# Tensor-decomposition-based unsupervised feature extraction applied to prostate cancer multiomics data

**Y-h. Taguchi[1] and T. Turki[2]**

[1]**Department of Physics, Chuo University, Tokyo 112-8551, Japan, tag@granular.com**
[2]**Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia, tturki@kau.edu.sa**

Corresponding author:
Y-h. Taguchi[1]

Email address: tag@granular.com

## ABSTRACT

The large *p* small *n* problem is a challenge without a de facto standard method available to it. In this study, we propose a tensor-decomposition (TD)-based unsupervised feature extraction (FE) formalism applied to multiomics datasets, in which the number of features is more than 100,000 whereas the number of samples is as small as about 100, hence constituting a typical large *p* small *n* problem. The proposed TD-based unsupervised FE outperformed other conventional supervised feature selection methods, such as random forest, categorical regression (also known as analysis of variance, or ANOVA), penalized linear discriminant analysis, and a unsupervised method, multiple non-negative matrix factorization, when applied not only to multiomics datasets but also to synthetic datasets. The genes selected by TD-based unsupervised FE were enriched in genes known to be related to tissues and transcription factors measured. TD-based unsupervised FE was demonstrated to be not only the superior feature selection method but also the method that can select biologically reliable genes. To our knowledge, this is the first study in which TD-based unsupervised FE has been successfully applied to the integration of this variety of multiomics measurements.

## 1 INTRODUCTION

The term "big data" often indicates a high number of instances as well as features (Richter and Khoshgoftaar, 2019; Awan et al., 2018). Typical big data comprise a few million images (photos) each composed of several million pixels. The number of features is as high as the number of instances, or often higher. Most recently developed machine-learning methods including deep learning (DL) aim to address this manner of problem (Nashaat et al., 2018; Shah et al., 2019; Bekkerman et al., 2011; Chatterjee et al., 2019; Richter and Khoshgoftaar, 2019; Ngiam and Khor, 2019; Santosh et al., 2020). Nonetheless, although less popular than these typical big data problems, there is another branch of big data problem known as the "large *p* small *n*" problem, or "high-dimensional data analysis". For these cases, the number of instances is typically smaller than the number of features (Ge et al., 2019; Wen et al., 2019; Yang et al., 2019). In this case, there can be many unique problems that do not exist for the typical big data problems mentioned above. For example, in discrimination tasks, because there are more features than instances, if we do not reduce the number of features, the accuracy will always be 100% for no other reason than overfitting. Alternatively, the so-called "curse of dimensionality" problem describes a shortage of instances covering sufficiently large spaces. Because of these difficulties, compared with the popular big data problems, the "large *p* small *n*" problem has rarely been investigated comprehensively.

This might not seem significant, as the "large *p* small *n*" problem is relatively rare and not as important. Nevertheless, dependent on the domains, the "large *p* small *n*" problem will inevitably become typical. For example, in genomic science, the typical number of features is as many as or more than that of genes (i.e., about $10^4$), whereas the number of samples is the number of patients, typically at most a few hundred. The length of the DNA sequence of the human genome is approximately $3 \times 10^9$, whereas the

human population is about $10^{10}$. Nevertheless, only a very small fraction of this can be considered in an individual study. Thus, in typical genomic studies, the ratio of the number of features to the number of samples can be in the region of $10^2$ or more.

In addition, there is yet another difficulty in genomic science: labels attributed to samples are often not numerical but categorical and may not even be ordered. This prevents us from applying most popular methods (e.g., lasso; (Friedman et al., 2010)) to the genomic datasets. Thus, applicable methods are very limited.

Nonetheless, Taguchi has proposed a very different method to the typical machine-learning methods that are applicable to large $p$ small $n$ problems: tensor-decomposition (TD)-based unsupervised feature extraction (FE) (Taguchi, 2020). In this method, a smaller number of representative features, referred to as singular value vectors, are generated with linear combinations of the original large number of features, without considering labeling. These singular value vectors are attributed to both samples and original features. By investigating the singular value vectors attributed to samples, we can identify which of them are useful (e.g., distinguishing between samples associated with distinct labeling). We can then select singular value vectors associated with those attributed to samples among those attributed to the original features. Finally, the selected singular value vectors attributed to the original features are used to evaluate the importance of the original features; because the singular value vectors are the linear combinations of original features, their coefficients of linear combinations can be used to evaluate the importance of individual original features. The original features with larger absolute values of coefficients are selected. As a result, we can avoid many difficulties in the "large $p$ small $n$" problem; for example, we can perform standard discriminant analysis when there is a higher number of features than instances.

In this paper, we propose a TD-based unsupervised FE formalism, applied to synthetic data that imitate typical genomic datasets. The proposed method is first compared with four conventional methods applicable to the "large $p$ small $n$" problem: random forest (RF), penalized linear discriminant analysis (LDA) (Witten and Tibshirani, 2011), categorical regression, and multiple non-negative matrix factorization (MNMF); the latter has been used especially often for analysis of multiomics datasets (Baldwin et al., 2020). These five methods are then also applied to a real dataset consisting of multiomics measurements of prostate cancers, a typical "large $p$ small $n$" dataset. The results for the synthetic and real datasets demonstrate the superiority of our proposed method when compared against the four conventional methods.

To our knowledge, this is the first study in which TD-based unsupervised FE has been successfully applied to the integration of this variety of multiomics measurements. Because the performance of TD-based unsupervised FE is highly data-type-dependent, it is challenging to estimate how many types of multiomics datasets (here, three transcription factor bindings as well as four histone modifications and one chromosomal state measurement, making eight types in total) can be successfully integrated with this method.

# 2 MATERIALS AND METHODS

## 2.1 Synthetic data

In the synthetic dataset, we assumed that there are 81 samples composed of three tissue subclasses (each class has three replicates) and three disease subclasses (each class also has three replicates). That is, the 81 samples are composed of all possible pairs of one of the nine tissue samples (= three tissues $\times$ three replicates) and one of the nine disease samples (= three tissues $\times$ three replicates). Typically, the three tissues are supposed to be, for example, heart, brain, and skin, while the three diseases are supposed to be prostate cancer, diabetes, and healthy controls; this means that we do not have any biological knowledge of which combinations are more likely to be associated with one another. The number of genes, which are features, is as many as $10^5$; although so-called protein-coding genes are at most a few tens of thousands, a higher number of non-coding genes have come to be considered. Unlike the synthetic data, in the real dataset to which the methods are applied in the later part of this paper, the number of features corresponds to the number of genomic regions with fixed DNA sequence length. Thus, it can be even higher than that of the synthetic data. Among these $10^5$ features, only the first 100 features are supposed to be associated with distinction between diseases and tissues. The purpose of the analysis is to identify these 100 features correctly based upon the available datasets in a fully data-driven approach. To emulate this situation, we

**Table 1.** Sumary of real dataset

| Omics | number of multiomics | number of tissues | number of biological replicates |
|---|---|---|---|
| AR | 2 | 2 | 3 |
| FOXA1 | 4 | 2 | 3 |
| HOXB13 | 4 | 2 | 3 |
| H3K27AC | 10 | 2 | 3 |
| H3K27me3 | 1 | 2 | 3 |
| H3K4me3 | 1 | 2 | 3 |
| K4me2 | 1 | 2 | 3 |
| ATAC | 1 | 2 | 3 |
| total | 24 | — | — |

introduced tensor $x_{ijk} \in \mathbb{R}^{10^4 \times 9 \times 9}$ as

$$x_{ijk} = \begin{cases} \left\lceil \frac{j}{3} \right\rceil \left( \left\lceil \frac{k}{3} \right\rceil + 3 \right) + \varepsilon_{ijk}, & i \leq 100 \\ \varepsilon_{ijk}, & i > 100 \end{cases} \tag{1}$$

where $\varepsilon_{ijk} \sim \mathcal{N}(1,0)$, $\mathcal{N}(\mu,\sigma)$ is a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$, and $\lceil x \rceil$ is a ceiling function that gives the smallest integer that is not less than $x$. The $x_{ijk}$ values are classified into nine subclasses dependent upon pairs $\left( \left\lceil \frac{j}{3} \right\rceil, \left\lceil \frac{k}{3} \right\rceil \right), 1 \leq \left\lceil \frac{j}{3} \right\rceil, \left\lceil \frac{k}{3} \right\rceil \leq 3$ because $1 \leq j,k \leq 9$.

## 2.2 Real dataset

The real dataset consists of multiomics measurements of prostate cancer retrieved from the Gene Expression Omnibus (GEO) (Edgar et al., 2002), using GEO ID GSE130408. The processed file named GSE130408_RAW.tar was downloaded and individual files with names starting with "GSM" were extracted from it. The individual files are composed of eight omics measurements: three transcription factor bindings, AR, FOXA1, and HOXB13; four histone modifications, H3K27AC, H3K27me3, H3K4me3, and K4me2; and ATAC-seq. Two additional subclasses, tumor and normal tissue, were attributed to each. Because the number of biological replicates varies from subclass to subclass, we randomly selected the following number of ChIP-seq subclasses: two AR, four FOXA1, four HOXB13, ten H3K27AC, one H3K27me3, one H3K4me3, one K4me2, and one ATAC-seq. In total, 24 multiomics measurement groups were constructed. Each multiomics measurement group is composed of six samples, comprising two tissue subclasses (tumor and normal prostate), each of which has three biological replicates (Table 1).

The values in each file were averaged over 25,000 DNA sequence intervals, giving a total number of 123,817, as $3 \times 10^9$, the total DNA length of the human genome, divided by 25,000 is equal to 120,000. The averaged values were treated as a dataset to be analyzed further.

## 2.3 Categorical regression

Categorical regression, also known as analysis of variance or ANOVA, is expressed as

$$x_{ijk} = \sum_{J=1}^{3} a_J \delta_{Jj} + \sum_{K=1}^{3} b_K \delta_{Kk}, \tag{2}$$

where $\delta_{Jj}$ and $\delta_{Kk}$ take 1 only when $j$ or $k$ belong to the $J$th tissue or $K$th disease subclass, respectively. $a_J$ and $b_K$ are regression coefficients (for synthetic data).

$$x_{ijkm} = \sum_{J=1}^{8} a_J \delta_{Jj} + \sum_{K=1}^{2} b_K \delta_{Kk}, \tag{3}$$

where $\delta_{Jj}$ and $\delta_{Kk}$ take 1 only when $j$ or $k$ belong to the $J$th multiomics measurement group or $K$th tissue ($K = 1$: tumor, $K = 2$: normal prostate) subclass, respectively. $a_J$ and $b_K$ are regression coefficients (for real data).

The computation was performed by the `ls` function in the base package in R (R Core Team, 2019). The computed $P$-values were adjusted by the BH criterion (Taguchi, 2020) and the $i$th features associated with adjusted $P$-values less than 0.01 were selected.

## 2.4 RF

RF was performed using the `randomForest` function implemented in the `randomForest` package (Liaw and Wiener, 2002) in R. Synthetic and real data were regarded as $3 \times 3 = 9$ and $8 \times 2 = 16$ subclasses, respectively, each of which corresponds to one of the $(J,K)$ pairs defined in Eqs. (2) or (3). Features included in the out-of-bag error were selected that had non-zero importance given by the `importance` function implemented in the `randomForest` package.

## 2.5 PenalizeLDA

PenalizedLDA was performed using the `PenalizedLDA.cv` function implemented in the `PenalizedLDA` package (Witten, 2015) in R. Synthetic and real data were regarded as $3 \times 3 = 9$ and $8 \times 2 = 16$ subclasses, respectively, each of which corresponds to one of the $(J,K)$ pairs defined in Eqs. (2) or (3). $\lambda$ was taken to be $0.01, 0.02, 0.03, 0.04$ for synthetic data. PenalizedLDA could not be performed for the real dataset because of the zero in-subclass standard deviations of some features (for more details, see (Witten and Tibshirani, 2011)).

## 2.6 TD-based unsupervised FE

Although the process was fully described in the recently published book (Taguchi, 2020), here we propose a variant, which is outlined briefly. First, a dataset must be formatted as a tensor. For synthetic data, $x_{ijk} \in \mathbb{R}^{10^5 \times 9 \times 9}$. For real data, $x_{ijkm} \in \mathbb{R}^{N \times 24 \times 2 \times 3}$ represents the averaged value of the $i$th interval of the $j$th omics measurement group of the $k$th tissue ($k = 1$: tumor, $k = 2$: normal prostate) of the $m$th biological replicates ($1 \leq m \leq 3$). Here, $N$ = 123,817 is the total number of genomic regions of 25,000 DNA sequence length.

Higher-order singular value decomposition (HOSVD) (Taguchi, 2020) was then applied to the tensor, obtaining

$$x_{ijk} = \sum_{\ell_1=1}^{9} \sum_{\ell_2=1}^{9} \sum_{\ell_3=1}^{10^5} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 i} \tag{4}$$

for synthetic data. $G \in \mathbb{R}^{9 \times 9 \times 10^5}$ is a core tensor and $u_{\ell_1 j} \in \mathbb{R}^{9 \times 9}, u_{\ell_2 k} \in \mathbb{R}^{9 \times 9}, u_{\ell_3 i} \in \mathbb{R}^{10^5 \times 10^5}$ are singular value matrices that are orthogonal. Conversely, for the real dataset,

$$x_{ijkm} = \sum_{\ell_1=1}^{24} \sum_{\ell_2=1}^{2} \sum_{\ell_3=1}^{3} \sum_{\ell_4=1}^{N} G(\ell_1 \ell_2 \ell_3 \ell_4) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 i}, \tag{5}$$

where $G \in \mathbb{R}^{24 \times 2 \times 3 \times N}$ is a core tensor and $u_{\ell_1 j} \in \mathbb{R}^{24 \times 24}, u_{\ell_2 k} \in \mathbb{R}^{2 \times 2}, u_{\ell_3 m} \in \mathbb{R}^{3 \times 3}, u_{\ell_4 i} \in \mathbb{R}^{N \times N}$ are singular value matrices that are orthogonal.

The first step is to check which singular value vectors attributed to sample subclasses ($u_{\ell_1 j}$ attributed to tissue subclasses and $u_{\ell k}$ to disease subclasses for the synthetic dataset, and $u_{\ell_1 j}$ attributed to multiomics measurement groups and $u_{\ell_2 k}$ to tissues for the real dataset) represent distinction between samples of interest. After identifying biologically interesting singular value vectors attributed to samples ($u_{\ell_1 j}$ and $u_{\ell_2 k}$ for the synthetic data, and $u_{\ell_1 j}$, $u_{\ell_2 k}$, and $u_{\ell_3 m}$ for the real data), we next attempt to find singular value vectors attributed to features ($u_{\ell_3 i}$ for the synthetic data and $u_{\ell_4 i}$ for the real data) that share core tensor $G$ with larger absolute values with those identified singular value vectors attributed to samples. For example, suppose that $\ell'_1, \ell'_2$ are selected for synthetic data and $\ell'_1, \ell'_2, \ell'_3$ are selected for real data. In this case we seek $\ell_3$ that has $G(\ell'_1 \ell'_2 \ell_3)$ with larger absolute values for the synthetic data and $\ell_4$ that has $G(\ell'_1 \ell'_2 \ell'_3 \ell_4)$ with larger absolute values for the synthetic data. Next, using the selected $u_{\ell_3 i}$ and $u_{\ell_4 i}$, $P$-values are attributed to the $i$th feature as

$$P_i = P_{\chi^2} \left[ > \left( \frac{u_{\ell_3 i}}{\sigma_{\ell_3}} \right)^2 \right] \tag{6}$$

for synthetic data and

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}}\right)^2\right] \tag{7}$$

for real data. $P_{\chi^2}[> x]$ is the cumulative probability of $\chi^2$ distribution that the argument is larger than $x$. $\sigma_{\ell_3}$ and $\sigma_{\ell_4}$ are the standard deviations. The computed $P$-values are corrected by the BH criterion and features associated with adjusted $P$-values less than 0.01 are selected.

### 2.7 MNMF

Suppose that we have multiple matrices sharing the same number of rows, i.e., $X_k \in \mathbb{R}^{N \times M_k}$. When there are no non-positive components in these matrices, we can apply MNMF as

$$\min_{Q,H_k} \sum_k ||X_k - QH_k||_F, \tag{8}$$

where $Q \in \mathbb{R}^{N \times n}$ and $H_k \in \mathbb{R}^{n \times M_k}$, the column and row vectors of which are $n$ latent variable vectors attributed to $i$s and samples, respectively, and there are no non-positive components in $Q$ and $H_k$s. $||\cdot||_F$ is the Frobenius norm. $n$ is an integer smaller than any $N$ or $M_k$. MNMF can be easily performed by applying non-negative matrix factorization (NMF) to the contraction matrix of $X_k$s, $X \in \mathbb{R}^{N \times \sum_k M_k}$, which is an unfolded matrix of tensors in this study (see Fig. 1 and below), because

$$\sum_k ||X_k - QH_k||_F = ||X - QH||_F, \tag{9}$$

where $H \in \mathbb{R}^{n \times \sum_k M_k}$ is a contraction matrix of $H_k$s. NMF was performed by the `nmf` function in the NMF package of R.

When MNMF is applied to the synthetic dataset, it is applied to the matrix $x_{i(j)} \in \mathbb{R}^{10^5 \times 81}$ generated by unfolding tensor $x_{ijk} \in \mathbb{R}^{10^5 \times 9 \times 9}$. When MNMF is applied to the real dataset, it is applied to the matrix $x_{i(jkm)} \in \mathbb{R}^{123817 \times 144}$ generated by unfolding tensor $x_{ijkm} \in \mathbb{R}^{123817 \times 24 \times 2 \times 3}$. Because NMF cannot be applied to matrices with non-positive values, $\min(x_{ijk})$ are extracted from $x_{ijk}$ so that NMF can be applied to matrices without non-positive values when applied to synthetic data. When NMF is applied to the real dataset, $1 \times 10^{-10}$ is added to $x_{ijkm}$ when $x_{ijkm} = 0$ (no negative values are included in $x_{ijkm}$).

## 3 RESULTS

Figure 1 shows the flowchart of analyses performed in this study.

### 3.1 Synthetic data

We applied RF, PenalizedLDA, categorical regression, and TD-based unsupervised FE to the synthetic dataset. When RF was applied to the synthetic dataset, there were up to 7289 features with non-zero importance, although the 100 features with subclass dependence were correctly selected. Thus, RF has the ability to select all the correct features. Nevertheless, because it selects too many false positives, it is ineffective. In contrast, when PenalizedLDA was applied to the synthetic dataset, the 100 features with subclass dependence were selected without any false positives being selected. Thus, PenalizedLDA can outperform RF. The only disadvantage is that we have to find the optimal $\lambda$, which is the weight coefficient of the L1 norm. Through trial and error, we found that when $\lambda$ is taken to be 0.02, 0.03, or 0.04, it achieves the ideal performance mentioned above. In contrast, if we select $\lambda$ as 0.01, it selects more than 200 features, although the 100 features with subclass dependence are correctly selected. If we do not have prior knowledge that the correct number is 100, it is impossible to tune the optimal parameter, because both $\lambda = 0.01$ and $\lambda = 0.04$ achieve perfect prediction accuracy of discrimination of the 81 samples into nine subclasses (on average, only one of the 81 samples is wrongly classified). Thus, in this sense, PenalizedLDA cannot be regarded as a perfect method, although it can achieve perfect performance if we can successfully tune the optimal parameter. Categorical regression is far superior to PenalizedLDA, as it achieves good performances regardless of the threshold $P$-value (Table 2). It exhibits a maximum of just 11 false positives even if the threshold $P$-values vary from 0.01 to 0.1. This is in significant contrast to

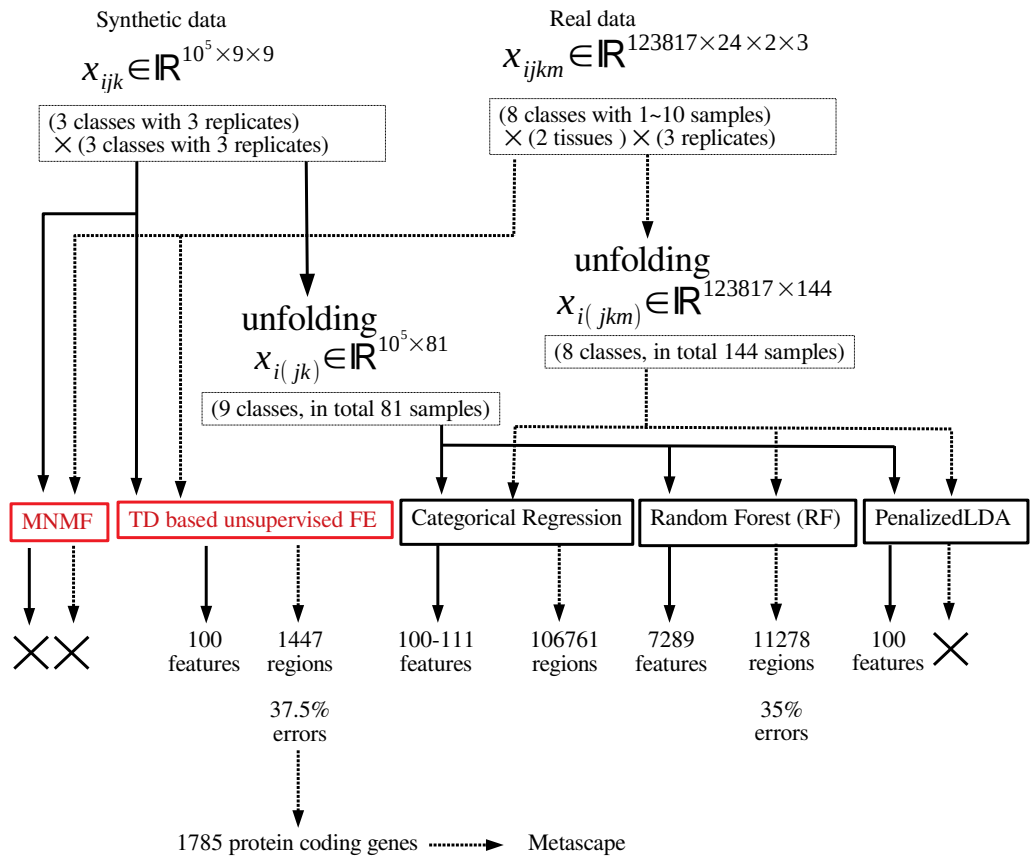


**Figure 1.** Flowchart of analyses performed in this study. Those in red and black are unsupervised and supervised feature selection, respectively.

**Table 2.** Confusion matrices when categorical regression is applied to synthetic data. $P$ is the threshold $P$-value. Features are selected when adjusted $P$-values less than the threshold P-values are associated with them.

| $P$ | 0.1 | | 0.05 | | 0.01 | | 0.001 | |
|---|---|---|---|---|---|---|---|---|
| | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ |
| not selected | 99889 | 0 | 99895 | 0 | 99899 | 0 | 99900 | 0 |
| selected | 11 | 100 | 5 | 100 | 1 | 100 | 0 | 100 |

PenalizedLDA, as the number of false positives varies from 0 to more than 100 when $\lambda$ slightly decreases from 0.02 to 0.01.

Nevertheless, TD-based unsupervised FE is demonstrated to outperform even categorical regression. Initially, after applying HOSVD to the synthetic dataset, we noticed that $u_{1j}$ and $u_{2k}$ were associated with distinction between the three subclasses (Fig. 2). Thus, we attempted to find which $\ell_3$ is associated with $G(1,1,\ell_3)$ of the largest absolute value (Fig. 3). From this, it was obvious that $G(1,1,1)$ had the largest absolute values. Equation (6) with $\ell_3 = 1$ was used to attribute $P_i$ to the $i$th feature. $P_i$ was corrected by the BH criterion and features associated with adjusted $P$-values less than the threshold values were selected (Table 3). As a result, TD-based unsupervised FE achieved perfect performances regardless of the threshold $P$-values.

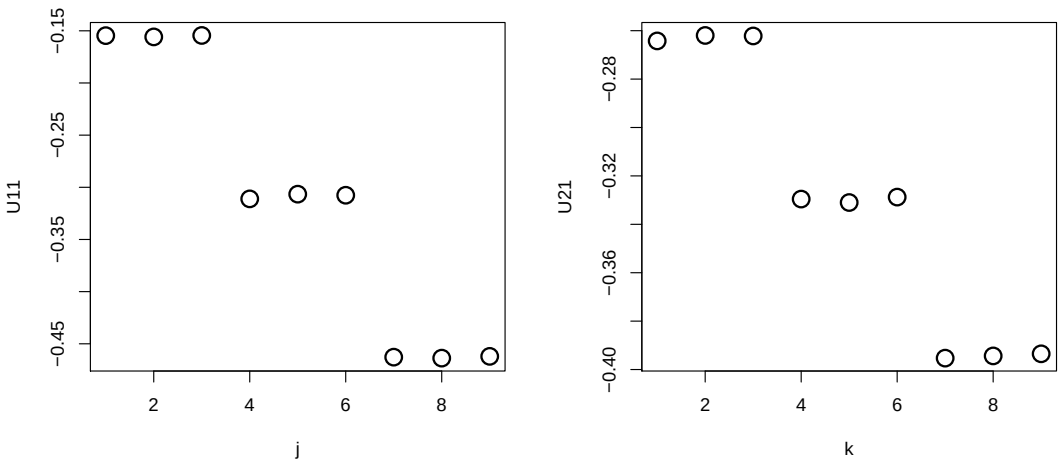


**Figure 2.** $u_{1j}$ and $u_{2k}$ obtained when HOSVD is applied to the synthetic dataset.

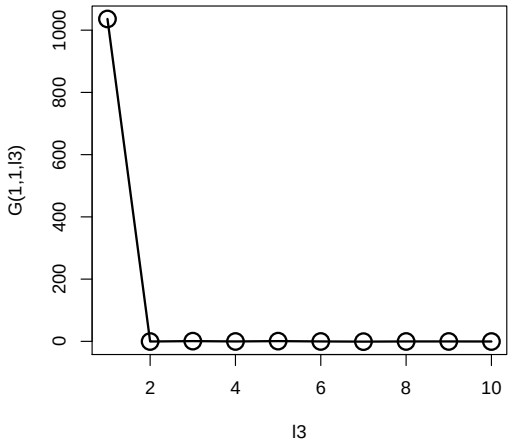


**Figure 3.** $G(1,1,\ell_3)$ obtained when HOSVD is applied to the synthetic dataset.

**Table 3.** Confusion matrices when TD-based unsupervised FE is applied to the synthetic data. $P$ is the threshold $P$-value. Features are selected when adjusted $P$-values less than the threshold P-values are associated with them. $\left\lceil \frac{j}{3} \right\rceil \left( \left\lceil \frac{k}{3} \right\rceil + 3 \right)$.

| $P$ | 0.1 | | 0.05 | | 0.01 | | 0.001 | |
|---|---|---|---|---|---|---|---|---|
| | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ | $i > 100$ | $i \leq 100$ |
| not selected | 99900 | 0 | 99900 | 0 | 99900 | 0 | 99900 | 0 |
| selected | 0 | 100 | 0 | 100 | 0 | 100 | 0 | 100 |

### 3.2 Real data

Next, we apply TD-based unsupervised FE to a real dataset, as it was shown to be the best of the four methods tested on synthetic data in the previous subsection. After applying HOSVD to $x_{ijkm}$, we realized that $u_{1j}$ and $u_{2j}$ have significant biological dependence upon multiomics measurement groups (Fig. 4). To explain why $u_{1j}$ and $u_{2j}$ are biologically reliable, we need to introduce genomic science briefly as follows:

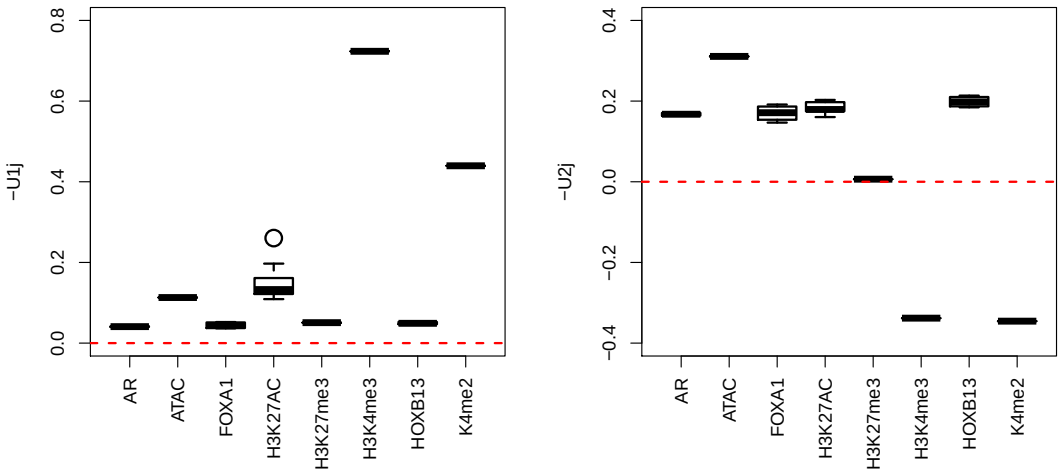


**Figure 4.** Boxplot of $u_{1j}$ and $u_{2j}$ when HOSVD is applied to the real dataset.

A genome, which is a sequence of DNA, stores the information to be translated into a protein. Because all cells store the same genome, the distinct functions and structures of individual cells are controlled by whichever parts of the genome are translated into proteins through transcription to RNA. This process is controlled by the various proteins that bind to genomes. In these measurements, the amounts of three such proteins, AR, FOXA1, and HOXB13, are measured. Another factor that controls this process is the small molecules that bind to histone. Histone is a core protein around which DNA wraps itself, and the tightness of this wrapping is the factor that controls the process from DNA to RNA. H3K27AC, H3K27me3, H3K4me3, and K4me2 are the measures of which small molecules bind to which parts of histone. Finally, ATAC-seq measures how "open" the DNA is. As DNA becomes more open, it can produce more RNA. Thus, the number of proteins that bind to DNA, the number of small molecules that bind to histone, and ATAC-seq are the measures of how many genomes are transcribed to RNA. The difficulty in multiomics analysis is to identify the combinations of proteins binding to DNA and small molecules binding to histone that can enhance the openness of DNA, which is assumed to be observed by ATAC-seq. As we do not know what combination is likely to be observed, the supervised approach is not easy to employ.

We now explain why $u_{1j}$ and $u_{2j}$ are biologically reliable. H3K4me3 and H3K27ac are known to be active/open chromatin marks, whereas H3K27me3 is known to be an inactive mark (Igolkina et al., 2019). $u_{1j}$ has higher values for H3K4me3 and H3K27ac but smaller values for H3K27me3, with larger values for ATAC-seq, which measures the amount of openness of genome. Thus, the dependences of $u_{1j}$ upon H3K4me3, H3K27ac, H3K27me3, and ATAC-seq are reliable. H3K4me2 (Pekowska et al., 2010) is also known to activate gene transcription in tissue-specific ways. The overall dependence of $u_{1j}$ upon multiomics measurement groups is reasonable. Nonetheless, three proteins, AR (Fujita and Nonomura, 2019), FOXA1 (Gerhardt et al., 2012), and HOXB13 (Navarro and Goldstein, 2018), are also known to mediate prostate cancer progression, and their enrichment corresponded with $u_{2j}$. We decided that $u_{1j}$ and $u_{2j}$ successfully captured the biological aspects of multiomics measurements.

In contrast, $u_{2k}$ represents the distinction between a tumor ($k = 1$) and normal prostate ($k = 2$), while $u_{1m}$ represents the independence of biological replicates (Fig. 5). As we are seeking multiomics measurements that are distinct between a tumor and normal prostate as well as common between biological replicates, $u_{2j}$ and $u_{1m}$ represent the properties that we seek. Although we can select either $\ell_1 = 1$ or $\ell_1 = 2$, as both are biologically reliable, we select $\ell_1 = 1$ because it is the first (i.e., mathematically primary) factor. To find the $u_{\ell_4 i}$ used for attributing $P$-values to genes $i$, we seek $G(1,2,1,\ell_4)$ with the largest absolute value (Fig. 6).

It is obvious that $G(1,2,1,8)$ has the largest absolute value. Hence, $u_{8i}$ is employed to assign $P$-values to the $i$th genomic regions using (7) with $\ell_4 = 8$. The $P_i$ values are corrected using the BH criterion,

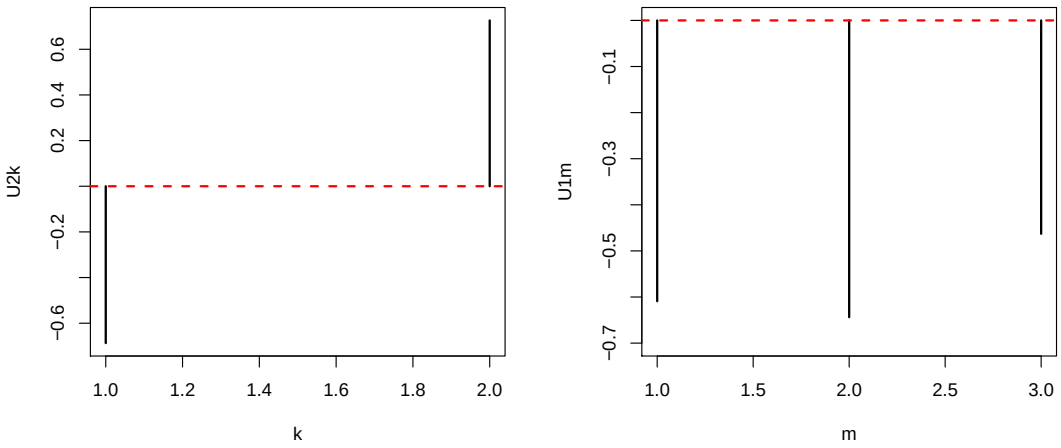


**Figure 5.** $u_{2k}$ and $u_{1m}$ when HOSVD is applied to the real dataset.

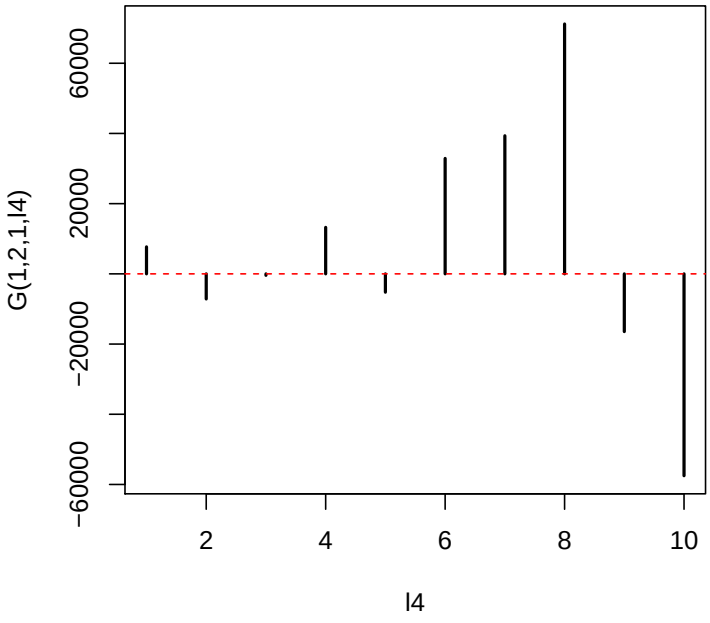


**Figure 6.** $G(1,2,1,\ell_4)$ when HOSVD is applied to the real dataset.

and 1447 genomic regions associated with adjusted $P_i$ less than 0.01 are selected from the total 123,817 genomic regions. This indicates that TD-based unsupervised FE selects as little as approximately 1% of genomics regions of the whole human genome.

Various biological evaluations of selected regions are possible, one of which is the number of protein-coding regions included in these 1447 selected regions. As the genomic regions that include protein-coding genes comprise less than a few percent of the human genome, and the total number of human protein-coding genes is up to $2 \times 10^4$, the expected number of protein-coding genes included in these 1447

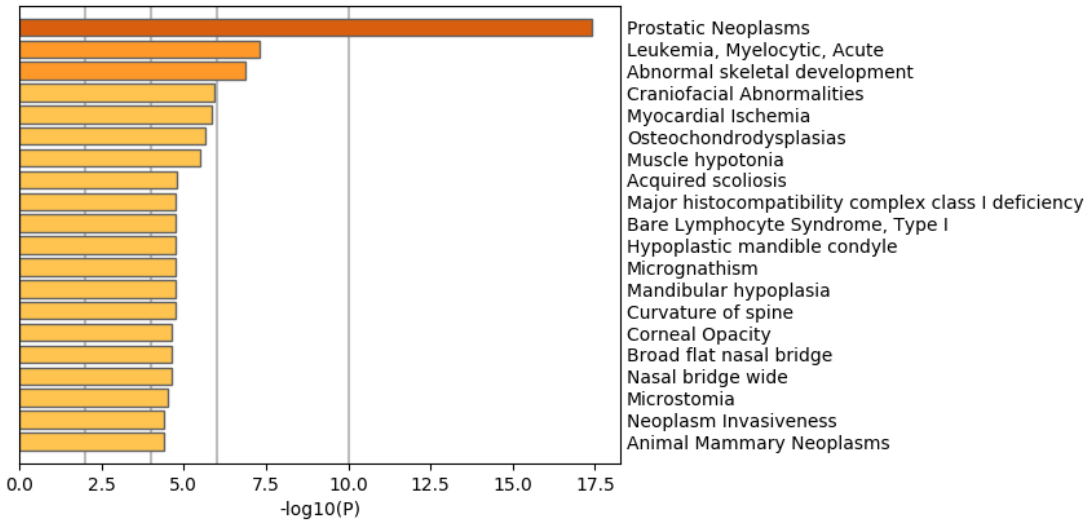


**Figure 7.** Summary of enrichment of the DisGeNet category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

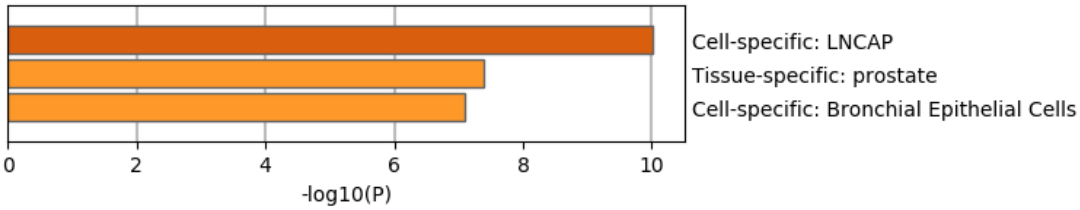


**Figure 8.** Summary of enrichment of the PaGenBase category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

selected genomic regions is at most a few hundred. Nevertheless, as many as 1785 protein-coding genes can be counted in these regions, which is much higher than expected. This indicates that TD-based unsupervised FE can select genomic regions that include protein-coding genes, correctly considering the altered multiomics variables between normal and tumor tissues.

Another evaluation of these selected 1447 genomic regions is the biological properties of the 1785 protein-coding genes included in these 1447 genomic regions. If biologically reliable protein-coding genes are selected, they should be enriched with the genes related to prostate cancer. To confirm this point, we uploaded these genes to Metascape (Zhou et al., 2019), which evaluates the enrichment of various biological terms with reducing redundancy. Figure 7 shows the results of the DisGeNet (Piñero et al., 2019) category of Metascape. It is obvious that prostatic neoplasms are not only top ranked but also outstandingly enriched. This suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Figure 8 shows the results of the PaGenBase (Pan et al., 2013) category of Metascape. It is obvious that LNCaP cells are not only top ranked but also outstandingly enriched. LNCAP (Horoszewicz et al., 1983) is a model cell line of a human prostatic carcinoma. This also suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Figure 9 shows the results of the TRRUST (Han et al., 2015) category of Metascape. It is obvious that the androgen receptor (AR) is not only top ranked but also outstandingly enriched. As mentioned above, AR is a protein that binds to the genome and mediates prostate cancer progression. This also suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Although there are more convincing results, the supplementary materials include the full report provided by Metascape.

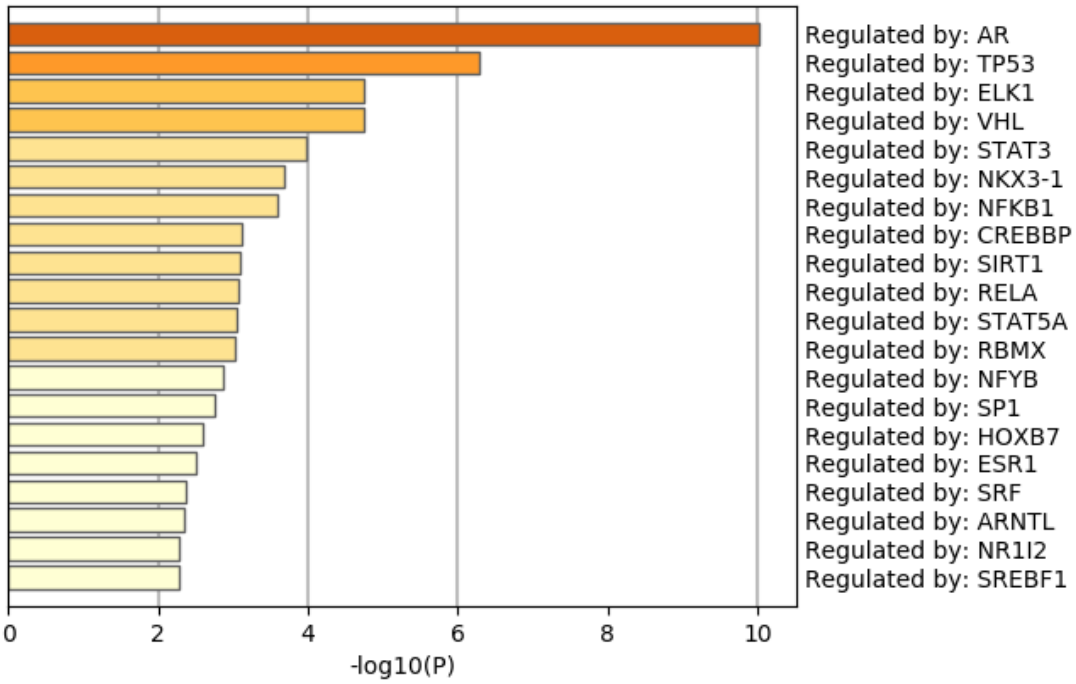


**Figure 9.** Summary of enrichment of the TRRUST category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

# DISCUSSION

## Synthetic data

To determine whether TD-based unsupervised FE outperformed the three supervised feature selections because of its unsupervised nature, we also applied an alternative unsupervised feature selection, MNMF, to the synthetic data, specifying $n = 3$ as the number of vectors each of which is composed of latent variables. Figures 10 and 11 show the latent variable vectors attributed to the samples and $i$s, respectively, computed by MNMF.

Compared with Figs. 2 and 3, it is unclear how we can utilize the latent variable vectors for feature selection. First, the latent variable vector attributed to the samples and coincident with $\left\lceil \frac{j}{3} \right\rceil \left( \left\lceil \frac{k}{3} \right\rceil + 3 \right)$ appears only in the third latent variable vector. In contrast, TD-based unsupervised FE has the first and second singular value vectors associated with $j$ and $k$. The correspondences with $j$ and $k$ of MNMF are even poorer than those of TD-based unsupervised FE (comparison between $u_{1j}$ and $u_{2k}$ in Fig. 2 and the third latent variable vector in Fig. 10). Moreover, even if we can identify the third latent variable vector attributed to the samples as corresponding with classifications (i.e., $j$ and $k$), we have no way to decide which latent variable vectors attributed to $i$ correspond to the third latent variable vector attributed to the samples, because there is no one-to-one correspondence between the latent variable vectors attributed to the samples and those attributed to the $i$s. In Eq. (8), the matrices composed of latent variables attributed to $i$s, $Q$, are multiplied with those composed of latent variables attributed to samples, $H_k$; thus, each of the three latent variable vectors attributed to $i$s (i.e., $n$ column vectors in $Q$) is multiplied with each of the three latent variable vectors attributed to samples (i.e., $n$ row vectors in $H_k$). If we cannot identify which latent variable vector attributed to $i$ corresponds to the third latent variable vectors attributed to samples that correspond with classification, we also cannot decide which latent variable vectors attributed to $i$s should be used to select $i$s that correspond with the classification. In general, the reason for employing NMF is because it can often provide sparse solution, which it failed to do in this situation because all components have non-zero values. Thus, TD-based unsupervised FE is clearly the best of the five methods tested on the synthetic dataset.

TD-based unsupervised FE also has the advantage of its CPU time (Table 6). TD-based unsupervised FE is the fastest of the five methods tested on the synthetic data. It can finish within a few seconds,

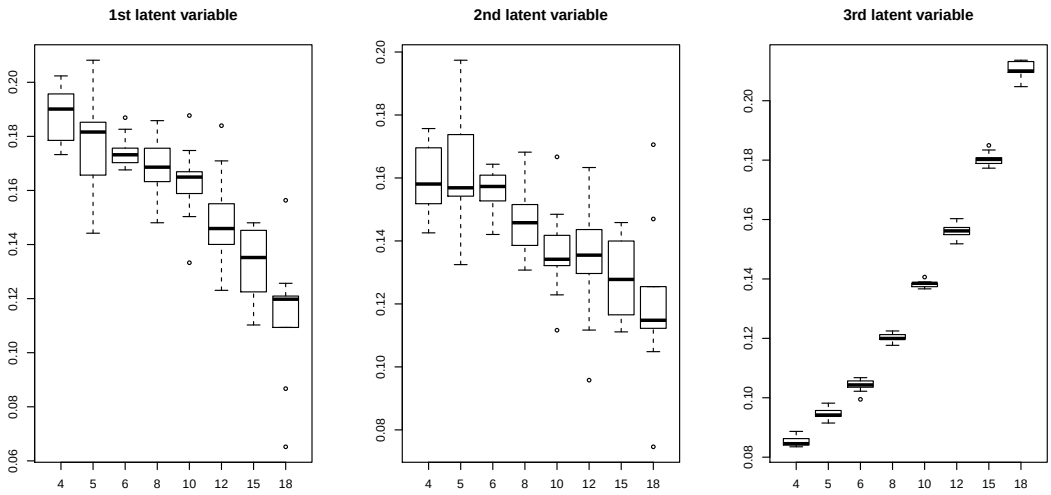


**Figure 10.** Boxplot of the first to third latent variable vectors attributed to the 81 samples. Horizontal axis: $\left\lceil \frac{j}{3} \right\rceil \left( \left\lceil \frac{k}{3} \right\rceil + 3 \right), 1 \leq j,k \leq 3$.

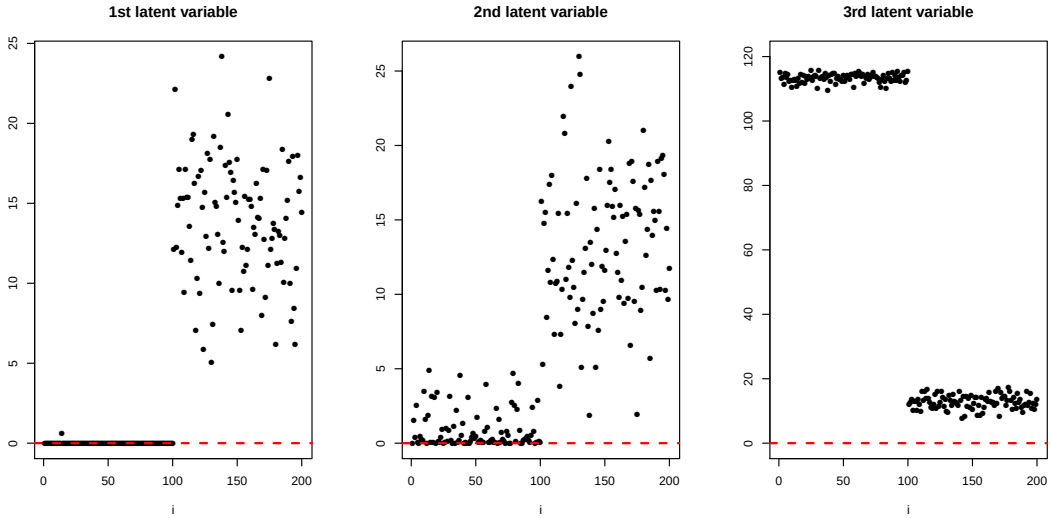


**Figure 11.** Scatter plots of the first to third latent variable vectors attributed to $i$ and i ($i \leq 200$ are shown). Horizontal red broken lines are base lines (zero).

whereas the others take longer, ranging from a few tens of seconds for categorical regression to a few tens of minutes for RF. In this sense, TD-based unsupervised FE is the fastest, most accurate, and most robust of the five methods tested on the synthetic dataset.

## Real data

We now compare the performances of TD-based unsupervised FE with those of the other four methods: RF, PenalizedLDA, categorical regression, and MNMF. First, RF was applied to the real dataset, which is classified into 16 categorical classes composed of any pairs of eight mulitiomics measurement groups and two tissues (tumor and normal prostate). RF results in as many as 11,278 genomic regions having non-zero importance. This number is almost eight times larger than the number of regions selected by TD-based unsupervised FE (1447). One of the evaluations of RF performance is its accuracy of discrimination, as RF is a supervised method. RF should have selected the minimum features that can successfully discriminate

**Table 4.** Confusion matrix for RF applied to the real dataset. N: normal prostate, T: tumor. The error rate is 35%.

| | AR | | ATAC | | FOXA1 | | H3K27AC | | H3K27me3 | | H3K4me3 | | HOXB13 | | K4me2 | | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | T | N | T | N | T | N | T | N | T | N | T | N | T | N | T | error |
| AR_N | 5 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/6 |
| AR_T | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| ATAC_N | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ATAC_T | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2/3 |
| FOXA1_N | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1/4 |
| FOXA1_T | 0 | 0 | 0 | 0 | 4 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 7/12 |
| H3K27AC_N | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2/15 |
| H3K27AC_T | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/5 |
| H3K27me3_N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2/3 |
| H3K27me3_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1/3 |
| H3K4me3_N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 1 |
| H3K4me3_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 |
| HOXB13_N | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 1 | 0 | 0 | 1/4 |
| HOXB13_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 11 | 0 | 0 | 1/12 |
| K4me2_N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 |
| K4me2_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 |

**Table 5.** Confusion matrix for LDA using $u_{1j}u_{2k}u_{3m}$ given by TD applied to 1447 genomic regions selected by TD-based unsupervised FE applied to the real dataset. N: normal prostate, T: tumor. The error rate is 37.5%.

| | AR | | ATAC | | FOXA1 | | H3K27AC | | H3K27me3 | | H3K4me3 | | HOXB13 | | K4me2 | | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | T | N | T | N | T | N | T | N | T | N | T | N | T | N | T | error |
| AR_N | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| AR_T | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| ATAC_N | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| ATAC_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| FOXA1_N | 0 | 0 | 0 | 0 | 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1/4 |
| FOXA1_T | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1/4 |
| H3K27AC_N | 0 | 0 | 2 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1/6 |
| H3K27AC_T | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 1/6 |
| H3K27me3_N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H3K27me3_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H3K4me3_N | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1/3 |
| H3K4me3_T | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1/3 |
| HOXB13_N | 0 | 0 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 1/2 |
| HOXB13_T | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 1/2 |
| K4me2_N | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| K4me2_T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

between the 16 classes. The error rate given by RF was as small as 0.35 (Table 4). This indicates that 65% of the 144 samples were correctly classified into the 16 classes (see Table 1 for details of how the 144 samples were classified into the 16 subclasses). To validate the discrimination performances of the 1447 genomic regions selected by TD-based unsupervised FE, we carried out the following. $u_{1j}, u_{2k}, u_{3k}$ were recomputed by HOSVD using only the 1447 genomic regions selected by TD-based unsupervised FE. The 144 samples were then discriminated using linear discriminant analysis, employing the product $u_{1j}u_{2k}u_{3k}$ as input vectors (Table 5 with leave-one-out cross-validation). The error rate was as small as 37.5%, similar to that of RF (0.35). Considering the fact that TD-based unsupervised FE employed only one-eighth of the number of genomic regions selected by RF, we concluded that TD-based unsupervised FE can outperform RF.

Next, we attempted to apply PenalizedLDA to the real dataset, but found that we could not. To perform LDA, all in-subclass variance must be non-zero. This condition was not fulfilled for the real data. Hence, PenalizedLDA cannot be employed to process real datasets.

Third, we applied categorical regression to the real dataset by regarding it as 16 categorical subclasses. The number of features associated with adjusted $P$-values less than 0.01 was as high as 106,761 among all 123,817 features. This indicates that categorical regression selected almost all features and is thus clearly ineffective. The reason categorical regression failed, despite being relatively successful (Table 2) when

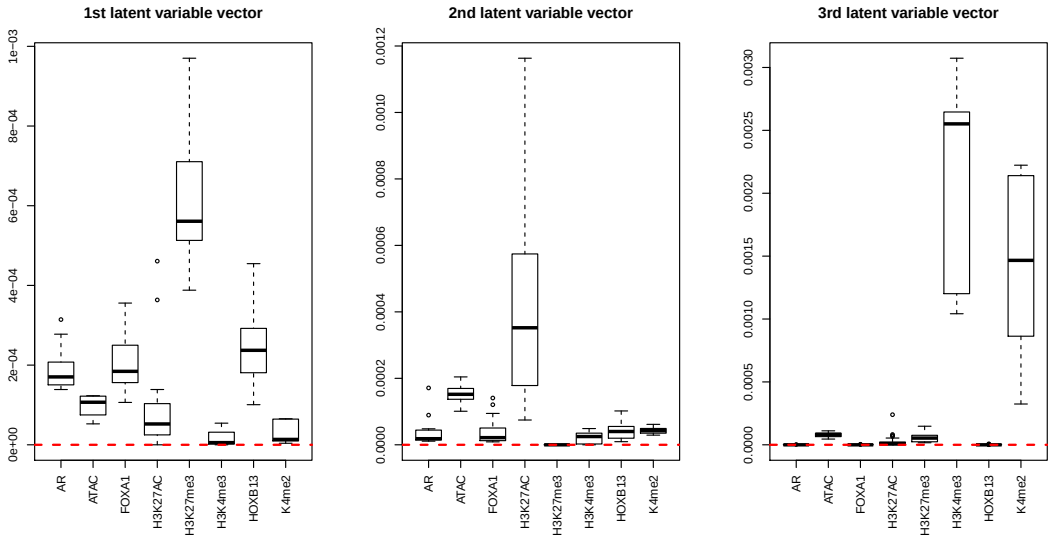


**Figure 12.** Boxplots of the first to third latent variable vectors attributed to 144 samples classified into eight subclasses, each of which corresponds to a multiomics measurement. The horizontal red broken lines are base lines (zero).

**Table 6.** CPU times of the five methods required for the synthetic and real datasets. The CPU time for PenalizedLDA is for testing four $\lambda$ values with between one and eight discriminant functions. The CPU time for an individual run is 18 s.

| | methods | TD based unsupervised FE | categorical regression | random forest | penalized LDA | MNMF |
|---|---|---|---|---|---|---|
| cpu time [sec] | synthetic data | 6 | 66 | 142 | 575 (18) | 328 |
| | real data | 20 | 87 | 321 | — | 223 |

applied to the synthetic dataset, is as follows. In the synthetic dataset, there was only one dependence upon subclasses. Selecting features with subclass dependence automatically gives us targeted features. In the real dataset, this cannot stand as it is. There are many variations of subclass dependence. In TD-based unsupervised FE, by assessing the dependence of singular value vectors upon subclasses (see Fig. 4), we can specify the type of dependence upon subclasses that should be considered. This allows us to identify a restricted number of biologically reliable genomic regions. Categorical regression cannot distinguish between various subclass dependences and can only identify genomic regions associated with any type of subclass dependence. This results in the identification of almost all genomic regions, as most are likely to be associated with a type of subclass dependence.

Finally, MNMF ($n = 3$), one of the state-of-the-art methods that are often applied to multiomics datasets (Baldwin et al., 2020), was applied to $x_{ijkm}$ as described in Materials and Methods. Fig. 12 shows the latent variable vectors attributed to 144 samples. Although the dependence upon eight multiomics measurements are similar to those of TD-based unsupervised FE, because the correspondence with eight types of measurement is clearly inferior to that of TD-based unsupervised FE (within class variances much larger than those of the singular value vectors obtained by TD-based unsupervised FE, Fig. 4), there is no reason to employ MNMF instead of TD-based unsupervised FE. An additional reason not to employ MNMF instead of TD-based unsupervised FE is that, although singular value vectors derived by HOSVD can be assumed to obey a Gaussian distribution, latent variables computed by MNMF are unlikely to because they do not take negative values. Thus, there is no way to attribute $P$-values to genes, as there is no suitable null hypothesis by which we can compute the $P$-values.

As a result, none of the four methods (RF, PenalizedLDA, categorical regression, or MNMF) can compete with TD-based unsupervised FE when applied to a real dataset.

In addition, the CPU time required is also the shortest compared to the other four methods when applied to the real dataset (Table 6). From this perspective, TD-based unsupervised FE outperforms the other methods.

## CONCLUSIONS

In this study, a TD-based unsupervised FE formalism is successfully applied for the first time to a variety of multiomics measurements for both synthetic and real datasets that represented a typical large $p$ small $n$ problem. The proposed method outperformed the conventional supervised feature selection methods of RF, categorical regression, PenalizedLDA, and MNMF, and was demonstrated to be superior not only in feature selection but also in selection of biologically reliable genes.

## ACKNOWLEDGMENT

This study was supported by KAKENHI, 20K12067, 20H04848, and 19H05270

## SUPPLEMENTARY DATA

Output from Metascape.

## REFERENCES

Awan, M. G., Eslami, T., and Saeed, F. (2018). GPU-daemon: GPU algorithm design, data management & optimization template for array based big omics data. *Computers in biology and medicine*, 101:163–173.

Baldwin, E., Han, J., Luo, W., Zhou, J., An, L., Liu, J., Zhang, H. H., and Li, H. (2020). On fusion methods for knowledge discovery from multi-omics datasets. *Computational and Structural Biotechnology Journal*, 18:509–517.

Bekkerman, R., Bilenko, M., and Langford, J. (2011). *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press.

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., and Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fujita, K. and Nonomura, N. (2019). Role of androgen receptor in prostate cancer: A review. *The World Journal of Men's Health*, 37(3):288.

Ge, J., Li, X., Jiang, H., Liu, H., Zhang, T., Wang, M., and Zhao, T. (2019). Picasso: A sparse learning library for high dimensional data analysis in r and python. *J. Mach. Learn. Res.*, 20:44–1.

Gerhardt, J., Montani, M., Wild, P., Beer, M., Huber, F., Hermanns, T., Müntener, M., and Kristiansen, G. (2012). FOXA1 promotes tumor progression in prostate cancer and represents a novel hallmark of castration-resistant prostate cancer. *The American Journal of Pathology*, 180(2):848–861.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C. Y., Cho, H. J., Kang, B., Shin, S., and Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5(1).

Horoszewicz, J. S., Leong, S. S., Kawinski, E., Karr, J. P., Rosenthal, H., Chu, T. M., Mirand, E. A., and Murphy, G. P. (1983). LNCaP model of human prostatic carcinoma. *Cancer Research*, 43(4):1809–1818.

Igolkina, A. A., Zinkevich, A., Karandasheva, K. O., Popov, A. A., Selifanova, M. V., Nikolaeva, D., Tkachev, V., Penzar, D., Nikitin, D. M., and Buzdin, A. (2019). H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 histone tags suggest distinct regulatory evolution of open and condensed chromatin landmarks. *Cells*, 8(9).

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Nashaat, M., Ghosh, A., Miller, J., Quader, S., Marston, C., and Puget, J.-F. (2018). Hybridization of active learning and data programming for labeling large industrial datasets. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 46–55. IEEE.

Navarro, H. I. and Goldstein, A. S. (2018). HoxB13 mediates AR-V7 activity in prostate cancer. *Proceedings of the National Academy of Sciences*, 115(26):6528–6529.

Ngiam, K. Y. and Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273.

Pan, J.-B., Hu, S.-C., Shi, D., Cai, M.-C., Li, Y.-B., Zou, Q., and Ji, Z.-L. (2013). PaGenBase: A pattern gene database for the global and dynamic understanding of gene function. *PLOS ONE*, 8(12):1–9.

Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Research*, 20(11):1493–1502.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richter, A. N. and Khoshgoftaar, T. M. (2019). Efficient learning from big data for cancer risk modeling: a case study with melanoma. *Computers in biology and medicine*, 110:29–39.

Santosh, T., Ramesh, D., and Reddy, D. (2020). LSTM based prediction of malaria abundances using big data. *Computers in Biology and Medicine*, page 103859.

Shah, R., Zhang, S., Lin, Y., and Wu, P. (2019). xSVM: Scalable distributed kernel support vector machine training. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 155–164. IEEE.

Taguchi, Y.-H. (2020). *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing.

Wen, F., Chu, L., Ying, R., and Liu, P. (2019). Fast and positive definite estimation of large covariance matrix for high-dimensional data analysis. *IEEE Transactions on Big Data*.

Witten, D. (2015). *penalizedLDA: Penalized Classification using Fisher's Linear Discriminant*. R package version 1.1.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Yang, S., Wen, J., Zhan, X., and Kifer, D. (2019). ET-lasso: a new efficient tuning of lasso-type regularization for high-dimensional data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 607–616.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., and Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10(1).