processes

MDPI

Article

CrossTx: Cross-Cell-Line Transcriptomic Signature Predictions

Panagiotis Chrysinas [1], Changyou Chen [2] and Rudiyanto Gunawan [1,*]

[1] Department of Chemical and Biological Engineering, University at Buffalo-SUNY, Buffalo, NY 14260, USA; pchrysin@buffalo.edu

[2] Department of Computer Science and Engineering, University at Buffalo-SUNY, Buffalo, NY 14260, USA; changyou@buffalo.edu

[*] Correspondence: rgunawan@buffalo.edu

Abstract: Predicting the cell response to drugs is central to drug discovery, drug repurposing, and personalized medicine. To this end, large datasets of drug signatures have been curated, most notably the Connectivity Map (CMap). A multitude of *in silico* approaches have also been formulated, but strategies for predicting drug signatures in unseen cells—cell lines not in the reference datasets—are still lacking. In this work, we developed a simple-yet-efficacious computational strategy, called CrossTx, for predicting the drug transcriptomic signatures of an unseen target cell line using drug transcriptome data of reference cell lines and unlabeled transcriptome data of the target cells. Our strategy involves the combination of Predictor and Corrector steps. The Predictor generates cell-line-agnostic drug signatures using the reference dataset, while the Corrector produces target-cell-specific drug signatures by projecting the signatures from the Predictor onto the transcriptomic latent space of the target cell line. Testing different Predictor–Corrector functions using the CMap revealed the combination of averaging (Mean) as a Predictor and Principal Component Analysis (PCA) followed by Autoencoder (AE) as a Corrector to be the best. Yet, using Mean as a Predictor and PCA as a Corrector achieved comparatively high accuracy with much lower computational requirements when compared to the best combination.

Keywords: gene expression; drug signature; drug repurposing; principal component analysis; autoencoder

Citation: Chrysinas, P.; Chen, C.; Gunawan, R. CrossTx: Cross-Cell-Line Transcriptomic Signature Predictions. *Processes* **2024**, *12*, 332. https://doi.org/10.3390/pr12020332

Academic Editor: Tong Gao

Received: 29 December 2023
Revised: 29 January 2024
Accepted: 31 January 2024
Published: 3 February 2024

1. Introduction

A critical step in the drug discovery process is identifying compounds that are able to elicit a desired activity toward disease-modifying targets. In this regard, data-driven strategies play an important role in mining and integrating the literature, knowledge base, and omics data (e.g., transcriptome) for prioritizing and matching drugs to molecular targets [1–5]. These strategies typically require an abundance of cellular signatures of drugs, preferably those from the specific human cell types or tissues that are affected by the disease. Of note is the Connectivity Map (CMap) dataset that contains 1.5 million human transcriptomic signatures from roughly 20,000 drug treatments and chemical perturbations for 71 different cell lines [6]. Although impressive in terms of its size, the majority of drug signatures in the CMap dataset are from immortalized cancer cell lines. Cancer cells are known to exhibit abnormalities in drug responses. Even among cancer cells, there exists a significant variability of molecular signatures in terms of their responses to drugs [6]. Complicating the matter further, cells' responses to drugs are known to be context-specific [7,8]. The question of whether the CMap signatures can be used to accurately predict drug responses in cell types and tissues of interest that are not among those in the dataset (i.e., unseen cells) is still unanswered. Despite the practical relevance of predicting drug response in an unseen cell line, this problem has not received much attention.

The drug signature prediction method of interest is illustrated in Figure 1a. The task involves predicting transcriptomic drug signatures in an unseen target cell line using drug signatures from reference cell lines and transcriptome data from the target cells. The transcriptome data of the target cells are considered unlabeled (i.e., no information about the original experiments is required). Thus, such data can be compiled from the literature and public databases (e.g., GEO database [9]) as well as from one's own experiments. The prediction in Figure 1a is akin to data imputation but with a key difference that makes most of the existing imputation algorithms previously developed for the CMap inapplicable: the data from the target cell line are unlabeled, and they do not necessarily come from drug treatments.

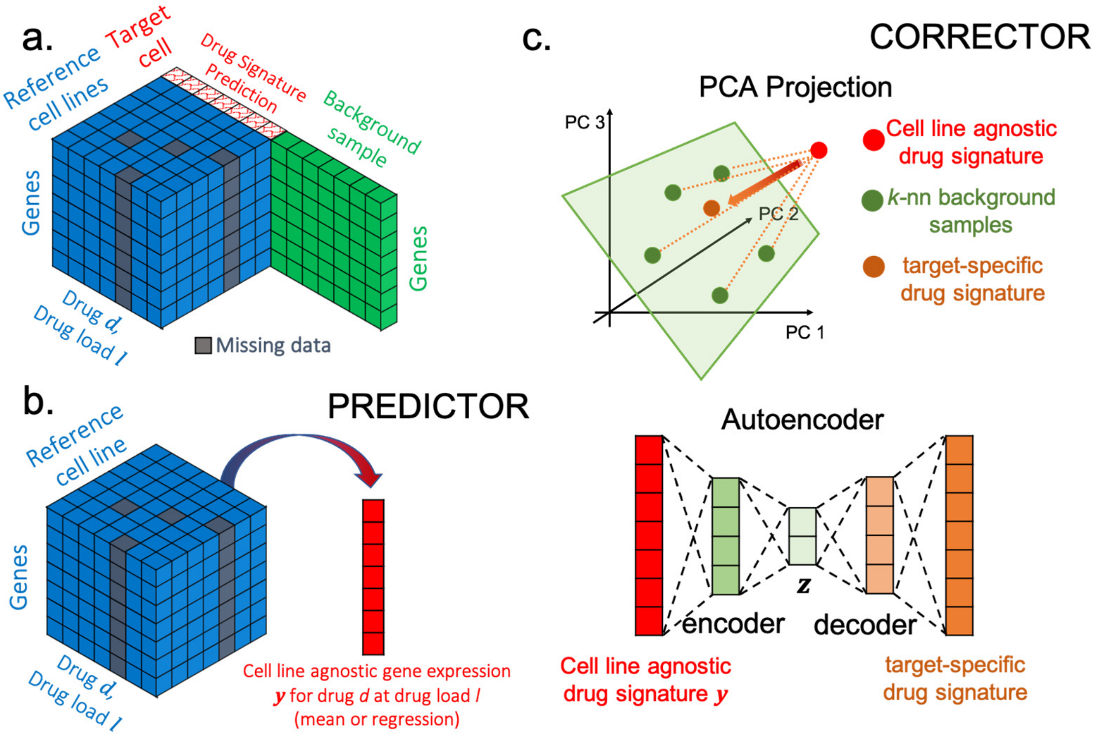

Figure 1. Transcriptomic drug signature prediction by CrossTx. (**a**). An overview of drug signature prediction in an unseen target cell line. Reference transcriptome signatures comprising gene expression data from different reference cell lines treated with various drugs and drug loads (shaded in blue) are displayed as a 3-D matrix. The combinations of drugs and drug loads may not be the same across the reference cell lines, displayed as missing data samples (shaded in grey). Background gene expression data of the target cell line (shaded in green) do not have any overlapping conditions (drug treatments) with the reference signatures. The drug signature slated to be predicted is highlighted in red. (**b**). The Predictor generates cell-line-agnostic gene expression via averaging or through a linear regression model. (**c**). The Corrector projects cell-line-agnostic signatures from the Predictor using PCA and/or AE, producing target-specific gene expression signatures.

The computational methods for drug signature prediction that have previously been developed and applied to the CMap dataset address a more common imputation problem; this problem is illustrated in Figure 1a by the 3-D matrix with missing samples. Several imputation strategies rely on the 3-D matrix representation of the CMap drug signature data, where genes, cell lines, and drug conditions make up the three axes. The simplest approach is averaging, which can be either 1-D (across reference cell lines for the same drug) or 2-D (across drugs from the same cell line and across reference cell lines for the same drug) [10]. More sophisticated strategies rely on tensor decompositions of a 3-D matrix [10,11]. Notably, the tensor decomposition method called TT-WOPT (Tensor-Train Weight Optimization) is able to generate drug signatures for unseen cells (referred to as 'missing cells' in the original publication) [11]. Another group of methods are based on a supervised learning strategy wherein a linear regression model or machine learning is trained to produce drug signatures of a target cell line given drug signatures from the reference cell lines. Some examples include GeneDNN, GeneGAN, GeneLASSO, and SampleLASSO [12]. GeneDNN and GeneGAN address a gene-wise data imputation, a wherein a deep neural network and a generative adversarial network model, respectively, are built for imputing missing expression data of select genes. Finally, GeneLASSO and SampleLASSO rely on LASSO regression modeling [13] to impute missing expression data for genes and samples, respectively. These supervised learning-based methods require training data that comprise drug signatures of the reference and target cell lines obtained under the same set of conditions.

Except for the TT-WOPT method, the drug signature prediction considered in this work cannot be addressed by the imputation methods mentioned above since there are no overlapping conditions between the reference dataset and the transcriptomic data of the target cells. Also, in the case of unseen cells, the TT-WOPT method was previously shown to yield poor accuracy [11]. To address the lack of viable algorithms, in this work, we developed CrossTx for cross-cell-line transcriptomic signature prediction. Considering the features of the CMap dataset, we considered drug load as a covariate in generating drug signature predictions. CrossTx is a simple-yet-efficacious method that involves two steps: Predictor and Corrector. Given a drug at a specific drug load, the Predictor produces cell-line-agnostic drug signatures by averaging or conducting a regression of the reference drug signatures. The Corrector is based on the idea of correcting cell-line-agnostic signatures from the Predictor by projecting those onto the transcriptomic latent space of the target cell line and in essence producing target-specific drug signatures. The transcriptomic latent space is constructed from the (unlabeled) gene expression data of the target cell line using Principal Component Analysis (PCA) and/or an Autoencoder (AE). Borrowing from machine learning, this latent space refers to the compressed representation of transcriptomic data [14]. For demonstration, we applied different Predictor–Corrector combinations to the CMap dataset using a Leave-One-Cell-line-Out (LOCO) procedure and assessed the accuracy of the predicted drug signatures using the Pearson correlation coefficient (ρ) and area under the receiver operating characteristics and precision–recall curve (AUROC and AUPR, respectively).

2. Materials and Methods

2.1. Prediction of Drug Signatures by CrossTx

The CrossTx comprises two components, a predictor and a corrector, that are applied in sequence. The Predictor (see Figure 1b) generates a cell-line-agnostic transcriptome signature for a given drug at a specific drug load—defined as the drug concentration multiplied by the treatment duration—via averaging or a linear regression model, hereon referred to as the ‘Mean’ and ‘Regression’ methods, respectively. Here, the same transcriptome signature is generated for a given drug and drug load regardless of the target cell lines. The Corrector subsequently projects the cell-line-agnostic drug signature from the Predictor to the transcriptomic latent space of the target cell line using PCA and/or AE, producing a target-specific signature (see Figure 1c). The details of the Predictor and Corrector steps are detailed below.

Given a drug and drug load, the Predictor produces a vector of cell-line-agnostic gene expression $\boldsymbol{y} \in \mathbb{R}^m$, where m is the number of genes. The Mean method produces the gene expression value $\boldsymbol{y} = \boldsymbol{\mu}_{d,l}$ by averaging the gene expression vectors in the reference dataset for the specified drug d and drug load l (see Figure 1b). An obvious drawback of the Mean method is that it is unable to produce any predictions when the reference dataset does not have samples for a drug for a specified drug load. In such a case, the Regression method should be used.

The Regression method relies on a linear regression model $r_{g,d}(l)$, in which the expression of gene g is the dependent variable and the drug load l is the independent variable, according to the following formulation:

$$r_{g,d}(l) = \beta_1^{g,d} l + \beta_0^{g,d}, \tag{1}$$

where $\beta_1^{g,d}$ and $\beta_0^{g,d}$ are the slope and intercept, respectively. Note that a linear regression model is built for each gene–drug g-d combination. The outcome is arranged into a vector of gene expression $\boldsymbol{y}$. In the application to the CMap dataset, the intercept was set to zero since the drug signatures in the dataset were normalized to have a zero mean (z-scores). The unknown slope was obtained using ordinary least squares method [15].

The first Corrector uses PCA as the projection method. The transcriptomic latent space of the target cell line is anticipated to be highly complex, limiting the ability of PCA to represent the complete latent space well. As illustrated in Figure 1c, only a subset of the gene expression data of the target cell line—samples nearest to the cell-line-agnostic expression from the Predictor $\boldsymbol{y}$—is used in the PCA projection. Specifically, a subset of k transcriptional profiles from the target cell line with the highest correlations with $\boldsymbol{y}$ are chosen for the projection (default k = 5). PCA is applied to the selected expression profiles from the target cells, and the top p principal components (PCs) are finally used for projection. The parameter p is set to the minimum number of PCs that cumulatively explain at least a given percentage of the total variance (default 80%). PCA projection onto the latent space is performed as follows:

$$\hat{\boldsymbol{y}}_{PCA} = \boldsymbol{W}\boldsymbol{W}^T(\boldsymbol{y} - \boldsymbol{m}) + \boldsymbol{m}, \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{m \times p}$ denotes the loading matrix for the selected PCs, $\boldsymbol{m} \in \mathbb{R}^m$ denotes the average expression of the selected target transcriptome profiles, and $\hat{\boldsymbol{y}}_{PCA}$ denotes the target-specific drug signature via PCA projection.

The second Corrector using an AE employs the entire set of transcriptome data to ascertain the transcriptomic latent representation of the target cell line. Autoencoders (AEs) and their many variants have been successfully applied to transcriptomic data for various purposes, ranging from dimensionality reduction to the imputation of missing values [16–19]. The AE architecture comprises an encoder and a decoder (see Figure 1c); each is a perceptron with one hidden layer using the same number of hidden nodes. In CrossTx, the AE employs the hyperbolic tangent (tanh) activation function for both the encoder and the decoder, and thus is able to account for non-linearity in the latent space projection. The number of nodes in the hidden layer n_{nodes}, the dimension of the latent representation n_z, the batch size n_{batch} (i.e., the size of data the subgrouping during the training), and the number of epochs n_{epoch} (i.e., the number of iterations during training) are hyperparameters that need to be optimized. The encoder and the decoder are trained together to minimize the Mean Squared Error Φ

$$\Phi = \frac{1}{N}\sum_{j=1}^{N} \left\| \widetilde{\boldsymbol{x}}_j - \boldsymbol{x}_j \right\|_{\boldsymbol{2}}^{\boldsymbol{2}}, \tag{3}$$

which has successfully been used for building AEs for gene expression data in other studies [20–22]. Here, $\boldsymbol{x}_j \in \mathbb{R}^m$ denotes a vector of gene expression of the target cell line, $\widetilde{\boldsymbol{x}}_j \in \mathbb{R}^m$ denotes the gene expression reconstructed by the AE, N is the number of samples, and $\|\cdot\|_{\boldsymbol{2}}$ denotes the L-2 norm. The output of the Predictor $\boldsymbol{y}$ is passed through the trained AE, as illustrated in Figure 1c, to produce the target-specific gene expression $\hat{\boldsymbol{y}}_{AE}$.

2.2. Application to the CMap Dataset

The CMap drug transcriptomic signature dataset comprises gene expression data for 978 landmark genes measured using the L1000 assay [6]. For assessing the performance of the proposed CrossTx, we employed the processed the CMap drug signatures from a previous study by Pham et al. [19]. This dataset was generated using a peak deconvolution procedure based on a Bayesian-based approach that generates more robust expression profiles [23]. This dataset includes seven cell lines with the most samples in the CMap, namely, MCF7, A375, HT29, PC3, HA1E, YAPC, and HELA, and samples from the six most-frequent drug concentrations taken after 24 h of treatment. The sample sizes for each cell line are given in Table 1.

Table 1. Sample sizes of the CMap dataset in CrossTx assessment.

Size	MCF7	A375	HT29	PC3	HA1E	YAPC	HELA
Total:	1000	817	596	833	792	456	800
Test [1]:							
Mean/Regression	250/253	354/377	263/272	240/246	267/280	209/210	173/174

[1] The number of test samples for the Mean method is different from that for the Regression method since the Mean method requires at least one sample in the reference data for a given drug load.

To train the AE-based corrector, we performed hyperparameter tuning using the Bayesian optimization method [24] and identified the following optimal settings: n_{nodes} of 150, n_z of 100, n_{batch} of 40, and n_{epoch} of 300. We trained the AE using the Adam optimizer [25] with a learning rate of 10^{-3}. The AE models were built using Keras (version 2.10.0) [26] with a TensorFlow backend (version 2.10.0) [27].

For the performance assessment, we used the Leave-One-Cell-line-Out (LOCO) procedure. Briefly, we selected one of the cell lines in the dataset to be the target cells, while the remaining six were assigned as the reference cell lines. We applied CrossTx using the CMap data for the reference cell lines for the Predictor and the transcriptome data from the selected target cell line for the Corrector. The above procedure was repeated for each of the cell lines in the dataset. We assessed the drug signature prediction accuracy (see Section 2.3) for the top 100 drugs in the dataset. The test sample size of the drug signature predictions for each target cell line is provided in Table 1. Since the Mean method is unable to produce cell-line-agnostic signature when the reference dataset does not have any samples from the drug at the drug load of interest, the test sample size for the Mean method is expectedly smaller than that for Regression. But the differences in the test sample sizes are small (with an average difference of 2.7%): the largest difference is 6.1% for cell line A375. Note that when predicting the gene expression signature for a drug treatment, all samples of the respective drug were excluded from the Corrector step; that is, the transcriptome data of the target cell line did not include any samples from the investigated drug.

2.3. Performance Scoring

The first accuracy score is the Pearson correlation coefficient ρ. Given the ground truth drug signature $\boldsymbol{y} \in \mathbb{R}^m$ and the predicted signature $\hat{\boldsymbol{y}} \in \mathbb{R}^m$, ρ is calculated as follows:

$$\rho = \frac{\sum_{i=1}^{m} \left(y_i - \mu_{y,i}\right)\left(\hat{y}_i - \mu_{\hat{y},i}\right)}{\sqrt{\sum_{i=1}^{m} \left(y_i - \mu_{y,i}\right)^2}\sqrt{\sum_{i=1}^{m} \left(\hat{y}_i - \mu_{\hat{y},i}\right)^2}}, \tag{4}$$

where $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_{\hat{y}}$ denote the average of the elements of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, respectively (i.e., $\boldsymbol{\mu}_y = \frac{1}{m}\sum_{i=1}^{m} y_i$). A higher ρ value suggests a more accurate prediction.

Another set of accuracy scores are the area under the Receiver Operating Characteristics (AUROC)—the curve of true the positive rate (TPR) vs. the false positive rate (FPR)—and the area under the precision–recall (AUPR) curve. These scores were computed for genes that were up- or downregulated by drugs separately. To compute the AUROC and AUPR of the upregulated (downregulated) genes, a ranked list of genes was created by sorting the genes based on the predicted gene expression in decreasing (increasing) order, spanning from the most positive (negative) to the most negative (positive). The ROC and PR curves were then generated by computing the numbers of true-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) results among the top k genes in the ranked list for an increasing k. TPs correspond to the intersection of the top k genes and the ground-truth gene set (up-downregulated genes), while FPs are those among the top k genes that are not in the ground-truth set. Genes that are not in the top k of the list but are in the ground-truth gene set are the set of FNs. Finally, genes that are not among the top k nor in the ground-truth set are TNs. For ROC, the TPR and FPR are computed as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{5}$$

$$FPR = \frac{FP}{FP + TN}. \tag{6}$$

The ROC is constructed by plotting TPR versus FPR for increasing k.

Meanwhile, for the PR curve, precision and recall are computed according to:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = TPR = \frac{TP}{TP + FN}. \tag{8}$$

As the name suggests, the PR curve is the plot of Precision versus Recall for increasing k. AUROC and AUPR values range between 0 and 1, where a value of 1 corresponds to a perfect prediction. Note that a random (binary) predictor has an expected value of AUROC equal to 0.5. The expected value of AUPR for a random prediction is equal to the proportion of positives (i.e., the number of genes in the ground-truth set over the total number of genes m). When the number genes in the ground truth set is much smaller than the total number of genes—that is, when classes are highly imbalanced—the AUPR is a more informative metric of accuracy than the AUROC as it accounts for the ratio between positives and negatives [28].

3. Results

CrossTx is a method for predicting the drug transcriptional signatures of an unseen target cell line given data from reference cell lines and gene expression data of the target cells. The transcriptome data of the target cells are unlabeled; that is, the conditions used when generating the data are not required nor used in the drug signature prediction. The drug signature prediction of interest is unlike the common imputation since the gene expression data of the reference cell lines and the transcriptome data of the target cell lines, highlighted in blue and green in Figure 1a, respectively, do not share any common conditions. The CrossTx method comprises two steps: a Predictor and a Corrector. The Predictor uses the reference dataset to produce cell-line-agnostic transcriptional signatures via averaging or linear regression (see Figure 1b). The Corrector leverages the transcriptome dataset of target cells to reconstruct the transcriptomic latent space of the target cell line using either PCA and/or an autoencoder (see Figure 1c) and projects the cell-line-agnostic signatures from the Predictor onto this latent space to produce target-specific signatures. In the following, we assess the performance of CrossTx and compare it with TT-WOPT [11]. As noted earlier, most existing imputation methods cannot be applied to the problem described in Figure 1a.

We evaluated the accuracy of the CrossTx predictions for drug signatures in seven cancer cell lines: MCF7, A375, HT29, PC3, HA1E, YAPC, and HELA. Preprocessed and filtered drug signatures were obtained from the study by Pham et al. [19]. For testing, we followed the LOCO procedure (see Section 2.2), where we systematically picked one cell line as the target and treated the remaining six cell lines as the reference. We generated drug signature predictions for 100 drugs with the most samples in the dataset (see Table 1). When making a signature prediction for a drug, we removed all samples of this drug from the transcriptome data of the target cells. For performance scoring, we employed the Pearson correlation ρ and the area under the ROC (AUROC) and the PR (AUPR) curve. When using the AUROC and AUPR, we performed accuracy assessments for the predictions of up- and downregulated genes separately. In total, ten different drug signatures generated by CrossTx were assessed, namely, two cell-line-agnostic signatures using either the Mean or the Regression method and eight different combinations of the Predictor–Corrector: (Mean or Regression) + (PCA or AE or PCA + AE or AE + PCA).

Figure 2 gives the performance scores of CrossTx and TT-WOPT for all drug signature predictions for different target cell lines. The scores for individual target cell lines are given in Table 2 (see Supplementary Tables S1–S5 for the full results). The results show that cell-line-agnostic drug signatures produced using the Mean and Regression methods have reasonable agreement with the ground-truth data, with correlations ρ averaging 0.59/0.55 (Mean/Regression), AUROCs at 0.79/0.77 for up- and 0.80/0.78 for downregulated genes, and AUPRs at 0.65/0.63 for up- and 0.67/0.65 for downregulated genes. In general, the Mean method provides better accuracy than the Regression method (p-values < 10^{-4}, two-sided paired t-test). Still, the Regression method has an advantage over the Mean method in that it can provide predictions for drug load values that are not in the reference dataset. The Corrector using the PCA projection improves the cell-line-agnostic drug signatures from the Mean (Mean + PCA) and Regression methods (Regression + PCA). In contrast, the Corrector using the AE provides improvements only for cell-line-agnostic signatures from the Mean method (Mean + AE). When starting with the relatively poorer cell-line-agnostic signatures from the Regression method, the AE projection (Regression + AE) drug signature predictions exhibit poorer accuracy than the Predictor.

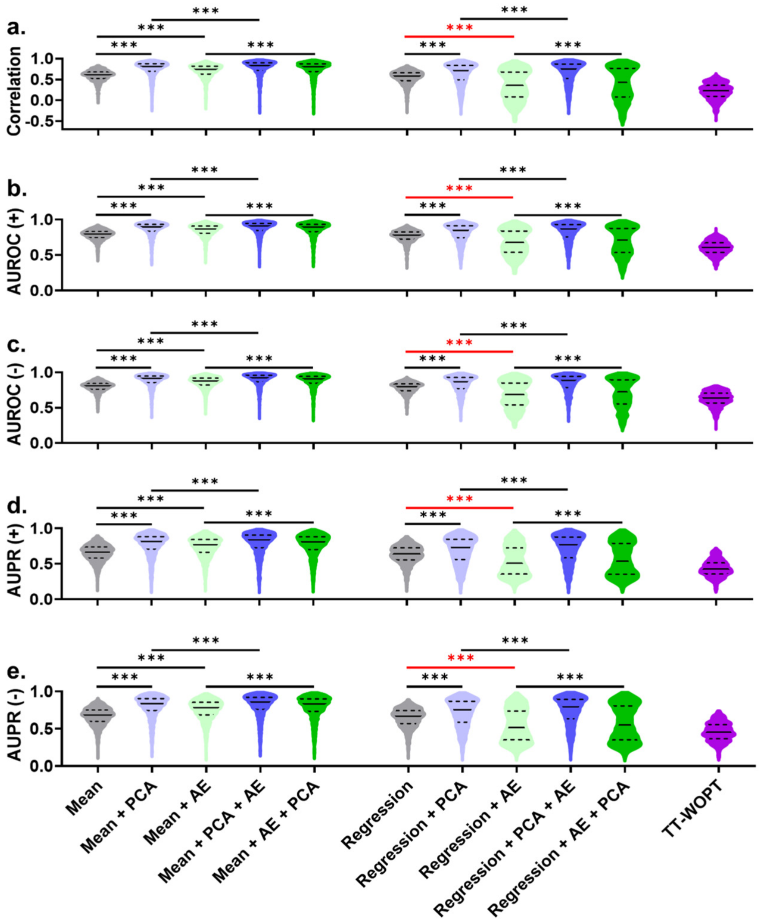

Figure 2. Accuracy of drug signature predictions by CrossTx and TT-WOPT. The accuracy of drug signature predictions was assessed using (**a**) Pearson correlation ρ, (**b**,**c**) AUROC, and (**d**,**e**) AUPR. AUROC (+)/AUROC (−) refer to values for up-/downregulated genes, respectively. Similarly, AUPR (+)/AUPR (−) refer to values for up-/downregulated genes, respectively. Solid lines correspond to medians, and dashed lines denote the inter-quartile range. Statistical significance was determined using two-sided paired t-test. Black *** indicates p-values < 10^{-3} for improved accuracy. Red *** indicates p-values < 10^{-3} for decreased accuracy.

Table 2. Accuracy of drug signature predictions by CrossTx. For each method, the accuracy of drug signature predictions is assessed using Pearson correlation ρ (first row), AUROC (second row), and AUPR (third row). The two values of AUROCs correspond to the accuracy of predicting up-/downregulated genes. Similarly, the two values of AUROCs correspond to the accuracy of predicting up-/downregulated genes. Bold values highlight the best-performing method for the respective accuracy metric. Statistical significance was established via two-sided paired *t*-test to assess the change in accuracy by adding a Corrector. For example, Mean + PCA was compared to Mean, while Mean + PCA + AE was compared to Mean + PCA. Note that the addition of a Corrector may degrade accuracy when using the AE method. *: *p*-values < 0.05.

Cell Method		MCF7	A375	HT29	PC3	HA1E	YAPC	HELA
Predictor: Mean (μ) + Corrector: PCA, AE, PCA + AE, AE + PCA								
μ	ρ AUROC AUPR	0.59 0.78/0.80 0.66/0.65	0.58 0.78/0.79 0.65/0.67	0.6 0.79/0.81 0.66/0.67	0.59 0.79/0.80 0.65/0.65	0.58 0.79/0.79 0.65/0.66	0.58 0.78/0.79 0.65/0.68	0.62 0.80/0.81 0.68/0.69
μ + PCA		0.76 * 0.87 */0.89 * 0.78 */0.79 *	0.77 * 0.88 */0.89 * 0.79 */0.82 *	0.77 * 0.87 */0.89 * 0.79 */0.81 *	0.75 * 0.87 */0.88 * 0.77 */0.78 *	0.74 * 0.86 */0.87 * 0.77 */0.79 *	0.72 * 0.85 */0.87 * 0.75 */0.79 *	0.76 * 0.87 */0.89 * 0.79 */0.81 *
μ + AE		0.72 * 0.85 */0.87 * 0.75 */0.76 *	0.67 * 0.84 */0.84 * 0.73 */0.74 *	0.73 * 0.86 */0.87 * 0.76 */0.78 *	0.71 * 0.85 */0.86 * 0.74 */0.74 *	0.67 * 0.84 */0.84 * 0.72 */0.72 *	0.69 * 0.84 */0.86 * 0.73 */0.76 *	0.76 * 0.88 */0.89 * 0.78 */0.80 *
μ + PCA + AE		**0.78 *** **0.88 */0.90 *** **0.79 */0.81 ***	**0.78 *** **0.89 */0.90 *** **0.80 */0.83 ***	**0.78 *** **0.88 */0.90 *** **0.80 */0.82 ***	**0.77 *** **0.88 */0.89 *** **0.79 */0.80 ***	**0.75 *** **0.87 */0.88 *** **0.78 */0.80 ***	**0.74 *** **0.86 */0.88 *** **0.78 */0.81 ***	**0.79 *** **0.89 */0.90 *** **0.81 */0.83 ***
μ + AE + PCA		0.75 * 0.87 */0.88 * 0.77 */0.79 *	0.76 * 0.87 */0.88 * 0.78 */0.80 *	0.76 * 0.87 */0.89 * 0.78 */0.81 *	0.73 * 0.86/0.87 * 0.76 */0.77 *	0.73 * 0.86 */0.87 * 0.76 */0.78 *	0.69 0.84/0.86 0.74/0.77	0.76 0.87/0.89 0.79/0.80
Predictor: Regression + Corrector: PCA, AE, PCA + AE, AE + PCA								
Regression		0.55 0.77/0.79 0.64/0.63	0.56 0.77/0.78 0.64/0.67	0.56 0.77/0.79 0.63/0.65	0.56 0.77/0.79 0.63/0.64	0.54 0.76/0.77 0.62/0.64	0.52 0.75/0.77 0.61/0.65	0.54 0.76/0.78 0.63/0.64
Regression + PCA		0.62 * 0.80 */0.83 * 0.67 */0.69 *	0.69 * 0.84 */0.86 * 0.74 */0.76 *	0.66 * 0.82 */0.84 * 0.71 */0.73 *	0.61 * 0.80 */0.82 * 0.66 */0.68 *	0.60 * 0.8 */0.81 * 0.66 */0.69 *	0.62 * 0.80 */0.83 * 0.68 */0.72 *	0.63 * 0.81 */0.83 * 0.69 */0.71 *
Regression + AE		0.46 * 0.73 */0.75 * 0.58 */0.59 *	0.25 * 0.63 */0.61 * 0.51 */0.50 *	0.44 * 0.72 */0.73 * 0.58 */0.58 *	0.35 * 0.67 */0.68 * 0.52 */0.52 *	0.23 * 0.62 */0.60 * 0.49 */0.48 *	0.38 * 0.69 */0.70 * 0.55 */0.56 *	0.49 * 0.74 */0.77 * 0.60 */0.63 *
Regression + PCA + AE		**0.66 *** **0.82 */0.85 *** **0.71 */0.73 ***	**0.72 *** **0.86 */0.87 *** **0.77 */0.78 ***	**0.68 *** **0.83 */0.85 *** **0.73 */0.76 ***	**0.63 *** **0.81 */0.83 *** **0.69 */0.70 ***	**0.61 *** **0.80 */0.82 *** **0.69 */0.71 ***	**0.66 *** **0.82 */0.85 *** **0.71 */0.75 ***	**0.66 *** **0.82 */0.85 *** **0.72 */0.74 ***
Regress + AE + PCA		0.45 0.73/0.75 0.58/0.58 *	0.28 * 0.64 */0.63 * 0.54 */0.53 *	0.55 * 0.76 */0.79 * 0.64 */0.66 *	0.39 * 0.69 */0.71 * 0.54 */0.55 *	0.23 0.62/0.61 0.50 */0.49 *	0.47 * 0.73 */0.74 * 0.59 */0.61 *	0.47 0.73/0.77 0.59/0.63
TT-WOPT								
TT-WOPT		0.31 0.65/0.69 0.48/0.50	0.19 0.59/0.60 0.43/0.46	0.18 0.59/0.61 0.43/0.43	0.05 0.52/0.55 0.35/0.37	0.36 0.68/0.69 0.52/0.53	0.18 0.57/0.62 0.41/0.45	0.24 0.62/0.64 0.44/0.48

The observed trend for the AE suggests that the AE projection method can be a viable Corrector strategy when the cell-line-agnostic signatures are relatively near the transcriptomic latent space (manifold) of the target cells. This observation motivated us to combine the two Corrector methods in series. We tested whether the accuracy of the PCA-projected drug signatures could be improved further by a second projection using the AE, a strategy referred to as PCA + AE. The basic premise here is that the PCA projection should have brought the cell-line-agnostic prediction close to the transcriptomic latent space of the target cells, for which the AE may offer further improvements. Indeed, in every target cell line considered (see Table 2) and regardless of the method used to generate the cell-line-agnostic signatures, the PCA + AE combination yielded significant accuracy improvements over the use of PCA alone. This improvement was also notable for every combination using the Regression method as the Predictor (see Table 2). For the sake of completeness, we also tested the AE + PCA combination as the Corrector, passing the cell-line-agnostic signatures to the AE first, and then applying the PCA projection to the output of the AE. In this case, the PCA projection did not always provide accuracy improvements for the drug signatures produced by the AE. Overall, the best-performing method was the Mean + PCA + AE combination. Lastly, the accuracy of the TT-WOPT-generated drug predictions was consistently poorer than that of the CrossTx predictions, even when compared with the drug signatures generated using the Mean method (p-values < 10^{-4}; two-sided paired t-test).

4. Discussion

Our work was motivated by the practical problem of predicting drug signature data in disease-relevant cells. Specifically, we addressed the challenge of leveraging large drug signature datasets, such as the CMap, to predict drug responses in unseen cell lines. Since distinct gene, signaling, and metabolic networks operate in different tissues and cell lines [29–31], one expects that the molecular signatures of a drug will exhibit cell-context specificity. For applications in drug discovery, drug repurposing, and precision medicine, there is a clear need for methods that are able to predict cell-line-specific drug response accurately and not just for cancer cells. To the best of our knowledge, none of the existing imputation algorithms, including those specifically developed for the CMap dataset, are applicable to the problem at hand. We developed CrossTx to address the gap in the available algorithms regarding the drug signature prediction above. Our method comprises two key components: a Predictor that generates cell-line-agnostic drug signatures using the reference data and a Corrector that produces target-specific drug signatures by projecting cell-line-agnostic signatures from the Predictor onto the transcriptomic latent space of the target cell line. Here, we tested two alternative methods for the Predictor, namely, Mean and Regression, and two different latent projections for the Corrector, namely, PCA and AE. However, our two-step strategy is fully generalizable to other algorithms. For example, nonlinear regression models can be employed as the Predictor, and other machine/deep learning models for latent space reconstruction and projection can be adopted as the Corrector. Lastly, while this study focuses on predicting transcriptomic responses to drugs, the task and our proposed method are also relevant for other important applications, for example, predicting aging-related transcriptomic changes in inaccessible or difficult-to-access tissues (e.g., the human brain) using gene expression data from more accessible cells (e.g., blood and adipose tissue). Furthermore, the Predictor–Corrector approach can also be applied to other omics datasets.

We tested the accuracy of ten different predictions produced by CrossTx using various combinations of Predictor–Corrector algorithms. We applied CrossTx to drug signatures from seven cell lines with the most samples in the CMap dataset. We found that averaging reference drug signatures (i.e., the Mean method) is a simple and efficacious method for the Predictor. When no reference samples at the specified drug load exist, the Regression method is a viable alternative method for a Predictor. Between the two Corrector methods, our tests showed that the PCA is superior to the AE not only in terms of the accuracy of the resulting drug signatures but also in its computational simplicity and robustness with respect to the accuracy of its input. As a Corrector, the AE projection induced improvements when the input signatures were close to the transcriptome latent space (manifold) of the target cell line. This observation is not surprising considering that the AE was trained using transcriptional data of the target cells. The combination of PCA and AE, wherein PCA-projected drug signatures were inputted to the AE, yielded the highest accuracy. But the reverse implementation—applying AE projection and then PCA—did not lead to superior performance using only the PCA projection. Finally, all drug signatures generated by CrossTx had higher accuracies than those produced by a tensor-decomposition-based method called TT-WOPT. Here, the TT-WOPT method did not use any transcriptome data of the target cell line for its prediction.

Despite the promising results achieved by CrossTx in predicting drug signatures in unseen cell lines, our approach is subject to certain limitations. Firstly, the efficacy of CrossTx is contingent upon the availability and quality of reference drug signature data. In scenarios where such data are sparse or of low quality, the performance of the Predictor component may be compromised, potentially diminishing the overall accuracy of the drug signature predictions. Secondly, the effectiveness of the Corrector step, which is designed to tailor drug signatures to specific cell lines, is significantly influenced by the characteristics of the transcriptomic latent space. If this latent space fails to capture essential biological variations due to limitations in the available gene expression data or the underlying PCA or AE models, the precision of CrossTx predictions could be adversely affected. Lastly, the emphasis on transcriptomic signatures in our work restricts CrossTx's utility for predicting drug effects that are primarily governed by non-transcriptional mechanisms. Future work could focus on expanding this methodology to include other omics datasets, thereby potentially mitigating these limitations and enhancing the predictive capacity and applicability of this method across a wider spectrum of biological contexts and drugs.

5. Conclusions

In summary, we presented a simple-yet-efficacious Predictor–Corrector strategy, called CrossTx, for cross-cell line/tissue transcriptome signature prediction. The method CrossTx was specifically developed for leveraging large reference datasets of drug transcriptomic signatures, such as the CMap, to predict drug response in unseen cells—cells that are in the reference dataset. When applied to the CMap data, CrossTX was able to produce highly accurate drug signatures, where the best combination was to use the Mean method for the Predictor and the PCA + AE as the Corrector. Nevertheless, the Mean + PCA combination was of note because of its simplicity and low computational cost while also being able to provide drug signature predictions with relatively high accuracy when compared to the best combination above. The simplicity of CrossTx is an advantage, enabling its adaptation to other related applications and the use of other Predictor and Corrector algorithms with little additional effort. CrossTx is available from https://github.com/cabsel/crosstx (accessed on 4 January 2023).

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/pr12020332/s1. Table S1: Pearson correlations of drug signature predictions by CrossTx and TT-WOPT for individual cell lines; Table S2: AUPRs of drug signature predictions by CrossTx using the Mean method as Predictor and TT-WOPT for individual cell lines; Table S3: AUPRs of drug signature predictions by CrossTx using the Regression method as Predictor and TT-WOPT for individual cell lines; Table S4: AUROCs of drug signature predictions by CrossTx using the Mean method as Predictor and TT-WOPT for individual cell lines; Table S5: AUROCs of drug signature predictions by CrossTx using the Regression method as Predictor and TT-WOPT for individual cell lines.

Author Contributions: Conceptualization, P.C. and R.G.; methodology, P.C., C.C., and R.G.; software, P.C.; validation, P.C. and R.G.; formal analysis, P.C.; investigation, P.C, C.C. and R.G.; resources, R.G.; data curation, P.C.; writing—original draft preparation, P.C.; writing—review and editing, R.G. and C.C.; visualization, P.C. and R.G.; supervision, R.G.; project administration, R.G.; funding acquisition, R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by University at Buffalo-SUNY, NSF HDR I-DIRSE-Ideas Labs (grant number 1940162), and National Institutes of Health (grant number 3R01HL103411-10S1).

Data Availability Statement: The L1000 drug signature (CMap) data used in this study were obtained from https://github.com/njpipeorgan/L1000-bayesian (accessed on 1 December 2021) [23].

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Louhimo, R.; Laakso, M.; Belitskin, D.; Klefstrom, J.; Lehtonen, R.; Hautaniemi, S. Data integration to prioritize drugs using genomics and curated data. *BioData Min.* **2016**, *9*, 21. [CrossRef] [PubMed]
2. Dudley, J.T.; Deshpande, T.; Butte, A.J. Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.* **2011**, *12*, 303–311. [CrossRef] [PubMed]
3. Jin, G.; Wong, S.T. Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines. *Drug Discov. Today* **2014**, *19*, 637–644. [CrossRef] [PubMed]
4. Kim, R.S.; Goossens, N.; Hoshida, Y. Use of big data in drug development for precision medicine. *Expert Rev. Precis. Med. Drug Dev.* **2016**, *1*, 245–253. [CrossRef] [PubMed]
5. Qian, T.; Zhu, S.; Hoshida, Y. Use of big data in drug development for precision medicine: An update. *Expert Rev. Precis. Med. Drug Dev.* **2019**, *4*, 189–200. [CrossRef]
6. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171*, 1437–1452e1417. [CrossRef]
7. Wang, Y.-Y.; Kang, H.; Xu, T.; Hao, L.; Bao, Y.; Jia, P. CeDR Atlas: A knowledgebase of cellular drug response. *Nucleic Acids Res.* **2021**, *50*, D1164–D1171. [CrossRef]
8. Zhao, W.; Dovas, A.; Spinazzi, E.F.; Levitin, H.M.; Banu, M.A.; Upadhyayula, P.; Sudhakar, T.; Marie, T.; Otten, M.L.; Sisti, M.B.; et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Med.* **2021**, *13*, 82. [CrossRef]
9. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef]
10. Hodos, R.; Zhang, P.; Lee, H.C.; Duan, Q.; Wang, Z.; Clark, N.R.; Ma'ayan, A.; Wang, F.; Kidd, B.; Hu, J.; et al. Cell-specific prediction and application of drug-induced gene expression profiles. In *Biocomputing 2018: Proceedings of the Pacific Symposium, 2018*; World Scientific Publishing Company: Singapore, 2018; Volume 23, pp. 32–43.
11. Iwata, M.; Yuan, L.; Zhao, Q.; Tabei, Y.; Berenger, F.; Sawada, R.; Akiyoshi, S.; Hamano, M.; Yamanishi, Y. Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm. *Bioinformatics* **2019**, *35*, i191–i199. [CrossRef]
12. Mancuso, C.A.; Canfield, J.L.; Singla, D.; Krishnan, A. A flexible, interpretable, and accurate approach for imputing the expression of unmeasured genes. *Nucleic Acids Res.* **2020**, *48*, e125. [CrossRef]
13. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
14. Liu, Y.; Jun, E.; Li, Q.; Heer, J. Latent Space Cartography: Visual Analysis of Vector Space Embeddings. *Comput. Graph. Forum* **2019**, *38*, 67–78. [CrossRef]
15. Beck, J.V.; Arnold, K.J.; Arnold, K.J. *Parameter Estimation in Engineering and Science*; Wiley: New York, NY, USA, 1977; pp. 213–327, 501.
16. Arisdakessian, C.; Poirion, O.; Yunits, B.; Zhu, X.; Garmire, L.X. DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.* **2019**, *20*, 211. [CrossRef] [PubMed]
17. Wang, D.; Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genom. Proteom. Bioinform.* **2018**, *16*, 320–331. [CrossRef] [PubMed]
18. Way, G.P.; Greene, C.S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Biocomputing 2018: Proceedings of the Pacific Symposium, 2018*; World Scientific Publishing Company: Singapore, 2018; Volume 23, pp. 80–91.
19. Pham, T.H.; Qiu, Y.; Zeng, J.; Xie, L.; Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat. Mach. Intell.* **2021**, *3*, 247–257. [CrossRef] [PubMed]
20. Lotfollahi, M.; Wolf, F.A.; Theis, F.J. scGen predicts single-cell perturbation responses. *Nat. Methods* **2019**, *16*, 715–721. [CrossRef]
21. Xie, R.; Wen, J.; Quitadamo, A.; Cheng, J.; Shi, X. A deep auto-encoder model for gene expression prediction. *BMC Genom.* **2017**, *18*, 845. [CrossRef] [PubMed]
22. Dincer, A.B.; Janizek, J.D.; Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **2020**, *36*, i573–i582. [CrossRef]
23. Qiu, Y.; Lu, T.; Lim, H.; Xie, L. A Bayesian approach to accurate and robust signature detection on LINCS L1000 data. *Bioinformatics* **2020**, *36*, 2787–2795. [CrossRef]
24. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv* **2012**, arXiv:1206.2944.
25. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
26. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 5 December 2021).
27. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Greg, S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.
28. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
29. Marbach, D.; Lamparter, D.; Quon, G.; Kellis, M.; Kutalik, Z.; Bergmann, S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **2016**, *13*, 366–370. [CrossRef] [PubMed]

30. Schultz, A.; Qutub, A.A. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. *PLoS Comput. Biol.* **2016**, *12*, e1004808. [CrossRef] [PubMed]
31. Sharma, S.; Petsalaki, E. Large-scale datasets uncovering cell signalling networks in cancer: Context matters. *Curr. Opin. Genet. Dev.* **2019**, *54*, 118–124. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.