OXFORD

Briefings in Bioinformatics, 2024, **25(6)**, bbae512
https://doi.org/10.1093/bib/bbae512
Problem Solving Protocol

Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping

Zuqi Li[1,2,3,‡], Sonja Katz[4,5,6,‡], Edoardo Saccenti[5], David W. Fardo[7,8], Peter Claes[2,9,10], Vitor A.P. Martins dos Santos[6,11], Kristel Van Steen[1,3], Gennady V. Roshchupkin[2,12,*]

[1]BIO3 - Laboratory for Systems Medicine, Department of Human Genetics, KU Leuven, Herestraat 49, 3000 Leuven, Belgium
[2]Medical Imaging Research Center, University Hospitals Leuven, Herestraat 49, 3000 Leuven, Belgium
[3]BIO3 - Laboratory for Systems Genetics, GIGA Molecular & Computational Biology, University of Liège, Avenue de l'Hôpital 11, 4000 Liège, Belgium
[4]Department of Radiology and Nuclear Medicine, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, Netherlands
[5]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, PO Box 8033, 6700 EJ Wageningen, Netherlands
[6]LifeGlimmer GmbH, Markelstraße 38, 12163 Berlin, Germany
[7]Department of Biostatistics, University of Kentucky, 111 Washington Avenue, Lexington, KY 40536, United States
[8]Sanders-Brown Center on Aging, University of Kentucky, 789 S Limestone, Lexington, KY 40536, United States
[9]Department of Human Genetics, KU Leuven, Herestraat 49, 3000 Leuven, Belgium
[10]Department of Electrical Engineering, ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[11]Laboratory of Bioprocess Engineering, WageningenUniversity & Research, PO Box 16, 6700 AA Wageningen, the Netherlands
[12]Department of Epidemiology, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, Netherlands

*Corresponding author. Department of Epidemiology, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, Netherlands. E-mail: g.roshchupkin@erasmusmc.nl
‡Zuqi Li and Sonja Katz contributed equally.

Abstract

Unsupervised learning, particularly clustering, plays a pivotal role in disease subtyping and patient stratification, especially with the abundance of large-scale multi-omics data. Deep learning models, such as variational autoencoders (VAEs), can enhance clustering algorithms by leveraging inter-individual heterogeneity. However, the impact of confounders—external factors unrelated to the condition, e.g. batch effect or age—on clustering is often overlooked, introducing bias and spurious biological conclusions. In this work, we introduce four novel VAE-based deconfounding frameworks tailored for clustering multi-omics data. These frameworks effectively mitigate confounding effects while preserving genuine biological patterns. The deconfounding strategies employed include (i) removal of latent features correlated with confounders, (ii) a conditional VAE, (iii) adversarial training, and (iv) adding a regularization term to the loss function. Using real-life multi-omics data from The Cancer Genome Atlas, we simulated various confounding effects (linear, nonlinear, categorical, mixed) and assessed model performance across 50 repetitions based on reconstruction error, clustering stability, and deconfounding efficacy. Our results demonstrate that our novel models, particularly the conditional multi-omics VAE (cXVAE), successfully handle simulated confounding effects and recover biologically driven clustering structures. cXVAE accurately identifies patient labels and unveils meaningful pathological associations among cancer types, validating deconfounded representations. Furthermore, our study suggests that some of the proposed strategies, such as adversarial training, prove insufficient in confounder removal. In summary, our study contributes by proposing innovative frameworks for simultaneous multi-omics data integration, dimensionality reduction, and deconfounding in clustering. Benchmarking on open-access data offers guidance to end-users, facilitating meaningful patient stratification for optimized precision medicine.

Keywords: deep learning; autoencoder; multi-omics; confounders; fairness; clustering

Introduction

Unsupervised learning, in particular clustering, focuses on sub-grouping individuals based on their intrinsic data structures, therefore playing an essential role in tasks like disease subtyping and patient stratification. In the realm of biology and medicine, where large-scale multi-omics data, including genomics, transcriptomics, and epigenomics, are prevalent, deep learning models can enhance clustering algorithms. Their ability to reduce the dimensionality of complex data allows clustering algorithms to more effectively explore the heterogeneity between patients. Underscoring the utility of deep learning models, in particular autoencoders, in terms of data integration, dimensionality reduction, and handling a multitude of heterogeneous input data, Simidjievski *et al.* [1] recently benchmarked various variational autoencoder (VAE) models for multi-omics data.

Although patient stratification with deep learning methods are gaining traction in genomic data applications, they are often susceptible to external influences that are unrelated to the condition of interest. One severe limitation is the entanglement of biologically meaningful signals with variables unrelated to the inherent structure that one is interested in, i.e. technical artifacts, random noise from measurements, or other biological factors such as sex, ethnicity, and age (Fig. 1a). These factors, referred to as confounders in the context of unsupervised learning, may

Received: March 12, 2024. **Revised:** September 20, 2024. **Accepted:** September 25, 2024
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

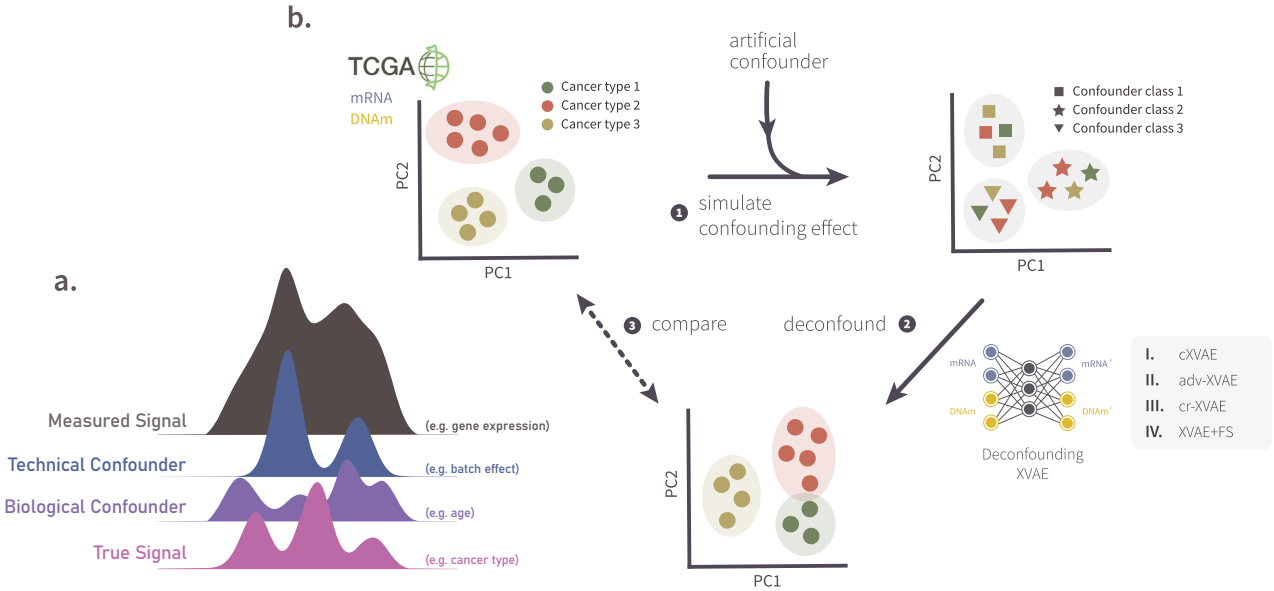

Figure 1. **Graphical illustration of** (**a**.) **confounded signal and** (**b**.) **workflow of this study**. (**a**.) A simplified graphical representation of a measured signal (gray) which is a mix of independent sources such as the true signal (pink), a biological confounder (purple), and a technical confounder (blue). Note the difficulty of extracting the true signal from the measured additive signals. (**b**.) Graphical summary of the work conducted in this study. (1) Based on multi-omics pan-cancer TCGA data (section 2.1) different confounding effects were simulated (section 2.2). (2) Subsequently, four different deconfounding VAE frameworks (section 2.4) were trained on the the artificially confounded data. (3) The obtained deconfounded data were compared with the original, un-confounded input data in terms of clustering stability and deconfounding capabilities (section 2.6).

cause clustering algorithms to form subgroups based on irrelevant signals, which may ultimately lead to spurious biological conclusions [2, 3].

Conventional strategies to account for and mitigate confounders involve training linear regression (LR) per feature against the confounder and taking the residual part during pre-processing [4] or mapping the latent embedding to an unconfounded space after model training [5]. Conditional variational autoencoders (cVAEs) have been used to create normative models considering confounding variables, such as age, for neurological disorders [6]. Dincer *et al.* [7] proposed adversarial training to derive expression embeddings devoid of confounding effects, expanded upon by the single-cell generative adversarial network for batch effect removal [8]. Liu *et al.* [9] used a regularization term in the autoencoder's loss function to minimize correlation between latent embeddings and confounding bias. Despite their methodological diversity, these methods have only been validated to work effectively on data from a single omics source and are not tailored toward disease subtyping and patient stratification.

To address this gap, we propose four novel VAE-based deconfounding frameworks for clustering of multi-omics data, utilizing the (i) removal of latent features correlated with confounders, (ii) a cVAE [6], (iii) adversarial training [7, 8], and (iv) adding a regularization term to the loss function [9] as deconfounding strategies. To objectively assess whether our models can remove out-of-interest signals and find a patient clustering unbiased by confounding signals, we applied and evaluated our models on gene expression and DNA methylation pan-cancer data from The Cancer Genome Atlas (TCGA) program which we augmented with artificial confounding effects. In total, we simulated four different types of confounders, including linear, nonlinear, categorical, and a mixture thereof, resembling realistic confounders such as age (linar, nonlinear) [10–13], BMI (nonlinear) [14], or batch effects (categorical) [2, 7].

The contribution of our study is as follows:

- Four novel multi-omics clustering models based on VAE and different deconfounding strategies are presented.
- We highlight that various deconfounding techniques address confounded clustering in distinct ways, often overlooked within the algorithm's framework.
- Different confounding effects are simulated on the real-life TCGA dataset to demonstrate the influence of confounders on clustering and underscore the necessity for deconfounding models.
- Readers are provided with guidelines detailing strengths and limitations of each approach, along with suggestions on selecting an appropriate framework fitting their purposes.

Materials and methods

Data collection and preprocessing

This study utilized data collected within TCGA project [15], encompassing gene expressions (mRNA) of 4333 patients and DNA methylations (DNAm) of 2940 patients across six different cancer types, including breast invasive carcinoma (BRCA), thyroid carcinoma (THCA), bladder urothelial carcinoma (BLCA), lung squamous cell carcinoma (LUSC), head and neck squamous cell carcinoma (HNSC), and kidney renal clear cell carcinoma (KIRC). We selected these six types of carcinoma for their relatively similar histology, compared with glioma, sarcoma, etc., and balanced sample sizes to better evaluate our proposed methods. Focusing solely on gene expression and DNA methylation omics, we aimed to optimize model performance and prevent overfitting. Additionally, selecting these omics introduced diversity in data formats and distributions, with gene expression ranging continuously and DNA methylation exhibiting a largely

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

bimodal beta distribution, enhancing the complexity and depth of our analysis.

Datasets were downloaded using the R package TCGAbiolinks [16]. The subsequent filtering step removed patients with (i) only a single data type available, (ii) missing clinical metadata, (iii) "american indian" or "alaska native" ancestry, and (iv) unknown tumor stage, resulting in a total of 2547 patients. The preprocessing of mRNA and DNAm data included the removal of probes (i) not shared across all cancer types, (ii) with missing values, and (iii) with 0 variance across all included patients, resulting in 58 456 mRNA and 232 088 DNAm features. To reduce the number of input features, we only considered the 2000 probes showing the largest variance across patients for each data type, resulting in a final data set of 2547 patients and 4000 features. This reduction strikes a balance between the number of features included and biological variability addressed and is in line with other clustering works on TCGA data [17]. TCGA–BRCA molecular subtype information for 724 (out of 731) patients was derived from the TCGA Pan-Cancer Atlas [18] *via* the cBioPortal [19].

Simulation of confounders

To imitate common confounding scenarios in real-life clustering applications we simulated linear, squared, categorical confounders, and a mixture thereof, resembling common confounders in real-life studies, e.g. ageing [12, 13, 20], BMI [14], or batch effects [2, 7]. These confounders hinder the true or biologically meaningful clustering by intrinsically affecting the data structure in an unwanted way and possibly leading to a confounded clustering. The goal of this simulation is to evaluate how well the multi-view clustering models perform in the presence of confounders, whose pattern and magnitude are artificially generated and hence measurable.

Here we denote the mRNA data as $X_1 \in \mathbb{R}^{n\times p}$ and DNAm data as $X_2 \in \mathbb{R}^{n\times q}$, where n, p, q are the number of patients, gene expressions, and DNA methylations, respectively. We first rescaled every mRNA and DNAm feature to the range $[0, 1]$ to avoid large ratios between the raw feature and the confounding effect. A visualization of all confounding effects can be found in the Supplementary Methods.

Linear confounder

We uniformly generated a random numeric confounder $\mathbf{c} \in \mathbb{R}^n$ with discrete values $\{0, 1, 2, 3, 4, 5\}$, leading to a confounder clustering of six classes. Its linear effect on each individual is $\mathbf{c} + 5$ and a random weight for each feature was multiplied with it:

$$X_1^* = X_1 + E_1^{\text{linear}} = X_1 + (\mathbf{c} + 5) \otimes \mathbf{w_1}; \tag{1}$$

$$X_2^* = X_2 + E_2^{\text{linear}} = X_2 + (\mathbf{c} + 5) \otimes \mathbf{w_2}, \tag{2}$$

where $\otimes$ denotes the outer product between two vectors, and $\mathbf{w_1} \in \mathbb{R}^p \sim U(0, 0.1)$, $\mathbf{w_2} \in \mathbb{R}^q \sim U(0, 0.2)$. We chose the uniform distribution of $\mathbf{w_1}$ to range from 0 to 0.1 so that the total linear effect would range from 0 to 1, having the same scale as X_1. We increased the upper bound of $\mathbf{w_2}$ to 0.2 due to our observation that X_2 is less sensitive to linear confounders.

Nonlinear confounder

Nonlinear effects were simulated in a similar way to linear effects. However, to mimic a nonlinear confounder, as observed in, e.g. the significant quadratic association between body mass index and colon cancer risk [14], we considered adding an element-wise squared confounding effect $\mathbf{c}^2$ on the features:

$$X_1^* = X_1 + E_1^{\text{square}} = X_1 + \mathbf{c}^2 \otimes \mathbf{w_1}; \tag{3}$$

$$X_2^* = X_2 + E_2^{\text{square}} = X_2 + \mathbf{c}^2 \otimes \mathbf{w_2}, \tag{4}$$

where $\mathbf{w_1} \sim U(0, 0.04)$, $\mathbf{w_2} \sim U(0, 0.04)$. The distribution of $\mathbf{w_1}$ and $\mathbf{w_2}$ was also determined based on the scale of X_1 and X_2.

Categorical confounder

The categorical confounding effect was achieved by shifting patients with the same confounder class to a distinctive direction in the feature space. More specifically, we first sampled six p-dimensional vectors for shifting mRNA data and six q-dimensional vectors for shifting DNAm data, both from $U(0, 1)$ and corresponding to six different confounder classes. The n patients were randomly assigned to each of the six categories. As a result, two matrices $C_1 \in \mathbb{R}^{n\times p}$ and $C_2 \in \mathbb{R}^{n\times q}$ denote the concatenation of shifting vectors of every patient for mRNA and DNAm, respectively. The categorical confounder is therefore the membership of all individuals in the six classes. A typical example of categorical confounders for clustering could be batch effects caused by collecting data from different centers [2, 7]. In the field of cancer subtyping, common categorical confounders include tumor stage and ethnicity [21, 22]. Then the confounded features were created via

$$X_1^* = X_1 + E_1^{\text{categ}} = X_1 + \mathbf{diag}(\mathbf{w}) \cdot C_1; \tag{5}$$

$$X_2^* = X_2 + E_2^{\text{categ}} = X_2 + \mathbf{diag}(\mathbf{w}) \cdot C_2, \tag{6}$$

where $\mathbf{diag}(\cdot)$ converts a vector into its corresponding diagonal matrix. Different from the case of a numeric confounder, the weight vector $\mathbf{w} \in \mathbb{R}^n \sim U(0, 1)$ of the categorical confounder indicates to what extent every patient was shifted so that patients would have various strength of association with their confounder class.

Mixed confounder types

Real-life data analyses are likely affected by multiple confounders of different kinds, for instance, many cancer studies correct for age, age squared, education, etc. jointly in their models [12, 13]. Here we simulated a mixed confounding effect of linear, nonlinear, and categorical confounders as described below:

$$X_1^* = X_1 + E_1^{\text{linear}} + E_1^{\text{square}} + E_1^{\text{categ}}; \tag{7}$$

$$X_2^* = X_2 + E_2^{\text{linear}} + E_2^{\text{square}} + E_2^{\text{categ}}, \tag{8}$$

where $E_1^{\text{linear}}, E_2^{\text{linear}}, E_1^{\text{square}}, E_2^{\text{square}}, E_1^{\text{categ}}, E_2^{\text{categ}}$ represent the second term in Formulae (1)–(6), respectively.

Variational autoencoder for data integration (XVAE)

A variety of different VAE architectures exist for the purpose of data integration, as extensively compared by Simidjievski *et al.* [1]. In this study, we utilize one architecture recommended by the respective authors, namely the X-shaped variational autoencoder (XVAE) (Fig. 2). This architecture merges the heterogeneous input data sources into a combined latent representation z by learning to reconstruct each source individually from the common representation. Here we consider only two data types of a single

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

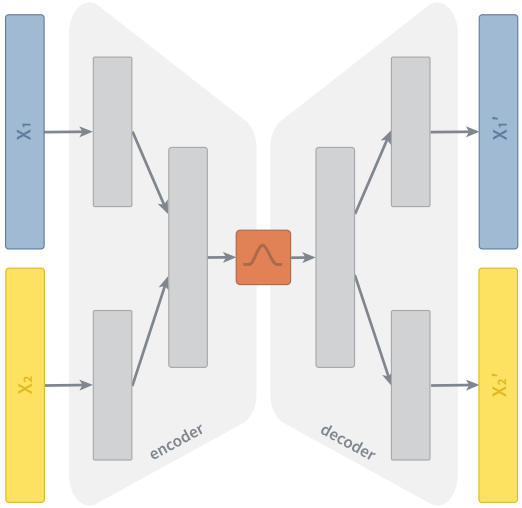

Figure 2. **Schematic representation of an XVAE.** The two input layers (X_1, X_2) denote the two omics dimension used in this study, namely gene expression and DNA methylation. The encoder consists of contiguous hidden layers, each with fewer nodes. We design the encoder of XVAE with a total of two layers prior to the latent embedding. In the first hidden layer, the dimension of each input entity is reduced individually. In the second hidden layer, input entities get fused into a combined layer. The latent embedding (red) represents the bottleneck of the XVAE with the minimum number of nodes. The decoder reversely mirrors the layer structure of the encoder, with the final layer featuring the same number of nodes as the input layer as it attempts to reconstruct (X_1', X_2') the original input from the latent embedding.

datapoint x_1 and x_2, and the loss function of XVAE is as follows:

$$L_{\text{XVAE}}(\phi, \theta; x_1, x_2) = -E_{z\sim q_\phi(z|x_1,x_2)}[\log p_\theta(x_1, x_2|z)] + \beta * \text{MMD}(q_\phi(z|x_1, x_2)||p(z)), \quad (9)$$

where $q_\phi(z|x_1, x_2)$ encodes the latent space as a probability distribution over the input variables (parameterized by ϕ) and $p_\theta(x_1, x_2|z)$ encodes the reconstruction of input variables as a probability distribution over the latent space (parameterized by θ). Following the originally proposed implementation, we use maximum mean discrepancy (MMD) as a regularization term to constrain the latent distribution q_ϕ to be a standard Gaussian distribution, balanced by the constant beta (β), which is set to 1 for all experiments. A more detailed description on autoencoders, as well as the XVAE architecture and training procedure, can be found in the Supplementary Methods.

Multi-omics deconfounding models

Here, we will first describe in section 2.4.1 the use of LR for confounder correction and PCA for dimensionality reduction, which we deem the "baseline model" due to their wide popularity. Then, we outline in sections 2.4.2–2.4.5 the four XVAE-based deconfounding models proposed in this study. Throughout this section we denote the confounder value of a single data point as c.

Baseline model: linear regression and PCA (LR+PCA)

Under the assumption that the effects of one or multiple confounders are linearly additive to the true signal of a feature, we build a LR model for the confounders against each mRNA or DNAm feature and then take their residuals as adjusted features. Subsequently, the adjusted features from the two data types are concatenated and their dimensionality is reduced via PCA (LR+PCA). We select the top 50 PCs explaining most of the variance of data to keep the embedding size identical to that of every XVAE-based model. The 50 PCs explaining the most variance of data are considered for the final clustering, for which KMeans with 10 random initializations is applied.

Conditional X-shaped variational autoencoder

cVAE [23] is a semi-supervised variation of VAE, which originally aims to fit the distribution of the high-dimensional output as a generative model conditioned on auxiliary variables. Lawry *et al.* [6] proposed to achieve deconfounding through a cVAE incorporating confounding variable information as auxiliary variables. We extend this initial idea to be able to handle multi-omics data by replacing the originally proposed VAE with the XVAE model, resulting in a conditional X-shaped variational autoencoder (cXVAE) architecture (Fig. 3A). We tested the integration of confounders at different levels of the cXAVE, including the input layer, the hidden layer that fuses multiple inputs, and the embedding. More details on cXVAE implementations can be found in the Supplementary Methods.

X-shaped variational autoencoder with adversarial training

The adversarial deconfounding autoencoder proposed by Dincer *et al.* [7] follows the idea of training two networks simultaneously—an autoencoder to generate a low dimensional embedding and an adversary multi-layer perceptron (MLP) trained to predict the confounder from said embedding (Fig. 3B). By adversarially training the two networks, i.e. the autoencoder aims to generate an embedding which cannot be used for confounder prediction by the MLP, it aims at generating embeddings that can encode biological information without encoding any confounding signal. As the original framework can only handle a single data type, we adapt it to work with multi-omics input by replacing its autoencoder with XVAE architecture. Details on architecture and training procedure of X-shaped variational autoencoder with adversarial training (adv-XVAE) can be found in Supplementary Methods.

X-shaped variational autoencoder with deconfounding regularization

Augmenting the loss function of deep learning models is an effective way to impose restrictions on the model or enforce learning of specific patterns. As an example, studies focused on disentangling the often highly correlated latent space of autoencoders impose constraints on the correlation between latent features by adding a penalty term to the loss function [9]. Inspired by this idea, we formulate a deconfounding regularization term aiming to reduce the degree of correlation between latent features and confounders. The regularized loss function becomes

$$L_{\text{cr-XVAE}}(\phi, \theta; x_1, x_2, c) = L_{\text{XVAE}}(\phi, \theta; x_1, x_2) + f(z, c), \quad (10)$$

where $f(z, c)$ denotes the joint association between latent features and confounders. More specifically, we implement two different association measurements: Pearson correlation and mutual information. Because Pearson correlation ranges from -1 (negatively correlated) to 1 (positively correlated) and both indicate strong relationship, we regularize only the magnitude of correlation by two methods, taking its absolute value or squared value. Because the confounder distribution needed for mutual information is usually unknown, we implement two methods to approximately

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

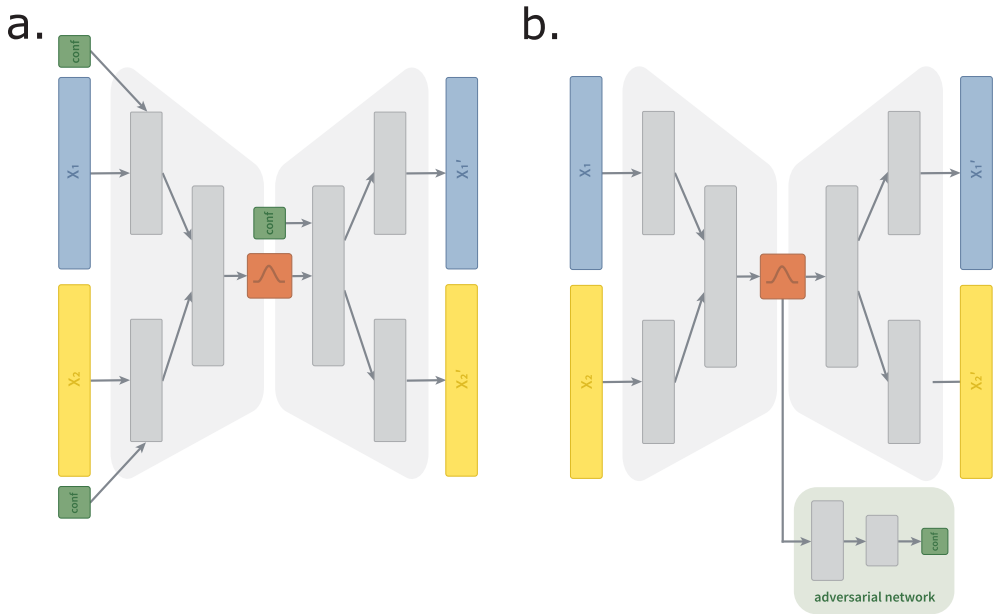

Figure 3. Schematic representation of (**a**.) cVAE and (**b**.) adv-XVAE. (**a**.) Depicts the cXVAE implementation termed *input + embed* due to the addition of confounders (green) in the first layer of the encoder and decoder. (**b**.) Depicts the adv-XVAE implementation termed *multiclass* due to the usage of only a single supervised adverarial network (light green) trained to predict confounders (green) using a multiclass prediction loss. X_1 and X_2 are the two omics dimensions, namely gene expression and DNA methylation, while X_1' and X_2' denote their respective reconstruction. More details and visualizations of other implementation can be found in the Supplementary Material.

compute mutual information as loss function, with differentiable histogram or kernel density estimate.

Feature selection by removing correlated latent features (XVAE+FS)

The removal of latent features correlated with confounders comes from the idea of *post hoc* interpretation of latent features [24]. To identify confounded latent features, we calculate the Pearson correlation between each latent variable and the confounder. For determining the threshold indicating which latent features are being removed from further analyses, we test two different approaches:

(1) *P -value cutoff*—the *P*-value of the Pearson correlation indicates the probability that the computed correlation is smaller than a random correlation between uncorrelated datasets. Latent features with a *P*-value <0.05 are excluded from analyses.
(2) *Absolute correlation coefficient*—Pearson correlation measures the linear relationship between two variables. Latent features exhibiting an absolute Pearson correlation of more than 0.3 (weak correlation) or 0.5 (strong correlation) are excluded.

Consensus clustering

Different from the baseline LR model which adopts KMeans on the deconfounded features for clustering, we apply consensus clustering on the latent features of each VAE-based deconfounding model. Here, consensus clustering takes the advantage of random sampling in a VAE and it aggregates the individual clustering of each embedding sampled from the latent distributions [25]. We first generate 50 embedding matrices for all the n samples, on each of which a KMeans clustering is performed. Subsequently, a consensus matrix $\bar{A} \in \mathbb{R}^{n\times n}$ is constructed from all the 50 clusterings:

$$\bar{A} = \frac{1}{50}\sum_{i=1}^{50} A_i, \tag{11}$$

where $A_i \in \mathbb{R}^{n\times n}$ is the binary matrix of each KMeans clustering indicating if two data points are assigned to the same cluster or not. Values of $\bar{A}$ are in the range [0,1], where 0 means the two corresponding samples are never clustered together in the 50 clusterings, while 1 means they are always in the same cluster. Finally, a spectral clustering is performed on the consensus matrix $\bar{A}$ to derive a stable clustering of the patients. To experiment on the potential impact of the number of embedding matrices, we rerun the model with various numbers (10, 50, 100, 150, 200) and compare the model performance (Supplementary Table 4).

Evaluation metrics

We apply each of the aforementioned models to the artificially confounded multi-omics dataset described in sections 2.1 and 2.2. Every model is evaluated in terms of their XVAE reconstruction accuracy, measured as the relative reconstruction error of inputs, their clustering stability, evaluated by the dispersion score of consensus clustering (CC), and deconfounding capabilities for clustering, estimated by calculating the adjusted Rand index (ARI) for true (cancer types) and confounder labels.

XVAE reconstruction accuracy

Model training is monitored through inspection of the validation loss. To evaluate reconstruction quality of the trained XVAE model, we compute the L2 relative error (RE) between the original input (x) and reconstructed data (x') for (i) each data type individually:

$$\mathrm{RE} = \frac{\sqrt{\sum_{i=1}^{n} \|x_i - x_i'\|^2}}{\sqrt{\sum_{i=1}^{n} \|x_i\|^2}}, \tag{12}$$

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

as well as (ii) for the combined data types:

$$\mathrm{RE} = \frac{\sum_{m=1}^{2} \sqrt{\sum_{i=1}^{n} \|x_{mi} - x'_{mi}\|^2}}{\sum_{m=1}^{2} \sqrt{\sum_{i=1}^{n} \|x_{mi}\|^2}}, \tag{13}$$

where $m = 1, 2$ indicates the two data types and n is the number of samples.

Clustering stability

Before assessing how well each model can derive a meaningful clustering, we want to first check if a model can stably cluster the samples. To achieve this goal, we employ the dispersion score to measure the internal stability of consensus clustering based on its consensus matrix $\bar{A}$:

$$\text{Dispersion} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (\bar{A}_{ij} - 0.5)^2 * 4}{n^2} \tag{14}$$

The dispersion score ranges from 0 to 1, where 1 shows a perfect stability that every value in $\bar{A}$ is either 0 or 1, i.e. no confusion among the clusterings, and the lower the less consensus among the clusterings.

Deconfounding capabilities

We compare our clustering with two different labels, the true one, namely cancer types, and the confounder. An ideal model should deconfound the features sufficiently while keeping the meaningful information for obtaining the true clustering. In other words, we expect a model with high ARI with the true label and low ARI with the confounder label. The association between true patient label and clusters obtained when modeling the original (unconfounded) data represents the best achievable clustering, with ARI value converging toward 1.

Similar to ARI, we compute another external clustering metric, the normalized mutual information (NMI), which measures the dependence between two clusterings. As it only shows complementary information to ARI, we record the NMI of every clustering model in Supplementary Table 1.

Implementation

For better stability and generalization, we train each model 50 times using (i) randomly sampled training and validation sets with a ratio of 80:20 and (ii) different seed of randomization.

Software

All of the models described in this study are built in Pytorch Lightning [26] and trained using the GPU units RTX 2080 Ti 11GB.

Results

cXVAE outperforms other considered deconfounding strategies in the presence of a single confounder

We simulated different types of confounding effects—linear, nonlinear (squared), and categorical—on the multi-omics TCGA pan-cancer dataset to benchmark a total of four deconfounding frameworks, namely XVAE with Pearson correlation feature selection (XVAE+FS), conditional XVAE (cXVAE), adversarial training with XVAE (adv-XVAE), and confounder-regularised XVAE (cr-XAVE) (see Methods for more details). We additionally included two baseline models to compare with (1) confounder correction with linear regression (LR+PCA) and (2) vanilla XVAE without any deconfounding (XVAE). To estimate the robustness of each method, each model was trained on 50 iterations of randomly sampled training and validation data (80:20 split) and random seed initialization.

All proposed deconfounding approaches were able to correct for a linear confounder, as denoted by the high ARI for true clustering and low ARI for confounder clustering (Table 1). Performances started to decline for nonlinear confounding problems, with cXVAE clearly outperforming other strategies. For nonlinear confounders we noted large ARI for confounder clustering across all strategies and simulation setups. This illustrates that, while good clustering performance for true labels were achieved, the full removal of unwanted signal was not easily achievable for all the models. Categorical confounding was perceived to be the most difficult, with all models except cXVAE exhibiting a high decrease in true clustering performance. Notably, X-shaped variational autoencoder with deconfounding regularization (cr-XVAE) and XVAE+FS were able to remove artificial confounders completely, however at the cost of simultaneously removing true clustering signal. adv-XVAE, which in theory should be a strategy well suited to deal with categorical problems, fails to consistently remove the categorical confounding effect. In general we noted a decline of reconstruction accuracy of models with increasing complexity of the confounder simulations.

To illustrate the deconfounding capabilities of several of the developed models, we examined their clustering results obtained on the TCGA dataset involving categorical confounders (Fig. 4). The UMAP plot of the original, unconfounded data (Fig. 4a) shows the distinct clustering of THCA, KIRC, and BRCA, while the clustering of BLCA, LUSC, and HNSC appears to be more entangled. The observed clustering was significantly obscured by the addition of an artificial categorical confounder, visible as the clustering being dictated by the confounder class rather than cancer type (Fig. 4b). An attempted deconfounding using a vanilla XVAE (Fig. 4c) or adv-XVAE (Fig. 4d) model displayed little improvement over the confounded clustering, demonstrating the models' inability to remove the artificially introduced signal. The cXVAE model, however, proved to be able to effectively mitigate the confounding effect, resulting in a clustering similar to the original, unconfounded data (Fig. 4e). In an attempt to investigate whether the deconfounding using cXVAE not only restores pan-cancer type but can also recover cancer subtypes, we examined the clustering of several TCGA-BRCA molecular subtypes, including Her2, LumA, Basal, LumB, and normal, before and after deconfounding (Supplementary Figure 5). This revealed that while molecular subtypes were completely masked by the simulated categorical confounders (Supplementary Figure 5, b), deconfounding with cXVAE could retrieve their original clustering (Supplementary Figure 5c).

In summary, across all confounder simulations, cXVAE clearly outperformed other deconfounding strategies in terms of clustering accuracy, deconfounding power, and model robustness. The ARI on true clustering obtained by cXVAE in all three scenarios reached around 0.7, which is very close to the performance of the vanilla XVAE on unconfounded data (0.731, see details in Supplementary Table 2).

A more detailed summary of the performances of each model can be found in Supplementary Table 1. The dimensionality reduction plots for models not displayed in Fig. 4, namely LR+PCA, XVAE+FS, and cr-XVAE can be found in Supplementary Figure 2. It illustrates that while XVAE+FS and cr-XVAE yield performances similar to XVAE and adv-XVAE, LR+PCA seems unable to distinguish the true signal from the confounder signal.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

Table 1. **Overview performances of deconfounding strategy for single confounder simulations.** Values are displayed as mean ± standard deviation of 50 runs with different parameter initialization and randomly sampled training and validation data. Models on the first column indicate the following deconfounding strategies and implementations thereof: LR followed by principal component analysis and KMeans clustering (LR+PCA), vanilla XVAE without any deconfounding (XVAE), XVAE with feature selection in the form of removing correlated latent features (XVAE+FS, correlation cutoff = 0.5), conditional XVAE (cXVAE, input + embedding), adversarial training with XVAE (adv-XVAE, multiclass MLP), confounder-regularised XVAE (cr-XAVE, squared correlation regularization). Reconstruction error: relative error in the reconstruction of X_1 and X_2 weighted eqally; CC dispersion: consensus clustering agreement over 50 iterations; True clustering: ARI of consensus clustering derived clusters with True label labels; Confounder clustering: ARI of consensus clustering derived clusters with simulated confounder labels

Linear				
			Clustering performance (ARI)	
	Reconstruction error	**CC dispersion**	**True label (cancer type)**	**Confounder**
LR+PCA	–	–	0.692	0.001
XVAE	0.246 (± 0.004)	0.844 (± 0.045)	0.506 (± 0.116)	0.151 (± 0.057)
XVAE+FS	0.245 (± 0.004)	0.860 (± 0.028)	0.571 (± 0.092)	0.008 (± 0.007)
cXVAE	**0.234 (± 0.003)**	**0.935 (± 0.023)**	**0.712 (±0.055)**	**0.001 (± 0.001)**
adv-XVAE	0.245 (± 0.004)	0.901 (± 0.032)	0.568 (± 0.070)	0.093 (± 0.051)
cr-XVAE	0.244 (± 0.003)	0.873 (± 0.028)	0.598 (± 0.074)	0.004 (± 0.004)
Nonlinear				
			Clustering performance (ARI)	
	Reconstruction error	**CC dispersion**	**True label (cancer type)**	**Confounder**
LR+PCA	–	–	0.391	0.215
XVAE	0.236 (± 0.003)	0.826 (± 0.042)	0.307 (± 0.141)	0.297 (± 0.071)
XVAE+FS	0.236 (± 0.003)	0.805 (± 0.040)	0.411 (± 0.142)	0.138 (±0.078)
cXVAE	**0.227 (± 0.002)**	**0.908 (± 0.031)**	**0.646 (± 0.079)**	**0.076 (± 0.074)**
adv-XVAE	0.238 (± 0.004)	0.942 (± 0.025)	0.568 (± 0.049)	0.194 (± 0.006)
cr-XVAE	0.235 (± 0.002)	0.852 (± 0.043)	0.478 (± 0.129)	0.154 (± 0.042)
Categorical				
			Clustering performance (ARI)	
	Reconstruction error	**CC dispersion**	**True label (cancer type)**	**Confounder**
LR+PCA	–	–	0.150	0.071
XVAE	0.216 (± 0.003)	0.762 (± 0.055)	0.330 (± 0.125)	0.048 (± 0.088)
XVAE+FS	0.216 (± 0.003)	0.787 (± 0.040)	0.361 (± 0.100)	0.010 (± 0.023)
cXVAE	**0.210 (± 0.002)**	**0.911 (± 0.033)**	**0.664 (± 0.070)**	**0.001 (± 0.001)**
adv-XVAE	0.217 (± 0.002)	0.764 (± 0.058)	0.240 (± 0.188)	0.156 (± 0.084)
cr-XVAE	0.216 (± 0.003)	0.813 (± 0.034)	0.368 (± 0.101)	0.001 (± 0.001)

Furthermore, the deconfounded clustering derived by cXVAE in the presence of multiple confounder is shown in Supplementary Figure 4, which indicates the capabilities of cXVAE to deconfound even in this complex scenario. While Table 1 depicts the best-performing implementation of each deconfounding model, we tested a number of possible implementations (see Methods), which we observed to have a notable impact on model performance (Supplementary Table 2). Therefore, we provide design recommendations for each deconfounding strategy in the Supplementary Results.

cXVAE is easily extendable to handle multiple confounders of mixed types

In a realistic setting datasets can be confounded by multiple confounders with different biasing effects. In an attempt to investigate how well-deconfounding strategies can handle more than one confounder, we simulated the parallel presence of three confounders of different effect, namely linear, nonlinear, and categorical (Table 2). In line with our observations with the single confounder simulations, cXVAE outperformed other models in terms of true clustering accuracy and deconfounding efficiency. While also other strategies like XVAE+FS, cr-XVAE, or LR+PCA were able to successfully remove all three simulated effects, they achieved this at the cost of true signal. adv-XVAE failed to fully remove confounders while also showing very low true clustering accuracy and can therefore be considered unsuitable for the task. We also noted that the decline in reconstruction accuracy with increasingly complex confounding situations is even more pronounced in multiple confounder settings.

cXVAE is able to retrieve biology-driven clustering from confounded data

To illustrate the deconfounding capabilities of cXVAE, the model that outperformed others across all four evaluation metrics in various confounding scenarios, we examined the clustering results obtained on the TCGA dataset involving categorical confounders (Fig. 4). The UMAP plot of latent features clearly showed that BRCA, THCA, and KIRC were well clustered by cXVAE, while BLCA, LUSC, and HNSC were still entangled.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

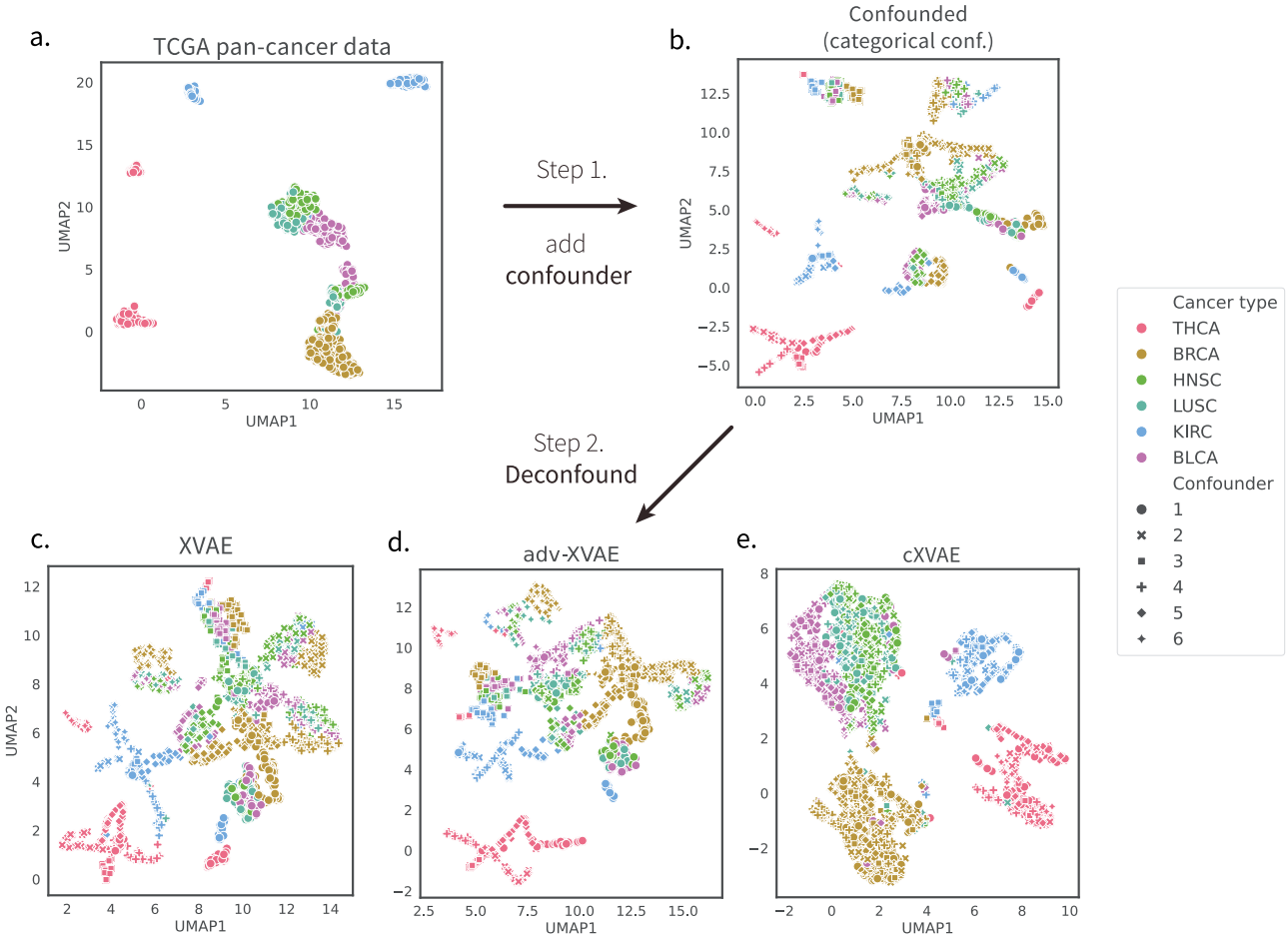

Figure 4. **Deconfounding behavior of several developed models.** Dimensionality reduction (UMAP) plot of the (**a**.) unconfounded TCGA pan-cancer data, (**b**.) categorically confounded data, as well as deconfounding using the (**c**.) vanilla XVAE (**d**.) adv-XVAE, and (**e**.) cXVAE. The marker colors indicate the true label labels (i.e. TCGA cancer types), while the marker shapes indicate the six classes (1–6) of the confounder (see section 2.2). The steps indicated describe the experimental order (see Fig. 1).

Table 2. Overview performances of deconfounding strategy in the presence of multiple confounders. Values are displayed as mean ± standard deviation of 50 runs with different parameter initialisation and randomly sampled training and validation data. For a detailed description of columns and models, please refer to Table 1.

Multiple confounders

			Clustering performance (ARI)			
	Reconstruction error	**CC dispersion**	**True label (cancer type)**	**Linear confounder**	**Squared confounder**	**Categorical confounder**
LR+PCA	–	–	0.215	0.001	0.014	0.001
XVAE	0.161 (± 0.003)	0.725 (± 0.043)	0.216 (± 0.089)	0.015 (± 0.019)	0.140 (± 0.043)	0.067 (± 0.048)
XVAE+FS	0.161 (± 0.003)	0.731 (± 0.037)	0.265 (± 0.085)	0.007 (± 0.009)	0.019 (± 0.030)	0.109 (± 0.057)
cXVAE	**0.146 (± 0.002)**	**0.905 (± 0.022)**	**0.634 (± 0.042)**	**0.001 (± 0.001)**	**0.001 (± 0.001)**	**0.001 (± 0.001)**
adv-XVAE	0.158 (± 0.004)	0.753 (± 0.066)	0.225 (± 0.120)	0.016 (± 0.023)	0.107 (± 0.052)	0.106 (± 0.051)
cr-XVAE	0.161 (± 0.003)	0.764 (± 0.031)	0.369 (± 0.064)	0.003 (± 0.002)	0.007 (± 0.010)	0.001 (± 0.001)

In summary, we found the deconfounding behavior of cXVAE to not only yield clusterings resembling those of the unconfounded data, but also be in line with the pathological and physiological differences between the pan-cancer types. BLCA arises from urothelial cells in the transitional epithelium, which can change from cuboidal to squamous form when stretched. Furthermore, squamous differentiation is by far the most common histological variant of urothelial carcinoma [27], indicating a close relationship between urothelial carcinoma and squamous cell carcinoma. Apart from BLCA, the overlap in clustering of LUSC and HNSC can be directly explained by their common origin of squamous cells, while BRCA, THCA, and KIRC are all carcinoma related to glandular cells [28]. Supporting the validity of our obtained cXVAE clustering, other multi-omics pan-cancer studies utilizing stacked VAEs [29], penalized matrix factorization [30], or supervised VAE [31] have retrieved similar cancer type clustering.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

Discussion

In this study, we addressed the possible harm of ignoring or inadequately handling confounders to clustering samples with (multi-)omics measurements. In epidemiology, a confounder is a variable that can affect the result of a study because it is related to both the exposure and the outcome being studied. Here, we extended the definition to unsupervised models for disease subtyping to indicate variables that can distort the relationship between inferred or predicted cluster membership and disease.

Extensive simulation revealed that cXVAE stands out as a versatile and accurate deconfounding approach. The applicability of conditional autoencoder to biological data to e.g. correct for batch effects [6] or disentangle confounders in functional magnetic resonance imaging [32] or microRNA data [33] has been shown before. However, by merging the principles of a conditional autoencoder with the framework of an autoencoder specifically tailored for the integration of multi-omics data, our research charts new frontiers in the domain of deconfounded patient stratification.

While adversarial training may offer an alternative flexible deconfounding approach, we confirm that optimization of model hyper-parameters is challenging [8]. Instability may become more pronounced in the presence of multiple confounders. This can be explained by the fact that adversarial networks were trained separately for each confounder, sequentially adding extra terms to its objective function (see Supplementary Table 2). In the literature, a statistical correlation loss has been proposed to replace the adversarial prediction loss in a adversarial training model [34], resembling our cr-XVAE model. The difference is that cr-XVAE directly computes the correlation between the VAE embedding and the confounder without an additional adversarial network.

Confounders can be dealt with in various ways. We implemented the most widely used strategy in association studies, namely regressing out the confounding effect from each feature with a linear model [4], as a baseline model to compare with. There are other methods working on the confounding issue for out-of-sample prediction based on, e.g. confound-isolating cross-validation [35] and feature selection in the embedding space [36]. However, as we are more interested in the in-sample heterogeneity, in this paper, we focus on deconfounding approaches that have shown success with unsupervised autoencoder models.

The identification of disease subtypes requires performing a clustering algorithm at some point. Instead of a joint training for reconstruction and clustering, we chose for a decoupled strategy to (1) avoid having too many terms in loss function to confuse training, and (2) reduce computation time and initialization settings with iteratively training clustering in a joint loss function. Consensus clustering furthermore has several advantages in data science including robustness, stability, interpretability, and flexibility, as it can be applied to various types of data and clustering algorithms. Our consensus clustering scheme adopts spectral clustering as its final step because the consensus matrix can be naturally viewed as a graph of all patients and the superior performance of spectral clustering has been shown on graphs [37]. We chose to sample and cluster the embedding for 50 iterations to construct the consensus matrix. Having more iterations can improve the robustness of consensus clustering but at the cost of computation time. We observed that the 50 individual clusterings are very consistent and increasing the number of iterations will not necessarily improve the final performance (Supplementary Table 4).

It remains a daunting task to generate data that adequately reflect the complexity of real-life cases. Therefore, one needs to be aware that simulations of confounders always represent simplifications of real observable effects. Note that the range of weight vectors $\mathbf{w_1}$ and $\mathbf{w_2}$ may have an important influence on how the data are confounded. A large weight will cause a stronger confounding effect and potentially increase difficulty in finding the true clustering. Currently, we set their values based on the scale of the original features to balance between the true signal and confounder signal. While this study is limited to the use of two data types, in principle the XVAE design utilized allows the integration of heterogeneous data from many more sources simultaneously. Additionally, since all evaluated deconfounding strategies share the same XVAE design as a foundation, we anticipate consistent training time and performance across different model architectures when scaling up dimensions. Although the multi-omics pan-cancer data available within TCGA are vast, this study utilized only a limited set of information. This presents a limitation and warrants that the biological conclusions of this study be interpreted with care. Future research will utilize the full pan-cancer TCGA data, enabling a more holistic interpretation of findings.

Conclusion

In this study, we presented four VAE-based multi-omics clustering models and their variations, following different deconfounding strategies. Their clustering and deconfounding performance was evaluated and compared with baseline models on the multi-omics pan-cancer dataset from TCGA with artificially generated confounding effects. The results showed both the necessity to adjust for confounders and that our novel models, cXVAE in particular, can effectively deal with the confounding effects and obtain the biologically meaningful clustering. We demonstrate that our multi-omics deconfounding VAE clustering models have big potential in delivering accurate patient subgrouping or disesase subtyping, ultimately enabling better personalized healthcare.

Key Points

- The biasing effect of confounders on clustering is a often overlooked, which may result in spurious biological conclusions.
- We present four novel multi-omics clustering models based on VAE and different deconfounding strategies.
- The impact of confounding effects on clustering is demonstrated using real-life TCGA dataset and showcases the effectiveness of deconfounding models in mitigating these influences.
- Readers are provided with guidelines with strengths and limitations of each deconfounding approach and suggestions on selecting an appropriate framework fitting their purposes.

Acknowledgements

The authors thank the supporters of this study, namely the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 860895

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

TranSYS. Furthermore we thank the members of the Computational Population Biology group at Erasmus Medical Center for their critical and creative input to this work. We also would like to extend our gratitude to the PhD candidates enrolled in the *"Frontières de l'Innovation en Recherche et Éducation"* (FIRE) doctoral school and thank them for their critical reviewing and feedback on the preprint of this study.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement [860895 to Z.L., S.K., and K.V.S.]; E. S. acknowledges the funding received from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalized medicine—smart combination of pre-clinical and clinical research with data and ICT solutions).

Conflicts of interest: The authors declare no competing interests.

Data and code availability

Our code is available at https://github.com/ZuqiLi/Multi-Omics-Deconfounding-VAE. The simulated data generated in the course of this study are available at Zenodo (https://doi.org/10.5281/zenodo.10458941).

Author contributions

Z.L. and S.K. designed, planned, carried out the practical aspects of this study, and wrote the manuscript. G.V.R. offered mentoring and scientific input throughout the work process. G.V.R., E.S., K.V.S., and V.MdS. provided critical revision of the manuscript. All authors approve of the final manuscript.

References

1. Simidjievski N, Bodnar C, Tariq I. *et al.* Variational autoencoders for cancer data integration: Design principles and computational practice. *Front Genet* 2019; **10**:1205. https://doi.org/10.3389/fgene.2019.01205.
2. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017; **35**:498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.
3. Čuklina J, Pedrioli PGA, Aebersold R. " Review of batch effects prevention, diagnostics, and correction approaches," In: R. Matthiesen, (ed.), *Mass Spectrometry Data Analysis in Proteomics, Methods in Molecular Biology*, pp. 373–387, New York, NY: Springer, 2020, https://doi.org/10.1007/978-1-4939-9744-2_16.
4. Pourhoseingholi M, Baghestani A, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench* 2012; **5**:79–83.
5. Radhakrishnan A, Friedman SF, Khurshid S. *et al.* Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nat Commun* 2023; **14**:2436. https://doi.org/10.1038/s41467-023-38125-0.
6. Lawry Aguila A, Chapman J, Janahi M. Conditional VAEs for confound removal and normative modelling of neurodegenerative diseases. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022, Lecture Notes in Computer Science*. Wang L, Dou Q, Fletcher PT. *et al.* (eds.), pp. 430–40. Cham: Springer Nature Switzerland, 2022.
7. Dincer AB, Janizek JD, Lee S-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* 2020; **36**:i573–82. https://doi.org/10.1093/bioinformatics/btaa796.
8. Bahrami M, Maitra M, Nagy C. *et al.* Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics* 2021; **37**:1345–51. https://doi.org/10.1093/bioinformatics/btaa976.
9. Liu X, Li B, Bron E. *et al.* " Projection-wise disentangling for fair and interpretable representation learning: Application to 3d facial shape analysis," In: Bruijne M. (ed.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Strasbourg, France, 2021. Proceedings, Part V 24. Springer International Publishing, New York City, United States.* MICCAI 2021. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-87240-3_78.
10. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* 2022; **8**:4. https://doi.org/10.1038/s41514-022-00085-y.
11. Kuzub N, Smialkovska V, Momot V. *et al.* Evaluation of epigenetic age based on dna methylation analysis of several cpg sites in ukrainian population. *Front Genet* 2022; **12**:772298. https://doi.org/10.3389/fgene.2021.772298.
12. Tu H, Wen CP, Tsai SP. *et al.* Cancer risk associated with chronic diseases and disease markers: Prospective cohort study. *BMJ* 2018; **360**:k134. https://doi.org/10.1136/bmj.k134.
13. Kartsonaki C, Pang Y, Millwood I. *et al.* Circulating proteins and risk of pancreatic cancer: A case-subcohort study among Chinese adults. *Int J Epidemiol* 2022; **51**:817–29. https://doi.org/10.1093/ije/dyab274.
14. Odegaard AO, Koh WP, Yu MC. *et al.* Body mass index and risk of colorectal cancer in chinese singaporeans: The Singapore chinese health study. *Cancer* 2011; **117**:3841–9. https://doi.org/10.1002/cncr.25936.
15. Weinstein JN, Collisson EA, Mills GB. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013; **45**:1113–20. https://doi.org/10.1038/ng.2764.
16. Colaprico A, Silva TC, Olsen C. *et al.* TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016; **44**:e71. https://doi.org/10.1093/nar/gkv1507.
17. Chen F, Zhang Y, Bossé D. *et al.* Pan-urologic cancer genomic subtypes that transcend tissue of origin. *Nat Commun* 2017; **8**:199. https://doi.org/10.1038/s41467-017-00289-x.
18. Hoadley KA, Yau C, Hinoue T. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 2018; **173**:291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.
19. Cerami E, Gao J, Dogrusoz U. *et al.* The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012; **2**:401–4. https://doi.org/10.1158/2159-8290.CD-12-0095.
20. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* 2022; **8**:1–15. https://doi.org/10.1038/s41514-022-00085-y.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024

21. Li R, Daniel R, Rachet B. How much do tumor stage and treatment explain socioeconomic inequalities in breast cancer survival? Applying causal mediation analysis to population-based data. *Eur J Epidemiol* 2016; **31**:603–11. https://doi.org/10.1007/s10654-016-0155-5.
22. Nipp R, Tramontano AC, Kong CY. *et al*. Disparities in cancer outcomes across age, sex, and race/ethnicity among patients with pancreatic cancer. *Cancer Med* 2018; **7**:525–35. https://doi.org/10.1002/cam4.1277.
23. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Cortes C, Lawrence N, Lee D. *et al*. (eds.), *Advances in Neural Information Processing Systems*, Vol. **28**. Curran Associates, Inc., New York, United States, 2015.
24. Fan YJ. Autoencoder node saliency: Selecting relevant latent representations. *Pattern Recognit* 2019; **88**:643–53. https://doi.org/10.1016/j.patcog.2018.12.015.
25. Kiselev V, Kirschner K, Schaub M. *et al*. Sc3: Consensus clustering of single-cell rna-seq data. *Nat Methods* 2017; **14**:483–6. https://doi.org/10.1038/nmeth.4236.
26. Falcon W, The PyTorch Lightning team. PyTorch lightning. 2019.
27. Kamat AM, Hahn NM, Efstathiou JA. *et al*. Bladder cancer. *Lancet* 2016; **388**:2796–810. https://doi.org/10.1016/S0140-6736(16)30512-8.
28. Horne TK, Cronje MJ. Cancer tissue classification, associated therapeutic implications and PDT as an alternative. *Anticancer Res* 2017; **37**:2785–807.
29. Uyar B, Ronen J, Franke V. *et al*. Multi-omics and deep learning provide a multifaceted view of cancer. *bioRxiv* 2021;2021–09.
30. González-Reymúndez A, Vázquez AI. Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin. *Sci Rep* 2020; **10**:8341. https://doi.org/10.1038/s41598-020-65119-5.
31. Zhang X, Zhang J, Sun K. *et al*. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In: Illhoi Y, Jinbo B, Xiaohua H (eds.), *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Curran Associates, Inc., New York, United States. pp. 765–9, 2019.
32. Wang X, Zhou R, Zhao K. *et al*. Normative Modeling via conditional Variational autoencoder and adversarial learning to identify brain dysfunction in Alzheimer's disease. In: Natasha L and Oscar A (eds.), *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. Curran Associates, Inc., New York, United States, pp. 1–4, 2023.
33. Yu T. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. *PLoS Comput Biol* 2022; **18**:e1009826. https://doi.org/10.1371/journal.pcbi.1009826.
34. Adeli E, Zhao Q, Pfefferbaum A. *et al*. Representation learning with statistical independence to mitigate bias. In: Kevin B and Gérard M (eds.), *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Curran Associates, Inc., New York, United States, pp. 2513–23, 2021.
35. Chyzhyk D, Varoquaux G, Milham M. *et al*. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* 2022; **11**:giac014. https://doi.org/10.1093/gigascience/giac014.
36. He D, Liu Q, Wu Y. *et al*. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nat Mach Intell* 2022; **4**:879–92. https://doi.org/10.1038/s42256-022-00541-0.
37. Rodriguez MZ, Comin CH, Casanova D. *et al*. Clustering algorithms: A comparative approach. *PloS One* 2019; **14**:e0210236. https://doi.org/10.1371/journal.pone.0210236.

Downloaded from https://academic.oup.com/bib/article/25/6/bbae512/7824239 by guest on 18 October 2024