npj | systems biology and applications

www.nature.com/npjsba

ARTICLE OPEN

GINv2.0: a comprehensive topological network integrating molecular interactions from multiple knowledge bases

Xiao Chang[1,6], Shen Yan[2,6], Yizheng Zhang[3,4], Yingchun Zhang[5], Luyang Li[3,4], Zhanyu Gao[3,4], Xuefei Lin[1] and Xu Chi[3] ✉

Knowledge bases have been instrumental in advancing biological research, facilitating pathway analysis and data visualization, which are now widely employed in the scientific community. Despite the establishment of several prominent knowledge bases focusing on signaling, metabolic networks, or both, integrating these networks into a unified topological network has proven to be challenging. The intricacy of molecular interactions and the diverse formats employed to store and display them contribute to the complexity of this task. In a prior study, we addressed this challenge by introducing a "meta-pathway" structure that integrated the advantages of the Simple Interaction Format (SIF) while accommodating reaction information. Nevertheless, the earlier Global Integrative Network (GIN) was limited to reliance on KEGG alone. Here, we present GIN version 2.0, which incorporates human molecular interaction data from ten distinct knowledge bases, including KEGG, Reactome, and HumanCyc, among others. We standardized the data structure, gene IDs, and chemical IDs, and conducted a comprehensive analysis of the consistency among the ten knowledge bases before combining all unified interactions into GINv2.0. Utilizing GINv2.0, we investigated the glycolysis process and its regulatory proteins, revealing coordinated regulations on glycolysis and autophagy, particularly under glucose starvation. The expanded scope and enhanced capabilities of GINv2.0 provide a valuable resource for comprehensive systems-level analyses in the field of biological research. GINv2.0 can be accessed at: https://github.com/BIGchix/GINv2.0.

npj Systems Biology and Applications (2024)10:4 ; https://doi.org/10.1038/s41540-024-00330-y

INTRODUCTION

Accumulation of evidence regarding molecular interactions in biological processes has paved the way for the construction of various biological networks, including signaling, Protein-Protein Interaction (PPI), metabolic, and gene regulatory networks, among others. These networks have found various applications, ranging from visualizing omics data[1,2] to enriching gene sets using topology[3], identifying functional modules[4], conducting causal analyses[5,6], and developing computational models to understand the effects of network perturbations on cellular states[7]. Moreover, recent efforts have been directed towards associating changes in biological networks with diseases, leading to the emergence of "disease maps"[8–11]. Undoubtedly, the comprehensiveness and accuracy of biological networks form the fundamental keys for their successful application in network-based research.

A number of popular knowledge bases, such as KEGG[12,13], Reactome[14], and BioCyc[15], hold valuable information on molecular interactions in biological processes. To represent the complex relationships between biological molecules, several languages have been developed, such as KGML, BioPAX[16], GPML[17], and SBML[18]. However, converting this information into a comprehensive topological network has been a challenging endeavor, especially when dealing with different types of networks, such as signaling and metabolic networks. These networks often utilize distinct definitions for nodes and edges, leading to confusion and potential misinterpretations.

For instance, in signaling networks, an edge starting from node A and ending in node B, i.e., "node A activates node B", typically implies that A is an enzyme, while B is the substrate and product of a post-transcriptional modification (PTM) reaction, resulting in the retention of the same names for both the substrate and product. In contrast, metabolic networks involve substantial changes in substrates, leading to the generation of products with new names. Therefore, in a metabolic network, an edge starting from A and ending in B, i.e., "node A generates node B", refers to A being the substrate, and B being the product in this reaction, which significantly differs from the definitions in signaling networks. Without unifying the definitions of the nodes and edges, direct integration of signaling and metabolic networks may introduce confusion and misguidance.

Various tools have been developed to read and parse these languages, with the ability to convert the information into the Simple Interaction Format (SIF)[1,2]. SIF is a semi-structured format, in which each line specifies a source node, a character string describing the type of the edge(s), and one or more target nodes. However, the conversions often work better for signaling networks than for metabolic networks, as multiple substrates in metabolic reactions can lead to the ambiguity of multiple participants. Consequently, information regarding "who participates which reaction" can be lost during the conversion process.

To address these challenges, knowledge bases often visualize networks with edges pointing to edges, such as KEGG, Reactome, and Wikipathways[19]. Although this visualization is user-friendly, it is not suitable to work with common network analysis algorithms and tools. With mounting evidence suggesting the importance of crosstalk between signaling and metabolic networks, there is an urgent need to integrate these networks into a global integrative network, termed "GIN". Efforts have been made, but mainly focus

[1]Department of Dermatology and Venereal Disease, Xuan Wu Hospital, Beijing 100053, China. [2]Agricultural Information Institute, Chinese Academy of Agricultural Science, Beijing 100081, China. [3]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China. [4]University of Chinese Academy of Sciences, Beijing 100049, China. [5]Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. [6]These authors contributed equally: Xiao Chang, Shen Yan. ✉email: chix@big.ac.cn

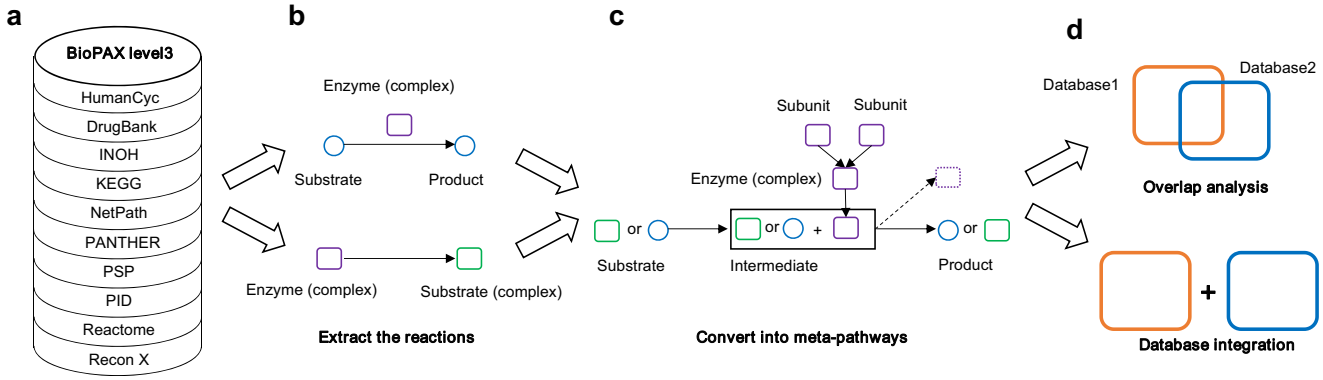

Fig. 1 Construction of global integrative network for human. The workflow consists of (**a**, **b**), the extraction of reactions from the BioPAX level3 (owl) files of 10 databases, (**c**) the conversion of the data into meta-pathways, and (**d**) the analysis of overlaps and the integration of the databases. The upper part of (**b**) represents a metabolic reaction and the lower part represents a signaling reaction.

on the visualization[20] or leveraging information from PPI network[21], leaving the signaling and metabolic networks topologically disconnected.

In this context, we propose a visualization layout called "meta-pathway" to fundamentally unify the topological structure of signaling and metabolic networks. To convert conventional pathways into meta-pathways, we introduce an intermediate node for each reaction in the pathways to represent a conceptual "intermediate" state of molecules in biochemical reactions. In most of biochemical reactions with multiple substrates or at least one enzyme, the substrate(s) and the enzyme need to get close enough to each other for the reactions to proceed, which forms the intermediate state. This intermediate state of the molecules is temporary, and will quickly be converted into products. Therefore, the intermediate nodes which come from the intermediate state of the molecules capture the relationships between molecules in real world, and enables both signaling and metabolic reactions to be considered as chemical reactions, facilitating storage in SIF-like format. By converting the pathways into meta-pathways and merging them, we have successfully built GINs for 7077 species based on KEGG[22].

In addition to KEGG, multiple biological knowledge bases offer valuable molecular interaction data across various aspects. In this study, we have converted molecular interaction data from ten different knowledge bases into the SIF format with intermediate nodes (referred to as SIFI). Subsequently, we conducted a thorough analysis of the consensus among these interactions before integrating the GINs into a single, comprehensive network, namely GIN for human version 2.0 (GINv2.0). Our results demonstrate that this version of GIN is currently one of the most comprehensive human databases of molecular interactions, allowing for straightforward visualization and interpretation of the crosstalk between signaling and metabolic networks, exemplified through a detailed examination of the glycolysis process and the related regulative proteins.

RESULTS

Conversion of BioPAX to SIFI

In our efforts to tackle the challenges of different knowledge base languages, we developed a R package named "SIFItools" to efficiently convert BioPAX level 3 owl files from various databases into SIFI format. OWL (Web Ontology Language) format is a powerful and expressive ontology language that allows users to define rich and complex relationships between entities. In the context of Biological Pathway Exchange (BioPAX) language, OWL is used to represent biological pathways and their components, such as molecules, interactions, and cellular processes, in a semantically meaningful way. With SIFItools, we firstly extracted biochemical reactions from the owl files of nine databases prepared by PathwayCommons[23], including HumanCyc, DrugBank[24], INOH[25], KEGG, NetPath[26], PANTHER[27], PhosphoSitePlus (PSP)[28], PID[29], and Recon X[30], as well as the owl file of Reactome from its official webpage (not from PathwayCommons). This facilitated the analysis of molecular interactions across multiple databases and laid the groundwork for building a comprehensive network for human cells (Fig. 1a, b). Then each of the reactions was converted into the structure of meta-pathway, introducing an intermediate node (Fig. 1c). After standardization of the 71 ID formats into seven, we analyzed the overlapping genes, chemicals and edges, then integrated all ten databases into one global integrative network, which we refer as GINv2.0 (Fig. 1d).

Notably, although SIFItools automated much of the curation process, manual curation was still necessary due to the diverse naming conventions and special characters used in different databases. In the process of manual curation, the most complicated task involved the conversion of internal IDs from each database's owl file to corresponding external gene or chemical IDs. This complexity arose from the fact that a single gene or chemical could have different internal IDs across various databases, each linked to one or more distinct external IDs. To overcome this challenge, we developed a two-step approach. Firstly, we constructed an ID mapping table using internal "XRef" links, enabling us to convert the internal IDs to external IDs from 71 different sources. Subsequently, we aggregated the external IDs from diverse sources into gene symbols and unified chemical ID types (UC_IDs), which includes CID[31], SID[31], CAS registry number, KEGG, HMDB[32], and ChEBI[33] (Supplementary Fig. 1). This method ensured consistency and standardization across the databases, facilitating seamless integration of the data in our subsequent analyses.

Consensus analysis of the databases

Conversion of BioPAX level3 into SIFI format generated networks varied in the number of edges and nodes, ranging from 873 nodes (NetPath) to 5614 nodes (Reactome) (Fig. 2a), and from 2444 edges (NetPath) to 29898 edges (Reactome) (Supplementary Fig. 2). Notably, the ratio between the quantity of genes and the number of chemicals exhibited variations across the databases. These variations accurately mirrored the distinct scopes of molecular interactions inherent to each individual database. For example, the SIFI format of NetPath and PSP exclusively contained human genes, while Recon X exclusively included chemical IDs (Fig. 2a). This distinction highlights the significance of our integrative approach in capturing a comprehensive picture of human molecular interactions.

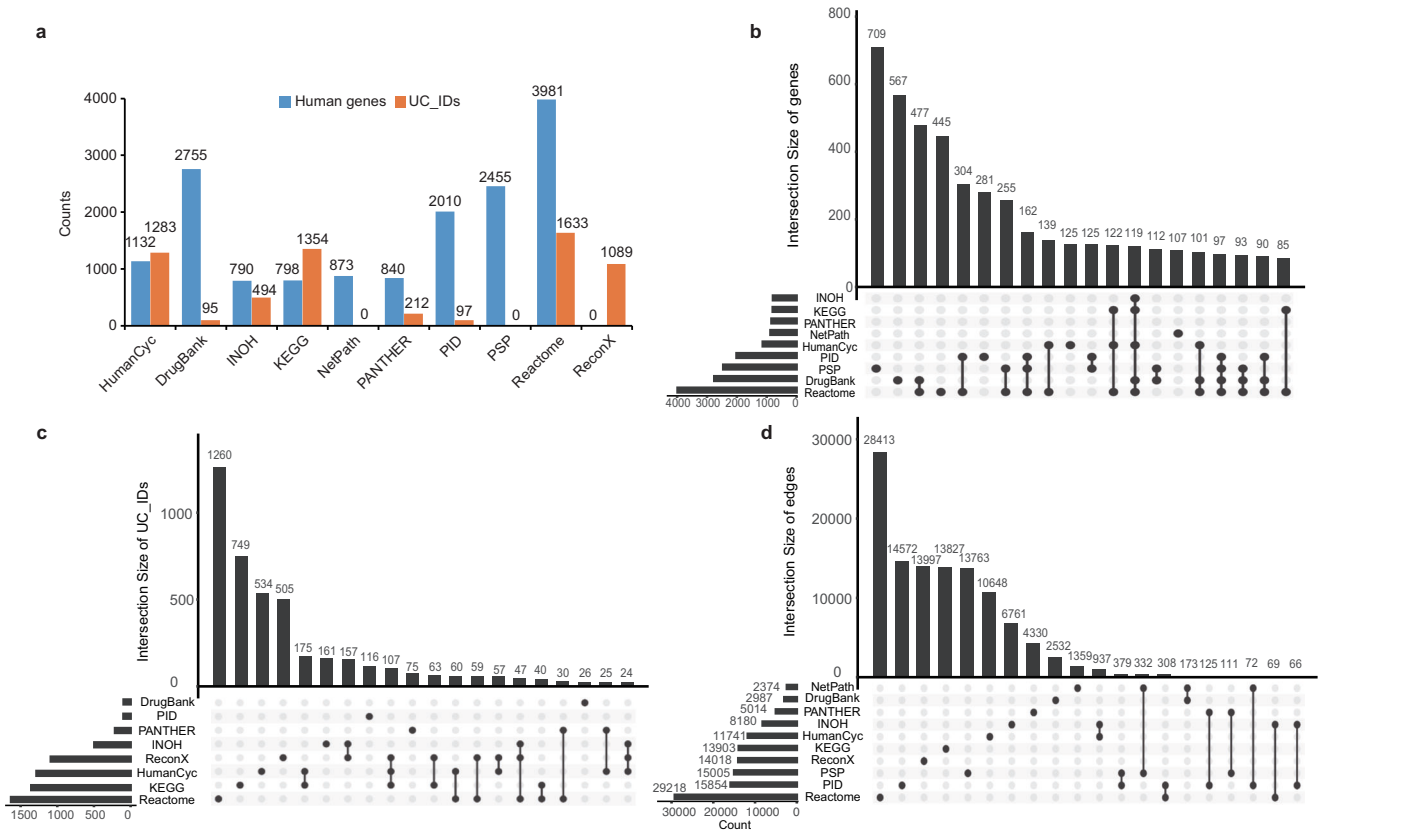

Fig. 2 The overlap analysis of the databases. a The number of human genes and chemical IDs in the ten databases. **b**–**d** Upset plots showing the overlap of the genes, chemicals and edges between the 10 databases. Each column in the matrix at the lower part of the plot shows the sources of the set whose number is displayed as a bar in the upper part of the plot.

Next, we conducted an analysis of the overlapping gene symbols (Fig. 2b), UC_IDs (Fig. 2c), and edges (Fig. 2d) among the ten databases. For clarity, Recon X was excluded from Fig. 2b due to its exclusive focus on chemicals. Similarly, NetPath and PSP were excluded from Fig. 2c. Our analysis revealed that the overlap of gene symbols was notably larger than the overlap of chemical IDs. For instance, in the case of Reactome, the number of unique gene symbols accounted for only 11.2% of its total gene symbols (445 out of 3981), whereas the number of unique chemical IDs represented 76.1% of its total chemical IDs (1243 out of 1633). Furthermore, we found that the overlap of interactions between databases was limited, with over 96.8% (110202 out of 113876) of the interactions being unique to each database for the majority of cases. This observation underscores the distinctiveness and database-specific nature of the interactions. The limited overlap of interactions highlights the importance of our integrative approach in leveraging data from multiple sources to build a comprehensive and interconnected network.

Integration of the ten databases

We merged the SIFI files from all ten databases to construct the raw global integrative network of human. Redundant edges were removed before importing the network into Cytoscape for visualization (Fig. 3a). The final GINv2.0 for human comprises 39,548 nodes and 113,876 edges, encompassing 6330 genes, 3579 chemical IDs, 3957 complexes, and 25,682 intermediate nodes. To facilitate further analysis, we utilized the Python package leidenalg[34] to cluster the network into distinct sub-networks. In Fig. 3a, we presented the top 20 sub-networks with the largest number of nodes. These sub-networks exhibit diverse compositions of genes, chemicals, and intermediates. Notably, most sub-networks are a mix of genes and chemicals; however, some sub-networks, such as clusters 3, 5, 6, 8, 11, 13, 15, 17, 18, and 19, are predominantly gene-driven, while others, like clusters 4 and 16, are primarily chemical-centric (Fig. 3b). This observation underscores the complex interplay between signaling networks and metabolic pathways, contributing to the complexity of the network.

Additionally, we calculated the topological network metrics, presented in Table 1. Notably, the node with the highest degree was water, followed by ATP and ADP. These findings indicate that water, ATP, and ADP are central participants in biological processes within human cells, aligning well with established knowledge in the field. To gain deeper insights into specific sub-networks, we conducted a focused examination of cluster 16. We identified several nodes with high degrees, including HIF1A, KDM1A, Succinic acid, Acetyl-CoA, Formaldehyde, CO2, NADH, and NAD+ (Fig. 3c).

To investigate the composition of the database sources of each cluster, we calculated the percentage of the edges contributed by different databases to each cluster (Fig. 3d). Our analysis showed that ReconX's data (only consists of chemicals) mainly presents in cluster 1 and cluster 10. For cluster 1, there are three major sources, ReconX, INOH, and HumanCyc. In cluster 10, the major sources are ReconX, Reactome, and HumanCyc. Similar results can be observed for PSP and NetPath. These evidences suggest that the databases focusing on only genes or chemicals are well mixed with other databases. On the other hand, KEGG and Reactome contribute the majority of edges of cluster 4 and cluster 16 respectively, which are chemical-centric, and cluster 15 which is gene-centric by Reactome. This suggests that these two comprehensive databases, KEGG and Reactome, who both cover signaling

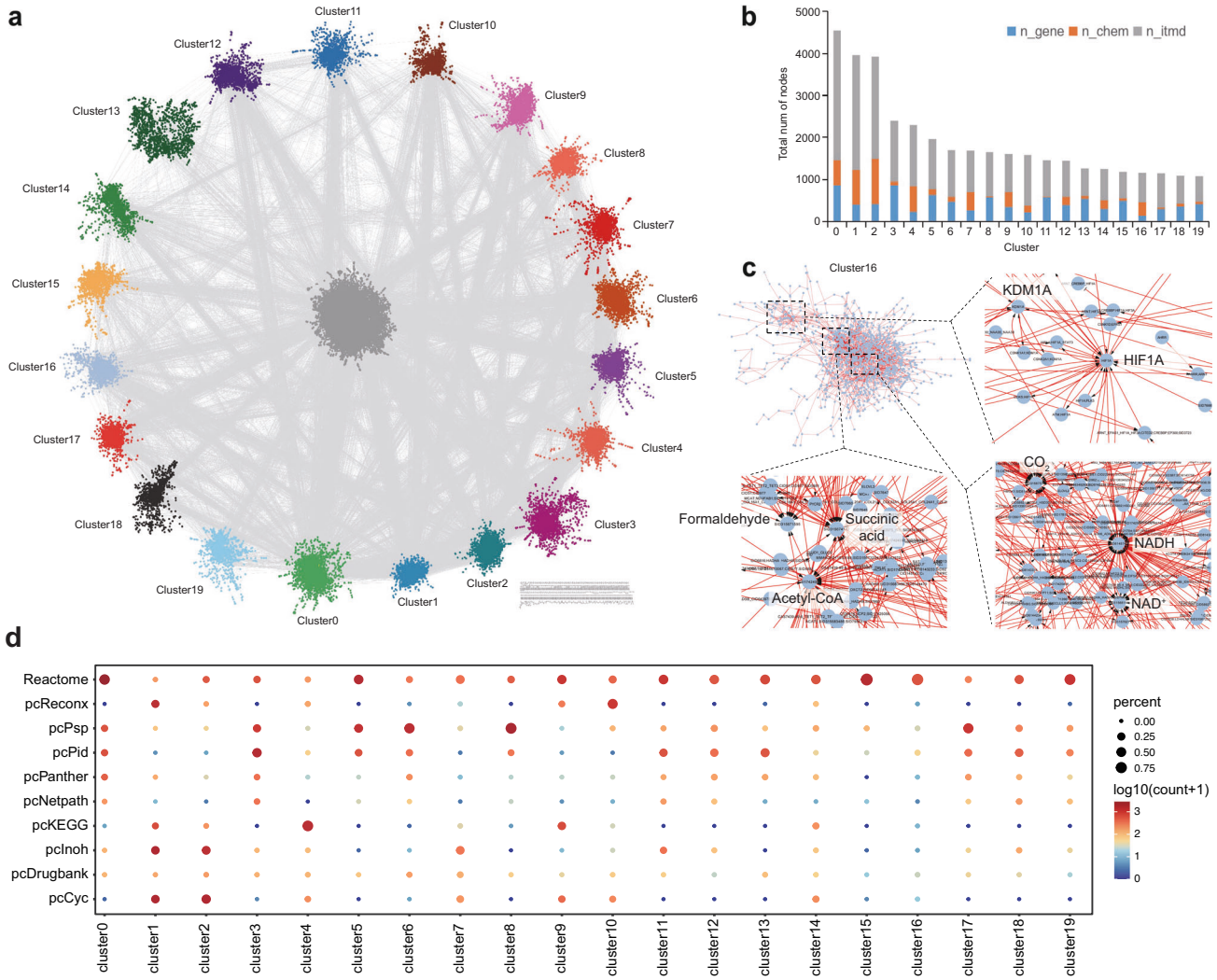

Fig. 3 GINv2.0 constructed from 10 databases. a Visualization of GINv2.0 by Cytoscape. The largest 20 sub-networks are shown in colors. The small and fragmented interactions which do not connect to the major network are shown on the right bottom. **b** The node composition of the top 20 sub-networks. n_itmd, number of intermediate nodes. **c** Zoomed visualization of cluster16. **d** Dot-plot showing the contribution of different databases to the edges of the clusters. The size of the circle represents the percentage of the edges. The color represents the log10 transformed counts of the overlapping edges.

and metabolic pathways, may have distinct scopes of signaling and metabolic reactions.

Regulation of glycolysis by signaling proteins

To demonstrate the practical application of GINv2.0 in analyzing signaling and metabolic networks, we extracted nodes representing the metabolites, intermediates, and protein enzymes involved in glycolysis, along with the proteins that regulate these enzymes. Subsequently, we visualized the network in Cytoscape (Fig. 4). Glycolysis is a fundamental cellular metabolic process that converts glucose to pyruvate, generating ATP and NADH. The GINv2.0 visualization of glycolysis clearly illustrates how enzymes are linked to metabolites through intermediate nodes. Moreover, each intermediate node represents a specific reaction, effectively circumventing ambiguity arising from multiple isozymes catalyzing the same reaction.

Subsequently, we focused on the incoming nodes of the enzymes involved in glycolysis, which provided insights into the proteins regulating this crucial metabolic pathway. Our analysis revealed that, out of the ten steps comprising glycolysis, seven steps were regulated by various kinases, including SRC[35,36], ULK1[37,38], AKT1[39], AKT2[40], PRKAA1[41], PRKCD[42], PAK1[43], MAPK1[44], MAPK8[45], GSK3B[46], PIM2[47], EGFR[43], and CDK6[48]. These findings highlight the complicated control mechanisms governing glycolysis, ensuring its harmonious coordination with the activation and inhibition of other cellular pathways, ultimately balancing energy production.

Notably, we found that ULK1 is a prominent positive regulator of HK1[37], PFKM[37], and ENO1[37,38], making it a pivotal protein in governing glycolysis based on the number of controlling enzymes. While SRC is known for its broad involvement in various cellular processes, ULK1 is well known for its essential role in initiating autophagy[49]. Building on this intriguing clue, we deeply explored the relationship between these kinases and autophagy regulation. Remarkably, seven out of the thirteen identified proteins were found to exhibit direct or indirect regulatory effects on autophagy. Protein Kinase AMP-Activated Catalytic Subunit Alpha 1 (PRKAA1), the catalytic subunit of AMPK, plays a crucial role in autophagy initiation under glucose deprivation by directly phosphorylating ULK1[41]. Additionally, AKT suppresses tuberous sclerosis complex proteins 1/2 (TSC1/2) through phosphorylation, leading to mTORC1 activation and subsequent autophagy inhibition[50]. Reports have shown that GSK3B promotes ULK1 acetylation by mediating KAT5/TIP60 phosphorylation during starvation[51].

Table 1. Top 20 nodes with highest degrees.

Id	Name	Degree
CID962	Water	3179
CID1038	H+	2925
CID6022	Adenosine-5′-Diphosphate	1933
CID5957	Adenosine-5′-Triphosphate	1498
SID8148096	ATP(4-)	655
CID977	Oxygen	524
CID87642	Coenzyme A	499
CID1004	Phosphoric Acid	490
CID5884	NADPH	486
SID85646635	ADP(3-)	397
CID1061	Phosphate Ion	383
SRC	*SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase*	**379**
SID99319226	NAD(1-)	345
SID111978360	Nucleoside Triphosphate(4-)	344
PRKACA	*Protein Kinase CAMP-Activated Catalytic Subunit Alpha*	**343**
CID923	Sodium Ion	342
CID4995	Diphosphate(2-)	336
CID21604869	beta-NADH	325
AKT1	*AKT Serine/Threonine Kinase 1*	**324**
CID280	Carbon Dioxide	299

The IDs in bold are protein kinases.

Furthermore, MAPK8 activates autophagy by mediating BCL2 phosphorylation, facilitating the dissociation of BCL2 from BECN1[52]. Finally, emerging evidence suggests that PIM2 is capable of phosphorylating HK2, thereby promoting autophagy under glucose deprivation[53].

Collectively, these findings indicate a synergistic regulation of glycolysis and autophagy, particularly under glucose-starved conditions, enriching the understanding of cellular adaptation to varying nutrient availability. In summary, our comprehensive network analysis empowers researchers with fresh perspectives on the cross-talk between metabolic and cellular regulatory networks, paving the way for deeper investigations into the underlying molecular complexities.

DISCUSSION

In this work, we compiled a much more comprehensive GIN for human compared with the previous version. The previous GIN[22] for human was built only upon KEGG, which includes 5145 genes and 1501 metabolites. In the present work, we compiled a new GIN for human from ten different databases, which involved 6330 genes and 3579 metabolites, with 23.0% and 138.4% increase, respectively. The new GIN for human is much more useful than the previous one, as the integration of various databases greatly enhances the comprehensiveness of the network. This is exemplified by the demonstration of the orchestrated regulation of autophagy and glucose metabolism under stress, which leveraged information from multiple databases.

We also offer a new tool for the conversion of BioPax level3 files into SIFI format. In our previous work, we built the GIN by a pipeline of perl scripts specifically written to parse KGML files. Since the use of KGML format is currently limited to KEGG, our

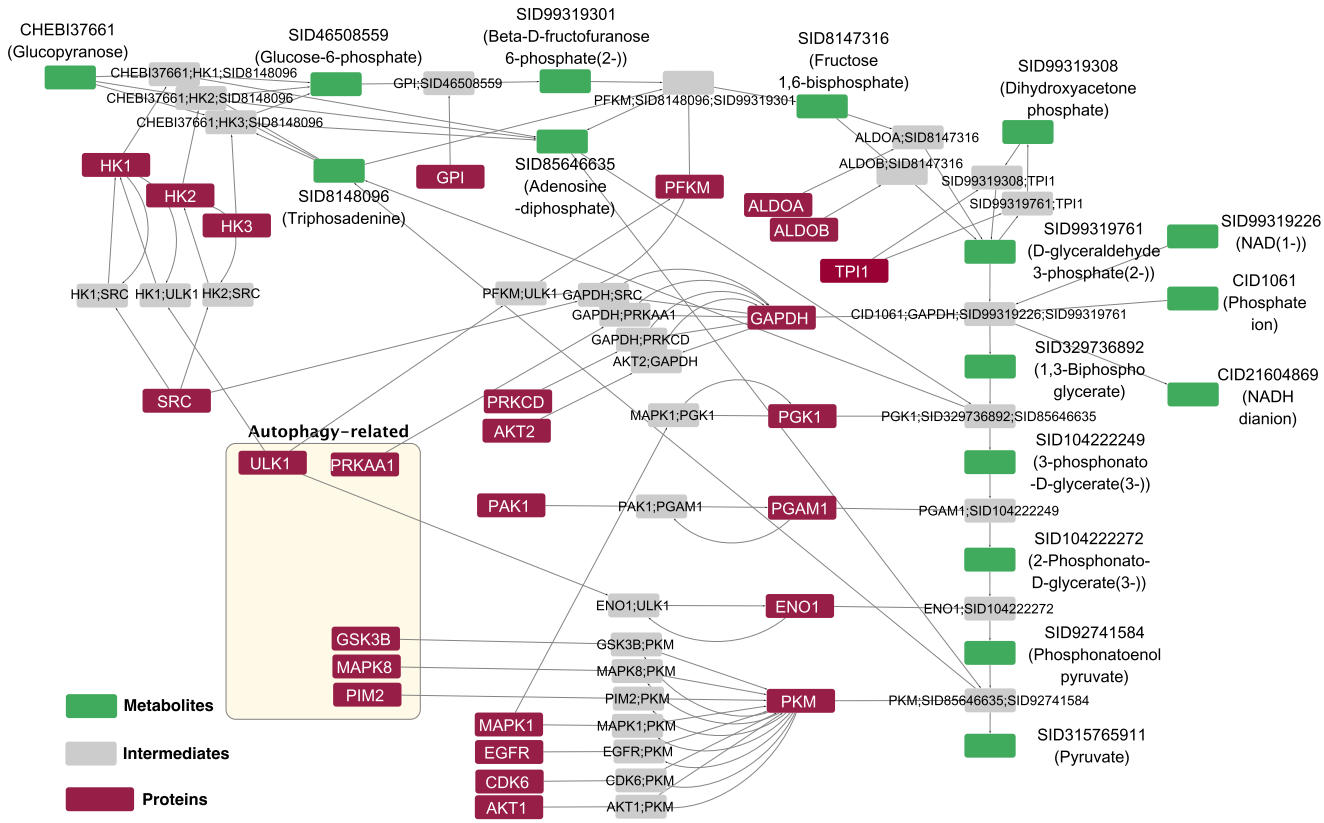

Fig. 4 Illustration of glycolysis and its regulative proteins. The regulative protein kinases which are associated with autophagy were clustered in a light-yellow box.

previous pipeline lacks the ability to process the files of other databases. In our present work, we construct a R package (SIFItools) which can convert BioPax level3 files into SIFI format with minimum manual curations required. Because many biological databases share their data in BioPax level3 format, our new package, SIFItools is more convenient and have much more potential applications when building GIN from databases.

Also in this work, we compared the overlapping information between the 10 databases. We were not able to conduct such analysis in our previous work since we only converted KEGG database into GIN. In our present work, we compared the molecular interactions of the 10 databases and surprisingly found that the overlaps between the databases were rare. Although this could partly due to the different focus of the databases, there is still a great proportion of unmatched ids, especially for metabolites. This could lead to confusing results when applying over-representation analysis (ORA) of pathways, as pointed out by another work[54].

The knowledge bases of pathways serve as repositories for capturing molecular interactions in both physiological and pathological contexts. While each database emphasizes distinct molecular interactions, synthesizing the collective insights from various sources can outline the comprehensive scope of these knowledge bases. However, the exploration of consensus among diverse knowledge bases has been limited, in part due to the varying data formats used by each database. Despite PathwayCommons' efforts to standardize data formats, the inherent features of XML format have posed challenges for direct cross-database comparisons. For example, in BioPAX level3, key information about a given reaction may be dispersed across properties such as "left", "right", "product", "controlled", "controller", or "cofactor." This distribution necessitates the extraction of reaction details from multiple attributes to facilitate comparison, thereby complicating and impeding the efficiency of the process. The introduction of meta-pathways and SIFI format has alleviated this predicament by structuring reaction information into a SIF-like three-column configuration. This transformation enables rapid comparisons between reactions, streamlining the comparative analysis.

We noticed that there are overlaps between the concepts of meta-pathway, SIFI format and GIN. To clarify the definitions of the three concepts: (1) Meta-pathway is the way of displaying pathways using intermediates to connect the substrates and products. (2) The GIN (Global Integrative Network) is a network combining the molecular interactions from all pathways. (3) SIFI (Simple Interaction Format with Intermediates) is a format we use to store the molecular interactions of meta-pathways and GIN. The differences between the three concepts are: meta-pathway is the component of GIN, while they can be both stored in SIFI format.

The consensus analysis of GINs generated from different databases highlighted significant diversity across the databases, particularly concerning the edges and nodes related to metabolites. This observed diversity could potentially rise from variations in the specific focus of each database or disparities in naming conventions. Such variations raise valid concerns regarding the reliability of metabolite enrichment analysis, aligning with findings from a recent investigation into the ORA of pathways leveraging metabolomics data. Notably, the authors of this study revealed significant disparities in ORA results when employing distinct databases, such as KEGG, Reactome and BioCyc[54], which may partially due to the inconsistency we found in our consensus analysis.

The credibility of the edges is also important for network analysis, since questionable edges will create misleading path when conducting path-related network analysis, as evidenced in our previous work[22]. In the comparative analysis of different databases, repeated edges may be more credible since it has been repeatedly validated by different databases. In fact, one of our original goals to compare the edges from different databases was to score the edges based on the number of repeats. However, with the analysis of the databases, we found that a large number of the non-redundant edges are the results of the variations of the scopes of databases. For example, in Fig. 4, the edges extracted from the PSP database are not found in any other databases, but all of these edges have credible sources of publications. This means that a large proportion of non-redundant edges may be credible. Based on this consideration, we excluded the analysis of the credibility of edges in our current work.

By analyzing GINv2.0, we found that the number of intermediate nodes was substantially larger than the combined count of both genes and metabolites. Since each intermediate node represents a distinct biochemical reaction, the number of genes/metabolites involved in a pathway, which is often used in conventional enrichment analysis such as GO, may not truly reflect the number of reactions associated with the pathway. For instance, consider a scenario where five genes are shared between the input gene set and a pathway gene set. While ORA and GSEA[55,56] might not distinguish whether these five genes participate in one single reaction or five distinct ones, the possibilities of significant associations between the input and pathway gene sets are distinct, judging by instinct. Thus, the intermediate nodes are likely a hidden layer reside between the genes/metabolites and pathways, which has not been investigated for enrichment analysis. The construction of GINs is therefore, a starting point for building the relations between genes/metabolites, intermediate states, and pathways, and further promote the improvement of gene set/pathway analysis.

The illustration of the glycolysis process and the regulative proteins underscores the benefits of the integration of multiple knowledge bases. Notably, we found that the core nodes and edges of the glycolysis process was primarily derived from KEGG, Reactome, HumanCyc, and INOH, while the regulatory interplays between kinases and glycolytic enzymes were from PSP. Individual GINs of any single databases were not able to provide such comprehensive view of molecular interactions. This demonstrates the necessity of database integration to forge a comprehensive and unified network.

In the current version of GIN (v2.0), the intermediate nodes are built for metabolic reactions and PTM reactions, but not for PPI. The reason for excluding PPI is that the GIN we built is a directed graph, but PPI networks are undirected, therefore, current PPI data does not fit for GIN we built. However, we are working on the solution to generate appropriate intermediate nodes for the complexes with multiple protein participants in PPI. With the flexibility of the meta-pathway's structure, other types of data regarding molecular interactions in cells, including the relations of transcription factors (TFs) and their targets, miRNAs and their targets, will soon be incorporated in GINs as well.

METHODS

Construction of the Global Integrative Network from ten databases

The owl files of BioPAX level 3 prepared by PathwayCommons were downloaded from https://www.pathwaycommons.org/archives/PC2/v12/. Specifically, we selected DrugBank, HumanCyc, INOH, KEGG, NetPath, PANTHER, PID, PhosphoSitePlus (PSP), and Recon X from PathwayCommons, which contain sufficient number of biochemical reactions extracted from the owl files. The BioPAX level 3 owl file of Reactome was acquired through https://reactome.org/download-data. The owl files were parsed by function "readBiopax" from R package rBiopaxParser[57], which generated a dataframe for each of the databases.

We built a R package to extract the reactions from the dataframe and convert them into SIFI format. Specifically, we first

extracted the reactions from classes of TransportWithBiochemicalReaction, Transport, BiochemicalReaction, ComplexAssembly, Degradation, and Conversion. Then we extracted the information of the enzymes from the classes of Catalysis, Control, and Modulation, and linked the enzymes with the reactions. These information was finally organized into one temporary table.

Subsequently, we created a component matching table designed to capture the relationships between proteins and complexes. We did not use the conventional name of the complexes; instead, we adopted a distinct approach wherein the complexes were systematically deconstructed into constituent proteins through recursive processes. Then the name of the complexes were given by concatenation of all the names of the components in alphabetical order, separated by underscores ("_") .

Next, we replaced the complex IDs in the reaction table with the name generated from the component names, and convert the reactions into SIFI format. The intermediate nodes were introduced during this conversion step. The names of the intermediate nodes were the concatenation of all the substrates and enzymes, separated by semicolon (";"). Note that the result of this step still used the local ID system for each owl file specifically which cannot be shared with other owl files.

Since the owl file of each database provides mapping relations between local IDs and commonly used (external) IDs, we replaced the local IDs with the external IDs suggested by each database. However, each databases has its own preference on the use of the ID sources, thus we had to uniform the sources of the IDs to ensure that the same gene/chemical got the same ID in different databases. Uniprot IDs were converted to gene symbols by R package biomaRt[58]. For metabolite IDs, We constructed a mapping table using R package metaboliteIDmapping[59], and used the strategy in Supplementary Fig. 1 to uniform the IDs with a preference of the sources. A tutorial for the conversion of KEGG's owl file to SIFI format can be found at https://github.com/BIGchix/SIFItools.

Finally, we concatenated all the SIFI files into one single file, and removed the redundant edges. The edges containing gene IDs of other species were removed.

Network analysis of human GINv2.0

The intersection results of genes, metabolites and edges were visualized by R package UpSetR[60]. The total network and the network of glycolysis were visualized by Cytoscape[1,2]. The community detection was performed by python package leidenalg[34], to efficiently work with large directed graph using the Leiden algorithm[34].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

We used the BioPAX level3 owl files prepared by PathwayCommons[23] which can be accessed here (https://www.pathwaycommons.org/archives/PC2/v12/). The BioPAX level 3 owl file of Reactome was acquired through https://reactome.org/download-data. The GINv2.0 generated in this work can be freely accessed from github: https://github.com/BIGchix/GINv2.0.

CODE AVAILABILITY

The R package "SIFItools" can be freely accessed and installed from github: https://github.com/BIGchix/SIFItools.

Received: 19 August 2023; Accepted: 2 January 2024;
Published online: 13 January 2024

REFERENCES

1. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504, https://doi.org/10.1101/gr.1239303 (2003).
2. Cline, M. S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382, https://doi.org/10.1038/nprot.2007.324 (2007).
3. Zito, A. et al. Gene set enrichment analysis of interaction networks weighted by node centrality. *Front. Genet.* **12**, 577623, https://doi.org/10.3389/fgene.2021.577623 (2021).
4. Loers, J. U. & Vermeirssen, V. SUBATOMIC: a SUbgraph BAsed mulTi-OMIcs clustering framework to analyze integrated multi-edge networks. *BMC Bioinforma.* **23**, 363–363, https://doi.org/10.1186/s12859-022-04908-3 (2022).
5. Catlett, N. L. et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinforma.* **14**, 340, https://doi.org/10.1186/1471-2105-14-340 (2013).
6. Krämer, A., Green, J., Pollard, J. Jr & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530, https://doi.org/10.1093/bioinformatics/btt703 (2014).
7. Azeloglu, E. U. & Iyengar, R. Signaling networks: information flow, computation, and decision making. *Cold Spring Harb. Perspect. Biol.* **7**, a005934–a005934, https://doi.org/10.1101/cshperspect.a005934 (2015).
8. Hoch, M. et al. Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps. *npj Syst. Biol. Appl.* **8**, 13, https://doi.org/10.1038/s41540-022-00222-z (2022).
9. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562, https://doi.org/10.1038/nrg.2017.38 (2017).
10. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLOS Comput. Biol.* **6**, e1000641, https://doi.org/10.1371/journal.pcbi.1000641 (2010).
11. Stoney, R., Robertson, D. L., Nenadic, G. & Schwartz, J.-M. Mapping biological process relationships and disease perturbations within a pathway network. *npj Syst. Biol. Appl.* **4**, 22, https://doi.org/10.1038/s41540-018-0055-2 (2018).
12. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
13. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592, https://doi.org/10.1093/nar/gkac963 (2023).
14. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503, https://doi.org/10.1093/nar/gkz1031 (2020).
15. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **6**, R2, https://doi.org/10.1186/gb-2004-6-1-r2 (2004).
16. Demir, E. et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **28**, 935–942, https://doi.org/10.1038/nbt.1666 (2010).
17. van Iersel, M. P. et al. Presenting and exploring biological pathways with PathVisio. *BMC Bioinforma.* **9**, 399, https://doi.org/10.1186/1471-2105-9-399 (2008).
18. Hucka, M. in *Encyclopedia of Systems Biology* (eds Dubitzky. W, Wolkenhauer. O, Cho, K-H, & Yokota. H) 2057–2063 (Springer New York, 2013).
19. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621, https://doi.org/10.1093/nar/gkaa1024 (2021).
20. Sompairac, N. et al. Metabolic and signalling network maps integration: application to cross-talk studies and omics data analysis in cancer. *BMC Bioinforma.* **20**, 140, https://doi.org/10.1186/s12859-019-2682-z (2019).
21. Bag, A. K. et al. Connecting signaling and metabolic pathways in EGF receptor-mediated oncogenesis of glioblastoma. *PLOS Comput. Biol.* **15**, e1007090, https://doi.org/10.1371/journal.pcbi.1007090 (2019).
22. Lin, Y., Yan, S., Chang, X., Qi, X. & Chi, X. The global integrative network: integration of signaling and metabolic pathways. *aBIOTECH* **3**, 281–291, https://doi.org/10.1007/s42994-022-00078-1 (2022).
23. Rodchenkov, I. et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **48**, D489–D497, https://doi.org/10.1093/nar/gkz946 (2020).
24. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082, https://doi.org/10.1093/nar/gkx1037 (2018).
25. Yamamoto, S. et al. INOH: ontology-based highly structured database of signal transduction pathways. *Database* **2011**, bar052, https://doi.org/10.1093/database/bar052 (2011).
26. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3, https://doi.org/10.1186/gb-2010-11-1-r3 (2010).
27. Thomas, P. D. et al. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22, https://doi.org/10.1002/pro.4218 (2022).

28. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520, https://doi.org/10.1093/nar/gku1267 (2015).
29. Schaefer, C. F. et al. PID: the pathway interaction database. *Nucleic Acids Res.* **37**, D674–D679, https://doi.org/10.1093/nar/gkn653 (2009).
30. Duarte, N. C. et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* **104**, 1777–1782 (2007).
31. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380, https://doi.org/10.1093/nar/gkac956 (2023).
32. Wishart, D. S. et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res.* **50**, D622–D631, https://doi.org/10.1093/nar/gkab1062 (2022).
33. Hastings, J. et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219, https://doi.org/10.1093/nar/gkv1031 (2016).
34. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233, https://doi.org/10.1038/s41598-019-41695-z (2019).
35. Zhang, J. et al. c-Src phosphorylation and activation of hexokinase promotes tumorigenesis and metastasis. *Nat. Commun.* **8**, 13732, https://doi.org/10.1038/ncomms13732 (2017).
36. Rush, J. et al. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23**, 94–101, https://doi.org/10.1038/nbt1046 (2005).
37. Li, T. Y. et al. ULK1/2 constitute a bifurcate node controlling glucose metabolic fluxes in addition to autophagy. *Mol. Cell* **62**, 359–370, https://doi.org/10.1016/j.molcel.2016.04.009 (2016).
38. Weber, C., Schreiber, T. B. & Daub, H. Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells. *J. Proteom.* **75**, 1343–1356, https://doi.org/10.1016/j.jprot.2011.11.004 (2012).
39. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292, https://doi.org/10.1038/nbt1240 (2006).
40. Huang, Q. et al. Akt2 Kinase Suppresses Glyceraldehyde-3-phosphate Dehydrogenase (GAPDH)-mediated apoptosis in ovarian cancer cells via phosphorylating GAPDH at Threonine 237 and Decreasing Its Nuclear Translocation *. *J. Biol. Chem.* **286**, 42211–42220, https://doi.org/10.1074/jbc.M111.296905 (2011).
41. Chang, C. et al. AMPK-Dependent Phosphorylation of GAPDH Triggers Sirt1 activation and is necessary for autophagy upon glucose starvation. *Mol. Cell* **60**, 930–940, https://doi.org/10.1016/j.molcel.2015.10.037 (2015).
42. Qvit, N., Joshi, A. U., Cunningham, A. D., Ferreira, J. C. B. & Mochly-Rosen, D. Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) Protein-Protein Interaction Inhibitor Reveals a Non-catalytic Role for GAPDH Oligomerization in Cell Death *. *J. Biol. Chem.* **291**, 13608–13621, https://doi.org/10.1074/jbc.M115.711630 (2016).
43. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62, https://doi.org/10.1038/nature18003 (2016).
44. Wang, Y. et al. O-GlcNAcylation destabilizes the active tetrameric PKM2 to promote the Warburg effect. *Proc. Natl. Acad. Sci.* **114**, 13732–13737, https://doi.org/10.1073/pnas.1704145115 (2017).
45. Iansante, V. et al. PARP14 promotes the Warburg effect in hepatocellular carcinoma by inhibiting JNK1-dependent PKM2 phosphorylation and activation. *Nat. Commun.* **6**, 7882, https://doi.org/10.1038/ncomms8882 (2015).
46. Sizemore, S. T. et al. Pyruvate kinase M2 regulates homologous recombination-mediated DNA double-strand break repair. *Cell Res.* **28**, 1090–1102, https://doi.org/10.1038/s41422-018-0086-7 (2018).
47. Yu, Z. et al. Proviral Insertion in Murine Lymphomas 2 (PIM2) oncogene phosphorylates pyruvate kinase M2 (PKM2) and promotes glycolysis in cancer cells. *J. Biol. Chem.* **288**, 35406–35416, https://doi.org/10.1074/jbc.M113.508226 (2013).
48. Liang, J. et al. PKM2 dephosphorylation by Cdc25A promotes the Warburg effect and tumorigenesis. *Nat. Commun.* **7**, 12431, https://doi.org/10.1038/ncomms12431 (2016).
49. Egan, D. F. et al. Phosphorylation of ULK1 (hATG1) by AMP-activated protein kinase connects energy sensing to mitophagy. *Science* **331**, 456–461, https://doi.org/10.1126/science.1196371 (2011).
50. Inoki, K., Li, Y., Zhu, T., Wu, J. & Guan, K.-L. TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling. *Nat. Cell Biol.* **4**, 648–657, https://doi.org/10.1038/ncb839 (2002).
51. Cheng, X. et al. Pacer Is a Mediator of mTORC1 and GSK3-TIP60 signaling in regulation of autophagosome maturation and lipid metabolism. *Mol. Cell* **73**, 788–802.e787, https://doi.org/10.1016/j.molcel.2018.12.017 (2019).
52. Wei, Y., Pattingre, S., Sinha, S., Bassik, M. & Levine, B. JNK1-Mediated Phosphorylation of Bcl-2 Regulates starvation-induced autophagy. *Mol. Cell* **30**, 678–688, https://doi.org/10.1016/j.molcel.2008.06.001 (2008).
53. Yang, T. et al. PIM2-mediated phosphorylation of hexokinase 2 is critical for tumor growth and paclitaxel resistance in breast cancer. *Oncogene* **37**, 5997–6009, https://doi.org/10.1038/s41388-018-0386-x (2018).
54. Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *PLOS Comput. Biol.* **17**, e1009105, https://doi.org/10.1371/journal.pcbi.1009105 (2021).
55. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550, https://doi.org/10.1073/pnas.0506580102 (2005).
56. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273, https://doi.org/10.1038/ng1180 (2003).
57. Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A. & Beißbarth, T. rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics* **29**, 520–522, https://doi.org/10.1093/bioinformatics/bts710 (2013).
58. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191, https://doi.org/10.1038/nprot.2009.97 (2009).
59. Canzler, S. metaboliteIDmapping: mapping of metabolite IDs from different sources. *R package version 0.99.10*, https://github.com/yigbt/metaboliteIDmapping (2022).
60. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940, https://doi.org/10.1093/bioinformatics/btx364 (2017).

ACKNOWLEDGEMENTS

We thank Yan Yan for her insightful discussion on SIFItools, and Siyi Su for his help on the glycolysis illustration.

AUTHOR CONTRIBUTIONS

Xu.C. and Xiao.C. conceived the study; S.Y. developed SIFItools; Xiao.C., Y.Z., L.L., Z.G. and X.L. collected data and performed conversion; Xu.C., S.Y. and Y.Z. analyzed the data; Y.Z. performed network visualization; Xu.C. wrote the manuscript; All authors helped in the writing of the manuscript. All authors approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00330-y.

Correspondence and requests for materials should be addressed to Xu Chi.