Genome-wide discovery of hidden genes mediating known drug-disease association using KDDANet

Hua Yu [1, 4,#*], Lu Lu [2, #], Ming Chen [3], Chen Li [2, *], Jin Zhang [1, *]

[1] Center for Stem Cell and Regenerative Medicine, Department of Basic Medical Sciences, and The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. Institute of Hematology, Zhejiang University, Hangzhou, Zhejiang, China.

[2] Department of Human Genetics, and Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China

[3] College of Life Sciences, Zhejiang University, Hangzhou, China.

[4] Lead contact.

[#] These authors contributed equally.

[*] Correspondence: huayu@zju.edu.cn, chenli2012@zju.edu.cn, zhgene@zju.edu.cn.

Abstract

Many of genes mediating Known Drug-Disease Association (KDDA) are escaped from experimental detection. Identifying of these genes (hidden genes) is of great significance for understanding disease pathogenesis and guiding drug repurposing. Here, we presented a novel computational tool, called KDDANet, for systematic and accurate uncovering the hidden genes mediating KDDA from the perspective of genome-wide functional gene interaction network. KDDANet demonstrated the competitive performances in both sensitivity and specificity of identifying genes in mediating KDDA in comparison to the existing state-of-the-art methods. Case studies on Alzheimer's disease (AD) and obesity uncovered the mechanistic relevance of KDDANet predictions. Furthermore, when applied with multiple types of cancer-omics datasets, KDDANet not only recapitulated known genes mediating KDDAs related to cancer, but also revealed novel candidates that offer new biological insights. Importantly, KDDANet can be used to discover the shared genes mediating multiple KDDAs. KDDANet can be accessed at http://www.kddanet.cn and the code can be freely downloaded at https://github.com/huayu1111/KDDANet.

Introduction

The conventional development of novel promising drugs for treating specific diseases is a time-consuming and efforts-costing process, including discovery of new chemical entities, target detection and verification, preclinical and clinical trials and so on [1]. Furthermore, the success rate for a new drug to be taken to market is very low, usually only 10% per year of drugs approved by FDA and thus prevents their use in actual therapy [1]. This results in that pharmaceutical research faces a decreasing productivity in drug development and a sustaining gap between therapeutic needs and available treatments [1]. Compared with traditional drug development, drug repositioning, i.e., finding the novel indications of existing drugs, offers the possibility for safer and faster drug development because of several steps of traditional drug development pipeline can be avoided during repurposing efforts [2]. Many successful cases of repositioned drugs have been shown: from Minoxidil, designed for treatment of hypertension and now indicated for hair loss [3], to Sildenafil, developed for patients with heart problems and repurposed for erectile dysfunction [4]. Yet, these examples are mainly based on clinical observations of secondary effects [5]. Thanks to the advance in next generation omics sequencing and qualification technologies, a large volume of biomedical data, for example the pharmacogenomics datasets produced by Connective Map (CMap) project, The Cancer Genome Atlas (TCGA) project, Cancer Cell Line Encyclopedia (CCLE) project, Genomics of Drug Sensitivity in Cancer 1000 human cancer cell lines (GDSC1000) project and Library of Integrated Network-based Cellular Signatures (LINCS) project, has been rapidly accumulated for enabling drug repurposing [6-14]. Based on these datasets, various computational methods have been designed for facilitating the process of drug repurposing (see Supplementary Note 1 for a mini review). To infer pharmacokinetic and pharmacodynamic drug-drug interactions and their associated recommendations, for example, Gottlieb et al. designed a similarity measure-based logistic regression classifier [15]. For predicting drug side-effects, Tatonetti et al. presented an adaptive

data-driven approach [16]. To identify novel drug combinations, Zhao et al. integrated the molecular and pharmacological data and developed a novel computational approach [17]. In addition, network-based method has also been employed to achieve the similar goal [18]. Interestingly, Kuenzi et al. developed a deep-learning model of visible neural network, called DrugCell, to predict drug response and synergy in human cancer cells [19]. For discovering novel drug indication, Gottlieb et al. developed PREDICT, which scored a possible drug-disease link by combining multiple drug-drug and disease-disease similarity measures [20]. With the similar motivation and aim, an unsupervised and unbiased network-based proximity measure has also been designed [21]. Moreover, Cheng et al. showed that the further integration of network proximity-based approach with large-scale patient-level longitudinal data can offer an effective platform for validating drug indications [22]. For predicting drug-target interaction, Paolini et al. presented a global mapping of pharmacological space and probabilistic models by integrating multiple medicinal chemistry data [23]. Different from the approach employed by Paolini et al., Campillos et al. used the phenotypic side-effect similarity to determine whether two drugs share a target [24]. Besides, we and others also employed machine learning and network integration approaches for achieving the same goal [25,26]. To discover disease related genes, Gottlieb et al. designed "PRINCIPLE", which employed classical network propagation algorithm [27]. Wu et al. developed CIPHER that integrated human protein-protein interactions, disease phenotype similarities, and known gene-phenotype associations to capture the complex relationships between phenotypes and genotypes [28]. Additionally, Ghiassian et al. identified disease gene modules based on systematic analysis of connectivity patterns of disease proteins in the human interactome [29]. Furthermore, Zhou et al. and Menche et al. constructed disease-symptom and disease-disease relationship networks based on the biomedical literature databases and incomplete interactome, respectively [30,31]. Excitingly, Hofree et al. developed a new computational approach which employed network-based stratification (NBS) to

uncover tumor subtypes by integrating somatic tumor genomes with gene networks [32]. Collectively, these methods have effectively exploited and integrated multi-level biomedical and omics data sources for understanding the pathology of diseases and mechanisms of drug actions and thus accelerated drug repurposing.

Theoretically, drug repurposing has been proposed based on two molecular aspects. I) On one hand, complex diseases often involve multiple genetic and environmental determinants, including multi-factor driven alterations and dysregulation of a series of genes [33,34], which will propagate and perturb certain biological processes by the interactions among molecules, leading to the onset of diseases. II) On the other hand, one drug can exert impacts on many targets and perturb multiple biological processes [34,35]. As a result, the genes of shared biological pathways in the cellular network manifested in certain disease state and induced by a known drug administration suggest potential drug repurposing [33-38]. However, many of these genes mediating Known Drug-Disease Associations (KDDAs) across various types of diseases have not yet been identified (see the Results section and Supplementary Figure S1a-S1b for supporting data of this claim). Therefore, developing the appropriate theoretical computational tools from the perspective of molecular interaction network to unveil the KDDA genes missed from experiments (hidden genes) is of great significance for understanding disease pathogenesis and guiding drug repurposing. The network-based computational methods have been designed for facilitating drug repurposing which linked drugs to targets or connected diseases to genes or associated drugs with diseases [18,21,22,26-29,39]. Nevertheless, the publicly available computational tools specially tailored for simultaneously bridging drugs, genes, and diseases have not been fully developed. To our knowledge, some computational tools have been designed to identify drug-gene-disease co-module. Kutalik et al. and Chen et al. developed Ping-Pong Algorithm (PPA) and Sparse Network-regularized Partial Least Square (SNPLS) to identify co-modules related to

specific cancer cell lines of NCI-60 and Cancer Genome Projects, respectively [40,41]. However, these two methods need to integrate gene-expression and drug-response data of cancer cell lines for constructing models and do not carry out prediction for other types of diseases. The other two methods of comCIPHER and DGPsubNet have been designed to identify the coherent subnetworks linking drugs and diseases (not limited in cancer) with the related genes [42,43]. The common shortcomings of these methods are the identified co-modules including multiple drugs and diseases and thus are unable to uncover genes specifically mediating individual KDDA. To this end, we designed a novel computational pipeline, called KDDANet, which uses known functional gene interaction network to identify hidden genes of cellular pathways mediating KDDA in a genome-wide scale. Our KDDANet pipeline depends on three existing network algorithms: minimum cost network flow optimization, depth first searching and graph clustering algorithm. The minimum cost network flow optimization has been effectively employed to identify cellular response subnetwork connecting genetic hits and differentially expressed genes, including components of the response that are otherwise hidden or missed from experiments [44]. KDDANet can be applied to two contexts: 1) uncovering hidden genes mediating the association between a query drug and its related disease (SDrTDi); 2) unveiling hidden genes mediating the association between a query disease and its related drug (SDiTDr). The computational procedure of KDDANet in SDrTDi context was showed in Fig. 1 (see Method for details). KDDANet first built a unified flow model by integrating query drug, genes, and all related diseases into a heterogeneous network (Fig. 1a). Then, the minimum cost flow optimization was designed and implemented to identify gene subnetwork mediating the association between the query drug to all its related diseases (Fig. 1b). Finally, depth first searching, and Markov clustering algorithm were adopted to further uncover gene modules mediating the association between the query drug and each its related disease (Fig. 1c-1e). The outputs of KDDANet were validated against existing literature and multi-omics' datasets (Fig. 1f). To apply

KDDANet in SDiTDr context, what the user need was just to simply rebuilt the unified flow network model (see Methods for details).

The key novelty of the KDDANet method lay in that the original designation of a unified flow network model and effective implementing multiple network-based algorithms on this unified flow network model. We demonstrated that KDDANet showed competitive performance in discovering known and novel genes mediating KDDA through a comparison with existing methods. KEGG pathway enrichment analysis on KDDANet resulting subnetworks and case studies on Alzheimer's disease (AD) and obesity further showed the mechanistic relevance of KDDANet predictions. Validated with multiple types of cancer-omics' datasets, KDDANet did not only revealed known genes mediating KDDAs associating drug with cancer, but also uncovered new candidates that offer novel biological insights. Particularly, our results demonstrated that KDDANet can reveal the shared genes mediating multiple KDDAs. These outcomes showed that importance of incorporating hidden genes in drug discovery pipelines. For facilitating biomedical researchers to explore the molecular mechanism of KDDA and guiding drug repurposing, an online web server, http://www.kddanet.cn, was provided for user to access the subnetwork of genes mediating KDDA, and the source codes of KDDANet were freely available at https://github.com/huayu1111/KDDANet. In summary, we developed an effective and universal computational tool and an online web source for accurate and systematic discovering hidden genes mediating KDDA and thus providing novel insights into mechanism basis of drug repurposing and disease treatments. We believed that KDDANet can provide additional contributions to the development of new therapies.

Results

Evaluation of the performance and general applicability of KDDANet method.

We first checked whether the potential genes mediating KDDA across various types of diseases have been experimentally identified. For a given KDDA, we proposed a hypothesis that the Known Drug Target Genes (KDTGs) and Known Disease Related Genes (KDRGs) should be highly overlapped if the genes mediating this KDDA have been fully identified. We analyzed the overlap between Known Drug Target Genes (KDTGs) and Known Disease Related Genes (KDRGs) of 53124 KDDAs obtained from Comparative Toxicogenomics Database (CTD) [12]. For a KDDA, we defined the overlap ratio as the number of shared genes between KDTGs and KDRGs divided by the number of total KDTG and KDRG genes. We observed that the most gene sets demonstrated extremely low overlap ratio, with the increasing of gene number, the overlap ratio was sharply decreased (Supplementary Figure 1a). We then checked the overlap ratio in different types of diseases and found the overlaps between KDTGs and KDRGs were small for all 19 disease types in our dataset (Supplementary Figure 1b). The discrepancy between KDTGs and KDRGs indicated that each gene set alone provided only a limited and biased view of KDDA, many of true genes in the cellular pathways mediating KDDA were not identified from the experiments but otherwise hidden. To address this, we designed a novel computational tool, KDDANet, which effectively integrated minimum cost flow optimization, combined with depth first searching and graph clustering to systematically discover the hidden genes of cellular pathways mediating KDDA (see Methods for details).

To examine whether KDDANet can capture true genes mediating KDDA, we introduced two new concepts: “known true KDDA genes” (KTKGs) and “novel true KDDA genes” (NTKGs). For a given KDDA, KTKGs was defined as the shared genes between KDTGs and KDRGs inputted for constructing KDDANet flow network model (see Dataset for details). To obtain NTKGs, we collected a set of drug’s non-target genes (A) from SMPDB 2.0 database [45] that were included in the drug’s ADME pathways and were responsible for mediating KDDA. Meanwhile, we collected a

recently updated set of disease-related genes (B) from DisGeNet v6.0 database [46], which were not included in KDDANet flow network model. Based on these, we defined, for each KDDA, the NTKGs as the shared genes between A and B. A parameter γ effected the size and quality of KDDANet output subnetwork, higher γ values will identify more gene links mediating KDDA but with lower confidence. Using gene set enrichment analysis, we observed that KDDANet can consistently and effectively capture the KTKGs and NTKGs mediating KDDA under different γ settings (see Supplementary Note 2, Supplementary Note 3, and Supplementary Figure 1c-1l for details).

Based on this, we compiled a standard set of positive and negative KDDA genes for each KDDA to unbiasedly evaluate the capability of KDDANet method on uncovering the true genes mediating KDDA using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves (see Methods for details). We observed that the performance of KDDANet was obviously better than random permutation across a widely settings of γ (Fig. 2a and Fig. 2b). We next selected γ = 6 for SDrTDi and γ = 8 for SDiTDr for subsequent evaluation (see Supplementary Note 4 for reasons). With this setting, we further compared KDDANet with other existing methods for drug-gene-disease co-module discovery, including SNPLS [40], PPA [41], comCHIPER [42] and DGPsubNet [43] (see Methods for the details of performance comparison). Among all the methods tested, KDDANet demonstrated the competitive performance with the averages of AUROC 0.733 and 0.715 and AUPRC 0.793 and 0.825 in SDrTDi and SDiTDr contexts, respectively (Fig. 2c and Table 1). These results suggested that KDDANet was an effective tool for achieving the goal of uncovering true genes mediating KDDA genome wide. We next tested whether KDDANet had the general application value across a variety of disease types and found that KDDANet can make effective capturing of true genes mediating KDDA for all 19 types of diseases (Supplementary Figure 2a-2d). We further tested KDDANet using different types of

networks, including HINT+HI2012, iRefIndex, MultiNet and STRINGv10 [94]. We found that the performances of KDDANet were consistent well across all these networks (Fig. 2d, Supplementary Figure 2e and Supplementary Figure 2f). Collectively, these results fully demonstrated that KDDANet was an effective and general computational tool for uncovering hidden genes mediating KDDA across broad types of diseases.

Mechanistic relevance of KDDANet predictions.

We examined whether the enriched pathways of KDDANet resulting subnetworks have the mechanistic relevance with KDDA by carrying out a global enrichment of all predicted KDDA subnetwork genes against 53 classical KEGG pathways. The obtained enrichment results can be validated by existing knowledge (see Supplementary Note 5 and Supplementary Figure 3a-3b for details). We aimed to provide two cases to intuitively describe the mechanistic relevance of KDDANet resulting subnetwork. Phylloquinone (DB01022)-Alzheimer's disease (AD) (104300) association has been reported in previous study [47]. A subnetwork including 46 genes and 44 links were uncovered mediating this association (Fig. 3a). Interestingly, for this subnetwork, two separated gene modules (M1 and M2) were detected (Fig. 3a). The AUROC and AUPRC values of for this KDDANet resulting subnetwork were 0.864 and 0.795, respectively (Supplementary Figure 3c). Two known targets of phylloquinone and 18 AD-related genes were identified in this subnetwork. Against with genome background, this subnetwork captured 4 NTKGs with ~86-fold enrichment and adjusted *p*-value of 9.716e-09 (Hypergeometric test and Bonferroni correction). The top 10 enriched KEGG terms of this subnetwork, such as Phospholipase D signaling pathway and Neurotrophin signaling pathway [48,49], were closely related with AD (Fig. 3b). The module M1 mainly functioned in Insulin signaling pathway, ErbB signaling pathway, FoxO signaling pathway and growth hormone synthesis, secretion, and action, which played important roles in neural system development and the onset and development of AD [50-52] (Supplementary Figure 3d). The enriched KEGG

pathways of M2 genes, including Complement and coagulation cascades, Glycolysis and AGE-RAGE signaling pathway in diabetic complications, were dysfunction in AD [53-55] (Supplementary Figure 3d). We further analyzed the published RNA-seq data to detect the expressional change of these two modules in normal individuals and AD patients [56]. We observed that the averaged expression level of M1 genes was significantly upregulated in AD (Supplementary Figure 3e). Interestingly, a predicted novel gene, STAT3, was obviously activated in AD patients (Fig. 3c). Two newest studies published in years 2019 and 2020 reported that STAT3 was a potential therapeutic target for cognitive impairment in AD [57,58]. Another predicted novel gene, GNAI1, was also obviously activated in AD patients. This gene was contained in the causal pathways associated with an imaging endophenotype characteristic of longitudinal structural change in the brains of patients with AD [59]. For module M2 genes, the obviously expressional changes between normal individuals and AD patients were not observed (Supplementary Figure 3e). However, we found that a predicted novel KDDA gene, GC, was activated in AD patients (Fig. 3c). This gene encoded Vitamin D Binding Protein, which was recently evidenced as a potential therapeutic agent for the treatment of AD [60].

The association between heparin (DB01109) and obesity (601665) was inferred by multiple genes as described in CTD database. For this association, KDDANet predicted a subnetwork containing 168 edges connecting 169 genes (Fig. 3d). We found that two known drug target genes and 67 disease-related genes were captured in this subnetwork. Particularly, 8 genes were the NTKGs with ~52-fold enrichment and adjusted *p*-value of 1.485e-10 (Hypergeometric test and Bonferroni correction). Three genes used to infer this KDDA, including ARK1, PARP1 and TNF, were also effectively captured in the resulting subnetwork. The top 10 enriched functions of this subnetwork were showed in Fig. 3e. As expected, Insulin resistance, Type II diabetes mellitus, Insulin signaling pathway and Adipocytokine signaling pathway

were the frequently reported events and molecular processes associated with obesity [61,62]. In addition, Proteoglycans, Lipolysis and AMPK signaling pathway were also highly related with Insulin resistance [63-65]. In consistent with this, the AUROC and AUPRC values of KDDANet for this subnetwork were 0.854 and 0.849, respectively (Supplementary Figure 3e). We next delineated the subnetwork into gene modules. The enriched functions of top 3 gene modules were demonstrated in Supplementary Figure 3f. The genes of module M1 mainly participated in Glycolysis and Carbon metabolism. Further analysis of public RNA-seq data of normal individuals and obesity patients [66,67] demonstrated the significantly repressed expression of M1 genes in patients with obesity, such as GADPH (Supplementary Figure 3g and 3h). This was consistent with the fact that enhancing the level of glycolysis reduced obesity [68]. The mainly enriched pathways of M2 genes were related to Insulin resistance, a frequently happened event in obesity patients. Interestingly, Cell adhesion molecules was a significantly enriched KEGG term of M2 genes (Supplementary Figure 3f), which was elevated in patients with obesity [69]. The M3 genes functioned in Complement and coagulation cascades and Platelet activation (Supplementary Figure 3f). As reported, these two terms were closely associated with obesity [70,71]. In support with these, the expression of genes in M2 and M3 were activated in patients with obesity (Supplementary Figure 3g and Supplementary Figure 3h). Together, these results indicated that KDDANet can serve as a useful tool to unveil the molecular basis of KDDA.

KDDANet provided novel molecular insights on KDDAs related to cancer.

Cancer was a frequently happened complex genetic disease caused by DNA abnormalities [72]. For this reason, substantial genetic, genomic and pharmacogenomics efforts, including TCGA, CCLE and GDSC1000 [7-9], have been undertaken to improve existing therapies or to guide early-phase clinical trials of compounds under development. With these efforts, an increasing amount of

available high-throughput data sets at both levels of genomic data and pharmacogenomics data were produced at recent years. In addition, COSMIC, the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer, collected a catalogue of genes with mutations that were causally implicated in cancer (https://cancer.sanger.ac.uk/cosmic). With these datasets, we observed that the genes of KDDANet resulting subnetworks mediating the associations between drugs and cancer were significantly enriched in COSMIC Cancer Gene Census, and these genes harbored more oncogenic alterations in tumor samples than randomly selected genes (Supplementary Note 6, Supplementary Figure 4a and Supplementary Figure 4b). Moreover, we found that oncogenic alterations of genes in KDDANet resulting subnetworks mediating the associations between drugs and cancer were more correlated with the responses of cancer cell lines under anti-cancer drug treatment than randomly selected genes (see Supplementary Note 5, Supplementary Figure 4c-4f). We provided two detailed examples to describe the potential values of KDDANet in revealing novel genes mediating the associations between drugs and cancer.

Sotalol (DB00489) was normally used to treat life threatening ventricular arrhytmias. It has been reported that sotalol was associated with decreased prostate cancer (176807) risk [73]. For sotalol-prostate cancer association, KDDANet predicted a subnetwork consisting of 31 genes and 28 links (Fig. 4a). By applying MCL with default parameters, this subnetwork was further decomposed into three gene modules, M1, M2 and M3. All three known target genes of sotalol were included in this subnetwork. Meanwhile, this subnetwork also captured 12 prostate cancer-related genes with a novel NTKG of PRKACA. The top 10 enriched KEGG terms of this subnetwork were showed in Fig. 4b. Among these, PI3K-Akt signaling pathway, MAPK signaling pathway and FoxO signaling pathway were related to cancer formation and development. The relationships between EGFR tyrosine kinase

inhibitor resistance, Relaxin signaling pathway, AGE-RAGE signaling pathway and prostate cancer have been widely investigated and reported in the previous works [74-76]. In addition, Autophagy and Focal adhesion were two widely observed processes in cancer [77,78]. As expected, the enriched KEGG signaling pathways of M1 were closely related to tumorigenesis (Supplementary Figure 4g). The expression of M1 genes were activated in tumor samples, such as an oncogene YWHAE (Supplementary Figure 4h and Supplementary Figure 4i). Interestingly, M1 captured that IGF1R, a gene encoding insulin-like growth factor receptor, harbored SNVs and CNVs in TCGA prostate cancer samples, and has been reported to be oncogenic genes of prostate cancer [79]. The genes of M2 were not enrich to any KEGG term and but have lower expression in TCGA tumor samples (Supplementary Figure 4h). Surprisingly, SPARC, a reported prostate cancer-related gene [80] with significantly lower expression in TCGA tumor samples was captured in this module (Supplementary Figure 4i). Moreover, we found that both SPARC and COL1A1 harbor SNVs and CNVs in TCGA prostate cancer samples and their expressions were obviously correlated to the survival of patients (Fig. 4c). The M3 genes mainly participated in ErbB signaling pathways, a biological process involved in prostate cancer progression [81]. The expression of M3 genes were not obviously changed in TCGA prostate cancer samples (Supplementary Figure 4h). However, we found that GRB2 was over-expressed in TCGA prostate cancer samples and obviously correlated with the survival of patients at the later stage of disease (Supplementary Figure 4i and Fig. 4c).

Another example was the association between nebularine (DB04440)-lung cancer (211980) that was inferred by ADA targeted by nebularine [82]. For this association, KDDANet predicted 237 genes connected by 238 links that constituted a subnetwork without apparent modular structure (Fig. 4d). The only known target gene ADA of nebularine was connected to 119 lung cancer-related genes and 117 predicted novel

KDDA genes in this subnetwork. As expected, KEGG enrichment demonstrated that the genes in this subnetwork were involved in various cancers (Fig. 4e). Interestingly, we found that two highly connected novel genes BYSL and BOP1 were significantly over-expressed in TCGA lung cancer tumor samples and their expression levels were obviously correlated with the survival of lung cancer patients (Fig. 4f). These results indicated that KDDANet not only captured known genes mediating KDDAs linking drug with cancer, but also uncovered novel candidates that offered novel biological insights.

KDDANet uncovered the shared genes mediating multiple KDDAs.

Comprehensive analysis above fully demonstrated that KDDANet can uncover true genes mediating individual KDDA. We further asked whether KDDANet can reveal the shared genes mediating multiple KDDAs. We answered this from two aspects as follow: I) Multiple Diseases associating with One Drug (MDiODr); II) Multiple Drugs associating with One Disease (MDrODi). Considering the practical merits for the first one analysis, we required multiple diseases belongs to the same type of diseases. To evaluate the capability of KDDANet for revealing the shared genes mediating multiple KDDAs, we produced meta-subnetworks by integrating multiple KDDANet resulting subnetworks for 12386 MDiODr combinations and 773 MDrODi combinations produced in SDrTDi context, and 12189 MDiODr combinations and 773 MDrODi combinations in SDiTDr context. The weight of an edge in the meta-subnetwork was defined as the number of KDDA resulting subnetworks containing this edge divided by the total number of KDDA resulting subnetworks. The higher weight value of a link in the meta-subnetwork indicated more conservation and commonality. Thus, we used the weight value to evaluate the capacity of KDDANet in unveiling the shared gene interactions mediating multiple KDDAs. We carried out a permutation test by producing random meta-subnetworks with the same number of edges for comparing with random background. As shown in Fig. 5a,

Supplementary Figure 5a and Supplementary Figure 5b, the weights of KDDANet meta-subnetworks were significantly higher than random meta-subnetworks across different types of diseases in SDrTDi context. This indicate that the genes tend to be shared in KDDANet meta-subnetworks than random one. We also conducted the same analysis in SDiTDr context and obtained the similar results (Fig. 5b, Supplementary Figure 5c and Supplementary Figure 5d). Collectively, these results indicated that KDDANet can effectively uncover the shared genes mediating multiple KDDAs.

We presented some examples for describing the capability of KDDANet in identifying shared genes mediating multiple KDDAs. For MDiODr, we selected two cases: I) profenamine (DB00392) and neurological disease associations and II) mirtazapine (DB00370) and cancer associations. As reported in CTD database, profenamine was associated with three neurological diseases, including Parkinsonian Disorders, Multiple Sclerosis and Alzheimer Disease. The shared meta-subnetwork contained 75 edges linking 43 unknown genes with three profenamine's target genes and 30 neurological disease-related genes (Fig. 5c). The top 10 enriched KEGG terms were demonstrated in Fig. 5d. A majority of enriched KEGG terms, such as Cholinergic synapse and Glutamatergic synapses, have been reported dysfunction in neurological disorders and diseases [83,84]. Interestingly, we found that GNAI2 and GNA11 were two mostly shared genes linking profenamine with neurological diseases. These two genes were recently discovered involving in the pathological pathways of neurological diseases [85,86]. Mirtazapine was associated with 9 types of cancers, including Colorectal Neoplasms, Breast Neoplasms, Neuroblastoma, Glioma, Urinary Bladder Neoplasms, Stomach Neoplasms, Esophageal Neoplasms, Lung Neoplasms and Prostatic Neoplasms. Supplementary Figure 5e showed the shared meta-subnetwork that included 97 edges connecting 21 mirtazapine's target genes with 33 cancer-related genes and 54 unknown genes. As expected, KEGG enrichment

demonstrated that these genes were involved in cancer-related signaling pathways and played important roles in various cancers (Supplementary Figure 5f). Intriguingly, DRD4 and GRB2 were two mostly shared genes mediating the associations between mirtazapine and cancers (Supplementary Figure 5e). These two genes were involved in the oncogenesis of multiple types of cancers [87,88].

For MDrODi, some interesting cases were also observed. For example, GRACILE syndrome, a metabolic disease, was associated with 13 drugs as recorded in CTD database. A shared meta-subnetwork including 66 genes and 61 edges were obtained for this disease (Fig. 5e). This meta-subnetwork contained 14 drug target genes and 13 GRACILE syndrome-related genes. The mostly shared gene was ATP5B, and the mostly significant enriched KEGG term was Oxidative phosphorylation (Fig. 5f). This was expected as GRACILE syndrome was a fatal inherited disorder caused by a mutation in an oxidative phosphorylation related gene, BCS1L[89]. It was also not surprising that the neurological diseases related genes were also enriched in this meta-subnetwork as patients with GRACILE syndrome had severe neurological problems[89]. Another example was the Keratoconus (148300), an ophthamological disease, which was associated with three different drugs, including acetaminophen, valproic acid and theophylline. Keratoconus and these three drugs shared a meta-subnetwork consisting of 54 genes and 51 edges (Supplementary Figure 5g). This meta-subnetwork included 7 Keratoconus-related genes, 11 drug target genes and 36 unknown genes. It was expected that HDAC2 had high weights with its partner genes in this meta-subnetwork as it was involved in notch signaling pathway which is downregulated in keratoconus [90]. Consistent with the associations between collagen genes and keratoconus [91], a novel collagens coding gene, COL1A1, was captured in this meta-subnetwork as a highly shared gene. The enriched KEGG terms of this meta-subnetwork included Hippo signaling pathway (Supplementary Figure 5h) which has been reported involving in keratoconus corneas [92]. Collectively, these

results indicated that KDDANet can discover the shared genes mediating multiple KDDAs. The highly shared unknown genes can serve as potential candidate targets for drug repurposing.

Discussion

To facilitate drug repurposing, various computational tools have been developed to uncover novel drug-disease associations [93]. However, the potential genes mediating KDDA have been still not fully explored. Unveiling the hidden genes (missed from experiments) mediating KDDA become a great challenge for guiding novel target discovery and drug repurposing. In this work, we developed a novel computational tool, KDDANet, which integrated minimum cost network flow optimization, depth first searching and graph clustering algorithm to reveal hidden genes and modules mediating KDDA. KDDANet allowed for a global and systematic exploration of the hidden genes mediating KDDA. We applied KDDANet to unravel the subnetworks of genes mediating 53124 KDDAs. The comprehensive and system-level evaluations fully demonstrated the effective prediction capability and general applicability of KDDANet. Case studies on both AD and obesity showed that the subnetworks of genes identified by KDDANet were reliable and useful. Further validated by integrating analysis of genomic, transcriptomic, pharmacogenomic and survival data on primary tumors and cancer cell lines highlighted that KDDANet captured novel candidates from interactome mediating the associations between drugs and different types of cancer. Based on these, we concluded that the inferred subnetworks mediating KDDA can serve as genome-wide molecular landscapes for guiding drug repurposing and disease treatment. Insights learned from our predictions would also enable to help researchers to design repurposing drugs to reverse disease phenotypes via targeting key genes in the subnetwork mediating KDDA. An important capability of KDDANet method was that it can reveal the shared genes mediating multiple KDDAs. This provided more valuable guides for drug repositioning

and disease treatment since the shared genes mediating multiple KDDAs were closely linked to the molecular basis of drug repurposing. We constructed a user-friendly online web tool (http://www.kddanet.cn), which allowed users to explore the subnetwork of genes mediating individual KDDA and the meta-subnetworks mediating multiple KDDAs (See Supplementary Note 7 as well as the Help section of our website for detailed description of the utility of an online web version of KDDANet). In summary, we presented a novel computational tool KDDANet and an online web source to decode the hidden genes mediating KDDA that had broad utility and application value in biomedical studies.

The hidden genes mediating KDDA predicted by KDDANet highlighted the power of integrative approaches to illuminate underexplored molecular processes mediating KDDA. In the future, the application value of our KDDANet tool in drug repurposing can be further improved from three aspects as follow: Firstly, the gene interactome used in KDDANet did not contain enough interactions between genome elements. In further studies, we would integrate other non-coding genome elements, especially long non-coding RNA and microRNA for constructing a comprehensive interactome. Secondly, integrating the biological networks from other omics layers, such as epigenomics, might have also further enhance the accuracy of KDDANet in discovering subnetworks and key genes mediating KDDA, and help us better to understand KDDA at multi-omics levels. The intrinsically capability of KDDANet to analyze large-scale heterogeneous interactome data containing tens of thousands of nodes and edges make it can well be suited to analyzing the accumulating data from 'multi-omics' technologies and biomedical research. Finally, KDDANet did not carry out predictions for KDDA when a drug's target genes were unknown or when a disease has not been related to any known gene. For this, we planned to calculate the similarity scores between drugs and the similarity scores between diseases, and then integrate drugs without any target gene and diseases without any related gene

to our flow network model by similarity scores.

Methods

Datasets

Five different types of gene networks, including HumanNet, HINT+HI2012, iRefIndex, MultiNet and STRINGv10, were used in our current study [94,95], in which the nodes were represented by gene ID and connected by bidirectional edges (Supplementary Data 1). Drugs and their target genes were obtained from DrugBank 5.0.9 database (http://www.drugbank.ca/). In this study, we selected 4861 drugs with at least one known target that was contained in the gene networks for further analysis. In total, 2196 Known Drug Target Genes (KDTGs) included in the gene networks were connected to these drugs by 12014 interactions (Supplementary Data 2). Known Disease Related Genes (KDRGs) and classification of diseases were obtained by manually collecting Human Disease Network (HDN) from the previous study of Human Disease Network (HDN) [33] and DisGeNET v5.0 database [96]. We focused on 1441 diseases with at least one related gene which was included in the gene networks for our study. In total, 16712 associations link these diseases to 1521 genes which were exist in the gene networks (Supplementary Data 3). The KDDAs were extracted from Comparative Toxicogenomics Database [12] (Supplementary Data 4). In this study, 53124 KDDAs were analyzed in which the drug had at least one target gene and the disease had at least one related gene contained in HumanNet. For simplicity and consistency, we converted different types of drugs, diseases, and gene nomenclatures to DrugBank drug ID, OMIM disease ID and NCBI Entrez gene ID for subsequent modeling and analysis.

The primary RNA-seq datasets of Alzheimer's disease (AD) patients, obesity patients and normal individuals were downloaded from NCBI GEO Datasets under accession number of GSE53697, GSE81965 and GSE63887. After the SRA files were gathered,

the archives were extracted and saved in FASTQ format using the SRA Toolkit. RNA-seq reads were trimmed using Trimmomatic software (http://www.usadellab.org/cms/?page=trimmomatic) with the following parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Version 0.36), and were further quality-filtered using FASTX Toolkit's fastq_quality_trimmer command (http://hannonlab.cshl.edu/fastx_toolkit/) (Version 0.0.13) with the minimum quality score 20 and minimum percent of 80% bases that had a quality score larger than this cutoff value. The high-quality reads were mapped to the hg38 genome by HISAT2, a fast and sensitive spliced alignment program for mapping RNA-seq reads, with -dta paramenter (http://daehwankimlab.github.io/hisat2/). PCR duplicate reads were removed using Picard tools (http://broadinstitute.github.io/picard/) and only uniquely mapped reads were kept for further analysis. The expression levels of genes were calculated by StringTie (http://ccb.jhu.edu/software/stringtie/) (Version v1.3.4d) with -e -B -G parameters using Release 29 (GRCh38.p12) gene annotations downloaded from GENCODE data portal (https://www.gencodegenes.org/). To obtain comparable expression abundance estimation for each gene, reads mapped to hg38 were counted as FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) based on their genome locations. Differential expression analysis of genes was performed by DESeq2 using the reads count matrix produced from a python script "prepDE.py" provided in StringTie website (http://ccb.jhu.edu/software/stringtie/).

TCGA cancer genomics datasets were directly downloaded via the UCSC Xena project data portal (https://xenabrowser.net/datapages/). As the DNA methylation datasets were quantified as beta value in the DNA probe level, we mapped Illumina Human Methylation 450 probe ID to gene name using HumanMethylation450 annotation file. If a gene was mapped by multiple probes, we considered the averaged signals of these probes as the methylation level of this gene. We used Wilcox signed rank test

for differential analysis of gene expression and DNA methylation of KDDANet resulting subnetwork genes mediating the associations between drug and cancer. For this analysis, we only used the tumor samples which have adjacent normal tissue samples as control. Genomics of Drug Sensitivity in Cancer 1000 (GDSC1000) cancer cell line pharmacogenomic datasets were downloaded from GDSC website (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html). Since this website provided only the CpG island methylation data, we downloaded the beta value matrix of probe-level DNA methylation from NCBI GEO Dataset under accession number of GSE68379 and then converted it to gene-level beta value matrix using the same method as TCGA DNA methylation data. Cancer Cell Line Encyclopedia (CCLE) pharmacogenomic datasets were directly downloaded from Broad Institute data portal (https://portals.broadinstitute.org/ccle/data).

Construction of a unified flow network model.

For each query drug and all its related diseases, the unified flow network model in SDrTDi context was built by integrating the query drug and all its related diseases into the gene network based on known drug-target relationships and gene-disease associations. As shown in Fig. 1, for the given query drug, we used it as source node (S) and integrated it into gene network by introducing the directed edges from it point to its target genes. For each its related disease, we mapped the disease to gene network by introducing the directed edges from its related genes point to it. We incorporated a sink node T and introduced a directed edge pointing from each disease to it. With these definitions, a unified flow network model was constructed as a complex heterogeneous graph $G = (V, E)$, where V was the set of vertices and E was the set of edges. This graph included two types of edges (bidirectional and directed) and three types of nodes (drugs, genes, and diseases). Each edge was assigned with a weight and a capacity. Flow goes from a source node to a sink node through the graph edges. The assigning scheme of weight and capacity was

illustrated as follow:

Weight and capacity assigning scheme for network edges.

Edges between gene nodes. Edges between gene nodes were weighted (W_{ij}) to reflect the probability that two genes g_i and g_j were functionally linked in the biological processes. The weight value between g_i and g_j was derived from a Bayesian statistics approach by integrating diverse functional genomics datasets [94]. Briefly, each dataset was benchmarked for its real capability of reconstructing known cellular pathways by measuring the likelihood that pairs of genes (linkages) were functionally connected conditioned on the experimental evidence, calculated as a log likelihood score *LLS* [97]:

$$LLS = \ln\left\{\frac{F(L\,|\,E)/{\sim}F(L\,|\,E)}{F(L)/{\sim}F(L)}\right\} \tag{1}$$

Where *F(L|E)* and *~F(L|E)* represented the observed numbers of linkages *L* appear in the given experiment *E* between functionally annotated human genes interacting within the same pathway and between different pathways, respectively, *F(L)* and *~F(L)* denoted the total observed numbers of linkages between all annotated human genes interacting within the same pathway and between different pathways, respectively. The weight value (*W*) of a given linkage between two genes was produced by combining LLS score of each dataset using the formula:

$$W = LLS_{best} + \sum_{i=1}^{N} \frac{LLS_i}{D \cdot i},\ for\ all\ L \geq T \tag{2}$$

Where LLS_{best} represented the highest *LLS* score for a linkage between two genes, *D* determined decay rate of the LLS score for additional evidence, and *i* was the order index of LLS scores for a given linkage between two genes, ranking starting from the second maximum LLS score with descending order of magnitude for all N remaining LLS scores. *T* represented a minimum threshold of LLS score to be considered. The values of *D* and *T* were empirically optimized to maximize overall performance on

known GO annotations measured by AUPRC [98].

Edges between drug and drug's target gene nodes: Edges between each drug and drug's target gene nodes were weighted (w_{Si}) to reflect the normalized reliability of the interaction between drug and target protein based on experimental and computational evidence. The weighting scheme was based on the predicted score of drug and target protein interaction [25]. The weight value (W_{Si}) was calculated as:

$$W_{Si} = \frac{P_i}{\sum_{j \in T} P_j} \tag{3}$$

Where T denoted the set of each drug's targets, P_i denotes the predicted score between drug S and target i, P_j denoted the predicted score between drug S and target j.

Edges between disease-related gene and disease nodes. Edges between disease-related gene and disease nodes were weighted (w_{jd}) to reflect the normalized reliability of the linkage between disease and gene based on experimental and computational evidence. The weighting scheme was based on the predicted score of gene-disease association derived from MAXIF algorithm [99]. We calculated the weight value (W_{jd}) as follow:

$$W_{jd} = \frac{F_i}{\sum_{j \in D} F_j} \tag{4}$$

Where D denoted the set of genes linked to disease d, F_i denoted the predicted score between gene i and disease d, P_j denoted the predicted score between gene j and disease d.

Edges between disease and sink nodes. For each edge linking each disease d to sink node T, we assigned it a same weight value $W_{dT} = 1/N$, where N denoted the number of diseases linking to sink node.

We further defined for each edge in this heterogeneous network a capacity value that limited the flow quantity. For each edge connecting the query drug *S* to its target gene *i*, we assigned it a capacity C_{Si} equal to W_{Si}. For each edge linking the disease-related gene *j* to disease *d*, we assigned it a capacity C_{jd} equal to $W_{jd.}$ For each edge linking the disease *d* to sink node *T*, we assigned it a capacity C_{dT} equal to $W_{dT.}$ For other edges, we assigned them a capacity $C_{ij} = 1$.

Minimum cost flow optimization algorithm.

With the purpose of identifying hidden genes mediating KDDA, we search for an possible solution that would (I) capture the subset of the query drug's target genes which closely modulate all its related diseases by the disease-related genes without restrict to the prior KDDA genes, (II) determine hidden genes that were likely to be part of cellular pathways connecting the query drug's target genes to all its related diseases but escaped detection by experiments, (III) give high priority to genes that lie on paths with highest probability connecting the query drug to all its related diseases without making constraints on the network structure. The rationality for proposing this solution including two aspects: I) this needed less computational time than that finding the highest probability subnetwork connects a query drug to each its related diseases at a time; II) this can effectively find the shared genes mediating multiple KDDAs. Inspired from the fact of water always flowing through a path of least resistance, we formulated this goal as a minimum cost flow optimization problem [44,100,101]. Cost was defined as the negative log of the probability of an edge. Hence, minimizing the cost gave preference to highest-probability paths. Given the unified flow network, this problem can be expressed as a linear programming formula that minimized the total cost of the flow network while diffusing the most flow from query drug node to hypothetical sink node. Let W_{ij}, F_{ij} and C_{ij} referred to the weight, flow, and capacity from node *i* to node *j*, respectively. The linear

programming formula can be written as follow.

Minimize $$\sum_{i\in V, j\in V}\left(-\log\left(W_{ij}\right) * F_{ij}\right) - \left(\gamma * \sum F_{Sv}\right) \quad (5)$$

Subject to $$\sum_{j\in V} F_{ij} - \sum_{j\in V} F_{ji} \quad \forall i \in V - \{S, T\} \quad (6)$$

$$F_{Sv} - \sum_{i\in D} F_{iT} = 0 \quad (7)$$

$$0 \leq F_{ij} \leq C_{ij} \quad (8)$$

Where *V* denoted a set of nodes included in the flow network; *S* denoted the query drug and *v* denoted its target gene. The parameter gamma (γ) controlled the size and the quality of the optimized subnetwork. The first component of this formula, $\sum_{i\in V, j\in V}\left(-\log\left(W_{ij}\right) * F_{ij}\right)$, ensured minimizing the network cost that given priority to obtain highest probability gene subnetwork, at the same time, the second component, $-\left(\gamma * \sum F_{Sv}\right)$, indicated maximizing the total flow across entire network. This optimization problem can be efficiently solved using primal simplex method provided in Mixed Integer Linear Programming (MILP) solver (http://lpsolve.sourceforge.net/). The solution $argmin_{f_{ij}>0} \sum_{i\in V, j\in V}\left(-\log\left(w_{ij}\right) * f_{ij}\right) - \left(\gamma * \sum f_{Sv}\right)$ obtained the highest probability subnetwork mediating the associations between query drug and all its related diseases.

Depth first searching and Markov clustering (MCL).

Once the highest probability subnetwork mediating the query drug (disease) with all its related diseases (drugs) was obtained, we implemented depth first searching [102] on this subgraph to find the subnetwork made up of all paths linking the query drug (disease) to each it's related disease (drug). All genes in the solution were ranked by the amount of flow they carry. The more flow that passes through a protein, the more important it was in mediating KDDA. After obtaining the subnetwork mediating individual KDDA, Markov clustering (MCL) (https://micans.org/mcl/) was employed to further discover gene modules mediating KDDA by using the flow quantities through edges of subnetwork as weight values.

Application KDDANet to SDiTDr context.

To apply KDDANet in SDiTDr context, the query disease, and a set of all its related drugs were mapped into gene network by disease-related genes and drug target genes. For constructing flow network model, the weights and capacities of network edges can be assigned using the similar method as described in SDrTDi context by substituting query drug with query disease and using query disease as source node (*S*), substituting query drug related diseases with query disease related drugs, substituting query drug's target genes with query disease related genes, substituting disease related genes with drug's target genes, and incorporating a sink node *T* and introducing a directed edge pointing from each drug to it. Implementing minimum cost flow optimization, depth first searching and Markov clustering (MCL) on the unified flow network was same as SDrTDi context.

Performance evaluation.

As the true genes mediating KDDA are poorly understood, there was no perfect way to assess the prediction results. The predictive performance of KDDANet was evaluated as follow: I) Based on the hypothesis that the larger functional similarity between a gene and known KDDA genes, the higher probability this gene was positive one mediating KDDA, we compiled a standard set of positive and negative KDDA genes for each KDDA resulting subnetwork to unbiasedly evaluate the performance of KDDANet using the following strategy:

I) For each gene in the KDDA resulting subnetwork, we first calculated the mean functional similarity scores of it with KDTGs and KDRGs by our previous published method using Gene Ontology (GO), KEGG pathways and InterPro annotation as functional terms [103].

II) We performed a permutation test by randomly producing KDTGs and KDRGs 1000 times to compute the empirical significance level of functional similarity. We selected

the genes having significant functional similarities with both KDTGs and KDRGs from KDDA resulting subnetwork as positive KDDA genes using the criterion that the similarity score was larger than 95th percentiles of the simulated background distributions. The other genes were considered as negative KDDA genes.

Based on this gold standard, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were produced and the areas of under curve of ROC and PR (AUROC and AUPRC) were calculated for gene list ranked by flow amount of each KDDANet resulting subnetwork. For comparison with PPA, SNPLS, comCHIPER and DGPsubNet, we used the default parameters as mentioned in their published papers. We considered that all genes contained in a co-module mediating KDDA of each drug and disease pair in the co-module. The genes were ranked by probability score mediating a KDDA. For PPA and SNPLS, the probability score of a gene *g* mediating the association between drug *i* and a type of cancer *j* was defined as averaged prediction score across all cell lines of this type of cancer [40,41]. For comCHIPER, the probability score of a gene *g* mediating the association between drug *i* and disease *j* was defined as the sum of the products of posterior indicator probabilities across all co-modules [42]. For DGPsubNet, we defined the probability score of a gene *g* mediating the association between drug *i* and disease *j* as the calculated *z-score* z_{ij} [43]. The areas of under curve of ROC and PR (AUROC and AUPR) were calculated using R package of PRROC. KDDANet resulting subnetwork visualization was carried out using Cytoscape software (https://cytoscape.org/). KEGG pathway enrichment analysis was performed by clusterProfiler R package [104]. The visualization of results was carried out in R software. Case studies were conducted in SDrTDi context unless specifically indicated.

Data Availability

The authors declare that data supporting the findings of this study are available within the paper and its supplementary information files.

Code availability

The code of KDDANet was freely available at https://github.com/huayu1111/KDDANet.

Acknowledgements

H.Y. was supported by Project funded by China Postdoctoral Science Foundation (Grant No. 2018M642441), Zhejiang Natural Science Foundation Projects of China (Grant No. LQ21C120002) and High-Performance Computing Platform in Center of Cryo-Electron Microscopy of Zhejiang University.

Competing Interests

The authors declare that there are no competing interests.

Author Contributions

H.Y. and L.L. conceived the original research plans, analyzed the data, developed software, and wrote the article; L.L and H.Y developed the web server. H.Y., supervised the experiments; H.Y. and L.L. contributed equally to this work. J.Z., C.L., and M.C. provided helpful suggestions and computational support. All authors revised and confirmed the paper.

References

1. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203-214, doi:10.1038/nrd3078 (2010).

2. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**, 673-683, doi:10.1038/nrd1468 (2004).

3. Bradley, D. Why big pharma needs to learn the three 'R's. *Nature Reviews Drug Discovery* **4**, 446, doi:10.1038/nrd1766 (2005).

4. Ghofrani, H. A., Osterloh, I. H. & Grimminger, F. Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery* **5**, 689-702, doi:10.1038/nrd2030 (2006).

5. Hurle, M. R. *et al.* Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology and Therapeutics* **93**, 335-341, doi:10.1038/clpt.2013.1 (2013).

6. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935, doi:10.1126/science.1132939 (2006).

7. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

8. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754, doi:10.1016/j.cell.2016.06.017 (2016).

9. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

10. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

11. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074-d1082, doi:10.1093/nar/gkx1037 (2018).

12. Davis, A. P. *et al.* Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research* **37**, D786-792, doi:10.1093/nar/gkn580 (2009).

13. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833-d839, doi:10.1093/nar/gkw943 (2017).

14. Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems* **6**, 13-24, doi:10.1016/j.cels.2017.11.001 (2018).

15. Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E. & Sharan, R. INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology* **8**,

592, doi:10.1038/msb.2012.26 (2012).

16. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Science Translational Medicine 4, 125ra31. doi: 10.1126/scitranslmed.3003377 (2012).

17. Zhao, X. M. *et al.* Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Computational Biology* **7**, e1002323, doi:10.1371/journal.pcbi.1002323 (2011).

18. Cheng, F., Kovács, I. A. & Barabási, A. L. Network-based prediction of drug combinations. *Nature Communications* **10**, 1197, doi:10.1038/s41467-019-09186-x (2019).

19. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* **38**, 672-684.e676, doi:10.1016/j.ccell.2020.09.014 (2020).

20. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology* **7**, 496, doi:10.1038/msb.2011.26 (2011).

21. Guney, E., Menche, J., Vidal, M. & Barábasi, A. L. Network-based in silico drug efficacy screening. *Nature Communications* **7**, 10331, doi:10.1038/ncomms10331 (2016).

22. Cheng, F. & Desai, R. J. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications* **9**, 2691, doi:10.1038/s41467-018-05116-5 (2018).

23. Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nature Biotechnology* **24**, 805-815, doi:10.1038/nbt1228 (2006).

24. Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263-266, doi:10.1126/science.1158140 (2008).

25. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data. *PLoS ONE* **7**(5): e37608, https://doi.org/10.1371/journal.pone.0037608 (2012).

26. Luo, Y., Zhao, X., Zhou, J. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communication* **8**, 573 (2017). https://doi.org/10.1038/s41467-017-00680-8.

27. Gottlieb, A., Magger, O., Berman, I., Ruppin, E. & Sharan, R. PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics* **27**, 3325-3326, doi:10.1093/bioinformatics/btr584 (2011).

28. Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Molecular Systems Biology* **4**, 189, doi:10.1038/msb.2008.27 (2008).

29. Ghiassian, S. D., Menche, J. & Barabási, A. L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the

human interactome. *PLoS Computational Biology* **11**, e1004120, doi:10.1371/journal.pcbi.1004120 (2015).

30. Zhou, X., Menche, J., Barabási, A. L. & Sharma, A. Human symptoms-disease network. *Nature Communications* **5**, 4212, doi:10.1038/ncomms5212 (2014).

31. Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601, doi:10.1126/science.1257601 (2015).

32. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature Methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).

33. Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).

34. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56-68, doi:10.1038/nrg2918 (2011).

35. Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabási, A. L. & Vidal, M. Drug-target network. *Nature Biotechnology* **25**, 1119-1126, doi:10.1038/nbt1338 (2007).

36. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4**, 682-690, doi:10.1038/nchembio.118 (2008).

37. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113, doi:10.1038/nrg1272 (2004).

38. Vidal, M., Cusick, M. E. & Barabási, A. L. Interactome networks and human disease. *Cell* **144**, 986-998, doi:10.1016/j.cell.2011.02.016 (2011).

39. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551-562, doi:10.1038/nrg.2017.38 (2017).

40. Chen, J. & Zhang, S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* **32**, 1724-1732, doi:10.1093/bioinformatics/btw059 (2016).

41. Kutalik, Z., Beckmann, J. S. & Bergmann, S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology* **26**, 531-539, doi:10.1038/nbt1397 (2008).

42. Zhao, S. & Li, S. A co-module approach for elucidating drug-disease associations and revealing their molecular basis. *Bioinformatics* **28**, 955-961, doi:10.1093/bioinformatics/bts057 (2012).

43. 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**, e146, doi:10.1038/psp.2014.44 (2014).

44. Yeger-Lotem, E. *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics* **41**, 316-323, doi:10.1038/ng.337

(2009).

45. Jewison, T. *et al.* SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Research* **42**, D478-484, doi:10.1093/nar/gkt1067 (2014).

46. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845-d855, doi:10.1093/nar/gkz1021 (2020).

47. Alisi, L. *et al.* The Relationships Between Vitamin K and Cognition: A Review of Current Evidence. *Frontiers in Neurology* **10**, 239, doi:10.3389/fneur.2019.00239 (2019).

48. Chen, X. Q., Sawa, M. & Mobley, W. C. Dysregulation of neurotrophin signaling in the pathogenesis of Alzheimer disease and of Alzheimer disease in Down syndrome. *Free Radical Biology & Medicine* **114**, 52-61, doi:10.1016/j.freeradbiomed.2017.10.341 (2018).

49. Oliveira, T. G. & Di Paolo, G. Phospholipase D in brain function and Alzheimer's disease. *Biochimica Et Biophysica Acta* **1801**, 799-805, doi:10.1016/j.bbalip.2010.04.004 (2010).

50. Buonanno, A. & Fischbach, G. D. Neuregulin and ErbB receptor signaling pathways in the nervous system. *Current Opinion in Neurobiology* **11**, 287-296, doi:10.1016/s0959-4388(00)00210-5 (2001).

51. Gomez, J. M. Growth hormone and insulin-like growth factor-I as an endocrine axis in Alzheimer's disease. *Endocrine, Metabolic & Immune Disorders Drug Targets* **8**, 143-151, doi:10.2174/187153008784534367 (2008).

52. Pardeshi, R. *et al.* Insulin signaling: An opportunistic target to minify the risk of Alzheimer's disease. *Psychoneuroendocrinology* **83**, 159-171, doi:10.1016/j.psyneuen.2017.05.004 (2017).

53. Yan, S. D., Bierhaus, A., Nawroth, P. P. & Stern, D. M. RAGE and Alzheimer's disease: a progression factor for amyloid-beta-induced cellular perturbation? *Journal of Alzheimer's Disease* **16**, 833-843, doi:10.3233/jad-2009-1030 (2009).

54. Vlassenko, A. G. & Raichle, M. E. Brain aerobic glycolysis functions and Alzheimer's disease. *Clinical and Translational Imaging* **3**, 27-37, doi:10.1007/s40336-014-0094-7 (2015).

55. Krance, S. H. *et al.* The complement cascade in Alzheimer's disease: a systematic review and meta-analysis. *Molecular Psychiatry*, doi:10.1038/s41380-019-0536-8 (2019).

56. Scheckel, C. *et al.* Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. *eLife* **5**, doi:10.7554/eLife.10421 (2016).

57. Reichenbach, N. *et al.* Inhibition of Stat3-mediated astrogliosis ameliorates pathology in an Alzheimer's disease model. *EMBO Molecular Medicine* **11**, doi:10.15252/emmm.201809665 (2019).

58. Choi, M., Kim, H., Yang, E. J. & Kim, H. S. Inhibition of STAT3 phosphorylation attenuates impairments in learning and memory in 5XFAD mice, an animal model of Alzheimer's disease. *Journal of Pharmacological Sciences*, doi:10.1016/j.jphs.2020.05.009 (2020).

59. Silver, M., Janousova, E., Hua, X., Thompson, P. M. & Montana, G. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage* **63**, 1681-1694, doi:10.1016/j.neuroimage.2012.08.002 (2012).

60. Zhang, H. *et al.* Impact of Vitamin D Binding Protein Levels on Alzheimer's Disease: A Mendelian Randomization Study. *Journal of Alzheimer's Disease* **74**, 991-998, doi:10.3233/jad-191051 (2020).

61. Kahn, B. B. & Flier, J. S. Obesity and insulin resistance. *The Journal of Clinical Investigation* **106**, 473-481, doi:10.1172/jci10842 (2000).

62. Taylor, V. H. & Macqueen, G. M. The Role of Adipokines in Understanding the Associations between Obesity and Depression. *Journal of Obesity* **2010**, doi:10.1155/2010/748048 (2010).

63. Langin, D. *et al.* Adipocyte lipases and defect of lipolysis in human obesity. *Diabetes* **54**, 3190-3197, doi:10.2337/diabetes.54.11.3190 (2005).

64. Olsson, U. *et al.* Changes in matrix proteoglycans induced by insulin and fatty acids in hepatic cells may contribute to dyslipidemia of insulin resistance. *Diabetes* **50**, 2126-2132, doi:10.2337/diabetes.50.9.2126 (2001).

65. Viollet, B. *et al.* Targeting the AMPK pathway for the treatment of Type 2 diabetes. *Frontiers in Bioscience* **14**, 3380-3400 (2009).

66. Väremo, L. *et al.* Type 2 diabetes and obesity induce similar transcriptional reprogramming in human myocytes. *Genome Medicine* **9**, 47, doi:10.1186/s13073-017-0432-2 (2017).

67. Väremo, L. *et al.* Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell Reports* **11**, 921-933, doi:10.1016/j.celrep.2015.04.010 (2015).

68. Wu, C. *et al.* Enhancing hepatic glycolysis reduces obesity: differential effects on lipogenesis depend on site of glycolytic modulation. *Cell Metabolism* **2**, 131-140, doi:10.1016/j.cmet.2005.07.003 (2005).

69. Isoppo de Souza, C. *et al.* Association of adipokines and adhesion molecules with indicators of obesity in women undergoing mammography screening. *Nutrition & Metabolism* **9**, 97, doi:10.1186/1743-7075-9-97 (2012).

70. Santilli, F., Vazzana, N., Liani, R., Guagnano, M. T. & Davì, G. Platelet activation in obesity and metabolic syndrome. *Obesity Reviews* **13**, 27-42, doi:10.1111/j.1467-789X.2011.00930.x (2012).

71. Zheng, L. Plasminogen: a potential target gene for dietary supplements and biomarker of the early stage of obesity by fatigue mice. *Biomedical Research* **28**, 4299-4304 (2017).

72. Vogelstein, B. *et al.* Cancer genome landscapes. *Science (New York, N.Y.)* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

73. Kaapu, K. J., Ahti, J., Tammela, T. L., Auvinen, A. & Murtola, T. J. Sotalol, but not digoxin is associated with decreased prostate cancer risk: A population-based case-control study.

International Journal of Cancer **137**, 1187-1195, doi:10.1002/ijc.29470 (2015).

74. Bao, J. M. *et al.* AGE/RAGE/Akt pathway contributes to prostate cancer cell proliferation by promoting Rb phosphorylation and degradation. *American Journal of Cancer Research* **5**, 1741-1750 (2015).

75. Neschadim, A., Summerlee, A. J. & Silvertown, J. D. Targeting the relaxin hormonal pathway in prostate cancer. *International Journal of Cancer* **137**, 2287-2295, doi:10.1002/ijc.29079 (2015).

76. Ozvegy-Laczka, C., Cserepes, J., Elkind, N. B. & Sarkadi, B. Tyrosine kinase inhibitor resistance in cancer: role of ABC multidrug transporters. *Drug Resistance Updates* **8**, 15-26, doi:10.1016/j.drup.2005.02.002 (2005).

77. Eke, I. & Cordes, N. Focal adhesion signaling and therapy resistance in cancer. *Seminars in Cancer Biology* **31**, 65-75, doi:10.1016/j.semcancer.2014.07.009 (2015).

78. Farrow, J. M., Yang, J. C. & Evans, C. P. Autophagy as a modulator and target in prostate cancer. *Nature Reviews Urology* **11**, 508-516, doi:10.1038/nrurol.2014.196 (2014).

79. Heidegger, I., Kern, J., Ofer, P., Klocker, H. & Massoner, P. Oncogenic functions of IGF1R and INSR in prostate cancer include enhanced tumor growth, cell migration and angiogenesis. *Oncotarget* **5**, 2723-2735, doi:10.18632/oncotarget.1884 (2014).

80. Tai, I. T. & Tang, M. J. SPARC in cancer biology: its role in cancer progression and potential for therapy. *Drug Resistance Update* **11**, 231-246, doi:10.1016/j.drup.2008.08.005 (2008).

81. Brizzolara, A. *et al.* The ErbB family and androgen receptor signaling are targets of Celecoxib in prostate cancer. *Cancer Letters* **400**, 9-17, doi:10.1016/j.canlet.2017.04.025 (2017).

82. Gannon, H. S., Zou, T., Kiessling, M. K. & Gao, G. F. Identification of ADAR1 adenosine deaminase dependency in a subset of cancer cells. *Nature Communication* **9**, 5450, doi:10.1038/s41467-018-07824-4 (2018).

83. Moretto, E., Murru, L., Martano, G., Sassone, J. & Passafaro, M. Glutamatergic synapses in neurodevelopmental disorders. *Progress in Neuropsychopharmacology & Biological Psychiatry* **84**, 328-342, doi:10.1016/j.pnpbp.2017.09.014 (2018).

84. Tata, A. M., Velluto, L., D'Angelo, C. & Reale, M. Cholinergic system dysfunction and neurodegenerative diseases: cause or effect? *CNS & Neurological Disorders Drug Targets* **13**, 1294-1303, doi:10.2174/1871527313666140917121132 (2014).

85. Zhang, M. *et al.* Genome-wide pathway-based association analysis identifies risk pathways associated with Parkinson's disease. *Neuroscience* **340**, 398-410, doi:10.1016/j.neuroscience.2016.11.004 (2017).

86. Zhang, Q. *et al.* Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. *Acta Neuropathologica Communications* **6**, 19, doi:10.1186/s40478-018-0524-2 (2018).

87. Ijaz, M. *et al.* The Role of Grb2 in Cancer and Peptides as Grb2 Antagonists. *Protein and*

Peptide Letters **24**, 1084-1095, doi:10.2174/0929866525666171123213148 (2018).

88. Weissenrieder, J. S., Neighbors, J. D., Mailman, R. B. & Hohl, R. J. Cancer and the Dopamine D(2) Receptor: A Pharmacological Perspective. *The Journal of Pharmacology and Experimental Therapeutics* **370**, 111-126, doi:10.1124/jpet.119.256818 (2019).

89. Visapää I, Fellman V, Vesa J, Dasvarma A, Hutton JL, Kumar V, Payne GS, Makarow M, Van Coster R, Taylor RW, Turnbull DM, Suomalainen A, Peltonen L. GRACILE syndrome, a lethal metabolic disorder with iron overload, is caused by a point mutation in BCS1L. *American Journal of Human Genetics* **71**, 863-876, doi: 10.1086/342773 (2002).

90. You, J., Corley, S. M., Wen, L. & Hodge, C. RNA-Seq analysis and comparison of corneal epithelium in keratoconus and myopia patients. *Scientific Reports* **8**, 389, doi:10.1038/s41598-017-18480-x (2018).

91. Bykhovskaya, Y., Margines, B. & Rabinowitz, Y. S. Genetics in Keratoconus: where are we? *Eye and Vision* **3**, 16, doi:10.1186/s40662-016-0047-5 (2016).

92. Kabza, M. *et al.* Collagen synthesis disruption and downregulation of core elements of TGF-β, Hippo, and Wnt pathways in keratoconus corneas. *European Journal of Human Genetics* **25**, 582-590, doi:10.1038/ejhg.2017.4 (2017).

93. Lu, L. & Yu, H. DR2DI: a powerful computational tool for predicting novel drug-disease associations. *Journal of Computer-Aided Molecular Design* **32**, 633-642, doi:10.1007/s10822-018-0117-y (2018).

94. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* **21**, 1109-1121, doi:10.1101/gr.118992.110 (2011).

95. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**, 106-114, doi:10.1038/ng.3168 (2015).

96. Piñero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, bav028, doi:10.1093/database/bav028 (2015).

97. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science* **306**, 1555-1558. doi: 10.1126/science.1099511 (2004).

98. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature Genetics* **40**, 181-188. doi: 10.1038/ng.2007.70 (2008).

99. Chen, Y., Jiang, T. & Jiang, R. Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics* **27**, i167-176, doi:10.1093/bioinformatics/btr213 (2011).

100. Huang, J., Liu, Y., Zhang, W., Yu, H. & Han, J. D. eResponseNet: a package prioritizing

candidate disease genes through cellular pathways. *Bioinformatics (Oxford, England)* **27**, 2319-2320, doi:10.1093/bioinformatics/btr380 (2011).

101. da Rocha, E. L., Ung, C. Y., McGehee, C. D., Correia, C. & Li, H. NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Research* **44**, e100, doi:10.1093/nar/gkw166 (2016).

102. Asano T. et al. Depth-First Search Using O(n) Bits. In: Ahn HK., Shin CS. (eds) Algorithms and Computation. ISAAC 2014. Lecture Notes in Computer Science, vol 8889. Springer, Cham. https://doi.org/10.1007/978-3-319-13075-0_44 (2014).

103. Yu, H., Lu, L., Jiao, B. & Liang, C. Systematic discovery of novel and valuable plant gene modules by large-scale RNA-seq samples. *Bioinformatics* **35**, 361-364, doi:10.1093/bioinformatics/bty642 (2019).

104. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

Figure Legends

Fig. 1: Schematic illustration of KDDANet computational pipeline in SDrTDi context. a) Mapping query drug (QD) and all its related diseases (RDs) into the weighted gene network (WGN) through known drug-target relationships and gene-disease associations. **b)** Constructing a unified flow network model for each QD and RDs (multiple KDDAs). **c)** Identifying the highest probability gene subnetwork mediating multiple KDDAs by minimum cost flow optimization. **d)** Identifying gene subnetwork mediating individual KDDA by depth first searching. **e)** Identifying gene interaction modules mediating individual KDDA by Markov clustering (MCL). **f)** Validating the prediction results of KDDANet by existing knowledgebases.

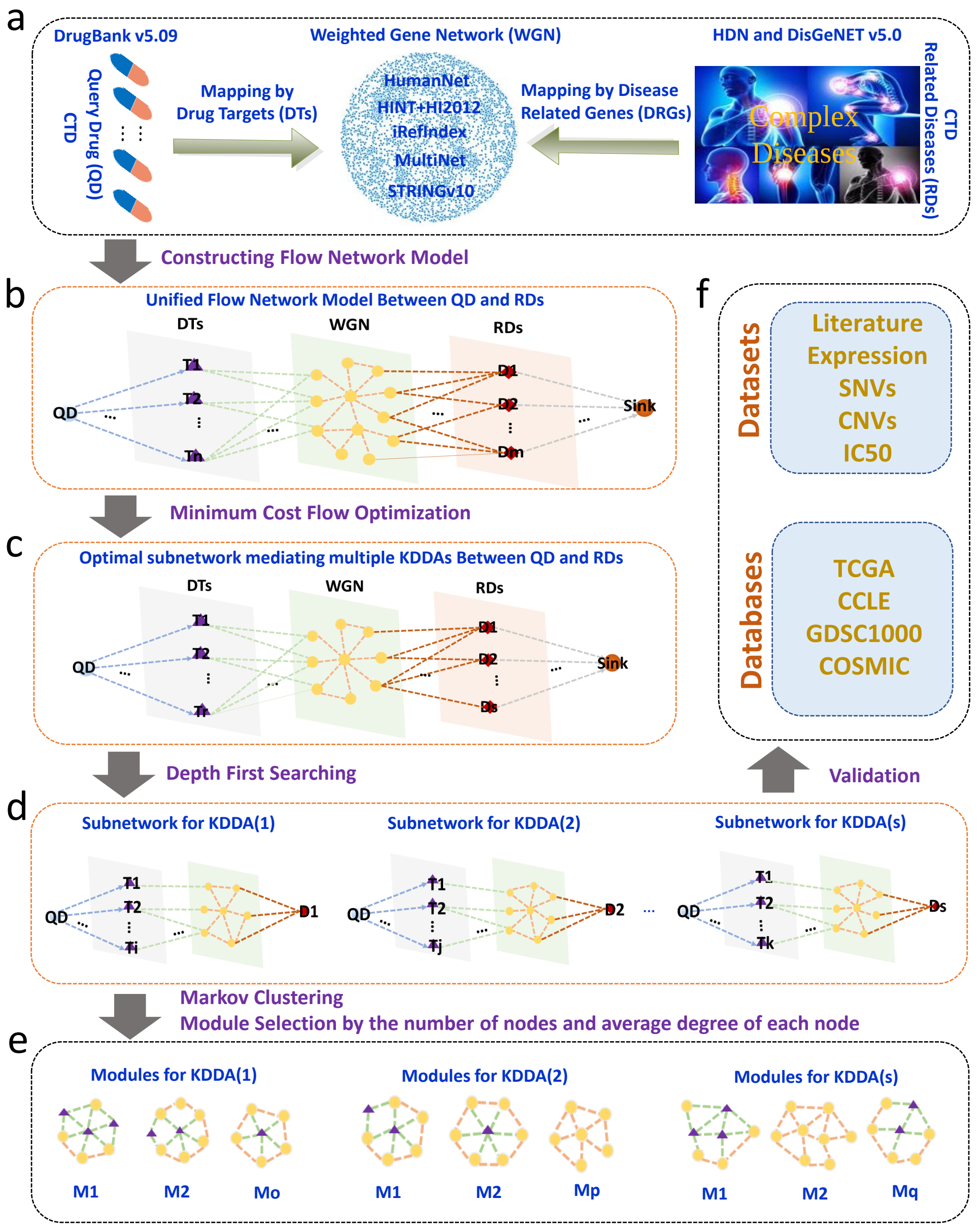
Fig. 2: Performance evaluation of KDDANet method. a) The density curve of AUROC and AUPRC of KDDANet with different γ setting and permutation test in SDrTDi context. **b)** The density curve of AUROC and AUPRC of KDDANet with different γ setting and permutation test in SDiTDr context. **c)** Comparison of AUROC and AUPRC of KDDANet with PPA, SNPLS, comCHIPER and DGPsubNet. **d)** AUROC and AUPRC of KDDANet with different types of gene interaction networks.

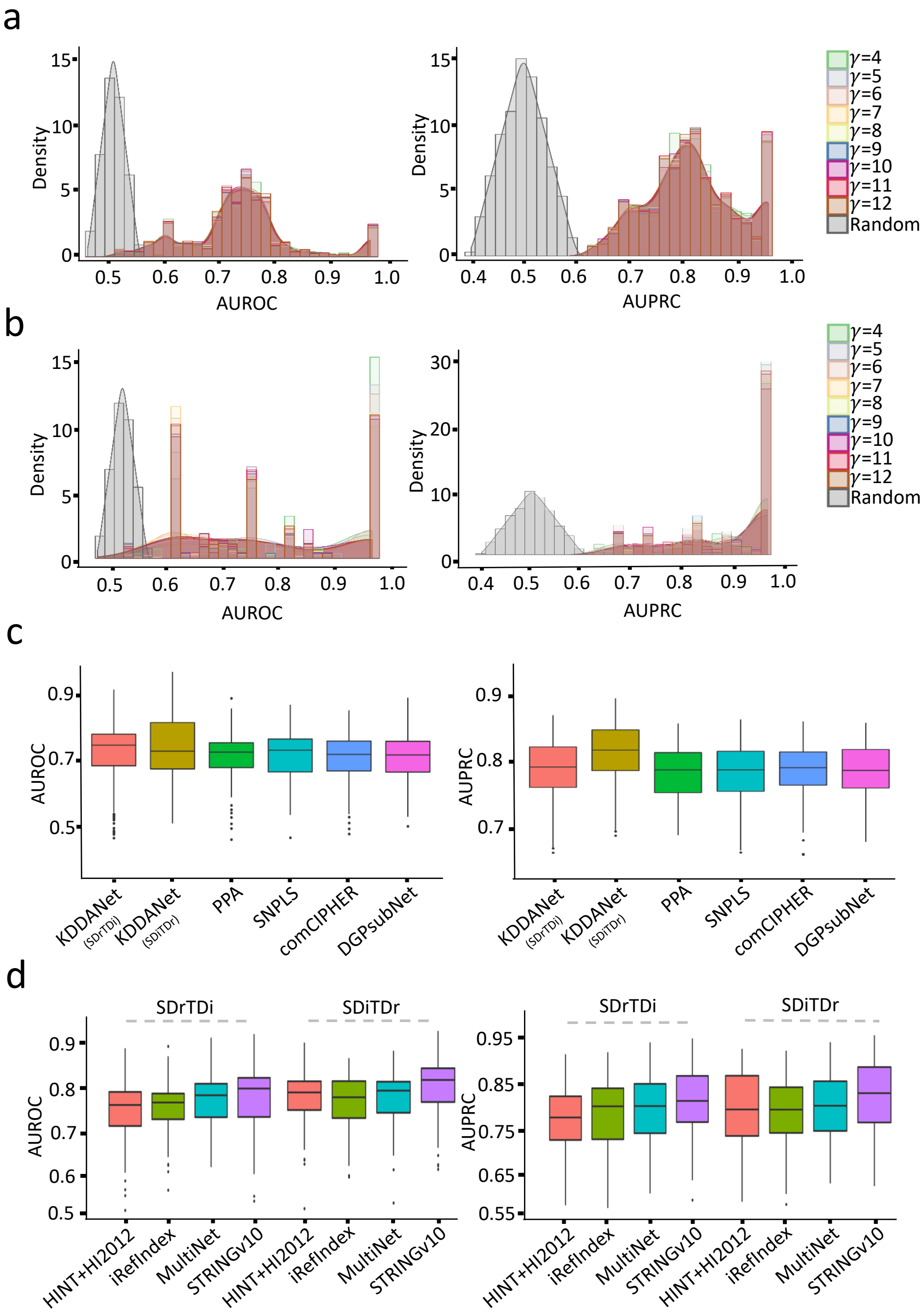
Fig. 3: Mechanistic relevance of KDDANet prediction results. a) KDDANet resulting subnetwork mediating phylloquinone (DB01022)-Alzheimer's disease (AD, 104300) association. **b)** Top 10 enriched KEGG terms of subnetwork genes mediating phylloquinone-AD association. **c)** Expression level of STAT3, GNAI1 and GC in normal individuals and AD patients, ** p-value < 0.01, calculated by Mann-Whitney U test. **d)** KDDANet resulting subnetwork for heparin (DB01109)-obesity (601665) association. **e)** Top 10 enriched KEGG terms of subnetwork genes mediating heparin-obesity association. **f)** ROC and PR curves of KDDANet gene subnetwork mediating heparin-obesity association. In the subnetworks, the size of a gene node was proportional to its network degree; The thickness of a network edge was proportional to its flow amount. Fragments Per Kilobase Of Exon Per Million Fragments Mapped, FPKM.

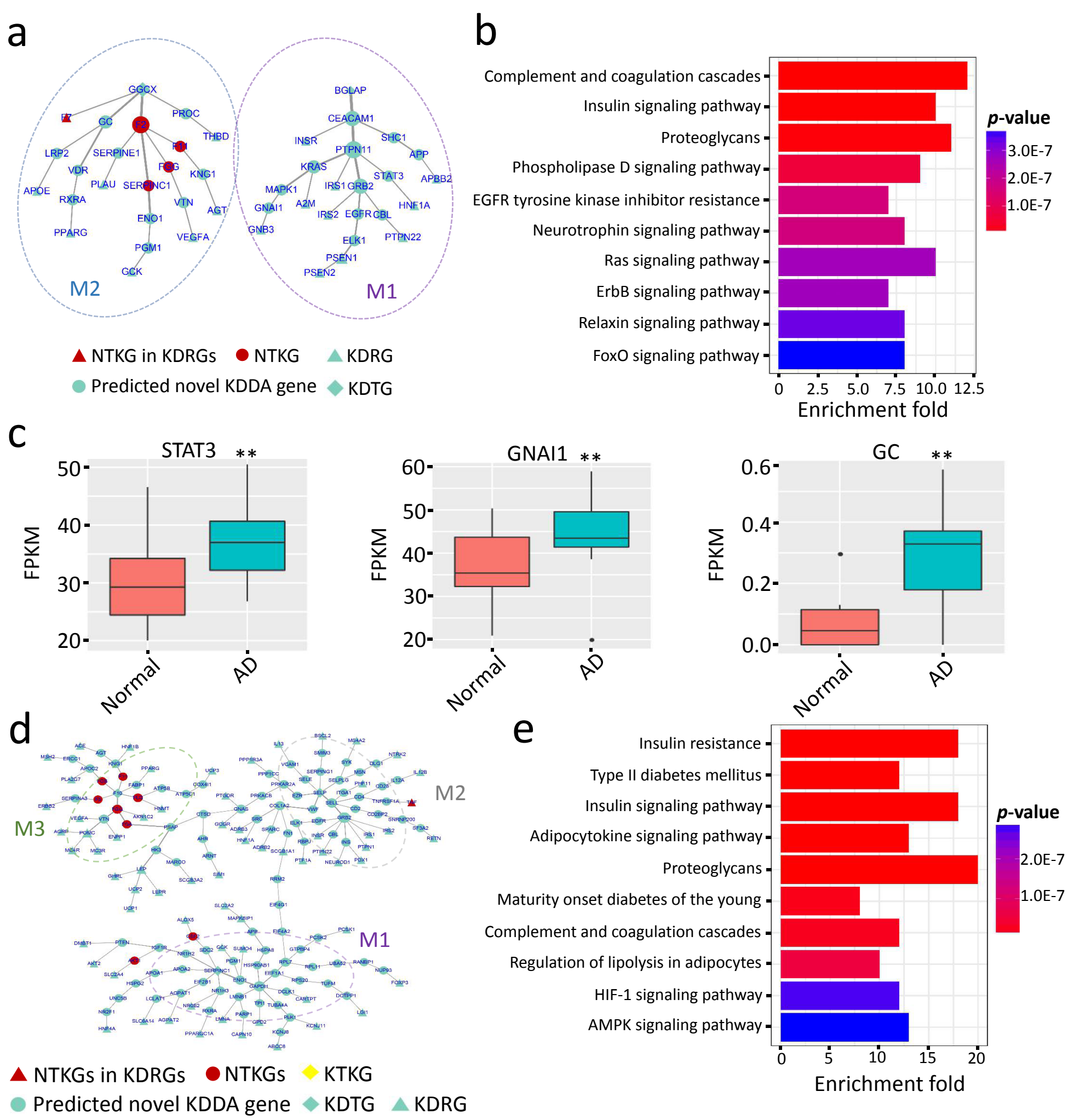
Fig. 4: KDDANet provided novel molecular insights on KDDAs related to cancer. a) KDDANet resulting subnetwork mediating sotalol (DB00489)-prostate cancer (176807) association. **b)** Top 10 enriched KEGG terms of subnetwork genes mediating sotalol-prostate cancer association. **c)** Gene expression-based survival analysis for CASP2, SPARC and GRB2 in patients with prostate cancer, obtained by GEPIA online analysis (http://gepia.cancer-pku.cn/about.html). **d)** KDDANet gene subnetwork mediating nebularine (DB04440)-lung cancer (211980) association. **e)** Top 10 enriched KEGG terms of subnetwork genes mediating nebularine-lung cancer association. **f)** Gene expression-based survival analysis for BYSL and BOP1 in patients with TCGA LUAD lung cancer (obtained by GEPIA online analysis (http://gepia.cancer-pku.cn/about.html) and their expression levels in lung cancer tumor samples and adjacent normal tissue samples, *** *p*-value < 0.001, calculated by Wilcox signed rank test. In the subnetworks, the size of a gene node was proportional to its network degree; The thickness of a network edge was proportional to its flow amount.

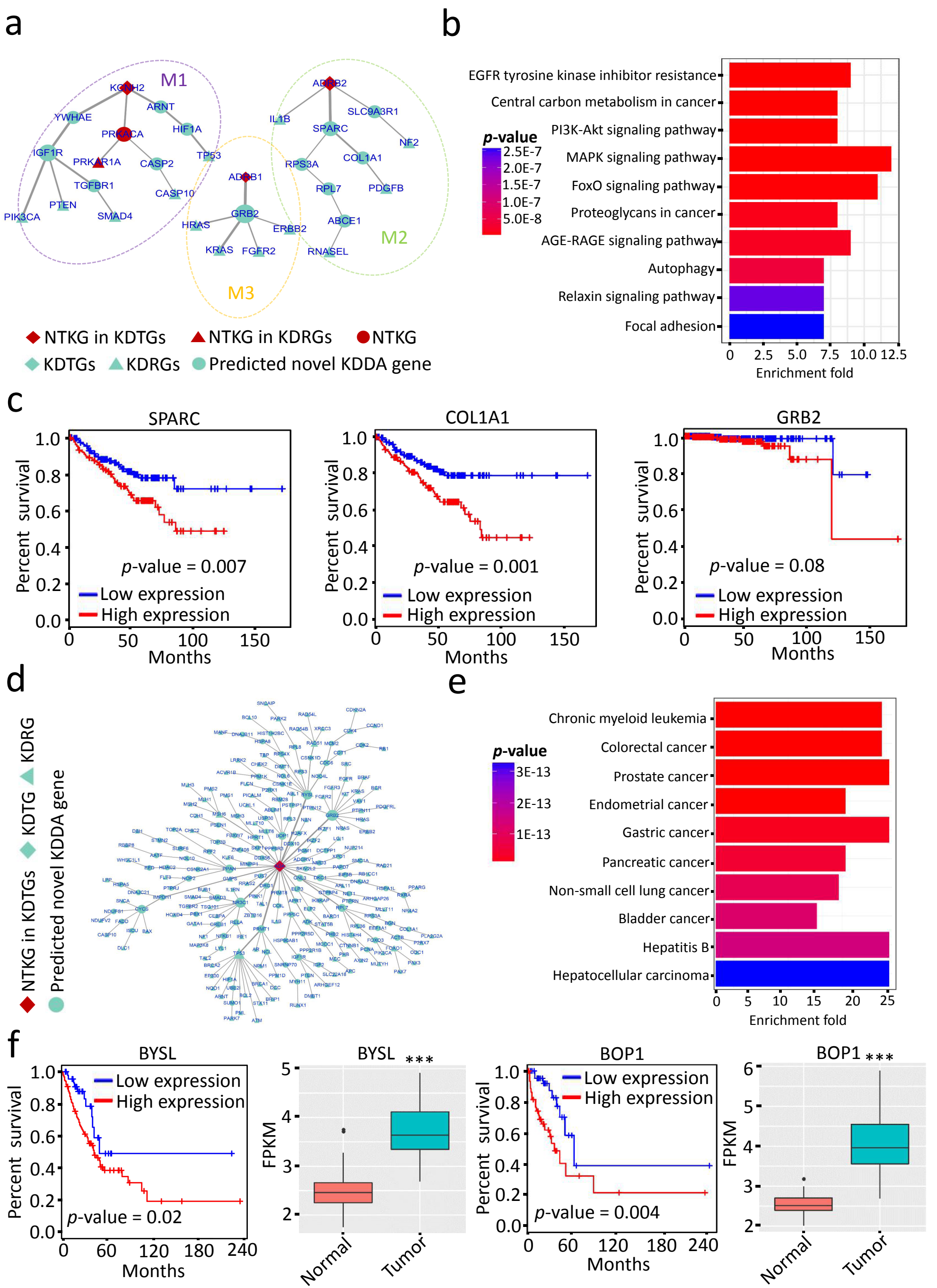
Fig. 5: KDDANet uncovered the shared genes mediating multiple KDDAs. a) Boxplots demonstrating the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDiODr and MDrODi in SDrTDi context, *** *p*-value < 0.001, calculated by Wilcox signed rank test. **b)** Boxplots demonstrating the distributions of weight values of KDDANet meta-subnetwork and random meta-subnetwork for MDiODr and MDrODi in SDiTDr context, *** *p*-value < 0.001, calculated by Wilcox signed rank test. **c)** Shared meta-subnetwork mediating profenamine (DB00392)-neurological disease associations. **d)** Top 10 enriched KEGG terms of shared meta-subnetwork genes mediating profenamine-neurological diseases associations. **e)** Shared meta-subnetwork mediating the associations between GRACILE syndrome (603358) and 13 drugs. **f)** Top 10 enriched KEGG terms of shared meta-subnetwork genes mediating the associations between GRACILE syndrome and multiple drugs; In the meta-subnetworks, the size of a gene node was

proportional to its network degree; the thickness of a network edge was proportional to its weight value.


a
DrugBank v5.09
Query Drug (QD)
CTD
Mapping by
Drug Targets (DTs)
Weighted Gene Network (WGN)
HumanNet
HINT+HI2012
iRefIndex
MultiNet
STRINGv10
Mapping by Disease
Related Genes (DRGs)
HDN and DisGeNET v5.0
Complex
Diseases
CTD
Related Diseases (RDs)
Constructing Flow Network Model
b
Unified Flow Network Model Between QD and RDs
DTs
WGN
RDs
QD
T1
T2
Tn
D1
D2
Dm
Sink
Minimum Cost Flow Optimization
c
Optimal subnetwork mediating multiple KDDAs Between QD and RDs
DTs
WGN
RDs
QD
T1
T2
Tr
D1
D2
Ds
Sink
Depth First Searching
Validation
f
Datasets
Literature
Expression
SNVs
CNVs
IC50
Databases
TCGA
CCLE
GDSC1000
COSMIC
d
Subnetwork for KDDA(1)
Subnetwork for KDDA(2)
Subnetwork for KDDA(s)
QD
T1
T2
Ti
D1
Tj
D2
Tk
Ds
Markov Clustering
Module Selection by the number of nodes and average degree of each node
e
Modules for KDDA(1)
Modules for KDDA(2)
Modules for KDDA(s)
M1
M2
Mo
M1
M2
Mp
M1
M2
Mq


a
Density
AUROC
AUPRC
γ=4
γ=5
γ=6
γ=7
γ=8
γ=9
γ=10
γ=11
γ=12
Random
b
Density
AUROC
AUPRC
γ=4
γ=5
γ=6
γ=7
γ=8
γ=9
γ=10
γ=11
γ=12
Random
c
AUROC
AUPRC
KDDANet (SDrTDi)
KDDANet (SDiTDr)
PPA
SNPLS
comCIPHER
DGPsubNet
d
SDrTDi
SDiTDr
AUROC
AUPRC
HINT+HI2012
iRefIndex
MultiNet
STRINGv10


a
GGCX
F7
GC
F2
PROC
THBD
LRP2
SERPINE1
F11
VDR
FGG
KNG1
APOE
PLAU
SERPINC1
RXRA
VTN
AGT
ENO1
PPARG
VEGFA
PGM1
GCK
M2
BGLAP
CEACAM1
INSR
SHC1
PTPN11
APP
KRAS
STAT3
APBB2
MAPK1
IRS1
GRB2
GNAI1
A2M
HNF1A
IRS2
EGFR
CBL
GNB3
ELK1
PTPN22
PSEN1
PSEN2
M1
NTKG in KDRGs
NTKG
KDRG
Predicted novel KDDA gene
KDTG
b
Complement and coagulation cascades
Insulin signaling pathway
Proteoglycans
Phospholipase D signaling pathway
EGFR tyrosine kinase inhibitor resistance
Neurotrophin signaling pathway
Ras signaling pathway
ErbB signaling pathway
Relaxin signaling pathway
FoxO signaling pathway
0
2.5
5.0
7.5
10.0
12.5
Enrichment fold
p-value
3.0E-7
2.0E-7
1.0E-7
c
STAT3
**
GNAI1
**
GC
**
FPKM
Normal
AD
d
M3
M2
M1
NTKGs in KDRGs
NTKGs
KTKG
Predicted novel KDDA gene
KDTG
KDRG
e
Insulin resistance
Type II diabetes mellitus
Insulin signaling pathway
Adipocytokine signaling pathway
Proteoglycans
Maturity onset diabetes of the young
Complement and coagulation cascades
Regulation of lipolysis in adipocytes
HIF-1 signaling pathway
AMPK signaling pathway
0
5
10
15
20
Enrichment fold
p-value
2.0E-7
1.0E-7
f
AUC=0.854
Sensitivity
1-Specificity
Threshold
AUC=0.8491
Precision
Recall
Threshold


a
M1
M2
M3
KCNH2
ARNT
YWHAE
HIF1A
PRKACA
IGF1R
PRKAR1A
CASP2
TP53
TGFBR1
CASP10
PTEN
PIK3CA
SMAD4
ADRB2
IL1B
SLC9A3R1
SPARC
NF2
RPS3A
COL1A1
RPL7
PDGFB
ADRB1
GRB2
HRAS
ERBB2
ABCE1
KRAS
FGFR2
RNASEL
NTKG in KDTGs
NTKG in KDRGs
NTKG
KDTGs
KDRGs
Predicted novel KDDA gene
b
EGFR tyrosine kinase inhibitor resistance
Central carbon metabolism in cancer
PI3K-Akt signaling pathway
MAPK signaling pathway
FoxO signaling pathway
Proteoglycans in cancer
AGE-RAGE signaling pathway
Autophagy
Relaxin signaling pathway
Focal adhesion
p-value
2.5E-7
2.0E-7
1.5E-7
1.0E-7
5.0E-8
0
2.5
5.0
7.5
10.0
12.5
Enrichment fold
c
SPARC
COL1A1
GRB2
Percent survival
1.0
0.8
0.6
0.4
0.2
0.0
p-value = 0.007
p-value = 0.001
p-value = 0.08
Low expression
High expression
0
50
100
150
Months
d
NTKG in KDTGs
KDTG
KDRG
Predicted novel KDDA gene
e
Chronic myeloid leukemia
Colorectal cancer
Prostate cancer
Endometrial cancer
Gastric cancer
Pancreatic cancer
Non-small cell lung cancer
Bladder cancer
Hepatitis B
Hepatocellular carcinoma
p-value
3E-13
2E-13
1E-13
0
5
10
15
20
25
Enrichment fold
f
BYSL
BOP1
Percent survival
Low expression
High expression
p-value = 0.02
p-value = 0.004
0
60
120
180
240
Months
FPKM
***
Normal
Tumor

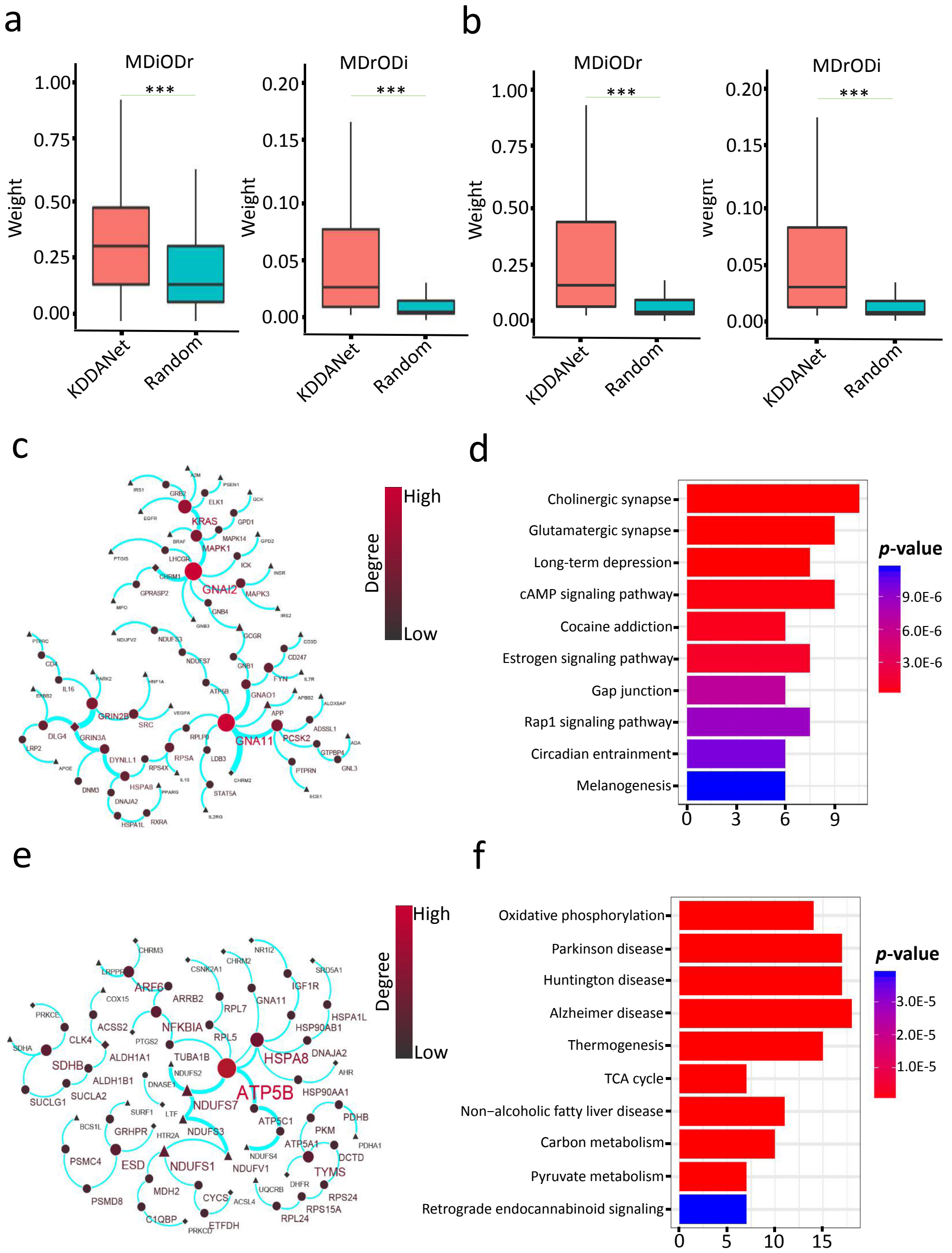

a
MDiODr
MDrODi
Weight
***
KDDANet
Random
b
c
KRAS
MAPK1
GNAI2
GNA11
GRIN2B
ATP5B
High
Low
Degree
d
Cholinergic synapse
Glutamatergic synapse
Long-term depression
cAMP signaling pathway
Cocaine addiction
Estrogen signaling pathway
Gap junction
Rap1 signaling pathway
Circadian entrainment
Melanogenesis
p-value
9.0E-6
6.0E-6
3.0E-6
e
ARF6
NFKBIA
HSPA8
ATP5B
NDUFS7
NDUFS1
ESD
f
Oxidative phosphorylation
Parkinson disease
Huntington disease
Alzheimer disease
Thermogenesis
TCA cycle
Non−alcoholic fatty liver disease
Carbon metabolism
Pyruvate metabolism
Retrograde endocannabinoid signaling
3.0E-5
2.0E-5
1.0E-5

Table 1. Average AUC values of different computational tools for identifying hidden genes mediating KDDA.

Method	KDDANet(SDrTDi)	KDDANet(SDiTDr)	SNPLS	PPA	comCHIPER	DGPsubNet
Average AUROC	0.733	0.715	0.712	0.713	0.703	0.706
Average AUPRC	0.793	0.825	0.775	0.776	0.778	0.777