A tensor decomposition-based integrated analysis applicable to multiple gene expression profiles without sample matching

Taguchi Y-h. (✉ tag@granular.com)

Chuo University

Turki Turki

King Abdulaziz University

Article

Keywords: tensor decomposition-based integrated analysis, multiple gene expression profiles

Posted Date: August 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-766884/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A tensor decomposition-based integrated analysis applicable to multiple gene expression profiles without sample matching

Y-h. Taguchi* and Turki Turki†

July 30, 2021

Abstract

The integrated analysis of multiple gene expression profiles measured in distinct studies is always problematic. Especially, missing sample matching and missing common labeling between distinct studies prevent the integration of multiple studies in fully data-driven and unsupervised manner. In this study, we propose a strategy enabling the integration of multiple gene expression profiles among multiple independent studies without either labeling or sample matching, using tensor decomposition-based unsupervised feature extraction. As an example, we applied this strategy to Alzheimer’s disease (AD)-related gene expression profiles that lack exact correspondence among samples as well as AD single-cell RNA-seq (scRNA-seq) data. We found that we could select biologically reasonable genes with integrated analysis. Overall, integrated gene expression profiles can function analogously to prior learning and/or transfer learning strategies in other machine learning applications. For scRNA-seq, the proposed approach was able to drastically reduce the required computational memory.

1 Introduction

Integrated analysis of gene expression profiles [1–4] is generally difficult. Although the primary purpose of integration of multiple gene expression profiles is to compensate for small sample sizes by integrating gene expression profiles measured in multiple studies, this process is generally not straightforward because there are usually no matching samples between them. If sample matching information is missing, when samples are associated with common labeling, e.g., healthy controls and patients, although we can simply group gene expression profiles based upon labeling, the requirements of this process are not always fulfilled, either. Thus, the establishment of general frameworks to integrate multiple gene expression profiles without either sample matching or common sample labeling would be of considerable benefit. Because of the basic requirements, such methods must be unsupervised, because otherwise the integration of multiple gene expression profiles that share nothing other than genes between them. In this study, we employ tensor decomposition for this purpose [5].

Let $x_{ij} \in \mathbb{R}^{N \times M}$ and $x_{ij'} \in \mathbb{R}^{N \times M'}$ be two gene expression profile data represented as matrices that share genes i but not samples j and j' . Previously [5], we proposed two strategies to integrate them using a tensor as

$$x_{ijj'} = x_{ij}x_{ij'} \in \mathbb{R}^{N \times M \times M'}, \quad (1)$$

or as a matrix given by taking summation of $x_{ijj'}$ over i ,

$$x_{jj'} = \sum_i x_{ijj'} \in \mathbb{R}^{M \times M'}, \quad (2)$$

*Department of Physics, Chuo University

†King Abdulaziz University, Department of Computer Science, Jeddah, 21589, Saudi Arabia.

and applied higher-order singular value decomposition [5] (HOSVD) to them (for a matrix, HOSVD is equivalent to a usual singular value decomposition (SVD)). Although we have also suggested to extend this strategy to more than two matrices as

$$x_{ij_1j_2\cdots j_K} = x_{ij_1}x_{ij_2}\cdots x_{ij_K}, \quad (3)$$

or that whose summation is taken over i ,

$$x_{ij_1j_2\cdots j_K} = \sum_{i=1}^N x_{ij_1j_2\cdots j_K}, \quad (4)$$

the effect of multiplication on outcomes is unclear, and it was successfully applied only a few times to $K = 3$ [6, 7] and has never been applied to data sets with $K > 3$. In addition, the size of the tensor grows exponentially as K increases, because the size of the tensor is as large as $N \times M_1 \times \cdots \times M_K$, where M_k is the number of samples in k th omics profile. In contrast, we have recently proposed [8] an alternative strategy to integrate omics datasets to construct a tensor as

$$x_{j_kj'_k} = \sum_i x_{ij_k}x_{ij'_k} \in \mathbb{R}^{M_k \times M_k \times K} \quad (5)$$

However, this is possible only when M_k is constant and common among K omics data set. Thus, it is not applicable to the case considered in this study, i.e., there are no samples matching among distinct studies.

To apply TD based unsupervised FE to the situation in which there are no matching samples among independent studies, we applied TD or SVD to individual profiles prior to integrated analyses and projected individual gene expression profiles onto the space spanned by the top L eigenvectors regardless of M_K . This enabled us to obtain matrices having common $N \times L$ structures independent of k and $N \times L \times K$ tensor to which TD can be applied. As may be observed below, this simple idea can be applied relatively effectively.

2 Results

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

2.1 Integrated analysis of data sets 1, 2 and 3

At first, to test whether this simple idea can integrate multiple gene expression profiles lacking sample matching, we tried to integrate three gene expression profiles, referred to as datasets 1, 2 and 3. Datasets 1 and 3 were composed of cell lines that model either healthy controls or AD, whereas dataset 2 comprised cell lines aiming to mutate AD associated genes, and thus was related to AD only indirectly. After obtaining a singular value vector $u_{\ell_2\ell}$ as described in §5.4.1, we first investigated whether they were associated with classifications. Singular value vectors attributed to the j_k th sample $\tilde{v}_{\ell_2j_k}^{[k]}, 1 \leq k \leq 3$ were computed from $u_{\ell_2\ell}$ by eq. (14), of which whether $\tilde{v}_{\ell_2j_k}^{[k]}$ was coincident with classification in any of datasets 1, 2, or 3 was investigated (Table 1 and Fig. 2). It is obvious that $\tilde{v}_{\ell_2j_k}^{[k]}, 1 \leq \ell_2 \leq 4$ were significantly correlated with classifications in any of datasets 1, 2, or 3. Next, we tried to find which u_{ℓ_1i} were associated with $u_{\ell_2\ell}, 1 \leq \ell_2 \leq 3$ which was coincident with classifications through $\tilde{v}_{\ell_2j_k}^{[k]}$ computed from $u_{\ell_2\ell}$ s with eq. (14). Then, we selected that $u_{\ell_1i}, 1 \leq \ell_1 \leq 5$ as those associated with $u_{\ell_2\ell}, 1 \leq \ell_2 \leq 3$ by investigating $G(\ell_1\ell_2\ell_3)$ in eq. (13) while fixing $1 \leq \ell_2 \leq 3$ (see Table S1). P_i s are attributed to i using eq. (17) while taking $\Omega_{\ell_1} = \{\ell_1 | 1 \leq \ell_1 \leq 5\}$. The obtained P_i s were corrected by BH criterion [5] and 565 i s (genes) associated with adjusted P_i s less than 0.01 were selected. To validate the selected 565 genes biologically, gene symbols associated with genes were uploaded to Enrichr [9]. We observed various enrichment

(Data S1). For example, the “KEGG 2021 Human” category listed six neurodegenerative diseases’ related pathways within the top ranked ten pathways (Table 2(A)). Similarly, the “Jensen Diseases” category listed four neurodegenerative diseases within the top ranked ten diseases, as well (Table 2(A)). Additionally, “Disease Perturbations from GEO down” and “Disease Perturbations from GEO up” listed similar neurodegenerative diseases within the top ranked diseases (Data S1). We noted from that “Allen Brain Atlas up” and “Allen Brain Atlas down” categories that many AD related brain regions were highly ranked (Data S1). Enrichment analyses using databases other than Enrichr also reported convincing results. The top- and the third-ranked terms in “GAD_DISEASE_CLASS” category of DAVID [10, 11] are PSYCH and NEUROLOGICAL, respectively (Table 2(A)). The top ten ranked terms in “GAD_DISEASE” category of DAVID included five neurodegenerative diseases (Table 2(A)). g:Profiler [12] listed 16 KEGG pathways as significantly enriched with gene symbols associated with 565 selected genes; 16 pathways included six neurodegenerative disease-related pathways (Table 2(A)). Thus, obviously TD-based unsupervised FE was used to integrate three gene expression profiles lacking sample matching, genes the expressions of which were associated with classifications in three independent studies were identified, and genes selected were enriched by various neurodegenerative disease-related terms for three biological databases. This suggests that TD based unsupervised FE is promising method to integrate gene expression profiles without sample matching.

2.2 Drug repositioning using the tensor obtained with data sets 1, 2 and 3

Previously [13], we have demonstrated that integrated analysis of gene expression profiles between model animals treated by various drugs and patients is useful to see which drug compounds are effective against diseases. In our previous study [13], we were able to integrate only one gene expression profile of disease with that of drug treatment because we employed eq. (1). In this study, we were able to integrate more gene expression profiles of diseases with those of drug treatment because of a novel framework introduced herein. As gene expression profiles of a disease, we employed that obtained in the previous subsection, i.e., the integration of dataset 1, 2 and 3, $x_{ilk}, 1 \leq k \leq 3$, and we employed dataset 4 as a gene expression profile of drug treatment. We got x_{il4} as described in §5.4.2 to be integrated with $x_{ilk}, 1 \leq k \leq 3$, obtained in the previous section and HOSVD was applied to the obtained $x_{ilk}, 1 \leq k \leq 4$.

At first, we sought to validate which singular value vectors, $u_{\ell_2\ell}$, were associated with classifications when they were converted back to $\tilde{v}_{\ell_2j_k}^{[k]}$ (Table 1 and Fig. 3). In contrast to the integrated analysis of datasets 1, 2 and 3, in this case we must consider up to $\ell_2 = 4$ to obtain singular value vectors associated with any of the classifications of datasets 1, 2 or 3. Then, we selected $u_{\ell_1i}, 1 \leq \ell_1 \leq 5$ as those associated with $u_{\ell_2\ell}, 1 \leq \ell_2 \leq 4$ by investigating $G(\ell_1, \ell_2, \ell_3)$ in eq. (13) while fixing $1 \leq \ell_2 \leq 4$ (see Table S2). P_i s are attributed to i using eq. (17) while taking $\Omega_{\ell_1} = \{\ell_1 | 1 \leq \ell_1 \leq 5\}$. The obtained P_i s were corrected by BH criterion and 544 i s (genes) associated with adjusted P_i s less than 0.01 were selected. To validate the selected 544 genes biologically, gene symbols associated with these 544 genes were uploaded to Enrichr. We observed various enrichment (Data S2). For example, the “KEGG 2021 Human” category listed six neurodegenerative diseases’-related pathways within the top ranked ten pathways (Table 2(B)). Similarly, the “Jensen Diseases” category listed four neurodegenerative diseases within the top ranked ten diseases, as well (Table 2(B)). In addition, “Disease Perturbations from GEO down” and “Disease Perturbations from GEO up” listed similar neurodegenerative diseases within the top ranked diseases (Data S2). Examining the “Allen Brain Atlas up” and “Allen Brain Atlas down” categories, many AD related brain regions were highly ranked (Data S2). Enrichment analyses using other databases than Enrichr also reported convincing results. The first- and fourth-ranked terms in “GAD_DISEASE_CLASS” category of DAVID were PSYCH and NEUROLOGICAL, respectively (Table 2(B)). The top ten ranked terms in “GAD_DISEASE” category of DAVID included four neurodegenerative diseases (Table 2(B)). g:Profiler listed 13 KEGG pathways as significantly enriched with gene symbols associated with 544 selected genes; 13 pathways included six neurodegenerative diseases related pathways (Table 2(B)). Thus, it is

clear that TD based unsupervised FE integrated four gene expression profiles lacking sample matching, identified genes whose expression was associated with classifications in three independent studies simultaneously, and selected genes were enriched by various neurodegenerative disease-related terms for three biological databases.

Next, we tried to identify how this strategy to rank drugs could be used to generate gene expression profiles in dataset 4. After computing $\tilde{u}_{\ell_2 j_4^{[1]} j_4^{[2]} j_4^{[3]}}^{[4]} \in \mathbb{R}^{94 \times 4 \times 3}$ by eq. (15), SVD was applied to $\tilde{u}_{\ell_2 j_4^{[1]} j_4^{[2]} 1}^{[4]} \in \mathbb{R}^{94 \times 4}$, and we obtained

$$\tilde{u}_{\ell_2 j_4^{[1]} j_4^{[2]} 1}^{[4]} = \sum_{\tilde{\ell}=1}^4 \lambda_{\tilde{\ell}} \tilde{u}_{\tilde{\ell} j_4^{[1]}}^{[\ell_2]} \tilde{u}_{\tilde{\ell} j_4^{[2]}}^{[\ell_2]} \quad (6)$$

where $\tilde{u}_{\tilde{\ell} j_4^{[1]}}^{[\ell_2]} \in \mathbb{R}^{94 \times 4}$ was a singular value vector attributed to drugs and $\tilde{u}_{\tilde{\ell} j_4^{[2]}}^{[\ell_2]} \in \mathbb{R}^{4 \times 4}$ was a singular value vector attributed to dose density. At first, we checked $\tilde{u}_{\tilde{\ell} j_4^{[2]}}^{[1]}$ while selecting $\ell_2 = 1$ to determine whether they were associated with dose dependence, and based on the above we have that they were dose-dependent. Then, we checked which drugs had larger $\left| \tilde{u}_{\tilde{\ell} j_4^{[1]}}^{[1]} \right|$. Table 3 shows the top ranked five drugs for $\tilde{\ell} = 1$ and 2, respectively. Because two of the five drugs were regarded as effective in the original study [14], drug repositioning seems to be successful.

2.3 Transfer learning using the tensor obtained with data sets 1, 2 and 3

One possible application of the integrated analysis of multiple gene expressions lacking sample matching is analogous to transfer learning (TL) [15]. In TL, pre-trained machine learning models are used to achieve better performance even with smaller sample sizes. Although deep learning (DL) is often used for TL, DL architectures are not suitable for gene expression profile processing; for example, Yifei et al [16] found that DL achieved better performance than linear regression (LR) only by less than seven percent. This may be attributed to the lack of structure of gene expression profiles; in contrast to images and sentences (or documents) to which DL was applied successfully, gene expressions are orderless vectors of real numbers. In contrast, TD was successfully applied to gene expression profiles [5] because tensor similarly does not consider the order of real numbers at all. Thus, TD is suitable architecture to process gene expression profiles.

To determine whether the integrated analysis of multiple gene expression profiles obtained in §2.1 could be used as a pre-trained system in TL, we combined dataset 5 with $x_{i\ell k}$, $1 \leq k \leq 3$ as a pre-trained system. Dataset 5 comprised gene expression profiles caused by overexpression of the ABCC1 gene that was recently recognized as an AD therapy target [17]. Usually, only gene expression profiles caused by overexpression of ABCC1 gene are analyzed and genes whose expression was altered by ABCC1 over-expression are selected. Then selected genes are then compared with genes whose expression is known to be altered by AD. Nevertheless, this procedure is somewhat indirect because no sample matching between gene expression profiles of AD and ABCC1 overexpression experiments exist. In our strategy, because we can directly compare these expression profiles, our strategy is more suitable for the comparison of gene expression profiles of AD and ABCC1 overexpression.

After getting TD as in eq. (13), we compute $\tilde{v}_{\ell_2 j_k}^{[k]}$, $1 \leq k \leq 3$ by eq. (14) to determine whether they are still coincident with the classification even after merging $x_{i\ell k}$, $1 \leq k \leq 3$ with $x_{i\ell 5}$. For dataset 5, after $\tilde{u}_{\ell_2 j_5^{[1]} j_5^{[2]} j_5^{[3]}}^{[5]}$ is computed by eq. (15), the correlation between six categories in dataset 5 described in §5.3.5 and $\tilde{u}_{1 j_5^{[1]}}^{[\ell_2]} \tilde{u}_{2 j_5^{[2]}}^{[\ell_2]} \tilde{u}_{1 j_5^{[3]}}^{[\ell_2]}$ computed by eq. (16) is computed (Table 1 and the results are shown in Fig. 4). Although we have considered up to $\ell_2 = 5$, $u_{\ell_2 \ell}$ s were coincident with classification after being converted to $\tilde{v}_{\ell_2 j_k}^{[k]}$. Thus, as expected, the data showed correlations between gene expression profiles of AD and ABCC1 overexpression through TD.

Then we selected $u_{\ell_1 i}$, $1 \leq \ell_1 \leq 5$ as those associated with $u_{\ell_2 \ell}$, $1 \leq \ell_2 \leq 5$ by investigating $G(\ell_1, \ell_2, \ell_3)$ in eq. (13) and fixing $\ell_2 \in \{1, 3, 4, 5\}$ (see Table S3). P_i s were attributed to i using eq. (17) while taking

$\Omega_{\ell_1} = \{\ell_1 | 1 \leq \ell_1 \leq 5\}$. The obtained P_i s were corrected by BH criterion and 660 *is* (genes) associated with adjusted P_i s less than 0.01 were selected. To validate the selected 660 genes biologically, gene symbols associated with genes were uploaded to Enrichr. We observed various enrichment (Data S3). For example, the “KEGG 2021 Human” category listed six neurodegenerative diseases’ (Table 2(C)) related pathways within the top ranked ten pathways. Similarly, the “Jensen Diseases” category listed four neurodegenerative diseases (Table 2(C)) within the top ten ranked diseases, as well. “Disease Perturbations from GEO down” and “Disease Perturbations from GEO down” listed similar neurodegenerative diseases within the top ranked diseases (Data S3). In “Allen Brain Atlas up” and “Allen Brain Atlas down” categories, many AD related brain regions were highly ranked (Data S3). Enrichment analyses using databases other than Enrichr also reported convincing results. The top tree ranked terms in the “GAD_DISEASE_CLASS” category of DAVID were PSYCH, NEUROLOGICAL, and AGING (Table 2(C)). The top ten ranked terms in the “GAD_DISEASE” category of DAVID included four neurodegenerative diseases (Table 2(C)). g:Profiler listed 16 KEGG pathways as significantly enriched with gene symbols associated with 633 selected genes; 16 pathways included six neurodegenerative disease-related pathways (Table 2(C)). Thus, we were able to successfully select genes whose expression was simultaneously altered by both AD and ABCC1 overexpression and biological properties evaluated by enrichment analysis is reasonable.

2.4 Integrated analysis of scRNA-seq data

In this subsection, we tried to apply our strategy to integrated analysis of scRNA-seq data set. Because individual experiments with scRNA-seq included $\sim 10^4$ cells, it was unclear how to integrate multiple scRNA-seq profiles. After obtaining TD as shown in eq. (13), where k was replaced with an integer $1 \leq c \leq 25$ representing the c th RNA-seq measurement, as described in §5.4.4, we focused $u_{\ell_3 c} \in \mathbb{R}^{25 \times 25}$ because c was divided into four groups, including either AD or healthy controls in either hippocampus or cortex brain regions. Next, we tried to determine which $u_{\ell_3 c}$ was associated with these four classes (Fig. 5). Then, we found that only u_{6c} was associated with these four classes; categorical regression was applied to $u_{\ell_3 c}$ for each ℓ_3 separately, P -values were computed and corrected by BH criterion. Only $\ell_3 = 6$ was associated with adjusted P -values less than 0.05. As a result, we found that $\sum_{\ell_2} G(6\ell_2 6)^2$ was the largest among $\sum_{\ell_2} G(\ell_1 \ell_2 6)^2$ where a summation was taken over ℓ_2 because there were no reasons to select one specific ℓ_2 (Table S4). P_i were attributed to *is* by eq. (17) with $\Omega_{\ell_1} = \{\ell_1 | 6\}$. P_i s were corrected and 177 *is* (genes) were associated with an adjusted P_i s of less than 0.01. Gene symbols associated with these 177 *is* were uploaded to Enrichr (Data S4). The results were very distinct from those in Table 2. No neurodegenerative disease terms were detected within the top ten ranked terms in either of the “KEGG 2021 HUMAN” “JENSEN DISEASES” categories. Instead, the top ten ranked terms in the “Human Gene Atlas” category contained several brain tissues (Table 4). In contrast, the “Disease Perturbations from GEO down” and “Disease Perturbations from GEO down” categories did include many diseases related to the brain (Table 4). Although these are only some examples, in spite of the simplicity of our strategy, it was able to list genes that are associated with tissue specificity as well as AD. Thus, it is obvious that our proposed strategy can deal with scRNA-seq very easily even without directly considering the massive number of cells in individual scRNA-seq measurements, because we consider only the top 10 singular value vectors within individual scRNA-seq gene expression profiles that include as many as 10^4 cells.

3 Discussions

It is evident that our implementation has, at least, two unique advantages. The first advantage is that it enables the integration of multiple gene expression profiles without sample matching. Datasets 1, 2, and 3 had sample sizes of nine, 23, and eight. Replacing these numbers of samples with those of singular value vectors, eight, we

can obtain a single tensor $x_{i\ell k}$ to which HOSVD is easily applied. It is also possible to evaluate the consistency of singular value vectors, $u_{\ell 2\ell}^{[k]}$, obtained by applying HOSVD to $x_{i\ell k}$, with classifications of individual data sets by converting $u_{\ell 2\ell}^{[k]}$ back to singular value vectors $\tilde{v}_{\ell 2j_k}^{[k]}$, with eq. (14), attributed to samples in individual datasets. Because this strategy is clearly applicable to the general number of samples in individual gene expression profiles, it is very useful. Its second advantage is a reduction in the computational memory required for analysis. When we applied this implementation to scRNA-seq data, we employed only the highest-ranked 10 singular value vectors computed by applying SVD to individual scRNA-seq data, rather than considering as many as $\sim 10^4$ cells. This drastically reduced the required memory by a factor of one hundred. In spite of the greatly reduced memory requirements, we were able to select genes associated with various significantly enriched biological terms related to tissue specificity as well as diseases, as expected. Because more cells are expected to be sequenced in individual experiments with scRNA-seq in the future, our strategy is recommended for a wide range of scRNA-seq analysis applications.

Figures 2, 3, and 4 show that the proposed strategy can relate gene expression profiles lacking sample matching with one another. For example, Figures 2 (A), (C) and (F) are interrelated with one another since they were commonly derived from $u_{1\ell}$ with eq. (14), and were related to $u_{\ell 1i}$ s. This means that the gene expression profiles shown in Figs. 2 (A), (C) and (F) should share gene dependence, represented by corresponding $u_{\ell 1i}$ s. Similarly, Figures 2 (D) and (G) derived from $u_{2\ell}$ should be interrelated with each other and Figs. 2 (B) and (E) derived from $u_{3\ell}$ should be interrelated with each other. These correspondences can help us interpret individual experiments. For example, Fig. 2(C) shows the expressions that were associated with distinction between PTPN6 suppression and control in WT whereas no distinction between PTPN6 suppression and control was observed in the CD33 KO cell line because it is related to Fig. 2(F), which represents the distinction between AD1 and control, but there was no distinction between AD2 and the control. Thus, it is likely that the AD1 cell line was associated with PTPN6 suppression or expression. Although this is only one example, the present strategy should have ability to relate two distinct experiments with each other. When we compare Fig. 4 with Fig. 2, we can understand the effect of ABCC1 overexpression. Possibly, it might be expected that overexpression of ABCC1 should be related to the distinction between AD and healthy control. Nevertheless, this was not the case. Although Figs. 4 (G) and (H) were related to distinction between ABCC1 over expression and controls, $u_{1\ell}$ or $u_{2\ell}$, from which $\tilde{v}_{1j_5}^{[5]}$ or $\tilde{v}_{2j_5}^{[5]}$ shown in Figs. 4 (G) and (H) were derived, and were not associated with anything in datasets 1, 2, and 3. Instead, Figs. 4 (A), (C) and (I) were commonly derived from $u_{4\ell}$; Figure 4 (I) represents not the effect of ABCC1 overexpression but the distinction between RNA sequences used for the treatment. It is coincident with that 2 (B) and (E), which appear very similar to 4 (A) and (C), were commonly derived from $u_{3\ell}$. This might help us to understand how ABCC1 can affect the progression of AD.

If we compare Fig. 2 with either Fig. 3 or Fig. 4, we can obtain further insight of these expression. For example, Figs. 3 (A) and (C) were derived from $u_{1\ell}$ as Figs. 2 (A) and (C). In contrast, although Fig. 2 (F) was also derived from $u_{1\ell}$, nothing associated with dataset 3 was derived from $u_{1\ell}$ in Fig. 3. This suggests that Fig. 2(F), which represents a distinction between AD1 and the control but no distinction between AD2 and the control, was not coincident with drug treatment. Possibly, there were no drugs that can affect AD1 but cannot affect AD2. Drug treatment might result in not the profiles represented by 2(F), but rather those represented by Fig. 3(G), i.e., monotonic expression starting from control through AD2 to AD1 cell lines if these cell lines are treated with these drugs. On the other hand, it is remarkable that both 2(C) and 3(C) represented similar profiles, i.e., a distinction between PTPN6 suppression and control in WT whereas no distinction between PTPN6 suppression and control was observed in the CD33 KO cell line. This suggests that these profiles are coincident with drug treatment. In other words, if these cell lines are treated with these drugs, this tendency will be maintained or enhanced.

To demonstrate the superior performance of our proposed strategy, we compared the performance to some conventional methods proposed to integrate multiple matrices. Unfortunately, most of methods designed for ap-

plication to genomic science require sample matching. For example, all of the nine methods listed in reference [18] require sample matching. intNMF [19] is also limited to datasets attributed to the same individuals. In contrast to these studies, although a review [20] comprehensively reported on integrated multi-view analyses, samples are implicitly assumed to be shared, and no integrated methods aim to integrate multiple matrices or tensors that do not share samples but only features were included. Although MINT [21] was proposed to integrate gene expression profiles sharing genes rather than samples, it assumed common labeling among multiple studies in a supervised learning framework, whereas its unsupervised framework required the common quantitative variables associated with all of samples included in the studies; thus, this method cannot be applied to the present task of investigating the correlations between latent variables and classification not shared between individual studies, as shown in Table 1. We could not find any implementations designed to integrate multiple gene expression profiles formatted as matrices or tensors sharing only genes and not samples. Thus, we were forced to seek suitable methods outside the genomic science. At first, we applied collective matrix factorization (CMF) [22], which was implemented as a CMF package in R [23], to the integration of datasets 1, 2, and 3. We computed four latent variables assuming a Poisson distribution for gene expression. No obtained latent variables were significantly correlated with classification in data sets 1, 2 and 3 (Table 5). Next, we tried group factor analysis (GFA) [24], which is also implemented as a GFA package in R. Although we computed five latent variables, none of them were correlated with the classification in dataset 1, although some of them were correlated with the classifications in datasets 2 and 3 (Table 5). Thus, these two advanced methods, CMF and GFA, failed to identify latent variables correlated with classifications of all three datasets, 1, 2, and 3, using integrated analysis. Finally, we tried simple concatenation. Ironically, SVD applied to 60617×40 contracted matrix composed of $x_{ij_1} \in \mathbb{R}^{60617 \times 9}$, $x_{ij_2} \in \mathbb{R}^{60617 \times 23}$, and $x_{ij_3} \in \mathbb{R}^{60617 \times 8}$, gave us the first singular value vector which was correlated with all of the classifications of the three datasets, although P -values for the correlation with classification of dataset 1 was only very slightly significant ($P = 0.05$) (Table 5). We were able to select 147 genes using eq. (17) by replacing $u_{\ell_1 i}$ with the first singular value vectors correlated with all of classifications of datasets 1, 2, and 3. To confirm the inferiority of simple concatenation toward TD-based unsupervised FE, although we evaluated their enrichment by uploading 147 genes to Enrichr, DAVID, and g:Profiler their enrichment was clearly poorer than those of genes selected by TD-based unsupervised FE (Table 2(D) and Data S5). Thus, we could not find any other methods that were comparable with or superior to TD-based unsupervised FE. Considering that two advanced methods, CMF and GFA, were inferior to even SVD applied to simple concatenated matrices, the methods developed for this purpose, i.e., integrated analysis of multiple matrices, seem to be inapplicable to the present purpose. Hence more advanced methods must be developed more suitable for the purpose of the present work, i.e., the integrated analysis of gene expression profiles that lack common matching samples.

An additional advantage of the proposed strategy may be noted as a bi-product. scRNA-seq data is known include many missing values, and matrix representation was previously employed to resolve this problem [25]. In the proposed strategy, although the employment SVD and HOSVD did not aim to fill missing values, it should naturally function as a mechanism to do so. This may be the reason the present strategy also works for scRNA-seq data as well without any specific additional modifications and implementations toward the treatment of scRNA-seq data set. Of note our proposed strategy was able to provide a platform to integrate typical gene expression profile measurements with small samples with scRNA-seq, which includes massive numbers of single cells, because both can be represented as a tensor form $x_{ijk} \in \mathbb{R}^{N \times L \times K}$ no matter how large number of cells are included in scRNA-seq. Because we could process scRNA-seq with $L = 10$ in this study, this might suggest some possibility that scRNA-seq does not provide us as large amount of information as expected by the number of single cells.

4 Conclusions

In this study, we have proposed a new strategy that can integrate multiple gene expression profiles that lack not only sample matching but also common labeling among them. This strategy has successfully integrated both AD gene expression profiles lacking sample matching and AD scRNA-seq datasets. The former was also used for drug repositioning and TL, whereas we were able to save massive amounts of computational memory with the latter. The proposed strategy appears useful for integrating gene expression profiles even if they lack both sample matching and common labeling among them.

5 Methods

5.1 Mathematical Formulation

In the following, we consider the cases where the number of genes, N , is much larger than the number of samples in the k th gene expression profile, M_k , as $N \gg M_k$. Suppose that we have K gene expression profiles, represented as either a matrix

$$x_{ij_k} \in \mathbb{R}^{N \times M_k} \quad (7)$$

or a $(S+1)$ -mode tensor

$$x_{i j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}} \in \mathbb{R}^{N \times M_k^{[1]} \times M_k^{[2]} \times \dots \times M_k^{[S]}}, \quad (8)$$

where $1 \leq k \leq K$. Then, SVD or HOSVD is applied to them and we get

$$x_{ij_k} = \sum_{\ell} u_{\ell i}^{[k]} \lambda_{\ell}^{[k]} v_{\ell j_k}^{[k]} \quad (9)$$

where $u_{\ell i}^{[k]} \in \mathbb{R}^{N \times M}$ are eigenvectors attributed to genes and $u_{\ell j}^{[k]} \in \mathbb{R}^{M \times M}$ are eigenvectors attributed to samples, $\lambda_{\ell}^{[k]}$ are eigenvalues, or

$$x_{i j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}} = \sum_{\ell_1 \ell_2 \dots \ell_S \ell_{S+1}} G\left(\ell_1^{[k]} \ell_2^{[k]} \dots \ell_S^{[k]} \ell_{S+1}^{[k]}\right) u_{\ell_1 j_k^{[1]}}^{[k]} u_{\ell_2 j_k^{[2]}}^{[k]} \dots u_{\ell_S j_k^{[S]}}^{[k]} u_{\ell_{S+1} i}^{[k]}, \quad (10)$$

where $G\left(\ell_1^{[k]} \ell_2^{[k]} \dots \ell_S^{[k]} \ell_{S+1}^{[k]}\right) \in \mathbb{R}^{M_k^{[1]} \times M_k^{[2]} \times \dots \times M_k^{[S]} \times N}$ is a core tensor representing a weight of products, $u_{\ell_1 j_k^{[1]}}^{[k]} u_{\ell_2 j_k^{[2]}}^{[k]} \dots u_{\ell_S j_k^{[S]}}^{[k]} u_{\ell_{S+1} i}^{[k]}$, $u_{\ell_s j_k^{[s]}}^{[k]} \in \mathbb{R}^{M_k^{[s]} \times M_k^{[s]}}$ and $u_{\ell_{S+1} i}^{[k]} \in \mathbb{R}^{N \times N}$ are singular value matrices and are orthogonal matrices. Then, we derived reduced matrices as either

$$x_{i \ell k} = \sum_{j_k=1}^{M_k} x_{ij_k} v_{\ell j_k}^{[k]}, 1 \leq \ell \leq L, \quad (11)$$

or

$$x_{i \ell k} = \sum_{j_k^{[1]}=1}^{M_k^{[1]}} \dots \sum_{j_k^{[S]}=1}^{M_k^{[S]}} x_{i j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}} u_{\ell_1 j_k^{[1]}}^{[k]} \dots u_{\ell_S j_k^{[S]}}^{[k]},$$

$$1 \leq \ell_s \leq L_k^{[s]} (\leq M_k^{[s]}), 1 \leq \ell \leq L = \prod_{s=1}^S L_k^{[s]}. \quad (12)$$

This implementation does involve a problem of note; for example, in eq. (9), simultaneous changes, $u_{\ell_i}^{[k]} \rightarrow -u_{\ell_i}^{[k]}$ and $v_{\ell_{j_k}}^{[k]} \rightarrow -v_{\ell_{j_k}}^{[k]}$, also satisfy eq. (9). This means that we cannot fix signs of $u_{\ell_i}^{[k]}$ and $v_{\ell_{j_k}}^{[k]}$. This does not present a problem provided that we are interested in individual gene expression profiles, but if we would like to compare multiple profiles, it does. For example, we compare multiple profiles in scRNA-seq data analysis in this study. Thus, we fix signs of $v_{\ell_{j_k}}^{[k]}$ such that the correlation coefficients between $u_{\ell_i}^{[1]}$ and $u_{\ell_i}^{[k]}$, $k \neq 1$, have positive values prior to the computations in eq. (11).

HOSVD was applied to $x_{i\ell k} \in \mathbb{R}^{N \times L \times K}$ and we obtained

$$x_{i\ell k} = \sum_{\ell_1=1}^N \sum_{\ell_2=1}^L \sum_{\ell_3=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 \ell} u_{\ell_3 k}, \quad (13)$$

where $G(\ell_1 \ell_2 \ell_3) \in \mathbb{R}^{N \times L \times K}$ is a core tensor and $u_{\ell_1 i} \in \mathbb{R}^{N \times N}$, $u_{\ell_2 \ell} \in \mathbb{R}^{L \times L}$ and $u_{\ell_3 k} \in \mathbb{R}^{K \times K}$ are singular value matrices and are orthogonal matrices. When the k th gene expression profile is a matrix, we also can occasionally convert $u_{\ell_2 \ell}$ back to the original notation as

$$\tilde{v}_{\ell_2 j_k}^{[k]} = \sum_{\ell=1}^L u_{\ell_2 \ell} v_{\ell j_k}^{[k]} \quad (14)$$

to evaluate the dependence of singular value vectors on samples j_k in the k th gene expression profiles. When k th gene expression profile is a tensor,

$$\tilde{u}_{\ell_2 j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}}^{[k]} = \sum_{\ell=1}^L u_{\ell_2 \ell} u_{\ell_1(\ell) j_k^{[1]}}^{[k]} u_{\ell_2(\ell) j_k^{[2]}}^{[k]} \dots u_{\ell_S(\ell) j_k^{[S]}}^{[k]}, \quad (15)$$

where $\ell_s(\ell)$ represents the correspondence between ℓ and ℓ_s in eq. (12). HOSVD was applied to $\tilde{u}_{\ell_2 j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}}^{[k]}$ and we obtained

$$\tilde{u}_{\ell_2 j_k^{[1]} j_k^{[2]} \dots j_k^{[S]}}^{[k]} = \sum_{\tilde{\ell}_1=1}^{L_k^{[1]}} \sum_{\tilde{\ell}_2=1}^{L_k^{[2]}} \dots \sum_{\tilde{\ell}_S=1}^{L_k^{[S]}} \tilde{G}^{[\ell_2]}(\tilde{\ell}_1 \tilde{\ell}_2 \dots \tilde{\ell}_S) \tilde{u}_{\tilde{\ell}_1 j_k^{[1]}}^{[\ell_2]} \tilde{u}_{\tilde{\ell}_2 j_k^{[2]}}^{[\ell_2]} \dots \tilde{u}_{\tilde{\ell}_S j_k^{[S]}}^{[\ell_2]} \quad (16)$$

and $\tilde{u}_{\tilde{\ell}_S j_k^{[S]}}$ s were used to evaluate the dependence of singular value vectors upon samples, $j_k^{[s]}$.

5.2 TD based unsupervised FE

Using the mathematical formulation introduced in §5.1, first, we investigate which $u_{\ell_3 k}$ has larger contributions to k th gene expression profile. Second, we investigate which of $\tilde{v}_{\ell_2 j_k}^{[k]}$ or $\tilde{u}_{\tilde{\ell}_S j_k^{[S]}}^{[\ell_2]}$ s are associated with the desired property, e.g., distinction between healthy control set and patients. Then, we check which of $G(\ell_1 \ell_2 \ell_3)$ has larger absolute values with fixing ℓ_2 and ℓ_3 . Finally, we select genes i by attributing P -values to i th genes while assuming $u_{\ell_1 i}$ obeys a Gaussian distribution,

$$P_i = P_{\chi^2} \left[> \sum_{\ell_1 \in \Omega_{\ell_1}} \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right)^2 \right], \quad (17)$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution where the argument is larger than x , σ_{ℓ_1} is the standard deviation, and the summation of ℓ_1 is taken over a set of ℓ_1 , Ω_{ℓ_1} , which is a set of ℓ_1 s selected as having larger $|G(\ell_1 \ell_2 \ell_3)|$ with fixed ℓ_2 and ℓ_3 . P_i s are corrected via BH criterion and i s associated with adjusted P -values less than 0.01 is selected.

5.3 Gene expression profiles

As denoted in the introduction, we specifically focus AD in this study. All gene expression profiles included in these six studies listed below were selected in GEO. Individual profiles were normalized so as to have zero mean and a standard deviation of one within individual profiles. This means that when gene expression profiles were formatted as a matrix $x_{ij_k} \in \mathbb{R}^{N \times M_k}$ representing expression of the i th gene at the j_k th sample,

$$\sum_{i=1}^N x_{ij_k} = 0, \quad (18)$$

$$\sum_{i=1}^N x_{ij_k}^2 = N, \quad (19)$$

and when gene expression profiles are formatted as a tensor, $x_{ij_k^{[1]} \dots j_k^{[S]}} \in \mathbb{R}^{N \times M_k^{[1]} \times \dots \times M_k^{[S]}}$, representing the expression of the i th gene at the sample annotated by indices $j_1, \dots, j_S$,

$$\sum_{i=1}^N x_{ij_k^{[1]} \dots j_k^{[S]}} = 0, \quad (20)$$

$$\sum_{i=1}^N x_{ij_k^{[1]} \dots j_k^{[S]}}^2 = N. \quad (21)$$

5.3.1 Data set 1: GSE160224

The dataset [26], which is denoted as dataset 1 (i.e., $k = 1$) in this study, was composed of as few as nine samples including the nondemented controls, three APP duplications and three isogenically-corrected iPSC lines. In this study, three control samples are treated as controls and the other six samples are treated as treated samples when the coincidence with singular value vectors or latent variables were investigated and genes whose expression is altered between these two classes are intended to be selected. Gene expression profiles were measured by RNA-seq technology and gene are annotated by Ensembl gene id. The number of genes whose expression was measured is as many as 58302, in contrast to the small number of samples. As a result, gene expression profiles were formatted as a matrix $x_{ij_1} \in \mathbb{R}^{58302 \times 9}$ representing expression of i th gene of the j_1 th sample.

5.3.2 Data set 2: GSE155567

This dataset [27], which is denoted as dataset 2 (i.e., $k = 2$) in this study, was composed of four classes, including THP1 macrophages after knockout of CD33 and/or knockdown (silencing) of PTPN6 (6 samples per class). Thus, in total, it includes as few as 24 samples. In this study, we did not specifically assume which classes were considered controls, but they are regarded as four categorical classes when coincidence with singular value vectors or latent variables are investigated. For unknown reasons, although gene expression profiles of one sample were missing, this was unlikely to have affected the outcome because it still included most (23 out of 24) samples. Gene expression profiles were measured by RNA-seq technology and genes were annotated by Ensembl gene id. The number of genes whose expression was measured was as many as 60617, in contrast to the small number of samples. As a result, gene expression profiles were formatted as a matrix $x_{ij_2} \in \mathbb{R}^{60617 \times 23}$ representing the expression of the i th gene of the j_2 th sample.

5.3.3 Data set 3: GSE162873

This dataset [28], which is denoted as dataset 3 (i.e., $k = 3$) in this study, was composed of eight samples, four of which were AD cell lines and the other four of which were normal cell lines. Because four AD cell lines were composed of two sets of two samples using two distinct cell lines, in this study, this data set is treated as including three categorical classes when coincidence with singular value vectors or latent variables were investigated; two samples were taken from the first AD cell line, two samples were taken from the second AD cell line, and four samples were taken from normal cell lines. Gene expression profiles were measured by RNA-seq technology and genes were annotated by Ensembl gene id. The number of genes whose expression was measured is as many as 47749, in contrast to the small number of samples. As a result, gene expression profiles were formatted as a matrix, $x_{ij_3} \in \mathbb{R}^{47749 \times 8}$ representing the expression of the i th gene of the j_3 th sample.

5.3.4 Data set 4: GSE164788

This dataset [14], which is denoted as dataset 4 (i.e., $k = 4$) in this study, was an *in-vitro* differentiated mixture of neuron and glial cells derived from the ReNcell VM neural progenitor cell line treated with 80 different compounds; mRNA levels were measured using RNA-seq. Among 80 compounds, we selected 94 drugs and drug combinations to which at least two doses were tested. For each dose, three biological replicates were provided (when more than three biological replicates were provided, we randomly selected three of them). Gene expression profiles were measured by RNA-seq technology and genes were annotated by Ensembl gene id. The number of genes whose expression was measured is as many as 28044. As a result, gene expression profiles were formatted as a tensor $x_{ij_4^{[1]}j_4^{[2]}j_4^{[3]}} \in \mathbb{R}^{28044 \times 94 \times 4 \times 3}$ representing the expression of the i th gene at the j_1 th drug combination, the j_2 th dose and j_3 th biological replicates.

5.3.5 Data set 5: GSE164642

This dataset [29], which is denoted as dataset 5 (i.e., $k = 5$) in this study, was composed of three sets of six samples; each set included two classes, each of which corresponded to either ABCC1 activated cells by distinct RNA or control cells, and included three biological replicates (thus 18 total samples). Gene expression profiles were measured by RNA-seq technology and genes were annotated by Ensembl gene id. Thus, they were treated as six categorical classes when coincidence with singular value vectors or latent variables was investigated. The number of genes whose expression was measured was as many as 58003, in contrast to the small number of samples. As a result, gene expression profiles were formatted as a tensor, $x_{ij_5^{[1]}j_5^{[2]}j_5^{[3]}} \in \mathbb{R}^{58003 \times 3 \times 2 \times 3}$ representing the expression of the i th gene at those treated by $j_5^{[1]}$ th RNA, $j_5^{[2]}$ th treatment ($j_5^{[2]} = 1$:control, $j_5^{[2]} = 2$:ABCC1 activated) and $j_5^{[3]}$ th biological replicates.

5.3.6 Data set 6: GSE163577

This was a scRNA-seq dataset [30], which is denoted as dataset 6 (i.e., $k = 6$) in this study, including both AD and healthy controls from 25 hippocampus and superior frontal cortex samples across 17 control and eight Alzheimer’s disease (AD) patients. Individual datasets of 25 datasets were formatted as a matrix, $x_{ij_6^{[c]}} \in \mathbb{R}^{33538 \times M_6^{[c]}}$, $1 \leq c \leq 25$, representing the expression of the i th gene at $j_6^{[c]}$ th cell within the c th scRNA-seq profile. The number of cells, $M_6^{[c]}$, in individual datasets varied and was roughly 10^4 .

5.4 Integrated analysis of gene expression profiles

5.4.1 Integrated analysis of data sets 1, 2 and 3

When we integrate the data sets 1, 2 and 3, we applied SVD to data sets 1, 2 and 3 and we got

$$x_{ij_1} = \sum_{\ell} u_{\ell i}^{[1]} \lambda_{\ell}^{[1]} v_{\ell j_1}^{[1]} \quad (22)$$

$$x_{ij_2} = \sum_{\ell} u_{\ell i}^{[2]} \lambda_{\ell}^{[2]} v_{\ell j_2}^{[2]} \quad (23)$$

$$x_{ij_3} = \sum_{\ell} u_{\ell i}^{[3]} \lambda_{\ell}^{[3]} v_{\ell j_3}^{[3]} \quad (24)$$

Then, we compute x_{ilk} using eq. (11) while setting $L = 8$ because $M_1 = 9, M_2 = 24, M_3 = 8$ and L cannot exceed M_k . Because the number of genes whose expression was measured differed among datasets 1, 2 and 3, we employed $N = 60617$ as the number of genes whose expression is measured in dataset 2 and the largest number of genes measured among datasets 1, 2, and 3. Then these three gene expression profiles are formatted as a tensor

$$x_{ilk} \in \mathbb{R}^{60617 \times 8 \times 3}, \quad (25)$$

where the missing expression in datasets 1 and 3 because of smaller number of genes whose expression was measured than that of dataset 2 is filled with zero. HOSVD was applied to x_{ilk} as in eq. (13).

5.4.2 Drug repositioning using the tensor obtained with data sets 1, 2 and 3

HOSVD was applied to $x_{ij_4^{[1]} j_4^{[2]} j_4^{[3]}} \in \mathbb{R}^{28044 \times 94 \times 4 \times 3}$. Then x_{il4} is computed by eq. (12) with $L_4^{[1]} = 4, L_4^{[2]} = 2$, and $L_4^{[3]} = 1$. The motivations of this choice were as follows. Biological replicates are expected to have common gene expression and the 1st singular value vector $u_{1j_4^{[3]}}$, is expected to have constant value regardless to $j_4^{[3]}$ based upon the previous studies. Thus, it is enough to consider only $u_{1j_4^{[3]}}$ for biological replicates. Then, two choices remain, including $L_4^{[1]} = 4, L_4^{[2]} = 2$ or $L_4^{[1]} = 2, L_4^{[2]} = 4$. Clearly, the former is reasonable, because the number of drug combinations, at 94, was much larger than that of dose density, which was four. Missing values are filled with zero. In addition to $x_{ilk}, 1 \leq k \leq 3$ obtained in the previous subsection, HOSVD was applied to $x_{ilk} \in \mathbb{R}^{60617 \times 8 \times 4}$ as in eq. (13).

5.4.3 Transfer learning using the tensor obtained with data sets 1, 2 and 3

To evaluate the performance of transfer learning, HOSVD was applied to $x_{ij_5^{[1]} j_5^{[2]} j_5^{[3]}} \in \mathbb{R}^{58003 \times 3 \times 2 \times 3}$. Then x_{il5} was computed by eq. (12) with $L_5^{[s]} = 2, 1 \leq s \leq 3$. HOSVD was applied to $x_{ilk} \in \mathbb{R}^{60617 \times 8 \times 4}$ obtained in the previous subsection as in eq. (13) after x_{il4} was replaced with x_{il5} .

5.4.4 Integrated analysis of scRNA-seq data

SVD was applied to $x_{ij_6^{[c]}}, 1 \leq c \leq 25$ one by one. Then, x_{ilc} was computed by eq. (11) while replacing k with c and $L = 10$. Then, HOSVD was applied to $x_{ilc} \in \mathbb{R}^{33538 \times 10 \times 25}$.

5.5 Methods to be compared

In the following, we tested three methods that assume the latent variables attributed to genes, is , and are also common among datasets 1, 2 and 3.

5.5.1 CMF

To perform CMF, we used $x_{ij_1} \in \mathbb{R}^{60617 \times 9}$, $x_{ij_2} \in \mathbb{R}^{60617 \times 23}$, and $x_{ij_3} \in \mathbb{R}^{60617 \times 8}$ such that they shared same number of genes (missing values were filled with zero) and were normalized to have zero mean and a standard deviation of one, as denoted in the beginning of this section. The structure of assumed modeling is given as

$$x_{ij_k} = \sum_{\ell=1}^L u_{\ell i} u_{\ell j_k}^{[k]} + b_i^{[k]} + b_{j_k}^{[k]} + \epsilon_{ij_k}^{[k]}, \quad (26)$$

where $u_{\ell i} \in \mathbb{R}^{L \times 60617}$, $u_{\ell j_k}^{[k]} \in \mathbb{R}^{L \times M_k}$, $b_i^{[k]} \in \mathbb{R}^{60617}$, $b_{j_k}^{[k]} \in \mathbb{R}^{M_k}$, $\epsilon_{ij_k}^{[k]} \in \mathbb{R}^{60617 \times M_k}$. We also employed the option that x_{ij_k} obeys a Poisson distribution because the negative signed binary distribution that is usually assumed for RNA-seq data set was not provided as an option. Before applying this model to x_{ij_k} s, some constants were added such that they do not take negative values, because Poisson distributions do not accept negative values. $L = 4$ because TD-based unsupervised FE was able to identify a singular value vector correlated with classification in datasets 1, 2 and 3 within the top four.

5.5.2 GFA

The datasets used were the same as those used in the trials using CMF. The structure of assumed model is given as

$$x_{ij_k} = \sum_{\ell=1}^L u_{\ell i} u_{\ell j_k}^{[k]} + \epsilon_{ij_k}^{[k]}. \quad (27)$$

The primary difference from CMF, which employed Bayesian inference, is that GFA does not assume the distribution of x_{ij_k} . $L = 5$ was assumed because this value was employed in the example in the tutorial of GFA and it is larger than the number of singular value vectors that were computed with TD based unsupervised FE and are correlated with classifications in datasets 1, 2 and three.

5.5.3 Simple concatenation

A matrix $x_{ij} \in \mathbb{R}^{60617 \times 40}$, where $40 = \sum_{k=1}^3 M_k$, was generated by concatenating three matrices, x_{ij_1} , x_{ij_2} , and x_{ij_3} , so as to share the row number; i.e.

$$x_{ij} = \begin{cases} x_{ij_1}, & j_1 = j, & 1 \leq j \leq 8 \\ x_{ij_2}, & j_2 = j - 8, & 9 \leq j \leq 32 \\ x_{ij_3}, & j_3 = j - 32, & 33 \leq j \leq 40 \end{cases} \quad (28)$$

Then, SVD was applied to x_{ij} and we obtained

$$x_{ij} = \sum_{\ell} u_{\ell i} \lambda_{\ell} v_{\ell j} \quad (29)$$

$v_{\ell j}$ s for $1 \leq j \leq 9$ were used for the latent variables attributed to nine samples in data set 1, $v_{\ell j}$ s for $10 \leq j \leq 32$ were used for the latent variables attributed to 23 samples in data set 2, and $v_{\ell j}$ s for $33 \leq j \leq 40$ were used for the latent variables attributed to eight samples in data set 3.

References

- [1] Chen Huang, Hailin Tang, Wenling Zhang, Xiaoling She, Qianjin Liao, Xiaoling Li, Minghua Wu, and Guiyuan Li. Integrated analysis of multiple gene expression profiling datasets revealed novel gene signatures and molecular markers in nasopharyngeal carcinoma. *Cancer Epidemiology and Prevention Biomarkers*, 21(1):166–175, 2012.
- [2] Pingzhao Hu, Celia MT Greenwood, and Joseph Beyene. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, 6(1):128, 2005.
- [3] Jung Kyoong Choi, Jong Young Choi, Dae Ghon Kim, Dong Wook Choi, Bu Yeo Kim, Kee Ho Lee, Young Il Yeom, Hyang Sook Yoo, Ook Joon Yoo, and Sangsoo Kim. Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Letters*, 565(1):93–100, 2004.
- [4] Zi-Yi Yang, Xiao-Ying Liu, Jun Shu, Hui Zhang, Yan-Qiong Ren, Zong-Ben Xu, and Yong Liang. Multi-view based integrative analysis of gene expression data for identifying biomarkers. *Scientific Reports*, 9(1), September 2019.
- [5] Y. h. Taguchi. *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing, 2020.
- [6] Y-h. Taguchi. Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing. *PLOS ONE*, 12(8):1–36, 08 2017.
- [7] Y-h. Taguchi. Multiomics data analysis using tensor decomposition based unsupervised feature extraction. In De-Shuang Huang, Vitoantonio Bevilacqua, and Prashan Premaratne, editors, *Intelligent Computing Theories and Application*, pages 565–574, Cham, 2019. Springer International Publishing.
- [8] YH. Taguchi and Turki Turki. Novel feature selection via kernel tensor decomposition for improved multi-omics data analysis. *bioRxiv*, 2021.
- [9] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma’ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 05 2016.
- [10] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, December 2008.
- [11] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 11 2008.
- [12] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 05 2019.
- [13] Y. h. Taguchi. Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. *Scientific Reports*, 7(1), October 2017.

- [14] Steve Rodriguez, Clemens Hug, Petar Todorov, Nienke Moret, Sarah A. Boswell, Kyle Evans, George Zhou, Nathan T. Johnson, Bradley T. Hyman, Peter K. Sorger, Mark W. Albers, and Artem Sokolov. Machine learning identifies candidates for drug repurposing in alzheimer’s disease. *Nature Communications*, 12(1), February 2021.
- [15] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), May 2016.
- [16] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 02 2016.
- [17] Ayman ElAli and Serge Rivest. The role of abcb1 and abca1 in beta-amyloid clearance at the neurovascular unit in alzheimer’s disease. *Frontiers in Physiology*, 4:45, 2013.
- [18] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1), January 2021.
- [19] Prabhakar Chalise and Brooke L. Fridley. Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. *PLOS ONE*, 12(5):1–18, 05 2017.
- [20] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 12 2016.
- [21] Florian Rohart, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, and Kim-Anh Lê Cao. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*, 18(1), February 2017.
- [22] Arto Klami, Guillaume Bouchard, and Abhishek Tripathi. Group-sparse embeddings in collective matrix factorization, 2014.
- [23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.
- [24] Eemeli Leppäaho, Muhammad Ammad ud din, and Samuel Kaski. Gfa: Exploratory analysis of multiple data sources with group factor analysis. *Journal of Machine Learning Research*, 18(39):1–5, 2017.
- [25] Yinlei Hu, Bin Li, Wen Zhang, Nianping Liu, Pengfei Cai, Falai Chen, and Kun Qu. WEDGE: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition. *Briefings in Bioinformatics*, 04 2021. bbab085.
- [26] Tao Ye, Yangyang Duan, Hayley W. S. Tsang, He Xu, Yuewen Chen, Han Cao, Yu Chen, Amy K. Y. Fu, and Nancy Y. Ip. Efficient manipulation of gene dosage in human iPSCs using CRISPR/cas9 nickases. *Communications Biology*, 4(1), February 2021.
- [27] Jannis Wißfeld, Ichiro Nozaki, Mona Mathews, Tamara Raschka, Christian Ebeling, Veit Hornung, Oliver Brüstle, and Harald Neumann. Deletion of Alzheimer’s disease-associated CD33 results in an inflammatory human microglia phenotype. *Glia*, 69(6):1393–1412, February 2021.
- [28] Roy Hanna, Anthony Flamier, Andrea Barabino, and Gilbert Bernier. G-quadruplexes originating from evolutionary conserved l1 elements interfere with neuronal gene expression in Alzheimer’s disease. *Nature Communications*, 12(1), March 2021.

- [29] Wayne M. Jepsen, Matthew De Both, Ashley L. Siniard, Keri Ramsey, Ignazio S. Piras, Marcus Naymik, Adrienne Henderson, and Matthew J. Huentelman. Adenosine triphosphate binding cassette subfamily c member 1 (ABCC1) overexpression reduces APP processing and increases alpha- versus beta-secretase activity, in vitro. *Biology Open*, January 2020.
- [30] Andrew C. Yang, Ryan T. Vest, Fabian Kern, Davis P. Lee, Christina A. Maat, Patricia M. Losada, Michelle B. Chen, Maayan Agam, Nicholas Schaum, Nathalie Khoury, Kruti Calcuttawala, Róbert Pálovics, Andrew Shin, Elizabeth Y. Wang, Jian Luo, David Gate, Julie A. Siegenthaler, M. Windy McNerney, Andreas Keller, and Tony Wyss-Coray. A human brain vascular atlas reveals diverse cell mediators of Alzheimer’s disease risk. *bioRxiv*, 2021.

Figure 1: Schematic figure of the present analyses. At first, datasets 1, 2, and 3 were formatted as a tensor (black solid line); then, they were formatted as a tensor together with dataset 4 for drug repositioning (red solid line), and were also formatted as a tensor together with dataset 5 (blue solid line). Dataset 6 was separately formatted as a tensor (green solid line). HOSVD was applied to the generated tensors (x_{ilk} or x_{ilc}). $u_{l_2 \ell}$ s were used to evaluate coincidence with classification (left part of the figure) for datasets 1, 2, 3, 4, and 5 after being converted to $\tilde{v}^{[k]}$, $1 \leq k \leq 3$ or $\tilde{u}^{[k]}$, $k = 4, 5$. $u_{l_3 c}$ was used for the evaluation of the distinction between AD and controls for dataset 6 (green broken line, left part). Which $u_{l_1 i}$ was used for P -value computation was decided in consideration of $G(l_1 l_2 l_3)$ and $u_{l_2 \ell}$. The selected $u_{l_1 i}$ s were used to attribute P -values to genes and genes were selected using P -values. Enrichment analyses were performed using selected genes by Enricher, DAVID, or g:Profiler (right part of the figure).

Figure 2: Boxplots of $\tilde{v}_{\ell_2 j_k}^{[k]}$ for integrated analysis of datasets 1, 2, and 3: Data set 1 ($k = 1$) (A) $\ell_2 = 1$, (B) $\ell_2 = 3$, Dataset 2 ($k = 2$) (C) $\ell_2 = 1$, (D) $\ell_2 = 2$, (E) $\ell_2 = 3$, Dataset 3 ($k = 3$) (F) $\ell_2 = 1$, (G) $\ell_2 = 2$, (H) $\ell_2 = 4$.

Figure 3: Boxplots of $\tilde{v}_{\ell_2 j_k}^{[k]}$ for drug repositioning: Dataset 1 ($k = 1$) (A) $\ell_2 = 1$, (B) $\ell_2 = 2$, Data set 2 ($k = 2$) (C) $\ell_2 = 1$, (D) $\ell_2 = 2$, (E) $\ell_2 = 3$, (F) $\ell_2 = 4$, Data set 3 ($k = 3$) (G) $\ell_2 = 4$.

Figure 4: Boxplots of $\tilde{v}_{\ell_2, j}^{[k]}$ for TL: Dataset 1 ($k = 1$) (A) $\ell_2 = 4$, Dataset 2 ($k = 2$) (B) $\ell_2 = 3$, (C) $\ell_2 = 4$, (D) $\ell_2 = 5$, Dataset 3 ($k = 3$) (E) $\ell_2 = 3$, (F) $\ell_2 = 5$, Dataset 5 ($k = 5$) (G) $\ell_2 = 1$, (H) $\ell_2 = 2$, (I) $\ell_2 = 4$

Figure 5: Boxplots of singular value vector, u_{6c} , computed by applying HOSVD to x_{i6c} for scRNA-seq analysis. “AD” and “C” correspond to the hippocampus of patients and healthy controls, respectively, “AD_Ctx” and “C_Ctx” represent the cortex of patients and healthy controls, respectively.

Table 1: Coincidence between $\tilde{v}_{\ell_2 j_k}^{[k]}$ and classification

Integration of data set 1, 2 and 3 (Figure 2)

ℓ_2	1	2	3	4
Data set 1	(A) 3.67×10^{-2}	—	(B) 3.70×10^{-2}	—
Data set 2	(C) 1.59×10^{-3}	(D) $< 2.2 \times 10^{-16}$	(E) $< 2.2 \times 10^{-16}$	—
Data set 3	(F) 3.81×10^{-4}	(G) 9.41×10^{-5}	—	(H) 4.16×10^{-5}

Drug repositioning (Figure 3)

ℓ_2	1	2	3	4
Data set 1	(A) 9.10×10^{-4}	(B) 7.03×10^{-3}	—	—
Data set 2	(C) 3.65×10^{-12}	(D) 4.98×10^{-14}	(E) 6.45×10^{-14}	(F) $< 2.2 \times 10^{-16}$
Data set 3	—	—	—	(G) 1.69×10^{-4}

Transfer learning (Figure 4)

ℓ_2	1	2	3	4	5
Data set 1	—	—	—	(A) 3.27×10^{-2}	—
Data set 2	—	—	(B) $< 2.2 \times 10^{-16}$	(C) $< 2.2 \times 10^{-16}$	(D) 2.00×10^{-12}
Data set 3	—	—	(E) 4.54×10^{-4}	—	(F) 2.45×10^{-5}
Data set 5	(G) $< 2.2 \times 10^{-16}$	(H) $< 2.2 \times 10^{-16}$	—	(I) $< 2.2 \times 10^{-16}$	—

Row P -values computed with t test for two classes (data set 1) or categorical regression for more classes (data set 2, 3 and 5). Upper characters in parentheses reveal the correspondence with Figs. 2, 3, and 4.

Table 2: Brain disease related terms within top ranked ten terms in various enrichment analyses

	(A) 565 genes		(B) 544 genes		(C) 660 genes		(D) 147 genes	
Term	rk	adjusted P-values	rk	adjusted P-values	rk	adjusted P-values	rk	adjusted P-values
KEGG 2021 Human, Enrichr								
Parkinson disease	3	9.77×10^{-16}	3	3.49×10^{-19}	3	3.72×10^{-15}	3	1.76×10^{-11}
Pathways of neurodegeneration	4	7.86×10^{-13}	4	1.76×10^{-16}	4	1.39×10^{-11}	8	5.76×10^{-9}
Amyotrophic lateral sclerosis	5	1.53×10^{-12}	6	3.75×10^{-14}	5	4.84×10^{-11}	6	4.97×10^{-9}
Alzheimer disease	7	3.62×10^{-11}	5	1.40×10^{-14}	6	2.58×10^{-10}	10	2.08×10^{-7}
Huntington disease	8	4.75×10^{-9}	8	2.41×10^{-10}	9	2.17×10^{-6}	—	—
Prion disease	9	4.77×10^{-9}	7	1.14×10^{-11}	8	4.51×10^{-8}	—	—
Jensen Diseases, Enrichr								
Frontotemporal dementia	2	2.36×10^{-4}	9	4.14×10^{-3}	6	3.41×10^{-3}	—	—
Dementia	3	2.36×10^{-4}	—	—	—	—	—	—
Neurodegenerative disease	8	1.32×10^{-3}	5	1.27×10^{-3}	2	4.07×10^{-4}	8	3.04×10^{-2}
Periventricular nodular heterotopia	9	1.32×10^{-3}	6	1.27×10^{-3}	10	3.42×10^{-3}	—	—
Cerebrovascular disease	—	—	7	2.65×10^{-3}	9	3.42×10^{-3}	—	—
Pick’s disease	—	—	—	—	7	3.41×10^{-3}	—	—
GAD_DISEASE_CLASS, DAVID								
PSYCH	1	2.03×10^{-9}	1	5.59×10^{-6}	1	7.14×10^{-9}	—	—
NEUROLOGICAL	3	6.77×10^{-5}	4	6.52×10^{-4}	2	3.04×10^{-7}	9	7.95×10^{-4}
AGING	—	—	—	—	3	2.18×10^{-6}	—	—
GAD_DISEASE, DAVID								
cognitive ability	2	2.94×10^{-13}	1	3.74×10^{-14}	2	3.00×10^{-12}	4	6.02×10^{-14}
Alzheimer Disease Ataxia	4	5.86×10^{-11}	3	1.63×10^{-11}	4	2.50×10^{-10}	3	3.01×10^{-14}
Diabetes Mellitus, Type 2								
Disease Obesity Parkinson								
Disease								
Alzheimer’s Disease	5	1.06×10^{-9}	5	5.66×10^{-8}	5	8.44×10^{-8}	7	9.82×10^{-9}
Parkinson’s Disease	6	1.14×10^{-7}	10	1.28×10^{-6}	9	3.74×10^{-6}	10	1.07×10^{-7}
Schizophrenia	8	4.35×10^{-7}	—	—	6	1.46×10^{-7}	—	—
KEGG, g:Profiler								
Parkinson disease	3	4.25×10^{-13}	3	1.06×10^{-14}	3	4.52×10^{-12}	3	1.59×10^{-7}
Amyotrophic lateral sclerosis	4	4.57×10^{-9}	6	2.54×10^{-9}	5	1.19×10^{-7}	7	6.78×10^{-5}
Pathways of neurodegeneration - multiple diseases	5	5.53×10^{-9}	4	5.41×10^{-11}	4	1.19×10^{-7}	9	2.58×10^{-4}
Alzheimer disease	6	2.77×10^{-8}	5	3.25×10^{-10}	6	3.07×10^{-7}	—	—
Prion disease	8	9.64×10^{-7}	7	5.38×10^{-8}	7	1.38×10^{-5}	10	6.29×10^{-4}
Huntington disease	9	1.57×10^{-6}	9	1.23×10^{-6}	9	5.84×10^{-4}	—	—

Terms are listed only when they are ranked upto top ranked ten terms. rk: rank, (A) Integrated analysis of data set 1, 2 and 3. (B) Drug repositioning (C) Transfer learning (D) Simple concatenation to which SVD was applied

Table 3: Top ranked drugs.

$\tilde{\ell}$	
1	"kw2449"* "tofacitinib"* "pazopanib" "zm-447439" "bx-912"
2	"(+)-jq1" "otssp167" "jwe-035"* "nintedanib"* "ponatinib"

Those asterisked are identified as effective drugs in the original study [14].

Table 4: Enricment analysis for scRNA-seq

Human Gene Atlas	adjusted				
Terms	P -values				
PrefrontalCortex	2.18×10^{-17}				
Amygdala	1.10×10^{-10}				
OccipitalLobe	1.12×10^{-5}				
pineal night	3.21×10^{-5}				
Fetalbrain	7.69×10^{-5}				
pineal day	4.27×10^{-4}				
SuperiorCervicalGanglion	7.48×10^{-4}				
Cerebellum	2.91×10^{-3}				
Wholebrain	8.13×10^{-3}				
CerebellumPeduncles	8.13×10^{-3}				
Disease	disease ID	species	GEO ID	sample	adj. P -values
Disease Perturbations from GEO down					
schizophrenia	DOID-5419	human	GSE25673	892	4.39×10^{-7}
Breast Cancer	C0006142		GSE1378	52	4.39×10^{-7}
Alzheimer's disease	DOID-10652		GSE5389	592	1.06×10^{-6}
Bipolar Disorder	C0005586		GSE5389	302	2.29×10^{-6}
schizophrenia	DOID-5419		GSE25673	891	2.81×10^{-6}
Alzheimer's disease	DOID-10652	mouse	GSE36980	520	3.46×10^{-6}
glaucoma associated with systemic syndromes	DOID-1686		GSE26299	488	2.06×10^{-18}
prostate cancer	DOID-10283		GSE26910	603	1.50×10^{-4}
astrocytoma	DOID-3069	human	GSE15824	860	2.01×10^{-4}
Huntington's disease	DOID-12858	mouse	GSE3621	704	2.22×10^{-4}
Disease Perturbations from GEO up					
Spinal Muscular Atrophy	C0026847	mouse	GSE10599	235	6.88×10^{-32}
multiple sclerosis	DOID-2377	human	GSE38010	737	6.72×10^{-7}
schizophrenia	DOID-5419		GSE25673	892	2.46×10^{-6}
multiple sclerosis	DOID-237		GSE38010	738	5.34×10^{-4}
schizophrenia	DOID-5419		GSE25673	891	7.63×10^{-4}
Schizophrenia	C0036341		GSE12649	261	1.72×10^{-4}
schizophrenia	DOID-5419		GSE49036	542	5.31×10^{-3}
Lewy body dementia	DOID-12217		GSE19587	1067	5.84×10^{-3}
polycystic ovary syndrome	DOID-11612		GSE48301	569	9.43×10^{-3}
Parkinson's disease	DOID-14330		GSE7621	941	1.32×10^{-3}

Table 5: Coincidence between latent variables and classification

CMF				
Latent variables	1st	2nd	3rd	4th
Data set 1	—	—	—	—
Data set 2	—	—	—	—
Data set 3	—	—	—	—
GFA				
Latent variables	1st	2nd	3rd	4th
Data set 1	—	—	—	—
Data set 2	1.94×10^{-6}	7.81×10^{-14}	4.15×10^{-9}	$< 2.2 \times 10^{-16}$
Data set 3	—	3.40×10^{-4}	1.26×10^{-2}	3.41×10^{-4}
Simple concatenation				
Latent variables	1st	2nd	3rd	4th
Data set 1	4.86×10^{-2}	—	—	—
Data set 2	6.88×10^{-5}	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
Data set 3	1.55×10^{-3}	8.23×10^{-7}	—	2.74×10^{-3}

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supptables.pdf](#)
- [DataS1.xlsx](#)
- [DataS2.xlsx](#)
- [DataS3.xlsx](#)
- [DataS4.xlsx](#)
- [DataS5.xlsx](#)
- [source.txt](#)
- [sample.csv](#)