Briefings in Bioinformatics, 22(6), 2021, 1–11

https://doi.org/10.1093/bib/bbab315
Problem Solving Protocol

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

Eloise Withnell[†], Xiaoyu Zhang[†], Kai Sun and Yike Guo

Corresponding author: Xiaoyu Zhang, Data Science Institute, Imperial College London, SW7 2AZ London, UK. Tel: +44 0207 594 8630; Fax: +44 0207 594 8630; E-mail: x.zhang18@imperial.ac.uk

[†]The first two authors contributed equally to this paper.

Abstract

The lack of explainability is one of the most prominent disadvantages of deep learning applications in omics. This ‘black box’ problem can undermine the credibility and limit the practical implementation of biomedical deep learning models. Here we present XOmiVAE, a variational autoencoder (VAE)-based interpretable deep learning model for cancer classification using high-dimensional omics data. XOmiVAE is capable of revealing the contribution of each gene and latent dimension for each classification prediction and the correlation between each gene and each latent dimension. It is also demonstrated that XOmiVAE can explain not only the supervised classification but also the unsupervised clustering results from the deep learning network. To the best of our knowledge, XOmiVAE is one of the first activation level-based interpretable deep learning models explaining novel clusters generated by VAE. The explainable results generated by XOmiVAE were validated by both the performance of downstream tasks and the biomedical knowledge. In our experiments, XOmiVAE explanations of deep learning-based cancer classification and clustering aligned with current domain knowledge including biological annotation and academic literature, which shows great potential for novel biomedical knowledge discovery from deep learning models.

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

Introduction

High-dimensional omics data (e.g. gene expression and DNA methylation) comprise up to hundreds of thousands of molecular features (e.g. gene and CpG site) for each sample. As the number of features is normally considerably larger than the number of samples for omics datasets, the genome-wide omics data analysis suffers from the ‘the curse of dimensionality’, which often leads to overfitting and impedes wider application. Therefore, performing feature selection and dimensionality reduction prior to the downstream analysis has become a common practice in omics data modelling and analysis [24]. Standard dimensionality reduction methods like principal component analysis [32] learn a linear transformation of the high-dimensional data, which struggles with the complicated non-linear patterns that are intractable to capture from omics data. Other non-linear methods such as t-distributed stochastic neighbor embedding [41] and uniform manifold approximation and projection [23] have become increasingly popular but still have limitations in terms of scalability.

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.

Submitted: 26 May 2021; **Received (in revised form):** 4 July 2021

Deep learning has proven to be a powerful methodology for capturing non-linear patterns from high-dimensional data [19]. Variational autoencoder (VAE) [17] is one of the emerging deep learning methods that have shown promise in embedding omics data to lower-dimensional latent space. With a classification downstream network, the VAE-based model is able to classify tumour samples and outperform other machine learning and deep learning methods [2, 15, 45, 46]. Among them, OmiVAE [46] is one of the first VAE-based multi-omics deep learning models for dimensionality reduction and tumour type classification. An accuracy of 97.49% was achieved for the classification of 33 pan-cancer tumour types and the normal control using gene expression and DNA methylation profiles from the Genomic Data Commons (GDC) dataset [12]. Similar to OmiVAE, DeePathology [2] applied two types of deep autoencoders, contractive autoencoder and VAE, with only the gene expression data from the GDC dataset, and reached accuracy of 95.2% for the same tumour type classification task. Hira *et al.* [15] adopted the architecture of OmiVAE with maximum mean discrepancy VAE and classified the molecular subtypes of ovarian cancer with an accuracy of 93.2–95.5%. Zhang *et al.* [45] synthesized previous models and developed a unified multi-task multi-omics deep learning framework named OmiEmbed, which supported dimensionality reduction, multi-omics integration, tumour type classification, phenotypic feature reconstruction and survival prediction. Despite the breakthrough of aforementioned work, a key limitation is prevalent among deep learning-based omics analysis methods. Most of these models are ‘black boxes’ with lack of explainability, as the contribution of each input feature and latent dimension towards the downstream prediction is obscured.

Various strategies have been proposed for interpreting deep learning models. Among them, the probing strategy, which inspects the structure and parameters learnt by a trained model, has been shown to be the most promising [3]. Probing strategies generally fall into one of three categories: connection weights-based, gradient-based and activation level-based approaches [25]. The connection weight-based approach sums the learnt weights between each input dimension and the output layer to quantify the contribution score of each feature [10, 28]. Way and Greene [43] and Bica *et al.* [4] adopted this probing strategy to explain the latent space of VAE on gene expression data. However, the connection weight-based approach can be limited or even misleading when positive and negative weights offset each other, when features do not have the same scale or when neurons with large weights are not activated [35]. In the gradient-based approach, contribution scores (or saliency) are measured by calculating the gradient when the inputs are perturbed [36]. Dincer *et al.* [8] applied a gradient-based approach, integrated gradients [39], to explain a VAE model for gene expression profiles. This approach overcomes limitations of the connection weights-based method. Despite this, it is inaccurate when small changes of the input do not effect the output [35]. The activation level-based approach conquers these drawbacks by comparing the feature activation level of an instance of interest and a reference instance [3]. An activation level-based method named layer-wise relevance propagation (LRP) has been used to explain a deep neural network for gene expression [13]. Nevertheless, LRP can produce incorrect results with model saturation [35]. Deep SHAP [22], which applies the key principles from DeepLIFT [35], has been used in a variety of biological applications [20, 40]. However, there is a lack of research on the application of Deep SHAP to interpret the latent space of VAE models and the VAE-based cancer classification using gene expression profiles.

Here we proposed explainable OmiVAE (XOmiVAE), a VAE-based explainable deep learning omics data analysis model for low-dimensional latent space extraction and cancer classification. OmiVAE took advantage of Deep SHAP [22] to provide the contribution score of each input molecular feature and omics latent dimension for the cancer classification prediction. Deep SHAP was selected as the interpretation approach of XOmiVAE due to its ability to provide more accurate explanations over other methods, which likely provides better signal-to-noise ratio in the top genes selected. With XOmiVAE, we are able to reveal the contribution of each gene towards the prediction of each tumour type using gene expression profiles. XOmiVAE can also explain unsupervised tumour type clusters produced by the VAE embedding part of the deep neural network. Additionally, we raised crucial issues to consider when interpreting deep learning models for tumour classification using the probing strategy. For instance, we demonstrate the importance of choosing reference samples that makes biological sense and the limitations of the connection weight-based approach to explain latent dimensions of VAE. The results generated by XOmiVAE were fully validated by both biomedical knowledge and the performance of downstream tasks for each tumour type. 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.

Methods

Datasets and pre-processing

The Cancer Genome Atlas Program (TCGA) [27] pan-cancer dataset, which comprise gene expression profiles of 33 various tumour types, was used in the experiment as a example to demonstrate the explainability of XOmiVAE. A total of 9081 samples from TCGA were selected for training and testing our proposed model, of which 407 were normal tissue samples. The TCGA dataset was downloaded from UCSC Xena data portal [11] (https://xenabrowser.net/datapages/, accessed on 1 May 2019). We followed the same omics data pre-processing step as OmiVAE [46] and OmiEmbed [45]. Genes targeting the Y chromosome, genes with zero expression level in all samples and genes with missing values (N/A) in more than 10% of the samples were removed to ensure the gene expression data were fair and clean across samples. Furthermore, the remaining N/A values that did not reach the 10% threshold were replaced by the expression level of corresponding genes. The fragments per kilobase of transcript per million mapped reads values were normalized to the unit interval of 0 to 1 to the meet input requirement of the network. The phenotype data of each sample were also downloaded from UCSC Xena, which is comprised of age and gender of the subjects and primary site and disease stage of the samples. The detailed cancer subtype information of each tumour sample was obtained from Sanchez-Vega *et al.* [33].

Explainable OmiVAE (XOmiVAE)

Based on vanilla OmiVAE, we proposed an interpretable deep learning model for cancer classification using high-dimensional omics data, named explainable OmiVAE, aka XOmiVAE. The overall architecture of XOmiVAE was illustrated in Figure 1. The input omics data, which were genome-wide gene expression profiles here, were first passed through a VAE embedding network to reduce the dimensionality of the input data from 58 043

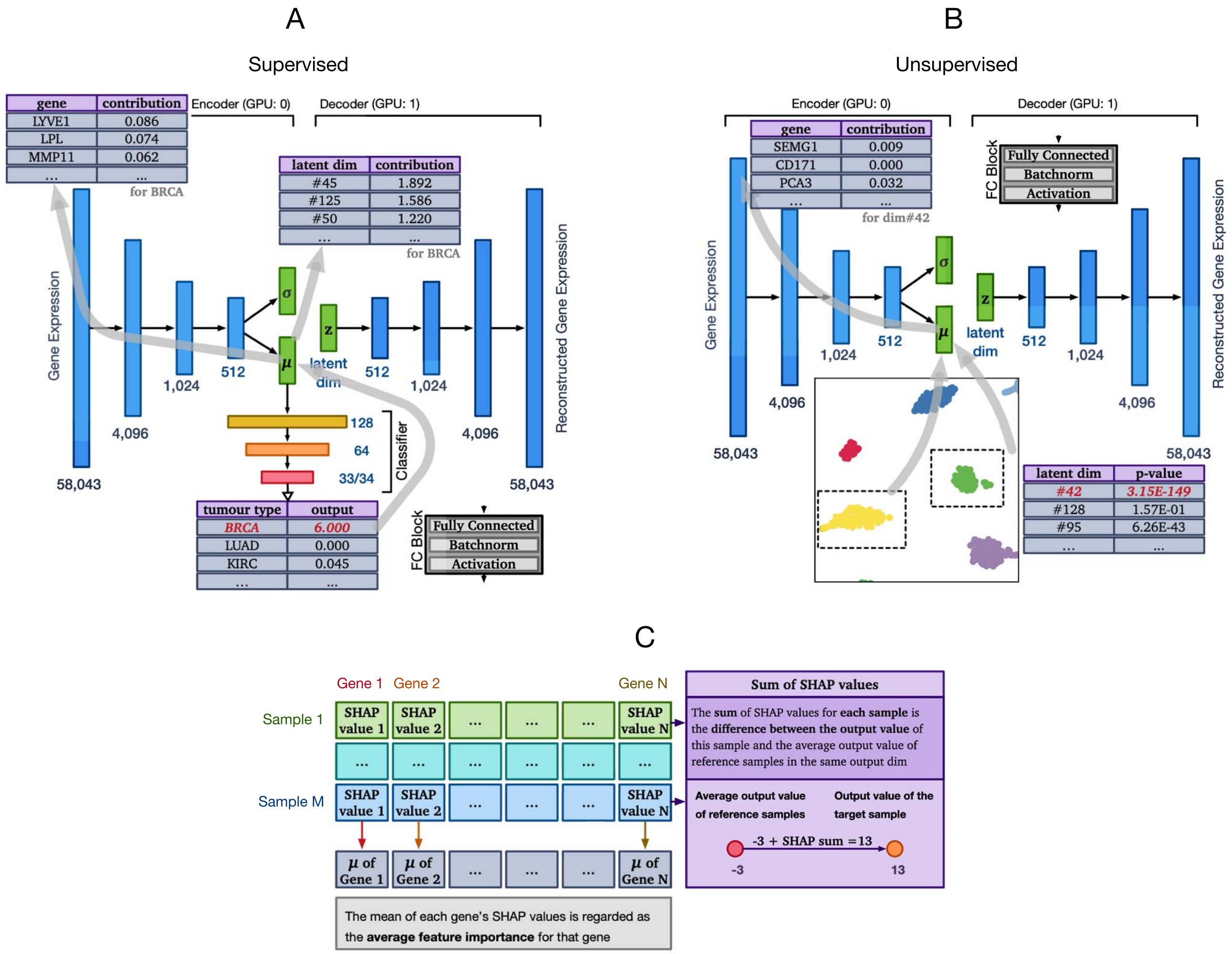

Figure 1. (**A**) Overall architecture of the XOmiVAE model in the supervised scenario. We can reveal the contribution score of each gene towards each cancer classification, the contribution score of each omics latent dimension learnt by VAE towards each cancer classification and the contribution score of each gene towards each omics latent dimension. The output values and contribution scores listed in the tables are just for demonstration. (**B**) Overall architecture of the XOmiVAE model in the unsupervised scenario. The importance of each omics latent dimension for separating two selected clusters can be obtained using the Welch's *t*-test. The contribution score of each gene can be revealed by the Deep SHAP explanation approach. The *P*-values and contribution scores listed in the tables are just for demonstration. (**C**) Illustration of how to appraise the contribution score of each gene. SHAP values were calculated for multiple samples of interest and then averaged to provide the average feature importance for each gene. To the right, we demonstrate that the SHAP values for each sample among different genes sum up to the difference between the average output value of the reference samples and the output value of the sample of interest on the same output dimension, which is another representation of the 'summation-to-delta' property.

to 128. The encoder of the embedding network contained two output vector, the mean vector $\boldsymbol{\mu}$ and the standard deviation vector $\boldsymbol{\sigma}$, which defined the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ of the latent variable $\mathbf{z}$ given the input omics data $\mathbf{x}$. In order to enable backpropagation for the sampling step, the reparameterization trick was applied according to Equation (1):

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

where $\boldsymbol{\epsilon}$ is a random variable sampled from a unit Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The VAE network of XOmiVAE was optimized by maximizing the variational evidence lower bound (ELBO) defined in Equation (2):

$$\mathrm{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right]. \tag{2}$$

$q_\phi(\mathbf{z}|\mathbf{x})$ is the variational distribution introduced to approximate the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$, where ϕ is the set of learnable parameters of the encoder and θ is the set of learnable parameters of the decoder. Equation (2) can further transform to Equation 3:

$$\mathrm{ELBO} = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\right) \tag{3}$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is the conditional distribution and D_{KL} is the Kullback–Leibler (KL) divergence between two probability distributions.

A three-layer classification neural network was attached to the bottleneck layer of the VAE deep embedding network for the tumour type classification downstream task. The latent vector $\boldsymbol{\mu}$ was fed to the classification network as the input and passed through two hidden layers with 128 neurons and 64 neurons, respectively, before the probability of each tumour type was obtained by the softmax activation function in the output layer. We defined the loss function of the classification network as

Table 1. Hyper-parameters used in the model

Hyper-parameter	Value
Latent dimension	128
Learning rate	1e-3
Batch size	32
Epoch number—unsupervised	50
Epoch number—supervised	100

the cross-entropy between the ground-truth tumour type y and predicted tumour type y′, as shown in Equation (4):

$$\mathcal{L}_{class} = CE(y, y'). \tag{4}$$

Thus, the overall loss function of the whole model was a weighted combination of the VAE loss $\mathcal{L}_{\text{VAE}}$ and the classification loss $\mathcal{L}_{class}$, which was defined in Equation (5):

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{\text{VAE}} + \beta\mathcal{L}_{class} \tag{5}$$

α and β weighted the two losses during training. The hyper-parameters used to train this model were listed in Table 1.

XOmiVAE has the ability to explain both the supervised tumour type classification results, which was illustrated in Figure 1A, and the unsupervised omics data clustering results, which was illustrated in Figure 1 (B). Based on the vanilla OmiVAE, we integrated the Deep SHAP explanation approach to XOmiVAE in a customized way. Deep SHAP inherited the key principle from DeepLIFT, which is the 'summation-to-delta' property. This property means that the sum of the attributions over the input equals the difference-from-reference of the output [35], which can be formalized by Equation (6):

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta o \tag{6}$$

where Δo is the difference between the output of the reference sample and the output of the target sample, which is $\Delta o = f(\mathbf{x}) - f(\mathbf{r})$, $\mathbf{x}$ is the target gene expression profile, $\mathbf{r}$ is the reference gene expression profile, $\Delta x_i = x_i - r_i$, i is the gene index and n is the number of genes used in the experiment [22]. Another representation of the 'summation-to-delta' property was demonstrated in Figure 1C. This property enables the calculation of the Shapley values, which indicate how to allocate contribution of each prediction result among the input features. Larger Shapley values, therefore, represent more important genes for the prediction of certain tumour type.

As for the implementation, a trained network was first passed to the Deep SHAP explainer object of XOmiVAE alongside the reference values to calculate the SHAP values from. The computation graph of the model was then able to effectively guide the explainer through the network to calculate the activation of neurons according to principles used by DeepLIFT. The original Deep SHAP was also modified to ensure it could take either the latent vector or the classification output vector as the output values for the contribution analysis. As recommended by Shrikumar *et al.* [35], we used the pre-activation output instead of the post-softmax probabilities to calculate feature contribution scores. For each prediction, n SHAP values corresponding to n genes or n latent dimensions were calculated 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, as shown in Figure 1C. This avoids the issue of positive and negative SHAP values offsetting each other when they were averaged across samples. To reveal the contribution of each omics latent dimension in unsupervised tasks, we calculated the mean and standard deviation of the latent vector values ($\boldsymbol{\mu}$ values) for the two groups of samples and applied a Welch's t-test to obtain the most statistically significant dimension that separates the two groups. The correlation between the each latent variable and each gene was obtained by backpropagating the latent vector through the Deep SHAP explainer object of XOmiVAE.

Bioinformatics analysis

To evaluate the contribution results obtained by XOmiVAE, we compared genes with high contribution scores with the differentially expressed genes (DEGs) between normal and cancer samples for each tumour type. We used an R Bioconductor package TCGAbiolinks [6] to conduct the differential gene expression analysis. The DEGs were selected according to the cut-off of 0.05 for the false discovery rate (FDR) adjusted P-value and the threshold of 3 for the absolute log_2 fold change. To reveal the biological implication of the top genes with high contribution scores, we used the Broad Institute's Gene Set Enrichment Analysis (GSEA) software to perform pathway enrichment analysis [38]. Additionally, we used the curated gene sets from online databases including the Gene Ontology (GO) [7], Kyoto Encyclopedia of Genes and Genomes [16] and the Reactome pathway database (Reactome) [9] to test subtypes pathways. g:Profiler [31] was used to obtain and visualise the top pathways. GeneCard [37] was used to obtain the specific gene set for each TCGA tumour type.

Results and discussion

Multi-level explanation of XOmiVAE

Most important genes for cancer classification

We trained XOmiVAE on the TCGA pan-cancer dataset and calculated the contribution score of each gene for the prediction of each tumour type. The model achieved high accuracy for differentiating between normal and tumour tissue. For instance, the classification accuracy of breast invasive carcinoma (BRCA) and normal breast tissue was 99.6% and 100%, respectively. The contribution scores followed a power-law distribution, which suggested the majority of input features (i.e. genes) were unimportant for cancer prediction (see Supplementary Figures S1 and S2). As an example, we illustrated the top 10 genes with the highest scores that contributed most to the BRCA prediction in Figure 2. This demonstrated the explainability of XOmiVAE by revealing the contribution of each input feature (i.e. gene). To validate whether the top genes found by XOmiVAE made biological sense, we analysed the biological function of the most and least important genes. The top genes are known to be related to BRCA. For example, the top 1 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 [18]. The 2nd most important gene, *AZGP1*, is associated with an aggressive breast cancer phenotype [29]. On the contrary, 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 top genes with their contribution scores

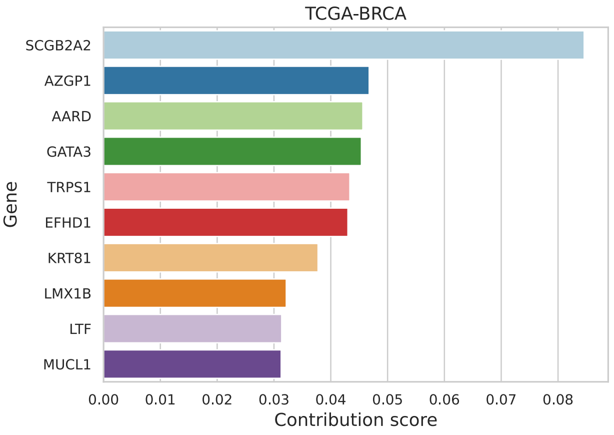

Figure 2. The top 10 genes for the prediction of breast invasive carcinoma (BRCA). Random samples were used as the reference.

Table 2. 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
3rd	**35**	**35**	**35**
4th	111	53	125
5th	42	103	45

The bold values in the table here are essential to indicate that all of the three subtypes shared the same 3rd important dimension.

for the other 32 tumour types was also obtained by XOmiVAE and shown in Supplementary Figures S3 to S6.

Most important dimensions for cancer classification

By passing an interim layer to the Deep SHAP explainer object of XOmiVAE, it is possible 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 are reduced into a latent representation with lower dimensionality, 128 dimensions in our scenario. Therefore, the input of the 1st 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 3rd one was shared (Table 2), which is therefore possible the dimension responsible for separating the kidney located tumours. Additionally, it is practicable to find the most associated genes and, therefore, the most related biological pathways to a specific dimension. We investigate the top 15 genes for the shared kidney dimension 35 as an example (see Supplementary Figure S12). These results can be obtained for every dimension and every tumour type.

Validation by biomedical knowledge

Biomedical meaning of the top genes

To validate the top genes revealed by XOmiVAE, we first compared the genes, as ranked by contribution, for each tumour type, with genes associated with the corresponding tumour type from GeneCard [37]. GeneCard was chosen due to its comprehensive disease gene sets, which are integrated from around 150 different web sources 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 (the most important gene) to 58 043 (the total number of genes). XOmiVAE were compared to different thresholds of a random sample of genes (averaged over 10 random seeds) and state-of-the-art methods, including Saliency [36], Input X Gradient (an extension of Saliency) and GradientSHAP [22]. The results were plotted as a ROC curve for the true positive rates (TPRs) and false positive rates (FPRs). The TPRs were calculated by

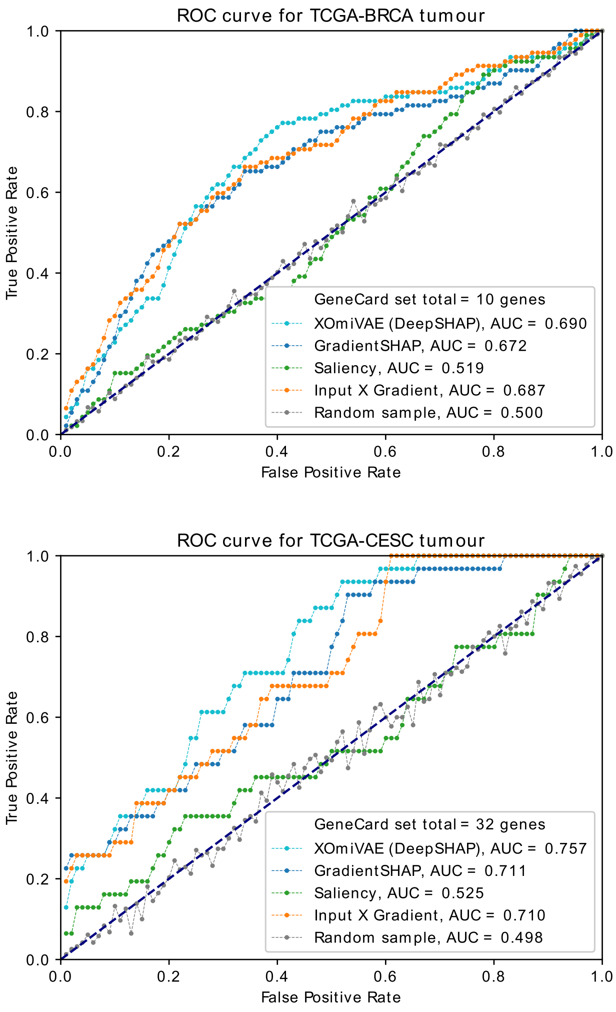

Figure 3. AUC-ROC curves of genes as ranked by the XOmiVAE importance scores, for breast invasive carcinoma (BRCA) and cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) tumour prediction, against the GeneSet gene list for the respective tumour type. State-of-the-art methods (i.e. Saliency, Input X Gradient and GradientSHAP) and a random selection of genes are used for comparison.

$$\frac{\textit{\# of top genes which are GeneCard disease genes}}{\textit{\# of GeneCard disease Genes}}. \tag{7}$$

And the FPR were calculated by

$$\frac{\textit{\# of top genes which are not GeneCard disease genes}}{\textit{\# of genes not associated with the GeneCards disease}}. \tag{8}$$

Table 3. The average AUC score across all 33 tumour types for the ranked gene importance scores compared to the GeneCard genes

Methods	Average AUC	Standard deviation
Saliency	0.5331	0.0839
GradientSHAP	0.7682	0.0578
Input X Gradient	0.7762	0.0767
XOmiVAE	**0.7950**	0.0673

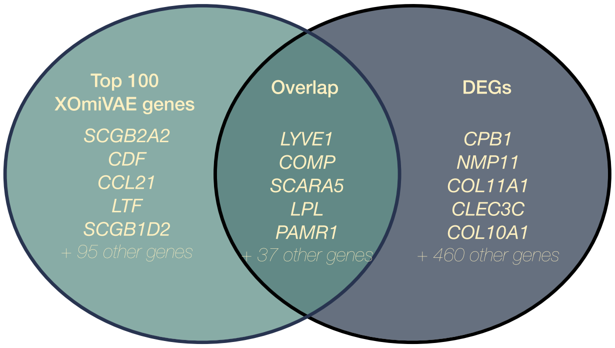

Figure 4. A Venn diagram representing the overlap between the DEGs and top contribution genes, highlighting a total of 42 DEGs found in the top 100 contribution genes.

A total of 21 tumour types had gene sets found in GeneCard and were therefore chosen for analysis. The ROC curves and AUC metrics are shown in 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. The average AUC metric among all 33 tumour types of XOmiVAE and three state-of-the-art methods was listed in Table 3. XOmiVAE outperformed all of the three state-of-the-art methods.

To further explore and understand the top genes revealed by XOmiVAE, they were evaluated using GSEA. We used g:Profiler [31], a web server for functional enrichment analysis, to identify the most significant GO terms enriched in the top genes for each tumour type. Supplementary Table S1 lists the GO terms that are significantly overrepresented in top BRCA genes. The most significant GO terms are closely related to the extracellular matrix organization, which is an area of intense interest in breast cancer research. [42] A break down of the pathways found from the other sources used in g:Profiler was shown in Supplementary Figure S10.

The top 100 most important genes for BRCA over normal breast tissue were compared with the 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 the kidney cancers KIRC, KIRP and KICH (Supplementary Figure S12). *APQ2* is the most important gene for that dimension for all three cancer subtypes, which is located in the apical cell membranes of the collecting duct principal cells in kidneys. Additionally, all of the other high ranking genes such as *UMOD*, *SCNN1G* and *SCNN1B* are all well-known genes associated with kidney functions [5, 14]. As another example, we also explain dimensions 42 and 73, the 1st and 2nd most important dimension for lung adenocarcinoma (LUAD) prediction, respectively, as shown in Table 4. The top genes were calculated using random training samples as the reference value, to show the most important genes for LUAD versus all the other sample types. We demonstrated that dimension 42 relies heavily on the immune response pathways, while 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 [34, 44].

We found that the most important input features for the latent dimensions varied according to the tumour type used for the analysis (Table 4). 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 [43] and Bica *et al.* [4], which show no specificity for different input samples and different prediction targets. Table 4 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 neurons 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 5). *XIST* is one of the key genes involved in the transcriptional silencing of one of the X chromosomes [47].

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 6, which indicates the contribution of the top genes obtained by XOmiVAE to the cancer classification tasks. Additionally, we calculated the average metrics for all the other tumour types except the target

Table 4. The biological pathways enriched for dimensions 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}

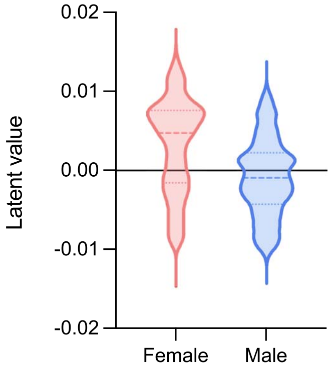

Figure 5. Violin plot of the latent dimension 78 for female and male samples.

Table 5. The top five genes for dimension 78 when separating female and male samples in the classification model of OmiVAE

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

one and found that while 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 top genes revealed 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 1st and the least important gene ranking 58,043rd, and selected 20 genes with the lowest sum rankings to retrain the model and calculate the overall accuracy (Table 7). We then compared it with the performance of a model trained by 20 random genes and a model trained by the overall 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 20 least important 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 scores for feature selection in model training with high-dimensional 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 8). We set the output of the target dimension to −1 when the output value was positive, and set the output of the target dimension to 1 when the output value was negative, based off a similar ablation approach by Morcos *et al.* [26]. 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 [35]. 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

Deep SHAP, similar to other activation level-based approaches, used reference samples as background to appraise the feature importance of each gene or latent dimension. The selection of

Table 6. The evaluation metrics of cancer classification using only the top 20 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 7. 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% ± 0.46%
Top 20 genes for all tumours	20	87.07% ± 0.38%
20 random genes	20	56.10% ± 0.24%
Bottom 20 genes for all tumours	20	1.68% ± 0.37%

Table 8. 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 change
1st	−11.9% ± 19.9%
2nd	−11.9% ± 20.2%
3rd	−2.9% ± 6.5%
Top three combined	−95.9% ± 2.0%
126th	0.0% ± 0.3%
127th	0.0% ± 0.0%
128th	0.0% ± 0.3%
Bottom three combined	0.0% ± 0.3%

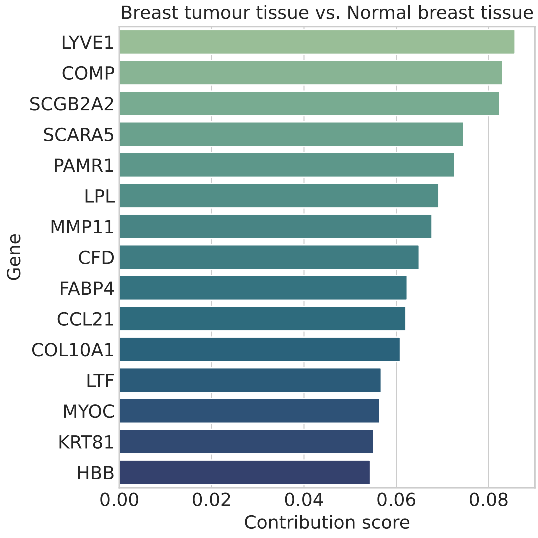

Figure 6. The top 15 genes obtained by XOmiVAE for the classification of BRCA using normal breast tissue samples as the reference.

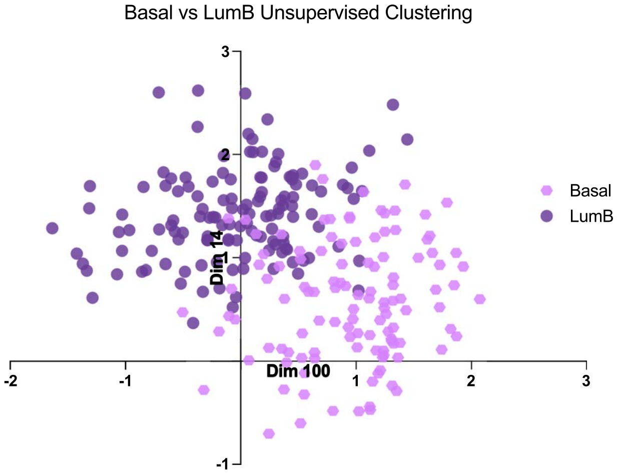

Figure 7. Top two dimensions for splitting Basal and LumB subtypes in the latent space.

reference samples is crucial for the explanation, since importance scores are calculated by comparing the activation level of neurons when a reference sample is fed to the network or when a target sample is fed to the network. One of the recommended choices for this reference 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 reference.

To further understand the effect of the reference, 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). Twenty-five 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 Figures 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

Table 9. 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 that 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 10. 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	1.15e-47
Genes up-regulated in bone relapse of breast cancer	39	4.25e-43
Genes down-regulated in ductal carcinoma versus normal ductal breast cells	24	2.6e-42
Genes down-regulated in nasopharyngeal carcinoma (NPC) positive for LMP1, a latent gene of Epstein–Barr virus (EBV)	29	1.85e-37
Genes up-regulated in breast cancer samples positive for ESR1 compared to the ESR1 negative tumours	26	1.82e-30

shown to be involved in BRCA progression from normal tissue [42]. 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 reference 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. Explaining the latent dimensions of VAEs would be crucial when it is important to understand the genes involved in subtype clustering of cancers that are yet to be defined, and labels that could be used for supervised learning are scarce. 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 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 [38], which includes pathways from experiments comparing the subtypes.

In Table 9, 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 9 we can see the top pathways also include the genes that differentiate between the *ESR1* negative and *ESR1* positive tumours. Table 10 shows the results when the three other BRCA subtypes (LumA, Her2 and Basal) are used as the reference samples when explaining subtype LumB. The results show that a larger range of subtype pathways are present in the most important features. These results prove 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 1st attempt at using an activation level-based explanation approach for clustering generated by autoencoders. Typically, differential gene expression methods, such as DESeq2 [21], are used to explain differences in clusters, which treats each gene as independent. More recent methods improve on this, such as global counterfactual explanation (GCE) [30] and gene relevance score (GRS) [1]. 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 presented an explainable VAE-based deep learning method for high-dimensional omics data analysis, named XOmiVAE. We illustrated 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 showed that it is practicable to explain the most important genes in an unsupervised network and therefore provide a method for explaining deep learning-based clustering. We evaluated the explanations of XOmiVAE and demonstrated that they make biological sense. Additionally, we offered important steps to consider when interpreting deep learning models for tumour classification. For example, we highlighted the importance of choosing reference samples that makes biological sense when explaining the model, and we disclosed the limitations of connection weight-based methods to explain latent dimensions. We believe XOmiVAE is a promising methodology that could help open the 'black box' and 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 score of each input molecular feature and omics latent dimension for each prediction.
- XOmiVAE is able to explain unsupervised clusters produced by VAEs without the need for labelling.
- 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 https://github.com/zhangxiaoyu11/XOmiVAE/. The TCGA pan-cancer dataset can be downloaded from the UCSC Xena data portal https://xenabrowser.net/datapages/.

Supplementary Data

Supplementary data are available online at GitHub https://github.com/zhangxiaoyu11/XOmiVAE/blob/main/documents/supplementary.pdf.

Funding

European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement (764281).

References

1. Angerer P, Fischer DS, Theis A, *et al.* Automatic identification of relevant genes from low-dimensional embeddings of single-cell RNA-seq data. *Bioinformatics* 2020; **36**(15): 4291–5.
2. Azarkhalili B, Saberi A, Chitsaz H, *et al.* DeePathology: deep multi-task learning for inferring molecular pathology from cancer transcriptome. *Sci Rep* 2019; **9**(1): 16526.
3. Azodi CB, Tang J, Shiu SH. Opening the black box: interpretable machine learning for geneticists. *Trends Genet* 2020; **36**(6): 442–55.
4. Bica I, Andrés-Terré H, Cvejic A, *et al.* Unsupervised generative and graph representation learning for modelling cell differentiation. *Sci Rep* 2020; **10**(1): 9790.
5. Carney EF. Evolving risks of umod variants. *Nat Rev Nephrol* 2016; **12**(5): 257–7.
6. Colaprico A, Silva TC, Olsen L, *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2015; **44**(8): e71–1.
7. Gene Ontology Consortium. The Gene Ontology (go) database and informatics resource. *Nucleic Acids Res* 2004; **32**(90001): D258–61.
8. Dincer AB, Celik S, Hiranuma N, *et al.* DeepProfile: deep learning of cancer molecular profiles for precision medicine. bioRxiv2018.
9. Fabregat A, Korninger F, Viteri G, *et al.* Reactome graph database: efficient access to complex pathway data. *PLoS Comput Biol* 2018; **14**:1–13.
10. Garson DG. Interpreting neural-network connection weights. *AI Expert* 1991; **6**(4): 46–51.
11. Goldman MJ, Craft B, Hastie M, *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020; **38**(6): 675–8.
12. Grossman RL, Heath AP, Ferretti V, *et al.* Toward a shared vision for cancer genomic data. *N Engl J Med* 2016; **375**(12): 1109–12.
13. Hanczar B, Zehraoui F, Issa T, *et al.* Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics* 2020; **21**(1): 501.
14. Hanukoglu I, Hanukoglu A. Epithelial sodium channel (ENaC) family: phylogeny, structure-function, tissue distribution, and associated inherited diseases. *Gene* 2016; **579**(2): 95–132.
15. Hira MT, Razzaque MA, Angione C, *et al.* Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep* 2021; **11**(1): 6265.
16. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**(1): 27–30.
17. Kingma DP, Welling M. Auto-encoding variational Bayes. In: *International Conference on Learning Representations (ICLR)*, 2014; Banff, AB, Canada, ICLR.
18. Lacroix M. Significance, detection and markers of disseminated breast cancer cells. *Endocr Relat Cancer* 2006; **13**(4): 1033–67.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**(7553): 436–44.
20. Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics* 2020; **21**(1): 146.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**(12): 550.
22. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, MIT Press. 2017, pp. 4768–77.
23. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv, abs/1802.03426, 2020.
24. Meng C, Zeleznik OA, Thallinger GG, *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016; **17**(4): 628–41.
25. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process* 2018; **73**:1–15.
26. Morcos AS, Barrett D, Rabinowitz NC, *et al.* On the importance of single directions for generalization. In: *International Conference on Learning Representations (ICLR)*; Vancouver, BC, Canada, ICLR. 2018.
27. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013; **45**(10): 1113–20.
28. Olden JD, Jackson DA. Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. *Ecol Model* 2002; **154**(1): 135–50.
29. Parris TZ, Kovács A, Aziz L, *et al.* Additive effect of the AZGP1, PIP, S100A8 and UBE2 molecular biomarkers improves outcome prediction in breast carcinoma. *Int J Cancer* 2014; **134**(7): 1617–29.

30. Plumb G, Terhorst J, Sankararaman S, *et al*. Explaining groups of points in low-dimensional representations. In: *Proceedings of the 37th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, Vienna, Austria, ACM. Vol. 119, 2020, pp. 7762–71
31. Raudvere U, Kolberg L, Kuzmin I, *et al*. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019; **47**(W1): W191–8.
32. Ringnér M. What is principal component analysis? *Nat Biotechnol* 2008; **26**(3): 303–4.
33. Sanchez-Vega F, Mina M, Armenia J, *et al*. Oncogenic signaling pathways in the cancer genome atlas. *Cell* 2018; **173**(2): 321–37.e10.
34. Schneider MA, Granzow M, Warth A, *et al*. Glycodelin: a new biomarker with immunomodulatory functions in non-small cell lung cancer. *Clin Cancer Res* 2015; **21**(15): 3529–40.
35. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, Sydney, Australia, ACM. Vol. 70, 2017, pp. 3145–53.
36. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: *Workshop at International Conference on Learning Representations (ICLR)*, 2014; Banff, AB, Canada, ICLR.
37. Stelzer G, Rosen N, Plaschkes I, *et al*. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinform* 2016; **54**(1): 1.30.1–1.30.33.
38. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005; **102**(43): 15545–50.
39. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *International Conference on Machine Learning (ICML)*, 2017; Sydney, Australia, ACM.
40. Tasaki S, Gaiteri C, Mostafavi S, *et al*. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* 2020; **2**(7): 376–86.
41. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**(86): 2579–605.
42. Walker C, Mojares E, Hernández ADR. Role of extracellular matrix in development and cancer progression. *Int J Mol Sci* 2018; **19**(10).
43. Way GP, Casey S. Greene Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In: *Biocomputing 2018*, Kohala Coast, HI, USA, World Scientific. 2018, pp. 80–91.
44. Yamamoto O, Takahashi H, Hirasawa M, *et al*. Surfactant protein gene expressions for detection of lung carcinoma cells in peripheral blood. *Respir Med* 2005; **99**(9): 1164–74.
45. Zhang X, Xing Y, Sun K, *et al*. OmiEmbed: a unified multitask deep learning framework for multi-omics data. *Cancers* 2021; **13**(12): 3047.
46. Zhang X, Zhang J, Sun K, *et al*. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, IEEE. 2019, pp. 765–9.
47. Zuccotti M, Monk M. Methylation of the mouse Xist gene in sperm and eggs correlates with imprinted xist expression and paternal x-inactivation. *Nat Genet* 1995; **9**(3): 316–20.