www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 24 April 2017
Accepted: 13 September 2017
Published online: 23 October 2017

Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets

Y.-h. Taguchi

Identifying drug target genes in gene expression profiles is not straightforward. Because a drug targets proteins and not mRNAs, the mRNA expression of drug target genes is not always altered. In addition, the interaction between a drug and protein can be context dependent; this means that simple drug incubation experiments on cell lines do not always reflect the real situation during active disease. In this paper, I applied tensor-decomposition-based unsupervised feature extraction to the integrated analysis using a mathematical product of gene expression in various diseases and gene expression in the DrugMatrix dataset, where comprehensive data on gene expression during various drug treatments of rats are reported. I found that this strategy, in a fully unsupervised manner, enables researchers to identify a combined set of genes and compounds that significantly overlap with gene and drug interactions identified in the past. As an example illustrating the usefulness of this strategy in drug discovery experiments, I considered cirrhosis, for which no effective drugs have ever been proposed. The present strategy identified two promising therapeutic-target genes, CYPOR and HNFA4; for their protein products, bezafibrate was identified as a promising candidate drug, supported by *in silico* docking analysis.

In silico drug discovery is an important task because experimental identification and verification of therapeutic compounds are time-consuming and expensive processes. There are two major trends in *in silico* drug discovery: the ligand-based approach[1] and structure-based approach[2]. The former is straightforward: new drug candidates are identified based upon the similarity with known drugs no matter how the similarity is evaluated. Although it is a powerful method, there are some drawbacks: if there are no known drugs for target proteins, then there is no way to find new drug candidates. Even if there are many known drugs for the target protein, it is hopeless to try to find compounds that are effective but without any similarity to known drugs. The second, structure-based approach, can address these weaknesses. It can identify new therapeutic compounds even without the information about known drugs. Of course, there are some drawbacks in this strategy too. If the target protein's structure is not known, it must be predicted prior to the drug discovery process. Even if the target protein's structure is known, because we need to numerically verify the binding affinity between the ligand compound and target protein – this work also requires a large amount of computational resources – structure-based *in silico* drug discovery is still far from easy to implement. In addition, the accuracy of prediction of protein structure and of ligand-binding structure is not very high[3]. Thus, it would be helpful to have an additional or alternative strategy for *in silico* drug discovery.

Recently, an alternative approach was proposed that is aimed at finding drug candidates computationally using gene expression profiles of cell lines treated with compounds[4,5]. This third approach is not straightforward. Firstly, because compounds target not mRNAs but proteins, mRNA expression of drug target proteins is not always

Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan. Correspondence and requests for materials should be addressed to Y.-h.T. (email: tag@granular.com)

affected. Therefore, direct identification of a drug target protein in gene expression data cannot be done. Secondly, gene expression alteration caused by treatment with a compound may be context dependent; in other words, in a cell line, the gene expression change caused by incubation with a compound may differ from that in diseases. To compensate for these difficulties, the gene expression signature strategy was developed[6,7]. In this approach, gene expression alteration profiles caused by treatment of a cell line with various drug candidates are compared with those of known drugs. If the profiles are similar, then new drug candidates are expected to function similarly to known drugs. Although this third strategy is useful, if there are no known drugs for the disease under study, then this approach cannot function, as in the case of ligand-based methods. In this paper, I propose a strategy that can infer drug candidates from profiles of drug treatment-associated gene expression without the information about known compounds for diseases. In this strategy, I employ the tensor decomposition (TD)-based unsupervised feature extraction (FE) approach, which is an extension of the recently proposed principal component analysis (PCA)-based unsupervised FE; PCA-based unsupervised FE has successfully solved various bioinformatic problems[8–28]. PCA-based unsupervised FE has also been often applied to integrated analysis. In their integrated approach, after PCA-based unsupervised FE is applied to an individual dataset separately, genes are selected from each of these datasets. Then, intersections among them are evaluated as gene sets carrying greater confidence. This strategy has been applied to the identification of genes associated with aberrant promoter methylation common among three autoimmune diseases[14], identification of genes commonly affected by two histone deacetylase inhibitors[11], and identification of reliable biomarkers considering microRNA, mRNA expression, and compounds together[20]. The advantage of these integrated analyses using PCA-based unsupervised FE compared with standard integrated analysis is that there is no need for a weight factor, which is necessary to integrate multiple datasets because PCA-based unsupervised FE was applied to an individual dataset separately. On the other hand, this strategy, i.e. integrated analysis by applying PCA-based unsupervised FE to an individual dataset separately also has some drawbacks: when there are no common gene sets, there is no way to proceed with the analysis although there are rarely no common gene sets. To compensate for this weakness without introducing weights, in the newly developed TD-based strategy here, tensors are generated using a mathematical product of a gene expression profile of drug-treated cell lines and of a gene expression profile of a disease. This approach does not require any weights that a linear combination inevitably requires. Then, pairs of compounds and genes are identified, where the gene's mRNA expression alteration is associated with drug-treated cell lines and is simultaneously coincident with such an alteration during disease progression. Drug target genes are further inferred based upon gene expression profiles when single-gene knockout or overexpression experiments are conducted. In this study, the inferred gene–compound interactions were found to significantly overlap with known gene–compound interactions.

Results

TD-based unsupervised FE is applied to a combined tensor.

In this study, to demonstrate the usefulness of the present strategy even when applied to various diseases, I consider application to multiple diseases as follows. In the text that follows, I describe detailed results for all the diseases analysed in this study. For a summary of this section, see Table 1 and Fig. 1.

Combined analysis of human diseases and rat tissues.

Here, I focused on multiple human diseases and rat tissues for the following reason. If the proposed strategy is tested on only a single combination of human diseases and rat tissues, then the probability of success is rather low. Therefore, it is better to test the strategy on as many combinations as possible. On the other hand, if the number of compounds tested on some rat tissue is not large enough, then this tissue is not suitable for the present study because the purpose of this study is to apply TD-based unsupervised FE to drug discovery; a smaller number of tested compounds means fewer opportunities for identification of an effective drug. Using this criterion, I selected five tissues, i.e. heart, brain, born narrow, kidney, and liver tissues, on which 22–315 drugs have been tested.

Next, I selected diseases with which the five tissues are analysed together. During this selection, diseases that are supposed to take place in the corresponding tissues should be chosen; otherwise, correct drug selection is unlikely. For heart tissue, I used heart failure. For brain tissue, I used post-traumatic stress disorder (PTSD). For bone narrow tissue, I selected acute lymphocytic leukaemia (ALL). For kidney tissue, I used diabetes and cancer. For liver tissue, I chose cirrhosis. The reason why I used two diseases with kidney tissue is to check whether distinct genes are identified for different diseases when I use distinct gene expression profiles.

Heart failure.

In gene expression profiles of the rat left ventricle (LV) treated with 218 drugs, I selected four time points (1/4, 1, 3, and 5 days after treatment); this is because there were substantially smaller numbers of compounds tested for other time points than for these four. On the other hand, human heart gene expression profiles represent 82 patients with idiopathic dilated cardiomyopathy, 95 patients with ischemic stroke, and 136 healthy controls. In these profiles, 3937 genes sharing gene symbols between humans and rats were considered. Then the generated tensor is

$$\tilde{x}_{j_1 j_2 j_3 i} = x_{j_1 j_2 i} \cdot x_{j_3 i}, \tag{1}$$

where $\tilde{x}_{j_1 j_2 j_3 i} \in \mathbb{R}^{218\times 4\times 313\times 3937}$, which represents a mathematical product of gene expression of the i th gene of LV treated with the j_1 th compound at the j_2 th time point after the drug treatment, $x_{j_1 j_2 i} \in \mathbb{R}^{218\times 4\times 3937}$, and gene expression of the j_3 th human heart, $x_{j_3 i} \in \mathbb{R}^{313\times 3937}$, respectively. Higher-order singular value decomposition (HOSVD) was applied to $\tilde{x}_{j_1 j_2 j_3 i}$, and then core tensor $G(\ell_1 \ell_2 \ell_3 \ell_4) \in \mathbb{R}^{218\times 4\times 313\times 3937}$, compound singular value matrix $u_{\ell_1 j_1} \in \mathbb{R}^{218\times 218}$, time point singular value matrix $u_{\ell_2 j_2} \in \mathbb{R}^{4\times 4}$, human sample singular value matrix $u_{\ell_3 j_3} \in \mathbb{R}^{313\times 313}$, and gene singular value matrix $u_{\ell_4 i} \in \mathbb{R}^{3937\times 3937}$ were obtained.

	tensors				
diseases	DrugMatrix	disease	generated	core tensor	singular value vectors
heart failure	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_4}$	$x_{j_3 i} \in \mathbb{R}^{N_3 \times N_4}$	$x_{j_1 j_2 j_3 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4}$	$G(\ell_1 \ell_2 \ell_3 \ell_4)$	$u_{\ell_k j_k}, k \leq 3, u_{\ell_4, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4) = (218, 4, 313, 3937)$
Selected				ℓ_1=2; ℓ_2=2; ℓ_3=2,3; ℓ_4=21,25,27,28,33,36,37,41,42,48	
PTSD rat model	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_6}$	$x_{j_3 j_k i}$, k=4,5 $\in \mathbb{R}^{N_3 \times N_k \times N_6}$	$x_{j_1 j_2 j_3 j_4 j_5 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4 \times N_5 \times N_6}$	$G(\ell_1 \ell_2 \ell_3 \ell_4 \ell_5 \ell_6)$	$u_{\ell_k j_k}, k \leq 5, u_{\ell_6, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4, N_5, N_6) = (22, 4, 2, 15, 15, 7501)$
Selected				ℓ_1=2; ℓ_2=2; ℓ_3=1; $\ell_4 = \ell_5$=3; ℓ_6=75,77,81,83,84,85,89,90,102	
ALL	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_5}$	$x_{j_3 j_4 i} \in \mathbb{R}^{N_3 \times N_4 \times N_5}$	$x_{j_1 j_2 j_3 j_4 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4 \times N_5}$	$G(\ell_1 \ell_2 \ell_3 \ell_4 \ell_5)$	$u_{\ell_k j_k}, k \leq 4, u_{\ell_5, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4, N_5) = (77, 4, 4, 74, 2597)$
Selected				ℓ_1=2,3,5,6,9,10;ℓ_2=3; ℓ_3=4; ℓ_5=1,2,3,5	
diabetes	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_4}$	$x_{j_3 i} \in \mathbb{R}^{N_3 \times N_4}$	$x_{j_1 j_2 j_3 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4}$	$G(\ell_1 \ell_2 \ell_3 \ell_4)$	$u_{\ell_k j_k}, k \leq 3, u_{\ell_4, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4) = (253, 4, 69, 3489)$
Selected				ℓ_1=2; ℓ_2=2; ℓ_3=1,4; ℓ_4=1,4	
renal carcinoma	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_4}$	$x_{j_3 i} \in \mathbb{R}^{N_3 \times N_4}$	$x_{j_1 j_2 j_3 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4}$	$G(\ell_1 \ell_2 \ell_3 \ell_4)$	$u_{\ell_k j_k}, k \leq 3, u_{\ell_4, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4) = (253, 4, 202, 4036)$
Selected				ℓ_1=2; ℓ_2=2; ℓ_3=13,15,30,33,35; ℓ_4=186,215,233,244,251,269,274,309,312,318	
cirrhosis	$\tilde{x}_{j_1 j_2 i} \in \mathbb{R}^{N_1 \times N_2 \times N_4}$	$x_{j_3 i} \in \mathbb{R}^{N_3 \times N_4}$	$x_{j_1 j_2 j_3 i}$ $\in \mathbb{R}^{N_1 \times N_2 \times N_3 \times N_4}$	$G(\ell_1 \ell_2 \ell_3 \ell_4)$	$u_{\ell_k j_k}, k \leq 3, u_{\ell_4, i} \in \mathbb{R}^{N_k \times N_k}$ $(N_1, N_2, N_3, N_4) = (355, 4, 216, 3961)$
Selected				ℓ_1=2; ℓ_2=2; ℓ_3=2,6;2 $\leq \ell_4 \leq$ 10	

Table 1. A summary of TDs and identification of various singular value vectors for identification of candidate drugs and genes used to find genes encoding drug target proteins. In all cases, ℓ_1 stands for singular value vectors of compounds, whereas ℓ_k with the last (largest) k denotes gene singular value vectors. k stands for singular value vectors of time points in DrugMatrix data. The remaining singular value vectors correspond to sample singular value vectors dependent on the properties of gene expression profiles of diseases. See also Fig. 1 for the corresponding data.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2 j_2}$, represents time dependence (Fig. 2). Then, I decided to use the second singular value vectors ($\ell_2 = 2$) for heart failure.

We also need to know which sample singular value vectors generated from gene expression of disease samples are associated with the difference between healthy controls and diseases. These data must be evaluated on a case-to-case basis because gene expression profiles of disease samples do not have any unified format.

As for heart failure, samples are composed of two disease classes and controls. Because it is obvious that the second and third singular value vectors of samples, $u_{\ell_3 \in (2,3), j_3}$, show a difference between patients and healthy controls (Fig. 3A), I decided to employ these two for the selection of singular value vectors for identification of genes and drugs in the analysis below. The associated *P*-values are also small enough even after Benjamini–Hochberg (BH) criterion corrections.

The next task is to identify singular value vectors of genes and drugs to be used for identification of genes and drugs. This is because I could find combinations of singular value vectors of time points and samples that have the required properties, i.e. an association with time dependence and a difference between diseased and normal samples, as suggested in Fig. 4. These selection procedures are based on the association with core tensors, $G(\ell_1, \ldots, \ell_m)$, that have greater absolute values and are associated with sample and time point singular value vectors identified above. At the moment, we do not have any specific criteria for how to handle large $G(\ell_1, \ldots, \ell_m)$ s. After analysing the top 20 $G(\ell_1, \ldots, \ell_m)$ s with larger absolute values, I realised the following. Gene singular value vectors associated with the top 20 $G(\ell_1, \ldots, \ell_m)$ s generally vary from one $G(\ell_1, \ldots, \ell_m)$ to another $G(\ell_1, \ldots, \ell_m)$. In other words, each $G(\ell_1, \ldots, \ell_m)$ was associated with distinct singular value vectors of genes. Thus, tentatively, I decided to use the top 10 $G(\ell_1, \ldots, \ell_m)$ s and gene singular value vectors associated with them. Because the selected genes based on these singular value vectors are not assumed to directly encode drug target proteins but are simply used for identifying genes encoding drug target proteins (see the region within the rectangle with rounded corners in Fig. 4), I presumed that these specific criteria do not drastically affect identification of genes encoding drug target proteins. In contrast, compound singular value vectors are more restricted. In other words, $G(\ell_1, \ldots, \ell_m)$ s are associated with common compound singular value vectors. Thus, the selection of compound singular value vectors used for drug selection is more robust.

As for heart failure, $G(\ell_1 \ell_2 \ell_3 \ell_4)$, $l_2 = 2, 2 \leq l_3 \leq 3$ were studied. Because absolute values of $G(\ell_1, \ldots, \ell_4)$ s gradually decrease as the rank of $G(\ell_1, \ldots, \ell_4)$ s decreases, there are no clear threshold values of $G(\ell_1, \ldots, \ell_4)$ above which $G(\ell_1, \ldots, \ell_4)$ s are used. Therefore, I selected the top 10 singular value vectors of genes, $\ell_4 = 21, 25, 27, 28, 33, 36, 37, 38, 41, 42$, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later identified by the enrichment analysis of these genes in terms of the

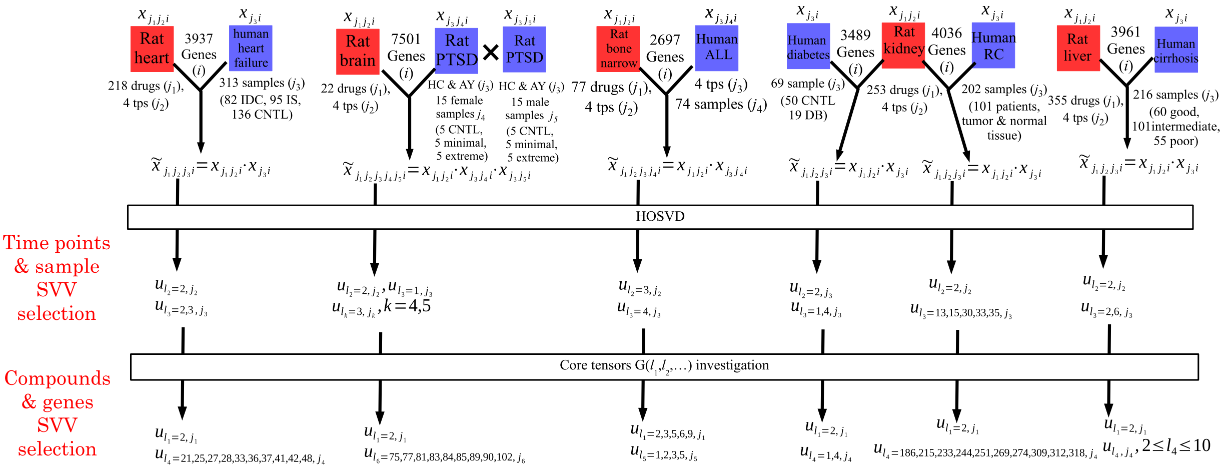
$x_{j_1j_2i}$
Rat heart
3937 Genes (i)
x_{j_3i}
human heart failure
218 drugs (j_1), 4 tps (j_2)
313 samples (j_3) (82 IDC, 95 IS, 136 CNTL)
$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i}$
$x_{j_1j_2i}$
Rat brain
7501 Genes (i)
$x_{j_3j_4i}$
Rat PTSD
×
$x_{j_3j_5i}$
Rat PTSD
22 drugs (j_1), 4 tps (j_2)
HC & AY (j_3) 15 female samples j_4 (5 CNTL, 5 minimal, 5 extreme)
HC & AY (j_3) 15 male samples j_5 (5 CNTL, 5 minimal, 5 extreme)
$\tilde{x}_{j_1j_2j_3j_4j_5i} = x_{j_1j_2i} \cdot x_{j_3j_4i} \cdot x_{j_3j_5i}$
$x_{j_1j_2i}$
Rat bone narrow
2697 Genes (i)
$x_{j_3j_4i}$
Human ALL
77 drugs (j_1), 4 tps (j_2)
4 tps (j_3) 74 samples (j_4)
$\tilde{x}_{j_1j_2j_3j_4i} = x_{j_1j_2i} \cdot x_{j_3j_4i}$
x_{j_3i}
Human diabetes
3489 Genes (i)
69 sample (j_3) (50 CNTL 19 DB)
$x_{j_1j_2i}$
Rat kidney
253 drugs (j_1), 4 tps (j_2)
4036 Genes (i)
x_{j_3i}
Human RC
202 samples (j_3) (101 patients, tumor & normal tissue)
$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i}$
$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i}$
$x_{j_1j_2i}$
Rat liver
3961 Genes (i)
x_{j_3i}
Human cirrhosis
355 drugs (j_1), 4 tps (j_2)
216 samples (j_3) (60 good, 101intermediate, 55 poor)
$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i}$
HOSVD
Time points & sample SVV selection
$u_{l_2=2,j_2}$ $u_{l_3=2,3,j_3}$
$u_{l_2=2,j_2}, u_{l_3=1,j_3}$ $u_{l_k=3,j_k}, k=4,5$
$u_{l_2=3,j_2}$ $u_{l_3=4,j_3}$
$u_{l_2=2,j_3}$ $u_{l_3=1,4,j_3}$
$u_{l_2=2,j_2}$ $u_{l_3=13,15,30,33,35,j_3}$
$u_{l_2=2,j_2}$ $u_{l_3=2,6,j_3}$
Core tensors G(l_1,l_2,...) investigation
Compounds & genes SVV selection
$u_{l_1=2,j_1}$ $u_{l_4=21,25,27,28,33,36,37,41,42,48,j_4}$
$u_{l_1=2,j_1}$ $u_{l_6=75,77,81,83,84,85,89,90,102,j_6}$
$u_{l_1=2,3,5,6,9,j_1}$ $u_{l_5=1,2,3,5,j_5}$
$u_{l_1=2,j_1}$ $u_{l_4=1,4,j_4}$
$u_{l_1=2,j_1}$ $u_{l_4=186,215,233,244,251,269,274,309,312,318,j_4}$
$u_{l_1=2,j_1}$ $u_{l_4,j_4}, 2 \leq l_4 \leq 10$

Figure 1. Schematics that illustrate the procedure of TD-based unsupervised FE applied to the various disease and DrugMatrix datasets. SVV: singular value vector.

genes associated with altered gene expression when genes encoding drug target proteins are perturbed. Nevertheless, because only the second singular value vector of compounds is associated with the top 20 $G(\ell_1, \ldots, \ell_4)$ s, I decided to employ the second singular value vectors of compounds ($\ell_1 = 2$) for use in drug identification later.

PTSD. In gene expression profiles of the rat brain treated with 22 drugs, I selected four time points (1/4, 1, 3, and 5 days after treatment); this is because there were substantially smaller numbers of compounds tested for other time points. Rat hippocampus and amygdala gene expression profiles of 15 males and 15 females are composed of five control samples, five minimal behavioural response samples, and five extreme behavioural response samples, respectively. In these profiles, 7501 genes sharing gene symbols between the two experiments were considered. Then, the generated tensor is

$$\tilde{x}_{j_1j_2j_3j_4j_5i} = x_{j_1j_2i} \cdot x_{j_3j_4i} \cdot x_{j_3j_5i}, \tag{2}$$

where $\tilde{x}_{j_1j_2j_3j_4j_5i} \in \mathbb{R}^{22\times4\times2\times15\times7501}$, which represents the mathematical product of gene expression of the i th gene of a brain treated with the j_1 th compound at the j_2 th time point after the drug treatment, $x_{j_1j_2i} \in \mathbb{R}^{22\times4\times7501}$, gene expression of the j_3 th brain region (hippocampus: $j_3 = 1$, and amygdala: $j_3 = 2$) of the j_4 th female sample and j_5 th male sample (control samples: $1 \leq j_4, j_5 \leq 5$, minimal behavioural response samples: $6 \leq j_4, j_5 \leq 10$, and extreme behavioural response samples: $11 \leq j_4, j_5 \leq 15$), $x_{j_3j_4i} \in \mathbb{R}^{2\times15\times7501}$ and $x_{j_3j_5i} \in \mathbb{R}^{2\times15\times7501}$, respectively. HOSVD was applied to $\tilde{x}_{j_1j_2j_3j_4j_5i}$, and then core tensor $G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6) \in \mathbb{R}^{22\times4\times2\times15\times7501}$, compound singular value matrix $u_{\ell_1j_1} \in \mathbb{R}^{22\times22}$, time point singular value matrix $u_{\ell_2j_2} \in \mathbb{R}^{4\times4}$, singular value matrix of a rat brain region $u_{\ell_3j_3} \in \mathbb{R}^{2\times2}$, rat sample singular value matrices $u_{\ell_4j_4} \in \mathbb{R}^{15\times15}$ and $u_{\ell_5j_5} \in \mathbb{R}^{15\times15}$, and gene singular value matrix $u_{\ell_6i} \in \mathbb{R}^{7501\times7501}$ were obtained.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2j_2}$, represents time dependence (Fig. 2). Then, I decided to use the second singular value vectors ($\ell_2 = 2$) for PTSD.

In relation to the rat model of PTSD, the set of samples is composed of 30 female and male samples, and each of these groups represents 15 hippocampus samples and 15 amygdala samples. Fifteen rats were subdivided into five controls, five minimal-behavioural-response samples, and five extreme-behavioural-response ones. Figure 3B shows the comparisons of sample singular value vectors $u_{\ell_kj_k}(k = 4, 5)$ between 10 male and female controls ($1 \leq j_k \leq 5$), 10 minimal-behavioural-response males and females ($6 \leq j_k \leq 10$), and 10 extreme-behavioural-response males and females ($11 \leq j_k \leq 15$), respectively. Next, the third singular value vector of samples, $u_{\ell_k=3,j_k}$, $k = 4, 5$, turned out to show a significant difference between controls and others. Thus, I decided to employ $u_{\ell_k=3,j_k}$, $k = 4, 5$ to determine which singular value vectors of genes and drugs should be used for selection of genes and drugs in the analysis below. Although the P-value associated with $u_{\ell_k=3,j_k}$, $k = 4, 5$ is relatively greater (weaker significance), this drawback is caused by the smaller number of samples. In actuality, the BH criterion correction does not make the P-value non-significant.

Regarding the PTSD rat model, $G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6)$, $\ell_2 = 2, \ell_4 = \ell_5 = 3$ were investigated. Given that absolute values of $G(\ell_1, \ldots, \ell_6)$ s again gradually decrease as the rank of $G(\ell_1, \ldots, \ell_6)$ s diminishes, there are no clear threshold values of $G(\ell_1, \ldots, \ell_6)$ above which $G(\ell_1, \ldots, \ell_6)$ s are employed. Therefore, I decided to select the top 10 singular value vectors of genes, $\ell_6 = 75,77,81,83,84,85,88,89,90,102$, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later inferred by the enrichment analysis of these

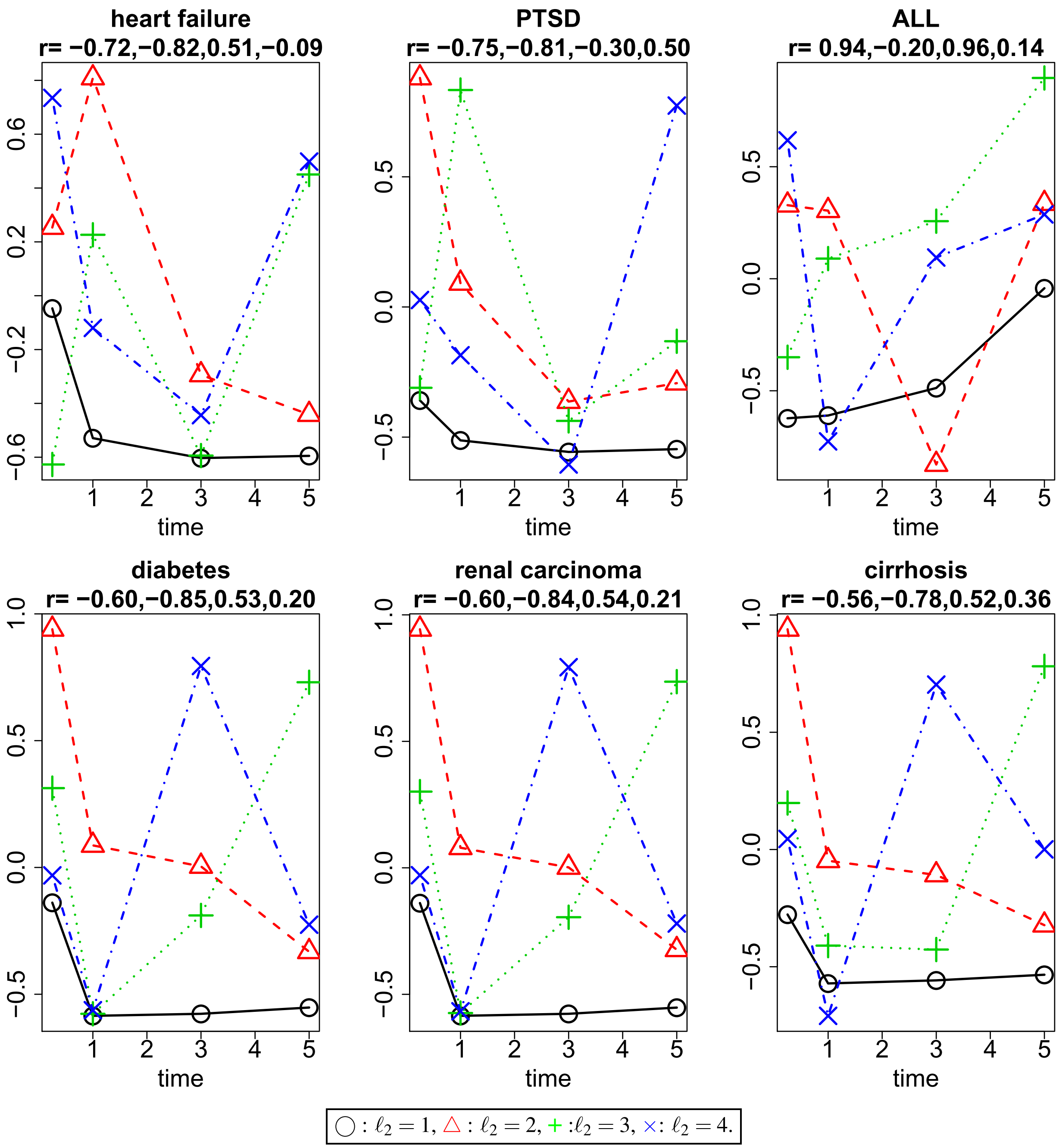

Figure 2. Time point singular value vectors. *r* represents Pearson's correlation coefficients between time points (1/4, 1, 3, and 5 days after a treatment) and the first to forth singular value vectors of time points, $x_{\ell_2 j_2}$, $1 \leq j_2, \ell_2 \leq 4$. Black open circles: $\ell_2 = 1$, red open triangles: $\ell_2 = 2$, green crosses: $\ell_2 = 3$, and blue crosses: $\ell_2 = 4$. $j_2 = 1, 2, 3, 4$ correspond to time points 1/4, 1, 3, and 5 days, respectively.

genes in terms of the genes associated with altered gene expression when genes encoding drug target protein are perturbed. Because only the second singular value vector of compounds ($\ell_1 = 2$) is associated with the top 20 $G(\ell_1, \ldots, \ell_6)$ s, I decided to use the second singular value vectors of compounds for drug identification later. Nonetheless, $\ell_3 = 1$ for all the 10 top-ranked $G(\ell_1, \ldots, \ell_6)$.

ALL. From gene expression profiles of the rat bone narrow treated with 77 drugs, I selected four time points (1/4, 1, 3, and 5 days after treatment); this is because there were substantially smaller numbers of compounds tested for other time points than for these four. On the other hand, gene expression profiles of bone narrow in human (child) ALL represents 74 ALL patients at 0, 8, 15, and 33 days after a remission induction therapy. In these profiles, 2597 genes sharing gene symbols between rats and humans were considered. Then, the generated tensor is

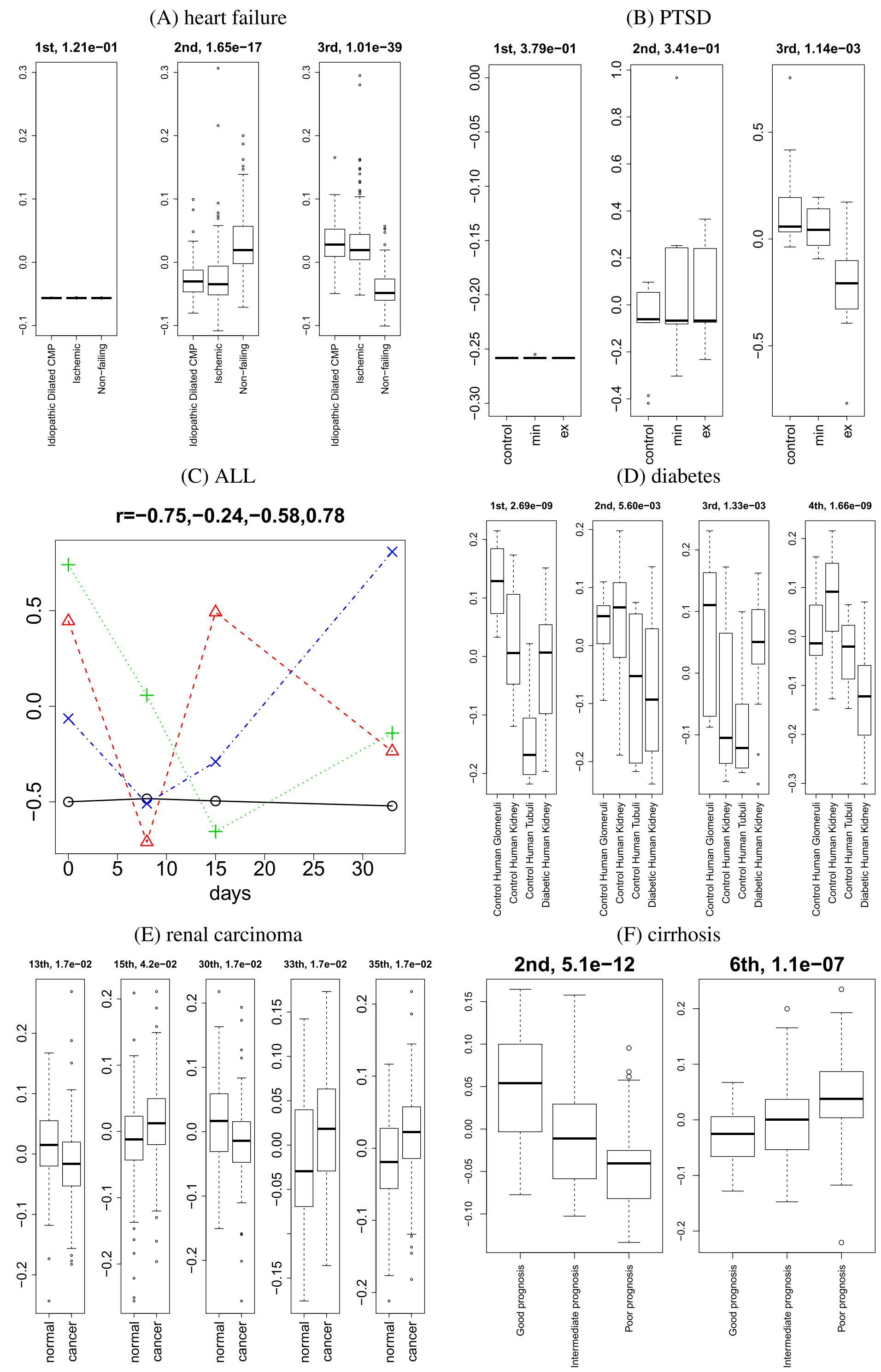

Figure 3. Boxplots [for (**A**), (**B**), (**D**), (**E**), and (**F**)] and time dependence (**C**) for sample singular value vectors. The numbers above boxplots are *P*-values for (**A**), (**B**), and (**D**) and adjusted *P*-vales for (**E**) and (**F**), computed by categorical regression (in other words, ANOVA). The numbers above time dependence (**C**) are correlation coefficients. (**C**) Black open circles: $\ell_3 = 1$, red open triangles: $\ell_3 = 2$, green crosses: $\ell_3 = 3$, and blue crosses: $\ell_3 = 4$. $j_3 = 1, 2, 3, 4$ correspond to time points 0, 8, 15, and 33 days, respectively.

$$\tilde{x}_{j_1j_2j_3j_4i} = x_{j_1j_2i} \cdot x_{j_3j_4i} \qquad (3)$$

where $\tilde{x}_{j_1j_2j_3j_4i} \in \mathbb{R}^{77\times4\times4\times74\times2597}$, which represents the mathematical product of gene expression of the i th gene of rat bone narrow treated with the j_1 th compound at the j_2 th time point after the drug treatment,

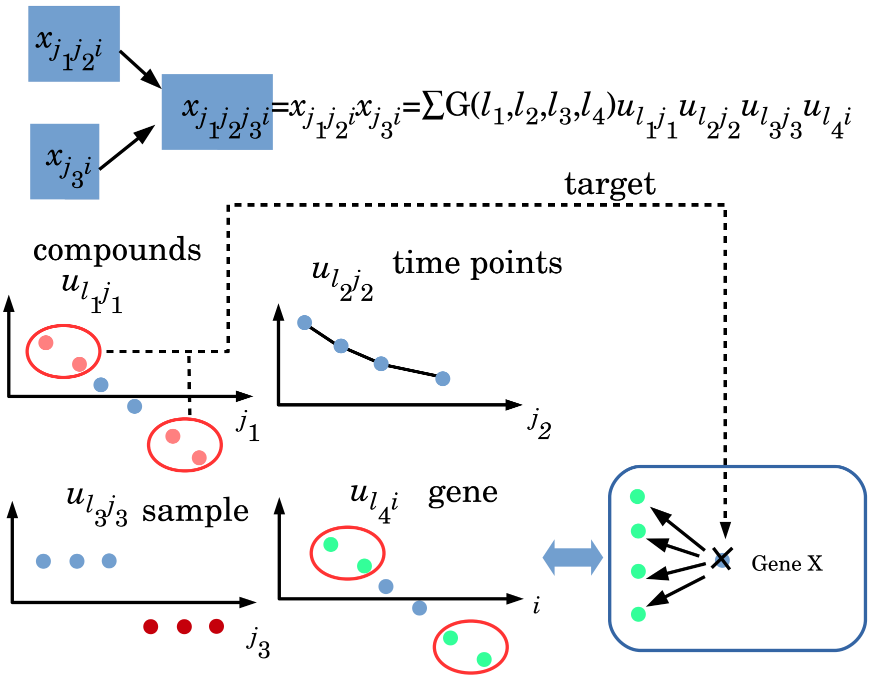

Figure 4. Intuitive illustration of the present strategy. Suppose there is a tensor, x_{j1j2i}, which represents the i th gene expression at the j_2 th time point after the j_1 th compound is given to a rat; these data are taken from the DrugMatrix[44] dataset. There is also a matrix, x_{j3i}, which represents the i th gene expression of the j_3 th sample; samples typically include disease samples and control samples. Tensor $\tilde{x}_{j_1j_2j_3i}$ was generated as a 'mathematical product' of x_{j1j2i} and x_{j3i}. Then, tensor $\tilde{x}_{j_1j_2j_3i}$ is decomposed, and singular value matrix of compounds $u_{\ell_1j_1}$, singular value matrix of time points $u_{\ell_2j_2}$, sample singular value matrix $u_{\ell_3j_3}$, and gene singular value matrix u_{ℓ_4i} are obtained. Among them, I selected the combinations of $\ell_k, 1 \leq k \leq 4$, which are simultaneously associated with all of the following: i) core tensor $G(\ell_1, \ell_2, \ell_3, \ell_4)$ with a large enough absolute value, ii) a singular value vector of time points, $u_{\ell_2j_2}$, whose value significantly varies with time, and iii) sample singular value vector $u_{\ell_3j_3}$. These parameters are different between a disease (red filled circles) and control samples (cyan filled circles). Finally, using gene singular value vector u_{ℓ_4i} and compound singular value vector $u_{\ell_1j_1}$, compounds (filled pink circles) and genes (filled light-green circles) associated with $G(\ell_1, \ell_2, \ell_3, \ell_4)$ s with large enough absolute values are selected. Next, if the selected genes are coincident with the genes associated with a significant alteration when gene X is knocked out (or overexpressed), then the compounds are assumed to target gene X.

$x_{j_1j_2i} \in \mathbb{R}^{77\times4\times2597}$ and ALL bone narrow gene expression on the j_3 th day after the remission induction therapy of the j_4 th sample, $x_{j_3j_4i} \in \mathbb{R}^{4\times74\times2597}$, respectively. HOSVD was applied to $\tilde{x}_{j_1j_2j_3j_4i}$, and then core tensor $G(\ell_1\ell_2\ell_3\ell_4\ell_5) \in \mathbb{R}^{77\times4\times4\times74\times2597}$, compound singular value matrix $u_{\ell_1j_1} \in \mathbb{R}^{77\times77}$, time point singular value matrix $u_{\ell_2j_2} \in \mathbb{R}^{4\times4}$, a singular value matrix of ALL days, $u_{\ell_3j_3} \in \mathbb{R}^{4\times4}$, a singular value matrix of ALL patients, $u_{\ell_4j_4} \in \mathbb{R}^{74\times74}$, and gene singular value matrix $u_{\ell_5i} \in \mathbb{R}^{2597\times2597}$ were obtained.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2j_2}$, represents time dependence (Fig. 2). After that, I decided to use the third vector ($\ell_2 = 3$) for ALL because this vector has the strongest correlations with days.

Regarding ALL, in contrast to the other diseases, the given samples represent not diseases and controls but rather time progression of gene expression after therapy. Then, because the fourth singular value vector of time points (Fig. 3C), $u_{\ell_3=4,j_3}$, has the strongest correlation with time points, I decided to use it for the further analysis in the text below. Although the first one, $u_{\ell_3=1,j_3}$, also shows a moderate correlation, because the gradient is much smaller, I did not select the first vector.

As for ALL, $G(\ell_1\ell_2\ell_3\ell_4\ell_5)$, $\ell_2 = 3$, $\ell_3 = 4$ were studied. Although $G(\ell_1, \ldots, \ell_5)$ with the largest absolute value is exceptionally big, lower-ranked $G(\ell_1, \ldots, \ell_5)$ s are not small enough to be ignored. Therefore, I decided to employ the top 10 $G(\ell_1, \ldots, \ell_5)$ s again. In contrast to the above two cases (heart failure and rat model of PTSD), both gene and compound singular value vectors associated with the top 20 $G(\ell_1, \ldots, \ell_5)$ s vary depending on $G(\ell_1, \ldots, \ell_5)$ s. Thus, I decided to select gene singular value vectors $\ell_5 = 1,2,3,5$ and compound singular value vectors $\ell_1 = 2, 3, 5, 6, 9, 10$, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later inferred by the enrichment analysis of these genes in terms of the genes associated with altered gene expression when genes encoding drug target proteins are perturbed. The reason why the number of vectors selected is less than 10, which is the number of $G(\ell_1, \ldots, \ell_5)$ s being analysed, is that each singular value vector is associated with more than one $G(\ell_1, \ldots, \ell_5)$.

Diabetes. In gene expression profiles of the rat kidney treated with 253 drugs, I selected four time points (1/4, 1, 3, and 5 days after treatment); this is because there were substantially smaller numbers of compounds tested for other time points than for these four. Kidney gene expression of 50 controls (25 whole kidneys, 12 tubuli, and 13 glomeruli) and 19 diabetic kidney samples were considered. In these data, 3489 genes sharing gene symbols between rats and humans were analysed. Then the generated tensor is

$$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i} \tag{4}$$

where $\tilde{x}_{j_1j_2j_3i} \in \mathbb{R}^{253\times4\times69\times3489}$, which represents the products of gene expression of the i th gene of the rat kidney treated with the j_1 th compound at the j_2 th time point after the drug treatment, $x_{j_1j_2i} \in \mathbb{R}^{253\times4\times3489}$ and gene expression of the j_3 th human kidney sample, $x_{j_3i} \in \mathbb{R}^{69\times3489}$, respectively. HOSVD was applied to $\tilde{x}_{j_1j_2j_3i}$, and then core tensor $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{253\times4\times69\times3489}$, compound singular value matrix $u_{\ell_1j_1} \in \mathbb{R}^{253\times253}$, time point singular value matrix $u_{\ell_2j_2} \in \mathbb{R}^{4\times4}$, human sample singular value matrix $u_{\ell_3j_3} \in \mathbb{R}^{69\times69}$, and gene singular value matrix $u_{\ell_4i} \in \mathbb{R}^{3489\times3489}$ were obtained.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2j_2}$, represents time dependence (Fig. 2). Then, I decided to use the second singular value vectors ($\ell_2 = 2$) for diabetes.

As for diabetes, the first four singular value vectors of samples, $u_{\ell_3j_3}$,$1 \le \ell_3 \le 4$, show a significant difference among four classes (Fig. 3D), which are composed of three healthy tissue classes and one diseased kidney class. Nonetheless, because the first and fourth vectors have greater significance, I decided to employ these two.

Regarding diabetes, $G(\ell_1\ell_2\ell_3\ell_4)$, $\ell_2 = 2$, $\ell_3 = 1{,}4$ were investigated. Top two $G(\ell_1, \ldots, \ell_4)$ s with larger absolute values are outstanding. Accordingly, I selected the first and fourth singular value vectors of genes ($\ell_4 = 1{,}4$) and the second vector of compounds ($\ell_1 = 2$) that are associated with the first two top-ranked $G(\ell_1, \ldots, \ell_4)$ s, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later inferred by the enrichment analysis of these genes relative to the genes associated with altered gene expression when genes encoding drug target proteins are perturbed.

Renal carcinoma. Rat kidney samples were also combined with human renal carcinoma samples, which represent 101 human patients (one tumour sample and one sample of adjacent non-tumourous renal tissue from each patient) because some correlation between diabetes and renal carcinoma has been reported[29]. If the present strategy works similarly for both diseases, then these results can strengthen its usefulness. Among genes included in rat kidney and renal carcinoma profiles, 4036 genes sharing gene symbols between rats and humans were considered. Accordingly, the generated tensor is

$$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i} \tag{5}$$

where $\tilde{x}_{j_1j_2j_3i} \in \mathbb{R}^{253\times4\times202\times4036}$, which represents the mathematical product of gene expression of the i th gene of the rat kidney treated with the j_1 th compound at the j_2 th time point after the drug treatment, $x_{j_1j_2i} \in \mathbb{R}^{253\times4\times4036}$, and gene expression of the j_3 th human kidney sample, $x_{j_3i} \in \mathbb{R}^{202\times4036}$, respectively. HOSVD was applied to $\tilde{x}_{j_1j_2j_3i}$, and then core tensor $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{253\times4\times202\times4036}$, compound singular value matrix $u_{\ell_1j_1} \in \mathbb{R}^{253\times253}$, time point singular value matrix $u_{\ell_2j_2} \in \mathbb{R}^{4\times4}$, human sample singular value matrix $u_{\ell_3j_3} \in \mathbb{R}^{202\times202}$, and gene singular value matrix $u_{\ell_4i} \in \mathbb{R}^{4036\times4036}$ were obtained.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2j_2}$, represents time dependence (Fig. 2). Then, I decided to use the second singular value vectors ($\ell_2 = 2$) for renal carcinoma.

Regarding renal carcinoma, the samples represent 101 patients, from whom one tumorous and one healthy tissue sample were taken. In this case, the first several singular value vectors do not show a significant difference between healthy and tumorous tissue samples. Thus, I performed BH corrections on all the *P*-values attributed to 202 singular value vectors of samples, $u_{\ell_3j_3} \in \mathbb{R}^{202\times202}$. Next, I found that five sample singular value vectors, $\ell_3 = 13, 15, 30, 33, 35$ (Fig. 3E), can be used to find singular value vectors of genes and drugs for identification of genes and drugs in the analysis below.

In relation to renal carcinoma, $G(\ell_1\ell_2\ell_3\ell_4)$, $\ell_2 = 2$,$\ell_3 = 13, 15, 30, 33, 35$ were analysed. Although only the second singular value vector of compounds ($\ell_1 = 2$) is associated with 20 top-ranked $G(\ell_1, \ldots, \ell_4)$ s with larger absolute values, various gene singular valued vectors are associated with 20 top-ranked $G(\ell_1, \ldots, \ell_4)$ s with larger absolute values. After that, gene singular value vectors $\ell_4 = 186, 215, 233, 244, 251, 269, 274, 309, 312, 318$ were selected, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later inferred by the enrichment analysis of these genes regarding the genes associated with altered gene expression when genes encoding drug target proteins are perturbed.

Cirrhosis. In gene expression profiles of the rat liver treated with 355 drugs, I selected four time points (1/4, 1, 3, and 5 days after treatment); this is because there were substantially smaller numbers of compounds tested for other time points. Human liver gene expression profiles of patients with cirrhosis are composed of 216 samples (60 correspond to a good prognosis, 101 to an intermediate prognosis, and 55 samples correspond to a poor prognosis). In these profiles, 3961 genes sharing gene symbols between rats and humans were considered. Then the generated tensor is

$$\tilde{x}_{j_1j_2j_3i} = x_{j_1j_2i} \cdot x_{j_3i} \tag{6}$$

where $\tilde{x}_{j_1j_2j_3i} \in \mathbb{R}^{355\times 4\times 216\times 3961}$, which represents the products of gene expression of the i th gene of the rat liver treated with the j_1 th compound at the j_2 th time point after the drug treatment, $x_{j_1j_2i} \in \mathbb{R}^{355\times 4\times 3961}$, and gene expression of the j_3 th human liver sample, $x_{j_3i} \in \mathbb{R}^{216\times 3961}$, respectively. HOSVD was applied to $\tilde{x}_{j_1j_2j_3i}$, and then core tensor $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{355\times 4\times 216\times 3961}$, compound singular value matrix $u_{\ell_1j_1} \in \mathbb{R}^{355\times 355}$, time point singular value matrix $u_{\ell_2j_2} \in \mathbb{R}^{4\times 4}$, human sample singular value matrix $u_{\ell_3j_3} \in \mathbb{R}^{216\times 216}$, and gene singular value matrix $u_{\ell_4i} \in \mathbb{R}^{3961\times 3961}$ were obtained.

Prior to selection of genes and compounds, we need to know which singular value vector of time points, $u_{\ell_2j_2}$, represents time dependence (Fig. 2). Then, I decided to use the second singular value vectors ($\ell_2 = 2$) for cirrhosis.

As for cirrhosis, the second and sixth singular value vectors of samples, $u_{\ell_3\in(2,6),j_3}$, show gradual and significant dependence on progression of diseases (Fig. 3F), from good to intermediate to poor prognosis. Accordingly, I decided to use these two. Because analysis of the sixth may be a bad approach, I performed BH criterion corrections on P-vales and confirmed that these two vectors are still significant.

As for cirrhosis, $G(\ell_1\ell_2\ell_3\ell_4)$, $\ell_2 = 2, \ell_3 = 2, 6$ were investigated. Although only the second singular value vector of compounds (ℓ_1=2) is again associated with 20 top-ranked $G(\ell_1, \ldots, \ell_4)$ s with larger absolute values, various gene singular value vectors are associated with 20 top-ranked $G(\ell_1, \ldots, \ell_4)$ s with larger absolute values. Next, gene singular value vectors $2 \le \ell_4 \le 10$ were selected, which are used for identification of genes with altered expression. Genes encoding drug target proteins are later inferred by the enrichment analysis of these genes regarding the genes associated with altered gene expression when genes encoding drug target proteins are perturbed.

Identification of drugs and drug target proteins.

By means of gene and compound singular value vectors u_{ℓ_ki} and $u_{\ell_1j_1}$, which were selected in the previous subsections and are listed in Table 1 and Fig. 1, drug candidates and genes encoding drug target proteins were identified as suggested in Fig. 4

Firstly, I considered candidate drugs (for the full list, see Table S2). Except for ALL, the second singular value vector of compounds, $u_{\ell_1=2,j_1}$, was identified. Thus, I decided to select outliers with it (Fig. 5). In the previous applications where PCA was used for identification of genes, I employed P-values attributed to each gene, assuming a χ^2 distribution. Unfortunately, this strategy could not be applied to the identification of candidate drugs. When trying to identify critical genes, I also observed gene expression of other, non-critical genes. Consequently, I could identify critical genes using comparisons with them. As for the drugs, however, drugs that are unlikely to affect the rat tissues being considered were not tested. Thus, in a sense, all the analysed drugs are seemingly critical. Therefore, I decided to identify drugs belonging to the cluster farthest from the origin. Because the definition of a cluster may be arbitrary, the selection of genes may not be unique (more detailed criteria and the list of selected drugs are presented in supplementary materials, Text S1).

Regarding identification of genes encoding drug target proteins, I utilised the following strategy as suggested in Fig. 4. Firstly, genes were selected using P-values computed from gene singular value vectors listed in Table 1 (P-values were calculated assuming χ^2 distributions for gene singular value vectors; genes associated with P-values adjusted by the BH criterion and less than 0.01 were selected; for the full list, see Table S2). Then, these genes were uploaded to Enrichr[30] to identify genes encoding drug target proteins via 'Single Gene Perturbations from GEO up' and 'Single Gene Perturbations from GEO down' categories. After that, genes associated with adjusted P-values less than 0.01 were selected as genes encoding drug target proteins (Supporting Data S1).

Biological evaluation of the selected compounds and genes.

At first, it was tested whether known drug target proteins were enriched among those identified by the present strategy. To obtain the list of known drug target proteins, I used DINIES[31]. Although it can also infer unknown interactions between drugs and proteins, by uploading a single drug using a DINIES search with parameters 'chemogenomic approach' and 'with learning on all DBs', I obtained the list of target proteins of individual drugs (Supporting Data S2). Next, target genes were collected for individual drugs, and the collected sets of target genes were generated and used for further analysis in the text below.

To evaluate significant overlaps between the set of target genes predicted by DINIES and those with gene expression profiles, I carried out Fisher's exact test and the uncorrected χ^2 test (Table 2). After that, I found that most of cases (10 out of 12 tables) are associated with significant P-values (ΔG) for both tests. Odds ratios are also generally approximately equal to three, which is usually considered large enough. Therefore, the proposed strategy improves identification of drug target proteins three-fold over random selection; this enhancement is large enough for effectiveness. Thus, it is obvious that this strategy is successful.

One may wonder whether Table 2 supports the usefulness of the present strategy because, generally, most genes identified as drug target genes by the present strategy are false positives although the inference and previous knowledge are significantly related. In this case, however, the situation is a little bit different from usual cases. For example, proteins inferred in the present study but not supported by previous knowledge may simply reflect the lack of experiments. In this case, these apparent false positives may turn out to be a true positive after suitable experimental validation. Alternatively, proteins supported by previous knowledge but not by the present

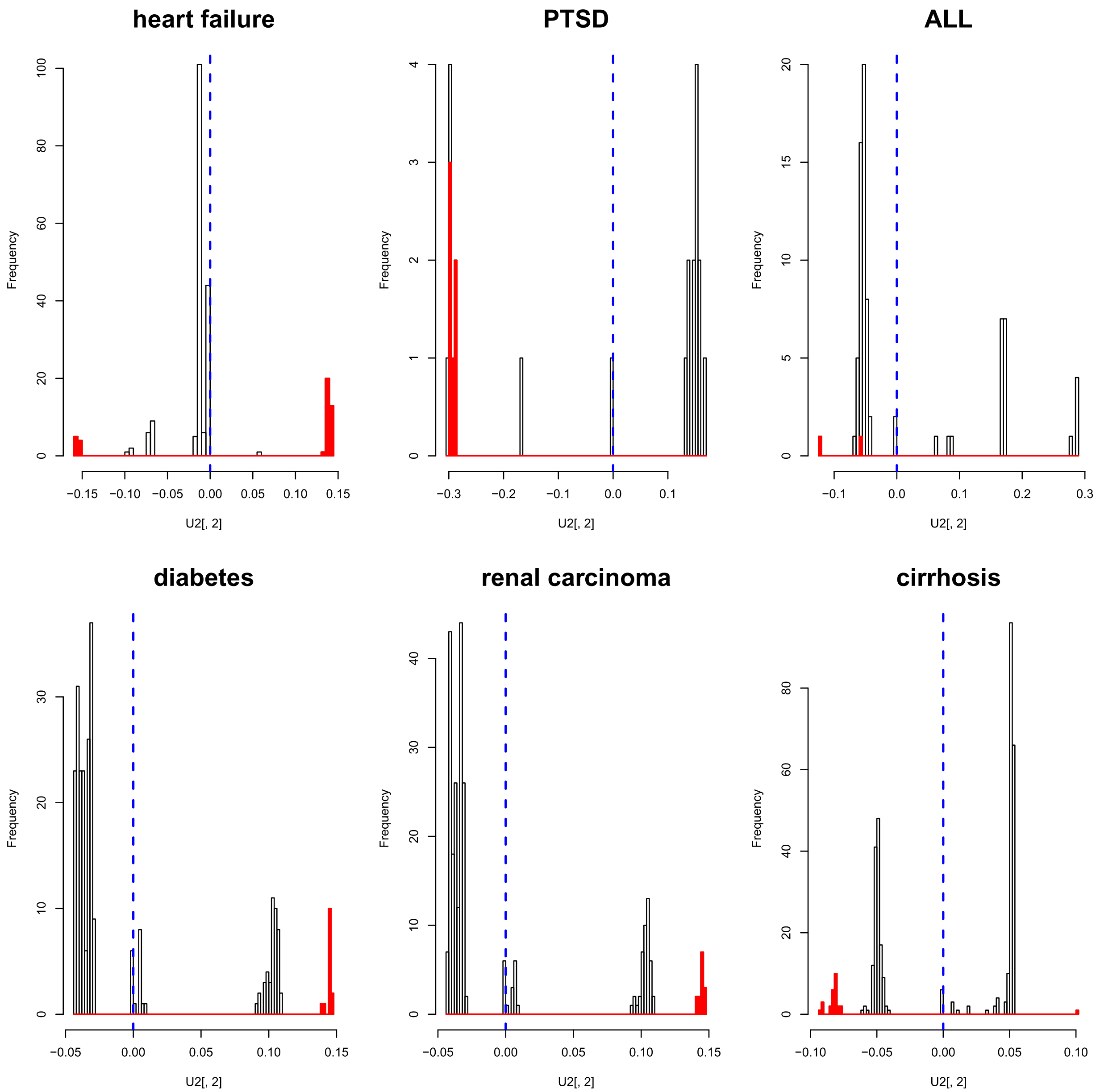

Figure 5. A histogram of the second singular value vectors of compounds. Red parts represent drugs selected as outliers, and vertical blue dashed lines are origins of the axes, by the distances from which the outliers were identified.

study (i.e. false negatives) may simply reflect the lack of perturbation experiments. Proteins not included in perturbation experiments cannot be listed as true positives in the present study. This means that more perturbation experiments and additional validation of protein–drug interactions should be performed to validate the present results more precisely.

Readers may also wonder whether this kind of identification is useful because we are dealing with many-to-many correspondence not one-to-one correspondence. Nevertheless, target gene identification by DINIES was largely found to overlap among multiple compounds. As shown in Fig. 6, a substantial proportion of genes was targeted by more than a single compound, except for ALL where only two compounds were selected. In this sense, the proposed strategy successfully identifies a set of genes that share, to some extent, the compounds targeting these genes.

Discussion

A question may arise whether the present study can be useful for real drug discovery experiments because there are more than a hundred genes listed for individual diseases (Table 2). To demonstrate the usefulness of the proposed strategy, in the text below, I specifically consider two genes, *CYPOR* and *HNF4A*, listed at the top for cirrhosis; CYPOR is top-ranked according to 'Single Gene Perturbations from GEO up', whereas HNF4A is

		Single Gene Perturbations from GEO up					Single Gene Perturbations from GEO down				
		F	T	P_F	$P_{\chi}{}^2$	RO	F	T	P_F	$P_{\chi}{}^2$	RO
heart failure	F	521	517	3.4×10^{-4}	3.9×10^{-4}	3.02	628	416	1.3×10^{-3}	7.3×10^{-4}	2.61
	T	13	39				19	33			
PTSD	F	500	560	3.8×10^{-2}	3.1×10^{-2}	2.67	532	529	6.1×10^{-3}	4.5×10^{-3}	3.81
	T	6	18				5	19			
ALL	F	979	89	2.7×10^{-1}	3.0×10^{-1}	2.19	1009	57	1.0×10^{0}	—	—
	T	10	2				12	0			
diabetes	F	889	177	1.2×10^{-2}	7.1×10^{-3}	3.00	936	130	3.6×10^{-4}	2.0×10^{-5}	5.13
	T	15	9				14	10			
renal carcinoma	F	847	219	2.0×10^{-2}	1.2×10^{-2}	2.75	895	169	4.3×10^{-2}	2.2×10^{-2}	2.64
	T	14	10				16	8			
cirrhosis	F	572	219	1.1×10^{-2}	8.1×10^{-3}	2.91	595	169	1.6×10^{-3}	1.1×10^{-3}	3.81
	T	8	10				7	8			

Table 2. Fisher's exact test (P_F) and the uncorrected χ^2 test ($P_{\chi}{}^2$) of known drug target proteins regarding the inference of the present study. Rows: known drug target proteins (DINIES). Columns: Inferred drug target proteins using 'Single Gene Perturbations from GEO up' or 'Single Gene Perturbations from GEO down'. OR: odds ratio.

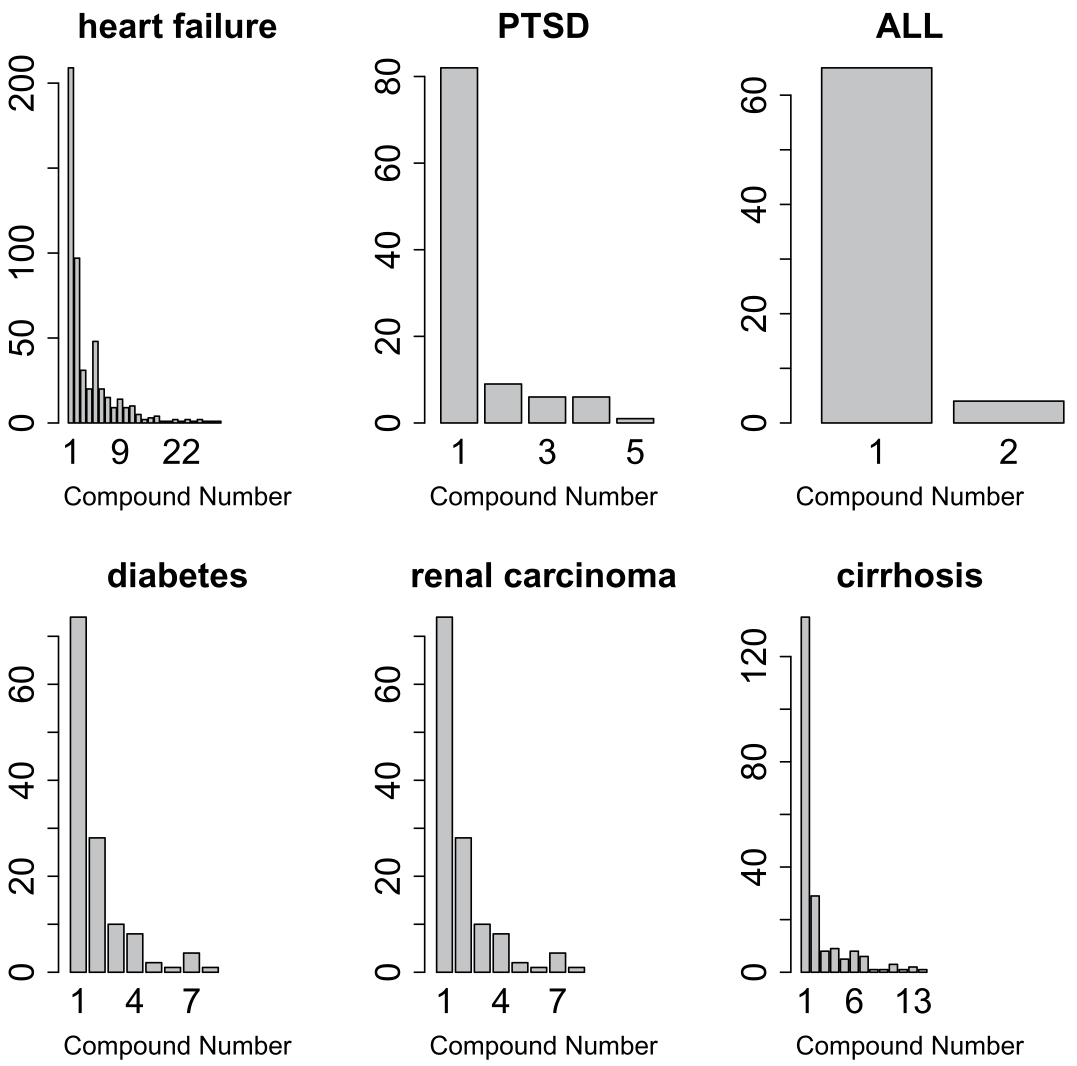

Figure 6. A histogram of the number of genes targeted by multiple compounds according to the results obtained after the identified drugs were uploaded to the DINIES server.

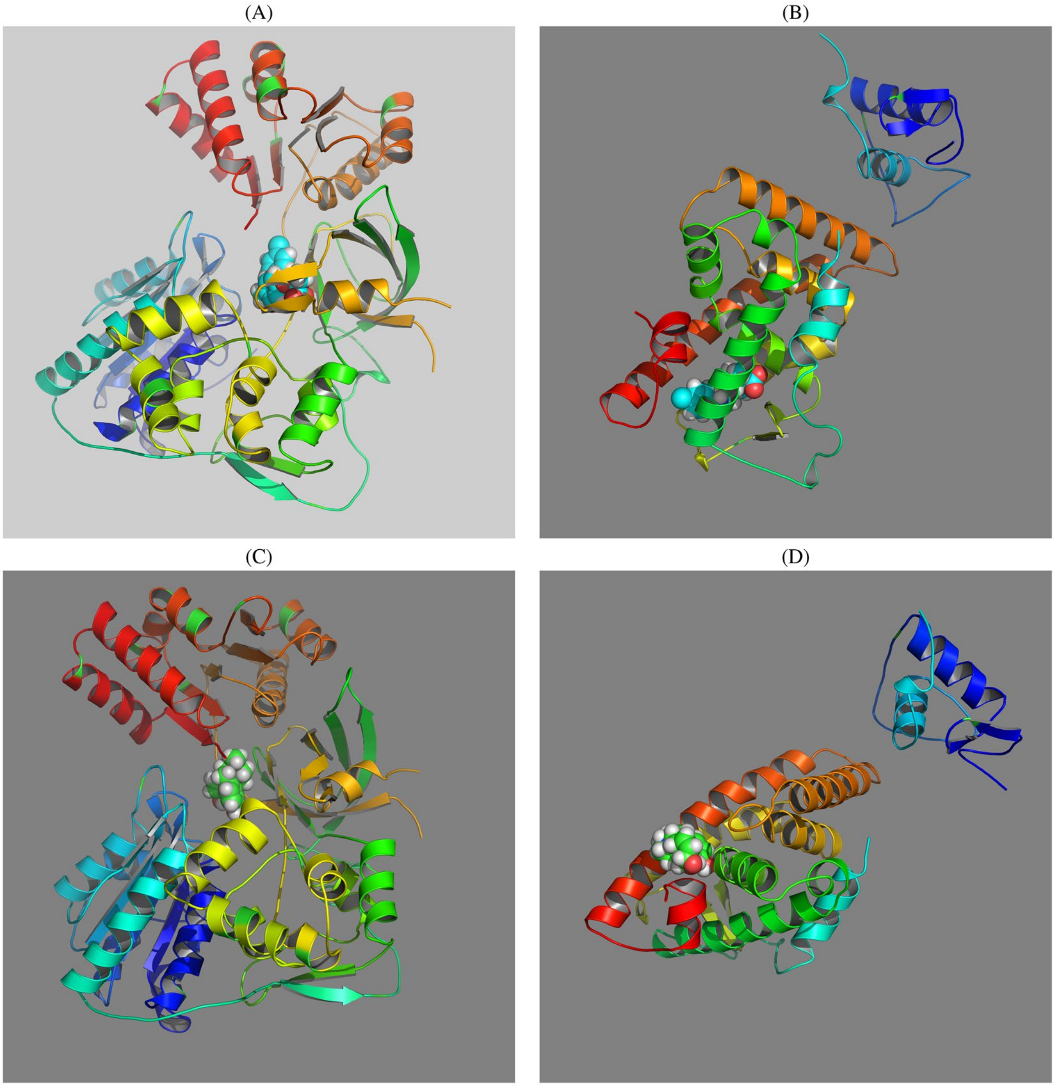

Figure 7. (**A**) Binding structure of bezafibrate toward CYPOR computed by SwissDock. (**B**) Binding structure of bezafibrate toward the ligand-binding domain of HNF4A computed by SwissDock. (**C**) Binding structure of morphine toward CYPOR computed by SwissDock. (**D**) Binding structure of morphine toward the ligand-binding domain of HNF4A computed by SwissDock.

top-ranked based on 'Single Gene Perturbations from GEO down'. At present, there are no drugs directly targeting cirrhosis; thus, the strategy will hold some promise if we can propose some candidate drugs based on this analysis.

Although the CYPOR protein has never been specifically studied in relation to cirrhosis, the activity of cytochrome P450 enzymes, to which CYPOR belongs, is regarded as important in this disease[32]. Accordingly, CYPOR may be an important therapeutic target in cirrhosis. Although no ligand compounds are known for CYPOR, bezafibrate was evaluated as a candidate ligand by *in silico* analysis (Fig. 7A). Its estimated binding energy, ΔG, is equal to 9.73 kcal, which corresponds to $K_i = exp(-(\Delta G)/(RT)) = 79$ nM ($T = 300$ K), which is small enough to consider bezafibrate a useful ligand ($R = 8.31$ is the gas constant). Because bezafibrate itself is a known CYP inhibitor[33], it is plausible that bezafibrate can be a drug for cirrhosis because of effects on CYPOR.

Although HNF4A has long been an orphan receptor, a ligand was identified recently[34]. Among the compounds identified by the present strategy, bezafibrate was estimated to bind to the ligand-binding domain of HNF4A with high affinity ($\Delta G = 9.44$ kcal, which corresponds to 0.13 μM at $T = 300$ K) by *in silico* analysis (Fig. 7B). Given that the structure of the ligand-binding domain is supposed to affect that of the DNA-binding domain[35] (even can enhance the binding), and HNF4A is a transcription factor that can affect cirrhosis[36], bezafibrate may also be a drug for cirrhosis from this point of view.

Although this is only one example and is limited to *in silico* analysis, if also considering that some researchers reported the efficiency of bezafibrate toward cirrhosis[37,38], it indicates that the findings here can be a starting point for identification of new candidate drugs for diseases.

To evaluate the validity of this *in silico* analysis, I tried to assess binding of a possible negative control, morphine, which is unlikely to bind to either CYPOR or HNF4A (Fig. 7C and D). Morphine turned out to have relatively weak binding affinities ($\Delta G = 8.14$ and 8.08 kcal, respectively, both of which correspond to 1.3 μM). This result suggested that our strategy successfully identified drugs that can bind to target proteins with relatively strong binding affinity.

One may wonder whether only ten fold difference (0.13 μM for bezafibrate vs 13 μM for morphine) is significant enough or not. Nonetheless, these values should be evaluated not as the absolute values, but as the relative binding affinity towards proteins. In order to access the general binding affinity towards proteins, I accessed ChEMBL[39] that stores vast number of experimentally validated compounds-protein interactions. Then I have found three records for morphine (CHEMBL409938), 0.32 nM with MOP, 11 nM with OPRD1, and 230 nM with OPRK1. On the other hand, I have found two records for bezafibrate (CHEMBL264374), 33 μM for FABP2 and 44 μM for FABP1. Thus, the best record of morphine in ChEMBL is more than 10^5 times better than that of bezafibrate. Thus, our finding that bezafibrate is ten times better than morphine is definitely remarkable.

Finally, I would like to comment on the related studies. To my knowledge, there are no existing unsupervised methods comparable to the proposed one. The reason is as follows. To perform analyses similar to the present one, in a sense, multi-view data analyses[40] are necessary, because two independent datasets, i.e. DrugMatrix and disease gene expression profiles, must be integrated. There are no existing unsupervised multi-view analytical methods by which more than one feature selection can be performed. For instance, Khan *et al.*[41] proposed multi-tensor decomposition (thus, by definition, an unsupervised method) and applied it to drug discovery; they could identify only compounds and could not identify any drug target genes because by means of their methodology, only features shared across multiple views (in their case compounds) can be screened. Although Li[42] proposed an integrated method that can identify drug–target protein interactions, the method is fully supervised because it cannot identify new drug–target protein interactions without any pre-knowledge.

Before closing the discussion, I would like to comment on the rationality of the present strategy. It is not obvious that applying TD to a mathematical product to obtain singular value vectors is guaranteed success because the relation between the mathematical product and the obtained singular value vectors is unclear. Nevertheless, the same problem had already existed when PCA was applied to gene expression. PCA is essentially matrix factorisation that approximates a matrix as a matrix product. This means that dimensionality of the resulting matrices can never be the same as that of the original gene expression matrix. As readers can see in eq. ((7)), the order of the products of singular value vectors is the same as the order of modes which can take any values. In this sense, dimensionality of the original matrix and that of the obtained matrix after application of TD to the original matrix can never match each other. Any kinds of matrix factorisation or TD simply represent assumptions that may become good approximations. If they work (in the present study, this word indicates that the resulting singular value vectors can be biologically interpretable), this means that the assumption is suitable. If not, it is not reliable. In this study, applying TD to a mathematical product clearly worked well, i.e. the obtained singular value vectors are biologically interpretable. Therefore, the proposed strategy is suitable for the present study and for the dataset being analysed, nothing more. I also would like to emphasize that the present strategy cannot fully replace the compounds or structure based approach that can screen huge number of compounds as many as millions. Until there are more gene expression profiles available, the present strategy must be regarded as only a supportive methodology toward these two main stream strategies.

Methods

Mathematics of TD.

In this subsection, I briefly discuss what TD is and how I apply it to the problem at hand. Suppose an m-mode tensor, X, whose components are denoted as $x_{j_1,\ldots,j_{m-1}i} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_m}$, represents expression of the i th gene under the j_k ($k = 1, \ldots, m-1, j_k = 1, \ldots, N_k$) conditions, examples of which are diseases, patients, tissues, and time points. Then, TD is defined as

$$x_{j_1,\ldots,j_{m-1}i} = \sum_{\ell_1,\ldots,\ell_m}^{N_1,\ldots,N_m} G(\ell_1, \ldots, \ell_m) \cdot u_{\ell_m i} \cdot \prod_{k=1}^{m-1} u_{\ell_k j_k} \tag{7}$$

where $G(\ell_1, \ldots, \ell_m) \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_m}$ is a core tensor and $u_{\ell_m i} \in \mathbb{R}^{N_m \times N_m}$ and $u_{\ell_k j_k} \in \mathbb{R}^{N_k \times N_k}$ are singular value matrices that are supposed to be orthogonal to each other, i.e.

$$\sum_{i=1}^{N_m} u_{\ell_m i} u_{\ell_{m'} i} = \delta_{\ell_m \ell_{m'}} \text{ and } \sum_{j_k=1}^{N_k} u_{\ell_k j_k} u_{\ell_{k'} j_k} = \delta_{\ell_k \ell_{k'}} \tag{8}$$

where $\delta_{\ell_m \ell_{m'}}$ and $\delta_{\ell_k \ell_{k'}}$ are Kronecker's delta. Because the number of $G(\ell_1 \ldots \ell_m)$ is assumed to be as large as that of $x_{j_1 \cdots jm-1i}$, this is obviously an overcomplete problem; therefore, there are no unique solutions. To solve TD uniquely, I specifically employed the HOSVD algorithm[43] that attempts to attain TD such that a smaller number of core tensors and singular value vectors can represent $x_{j_1 \cdots jm-1i}$ as much as possible. All the tensors are standardised such that $\sum_i x_{j_1,\ldots,jm-1i} = 0, \sum_i x_{j_1,\ldots,jm-1i}{}^2 = N_m$ before TD is performed in the present study.

The advantages over more popular TD – parallel factor analysis (PARAFAC)[43] – are as follows. First of all, PARAFAC is NP-complete; in other words, there are no known algorithms that derive PARAFAC with polynomial time. Especially, in the present analysis, singular value vectors associated with a smaller contribution were often useful; ℓ_k is often greater than 100 (see Table 1). Thus, rapid convergence of computation is required, which is not achieved by PARAFAC. Secondly, the advantage of PARAFAC is that it provides one-to-one relations between singular value vectors having common ℓ_k. By contrast, as shown in Table 1, the observed correspondence between singular value vectors is not one to one but often many to many, in which case, HOSVD was more suitable. For these two reasons, I decided to conduct HOSVD instead of more popular PARAFAC.

Tensor generation for integrated analysis. Often, there is a set of gene expression profiles of human cell lines or model animals treated with various compounds at multiple dose densities. DrugMatrix[44] and LINCS[45] are good examples although the former comprises only temporal gene expression after treatment with various drugs. Nonetheless, it is not easy to infer a drug's action on diseases by means of only these gene expression profiles. Some kind of integrated analysis with disease gene expression profiles is necessary, but it is not so straightforward. Candidate drugs should satisfy these conditions:

- Gene expression in these profiles must significantly decrease or increase with the increasing dose density of a compound.
- Gene expression alteration caused by drug treatment must be significantly coincident with that associated with disease progression.

How these two independent significance values can be evaluated is unclear. For example, we can have two sets of significant gene expression alterations of the i th gene, $\{\Delta x_i\}$, caused by drug treatment and those of the i' th gene, $\{\Delta x'_{i'}\}$, during disease progression, respectively. Firstly, we need to test whether the two sets of genes significantly overlap. Next, when there is a significant overlap, we have to determine whether these two gene expression alteration profiles correlate significantly. Furthermore, because the analysis is usually conducted among multiple compounds, all the significance evaluation must be corrected based upon a multiple-comparison criterion. This is obviously a complicated and unpromising strategy.

On the other hand, if we can have gene expression profiles expressed via a tensor, $x_{j_1 \cdots jm-1i}$, where j_k, $k = 1, \ldots, m-1$ corresponds to drug candidates, dose density, and disease progression, we can easily evaluate a candidate drug using TD: eq. ((7)). If there are $u_{\ell_k j_k}$ values that represent significant dependence on dose densities and disease progression, then genes' and compounds' singular value vectors – that share core tensor $G(\ell_1, \ldots, \ell_m)$ having larger absolute values with these $u_{\ell_k j_k}$ s – can be used for the selection of genes as well as compounds as follows.

Suppose $\{\ell_k\}$ is a set of indices of genes' or compounds' singular value vectors that are associated with significant dose density dependence as well as dependence on disease progression. Genes and compounds can be identified that are associated with significant singular value vector components. For this purpose, P-values are attributed to each i th gene and j_k th compound assuming a χ^2 distribution,

$$P_i = P_{\chi^2}\left[> \sum_{\{\ell_m\}} \left(\frac{u_{\ell_m i}}{\sigma_{\ell_m}}\right)^2\right] \quad or \quad P_{j_k} = P_{\chi^2}\left[> \sum_{\{\ell_k\}} \left(\frac{u_{\ell_k j_k}}{\sigma_{\ell_k}}\right)^2\right] \tag{9}$$

where $P_{\chi 2}[>x]$ is the cumulative probability that the argument is greater than x assuming the χ^2 distribution, and σ_{ℓ_m} and σ_{ℓ_k} are standard deviations. After adjustment of P-values via the BH criterion[46], genes and compounds that have significant P-values, i.e. less than 0.01, are selected as those contributing to the specified singular value vectors. Nonetheless, because such a tensor can be obtained only when drug treatment is performed on patients, this strategy is useless; if we can test drug efficiency directly on patients, then there is no need for *in silico* drug discovery. To overcome this obstacle, I replace $x_{j_1 \cdots jm-1i}$ with a 'mathematical product'; hereafter, this means that each component is defined as a product of two components, i.e.

$$\tilde{x}_{j_1 \ldots j_{m-1} i} = x_{j_1 \ldots j_{m'} i} \cdot x_{j_1 \ldots j_{m''} i}, \tag{10}$$

where $x_{j_1 \ldots j_{m'} i} \in \mathbb{R}^{N_1 \times \cdots \times N_{m'}}$ is gene expression for the drug treatment of cell lines or model animals, whereas $x_{j_1 \ldots j_{m''} i} \in \mathbb{R}^{N_1 \times \cdots \times N_{m''}}$ is gene expression for the patients ($m - 1 = m' + m''$). Because these two datasets can be obtained independently, we can test any kind of combinations of drug treatments and diseases even after all measurements were finalised. It is also obvious that this strategy retains the advantages of integrated analysis using PCA-based unsupervised FE because it still does not require assignment of any weights to each gene's expression. On the other hand, it compensates the shortcoming of the integrated analysis of PCA-based unsupervised FE. Because this strategy identifies genes by considering expression of two genes in an integrated manner, there is no need to use common sets among multiple genes selected within an individual dataset as in the PCA based unsupervised FE.

Explanation of the present strategy. Here is an accessible description of the present strategy (Fig. 4). It is obvious that the success of this seemingly complicated method heavily depends upon whether I can easily (or usually, in other words) obtain such preferable combinations of singular value vectors that have the properties illustrated in Fig. 4. Although this situation is unlikely to happen often, it seems to be possible in many cases as shown in the text above.

Gene expression profiles. *Heart failure.* Gene expression profiles of the heart for drug treatments of rats were retrieved from DrugMatrix under the gene expression omnibus (GEO) ID GSE59905, whereas human gene expression for heart failure was retrieved from GEO ID 57345. For both datasets, expression files of genes GSE57345-GPL11532_series_matrix.txt.gz, GSE59905-GPL5426_series_matrix.txt.gz, and GSE59905-GPL5425_series_matrix.txt.gz were directly downloaded from the series matrix.

The rat model of PTSD. Gene expression profiles of the brain for drug treatments of rats were retrieved from DrugMatrix under GEO ID GSE59895, whereas amygdala and hippocampus gene expression for the rat model of

PTSD was taken from GEO ID GSE60304. For both datasets, expression files of genes GSE60304_series_matrix.txt.gz, GSE59895-GPL5425_series_matrix.txt.gz, and GSE59895-GPL5426_series_matrix.txt.gz were directly downloaded from the series matrix.

ALL. Gene expression profiles of bone marrow for drug treatments of rats were retrieved from DrugMatrix under GEO ID GSE59894, and ALL human bone marrow gene expression was taken from GEO ID GSE67684. For both datasets, expression files of genes GSE67684-GPL570_series_matrix.txt.gz, GSE67684-GPL96_series_matrix.txt.gz, GSE59894-GPL5425_series_matrix.txt.gz, and GSE59894-GPL5426_series_matrix.txt.gz were directly downloaded from the series matrix.

Diabetes and renal cancer. Gene expression profiles of kidneys for drug treatments of rats were retrieved from DrugMatrix under GEO ID GSE59913, whereas gene expression for diabetic human kidneys and renal cancer was obtained from GEO ID GSE30122 and GSE40435, respectively. For these datasets, expression files of genes GSE30122_series_matrix.txt.gz, GSE40435_series_matrix.txt.gz, GSE59913-GPL5424_series_matrix.txt.gz, GSE59913-GPL5425_series_matrix.txt.gz, and GSE59913-GPL5426_series_matrix.txt.gz were directly downloaded from the series matrix.

Cirrhosis. Gene expression profiles of the liver for drug treatments of rats were retrieved from DrugMatrix under GEO ID GSE59923, whereas gene expression for cirrhosis of the human liver was obtained from GEO ID GSE15654. For both datasets, GSE15654_series_matrix.txt.gz, GSE59923-GPL5424_series_matrix.txt.gz, GSE59923-GPL5425_series_matrix.txt.gz, and GSE59923-GPL5426_series_matrix.txt.gz were directly downloaded from the series matrix.

An enrichment analysis server. Gene symbols were uploaded to Enrichr[30]. Then, 'Single Gene Perturbations from GEO up' and 'Single Gene Perturbations from GEO down' categories were downloaded.

Statistical analysis. All the statistical analyses were conducted in the R software. HOSVD was carried out using the hosvd function in the rTensor package.

***In silico* evaluation of candidate drugs using SwissDock.** SwissDock[47] was used for the *in silico* evaluation of binding affinity of the identified drugs for a prospective protein. For CYPOR, Protein Data Bank[48] (PDB) ID 1J9Z_A[49] was used to retrieve protein structure, and for HNF4A, PDB ID 4IQR_A[50] was used for this purpose. Compound structures provided by SwissDock when I searched by drug names, bezafibrate or morphine, were analysed. All the default options were chosen, and the binding structures associated with the smallest ΔG were selected.

References

1. Favia, A. D. Theoretical and computational approaches to ligand-based drug discovery. *Front Biosci (Landmark Ed)* **16**, 1276–1290 (2011).
2. Lionta, E. *et al*. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem* **14**, 1923–1938 (2014).
3. Kim, R. & Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J Comput Chem* **29**, 1316–1331 (2008).
4. Liu, C. *et al*. Compound signature detection on LINCS L1000 big data. *Mol Biosyst* **11**, 714–722 (2015).
5. Hizukuri, Y., Sawada, R. & Yamanishi, Y. Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. *BMC Med Genomics* **8**, 82 (2015).
6. Stegmaier, K. *et al*. Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation. *Nat. Genet.* **36**, 257–263 (2004).
7. Lee, B. K. *et al*. DeSigN: connecting gene expression with therapeutics for drug repurposing and development. *BMC Genomics* **18**, 934 (2017).
8. Taguchi, Y. H., Iwadate, M., Umeyama, H. & Murakami, Y. Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis. In *Computational Methods with Applications in Bioinformatics Analysis*, chap. 8, 153–182 (World Scientific, 2017).
9. Taguchi, Y. H. Principal Components Analysis Based Unsupervised Feature Extraction Applied to Gene Expression Analysis of Blood from Dengue Haemorrhagic Fever Patients. *Sci Rep* **7**, 44016 (2017).
10. Taguchi, Y. H. microRNA-mRNA Interaction Identification in Wilms Tumor Using Principal Component Analysis Based Unsupervised Feature Extraction. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, 71–78 (2016).
11. Taguchi, Y.-H. Principal component analysis based unsupervised feature extraction applied to publicly available gene expression profiles provides new insights into the mechanisms of action of histone deacetylase inhibitors. *Neuroepigenetics* **8**, 1–18 (2016).
12. Taguchi, Y. H., Iwadate, M. & Umeyama, H. Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinformatics* **16**, 139 (2015).
13. Taguchi, Y. H. & Okamoto, A. Principal component analysis for bacterial proteomic analysis. In Shibuya, T., Kashima, H., Sese, J. & Ahmad, S. (eds) *Pattern Recognition in Bioinformatics*, vol. 7632 of *LNCS*, 141–152 (Springer International Publishing, Heidelberg, 2012).
14. Ishida, S., Umeyama, H., Iwadate, M. & Taguchi, Y. H. Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery. *Protein Pept. Lett.* **21**, 828–39 (2014).
15. Kinoshita, R., Iwadate, M., Umeyama, H. & Taguchi, Y. H. Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets. *BMC Syst Biol* **8**(Suppl 1), S4 (2014).
16. Taguchi, Y. H. & Murakami, Y. Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. *PLoS ONE* **8**, e66714 (2013).
17. Taguchi, Y. H. & Murakami, Y. Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases? *BMC Res Notes* **7**, 581 (2014).
18. Murakami, Y. *et al*. Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease. *PLoS ONE* **7**, e48366 (2012).
19. Murakami, Y. *et al*. Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray. *PLoS ONE* **9**, e106314 (2014).
20. Murakami, Y. *et al*. Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. *Sci Rep* **5**, 16294 (2015).

21. Umeyama, H., Iwadate, M. & Taguchi, Y. H. TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. *BMC Genomics* **15**(Suppl 9), S2 (2014).
22. Taguchi, Y. H., Iwadate, M. & Umeyama, H. Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, 1–10 (2015).
23. Taguchi, Y. H., Iwadate, M., Umeyama, H., Murakami, Y. & Okamoto, A. Heuristic principal component analysis-aased unsupervised feature extraction and its application to bioinformatics. In Wang, B., Li, R. & Perrizo, W. (eds) *Big Data Analytics in Bioinformatics and Healthcare*, 138–162 (2015).
24. Taguchi, Y. H. Integrative analysis of gene expression and promoter methylation during reprogramming of a non-small-cell lung cancer cell line using principal component analysis-based unsupervised feature extraction. In Huang, D.-S., Han, K. & Gromiha, M. (eds.) *Intelligent Computing in Bioinformatics*, vol. 8590 of *LNCS*, 445–455 (Springer International Publishing, Heidelberg, 2014).
25. Taguchi, Y. H. Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage. *BMC Bioinformatics* **16**(Suppl 18), S16 (2015).
26. Taguchi, Y. H. Identification of More Feasible MicroRNA-mRNA Interactions within Multiple Cancers Using Principal Component Analysis Based Unsupervised Feature Extraction. *Int J Mol Sci* **17**, E696 (2016).
27. Taguchi, Y. H. Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression. *BioData Min* **9**, 22 (2016).
28. Taguchi, Y. H., Iwadate, M. & Umeyama, H. SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer. *BMC Med Genomics* **9**(Suppl 1), 28 (2016).
29. Habib, S. L., Prihoda, T. J., Luna, M. & Werner, S. A. Diabetes and risk of renal cell carcinoma. *J Cancer* **3**, 42–48 (2012).
30. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97 (2016).
31. Yamanishi, Y. *et al.* DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.* **42**, 39–45 (2014).
32. Parkinson, A., Mudra, D. R., Johnson, C., Dwyer, A. & Carroll, K. M. The effects of gender, age, ethnicity, and liver cirrhosis on cytochrome P450 enzyme activity in human liver microsomes and inducibility in cultured human hepatocytes. *Toxicol. Appl. Pharmacol.* **199**, 193–209 (2004).
33. Makia, N. L. & Goldstein, J. A. CYP2C8 Is a Novel Target of Peroxisome Proliferator-Activated Receptor α in Human Liver. *Mol. Pharmacol.* **89**, 154–164 (2016).
34. Yuan, X. *et al.* Identification of an endogenous ligand bound to a native orphan nuclear receptor. *PLoS ONE* **4**, e5609 (2009).
35. Huang, W., Greene, G. L., Ravikumar, K. M. & Yang, S. Cross-talk between the ligand- and DNA-binding domains of estrogen receptor. *Proteins* **81**, 1900–1909 (2013).
36. Ning, B. F. *et al.* Hepatocyte nuclear factor 4 alpha suppresses the development of hepatocellular carcinoma. *Cancer Res.* **70**, 7640–7651 (2010).
37. KANEKO, A. *et al.* Efficacy and problems in bezafibrate treatment for PBC. *Kanzo* **46**, 200–207 (2005).
38. Yin, Q. *et al.* Systematic review and meta-analysis: bezafibrate in patients with primary biliary cirrhosis. *Drug Des Devel Ther* **9**, 5407–5419 (2015).
39. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Research* **42**, D1083–D1090, doi:10.1093 (2013).
40. Li, Y., Wu, F. X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinformatics* (2016).
41. Khan, S. A., Leppäaho, E. & Kaski, S. Bayesian multi-tensor factorization. *Machine Learning* **105**, 233–253 (2016).
42. Li, L. Mpgraph: multi-view penalised graph clustering for predicting drug-target interactions. *IET Systems Biology* **8**, 67–73 (2014).
43. Lathauwer, L. D., Moor, B. D. & Vandewalle, J. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* **21**, 1253–1278 (2000).
44. National Toxicology Program. DrugMatrix. https://ntp.niehs.nih.gov/drugmatrix/index.html (2010).
45. Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–460 (2014).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
47. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **39**, W270–277 (2011).
48. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2017).
49. Hubbard, P. A., Shen, A. L., Paschke, R., Kasper, C. B. & Kim, J. J. NADPH-cytochrome P450 oxidoreductase. Structural basis for hydride and electron transfer. *J. Biol. Chem.* **276**, 29163–29170 (2001).
50. Chandra, V. *et al.* Multidomain integration in the structure of the HNF-4α nuclear receptor complex. *Nature* **495**, 394–398 (2013).

Acknowledgements

This study was supported by KAKENHI 17K00417 and Chuo University grant for special research projects.

Author Contributions

Y.H.T. planned the study, performed the analyses, and wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-13003-0.

Competing Interests: The author declares that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.