Pre-print, 2024, pp. 1–7

doi: DOI HERE
Paper

PAPER

Multi-task deep latent spaces for cancer survival and drug sensitivity prediction

Teemu J. Rintala,[1] Francesco Napolitano[2] and Vittorio Fortino[1,*]

[1]Institute of Biomedicine, University of Eastern Finland, Yliopistonranta 8, 70210 Kuopio, Finland and [2]Department of Science and Technology, University of Sannio, 82100 Benevento, Italy
[*]Corresponding author. vittorio.fortino@uef.fi

Abstract

Motivation: Cancer is a very heterogeneous disease that can be difficult to treat without addressing the specific mechanisms driving tumour progression in a given patient. High-throughput screening and sequencing data from cancer cell-lines has driven many developments in drug development, however, there are important aspects crucial to precision medicine that are often overlooked, namely the inherent differences between tumours in patients and the cell-lines used to model them *in vitro*. Recent developments in transfer learning methods for patient and cell-line data have shown progress in translating results from cell-lines to individual patients *in silico*. However, transfer learning can be forceful and there is a risk that clinically relevant patterns in the omics profiles of patients are lost in the process.
Results: We present MODAE, a novel deep learning algorithm to integrate omics profiles from cell-lines and patients for the purposes of exploring precision medicine opportunities. MODAE implements patient survival prediction as an additional task in a drug-sensitivity transfer learning schema and aims to balance autoencoding, domain adaptation, drug-sensitivity prediction, and survival prediction objectives in order to better preserve the heterogeneity between patients that is relevant to survival. While burdened with these additional tasks, MODAE performed on par with baseline survival models, but struggled in the drug-sensitivity prediction task. Nevertheless, these preliminary results were promising and show that MODAE provides a novel AI-based method for prioritizing drug treatments for high-risk patients.
Availability: `https://github.com/UEFBiomedicalInformaticsLab/MODAE`

Key words: precision medicine, multi-task deep-learning, cancer transcriptomics, domain adaptation, drug sensitivity analysis, survival analysis

Introduction

High-throughput sequencing of omics data can be used to characterize the molecular state of cells and tissues in both patient tumors and cancer cell-lines that are used to model them. Very large patient cohorts such as the cancer genome atlas (TCGA) have been used to study the molecular mechanisms underlying cancer progression and the remarkable heterogeneity of tumors, leading to the development of precision medicine approaches that target specific genomic aberrations to improve treatment outcomes (Chin et al., 2011). On the other hand, cancer cell-lines have been widely screened *in vitro* for sensitivity with large panels of anti-cancer compounds and thoroughly characterized in projects such as CCLE (Ghandi et al., 2019; Barretina et al., 2012), GDSC (Yang et al., 2013), and CTRP (Seashore-Ludlow et al., 2015; Basu et al., 2013). Large collections of screening results together with comprehensive molecular profiles of the cell-lines provide incredible resources for data-mining drug mechanisms and targeted therapies. However, the heterogeneity of cancer aberrations and the intrinsic differences between cancer cell lines and patient tumors has made it difficult to identify appropriate models for specific (sub-)types of cancer (Salvadores et al., 2020). In order to translate drug screening results to patient tumors it is very important to properly account for these differences (e.g. by matching each patient tumour to a cell-line model based on the similarity of their omics profiles). However, instead of directly establishing the best representative model for each cancer, the issue could be addressed as a domain adaptation problem to translate predictive models learned on cell-line omics profiles to the domain of patient tumor omics profiles. Recently, deep learning approaches have been shown to outperform traditional machine learning and domain adaptation techniques in this complex task (He et al., 2022). Accurately translated sensitivity screens could help prioritize drugs to a given group of omics profiles and reduce drug development costs for precision oncology.

To address the problem of transfer learning, Dincer et al. (2020) implemented an adversarial extension of the autoencoder model (AD-AE) to obtain a representation of data free from confounding factors between datasets. CODE-AE-ADV by He

© The Author 2024.

et al. (2022) used a more robust version of the adversarial autoencoder model by utilizing the Wasserstein distance as inspired by WGAN (Arjovsky et al., 2017; Gulrajani et al., 2017). CODE-AE-ADV also implemented shared and private embeddings for learning separate latent representations of the confounders in addition to the deconfounded representation shared between data domains. The goal of CODE-AE-ADV was to obtain a representation that contains only the shared characteristics of both cell-line and patient tumor omics profiles and to use this representation to learn a transferable model of drug sensitivity prediction. In addition to bulk omics, a similar transfer learning schema has also been applied to single-cell drug sensitivity by Chen et al. (2022). However, focusing solely on optimizing drug sensitivity may skew the shared embedding space towards features more informative to cancer cell lines than to patient cases. For a comprehensive cancer precision medicine approach, it is essential that the genomic embedding space that connects cell-line and patient tumor omics profiles, remains relevant to cancer patients. This can be achieved by incorporating additional learning tasks focusing on clinical aspects, such as survival data or molecular subtypes. This ensures the genomic data space accurately reflects patient information as well as cancer cell line information, potentially facilitating the identification of clinically actionable cancer subtypes. By studying patients with similar genomic profiles and drug sensitivities, and linking these to specific survival outcomes, we can attain a true precision medicine approach. To integrate these disparate objectives within machine learning we can use multi-task learning (MTL) approaches. MTL can be used to learn multiple tasks simultaneously instead of building separate models for each task or learning in stages. The combination of autoencoders for feature extraction and predictive models as part of MTL has been previously utilized for drug-sensitivity prediction (Chen et al., 2022) as well as survival prediction (Vo et al., 2021) separately in the context of cancer omics. In this paper we implement a novel end-to-end deep learning architecture extending the CODE-AE-ADV to include patient survival prediction and a joint learning scheme to learn a latent space that is suitable for prioritizing drugs for groups of patients based on their omics profiles.

Methods

MODAE

Autoencoder

The autoencoder consists of an encoder module $E : \mathbb{R}^d \to \mathbb{R}^p$ and a decoder module $D : \mathbb{R}^p \to \mathbb{R}^d$. We denote the input omics with $\mathbf{x}$, the bottle-neck layer as $\mathbf{z} = E(\mathbf{x})$ and the reconstructed omics as $\hat{\mathbf{x}} = D(\mathbf{z})$. The purpose of the autoencoder is to find a low dimensional representation $\mathbf{z} \in \mathbb{R}^p$ of the data in an unsupervied manner by minimizing the reconstruction loss

$$L_{\text{rec}}(\mathbf{X}) = \frac{1}{nd} \sum_{i=1}^{n} \left\| \hat{\mathbf{x}}^{(\mathbf{i})} - \mathbf{x}^{(i)} \right\|^2 \tag{1}$$

The input dimension d is included in the denominator in order to normalize the scales of different objectives for the multi-task learning step (Equation 8).

Adversarial deconfounding autoencoder

To separate learn separate representations for confounders, we split the vector $\mathbf{z}$ into three parts: $\mathbf{z}_s$ (shared), $\mathbf{z}_p$ (patient tissue), and $\mathbf{z}_c$ (cell-line). In contrast to He et al. (2022), we use a single encoder to learn all these representations by setting parts of $\mathbf{z}$ to $\mathbf{0}$ by applying the following function f before the decoder (i.e., $\hat{\mathbf{x}} = D(f(\mathbf{z}))$):

$$f_j(z_j, s) = \begin{cases} z_j & \text{if } j \in U \\ z_j & \text{if } j \in V \wedge s = 0 \\ z_j & \text{if } j \in W \wedge s = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where U is the set of indices used for the shared representation, V is the set of indices for patients, and W is the same for cell-lines, and where $s \in \{0, 1\}$ is an indicator for whether a given sample is from a cell-line (and not from patient tissue). Supervised tasks use the shared part and we denote $q = |U|$ as its dimension. An adversary aims to distinguish cell-lines from patients and the WGAN inspired autoencoder is forced to learn representations that minimize differences between the two data domains, which should lead to more reliable transfer of predictions. The adversary and the autoencoder are trained in alternating steps. In the first step the critic network $C : \mathbb{R}^q \to \mathbb{R}$ is trained to minimize the adversarial loss

$$L_{\text{adv}}(\mathbf{Z}_s, \mathbf{s}) = \frac{1}{n} \sum_{i=1}^{n} C(\mathbf{z}_s^{(i)})(1 - 2s_i). \tag{3}$$

Then, the autoencoder is trained to minimize the sum of L_{rec} and the deconfounding loss

$$L_{\text{dc}}(\mathbf{Z}_s, \mathbf{s}) = \frac{1}{\sum_{i=1}^{n} s_i} \sum_{i=1}^{n} C(\mathbf{z}_s^{(i)}) s_i \tag{4}$$

which is equivalent to the cell-line specific term of L_{adv} with the sign reversed. L_{adv} is an approximation of the Wasserstein distance between the distributions of the two sets of points in the latent representation (Arjovsky et al., 2017). In simpler terms, the critic or dataset detector network C is trained to estimate the patient-likeness of a given sample while the deconfounding autoencoder tries to minimize the distance of the cell-line sample distribution from the patient sample distribution. Furthermore, in order to satisfy the continuity assumptions of the WGAN model, we use a gradient penalty on $\nabla C(\mathbf{z}_s)$ similarly to He et al. (2022). However, our formulation of L_{adv} and L_{bc} differs from the equivalent $\mathcal{L}_{\text{critic}}$ and $\mathcal{L}_{\text{gen}}$ in that we use the shared representation $\mathbf{z}_s$ while He et al. (2022) use a concatenation of shared and private representations and in our case the deconfounding loss is calculated on cell-lines instead of patient tissues.

Deep Cox regression for survival modeling

A deep learning extension of the non-parametric Cox regression model can be obtained with the the hazard function:

$$\lambda(\mathbf{x}) \propto e^{g(\mathbf{x})} = e^{A(f_s(E(\mathbf{x})), \mathbf{k})}, \tag{5}$$

where $A : \mathbb{R}^q \to \mathbb{R}$ is a survival risk prediction network, $f_s : \mathbb{R}^p \to \mathbb{R}^q$ is a function that extracts the shared representation from $\mathbf{z}$, and $\mathbf{k}$ is an optional vector of additional input covariates, such as patient age or sex. We used Efron's method (Efron, 1977) to account for tied survival times which can be quite common for large datasets or aggressive cancers. Hence, the loss function for training A is based on the adjusted Cox partial log-likelihood

$$L_{\text{surv}} = \frac{1}{\sum_{h \in H} |h|} \sum_{i=1}^{|H|} \left[\sum_{j \in h_i} g(\mathbf{x}^{(j)}) - r(X, h_i, T) \right], \quad (6)$$

$$r(X, h_i, T) = \sum_{l=0}^{|h_i|-1} \ln \left(\sum_{j:T_j \geq t_i} e^{g(\mathbf{x}^{(j)})} - \frac{l}{|h_i|} \sum_{j \in h_i} e^{g(\mathbf{x}^{(j)})} \right),$$

where $h_i \in H$ is the set of indices of all events for which $T_j = t_i \forall j \in h_i$, H contains the sets of events for each unique time point t_i, and $\mathbf{x}^{(i)}$ is the vector of omics data for patient i. To improve the numerical stability of the calculations we utilized the log-sum-exp trick. The likelihood was divided by the number of observed events to maintain consistent scaling of the losses.

Deep regression for drug sensitivity modeling

Drug sensitivity prediction was addressed as a multiple output regression problem. Since every drug-cell-line pair was not screened in CTRP, losses were computed on non-missing values only. In other words we minimize the loss

$$L_{\text{drug}} = \frac{1}{|O(\mathbf{R})|} \sum_{i,j \in O(\mathbf{R})} \|\hat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|^2, \quad (7)$$

where $\mathbf{R}$ is a matrix of observed drug data that may contain missing values, O is a function that returns the set of indices of observed cell-line drug pairs, and $\hat{\mathbf{R}}$ contains drug-sensitivity vectors $\mathbf{r}^{(i)} = B(\mathbf{z}_{\mathbf{s}}^{(\mathbf{i})})$ computed for each cell-line i with the drug-sensitivity network $B : \mathbb{R}^q \to \mathbb{R}^m$, where m is the number of drugs.

Joint multi-task learning

To simultaneously learn to represent the molecular state, to find a shared deconfounded latent space, and to predict both patient survival and cell-line drug sensitivity from omics data we define the total loss as a weighted sum of the losses used for individual tasks as follows:

$$L_{\text{MTL}} = \alpha_1 L_{\text{rec}} + \alpha_2 L_{\text{dc}} + \alpha_3 L_{\text{surv}} + \alpha_4 L_{\text{drug}}. \quad (8)$$

The weights α_{1-4} can be used to adjust the task priority and were tuned as hyper-parameters to explore different trade-offs as described in section 3.2.

To ensure robust training for the multi-task model, we implemented a pre-training procedure for each network module that corresponds to a specific task (Supplementary Algorithm 1). The autoencoder, critic network, deconfounding autoencoder, patient-survival network, and drug-sensitivity network can be trained in order before running the joint learning that updates the full model at each step (Supplementary Algorithm 2). In practice, skipping the first two steps and starting from the deconfounding autoencoder training worked better than pre-training each step. Pre-training of the deconfounding autoencoder was continued for 300 epochs, followed by 100 epochs each of patient survival and cell-line drug-sensitivity module training, after which the joint multi-task training was run for 300 epochs. Algorithm 1 shows the joint learning procedure which samples equally sized mini-batches from both datasets and updates the weights of all MODAE modules based on Equation 8. Since the number of patients and cell-lines can be different, we employed an oversampling strategy to generate the same number of mini-batches from both sample types for each epoch. Supplementary Figure S1 shows the constituent losses for each epoch across training phases and cross-validation folds for a model selected by parameter search. We used the scaled exponential linear units (Klambauer et al., 2017) to increase the robustness of the joint model in place of more conventional batch-normalization. Furthermore, L2-norm regularization was applied on all the model weights, and dropout was applied after all layers except outputs and bottle-neck. The critic network C was trained in epochs that alternate with the autoencoder epoch in pre-training and multi-task epoch in joint training. Additionally, for each epoch of the deconfounding and joint training the adversary was trained for 5 epochs and used the Adam optimizer with recommended momentum parameters $\beta_1 = 0, \beta_2 = 0.9$ as suggested by Arjovsky et al. (2017); Gulrajani et al. (2017).

Data and pre-processing

To perform a case-study on real patient data, we used cancer tissue data from the Swedish SCANB breast cancer cohort (Staaf et al., 2022) as a discovery dataset and the TCGA BRCA cohort as a validation dataset. To simplify the analysis, we focused on RNA-Seq data since the transcriptome has been shown to be the most relevant omic for drug sensitivity prediction (Costello et al., 2014). The SCANB datat was already pre-processed to fragments per kilobase per million (FPKM) while pre-processed TCGA data downloaded with the *curatedTCGAData* R-package (Ramos et al., 2020) was in upper-quartile (UQ) normalized transcripts per million (TPM). To harmonize them, SCANB data was normalized to UQ TPM using the *edgeR* R-package and then both were log-transformed to $\log_2(\text{TPM} + 1)$. Cell-line transcriptome profiles were acquired from the DepMap portal (Corsello et al., 2020) which were already equivalently processed. Genes with average log-expression less than 1.0 in either the patient and cell-line dataset were excluded and the values in each dataset were independently centered and scaled to mean 0 and standard deviation 1. This yielded gene-expression datasets of 12021 genes, 9206 patients for discovery, 1093 patients for validation, and 1293 cell-lines for discovery.

Algorithm 1 Multi-task training of MODAE

Input: $\mathbf{X}_p$ patient tissue omics,
$\mathbf{X}_c$ cell-line omics,
$\mathbf{t}$ patient survival time,
$\mathbf{e}$ patient survival event indicator,
$\mathbf{K}$ patient clinical data (e.g. age),
$\mathbf{R}$ cell-line drug-sensitivity matrix

Require: N_{epochs}; network modules: A, B, C, D, E

```
for epoch = 1 to N_epochs do
    for X_mini-p, X_mini-c, t_b, e_b, K_b, R_b ∈ mini-batches(X_p, X_c, t, e, K, R) do
        Z_p = E(X_mini-p)
        Z_c = E(X_mini-c)
        Z = concatenate_samples(Z_p, Z_c)
        s is an indicator vector for z^(i) ∈ Z_c
        Z_augmented = concatenate_features(Z_p, K_b)
        Update A, B, D, E with L_MTL
    end for
    for adversarial epoch = 1 to N_adv_per_AE do
        run CRITIC TRAINING EPOCH
    end for
end for
```

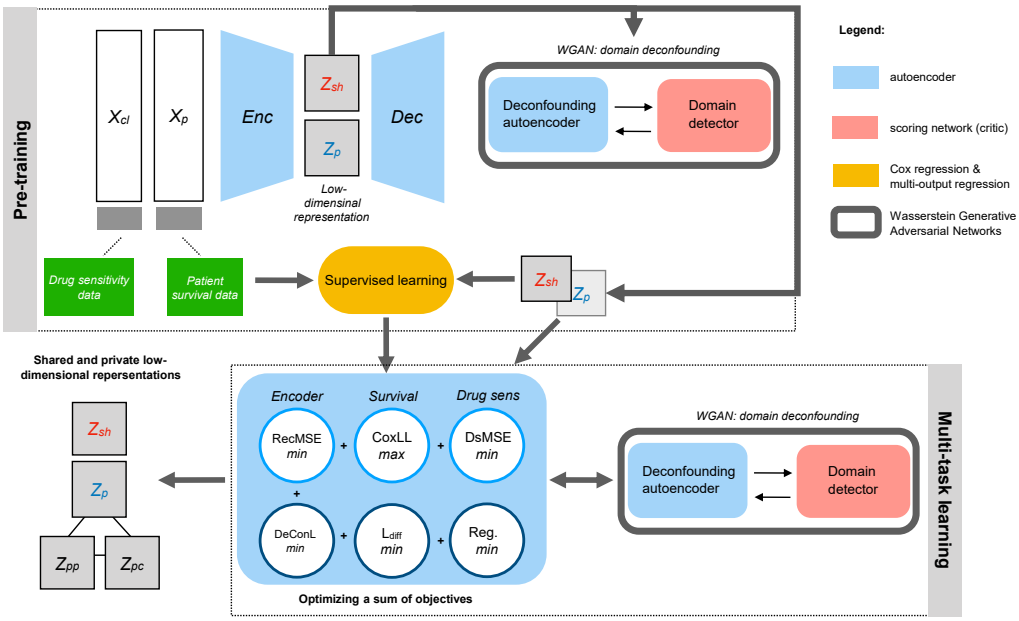

Fig. 1. MODAE model diagram showing the essential structure and training steps. Cell-line omics profiles X_{cl} and patient omics profiles X_p are first encoded into two parts, the shared representation z_{sh} and private representation z_p which are further divided into private cell-line and private patient representations z_{pc} and z_{pp}, respectively. The WGAN-based adversarial training adjusts the autoencoder with the help of a domain detector. Deconfounded shared representations are used for supervised training using drug-sensitivity and survival data. In a final training stage all tasks are trained simultaneously by optimizing a sum of objectives.

Survival data was obtained from Liu et al. (2018) who re-analysed TCGA clinical data to provide more accurate survival information. For drug sensitivity, we used CTRP which contains one of the most comprehensive drug-sensitivity datasets spanning 545 different drugs screened across 841 cell-lines of which 755 could be matched with transcriptomic profiles in CCLE (Seashore-Ludlow et al., 2015; Basu et al., 2013). CTRP data was downloaded using the *PharmacoGx* R-package (Smirnov et al., 2016). We decided to use the area above the curve (AAC) as the target drug-sensitivity value, since it is bounded in the range $[0, 1]$ and is defined even when some other common alternatives, such as IC50, are not. Performance was evaluated with 5-fold cross-validation stratified with respect to the survival event and time in the case of patient data. Since the Cox model depends on the order of the events (Equation 6), event times were split into 10 equally sized quantiles after which the samples were evenly distributed into the folds based on the time quantile and event status. In order to account for the differences in time elapsed since TCGA and SCANB studies were conducted, the maximum follow-up time was set to 3000.

Results

Overview of MODAE

We developed MODAE to integrate and deconfound patient tissue and cell-line samples for survival and drug sensitivity related feature extraction in a hybrid unsupervised-supervised fashion (See Figure 1). Since survival and drug sensitivity are specific to their corresponding data domain, task specific models must be adapted to new target domains. MODAE achieves this by learning a shared representation of the molecular data which the task specific models can influence to mold it to be more useful for survival and drug sensitivity prediction. Differently from CODE-AE, we treat drug sensitivity prediction as a regression problem and perform multi-task learning with supervised tasks together with the reconstruction error.

Trade-offs between evaluation metrics in random hyper-parameter search

To search for the optimal hyper-parameters of MODAE, 100 iterations of random parameter search were run. For the internal evaluation using 5-fold cross-validation, the parameter search was carried out in a nested fashion by using one of the training folds as a validation fold. Then for external validation, the hyper-parameters were selected based on the 5-fold cross-validation results. In either case, the selection of the best hyper-parameters is a multi-objective optimization problem. MODAE aims to optimize four different objectives: omics profile reconstruction, decondounding, survival prediction, and drug sensitivity prediction. The random parameter-search was intended to explore different trade-offs between these objectives by adjusting the learning rate, regularization, and number of hidden units associated with each module of MODAE as well as the weights assigned to the different objectives in Equation 8. Figure 2 shows the 5-fold average performance of each hyper-parameter setting: reconstruction error in negative $log_{10}MSE$ (separate for patient and cell-line data), deconfounding performance in terms of AUROC achieved by a linear SVM in distinguishing the shared embeddings corresponding to different datasets, survival concordance in Harrell's C-index (Harrell, 1982), and 90th percentile sensitivity prediction in R^2. We evaluated each drug prediction separately and focus only on top 10% of drug sensitivity predictions because we wanted to focus on drugs for which the models could explain a large portion of variance across the cell-lines.

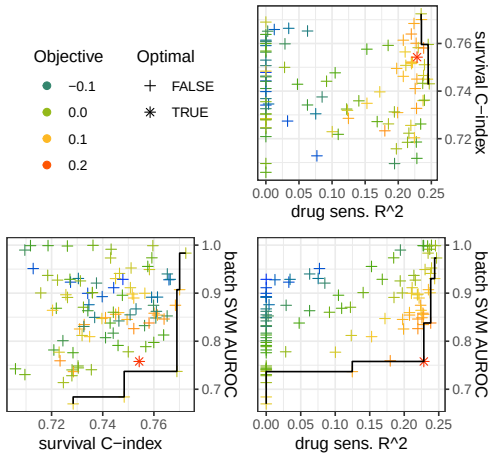

Fig. 2. Pairwise scatter plots of key evaluation metrics. Each point represents one random parameter search result. The points are colored by the average sum of the test set survival concordance-index, top 10% drug sensitivity R^2 quantile, and batch-effect defined as negative AUROC of a data domain detecting linear SVM. Drug-sensitivity performance in the plot was truncated at 0 since a portion of settings resulted in negative R^2 suggesting a failure to learn useful models. The black line goes through points that are on the first Pareto front of each pairwise comparison. The optimal setting used for downstream analysis is highlighted by using an asterisk in place of a cross.

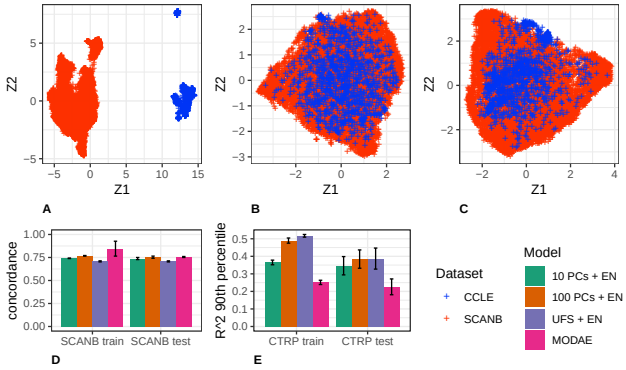

Fig. 3. **A)** UMAP visualization of the concatenated standardized gene-expression matrices corresponding to SCANB and CCLE. **B)** UMAP visualization of the shared embeddings of SCANB and CCLE samples after pre-training the deconfounding autoencoder. **C)** UMAP visualization of the shared embeddings of SCANB and CCLE samples after joint multi-task learning. **D)** Mean survival concordance index of the proposed model and baselines cross-validated in SCANB. **E)** Mean of the 90th percentile R^2 of 545 drug sensitivity predictions cross validated in CTRP. Error bars represent standard deviation. PC: principal component; EN: elastic net regression; UFS: univariate feature selection; R^2: coefficient of determination.

Settings that were better in at least one metric compared to the others form a Pareto front which indicates results with the best trade-offs. The model corresponding to the highest sum of these metrics was compared to baseline models in cross-validation and external datasets.

Baseline models

As baseline, we used elastic net regression with simple feature selection and PCA. For drug-sensitivity prediction, each baseline drug model was trained independently and implemented with the *scikit-learn* Python module (Pedregosa et al., 2011). For survival we used the penalized Cox model implementation provided in the *scikit-survival* Python module (Pölsterl, 2020). The elastic net penalty and L1-ratio hyper-parameters were optimized via grid-search in nested cross-validation with the ranges $[10^{-3}, 10^{-1}]$ and $[0.1, 0.99]$, respectively. Feature selection was based on univariate regression models and the top-k approach, where the number of top predictors k was tuned as a hyper-parameter in the range $[100, 1000]$. These models were trained independently of each other and are therefore not considered multi-task learners, but their purpose is to provide context for the performance of MODAE.

MODAE embeddings and performance comparison

Figures 3A-C show UMAP (McInnes et al., 2018) visualizations of the differences between cell-lines and patients in the original gene-expression, the deconfounded shared representation learned in pre-training, and the final shared representation learned while compromising between multiple tasks. The first plot shows a clear difference between the datasets with no sample from either dataset overlapping the other. Meanwhile the second plot demonstrates the power of the deconfounding autoencoder where the samples from CCLE evenly overlap the samples from SCANB. The third plot shows that the compromise with multi-task learning does not unravel the deconfounding effect learned during pre-training. Here all results are evaluated with the nested parameter search (i.e. the selected hyper-parameters can be different between folds). Figure 3D shows the survival concordance index of the models cross-validated in TCGA. All the models achieved concordance index around 0.75 with PCA based models performing better than univariate feature selection (UFS) and MODAE matching the best model in test set performance while showing slightly higher performance in the training set, but not to a large degree, which suggests that the model is adequately regularized to avoid overfitting. This result shows that MODAE achieves respectable performance even when burdened with deconfounding the predictive representation of differences between patient and cell-line profiles. Figure 3E shows the drug sensitivity prediction performance. Out of the three elastic net models, UFS performed the best, while 100 PCs was better than 10 PCs. MODAE performed significantly worse than the baselines, which suggests that the deconfounding process is hindering drug-sensitivity prediction. The 100 PC and UFS baseline models showed a significant degree of overfitting that was not present in MODAE. While the results shown here represent the average across best settings in each fold, it is consistent with Figure 2 which shows that no individual setting of MODAE performed significantly better than the search average.

Prediction performance on external survival dataset

We used the TCGA BRCA dataset as an external test set for survival predictions to see if the model predictions could be generalized to other datasets. The TCGA BRCA is a much smaller dataset but it has similar overall patient survival rate as the SCANB cohort with 16.3% in TCGA and 16.7% in SCANB. Since TCGA has much longer maximum follow up time, all observations beyond 3000 days were set to censored in both datasets to equalize both studies. Figure 4 shows the Harrell concordance index of MODAE as well as baseline models in the

training set (SCANB) and the test set (TCGA). The external validation results reflect the cross-validation results with 100 PC EN and MODAE performing the best. The MODAE concordance index on the external dataset (0.710) was only slightly lower than on the internal test set (0.755) suggesting some losses in generalization. We also checked the proportional hazard assumption of MODAE by using the *survminer*-package in R. Supplementary Figure S2 shows the summary of tests on the last hidden layer of the survival risk module of MODAE, which indicated that the values observed at the hidden layers were not dependent on time and hence the proportional hazard assumption held for the final model. Due to differences in clinical data not all relevant covariates could be considered while testing the model. Additionally, the concordance index is based on the order of risk scores and does not account for shifts in predictions. Therefore, to further assess the predicted risk scores, we stratified the patients according to the predicted risk scores given by MODAE by calculating three even quantiles (high, mid, low) of the SCANB risk scores and stratifying TCGA patients using the same quantiles. Figure 5 shows the hazard ratios of a CoxPH model fitted with the additional clinical variables and the risk category in SCANB and Figure 6 shows the same for TCGA. In SCANB the group indicators, age z-score, and tumor size z-score were all significant while of the Nottingham histological grading only the third grade was significantly more predictive than the first grade. In TCGA only the high risk group from MODAE and the advanced tumor stages III and IV were significant compared to mid-risk and stage-I respectively. Notably the coefficients associated with the low-risk and high-risk groups were similar between the datasets, but the larger size of the SCANB cohort resulted in more significant p-values.

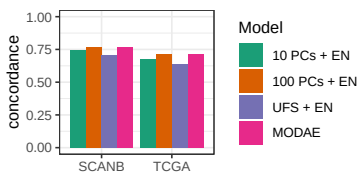

Fig. 4. Survival prediction concordance index of MODAE and baseline models trained with SCANB and tested with TCGA BRCA.

Assessing transfer learning of drug sensitivity predictions

The expression of the target genes modulated by a given drug is expected to be dysregulated. Similarly to He et al. (2022), to assess the agreement of the predictions with this expectation, we divided the patients into resistant and sensitive groups based on the sensitivity predicted by MODAE and compared the target-gene expression between them. The resistant and sensitive groups were defined as sensitivity z-score below -1 or above 1, respectively. 96% of gene-targets were dysregulated in breast cancer patient data with significance level 0.05, which suggests agreement with sensitivity predictions. However, only 22% of them had a fold-change over 1.5, which may suggest some false positives, since the sheer number of patients in SCANB can lend significance even to small differences. Supplementary File 1 contains the log-fold changes between groups and t-test p-values for every drug and their targets listed in DrugBank.

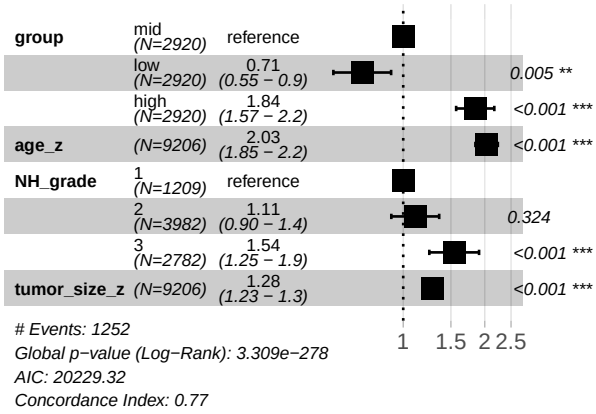

Fig. 5. Hazard ratios of a CoxPH model with additional SCANB clinical variables and risk groups based on MODAE predictions.

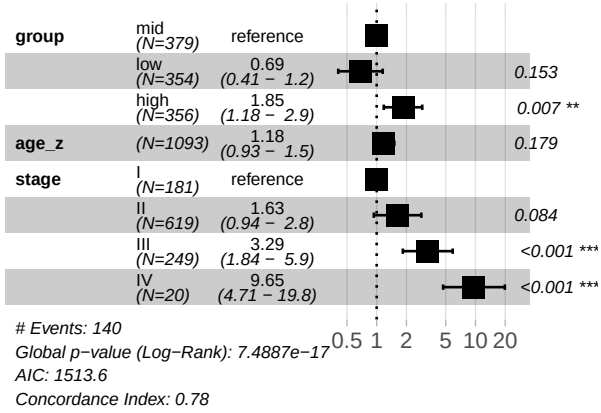

Fig. 6. Hazard ratios of a CoxPH model with additional TCGA BRCA clinical variables and risk groups based on MODAE predictions.

Discussion

In this article we introduced a novel deep learning model to learn latent representations of omics data for predicting cancer patient survival and transfer of drug sensitivity predictions from cell-line screens to patients while adjusting for differences between the molecular profiles of these two sample types. We applied our model to a case study of breast cancer patients to illustrate its usefulness for precision medicine and patient stratification. The model performed well with survival prediction in cross-validation as well as external validation, suggesting that we succeeded in recapitulating clinically relevant patient information from their genomic profiles. However the drug sensitivity predictions on cell-lines were less accurate compared to simple linear models. A major limitation of this breast cancer case-study is the domain adaptation with pan-cancer cell-lines in the CCLE dataset. We wanted to focus on breast cancer, since the SCANB dataset is an unparalleled resource for studying a single cancer type and TCGA BRCA is the single largest cancer dataset within TCGA. Unfortunately there were not enough breast cancer cell-lines profiles available in CCLE to train reliable drug-sensitivity prediction models. In theory, the deconfounding autoencoder

can learn to represent the other cancer types by using the private embeddings even if the other dataset only contains one type, however, the drug-sensitivity predictions are based on the shared embedding where cell-lines of different cancer types are all forced to a representation similar to breast cancer tissues. While the results are somewhat preliminary, they demonstrate that the model has the potential to aid in the discovery of new precision medicine therapies. In the future we aim to improve the drug-sensitivity model and extend the case-study to pan-cancer TCGA. Another interesting direction would be the use of drug-features (e.g., based on their chemical structures) since this could lead to targeted drug design for specific patient groups. Nevertheless, the proposed model could already be used to prioritize repositioning of drugs for high-risk patients.

Competing interests

The authors declare no competing interests.

Author contributions

Conceptualization and methodology T.J.R., F.N. and V.F. ; software and data preparation T.J.R.; writing and editing, T.J.R. and V.F.; supervision and project administration, V.F. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

Computations were performed on infrastructure hosted by the Bioinformatics Center, University of Eastern Finland and the Finnish IT center for science (CSC). This work was supported by the Academy of Finland [grant numbers 336275, 332510], the Jane and Aatos Erkko Foundation, Sigrid Jusélius Foundation, and the Finnish Cultural Foundation [grant number 00230994].

Software availability

`https://github.com/UEFBiomedicalInformaticsLab/MODAE`

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv*, 2017. doi: 10.48550/arxiv.1701.07875.

J. Barretina, G. Caponigro, N. Stransky et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, mar 2012. ISSN 1476-4687.

A. Basu, N. E. Bodycombe, J. H. Cheah et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, aug 2013.

J. Chen, X. Wang, A. Ma et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nature Communications*, 13(1):6494, oct 2022. ISSN 2041-1723.

L. Chin, J. N. Andersen, and P. A. Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 17(3):297–303, mar 2011.

S. M. Corsello, R. T. Nagari, R. D. Spangler et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, feb 2020. ISSN 2662-1347.

J. C. Costello, L. M. Heiser, E. Georgii et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, dec 2014.

A. B. Dincer, J. D. Janizek, and S.-I. Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Suppl_2): i573–i582, dec 2020. ISSN 1367-4803.

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72 (359):557, sep 1977. ISSN 01621459.

M. Ghandi, F. W. Huang, J. Jané-Valbuena et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, may 2019. ISSN 0028-0836.

I. Gulrajani, F. Ahmed, M. Arjovsky et al. Improved training of wasserstein GANs. *Advances in Neural Information Processing Systems*, 2017.

F. E. Harrell. Evaluating the yield of medical tests. *The Journal of the American Medical Association*, 247(18):2543, may 1982. ISSN 0098-7484.

D. He, Q. Liu, Y. Wu, and L. Xie. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, oct 2022. ISSN 2522-5839.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *arXiv*, jun 2017. doi: 10.48550/arxiv.1706.02515.

J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, apr 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.02.052.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, feb 2018. URL `https://arxiv.org/abs/1802.03426`.

F. Pedregosa, G. Varoquaux, A. Gramfort et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011.

S. Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

M. Ramos, L. Geistlinger, S. Oh et al. Multiomic integration of public oncology databases in bioconductor. *JCO clinical cancer informatics*, 4:958–971, oct 2020.

M. Salvadores, F. Fuster-Tormo, and F. Supek. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*, 6(27), jul 2020. ISSN 2375-2548.

B. Seashore-Ludlow, M. G. Rees, J. H. Cheah et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer discovery*, 5(11):1210–1223, nov 2015.

P. Smirnov, Z. Safikhani, N. El-Hachem et al. PharmacoGx: an r package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, apr 2016.

J. Staaf, J. Häkkinen, C. Hegardt et al. RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer. *NPJ breast cancer*, 8(1):94, aug 2022.

T.-H. Vo, G.-S. Lee, H.-J. Yang et al. Survival prediction of lung cancer using small-size clinical data with a multiple task variational autoencoder. *Electronics*, 10(12):1396, jun 2021. ISSN 2079-9292.

W. Yang, J. Soares, P. Greninger et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(Database issue):D955–61, jan 2013.