Briefings in Bioinformatics, 2022, **23(6)**, 1–13

https://doi.org/10.1093/bib/bbac469

Advance access publication date 8 November 2022

Problem Solving Protocol

Deep learning identifies explainable reasoning paths of mechanism of action for drug repurposing from multilayer biological network

Jiannan Yang, Zhen Li, William Ka Kei Wu, Shi Yu, Zhongzhi Xu, Qian Chu and Qingpeng Zhang

Corresponding authors: Qingpeng Zhang, City University of Hong Kong, Hong Kong SAR, China, City University of Hong Kong Shenzhen Research Institute, Shenzhen, China, E-mail: qingpeng.zhang@cityu.edu.hk; Qian Chu, Department of Thoracic Oncology, Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China, E-mail: qianchu@tjh.tjmu.edu.cn

Abstract

The discovery and repurposing of drugs require a deep understanding of the mechanism of drug action (MODA). Existing computational methods mainly model MODA with the protein–protein interaction (PPI) network. However, the molecular interactions of drugs in the human body are far beyond PPIs. Additionally, the lack of interpretability of these models hinders their practicability. We propose an interpretable deep learning-based path-reasoning framework (iDPath) for drug discovery and repurposing by capturing MODA on by far the most comprehensive multilayer biological network consisting of the complex high-dimensional molecular interactions between genes, proteins and chemicals. Experiments show that iDPath outperforms state-of-the-art machine learning methods on a general drug repurposing task. Further investigations demonstrate that iDPath can identify explicit critical paths that are consistent with clinical evidence. To demonstrate the practical value of iDPath, we apply it to the identification of potential drugs for treating prostate cancer and hypertension. Results show that iDPath can discover new FDA-approved drugs. This research provides a novel interpretable artificial intelligence perspective on drug discovery.

Keywords: mechanism of drug action, interpretable deep learning, drug repurposing

Introduction

Artificial intelligence has recently shown the huge potential to subvert the typical drug discovery process [1]. Scientists are using deep learning technologies to discover candidate drugs for the treatment of COVID-19 [2–4], Alzheimer's disease [5], cancers [6] and so on. Among various applications, drug repurposing, the identification of the new use of approved or investigational drugs that are outside of the original medical indication, can shorten the time of drug development while ensuring safety and thus attracts attention from drug discovery communities and industries [7, 8]. Existing computational approaches mainly study this problem from biological and clinical perspectives [9]. A common approach in the clinical view is using electronic health records to discover the efficacy of drugs on a specific population [10] or emulating clinical trials on real-world patient data [11]. In the biological computational approaches, molecular docking [12], genetic association [13] and so on [14] are the common techniques to identify drug repurposing.

With the development of high-throughput omics technologies, the detailed characterization of the molecular interactions of drugs in the human body became possible [15]. The protein–protein interaction (PPI) network serves as a 'skeleton' for the body's signaling circuitry [16] and shows tremendous power in guiding drug discovery [17–20]. A series of studies explored the potential of mining the network properties of drugs in the PPI network in synergistic drug combination identification [17] and drug repurposing [18]. A recent study introduced advanced graph-based deep learning approaches to the identification of anti-cancer drug combinations by learning the graphic representations of the PPI network [21]. However, the molecular network in the human body is not limited to the PPI network, the gene regulatory mechanisms [22], the binding work of the proteins and chemicals [23] and the interactions of the chemicals [24, 25] also play a role in the mechanism of drug action (MODA). These processes rely on the drug's interactions with various proteins and chemicals in the human body [26, 27]. Usually, the MODA is described by

Jiannan Yang is a data scientist and a PhD candidate at City University of Hong Kong. His research interests are interpretable deep learning, drug design and disease prediction.

Zhen Li is a radiologist and a professor at Tongji Hospital, Huazhong University of Science and Technology. He has extensive experience in abdominal imaging and interventional radiology research.

William Ka Kei Wu is an associate professor at Chinese University of Hong Kong. He has extensive experience in evolutionary genomics, computational modeling of gene regulation and analysis of high-throughput genomic datasets.

Shi Yu is a biologist at the University of Southern California. He has extensive experience in plant circadian clock research.

Zhongzhi Xu is a data scientist with extensive experience in medical data analytics.

Qian Chu is an oncologist and a professor at Tongji Hospital, Huazhong University of Science and Technology. She has extensive experience in breast cancer research.

Qingpeng Zhang is an associate professor at City University of Hong Kong. He has extensive experience in healthcare data analytics, medical informatics, AI in drug discovery and network science.

Received: May 12, 2022. **Revised:** September 7, 2022. **Accepted:** September 29, 2022

biological pathways, a series of biochemical and molecular steps to achieve a specific function or to produce a certain product [28]. Such biological pathways can be naturally denoted as a series of paths in the biological network. Furthermore, instead of targeting specific proteins, some drugs need to take further chemical reactions to be effective [29]. For example, cytarabine, an important drug in the treatment of acute myeloid leukemia [30], must be phosphorylated intracellularly to a nucleotide (cytarabine 5′-triphosphate, Ara-CTP) before it can exert its cytotoxic effect [31].

Previous machine learning approaches [32–37] have introduced multilayer information to drug repurposing. For example, Napolitano *et al.* [32] integrated the drug's chemical information, PPI network and correlation of gene expression patterns after treatment together. By integrating multiple layers of information, these studies enhanced the drug repurposing prediction performance [32–34, 36, 37] and investigated the robustness of the system [38]. However, such multilayer information has not been fully utilized to characterize the MODAs. Furthermore, nearly all of these machine learning models are *black-box*. The lack of model interpretability hinders machine learning's potential in practical drug discovery tasks. The need for explainable machine learning models led to the development of a series of novel neural network architectures, such as attribution methods [39, 40] and knowledge-graph-based models [41, 42]. These models have been further applied in healthcare [43–46], such as using biological-informed neural networks to identify anti-cancer drug combinations [45] and predict disease risk based on comorbidity network [46]. However, whether these explainable modules can accelerate drug discovery and further enhance the knowledge of the MODA is unknown.

To fill the aforementioned research gaps, based on our previous research on interpretable machine learning [4, 21, 46, 47], we propose the interpretable Deep learning-based Path-reasoning framework for drug repurposing (iDPath), which captures the MODA by identifying the critical paths from drugs to diseases in the human body. To accurately characterize the MODA, we build a comprehensive multilayer biological network instead of using the PPI network alone. The multilayer biological network is the integration of a gene regulatory layer, a PPI layer, a protein–chemical interaction (PCI) layer and a chemical–chemical interaction (CCI) layer, integrated with the drug and diseases-related information. Starting from this multilayer biological network, iDPath utilizes a graph convolutional network (GCN) module to capture the global connectivity information of the human molecular network and a long–short-term memory (LSTM) neural network module to capture the detailed mechanisms of drug action based on the shortest paths between drugs and diseases. Furthermore, iDPath introduces two attention modules, namely the path attention and the node attention, to enhance model interpretability. Experiments with drug screen data demonstrate the superior performance of iDPath in a general drug repurposing task featuring 1993 drugs and 2794 diseases. Further investigations demonstrate that iDPath can identify explicit critical paths that are consistent with clinical evidence. To demonstrate the practical value of iDPath, we apply it to identify potential drugs for the treatment of prostate cancer and hypertension. Results show that iDPath can successfully discover new FDA-approved drugs. These results indicate that iDPath can facilitate drug discovery and repurposing and has the potential to address other computational chemistry and biology tasks involving the understanding of the molecular interactions in the human body.

Methods

In this section, we describe the datasets and the proposed iDPath model, as well as baseline models for drug repurposing, including DeepWalk, GCN, LSTM networks and knowledge-aware path recurrent network (KPRN).

Data

Multilayer biological network

Gene regulatory network (GRN) layer

The GRN is adopted from RegNetwork [48], which collects the experimentally validated and the predicted regulations based on the transcription factor (TF) binding sites. The edges in RegNetwork start from TF and microRNA (miRNA) and target the regulated genes. In total, RegNetwork provides us with 369 277 gene regulations between 1456 TFs, 1904 miRNAs and 19 719 genes.

PPI layer

The PPI network consists of information from two sources. The first dataset, STRING dataset [49], is the most comprehensive database of known and predicted PPIs till now, with more than 1380 million PPIs among over 9 million proteins. We only keep the PPIs in the human body (*Homo sapiens*) and at high confidence or better (confidence level > 0.7). Another PPI dataset is the human interactome built by Cheng *et al.* [18]. This dataset is harnessed from multiple databases with experimental evidence. After preprocessing, our PPI network contains 614 970 interactions connected by 13 758 proteins.

PCI layer

We obtain a PCI network by curating from the STITCH database [50], which is the most comprehensive database of known and predicted interactions between chemicals and proteins till now. We select PCIs in the human body (*H. sapiens*) at high confidence or better (confidence level > 0.7). The processed PCI network consists of 203 551 interactions among 9393 proteins and 73 199 chemicals.

CCI layer

CCI network is curated from STITCH [50] database and further processed by selecting CCIs in the human body (*H. sapiens*) at high confidence or better (confidence level > 0.7). The processed CCI network has 396 284 interactions among 107 055 chemicals.

Constructing the multilayer biological network

We construct a multilayer biological network by mapping all the entities in GRN, PPI, PCI and CCI to the same nomenclature. The proteins are named by their encoded genes and the miRNAs are mapped to their corresponding genes by BioMart [51]. All the genes are encoded to their Entrez IDs [52]. All the chemicals are denoted by their PubChem CIDs (Compound ID number) [53].

Therapeutic drug–disease pairs

For the drug repurposing task, we collect therapeutic drug–disease pairs from the Therapeutic Target Database (TTD) [54], which provides the known and explored therapeutic protein and nucleic acid targets, the targeted disease and the pathway information of tens of thousands of drugs. We only keep the FDA-approved drugs in TTD and map them to PubChem CID to be consistent with the chemicals in the multilayer biological network. The diseases in TTD are in the ICD-11 coding system and are mapped to their corresponding ICD-10 codes. The cleaned dataset of therapeutic drug–disease pairs includes 1993 drugs and 2794 diseases and constitutes 19 500 pairs.

Drug–Protein associations and drug–chemical associations

We collect drug–protein associations from four datasets: the PCI network from STITCH [50], the drug–protein associations built by Cheng *et al*. [18], the TTD [54] and DrugBank [55]. The drug–chemical associations are extracted from STITCH by selecting the compounds that are drugs. DrugBank is a commonly used database containing comprehensive molecular information about drugs, their mechanisms, interactions and targets. The aggregated dataset contains 85 305 drug–protein associations between 20 405 drugs and 7796 proteins, 83 271 drug–chemical associations between 4630 drugs and 12 042 chemicals. All the drugs and chemicals are denoted by their PubChem CIDs, and all the proteins are represented by their encoded genes using Entrez ID.

Disease–Gene associations and disease–miRNA associations

The disease–gene associations include genes and variants associated with human diseases, curated from DisGeNET [56] by selecting expert-curated repositories. The miRNAs associated with human diseases come from the Human microRNA Disease Database [57], which is a database about curated experiment-supported evidence for human miRNA and disease associations. All the genes, variants and miRNAs are mapped to Entrez IDs, and diseases are mapped to ICD-10 codes. After processing, we have 230 837 associations among 7559 genes, 6830 variants and 705 miRNAs with 5602 diseases.

Overall architecture of iDPath

The iDPath framework for drug repurposing is presented in Figure 1. The MODA-related biological paths are identified by the shortest paths between the targets of drugs and diseases (Figure 1B) in the multilayer biological network (Figure 1A). To learn the global connectivity information of the multilayer biological network, iDPath first utilizes a three-layer GCN to learn the embeddings of associated nodes. Then, to capture the detailed MODA patterns, the embeddings of the nodes along the shortest paths between a drug and a disease are fed into an LSTM module to model their sequential dependencies. iDPath also introduces two attention modules to aggregate the embeddings of nodes and paths—path attention and node attention. These two attention modules are capable of discriminating the contribution of different nodes to one MODA-related biological path as well as the contribution of different paths to the final prediction.

GCN to capture the global connectivity information of the multilayer biological network

With the uniform nomenclature of the nodes, we aggregate the drug-related information (drug–protein associations, drug–chemical associations), the disease-related information (disease–gene associations and disease–miRNA associations) and the multilayer biological network to the network $G = (V, E)$, where V and E are the node set and edge set, respectively, iDPath introduces a three-layer GCN, following a spatial-based GCN architecture [58], to encode the global topological information of the multilayer biological network. Suppose there are N nodes in total, and the initial embeddings of these nodes are $E \in \mathbb{R}^{N\times d}$ (d is the dimension of the embedding), and the adjacency matrix of network G is A, the computation formula of layer l of GCN is shown below:

$$\mathrm{E}^{l+1} = \sigma_G \left(D^{-1} \left(A + I \right) D^{-1} \mathrm{E}^{l} \mathrm{W}^{l} \right), \tag{1}$$

where D is the diagonal node degree matrix, I is the identity matrix, σ_G is the activation function (relu) and W^l is the learning weights at layer l.

MODA-related biological paths

The MODA is dependent on the interactions of drugs with molecules in the human body, which can be represented as a series of paths in the multilayer biological network [28]. To accurately model the effects of drugs, we need to identify informative paths to represent the MODA in an efficient way. We prioritize the shortest paths because the shorter distance between drug and disease is found to be associated with higher chance of the therapeutic effect [17, 18, 59]. We adopted GPU-accelerated *sssp* algorithm implemented by NVidia's cuGraph package to identify the shortest paths [60]. For a drug and a disease, the shortest paths between them are connected by their associated nodes in the multilayer biological network. Given a drug and a disease, the shortest paths between this pair form a set $PATH = \{path_1, path_2, \ldots, path_L\}$, where $path_i = \{node_{drug} \rightarrow node_{m_1} \rightarrow node_{m_2} \rightarrow \cdots \rightarrow node_{disease}\}$, $node_{m_1}$ and $node_{m_2}$ denote the middle nodes of one path, and L is the number of shortest paths. Since L differs among drug–disease pairs, we choose a fixed value for L by randomly sampling from the shortest paths set $PATH$.

LSTM layer

Given a drug–disease pair, the embeddings E_{GCN} generated by the GCN and the shortest path set $PATH$, we employ LSTM [61] to encode both long-term and short-term dependencies in a MODA-related biological path. Such sequential dependencies are crucial to the model intelligibility. Meanwhile, we introduce the type of nodes to strengthen the model's capability of identifying different nodes. Here, we consider four types: protein (gene), chemical, drug and disease. Their embeddings $E_{TYPE} \in \mathbb{R}^{4\times d}$ are randomly initialized. Therefore, given one node $node_j$ of one path $path_p$, the input to LSTM is the concatenation of its GCN embedding $e_j = E_{GCN}[node_j]$ and type embedding $e'_j = E_{TYPE}[node_j]$, that is:

$$\mathrm{x}_j = e_j \bigoplus e'_j, \tag{2}$$

where $\bigoplus$ denotes the concatenation operation along the row axis. The hidden state h_{j-1} generated by the previous node $node_{j-1}$ in the same path and x_j are used to learn the hidden state of the input of the next node $node_j$, which is defined as follows:

$$\begin{aligned}
\mathrm{i}_j &= \sigma \left(\mathrm{W}_{\mathrm{i}} \mathrm{x}_j + \mathrm{W}_h h_{j-1} + b_{\mathrm{i}} \right), \\
f_j &= \sigma \left(\mathrm{W}_f \mathrm{x}_j + \mathrm{W}_h h_{j-1} + b_f \right), \\
g_j &= \tanh \left(\mathrm{W}_g \mathrm{x}_j + \mathrm{W}_h h_{j-1} + b_g \right), \\
\mathrm{o}_j &= \sigma \left(\mathrm{W}_{\mathrm{o}} \mathrm{x}_j + \mathrm{W}_h h_{j-1} + b_{\mathrm{o}} \right), \\
\mathrm{c}_j &= f_j \bigodot \mathrm{c}_{j-1} + \mathrm{i}_j \bigodot g_j, \\
h_j &= \mathrm{o}_j \tanh \left(\mathrm{c}_j \right),
\end{aligned} \tag{3}$$

where i_j, f_j, g_j and o_j are the input, forget, cell and output gates, respectively; $\mathrm{c}_j \in \mathbb{R}^{d'}$ and $h_j \in \mathbb{R}^{d'}$ are the cell state and hidden state at path step j, and d' is the dimension of the output; W_{i}, W_f, W_f, W_{o} and W_h are learnable weights, and b_{i}, b_f, b_g and b_{o} are bias; σ denotes the sigmoid function and $\bigodot$ is the Hadamard product,

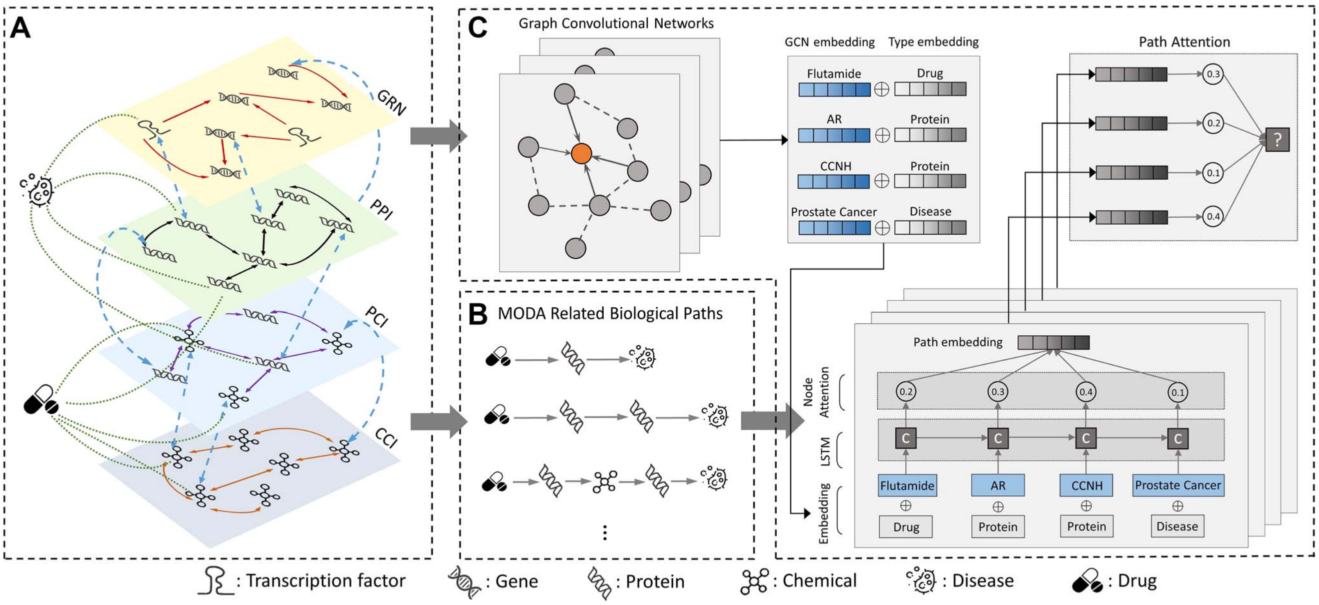

Figure 1. The framework of iDPath on drug repurposing tasks. (**A**) The multilayer biological network consists of four layers: GRN layer (one-way red arrows), PPI layer (two-way black arrows), PCI layer (two-way purple arrows) and CCI layer (two-way orange arrows). The blue two-way dashed arrows represent that the two corresponding nodes in different layers are identical. The nodes associated with the drugs and diseases are marked by green dashed lines. (**B**) The schematic representation of the MODA-related biological paths. The MODA-related biological paths are identified by the shortest paths between drug and disease generated in the multilayer biological network. Since the targets of drugs and diseases are proteins, all the shortest paths have the form of <*drug–protein–*...*–protein–disease*>. (**C**) The schematic representation of the algorithm: the multilayer biological network is fed into three-layer GCN to learn the embeddings of all nodes. The GCN embeddings of nodes along the shortest path between one drug and one disease are fed into an LSTM module to learn the sequential dependencies. Node attention and path attention modules are introduced to aggregate the embeddings of the nodes and paths. The final prediction is the probability that one drug is effective for one disease.

which is the element-wise multiplication of two matrixes. The hidden states h_j for each path step are aggregated to the attention modules for the representation of the whole path and final prediction. Since the shortest paths are not of equal length, we borrow the padding method as follows [62]. Suppose the maximum length of one path is set to l_{max}, for the paths that are shorter than l_{max}, we use a padding value *pad* (such as 0) to fill the path, and the following processing will ignore these padding positions not to affect the performance.

Node attention and path attention

Attention mechanism [63] is widely used in various deep learning tasks to enhance model intelligibility [21]. In this study, we introduced two attention modules to separately learn the importance of different nodes to one MODA-related biological path as well as the importance of different paths to the final prediction.

Node attention

For one shortest path $path_p$ between one drug–disease pair, the hidden states of each node generated by LSTM layers are $H_p \in \mathbb{R}^{l_{max} \times d'}$ where $H_p = \{h_1, h_2, \ldots, h_{pad}, h_{pad}\}$ and h_{pad} are the hidden states of the padding step. We first transfer all the hidden values at the padding positions to negative infinity for the following Softmax transformation. Then, we applied a linear layer with the Softmax activation to aggregate the embeddings to one numeric value denoting the importance (weight). That is

$$\Omega_p = H_p W_n, \tag{4}$$

where $W_n \in \mathbb{R}^{d' \times 1}$ is the learnable parameter, $\Omega_p \in \mathbb{R}^{l_{max} \times 1}$ denotes the weights of each node in path $path_p$, where $\Omega_p = \{\omega_1, \omega_2, \ldots, \omega_{l_{max}}\}$. For one node j in $path_p$, its weight is computed as follows:

$$\hat{\omega}_j = \frac{e^{\omega_j}}{\sum_{k=1}^{l_{max}} e^{\omega_k}}. \tag{5}$$

Then, we aggregate the hidden states of these nodes weighted by $\hat{\omega}_i$ to get the embedding of $path_p$:

$$e_{path_p} = \sum_{k=1}^{l_{max}} \hat{\omega}_k h_k. \tag{6}$$

Path attention

After the aggregation operation, for one drug–disease pair, the embeddings of their shortest paths are: $E_{PATH} = \{e_{path_1}, e_{path_2}, \ldots, e_{path_L}\}$. The path attention layer is similar to the node attention:

$$\begin{gathered} \Omega_{PATH} = e_{PATH} W_p, \\ \hat{\omega}_{path_p} = \frac{e^{\omega_{path_k}}}{\sum_{k=1}^{L} e^{\omega_{path_k}}}, \\ \hat{y} = \sigma_p \left(\sum_{k=1}^{L} e_{path_k} \hat{\omega}_{path_k} \right), \end{gathered} \tag{7}$$

where σ_p is the sigmoid function. The final prediction $\hat{y}$ is the aggregation of the weighted embedding of each path, indicating the probability that the drug is effective in treating the disease.

Objective function

In this study, we treat the training of iDPath as a binary classification task following common practice. That is, besides all the therapeutic drug–disease pairs (marked as 1), we introduce the negative sampling to get an equal number of non-therapeutic drug–disease pairs (marked as 0). The objective function used by iDPath is the binary-cross-entropy loss with l_2 regularization:

$$\mathbb{L} = \sum \left(-y \log \hat{y} - (1-y)\log(1-\hat{y})\right) + \lambda ||\Theta||_2^2, \tag{8}$$

where Θ is the set of parameters to be learned in iDPath, λ is the l_2 regularizer to prevent over-fitting, $\left|\left|\Theta\right|\right|_2^2$ is the square of the second norm of Θ.

Baselines

In this section, we describe a set of baseline models to compare with, including DeepWalk, GCN, LSTM and KPRN. We fed baseline models as much information as possible to have a fair comparison. The two path-based models (KPRN and LSTM) utilized the same input as iDPath. For GCN and DeepWalk, we used the drug targets and disease-related genes as input to train both models, but not the paths because these two models cannot handle sequential data naturally.

DeepWalk

DeepWalk [64] is a widely used graph embedding approach by modeling a stream of short random walks and has already been introduced to several drug-related tasks, such as drug–target identification [65].

GCN

GCN [58] is widely applied to many drug-related tasks, such as drug discovery using the drug's Smiles features [66] and anticancer drug combination identification [21]. As a component of iDPath, we apply GCN individually to test its performance on the drug repurposing task. Specifically, besides the basic graph convolutional layer, we introduce a fully connected layer to combine the embeddings of drug and disease for the final prediction.

LSTM

LSTM has been applied to drug discovery [67, 68]. We apply a vanilla LSTM network and use the last hidden states of each path as its representation. A two-layer fully connected layer following the LSTM layer is employed to generate the final prediction.

KPRN

KPRN [42] is an advanced path-based model for a reasonable recommendation based on a knowledge graph. We apply KPRN to the drug repurposing task by feeding it the same input as iDPath.

Performance evaluation and experiment setup

The training of iDPath is a binary classification task: given one drug–disease pair, we feed all the shortest paths between this drug and disease into iDPath. And iDPath will generate one value indicating the probability that this drug has a therapeutic effect on this disease. All models are trained on this binary classification task, and we utilize commonly used metrics to evaluate and fine-tune the models. The metrics used in the binary classification task include accuracy, recall, the area under the receiver operating characteristic curve (AUROC), and the area under the precision–recall curve (AUPRC). The drug repurposing task can be viewed as a recommendation task: for each disease, we go through all the available drugs in our dataset and use the model to calculate the probability that one drug is effective in treating the disease and then rank all the drugs based on the probability. We introduce two commonly used metrics in the recommendation system, *NDCG@K* and *Hit@K*. In addition, we also trained a *shuffled random model*, an iDPath variant trained on a randomly edge-shuffled multi-layer network as a baseline. Details of these metrics, the computing facilities and the experiment setup are listed in the supplementary information. The code and results of all drug–disease pairs are available on the GitHub page.

Results and discussions

iDPath consistently outperforms baselines in drug repurposing

In general, baseline models can be classified into two categories: graph-embedding-based models (GCN and DeepWalk) and path-based models (LSTM and KPRN). iDPath presents a modeling framework that combines graph-embedding-based and path-based approaches. We compare the performance of iDPath with baselines in the drug repurposing problem. As shown in Figure 2A and C, iDPath outperforms all the baselines with an AUPRC of 0.97. In detail, iDPath achieves a 91.51% true-negative rate (TN) and 91.23% true-positive rate (TP) in the test dataset, indicating that only <10.00% of drug–disease pairs have not been correctly classified (Figure 2B).

In addition, the poor performance of the shuffled random model (AUPRC 0.76) demonstrates the importance of learning on the multi-layer biological network with correct biological interactions. The utilization of the shortest paths can significantly improve the performance, as demonstrated by the superior performance of iDPath and path-based models over graph-embedding-based models (Figure 2A and C). These results indicate that the extracted MODA-related biological paths have pharmacological relevance.

Incorporating multiple biological network layers improves the prediction performance

We further investigate the performance of iDPath with different biological network layers. Existing studies mainly make predictions using the PPI network alone [17, 18]. Here, we evaluate the performance of iDPath with only one layer (PPI, GRN or PCI). Note that CCI cannot be directly linked to diseases, so we do not evaluate the model with only the CCI layer. As shown in Figure 2D, the full multilayer biological network can improve the performance of iDPath. Comparing individual networks, PPI performs the best, followed by GRN, and finally PCI. Note that the iDPath variants combining two or three layers cannot compete with the model with all four layers. Then, we investigate the proportion of nodes and interactions at each network layer in the identified MODA-related biological paths (Figure 3), to examine the impact of different network layers on iDPath performance. We find that nodes and interactions in the PPI layer are not the most prevalent in the identified paths. Instead, GRN nodes and PCI interactions are more common. Combining these results with the dominating role of the PPI network in prediction performance, we find that although the connectivity in the PPI network can capture the key relationships between drugs and diseases, it requires additional information at the GRN, PCI and CCI layers to further reveal the hidden biological paths related to MODA. By revealing these hidden paths, iDPath achieves higher prediction accuracy. The full GRN–PPI–PCI–CCI network had the best performance (Figure 2D and Supplementary Figure S5 in supplementary information) because the additional PCI and CCI layers provide a more comprehensive characterization of the signaling circuitry. However, adding only one layer of either PCI or CCI will introduce bias toward the corresponding biochemical processes.

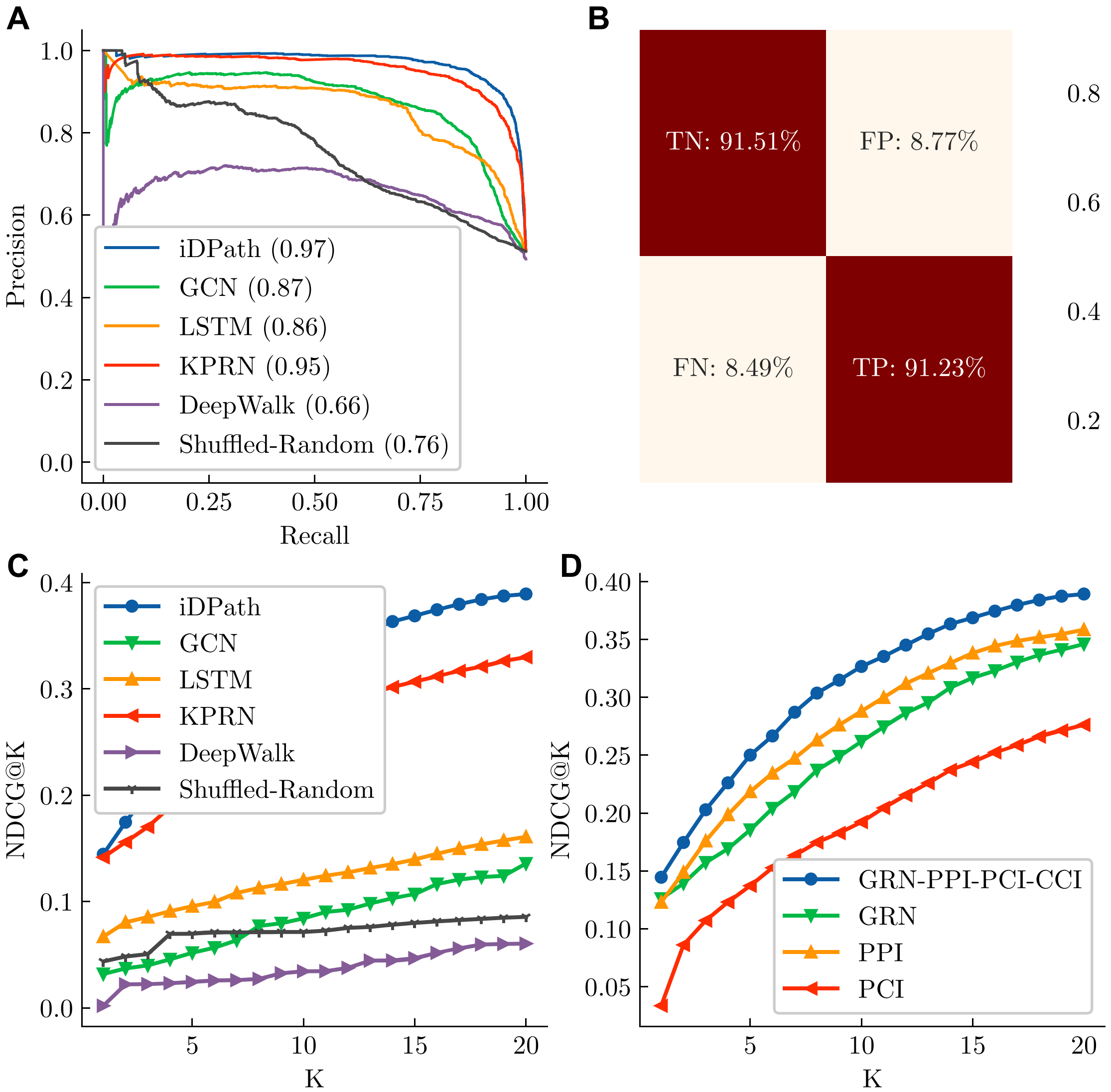

Figure 2. Performance of iDPath. (**A**) The precision–recall (PR) curves of all the models on the testing set, the values in the bracket denote the AUPRC. (**B**) The TN rate, false-negative rate, false-positive rate and TP rate of iDPath on the testing set. (**C**) The performance (NDCG@K) of all the models on the drug recommendation task on the testing set. These models are trained on the binary classification task and used to generate the repurposing probabilities of all the drugs on different diseases in the testing set. (**D**) The performance of iDPath with different biological network layers. Here GRN–PPI–PCI–CCI denotes the multilayer biological network generated by these four networks, GRN denotes using gene regulatory network alone, and the same goes for PPI and PCI. The K values in c and d denote the top K drugs used to compute for *NDCG*.

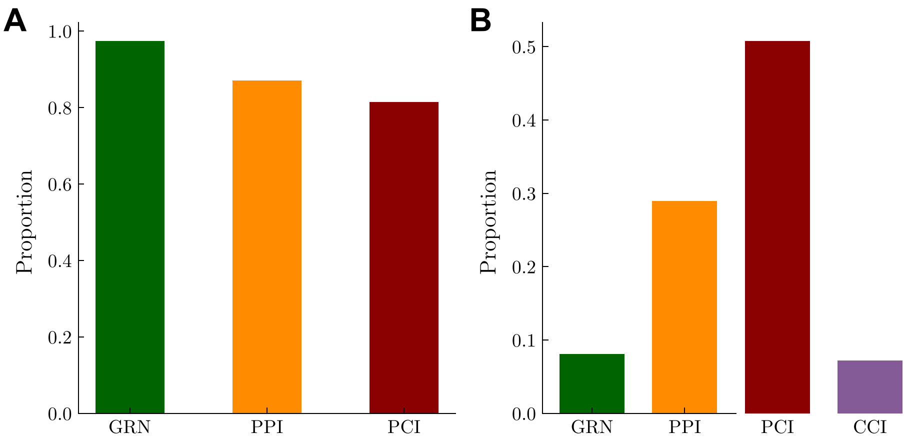

Figure 3. The proportion of nodes (**A**) and interactions (**B**) at each network layer in the identified MODA-related biological paths.

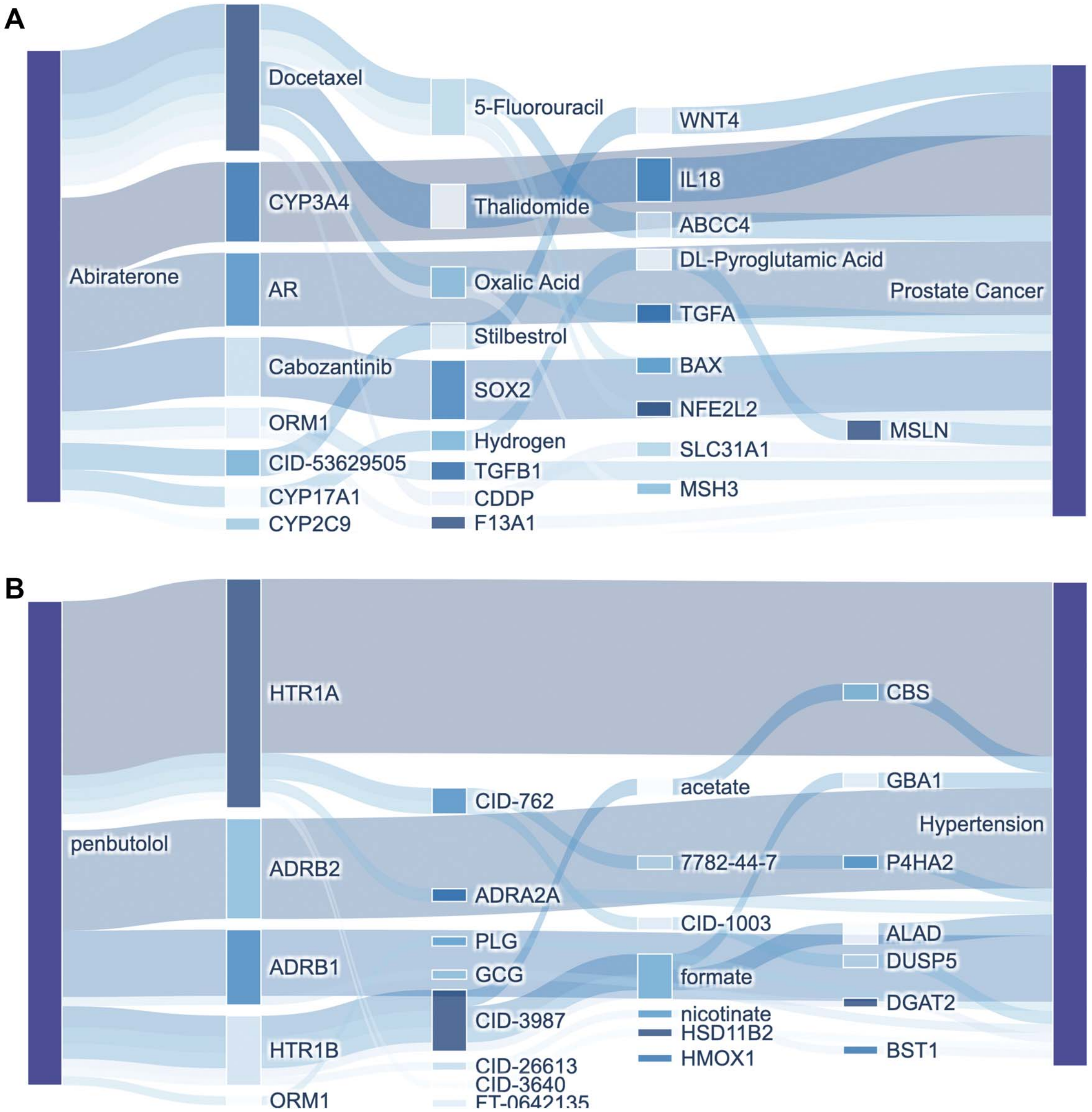

Figure 4. Interpretation of the MODA-related paths connecting abiraterone and prostate cancer and those connecting penbutolol and hypertension. (**A**) The Sankey diagram of the critical paths connecting abiraterone and prostate cancer identified by iDPath. (**B**) The Sankey diagram of the critical paths connecting penbutolol and hypertension identified by iDPath. The density of edge colors is determined by the path attention module. Edges with darker colors are more important. The density of node colors is determined by the node paths generated by the node attention module. Nodes with darker colors are more important.

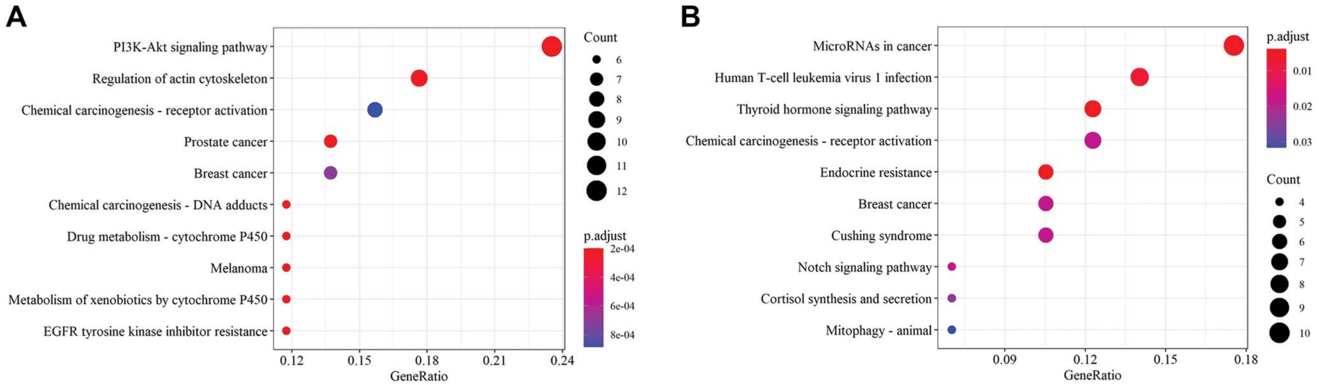

Figure 5. KEGG pathway enrichment analysis of the paths between abiraterone and prostate cancer. a and b are the dotplots of the KEGG pathway enrichment analysis for the proteins that existed in the top-50 paths and bottom-50 paths between abiraterone and prostate cancer, respectively.

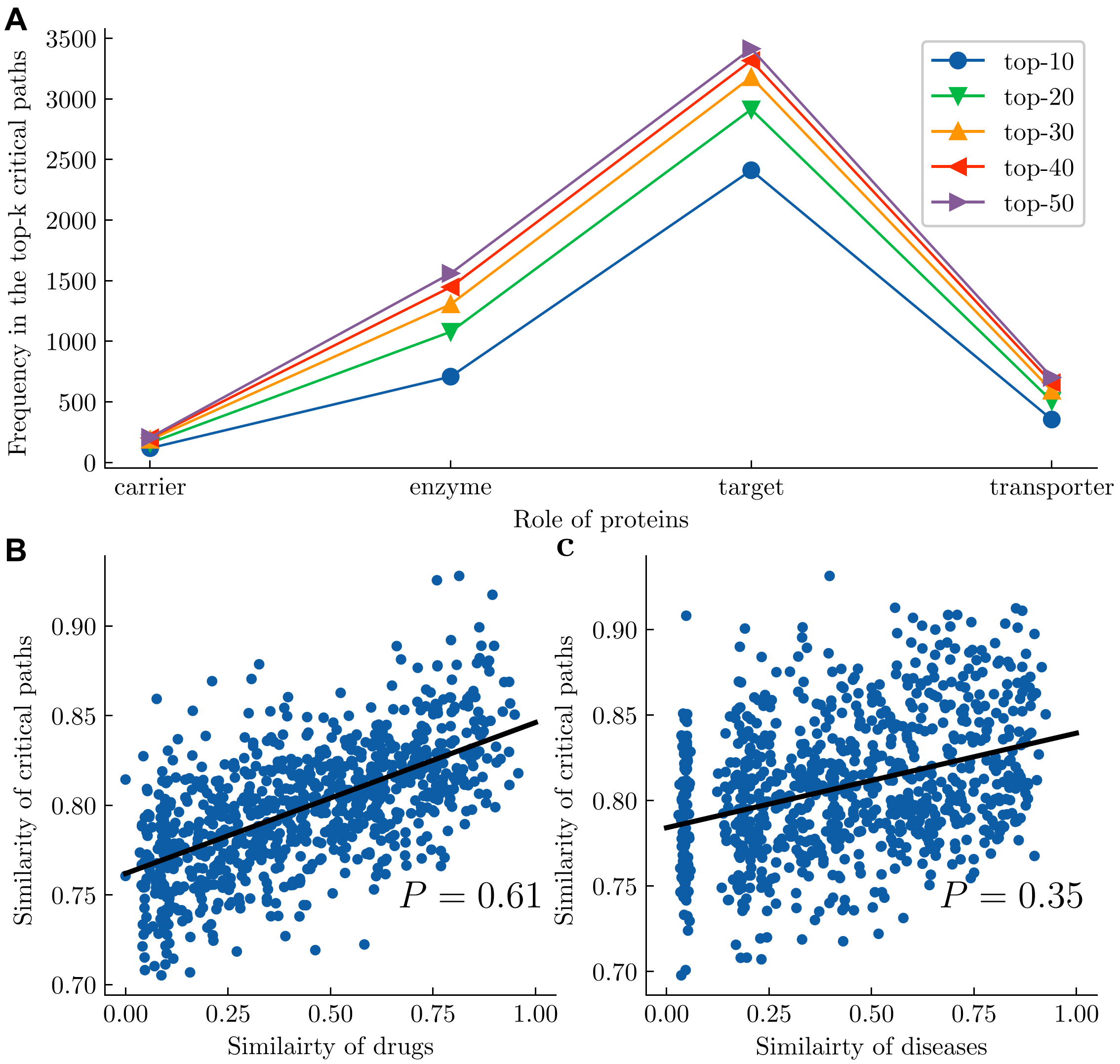

Figure 6. (**A**) The distribution of the role of drug-related proteins in top-k critical paths. (**B**) The relationship between the similarity of drugs and the similarity of critical paths. (**C**) The relationship between the similarity of diseases and the similarity of critical paths. *P* is the Pearson correlation coefficient. The *P*-values in **B** and **C** are both <0.0001.

iDPath identified the critical paths related to MODA

To investigate whether the identified critical paths are representative of MODA, we visualize the critical paths of correctly classified drugs for prostate cancer and hypertension. Figure 4A and B show one example of abiraterone (anti-prostate cancer drug) and penbutolol (anti-hypertension drug). Here we define the critical paths as the top 50 paths ranked by their weights identified by the path attention module. The top 15 paths are presented in Figure 4.

As shown in Figure 4A, among 256 shortest paths between abiraterone and prostate cancer, iDPath prefers the paths traversing through the gene targeted by both abiraterone and prostate cancer, such as abiraterone → CYP3A4 → prostate cancer and abiraterone → AR → prostate cancer. Previous studies have shown that abiraterone is a moderate inhibitor of CYP3A4 [69], and CYP3A4 is associated with oxidative deactivation of testosterone, which is the etiology of prostate cancer [70]. Androgen receptor (AR) is highly relevant to the growth and differentiation of prostate cancer [71], and abiraterone inhibits androgen biosynthesis to control the progression of prostate cancer [72]. Abiraterone is found to be an inhibitor of CYP17A1 [73], which has also been identified by iDPath. Specifically, the path abiraterone → CYP17A1 → Hydrogen → DL-Pyroglutamic Acid → MSLN → prostate cancer contributes to the prediction the most among all the CYP17A1-related paths, which is also consistent with previous biological studies [74]. In conclusion, the critical paths identified by iDPath represent the biological pathways, which represent the cascade of molecular interactions triggering the drug action. While they do not exclude other MODAs exerted by the drug, the identified critical paths suggest a greater probability.

As shown in Figure 4B, iDPath identified critical paths between penbutolol and hypertension, such as penbutolol → HTR1A → hypertension, which are consistent with clinical trial studies [75, 76]. Due to the length limit, we briefly introduce the results of hypertension in the main text and present the detailed discussions in the supplementary information.

To further validate that the paths with higher weights are more relevant to the progression of prostate cancer, we perform the KEGG Pathway enrichment analysis [77] on the proteins of the top 50 paths and bottom 50 paths for the abiraterone–prostate cancer pair. As shown in Figure 5A and B, the paths with higher

Table 1. The top-3 critical paths and top-3 KEGG pathways of the potential drugs for the treatment of prostate cancer

Drug	Critical paths (Top-3)	KEGG pathways (Top-3)
Dutasteride	SRD5A2	PI3K-Akt signaling pathway
	ORM1→TMSB4X	Lipid and atherosclerosis
	SRD5A1→UGT2B17	Hepatitis B
Aspirin	PLAUR	Proteoglycans in cancer
	FASLG	Hepatitis B
	TGFB1	Lipid and atherosclerosis
Erlotinib	STAT3	EGFR tyrosine kinase inhibitor resistance
	CYP3A5	Chemical carcinogenesis—receptor activation
	STAT3→NT5C2	Prostate cancer
Nicergoline	ADRA1A	Steroid hormone biosynthesis
	ARRA1A→Hydrogen→Azelaic Acid→SRD5A2	Prostate cancer
	ADRA1A→Triphosadenine→Diethylstilbestrol→SLC30A3	Cysteine and methionine metabolism
Acetohydroxamic acid	MMP13	Human T-cell leukemia virus 1 infection
	MMP8→KLK2	Prostate cancer
	MMP13→MMP7	Human cytomegalovirus infection
Midostaurin	RET	PI3K-Akt signaling pathway
	AURKB	Chemical carcinogenesis—receptor activation
	CYP3A5	Chemical carcinogenesis—DNA adducts
Apalutamide	Enzalutamide→CYP3A5	Prostate cancer
	ABCB1	Chemical carcinogenesis—DNA adducts
	Abiraterone→SULT2A1	Metabolism of xenobiotics by cytochrome P450
Atorvastatin	ABCC4	Chemical carcinogenesis—DNA adducts
	CYP3A5	Drug metabolism—cytochrome P450
	CYP2C19	Metabolism of xenobiotics by cytochrome P450
Carisoprodol	CYP2C19	Arachidonic acid metabolism
	CYP2C19→Hydrogen→Glycerol→FERMT2	Chemical carcinogenesis—DNA adducts
	Oxicone→Chloride ion→ITGAV	Drug metabolism—cytochrome P450
Oxcarbazepine	AKR1C3	Steroid hormone biosynthesis
	CYP2C19	Chemical carcinogenesis—DNA adducts
	ABCB1	Chemical carcinogenesis—reactive oxygen species

These drugs are ranked top 10 by iDPath among all the FDA-approved drugs used in this study. The head (drug) and tail (prostate cancer) of these critical paths are ignored due to the limit of space. The top-3 critical paths are determined by the weights generated by the path attention module. The KEGG pathways are identified by KEGG enrichment analysis on the proteins existed in the top-50 critical paths and ranked by *P*-adjust values.

weights focus on the P13K–Akt signaling pathway [78], regulation of actin cytoskeleton [79], prostate cancer pathway and so on, which are highly related to the progression of prostate cancer. For example, the activation of P13K–Akt signaling pathway appears to be characteristic of many aggressive prostate cancers and is more frequently observed as prostate cancer progresses toward a resistant and metastatic disease [78]. In contrast, the paths with lower weights (Figure 5B) are more enriched in the pathways related to other cancers or more general cancer progression, not specific to prostate cancer.

We investigated the roles of drug-related proteins in the identified critical paths by counting the frequency of proteins with different roles in the top-k critical paths. As show in Figure 6A, we found that the proteins in the identified critical paths are mainly disease *targets*, followed by *enzyme, transporter* and *carrier*, which is consistent with the principles of drug design and discovery [80]. Specifically, we also investigated the roles of drug-related proteins in abiraterone (Figure 4A) and penbutolol (Figure 4B), and the results are consistent with Figure 6A. For example, for abiraterone, we found CYP3A4 and AR are both in the high-weight paths, which are all commonly used targets for prostate cancer [70, 71]. We further investigated the relationship between the similarity of drugs (diseases) and the similarity of critical paths (see supplementary information for more details). As shown in Figure 6B and C, we found similar drugs or diseases have similar critical paths, indicating that similar drugs or diseases have similar MODA.

iDPath identified the critical paths to uncover the synergistic effect of drug combinations

iDPath represents a multilayer network approach to understanding the MODA. The interactions between proteins and chemicals and the interactions among chemicals can reveal more detailed therapeutic effects of individual drugs and potential drug combinations. Among the 16 interacted chemicals of abiraterone in our dataset, docetaxel and cabozantinib are identified as the most relevant contributors to the positive treatment effects of abiraterone on prostate cancer (Figure 4A). We notice that both docetaxel and cabozantinib show distinct molecular interactions targeting on prostate cancer while being used together with abiraterone. The combination of docetaxel and abiraterone can significantly improve radiographic progression-free survival for patients with metastatic castration-sensitive prostate cancer [81]. Cabozantinib enhances the anti-prostate cancer activity of abiraterone by inhibiting abiraterone's upregulation of IGFIR phosphorylation [82]. The identification of these combinations shows that iDPath has the capability to herald the synergistic drug combinations, even iDPath is not explicitly trained to perform this task.

Drug repurposing for prostate cancer

To demonstrate iDPath's utility in the real-world setting, we apply it to the discovery of potential drugs for treating prostate cancer among 1080 FDA-approved drugs that have not been labeled as therapeutic drugs for prostate cancer in our dataset.

We found that compared to the bottom-ranked drugs, the top-ranked drugs are more similar to the FDA-approved drugs (Supplementary Figure S7), indicating that iDPath identified potential drugs for treating prostate cancer. The 10 drugs with the highest score, together with their top-3 critical paths and top-3 KEGG pathways, are shown in Table 1. Among the 10 drugs identified for prostate cancer, six drugs have already been proved effective in previous studies, including dutasteride [83], aspirin [84], erlotinib [85], midostaurin [86], apalutamide [87] and atorvastatin [88]. The critical paths identified by iDPath shown in Table 1 are also consistent with drugs' MODA. For example, dutasteride is a medication primarily used to treat the symptoms of an enlarged prostate, shows therapeutic effects on prostate cancer by inhibiting dual 5α-reductase inhibitors (both SRD5A1 and SRD5A2) [89]. Aspirin is found to trigger cancer cell apoptosis by inducing the secretion of TGF-β1 (TGFB1) [90]. Apalutamide has recently been approved for the treatment of prostate cancer [91], but has not been labeled in our dataset. Specifically, iDPath finds that the most relevant paths for the efficacy of apalutamide are through enzalutamide or abiraterone (both are FDA-approved drugs for the treatment of prostate cancer and labeled as therapeutic in our dataset), where the combination with abiraterone has already been proved synergistic in a recent study [92]. For other drugs identified as therapeutic but not officially approved, the KEGG pathway enrichment analysis shows that the proteins that existed in their critical paths enriched in prostate cancer-related pathways, such as PI3K–Akt signaling pathway [93] and prostate cancer pathway.

Conclusion

In this study, we propose iDPath, an advanced deep learning framework to identify explainable biological paths to characterize the MODAs and predict the drugs that can be repurposed for treating certain diseases. iDPath is built on a multilayer biological network consisting of GRN, PPI, PCI and CCI networks. The proposed model achieves superior prediction performance compared with state-of-the-art models on a general drug repurposing task. Furthermore, we find that extending the PPI network to a multilayer biological network of the human body can significantly improve the prediction performance in drug repurposing. We investigate the identified critical paths of drugs for treating prostate cancer and hypertension and find that the critical paths are consistent with the known mechanism of the drug action. Then, we apply iDPath to the challenging problem of identifying potential drugs for the treatment of prostate cancer. Results show that iDPath can effectively identify the newly approved drugs not recorded by the database. We believe iDPath can bring revelation to the explainable deep learning technologies to drug discovery. As a deep learning approach, iDPath is limited to *in silico* study, which can be extended by *in vitro* and *in vivo* experiments to further validate its practical value and consistency with clinical evidence in future studies. In addition, the identified paths may contain rich biological knowledge beyond this study, such as some popular paths may be associated with common mechanisms of action in a class of diseases, which is worth further study.

Key Points

- A comprehensive multilayer biological network beyond protein–protein interactions is introduced to accurately characterize the mechanism of drug action.
- We propose an interpretable deep learning framework—iDPath to model the pathways of drugs by identifying explainable biological paths from drug targets to disease targets in the multilayer biological network of the human body.
- The superior performance of iDPath is verified by experiments on a general drug repurposing task.
- The model interpretability and credibility of iDPath is further validated on drugs treating prostate cancer and hypertension.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The data used to train iDPath and its source code and usage instructions are available in Github (https://github.com/JasonJYang/iDPath).

Author contributions statement

J.Y. and Q. Z.: study concept and design, development of methodology, writing of the manuscript; J.Y.: acquisition of samples and data, analysis and interpretation of data; Z.L., S.Y., Z.X., W.K.K.W. and Q.C.: interpretation of data; Q.Z.: study supervision and funding acquisition.

Funding

This work was supported by National Natural Science Foundation of China (71972164, 71672163, 62131009, and 82071889); Innovation and Technology Fund of Innovation and Technology Commission of Hong Kong (MHP/081/19); National Key Research and Development Program of China, Ministry of Science and Technology of China (2019YFE0198600).

References

1. Fleming N. How artificial intelligence is changing drug discovery. *Nature* 2018;**557**:S55–5, 7.
2. Pham T-H, Qiu Y, Zeng J, *et al.* A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell* 2021;**3**:247–57.
3. Jin W, Stokes JM, Eastman RT, *et al.* Deep learning identifies synergistic drug combinations for treating COVID-19. *Proc Natl Acad Sci* 2021;**118**:e2105070118.
4. Yan VK, Li X, Ye X, *et al.* Drug repurposing for the treatment of COVID-19: a knowledge graph approach. *Adv Ther* 2021;**4**:2100055.
5. Rodriguez S, Hug C, Todorov P, *et al.* Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat Commun* 2021;**12**:1–13.
6. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform* 2021;**22**:360–79.
7. Zhou Y, Wang F, Tang J, *et al.* Artificial intelligence in COVID-19 drug repurposing. *Lancet Digital Health* 2020;**2**:e667–76.

8. Sanseau P, Koehler J. Editorial: computational methods for drug repurposing. *Brief Bioinform* 2011;**12**:301–2.
9. Pushpakom S, Iorio F, Eyers PA, *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**: 41–58.
10. Xu H, Aldrich MC, Chen Q, *et al.* Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *J Am Med Inform Assoc* 2015;**22**:179–91.
11. Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat Mach Intell* 2021;**3**:68–75.
12. Dakshanamurthy S, Issa NT, Assefnia S, *et al.* Predicting new indications for approved drugs using a proteochemometric method. *J Med Chem* 2012;**55**:6832–48.
13. Sanseau P, Agarwal P, Barnes MR, *et al.* Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* 2012;**30**: 317–20.
14. Greene CS, Krishnan A, Wong AK, *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015;**47**:569–76.
15. Yang X, Kui L, Tang M, *et al.* High-throughput transcriptome profiling in drug and biomarker discovery. *Front Genet* 2020;**11**:19.
16. Silverbush D, Sharan R. A systematic approach to orient the human protein–protein interaction network. *Nat Commun* 2019;**10**:1–9.
17. Cheng F, Desai RJ, Handy DE, *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018;**9**:1–12.
18. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nat Commun* 2019;**10**:1–11.
19. Gysi DM, Do Valle Í, Zitnik M, *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc Natl Acad Sci* 2021;**118**.
20. Zhou Y, Hou Y, Shen J, *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;**6**: 1–18.
21. Yang J, Xu Z, Wu WKK, *et al.* GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction. *J Am Med Inform Assoc* 2021;**28**:2336–45.
22. Lopes-Ramos CM, Kuijjer ML, Ogino S, *et al.* Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer Res* 2018;**78**:5538–47.
23. Kalinina OV, Wichmann O, Apic G, *et al.* Combinations of protein-chemical complex structures reveal new targets for established drugs. *PLoS Comput Biol* 2011;**7**:e1002043.
24. Hu L-L, Chen C, Huang T, *et al.* Predicting biological functions of compounds based on chemical-chemical interactions. *PLoS One* 2011;**6**:e29491.
25. Liu X, Pan L. Detection of driver metabolites in the human liver metabolic network using structural controllability analysis. *BMC Syst Biol* 2014;**8**:51–17.
26. Zhou W, Wang Y, Lu A, *et al.* Systems pharmacology in small molecular drug discovery. *Int J Mol Sci* 2016;**17**:246.
27. Issa NT, Wathieu H, Ojo A, *et al.* Drug metabolism in preclinical drug development: a survey of the discovery process, toxicology, and computational tools. *Curr Drug Metab* 2017;**18**:556–65.
28. Sun YV. Integration of biological networks and pathways with genetic association studies. *Hum Genet* 2012;**131**:1677–86.
29. Snape TJ, Astles AM, Davies J. Understanding the chemical basis of drug stability and degradation. *Pharm J* 2010;**285**:416–7.
30. Löwenberg B, Pabst T, Vellenga E, *et al.* Cytarabine dose for acute myeloid leukemia. *N Engl J Med* 2011;**364**:1027–36.
31. Hamada A, Kawaguchi T, Nakano M. Clinical pharmacokinetics of cytarabine formulations. *Clin Pharmacokinet* 2002;**41**:705–18.
32. Napolitano F, Zhao Y, Moreira VM, *et al.* Drug repositioning: a machine-learning approach through data integration. *J Chem* 2013;**5**:1–9.
33. Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics* 2020;**36**:i525–33.
34. Wang W, Yang S, Zhang X, *et al.* Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;**30**:2923–30.
35. Zhang F, Wang M, Xi J, *et al.* A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**:1–9.
36. Liu H, Song Y, Guan J, *et al.* Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinform* 2016;**17**:269–77.
37. Yang M, Wu G, Zhao Q, *et al.* Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief Bioinform* 2021;**22**:bbaa267.
38. Liu X, Maiorino E, Halu A, *et al.* Robustness and lethality in multilayer biological molecular networks. *Nat Commun* 2020;**11**: 1–12.
39. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *International Conference on Machine Learning*. 2017, 3145–53. PMLR, Sydney, Australia.
40. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, United States. 2016, 1135–44.
41. Wang HW, Zhang FZ, Wang JL *et al.* RippleNet: propagating user preferences on the knowledge graph for recommender systems, Cikm'18: *Proceedings of the 27th Acm International Conference on Information and Knowledge Management*. Association for Computing Machinery New York, NY, United States. 2018: 417–26.
42. Wang X,Wang DX, Xu CR et al. Explainable reasoning over knowledge graphs for recommendation. *AAAI-19: Proceedings of the Thirty-Third AAAI conference on artificial intelligence*. 2019:5329–36. Association for the Advancement of Artificial Intelligence 2275 East Bayshore Road, Suite 160 Palo Alto, California, United States.
43. Elmarakeby HA, Hwang J, Arafeh R, *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* 2021;**598**:348–52.
44. Ma JZ, Yu MK, Fong S, *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* 2018;**15**:290–8.
45. Kuenzi BM, Park J, Fong SH, *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020;**38**:672–684.e6.
46. Xu ZZ, Zhang J, Zhang QP, *et al.* A comorbidity knowledge-aware model for disease prognostic prediction. *IEEE Trans Cybernet* 2021;**52**:9809–9819.
47. Guo M, Xu Z, Zhang Q, *et al.* Deciphering feature effects on decision-making in ordinal regression problems: an explainable ordinal factorization model. *ACM Trans Knowl Discov Data (TKDD)* 2021;**16**:1–26.
48. Liu Z-P, Wu C, Miao H, *et al.* RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015;**2015**:bav095.

49. Szklarczyk D, Gable AL, Nastou KC, *et al.* Correction to 'The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets'. *Nucleic Acids Res* 2021;**49**:10800–0.
50. Szklarczyk D, Santos A, von Mering C, *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**:D380–4.
51. BioMart KA. BioMart: driving a paradigm change in biological data management. *Database* 2011;**2011**:bar049.
52. Maglott D, Ostell J, Pruitt KD, *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2011;**39**:D52–7.
53. Kim S, Chen J, Cheng TJ, *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**:D1102–9.
54. Zhou Y, Zhang YT, Lian XC, *et al.* Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res* 2022;**50**: D1398–407.
55. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
56. Pinero J, Bravo A, Queralt-Rosinach N, *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–9.
57. Huang Z, Shi JC, Gao YX, *et al.* HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019;**47**:D1013–7.
58. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 2016.
59. Ren Y, Ay A, Kahveci T. Shortest path counting in probabilistic biological networks. *BMC Bioinform* 2018;**19**:1–19.
60. Hricik T, Bader D, Green O. *Using RAPIDS AI to accelerate graph data science workflows*. In: *2020 IEEE High Performance Extreme Computing Conference (HPEC)*. 2020, 1–4. IEEE, Manhattan, New York, United States.
61. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
62. Dwarampudi M, Reddy N. Effects of padding on LSTMs and CNNs. arXiv preprint arXiv:1903.07288. 2019.
63. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inform Process Syst* 2017;**30**:5998–6008.
64. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, United States. 2014, p. 701–10.
65. Chen Z-H, You Z-H, Guo Z-H, *et al.* Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model. *Front Bioeng Biotechnol* 2020;**8**:338.
66. Sakai M, Nagayasu K, Shibui N, *et al.* Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci Rep* 2021;**11**:1–14.
67. Xu Z, Wang S, Zhu F *et al.* Seq2seq fingerprint: an unsupervised deep molecular embedding for drug discovery. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. Association for Computing Machinery, New York, NY, United States. 2017, 285–94.
68. Santiso S, Perez A, Casillas A. Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery. *IEEE J Biomed Health Inform* 2019;**23**:2148–55.
69. Benoist GE, Hendriks RJ, Mulders PFA, *et al.* Pharmacokinetic aspects of the two novel oral drugs used for metastatic castration-resistant prostate cancer: abiraterone acetate and enzalutamide. *Clin Pharmacokinet* 2016;**55**:1369–80.
70. Zeigler-Johnson C. CYP3A4: a potential prostate cancer risk factor for high-risk groups. *Clin J Oncol Nurs* 2001;**5**:153–4.
71. Fujita K, Nonomura N. Role of androgen receptor in prostate cancer: a review. *World J Mens Health* 2019;**37**:288–95.
72. Jentzmik F, Azoitei A, Zengerling F, *et al.* Androgen receptor aberrations in the era of abiraterone and enzalutamide. *World J Urol* 2016;**34**:297–303.
73. Malikova J, Brixius-Anderko S, Udhanea SS, *et al.* CYP17A1 inhibitor abiraterone, an anti-prostate cancer drug, also inhibits the 21-hydroxylase activity of CYP21A2. *J Steroid Biochem Mol Biol* 2017;**174**:192–200.
74. DeVore NM, Scott EE. Structures of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* 2012;**482**:116–9.
75. Langlois M, Brémont B, Rousselle D, *et al.* Structural analysis by the comparative molecular field analysis method of the affinity of β-adrenoreceptor blocking agents for 5-HT1A and 5-HT1B receptors. *Eur J Pharmacol Mol Pharmacol* 1993;**244**:77–87.
76. Saxena PR, Villalón CM. Cardiovascular effects of serotonin agonists and antagonists. *J Cardiovasc Pharmacol* 1990;**15**:S17–34.
77. Ogata H, Goto S, Fujibuchi W, *et al.* Computation with the KEGG pathway database. *Biosystems* 1998;**47**:119–28.
78. Toren P, Zoubeidi A. Targeting the PI3K/Akt pathway in prostate cancer: challenges and opportunities (review). *Int J Oncol* 2014;**45**: 1793–801.
79. Yamaguchi H, Condeelis J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. *BBA-Mol Cell Res* 2007;**1773**:642–52.
80. Anderson AC. The process of structure-based drug design. *Chem Biol* 2003;**10**:787–97.
81. Fizazi K, Maldonado X, Foulon S, *et al. A Phase 3 Trial With a 2x2 Factorial Design of Abiraterone Acetate Plus Prednisone and/or Local Radiotherapy in Men With De Novo Metastatic Castration-Sensitive Prostate Cancer (mCSPC): First Results of PEACE-1*. Vol 39, pp. 5000–5000. Journal of Clinical Oncology, Alexandria, VA, USA, 2021.
82. Wang XD, Huang Y, Christie A, *et al.* Cabozantinib inhibits abiraterone's upregulation of IGFIR phosphorylation and enhances its anti-prostate cancer activity. *Clin Cancer Res* 2015;**21**: 5578–87.
83. Andriole GL, Bostwick DG, Brawley OW, *et al.* Effect of dutasteride on the risk of prostate cancer. *N Engl J Med* 2010;**362**:1192–202.
84. Joshi S, Murphy E, Olaniyi P, *et al.* The multiple effects of aspirin in prostate cancer patients. *Cancer Treat Res Commun* 2021;**26**:100267.
85. Gravis G, Goncalves A, Bladou F, *et al.* Monocentric evaluation of erlotinib in advanced prostate cancer. *J Clin Oncol* 2007;**25**: 15569.
86. Krishnappa K, Mallesh NK, Sharma SC, *et al.* Midostaurin inhibits hormone-refractory prostate cancer PC-3 cells by modulating nPKCs and AP-1 transcription factors and their target genes involved in cell cycle. *Front Biol* 2017;**12**:421–9.
87. Smith MR, Saad F, Chowdhury S, *et al.* Apalutamide treatment and metastasis-free survival in prostate cancer. *N Engl J Med* 2018;**378**:1408–18.
88. Khosropanah I, Falahatkar S, Farhat B, *et al.* Assessment of atorvastatin effectiveness on serum PSA level in hypercholesterolemic males. *Acta Med Iran* 2011;**49**:789–94.
89. Festuccia C, Gravina GL, Muzi P, *et al.* Effects of dutasteride on prostate carcinoma primary cultures: a comparative study with finasteride and MK386. *J Urol* 2008;**180**:367–72.

90. Wang Y, Du C, Zhang N, *et al.* TGF-β1 mediates the effects of aspirin on colonic tumor cell proliferation and apoptosis. *Oncol Lett* 2018;**15**:5903–9.
91. Al-Salama ZT. Apalutamide: first global approval. *Drugs* 2018;**78**:699–705.
92. Saad F, Efstathiou E, Attard G, *et al.* Apalutamide plus abiraterone acetate and prednisone versus placebo plus abiraterone and prednisone in metastatic, castration-resistant prostate cancer (ACIS): a randomised, placebo-controlled, double-blind, multinational, phase 3 study. *Lancet Oncology* 2021;**22**:1541–59.
93. Shorning BY, Dass MS, Smalley MJ, *et al.* The PI3K-AKT-mTOR pathway and prostate cancer: at the crossroads of AR, MAPK, and WNT signaling. *Int J Mol Sci* 2020;**21**:4507.