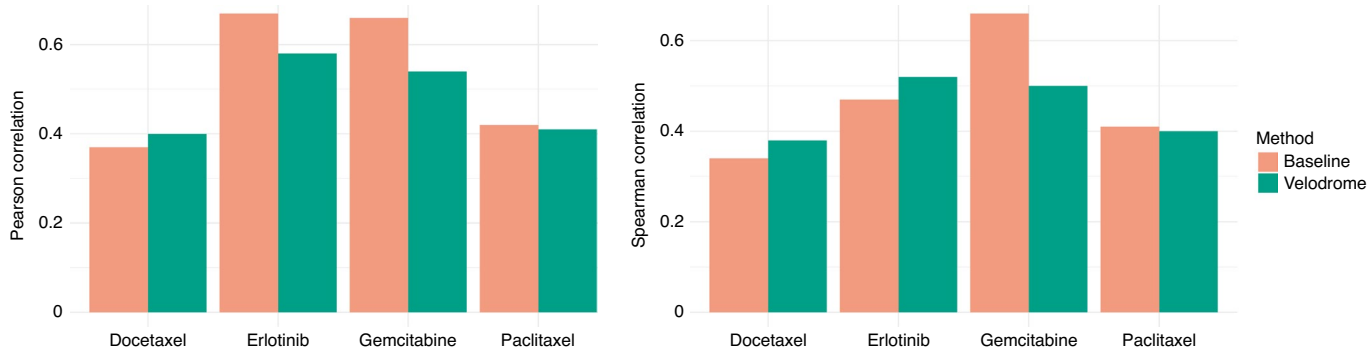


Fig. 4 | Comparisons of Velodrome predictions to the baseline correlation in terms of Pearson and Spearman correlations. The baseline correlation is obtained by calculating the correlation of cell lines in common between training and test data for each of the studied drugs. This baseline correlation shows how consistent the training and test data are in terms of AAC.

takes the gene expression of CTRPv2 and GDSCv2 samples and maps them to a shared feature space, and domain-specific predictors (for example, 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 unlabelled 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.

We add a consistency loss to incorporate unlabelled patient samples. The idea is to first extract features from patient samples using the feature extractor and then assign pseudolabels to them by utilizing the predictors associated with CTRPv2 and GDSCv2. The consistency loss takes the pseudolabels (that is, 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 (for example, 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 labelled and unlabelled 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 unlabelled 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 labelled 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.

We compared Velodrome against the state-of-the-art methods of each approach to evaluate its performance. For the first, we compared Velodrome with Mean Teacher[40], which is the state-of-the-art deep neural network for semi-supervised learning[41]. For the second, we compared it with PRECISE as a non-deep learning method based on subspace alignment. Finally, we compared Velodrome with Ridge-ERM (ridge regression) as a non-deep learning baseline, and DeepAll-ERM as a deep learning baseline, both of which are categorized as methods of empirical risk minimization (ERM). In an extensive benchmark in the context of computer vision, ERM methods tended to achieve state-of-the-art performance for out-of-distribution generalization[30]. They are trained in a supervised fashion by merging all available labelled input domains.

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 Spearman correlations on this dataset. On average ± s.d. over all drugs, DeepAll-ERM achieved the best performance (0.52 ± 0.09 and 0.48 ± 0.09 for Pearson and Spearman correlation coefficients, respectively; Fig. 2d). Velodrome achieved the second best performance (0.48 ± 0.09 and 0.45 ± 0.07 for Pearson and Spearman correlation coefficients, respectively; Fig. 2a,d). Ridge-ERM (0.46 ± 0.07; Fig. 2a,d) and Mean Teacher (0.43 ± 0.07; Fig. 2a,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 (Fig. 2a).

Velodrome makes accurate predictions for PDX 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 ± s.d. over all drugs, Velodrome achieved the best performance compared with the baselines (0.69 ± 0.21 in AUROC and 0.43 ± 0.26 in AUPR; Fig. 2b,d). PRECISE and DeepAll-ERM obtained the second best performance in terms of AUROC (0.67 ± 0.14; Fig. 2b,d) and AUPR (0.42 ± 0.23; Fig. 2b,d), respectively. Similarly, DeepAll and PRECISE had the third best performance in terms of AUROC (0.63 ± 0.19; Fig. 2b,d) and AUPR (0.41 ± 0.24; Fig. 2b,d), respectively. These results indicate that utilizing both labelled and unlabelled 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 ± s.d. over all drugs, Velodrome achieved the best performance compared

to the baselines and outperformed the competitors (0.64 ± 0.11 in AUROC and 0.77 ± 0.19 in AUPR; Fig. 2c,d). Mean Teacher obtained the second best performance (0.59 ± 0.21 in AUROC and 0.69 ± 0.23 in AUPR; Fig. 2c,d) and PRECISE had the third best performance (0.54 ± 0.1 in AUROC and 0.68 ± 0.18 in AUPR; Fig. 2c,d). Interestingly, these three top-performing methods take for inputs both labelled and unlabelled samples, in contrast to other baselines that consider only labelled samples. These results indicate that incorporating unlabelled patient data along with labelled data notably improves the generalization performance on patients; however, the results also demonstrate the advantage of learning features that are domain- and hypothesis-invariant for out-of-distribution generalization, as PRECISE 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 hyperparameters 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 retrained all of the models from scratch using the same settings with ten different random seeds and reported mean ± s.d. for each method (Fig. 3a) over all runs and drugs. 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 correlations 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 with baselines particularly on patients and cell lines.

The complete version of Velodrome exhibits 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 the complete version of Velodrome outperforms its variants on average ± s.d. (over all drugs for ten independent runs), indicating the added value of both alignment and consistency losses (Fig. 3b). 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[34].

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[11,12,20,42]. 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 (Extended Data Fig. 1).

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. We then 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 exhibits 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. BCL2 was not significant ($P > 0.05$) for Docetaxel after hypothesis correction at level $\alpha = 0.05$.

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 the prostate as, unlike the kidney, it is a under-represented tissue type 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 exhibits 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. In this cohort, BCL2 was not significant ($P > 0.05$) for Docetaxel after hypothesis correction at level $\alpha = 0.05$.

Velodrome generalizes to new tissue types. Finally, we trained Velodrome and the baseline methods only on samples (cell lines and patients) that originated from solid tissue types, as non-solid tissues such as haematopoietic and lymphoid have different molecular and pharmacological profiles compared with solid tissues[43]. We therefore 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 that originate from non-solid tissues in the gCSI cell-line dataset and evaluated performance in terms of the Pearson correlation between the predictions and the actual AAC values, as well as between the predictions and reported two-tailed P-values.

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 that originated from non-solid tissues in CTRPv2 and GDSCv2 datasets and tested this predictor on non-solid samples of the gCSI dataset. We therefore 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 (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.

Within-tissue generalization. We expanded on the within-tissue performance of Velodrome to further study the generalization capability of the trained models. We compared Velodrome predictions with the AAC baseline correlation in the utilized datasets. To obtain this baseline correlation, we selected cell lines with solid tissues that are common between our training (cell lines in CTRPv2 and GDSCv2) and test data (cell lines in gCSI). This baseline correlation indicates how much the data agree with each other in terms of drug response (AAC). We expect to see a comparable correlation for Velodrome predictions if the given model is accurate enough. Our results showed that for the majority of the drugs, Velodrome achieved comparable performance compared to the baseline correlation in terms of Pearson and Spearman correlations (Fig. 4). This reconfirms our previous results on generalization within solid tissues.

Cross-tissue generalization. We also expanded the cross-tissue generalization analysis by repeating a similar association study of well- and under-represented tissues but this time for two TCGA cohorts with non-solid tissue types. For that, we obtained TCGA-LAML and TCGA-DLBC datasets for acute myeloid leukaemia and diffuse large B-cell lymphoma, respectively. Similar to TCGA-KIRC and TCGA-PRAD, Velodrome predictions for TCGA-LAML and TCGA-DLBC patients demonstrated significant associations with known target genes of the studied drugs.

For Docetaxel, MAP2 showed a statistically significant association in TCGA-LAML ($P = 10^{-5}$) and TCGA-DLBC ($P = 0.02$) but BCL2 was not significant in these cohorts. For Erlotinib, EGFR and ERBB2 showed statistically significant associations for TCGA-LAML (both $P < 10^{-6}$) and these two genes were also significant in TCGA-DLBC (8×10^{-6} and 4×10^{-7}, respectively). For Gemcitabine, CMPK1 exhibits significant association ($P < 10^{-6}$) in both cohorts. Finally, for Paclitaxel, only BCL2 ($P = 4 \times 10^{-4}$) had significant associations for patients in the TCGA-LAML cohort and we did not observe any significant associations for TCGA-DLBC. Our results demonstrate that associations were not significant for Paclitaxel for non-solid patient data, which aligns with the previous results where Velodrome did not generalize to non-solid cell lines for Paclitaxel. For Docetaxel, *P*-values are larger than what we observed for solid tumours, which again aligns with the previous results where Velodrome did not generalize to non-solid cell lines; however, we observed significant associations for Erlotinib and Gemcitabine, which is consistent with Velodrome accurate performance on non-solid cell lines. Overall, these new results also reconfirm cross-tissue generalization of Velodrome models.

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 and survival[44]. More importantly, the expression of BCL2 may have an antiapoptotic activity against androgen, which is a key player in prostate cancer[45]. Similarly, BCL2 can also act as an oncoprotein in kidney cancer[46] and therapeutics roles.

As another example, microtubule-associated proteins including MAP2 have also been associated with different cancers including prostate[47] and kidney cancers[48]. 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[49]. Prostate cancer progression and lethal outcome have been associated with metabolic signalling pathways and CMPK1 (it mediates the mechanism of action for Gemcitabine) was shown to be highly expressed in prostate cancer patients[50]. A combination of Gemcitabine and other chemotherapy agents has shown to be effective for a subtype of kidney cancer[51]. Finally, EGFR and ERBB2 have been associated with different cancer types including prostate[52] and kidney[53] and they both showed therapeutic opportunities and increase in survival[54].

From the computational point of view, it has been shown that methods of ERM are highly competitive for supervised domain generalization[30]. It was therefore 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 labelled and unlabelled samples and were therefore 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, as these tissue types were well-represented in our cell-line datasets (Extended Data Fig. 1) and the four studied drugs were treatment options for these cell lines. This selection increases the relevancy of labelled (cell lines) and unlabelled (TCGA patients) data. Relevancy has been shown to improve semi-supervised learning performance even when both labelled and unlabelled datasets are imbalanced[41], which is the case for drug response prediction.

Although methods of adversarial domain adaptation have shown great performance in different applications, especially computer vision[55], we did not consider them as baselines because they were clearly outperformed by PRECISE (which we do use as baseline) in a recent study[12].

Although gene expression data have been shown many times to be the most effective genomic data type for drug response prediction[9,56], in principle Velodrome can be extended to incorporate other omics data types. Especially promising are proteomics data and germline variants, 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 further information about the drug—such as the chemical representation—to improve performance[57]. 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[21] to offer explainability and transparency. A major limitation of our work is the output space discrepancy between cell lines, PDX samples and patients, as on cell lines the drug response is measured on the basis of the concentration of the drug, whereas on PDX samples and patients the response is measured based on the change in the tumour volume after treatment. A recent method[12] adjusts for this output space discrepancy and improves the prediction performance, but it requires access to the target domain during training that violates the assumption of out-of-distribution generalization. In this work we used AAC as a 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 displayed that adjusting concentration ranges across different datasets improves the prediction performance[58,59], but we did not consider this adjustment as it reduces the sample size substantially.

Although drug response prediction was the driving problem of the Velodrome method, in principle it is also applicable to other problems, especially in clinical settings. For example, prostate cancer has steady progression, meaning datasets for this cancer have a limited number of labelled samples with respect to long-term outcomes such as metastasis but more unlabelled samples[60]. This makes Velodrome applicable to metastasis prediction via labelled and unlabelled gene expression data. Another example is the case of rare cancers for which obtaining large training datasets is extremely difficult. Velodrome demonstrated cross-tissue generalization, which makes it applicable to these rare tissue types. Both of these cases can be promising future directions.

Indeed, our results suggest that owing to its ability for cross-tissue generalization, Velodrome has a potential to predict the response even for cancer types under-represented or completely absent in the training data. The cross-tissue generalization is particularly interesting as some previous studies have shown basic differences in gene expression between solid and non-solid tissues[61], however, our empirical results demonstrated that Velodrome models trained on solid tissues generalize to non-solid tissues in cell lines and patients. This may suggest shared component(s) in basal transcription factors. One example is cyclin-dependent kinases, which are targets for global transcription inhibition in numerous drugs[62]. Interestingly, some of these drugs are treatments for both solid and non-solid cancers.

Another explanation for cross-tissue generalization can be the presence of drivers shared across cancer types. Different studies have identified pan-cancer driver genes across solid and non-solid tissues such as TP53, ERBB2, EGFR or EML4-ALK[63,64]. Interestingly, ERBB2 and EGFR are known to be associated with Erlotinib (one of the drugs we studied), which demonstrated cross-tissue generalization to non-solid tissues. EGFR activation drives the development of most head and neck tumours, and in a large proportion of glioblastoma and lung adenocarcinoma cases[65]. More recently, overexpression of EGFR was reported in a subset of acute myeloid leukaemia patients with poor prognosis[66]. A related receptor tyrosine kinase ERBB2 is frequently activated in breast cancer and, more rarely, in several other tumour types[67]. ERBB2 amplification and activation are associated with poor prognosis and the resistance to anti-EGFR treatments, including Erlotinib[68]. Later, putative driver ERBB2 mutations were found in three leukaemia patients[64]. Their effect on proliferation rate and sensitivity to ERBB inhibitors was shown in cellular assays[64].

We would like to note that the promising Velodrome results should be interpreted with caution due to the lack of drug response values in our TCGA association analyses. Direct experimental validation of cross-tissue generalization and Velodrome predictions is thus an avenue of future research.

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 (that is, 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 analysed 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 (that is, kidney cancer) and a tissue under-represented in our source domains (that is, prostate cancer). Finally, we also showed that Velodrome generalizes to new tissue types that were completely absent in the source domains. All of 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 ORCESTRA[69], which stores pharmacogenomics datasets in PharmacoSet 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 version 2.0.5 (ref. [70]) and are also comparable in terms of gene expression data, which was preprocessed via Kallisto_0.46.1 (ref. [71]). 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/) on 28 January 2016. Expression values were converted to transcripts per million and log2-transformed. The PDX and clinical trial datasets were preprocessed similar to the approach described in ref. [20]. The accession code is GSE25065 for Docetaxel and Paclitaxel patient data, whereas the accession code is GSE33072 for Erlotinib.

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 2,128 genes obtained from ref. [72]. After preprocessing, all of the available datasets for each drug had the same number of genes (Supplementary Table 1).

The Velodrome method. We propose Velodrome, a method of drug response prediction using labelled and unlabelled 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 labelled and unlabelled data. We designed the objective function of Velodrome to meet these requirements by combining three loss functions: (1) a standard supervised loss based on labelled 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 unlabelled data.

Problem definition. Following the notation of ref. [16], a domain D is defined by a raw input space $\mathbf{X}$, a probability distribution $p(X)$ and a corresponding dataset $X=\{x_1,x_2,\ldots,x_n\}$ 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 labelled and unlabelled 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 of D_i) as input and makes predictions (of the drug response) $\overline{Y_i}$ for this source domain; θ and $\{\phi_i\}_{i=1}^{n_l}$ are being optimized using an objective function $J\left(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}\right) = l\left(D^L, \theta, \{\phi_i\}_{i=1}^{n_l}\right) + \Omega\left(D^L, D^U, \theta, \{\phi_i\}_{i=1}^{n_l}\right)$, with a supervised loss $l(.)$ and some regularization terms $\Omega(.)$.

In drug response prediction, we have access to labelled source domains such as cell-line datasets and unlabelled 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 (θ and $\{\phi_i\}_{i=1}^{n_l}$) 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 labelled and unlabelled source domains:

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

where Z_i^j denotes the features extracted by the feature extractor $F_\theta(.)$ from X_i^j, the samples obtained from the ith domain of type j (labelled or unlabelled). 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 labelled 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 ith labelled source domain obtained from predictor $G_{\phi_i}^l(.)$ associated with the ith labelled source domain and parameterized by ϕ_i. These predictions will be utilized to optimize the parameters of the feature extractor and the ith predictor.

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

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

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

Alignment loss. Optimizing the parameters of the Velodrome model using only the supervised loss is likely to lead to overfitting to the labelled source domains. We therefore 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[30,73]. The CORAL loss is defined as follows:

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

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:

$$C(Z) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \underline{Z_i}\right)\left(Z_i - \underline{Z_i}\right)^T,$$

where, n is the number of samples and $\underline{Z}$ 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[73].

Consistency loss. Aligning the extracted features of the different domains imposes a strict constraint on learning an invariant latent representation as 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, that is, 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 unlabelled source domain. This consistency loss is defined as follows:

$$\text{CON}\left(D^u, \theta, \{\phi_i\}_{i=1}^{n_l}\right) = \sum_{i,j\neq i}\left\|G_{\phi_i}^l\left(Z^u\right) - G_{\phi_j}^l\left(Z^u\right)\right\|_2^2,$$

where, Z^u are extracted features for samples in a given unlabelled source domain

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

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

where, λ_1 and $\lambda_2 = 1 - \lambda_1$ denote the regularization coefficients for the CORAL loss and consistency loss, respectively. The function $\Omega(.)$ that we defined in the problem definition is given by $\Omega(.) = \lambda_1\text{CORAL}(.) + \lambda_2\text{CON}(.)$; λ_1 and λ_2 control balancing between learning domain- and hypothesis-invariant representations, as the alignment and consistency losses ensure learning domain- and hypothesis-invariant representations, respectively. 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:

$$\overline{y^t} = \sum\nolimits_i w_i G_{\phi_i}\left(F_\theta\left(x^t\right)\right),$$

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 that the final prediction will be a result of a weighted average of all predictors, and more accurate predictors will have higher weights.

Implementation detail. *Hyperparameters.* We considered a wide range of values for each hyperparameter of the Velodrome model and optimized these values via a random search separately for each drug. The sets of values considered are as follows:

$$\text{Epoch} = [10, 50, 100, 200],$$

$$\text{Learning rate (LR)} = [0.0001, 0.001, 0.01, 0.0005, 0.005, 0.05]$$

$$\text{Dropout (DR)} = [0.1, 0.3, 0.5, 0.8],$$

$$\text{Weight decay (WD)} = [0.001, 0.0001, 0.01, 0.05, 0.005, 0.0005],$$

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

$$\text{Minbatch size (MB)} = [17, 33, 65, 129]$$

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 labelled cell-line datasets (CTRPv2 and GDSCv2) into training and validation, and considered 90% for training and 10% for validation. We merged the training 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 hyperparameters.

Velodrome architecture. We followed previous works and designed predefined architectures (denoted by HD) for Velodrome[43,74]. 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 $128\times128\times128$ and the last architecture has four hidden layers with the size $64\times64\times64\times64$. We considered a batch normalization layer followed by an activation function (for which we considered the Relu, Tanh and Sigmoid functions) as well as a dropout after the activation function for each hidden layer. The predictors have only one layer HD×1, where HD denotes the size of the last layer in the feature extractor. The final hyperparameter 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 rerunning and the ablation study of the trained models, we considered these random values for the random seed:

$$\text{Seed} = [1, 21, 42, 84, 168, 336, 672, 1344, 2688, 5376].$$

We used 42 for the majority of the analyses in the paper (as 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 at https://github.com/bhklab/PGx_Guidelines; for PRECISE, we used the existing implementations at 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[75]. We did not perform statistical tests on the performance of Velodrome compared to the other methods to evaluate level of significance. For the association studies of well- or under-represented tissues types, we performed a two-tailed *t*-test on the regression coefficients with the null hypothesis that a given coefficient (corresponding to a gene) is zero and has no significant association with drug response. For the correlation analysis on the unseen tissues (non-solid tissue types), we employed a two-tailed test, testing the specific null hypothesis that the population correlation is zero against a two-tailed alternative.

Data availability

All the final preprocessed data employed in this paper are publicly available here: https://zenodo.org/record/4793442#.YK1HVqhKiUk (ref. [76]). All the raw data before preprocessing are also publicly available as follows: (1) cell-line datasets with gene expression and drug response data, including CTRPv2, GDSCv2 and gCSI, were downloaded from ORCESTRA[69]; (2) TCGA cohorts with gene expression data were downloaded from Firehose (http://gdac.broadinstitute.org/) on 28 January 2016. Drug response data for TCGA cohorts was obtained from ref. [39]; (3) PDX datasets (gene expression with drug response data) were obtained from the Supplementary Information of ref. [3]; (4) Patient dataset (gene expression with drug response data) were obtained from the accession codes GSE25065 (Docetaxel and Paclitaxel) and GSE33072 (Erlotinib). Source data are provided with this paper.

Code availability

All the codes, model objects and supplementary material used to run and reproduce our experimental results are publicly available at https://github.com/hosseinshn/Velodrome (ref. [77]). We also provided a conda environment to ensure version compatibility for future users.

Received: 31 May 2021; Accepted: 28 September 2021;
Published online: 11 November 2021

References

1. Marquart, J., Chen, E. Y. & Prasad, V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. *JAMA Oncol.* **4**, 1093–1098 (2018).
2. Pal, S. K. et al. Clinical cancer advances 2019: annual report on progress against cancer from the American society of clinical oncology. *J. Clin. Oncol.* **37**, 834–849 (2019).
3. Gao, H. et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
4. Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
5. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
6. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
7. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
8. Klijn, C. et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).
9. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
10. Haverty, P. M. et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
11. Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519 (2019).
12. Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C. & Ester, M. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. *Bioinformatics* **36**, i380–i388 (2020).
13. Haibe-Kains, B. et al. Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
14. Mpindi, J. P. et al. Consistency in drug response profiling. *Nature* **540**, E5–E6 (2016).
15. Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J. & Huang, R. S. Consistency in large pharmacogenomic studies. *Nature* **540**, E1–E2 (2016).
16. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
17. Neyshabur, B., Sedghi, H. & Zhang, C. What is being transferred in transfer learning? In *34th Conference on Neural Information Processing Systems* (NeurIPS, 2020).
18. Raghu, M. et al. Transfusion: understanding transfer learning for medical imaging. In *33rd Conference on Neural Information Processing System* (eds, Wallach, H. et al.) 3347–3357 (Curran Associates, 2019).
19. Hu, J. et al. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.* **2**, 607–618 (2020).
20. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019).
21. Snow, O. et al. Interpretable Drug Response Prediction using a Knowledge-based Neural Network. In *Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021).
22. Kuenzi, B. M. et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020).
23. Mourragui, S. et al. Predicting clinical drug response from model systems by non-linear subspace-based transfer learning. Preprint at https://www.biorxiv.org/content/10.1101/2020.06.29.177139v3 (2020).
24. Ma, J. et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
25. Zhu, Y. et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci. Rep.* **10**, 18040 (2020).
26. Salvadores, M., Fuster-Tormo, F. & Supek, F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Sci. Adv.* **6**, aba1862 (2020).
27. Najgebauer, H. et al. CELLector: genomics-guided selection of cancer in vitro models. *Cell Syst.* **10**, 424–432.e6 (2020).
28. Peres da Silva, R., Suphavilai, C. & Nagarajan, N. TUGDA: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from in vitro to in vivo settings. *Bioinformatics* **37**, i76–i83 (2021).
29. Warren, A. et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 22 (2021).
30. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations* (2021).
31. Wang, J. et al. Generalizing to unseen domains: a survey on domain generalization. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence* (2021).
32. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain generalization: a survey. Preprint at https://arxiv.org/abs/2103.02503 (2021).
33. Zhang, H. et al. An empirical framework for domain generalization in clinical settings. In *Proc. Conference on Health, Inference, and Learning* (ACM, 2021); https://doi.org/10.1145/3450439.3451878
34. Zhao, S., Gong, M., Liu, T., Fu, H. & Tao, D. Domain generalization via entropy regularization. In *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2020).
35. Wang, Z., Loog, M. & van Gemert, J. Respecting domain relations: hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition* 9756–9763 (ICPR, 2021).
36. Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
37. Schwartz, L. H. et al. RECIST 1.1—update and clarification: from the RECIST committee. *Eur. J. Cancer* **62**, 132–137 (2016).
38. Hatzis, C. et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
39. Ding, Z., Zu, S. & Gu, J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* **32**, 2891–2895 (2016).
40. Tarvainen, A. & Valpola, H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *31st Conference on Neural Information Processing Systems* (2017).
41. Yang, Y. & Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. In *Conference on Neural Information Processing Systems* (2020).
42. Geeleher, P. et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res.* **27**, 1743–1751 (2017).

43. Noghabi, H. S. et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings Bioinformatics* https://doi.org/10.1093/bib/bbab294 (2021).
44. Renner, W., Langsenlehner, U., Krenn-Pilko, S., Eder, P. & Langsenlehner, T. BCL2 genotypes and prostate cancer survival. *Strahlenther. Onkol.* **193**, 466–471 (2017).
45. Chaudhary, K. S., Abel, P. D. & Lalani, E. N. Role of the Bcl-2 gene family in prostate cancer progression and its implications for therapeutic intervention. *Environ. Health Perspect.* **107**, 49–57 (1999).
46. Paraf, F., Gogusev, J., Chrétien, Y. & Droz, D. Expression of Bcl-2 oncoprotein in renal cell tumours. *J. Pathol.* **177**, 247–252 (1995).
47. Bhat, K. M. R. & Setaluri, V. Microtubule-associated proteins as targets in cancer chemotherapy. *Clin. Cancer Res.* **13**, 2849–2854 (2007).
48. He, Z., Liu, H., Moch, H. & Simon, H.-U. Machine learning with autophagy-related proteins for discriminating renal cell carcinoma subtypes. *Sci. Rep.* **10**, 720 (2020).
49. Martin, S. K., Kamelgarn, M. & Kyprianou, N. Cytoskeleton targeting value in prostate cancer treatment. *Am. J. Clin. Exp. Urol.* **2**, 15–26 (2014).
50. Kelly, R. S. et al. The role of tumor metabolism as a driver of prostate cancer progression and lethal disease: results from a nested case-control study. *Cancer Metab.* **4**, 22 (2016).
51. Numakura, K. et al. Successful mammalian target of rapamycin inhibitor maintenance therapy following induction chemotherapy with gemcitabine and doxorubicin for metastatic sarcomatoid renal cell carcinoma. *Oncol. Lett.* **8**, 464–466 (2014).
52. Pignon, J.-C. et al. Androgen receptor controls EGFR and ERBB2 gene expression at different levels in prostate cancer cell lines. *Cancer Res.* **69**, 2941–2949 (2009).
53. Reid, A., Vidal, L., Shaw, H. & de Bono, J. Dual inhibition of ErbB1 (EGFR/HER1) and ErbB2 (HER2/neu). *Eur. J. Cancer* **43**, 481–489 (2007).
54. Gordon, M. S. et al. Phase II study of Erlotinib in patients with locally advanced or metastatic papillary histology renal cell cancer: SWOG S0317. *J. Clin. Oncol.* **27**, 5788–5793 (2009).
55. Chen, Y.-H. et al. No more discrimination: cross city adaptation of road scene segmenters. In *Proc. IEEE International Conference on Computer Vision* 1992–2001 (IEEE, 2017).
56. Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
57. Jiang, Y., Rensi, S., Wang, S. & Altman, R. B. DrugOrchestra: jointly predicting drug response, targets, and side effects via deep multi-task learning. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.17.385757v1 (2020).
58. Pozdeyev, N. et al. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* **7**, 51619–51625 (2016).
59. Xia F, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Brief. Bioinform.* (2021).
60. Sharifi-Noghabi, H., Liu, Y., Erho, N. & Shrestha, R. Deep genomic signature for early metastasis prediction in prostate cancer. Preprint at https://www.biorxiv.org/content/10.1101/276055v2 (2019).
61. Torrente, A. et al. Identification of cancer related genes using a comprehensive map of human gene expression. *PLoS ONE* **11**, e0157484 (2016).
62. Villicaña, C., Cruz, G. & Zurita, M. The basal transcription machinery as a target for cancer therapy. *Cancer Cell Int.* **14**, 18 (2014).
63. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
64. Joshi, S. K. et al. ERBB2/HER2 mutations are transforming and therapeutically targetable in leukemia. *Leukemia* **34**, 2798–2804 (2020).
65. Thomas, R. & Weihua, Z. Rethink of EGFR in cancer with its kinase independent function on board. *Front. Oncol.* **9**, 800 (2019).
66. Nath, S. et al. The prognostic impact of epidermal growth factor receptor (EGFR) in patients with acute myeloid leukaemia. *Indian J. Hematol. Blood Transfus.* **36**, 749–753 (2020).
67. Iqbal, N. & Iqbal, N. Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. *Molecular Biol. Int.* **2014**, 1–9 (2014).
68. Goss, G. D. et al. Association of ERBB mutations with clinical outcomes of Afatinib- or Erlotinib-treated patients with lung squamous cell carcinoma: Secondary analysis of the LUX-lung 8 randomized clinical trial. *JAMA Oncol.* **4**, 1189–1197 (2018).
69. Mammoliti, A. et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. Nature Communications volume 12, Article number: 5797 (2021).
70. Smirnov, P. et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).
71. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Erratum: near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).
72. Manica, M. et al. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).
73. Sun, B. & Saenko, K. Deep CORAL: correlation alignment for deep domain adaptation. In *Computer Vision—ECCV 2016 Workshops* 443–450 (Springer, 2016).
74. Sakellaropoulos, T. et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep.* **29**, 3367–3373.e4 (2019).
75. Smirnov, P. et al. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucl. Acids Res.* **46**, D994–D1002 (2018).
76. Sarifi-Noghabi, H,. Harjandi, P. A., Zolotareva, O., Collins, C. C. & Ester, M. *Velodrome: Out-of-Distribution Generalization from Labeled and Unlabeled Gene Expression Data for Drug Response Prediction* (Zenodo, 2021); https://doi.org/10.5281/zenodo.4793442
77. Sharifi-Noghabi, H. *Code Repository hosseinshn/Velodrome: DOI (v1.0.0)* (Zenodo, 2021); https://doi.org/10.5281/zenodo.5164625

Acknowledgements

We would like to thank H. Asghari (Ocean Genomics) and S. 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. 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.).

Author contributions

H.S-.N. and M.E. conceived the study concept and design. H.S-.N. was responsible for the deep learning design, implementations and analysis. H.S-.N. and O.Z. performed data preprocessing, analysis and interpretation. H.S-.N. and P.A.H. performed the experiments. H.S-N., P.A.H. and O.Z. analysed and interpreted the results. C.C.C. and M.E. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00408-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00408-w.

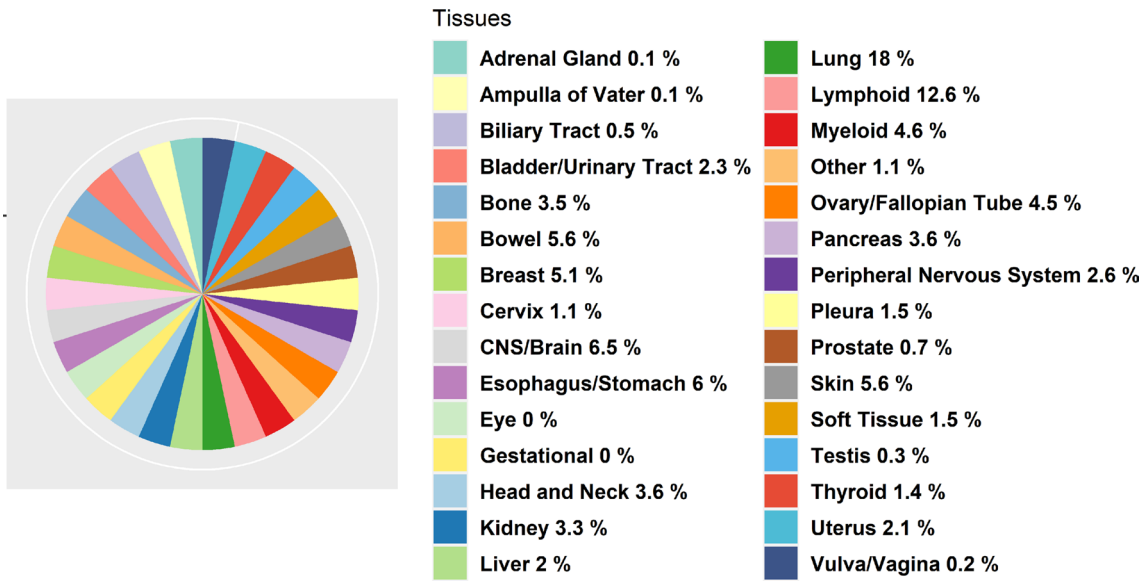
Correspondence and requests for materials should be addressed to Martin Ester.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021



Extended Data Fig. 1 | The percentage of tissue types in CTRPv2 and GDSCv2 cell line datasets combined.