Supplementary materials for

Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm

Michio Iwata[1], Longhao Yuan[2,3], Qibin Zhao[3,4], Yasuo Tabei[3], Francois Berenger[1], Ryusuke Sawada[1], Sayaka Akiyoshi[5], and Yoshihiro Yamanishi[1,6,*]

1. Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan. 2. Graduate School of Engineering, Saitama Institute of Technology, 1690 Fusaiji, Fukaya, Saitama 369-0293, Japan. 3. RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan. 4. School of Automation, Guangdong University of Technology, Panyu, Guangzhou, Guangdong, China. 5. Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan. 6. PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan.

*corresponding author: Yoshihiro Yamanishi (yamani@bio.kyutech.ac.jp)

Methods

CP-WOPT algorithm for data completion

We compare the performance of TT-WOPT algorithm with that of CANDECOMP/PARAFAC weighted optimization (CP-WOPT) that analyzes a real-valued tensor, $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, with missing entries [1]. The index of the missing entries can be recorded by a weight tensor ($\underline{W}$), the size of which is same as that of $\underline{X}$. Each entry of $\underline{W}$ satisfies the following conditions:

$$w_{i_1 i_2 \ldots i_N} = \begin{cases} 0 & \text{if } x_{i_1 i_2 \ldots i_N} \text{ is a missing entry,} \\ 1 & \text{if } x_{i_1 i_2 \ldots i_N} \text{ is an observed entry.} \end{cases}$$

CP decomposition decomposes a tensor into a sequence of matrices. The CP decomposition of the tensor $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ can be expressed as follows:

$$\underline{X} = \langle\langle \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, ..., \mathbf{A}^{(N)} \rangle\rangle,$$

where $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, ..., \mathbf{A}^{(N)}$ is a sequence of matrices of size $I_1 \times R$, $I_2 \times R$, ..., $I_N \times R$, respectively. The R is referred to as CP-ranks, which can limit the size of each matrix. Each element of tensor $\underline{X}$ can be written in the following index form:

$$x_{i_1 i_2 \ldots i_N} = \sum_{r=1}^{R} \prod_{n=1}^{N} a_{i_n r}^{(n)},$$

where $a_{i_n r}^{(n)}$ is the the (i_n, r)-th element of the n-th matrix.

In the optimization algorithm, the objective variables are the elements of all matrices. Here, the objective function can be written as follows:

$$f(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, ..., \mathbf{A}^{(N)}) = \frac{1}{2} \|(\underline{Y} - \underline{Z})\|^2,$$

where $\underline{Y} = \underline{W} * \underline{X}$ and $\underline{Z} = \underline{W} * \langle\langle \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, ..., \mathbf{A}^{(N)} \rangle\rangle$ ($*$ is the Hadamard product; [2]).

For $n = 1, ..., N$, the partial derivatives of the objective function with respect to the n-th matrix$\mathbf{A}^{(n)}$ can be expressed as follows:

$$\frac{\partial f}{\partial \mathbf{A}^{(n)}} = (\mathbf{Z}_{(n)} - \mathbf{Y}_{(n)}) \mathbf{A}^{(-n)},$$

where

$$\mathbf{A}^{(-n)} = \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)}.$$

The symbol $\odot$ denotes the Khatri-Rao product [3].

After the objective function and the derivation of gradient are obtained, we can solve the optimization problem by any optimization algorithms based on gradient descent method [4]. In this study, the maximum iteration number was set to 300 as the stop criteria for optimization.

Multitask learning method for drug indication prediction

We address the problem of therapeutic indications prediction by focusing on drugs. Note that there are a number of candidates for diseases, and different diseases may have common characteristics in terms of molecular mechanisms. The same drugs are sometimes used for multiple diseases. Thus, we propose formulating the problem in the framework of supervised multiple label prediction.

Suppose that there are M diseases and we are given P drugs. We consider predicting which diseases would be treated by a drug, that is, the i-th drug. Each drug is represented by a d-dimensional feature vector as $\boldsymbol{x}_i$ in this study, where $\boldsymbol{x}_i$ was obtained by averaging the multiple signatures from different cell lines.

We constructed a learning set of drug–disease pairs that are pairs given in drug–disease associations (see the Materials section for more details). There are M candidates for diseases, and each drug in the learning set is assigned a binary class label representing the m-th disease ($m = 1, 2, ..., M$). Let $y_{m,i} \in \{0, 1\}$ be the class label for the m-th disease assigned to the i-th drug, where $y_{m,i} = 1$ means that the i-th drug is used for the m-th disease, and $y_{m,i} = 0$ means that the i-th drug is not used for the m-th disease.

We construct a predictive model to predict whether the i-th drug would be used for the m-th disease ($m = 1, 2, ..., M$). Linear models are a useful tool to analyze extremely high-dimensional data for both prediction and feature extraction tasks. Thus, we adopt a linear function defined as $f_m = \boldsymbol{w}_m^{\mathrm{T}} \boldsymbol{x}_i$, where $\boldsymbol{w}_m$ is a d-dimensional weight vector for the m-th disease. We represent a set of M model weights by a $d \times M$ matrix defined as $\mathbf{W} \coloneqq [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M]$ and estimate the weight matrix $\mathbf{W}$ by minimizing an objective function based on the learning set.

To overcome the scarcity of existing knowledge concerning relationships between drugs and diseases, we propose learning individual predictive models $f_1, f_2, ..., f_M$ jointly, sharing information across M diseases.

We attempt to estimate all of the weight vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M$ jointly in the models by minimizing the logistic loss as follows:

$$R(\mathbf{W}) = \sum_{m=1}^{M} \sum_{i=1}^{P} \log(1 + \exp(-y_{m,i} \boldsymbol{w}_m^{\mathrm{T}} \boldsymbol{x}_i)).$$

We introduce a regularization term $\Omega(\mathbf{W})$ to the loss function in order to enhance the generalization properties. Thus, the optimization problem is written as follows:

$$\min_{\mathbf{W}} R(\mathbf{W}) + \Omega(\mathbf{W}). \tag{1}$$

Here we introduce two regularization terms. First, we use a standard ridge regularization term to avoid the over-fitting problem, which is defined as

$$\Omega_r \coloneqq \frac{1}{2} \mathrm{T}r(\mathbf{W}\mathbf{W}^{\mathrm{T}}).$$

Second, we design another regularization term reflecting the similarities among diseases. In this study we evaluate the similarity among diseases using the Jaccard

coefficient and construct an $M \times M$ similarity matrix $\mathbf{S}$ for diseases in which each element $\mathbf{S}_{i,j}$ is a similarity score between the i-th and j-th diseases (see section 2.2 for more details). Then, we introduce the following regularization term:

$$\Omega_s(\mathbf{W}) \coloneqq \frac{1}{4}\sum_{l=1}^{M}\sum_{m=1}^{M}\mathbf{S}_{l,m}\left\|\frac{\mathbf{w}_l}{\sqrt{\mathbf{K}_{l,l}}} - \frac{\mathbf{w}_m}{\sqrt{\mathbf{K}_{m,m}}}\right\| = \frac{1}{2}\mathrm{T}r(\mathbf{W}\mathbf{L}_s\mathbf{W}^\intercal),$$

where $\|\cdot\|$ is the Euclidean norm, $\mathbf{K}$ is a diagonal matrix defined as $\mathbf{K}_{l,l} \coloneqq \sum_{m=1}^{M}\mathbf{S}_{l,m}$, and $\mathbf{L}_s$ is a symmetric normalized Laplacian defined as $\mathbf{K}^{-1/2}(\mathbf{K}-\mathbf{S})\mathbf{K}^{-1/2}$. The regularization term $\Omega_s(\mathbf{W})$ has the effect of bringing the weight vectors $\mathbf{w}_i$ and $\mathbf{w}_j$ close to each other if $\mathbf{S}_{l,m}$ is high.

Finally, we introduce the following regularization term in the optimization problem (1):

$$\Omega(\mathbf{W}) \coloneqq \lambda_s\Omega_s(\mathbf{W}) + \lambda_r\Omega_r(\mathbf{W}),$$

where $\lambda_s \geq 0$ and $\lambda_r \geq 0$ are hyper-parameters to control the strength of the regularization terms Ω_s and Ω_r, respectively.

Results

A large-scale prediction of new therapeutic indications

We performed a comprehensive prediction of unknown therapeutic indications of 1,483 drugs. For these drugs, the gene expression data are available in the LINCS database. We used all known drug–disease associations as a learning dataset and predicted new drug therapeutic indications by the multitask learning method with tensor decomposition. Here, the possible therapeutic indications were related to 79 diseases.

Supplementary Figure 3 shows the distribution of drugs repositioned from the original disease class to other disease classes based on the predicted therapeutic indications of drugs. Diseases are classified according to the 10th revision of the International Classification of Diseases (ICD-10; [5]) disease chapters. The prediction resulted in the largest number of drugs that were possibly repositioned from chapter I of the ICD-10 (certain infectious and parasitic diseases) to chapter II of the ICD-10 (neoplasms) and *vice versa*, followed by possible drug repositioning from chapter II of the ICD-10 (neoplasms) to chapter IV of the ICD-10 (endocrine, nutritional, and metabolic diseases) and *vice versa*. These results suggest that the proposed approach for a large-scale prediction can provide new therapeutic indications for a wide range of diseases.

Supplementary Figure 4 shows the network of drug–disease associations that are predicted by only the multitask learning method with the tensor decomposition. Here, the associations are shown by focusing on drugs repositioned from the original disease class to other disease classes based on the new therapeutic indications of drugs. For example, niclosamide (D00436), an anthelmintic drug, was predicted to have therapeutic efficacy in adult T-cell leukemia. Adult T-cell leukemia and lymphoma (ATL) is a highly aggressive form of hematological malignancy and is caused by chronic infection with the human T-cell leukemia virus type 1 (HTLV-1). Researchers reported that niclosamide induced apoptosis of HTLV-1-transformed T cells [6]. This implies

that, via a large-scale analysis, finding the therapeutic indications of drugs approved for various diseases is possible.

References

[1] Acar, E., Kolda, T.G., Dunlavy, D.M., & Morup, M. Scalable tensor factorizations for incomplete data. *Chemometr. Intell. Lab. Syst.* **106,** 41–56 (2011).

[2] Kolda T.G. & Bader, B.W. Tensor decomposition and applications. *SIAM Rev.* **51,** 455–500 (2009).

[3] Khatri, C.G. & Rao, C.R. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhya: Indian J. Statistics, Series A* **30,** 167–180 (1968).

[4] Nocedal, J. & Wright, S. *Numerical Optimization.* Springer Science & Business Media, NewYork (2006).

[5] World Health Organization. *The ICD-10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines.* World Health Organization, Geneva, Switzerland (1992).

[6] Xiang, D., Yuan, Y., Chen, L., Liu, X., Belani, C. & Cheng, H. Niclosamide, an anti-helminthic molecule, downregulates the retroviral oncoprotein Tax and pro-survival Bcl-2 proteins in HTLV-1-transformed T lymphocytes. *Biochem. Biophys. Res. Commun.* **464,** 221–228 (2015).

Supplementary Table 1. Performance evaluation of data completion by tensor decomposition algorithms for third-order transcriptome data (drugs, genes, and cell lines) with different rates of artificial missing values. Missing values were generated by the "random missing" strategy. Relative standard errors (RSEs) between the original and reconstructed data from tensor decomposition were calculated for (a) all values and (b) missing values only. The proposed TT-WOPT method and the baseline CP-WOPT method are denoted as TT and CP, respectively. The optimized tensor ranks are shown for each method. Artificially generated missing rates of 10%, 50%, and 90% were tested. Cell lines are listed in order of increasing original missing rates.

	artificial missing rate														
	10%					50%					90%				
	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks
(a) RSEs for all values															
total cell lines	-	0.0739	30	**0.0674**	{1, 30, 30, 1}	-	0.0778	20	**0.0683**	{1, 30, 30, 1}	-	0.0790	30	**0.0720**	{1, 30, 30, 1}
MCF7	-	0.0614	30	**0.0553**	{1, 30, 30, 1}	-	0.0657	20	**0.0562**	{1, 30, 30, 1}	-	0.0680	30	**0.0599**	{1, 30, 30, 1}
PC3	-	0.0644	30	**0.0576**	{1, 30, 30, 1}	-	0.0673	20	**0.0598**	{1, 30, 30, 1}	-	0.0698	30	**0.0648**	{1, 30, 30, 1}
A375	-	0.0850	30	**0.0730**	{1, 30, 30, 1}	-	0.0904	20	**0.0757**	{1, 30, 30, 1}	-	0.0927	30	**0.0866**	{1, 30, 30, 1}
HA1E	-	0.0759	30	**0.0667**	{1, 30, 30, 1}	-	0.0798	20	**0.0688**	{1, 30, 30, 1}	-	0.0820	30	**0.0753**	{1, 30, 30, 1}
HT29	-	0.0770	30	**0.0677**	{1, 30, 30, 1}	-	0.0814	20	**0.0702**	{1, 30, 30, 1}	-	0.0832	30	**0.0785**	{1, 30, 30, 1}
A549	-	0.0749	30	**0.0681**	{1, 30, 30, 1}	-	0.0786	20	**0.0689**	{1, 30, 30, 1}	-	0.0794	30	**0.0724**	{1, 30, 30, 1}
VCAP	-	0.0701	30	**0.0633**	{1, 30, 30, 1}	-	0.0731	20	**0.0647**	{1, 30, 30, 1}	-	0.0740	30	**0.0685**	{1, 30, 30, 1}
YAPC	-	0.0787	30	**0.0716**	{1, 30, 30, 1}	-	0.0826	20	**0.0735**	{1, 30, 30, 1}	-	0.0842	30	**0.0792**	{1, 30, 30, 1}
HELA	-	0.0760	30	**0.0704**	{1, 30, 30, 1}	-	0.0799	20	**0.0717**	{1, 30, 30, 1}	-	0.0812	30	**0.0754**	{1, 30, 30, 1}
HCC515	-	0.0885	30	**0.0786**	{1, 30, 30, 1}	-	0.0929	20	**0.0807**	{1, 30, 30, 1}	-	0.0939	30	**0.0870**	{1, 30, 30, 1}
HEPG2	-	0.0835	30	**0.0788**	{1, 30, 30, 1}	-	0.0866	20	**0.0789**	{1, 30, 30, 1}	-	0.0870	30	**0.0809**	{1, 30, 30, 1}
HS578T	-	0.0677	30	**0.0640**	{1, 30, 30, 1}	-	0.0714	20	**0.0635**	{1, 30, 30, 1}	-	0.0716	30	**0.0625**	{1, 30, 30, 1}
MCF10A	-	0.0693	30	**0.0653**	{1, 30, 30, 1}	-	0.0735	20	**0.0647**	{1, 30, 30, 1}	-	0.0737	30	**0.0638**	{1, 30, 30, 1}
MDAMB231	-	0.0683	30	**0.0647**	{1, 30, 30, 1}	-	0.0722	20	**0.0641**	{1, 30, 30, 1}	-	0.0725	30	**0.0633**	{1, 30, 30, 1}
SKBR3	-	0.0678	30	**0.0639**	{1, 30, 30, 1}	-	0.0718	20	**0.0633**	{1, 30, 30, 1}	-	0.0721	30	**0.0626**	{1, 30, 30, 1}
BT20	-	0.0679	30	**0.0639**	{1, 30, 30, 1}	-	0.0719	20	**0.0633**	{1, 30, 30, 1}	-	0.0721	30	**0.0624**	{1, 30, 30, 1}
(b) RSEs for missing values															
total cell lines	0.0750	0.0765	30	**0.0694**	{1, 30, 30, 1}	0.0837	0.0798	20	**0.0716**	{1, 30, 30, 1}	NA	0.0820	30	**0.0776**	{1, 30, 30, 1}
MCF7	0.0634	0.0616	30	**0.0568**	{1, 30, 30, 1}	0.0735	0.0658	20	**0.0574**	{1, 30, 30, 1}	NA	0.0681	30	**0.0604**	{1, 30, 30, 1}
PC3	0.0648	0.0650	30	**0.0592**	{1, 30, 30, 1}	0.0742	0.0673	20	**0.0614**	{1, 30, 30, 1}	NA	0.0699	30	**0.0655**	{1, 30, 30, 1}
A375	0.0832	0.0862	30	**0.0764**	{1, 30, 30, 1}	0.0929	0.0906	20	**0.0788**	{1, 30, 30, 1}	NA	0.0930	30	**0.0881**	{1, 20, 20, 1}
HA1E	0.0744	0.0759	30	**0.0681**	{1, 30, 30, 1}	0.0842	0.0796	20	**0.0707**	{1, 30, 30, 1}	NA	0.0819	30	**0.0764**	{1, 30, 30, 1}
HT29	0.0773	0.0777	30	**0.0703**	{1, 30, 30, 1}	0.0853	0.0810	20	**0.0726**	{1, 30, 30, 1}	NA	0.0831	30	**0.0797**	{1, 20, 20, 1}
A549	0.0755	0.0785	30	**0.0708**	{1, 30, 30, 1}	0.0833	0.0812	20	**0.0718**	{1, 30, 30, 1}	NA	0.0822	30	**0.0770**	{1, 30, 30, 1}
VCAP	0.0643	0.0710	20	**0.0632**	{1, 30, 30, 1}	0.0703	0.0723	20	**0.0662**	{1, 30, 30, 1}	NA	0.0740	30	**0.0717**	{1, 30, 30, 1}
YAPC	0.0728	0.0738	30	**0.0679**	{1, 30, 30, 1}	0.0840	0.0786	20	**0.0718**	{1, 30, 30, 1}	NA	0.0810	30	**0.0782**	{1, 10, 10, 1}
HELA	0.0701	0.0715	30	**0.0666**	{1, 30, 30, 1}	0.0800	0.0749	20	**0.0693**	{1, 30, 30, 1}	NA	0.0772	30	**0.0739**	{1, 20, 20, 1}
HCC515	0.0986	0.0994	30	**0.0893**	{1, 30, 30, 1}	0.1068	0.1039	20	**0.0926**	{1, 30, 30, 1}	NA	0.1049	30	**0.1000**	{1, 20, 20, 1}
HEPG2	0.0948	0.0954	20	**0.0907**	{1, 30, 30, 1}	0.1012	0.0978	20	**0.0914**	{1, 30, 30, 1}	NA	0.0990	30	**0.0958**	{1, 30, 30, 1}
HS578T	0.0407	0.0420	30	**0.0403**	{1, 30, 30, 1}	0.0431	0.0432	20	**0.0412**	{1, 30, 30, 1}	NA	0.0445	10	**0.0431**	{1, 10, 10, 1}
MCF10A	0.0480	0.0476	10	**0.0455**	{1, 20, 20, 1}	0.0496	0.0482	20	**0.0455**	{1, 20, 20, 1}	NA	0.0496	10	**0.0476**	{1, 20, 20, 1}
MDAMB231	0.0432	0.0440	10	**0.0429**	{1, 30, 30, 1}	0.0475	0.0467	20	**0.0434**	{1, 30, 30, 1}	NA	0.0490	10	**0.0456**	{1, 30, 30, 1}
SKBR3	**0.0415**	0.0440	10	0.0416	{1, 30, 30, 1}	0.0426	0.0432	10	**0.0417**	{1, 10, 10, 1}	NA	0.0450	10	**0.0426**	{1, 10, 10, 1}
BT20	0.0433	0.0441	30	**0.0419**	{1, 30, 30, 1}	0.0443	0.0443	20	**0.0426**	{1, 10, 10, 1}	NA	0.0468	10	**0.0439**	{1, 20, 20, 1}

Supplementary Table 2. Performance evaluation of data completion by tensor decomposition algorithms for fourth-order transcriptome data (drugs, genes, cell lines, and time points) with artificial missing values. Missing values were generated by the "random missing" strategy. Relative standard errors (RSEs) between the original and reconstructed data from tensor decomposition were calculated for (a) all values and (b) missing values only. The proposed TT-WOPT method and the baseline CP-WOPT method are denoted as TT and CP, respectively. The optimized tensor ranks are shown for each method. Artificially generated missing rates of 10%, 50%, and 90% were tested. Cell lines are listed in order of increasing original missing rates.

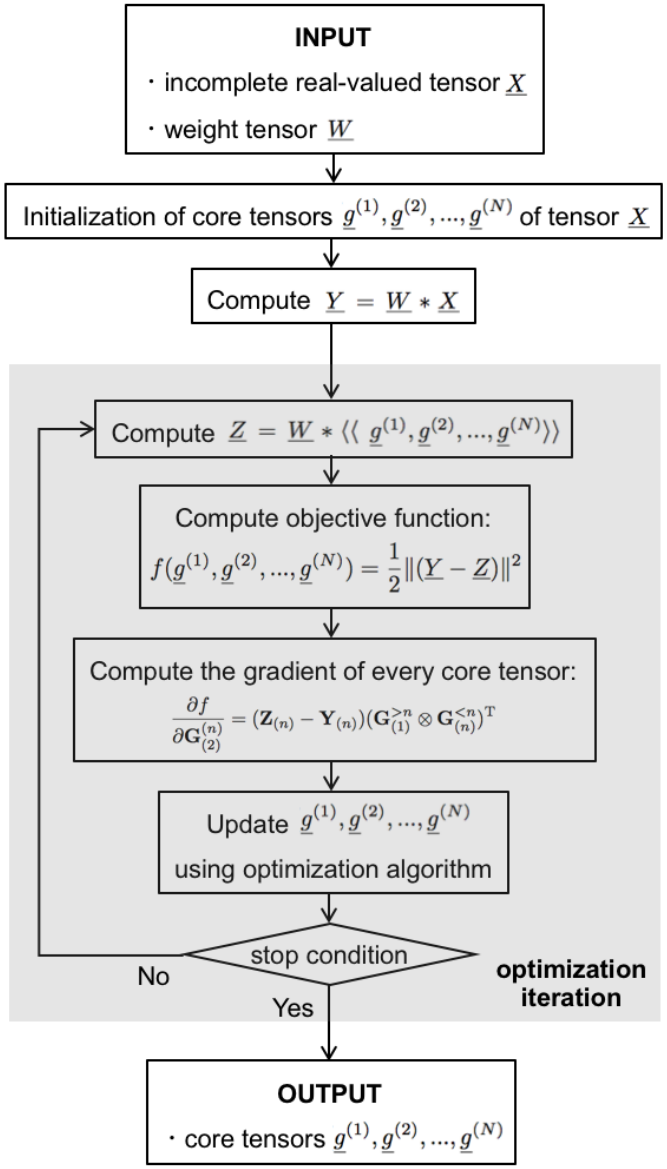
	artificial missing rate														
	10%					50%					90%				
	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	standard imputation	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks
(a) RSEs for all values															
total cell lines	-	0.00245	20	**0.00210**	{1, 20, 20, 20, 1}	-	0.00234	10	**0.00214**	{1, 20, 20, 20, 1}	-	0.00282	10	**0.00214**	{1, 20, 20, 20, 1}
MCF7	-	0.0023	30	**0.00181**	{1, 20, 20, 20, 1}	-	0.00221	10	**0.00186**	{1, 20, 20, 20, 1}	-	0.00288	10	**0.00189**	{1, 20, 20, 20, 1}
PC3	-	0.00208	20	**0.00170**	{1, 10, 10, 10, 1}	-	0.00194	10	**0.00169**	{1, 10, 10, 10, 1}	-	0.00265	10	**0.00172**	{1, 10, 10, 10, 1}
A375	-	0.00252	20	**0.00213**	{1, 20, 20, 20, 1}	-	0.00236	10	**0.00216**	{1, 20, 20, 20, 1}	-	0.00298	10	**0.00217**	{1, 20, 20, 20, 1}
HA1E	-	0.00274	30	**0.00245**	{1, 20, 20, 20, 1}	-	0.00271	10	**0.00248**	{1, 20, 20, 20, 1}	-	0.00311	10	**0.00248**	{1, 20, 20, 20, 1}
HT29	-	0.00234	30	**0.00201**	{1, 20, 20, 20, 1}	-	0.00229	10	**0.00206**	{1, 20, 20, 20, 1}	-	0.00279	10	**0.00205**	{1, 20, 20, 20, 1}
A549	-	0.00285	20	**0.00255**	{1, 20, 20, 20, 1}	-	0.00284	10	**0.00259**	{1, 20, 20, 20, 1}	-	0.00314	10	**0.00258**	{1, 20, 20, 20, 1}
VCAP	-	0.00255	30	**0.00224**	{1, 20, 20, 20, 1}	-	0.00259	10	**0.00228**	{1, 20, 20, 20, 1}	-	0.00293	10	**0.00228**	{1, 20, 20, 20, 1}
YAPC	-	0.00249	30	**0.00217**	{1, 20, 20, 20, 1}	-	0.00235	10	**0.00222**	{1, 20, 20, 20, 1}	-	0.00283	10	**0.00221**	{1, 20, 20, 20, 1}
HELA	-	0.00232	20	**0.00203**	{1, 20, 20, 20, 1}	-	0.00228	10	**0.00208**	{1, 20, 20, 20, 1}	-	0.00265	20	**0.00207**	{1, 20, 20, 20, 1}
HCC515	-	0.00283	20	**0.00249**	{1, 20, 20, 20, 1}	-	0.00264	10	**0.00253**	{1, 20, 20, 20, 1}	-	0.00323	10	**0.00252**	{1, 20, 20, 20, 1}
HEPG2	-	0.00249	30	**0.00218**	{1, 20, 20, 20, 1}	-	0.00234	10	**0.00222**	{1, 20, 20, 20, 1}	-	0.00292	10	**0.00221**	{1, 20, 20, 20, 1}
HS578T	-	0.00212	20	**0.00181**	{1, 20, 20, 20, 1}	-	0.00198	10	**0.00186**	{1, 20, 20, 20, 1}	-	0.00243	10	**0.00185**	{1, 20, 20, 20, 1}
MCF10A	-	0.00214	20	**0.00181**	{1, 20, 20, 20, 1}	-	0.00204	10	**0.00186**	{1, 20, 20, 20, 1}	-	0.00247	10	**0.00185**	{1, 20, 20, 20, 1}
MDAMB231	-	0.0023	20	**0.00201**	{1, 20, 20, 20, 1}	-	0.00220	10	**0.00206**	{1, 20, 20, 20, 1}	-	0.00261	10	**0.00204**	{1, 20, 20, 20, 1}
SKBR3	-	0.00246	20	**0.00216**	{1, 20, 20, 20, 1}	-	0.00241	10	**0.00221**	{1, 20, 20, 20, 1}	-	0.00278	10	**0.00220**	{1, 20, 20, 20, 1}
BT20	-	0.00215	20	**0.00181**	{1, 20, 20, 20, 1}	-	0.00203	10	**0.00186**	{1, 20, 20, 20, 1}	-	0.00243	10	**0.00185**	{1, 20, 20, 20, 1}
(b) RSEs for missing values															
total cell lines	0.00271	0.00308	30	**0.00266**	{1, 10, 10, 10, 1}	0.00276	0.00299	10	**0.00269**	{1, 20, 20, 20, 1}	NA	0.0036	10	**0.00276**	{1, 10, 10, 10, 1}
MCF7	0.00195	0.00314	30	**0.00189**	{1, 10, 10, 10, 1}	0.00242	0.00275	10	**0.00236**	{1, 10, 10, 10, 1}	NA	0.00366	10	**0.00263**	{1, 10, 10, 10, 1}
PC3	0.00243	0.00266	30	**0.00222**	{1, 30, 30, 30, 1}	0.00264	0.00286	10	**0.00237**	{1, 30, 30, 30, 1}	NA	0.00362	10	**0.00257**	{1, 30, 30, 30, 1}
A375	0.00288	0.00316	20	**0.00286**	{1, 20, 20, 20, 1}	0.00284	0.00302	10	**0.00261**	{1, 30, 30, 30, 1}	NA	0.00354	10	**0.00254**	{1, 20, 20, 20, 1}
HA1E	0.00331	0.00319	30	**0.00293**	{1, 30, 30, 30, 1}	0.00293	0.00317	10	**0.00276**	{1, 20, 20, 20, 1}	NA	0.00368	10	**0.00285**	{1, 30, 30, 30, 1}
HT29	0.00217	0.00215	30	**0.00178**	{1, 30, 30, 30, 1}	0.00195	0.00225	10	**0.00192**	{1, 10, 10, 10, 1}	NA	0.00297	10	**0.00203**	{1, 20, 20, 20, 1}
A549	0.00273	0.00272	30	**0.00241**	{1, 30, 30, 30, 1}	0.00327	0.00351	10	**0.00322**	{1, 10, 10, 10, 1}	NA	0.00390	10	**0.00333**	{1, 10, 10, 10, 1}
VCAP	0.00277	0.0036	30	**0.00273**	{1, 10, 10, 10, 1}	0.00310	0.00334	10	**0.00297**	{1, 20, 20, 20, 1}	NA	0.00371	10	**0.00301**	{1, 30, 30, 30, 1}
YAPC	0.00357	0.00371	30	**0.00344**	{1, 30, 30, 30, 1}	0.00367	0.00367	10	**0.00319**	{1, 20, 20, 20, 1}	NA	0.00406	10	**0.00355**	{1, 20, 20, 20, 1}
HELA	0.00429	0.00348	20	**0.00325**	{1, 20, 20, 20, 1}	0.00396	0.00398	10	**0.00372**	{1, 30, 30, 30, 1}	NA	0.00424	10	**0.00378**	{1, 20, 20, 20, 1}
HCC515	0.00230	0.00225	20	**0.00155**	{1, 20, 20, 20, 1}	0.00210	0.00225	10	**0.00199**	{1, 20, 20, 20, 1}	NA	0.00310	10	**0.00209**	{1, 30, 30, 30, 1}
HEPG2	0.00105	0.00173	20	**0.00100**	{1, 10, 10, 10, 1}	0.00142	0.00169	10	**0.00138**	{1, 10, 10, 10, 1}	NA	0.00303	10	**0.00163**	{1, 30, 30, 30, 1}
HS578T	0.00188	0.00177	30	**0.00074**	{1, 30, 30, 30, 1}	0.00108	0.00196	10	**0.00102**	{1, 10, 10, 10, 1}	NA	0.00312	10	**0.00136**	{1, 10, 10, 10, 1}
MCF10A	0.00160	0.00159	30	**0.00044**	{1, 20, 20, 20, 1}	0.00088	0.00169	10	**0.00081**	{1, 30, 30, 30, 1}	NA	0.00316	10	**0.00106**	{1, 10, 10, 10, 1}
MDAMB231	0.00061	0.00172	30	**0.00049**	{1, 30, 30, 30, 1}	0.00076	0.00205	10	**0.00066**	{1, 20, 20, 20, 1}	NA	0.00281	10	**0.00078**	{1, 10, 10, 10, 1}
SKBR3	**0.00035**	0.00173	30	0.00036	{1, 30, 30, 30, 1}	0.00116	0.00219	10	**0.00112**	{1, 20, 20, 20, 1}	NA	0.00289	20	**0.00108**	{1, 20, 20, 20, 1}
BT20	**0.00069**	0.00176	30	0.00083	{1, 20, 20, 20, 1}	0.00094	0.00198	10	**0.00071**	{1, 20, 20, 20, 1}	NA	0.00299	20	**0.00115**	{1, 20, 20, 20, 1}

Supplementary Table 3. Performance evaluation of data completion by tensor decomposition algorithms for third-order transcriptome data (drugs, genes, and cell lines) with different rates of artificial missing values. Missing values were generated by the "cell-based missing" strategy. Relative standard errors (RSEs) between the original and reconstructed data from tensor decomposition were calculated for missing values only. The proposed TT-WOPT method and the baseline CP-WOPT method are denoted as TT and CP, respectively. The optimized tensor ranks are shown for each method. Cell lines are listed in order of increasing original missing rates.

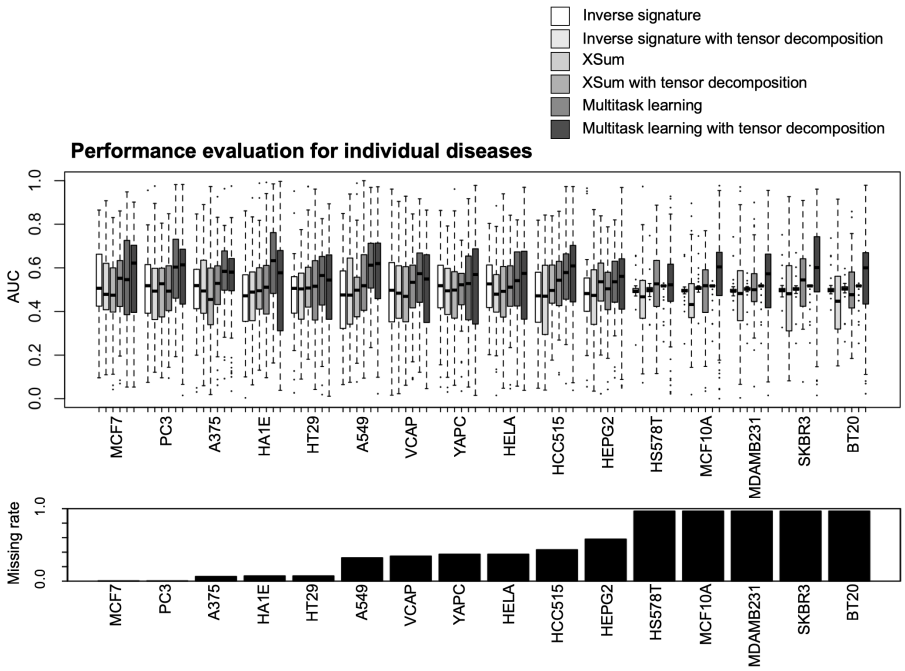
artificial	**(a) RSEs for all values**				**(b) RSEs for missing values**			
missing cell	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks
MCF7	0.1811	30	**0.1523**	{1, 30, 30, 1}	0.6673	30	**0.5498**	{1, 30, 30, 1}
PC3	0.2170	20	**0.1525**	{1, 30, 30, 1}	0.8199	20	**0.5514**	{1, 30, 30, 1}
A375	0.2216	10	**0.1511**	{1, 30, 30, 1}	0.8122	10	**0.5459**	{1, 30, 30, 1}
HA1E	0.2495	20	**0.1539**	{1, 30, 30, 1}	0.9562	20	**0.5583**	{1, 30, 30, 1}
HT29	0.2577	30	**0.1551**	{1, 30, 30, 1}	0.9910	30	**0.5638**	{1, 30, 30, 1}
A549	0.2401	10	**0.1529**	{1, 30, 30, 1}	0.9157	10	**0.5537**	{1, 30, 30, 1}
VCAP	0.2196	20	**0.1531**	{1, 30, 30, 1}	0.8329	20	**0.5549**	{1, 30, 30, 1}
YAPC	0.2604	20	**0.1530**	{1, 30, 30, 1}	1.0015	20	**0.5547**	{1, 30, 30, 1}
HELA	0.2695	20	**0.1540**	{1, 30, 30, 1}	1.0390	20	**0.5590**	{1, 30, 30, 1}
HCC515	0.2109	30	**0.1528**	{1, 30, 30, 1}	0.7910	30	**0.5541**	{1, 30, 30, 1}
HEPG2	0.1657	20	**0.1564**	{1, 30, 30, 1}	**0.5855**	20	0.5696	{1, 30, 30, 1}
HS578T	0.2281	30	**0.1517**	{1, 30, 30, 1}	0.8655	30	**0.5476**	{1, 30, 30, 1}
MCF10A	0.2157	20	**0.1508**	{1, 30, 30, 1}	0.8139	20	**0.5439**	{1, 30, 30, 1}
MDAMB231	0.2134	20	**0.1537**	{1, 30, 30, 1}	0.8029	20	**0.5571**	{1, 30, 30, 1}
SKBR3	0.2208	20	**0.1546**	{1, 30, 30, 1}	0.8307	20	**0.5609**	{1, 30, 30, 1}
BT20	0.2238	30	**0.1538**	{1, 30, 30, 1}	0.8500	30	**0.5574**	{1, 30, 30, 1}

Supplementary Table 4. Performance evaluation of data completion by tensor decomposition algorithms for fourth-order transcriptome data (drugs, genes, cell lines, and time points) with artificial missing values. Missing values were generated by the "cell-based missing" strategy. Relative standard errors (RSEs) between the original and reconstructed data from tensor decomposition were calculated for (a) all values and (b) missing values only. The proposed TT-WOPT method and the baseline CP-WOPT method are denoted as TT and CP, respectively. The optimized tensor ranks are shown for each method. Cell lines are listed in order of increasing original missing rates.

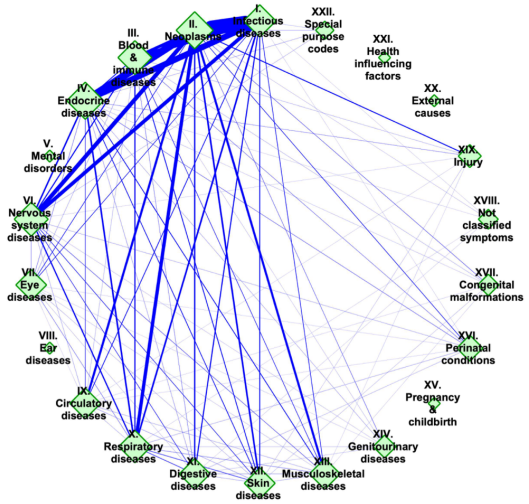
artificial	**(a) RSEs for all values**				**(b) RSEs for missing values**			
missing cell	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks	CP (baseline)	CP-ranks	TT (proposed)	TT-ranks
MCF7	0.2693	30	**0.0071**	{1, 30, 30, 30, 1}	1.0749	20	**0.0266**	{1, 30, 30, 30, 1}
PC3	0.2215	20	**0.0064**	{1, 30, 30, 30, 1}	0.8859	20	**0.0236**	{1, 30, 30, 30, 1}
A375	0.1811	20	**0.0122**	{1, 30, 30, 30, 1}	0.7245	20	**0.0481**	{1, 30, 30, 30, 1}
HA1E	0.2568	10	**0.0052**	{1, 30, 30, 30, 1}	1.0273	10	**0.0173**	{1, 30, 30, 30, 1}
HT29	0.295	10	**0.0056**	{1, 30, 30, 30, 1}	1.1522	20	**0.0198**	{1, 30, 30, 30, 1}
A549	0.2222	20	**0.0111**	{1, 30, 30, 30, 1}	0.8887	20	**0.0436**	{1, 30, 30, 30, 1}
VCAP	0.1543	10	**0.0115**	{1, 30, 30, 30, 1}	0.6172	10	**0.0452**	{1, 30, 30, 30, 1}
YAPC	0.1838	30	**0.0055**	{1, 30, 30, 30, 1}	0.7352	30	**0.0198**	{1, 30, 30, 30, 1}
HELA	0.2073	20	**0.0098**	{1, 30, 30, 30, 1}	0.8291	20	**0.0380**	{1, 30, 30, 30, 1}
HCC515	0.3141	30	**0.0048**	{1, 30, 30, 30, 1}	1.0315	10	**0.0171**	{1, 30, 30, 30, 1}
HEPG2	0.2077	30	**0.0051**	{1, 30, 30, 30, 1}	0.8308	30	**0.0175**	{1, 30, 30, 30, 1}
HS578T	0.1887	30	**0.0101**	{1, 30, 30, 30, 1}	0.7548	30	**0.0395**	{1, 30, 30, 30, 1}
MCF10A	0.1678	20	**0.0108**	{1, 30, 30, 30, 1}	0.6713	20	**0.0421**	{1, 30, 30, 30, 1}
MDAMB231	0.2241	30	**0.0053**	{1, 30, 30, 30, 1}	0.8964	30	**0.0191**	{1, 30, 30, 30, 1}
SKBR3	0.2164	20	**0.0108**	{1, 30, 30, 30, 1}	0.8654	20	**0.0423**	{1, 30, 30, 30, 1}
BT20	0.2711	30	**0.0052**	{1, 30, 30, 30, 1}	1.0127	20	**0.0178**	{1, 30, 30, 30, 1}



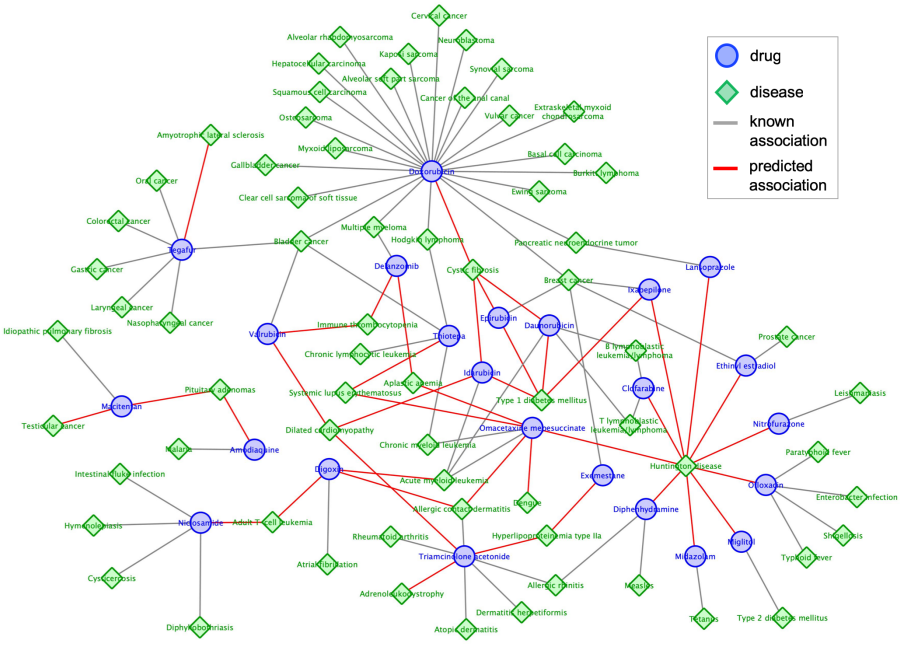
Supplementary Figure 1. Flow diagram of the tensor-train weighted optimization (TT-WOPT) algorithm.



Supplementary Figure 2. Performance comparison on drug indication prediction among the inverse signature, XSum, and multitask learning methods with and without tensor decomposition. The top panel shows the distribution of AUC scores calculated using all prediction scores for individual diseases. The bottom panel shows the missing rate in each cell line. Cell lines are listed in increasing order of missing rates.



Supplementary Figure 3. Distribution of drugs repositioned from the original disease class to other disease classes. Nodes (indicated by gray diamonds) represent ICD-10 disease chapters (shown with the chapter number and short chapter name). Edges (indicated by blue lines) indicate potential correlations between diseases according to the new therapeutic indications of drugs. Node size indicates the sum of the edges of each node. Edge width indicates the number of drugs repositioned between two disease chapters. The chapters are as follows: Chapter I: certain infectious and parasitic diseases (A00–B99). Chapter II: neoplasms (C00–D48). Chapter III: diseases of the blood, blood-forming organs, and certain disorders involving immune mechanisms (D50–D89). Chapter IV: endocrine, nutritional, and metabolic diseases (E00–E90). Chapter V: mental and behavioral disorders (F00–F99). Chapter VI: diseases of the nervous system (G00–G99). Chapter VII: diseases of the eye and adnexa (H00–H59). Chapter VIII: diseases of the ear and mastoid process (H60–H95). Chapter IX: diseases of the circulatory system (I00–I99). Chapter X: diseases of the respiratory system (J00–J99). Chapter XI: diseases of the digestive system (K00–K93). Chapter XII: diseases of the skin and subcutaneous tissue (L00–L99). Chapter XIII: diseases of the musculoskeletal system and connective tissue (M00–M99). Chapter XIV: diseases of the genitourinary system (N00–N99). Chapter XV: pregnancy, childbirth, and the puerperium (O00–O99). Chapter XVI: certain conditions originating in the perinatal period (P00–P96). Chapter XVII: congenital malformations, deformations; and chromosomal abnormalities (Q00–Q99). Chapter XVIII: symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified (R00–R99). Chapter XIX: injury, poisoning, and certain other consequences of external causes (S00–T98). Chapter XX: external causes of morbidity and mortality (V01–Y98). Chapter XXI: factors influencing health status and contact with health services (Z00–Z99). Chapter XXII: codes for special purposes (U00–U99).



Supplementary Figure 4. Drug–disease association network predicted using the multitask learning method with tensor decomposition. Blue circles and green diamonds denote drugs and diseases, respectively. Gray and red lines denote known and predicted associations, respectively.