arXiv, 2021, pp. 1–11

Preprint
XOmiVAE
Problem Solving Protocol

PROBLEM SOLVING PROTOCOL

XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data

Eloise Withnell,[1,2] Xiaoyu Zhang,[1,*] Kai Sun[1] and Yike Guo[1,3]

[1]Data Science Institute, Imperial College London, SW7 2AZ, London, UK, [2]Department of Health Informatics, University College London, WC1E 6BT, London, UK and [3]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[*]Corresponding author. x.zhang18@imperial.ac.uk

Abstract

Deep learning based approaches have proven promising to model omics data. However, one of the current limitations compared to statistical and traditional machine learning approaches is the lack of explainability, which not only reduces the reliability, but limits the potential for acquiring novel knowledge from unpicking the "black-box" models. Here we present XOmiVAE, a novel interpretable deep learning model for cancer classification using high-dimensional omics data. XOmiVAE is able to obtain contribution values of each gene and latent dimension for a specific prediction, and the correlation between genes and the latent dimensions. It is also revealed that XOmiVAE can explain both the supervised classification and the unsupervised clustering results from the deep learning network. To the best of our knowledge, XOmiVAE is one of the first activated-based deep learning interpretation method to explain novel clusters generated by variational autoencoders. The results generated by XOmiVAE were validated by both the biomedical knowledge and the performance of downstream tasks. XOmiVAE explanations of deep learning based cancer classification and clustering aligned with current domain knowledge including biological annotation and literature, which shows great potential for novel biomedical knowledge discovery from deep learning models. The top XOmiVAE selected genes and dimensions shown significant influence to the performance of cancer classification. Additionally, we offer important steps to consider when interpreting deep learning models for tumour classification. For instance, we demonstrate the importance of choosing background samples that makes biological sense and the limitations of connection weight based methods to explain latent dimensions.

Key words: explainable artificial intelligence, deep learning, cancer classification, omics data, gene expression

Introduction

Omics data, such as DNA methylation and gene expression data, is high dimensional, with upwards of tens of thousands of molecular features per sample. As the number of features is considerably larger than the number of labels, analysis of omics data suffers from the "the curse of dimensionality", which often leads to overfitting. Therefore, performing dimensionality reduction prior to the downstream analysis has become a standard approach in omics data modelling and analysis [Meng et al., 2016]. Standard techniques like principal component analysis (PCA) learn a linear transformation of the data, which struggle to accurately capture complex non-linear patterns typically found within omics data. Non-linear dimensionality reduction methods such as t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten and Hinton, 2008] have become increasingly popular, but still have limitations in terms of scalability.

Deep learning methods have proven to be powerful techniques at capturing non-linear patterns in high-dimensional data [LeCun et al., 2015]. Variational autoencoders (VAEs) [Kingma and Welling, 2014] are recent deep learning methods that have shown promise at modelling omics data in lower dimensions and shown to increase the accuracy for cancer classification tasks compared to other deep learning methods [Zhang et al., 2019, Azarkhalili et al., 2019, Zhang et al., 2021]. Among them, OmiVAE [Zhang et al., 2019] is a model for low dimension latent space extraction and multi-class classification using both gene expression and methylation data. It combined a VAE with a neural network classifier for feature extraction and classification. Using the Pan-cancer multi-omics

dataset from The Cancer Genome Atlas (TCGA), it achieved better performance than other existing methods. Despite the advancements, a key limitation of this kind of methods is that they are "black box" models, as the contribution of each input feature and latent dimension towards the downstream prediction are unknown.

Various approaches have been developed to improve the explainability of deep learning models. One of the most promising approaches are probing methods, which inspect the structure and parameters in a trained model [Azodi et al., 2020]. There are three different types of probing techniques: weight-based, gradient-based and activation level-based [Azodi et al., 2020]. Among them, activation level-based approaches, such as layer wise propagation (LRP) [Hanczar et al., 2020], DeepLIFT [Shrikumar et al., 2017] and Deep SHAP [Lundberg and Lee, 2017a], are considered to be more promising. They overcome the misleading results produced by weight based approaches, such as when weights cancel each other due to positive and negative connections, when features do not have the same scale, and when a neuron has a large weight value but it is not activated [Azodi et al., 2020]. Weight-based approaches have previously been used to explain VAEs using gene expression data [Way and Greene, 2017, Bica et al., 2019]. Additionally, activation level-based approaches overcome the limitations of gradient-based approaches [Azodi et al., 2020], as they are accurate even when small changes of the input value do not effect the output. LRP has been used to explain the predictions from a deep neural network for gene expression [Hanczar et al., 2020]. However, LRP can produce misleading results with model saturation [Shrikumar et al., 2017]. Deep SHAP, which uses the key principles from DeepLIFT, has been used in a variety of biological applications [Tasaki et al., 2020, Lemsara et al., 2020]. There is a lack of research on the use of Deep SHAP to interpret the latent space of a variational autoencoder.

Here we propose explainable OmiVAE (XOmiVAE), which is a VAE-based explainable deep learning omics data analysis model for low dimensional latent space extraction and cancer classification. In addition, it took advantage of Deep SHAP [Lundberg and Lee, 2017a] to provide the the feature importance scores for a prediction. Deep SHAP was selected as the explanation technique due to its ability to provide more accurate explanations over other interpretation methods [Lundberg and Lee, 2017a], and therefore likely provides better signal-to-noise ratio in the top genes selected. XOmiVAE provides contribution of each input molecular feature and latent dimension to the prediction. Additionally, XOmiVAE can explain unsupervised clusters produced by the VAE clustering part of the network. Explanations of the downstream prediction were evaluated by biological annotation and literature, which aligned with current domain knowledge. XOmiVAE shows great potential for novel biomedical knowledge discovery from deep learning models.

Methods

Datasets and Pre-processing

The Cancer Genome Atlas Program (TCGA) [Weinstein et al., 2013] pan-cancer dataset, which includes 33 various tumour types for gene expression, was used for the experiment. A total of 9,081 samples from TCGA were selected for training and testing our proposed model, of which 407 were normal tissue samples. The data was downloaded from UCSC Xena [Goldman et al., 2018] [1] and the pre-processing step followed Zhang et al. [2019]. Genes on the Y chromosome (n=594) and genes with zero expression level (n=1,904) were removed to ensure the expression data was fair and clean across tumour samples. Probes with missing values (N/A) in more than 10% of the samples were also removed, and the remaining N/A values that did not meet this cut off were replaced by the mean of the corresponding gene expression data. The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) were normalised to the unit interval [0, 1].

The phenotype data of each sample was also downloaded from UCSC Xena, which included the origin of the tissue, the subject's gender, age and other variables that were used in the downstream analysis. The information of the cancer subtype of each tumour sample was obtained from Sanchez-Vega *et al.* [Sanchez-Vega et al., 2018].

Explainable OmiVAE (XOmiVAE)

Figure 1 provides an overview of the architecture of the proposed XOmiVAE model. The omics data was first passed through a VAE, to reduce the number of dimensions of the input features to 128. The latent space contained two output layers, the mean μ and the standard deviation σ of a Gaussian distribution $\mathcal{N}(\mu,\sigma)$ given input sample x. To enable backpropagation, reparameterisation was applied to μ and σ,

$$z = \mu + \sigma\epsilon \tag{1}$$

where ϵ is a random variable sampled from a normal distribution. We optimised the variational lower bound when training the VAE,

$$E_{z\sim q_\Phi(z|x)} log p_\theta(x \mid z) - D_{KL}(q_\phi(z \mid x) \parallel p_\theta(z)). \tag{2}$$

Here, $q_\phi(z \mid x)$ is the variational distribution introduced to approximate $p_\theta(z \mid x)$, where θ is the set of learnable parameters of the encoder, z is the latent variable generated from the prior distribution and x is the input sample. $p_\phi(x \mid z)$ is the conditional distribution. The Kullback Leibler (KL) divergence between two probability distributions is denoted by D_{KL}.

The VAE was connected to a classification network using the μ of the learnt distributions as the input of the first classification hidden layer. Then, the data was passed through a hidden layer of 64 dimensions, before the probability of the labels were outputted using softmax activation. We defined the loss function for the classification network as the cross-entropy between the predicted tissue labels and the true labels.

The overall loss function for the model was a combination of both the VAE loss $\mathcal{L}_{VAE}$ and the classification loss $\mathcal{L}_{Class}$,

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{VAE} + \beta\mathcal{L}_{Class} \tag{3}$$

where α and β weight the two losses during training. We used the hyper-parameters as defined in [Zhang et al., 2019] unless otherwise specified (Supplementary Table S1).

In our proposed XOmiVAE model, we integrated Deep SHAP to OmiVAE. Deep SHAP uses a distribution of background samples to explain the predictions [Lundberg and Lee, 2017a]. The choice of background samples is important as

[1] `https://xenabrowser.net/datapages/`, accessed in July 2020

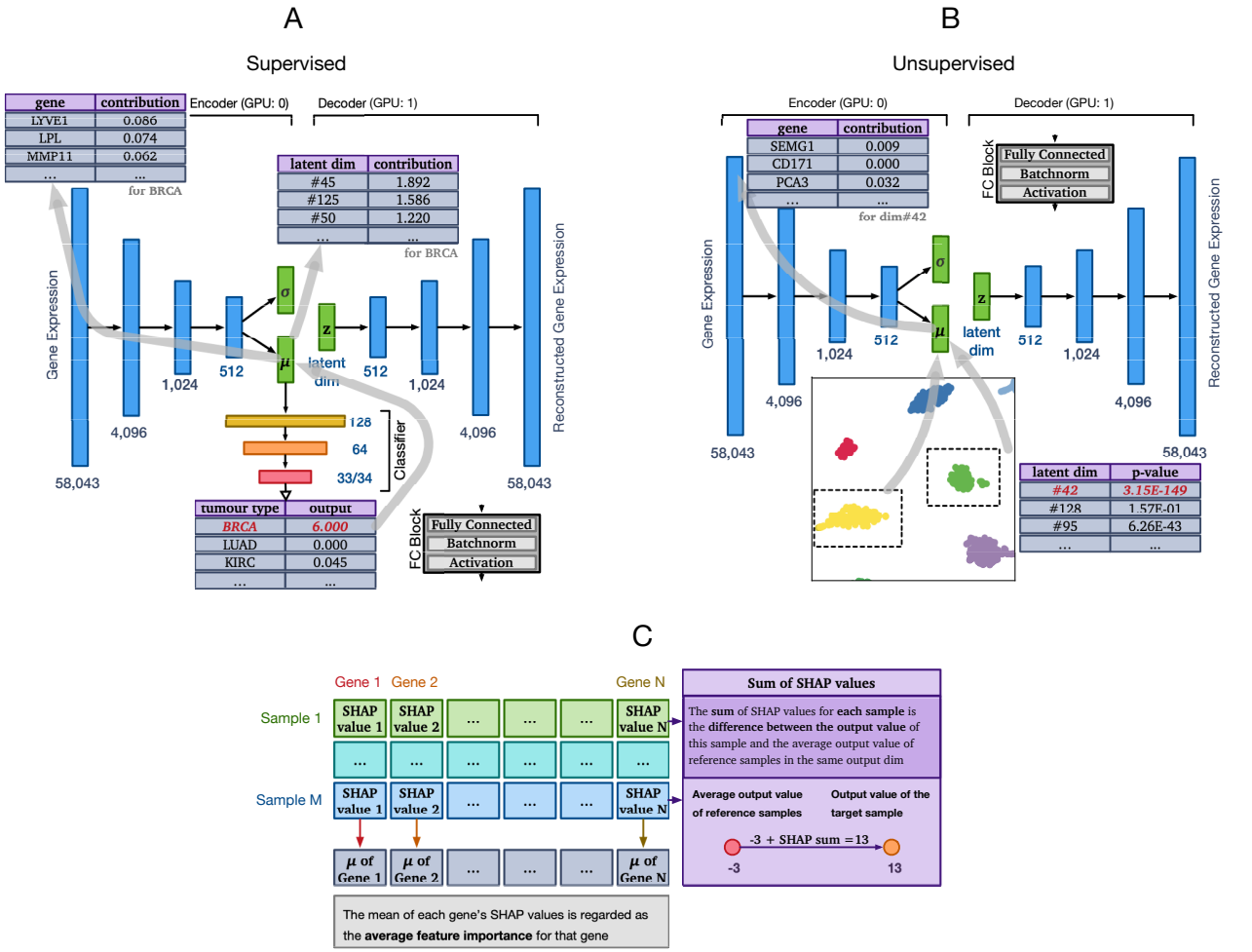

Fig. 1. (A) Overall architecture of the XOmiVAE model in the supervised scenario. The most important dimension for a prediction, the most important genes for a dimension and the most important genes for a prediction can be returned using Deep SHAP. The output values and contribution scores illustrated here are for demonstration. (B) Overall architecture of the XOmiVAE model in the unsupervised scenario. The most important dimensions used to separate two chosen clusters are found using a statistical test. The most important genes can then be returned using Deep SHAP. The p-values and contribution scores illustrated here are for demonstration. (C) An overview of the approach taken to process the SHAP values. SHAP values were calculated for multiple samples of interest and then averaged to provide the average feature importance for the gene. To the right, we demonstrate that the SHAP values for a sample sum up to the difference between the predicted output for the background and the output for the target sample.

Deep SHAP explains the difference between the output of target samples and background samples, detailed in Figure 1 (C). We chose either a random selection of samples or relevant samples to the target biomedical problem as the background samples. This choice depended on whether we were interested in the most important genes for a prediction when comparing all tumour types, or the most important genes to separate one specific set of samples from another. Where possible, 100 samples were used as the background, as recommended by Lundberg and Lee [2017b]. When there were less than 100 relevant samples with certain phenotype, we used the maximum number of samples available. We tested this assumption and the explanations were consistent between the different random samples used for the background.

XOmiVAE has the ability to provide explanations for both supervised (Figure 1 (A)) and unsupervised (Figure 1 (B)) learning. To achieve this, we adjusted the original Deep SHAP explanation technique to ensure it could take either the latent vector, or the output values of the classification network. As recommended by Shrikumar et al. [2017], we use the logits as opposed to the post-softmax probabilities, to calculate feature importance scores. DeepSHAP uses the key ideas from DeepLIFT, primarily the "summation-to-delta" property (the sum of the attributions over the input equals the difference from the reference of the output) [Shrikumar et al., 2017], which is represented by,

$$\sum_{n}^{i=1} C\Delta x_i \Delta o = \Delta o, \tag{4}$$

where r is the reference value (background value) chosen by the user, x is the input sample, $o = f(x)$ is the model output, $\Delta o = f(x) - f(r)$ is the difference between model output for the input sample and reference sample, and $\Delta x_i = x_i - r_i$ is the difference between the input and reference value. This enables the calculation of the Shapley values (feature importance

scores) for each gene, which indicate how to distribute the prediction of the model between the features. Large Shapley values therefore represent important features in the prediction of the model.

To explain the most important latent dimension for a prediction, we passed the bottleneck layer (μ value) to the Deep SHAP explainer, and the classification output values were backpropagated to assign a contribution score to each latent dimension. As for latent dimension contribution in unsupervised tasks, we calculated the mean and standard deviation of the latent vector values (μ value) for the two samples of interest and applied a Welch's t-test to obtain the most statistically significant dimension that separates the two samples. The correlation between the latent variables and the input molecular features was determined by backpropagating the latent vector and getting the contribution score for each pair of gene and latent dimension.

For each prediction, n SHAP values corresponding to n genes or latent dimensions were calculated in the XOmiVAE model to determine the contribution. The absolute values of SHAP values for each feature were averaged over a group of samples with the same label to indicate the overall contribution for each feature (Figure 1 (C)). This avoids the issue of scores canceling each other out for important features when averaged across samples. The sum of SHAP values for a particular sample represents the difference between the background output value and the output value of this sample (Figure 1 (C)).

Biostatistical analysis

To assess the results returned from XOmiVAE we compared the most important genes (i.e. the genes with highest absolute SHAP values) to the differentially expressed genes (DEGs) between the normal and cancer tissue. We used TCGAbiolinks [Colaprico et al., 2015], an R package for the differential expression analysis. The DEGs were selected with a false discovery rate (FDR) adjusted p-value cut-off of 0.05 and absolute log_2 fold change threshold of $>= 3$. To gain mechanistic insight into the most important genes detected by XOmiVAE for cancer classification, we used the Broad Institute's Gene Set Enrichment Analysis (GSEA) software [Subramanian et al., 2005] to perform pathway enrichment analysis. Additionally, we used the curated gene sets from online databases to test for subtypes pathways, including the gene ontology (GO) [Consortium, 2004], Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Gotot, 2000], and the reactome pathway database (reactome) [Fabregat et al., 2018]. g:Profiler [Raudvere et al., 2019] was used to obtain and visualise the top pathways found. GeneCard [Stelzer et al., 2016] was used to obtain specific TCGA tumour gene sets.

Results and discussion

Multi-level explanation of XOmiVAE

Most important genes for cancer classification

We trained XOmiVAE on the TCGA pan-cancer data and computed the contribution values for each gene, for each tumour prediction. The model achieves high accuracy for differentiating between normal and tumour tissue, e.g. BRCA (99.6%) and normal breast tissue (100%), and hence the model is able to pick up on cancer specific differences between the tissues. The contribution values followed a power-law distribution, which suggested the majority of input features

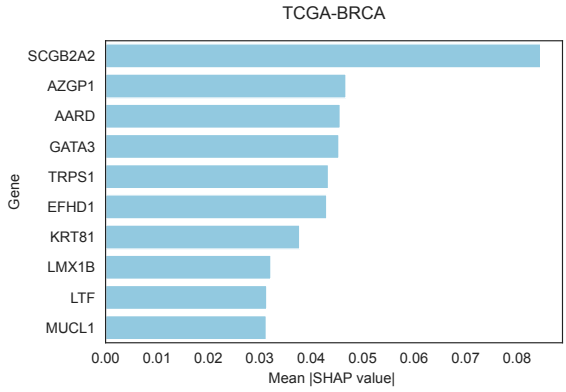

Fig. 2. The top 10 genes for the prediction of breast invasive carcinoma (BRCA) as an example of the most important genes for tumour prediction that we can obtain from XOmiVAE. Random training samples were used as the reference.

(*i.e.* genes) were unimportant for tumour prediction (see Supplementary Figure S1-S2). As an example, we illustrated the top 10 genes (*i.e.* genes with highest contribution values), that contributed most to Breast invasive carcinoma (BRCA) classification, Figure 2. This demonstrated the explanation of input features in XOmiVAE. To validate whether the important genes found by XOmiVAE made biological sense, we analysed the biological function of the most and least important genes and found that the most decisive genes are known to be important to BRCA. For example, the top gene for BRCA, *SCGB2A2*, which codes for the protein Mammaglobin A, is highly specific of breast tissue and increasingly being used as a marker for breast cancer [Lacroix, 2006]. The second most important gene, *AZGP1*, is associated with an aggressive breast cancer phenotype [Parris et al., 2013]. The 20 least important genes are either non-coding RNAs or pseudogenes with minor biological function, which are reasonable to be irrelevant when distinguishing breast tumour from normal breast tissue. A list of ranked genes with their contribution values (SHAP) can be obtained for each of the 32 tumour types predicted by XOmiVAE (see Supplementary Figure S3 to S6).

Most important dimensions for cancer classification

By passing in an interim layer to Deep SHAP explainer, it is possible for XOmiVAE to obtain the most important neuron for a prediction in a specific layer. In the case of OmiVAE, the most intriguing interim layer to explain is the bottleneck layer, where the high dimensional gene expression data is reduced into a latent representation with lower dimensionality (128 dimensions). Therefore, the input of the first layer in the classification network was intercepted and explained using XOmiVAE. As an example, we show the top dimensions for different subtypes of kidney tumours: kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP). The top two dimensions are different among kidney tumour subtypes and the third one was shared (Table 1), which is therefore possible the dimension responsible for separating the kidney located tumours. Additionally, it is possible to find the most associated genes and, therefore, the most related biological pathways to a specific dimension. We investigate the

Table 1. The top dimensions for kidney tumour subtypes: KICH, KIRC and KIRP.

Dimension Rank	Kidney Cancer Subtypes		
	KICH	KIRC	KIRP
1st	45	20	42
2nd	50	83	67
3th	**35**	**35**	**35**
4th	111	53	125
5th	42	103	45

top 15 genes for the shared kidney dimension 35 as an example (Supplementary Figure S12). These results can be obtained for each dimension for each tumour type.

Validation by biomedical knowledge

Biomedical meaning of important genes

To first validate the genes returned by XOmiVAE, we compared the genes, as ranked by importance, for each tumour type, with genes associated with that tumour type from GeneCard [Stelzer et al., 2016]. GeneCard was chosen due to its comprehensive disease gene sets, which are obtained from 150 different sources [Stelzer et al., 2016], and therefore covered the majority of tumour types in our analysis. We selected the genes to compare at 100 different thresholds, spaced evenly from 1 (most important gene) to 58,043 (total number of genes). The results were also compared to different thresholds of a random sample of genes (averaged over 10 random seeds). The results were plotted as a ROC curve for the True Positive Rates (TPR) and False Positive Rates (FPR). The TPR were calculated by,

$$\frac{\#\ of\ top\ genes\ which\ are\ GeneCard\ disease\ genes}{\#\ of\ GeneCard\ disease\ Genes}. \quad (5)$$

And the FPR were calculated by,

$$\frac{\#\ of\ top\ genes\ which\ are\ not\ GeneCard\ disease\ genes}{\#\ of\ genes\ not\ associated\ with\ the\ GeneCards\ disease}. \quad (6)$$

21 tumour types had gene sets found in GeneCard and were therefore chosen for analysis. The ROC curves and AUC metrics are shown Supplementary Figures S7 to S9. Two example ROC curves are illustrated in Figure 3. All 33 tumour types had an AUC metric considerably higher than the random samples which suggests that the most important genes returned by XOmiVAE are biologically relevant.

To further explore and understand the decisive genes found by XOmiVAE, the top genes for each tumour classification were evaluated using gene set enrichment analysis. We used g:Profiler, a web server for functional enrichment analysis [Raudvere et al., 2019], to identify the most significant GO terms enriched in the top tumour genes. Supplementary Table S2 lists the GO terms that are significantly overrepresented in top BRCA genes. The most significant GO terms related to the extracellular matrix organisation, an area of focus within breast cancer research [Walker et al., 2018]. A break down of the pathways found from the other sources used in g:Profiler is shown in Supplementary Figure S10.

The top 100 most important genes for BRCA classification over normal breast tissue were compared with the differentially

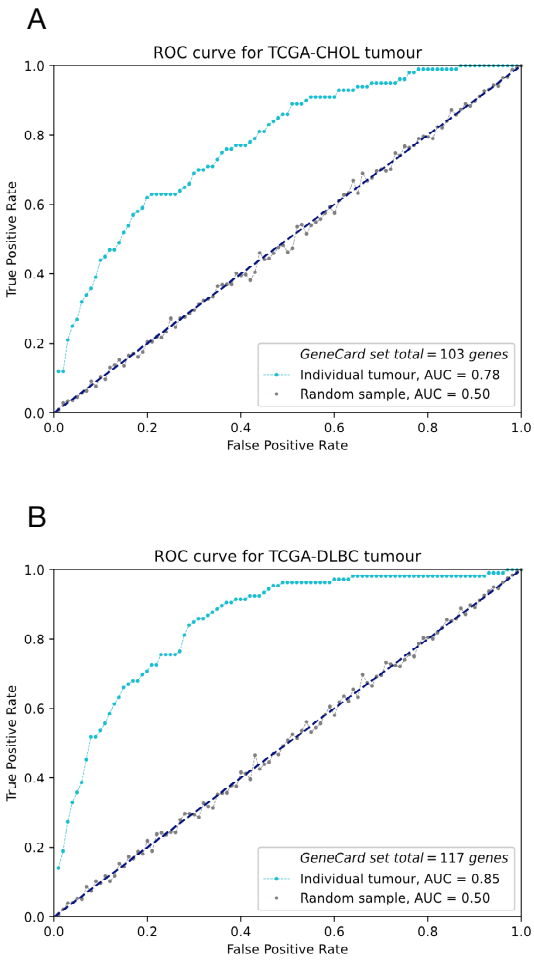

Fig. 3. AUC-ROC curves for the genes as ranked by importance, for (A) cholangiocarcinoma (CHOL) and (B) lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) tumour prediction against the GeneSet gene list for the respective tumour type. A random sample of genes is used as a comparison. 100 thresholds of genes were used, spaced evenly from 1 (most important gene) to 58,043 (total number of genes).

expressed genes (DEGs) between the target tumour and normal tissue. This helps ascertain the similarity between top genes found by XOmiVAE and DEGs obtained by the traditional statistical method. We find that there is an overlap of 48 out of the 100 top contribution genes when comparing BRCA versus normal breast tissue as an example (Figure 4). The top DEGs were chosen according to the threshold of $FDR < 0.05$ and $|LogFC| >= 3$ (see Supplementary Table S2 for details). The top genes obtained by XOmiVAE do not solely include DEGs, likely because the model has to ensure that the genes chosen for classification are different between cancers. Therefore, the DEGs that are common between cancers are not chosen as important features.

Biomedical meaning of important dimensions

To further understand the most important dimensions involved in tumour prediction, we analysed the biological meaning of the key genes used by the dimensions. As an example, we analyse the highest shared dimension (*i.e.* dimension 35) in

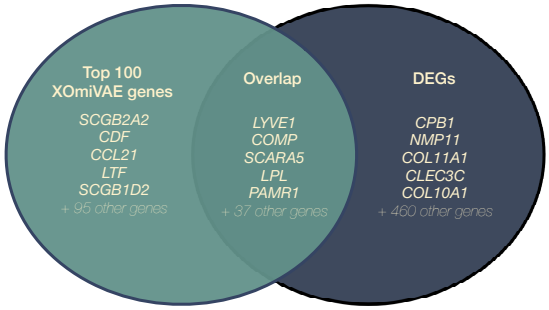

Fig. 4. A Venn diagram representing the overlap between the DEGs and top contribution genes, highlighting a total of 42 DEGs found in the top 100 contribution genes.

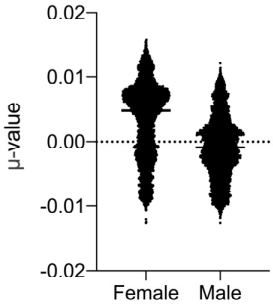

Fig. 5. Supervised latent values for female and male tissue samples. The dimension with the largest significance (p-value=3.6e-249) was chosen.

the kidney cancers KIRC, KIRP and KICH (Supplementary Figure S12). *APQ2* is the most important gene for that dimension for all three cancer subtypes. *APQ2* is located in the apical cell membranes of the collecting duct principal cells in kidneys. Additionally, all other high ranking genes such as *UMOD*, *SCNN1G* and *SCNN1B* are all well known genes associated with kidney functions [Carney, 2016, Hanukoglu and Hanukoglu, 2016]. As another example, we also explain dimension 42 and 73, the 1^{st} and 2^{nd} most important dimension for lung adenocarcinoma (LUAD) classification respectively, as shown in Table 2. The top genes were calculated using random training samples as the background value, to show the most important genes for LUAD versus all the other sample types. We show that dimension 42 relies heavily on the immune response pathways, whilst dimension 73 relates to the developmental process, albeit with one highly significant immune response pathway. The top gene for dimension 73 is pulmonary-associated surfactant protein C (*SPC*), a surfactant protein essential for lung function, and the top gene for dimension 42 is progestagen associated endometrial protein (*PAEP*), an immune system modulator, both of which have been implicated in LUAD [Schneider et al., 2015, Yamamotot et al., 2005].

We found that the most important input features for the latent dimensions varied according to the tumour type used for the analysis (Table 2). This demonstrates a possible limitation of previous methods explaining gene expression classification networks using solely a connection weight approach, for example by Way and Greene [2017] and Bica et al. [2019], which show no specificity for different input samples and different prediction targets. Table 2 shows that for BRCA, dimension 42 uses the genes related to blood vessels, and dimension 73 relies on the embryonic genes. However, this contrasts with the most important pathways that these dimensions used for LUAD classification. XOmiVAE is able to capture this as it detects the activation of a neuron using Deep SHAP, as opposed to solely the weights involved.

To further understand the latent space of the classification network, we tested whether there was a dimension that separated between female and male tissue samples. We observed a large statistical difference (p value $= 3.6 \times 10^{-249}$) between genders on dimension 78 in the classification model (Figure 5). To understand how dimension 78 captured gender, Deep SHAP was used to explain the genes involved. We found that *XIST*, a gene on chromosome X, was within the top 5 genes of the dimension 78 (Table 3). *XIST* is one of the key genes involved in the transcriptional silencing of one of the X chromosomes [Zuccotti and Monk, 1995].

Validation by the performance of downstream tasks

Influence of important genes for model performance

To further evaluate the results, we compared the classification performance of models using the top 20 XOmiVAE genes or 20 random genes for each target tumour type of interest. Four metrics including the F1-score (F1), Positive Predictive Value (PPV), True Positive Rate (TPR) and Area Under the Curve (AUC) were applied, and the performance of models using 20 random genes was averaged over 10 random seeds. A highly significant performance difference can be observed in Table 4, which indicates the importance and the contribution to cancer classification tasks of the top genes obtained by XOmiVAE. Additionally, we calculated the average metrics for all other tumour types except the target one and found that whilst there was also an increase in metrics from the randomly selected genes, it was not as significant as the increase for the target tumour type. This suggests that the most important genes learnt by XOmiVAE are specific for certain target tumour type.

To approximate the most important genes for the overall model, we summed the ranking of genes for each tumour type, with the most important gene having a ranking of 1^{st} and the least important gene ranking $58,043^{th}$, and selected 20 genes with the top 20 lowest sum rankings to retrain the model and calculate the overall accuracy (Table 5). We then compared it with the performance of a model trained by 20 random genes and a model trained by the overall top 20 least important genes with the highest ranking sums. Using the 20 most important genes we observed a significant improvement in accuracy over using a random selection of 20 genes. Additionally, we found that the least important 20 genes caused a large decrease in accuracy compared to a random selection of genes. These results suggest a possible role of using the XOmiVAE contribution values for feature selection in model training with high dimension omics data.

Influence of important dimensions for model performance

To understand whether XOmiVAE accurately detected the most and least important dimensions in the latent space, we evaluated the effect of knocking out the most important dimensions, Table 6. We set the output of the target dimension to -1 when the output value was positive, and set the output

Table 2. The biological pathways enriched for dimension 42 and 73 when classifying BRCA and LUAD.

Dimension ID	Tumour Type	GO Biological Process	FDR adjusted p-value
42	LUAD	Humoral immune response	1.8×10^{-8}
		Response to bacterium	2.0×10^{-8}
		Response to stimulus	2.0×10^{-7}
		Immune system process	2.5×10^{-6}
		Response to other organism	3.7×10^{-6}
	BRCA	Circulatory system process	4.7×10^{-7}
		Blood circulation	1.3×10^{-6}
		Developmental process	4.6×10^{-5}
		Regulation of blood pressure	2.0×10^{-5}
		Humoral immune response	2.5×10^{-5}
73	LUAD	Response to external stimulus	2.4×10^{-5}
		Response to bacterium	4.5×10^{-5}
		Anatomical structure morphogenesis	5.4×10^{-5}
		Tube development	7.4×10^{-5}
		Response to biotic stimulus	1.8×10^{-4}
	BRCA	Anterior/posterior pattern specification	1.2×10^{-7}
		Embryonic morphogenesis	1.5×10^{-6}
		Embryo development	2.0×10^{-6}
		Embryonic skeletal system morphogenesis	2.3×10^{-6}
		Anatomical structure development	2.3×10^{-6}

Table 3. The top five genes for dimension 78 when separating female and male samples in the classification model of OmiVAE.

Gene	Contribution value	Chromosome
CLDN3	0.00031	chr7
SLPI	0.00031	chr20
WFDC2	0.00031	chr20
XIST	0.00030	chrX
MMP1	0.00029	chr11

of the target dimension to 1 when the output value was negative, based off a similar ablation approach by Morcos et al. [2018]. This ensures that the output is perturbed from the original value. Individually, the most important dimensions did not have a large effect when ablated, which is likely due to model saturation, a feature of neural networks that Deep SHAP addresses whereas other interpretability techniques fail to capture [Shrikumar et al., 2017]. When the top dimensions combined were ablated, the classification accuracy fell to 0. This is in contrast to the least important dimensions, which did not have any effect on the network when knocked out, individually or combined. This provides evidence to support the most and least important dimensions obtained by XOmiVAE.

Different results depending on reference chosen

The SHAP library requires a background sample to determine the most important features for a set of predictions. One of the recommended choices for this background sample is a random sample from the training set. However, we can also choose samples with certain phenotype as the reference to compare with for certain prediction rather than using a random selection of the training data, which can be more informative in some cases. For example when explaining the important genes to differentiate gender we use samples from the opposite gender as the background.

To further understand the effect of the background, we compared the important genes involved in BRCA classification using both a random selection from the training set and the normal breast tissue samples, Figure 6. 25 of the top 50 XOmiVAE genes were shared between the two reference selection methods. To gain a clearer understanding of the different biological pathways enriched from the top genes when using the two different reference samples sets, we compared the g:Profiler pathway enrichment results (Supplementary Figure S10 and S11). There is a decreased enrichment of extracellular pathways when using a set of random training data to explain the BRCA predictions. As alluded earlier, extracellular pathways have been shown to be involved in BRCA progression from normal tissue [Walker et al., 2018]. It is possible that when using normal breast tissue as reference samples, the specific genes that lead to breast cancer are more pronounced, as opposed to also relying on breast tissue genes as would be the case when differentiating BRCA from all the other predictions. Therefore, it is shown that XOmiVAE is able to gain a more focused understanding of the most important genes for a tumour type by selecting the appropriate background samples.

Explaining unsupervised clustering results

As an example of explaining the unsupervised clustering results, we used Basal-like (Basal) and Luminal B (LumB) breast tumour subtypes. However, explaining the dimensions of VAEs would usually be important for when it is important to understand the genes involved in subtype clustering of cancers that are yet to be defined, and therefore lack labels that could be used for supervised tasks. Figure 7 shows the two most decisive dimensions splitting the subtypes. As the most statistically significant dimension for separating the two subtypes was dimension 100, we evaluated the enriched pathways when this dimension is used to separate Basal and LumB. Here, the μ value for a subtype (LumB) was treated as the output and backpropagated through the network using

Table 4. The evaluation metrics of cancer classification using only the top 20 most important genes obtained by XOmiVAE (columns 1 and 3) or 20 random genes chosen from the overall gene set of 58,043 features (columns 2 and 4). The metrics for each individual tumour type of interest are shown in columns 1 and 2, and the metrics for all of the other tumour types (except the target one) are shown in columns 3 and 4. The results were averaged among all 33 target tumour types and 10 random seeds.

	Average metric across all 33 tumour types			
	Target tumour trained by top 20 XOmiVAE genes	Target tumour trained by 20 random genes	All other tumours trained by top 20 XOmiVAE genes of the target tumour	All other tumours trained by 20 random genes of the target tumour
F1	0.90 ± 0.11	0.46 ± 0.21	0.66 ± 0.11	0.48 ± 0.01
PPV	0.91 ± 0.11	0.48 ± 0.20	0.69 ± 0.08	0.50 ± 0.01
TPR	0.91 ± 0.10	0.66 ± 0.11	0.48 ± 0.01	0.48 ± 0.01
AUC	0.94 ± 0.07	0.67 ± 0.11	0.83 ± 0.06	0.68 ± 0.00

Table 5. The accuracy of XOmiVAE using the full gene set, the top 20 contribution genes for all tumours, 20 random genes and the bottom 20 contribution genes for all tumours.

Gene set	N	Overall accuracy
Full gene set	58,043	$96.85\% \pm 0.46\%$
Top 20 genes for all tumours	20	$87.07\% \pm 0.38\%$
20 random genes	20	$56.10\% \pm 0.24\%$
Bottom 20 genes for all tumours	20	$1.68\% \pm 0.37\%$

Table 6. The accuracy difference for each tumour type when the most important and least important dimensions were individually or together removed from the network. Values represent the mean and standard deviation of the accuracy difference among 33 tumour types.

Ablated dimension	Accuracy difference
1st	$-11.9\% \pm 19.9\%$
2nd	$-11.9\% \pm 20.2\%$
3rd	$-2.9\% \pm 6.5\%$
Top three combined	$-95.9\% \pm 2.0\%$
126th	$0.0\% \pm 0.3\%$
127th	$0.0\% \pm 0.0\%$
128th	$0.0\% \pm 0.3\%$
Bottom three combined	$0.0\% \pm 0.3\%$

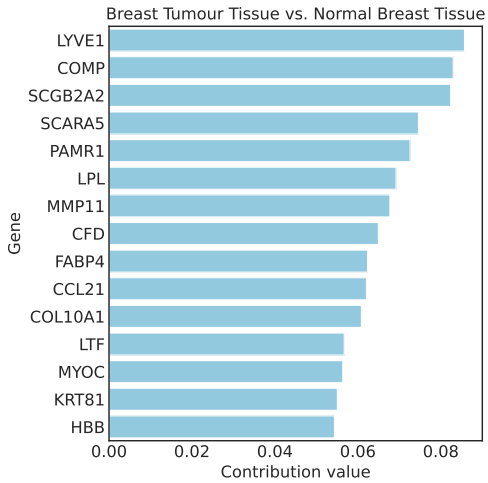

Fig. 6. The top 15 genes obtained by XOmiVAE for the classification of BRCA using normal breast tissue samples as the reference.

Deep SHAP, and compared to the other subtype (Basal) as the reference. As we were interested in validating whether the model can explain the subtype specific pathways, we evaluated the top 100 genes using the Broad Institute's curated pathway database [Subramanian et al., 2005], which includes pathways from experiments comparing the subtypes.

In Table 7 we can see the pathways are highly specific for the subtypes. A key differentiating feature between the subtypes is that LumB is estrogen-receptor (*ESR1*) positive, and Basal is *ESR1* negative and in Table 7 we can see the top pathways also include the genes that differentiate between the *ESR1* negative and *ESR1* positive tumours. Table 8 shows the results when the three other BRCA subtypes (LumA, Her2 and Basal) are used as the background samples when explaining subtype LumB. The results show that a larger range of subtype pathways are present in the most important features. These results proves that it is a useful method of being able to obtain the unique genes for one subtype versus multiple other subtypes.

This is, to the best of our knowledge, the first attempt at using an activation-based explanation method for clustering generated by autoencoders. Typically, differential gene expression methods, such as DESeq2 [Love et al., 2014], is used to explain differences in clusters, which treats each gene as independent. More recent techniques improve on this, such as Global Counterfactual Explanation (GCE) [Plumb et al., 2020] and Gene Relevance Score (GRS) [Angerer et al., 2020]. However, GCE requires a linear embedding, and the embedding of GRS is constrained to ensure the gradients are easy to calculate. XOmiVAE allows for a non-linear embedding, and becomes one of the first activated-based deep learning interpretation method to explain novel clusters generated by VAEs.

Conclusion

Here we present an explainable variational autoencoder based deep learning method for high dimensional omics data analysis named XOmiVAE. We illustrate that it is possible to explain the supervised task of the network and obtain the most important genes and dimensions for a prediction. We also show that it is possible explain the most important genes in the unsupervised part of the network, and therefore provide a method of explaining deep learning based clustering. We evaluate the explanations and show that they make biological sense. Additionally, we offer important steps to consider when interpreting deep learning models for tumour classification. For

Table 7. The top pathways for differentiating LumB and Basal (BRCA subtypes), using the Broad Institute's curated pathway database.

Pathway	Genes in Overlap	P-value
Genes up-regulated in breast cancer samples positive for ESR1 compared to the ESR1 negative tumours	27	1.15e-47
Genes down-regulated in basal subtype of breast cancer samples	39	4.25e-43
Genes up-regulated in bone relapse of breast cancer	24	2.6e-42
Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR basal (ESR1- AR-) and luminal (ESR1+ AR+)	29	1.85e-37
Genes up-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	26	1.82e-30

Table 8. The top pathways for differentiating between LumB and the other three subtypes (Basal, LumA and Her2), using the Broad Institute's curated pathway database.

Pathway	Genes in Overlap	P-value
Genes down-regulated in basal subtype of breast cancer samples.	27	1.15 e-47
Genes up-regulated in bone relapse of breast cancer.	39	4.25 e-43
Genes down-regulated in ductal carcinoma vs normal ductal breast cells.	24	2.6 e-42
Genes down-regulated in nasopharyngeal carcinoma (NPC) positive for LMP1, a latent gene of Epstein-Barr virus (EBV).	29	1.85 e-37
Genes up-regulated in breast cancer samples positive for ESR1 compared to the ESR1 negative tumours.	26	1.82 e-30

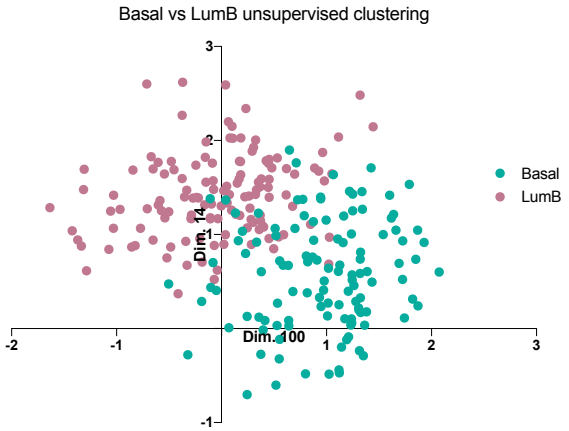

Fig. 7. The two most important dimensions for splitting Basal and LumB subtypes in the latent space.

example, we show the importance of choosing a background sample that makes biological sense when explaining the model, and we show the limitations of connection weight based methods to explain latent dimensions. We believe XOmiVAE is a promising technique that could help discover novel biomedical knowledge from deep learning models.

Key Points

- XOmiVAE is a novel interpretable deep learning model for cancer classification using high-dimensional omics data.
- XOmiVAE provides contribution of each input molecular feature and latent dimension to the prediction.
- XOmiVAE is able to explain unsupervised clusters produced by the VAE clustering part of the network.
- XOmiVAE explanations of the downstream prediction were evaluated by biological annotation and literature, which aligned with current domain knowledge.
- XOmiVAE shows great potential for novel biomedical knowledge discovery from deep learning models.

Availability

The source code have been made publicly available on GitHub[2]. The TCGA pan-cancer dataset can be downloaded from the UCSC Xena data portal[3].

Supplementary data

Supplementary is available at GitHub[4].

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [764281].

[2] https://github.com/zhangxiaoyu11/XOmiVAE/

[3] https://xenabrowser.net/datapages/

[4] https://github.com/zhangxiaoyu11/XOmiVAE/blob/main/documents/supplementary.pdf

Competing interests

The authors declare that they have no conflict of interest.

References

P. Angerer, D. S. Fischer, F. J. Theis, A. Scialdone, and C. Marr. Automatic identification of relevant genes from low-dimensional embeddings of single cell rnaseq data. 2020. doi: 10.1101/2020.03.21.000398.

B. Azarkhalili, A. Saberi, H. Chitsaz, and A. Sharifi-Zarchi. Deepathology: Deep multi-task learning for inferring molecular pathology from cancer transcriptome. *Scientific Reports*, 9, 2019.

C. B. Azodi, J. Tang, and S.-H. Shiu. Opening the black box: Interpretable machine learning for geneticists. 2020. doi: 10.20944/preprints202002.0239.v1.

I. Bica, H. Andrés-Terré, A. Cvejic, and P. Liò. Unsupervised generative and graph representation learning for modelling cell differentiation. *Scientific Reports*, 2019. doi: 10.1101/801605.

E. F. Carney. Evolving risks of umod variants. *Nature Reviews Nephrology*, 12(5):257–257, 2016. doi: 10.1038/nrneph.2016.46.

A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, and et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, 44(8), 2015. doi: 10.1093/nar/gkv1507.

G. O. Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32(90001), 2004. doi: 10.1093/nar/gkh036.

A. Fabregat, F. Korninger, G. Viteri, K. Sidiropoulos, P. Marin-Garcia, P. Ping, G. Wu, L. Stein, P. D'Eustachio, H. Hermjakob, and et al. Reactome graph database: Efficient access to complex pathway data. *PLOS Computational Biology*, 14(1), 2018. doi: 10.1371/journal.pcbi.1005968.

M. Goldman, B. Craft, A. Kamath, A. N. Brooks, J. Zhu, and D. Haussler. The ucsc xena platform for cancer genomics data visualization and interpretation. *bioRxiv*, 2018.

B. Hanczar, F. Zehraoui, T. Issa, and M. Arles. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics volume*, page 501, Nov 2020. doi: https://doi.org/10.1186/s12859-020-03836-4.

I. Hanukoglu and A. Hanukoglu. Epithelial sodium channel (enac) family: Phylogeny, structure-function, tissue distribution, and associated inherited diseases. *Gene*, page 95–132, Apr 2016. doi: 10.1016/j.gene.2015.12.061.

M. Kanehisa and S. Gotot. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations, ICLR*, 2014.

M. Lacroix. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-Related Cancer*, 13(4):1033–1067, 2006. doi: 10.1677/erc-06-0001.

LeCun, Bengio, and Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

A. Lemsara, S. Ouadfel, and H. Fröhlich. Pathme: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics*, 21(1), 2020. doi: 10.1186/s12859-020-3465-2.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), 2014. doi: 10.1186/s13059-014-0550-8.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017a.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 4768–4777, 2017b.

C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, Jul 2016. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4945831/`.

A. Morcos, D. Barrett, and N. Rabinowitz. On the importance of single directions for generalization. *International Conference on Learning Representations*, 2018.

T. Z. Parris, A. Kovács, L. Aziz, S. Hajizadeh, S. Nemes, M. Semaan, E. Forssell-Aronsson, P. Karlsson, and K. Helou. Additive effect of the azgp1, pip, s100a8 and ube2c molecular biomarkers improves outcome prediction in breast carcinoma. *International Journal of Cancer*, 134(7):1617–1629, 2013. doi: 10.1002/ijc.28497.

G. Plumb, J. Terhorst, S. Sankararaman, and A. Talwalkar. Explaining groups of points in low-dimensional representations. *ICML*, 2020.

U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1), 2019. doi: 10.1093/nar/gkz369.

F. Sanchez-Vega, M. Mina, J. Armenia, W. K. Chatila, A. Luna, K. C. La, S. Dimitriadoy, D. L. Liu, H. S. Kantheti, S. Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

M. A. Schneider, M. Granzow, A. Warth, P. A. Schnabel, M. Thomas, F. J. Herth, H. Dienemann, T. Muley, and M. Meister. Glycodelin: A new biomarker with immunomodulatory functions in non–small cell lung cancer. *Clinical Cancer Research*, 21(15):3529–3540, 2015. doi: 10.1158/1078-0432.ccr-14-2464.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, and et al. The genecards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1), 2016. doi: 10.1002/cpbi.5.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

S. Tasaki, C. Gaiteri, S. Mostafavi, and Y. Wang. Deep learning decodes the principles of differential gene expression. *Nature Machine Intelligence*, 2(7):376–386, 2020. doi: 10.1038/s42256-020-0201-6.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, page 2579–2605, Nov 2008.

C. Walker, E. Mojares, and A. Hernández. Role of extracellular matrix in development and cancer progression. *Int J Mol Sci*, Oct 2018. doi: 10.3390/ijms19103028.

G. P. Way and C. S. Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. 2017. doi: 10.1101/174474.

J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.

O. Yamamotot, H. Takahashi, M. Kirasawa, and H. Chiba. Surfactant protein gene expressions for detection of lung carcinoma cells in peripheral blood. *Respiratory Medicine*, Sep 2005. doi: 10.1016/j.rmed.2005.02.009.

X. Zhang, J. Zhang, K. Sun, X. Yang, C. Dai, and Y. Guo. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769, 2019.

X. Zhang, Y. Xing, K. Sun, and Y. Guo. Omiembed: a unified multi-task deep learning framework for multi-omics data. *arXiv preprint arXiv:2102.02669*, 2021.

M. Zuccotti and M. Monk. Methylation of the mouse xist gene in sperm and eggs correlates with imprinted xist expression and paternal x–inactivation. *Nature Genetics*, 9(3):316–320, 1995. doi: 10.1038/ng0395-316.

Eloise Withnell is currently a PhD candidate at Department of Health Informatics, University College London, London, UK.

Xiaoyu Zhang is currently a PhD candidate at Data Science Institute, Imperial College London, London, UK.

Kai Sun is currently the acting operations manager of Data Science Institute, Imperial College London, London, UK.

Yike Guo is currently the co-director of Data Science Institute, Imperial College London, London, UK and vice-president of Hong Kong Baptist University, Hong Kong, China.