Journal of Mathematical Chemistry
https://doi.org/10.1007/s10910-023-01511-6

REVIEW

Applications of graph theory in studying protein structure, dynamics, and interactions

Ziyun Zhou[1,2] · **Guang Hu**[1,2]

Received: 20 June 2023 / Accepted: 19 August 2023

Abstract

Being a core tool, graph theory, as a mathematical formalism in mathematical chemistry, has become an essential approach for studying the complex behavior and interactions in protein systems. Here, we review recent advances in the field, particularly in our group, including the methods developed to access protein functions and their applications in disease biology. First, we provide the necessary background and definitions of graph-based structures and network centralities, and methodologies developed at the node-, subgraph- and pathway-levels. We then review the applications of how to use these algorithms to gain new biological insights, ranging from protein structures to protein dynamics, and interactions for linking genotypes and phenotypes. Furthermore, we discuss immediate challenges in the multilayer network, which is more realistic in the biological world, and hope to draw increasing attention from mathematicians, especially graph theorists, to reveal the basic principles of "networks of networks".

Keywords Protein structure graph · Protein-protein interaction network · Systems biology · Allosteric regulation · Drug target

1 Introduction

Graph theory has long been established as a backbone for describing the structure and function of various natural and artificial systems [1]. Prominent examples in mathematical chemistry consider chemical molecules as graphs of atoms and bonds [2]. As such, different graph structures, including acyclic, cyclic, and polycyclic graphs [3], along with various simple centrality measures in graphs, including

✉ Guang Hu
huguang@suda.edu.cn

[1] Center for Systems Biology, Department of Bioinformatics, School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, China

[2] Jiangsu Province Engineering Research Center of Precision Diagnostics and Therapeutics Development, Soochow University, Suzhou 215123, China

Published online: 04 September 2023

centrality based on “resistance distance” [4] and cumulative centrality index [5], are able to rank molecular chemical graphs, capture chemical similarity, and further quantify structure-activity relationships.

Benefiting from the advances of graph theory and high-throughput biomedical technologies, recent years have seen the emergence of a new paradigm- systems biology- as an advanced graph theory to study biological systems [6]. From the viewpoint of systems biology, biological systems are considered as complex networks, such as protein structure networks [7], protein-protein interaction networks [8], drug-target networks [9], co-expression networks [10]. Specifically, a complex network refers to the connectivity pattern between elements, and a “complex node” refers to the nonlinear behavior of individual elements [11]. These networks are graph-theoretical constructs composed of nodes and edges that aim to describe the integrated state of a biological system. For example, the protein structure graph (PSG) is considered as an emerging paradigm in chemistry, lying somewhere between mathematical chemistry and systems biology [12].

Similar to chemical structure networks as regular graphs, the structure of biological networks is also not random but forms a specific distribution. Network analysis using graph theory principles offers insights into biological problems by different measures, covering proteins and other biological entities, such as diseases and cell lines [13]. Moreover, biological networks are always organized into modular structures (subgraphs) to perform biological functions in a collective manner. For example, the violin model based on graph theory makes use of graph-based measure calculation and subgraph analysis to understand information flow through PSGs [14]. At the protein-protein interaction network (PPIN) level, the usage of topographic measures can help to predict disease markers and potential drug targets [15], and their modular intrinsic property can guide the search for elucidation underlying disease relationships by the construction of multiscale networks [16, 17] and suggest *drug repurposing* by considering *diseases* as perturbations of functional communities [18].

In this review, we will give the basic knowledge of graph theory in systems biology: network centralities and topological structures. For applications, the development in our group will be introduced for quantifying protein allostery to first, find the signal transmission routes based on protein structures and dynamics, and second, identify key proteins in PPINs as potential drug targets. The review will end with immediate challenges in the field.

2 Preliminaries

This section, therefore, gives preliminaries of network measures and network topologies that have been adopted in studying protein structures and interactions. Biological networks are different graphs $G = (V, E)$, whose *nodes* or *vertices* V are defined as biological entities (residues, genes, proteins, or drugs), and their *links* or *edges* E are defined as (physical, biochemical, or functional) interactions [19, 20]. Commonly, the topological structure and dynamical properties of a biological network can be analyzed by graph theory-based centralities in terms of three different levels, and the most comprehensive centrality measures can be found at: https://www.centiserver.org/.

First, at the node-level, some commonly used centralities have been proposed [21], including degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), clustering coefficient (C), and eigenvector centrality (EC), and so on. These node-level centralities in PSGs can be used to ascertain the residue-wise contribution to allosteric communications [22], and have become a principal method for identifying essential proteins in PPINs [23].

Second, pairwise features for edges E, such as mutual neighbors (whether a source and target are connected via an edge), were calculated to predict important interactions at the edge-level. For instance, six pair-wise features were integrated into an artificial intelligence-based model to predict synthetic lethal interactions as conserved patterns in PPINs [24].

Given a network G, a community M is defined as a set of distinct nodes:$M = \left\{v_1, v_2, \cdots, v_n\right\}$, where n is the number of nodes in the network. Sometimes, communities in a network are also called subgraphs within a graph. Many methods have been proposed to identify communities or subgraphs in biological networks, such as fast greedy, matrix-eigenvector-based, edge-betweenness-based, and land multilevel modularity optimization algorithms [25, 26]. Subgraph mining provides a method to detect functional communities for both protein structures and interactions [27].

Third, only a few community-level centralities were developed, and the most commonly used is the modularity Q, which measures the fraction of edges in the network connecting vertices within the same community and then subtracts from this fraction its expected value in a network with the same partition scheme over randomly connected nodes [28]. Recently, $\mathrm{CR}_{\mathrm{ANK}}$ is a matrix that can efficiently evaluate the robustness and magnitude of the structural features of each community and then combines these features (community likelihood, community density, community boundary, and community allegiance) to prioritize network communities [29]. From the biological perspective, important communities might be functional modules or pathways, which can represent sets of proteins or genes involved in a specific cellular process [30].

Regarding the global topology, biological networks are normally classified into two types: small-world and scale-free networks, which are defined as follows.

Definition 1 *A small-world network is defined as a graph G (V, E) rewired from a ring lattice with its average distance increasing logarithmically (*Fig. 1*), while the network size n and the average clustering coefficient (*$\langle C\rangle$*) is significantly large* [31].

Specifically, PSGs belong to small-world graphs, in which most nodes are not nearest neighbors but are linked by short paths on the network. Therefore, the small world properties of PSGs show that the efficient and functional flow of information within a protein is related to its allosteric regulation properties [12].

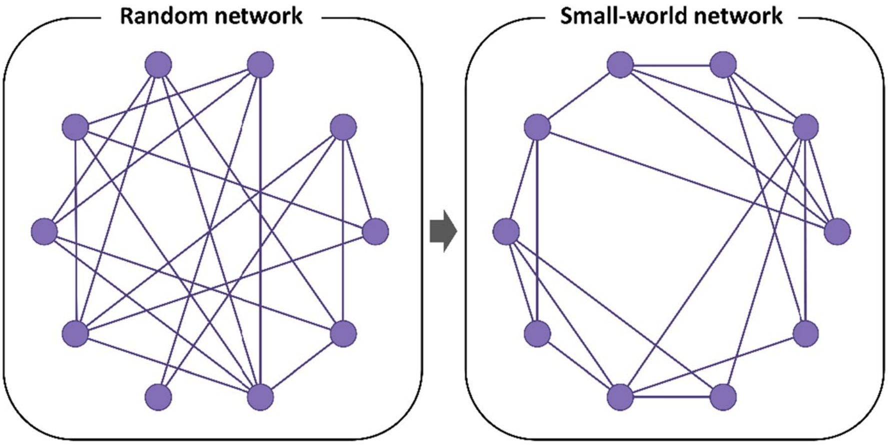

Fig. 1 A random network and its rewired small-world network, with the same node numbers but different network topologies

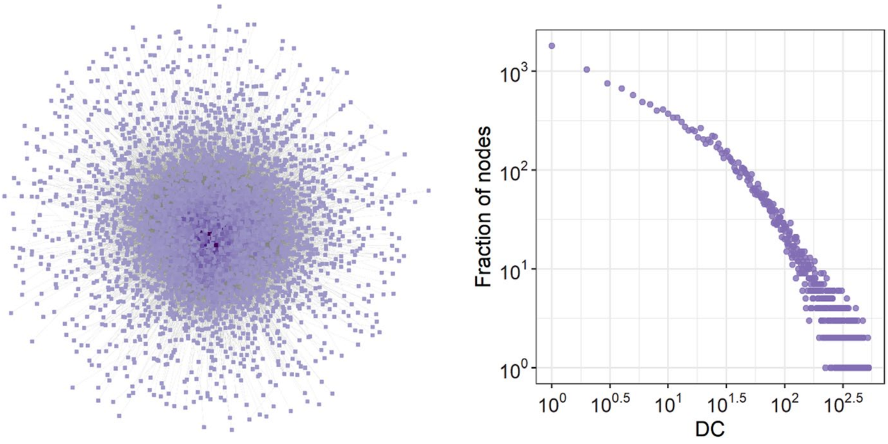

Fig. 2 A PPIN and its degree distribution, showing the scale-free property

Definition 2 *For a complex network, the degree of a node v_i(denoted DC_i, mathematically, is defined as the number of links to v_i, with $a_{i,j}$being the affinity or interaction strength of a link. Many biological networks exhibit the scale-free* [32] *(or scale-free-like) characteristic, where the node degree follows a power-law distribution, and the degree distribution P(DC) follows, where λ is called the degree exponent.*

Biological networks, including PPINs, are always scale-free (Fig. 2). In this kind of graph, many genes or proteins have few links and sporadic nodes with a more significant number of interactions, and the network hubs tend to be the essential parts, which would be critically important in disease progression, cellular dysfunction, and so on [33].

More detailed definitions of topological descriptors and network types can be found in an explorative study [34], which help us to understand the organization principles of biological networks.

3 Development of methodology

In this section, we briefly introduce the recent advances in developing new graph theory-based methods of protein structures and interactions in our group.

Definition 3 *The amino acid contact energy network (AACEN) is a kind of PSG, in which nodes (V) are residues and edges (E) exist when the environment-dependent residue contact energy (ERCE) is smaller than zero* [35]. *Mathematically, AACEN is defined based on the adjacent matrix AM as follows*:

$$AM_{ij} = \begin{cases} 0, e_{ij} \geq 0 \\ 1, e_{ij} < 0 \end{cases} \tag{1}$$

and

$$e_{ij} = -\ln\left(\frac{N_{ij}N_{00}}{N_{i0}N_{j0}}\frac{C_{i0}C_{j0}}{C_{ij}C_{00}}\right) \tag{2}$$

where e_{ij} represents the contact energy between vertices i and j, N_{ij}, N_{i0}, N_{j0}, and N_{00} are the contact numbers from the known structures, and C_{ij}, C_{i0}, C_{j0}, and C_{00} are the corresponding parameters expected in a reference state.

Definition 4 *Based on AACEN, the node-weighted amino acid contact energy network (NACEN) was further defined when the node in AACEN weighted by six features, namely, relative solvent accessibility, mass, hydrophobicity, polarity, flexibility, and JSD conservation score* [36].

Definition 5 *For a PPIN (G), the RN (SVM-RFE and Network topological) score was defined to rank the importance of protein nodes (V) in the PPIN* [37].

First, the SVM recursive feature elimination (SVM-RFE) score (R_s) was calculated according to the following formula:

$$R_s = \frac{1 + n - r_i}{n} \tag{3}$$

where n is the number of proteins and r_i is the rank of protein i.
Second, the RN score for each protein was determined by incorporating two commonly used network parameters, the degree DC, and the average shortest path length, with R_s as the following formula:

$$RNs = \frac{DC * R_s}{L} \tag{4}$$

where L is the average shortest path length between this protein and all other proteins in the network.

Definition 6 *The topological-functional connection (TFC) score was proposed to rank the PPI edges (E) by integrating the edge betweenness and gene ontology (GO) semantic similarity* [38], *defined as*

$$TFC = \sum\nolimits_{e=1}^{N} \frac{T_e^* + F_e}{|T_e^* + F_e - 2|} * 100 \tag{5}$$

$$T_e^* = \frac{T_e - Min_T}{Max_T - Min_T} \tag{6}$$

where N is the number of interactions, and T_e and F_e represent the edge betweenness and gene ontology (GO) semantic similarity of interaction E.

Definition 7 *Node and Edge Prioritization-based Community Analysis (ne-PCA)* [38] *is defined as a network modularization method by detecting functional models based on robust communities based on Girvan-Newman (GN)* [39] *and Label Propagation Algorithm (LPA)* [40] *that was applied to the PPIN weighted by TFC.*

Definition 8 *The Proximity Score of Vertex to Network (PS-V2N)* [41] *is another network parameter to prioritize the importance of a node in one set affecting another set in the community network. The PS-V2N of vertices $v_i$is defined as*:

$$PS\left(v_i, B\right) = \frac{1}{n-1}\left(\sum_{j=1, i \neq j}^{n} \frac{1}{d\left(v_i, v_j\right)} + \frac{1}{(r+1) * M_G} + \delta\right) \tag{7}$$

$$\delta = \begin{cases} 0, if v_i \notin A \cap B \\ \frac{1}{|B| * \sum_{s,t \in B} d\left(v_s, v_t\right)}, if v_i \in A \cap B \end{cases} \tag{8}$$

where $v_i \in A$ and $v_j \in B$, n is the number of vertices in subgraph M, $d(v_i, v_j)$ is the shortest path length between v_i and v_j, and M_G is the diameter of M. As such, the score measures the influence of vertices in set B on set A based on the score of the shortest path from set A to set B in the community network.

All codes for methods of definitions 3–8 are listed in Table 1.

The methods described above have been applied to study protein structures, dynamics, and interactions: (1) the comparison of protein structure and functions with different topologies, (2) the identification of key residues and pathways involved in allosteric regulation, (3) the prediction of disease mutations and post-translational modifications (PTMs) based on machine learning of network

Table 1 Some graph theory-based methods developed in our group

Methods	Networks	Links	References
NACEN	PSG	http://sysbio.suda.edu.cn/NACEN/index.html	[35]
AACEN	PSG		[36]
RN	PPIN	https://github.com/CSB-SUDA/RNs	[37]
TFC	PPIN	https://github.com/CSB-SUDA/ne-PCA	[38]
PS-V2N	PPIN	https://github.com/CSB-SUDA/PS-V2N	[41]

parameters, and (4) the analysis of PPINs based on the three levels of newly developed centralities, that is, node-based, subgraph-based, and pathway-based approaches.

4 Applications of graph theory in studying protein topology and dynamics

Using graph theory-based approaches toward understanding the structure-function relationship in proteins began in the late 1990s [42]. Since then, several types of network models for protein structures have been proposed, including protein structure networks [43], protein contact networks [44], and amino acid networks [45]. Here, we use the name of PSGs. With the development of AlphaFold2, a large amount of protein structure data would bring great challenges to the field [46]. Of note, protein structures are dynamic, while a single conformation is a static 3D reconstruction, resulting in a partial representation of physiological and/or disease dynamics. Combined with traditional protein dynamic methods, such as molecular dynamic (MD) simulation and normal mode analysis, network-based modeling has also provided an array of powerful tools to quantify protein dynamics, and further dissection of underlying mechanisms of protein function and interactions in the cellular environment, such as allosteric regulatory mechanisms in protein systems [47].

4.1 Protein topological comparison

Topology is an essential aspect of protein structure [48, 49], and protein structures adopt discrete topologies with globular, cylindrical (transmembrane proteins) or toroidal shapes (Fig. 3). Thus, a structure-based network is a useful tool to describe and compare the topologies of proteins, which provide insight into protein functions.

By focusing on two protein examples with cylindrical and toroidal topologies [50], we first employed characteristic path lengths, clustering coefficients, and diameters to investigate their global topology parameters such as small-world properties and packing density, as well as used network centralities for the subgraph of the hydrophobic pocket in Dronpa to describe its detailed topology and the photoswitching activity of the chromophore.

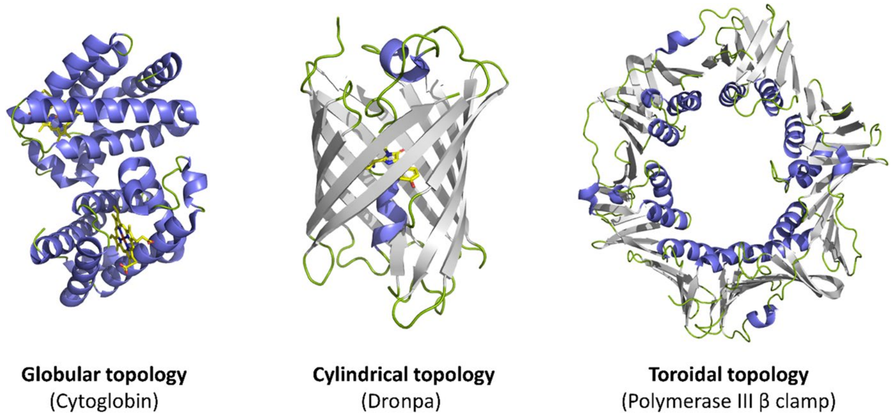

Fig. 3 Protein structures for Cytoglobin (PDB code: 1UMO) with globular topology, Dropa (PDB code: 2IE2) with cylindrical topology, and Polymerase III β clamp (PDB code: 2POL) with toroidal topology

AACENs (Definition 3) are constructed and extended to large protein families with three different topologies [51]. The network comparison results showed that globular proteins have the highest network density, average closeness, and system vulnerability, while toroidal proteins have the lowest values of these parameters. Transmembrane proteins are found to have significantly higher assortativity values than globular and toroidal proteins. In addition, the ratio $< C > / C_{random}$ indicated that the toroidal topology is the most correlated with small-world properties, compared with the cylindrical topology.

By extracting subgraphs only including interfacial residues between different chains, it may provide a simple but straightforward method to identify hot spots of toroidal proteins. Therefore, a deeper inspection of the topology-dynamics-function relationships of proteins is needed. The investigation of how the toroidal topology adapted its collective motion to control substrate binding is such a work toward this aim [52].

4.2 Protein allosteric regulation

As the “second secret of life”, allosteric regulation of proteins describes the biological process of signaling transformation within a protein through mutations, PTMs, and small molecule binding [53]. Graph theory-based methods are increasingly being used to understand information flow through the construction of PSGs. In the past few years, we have applied several conformation-based network descriptors to capture network signaling efficiency and explain allostery in terms of signal transmission, for several important protein systems (Fig. 4).

DNA methyltransferases (DNMTs) not only play key roles in epigenetic gene regulation, but also serve as emerging targets for several diseases, especially cancers. Allosteric regulation of DNMTs continues to be an engaging research topic for the scientific community [54]. In the first case, we applied the protein structure network method to investigate the intrinsic dynamics and allosteric properties

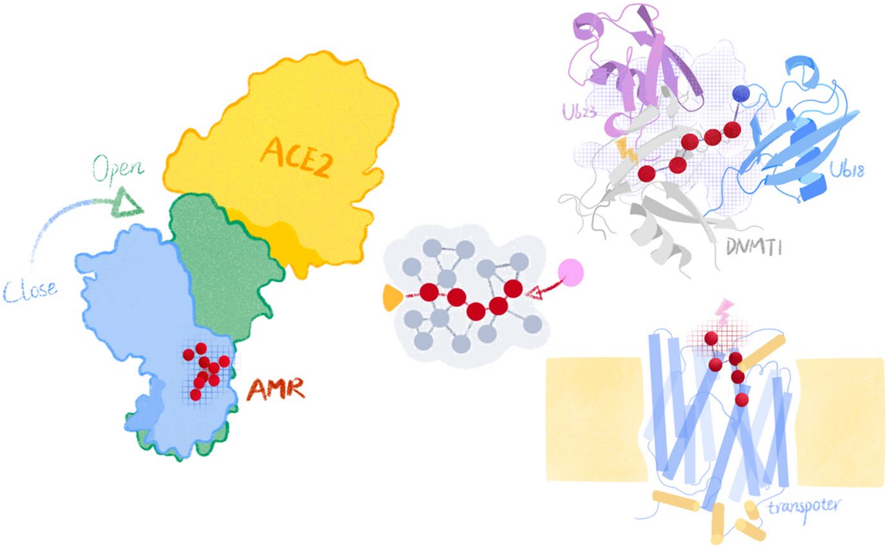

Fig. 4 The PSG based on a protein structure provides a network model to study allosteric regulations through the identification of allosteric sites and the quantification of signal transmission

of DNMT3A resolved in autoinhibitory and active states [55]. The betweenness centrality highlights the critical residue pairs for inter-subunit communication, suggesting the pivotal role of the dimer interface in interdomain allostery. For the DNMT1 monomer, the integration of PSG and MD analysis revealed both intra and inter-domain allosteric communications [56]. In particular, the TRD interface along the inter-domain communication pathways was predicted to be an allosteric pocket for further drug targeting. Then, we extended the network modeling method to examine the allosteric regulation of DNMT1-ubiquitylated H3 (H3Ub)/ubiquitin specific peptidase 7 (USP7) complexes, and some more PPIN-based communication pathways and allosteric pockets were found [57].

The second case is the SARS-CoV-2 spike protein (S protein), which is the most promising protein target for COVID-19. Although there are a large number of studies on the molecular modeling, simulations, and predictions of potential sites of the S protein for targeting drugs, vaccines, and antibodies [58], the study of allosteric regulation of the S protein may provide new insight into further drug discovery. Based on the structure of the S protein complex with angiotensin-converting enzyme 2 (S-ACE2 complex), subgraph analysis using a PSG predicted a specific region located at the fusion peptide that can act as an allosteric modulation region (AMR) [59]. Applying the same subgraph analysis to the S-ACE2 complex in both closed and open states, the position of AMR is conserved but shows a higher drug response in the open state, suggesting the importance of AMR in maintaining the S protein in the closed state [60]. Therefore, the detection of AMR can be used to elucidate the molecular pathways that can be targeted with allosteric drugs to weaken the S-ACE2 interaction and, thus, reduce viral infectivity.

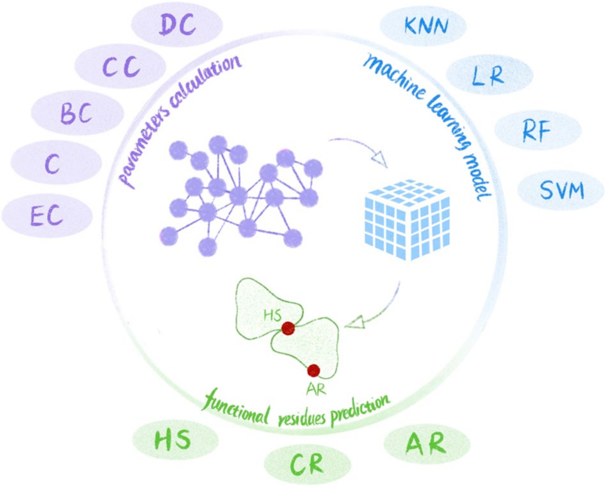

Fig. 5 Framework of learning PSG-based features by different MLs for predicting functional sites

The third case is transport proteins. The special structural feature of transport proteins is that using tunnels to transport ligands and perform molecular functions. We have recently applied three types of PSGs to quantify allosteric regulation regarding tunnels. In the first type of PSG for CYP17A1, nodes are C_β (C_α for glycine) atoms of each residue connected by edges within the cutoff distance of 6.5 Å. Two network parameters, the shortest path length (*L*) and *BC* of each residue, were calculated based on MD simulation to quantify allosteric regulation of tunnels upon different ligand binding [61]. In the second type of PSG for the putative pentose transporter (CCM_06358 gene), the topology is based on individual amino acids as nodes that are connected by edges over the non-covalent interaction between the main chain (*mc*) and side chain (*sc)*, including van der Waals contacts, hydrogen bonds, overlaps of van der Waals radii, and a combination of any of the previous three interactions [62]. AACENs for 85 sugar transporter proteins were further constructed [63]. In all three types of PSGs, BC was found to be a good network indicator for identifying allosteric sites and thus describing allosteric communication pathways.

Overall, based on graph theory, network model analysis based on network parameters and subgraph analysis has been applied to quantify allosteric regulation to unravel the workings of different protein systems, including DNMTs, S protein, and transport proteins.

4.3 Machine learning models

The development of network parameters provides a wealth of raw materials for further understanding biological functions [64]. Leveraging network-based descriptions in machine learning (ML) frameworks can facilitate functional inferences, and assessing pathogenicity or predicting functional PTMs is an exciting frontier [65]. We list below such recent developments in our group (Fig. 5).

To predict functional residues including hot spots, catalytic residues, and allosteric residues, machine learning predictors were built by the following steps [36]. *Step (1)* NACENs (Definition 4) for proteins/protein complexes were constructed, and then a total of 21 network parameters were calculated. *Step (2)* Utilizing these parameters as input features, the performance of five ML models was compared, including support vector machine (SVM), random forest (RF), logistic regression (LR), and K nearest neighbors (KNN) models. *Step (3)* The best ML predictors for specific functional residues were selected based on performance. Based on network parameters from AACEN (Definition 3), the relationship between the genotype of mutations and the disease phenotypes was established [66]. The RF model indicated that the change in BC is the best indicator for classifying mild and severe ALPL mutations in hypophosphatasia. According to the biological meaning of BC, we suggested that severe ALPL mutations may have long-range allosteric effects through the protein-protein interface, which has been validated by the MD simulation.

As another important functional sites, PTM sites play a particularly important role in signal transformation, and thus regulate cellular function and disease pathogenesis. For the prediction of PTMs and the estimation of their functions, in addition to the above ML models, some artificial intelligence (AI)- based models were employed. A fully connected neural network (FCNN) deep learning model with network parameter features could predict PTMs in the kinase family [67]. By applying the method to the case of c-Src kinase, a potential druggable PTM pocket and its covalent inhibitor DC-Srci-6668 were successfully detected. Furthermore, two network parameter-based deep learning models, namely, cDL-PAU and cDL-FuncPhos, were proposed to elucidate the molecular basis and underlying functional landscape of PTMs [68]. cDL-PAU achieved satisfactory area under the curve (AUC) scores of 0.804–0.888 for predicting phosphorylation, acetylation, and ubiquitination (PAU) sites, while cDL-FuncPhos achieved an AUC value of 0.771 for predicting functional phosphorylation (FuncPhos) sites, displaying reliable improvements. By extending network parameter features in the Pre-W net to study PTM cross-talk, PPICT as a novel integrated deep neural network was proposed and showed good performance [69].

In summary, these graph-based descriptors provide a scalable and intuitive means to capture complex relationships between residues and have shown state-of-the-art performance in functional residue prediction.

5 Applications of graph theory in studying PPINs

In the post-genomic era, proteomics has achieved significant theoretical and practical advances with the development of high-throughput technologies. In particular, the rapid accumulation of protein-protein interaction data provides a foundation for constructing PPINs, which can furnish a new perspective for understanding cellular organizations, processes, and functions using graph theory [70]. In our group, we have also developed PPIN-based methods on three levels (Fig. 6): nodes, subgraphs, and pathways, as well as demonstrated their prospective utilities.

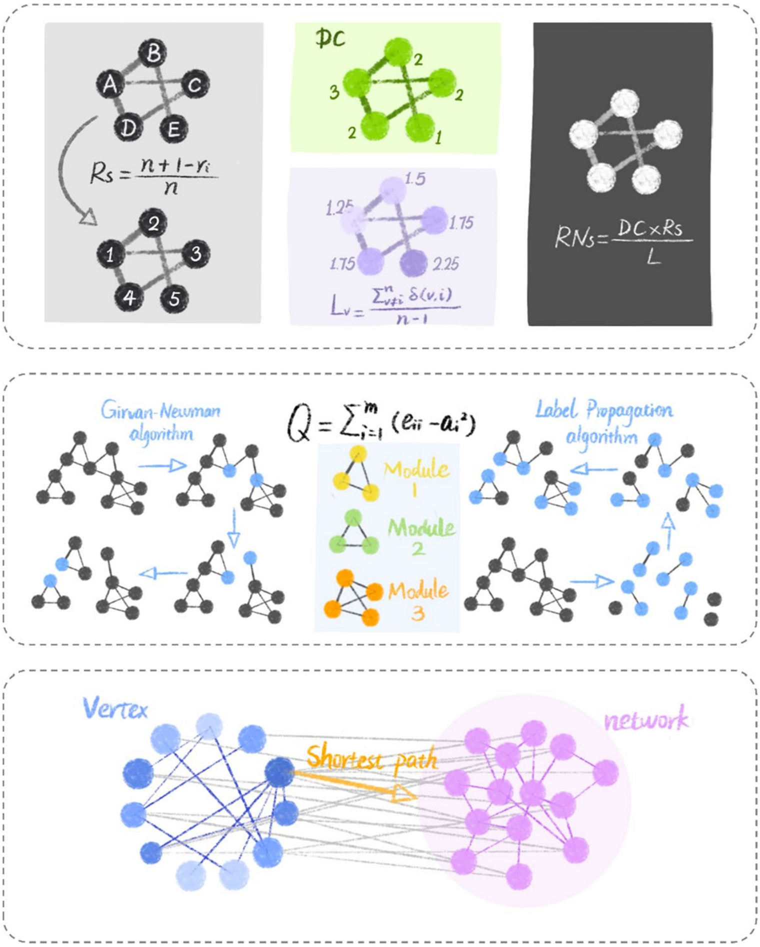
B
A
C
D
E
$R_s = \frac{n+1-r_i}{n}$
2
1
3
4
5
DC
2
3
2
2
1
1.5
1.25
1.75
1.75
2.25
$L_v = \frac{\sum_{v \neq i}^{n} \delta(v,i)}{n-1}$
$RNs = \frac{DC \times Rs}{L}$
Girvan-Newman algorithm
$Q = \sum_{i=1}^{m} (e_{ii} - a_i^2)$
Label Propagation algorithm
Module 1
Module 2
Module 3
Vertex
Shortest path
network

Fig. 6 Three levels of PPIN analysis: the node level (*RNs*), the subgraph level, and the pathway level (*PS-V2N*).

5.1 Node centrality

Conventionally, PPINs can be modeled via graphs whose nodes represent proteins and whose edges connect pairs of interacting proteins. Depending on the scale-free topology of PPINs, a small number of proteins called "hubs" are most important for maintaining its global connectivity; this is also known as the centrality-lethality rule [71]. However, the limitation is that the current methods for the analysis of PPIN just by pure topological properties. To improve the performance, we developed a new node centrality (Definition 5), called *RNs* [37], which integrates gene expression and network topological information to identify disease-associated genes. As such, in addition to degree *DC* and the shortest path length *L*, the new node centrality not only considers the topological information of the PPIN, but also includes the cancer status of each gene. Thus, *RNs* are node centrality that have potential biological meaning.

By applying this method to analyze three independent expression microarray datasets of pancreatic ductal adenocarcinoma (PDAC) patients, 17 nodes were first predicted as disease genes. The following "druggability" prediction and protein-drug interactions suggested that two integrins, ITGAV and ITGA2, are two potential drug targets in PDAC. More recently, the *RNs* centrality has been applied to predict key genes as biomarkers for gastric cancer progression. Importantly, the biological functions of key genes were further confirmed by a series of experimental studies and clinical data [72].

5.2 Subgraph analysis

Except for nodes, the topological representation of a PPIN is useful for the prediction of novel interactions and the identification of functional modules. In principle, if two or more proteins have many common partners in the network they tend to function in similar biological processes. Mathematically, subgraph analysis helps in finding functional modules or communities.

Based on PPINs, we proposed a knowledge-guided and network-based integration method, called the node and edge Prioritization-based Community Analysis (ne-PCA), to identify functional modules [38]. For the construction of a PPIN, two kinds of nodes are needed. One is seed genes for a particular disease collected from databases, and the other is analyzed from omics data. Using seed nodes as references, node scores were assigned to other nodes by performing a random walk with restart algorithm. In this method framework, two subgraph analysis were performed including ranking edges based on *TFC* (Definition 6*)* and detecting communities by combining the Girvan-Newman (GN) algorithm and Label Propagation analysis (LPA). *TFC* is a score that can be used to identify key protein interactions by integrating network topology (edge betweenness) and biological characteristics (Gene Ontology), which supplement missing functions in traditional network information flow.

This method has been used to analyze two independent expression microarray datasets for non-small-cell lung cancer (NSCLC). The subgraph analysis detected two important functional modules: the CCNB1-mediated subgraph in the largest community provides a modular biomarker and the second community serves as a drug regulatory module. By the *TFC* score ranking, some PPIs including GNG11-CXCR2, GNG11-CXCL3, and GNG11-PPBP were suggested as candidate therapeutic targets for treating NSCLC. Finally, the functional significance of key genes and important subgraphs was further verified by published experimental data and protein structural modeling.

The subgraph analysis can also be extended to the dynamic process of disease, such as cancer progression. To this aim, the detection and comparison of subgraphs from PPINs associated with different disease states highlighted key factors in their dynamic processes [73]. The comparison of subgraphs, also called "differential modular analysis", can help to detect inter-modular edges, which are important PPIs for network rewiring. By applying the method to study the lymph node metastasis (LNM) process in breast cancer, functional enrichment analysis showed that inter-modular edges play important roles in cancer metastasis and invasion, and in metastasis hallmarks. Considering date hubs as nodes and inter-modular edges as new edges, an important subgraph that contained most of the dynamic properties of LNM was constructed. We expect that the analysis of such a dynamic subgraph will provide more applications in the understanding of complex diseases and the identification of therapeutic targets.

5.3 Integrative network analysis

Elucidating the complex relationships of biological networks at local levels is not enough; thus, we proposed an integrated network analysis (INA) pipeline for generating and analyzing comprehensive networks from different "Omics Data". Network analysis is implemented from the local level to the global level by integrating our developed methods. In the first step, network construction can be an alternative, using PPI interaction data from various databases such as STRING as edges [74], and the nodes can be DEPs, proteins ranked by *RNs*, or collected seed genes. As such, along the network construction, the node level-based analysis was also performed. The second step contains two levels of analysis, that is, the ranking of PPIs based on *TFC* score on the edge level and the identification of functional modules based on subgraph level analysis. Moreover, at the pathway level, *PS-V2N* (Definition 8) was employed to find the topological shortest pathway between one kind of nodes (or studied genes/proteins) to seed genes, and then prioritize nodes with a new shortest path-based score [41]. INA is freely available at https://github.com/CSB-SUDA/INA.

We applied integrative network analysis to reveal the molecular mechanism of primary and recurrent thymic epithelial tumors (TETs) based on proteomics data [75]. Three kinds of PPINs were constructed: the primary PPIN including 1016 nodes, 9 differentially expressed proteins (DEPs) and 9557 PPI edges, the recurrent TET network based on DEPs only including 13 nodes and 17 edges, and the recurrent TET network based on RNs containing 50 nodes and 654 edges. Then, the subgraph combined with functional enrichment analysis revealed that primary and recurrent TETs shared certain common molecular mechanisms, including a spliceosome module consisting of RNA splicing and RNA processing, but the recurrent TET was specifically related to the ribosome pathway. Finally, by applying the PS-V2N algorithm from the node in the identified subgraphs to the seed module, the ribonucleoprotein hnRNPA2B1 was identified as a key node that may serve as a potential target for recurrent TET therapy.

In summary, we have demonstrated an application of using integrative network analysis to identify drug targets. We expect our integrative method to be extended to study more disease biology questions, particularly the understanding of disease progression, clinical subtypes, and their genotype-phenotype relationship.

6 Future challenges

Biological systems are by nature multiscale, consisting of subsystems that factor into progressively smaller units in a deeply hierarchical structure. At any level of the hierarchy, a multilayer network paradigm can be applied to characterize the corresponding biological units and their relations, resulting in large networks of physical or functional proximities, e.g., proximities of amino acids within a protein, of proteins within a complex, or of cell types within a tissue [76]. Although recent advances in studying the structure and dynamics of such multilayer networks have enabled a better understanding of the complexity in biological systems [77], only a minuscule fraction of this new domain has been explored.

In 2021, Buphamalai et al. [16] proposed that there are six major biological scales between genotype and phenotype. Different edges in each layer are defined. Edges are defined by genetic interactions at the genetic scale. At the transcriptome scale, interactions represent co-expression, that is, co-variability of gene transcription levels indicates higher-level regulatory mechanisms. At the proteome and pathway scales, links represent physical interactions between gene products and pathway co-membership derived from databases, respectively. At the biological processes and molecular function level, edges are biological processes and molecular functions derived from the Gene Ontology. At the top level, the phenotypic scale, links represent similarity in annotated phenotypes derived from the Mammalian and Human Phenotype Ontologies. This work opens a door for studying the relationship between genotype and phenotype by the method of network modeling of biological organization at various levels. One popular approach is to treat multilayer networks as different "features", and then use machine learning models to study the statistical associations among features and to predict any range of biological outputs [78]. However, most machine-learning approaches are ‘‘black boxes’’, and thus, the complex causal relationship between different features from which level and the outside of biology is uninterpretable. The bottom-up development of a theoretical method is needed.

It is now clear that multilayer networks (or "networks of networks", Fig. 7) are new forms of biological systems [79], and we propose future challenges in this area that deserve special encouragement, especially in the mathematical context.

First, two network models were mainly introduced for the study of protein structures, dynamics, and interactions. Nodes in PSGs and PPIs are amino acids and proteins, while the definition of their edges is more diverse. In our previous works, we treated two network models independently, which is clearly a limitation. How to integrate PSGs and PPINs to build a multilayer network and find some shared network parameters between two levels is a long question in our group.

A second open question regards the analysis of multilayer networks. Based on the PPI network, we proposed an integrative network analysis including node, subgraph, and pathway analysis. These kinds of analyses are also a useful topological way to mine information from multilayer networks. As we know, however, the alignment of hierarchies of nodes, edges, and subgraphs has its own set of challenges.

Some multilayer networks have been constructed from omics data. For example, two notable multiscale maps of protein systems are proposed. NeST is a

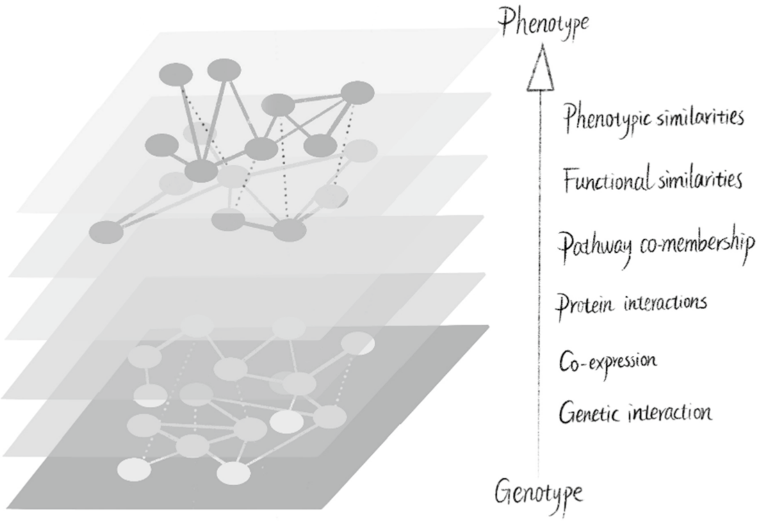

Fig. 7 The architecture of multilayer networks linking genotype to phenotype

comprehensive map of cancer protein systems integrating cancer mutations and multi-omic interaction data at multiple scales [80]. MuSIC 1.0 is another hierarchical map of human cell integrating immunofluorescence images in the Human Protein Atlas and PPI data in BioPlex [81]. How to develop and use graph theory-based methods to analyze biological systems at multiple scales to approach their underlying functions is a very direct challenge.

All in all, we now envision a new scenario of how to extend graph theory in mathematical chemistry to systems biology, by quantifying complex biological systems in terms of not simple arithmetic operations in “networks of networks”.

Author contributions ZZ and GH: wrote the main manuscript text and prepared all figures. Both authors reviewed the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (32271292, 31872723), and a Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

References

1. P. Velickovic, Curr. Opin. Struct. Biol. **79**, 102538 (2023)
2. P. Csermely, T. Korcsmaros, H.J.M. Kiss, G. London, R. Nussinov, Pharmacol. Ther. **138**, 333 (2013)
3. M. Randic, M. Novic, M. Vracko, D. Plavsic, J. Comput. Chem. **34**, 2514 (2013)
4. D.J. Klein, J. Math. Chem. **47**, 1209 (2010)
5. P. Nirmala, R. Nadarajan, J. Mol. Struct. **1247**, 131354 (2022)
6. C. Liu, Y.F. Ma, J. Zhao, R. Nussinov, Y.C. Zhang, F.X. Cheng, Z.K. Zhang, Phys. Rep. **846**, 1 (2020)
7. W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, B. Shen, Amino Acids. **46**, 1419 (2014)
8. W. Yan, D. Zhang, C. Shen, Z. Liang, G. Hu, Curr. Top. Med. Chem. **18**, 1031 (2018)
9. A. Badkas, S. De Landtsheer, T. Sauter, Brief. Bioinform. **22**, bbaa357 (2021)
10. A. Savino, P. Provero, V. Poli, Int. J. Mol. Sci. **21**, 24 (2020)
11. R.M. D'Souza, M. di Bernardo, Y.Y. Liu, Nat. Rev. Phys. **5**, 250 (2023)
12. L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, A. Giuliani, Chem. Rev. **113**, 1598 (2013)
13. A. Fernandez-Torras, A. Comajuncosa-Creus, M. Duran-Frigola, P. Aloy, Curr. Opin. Chem. Biol. **66**, 102090 (2022)
14. I. Rivalta, V.S. Batista, Methods Mol. Biol. **2253**, 137 (2021)
15. C. Fotis, A. Antoranz, D. Hatziavramidis, T. Sakellaropoulos, L.G. Alexopoulos, Drug. Discov. Today. **23**, 626 (2018)
16. P. Buphamalai, T. Kokotovic, V. Nagy, J. Menche, Nat. Commun. **12**, 6306 (2021)
17. C. Ruiz, M. Zitnik, J. Leskovec, Nat. Commun. **12**, 1796 (2021)
18. S. Sadegh, J. Skelton, E. Anastasi, J. Bernett, D.B. Blumenthal, G. Galindez, M. Salgado-Albarran, O. Lazareva, K. Flanagan, S. Cockell, C. Nogales, A.I. Casas, H. Schmidt, J. Baumbach, A. Wipat, T. Kacprowski, Nat. Commun. **12**, 6848 (2021)
19. Y. You, X. Lai, Y. Pan, H. Zheng, J. Vera, S. Liu, S. Deng, L. Zhang, Signal. Transduct. Target. Ther. **7**, 156 (2022)
20. M. Recanatini, C. Cabrelle, J. Med. Chem. **63**, 8653 (2020)
21. N.T. Doncheva, Y. Assenov, F.S. Domingues, M. Albrecht, Nat. Protoc. **7**, 670 (2012)
22. L.K. Madan, C.L. Welsh, A.P. Kornev, S.S. Taylor, J. Chem. Phys. **158**, 081001 (2023)
23. M. Ashtiani, A. Salehzadeh-Yazdi, Z. Razaghi-Moghadam, H. Hennig, O. Wolkenhauer, M. Mirzaie, M. Jafari, BMC Syst. Biol. **12**, 80 (2018)
24. G. Benstead-Hume, X. Chen, S.R. Hopkins, K.A. Lane, J.A. Downs, F.M.G. Pearl, PLoS Comput. Biol. **15**, e1006888 (2019)
25. S. Wu, D.J. Chen, M.P. Snyder, Curr. Opin. Chem. Biol. **66**, 102101 (2022)
26. A. Singhal, S. Cao, C. Churas, D. Pratt, S. Fortunato, F. Zheng, T. Ideker, PLoS Comput. Biol. **16**, e1008239 (2020)
27. T.K. Saha, A. Katebi, W. Dhifli, M.A. Hasan, IEEE/ACM Trans. Comput. Biol. Bioinform. **16**, 1537 (2019)
28. S.A. Alcala-Corona, S. Sandoval-Motta, J. Espinal-Enriquez, E. Hernandez-Lemus, Front. Genet. **12**, 701331 (2021)
29. M. Zitnik, R. Sosic, J. Leskovec, Nat. Commun. **9**, 2544 (2018)
30. E.L. Huttlin, R.J. Bruckner, J.A. Paulo, J.R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M.P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A.E. White, D.K. Schweppe, R. Rad, B.K. Erickson, R.A. Obar, K.G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S.P. Gygi, Nature **545**, 505 (2017)
31. D.J. Watts, S.H. Strogatz, Nature. **393**, 440 (1998)
32. R. Albert, H. Jeong, A.L. Barabasi, Nature. **406**, 378 (2000)
33. G. Ruiz Amores, A. Martinez-Antonio, Funct. Integr. Genomics **22**, 1433 (2022)
34. H. Kohestani, A. Giuliani, Biosystems. **141**, 31 (2016)
35. W. Yan, M. Sun, G. Hu, J. Zhou, W. Zhang, J. Chen, B. Chen, B. Shen, J. Theor. Biol. **355**, 95 (2014)
36. W.Y. Yan, G. Hu, Z.J. Liang, J.H. Zhou, Y. Yang, J.J. Chen, B.R. Shen, J. Chem. Inf. Model. **58**, 2024 (2018)
37. W. Yan, X. Liu, Y. Wang, S. Han, F. Wang, X. Liu, F. Xiao, G. Hu, Front. Pharmacol. **11**, 534 (2020)

38. F. Wang, S. Han, J. Yang, W. Yan, G. Hu, Cells. **10**, 402 (2021)
39. M. Girvan, M.E. Newman, Proc. Natl. Acad. Sci. U. S. A. **98**, 7821 (2002)
40. U.N. Raghavan, R. Albert, S. Kumara, Phys. Rev. E. **76**, 036106 (2007)
41. J. Yang, H. Li, F. Wang, F. Xiao, W. Yan, G. Hu, ACS Chem. Neurosci. **12**, 917 (2021)
42. N. Kannan, S. Vishveshwara, J. Mol. Biol. **292**, 441 (1999)
43. V. Gadiyaram, S. Vishveshwara, S. Vishveshwara, J. Chem. Inf. Model. **59**, 1715 (2019)
44. L. Di Paola, A. Giuliani, Curr. Opin. Struct. Biol. **31**, 43 (2015)
45. L. Vuillon, C. Lesieur, Curr. Opin. Struct. Biol. **31**, 1 (2015)
46. W.Y. Yan, G. Hu, Curr. Bioinform. **17**, 493 (2022)
47. Z. Liang, G.M. Verkhivker, G. Hu, Brief. Bioinform. **21**, 815 (2020)
48. G. Hu, S. Michielssens, S.L. Moors, A. Ceulemans, J. Mol. Graph Model. **34**, 28 (2012)
49. G. Hu, S. Michielssens, S.L. Moors, A. Ceulemans, J. Chem. Inf. Model. **51**, 2361 (2011)
50. G. Hu, W.Y. Yan, J.H. Zhou, B.R. Shen, J. Theor. Biol. **348**, 55 (2014)
51. W.Y. Yan, G. Hu, B.R. Shen, Curr. Bioinform. **11**, 480 (2016)
52. H. Li, P. Doruker, G. Hu, I. Bahar, Biophys. J. **118**, 1782 (2020)
53. A.W. Fenton, Trends Biochem. Sci. **33**, 420 (2008)
54. Z. Liang, Y. Zhu, X. Liu, G. Hu, Adv. Protein Chem. Struct. Biol. **121**, 49 (2020)
55. Z. Liang, J. Hu, W. Yan, H. Jiang, G. Hu, C. Luo, Biochim. Biophys. Acta Gen. Subj. **1862**, 1667 (2018)
56. Z.J. Liang, Y. Zhu, J. Long, F. Ye, G. Hu, Comput. Struct. Biotechnol. J. **18**, 749 (2020)
57. Y. Zhu, F. Ye, Z.Y. Zhou, W.L. Liu, Z.J. Liang, G. Hu, Molecules. **26**, 5153 (2021)
58. K. Gao, R. Wang, J. Chen, L. Cheng, J. Frishcosy, Y. Huzumi, Y. Qiu, T. Schluckbier, X. Wei, G.W. Wei, Chem. Rev. **122**, 11287 (2022)
59. L. Di Paola, H. Hadi-Alijanvand, X. Song, G. Hu, A. Giuliani, J. Proteome Res. **19**, 4576 (2020)
60. H. Hadi-Alijanvand, L. Di Paola, G. Hu, D.M. Leitner, G.M. Verkhivker, P.X. Sun, H. Poudel, A. Giuliani, Acs Omega. **7**, 17024 (2022)
61. F. Xiao, X.Y. Song, P.Y. Tian, M. Gan, G.M. Verkhivker, G. Hu, J. Chem. Inf. Model. **60**, 3632 (2020)
62. K. Sirithep, F. Xiao, N. Raethong, Y. Zhang, K. Laoteng, G. Hu, W. Vongsangnak, Cells. **9**, 401 (2020)
63. X. Liu, H.Y. Zhang, Z.Y. Zhou, P. Prabhakaran, W. Vongsangnak, G. Hu, F. Xiao, Phys. Chem. Chem. Phys. **25**, 14311 (2023)
64. S. Jin, X. Zeng, F. Xia, W. Huang, X. Liu, Brief. Bioinform. **22**, 1902 (2021)
65. A. Banerjee, S. Saha, N.C. Tvedt, L.W. Yang, I. Bahar, Curr. Opin. Struct. Biol. **78**, 102517 (2023)
66. F. Xiao, Z. Zhou, X. Song, M. Gan, J. Long, G. Verkhivker, G. Hu, PLoS Comput. Biol. **18**, e1010009 (2022)
67. H. Zhang, J. He, G. Hu, F. Zhu, H. Jiang, J. Gao, H. Zhou, H. Lin, Y. Wang, K. Chen, F. Meng, M. Hao, K. Zhao, C. Luo, Z. Liang, J. Med. Chem. **64**, 15111 (2021)
68. F. Zhu, S. Yang, F. Meng, Y. Zheng, X. Ku, C. Luo, G. Hu, Z. Liang, J. Chem. Inf. Model. **62**, 3331 (2022)
69. F. Zhu, L. Deng, Y. Dai, G. Zhang, F. Meng, C. Luo, G. Hu, Z. Liang, Brief. Bioinform. **24**, 2 (2023)
70. X.M. Meng, W.K. Li, X.Q. Peng, Y.H. Li, M. Li, Front. Comput. Sci. **15**, 156902 (2021)
71. H. Ahmed, T.C. Howton, Y.L. Sun, N. Weinberger, Y. Belkhadir, M.S. Mukhtar, Nat. Commun. **9**, 2312 (2018)
72. W. Yan, Y. Chen, G. Hu, T. Shi, X. Liu, J. Li, L. Sun, F. Qian, W. Chen, J. Transl Med. **21**, 163 (2023)
73. X. Liu, B. Yang, X. Huang, W. Yan, Y. Zhang, G. Hu, Interdiscip Sci. Comput. Life Sci. (2023). https://doi.org/10.1007/s12539-023-00568-w
74. D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N.T. Doncheva, J.H. Morris, P. Bork, L.J. Jensen, C.V. Mering, Nucleic. Acids. Res. **47**, D607 (2019)
75. Z.Y. Zhou, Y. Lu, Z.T. Gu, Q.L. Sun, W.T. Fang, W. Yan, X. Ku, Z.J. Liang, G. Hu, Comput. Biol. Med. **155**, 106665 (2023)
76. L.V. Schaffer, T. Ideker, Cell. Syst. **12**, 622 (2021)
77. S. Chaudhuri, A. Srivastava, J. Biosci. **47**, 55 (2022)
78. D.M. Camacho, K.M. Collins, R.K. Powers, J.C. Costello, J.J. Collins, Cell. **173**, 1581 (2018)
79. V.N. Uversky, A. Giuliani, Front. Genet. **12**, 706260 (2021)

80. F. Zheng, M.R. Kelly, D.J. Ramms, M.L. Heintschel, K. Tao, B. Tutuncuoglu, J.J. Lee, K. Ono, H. Foussard, M. Chen, K.A. Herrington, E. Silva, S.N. Liu, J. Chen, C. Churas, N. Wilson, A. Kratz, R.T. Pillich, D.N. Patel, J. Park, B. Kuenzi, M.K. Yu, K. Licon, D. Pratt, J.F. Kreisberg, M. Kim, D.L. Swaney, X. Nan, S.I. Fraley, J.S. Gutkind, T. Ideker, Science **374**, eabf3067 (2021)
81. Y. Qin, E.L. Huttlin, C.F. Winsnes, M.L. Gosztyla, L. Wacheul, M.R. Kelly, S.M. Blue, F. Zheng, M. Chen, L.V. Schaffer, K. Licon, A. Backstrom, L.P. Vaites, J.J. Lee, W. Ouyang, S.N. Liu, T. Zhang, E. Silva, J. Park, A. Pitea, J.F. Kreisberg, S.P. Gygi, J. Ma, J.W. Harper, G.W. Yeo, D.L.J. Lafontaine, E. Lundberg, T. Ideker, Nature **600**, 536 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.