Novel artificial intelligence-based identification of drug-gene-disease interaction using protein-protein interaction

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

[1*]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan.
[2]Department of Computer Science, King Abdulaziz University, Street, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): tag@granular.com;
Contributing authors: tturki@kau.edu.sa;

Abstract

The evaluation of drug-gene-disease interactions is key for the identification of drugs effective against disease. However, at present, drugs that are effective against genes that are critical for disease are difficult to identify. Following a disease-centric approach, there is a need to identify genes critical to disease function and find drugs that are effective against them. By contrast, following a drug-centric approach comprises identifying the genes targeted by drugs, and then the diseases in which the identified genes are critical. Both of these processes are complex. Using a gene-centric approach, whereby we identify genes that are effective against the disease and can be targeted by drugs, is much easier. However, how such sets of genes can be identified without specifying either the target diseases or drugs is not known. In this study, a novel artificial intelligence-based approach that employs unsupervised methods and identifies genes without specifying neither diseases nor drugs is presented. To evaluate its feasibility, we applied tensor decomposition (TD)-based unsupervised feature extraction (FE) to perform drug repositioning from protein-protein interactions (PPI) without any other information. Proteins selected by TD-based unsupervised FE include many genes related to cancers, as well as drugs that target the selected proteins. Thus, we were able to identify cancer drugs using only PPI. Because the selected proteins had more interactions, we replaced the selected proteins with hub proteins and found that hub proteins themselves could be used for drug repositioning. In contrast to hub proteins, which can only identify cancer drugs, TD-based unsupervised FE enables the identification of drugs for other diseases.

In addition, TD-based unsupervised FE can be used to identify drugs that are effective in *in vivo* experiments, which is difficult when hub proteins are used. In conclusion, TD-based unsupervised FE is a useful tool for drug repositioning using only PPI without other information.

Keywords: protein-protein interaction, drug repositioning, artificial intelligence, unsupervised learning, tensor decomposition

1 Introduction

The identification of drug-gene-disease information is critical for drug repositioning. Nevertheless, due to the fact that this does not involve a simple identification of paired information but rather of triplet information, identifying drug-gene-disease information is not an easy task. The identification of drug-gene-disease information is often performed in two steps. From the perspective of the genes (i.e. drug-centric approach), the identification of drug-gene relationships is performed first, followed by the extraction of gene-disease information by identifying diseases related to the selected genes. By contrast, when starting from the perspective of the diseases (i.e. disease-centric approach), the identification of gene-disease relations is performed first, followed by the extraction of drug-gene information via the identification of drugs that target the genes selected. Since the existence of a significant relationship at the second stage is not guaranteed (i.e., gene-disease information for the drug-centric approach or drug-gene information for the disease-centric approach), the identification of drug-gene-disease information is much more difficult than the identification of only gene-disease information or drug-gene information.

Several studies have reported on such drug-gene-disease relationships. Zickenrott et al. [1] attempted to predict disease–gene–drug relationships using differential network analysis, while Wong et al. [2] searched for gene-drug-disease interactions in pharmacogenomics by using GeneDive. Yu et al. [3] predicted drugs with opposing effects on disease genes using a directed network, Sun [4] investigated gene-gene, drug-drug, and disease-disease networks and studied their relationships, and Qahwaji et al. [5] reviewed the genetic approaches to drug development and therapy. Furthemore, Iida et al. [6] investigated network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets. Lastly, Quan et al. [7] considered genetic disease genes as promising sources of drug targets. While these are only some examples, all focused on diseases, drugs, or both, and no studies have focused on genes without considering drugs and diseases. Thus, the relevant literature is not free from the use of a two-stage approach.

To address this difficulty, linear algebra has often been used to identify drug-gene-disease information, since it enables the identification of drug-gene-disease information in neither a drug-centric nor disease-centric centric-manner. Wang et al. [8] applied matrix factorization to gene expression matrices for drug and disease treatments, while Kim and Cho [9] employed tensor decomposition (TD) to extract drug-gene-disease information starting from the product of ID embedding vectors for drugs, genes, and

diseases. In these approaches, because the identification of drug-gene-disease information is performed without the need for a two-stage approach, these approaches can often avoid the difficulty associated with this approach.

Another advantage of these linear-algebra-based approaches is that they are fully unsupervised. Therein, there is no need for any drug- or disease-specific information (e.g., differential expression between healthy controls and patients), and we are free to identify drug-gene-disease information in a fully data-driven manner. We recently applied a TD-based unsupervised FE [10, 11] to integrate PPI with gene expression in cancer [12]. We found that integrated analysis enhanced the coincidence with clinical labels. In this study, we also found that PPI only contained information related to cancers, even though PPI themselves are unlikely to be associated with cancers. In this paper, we attempt to determine the degree of the relationship between PPI and cancers using TD-based unsupervised FE without integrating other information (e.g. gene expression). As a result, we found that applying TD-based unsupervised FE to PPI enabled the identification of cancer-related genes that were also used for drug repositioning. After identifying that the selected proteins were likely to be hub proteins in PPI, we replaced the proteins selected via TD-based unsupervised FE with hub proteins and found that the hub proteins could be used for drug repositioning. The distinction between hub proteins and those selected using TD-based unsupervised FE was the strength of the correlation among the interactions; that is, proteins selected using TD-based unsupervised FE were found to exhibit a greater number of shared interactions. Only proteins selected using TD-based unsupervised FE had hits in the *in vivo*-based drug database DrugMatrix. The identification of proteins that not only have more interactions (i.e., hub proteins) but also more shared interactions may be the key to identifying more promising proteins for drug repositioning.

2 Results

Figure 1 illustrates the analysis conducted in this study.

2.1 TD-based unsupervised FE

The first matrix to be integrated was human PPI, $n^1_{ii'} \in \mathbb{R}^{24875\times24875}$ (for BioGRID) or $n^1_{ii'} \in \mathbb{R}^{4901\times4901}$ (for DIP), and the second was mouse PPI, $n^2_{ii'} \in \mathbb{R}^{16645\times16645}$ (for BioGRID) or $n^2_{ii'} \in \mathbb{R}^{2342\times2342}$ (for DIP). Since the number of common protein, N_{common}, is 9688 (for BioGRID) or 1097 (for DIP), and the number of orthogonal protein, N_{ortho}, is 4026 (for BioGRID) or and 432 (for DIP), the resulting integrated tensor is $n_{ii'k} \in \mathbb{R}^{N\times N\times 2}$ where $N = 24875 + 16645 - 9688 - 4026 = 27806$ (for BioGRID) or $N = 4901 + 2342 - 1097 - 432 = 5714$ (for DIP). After applying HOSVD to $n_{ii'k}$, we obtained eq. (7). P_is were attributed to is usingu_{2i} (for BioGRID and DIP) or u_{3i} (only for DIP) using Eq. (6) with replacing $u_{\ell i}$ with $u_{\ell_1 i}$ and are corrected using the BH criterion [10, 11] (The reason why u_{3i} was considered only for DIP was because genes selected by u_{3i} for DIP share the same enriched diseases, cancers, with those by u_{2i}, which was not the case for BioGRID. For more details, see the latter part of this paper.) Thus, 195 (using u_{2i} for BioGRID) and 196 (using u_{2i} for DIP) and 59 (using u_{3i} for DIP) proteins were associated with adjusted P-values less than

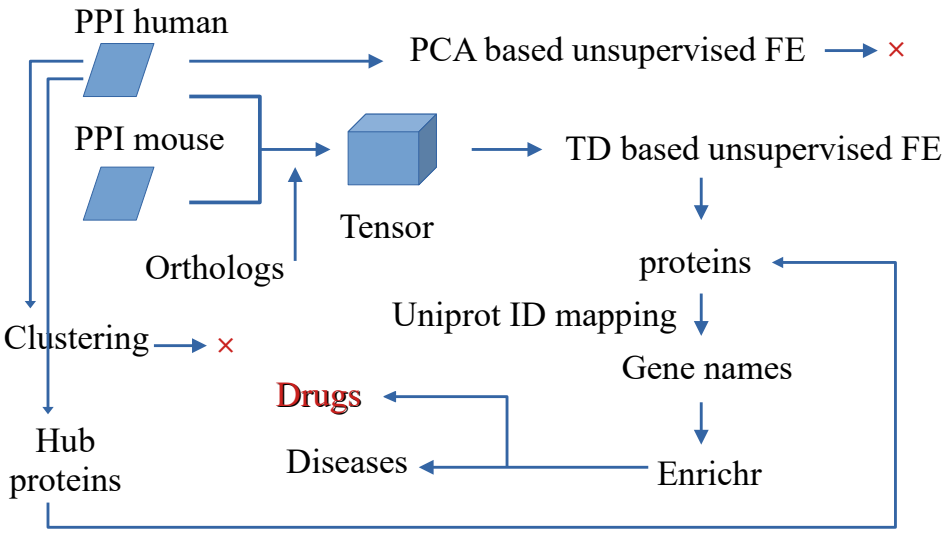

Fig. 1 Flowchart of the analyses in this study. PCA-based unsupervised FE was applied to human PPI, but failed. TD-based unsupervised FE was applied to tensor generated from human and mouse PPI. Gene names associated with the identified proteins were uploaded to Enrichr to identify associated diseases and drugs. Hub proteins and proteins selected via cluster analyses were tested and used for the identification of diseases and drugs.

0.01. The Uniprot accession numbers associated with these proteins were converted to 217 (using u_{2i} for BioGRID), 193 (using u_{2i} for DIP), and 57 (using u_{3i} for DIP) gene names by Uniprot ID mapping, respectively (see Supplementary Information for the list of proteins and gene names). 217, 193, and 57 gene names were uploaded to Enrichr [13] for evaluation purposes.

Next, the types of diseases associated with three sets of gene names were identified. Firstly, we considered the category of "Jensen Diseases". For the 217 gene names selected by u_{2i} on BioGrid (Table 1), not only were there highly significant diseases, but most were cancers or tumors ("cancer," "stomach cancer," "adenoma," "immune system cancer," "ovarian cancer," "ductal carcinoma in situ," "esophageal carcinoma," and "lymphoid leukemia").

For the 193 gene names selected by u_{2i} for the DIP (see Table 2), not only were there highly significant diseases, but more than half were cancers or tumors ("lymphoid leukemia," "immune system cancer," "cancer," "biliary tract cancer," and "ovarian cancer"). Additionally, "familial adenomatous polyposis" is known to develop into cancer [14]. For the 57 gene names selected by u_{2i} for DIP (Table 3), not only were there highly significant diseases, but more than half were cancer or tumors ("ovarian cancer," "immune system cancer," "breast cancer," "biliary tract cancer," "ductal carcinoma in situ," and "hereditary breast ovarian cancer"). In addition, "Li-Fraumeni syndrome" is known to be the cause of several cancers [15].

Next, we considered the "OMIM Diseases" category of Enrichr.

Table 1 Top 10 diseases in the "Jensen Diseases" category of Enrichr for 217 gene names selected by u_{2i} for BioGrid

Term	Overlap	P-value	Adjusted P-value
Cancer	16/300	1.93×10^{-7}	4.53×10^{-5}
Stomach cancer	5/14	2.65×10^{-7}	4.53×10^{-5}
DOID:9917	6/30	7.27×10^{-7}	7.77×10^{-5}
Adenoma	4/8	9.12×10^{-7}	7.77×10^{-5}
Immune system cancer	10/137	2.54×10^{-6}	1.73×10^{-4}
Upper respiratory tract disease	4/13	8.93×10^{-6}	5.07×10^{-4}
Ovarian cancer	6/51	1.82×10^{-5}	8.89×10^{-4}
Ductal carcinoma in situ	4/18	3.66×10^{-5}	1.48×10^{-3}
Esophageal carcinoma	3/7	4.27×10^{-5}	1.48×10^{-3}
Lymphoid leukemia	9/152	4.33×10^{-5}	1.48×10^{-3}

Table 2 Top 10 diseases in the "Jensen Diseases" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Lymphoid leukemia	13/152	2.51×10^{-9}	8.33×10^{-7}
Immune system cancer	11/137	8.40×10^{-8}	1.39×10^{-5}
Cancer	15/300	2.07×10^{-7}	2.29×10^{-5}
Familial adenomatous polyposis	5/25	3.43×10^{-6}	2.85×10^{-4}
Upper respiratory tract disease	4/13	5.39×10^{-6}	3.58×10^{-4}
Biliary tract cancer	4/14	7.49×10^{-6}	4.14×10^{-4}
Neurodegenerative disease	11/293	1.24×10^{-4}	5.15×10^{-3}
Ovarian cancer	5/51	1.24×10^{-4}	5.15×10^{-3}
Systemic lupus erythematosus	6/83	1.44×10^{-4}	5.31×10^{-3}
Ataxia telangiectasia	4/30	1.82×10^{-4}	6.04×10^{-3}

For the 217 gene names selected by u_{2i} for BioGrid (Table 4), six diseases were cancer or tumors ("breast cancer," "ovarian cancer," "thyroid carcinoma," "prostate cancer," "colorectal cancer," "pancreatic cancer," and "gastric cancer"), although stastical (i.e., associated with adjusted P-values less than 0.05). significant was only observed for two. In contrast to BioGRID, which failed to identify sufficiently large significant diseases, 193 gene names selected by u_{2i} for DIP (Table 5), not only were there highly significant diseases, but four diseases were related to cancers or tumors ("colorectal cancer," "breast cancer," "leukemia," and "lymphoma"). Additionally, "Fanconi anemia" is also often associated with cancer [16]. For 57 gene names selected by u_{3i} for DIP (Table 6), not only there were highly significant diseases, but also more than half of the diseases were cancers or tumors ("pancreatic cancer," "ovarian cancer," "colorectal cancer," "breast cancer," "prostate cancer," and "melanoma"). In addition, "Fanconi anemia" was also identified. In conclusion, these three sets of proteins were found to be highly related to cancers and tumors regardless of the PPI datasets used, excluding the 217 gene names selected by u_{2i} for BioGRID.

Next, we considered potential drug repositioning using these gene sets. Multiple categories can be used for drug repositioning in Enrichr: "LINCS L1000 Chem Pert Consensus Sigs," "DSigDB," "DrugMatrix," "Drug Perturbations from GEO down,"

Table 3 Top 10 diseases in the "Jensen Diseases" category of Enrichr for 57 gene names selected by u_{3i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Li-Fraumeni syndrome	5/12	9.26×10^{-11}	1.38×10^{-8}
Ovarian cancer	6/51	4.77×10^{-9}	3.56×10^{-7}
Immune system cancer	6/137	1.82×10^{-6}	8.19×10^{-5}
Breast cancer	9/434	2.20×10^{-6}	8.19×10^{-5}
Lymphoid leukemia	6/152	3.34×10^{-6}	9.94×10^{-5}
Biliary tract cancer	3/14	6.63×10^{-6}	1.65×10^{-4}
Ductal carcinoma in situ	3/18	1.48×10^{-5}	3.14×10^{-4}
Nijmegen breakage syndrome	3/22	2.76×10^{-5}	5.15×10^{-4}
ENSP00000256078	3/28	5.81×10^{-5}	8.91×10^{-4}
Hereditary breast ovarian cancer	2/5	7.12×10^{-5}	8.91×10^{-4}

Table 4 Top 10 diseases in the "OMIM Diseases" category of Enrichr for 217 gene names selected by u_{2i} for BioGrid

Term	Overlap	P-value	Adjusted P-value
Parkinson disease	3/22	1.67×10^{-3}	4.33×10^{-2}
Breast cancer	3/28	3.38×10^{-3}	4.39×10^{-2}
Dementia	2/12	7.20×10^{-3}	5.49×10^{-2}
Ovarian cancer	2/13	8.45×10^{-3}	5.49×10^{-2}
Thyroid carcinoma	2/17	1.43×10^{-2}	7.44×10^{-2}
Autism	2/19	1.77×10^{-2}	7.69×10^{-2}
Prostate cancer	2/30	4.18×10^{-2}	1.55×10^{-1}
Colorectal cancer	2/40	6.99×10^{-2}	2.27×10^{-1}
Pancreatic cancer	1/11	1.13×10^{-1}	2.30×10^{-1}
Gastric cancer	1/11	1.13×10^{-1}	2.30×10^{-1}

and "Drug Perturbations from GEO up." All of these categories return a list of compounds that are supposed to significantly target a set of uploaded genes. Because three sets of gene names are supposed to be deeply related to cancers and tumors, drugs that target these genes can be used to treat tumors and cancers. A list of the top 10 compounds in the "LINCS L1000 Chem Pert Consensus Sigs" category for 217 gene names selected by u_{2i} for BioGrid and 193 gene names selected by u_{2i} for DIP (no significant hit for 57 gene names selected by u_{3i} for DIP is provided in Tables 7 and 8). Table 9, 10 and 11 list the top 10 compounds in the "DSigDB" category for 217 gene names selected by u_{2i} for BioGRID, 193 gene names selected by u_{2i} for DIP, and 57 gene names selected by u_{3i} for DIP. Table s 12 and 13 list the top 10 compounds in the "DrugMatrix" category for 217 gene names selected by u_{2i} for BioGRID and 193 gene names selected by u_{2i} for DIP (no significant hits for 57 gene names selected by u_{3i} for DIP). Table 14, 15 and 16 list the top 10 compounds in the "Drug Perturbations from GEO down" category for 217 gene names selected by u_{2i} for BioGRID, 193 gene names selected by u_{2i} for DIP, and 57 gene names selected by u_{3i} for DIP.

Table 17, 18, and 19 list the top 10 compounds in the "Drug Perturbations from GEO up" category for 217 gene names selected by u_{2i} for BioGRID, 193 gene names selected by u_{2i} for DIP, and 57 gene names selected by u_{3i} for DIP, respectively.

Table 5 Top 10 diseases in the "OMIM Diseases" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Colorectal cancer	7/40	9.27×10^{-8}	2.97×10^{-6}
Parkinson disease	5/22	1.74×10^{-6}	2.78×10^{-5}
Ectodermal dysplasia	3/11	1.34×10^{-4}	1.43×10^{-3}
Breast cancer	3/28	2.36×10^{-3}	1.51×10^{-2}
Immunodeficiency	3/28	2.36×10^{-3}	1.51×10^{-2}
Fanconi anemia	2/13	6.60×10^{-3}	3.03×10^{-2}
Leukemia	4/78	6.64×10^{-3}	3.03×10^{-2}
Myocardial infarction	2/16	9.97×10^{-3}	3.99×10^{-2}
Orofacial cleft	2/19	1.39×10^{-2}	4.96×10^{-2}
Lymphoma	2/22	1.85×10^{-2}	5.92×10^{-2}

Table 6 Top 10 diseases in the "OMIM Diseases" category of Enrichr for 57 gene names selected by u_{3i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Pancreatic cancer	3/11	3.02×10^{-6}	3.13×10^{-5}
Ovarian cancer	3/13	5.22×10^{-6}	3.13×10^{-5}
Colorectal cancer	3/40	1.71×10^{-4}	6.85×10^{-4}
Orofacial cleft	2/19	1.19×10^{-3}	3.56×10^{-3}
Breast cancer	2/28	2.59×10^{-3}	5.93×10^{-3}
Prostate cancer	2/30	2.97×10^{-3}	5.93×10^{-3}
Ectodermal dysplasia	1/11	2.93×10^{-2}	4.61×10^{-2}
Fanconi anemia	1/13	3.45×10^{-2}	4.61×10^{-2}
Melanoma	1/13	3.45×10^{-2}	4.61×10^{-2}
Immunodeficiency	1/28	7.30×10^{-2}	8.75×10^{-2}

As can be seen above, although it is not true for all categories, three sets of gene names were often associated with lists of drugs that significantly target these sets of gene names. Thus, we can conclude that our proposed gene-centric drug repositioning strategy can be performed employing TD-based unsupervised FE to select a set of proteins based on PPI networks , regardless of the PPI dataset employed.

2.2 PCA-based unsupervised FE

One might wonder why we needed to integrate human and mouse PPI. Simply using only human PPI might result in a similar or even better performance. To address this problem, we applied PCA-based unsupervised FE [10, 11] to human PPI and attempted to obtain a set of proteins associated with adjusted P-values less than 0.01. Nevertheless, when considering the u_{2i} for BioGRID, although 158 proteins were associated with adjusted P-values less than 0.01, the "DrugMatrix" category failed to identify enriched drug for gene names associated with these proteins. Similarly, when considering u_{2i} for DIP, only one protein was found to be associated with adjusted P-values of less than 0.01. This results suggest that TD-based unsupervised FE is superior to PCA-based unsupervised FE and thereby worth testing.

Table 7 Top 10 drugs in the "LINCS L1000 Chem Pert Consensus Sigs" category of Enrichr for 217 gene names selected by u_{2i} for BioGrid.

Term	Overlap	P-value	Adjusted P-value
Nutlin-3 Down	15/240	6.03×10^{-8}	5.18×10^{-4}
Amisulpride Up	15/249	9.78×10^{-8}	5.18×10^{-4}
Pyrvinium-Pamoate Down	14/243	4.54×10^{-7}	1.35×10^{-3}
Actinomycin-D Down	14/248	5.80×10^{-7}	1.35×10^{-3}
AR-C133057XX Up	14/250	6.39×10^{-7}	1.35×10^{-3}
CVF-SUMO-11 Up	13/241	2.45×10^{-6}	2.97×10^{-3}
KPT-330 Down	13/243	2.69×10^{-6}	2.97×10^{-3}
TAS-103 Down	13/243	2.69×10^{-6}	2.97×10^{-3}
XL-888 Up	13/245	2.94×10^{-6}	2.97×10^{-3}
APO-866 Up	13/246	3.08×10^{-6}	2.97×10^{-3}

Table 8 Top 10 drugs in the "LINCS L1000 Chem Pert Consensus Sigs" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
KU-0060648 Up	14/236	6.59×10^{-8}	3.41×10^{-4}
Okadaic-Acid Up	14/246	1.10×10^{-7}	3.41×10^{-4}
Palmatine-Chloride Down	14/247	1.16×10^{-7}	3.41×10^{-4}
YL-55 Up	14/249	1.28×10^{-7}	3.41×10^{-4}
CHEMBL-399379 Up	13/231	3.61×10^{-7}	6.87×10^{-4}
LY-2606368 Up	13/233	3.98×10^{-7}	6.87×10^{-4}
RS-102221 Down	13/242	6.13×10^{-7}	6.87×10^{-4}
4-HQN Down	13/242	6.13×10^{-7}	6.87×10^{-4}
Butalbital Down	13/244	6.73×10^{-7}	6.87×10^{-4}
TL-HRAS-61 Up	13/247	7.72×10^{-7}	6.87×10^{-4}

2.3 Cluster analyses

However, another concern is whether TD-based unsupervised FE is required when selecting a set of proteins since several methods exist with which to select sub-clusters within large networks. Such methods can replace TD-based unsupervised FE in the selection of a set of proteins. Following Zhao et al [17], we employed three methods, multi-level [18], label-progration [19], and edge-betweenness [20], which enabled us to identify sub-clusters. When considering PPI human for BioGrid, edge-betweenness failed to converge after waiting more than eight hours (i.e longer than the ten minutes required for TD-based unsupervised FE, which includes the time consuming process of generating tensor) $n_{ii'k} \in \mathbb{R}^{N \times N \times 2}$), and the first and the second clusters generated by multi-level were too large (6406 and 2776). The first and the second clusters generated by label-progration were 24501 and 1, which was not reasonable at all. However, when human PPI for DIP was considered, the results improved. Table 20 lists the performances for the identification of diseases and drugs significantly associated with selected proteins in individual identified sub-clusters (label-progration could not identify large enough clusters to be used for enrichment analyses). It is clear that the tensor method outperforms the clustering methods , even for DIP. In conclusion, in

Table 9 Top 10 drugs in the "DSigDB" category of Enrichr for 217 gene names selected by u_{2i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
troglitazone CTD 00002415	30/651	2.31×10^{-11}	6.72×10^{-8}
Vorinostat CTD 00003560	24/425	4.48×10^{-11}	6.72×10^{-8}
MG-132 CTD 00002789	15/173	7.04×10^{-10}	6.01×10^{-7}
resveratrol CTD 00002483	46/1601	8.01×10^{-10}	6.01×10^{-7}
N-Acetyl-L-cysteine CTD 00005305	19/315	1.77×10^{-9}	1.06×10^{-6}
doxorubicin CTD 00005874	29/750	3.00×10^{-9}	1.50×10^{-6}
Bortezomib CTD 00003736	30/878	2.61×10^{-8}	1.12×10^{-5}
etoposide CTD 00005948	21/461	3.43×10^{-8}	1.15×10^{-5}
piroxicam CTD 00006571	23/549	3.45×10^{-8}	1.15×10^{-5}
Arsenenous acid CTD 00000922	37/1283	4.47×10^{-8}	1.34×10^{-5}

Table 10 Top 10 drugs in the "DSigDB" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
curcumin CTD 00000663	34/528	8.71×10^{-19}	2.63×10^{-15}
doxorubicin CTD 00005874	38/750	2.02×10^{-17}	3.05×10^{-14}
N-Acetyl-L-cysteine CTD 00005305	25/315	4.17×10^{-16}	4.19×10^{-13}
Arsenenous acid CTD 00000922	47/1283	6.07×10^{-16}	4.58×10^{-13}
MG-132 CTD 00002789	19/173	4.61×10^{-15}	2.78×10^{-12}
wortmannin CTD 00000504	17/146	5.05×10^{-14}	2.54×10^{-11}
LY 294002 CTD 00003061	20/227	6.01×10^{-14}	2.59×10^{-11}
ZINC CTD 00007011	50/1642	8.50×10^{-14}	3.21×10^{-11}
5-Fluorouracil CTD 00005987	42/1202	1.48×10^{-13}	4.96×10^{-11}
etoposide CTD 00005948	25/461	2.38×10^{-12}	7.18×10^{-10}

contrast to conventional cluster analysis in which cluster sizes varied heavily depending on the PPI dataset used, using TD-based unsupervised FE can provide a stable and reasonable size for a set of genes. Thus, TD-based unsupervised FE appears to be superior to conventional cluster analysis.

2.4 Other species

Although we only used human and mouse PPI, it is possible to employ other combinations because BioGRID and DIP to include more species-specific PPI. To validate this, we initially examined a combination of human and rat PPI for BioGRID. However, this outcome was not very promising. Next, 271 proteins associated with adjusted P-values less than 0.01 were selected, and the gene names associated with these proteins were uploaded. The "Jensen Diseases" and "OMIM Diseases" categories identified two and zero diseases enriched with the uploaded gene names, respectively. Because the identification of enriched diseases is the starting point, if the process fails at this point, there are no way to proceed. The reason for this is the inadequateness of rat PPI, which includes too few PPI; integrating human and rat PPI results in the wrong conclusion, namely that missing PPI in rats indicates a lack of interactions, although this may simply mean that PPI in rats has not been thoroughly investigated. Thus,

Table 11 Top 10 drugs in the "DSigDB" category of Enrichr for 57 gene names selected by u_{3i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Vorinostat CTD 00003560	23/425	1.11×10^{-24}	2.69×10^{-21}
FENRETINIDE CTD 00007166	16/228	6.88×10^{-19}	8.36×10^{-16}
curcumin CTD 00000663	20/528	2.65×10^{-18}	2.15×10^{-15}
resveratrol CTD 00002483	28/1601	3.80×10^{-17}	2.31×10^{-14}
doxorubicin CTD 00005874	21/750	1.40×10^{-16}	6.79×10^{-14}
aspirin CTD 00005447	19/561	1.77×10^{-16}	7.17×10^{-14}
bay 11-7082 CTD 00003959	11/83	2.62×10^{-16}	9.09×10^{-14}
Arsenenous acid CTD 00000922	25/1283	3.34×10^{-16}	1.02×10^{-13}
tamoxifen CTD 00006827	20/802	8.12×10^{-15}	2.12×10^{-12}
Leptomycin B CTD 00001805	9/50	8.70×10^{-15}	2.12×10^{-12}

Table 12 Top 10 drugs in the "DrugMatrix" category of Enrichr for 217 gene names selected by u_{2i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
Azauridine-1500 mg/kg in CMC-Rat-Bone marrow-3d-up	16/311	3.13×10^{-7}	1.89×10^{-3}
Lead (II) Acetate-200 mg/kg in Saline-Rat-Spleen-0.25d-up	16/330	6.91×10^{-7}	1.89×10^{-3}
Famciclovir-112 mg/kg in Saline-Rat-Bone marrow-1d-dn	15/291	7.23×10^{-7}	1.89×10^{-3}
44'-Methylenedianiline-296 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	15/304	1.25×10^{-6}	2.45×10^{-3}
Rabeprazole-1024 mg/kg in Water-Rat-Liver-1d-up	14/279	2.34×10^{-6}	3.43×10^{-3}
Carbon Tetrachloride-3178 mg/kg in Saline-Rat-Liver-1d-up	15/326	2.96×10^{-6}	3.43×10^{-3}
Diethylstilbestrol-280 mg/kg in Corn Oil-Rat-Liver-1d-up	14/290	3.67×10^{-6}	3.43×10^{-3}
Clobetasol Propionate-17 mg/kg in Corn Oil-Rat-Bone marrow-3d-dn	14/294	4.30×10^{-6}	3.43×10^{-3}
6-Mercaptopurine-25 mg/kg in Corn Oil-Rat-Liver-0.25d-up	14/299	5.22×10^{-6}	3.43×10^{-3}
Cyclosporin A-70 mg/kg in Corn Oil-Rat-Spleen-0.25d-up	15/347	6.31×10^{-6}	3.43×10^{-3}

combinations other than that between human and mouse PPI will not work until a more comprehensive PPI for species close to humans can be obtained.

In conclusion, PPI can be a useful source of information for drug repositioning when used in combination with TD.

Table 13 Top 10 drugs in the "DrugMatrix" category of Enrichr for 193 gene names selected by u_{2i} for DIP .

Term	Overlap	P-value	Adjusted P-value
Ceramide C2-44 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	16/311	5.40×10^{-8}	4.14×10^{-4}
Ceramide C2-44 uM in DMSO-Rat-Primary rat hepatocytes-1d-up	13/280	3.12×10^{-6}	1.19×10^{-2}
Cerivastatin-7 mg/kg in Corn Oil-Rat-Skeletal muscle-5d-up	12/274	1.35×10^{-5}	3.46×10^{-2}
Methyl Salicylate-444 mg/kg in Corn Oil-Rat-Kidney-3d-up	12/290	2.37×10^{-5}	4.37×10^{-2}
Lithocholic Acid-74 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	12/300	3.31×10^{-5}	4.37×10^{-2}
Stavudine-1120 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	12/301	3.42×10^{-5}	4.37×10^{-2}
Carbon Tetrachloride-400 mg/kg in Corn Oil-Rat-Liver-0.25d-up	11/280	8.30×10^{-5}	9.09×10^{-2}
Microcystin-LR-0.07 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	11/309	1.97×10^{-4}	1.66×10^{-1}
Pantoprazole-650 uM in DMSO-Rat-Primary rat hepatocytes-0.67d-up	11/310	2.03×10^{-4}	1.66×10^{-1}
Daunorubicin-1.5 uM in DMSO-Rat-Primary rat hepatocytes-1d-up	11/324	2.96×10^{-4}	1.66×10^{-1}

Table 14 Top 10 drugs in the "Drug Perturbations from GEO down" category of Enrichr for 217 gene names selected by u_{2i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
cytarabine, EC50, 1 d 6253 human GSE6930 sample 3373	24/339	3.73×10^{-13}	3.14×10^{-10}
ARC (NSC 188491) 302576 human GSE13477 sample 2503	22/329	1.16×10^{-11}	4.16×10^{-9}
doxorubicin, EC50, 1 d 31703 human GSE6930 sample 3260	21/300	1.48×10^{-11}	4.16×10^{-9}
lapatinib DB01259 human GSE38376 sample 2586	20/291	6.43×10^{-11}	1.35×10^{-8}
lapatinib DB01259 human GSE38376 sample 2585	20/301	1.18×10^{-10}	1.98×10^{-8}
EPZ004777 56962336 human GSE29828 sample 2650	22/376	1.54×10^{-10}	2.17×10^{-8}
sangivamycin (NSC 65346) 14978 human GSE13477 sample 2504	20/330	5.95×10^{-10}	7.17×10^{-8}
bexarotene DB00307 human GSE6914 sample 2680	23/453	9.32×10^{-10}	9.82×10^{-8}
cytarabine, EC50, 5 d 6253 human GSE6930 sample 3375	23/468	1.74×10^{-9}	1.63×10^{-7}

Table 15 Top 10 drugs in the "Drug Perturbations from GEO down" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
dactinomycin DB00970 mouse GSE5324 sample 2701	22/522	6.08×10^{-9}	5.20×10^{-6}
dactinomycin DB00970 mouse GSE5324 sample 2700	22/544	1.28×10^{-8}	5.48×10^{-6}
cisplatin DB00515 human GSE6410 sample 2532	15/315	3.86×10^{-7}	1.10×10^{-4}
SODIUM BUTYRATE 5222465 mouse GSE4410 sample 3582	14/289	7.78×10^{-7}	1.33×10^{-4}
Etanercept DB00005 human GSE7524 sample 3295	16/379	7.80×10^{-7}	1.33×10^{-4}
Curcumin CID 969516 human GSE10896 sample 3259	17/458	2.04×10^{-6}	2.91×10^{-4}
actinomycin D 2019 human GSE6400 sample 3098	14/323	2.88×10^{-6}	3.52×10^{-4}
SODIUM BUTYRATE 5222465 mouse GSE4410 sample 3584	13/343	2.72×10^{-5}	2.91×10^{-3}
lapatinib DB01259 human GSE38376 sample 2585	12/301	3.42×10^{-5}	3.25×10^{-3}
HYPOCHLOROUS ACID 24341 human GSE11630 sample 3202	11/289	1.10×10^{-4}	9.07×10^{-3}

Table 16 Top 10 drugs in the "Drug Perturbations from GEO down" category of Enrichr for 57 gene names selected by u_{3i} for DIP.

Term	Overlap	P-value	Adjusted P-value
trastuzumab DB00072 human GSE31432 sample 3123	7/280	9.87×10^{-6}	2.69×10^{-3}
ZINC ACETATE 11192 human GSE2964 sample 3589	6/189	1.17×10^{-5}	2.69×10^{-3}
lapatinib DB01259 human GSE38376 sample 2585	7/301	1.58×10^{-5}	2.69×10^{-3}
doxycycline DB00254 human GSE2624 sample 3074	8/425	1.72×10^{-5}	2.69×10^{-3}
genistein 5280961 mouse GSE2889 sample 3606	6/247	5.25×10^{-5}	6.57×10^{-3}
doxycycline DB00254 human GSE2624 sample 3077	7/391	8.30×10^{-5}	8.37×10^{-3}
2,3,7,8-tetrachlorodibenzo-p-dioxin 10852289 mouse GSE2812 sample 3602	7/400	9.57×10^{-5}	8.37×10^{-3}
Bisphenol A 6623 human GSE17624 sample 2654	6/281	1.07×10^{-4}	8.37×10^{-3}
lapatinib DB01259 human GSE38376 sample 2586	6/291	1.29×10^{-4}	9.00×10^{-3}
estradiol 5757 mouse GSE2889 sample 3609	7/433	1.56×10^{-4}	9.78×10^{-3}

3 Discussion

To understand the set of genes selected in this analysis, we computed the mean number of bindings of individual proteins, defined as

$$\langle n_{ii'1} \rangle_i = \frac{1}{N} \sum_{i'=1}^{N} n_{ii'1} \tag{1}$$

Table 17 Top 10 drugs in the "Drug Perturbations from GEO up" category of Enrichr for 217 gene names selected by u_{2i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
Dmnq (2,3-Dimethoxy-1,4-naphthoquinone) 3136 human GSE6907 sample 3527	20/321	3.67×10^{-10}	1.68×10^{-7}
Arachidonic acid DB04557 human GSE3737 sample 3171	22/395	3.93×10^{-10}	1.68×10^{-7}
NICKEL 935 human GSE6907 sample 3531	18/288	2.75×10^{-9}	6.94×10^{-7}
MT19c compound SID 134222379 human GSE23616 sample 3352	18/291	3.24×10^{-9}	6.94×10^{-7}
quercetin 5280343 human GSE7259 sample 3415	19/336	5.10×10^{-9}	8.76×10^{-7}
ARSENIC 5359596 human GSE6907 sample 3529	20/394	1.24×10^{-8}	1.77×10^{-6}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3042	19/362	1.71×10^{-8}	1.94×10^{-6}
MLN4924 73014187 human GSE30531 sample 2644	18/325	1.81×10^{-8}	1.94×10^{-6}
nicotine 942 human GSE6264 sample 3064	19/370	2.42×10^{-8}	2.16×10^{-6}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3045	18/332	2.51×10^{-8}	2.16×10^{-6}

Table 18 Top 10 drugs in the "Drug Perturbations from GEO up" category of Enrichr for 193 gene names selected by u_{2i} for DIP.

Term	Overlap	P-value	Adjusted P-value
bortezomib DB00188 human GSE30931 sample 2686	18/364	1.41×10^{-8}	1.23×10^{-5}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3045	15/332	7.50×10^{-7}	2.53×10^{-4}
phenol 996 human GSE6907 sample 3528	15/336	8.72×10^{-7}	2.53×10^{-4}
resveratrol DB02709 human GSE36930 sample 3498	15/349	1.40×10^{-6}	2.96×10^{-4}
Promegestone 36709 human GSE67561 sample 3694	13/265	1.70×10^{-6}	2.96×10^{-4}
estradiol 5757 mouse GSE2195 sample 3619	13/271	2.17×10^{-6}	3.16×10^{-4}
Bisphenol A 6623 mouse GSE4650 sample 3574	11/202	4.02×10^{-6}	5.01×10^{-4}
pioglitazone DB01132 rat GSE20219 sample 2794	13/292	4.92×10^{-6}	5.37×10^{-4}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3044	13/300	6.60×10^{-6}	6.39×10^{-4}
NVP-TNKS656 (tankyrase inhibitor) 71227201 human GSE55624 sample 3247	13/307	8.45×10^{-6}	7.37×10^{-4}

and compared $\langle n_{ii'1} \rangle_i$ between the selected and non-selected proteins for the human PPI. As can be seen in Table 21, $\langle n_{ii'1} \rangle_i$ for 217 proteins selected by u_{2i} for BioGRID, 195 proteins selected by u_{2i} for DIP, and 55 proteins selected by u_{3i} for DIP were always significantly larger than those of other (i.e. not selected) proteins. These results suggest that TD-based unsupervised FE selects proteins with more interactions in a fully data-driven and unsupervised manner.

Since proteins selected via TD-based unsupervised FE have more interactions, we proceeded to verify whether proteins were being selected for use in drug repositioning simply due to their greater number of interactions (hereafter, denoted as hub proteins).

Table 19 Top 10 drugs in the "Drug Perturbations from GEO up" category of Enrichr for 57 gene names selected by u_{3i} for DIP.

Term	Overlap	P-value	Adjusted P-value
cisplatin DB00515 mouse GSE6206 sample 3412	8/279	7.74×10^{-7}	3.96×10^{-4}
Bisphenol A 6623 mouse GSE4650 sample 3574	7/202	1.14×10^{-6}	3.96×10^{-4}
lipopolysaccharide (LPS) 11970143 mouse GSE23639 sample 3443	7/261	6.24×10^{-6}	1.44×10^{-3}
Mehp 20393 rat GSE4514 sample 3578	7/293	1.32×10^{-5}	1.84×10^{-3}
Hydrogen Peroxide 784 mouse GSE3078 sample 3599	6/199	1.56×10^{-5}	1.84×10^{-3}
levetiracetam 5284583 rat GSE2880 sample 2669	6/201	1.65×10^{-5}	1.84×10^{-3}
2,3,7,8-tetrachlorodibenzo-p-dioxin 10852289 human GSE24193 sample 3189	7/309	1.87×10^{-5}	1.84×10^{-3}
Triiodothyronine-[13C6] hydrochloride (T3 thyronine) 53442265 human GSE29159 sample 3043	7/320	2.34×10^{-5}	2.02×10^{-3}
Bisphenol A 6623 mouse GSE4650 sample 3575	6/225	3.12×10^{-5}	2.40×10^{-3}
Phosgene 6371 mouse GSE2565 sample 3611	6/243	4.80×10^{-5}	3.32×10^{-3}

Table 20 Performances of various protein selection methods for DIP. ◯s denotes cases with more than or equal to 10 drugs/associated with adjusted P-values less than 0.05. Otherwise, the corresponding practical numbers are presented.

Methods	Tensor		multi-level		edge-betweenness	
Order	2nd	3rd	1st	2nd	1st	2nd
# of proteins	195	55	144	138	39	82
# of genes	193	57	128	164	30	111
Jensen Diseases	◯	◯	◯	◯	7	5
OMIM Disease	9	9	8	0	0	0
LINCS L1000 Chem Pert Consensus Sigs	◯	0	4	1	0	◯
DSigDB	◯	◯	◯	◯	0	◯
DrugMatrix	6	0	0	0	0	0
Drug Perturbations from GEO down	◯	◯	0	0	0	2
Drug Perturbations from GEO up	◯	◯	6	0	1	0

To verify this, we selected the top 200 and 50 proteins with more interactions (these numbers were selected to be close to the number of proteins selected using u_{2i} and u_{3i}. See Supplementary Information for a list of proteins and gene names) using human PPI. Table 22 lists the confusion matrices for the selected proteins. Although they significantly overlapped because the majority were not shared, distinct sets of proteins were identified.

We uploaded the associated gene names to Enrichr to evaluate the top 200 and 50 hub proteins. Table 23 lists the performance achieved by the hub proteins. (for details on the top diseases and drugs in individual categories: see Table S1 to S7 for the top 200 hub proteins for BioGRID, Table S8 to S14 for the top 200 hub proteins for DIP, and Table S15 to S21 for the top 50 hub proteins for DIP). Their performances were essentially the same as those of TD-based unsupervised FE, excluding DrugMatrix for DIP.

Table 21 To compare $\langle n_{ii'1} \rangle_i$ between the selected and not selected proteins when using various $u_{\ell_1 i}$ to select proteins, P-values were computed using the t test based on the alternative hypothesis that the mean $\langle n_{ii'1} \rangle_i$ of the selected proteins was larger that that of not selected proteins.

$u_{\ell_1 i}$s used	u_{2i}		u_{3i}		u_{6i}	
	Selected	Not selected	Selected	Not selected	Selected	Not selected
BioGRID						
Mean $\langle n_{ii'1} \rangle_i$	3.63×10^{-2}	1.84×10^{-3}	1.97×10^{-2}	1.78×10^{-3}	—	—
P-value	5.32×10^{-37}		2.53×10^{-59}		—	
DIP						
Mean $\langle n_{ii'1} \rangle_i$	1.12×10^{-3}	4.17×10^{-4}	3.12×10^{-3}	4.13×10^{-4}	2.14×10^{-3}	4.26×10^{-4}
P-value	2.4×10^{-4}		5.3×10^{-5}		2.04×10^{-13}	

Table 22 Confusion matrix of selected proteins between TD-based unsupervised FE and hub proteins.

TD based unsupervised FE		Hub proteins Top 200	
		Not selected	Selected
BioGRID			
Using u_{2i}	Not selected	27526	80
	Selected	85	115
	Odds ratio	465.58	
	P-value	2.88×10^{-209}	
DIP			
Using u_{2i}	Not selected	5353	168
	Selected	165	28
	Odds ratio	5.4	
	P-value	4.55×10^{-11}	
		Top 50	
		Not selected	Selected
Using u_{3i}	Not selected	5625	42
	Selected	30	17
	Odds ratio	75.4	
	P-value	3.04×10^{-23}	

Comparing the performance of TD-based unsupervised FE with hub proteins, to the best of our knowledge, there are no other studies that use hub proteins derived from only PPI for drug repositioning without using other information. This is because hub proteins derived from PPI themselves, without any other information, cannot be used for drug repositioning for cancer. This was clear from the fact that proteins selected by TD-based unsupervised FE were hub proteins. Without TD-based unsupervised FE, we could not identify hub proteins derived only from PPI that could be used for drug repositioning for cancer. Secondly, the hub proteins for DIP failed to identify hits in DrugMatrix. DrugMatrix is an *in vivo* specific database. Thus, the hub

Table 23 Performances achieved by hub proteins. ◯s denote case with more than or equal to 10 drugs/associated with adjusted P-values less than 0.05. Otherwise, the corresponding practical numbers are presented.

Methods	Tesnor					Hub proteins		
	BioGRID		DIP			BioGRID	DIP	
Order	2nd	3rd	2nd	3rd	6th	Top 200	Top 200	Top 50
# of proteins	195	475	195	55	38	199	179	45
# of genes	217	503	193	57	41	229	190	59
Jensen Diseases	◯	6	◯	◯	◯	◯	◯	◯
OMIM Disease	2	3	9	9	◯	4	◯	6
LINCS L1000 Chem Pert Consensus Sigs	◯	◯	◯	0	6	◯	◯	0
DSigDB	◯	◯	◯	◯	◯	◯	◯	◯
DrugMatrix	◯	◯	6	0	0	◯	0	1
Drug Perturbations from GEO down	◯	◯	◯	◯	◯	◯	◯	◯
Drug Perturbations from GEO up	◯	◯	◯	◯	◯	◯	◯	◯

proteins for DIP do not have the ability to identify drugs that may be useful in *in vivo* experiments. Thus, even if hub proteins can achieve a performance similar to that of the *in vitro* experiments, TD-based unsupervised FE is still useful, at least for DIP. Third, using hub proteins, we cannot target diseases other than cancers because hub proteins are enriched mainly in cancers. For the top 200 hub proteins for BioGRID, seven cancers were identified ("stomach cancer," "adenoma," "cancer," "immune system cancer," "esophageal carcinoma," "intestinal benign neoplasm," and "lymphoid leukemia") within the top ranked diseases in the "Jensen Diseases" category and as many as five cancers ("ovarian cancer," "breast cancer," "colorectal cancer," "thyroid carcinoma," and "prostate cancer") within the top ranked diseases in the "OMIM Diseases" category (Table S1 and S2). For the top 200 hub proteins for DIP, as many as six cancers were identified ("cancer," "lymphoid leukemia," "intestinal benign neoplasm," "immune system cancer," "biliary tract cancer," and "stomach cancer") within the top ranked diseases in the "Jensen Diseases" category and as many as seven ("colorectal cancer," "ovarian cancer," "breast cancer," "leukemia," "lung cancer," "melanoma," and "pancreatic cancer") within the top ranked diseases in the "OMIM Diseases" category (Table S8 and S9). For the top 50 hub proteins for DIP, as many as five cancers ("cancer," "DOID:9917" (pleural cancer), "thyroid cancer", "esophageal carcinoma", and "intestinal benign neoplasms") were identified within the top ranked diseases in the "Jensen Diseases" category and as many as five ("colorectal cancer," "ovarian cancer," "breast cancer," "lung cancer," and "pancreatic cancer") within the top ranked diseases in the "OMIM Diseases" category (Table S15 and S16). Nevertheless, because TD-based unsupervised FE has the freedom to select $u_{\ell_1 i}$, it can provide a set of proteins enriched in diseases other than cancer. Table 24 and 25 list the diseases enriched in proteins selected by u_{3i} for BioGRID, while Table 26 and 27 list the diseases enriched in proteins selected by u_{6i} for DIP (see Supplementary Information for the list of proteins and gene names); they are mainly distinct from cancers (only one cancer ("stomach cancer") for BioGRID and only one cancer ("cancer") for DIP within the top 10 diseases in the "Jensen Diseases" category (Tables 26 and 26)

and Three cancers ("thyroid carcinoma," "gastric cancer," and "ovarian cancer ") for BioGRID and three cancers ("pancreatic cancer," "melanoma," and "lung cancer") for DIP were identified within the top 10 diseases in the "OMIM Diseases" category (Table s 25 and 27)). For these cancers, sufficient drugs exist (for detailed drug names, see Table S22–S31). Thus, the results of TD-based unsupervised FE are useful even if hub proteins can be used to identify drugs for cancers.

Table 24 Top 10 diseases in the "Jensen Diseases" category of Enrichr for gene names selected by u_{3i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
Diamond-Blackfan anemia	12/15	2.39×10^{-17}	1.35×10^{-14}
Bowen-Conradi syndrome	6/15	1.01×10^{-6}	2.86×10^{-4}
Spondylolysis	4/8	2.55×10^{-5}	4.81×10^{-3}
Frontotemporal dementia	7/39	4.67×10^{-5}	6.60×10^{-3}
Paronychia	3/6	2.99×10^{-4}	3.05×10^{-2}
Stomach cancer	4/14	3.24×10^{-4}	3.05×10^{-2}
Dent disease	4/17	7.25×10^{-4}	5.85×10^{-2}
Vaccinia	5/32	1.13×10^{-3}	6.01×10^{-2}
Spinal muscular atrophy	4/19	1.13×10^{-3}	6.01×10^{-2}
Spinocerebellar ataxia type 2	4/19	1.13×10^{-3}	6.01×10^{-2}

Table 25 Top 10 diseases in the "OMIM Diseases" category of Enrichr for gene names selected by u_{3i} for BioGRID.

Term	Overlap	P-value	Adjusted P-value
thyroid carcinoma	4/17	7.25×10^{-4}	1.52×10^{-2}
anemia	7/61	8.24×10^{-4}	1.52×10^{-2}
dementia	3/12	2.94×10^{-3}	3.62×10^{-2}
lateral sclerosis	3/19	1.14×10^{-2}	1.05×10^{-1}
gastric cancer	2/11	2.99×10^{-2}	2.21×10^{-1}
ovarian cancer	2/13	4.10×10^{-2}	2.53×10^{-1}
diabetes mellitus, type 2	2/32	1.92×10^{-1}	7.17×10^{-1}
alopecia	1/11	2.44×10^{-1}	7.17×10^{-1}
psoriasis	1/12	2.63×10^{-1}	7.17×10^{-1}
arrhythmogenic right ventricular dysplasia	1/13	2.82×10^{-1}	7.17×10^{-1}

It is worth considering the differences between hub proteins and the proteins selected by TD-based unsupervised FE if they are not identical. To determine this, the mean correlation coefficients between proteins

$$\langle r \rangle = \frac{2}{N(N-1)} \sum_{i" \neq i'} r(n_{ii'1}, n_{ii"1}) \tag{2}$$

where $r(n_{ii'1}, n_{ii"1})$ is Pearson's correlation coefficient between $\{n_{ii'1} | 1 \leq i \leq N\}$ and $\{n_{ii"1} | 1 \leq i \leq N\}$, were computed with human PPI. Table 28 lists the values of $\langle r \rangle$.

Table 26 Top 10 diseases in the "Jensen Diseases" category of Enrichr for gene names selected by u_{6i} for DIP.

Term	Overlap	P-value	Adjusted P-value
Histoplasmosis	3/17	4.57×10^{-6}	6.86×10^{-4}
Hypohidrotic ectodermal dysplasia	2/12	2.42×10^{-4}	1.81×10^{-2}
Arthritis	4/186	4.62×10^{-4}	2.31×10^{-2}
Progressive supranuclear palsy	2/24	9.95×10^{-4}	3.73×10^{-2}
Lung disease	3/119	1.61×10^{-3}	3.80×10^{-2}
Wolf-Hirschhorn syndrome	2/31	1.66×10^{-3}	3.80×10^{-2}
Chagas disease	2/32	1.77×10^{-3}	3.80×10^{-2}
Exanthem	2/36	2.24×10^{-3}	4.20×10^{-2}
Cancer	4/300	2.70×10^{-3}	4.50×10^{-2}
Gastritis	2/43	3.18×10^{-3}	4.54×10^{-2}

Table 27 Top 10 diseases in the "OMIM Diseases" category of Enrichr for gene names selected by u_{6i} for DIP.

Term	Overlap	P-value	Adjusted P-value
ectodermal dysplasia	2/11	2.02×10^{-4}	2.22×10^{-3}
immunodeficiency	2/28	1.36×10^{-3}	7.46×10^{-3}
pancreatic cancer	1/11	2.12×10^{-2}	4.23×10^{-2}
dementia	1/12	2.32×10^{-2}	4.23×10^{-2}
melanoma	1/13	2.51×10^{-2}	4.23×10^{-2}
adenoma	1/14	2.70×10^{-2}	4.23×10^{-2}
malaria	1/14	2.70×10^{-2}	4.23×10^{-2}
migraine	1/16	3.08×10^{-2}	4.23×10^{-2}
orofacial cleft	1/19	3.64×10^{-2}	4.39×10^{-2}
lung cancer	1/21	4.02×10^{-2}	4.39×10^{-2}

Because proteins selected by TD-based unsupervised FE for DIP had a substantially larger $\langle r \rangle$, the distinctions between the proteins selected by TD-based unsupervised FE and the hub proteins were correlated. The fact that a larger $\langle r \rangle$ means that proteins share the proteins with which they interact suggests that proteins that share interactions may be the key to identifying effective drugs in *in vivo* experiments. In reality, the top 200 hub proteins for BioGRID have a larger $\langle r \rangle$ and many enrichments in the "DrugMatrix" category (Table 12). Thus, we can guarantee a high correlation for hub proteins without consulting TD-based unsupervised FE, which allows us to identify genes that can be used to identify effective drugs, even in *in vivo* experiments.

Because we employed a gene-centric strategy, the drugs and diseases identified were associated with common genes. Nevertheless, one might wonder whether this always guarantees the effectiveness of the identified drugs against the diseases. The so-called docking simulation is not a good idea for validation because Enrichr is used to relate genes to drugs or diseases in order to identify relationships based on gene expression. Thus, this does not always mean that the selected genes are direct targets of drugs, but simply those whose expression is altered by drug treatment, since the genes are in the downstream pathway. However, to the best of our knowledge, search engines specifically for studies on drug-disease pairs within the proposed list of drugs and diseases do not exist. Alternatively, to achieve this, we employed a large language

Table 28 $\langle r \rangle$s and P-values computed using a one way t test

	$\langle r \rangle$	P-values
BioGRID		
Proteins selected using u_{2i}	$1,97 \times 10^{-1}$	0
Proteins selected using u_{3i}	1.60×10^{-1}	0
Top 200 proteins	2.02×10^{-1}	0
DIP		
Proteins selected using u_{2i}	3.29×10^{-1}	0
Proteins selected using u_{3i}	9.64×10^{-2}	7.70×10^{-211}
Proteins selected using u_{6i}	8.14×10^{-2}	5.84×10^{-70}
Top 200 proteins	7.49×10^{-3}	3.97×10^{-128}
Top 50 proteins	1.19×10^{-3}	6.84×10^{-18}

model (LLM). We sought to identify pairs of drugs and diseases for which research has been reported. To avoid the wrong relationships being reported by LLM due to hallucinations, we verified whether other papers existed in support of the relationship reported our LLM.

Table 29 lists the reported and validated combination of drugs and diseases between the "Jensen Diseases" category by u_{3i} for BioGRID (Table 24) and the "DSigDB" category or "DrugMatrix" category by u_{3i} for BioGRID (Table S23 or S24). The reason we employ this is simply because it is difficult to validate the correspondence between diseases and drugs if diseases are mostly composed of cancers because many cancers share effective drugs. Therefore, it is preferable to use a set of diseases other than cancer. It is evident that the combinations of drugs and diseases reported by

Table 29 Diseases and drugs whose relation is reported by LLM. Their corresponding ranks in categories appear in parentheses.

DSigDB	Jensen Disease	Reference
Verteporfin (1st)	Stomach Cancer (6th)	[21]
Clindamycin (2nd)	Paronychia (5th)	[22]
Captopril (3rd)	Frontotemporal Dementia (4th)	[23]
Glibenclamide (7th)	Spinal Muscular Atrophy (9th)	[24]
Puromycin (8th)	Spinocerebellar Ataxia Type 2 (10th)	[25]
DrugMatrix		
Mitomycin C (8th)	Stomach Cancer (6th)	[26]
Pravastatin (7th)	Frontotemporal Dementia (4h)	[27]
Miconazole (2nd)	Paronychia (6th)	[28]
NN-Dimethylformamide (1st)	Various Cancers (6th)	[29]

TD-based unsupervised FE are associated with previously published studies, despite the diverse types of diseases. Thus, we concluded that the pairs of diseases and drugs identified in this study are likely to include promising pairs.

4 Methods

4.1 Protein-protein interactions (PPI)

4.1.1 BioGRID

The PPI data were downloaded from BioGRID [30]. PPI classified by species, BIOGRID-ORGANISM-4.4.236.tab3.zip, were downloaded. Three species-specific files, BIOGRID-ORGANISM-Homo_sapiens-4.4.236.tab3.txt, BIOGRID-ORGANISM-Mus_musculus-4.4.236.tab3.txt, and BIOGRID-ORGANISM-Rattus_norvegicus-4.4.236.tab3.txt were extracted for analysis.

4.1.2 DIP

The PPI data were also downloaded from DIP [31]. The datasets used were species-specific sets for *Homo Sapiens*, Hsapi20170205, and *Mus musculus*, Mmusc20170205. "full [.gz]" were downloaded as tab-limited files.

4.1.3 Tensor format

The PPI files were further loaded into R using the `read_csv` command as $n^k_{ii'} \in \mathbb{R}^{N_k \times N_k}$ ($k = 1$: human, $k = 2$: mouse, , and $k = 3$: rat (only for BioGRID)), which takes 1 when the ith and i'th proteins interact with each other; otherwise, 0. As these matrices were sparse, they were stored in a sparse matrix format using the Matrix [32] package.

4.2 Principal component analysis (PCA)-based unsupervised feature extraction (FE)

Singular value decomposition was applied to the `irlba` function in the irlba package [33] for human PPI $n^1_{ii'}$. As a result, we obtained

$$n^1_{ii'} = \sum_{\ell=1}^{L} \lambda_\ell u_{\ell i} u_{\ell i'} \tag{3}$$

where λ_ℓ is a singular value, and $u_{\ell i} \in \mathbb{R}^{N \times N}$ is the singular value vector and orthogonal vector.

P-values are attributed to ith protein as

$$\langle u_{\ell i} \rangle = \frac{1}{N} \sum_{i=1}^{N} u_{\ell i} \tag{4}$$

$$\sigma_\ell = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(u_{\ell i} - \langle u_{\ell i} \rangle \right)^2} \tag{5}$$

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

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution, the argument is larger than x and $\langle u_{\ell i} \rangle$, and σ_ℓ are the mean and standard deviation, respectively. Thus, we assumed the following null hypothesis: $u_{\ll i}$ follows a Gaussian distribution. P-values were corrected using the BH criterion, and proteins associated with adjusted P-values less than 0.01 were selected.

4.3 Tensor decomposition (TD)-based unsupervised FE

To apply TD-based unsupervised FE to PPI, an integrated tensor that stores multiple PPIs must be constructed. Suppose we have two PPI matrices, $n^1_{ii'} \in \mathbb{R}^{N_1 \times N_1}$ and $n^2_{ii'} \in \mathbb{R}^{N_2 \times N_2}$. It is also assumed that there are N_{common} common proteins and N_{ortho} orthogonal proteins between the two datasets. Then, we merge these datasets into one tensor, $n_{ii'2} \in \mathbb{R}^{N \times N \times 2}$ where $N = N_1 + N_2 - N_{\text{common}} - N_{\text{otrho}}$ as shown in (Fig. 2). $n^k_{ii'}, (1 \leq i, i' \leq N_{\text{common}} + N_{\text{ortho}})$ are placed in $n_{ii'k}$ as is (yellow and

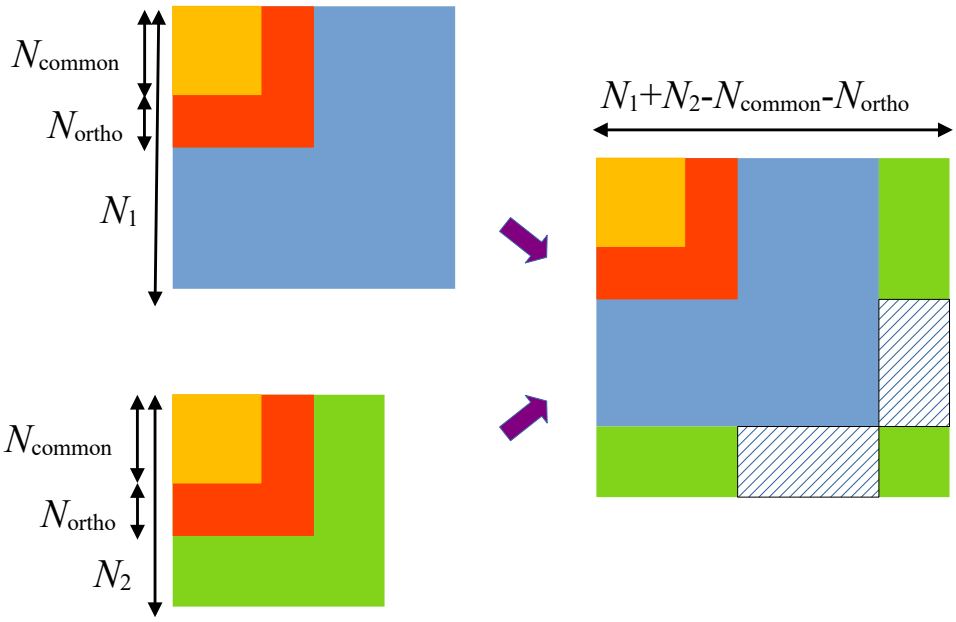

Fig. 2 How to merge two PPI matrices into one tensor.

red regions in Fig. 2). $n^1_{ii'}, (N_{\text{common}} + N_{\text{ortho}} < i, i' \leq N_1)$ are placed in $n_{ii'1}$ (the blue region in Fig. 2). $n^2_{ii'}, (N_{\text{common}} + N_{\text{ortho}} < i, i' \leq N_2)$ are placed in $n_{ii'2}$ in a fragmented manner (green region in Fig. 2). As a result, the blue region of $n_{ii'2}$ and green region of $n_{ii'1}$ are blank. The shaded region is blank in $n_{ii'k}$.

HOSVD was applied to $n_{ii'k}$, resulting in

$$n_{ii'k} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{N} \sum_{\ell_3=1}^{2} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 i'} u_{\ell_3 k} \tag{7}$$

where G is a core tensor representing the weight (contribution) of $u_{\ell_1 i} u_{\ell_2 i'} u_{\ell_3 k}$ to $n_{ii'2}$ and $u_{\ell_1 i} = u_{\ell_2 i'} \in \mathbb{R}^{N \times N}$ and $u_{\ell_3 k} \in \mathbb{R}^{2 \times 2}$ are singular value and orthogonal matrices, respectively. The attribution of P_is to is and the selection of proteins were performed by replacing $u_{\ell i}$ where $u_{\ell_1 i}$ in eq. (6).

HOSVD was performed using the `irlba` function by applying SVD to unfolded matrices because the usual R function that can perform HOSVD does not accept a sparse matrix as input.

4.4 Identification of orthogonal proteins

Orthologs between human and mouse were retrieved by querying "((organism_id:10090) OR ((organism_id:9606) AND (reviewed:true))) AND (database:orthodb)" in uniport search (https://www.uniprot.org/). Human Uniprot accessions were converted to gene name as "XXX_HUMAN" using the above retrieved information. Then, after seeking "XXX_MOUSE" in the above retrieved information (using the first hit if multiple hits are found), "XXX_MOUSE" was converted into the corresponding mouse Uniprot accession numbers in the information retrieved above. This results in the corresponding table of Uniprot ortholog (the above retrieved information is provided as Supplementary Material).

4.5 Gene ID conversion

The obtained Uniprot accession numbers attributed to proteins were converted to gene names using Uniprot ID mapping (https://www.uniprot.org/id-mapping).

4.6 Relating genes to drugs and diseases through enrichment analysis

Enrichr enables the validation of overlaps between two sets of genes using statistical tests that measure the probability that the observed overlap occurs by chance (the so-called P-value). If P-values are sufficiently small, the two sets of genes are significantly related. In Enrichr, one set of genes was provided by the researchers, and its overlap with the prepared sets of genes was evaluated. If the uploaded set of genes significantly overlaps with one of the prepared sets of genes, the uploaded set of genes can be said to be associated with the properties of the overlapping set of genes. In this analysis, we considered two sets of genes, diseases and drugs. For the disease gene sets, the uploaded set of genes was evaluated if it overlapped with a set of genes known to be related to diseases. The "Jensen Diseases" and "OMIM disease" are independent categories that collect the genes related to diseases. Thus, if the uploaded gene sets have significant overlap with sets of genes in either "Jensen Diseases" or "OMIM disease," we can regard the uploaded gene sets as being related to diseases. For drug

gene sets, the uploaded set of genes was evaluated if it overlapped with a set of genes whose expression was altered by drug treatment. If the uploaded set of genes has a significant overlap with the gene sets whose expression is known to be altered by some drugs, we can consider that the expression of the uploaded gene set is also altered by drug treatment. Thus, if the uploaded gene set significantly overlaps with the drug and disease gene sets simultaneously, we can expect that drug treatment can significantly alter the expression of genes whose expression is known to be altered by drug treatment. Consequently, drugs related to diseases through genes are potential drug compounds, although they are not always effective against specific diseases.

4.7 Query of drug-disease relation by LLM

The actual queries were performed using Microsoft Copilot. The basic structure of the prompt is provided in Supplementary Information.

The actual query prompts and LLMs replies are accessible through the following two URLs: https://copilot.microsoft.com/sl/dCHesf9R01c and https://copilot.microsoft.com/sl/jCQxrKJzZoO.

5 Conclusions

In this study, we demonstrated the usefulness of drug repositioning TD-based unsupervised FE applied to PPI. Although it is unlikely that only the information retrieved from PPI will be useful for disease-specific drug repositioning, our findings indicate that it works in a practical sense. Thus, TD-based unsupervised FE applied to PPI is likely to be useful for drug repositioning. Further studies will be needed to understand the extent to which this strategy is effective.

Author Contributions

Y.-H.T. planned the study and performed the analyses. Y.-H.T. and T.T. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All the authors have read and agreed to the published version of this manuscript.

Funding

This work was supported by KAKENHI [grant number 24K15168] and Chuo Universiy Tokutei Kadai kenkyuu.

Data and Availability

All the data used in this study can be downloaded from BioGRID [30] and DIP [31].

Declarations

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest.

References

[1] S, Z., V, E.A., B, B.U., A, d.S.: Prediction of disease–gene–drug relationships following a differential network analysis. Cell Death & Disease **7**(1), 2040–2040 (2016) https://doi.org/10.1038/cddis.2015.393

[2] Wong, M., Previde, P., Cole, J., Thomas, B., Laxmeshwar, N., Mallory, E., Lever, J., Petkovic, D., Altman, R.B., Kulkarni, A.: Search and visualization of gene-drug-disease interactions for pharmacogenomics and precision medicine research using genedive. Journal of Biomedical Informatics **117**, 103732 (2021) https://doi.org/10.1016/j.jbi.2021.103732

[3] Yu, H., Choo, S., Park, J., Jung, J., Kang, Y., Lee, D.: Prediction of drugs having opposite effects on disease genes in a directed network. BMC Systems Biology **10**(Suppl 1), 2 (2016) https://doi.org/10.1186/s12918-015-0243-2

[4] Sun, P.G.: The human drug–disease–gene network. Information Sciences **306**, 70–80 (2015) https://doi.org/10.1016/j.ins.2015.01.036

[5] Qahwaji, R., Ashankyty, I., Sannan, N.S., Hazzazi, M.S., Basabrain, A.A., Mobashir, M.: Pharmacogenomics: A genetic approach to drug development and therapy. Pharmaceuticals **17**(7) (2024) https://doi.org/10.3390/ph17070940

[6] Iida, M., Iwata, M., Yamanishi, Y.: Network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets. Bioinformatics **36**(Supplement_1), 516–524 (2020) https://doi.org/10.1093/bioinformatics/btaa439 https://academic.oup.com/bioinformatics/article-pdf/36/Supplement_1/i516/56702295/bioinformatics_36_supplement1_i516.pdf

[7] Quan, Y., Luo, Z.-H., Yang, Q.-Y., Li, J., Zhu, Q., Liu, Y.-M., Lv, B.-M., Cui, Z.-J., Qin, X., Xu, Y.-H., Zhu, L.-D., Zhang, H.-Y.: Systems chemical genetics-based drug discovery: Prioritizing agents targeting multiple/reliable disease-associated genes as drug candidates. Frontiers in Genetics **10** (2019) https://doi.org/10.3389/fgene.2019.00474

[8] Wang, L., Wang, Y., Hu, Q., Li, S.: Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. CPT: Pharmacometrics & Systems Pharmacology **3**(11), 146 (2014) https://doi.org/10.1038/psp.2014.44 https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1038/psp.2014.44

[9] Kim, Y., Cho, Y.-R.: Predicting drug–gene–disease associations by tensor decomposition for network-based computational drug repositioning. Biomedicines **11**(7) (2023) https://doi.org/10.3390/biomedicines11071998

[10] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 1st edn. Springer, ??? (2020). https://doi.org/10.1007/978-3-030-22456-1 . http://dx.doi.org/10.1007/978-3-030-22456-1

[11] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 2nd edn. Springer, ??? (2024)

[12] Taguchi, Y.-H., Turki, T.: Integrated analysis of gene expression and protein–protein interaction with tensor decomposition. Mathematics **11**(17) (2023) https://doi.org/10.3390/math11173655

[13] Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., Jeon, M., Ma'ayan, A.: Gene set knowledge discovery with enrichr. Current Protocols **1**(3), 90 (2021) https://doi.org/10.1002/cpz1.90 https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.90

[14] Carr, S., Kasi, A.: Familial Adenomatous Polyposis. StatPearls Publishing, Treasure Island (FL) (2024). Updated 2023 Feb 25. https://www.ncbi.nlm.nih.gov/books/NBK538233/

[15] Aedma, S.K., Kasi, A.: Li-Fraumeni Syndrome. StatPearls Publishing, Treasure Island (FL) (2024). PMID: 30335319. https://pubmed.ncbi.nlm.nih.gov/30335319/

[16] Bhandari, J., Thada, P.K., Puckett, Y.: Fanconi Anemia. StatPearls Publishing, Treasure Island (FL) (2024). Last Update: August 10, 2022. https://www.ncbi.nlm.nih.gov/books/NBK559133/

[17] Zhao, Q., Zhang, Y., Shao, S., Sun, Y., Lin, Z.: Identification of hub genes and biological pathways in hepatocellular carcinoma by integrated bioinformatics analysis. PeerJ **9**, 10594 (2021) https://doi.org/10.7717/peerj.10594

[18] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[19] Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007) https://doi.org/10.1103/PhysRevE.76.036106

[20] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004) https://doi.org/10.1103/PhysRevE.69.026113

[21] Kang, M.-H., Seok Jeong, G., Smoot, D.T., Ashktorab, H., Mo Hwang, C., Sik Kim, B., Sung Kim, H., Park, Y.-Y.: Verteporfin inhibits gastric cancer cell growth by suppressing adhesion molecule fat1. Oncotarget **8**(58), 98887–98897 (2017) https://doi.org/10.18632/oncotarget.21946

[22] Rigopoulos, D., Larios, G., Gregoriou, S., Alevizos, A.: Acute and chronic paronychia. Am. Fam. Physician **77**(3), 339–346 (2008)

[23] Arjmand Abbassi, Y., Mohammadi, M.T., Sarami Foroshani, M., Raouf Sarshoori, J.: Captopril and valsartan may improve cognitive function through potentiation of the brain antioxidant defense system and attenuation of oxidative/nitrosative damage in stz-induced dementia in rat. Adv Pharm Bull **6**(4), 531–539 (2016) https://doi.org/10.15171/apb.2016.067 https://apb.tbzmed.ac.ir/PDF/APB_4168_20160423171936

[24] Michela, C., Antonietta, M., Domenico, T.: Effects of the antidiabetic drugs on the age-related atrophy and sarcopenia associated with diabetes type ii. Current Diabetes Reviews **10**(4), 231–237 (2014) https://doi.org/10.2174/1573399810666140918121022

[25] Ohno, T., Nakane, T., Akase, T., Kurasawa, H., Aizawa, Y.: Development of an isogenic human cell trio that models polyglutamine disease. Genes & Genetic Systems **98**(4), 179–189 (2023) https://doi.org/10.1266/ggs.22-00030

[26] Giuliani, F., Molica, S., Maiello, E., Battaglia, C., Gebbia, V., Bisceglie, M.D., Vinciarelli, G., Gebbia, N., Colucci, G.: Irinotecan (cpt-11) and mitomycin-c (mmc) as second-line therapy in advanced gastric cancer: A phase ii study of the gruppo oncologico dell' italia meridionale (prot. 2106). American Journal of Clinical Oncology **28**(6), 581–585 (2005) https://doi.org/10.1097/01.coc.0000190398.52142.7f

[27] Zhu, X.-C., Dai, W.-Z., Ma, T.: Overview the effect of statin therapy on dementia risk, cognitive changes and its pathologic change: a systematic review and meta-analysis. Annals of Translational Medicine **6**(22) (2018)

[28] Billingsley, E., Vidimos, A.: Paronychia Treatment & Management. Medscape, 1106062 (2022). https://emedicine.medscape.com/article/1106062-treatment?form=fpf

[29] Yoon, J.-H., Yoo, C.-I., Ahn, Y.-S.: N,n-dimethylformamide: evidence of carcinogenicity from national representative cohort study in south korea. Scandinavian Journal of Work, Environment & Health (4), 396–401 (2019) https://doi.org/10.5271/sjweh.3802

[30] Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., Tyers, M.: The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science **30**(1), 187–200 (2021) https://doi.org/10.1002/pro.3978 https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3978

[31] Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: DIP: the Database of Interacting Proteins. Nucleic Acids Research **28**(1), 289–291 (2000) https://doi.org/10.1093/nar/28.1.289 https://academic.oup.com/nar/article-pdf/28/1/289/9895160/280289.pdf

[32] Bates, D., Maechler, M., Jagan, M.: Matrix: Sparse and Dense Matrix Classes and Methods. (2024). R package version 1.7-0. https://CRAN.R-project.org/package=Matrix

[33] Baglama, J., Reichel, L., Lewis, B.W.: Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. (2022). R package version 2.3.5.1. https://CRAN.R-project.org/package=irlba