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

Michio Iwata¹, Longhao Yuan²ⁿ, Qibin Zhao³ⁿ, Yasuo Tabei³, Francois Berenger¹, Ryusuke Sawada¹, Sayaka Akiyoshi⁵, and Yoshihiro Yamanishi¹ⁿ,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, $X \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, with missing entries [1]. The index of the missing entries can be recorded by a weight tensor ($W$), the size of which is same as that of $X$. Each entry of $W$ satisfies the following conditions:

$$W_{i_1i_2\ldots i_N} = \begin{cases} 
0 & \text{if } x_{i_1i_2\ldots i_N} \text{ is a missing entry,} \\
1 & \text{if } x_{i_1i_2\ldots i_N} \text{ is an observed entry.}
\end{cases}$$

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

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

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

$$x_{i_1i_2\ldots i_N} = \sum_{r=1}^{R} \prod_{n=1}^{N} a_{i_n}^{(n)} r,$$

where $a_{i_n}^{(n)}$ is 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:

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

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

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

$$\frac{\partial f}{\partial A^{(n)}} = (Z_{(n)} - Y_{(n)})A^{(-n)},$$

where

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

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 $x_i$ in this study, where $x_i$ 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 ($m = 1, 2, \ldots, M$). 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 ($m = 1, 2, \ldots, M$). 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 $f_m = w_m^T x_i$, where $w_m$ 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 $W := [w_1, w_2, \ldots, w_M]$ and estimate the weight matrix $W$ 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 $f_1, f_2, \ldots, f_M$ jointly, sharing information across $M$ diseases.

We attempt to estimate all of the weight vectors $w_1, w_2, \ldots, w_M$ jointly in the models by minimizing the logistic loss as follows:

$$R(W) = \sum_{m=1}^{M} \sum_{i=1}^{P} \log(1 + \exp(-y_{m,i}w_m^T x_i)).$$

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

$$\min_W R(W) + \Omega(W). \quad (1)$$

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

$$\Omega_r := \frac{1}{2} \text{tr}(WW^T).$$

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 $S$ for diseases in which each element $S_{ij}$ 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:

$$
\Omega_s(W) := \frac{1}{4 \sum_{i=1}^{M} \sum_{m=1}^{M} S_{im} \left\| \frac{w_i}{\sqrt{K_{ii}}} - \frac{w_m}{\sqrt{K_{mm}}} \right\| ^2 + \frac{1}{2} tr(WL_sW^T),
$$

where $\| \cdot \|$ is the Euclidean norm, $K$ is a diagonal matrix defined as $K_{ij} := \sum_{m=1}^{M} S_{im}$, and $L_s$ is a symmetric normalized Laplacian defined as $K^{-1/2}(K - S)K^{-1/2}$. The regularization term $\Omega_s(W)$ has the effect of bringing the weight vectors $w_i$ and $w_j$ close to each other if $S_{ij}$ is high.

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

$$
\Omega(W) := \lambda_s \Omega_s(W) + \lambda_r \Omega_r(W),
$$

where $\lambda_s \geq 0$ and $\lambda_r \geq 0$ are hyper-parameters to control the strength of the regularization terms $\Omega_s$ and $\Omega_r$, 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, New York (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.

| Cell line | 10% RSE | 50% RSE | 90% RSE |
|-----------|---------|---------|---------|
| MCF7      | 0.0683  | 0.0614  | 0.0657  |
| PC3       | 0.0686  | 0.0675  | 0.0808  |
| A375      | 0.0590  | 0.0606  | 0.0827  |
| H292      | 0.0780  | 0.0814  | 0.0852  |
| A549      | 0.0780  | 0.0814  | 0.0852  |
| H182      | 0.0760  | 0.0790  | 0.0812  |
| HEPG2     | 0.0855  | 0.0866  | 0.0870  |
| HUH7      | 0.0677  | 0.0674  | 0.0726  |
| HCC515    | 0.0695  | 0.0719  | 0.0725  |
| SKBR3     | 0.0679  | 0.0719  | 0.0721  |
| MDA-MB231 | 0.0669  | 0.0679  | 0.0701  |
| BT20      | 0.0679  | 0.0679  | 0.0701  |

Note: The table shows the relative standard errors (RSEs) for different missing rates and tensor decomposition methods.
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) total cell lines, (b) drug sets, (c) gene sets, and (d) time points.

| Drug Set | Gene Set | Cell Line | Time Points |
|----------|----------|-----------|-------------|
| total cell lines | 0.00284 | 20 | 0.00218 (1, 20, 20, 1) |
| MCF7 | 0.0023 | 30 | 0.00081 (1, 20, 20, 1) |
| PC3 | 0.00286 | 20 | 0.00178 (1, 20, 20, 1) |
| A375 | 0.00252 | 20 | 0.00213 (1, 20, 20, 1) |
| H1299 | 0.00256 | 30 | 0.00245 (1, 20, 20, 1) |
| HT29 | 0.00324 | 30 | 0.00201 (1, 20, 20, 1) |
| A549 | 0.00285 | 20 | 0.00235 (1, 20, 20, 1) |
| VCAP | 0.00362 | 30 | 0.00221 (1, 20, 20, 1) |
| YAPC | 0.00289 | 30 | 0.00217 (1, 20, 20, 1) |
| HELA | 0.00282 | 20 | 0.00203 (1, 20, 20, 1) |
| HCC70 | 0.00253 | 20 | 0.00214 (1, 20, 20, 1) |
| HEDG2 | 0.00309 | 30 | 0.00188 (1, 20, 20, 1) |
| HEST7E | 0.00322 | 30 | 0.00201 (1, 20, 20, 1) |
| MDA435A | 0.00368 | 30 | 0.00221 (1, 20, 20, 1) |
| SKBR3 | 0.00362 | 20 | 0.00258 (1, 20, 20, 1) |
| MDA231 | 0.00357 | 30 | 0.00245 (1, 20, 20, 1) |
| SKM1 | 0.00365 | 30 | 0.00217 (1, 20, 20, 1) |
| MDA361 | 0.00323 | 20 | 0.00211 (1, 20, 20, 1) |
| SKBR3 | 0.00321 | 20 | 0.00217 (1, 20, 20, 1) |
| HCC70 | 0.00361 | 20 | 0.00217 (1, 20, 20, 1) |
| SKM1 | 0.00361 | 20 | 0.00217 (1, 20, 20, 1) |
| MDA361 | 0.00361 | 20 | 0.00217 (1, 20, 20, 1) |

For RSEs for missing values:

| Drug Set | Gene Set | Cell Line | Time Points |
|----------|----------|-----------|-------------|
| total cell lines | 0.00284 | 20 | 0.00218 (1, 20, 20, 1) |
| MCF7 | 0.0023 | 30 | 0.00081 (1, 20, 20, 1) |
| PC3 | 0.00286 | 20 | 0.00178 (1, 20, 20, 1) |
| A375 | 0.00252 | 20 | 0.00213 (1, 20, 20, 1) |
| H1299 | 0.00256 | 30 | 0.00245 (1, 20, 20, 1) |
| HT29 | 0.00324 | 30 | 0.00201 (1, 20, 20, 1) |
| A549 | 0.00285 | 20 | 0.00235 (1, 20, 20, 1) |
| VCAP | 0.00362 | 30 | 0.00221 (1, 20, 20, 1) |
| YAPC | 0.00289 | 30 | 0.00217 (1, 20, 20, 1) |
| HELA | 0.00282 | 20 | 0.00203 (1, 20, 20, 1) |
| HCC70 | 0.00253 | 20 | 0.00214 (1, 20, 20, 1) |
| HEDG2 | 0.00309 | 30 | 0.00188 (1, 20, 20, 1) |
| HEST7E | 0.00322 | 30 | 0.00201 (1, 20, 20, 1) |
| MDA435A | 0.00368 | 30 | 0.00221 (1, 20, 20, 1) |
| SKBR3 | 0.00362 | 20 | 0.00258 (1, 20, 20, 1) |
| MDA231 | 0.00357 | 30 | 0.00245 (1, 20, 20, 1) |
| SKM1 | 0.00365 | 30 | 0.00217 (1, 20, 20, 1) |
| MDA361 | 0.00323 | 20 | 0.00211 (1, 20, 20, 1) |
| SKBR3 | 0.00321 | 20 | 0.00217 (1, 20, 20, 1) |
| HCC70 | 0.00361 | 20 | 0.00217 (1, 20, 20, 1) |
| SKM1 | 0.00361 | 20 | 0.00217 (1, 20, 20, 1) |
| MDA361 | 0.00361 | 20 | 0.00217 (1, 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 missing cell | (a) RSEs for all values | (b) RSEs for missing values |
|------------------------|-------------------------|----------------------------|
|                        | CP (baseline) | CP-ranks | TT (proposed) | TT-ranks | CP (baseline) | CP-ranks | TT (proposed) | TT-ranks |
| MCF7                   | 0.1811        | 0.1523   | [1, 30, 30, 1] | 0.6673  | 30           | 0.5496   | [1, 30, 30, 1] |
| PC3                    | 0.2170        | 0.1525   | [1, 30, 30, 1] | 0.8199  | 20           | 0.5514   | [1, 30, 30, 1] |
| A575                   | 0.2216        | 0.1531   | [1, 30, 30, 1] | 0.8122  | 10           | 0.5410   | [1, 30, 30, 1] |
| HA1E                   | 0.2495        | 0.1539   | [1, 30, 30, 1] | 0.9562  | 20           | 0.5593   | [1, 30, 30, 1] |
| HT29                   | 0.2577        | 0.1550   | [1, 30, 30, 1] | 0.9910  | 30           | 0.5538   | [1, 30, 30, 1] |
| A540                   | 0.2404        | 0.1529   | [1, 30, 30, 1] | 0.9157  | 10           | 0.5537   | [1, 30, 30, 1] |
| VCAP                   | 0.2196        | 0.1531   | [1, 30, 30, 1] | 0.8329  | 20           | 0.5549   | [1, 30, 30, 1] |
| YAPC                   | 0.2604        | 0.1530   | [1, 30, 30, 1] | 1.0015  | 20           | 0.5547   | [1, 30, 30, 1] |
| HELA                   | 0.2695        | 0.1540   | [1, 30, 30, 1] | 1.0390  | 20           | 0.5590   | [1, 30, 30, 1] |
| HCC515                 | 0.2109        | 0.1528   | [1, 30, 30, 1] | 0.7910  | 30           | 0.5541   | [1, 30, 30, 1] |
| HEPG2                  | 0.1657        | 0.1564   | [1, 30, 30, 1] | 0.5855  | 20           | 0.5696   | [1, 30, 30, 1] |
| HS578T                 | 0.2281        | 0.1517   | [1, 30, 30, 1] | 0.8655  | 30           | 0.5476   | [1, 30, 30, 1] |
| MCF10A                 | 0.2157        | 0.1508   | [1, 30, 30, 1] | 0.8139  | 20           | 0.5439   | [1, 30, 30, 1] |
| MDAMB231               | 0.2134        | 0.1537   | [1, 30, 30, 1] | 0.8029  | 20           | 0.5571   | [1, 30, 30, 1] |
| SKBR3                  | 0.2208        | 0.1546   | [1, 30, 30, 1] | 0.8307  | 20           | 0.5609   | [1, 30, 30, 1] |
| BT20                   | 0.2238        | 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 missing cell | CP (baseline) | (a) RSEs for all values | CP-ranks | TT (proposed) | TT-ranks | (b) RSEs for missing values | CP-ranks | TT (proposed) | TT-ranks |
|------------------------|--------------|-------------------------|----------|---------------|----------|---------------------------|----------|---------------|----------|
| MCF7                   | 0.2693       | 0.0071                  | [1, 30, 30, 30, 1] | 1.0749   | 20         | 0.0266                    | [1, 30, 30, 30, 1] |
| PC3                    | 0.2215       | 0.0064                  | [1, 30, 30, 30, 1] | 0.8859   | 20         | 0.0236                    | [1, 30, 30, 30, 1] |
| A375                   | 0.1811       | 0.0122                  | [1, 30, 30, 30, 1] | 0.7245   | 20         | 0.0481                    | [1, 30, 30, 30, 1] |
| HA1E                   | 0.2568       | 0.0052                  | [1, 30, 30, 30, 1] | 1.0273   | 10         | 0.0173                    | [1, 30, 30, 30, 1] |
| HT29                   | 0.295        | 0.0056                  | [1, 30, 30, 30, 1] | 1.1522   | 20         | 0.0198                    | [1, 30, 30, 30, 1] |
| A549                   | 0.2222       | 0.0111                  | [1, 30, 30, 30, 1] | 0.8887   | 20         | 0.0436                    | [1, 30, 30, 30, 1] |
| VCAP                   | 0.1543       | 0.0115                  | [1, 30, 30, 30, 1] | 0.6172   | 10         | 0.0452                    | [1, 30, 30, 30, 1] |
| YAPC                   | 0.1838       | 0.00055                 | [1, 30, 30, 30, 1] | 0.7352   | 30         | 0.0198                    | [1, 30, 30, 30, 1] |
| HELA                   | 0.2073       | 0.0098                  | [1, 30, 30, 30, 1] | 0.8291   | 20         | 0.0380                    | [1, 30, 30, 30, 1] |
| HCCS15                 | 0.3144       | 0.0048                  | [1, 30, 30, 30, 1] | 1.0315   | 10         | 0.0171                    | [1, 30, 30, 30, 1] |
| HEPG2                  | 0.2077       | 0.0051                  | [1, 30, 30, 30, 1] | 0.8308   | 30         | 0.0175                    | [1, 30, 30, 30, 1] |
| HS778T                 | 0.1887       | 0.0101                  | [1, 30, 30, 30, 1] | 0.7548   | 30         | 0.0395                    | [1, 30, 30, 30, 1] |
| MCF10A                 | 0.1678       | 0.0108                  | [1, 30, 30, 30, 1] | 0.6713   | 20         | 0.0423                    | [1, 30, 30, 30, 1] |
| MDAMB231               | 0.2244       | 0.0053                  | [1, 30, 30, 30, 1] | 0.8964   | 30         | 0.0191                    | [1, 30, 30, 30, 1] |
| SKBR3                  | 0.2164       | 0.0108                  | [1, 30, 30, 30, 1] | 0.8654   | 20         | 0.0423                    | [1, 30, 30, 30, 1] |
| BT20                   | 0.2711       | 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.