REVIEW ARTICLE OPEN

Artificial intelligence in cancer target identification and drug discovery

Yujie You[1], Xin Lai[2], Yi Pan[3], Huiru Zheng[4], Julio Vera[2], Suran Liu[1], Senyi Deng[5]✉ and Le Zhang[1,6,7]✉

Artificial intelligence is an advanced method to identify novel anticancer targets and discover novel drugs from biology networks because the networks can effectively preserve and quantify the interaction between components of cell systems underlying human diseases such as cancer. Here, we review and discuss how to employ artificial intelligence approaches to identify novel anticancer targets and discover drugs. First, we describe the scope of artificial intelligence biology analysis for novel anticancer target investigations. Second, we review and discuss the basic principles and theory of commonly used network-based and machine learning-based artificial intelligence algorithms. Finally, we showcase the applications of artificial intelligence approaches in cancer target identification and drug discovery. Taken together, the artificial intelligence models have provided us with a quantitative framework to study the relationship between network characteristics and cancer, thereby leading to the identification of potential anticancer targets and the discovery of novel drug candidates.

Signal Transduction and Targeted Therapy (2022)7:156 ; https://doi.org/10.1038/s41392-022-00994-0

INTRODUCTION

As one of the cutting-edge cancer treatments, targeted drug therapy has the advantages of high efficiency, few side effects, and low drug resistance for patients[1]. However, there are several drawbacks to the existing targeted therapies, such as a few druggable targets[2], ineffective coverage of the patient population, and the lack of alternative responses to drug resistance in patients[1]. Therefore, identifying novel therapeutic targets and evaluating their druggability[3,4] becomes the current cancer research focus of targeted drug therapy.

Since we have difficulty in comprehensively understanding the pathogenesis of cancer due to the complexity of the disease[5], most of the current targeted drugs are developed based on the experimentally validated hypothesis that can explain a possible mechanism underlying carcinogenesis but ignore other facts of the disease[6]. As a result, these therapies could have undesired impacts on normal tissues and even provoke serious side effects for patients[7,8].

To elucidate the molecular mechanisms underlying cancer genesis, interactome data can be comprised and modelled in network structures in which components are biological entities (e.g., genes, proteins, mRNAs, and metabolites) and edges are associations/interactions between them (e.g., gene co-expression, signalling transduction, gene regulation, and physical interaction between proteins[9–14]). Artificial intelligence biology analysis algorithms are effective method to process the biological network data, which build machines or programs to simulate human intelligence, so as to implement classification, clustering and prediction tasks in biological network[15]. Therefore, artificial intelligence algorithms can effectively tackle the complexity of cancer that arises from interactions between genes and their products[16,17] in biological network structures, so as to improve our understanding of carcinogenesis[11,12,18–22] and explore novel anticancer targets[23–29].

Over the past few decades, we have seen a fast development of artificial intelligence biology analysis algorithms. To make this study easy to understand, we not only divide these artificial intelligence algorithms into network-based biology analysis algorithm and machine learning-based (ML-based) biology analysis algorithm according to the data of biological network structure, but also employ Fig. 1 to describe the historical milestone for these artificial intelligence biology analysis algorithms.

On the one hand, network-based biology analysis algorithms provide a variety of alternative network approaches to identify cancer targets. More importantly, various network-based biology analysis algorithms can investigate network data from different perspectives, therefore they can compensate each other to provide accurate biological explanations[30].

On the other hand, ML-based biology analysis[31–33] not only can efficiently handle high throughput, heterogeneous, and complex molecular data, but also can mine the feature or relationship in the biological networks. Thus, we should develop more ML-based biology analysis algorithms to provide such advanced biology

[1]College of Computer Science, Sichuan University, Chengdu 610065, China; [2]Laboratory of Systems Tumor Immunology, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Universitätsklinikum Erlangen, Erlangen 91052, Germany; [3]Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Room D513, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen 518055, China; [4]School of Computing, Ulster University, Belfast BT15 1ED, UK; [5]Institute of Thoracic Oncology, Department of Thoracic Surgery, West China Hospital, Sichuan University, Chengdu 610065, China; [6]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China and [7]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
Correspondence: Senyi Deng (senyi_deng@scu.edu.cn) or Le Zhang (zhangle06@scu.edu.cn)
These authors contributed equally: Yujie You, Xin Lai.

Received: 4 July 2021 Revised: 14 March 2022 Accepted: 5 April 2022
Published online: 10 May 2022

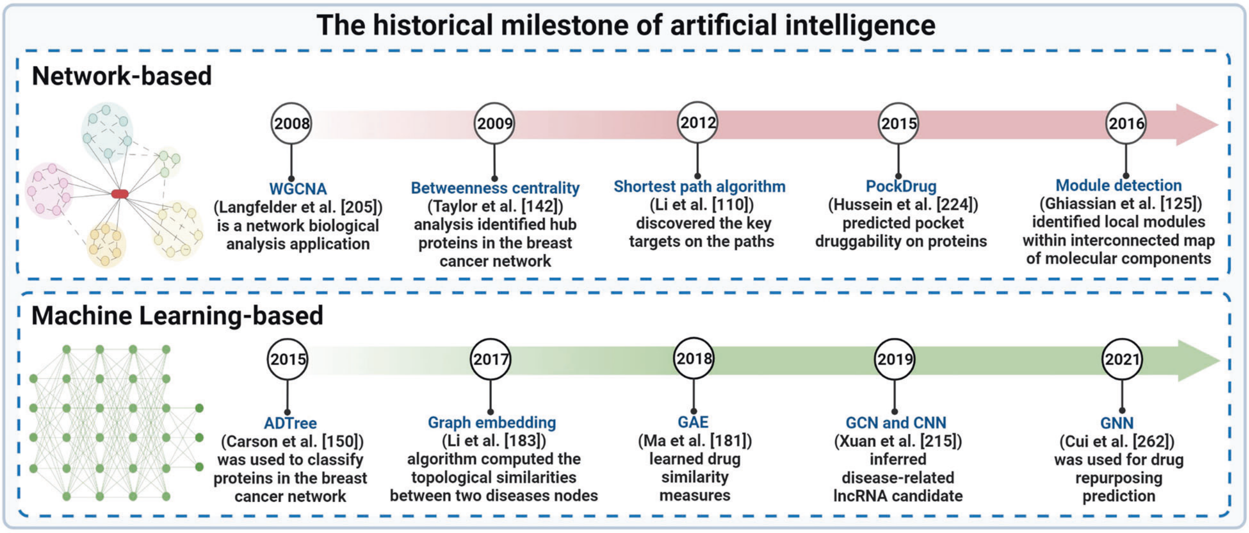

Fig. 1 The historical milestones of network-based and ML-based biology analysis. (Created with BioRender.com)

analyses that can allow precise target identification and drug discovery for cancer.

Although artificial intelligence biology analysis has been widely used to improve our understanding of carcinogenesis, to the best of our knowledge, there is no systematic review that introduces the scope of related research and explains the network-based and the ML-based biology analysis algorithms to identify novel anticancer targets and discover drugs. Therefore, in the next section, we will describe the scope of artificial intelligence biology analysis for novel anticancer targets investigation. In the third section, we will introduce the basic principles and theory of commonly used artificial intelligence biology analysis algorithms. Then, we will briefly review and discuss studies that utilize network-based and ML-based biology analysis for cancer target identification and drug discovery. Finally, we will summarize the content of the article, discuss the limitations and challenges faced by the community, and point out the potential of artificial intelligence biology analysis to identify the therapeutic targets and discover drugs for cancer.

THE SCOPE OF ARTIFICIAL INTELLIGENCE BIOLOGY ANALYSIS FOR NOVEL ANTICANCER TARGET INVESTIGATIONS

Recently, the rapid development of cancer-related multiomics technologies[34–36] has been one of the most important factors for artificial intelligence biology analysis to explore novel anticancer targets[37–39]. Figure 2 classifies these technologies into five aspects: epigenetics, genomics, proteomics, metabolomics, and multiomics integration analysis. Furthermore, Table 1 lists the related major diseases, drug targets, genomics, and network databases commonly used in multiomics integration analysis for these five aspects. Next, we will detail these five aspects.

Epigenetics analyses the reversal modifications of DNA or DNA-related proteins[54]. These modifications affect gene expression without changing the DNA sequence[54]. Investigating epigenetic data through artificial intelligence is not only important for elucidating fundamental mechanisms of cancer but also necessary for the design of targeted therapeutics. For example, Wilson et al.[55] took advantage of information-rich transcriptomic and epigenetic data to study regulatory networks surrounding histone lysine demethylation and highlighted the importance of epigenetic regulators in mitogenic control and their potential as therapeutic targets, which showed that epigenetic regulators such as KDM1A, KDM3A, EZH2, and DOT1L[56] are critical in oncogenesis and drug resistance.

Genomics aims to characterize the function of every genomic element of an organism by using genome-scale assays such as genome sequencing[57]. Applications of genomics include finding associations between genotype and phenotype[58], discovering biomarkers for patient stratification[59], predicting the function of genes[60] and charting biochemically active genomic regions such as transcriptional enhancers[49]. Recent developments in network-based biology analysis methods, such as sequence-similarity networks, genome networks, and gene family networks, have significantly improved the usability of molecular datasets in comparative genomics analysis[61]. These network methods collect expression and interaction data in the beginning and then transform them into interpretable biological processes[62,63], leading to the identification of tumour subtypes and the discovery of drug targets[64].

For example, Medi et al.[65] integrated gene expression profiles into genome-scale molecular networks to identify novel therapeutic targets for cervical cancer, including receptors, microRNAs (miRNAs), transcription factors (TFs), proteins (e.g., CRYAB, CDK1, PARP1, WNK1, GSK3B, and KAT2B), and metabolites (arachidonic acids). Laura et al.[66] developed a network-based biology analysis workflow that integrates different layers of genomic information, including transcription factor cotargeting, miRNA cotargeting, protein–protein interaction and gene coexpression, into a biological network. Then, the authors applied a consensus clustering algorithm (An ML-based biology analysis algorithm that divide the network into sub-modules with different functions)[67–73] on identified network communities to discover cancer driver genes, which demonstrated that F11R, HDGF, PRCC, ATF3, BTG2, and CD46 could be oncogenes and promising markers for pancreatic cancer.

For proteomics, proteomic experiments are performed for annotation and correlation of genome sequences, quantitation of protein abundance, detection of posttranslational modifications, and identification of protein-protein interactions (PPIs)[74]. PPIs not only play fundamental roles in structuring and mediating biological processes but also have been widely used for proteomics data analysis[75]. For example, Vinayagam et al.[37] analysed the human PPI interaction network to identify indispensable proteins that affect the controllability of the network with control theory[76], which shows that if a system can be driven from any initial state to any desired final state in finite time with a suitable choice of inputs, the system is controllable. By changing the number of driver nodes in the network upon removal of that protein, the hub can be classified as “indispensable” “neutral” or

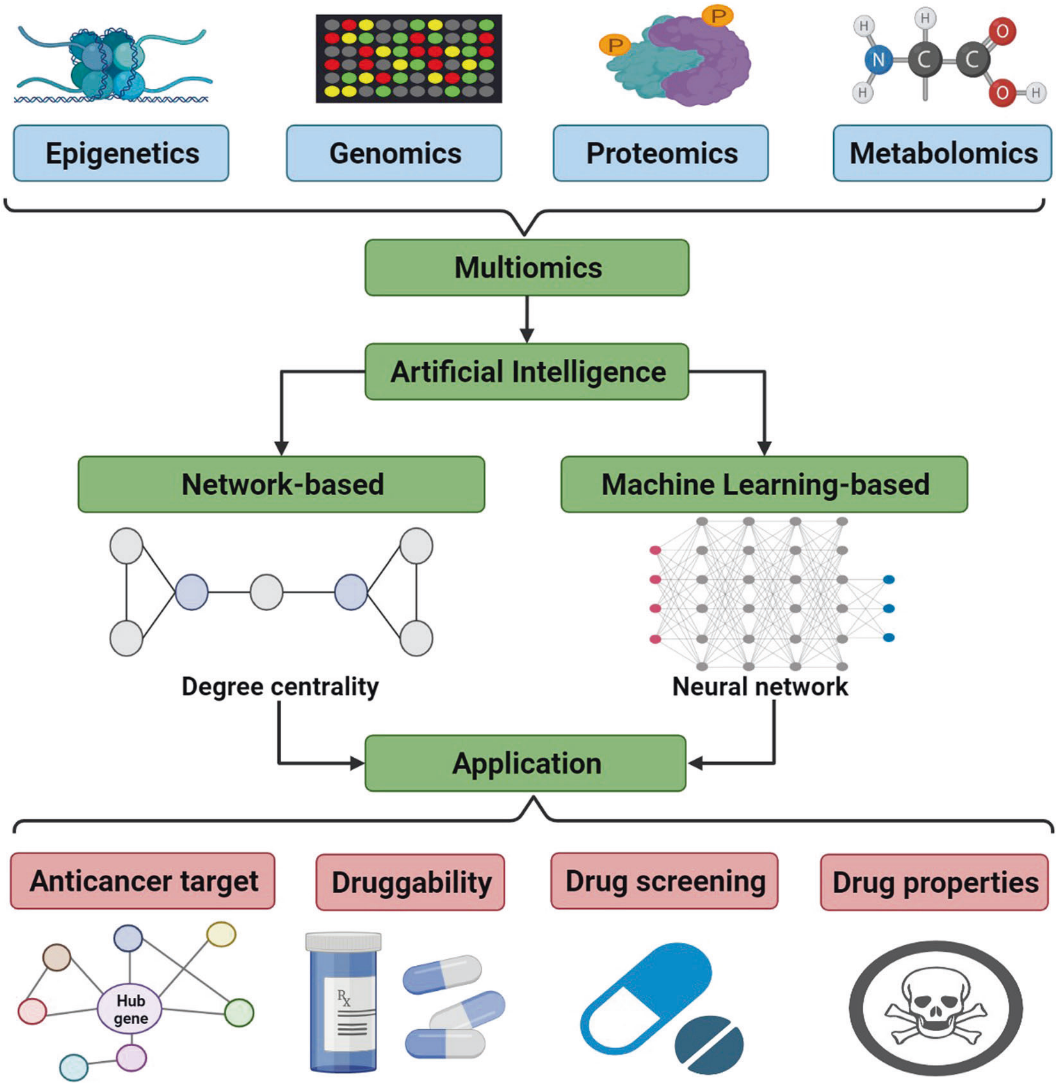

Fig. 2 Artificial intelligence to integrate multiomics data (e.g., epigenetics, genomics, proteomics, and metabolomics) for cancer therapeutic targets identification. (Created with BioRender.com)

"dispensable", which correlates with increasing, no effect, or decreasing the number of driver nodes in the network upon removal of the key protein. The evidence shows that these indispensable proteins are primary targets of disease-causing mutations, viruses, and drugs.

Furthermore, analysing data from 1,547 cancer patients revealed 56 indispensable genes in nine cancers. 46 of these genes were associated with cancer for the first time, demonstrating the ability of intelligent network controllability analysis to identify novel disease genes and potential drug targets[77]. Moreover, Valle et al.[78] developed a network-based biology analysis framework to compute the proximity between polyphenol targets and disease proteins. The calculated results indicated that the diseases whose proteins are proximal to polyphenol targets have significant gene expression changes, while the diseases whose proteins are distal to polyphenol targets have no such change. The network relationship between disease proteins and polyphenol targets provides not only a computing method to reveal the effect of polyphenols on diseases but also a basis to identify novel anticancer targets.

Metabolomics is routinely applied for biomarker discovery by profiling metabolites in biofluids, cells and tissues[34]. Because of the inherent sensitivity of biotechnology, subtle alterations in metabolic pathways can be detected to provide insights into the mechanisms that underlie various physiological conditions and cancer processing[34]. Owing to innovative developments in network biology, researchers employ biological networks to perform metabolomic analyses and provide us with a systems-level understanding of the role that metabolites play in cancer.

For example, Basler et al.[79] proposed an effective network-based biology analysis framework for the systematic study of flow control and identification of driver reactions in large-scale metabolic networks. They found that the driver reactions were under complex cellular regulation in Escherichia coli, suggesting their preeminent role in facilitating cellular control. Correlation statistics indicate that the driven response plays an important role in inhibiting tumour growth and represents a potential therapeutic target.

For multiomics integration analysis, addressing the complexity of tumour-host interactions requires an approach to handle integrative omics data[80]. Compared to single omics studies, multiomics data provide researchers with various and interconnected molecular profiles to study carcinogenesis[80]. Thus, integrated multiomics datasets in a network structure to artificial

Table 1. Commonly used repositories related to human diseases, drug targets, genomics, and biological networks

Database name	Description	Web link	Ref
Disease			
Online Mendelian Inheritance in Man (OMIM)	A comprehensive, authoritative, and timely knowledgebase of human genes and genetic disorders	http://www.omim.org/	40
Pathologisch Anatomisch Landelijk Geautomatiseerd Archief (PALGA)	A database of histopathology and cytopathology was stored.	https://www.palga.nl	41
Drug Target			
DrugBank	DrugBank is a web-enabled database containing comprehensive molecular information about drugs, their mechanisms, their interactions, and their targets.	https://www.drugbank.ca/	42
Therapeutic Targets Database (TTD)	A database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, etc.	http://db.idrblab.net/ttd/	43
PubChem	PubChem is an open repository for chemical structures and their biological test results.	http://pubchem.ncbi.nlm.nih.gov	44
ChEMBL	ChEMBL is an open data database containing binding, functional and ADMET information for many drug-like bioactive compounds.	https://www.ebi.ac.uk/chembldb	45
Genomics Data			
Gene Expression Omnibus (GEO)	GEO is a public functional genomics data repository. Array- and sequence-based data are accepted.	https://www.ncbi.nlm.nih.gov/geo/	46
The Cancer Genome Atlas (TCGA)	TCGA contains clinical data of various human cancers, genomic mutations, mRNA expression, miRNA expression, methylation, etc.	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga	47
Cancer Cell Line Encyclopedia (CCLE)	A compilation of gene expression, chromosomal copy number and massively parallel sequencing data from 947 human cancer cell lines.	https://sites.broadinstitute.org/ccle	48
ENCyclopedia Of DNA Elements (ENCODE)	ENCODE has systematically mapped regions of transcription, transcription factor association, chromatin structure, and histone modification.	https://www.encodeproject.org/	49
Catalogue Of Somatic Mutations In Cancer (COSMIC)	COSMIC curates comprehensive information on somatic mutations in human cancer.	http://www.sanger.ac.uk/cosmic	50
Biological Network			
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	A database of known and predicted protein interactions	http://string-db.org/	51
Gene Ontology (GO)	The world's largest source of information on the functions of genes.	http://www.geneontology.org/	52
Kyoto Encyclopedia of Genes and Genomes (KEGG)	A collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances	http://www.genome.jp/kegg/	53

intelligence biology analysis has emerged as a powerful tool to fully appreciate the complex interlayer regulatory interactions in cancer progression. Such an approach allows us to benefit from prior information that can be summarized and presented in networks, thereby providing us with insights into carcinogenesis from an overall perspective[81].

For example, Gov et al.[82] first performed comparative analyses of transcriptome data, and then identified common and tissue-specific reporter biomolecules such as genes, receptors, membrane proteins, TFs, and miRNAs. Second, they used the interactions among receptors, TFs, miRNAs, and their targeted DEGs to reconstruct a tissue-specific network for ovarian cancer and used network-based biology methods to identify interaction hubs. Finally, GATA2 and miR-124-3p were identified as hub nodes, suggesting that they are potential biomarkers for ovarian cancer.

THE PRINCIPLES AND THEORIES FOR COMMONLY USED ARTIFICIAL INTELLIGENCE BIOLOGY ANALYSIS ALGORITHMS

This study divides these commonly used artificial intelligence biology analysis algorithms into two categories. One is network-based biology analysis algorithm, including shortest path[83], module detection[84], and network centrality[85]; the other is ML-based biology analysis algorithm including decision tree[86–88] and deep learning models[89–91].

The principles and theory of network-based biology analysis algorithms

Biological networks are efficient in integrating complicated biological data, because they can capture the property of biological entities and their relationships[92]. Mathematically, a network can be represented as a graph $G = (V, E)$ where V and E are a set of nodes (vertices) and edges, respectively. Nodes in biological networks can represent proteins, genes, diseases, and drugs and edges in the network represent various biochemical physical or functional interactions between nodes. Therefore, network-based biology analysis algorithms focuses on identifying therapeutic targets and discovery of novel drugs for cancer from molecular networks such as protein-protein interaction networks[75], gene regulatory networks[93], metabolic networks[94], and drug-drug interaction networks[95].

Computational biologists have developed several network-based biology analysis algorithms to effectively process and

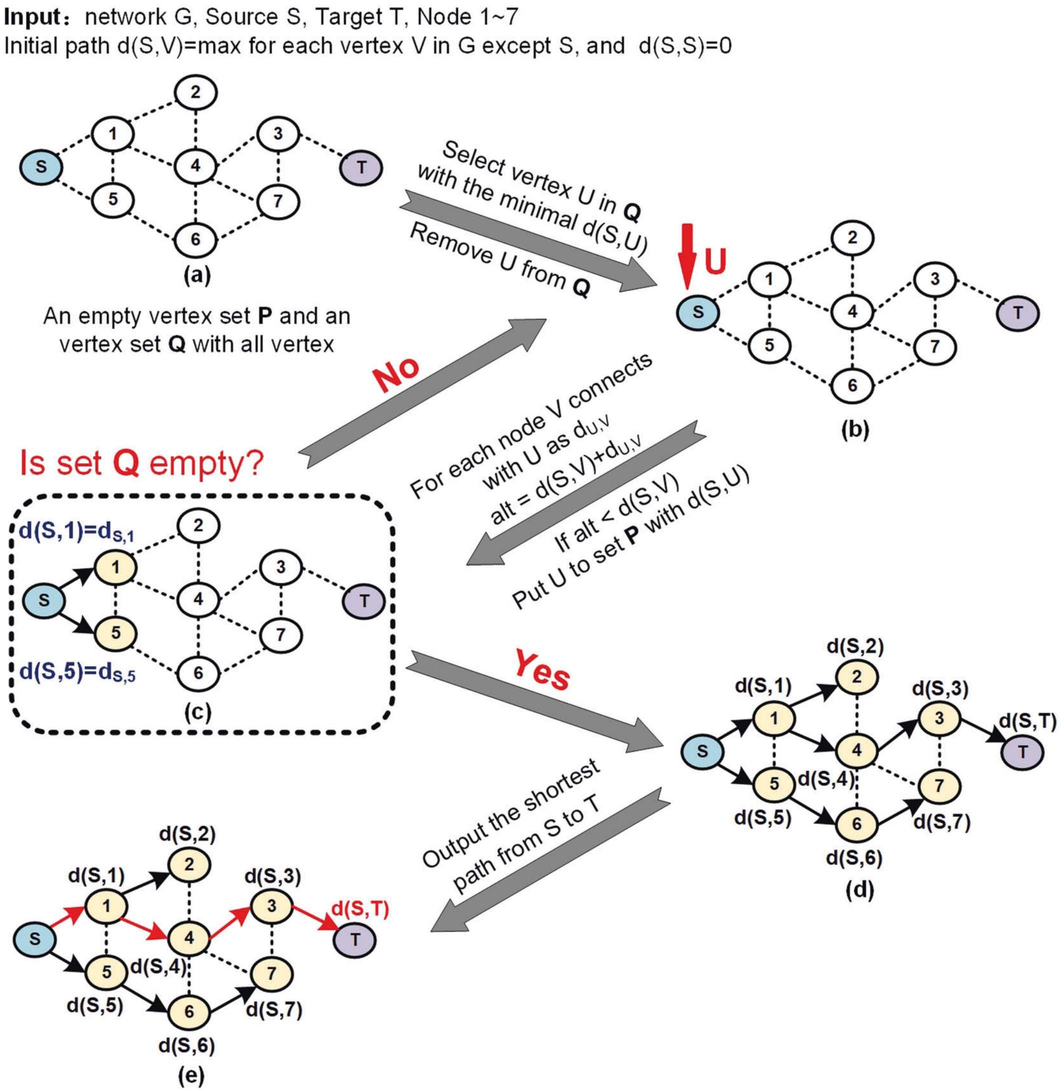
Input：network G, Source S, Target T, Node 1~7
Initial path d(S,V)=max for each vertex V in G except S, and d(S,S)=0
(a)
An empty vertex set P and an vertex set Q with all vertex
Select vertex U in Q with the minimal d(S,U)
Remove U from Q
U
(b)
No
Is set Q empty?
For each node V connects with U as d_U,V
alt = d(S,V)+d_U,V
If alt < d(S,V)
Put U to set P with d(S,U)
d(S,1)=d_S,1
d(S,5)=d_S,5
(c)
Yes
d(S,1)
d(S,2)
d(S,3)
d(S,T)
d(S,4)
d(S,5)
d(S,6)
d(S,7)
(d)
Output the shortest path from S to T
(e)
S
T
1
2
3
4
5
6
7

Fig. 3 The flow chart of the shortest path algorithm. The red paths in the bottom network are the identified shortest path from node *S* to *T*

analyze non-ordered or non-Euclidean data in biological networks, which can perform tasks such as link prediction[96], node ranking[85], network propagation[97], network modularization[98], and network control[99]. Here, we briefly review and discuss the shortest path algorithm, module detection algorithm, and node prioritization methods using node centrality in identifying cancer therapeutic targets and discovering drugs.

Tthe shortest path algorithm. The shortest path algorithm, one of network link algorithm, is used to intelligently identify the shortest connection between two genes or proteins in a graphical model that represents a cellular network[100,101]. The algorithm is illustrated in Fig. 3 and Algorithm 1. The shortest distance for a given network is calculated by Eq. (1):

$$d(S,T) = \min_{K \in V} d(S,K) + d_{K,T} \tag{1}$$

Here, *S* and *T* stand for the source and target node, respectively. *d (S,T)* is the length of the shortest path from node *S* to *T*. *V* is a set of network nodes. *K* stands for a node in the network, and $d_{K,T}$ represents the lengths of possible paths connecting nodes *K* and *T*.

Algorithm 1. The shortest path algorithm[102]

```
Input: Network G, Source S, Target T, Nodes
create an empty set P and a set Q contains all nodes
for each vertex V in Network:
    $d(S,V) \leftarrow$ infinity
$d(S,S) \leftarrow 0$
do:
    $U \leftarrow$ vertex in Q with minimal $d(S,U)$
    remove U from Q
    for each vertex V in Q that is connected with U:
        $alt \leftarrow d(S,U) + d_{U,V}$
        if $alt < d(S,V)$:
            $d(S,V) \leftarrow alt$
            add U to the set P
until Q is empty
Output: the shortest path from S to T
```

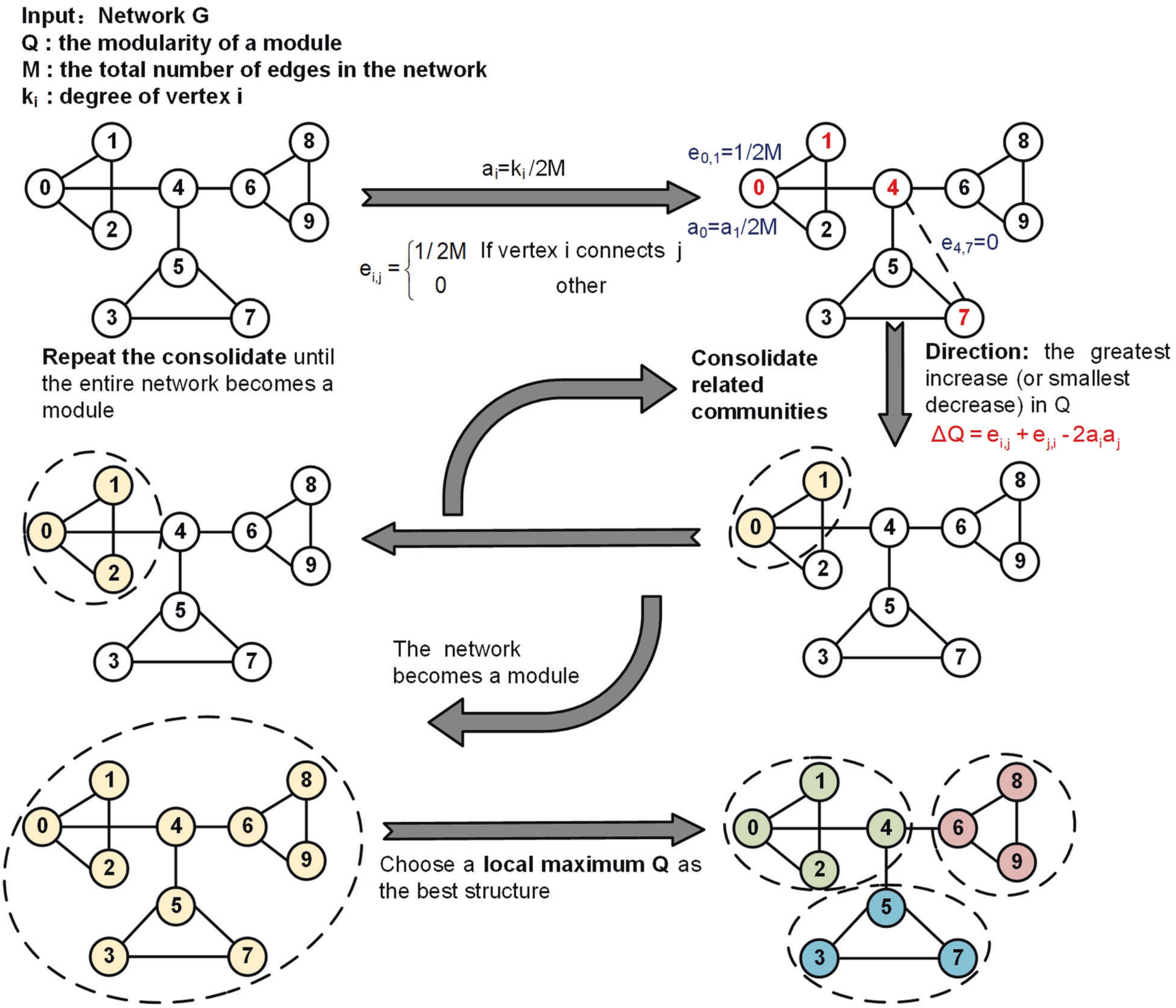
Input：Network G
Q : the modularity of a module
M : the total number of edges in the network
k_i : degree of vertex i
$a_i=k_i/2M$
$e_{i,j}=\begin{cases}1/2M & \text{If vertex i connects j}\\ 0 & \text{other}\end{cases}$
$e_{0,1}=1/2M$
$a_0=a_1/2M$
$e_{4,7}=0$
Repeat the consolidate until the entire network becomes a module
Consolidate related communities
Direction: the greatest increase (or smallest decrease) in Q
$\Delta Q=e_{i,j}+e_{j,i}-2a_ia_j$
The network becomes a module
Choose a local maximum Q as the best structure
0
1
2
3
4
5
6
7
8
9

Fig. 4 The flow chart of the module detection algorithm

The shortest path algorithm has been widely used to determine regulatory paths in cancer networks[103,104] and then discover the key targets on the paths[105]. For example, Li et al.[106] first identified a set of six genes that can distinguish colorectal tumours from normal adjacent tissues using the maximum relevance minimum redundancy approach[107]. The method ranks genes according to their relevance to the class of samples concerned while considering the redundancy of genes. Those genes that had the best trade-off between the maximum relevance to the sample class and the minimum redundancy were considered “good” biomarkers. Then, the authors applied the shortest path algorithm among the six genes in a PPI network underlying cancer and identified 15 shortest paths between any two genes of the gene set. Last, they found 35 genes on the identified shortest paths and ranked them according to their betweenness[108]. The results showed that androgen receptor (AR), a ligand-dependent transcription factor, is ranked as the top gene, suggesting its involvement in colon carcinogenesis through regulating the proliferation and differentiation of tumour cells[109].

Additionally, Chen et al.[105] used a network-based biology analysis method, SAM (Significance Analysis of Microarrays)[110], to analyse omics data and identified 153 differentially methylated CpG sites and differentially expressed molecules, including 42 miRNAs and 1,373 protein-coding genes. The authors first used the differentially expressed genes from the STRING database[111] to construct a PPI network. Then, they searched all the shortest paths connecting dysfunctional genes to identify potential cancer driver genes. Next, they ranked the genes by a permutation test and their network properties, such as betweenness and interaction scores. The top-ranking genes at different levels (i.e., methylation level, miRNA level, mutation level, and mRNA level) were regarded as driver genes of lung adenocarcinoma. Among these cancer driver genes, some appeared to be top candidates at different levels, suggesting their multifaceted contribution to lung carcinogenesis.

Above all, the shortest path algorithms[100,101] can help us efficiently identify regulatory paths in networks, allowing us to identify potential genes that are proximate to known cancer genes and thereby important for tumorigenesis. However, due to the complexity of the disease, potential cancer genes are not always on the identified shortest paths[106], revealing the limitations of such algorithms. To resolve this issue, Lu et al.[112] proposed a random walk with restart algorithm method and identified 298 potential CRC-associated genes, which is more

Table 2. The formula to compute degree centrality, coreness centrality, betweenness centrality and eigenvector centrality

Node centrality	Formula	Description	Eq.
Degree centrality	$C_D(i) = d_i$	d_i is the degree of vertex i.	(3)
Coreness centrality	$C_C(i) = \sum_{j \in N(i)} ks(j)$	Vertex j belongs to the neighbours of vertex i, ks(j) is the k-shell index of vertex j.	(4)
Betweenness centrality	$C_B(i) = \sum_{j<k} g_{j,k}(i)/g_{j,k}$	$g_{j,k}$ is the number of all shortest paths between j and k, $g_{j,k}(i)$ is the number of shortest paths between j and k containing i.	(5)
Eigenvector centrality	$C_E(i) = \frac{1}{\lambda}\sum_{j \in G} a_{i,j}x_j$	if vertex i is linked to vertex j, $a_{i,j} = 1$, x_j is the degree of vertex j, λ is a constant.	(6)

effective and accurate than the shortest path algorithm proposed by Li et al.[106]. In particular, the computing efficacy of the shortest path algorithm could be compromised by large networks and their search strategies[112].

The module detection algorithm. Cancers usually result from disruption of interactions of key regulatory genes with their partners[81,113]. Module detection algorithms[114], one of network propagation algorithm, identify communities of cancer genes in complex networks[115] by analysing their topological structures (Fig. 4 and Algorithm 2). Here, we explain and illustrate the commonly used modularity maximization algorithm[116], which identifies network modules with the maximum modularity coefficients by Eq. 2.

$$Q = \frac{1}{2M}\sum_{i,j \in V}\left[A_{ij} - P_{ij}\right] \cdot \delta_{C_i,C_j} \tag{2}$$

where Q represents the modularity coefficient of an identified module, M is the total number of edges in the network, A_{ij} is the adjacency matrix, and P_{ij} represents the expected number of edges between nodes i and j. C_i or C_j represents the module to which node i or node j belongs. If i and j belong to the same module, $\delta_{C_i,C_j} = 1$; otherwise, $\delta_{C_i,C_j} = 0$. The identified modules are a group of genes that are supposed to have a similar biological function, such as promoting or inhibiting tumourigenesis.

Algorithm 2. Module detection algorithm.

```
Input: Network G
M ← the total number of edges in the Network
for each vertex i in Network:
    i ← a single module
    k_i ← degree of vertex i
    a_i ← k_i/2 M
for each edge in Network:
    if vertex i connects j:
        e_{i,j} ← 1/2 M
    else:
        e_{i,j} ← 0
do:
    $\Delta Q \leftarrow e_{i,j} + e_{j,i} - 2a_ia_j$
    consolidate related communities
    direction ← the greatest increase (or smallest decrease) in Q
until the entire network becomes a module
Output: the module with a local maximum Q
```

Currently, many researchers employ module detection algorithms to intelligently identify potential therapeutic targets for cancer[117–119]. For example, Ghiassian et al.[120] used the DIseAse MOdule Detection (DIAMOnD) method[121] to identify the local modules within the interconnected map of molecular components. They found that disease-related genes were significantly enriched in highly overlapping modules, which indicated that the predicted modules may help identify new anticancer targets. Of note, since the results of module detection algorithms depend mainly on network structures, the identified modules may vary for the same disease network with slightly different topology[85,117].

Since potential drug targets may exist in different network modules, we can make use of the correlation between modules to identify reliable cancer treatment targets[81]. Therefore, Wang et al.[122] proposed the seed connector algorithm (adding a few extra hidden nodes as much as possible to link disease proteins) by considering the interactions among cancer-associated proteins. First, this algorithm starts with known seed proteins and induces a loosely connected subnetwork consisting of only seed proteins. Second, Wang et al. sequentially select such proteins as seed connectors that maximally increase the size of the largest connected component of the subnetwork until there is no additional protein that can be selected as a seed connector. Finally, the cancer modules are pinpointed.

While these aforementioned algorithms[122–124] can intelligently identify meaningful functional modules from network topologies, it may be difficult to capture disease modules[125]. One possible reason is that disease proteins do not constitute particularly densely connected subgraphs but agglomerate in specific large regions of the network. For this reason, Tripathi et al.[126] considered analysing the patterns of connectivity in a disease module to be an effective way to understand the properties of disease modules.

The node centrality. Node centrality measures the importance of nodes and is suitable to intelligently locate key nodes with important biological functions for network biology[127].

Usually, we listed four types of node centrality as follows: (1) As the simplest form of network centrality, degree centrality is the number of nodes directly connected to the network[127,128]; (2) Coreness centrality considers both the degree of nodes and their positions in a network[129]; (3) Betweenness centrality of a node is the probability for the shortest path between two randomly chosen nodes to go through that node, and it determines the actor that controls information among other nodes by connecting paths[130]; (4) Eigenvector centrality[131] not only considers the number of edges and the position of nodes but also the impact of adjacent nodes on the interactive network.

Table 2 shows the formulas for node centrality computing. Figure 5(a–d) illustrates the above four types of node centrality, and Algorithm 3 presents the pseudocode to compute four types of node centrality.

Algorithm 3. The algorithm of degree centrality, coreness centrality, betweenness centrality and eigenvector centrality.

```
function1 Degree centrality:
  Input: Network G
  for each vertex i in Network:
    $d_i$ ← the number of ties that vertex i has
    $C_D(i)=d_i$
  Output: $C_D(i)$
function2 Coreness centrality:
  Input: Network G
  for each vertex i in Network:
    N(i) ← the set of the neighbours adjacent to vertex i
    for each vertex j in N(i):
      ks(j) ← the k-shell index of vertex j
    $C_C(i) \leftarrow C_C(i) + ks(j)$
  Output: $C_C(i)$
function3 Betweenness centrality:
  Input: Network G
  for each vertex i in Network:
    for each vertex j in Network:
      for each vertex k in Network:
        if j < k:
          $g_{j,k}$ ← number of all shortest paths between j and k
          $g_{j,k}(i)$ ← number of shortest paths between j and k containing i
          $C_B(i) \leftarrow C_B(i) + g_{j,k}(i)/g_{j,k}$
  Output: $C_B(i)$
function4 Eigenvector centrality:
  Input: Network G
  for each vertex i in Network:
      for each vertex j in Network:
        if vertex i is linked to vertex j:
          $a_{i,j}=1$
        else:
          $a_{i,j}=0$
        $x_j$ ← the degree of vertex j
        $C_E(i) \leftarrow C_E(i) + 1/\lambda \cdot a_{i,j}x_j$
  Output: $C_E(i)$
```

As described in Fig. 5(a) and Eq. 3, the degree centrality of node 2 is 3 (C_D (2) = 3) because node 2 interacts with nodes 0, 1, and 3. We demonstrated that highly connected nodes or hubs are more likely to be essential[127]. Because the more direct connections a node has, the greater the impact that the node can exert on the network[132], we can utilize the degree centrality of nodes to identify cancer therapeutic targets.

For example, Zhang et al.[133] predicted that hypoxia inducible factor-1α (HIF-1α) and prolyl 4-hydroxylase beta polypeptide (P4HB) may be considered potential biomarkers of gastric cancer by constructing a PPI network. Nevertheless, not only Jalili et al.[130] suggested that high connectivity does not necessarily imply its essentiality, but also Kitsak et al.[129] argued that the location of nodes is more significant than the immediate neighbours to evaluate its spreading influence because degree centrality considers only direct interactions of a node but not its impact on other nodes, resulting in low accuracy for target prediction compared to other methods such as coreness centrality[134].

As shown in Fig. 5(b) and Eq. 4, the coreness centrality of node 3 is 8 (C_C (3) = 8) because the neighbours adjacent to the labelled vertex (3) are vertex (1), vertex (2), vertex (4) and vertex (5), and these four nodes belong to a 2-shell. Coreness centrality is an advanced form of node centrality because it considers both the degree of nodes and their positions in a network to quantify the importance of nodes in a network[129]. A node with a greater coreness means that the node is located in a more central place and is much more influential in network propagation than the nodes with high-degree but less coreness[129]. Among them, the most classic method to calculate the coreness centrality of network nodes is the k-core decomposition method[135], which decomposes the network iteratively according to the remaining degree of the nodes.

For instance, Li et al.[136] employed the k-core decomposition method to obtain the coreness of the PPI network. Subsequently, the targets were screened for topological importance. Then, the major hubs in the hub interaction network were determined, and a total of 62 major hubs were identified, including 11 indirubin (EGFR, JAK2, ERBB2, CHUK, CDK5, KIF11, DRD2, CDK3, HTR1A, JAK3 and TYK2) and derivative targets and 51 differentially expressed genes (DEGs) for imatinib resistance. These 11 major hubs were closely related to DEGs that were resistant to imatinib. Indirubin and its derivatives may inhibit imatinib resistance through the regulation of these genes to treat chronic myeloid leukaemia (CML).

Described by Fig. 5(c) and Eq. 5, the betweenness centrality of node 1 is 3.5 (C_B (1) = 3.5) because there are four node pairs contributing to node one ($g_{0,2}(1)/g_{0,2}(1)$ = 1, $g_{0,3}(1)/g_{0,3}$ = 1, $g_{0,4}(1)$ / $g_{0,4}$ = 1, and $g_{2,3}(1)/g_{2,3}$ =0.5). Betweenness centrality is based upon the frequency with which a node lies between the shortest path of all other possible pairs of nodes within a network and identifies the gatekeepers that control communication of nodes in the network[130].

For example, Taylor et al.[137] used betweenness centrality analysis to identify intermodular hub proteins and intramodular hub proteins in the breast cancer network. The identified proteins may serve as an indicator of breast cancer prognosis. Moreover, Raman et al.[138] computed degree, betweenness, and closeness indices in PPI networks for 20 organisms and showed that the degree and betweenness centralities of nodes correlate with their lethality in many organisms.

As described in Fig. 5(d) and Eq. 6, the eigenvector centrality of node 1 is 3 (C_E (1) = 3) because node 1 is connected to nodes 0, 2 and 3 ($a_{1,0}$, $a_{1,2}$ and $a_{1,3}$ equal 1, respectively), and the degree of x_0, x_2 and x_3 equals 1, respectively. Eigenvector centrality considers not only the number of edges and the position of nodes but also the impact of adjacent nodes on a network.

For example, Mallik et al.[139] first identified differentially expressed and methylated genes in uterine leiomyoma tumours and then found TFs and miRNAs that regulate the expression of these genes. Subsequently, they reconstructed a network that comprised the genes, TFs, and miRNAs and then used eigenvector centrality to identify potential biomarkers. They specified that PTGS2 and TACSTD2 are potential novel biomarkers, since both genes are downregulated and hypermethylated in the tumour.

Moreover, several researchers have attempted to integrate more than one centrality index to increase the efficiency of the node centrality algorithm. For instance, Chen et al.[140] used the differentially expressed proteins of prostate cancer (PC) to construct a PPI network. Then, they integrated the connectivity degree, betweenness centrality, and closeness centrality of nodes to evaluate critical nodes to identify the core module of the PPI network. Finally, they identified SLC2A4 and TUBB2C as important proteins regulating the pathogenesis of cancer, suggesting the proteins involved in biological processes and pathways as potential targets for PC diagnosis and treatment. In addition, Aamri et al.[141] constructed a gene-gene-interaction network for the entire human genome and then applied betweenness, closeness, eigenvector, and degree centrality metrics to rank the central genes of the network to identify

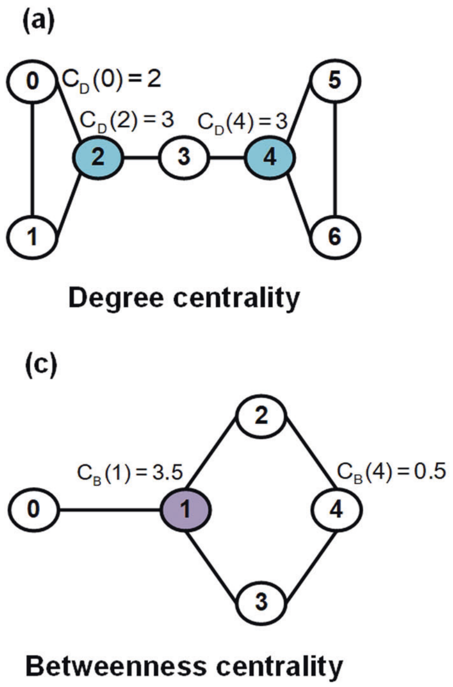

Fig. 5 **Four types of node centralities of biological networks.** (a) Degree centrality; (b) Coreness centrality; (c) Betweenness centrality; (d) Eigenvector centrality

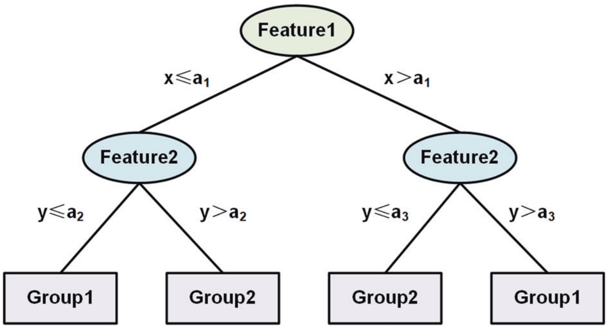

Fig. 6 An illustration of a simple decision tree model

possible cancer-related genes. The results showed that the average precision for identifying breast, prostate, and lung cancer genes varied between 80–100%.

Although highly connected nodes in the network architecture are essential, recent studies point out that integrating the prior knowledge of cancer into centrality indices can accurately identify anticancer targets[130]. For this reason, Jiang et al.[142] developed a network-based biology analysis method, named NEST, which predicts essential proteins according to the expression levels of their interacting partners in a network. Additionally, the results showed that NEST significantly outperformed the classic centralities on gene essentiality prediction and functional screen result enhancement.

Machine learning-based biology analysis algorithms

Machine learning (ML) algorithm is a subset of AI algorithms that can learn from data, therefore removing the need for explicit instructions on how to do certain tasks[15]. The key to identify therapeutic targets and discover drugs using ML-based biology analysis is to make use of network features in biological networks. The network features include the topological features (such as node centrality, interaction, local structure, subgraph, network propagation results, and network-based structure similarities) and the biological information that is embedded in network nodes (such as the gene expression profile, gene mutation frequency, and gene functional annotation).

Here, we introduce two classical ML-based algorithms: one is the decision tree algorithm, which selects significant topological features for cancer; the other is deep learning, which uses the network features to identify cancer targets and discover drugs.

The decision tree algorithm. A decision tree is a supervised classification algorithm[143] with three steps: feature selection, decision tree generation, and decision tree pruning[86–88]. Figure 6 shows how to classify a set of samples into two groups using the decision tree algorithm.

In the network-based biology analysis, network topology features[88] are usually integrated into a decision tree to classify gene-phenotype associations for cancers[144–146] to select significant topological features for cancer.

For instance, Ramadan et al.[147] extracted thirteen network topological features (Table 3) from a publicly available gene co-expression network and a PPI network of breast cancer. Then, to assess the significance of topological measurements associated with breast cancer, they used Decision Tree Bagger[156] to classify breast cancer gene-phenotype associations. The importance of each topological measure was then evaluated using a score that combines the accuracy of breast cancer classification and the Gini index[148] (Table 3). The computed scores of the top five identified features (i.e., structural holes, node degree, node coreness, *k*-Step Markov and subgraph) outperformed the others, and they were selected as key features for the classification of breast cancer phenotype-gene associations.

Although the decision tree algorithm can help us select key network features, it usually has the overfitting problem when too many features exist in the network[157], which significantly decreases the classification and prediction on independent testing[157].

Table 3. Thirteen network topological features for decision tree classification[147]. The score is a combination of the classification accuracy and the Gini index[148]

Topological measures	Concept	Score
Structural holes[149]	Rank nodes by their connectivity and lack of redundancy	13.37
Node degree	The number of connections of a node	13.36
Node coreness	Considers both the degree of nodes and their positions in a network	12.05
k-Step Markov[150]	The probability that a random walk of length k makes the system reach a certain vertex	10.47
Subgraph[151]	The number of times a given vertex participates in different connected subgraphs of a network	10.36
Within–module z-score[152]	Measure how nodes are related.	8.88
Katz status index[153]	Rank a vertex as highly important if many nodes are connected to it.	8.64
Closeness	The average length of the shortest path between nodes	8.18
Proximity prestige	The average shortest path length of a node	8.12
Eigenvector centrality	The influence of directly adjacent nodes on central node	8.09
Betweenness	A node acts as a bridge along the shortest path between two other nodes	7.93
Bary centre score[154]	Rank the nodes by the total shortest path of the vertex	5.70
Clustering coefficient[155]	Measure the degree of cohesiveness	0.15

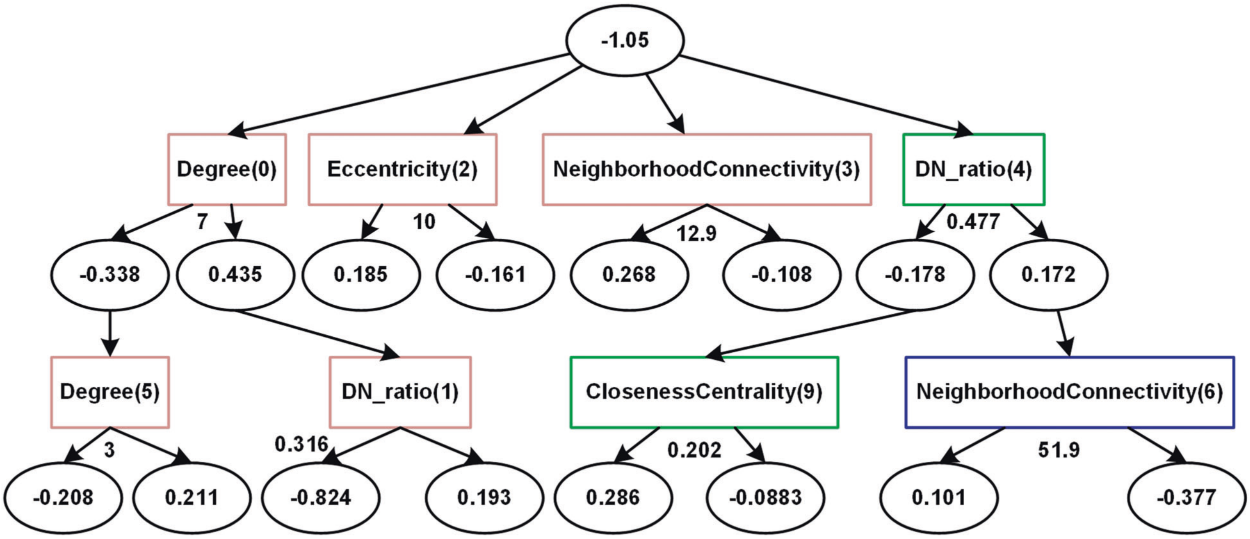

Fig. 7 An example of an ADTree model. The root nodes indicate the ratio between positive and negative class examples. The numbers in parentheses within each decision node (rectangles) indicate the order in which the rule was found. The amount of node conservation between each of the trees is indicated by the colour of the box. Ovals (prediction nodes) contain the value for the weighted vote. The numbers next to the arrows correspond to the threshold for the prediction

At present, there are two commonly used methods to resolve overfitting caused by the decision tree algorithm. One method is using dimension reduction[157] and pruning strategy[86] to improve the classification accuracy by feature reduction; the other is employing the random forest algorithm[158], an ensemble algorithm with multiple decision trees. The random forest algorithm adopts a bagging strategy, which has higher accuracy and reliability than the classical decision tree algorithm[159].

For example, Toth et al.[160] used the random forest algorithm to predict the aggressive behaviour of prostate cancer. Their methylation-based classifier demonstrated excellent performance in discriminating prognosis subgroups of the test set (Kaplan-Meier survival analyses with log-rank p value < 0.0001) with an AUC value of 0.95[161] for the sensitivity analysis. Finally, the experimental verification showed that the loss of ZIC2 protein expression was associated with poor prognosis and correlated with a significantly shorter time to biochemical recurrence.

In addition to the overfitting problem, it is difficult for decision trees to visualize the complicated classification procedure[146]. Recently, the alternating decision tree (ADTree)[162] has made the classification procedure intuitive and easy to understand by adding an intuitive graphical model, and the algorithm builds decision trees over a user-defined number of iterations using confidence-rated boosting, so it returns both a class label and a score that measures confidence in the classification, as shown in Fig. 7 and Algorithm 4.

For example, Carson et al.[146] used ADTree to classify proteins in a breast cancer network. As indicated in Fig. 7, the most effective attributes to distinguish disease and non-disease proteins are node degree, disease neighbour ratio, eccentricity, and neighbourhood connectivity, which was proven by Hao et al.[163] and Zhang et al.[164].

Table 4. Commonly used neural networks in ML-based biology analysis

Model	Characteristic	Application scenarios
Non-graph Neural Network		
DNN	Deep neural network (DNN), also called multi-layer perceptron, is a neural network with multi-layer hidden layer.	[169]
CNN	Convolutional neural network (CNN) obtains local information between input data by convolution.	[170]
Graph-based Neural Network		
GCN	Graph convolutional network (GCN) applied cconvolution in networks to obtain local information between nodes and neighbour nodes.	[171]
GAE	Graph autoencoder (GAE) uses autoencoder to extract the embedded features of the network.	[172]
GAN	Graph attention network (GAN) uses attention mechanism instead of convolution to obtain local or global information between nodes.	[173]
DeepWalk	DeepWalk is a network embedding model, which can represent the attributes of graph nodes as low dimensional and dense eigenvectors.	[174]

Algorithm 4. The algorithm of ADTree model[165]

```
Input: labelled dataset
root node ← the bias in the dataset
for each decision node in the tree:
  $a_i$ ← attribute value
  $t_i$ ← threshold
for each decision node in the tree
  if (the decision node has a parent node):
    if $a_i \geq t_i$:
      return the score of the prediction node for the left path
    else:
      return the score of the prediction node for the right path
  else:
    return 0
s ← the sum of all scores acquired
if $s > 0$:
  Output: the positive class
else:
  Output: the negative class
```

Although the decision tree, random forest and ADTree[86–88,158] demonstrate the tendency to identify such proteins that are well annotated and studied for cancer, these methods are subject to producing local optimal solutions. Therefore, Chen et al.[143] proposed using the decision tree classifier based on particle swarm optimization[166] to avoid falling into the trap of local minima by adding randomness to optimize the number of features and detection accuracy of cancer treatment targets. Furthermore, the gradient boosting decision tree[167] is a very flexible and scalable method to classify network nodes for future study.

The deep learning algorithms. Deep learning is a subfield of machine learning, and the origin of neural networks sets the stage for the emergence of deep learning models[168]. Deep learning model is a neural network composed of complex structures and nonlinear transformations[90,91] that attempts to model high-level abstractions of data using multilayer neurons. Through training and iteratively updating its hyperparameters (Eq. 7), the initial low-level feature representation (such as topological features and biological information) of samples is transformed into the high-level representation that shows the distinction between samples. The strength of deep learning is its ability to detect complex patterns in data, making it suitable to interrogate the biological networks that consist of complex, interdependent relationships among genes.

$$W_k \rightarrow W_{k+1} = W_k - \eta \frac{\partial C}{\partial W_k} \tag{7}$$

W, *k*, and *C* are the weight, iteration, learning rate, and loss function, respectively.

Currently, there are many neural network models and complex functions for ML-based biology analysis. In this paper, we only present several commonly used neural networks (Table 4). Benefiting from the strong ability of neural networks in mining complex information on links or nodes, deep learning is a suitable method to identify potential cancer targets and discover drugs for cancer treatment in complex biological networks[175]. For example, Selvaraj et al.[176] searched for therapeutic targets for lung adenocarcinoma in a network of protein-protein and protein-drug interactions and employed a neural network to identify candidate drugs, where phosphothreonine is predicted via molecular dynamics simulations to target the hub node MAPK1 in the network.

Currently, artificial intelligence biology analysis has benefited from the utilization of graph-based neural networks instead of commonly used non-graph neural networks such as CNN[170] or DNN[169], because graph-based neural networks can take the biological network structure as the input directly, learn an embedding that contains information about the neighbourhood of a target node in a graph, and analyse the biological network with neural networks technology. Figure 8 illustrates the basic flowchart of graph-based neural networks for the investigation of different properties of biological networks.

There are two advantages in using graph-based neural networks to identify cancer targets or discover drugs from biological networks.

1. Feature representation. Graph embedding[177] is the core method to extract features in graph-based neural networks, which represent network nodes as a low-dimensional vector representation, preserving both network topology and node content information[178]. For example, Li et al[174] proposed a similarity-based miRNA-disease prediction method that used DeepWalk, a graph embedding algorithm, to compute the topological similarities between two diseases nodes. The model extracts the disease node features in the disease-disease network based on the random walk algorithm, and significantly enhances the prediction performance by utilizing global network association information. For diseases nodes with similar features, if one of the diseases is

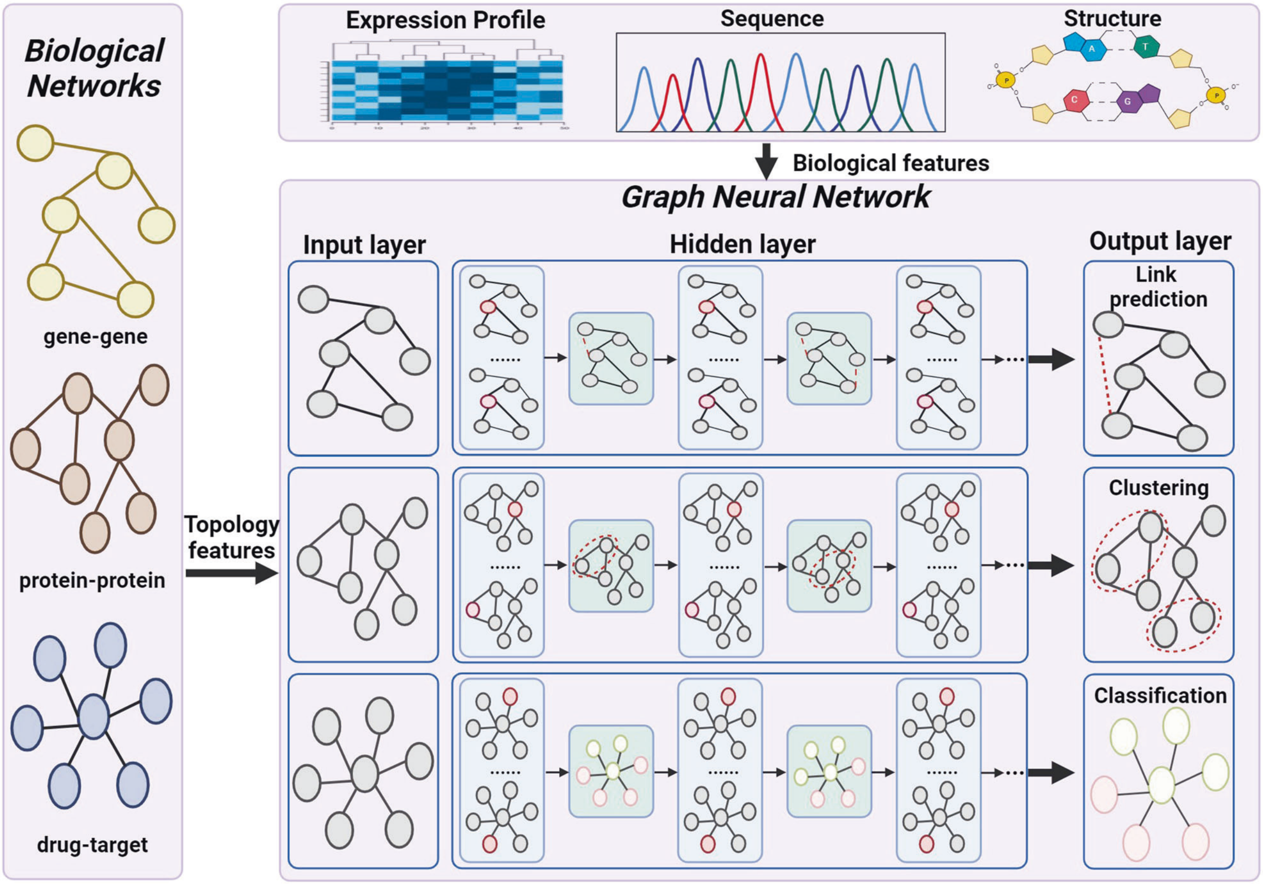

Fig. 8 The illustration of graph-based neural networks for ML-based biology analysis. The graph-based neural networks take the topology of the biological networks data (such as gene-gene networks, protein-protein networks and drug-target networks) as input data. And then, the graph-based neural network realizes the functions of link prediction, classification and clustering by analyzing the biological information in the network topology. (Created with BioRender.com)

associated with miRNA, the other is predicted to be associated with the miRNA.

In addition, Zheng et al.[179] proposed an attention-based graph neural networks (attention mechanism assigns different weight parameters to different targets through learning, so as to consider the importance of key targets locally and globally[180]) to learn the graph embedding feature (association scores) from piRNA-disease association network. The results showed that the predicted scores of piRNA-disease associations are positively correlated with the association probability between a piRNA and a disease, suggesting that piRNAs with closer distances to tumour genes in the network are more likely to be therapeutic targets of cancer.

2. Feature integration, which integrates the heterogeneous, noisy, nonlinear-related biological network information (such as node similarity, node interactions, upstream and downstream relationships) multi-views (such as drug molecular structures and drugs' indications)[181]. For example, Ma et al.[172] proposed a novel graph autoencoders model (GAE) to learn accurate and interpretable drug similarity measures from multiple types of drug properties. The GAE uses attention mechanism[180] to integrate multi-view (multiple types of drug properties) from drug-drug interactions network and determines the weights for each view with respect to the similarity measure tasks for better explaining the contribution of drug properties to drug similarity. Due to the ability to integrate network data from multi-views and autoencoder structures, GAE can resist the noise interference in the data. Thus, graph-based neural networks are more robust and reliable in most application scenarios[182].

Overall, deep learning can comprehensively explore features such as node degree, edge length, and module in biological networks[83–85,183] to provide an accurate prediction for drug targets of cancer through artificial intelligence of multiomics data in complex biology networks[184]. However, there are still two key issues to be addressed. One is the interpretability of the models, which is critical for clinical adoption[185]. The other is how to demonstrate the generalizability of the approach[185] and validate these approaches in the context of multi-institutional datasets. Therefore, these issues are actively being tackled from model interpretation, extraction of biological insights[186] and model reproducibility[187].

THE ARTIFICIAL INTELLIGENCE BIOLOGY ANALYSIS FOR BIOMEDICAL APPLICATIONS

Because the wide and easy accessibility of high-throughput data in oncology has provided the basis for developing novel artificial intelligence methods and validating their capability to identify therapeutic targets, this section will focus on reviewing the biomedical applications from four perspectives. First, we present the artificial intelligence applications to identify novel anticancer targets. Second, we present the artificial intelligence applications

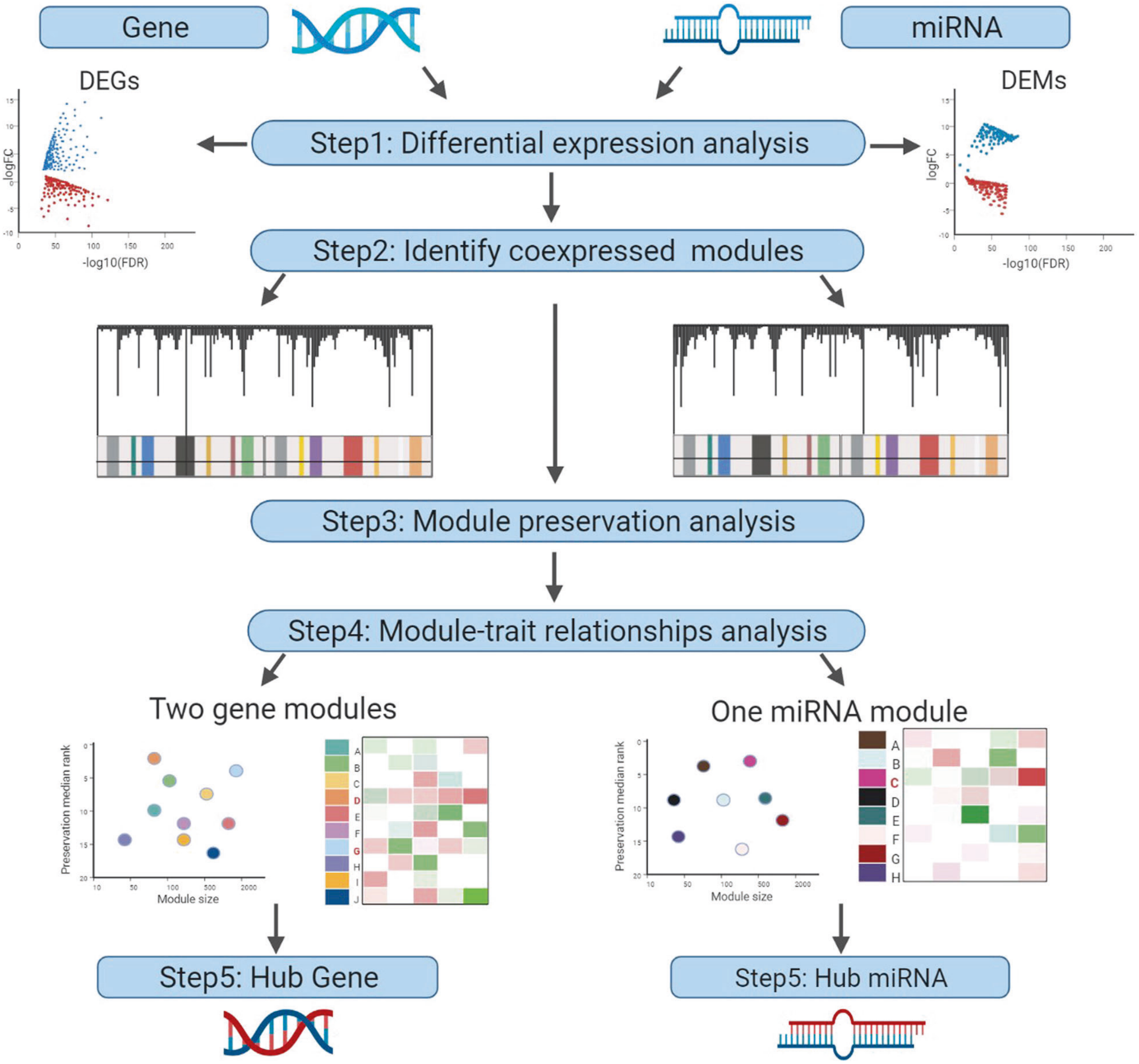
Gene
miRNA
DEGs
DEMs
logFC
-log10(FDR)
Step1: Differential expression analysis
Step2: Identify coexpressed modules
Step3: Module preservation analysis
Step4: Module-trait relationships analysis
Two gene modules
One miRNA module
Preservation median rank
Module size
Step5: Hub Gene
Step5: Hub miRNA

Fig. 9 The workflow to identify novel anticancer targets by network-based. (Created with BioRender.com)

to evaluate the druggability of potential target genes. Third, we show the artificial intelligence applications for drug discovery. Fourth, we show the artificial intelligence applications for drug property prediction.

Identification of novel anticancer targets

Artificial intelligence biology analysis applications[188] usually use omics data to build networks and identify co-expression modules of genes, proteins, metabolites, critical pathways between molecules, and key molecules in biological networks[189]. This study will introduce these applications from two perspectives: one is network-based biology analysis applications, and the other is ML-based biology analysis applications.

Network-based artificial intelligence for identifying novel anticancer targets. Network-based biology analysis applications firstly reconstruct networks by computing differential expressions of molecules and their correlations[190–193]. Then, gene set enrichment analysis are performed to identify network modules with different biological functions[194]. Finally, the identified network modules are used to discover key genes that are potential therapeutic targets (or biomarkers) for cancer. Here, we show the key target identification procedure by network-based biology analysis applications as follows.

WGCNA[195] is a commonly used network-based biology analysis application that uses various gene expression matrices as input. Then, WGCNA outputs different gene network modules and the core genes in the biological network. For example, Zhou et al.[196] used WGCNA to analyse colorectal cancer data from TCGA (Fig. 9), which demonstrated that 11 hub genes and 5 hub miRNAs have predictive power for the prognosis of colorectal cancer patients by the following steps.

In Step 1, the correlation between all pairs of genes and miRNAs by differential gene expression analysis was calculated, and two similarity matrices were constructed. In Step 2, the adjacency matrix, which comes from similarity matrices, is transformed into a topological overlap matrix (TOM) by using TOM similarity, and then the coexpressed gene and miRNA modules are identified by using dynamic tree cutting[197]. In Step 3, after module preservation analysis, six gene modules were found to have strong stability, and one miRNA module was found to have low stability. In Step 4, they performed module-trait relationship analysis to further validate the module–clinical trait relationships, and two pathological stage-related gene modules and one pathological stage-related miRNA module were identified. In Step 5, hub genes and hub miRNAs were identified by calculating the module membership and gene significance.

Though network-based biology analysis methods are useful in identifying anticancer targets, they have some limitations, such as they cannot effectively handle multiomics data, leading to high false-positive rates of identified targets[42]. Developing comprehensive network-based biology analysis applications may resolve the problems and increase the precision for predicting cancer biomarkers[198].

For example, Lai et al.[199] deployed an integrated approach that combined network-based algorithms and RNA sequencing data to delineate miRNA-based strategies that enhanced DC (dendritic cell)-elicited immune responses. First, the authors performed RNA sequencing to obtain the protein-coding genes and miRNAs in relation to standard DCs. Then, they analysed miRNA-gene interactions at the pathway level and reconstructed regulatory networks underlying the immunological functions of DCs. Finally, they performed network-based prioritization of miRNAs by combining their expression profiles and strength of association with other protein-coding genes. Their analysis identified dozens of promising miRNA candidates, of which miR-15a and miR-16 are the most promising ones for increasing the immunogenic potency of DCs and therefore improving DC-based immunotherapy against cancer.

In summary, we consider that an increasing number of network-based biology analysis applications will be developed for novel anticancer targets identification in the distant future.

ML-based artificial intelligence for identifying novel anticancer targets. ML-based biology network analysis applications are applied to interrogate the large, complex data and thus iden tifying reliable potential novel targets as effective treatments of human diseases[200]. These ML-based biology analysis applications for novel anticancer targets identification consist of classification[201], clustering[202], neural networks[203,204], and so on[205]. Here, due to the limit space of the review, we only focus on the ML-based biology network analysis applications for classifications and graph-based neural networks.

ML-based biology network analysis applications for classifications identify key targets by determining the key factors of classifications[206]. It considers specific biomarkers (such as gene or protein nodes) of the defined classes as key targets[206]. Recently, the classification-based applications and molecular profiling[207], use genome-wide gene transcription profiles, protein expression profiles and/or mutational landscapes to make a more accurate classification of tumor subtypes and identify biomarkers for specific tumor types.

For example, Sinkala et al,[208] applied classification analysis on networks to reveal subtypes of pancreatic cancer and their molecular characteristics. Firstly, the authors employed K-means clustering to the reverse phase protein array (RPPA), determined proteomics data with 45 high-purity pancreatic cancer samples, and then identified two clusters of samples.

Secondly, they compared their clustering results to other subtypes that have been reported in the literature for various other molecular data types (such as DNA methylation status, protein expression levels and expression levels of mRNAs and miRNAs), and then applied the similarity network fusion (SNF) to identify two-cluster and three-cluster solutions comprised 25 and 20 tumors. The SNF method solves the disparate clustering problem by constructing similarity networks of samples for each available molecular data type and then efficiently fuses these into one network that represents clustering based on all the underlying data.

Thirdly, they applied proteomics-based signaling pathway analysis to distinguish disease subtypes and found that, for tumors of the two major pancreatic cancer subtypes, oncogenesis may be primarily driven by perturbation in either SMAD4 or mTOR signaling pathways. Furthermore, they performed gene set enrichment analysis using the Gene Ontology database[52] and found that pancreatic cancer subtypes classified by mRNA expression levels and DNA methylation statuses show differences in molecular functions in terms of mRNA.

Finally, given that different types of molecular data yield different patterns of tumor clustering, they attempted to identify a list of biomarkers that can differentiate the two tumor subtypes. Using neighborhood component analysis, they identified biomarker sets comprising 50 mRNAs, 49 methylated genes, 14 proteins, and 20 miRNAs. Subsequently, they separately applied hierarchical clustering using each type of the molecular data and successfully reproduced the two pancreatic cancer subtypes.

For graph-based neural networks, they take advantage of not only making use of the correlation among samples described by similar networks, but also message passing between targets and neighbors to improve the accuracy of targets identification[209].

For example, to the best of our knowledge, the MOGONET proposed by Wang et al.[203] is the first to make use of both graph convolution networks (GCNs) and cross-omics relationships in the label space for effective multiomics integration in biomedical data classification tasks. The specific process is as follows:

Firstly, they constructed a weighted sample similarity network for each type of omics data using cosine similarity. Taking both the omics features and the corresponding similarity network as the input, a GCN is trained for each type of omics data to predict class labels.

Secondly, the predictions generated by each omics data-specific GCN are further utilized to construct a new tensor, named cross-omics discovery tensor, which can reflect the cross-omics label correlations.

Finally, the cross-omics discovery tensor is forwarded to VCDN (view correlation discovery network) to explore the latent correlations across different omics data for final label prediction. Because the importance of a feature to the classification task can be measured by the performance decrease after removing individual features. Therefore, they used this method on the test data set to quantify and rank the contribution of each feature of different omics data to the prediction. Using the method, they identified top-ranking features as biomarkers for breast cancer.

In addition, Xuan et al.[204] proposed a novel method based on the graph convolutional network and convolutional neural network (GCNLDA) to infer disease-related lncRNA candidates. First, they developed a network that is comprised of lncRNA, disease, and miRNA nodes. Then, they developed an embedding matrix of lncRNA-disease node pairs with respect to the biological premises. Then, they employed a convolutional neural network to explore various connections related to lncRNA-disease on node pair embedding. Finally, they learned the local network representations of lncRNA-disease pairs by deeply integrating the graph convolution autoencoder into topological lncRNA-disease-miRNA heterogeneous networks. Cross-validation confirmed that GCNLDA outperforms other state-of-the-art methods in terms of both AUC and AUPR[161]. Case studies[204] on stomach cancer, osteosarcoma and lung cancer confirmed that GCNLDA effectively discovered potential lncRNA-disease associations. Therefore, GCNLDA is becoming an effective tool to screen reliable candidates for lncRNA-disease association validation with the help of biological experiments.

In summary, we consider that an increasing number of ML-based biology analysis applications will be developed to identify novel anticancer targets with the development of deep learning in the future.

Evaluation of the druggability of potential targets

Druggability is a concept that assesses whether a drug can bind to a protein to alter its activity[3,4]. The human proteome has

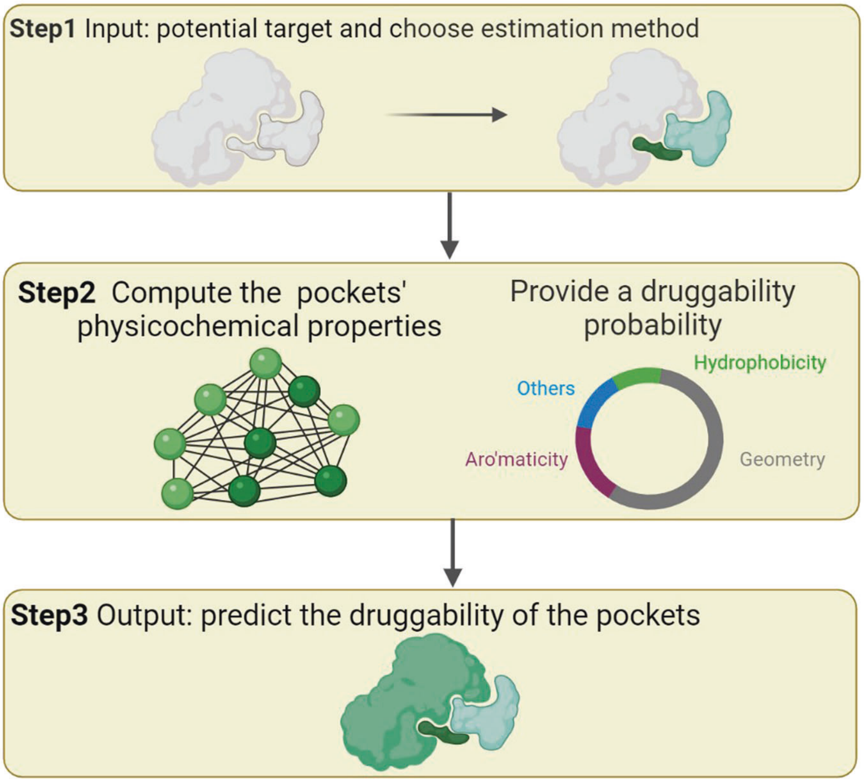
Step1 Input: potential target and choose estimation method
Step2 Compute the pockets' physicochemical properties
Provide a druggability probability
Hydrophobicity
Others
Aro'maticity
Geometry
Step3 Output: predict the druggability of the pockets

Fig. 10 The workflow to evaluate the druggability of potential target proteins. (Created with BioRender.com)

approximately 6,000 to 8,000 potential pharmacological targets, but only a small fraction can be targeted by drugs[7,210]. Therefore, it is important for us to evaluate druggability after finding novel anticancer targets. This study will introduce these applications from two perspectives: one is network-based biology analysis applications, and the other is ML-based biology analysis applications.

Network-based artificial intelligence for evaluating the druggability of potential targets. The druggability evaluating approach requires a long development cycle and high financial cost for the 3D structures of protein analysis[211], while network-based biology analysis application provides an alternative methods to accelerate the evaluation procedure for the druggability of potential targets[212].

Described by Fig. 10, PockDrug is a novel web server that is employed to predict pocket druggability on proteins and queried for a protein or a set of proteins[213]. For example, Yang et al.[214] constructed a protein–protein interaction network for thyroid cancer and identified three key targets, HEY2, TNIK, and LRP4. Then, they used PockDrug to predict whether HEY2, TNIK, or LRP4 have targetable pockets for drugs in the following three steps.

In Step 1, they inputted the potential target and located pocket estimation methods. In Step 2, they predicted the druggability of the pockets by computing the physicochemical properties of the target pockets. In Step 3, they screened three hub genes, HEY2, TNIK, and LRP4. Based on the predictions, TNIK, which has 8 out of 538 residues, has an average druggability probability greater than 0.5 and thus was considered a druggable pocket for thyroid cancer.

In short, with the in-depth study of protein pocket, an increasing number of network-based biology analysis applications are developed to accurately evaluate the druggability of anticancer targets, providing reliable druggable targets for cancer treatment.

ML-based artificial intelligence for evaluating the druggability of potential targets. These ML-based biology analysis applications for evaluating the druggability of potential targets consist of protein structure modeling and drug-target affinity analysis. Previously, traditional analysis of protein structure modeling required considerable time and financial cost[211], which greatly limited the traditional application of PockDrug since it is heavily dependent on an accurate 3D protein structure. Recent ML-based biology analysis applications have focused on developing methods to predict the 3D structure of a protein from its genetic sequence, also known as the protein folding problem. The cutting-edge ML-based modelling method[215–217] can generate 3D protein structures with high accuracy and efficiency, which makes it possible for PockDrug to be widely used.

For example, Yang et al.[218] developed the trRosetta algorithm, which fast and accurately predicts protein structures based on energy minimizations with restrained trRosetta. They employ a deep residual neural network to predict the restrained trRosetta, which consists of inter-residue distance and orientation distributions. Since trRosetta outperforms all previously protein modelling methods in benchmark tests on CASP13-[219] and CAMEO-[220] derived sets, it turns out that trRosetta can accurately predict protein structure. Furthermore, Senior et al.[221] developed Alphafold to predict protein structures from amino acid sequences. First, Alphafold predicts the distances between pairs of residues by training a neural network to analyse the covariation of homologous sequences. Then, Alphafold constructs a potential mean force that accurately describes the shape of a protein. Finally, Alphafold optimizes the protein structure by a gradient descent algorithm. Because AlphaFold can predict protein structure with high accuracy even for such sequences with fewer homologous sequences, we consider that AlphaFold makes great progress in protein-structure prediction.

ML-based biology analysis applications for drug-target affinity (DTA) analysis application estimates the interaction strength of novel drug–target pairs based on previous studies to evaluate the druggability of targets[222].

Compared with other methods, such as molecular docking[223] and collaborative filtering[224], graph-based neural networks are more effective in DTA prediction, because graph-based models facilitate the learning by considering both drug structure and drug-target interaction information instead of representing the drugs as string, as string sequences may lose the structural information of the molecule and may impair the predictive power of models[225].

For example, Nguyen et al.[225] is the first to use GNN for predicting DTA. The authors proposed GraphDTA, a new neural network model for regression tasks, which takes the drug-target pair as the input and outputs the continuous measurement of the binding affinity of the pair.

In detail, for the input drug-target pair, the protein targets are represented as sequence information instead of the molecular diagram of tertiary structure. While the drug compounds are represented as network graphs of atomic interaction, where each node is an eigenvector that represents five kinds of information: the atom symbol, the number of adjacent atoms, the number of adjacent hydrogens, the implicit value of the atom, and whether the atom is in an aromatic structure. For the output, GraphDTA combined the drug-target pair feature information to predict the continuous measurement of the binding affinity of the drug-target pair.

Through a multivariable statistical analysis of GraphDTA's output data from hidden layers, the authors have two conclusions. One is to identify the correlations between hidden node activations and domain-specific drug annotations, such as the number of aliphatic hydroxyl groups, which suggests that the graph neural network can automatically assign importance to well-defined chemical features without any prior knowledge. The other is that the model makes it easier to extract features from drugs with obvious molecular structure patterns to achieve high-precision predictions. Especially, drugs that do not have an obvious molecular structure pattern are more difficult to predict.

In short, with the development of deep learning, an increasing number of ML-based biology analysis applications can quickly and accurately evaluate the druggability of anticancer targets, providing reliable druggable targets for cancer treatment and reducing the time and financial costs of experiments.

Drug discovery

After evaluating the druggability of potential targets, it is essential to discover the drugs that interact with the potential therapeutic targets. As complex or concomitant diseases may usually require treatment with multiple drugs, but the use of multiple drugs will increase the risk of side effects[200], it is very essential for drug discovery to predict the interactions between drug-target and drug-drug.

This study will introduce these applications from two perspectives as the above section: one is network-based biology analysis applications, and the other is ML-based biology analysis applications.

Network-based artificial intelligence for drug discovery. These network-based analysis applications for drug discovery consist of drug screening and drug repurposing. Drug screening is a process that potential drugs are identified and optimized before selecting a candidate drug to progress to clinical trials[226]. Since screening drugs through biological experiment is quite laborious, expensive, and time-consuming[226], network-based biology analysis application becomes an alternative way for efficiently drugs screening.

Identifying drug-target interactions (DTIs) is crucial for drug screening. Especially, novel DTIs can be employed to look for the novel anticancer drugs with known targets[227].

The network-based biology analysis applications for DTI prediction are usually based on guilt-by-association principle that a protein may be a target for a drug if many of the protein's neighbors in the interaction network are targets of the drug[228]. Based on this principle, we classify the network-based biology analysis applications for predicting DTI into two categories.

One is 'top-down', which is from observable characteristics, such as side-effects or the diseases treated by a drug, to the interaction. For example, Campillos et al.[229] used the physiological effect information from side effect similarity networks between entities for DTI prediction to predict whether two molecules could interact.

The other is 'bottom-up', which is from molecular features, such as protein structure, to interactions. For example, Feng et al.[230] and Lee et al.[231] predicted DTI based on the proteins in protein-protein interaction networks with similar property features that may interact with the same drug.

Drug repurposing, also known as drug repositioning, is another drug discovery application. It refers to a method that identifies new indications for approved drugs or drug candidates which have failed in the development phase[232]. Compared to the drug screening process, since drug repurposing can significantly reduce the drug development period and costs[233], it is a better application to discover anticancer drugs.

The network-based biology analysis applications are efficient to carry out drug repurposing analysis, because the constructed drug similarity networks contain the similarity, interaction or linkages between drugs, diseases, and targets. Here, we introduce four major network-based biology analysis applications of drug repurposing[234–241] as follows.

The first network-based biology analysis application of drug repurposing quantifies the similarities or relationships for known drug-disease associations, and then uses regression models or statistical models to predict novel drug-disease associations[234,235]. For example, Cheng et al.[242] presented a network-based drug repurposing tool, which can accurately predicts drug responses in cancer cell lines by integrating human protein-protein interactome with transcriptome profiles, whole-exome sequencing, drug-target interactions and drug-induced microarray data.

The second network-based biology analysis application of drug repurposing infers new indications of drugs through analyzing information flow or performing random walks on drug-disease association networks[236–238]. For example, Luo et al.[243] proposed a novel random walk method to measure the similarity of drugs and diseases respectively by the drugs properties and diseases properties, so as to predict potential indications of drugs.

The third network-based biology analysis application of drug repurposing, named individualized Network-based Co-Mutation, quantifies putative genetic interactions in cancer and it can be used to identify candidate therapeutic pathways for cancer[239]. For example, Cheng et al.[244] used the approach to identify potential targets or new indications of existing cancer drugs that directly target significantly mutated genes or their neighbor genes in the human PPI interaction network.

The fourth network-based biology analysis application of drug repurposing can be realized directly through calculating the adjacency matrix of drug and disease network[240,241]. Based on this method, Luo et al.[245] utilized the matrix completion algorithm to fills out the unknown entries in the drug–disease matrix by constructing a low-rank matrix approximation. New drug–disease associations will be screened by the predicted fill value.

Taken together, the network-based drug screening and repurposing applications provide researchers a lot of alternative approaches for quickly anticancer drugs discovery.

ML-based artificial intelligence for drug discovery. Currently, ML-based biology analysis applications have been employed to carry out drug screening and drug repurposing. For drug screening, previous studies have shown that network-based biology analysis applications can only screen the neighbour proteins of known

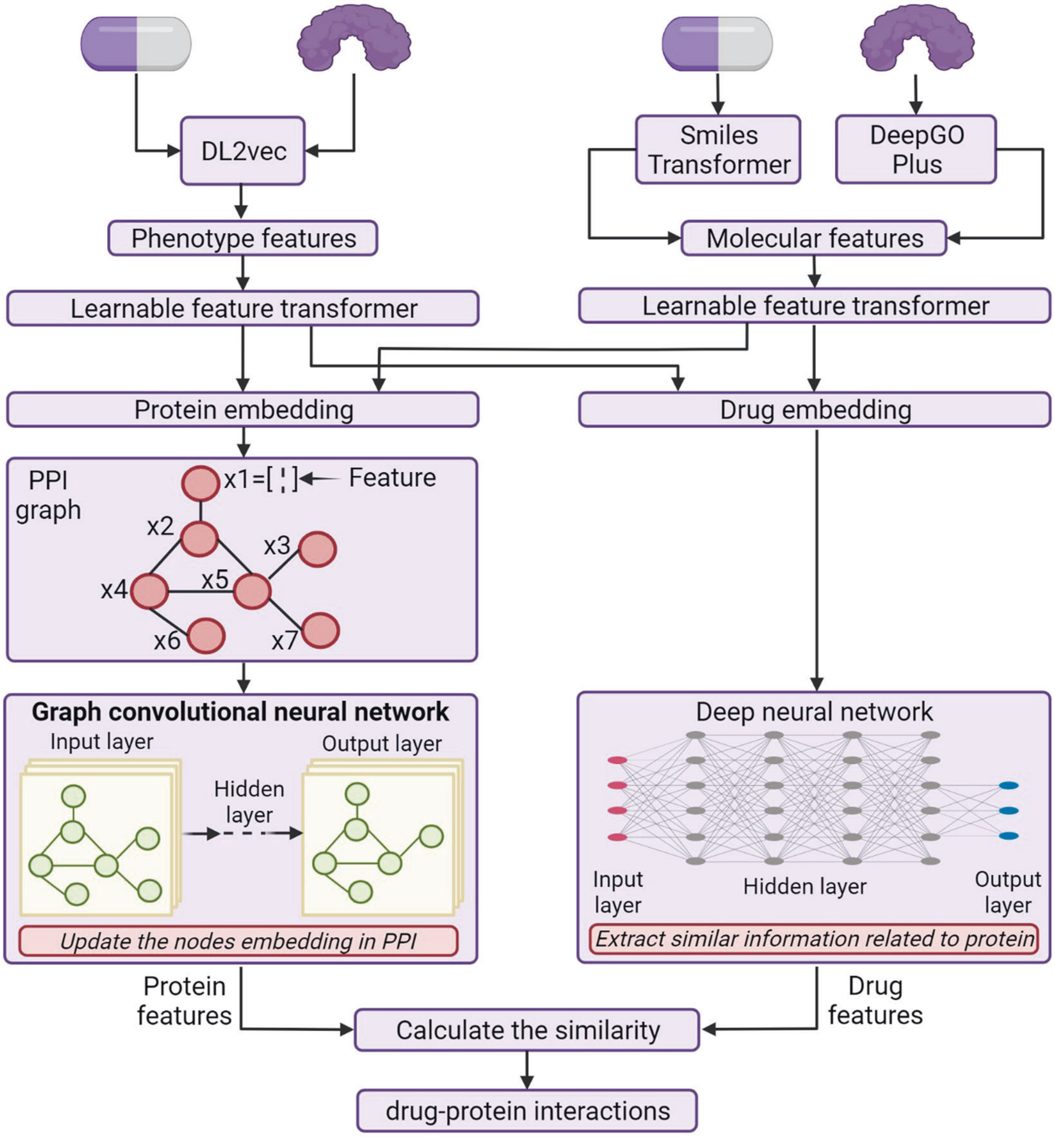

Fig. 11 The graph-based neural network for DTI prediction by combining both bottom-up and top-down biology analysis approaches. (Created with BioRender.com)

targets, while drug-protein interactions may dysregulate the targets' interacting neighbours[227] resulting in high false positive prediction results. ML-based biology analysis applications, such as graph-based neural network, have the advantage of integrated features that combine both 'bottom-up'[229] and 'top-down'[230] approaches to reduce the high false positive prediction results.

For example, Hinnerichs et al.[227] developed the DTI-Voodoo that combines molecular features and phenotypes information with an interaction network using graph neural networks to predict drug-protein interactions (Fig. 11).

Firstly, the model takes the two features, phenotypes features and molecular features, as input. To extracted phenotypes features, they utilized DL2Vec[246] to obtain ontology-based representations. DL2vec constructs a PPI network by introducing nodes for each ontology class and edges for ontology axioms, followed by random walks starting from each node in the graph to generate representations that enable encoding drug effects or protein functions while preserving their semantic neighborhood within that graph. To extract molecular features, they utilized SmilesTransformer[247] to capture the molecular organization of each drug from molecular structures of drugs and utilized DeepGOPlus[248] to capture protein molecular features from protein amino acid sequences.

Secondly, they used two learnable feature transformer models to investigate the latent relationship between phenotypes features and molecular features. According to relationship information, the transformer model, which input the phenotypes features, will output the protein embedding for PPI networks (the top-down approach), and the other transformer model, which input the molecular features, will output drug embedding (the bottom-up approach).

Finally, a DNN was used to extract similar information related to protein from drug embedding, while a GCN is used to update the nodes embedding in PPI networks. Then both protein features and both drugs' features are combined to calculate the similarity by cosine similarity. Since DTI-Voodoo performs well, it demonstrated that graph-based neural networks are good at identifying novel drug-protein interactions.

For drug repurposing, graph-based neural networks take the advantage of feature representation, which can not only utilize the drug-drug links information, but also the features between drug-cancer pairs.

Table 5. The brief description of the ADMET properties[256]

Property	Description
Absorption	The ability of a drug that cross membranes of many cell to reach its site of action, when drug is administered via oral ingestion.
Distribution	After absorption or systemic administration into the bloodstream, a drug is distributed to its site of action through the circulatory systems.
Metabolism	The process of chemically converting a drug to a metabolite is called metabolism or biotransformation.
Excretion	The collective term used for irreversibly removing a drug from the body
Toxicity	The extent to which a drug damages an entire organism, an organism's substructure, or an organ.

For example, Cui et al.[249] proposed GraphRepur, a model for drug repurposing prediction based on graph neural networks. Firstly, the authors collected the drug-induced gene expression data from the LINCS project[250] as well as the drug-drug links information from the STITCH database[251]. Secondly, to obtain the signature of drugs, they identified differentially expressed genes for breast cancer and used the drug-induced genes from LINCS as drug signatures. Thirdly, based on the drug-drug links information from the STITCH database and drug signatures, they constructed a drug-drug links graph with drug signatures as node features. Fourthly, they input drug signatures and drug-drug links information into GraphRepur, and then the model computes scores for drugs that can be repurposed for treating breast cancer. Finally, the authors validated some predictive drugs for breast cancer using experimental data from the literature and showed that the model has significantly better performances than others, such as GCN, DNN, and random forest, in drug repurposing. using published studies.

Furthermore, the authors summarize three conclusions. The first conclusion is that the drug-drug links information plays an important role in studying drug repurposing. The second conclusion is that if such a network with fewer isolated nodes can provide a lot of network topology information, it will significantly improve the prediction performance of graph neural networks. The third is that the drug-induced genetic feature help to improve the DTI prediction accuracy of graph neural network.

Taken together, with the development of graph-based neural networks, an increasing number of ML-based drug screening and repurposing applications can quickly and accurately discover anticancer drugs, reducing the time and financial costs of experiments.

Drug properties prediction

ADMET properties prediction. As discussed in section 4.3 (drug discovery step), after we have a list of drug molecules showing high affinity with the therapeutic target, it is necessary to investigate the properties of these candidates' drugs[252–255]. Since the prediction of drug properties usually adopts the ML-based methods, this study mainly reviews the ML-based biology analysis applications for drug properties prediction such as the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of chemical compounds[256]. Table 5 briefly described the ADMET properties.

ADMET properties prediction can be considered as a classification or regression problem. Because of the strong ability of feature representation[177], graph-based neural networks can capture the drug descriptors (the physicochemical properties, molecular representations, and drug-like properties of molecules) from the drug fingerprints (the substructure features of a molecule)[257], so as to predict ADMET properties by classification or regression algorithm (Fig. 12)[258].

For example, Duvenaud et al.[259] proposed a graph convolution network to learn drug molecular fingerprints, which shows better performance than the state-of-the-art circular fingerprint method for ADMET properties prediction. After that, more and more scientists have used graph-based neural networks to predict the ADMET properties of drug molecules.

For example, Liu et al.[171] proposed Chemi-Net, which utilizes GCN for ADMET properties prediction. They set the characterization of the atoms of the drug molecule and the relationship between atoms as the input of the Chemi-Net, while the output of Chemi-Net is the ADMET properties prediction of drug molecules. The predictive process of Chemi-Net is as follows.

Firstly, the model projects the assembling of the atoms and atom pair descriptors (features between atomic pairs)[257] onto a 3D space to obtain a drug molecule-shaped graph structure. Secondly, Chemi-Net carries out a series of graph convolution operations to output a single fixed-sized molecule embedding. Finally, they obtain accurate ADMET properties predictions of drugs after passing the molecule embedding representation through fully connected layers.

In summary, we consider that more artificial intelligence models for drug properties prediction will be developed in the distant future.

The drug properties application in clinical trial. Since there have been a large number of applications based on artificial intelligence to study the properties of drugs, it still takes on average 10–15 years and 1.5–2.0 billion to bring a new drug to market[260]. One of the main stumbling blocks is the high failure rate of clinical trials. Therefore, some research are committed to the application of artificial intelligence for clinical trial design.

For example, Shah et al[261] construct an artificial intelligence system that made use of the 'self-learning' deep reinforcement learning technology to looks at treatment regimens currently in use, and iteratively adjusts the doses. Therefore, the system can determine the fewest, smallest doses that could still shrink brain tumors, reduce toxicity and eventually find an optimal treatment plan with the lowest possible potency and frequency of doses that should still reduce tumor sizes to a degree comparable to that of traditional regimens. In simulated trials of 50 patients, the system designed treatment cycles that reduced the potency to less than a half of all the doses while maintaining the same tumor-shrinking potential.

In conclusion, we believe that with the development of artificial intelligence applications for drug property prediction, these applications will provide better help for clinical trial.

DISCUSSION AND CONCLUSIONS

Modelling of cellular networks underlying cancer has provided us with a quantitative framework to investigate the link between network properties and the disease by artificial intelligence biology analysis, thereby leading to the discovery of potential novel anticancer targets and drugs[23–29]. However, there is no systematic review that introduces artificial intelligence biology analysis in cancer target identification and drug discovery. For this reason, this study briefly reviewed the scope of artificial intelligence biology analysis to explore new anticancer

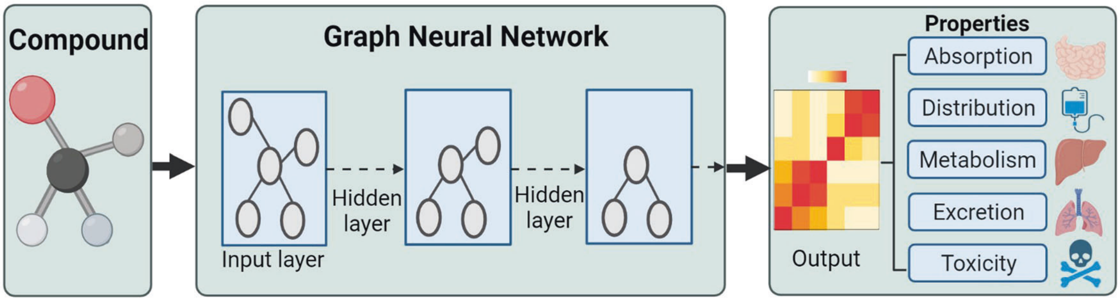

Fig. 12 The graph-based neural network capture the features related to drug properties from drug molecular structure to predict ADMET properties of drugs. (Created with BioRender.com)

targets[34,54,57,74,80], the principles and theory of commonly used artificial intelligence biology analysis algorithms[83–91], and the artificial intelligence applications for artificial intelligence biology analysis[42,195,213].

The scope of artificial intelligence analysis to explore novel anticancer targets consists of epigenetics[54], genomics[57], proteomics[74], metabolomics[34], etc. Since it is not accurate to have anticancer targets by single omics studies, we have to employ artificial intelligence biology analysis to effectively integrate multiple omics data and tackle the complexity of cancer that arises from interactions between genes and their products[16,17] and improve our understanding of carcinogenesis[23–29]. Therefore, how to employ artificial intelligence biology analysis algorithms to integrate multiomics data and identify novel anticancer targets will be an important future study direction.

Next, we introduced two categories of commonly used artificial intelligence algorithms. One is network-based biology analysis algorithms and the other is ML-based biology analysis algorithms. We here discuss their limitations and advantages.

The network-based biology analysis algorithms usually are comprised of shortest path[83], module detection[84] and network centrality[85], which have three major advantages: First, they provide a variety of alternative approaches to identify cancer targets, and different algorithms can compensate each other to identify targets from various perspectives, therefore providing new biological explanations[30]; Second, since they are not limited by the scale of the network, they are good at dealing with the case of small sample network; Third, prior biological knowledge and experience could be conveniently integrated into network-based biology analysis algorithms to make them interpretable.

However, previous studies also show two major shortcomings for the network-based algorithms: First, the current biological network data are biased toward much-studied targets[262]. Since previous studies have paid much attention to these targets, the network-based algorithms will more likely identify these well-studied targets than others due to the data bias[262]. Second, most algorithms only use the topological information of the biological network, but neglect the association between cell function or phenotypes and topological features (such as centrality-based algorithms that are discussed in Section 3.1.2).

ML-based biology analysis algorithms are usually comprised of decision trees[86–88] and deep learning[89–91], which have two major advantages.

One is feature learning and detection[177,181], which employ sophisticated neural network architectures to link up features of biological networks and characterize their relationships. Subsequently, they iteratively train the model to detect such features that are hard to be detected by network-based biology analysis algorithms.

The other is their ability to effectively integrate large and diverse data. It is possible for ML-based networks biology analysis algorithms to integrate multiomics biological network data and identify novel targets[263], because of the fast development of deep learning models and the easy access to high-throughput biological.

Although employing ML-based algorithms greatly benefits the target identification and drug discovery for cancer treatment[174], we still have three major challenges to overcome.

The first challenge is the lack of consistent data for validation[33]. Although the recent advances in biotechnologies have enabled the fast generation of massive biomedical data, such data often suffer from inconsistency in production and information missing in annotation, resulting in the lack of reliable and consistent data for validating deep learning models[264].

The second challenge is the integration of heterogeneous information[103]. Although deep learning models facilitate the integration of multimodal biological data, it is still difficult to build up a universal deep learning model due to the lack of biological domain knowledge[200].

The third challenge is hard to provide interpretability of deep learning models[185]. However, a recent study sheds a light to resolve the issue through a combination of a disease network with a neural network to characterize the mechanism of melanoma[263]. In addition, graphs-based neural networks can improve the interpretability of deep learning models[265].

In the last section of the study, we have reviewed the applications of artificial intelligence biology analysis for cancer therapy from four perspectives: novel anticancer targets identification[189], evaluating the druggability of potential targets[3,4], drug discovery[200], and drug properties prediction[252–255].

First, we presented several widely used applications to identify novel anticancer targets. However, exemplified by WGCNA[195], these network-based biology analysis applications not only requires high computing costs to reconstruct gene co-expression networks[42] but also has difficulty in accurately locating effective network nodes. Although ML-based biology analysis applications employ collaborative modelling by neighbourhood nodes information to reduce the computational cost and improve the predictive accuracy for anticancer targets, biological networks still have data bias[262], resulting in most of the identified targets by current applications already have been reported in previous studies. Therefore, how to develop such an efficient feature selection application that can solve the data bias problem will be appealing for novel therapeutic anticancer target identification[266–268] in the distant future.

Second, we introduce several widely used applications to evaluate the druggability of potential targets. For example, PockDrug is usually used to predict druggable pockets on proteins[213]. Although trRosetta[218] and Alphafold[221] offer opportunities for Pockdrug to evaluate the pharmaceuticals of potential targets, Pockdrug neither accurately predicts druggability due to the complexity of protein structure[269–271] nor costs low efforts to validate through biological experiments[272,273]. Nevertheless, since DTA prediction can quickly provide reliable druggable targets for cancer care with low financial costs[211], it is potential to develop the related efficient artificial intelligence biology analysis applications for DTA prediction in the distant future.

Third, we investigated several widely used applications for drug discovery, which consists of drug screening and drug repurposing.

For drug screening, identifying drug-target interactions (DTIs) is a crucial step. Since network-based biology analysis applications for DTI prediction are usually based on the guilt-by-association principle[228], it can only predict the interacting neighbors of known cancer targets. Currently, ML-based biology analysis applications can extend the predictions to downstream consequences[227], thereby screening out more possible anticancer drugs.

For drug repurposing[232], there are four commonly used network-based biology analysis applications[234–241] that integrate the similarities among various drugs but ignore prior knowledge. However, ML-based biology analysis applications not only can take advantage of the similarity among drugs, but also can integrate drug properties to improve the accuracy of drug repurposing.

Fourth, we introduce widely used applications for drug properties prediction. For example, graph convolution networks, which have a strong ability of feature representation[177], can capture the features related to ADMET properties of drugs from their molecular structures. Therefore, it is becoming a popular method to predict drug properties by integrating drug molecular structures and drug clinical phenotype for drug properties prediction through graph convolution networks[274]. Here, we wish once more and more artificial intelligence biology analysis models are developed to capture the features related to ADMET properties from the drug molecular structure, to improve the success rate of clinical trials.

In summary, although we have reviewed and discussed many artificial intelligence algorithms and corresponding applications for novel anticancer target identification and drug discovery, this review is still too brief to cover the entire research area. However, because artificial intelligence algorithms are effective in exploring new anticancer targets and discovering drugs, we wish this review could offer valuable enlightenments for interested researchers to develop an understanding of the principles behind artificial intelligence biology analysis in cancer target identification and drug discovery. Moreover, we wish that our perspective on artificial intelligence and related applications will provide the pathway for further advancement in the field.

AUTHOR CONTRIBUTIONS

Y.Y. and X.L. contributed equally to this work. Y.Y., X.L., Y.P., H.Z., J.V., S.L., S.D. and L.Z. contributed to writing and revising the paper. X.L., S.D., and L.Z. supervised the research. All authors have read and approved the article.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

REFERENCES

1. Shabani, M. & Hojjat-Farsangi, M. Targeting receptor tyrosine kinases using monoclonal antibodies: the most specific tools for targeted-based cancer therapy. *Curr. Drug Targets* **17**, 1687–1703 (2016).
2. Paananen, J. & Fortino, V. An omics perspective on drug target discovery platforms. *Brief. Bioinform* **21**, 1937–1953 (2019).
3. Hopkins, A. L. & Groom, C. R. Opinion: The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
4. Bushweller, J. H. Targeting transcription factors in cancer—from undruggable to reality. *Nat. Rev. Cancer* **19**, 611–624 (2019).
5. Colaprico, A. et al. Interpreting pathways to discover cancer driver genes with Moonlight. *Nat. Commun.* **11**, 69 (2020).
6. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug development in the era of precision medicine. *Nat. Rev. Drug Discov.* **17**, 183–196 (2018).
7. Manzari, M. T. et al. Targeted drug delivery strategies for precision medicines. *Nat. Rev. Mater.* **6**, 351–370 (2021).
8. Rosenblum, D., Joshi, N., Tao, W., Karp, J. M. & Peer, D. Progress and challenges towards targeted delivery of cancer therapeutics. *Nat. Commun.* **9**, 1410 (2018).
9. Song, H. et al. Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. *Neurocomputing* **469**, 354–365 (2022).
10. Zhang, L. et al. MCDB: a comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction. *Acta Pharm. Sin. B* **11**, 3092–3104 (2021).
11. Gao, J., Liu, P., Liu, G. D. & Zhang, L. Robust needle localization and enhancement algorithm for ultrasound by deep learning and beam steering methods. *J. Comput. Sci. Technol.* **36**, 334–346 (2021).
12. Liu, G. D., Li, Y. C., Zhang, W. & Zhang, L. A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Eng.-Prc* **6**, 462–467 (2020).
13. Zhang, L., Bai, W., Yuan, N. & Du, Z. Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput. Biol.* **15**, e1007069 (2019).
14. Zhang, L. & Zhang, S. Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: "Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition" by Qian Wang et al. *Phys. Life Rev.* **20**, 140–142 (2017).
15. Zhou, Y., Wang, F., Tang, J., Nussinov, R. & Cheng, F. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit. Health* **2**, e667–e676 (2020).
16. Suhail, Y. et al. Systems biology of cancer metastasis. *Cell Syst.* **9**, 109–127 (2019).
17. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
18. Lv, J., Deng, S. & Zhang, L. A review of artificial intelligence applications for antimicrobial resistance. *Biosaf. Health* **3**, 22–31 (2021).
19. Wu, W. et al. Exploring the dynamics and interplay of human papillomavirus and cervical tumorigenesis by integrating biological data into a mathematical model. *BMC Bioinform.* **21**, 152 (2020).
20. Xiao, M. et al. 2019nCoVAS: developing the web service for epidemic transmission prediction, genome analysis, and psychological stress assessment for 2019-nCoV. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**, 1250–1261 (2021).
21. Xiao, M., Yang, X., Yu, J. & Zhang, L. CGIDLA: developing the web server for CpG island related density and LAUPs (Lineage-Associated Underrepresented Permutations) study. *IEEE/ACM Trans. Comput Biol. Bioinform* **17**, 2148–2154 (2020).
22. Zhao, J., Cao, Y. & Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* **18**, 417–426 (2020).
23. Chen, Y. et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
24. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **17**, 615–629 (2016).
25. Ideker, T. & Nussinov, R. Network approaches and applications in biology. *PLoS Comput. Biol.* **13**, e1005771 (2017).
26. Lai, X. et al. MiR-205-5p and miR-342-3p cooperate in the repression of the E2F1 transcription factor in the context of anticancer chemotherapy resistance. *Theranostics* **8**, 1106 (2018).
27. Lai, X., Eberhardt, M., Schmitz, U. & Vera, J. Systems biology-based investigation of cooperating microRNAs as monotherapy or adjuvant therapy in cancer. *Nucleic Acids Res.* **47**, 7753–7766 (2019).
28. Seyfried, N. T. et al. A multi-network approach identifies protein-specific co-expression in asymptomatic and symptomatic Alzheimer's disease. *Cell Syst.* **4**, 60–72.e64 (2017).
29. Vidal, M., Cusick, Michael, E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
30. Chen, L. & Wu, J. Bio-network medicine. *J. Mol. Cell Biol.* **7**, 185–186 (2015).
31. Ghanat Bari, M., Ung, C. Y., Zhang, C., Zhu, S. & Li, H. Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. *Sci. Rep.* **7**, 6993 (2017).
32. Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Brief. Bioinform.* **22**, 1515–1530 (2021).
33. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
34. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
35. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).
36. Kim, H. & Kim, Y.-M. Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Sci. Rep.* **8**, 6041 (2018).
37. Vinayagam, A. et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl Acad. Sci. USA* **113**, 4976–4981 (2016).
38. do Valle, Í. F. et al. Network integration of multi-tumour omics data suggests novel targeting strategies. *Nat. Commun.* **9**, 1–10 (2018).
39. Yang, K. et al. A comprehensive analysis of metabolomics and transcriptomics in cervical cancer. *Sci. Rep.* **7**, 43353 (2017).

40. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
41. Casparie, M. et al. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Cell Oncol.* **29**, 19–24 (2007).
42. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2017).
43. Wang, Y. et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res*. **48**, D1031–D1041 (2019).
44. Wang, Y. et al. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37**, W623–W633 (2009).
45. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2011).
46. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
47. McLendon, R. et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
48. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
49. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
50. Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2010).
51. Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2020).
52. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
53. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
54. Perakakis, N., Yazdani, A., Karniadakis, G. E. & Mantzoros, C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* **87**, A1–A9 (2018).
55. Wilson, S. & Filipp, F. V. A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer. *NPJ Syst. Biol. Appl.* **4**, 24 (2018).
56. Filipp, F. V. Crosstalk between epigenetics and metabolism—Yin and Yang of histone demethylases and methyltransferases in cancer. *Brief. Funct. Genom.* **16**, 320–325 (2017).
57. Holmes, M. V., Richardson, T. G., Ference, B. A., Davies, N. M. & Davey Smith, G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat. Rev. Cardiol.* **18**, 435–453 (2021).
58. Ozaki, K. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).
59. Golub, T. R. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
60. Oliver, S. Proteomics: guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
61. Lanza, V. F., Baquero, F., Cruz, F. D. L. & Coque, T. M. AccNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. *Bioinformatics* **33**, btw601 (2016).
62. Fernandes, E. G., Lombardi, A., Solaro, R. & Chiellini, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
63. Escala-Garcia, M., Abraham, J. & Andrulis, I. L. et al. A network analysis to identify mediators of germline-driven differences in breast cancer prognosis. *Nat. Commun.* **11**, 312 (2020).
64. Pidò, S., Ceddia, G. & Masseroli, M, MM. Computational analysis of fused co-expression networks for the identification of candidate cancer gene biomarkers. *NPJ Syst. Biol. Appl.* **7**, 17 (2021).
65. Medi, K., Kazim, Y. A. & Craig, M. Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. *PLoS One* **13**, e0200717 (2018).
66. Cantini, L., Medico, E., Fortunato, S. & Caselle, M. Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep.* **5**, 17386 (2015).
67. Zhang, L., Dai, Z., Yu, J. & Xiao, M. CpG-island-based annotation and analysis of human housekeeping genes. *Brief. Bioinform.* **22**, 515–525 (2021).
68. Zhang, L. et al. Computed tomography angiography-based analysis of high-risk intracerebral haemorrhage patients by employing a mathematical model. *BMC Bioinform.* **20**, 193 (2019).
69. Zhang, L. et al. EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. *J. Mol. Cell Biol.* **9**, 477–488 (2017).
70. Zhang, L. et al. Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. *Nanoscale* **8**, 14877–14887 (2016).
71. Zhang, L., Xiao, M., Zhou, J. & Yu, J. Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). *Bioinformatics* **34**, 3624–3630 (2018).
72. Zhang, L. et al. Building up a robust risk mathematical platform to predict colorectal cancer. *Complexity* **2017**, 8917258 (2017).
73. Zhang, L. et al. Bioinformatic analysis of chromatin organization and biased expression of duplicated genes between two poplars with a common whole-genome duplication. *Horticult. Res.* **8**, 62 (2021).
74. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
75. Li, Z., Ivanov, A. A. & AL, e The OncoPPi network of cancer-focused protein–protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.* **8**, 14356 (2017).
76. Kalman, R. E. Mathematical description of linear dynamical systems. *J. Soc. Ind. Appl. Math. Ser. A Control* **1**, 152–192 (1963).
77. Ravindran, V., Sunitha, V. & Bagler, G. Identification of critical regulatory genes in cancer signaling network using controllability analysis. *Phys. A: Stat. Mech. Appl.* **474**, 134–143 (2017).
78. do Valle, I. F. et al. Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols. *Nat. Food* **2**, 143–155 (2021).
79. Basler, G., Nikoloski, Z., Larhlimi, A., Barabási, A.-L. & Liu, Y.-Y. Control of fluxes in metabolic networks. *Genome Res.* **26**, 956–968 (2016).
80. Chakraborty, S., Hosen, M. I., Ahmed, M. & Shekhar, H. U. Onco-Multi-OMICS approach: a new frontier in cancer research. *Biomed. Res. Int.* **2018**, 9836256 (2018).
81. Zhang, C. et al. The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Med. Oncol.* **34**, 101 (2017).
82. Gov, E., Kori, M. & Arga, K. Y. Multiomics analysis of tumor microenvironment reveals Gata2 and miRNA-124-3p as potential novel biomarkers in ovarian cancer. *OMICS* **21**, 603–615 (2017).
83. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
84. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2009).
85. Lü, L. et al. Vital nodes identification in complex networks. *Phys. Rep.* **650**, 1–63 (2016).
86. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
87. Loh, W.-Y. Classification and regression trees. *Phys. Rep.* **1**, 14–23 (2011).
88. Cao, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* **2**, 500–508 (2020).
89. Nordhausen & Klaus An introduction to statistical learning—with applications in R by Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani. *Int. Stat. Rev.* **82**, 156–157 (2014).
90. Hao, X., Zhang, G. & Ma, S. Deep learning. *Int. J. Semantic Comput.* **10**, 417–439 (2016).
91. Lecun, Y., Bengio, Y. & Hinton, G. E. Deep learning. *Nature* **521**, 436–444 (2015).
92. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
93. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
94. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193 (2002).
95. T-M, H. Architecture of the drug-drug interaction network. *J. Clin. Pharm. Ther.* **36**, 135–143 (2011).
96. Martinez, V., Berzal, F. & Cubero, J. C. A survey of link prediction in complex networks. *ACM Comput. Surv.* **49**, 69.61–69.33 (2017).
97. Hens, C., Harush, U., Haber, S., Cohen, R. & Barzel, B. Spatiotemporal signal propagation in complex networks. *Nat. Phys.* **15**, 403–412 (2019).
98. Lazareva, O., Baumbach, J., List, M. & Blumenthal, D. B. On the limits of active module identification. *Brief. Bioinform.* **22**, bbab066 (2021).
99. Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L. Controllability of complex networks. *Nature* **473**, 167–173 (2011).
100. Abhik, S. & Wild, D. J. Netpredictor: R and Shiny package to perform drug-target network analysis and prediction of missing links. *BMC Bioinform.* **19**, 265 (2018).

101. Kuperstein, I. et al. The shortest path is not the one you know: application of biological network resources in precision oncology research. *Mutagenesis* **30**, 191–204 (2015).
102. Rabbani, M. & Kazemi, S. Solving uncapacitated multiple allocation p-hub center problem by Dijkstra's algorithm-based genetic algorithm and simulated annealing. *Int. J. Ind. Eng. Comput.* **6**, 405–418 (2015).
103. Li, Z. et al. Identifying novel genes and chemicals related to nasopharyngeal cancer in a heterogeneous network. *Sci. Rep.* **6**, 25515 (2016).
104. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.* **12**, 1796 (2021).
105. Chen, L. et al. Identification of novel candidate drivers connecting different dysfunctional levels for lung adenocarcinoma using protein-protein interactions and a shortest path approach. *Sci. Rep.* **6**, 29849 (2016).
106. Li, B.-Q., Huang, T., Liu, L., Cai, Y.-D. & Chou, K.-C. Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. *PLoS One* **7**, e33393 (2012).
107. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
108. Barthélemy, M. Betweenness centrality in large complex networks. *Eur. Phys. J. B* **38**, 163–168 (2004).
109. Maclean, H. E., Warne, G. L. & Zajac, J. D. Localization of functional domains in the androgen receptor. *J. Steroid Biochem. Mol. Biol.* **62**, 233–242 (1997).
110. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116 (2001).
111. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2018).
112. Lu, S., Zhu, Z.-G. & Lu, W.-C. Inferring novel genes related to colorectal cancer via random walk with restart algorithm. *Gene Ther.* **26**, 373–385 (2019).
113. Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601–1257601 (2015).
114. Choobdar, S. et al. Assessment of network module identification across complex diseases. *Nat. Methods* **16**, 843–852 (2019).
115. Fortunato, S. & Hric, D. Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44 (2016).
116. Newman, M. E. J. Communities, modules and large-scale structure in networks. *Nat. Phys.* **8**, 25–31 (2012).
117. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
118. Silverbush, D. et al. Simultaneous integration of multi-omics data improves the identification of cancer driver modules. *Cell Syst.* **8**, 456–466.e455 (2019).
119. Hossain, S. M. M., Halsana, A. A., Khatun, L., Ray, S. & Mukhopadhyay, A. Discovering key transcriptomic regulators in pancreatic ductal adenocarcinoma using Dirichlet process Gaussian mixture model. *Sci. Rep.* **11**, 7853 (2021).
120. Ghiassian, S. D. et al. Endophenotype network models: common core of complex diseases. *Sci. Rep.* **6**, 27414 (2016).
121. 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 Comput. Biol.* **11**, e1004120 (2015).
122. Wang, R.-S. & Loscalzo, J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J. Mol. Biol.* **430**, 2939–2950 (2018).
123. Wang, Q., Yu, H., Zhao, Z. & Jia, P. EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics* **31**, 2591–2594 (2015).
124. Zhang, Y. et al. A gene module identification algorithm and its applications to identify gene modules and key genes of hepatocellular carcinoma. *Sci. Rep.* **11**, 5517 (2021).
125. Paci, P. et al. Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Syst. Biol. Appl.* **7**, 3 (2021).
126. Tripathi, B., Parthasarathy, S., Sinha, H., Raman, K. & Ravindran, B. Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Front. Genet.* **10**, 164 (2019).
127. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
128. Mangangcha, I. R., Malik, M. Z., Küçük, Ö., Ali, S. & Singh, R. K. B. Identification of key regulators in prostate cancer from gene expression datasets of patients. *Sci. Rep.* **9**, 16420 (2019).
129. Kitsak, M. et al. Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
130. Jalili, M. et al. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Front. Physiol.* **7**, 375 (2016).
131. Pastor-Satorras, R. & Castellano, C. Distinct types of eigenvector localization in networks. *Sci. Rep.* **6**, 18847 (2016).
132. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
133. Zhang, J. et al. P4HB, a novel hypoxia target gene related to gastric cancer invasion and metastasis. *Biomed. Res. Int.* **2019**, 9749751 (2019).
134. Ahajjam, S. & Badir, H. Identification of influential spreaders in complex networks using HybridRank algorithm. *Sci. Rep.* **8**, 11932 (2018).
135. Malliaros, F. D., Rossi, M.-E. G. & Vazirgiannis, M. Locating influential nodes in complex networks. *Sci. Rep.* **6**, 19307 (2016).
136. Li, H. et al. Deciphering the mechanism of Indirubin and its derivatives in the inhibition of Imatinib resistance using a "drug target prediction-gene microarray analysis-protein network construction" strategy. *BMC Complement. Alter. Med.* **19**, 75 (2019).
137. Taylor, I. W. et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27**, 199–204 (2009).
138. Raman, K., Damaraju, N. & Joshi, G. K. The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Syst. Synth. Biol.* **8**, 73–81 (2014).
139. Mallik, S. & Maulik, U. MiRNA-TF-gene network analysis through ranking of biomolecules for multi-informative uterine leiomyoma dataset. *J. Biomed. Infor.* **57**, 308–319 (2015).
140. Chen, C. et al. Construction and analysis of protein-protein interaction networks based on proteomics data of prostate cancer. *Int. J. Mol. Med.* **37**, 1576–1586 (2016).
141. Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M. & Homouz, D. Analyzing a co-occurrence gene-interaction network to identify disease-gene association. *BMC Bioinform.* **20**, 70 (2019).
142. Jiang, P. et al. Network analysis of gene essentiality in functional genomics experiments. *Genome Biol.* **16**, 1–10 (2015).
143. Chen, K.-H., Wang, K.-J., Wang, K.-M. & Angelia, M.-A. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Appl. Soft Comput.* **24**, 773–780 (2014).
144. Chen, K.-H. et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinform.* **15**, 49 (2014).
145. Li, Y., Tang, X.-Q., Bai, Z. & Dai, X. Exploring the intrinsic differences among breast tumor subtypes defined using immunohistochemistry markers based on the decision tree. *Sci. Rep.* **6**, 1–13 (2016).
146. Carson, M. B. & Lu, H. Network-based prediction and knowledge mining of disease genes. *BMC Med. Genom.* **8**, S9 (2015).
147. Ramadan, E., Alinsaif, S. & Hassan, M. R. Network topology measures for identifying disease-gene association in breast cancer. *BMC Bioinform.* **17**, 274 (2016).
148. Lerman, R. I. & Yitzhaki, S. A note on the calculation and interpretation of the Gini index. *Econ. Lett.* **15**, 363–368 (1984).
149. Burt, R. S. Structural holes and good ideas. *Am. J. Sociol.* **110**, 349–399 (2004).
150. Ye, N., Zhang, Y., Wang, R. & Malekian, R. Vehicle trajectory prediction based on Hidden Markov Model. *KSII Trans. Internet Infor. Syst.* **10**, 3150–3170 (2016).
151. Ernesto, E. Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* **71**, 056103 (2005).
152. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
153. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
154. Towfic, F. et al. Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinform.* **11**, S7 (2010).
155. Soffer, S. N. & Vázquez, A. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 057101 (2005).
156. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
157. Ezzat, A., Wu, M., Li, X. L. & Kwoh, C. K. Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* **129**, 81 (2017).
158. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
159. Sarica, A., Cerasa, A. & Quattrone, A. Random Forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front. Aging Neurosci.* **9**, 329 (2017).
160. Toth, R., Schiffmann, H., Hube-Magg, C., Büscheck, F. & Gerhuser, C. Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin. Epigenet.* **11**, 148 (2019).
161. Jin, H. & Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **17**, 299–310 (2005).
162. Kingsford, C. & Salzberg, S. L. What are decision trees? *Nat. Biotechnol.* **26**, 1011–1013 (2008).
163. Hao, D. & Li, C. The dichotomy in degree correlation of biological networks. *PLoS One* **6**, e28322 (2011).

164. Zhang, Q., Wang, F. Y., Zeng, D. & Wang, T. Understanding crowd-powered search groups: a social network perspective. *PLoS One* **7**, 1–16 (2012).
165. Freund, Y. & Mason, L. The Alternating Decision Tree Learning Algorithm. In *Proc. Sixteenth International Conference on Machine Learning*, 124–133 (1999).
166. Zhang, L. et al. Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. *Bioinformatics* **37**, 1554–1561 (2021).
167. Tabrizchi, H., Tabrizchi, M. & Tabrizchi, H. Breast cancer diagnosis using a multiverse optimizer-based gradient boosting decision tree. *SN Appl. Sci.* **2**, 1–19 (2020).
168. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
169. Liu, H., Zhang, W., Song, Y., Deng, L. & Zhou, S. HNet-DNN: inferring new drug–disease associations with deep neural network based on heterogeneous network features. *J. Chem. Inform. Modeling* **60**, 2367–2376 (2020).
170. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inform. Modeling* **57**, 942–957 (2017).
171. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol. Pharm.* **14**, 4462–4475 (2017).
172. Ma, T., Xiao, C., Zhou, J. & Wang, F. Drug similarity integration through attentive multi-view graph auto-encoders. In *Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3477–3483, https://doi.org/10.24963/ijcai.2018/483 (2018).
173. Lan, W. et al. GANLDA: graph attention network for lncRNA-disease associations prediction. *Neurocomputing* **469**, 384–393 (2022).
174. Li, G. et al. Predicting MicroRNA-disease associations using network topological similarity based on DeepWalk. *IEEE Access* **5**, 24032–24039 (2017).
175. Webb, S. Deep learning for biology. *Nature* **554**, 555–557 (2018).
176. Selvaraj, G. et al. Identification of target gene and prognostic evaluation for lung adenocarcinoma using gene expression meta-analysis, network analysis and neural network algorithms. *J. Biomed. Inform.* **86**, 120–134 (2018).
177. Goyal, P. & Ferrara, E. Graph embedding techniques, applications, and performance: a survey. *Knowl.-Based Syst.* **151**, 78–94 (2017).
178. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn Syst.* **32**, 4–24 (2021).
179. Zheng, K., You, Z.-H., Wang, L., Wong, L. & Chen, Z.-H. Inferring disease-associated Piwi-interacting RNAs via graph attention networks. 239–250, (2020).
180. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems*, 5998–6008 (2017).
181. Singh, M., Singh, R. & Ross, A. A comprehensive overview of biometric fusion. *Inform. Fusion* **52**, 187–205 (2019).
182. Shi, Z., Zhang, H., Jin, C., Quan, X. & Yin, Y. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinform.* **22**, 136 (2021).
183. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
184. Kim, D. et al. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *J. Am. Med. Inform. Assoc.* **24**, 577–587 (2016).
185. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
186. Alex, F., Song, J. S. & Ilya, I. Maximum entropy methods for extracting the learned features of deep neural networks. *PLoS Comput. Biol.* **13**, e1005836- (2017).
187. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
188. Ozerov, I. V. et al. In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. *Nat. Commun.* **7**, 13427 (2016).
189. Xia, J., Benner, M. J. & Hancock, R. E. NetworkAnalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic Acids Res.* **42**, W167–W174 (2014).
190. Hernández-de-Diego, R. et al. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* **46**, W503–W509 (2018).
191. Tuhkuri, A. et al. Patients with early-stage oropharyngeal cancer can be identified with label-free serum proteomics. *Br. J. Cancer* **119**, 200–212 (2018).
192. Abbas, S. Z., Qadir, M. I. & Muhammad, S. A. Systems-level differential gene expression analysis reveals new genetic variants of oral cancer. *Sci. Rep.* **10**, 14667 (2020).
193. Ren, G. & Liu, Z. NetCAD: a network analysis tool for coronary artery disease-associated PPI network. *Bioinformatics* **29**, 279–280 (2012).
194. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).
195. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
196. Xian-Guo, Z. et al. Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis. *Onco Targets Ther.* **11**, 2815–2830 (2018).
197. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
198. Wang, A. et al. Cell adhesion-related molecules play a key role in renal cancer progression by multinetwork analysis. *Biomed. Res. Int.* **2019**, 2325765 (2019).
199. Lai, X. et al. Network- and systems-based re-engineering of dendritic cells with non-coding RNAs for cancer immunotherapy. *Theranostics* **11**, 1412–1428 (2021).
200. Jin, S., Zeng, X., Xia, F., Huang, W. & Liu, X. Application of deep learning methods in biological networks. *Brief. Bioinform.* **22**, 1902–1917 (2020).
201. Zhu, Y., Shen, X. & Pan, W. Network-based support vector machine for classification of microarray samples. *BMC Bioinform.* **10**, S21 (2009).
202. Sanchez, R. & Mackenzie, S. A. Integrative network analysis of differentially methylated and expressed genes for biomarker identification in leukemia. *Sci. Rep.* **10**, 2123 (2020).
203. Wang, T. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* **12**, 3445 (2021).
204. Xuan, P., Zhang, P. S., Liu, T., Sun, Y. & Graph, H. Convolutional network and convolutional neural network based method for predicting lncrna-disease associations. *Cells* **8**, 1012 (2019). Aug 30.
205. Wu, M.-Y. et al. Regularized logistic regression with network-based pairwise interaction for biomarker identification in breast cancer. *BMC Bioinform.* **17**, 108 (2016).
206. Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* **17**, 595–610 (2013).
207. Gigliotti, B. J., Russell, M. D., Shonka, D. & Stathatos, N. Fine-needle aspiration and molecular analysis. *Surgery of the Thyroid and Parathyroid Glands (Third Edition)*, 118–131, https://doi.org/10.1016/B978-0-323-66127-0.00012-0 (2021).
208. Sinkala, M., Mulder, N. & Martin, D. Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Sci. Rep.* **10**, 1212 (2020).
209. Kaczmarek, E. et al. Multi-Omic graph transformers for cancer classification and interpretation. In Proc. *Pacific Symposium on Biocomput*ing **27**, 373–384, https://doi.org/10.1142/9789811250477_0034.
210. Vermeulen, M. & Lelie, N. The current status of nucleic acid amplification technology in transfusion-transmitted infectious disease testing. *ISBT Sci. Ser.* **11**, 123–128 (2016).
211. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
212. Kandoi, G., Acencio, M. L. & Lemke, N. Prediction of druggable proteins using machine learning and systems biology: a mini-review. *Front. Physiol.* **6**, 366 (2015).
213. Hussein, H. A. et al. PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res.* **43**, W436–W442 (2015).
214. Yang, Y.-F., Yu, B., Zhang, X.-X. & Zhu, Y.-H. Identification of TNIK as a novel potential drug target in thyroid cancer based on protein druggability prediction. *Medicines* **100**, e25541–e25541 (2021).
215. Sheng, W. et al. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
216. Ovchinnikov, S. et al. Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).
217. Wang, T. et al. Improved fragment sampling for ab initio protein structure prediction using deep neural networks. *Nat. Mach. Intell.* **1**, 347–355 (2019).
218. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
219. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins* **87**, 1011–1020 (2019).
220. Haas, J. et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **86**, 387–398 (2018).
221. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
222. Shim, J., Hong, Z.-Y., Sohn, I. & Hwang, C. Prediction of drug–target binding affinity using similarity-based convolutional neural network. *Sci. Rep.* **11**, 4416 (2021).
223. Liu, B., He, H., Luo, H., Zhang, T. & Jiang, J. Artificial intelligence and big data facilitated targeted drug discovery. *Stroke Vasc. Neurol.* **4**, 206–213 (2019).

224. He, T., Heidemeyer, M., Ban, F., Cherkasov, A. & Ester, M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminform.* **9**, 24 (2017).
225. Nguyen, T. et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2020).
226. Roda, A., Guardigli, M., Pasini, P. & Mirasoli, M. Bioluminescence and chemiluminescence in drug screening. *Anal. Bioanal. Chem.* **377**, 826–833 (2003).
227. Hinnerichs, T. & Hoehndorf, R. DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions. *Bioinformatics* **37**, 4835–4843 (2021).
228. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–602 (2000).
229. Monica, C. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
230. Feng, Y., Wang, Q. & Wang, T. Drug target protein-protein interaction networks: a systematic perspective. *Biomed. Res. Int.* **2017**, 1289259 (2017).
231. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **15**, e1007129 (2019).
232. Parvathaneni, V., Kulkarni, N. S., Muth, A. & Gupta, V. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug Discov. Today* **24**, 2076–2085 (2019).
233. Pritchard, J. E., O'Mara, T. A. & Glubb, D. M. Enhancing the promise of drug repositioning through genetics. *Front. Pharm.* **8**, 896 (2017).
234. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **7**, 496 (2011).
235. Iwata, H., Sawada, R., Mizutani, S. & Yamanishi, Y. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J. Chem. Inform. Model.* **55**, 446–459 (2015).
236. Liu, H., Song, Y., Guan, J., Luo, L. & Zhuang, Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinform.* **17**, 539 (2016).
237. Luo, H. et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* **32**, 2664–2671 (2016).
238. Wang, W., Yang, S., Zhang, X. & Li, J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* **30**, 2923–2930 (2014).
239. Liu, C. et al. Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6,700 cancer genomes. *PLoS Comput. Biol.* **16**, e1007701 (2020).
240. Yang, M., Luo, H., Li, Y. & Wang, J. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* **35**, i455–i463 (2019).
241. Yang, M., Luo, H., Li, Y., Wu, F.-X. & Wang, J. Overlap matrix completion for predicting drug-associated indications. *PLoS Comput. Biol.* **15**, e1007541 (2019).
242. Cheng, F. et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.* **10**, 3476 (2019).
243. Luo, H. et al. Drug repositioning based on comprehensive similarity measures and Bi-Random Walk algorithm. *Bioinformatics* **32**, btw228 (2016).
244. Feixiong, C., Junfei, Z., Michaela, F. & Zhongming, Z. A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. *J. Am. Med. Inform. Assoc.* **23**, 681–691 (2016).
245. Luo, H. et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* **34**, 1904–1912 (2018).
246. Chen, J., Althagafi, A. & Hoehndorf, R. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics* **37**, 853–860 (2020).
247. Honda, S., Shi, S. & Ueda, H. R. SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. *CoRR* abs/1911.04738 (2019).
248. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2019).
249. Cui, C. et al. Drug repurposing against breast cancer by integrating drug-exposure expression profiles and drug–drug links based on graph neural network. *Bioinformatics* **37**, 2930–2937 (2021).
250. Aravind, S. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e1417 (2017).
251. Szklarczyk, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2014).
252. Huang, L.-C., Wu, X. & Chen, J. Y. Predicting adverse side effects of drugs. *BMC Genom.* **12**, S11 (2011).
253. Arrowsmith & John Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* **10**, 87 (2011).
254. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2016).
255. Shaked, I., Oberhardt, M. A., Atias, N., Sharan, R. & Ruppin, E. Metabolic network prediction of drug side effects. *Cell Syst.* **2**, 209–213 (2016).
256. Zhong, H. A. ADMET properties: overview and current topics. *Drug Design: Principles and Applications*, 113–133, https://doi.org/10.1007/978-981-10-5187-6_8 (2017).
257. Lei, T. et al. ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *J. Cheminform.* **8**, 6 (2016).
258. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22**, 69–77 (2003).
259. Duvenaud, D. et al. ConvolutioNal networks on graphs for learning molecular fingerprints. *Adv. Neural Inform. Process. Syst.* **13**, 2224–2232 (2015).
260. Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial intelligence for clinical trial design. *Trends Pharm. Sci.* **40**, 577–591 (2019).
261. Yauney, G. & Shah, P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. *Proc. 3rd Mach. Learn. Healthc. Conf.* **85**, 161–226 (2018).
262. Cusick, M. E. et al. Literature-curated protein interaction datasets. *Nat. Methods* **6**, 934–935 (2009).
263. Lai, X. et al. A disease network-based deep learning approach for characterizing melanoma. *Int. J. Cancer* **150**, 1029–1044 (2022).
264. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
265. Mohamed, S. K., Nováček, V. & Nounu, A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* **36**, 603–610 (2019).
266. Zhang, Q. et al. Deep learning based classification of breast tumors with shear-wave elastography. *Ultrasonics* **72**, 150–157 (2016).
267. Takahashi, Y. et al. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Transl. Psychiatry* **10**, 157 (2020).
268. Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.* **3**, 513–526 (2021).
269. Wu, F., Ma, C. & Tan, C. Network motifs modulate druggability of cellular targets. *Sci. Rep.* **6**, 36626 (2016).
270. Abi Hussein, H. et al. Global vision of druggability issues: applications and perspectives. *Drug Discov. Today* **22**, 404–415 (2017).
271. Hiba Abi, H. PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res.* **43**, W436–442 (2015).
272. Zhang, A. et al. Discovery and verification of the potential targets from bioactive molecules by network pharmacology-based target prediction combined with high-throughput metabolomics. *RSC Adv.* **7**, 51069–51078 (2017).
273. Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A. & Ragan, M. A. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med.* **4**, 41–41 (2012).
274. Zheng, Y., Peng, H., Ghosh, S., Lan, C. & Li, J. Inverse similarity and reliable negative samples for drug side-effect prediction. *BMC Bioinform.* **19**, 554 (2019).