



# Knowledge Graph Construction for Molecular Interaction Exploration

Wenzhong Huang 

Biomass Research Center, Hainan Institute of Tropical Agricultural Resources, Sanya, 572025, Hainan, China

 Corresponding author: [wenzhong.huang@hitar.org](mailto:wenzhong.huang@hitar.org)Computational Molecular Biology, 2025, Vol.15, No.4 doi: [10.5376/cmb.2025.15.0017](https://doi.org/10.5376/cmb.2025.15.0017)

Received: 12 May, 2025

Accepted: 23 Jun., 2025

Published: 15 Jul., 2025

**Copyright** © 2025 Huang, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Preferred citation for this article:**Huang W.Z., 2025, Knowledge graph construction for molecular interaction exploration, Computational Molecular Biology, 15(4): 171-182 (doi: [10.5376/cmb.2025.15.0017](https://doi.org/10.5376/cmb.2025.15.0017))

**Abstract** In recent years, knowledge graph technology has emerged in bioinformatics, providing new ideas for the study of interaction relationships at the molecular level. This research focuses on the construction and analysis of the "Molecular Interaction Knowledge Graph", including the integration and preprocessing of data sources, the construction methods of the knowledge graph, the representation and analysis techniques of the graph, as well as the case study and system implementation of the protein-protein interaction knowledge graph. The research first sorted out the current application status of knowledge graphs in bioinformatics, and clarified the background significance and innovation points of constructing molecular interaction knowledge graphs. Subsequently, the standardization and entity semantic normalization strategies for multi-source biological data were discussed, and the modeling methods for entities and relationships as well as the automated construction process were proposed. In terms of graph analysis, key technologies such as knowledge representation learning, network topology analysis, semantic reasoning and relationship prediction are reviewed. Through the case of protein-protein interaction mapping, the specific process of mapping construction, visualization results and biological verification are presented, and the biological significance of the conclusions obtained is discussed. Finally, the current challenges in the field of molecular interaction knowledge graphs, such as data heterogeneity, model interpretability and knowledge uncertainty, are summarized, and the future development directions are prospected. The research work is expected to provide a solid knowledge support for promoting the systematic analysis of complex molecular networks and biomedical discoveries.

**Keywords** Molecular interaction; Knowledge graph; Bioinformatics; Protein-protein interactions; Knowledge representation learning

## 1 Introduction

The normal operation of biological systems largely depends on the intricate connections among molecules - protein interactions, gene regulation, metabolic reaction pathways. They are like intricate networks. Understanding these relationships is not merely about "clarifying the principles", but also about knowing why diseases occur and how drugs work. However, in the past, experiments were conducted one by one to verify, which was time-consuming and costly, and it was often difficult to see the whole picture clearly. Later, with the advent of big data and artificial intelligence, a method called "knowledge graph" was used to connect these scattered pieces of information (MacLean, 2021). It doesn't focus on fancy algorithms. The core is actually very simple: drawing all kinds of molecules, genes and their relationships into one diagram, so that the machine can understand "who is related to whom". Nowadays, many studies have found that such maps can be useful in drug discovery, target prediction, and even side effect analysis - it's equivalent to adding "common sense" to the model (Zhou et al., 2024). Therefore, constructing and researching knowledge graphs of molecular interactions not only enables us to have a more comprehensive understanding of life systems, but also provides new ideas and support for new drug development and disease diagnosis and treatment.

At first, the concept of "knowledge graph" did not emerge in the field of scientific research, but originated from the technology that Google used to improve search results. Unexpectedly, a few years later, it shone brightly in biomedical research and became an important tool for integrating complex data and assisting in analysis (Nicholson et al., 2020). Nowadays, almost all research directions related to biological data are attempting to use it to sort out those seemingly disordered information. Some researchers use it to integrate information such as genes, proteins, and compounds from different databases onto a single large graph. For instance, RNA-related

knowledge graphs almost cover all known RNA interactions. Some people also use it to explore new uses of drugs. By connecting drugs, targets and disease nodes, algorithms can discover potential new relationships (Zhou et al., 2024). In clinical diagnosis, it can also help doctors quickly match symptoms with diseases, improving the efficiency of diagnosis and treatment. A more complex approach is to weave multi-layer data such as genomes and proteomes into heterogeneous networks, use maps to analyze key regulatory relationships, and even automatically absorb knowledge from literature through text mining. Although knowledge graphs have been widely applied, their construction efficiency, update speed and scalability are still insufficient. These problems are precisely the key issues that need to be overcome in the future.

Although knowledge graphs have achieved a lot in bioinformatics, when it comes to molecular interactions, there are still quite a few problems. Data comes from all directions and in various formats. How can we make this information speak "in the same language"? For instance, for core entities like proteins, semantic standardization is always misaligned. If not done well, it will lead to a mess in subsequent analyses. There is an even more intractable question - can the established atlas really help us infer new molecular relationships? These are precisely the things that this study aims to explore. First, we designed a set of construction plans, taking into account data cleaning, entity modeling and storage architecture, so that proteins, small molecules and the like can all have clear positions in the map. Then explore methods such as graph representation learning, network analysis, and semantic reasoning to see if they can help discover new interaction clues; Take protein-protein interactions as an example again, create a case map, and verify the reasoning results through experiments and literature. Finally, a prototype system was set up to make the spectra visible and interactive, which is convenient for both drug mechanism and disease network research. Compared with our predecessors, we pay more attention to the integration and exploration at the detailed level, connecting theory, algorithms and applications into a line, hoping to take the research on molecular interactions one step further.

## **2 Data Sources and Preprocessing**

### **2.1 Data types and sources**

To build a reliable knowledge graph of molecular interactions, the first step is often not modeling but "retrieving data". Genes, proteins, small molecules, pathways, and diseases - these pieces of information are scattered in different places, some in databases and some hidden in papers. For protein-protein interactions, the ones that people often look up include IntAct, STRING, and BioGRID. The relationship between drugs and targets is mostly derived from Drugbanks or ChEMBL. As for functional annotations and semantic systems, GO and UMLS are almost unavoidable sources. Of course, these data were never readily available. The naming conventions of different databases vary. The same protein may be called differently in different databases. It is necessary to unify the identification first. Some mutual records of literature have not been reproduced, and their credibility needs to be rated. The types of interactions we focus on mainly include protein-protein, protein-small molecule, and gene regulation, etc. Therefore, when collecting, we not only need to capture the database but also rely on text mining to extract the description of "who interacts with whom" from PubMed (Figure 1) (Feng et al., 2022). Structural resources can also come in handy, such as finding clues to protein complexes from the eutectic structure of PDB. Overall, this process is more like a "jigsaw puzzle": the data from different companies vary greatly, and only through repeated comparisons and cross-verifications can the entire chart be both comprehensive and reliable (Zhou et al., 2024).

### **2.2 Data standardization and cleaning**

Obtaining the initial data doesn't mean the work is over. The truly troublesome part often comes later - cleaning and standardization. Data from different sources are like speaking different dialects, with distinct names, formats, and symbols. Without uniformity, it's simply impossible to proceed further. Usually, we first sort out the naming conventions. For instance, proteins are labeled with UniProt ID, genes with NCBI Gene ID, and small molecules with DrugBank or ChEBI numbers. This way, there won't be any confusion during the merging process. Next, it is necessary to give these concepts a "home", using an ontology system like GO to define functions, categories, and levels. Relationships also need to be defined in a unified way. Protein interactions should be unified as "interacts\_with", and those with specific directions should be clearly labeled. The cleaning stage is more like a

meticulous job: duplicates need to be removed, low-credibility data should be filtered out, and missing attributes should be filled in - even filling in "unknown" first is better than leaving it blank. Finally, convert all formats into a unified structure, such as RDF or CSV, to facilitate the import into the database (Mavridis et al., 2025). The results we have sorted out this time contain approximately tens of thousands of protein entities and tens of thousands of interaction relationships, each accompanied by a source identifier. The data processed in this way is clean and uniform, and can finally support the subsequent construction of the graph (Schulz et al., 2013).

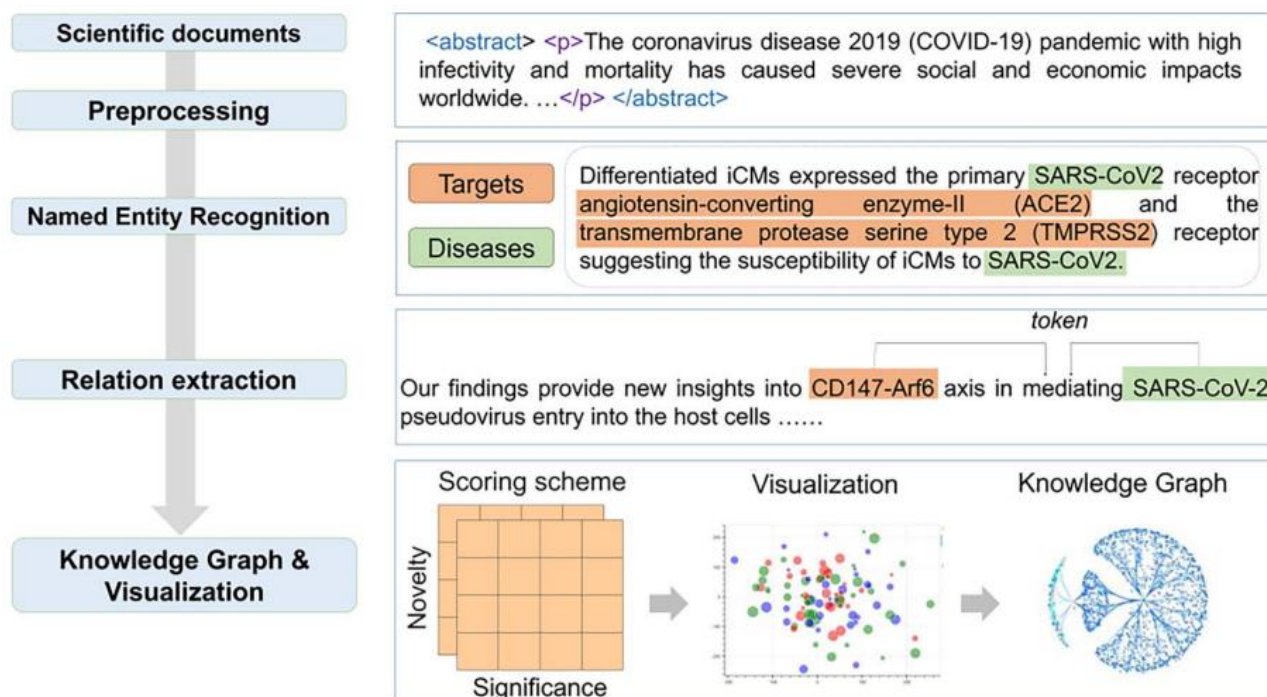


Figure 1 Architecture of the e-TSN web application (Adopted from Feng et al., 2022)

Image caption: The workflow involves several stages of scientific documents download, preprocessing, named entity recognition, relation extraction, knowledge discovery and visualization (Adopted from Feng et al., 2022)

### 2.3 Entity recognition and semantic normalization

When dealing with raw text data, the first problem one often has to confront is not the algorithm but rather the question of "what to call it". In different papers and databases, a protein may have several names - they need to be recognized first and then unified. The step of named entity recognition is to enable the system to automatically extract these biological names from the text and the relationships between them. For instance, in the sentence "P53 directly interacts with MDM2", two proteins and their interaction can be identified. Deep learning models are particularly useful here, especially those that can handle confusing aliases and irregular formats (Habibi et al., 2017). However, recognition is just the beginning; the real challenge lies in normalization. Like "TP53", "p53 protein" and "tumor suppressor protein 53", they all refer to the same thing - TP53. To make the graph recognize this, it is usually necessary to check the standard library, such as UniProt or NCBI, and match it with a unified ID. The same is true for compounds. One name may correspond to multiple trade names or chemical names. The naming of relationships should also be unified. It is best to group "inhibition" and "negative regulation" into the same category. As for those easily confused names, such as "APC", which are sometimes genes and sometimes protein complexes, one can only make a judgment by combining the context or the existing structure of the map. Only after all these have been processed can the data be considered "clean" and be firmly transformed into nodes and edges in the graph (Sung et al., 2022).

## 3 Knowledge Graph Construction Methods

### 3.1 Entity and relationship modeling

After the data is sorted out, the "framework" still needs to be set up. In the molecular interaction map, nodes usually include proteins, genes, small molecules, diseases, etc. The coding relationship between genes and

proteins should be marked. Each type of entity has its own attributes, such as the function and sequence of proteins, and the target and application of drugs. The core of the relationship is "interaction", but there may also be directional relationships such as "activation", "inhibition", and "combination". If it is too detailed, the data will become sparse. Therefore, it is often first unified as "interaction", and then further subdivided during reasoning. Cross-layer relationships can also be added, such as "gene mutations cause diseases" and "drugs treat diseases". Some relationships are aimless, while others have a direction. To complete the information, the reasoning relationship of "proteins of the same family" can also be added to help the model discover similarities (Zhou et al., 2024). The final graph framework is clear, facilitating subsequent expansion and analysis (Taneja et al., 2022).

### 3.2 Graph architecture design and storage model

After determining the entities and relationships, the next step is to figure out a way to "fit" these things in. There are generally two approaches: either follow the semantic web system of RDF or use the attribute graph model of graph databases. RDF emphasizes norms, with clear triples and the ability to directly use existing ontologies for semantic reasoning. However, once the data volume is large and the relationships are complex, the speed slows down. Property graphs are more like practical players. Nodes and edges can all be equipped with properties, making path lookup and centrality calculation fast. Many databases (such as Neo4j) rely on them. Considering that the molecular interaction graph has to handle tens of thousands of relationships, we prefer to choose the attribute graph. However, we cannot completely abandon semantics either. Therefore, adding ontology labels to the node attributes can be regarded as a compromise. In actual operation, each Protein node is labeled with a type, such as ' :Protein ', along with attributes like name and species. Information can also be hung on the edge, such as (BRCA1)-[:INTERACTS\_WITH {pmid:123456, method:"Y2H"}]->(PALB2), which not only shows who is interacting with whom but also knows the source of the evidence. Such an architecture is intuitive, flexible and convenient for expansion to larger diagrams in the future (Figure 2) (Tomaszuk et al., 2020; Alocci et al., 2015).

### 3.3 Automate the build process

It is almost impossible to manually sort out tens of thousands of pieces of information on molecular interactions bit by bit. Not only is it slow, but it is also prone to errors. So we simply set up an automated process to enable the knowledge graph to "grow" by itself. The entire process is roughly divided into four steps: first, capture the data; then, clean and transform it; next, import it into the graph database; and finally, check it once. The step of extracting data is the most complicated. You have to write scripts to call the API, pull protein data from NCBI and UniProt, crawl abstractions from PubMed, and then run NER models in batches to extract interaction information. After obtaining the raw data, it is necessary to convert the format, turning them into nodes and edges, and also perform normalization at the same time, such as merging synonyms into one entity. The conversion results will first be saved in the CSV file so that people can check them at any time. Next, import Neo4j, load nodes and relationships with batch tools, and configure indexes to improve query speed. After the graph is completed, it still needs to be verified whether there are any isolated nodes, whether the relationships are correct, and whether the PMids are accurately connected. If they are not up to standard, we will go back and change the rules to run again. If this cycle is repeated several times, the graph will become cleaner and cleaner. Now, after running one round, it only takes a few hours from the raw data to the formation of the graph. The relationships are almost all traceable, and updates are much more convenient (Clancy et al., 2019; Li et al., 2020).

## 4 Graph Representation and Analysis Methods

### 4.1 Knowledge representation learning

Building a map is just the beginning. The truly interesting part lies in whether one can "learn" something from it. We hope that these molecules and relationships are not just nodes and lines, but can be transformed into numbers that machines can understand. The approach is not complicated. To put it simply, it is to transform entities and relationships into vectors. This step is called embedding. After being converted into vectors, the model can use them to calculate similarities and make predictions. Different algorithms have different approaches. Some treat relations as translations (such as TransE), while others use ComplEx number Spaces to represent complex relations (like complex, RotatE) (Sun et al., 2018). There are many types of relationships in molecular interaction graphs, and they often have directionality. Therefore, we tend to choose models that can handle asymmetric



relationships. During training, some "fake data" also needs to be fed, such as randomly paired proteins, to help the model learn to distinguish between true interactions and false interactions. After training, proteins of the same type will automatically cluster together, and those with similar functions will also be closer, indicating that the structure and semantics have been embedded. After that, these vectors can still be used - for example, to find similar genes, predict new interactions, and conduct classification analysis. We also try to make the model "clearer", combining path or rule information to make the prediction more interpretable (Hu et al., 2024). Ultimately, this vectorized representation transforms the knowledge graph from merely a stack of information into a knowledge network that can be truly understood by the model.

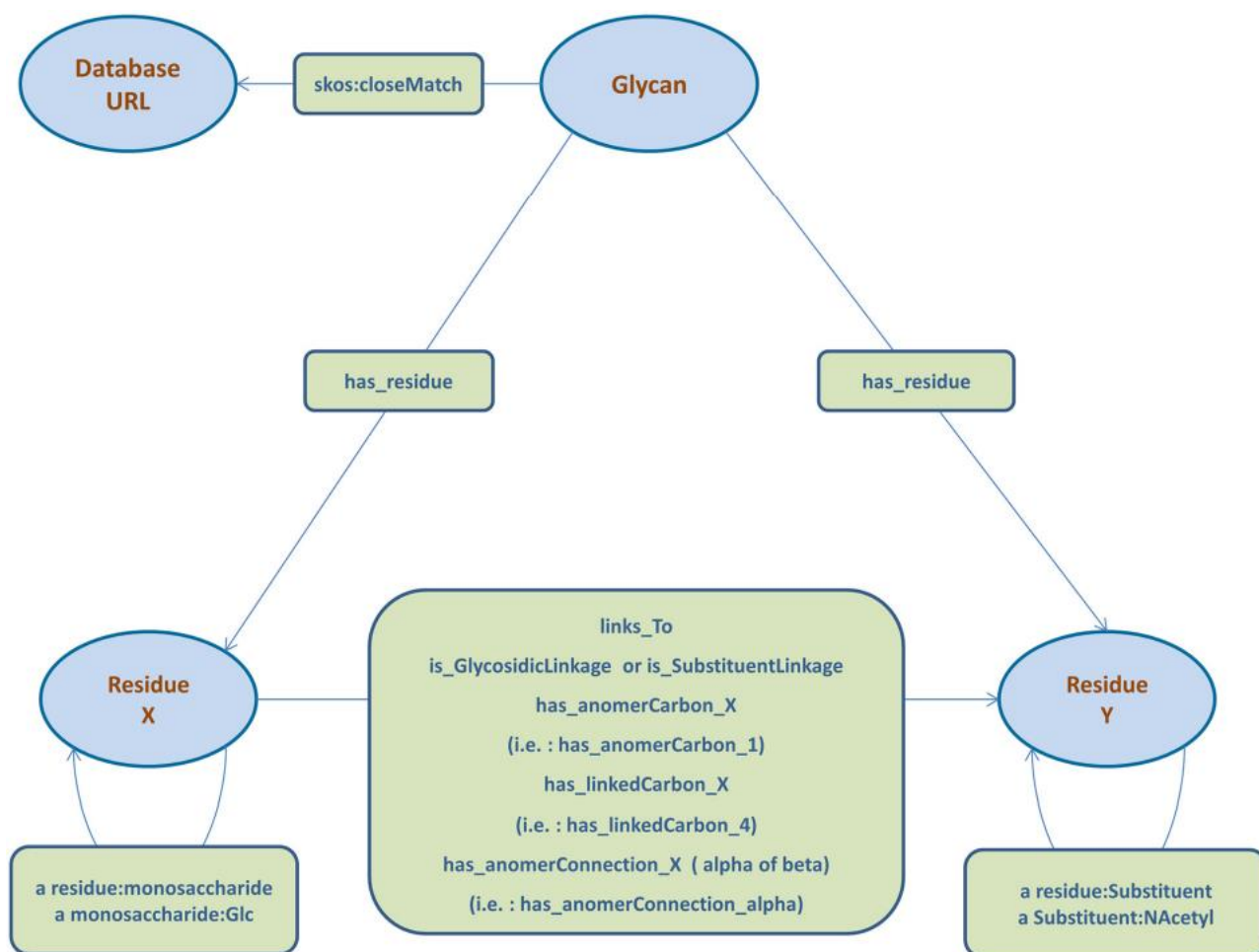


Figure 2 Ontology overview (Adopted from Alocci et al., 2015)

Image caption: Overview of the ontology developed for translating glycan structures into RDF/semantic triples. The figure shows all the predicates and the entities used for defining a glycan structures into the RDF triple store (Adopted from Alocci et al., 2015)

## 4.2 Network topology analysis

The molecular interaction knowledge graph is actually a vast network. Just looking at the nodes and lines doesn't make much sense; one has to find ways to read out the structural rules within it. We start with topological analysis and first calculate the importance of nodes, that is, who is more "critical" in the network. Some proteins are highly linked and have a high degree of centrality, often serving as hub molecules. There are also some that have few connections but are in special positions, connecting different modules like Bridges. Once such "bottleneck" proteins malfunction, the entire pathway may be affected. In addition to individual nodes, we also examined the group structure and used community detection algorithms to divide the network into compact small clusters. The results are quite interesting. For instance, some modules are entirely composed of immune-related proteins, while others focus on cell cycle regulation. Path analysis is more like looking for clues in a diagram - two seemingly unrelated molecules can sometimes connect after just two or three steps, and this might be a new regulatory

pathway. We even found some potential associations in the disease subplots, such as the indirect connection of Tau and GSK3 $\beta$  through signaling pathways. Overall, these topological features reveal the hierarchy and patterns within the network and also provide many new hypotheses for subsequent biological experiments (Mall et al., 2017; Seoane, 2024).

### 4.3 Semantic reasoning and relationship prediction

The beauty of a knowledge graph lies not in how much data it stores, but in what it can "guess". It is not merely a warehouse; it is more like a reasoning system. In the molecular interaction atlas, we hope it can identify missing connections on its own, such as predicting protein interactions or drug effects that have not yet been verified. There are generally two approaches: relying on rules or relying on models. Rule-based reasoning is logical. For instance, "If X activates Y and Y inhibits Z, then X might indirectly affect Z." Or perhaps A and B interact with each other, and B and C interact with each other. Then A and C might be on the same path. Such rules are clear but incomplete and are also easily disturbed by noise. So more people choose statistical methods - turning all the molecules into vectors to calculate which combinations are the most "compatible". We also predicted new interactions on the protein map in this way and compared them with the experimental data. Some of them were even confirmed by the literature. Later, we tried graph neural networks, enabling the model to aggregate neighbor information in subgraphs, making more accurate predictions and even telling you "why" - for instance, both A and B are connected to the key protein C (Kishan et al., 2020; Liu et al., 2021). Reasoning does not equal discovery. Predictions will eventually be verified, but it can help us pick out the most promising few from tens of thousands of molecules, greatly saving experimental energy.

## 5 Case Study: Knowledge Graph of Protein-Protein Interactions

### 5.1 Case selection and dataset description

To see if this atlas construction method is really effective or not, we selected protein-protein interactions (PPI) as the test case. The PPI network is the most crucial part of molecular interactions, featuring dense information and numerous connections. It is involved in disease mechanisms and drug effects, and almost every systems biology research will encounter it. Nowadays, the PPI map of human beings has been developed quite extensively, but this time we do not aim for completeness; we only aim for accuracy. The data mainly comes from several common libraries: extracting human interaction records from BioGRID and selecting highly reliable data that has been supported by multiple experiments or multiple literature articles (Oughtred et al., 2020); The data of STRING is also used as a supplement, but only the part with a higher comprehensive score is selected (Popik et al., 2014). The annotation information of proteins was captured from UniProt and GO, while the disease associations were obtained from OMIM and DisGeNET. Finally, a bit of literature mining results were added, and some new interactions that were not yet included in the database were supplemented. When integrated, the map contains approximately 15,000 protein nodes, over a thousand disease nodes, and several thousand high-quality interaction relationships. We value credibility more than quantity. Interactions with weak evidence will not be included for the time being. For the convenience of calculation, this time the undirected interaction network is mainly analyzed, and the direction of regulation is not subdivided for the time being. All data have undergone automatic process cleaning and standardization, uniformly identified by UniProt ID, and finally imported into the graph database. The entire process can now be replicated.

### 5.2 Analysis process and visualization results

After constructing the protein interaction map, we first conducted an overall statistics: approximately 15,000 protein nodes, over a thousand disease nodes, and more than 60,000 interactions, presenting a typical scale-free distribution. A few proteins, such as p53 and EGFR, have extremely large connections. The network is almost fully connected, indicating that the data integration is good. In the centrality analysis, signal hubs such as TP53, MYC, and AKT1 rank among the top. Further examination of the MAPK pathway reveals that the core proteins form a tight subnet with a clear structure. When magnifying the submap of the apoptotic pathway, the interaction patterns of the caspase family, Bcl-2 proteins, etc. are consistent with the classical pathway, but unexpected nodes such as SIRT1 also appear, suggesting potential new regulation. Link prediction also discovered several possible new interactions, some of which have been confirmed by the literature (López-Cortés et al., 2018; Kim et al.,

2021). There is another small group of unknown functional proteins that interact closely or are involved in chromatin remodeling. These results indicate that the graph not only integrates effectively but also inspires new discoveries.

### 5.3 Biological validation and result discussion

The graph has been built and the analysis results have come out, but whether it is reliable or not still needs to be verified through experiments. Let's first take a look at the "key proteins" unearthed through topological analysis. The top 20 results show that almost all of them are familiar faces - TP53, EGFR, and AKT1 are all included, indicating that the algorithm didn't make wild guesses. Interestingly, one of the originally less well-known proteins has unexpectedly become a bridge for the signaling pathway. After we knocked it down, the activity of both pathways dropped, which actually confirmed its crucial position. Then, we selected several pairs of new interactions predicted by the model for verification. We measured two pairs using the Co-IP experiment. One pair successfully detected the binding signal, while the other pair did not see any bands, possibly due to incorrect conditions. The successful couple has already been added to the map by us. Looking at the module level again, a group of mitochondrial proteins have aggregated into distinct subnets with consistent functions. The modules of those five unknown proteins are intriguing. We checked the co-expression data and found that they were always upregulated simultaneously. We also saw in a preprint that two of them did indeed appear in the same complex. Although these are just clues, they are sufficient to show that the direction of the graph prediction is valuable. Overall, it can not only reproduce known patterns but also generate new hypotheses. However, it is still a bit far from being "completely reliable", and experimental verification remains a crucial step (Collura et al., 2007; Zhan et al., 2024).

## 6 System Implementation and Application

### 6.1 System architecture and visual interface design

To make it more convenient for researchers to use this knowledge graph of molecular interactions, we have developed a prototype system, which can be regarded as integrating the "graph" into an interactive interface. The system architecture is not complicated, with the front-end and back-end separated: the back-end uses Neo4j and Flask, and the front-end uses Vue with D3. The database holds nodes and relationships, and the API is responsible for data retrieval, such as inputting `/query? protein=BRCA1` can return the list of proteins that interact with it. Users don't need to write query statements; they just need to click on the interface. The front end looks like a rotatable mind map, with proteins as dots, diseases as squares, and small molecules as triangles. The lines are connected, and by hovering, the relationships can be seen. Want to keep going? Just tap the node to expand. Too much information can also be filtered out, such as only looking at experimentally verified interactions. The sidebar can also display the centrality value and module affiliation, and the thickness of the edge represents the strength of the evidence. If you want to see if there is a connection between two proteins, just input them to highlight the shortest path. Click on any side and you can also see the literature information and PMID link. The prediction relationship is also clearly marked, along with the reasoning path. Several biologists tried it out and found the operation natural, as easy as browsing the web. Although the system is a prototype, it has stable performance, a fast response, and also leaves sufficient room for future expansion (Peng et al., 2022; Glen et al., 2025).

### 6.2 Application scenarios

The uses of the molecular interaction knowledge graph are actually quite diverse, and different people can do different things with it. For researchers, it is more like a knowledge map that can "come alive". In the past, to check the interaction of a protein, one had to switch back and forth between several databases. Now, just search for the name directly, and all the online, annotation and literature information will be clearly visible on one screen. For instance, when studying a new gene, one can simply click on the node to see which known carcinogenic proteins it is connected to, thereby inferring its possible pathways. For doctors, the atlas can also help identify the cause of the disease. When the patient's mutant gene is introduced, the system will mark the intersection points in the network and indicate which signaling pathways may have problems. It can also be used in new drug development. Researchers can look for potential targets in disease networks or explore new indications with

existing drugs. In education, it is also a useful teaching tool, allowing students to directly observe how signals are transmitted layer by layer. Data engineers can also use it as a graph analysis platform to extract features, calculate network metrics, and even connect to AI systems for question-and-answer. Overall, this atlas can not only be searched, viewed and "thought about", but also will be indispensable in the future, whether in scientific research, clinical practice or teaching (Lu et al., 2025).

### 6.3 Scalability and interoperability

To ensure that this knowledge graph system can survive for a long time, we left as much leeway as possible during the design process. The data part is the most prone to change - new molecular types and new interaction relationships can emerge at any time, so the system is not dead. To add content, simply give the node a new label and import new data. There is no need to change the architecture. The ETL process can also be run repeatedly, just like regularly updating BioGRID data, with the difference parts directly imported incrementally. If the data keeps piling up larger and larger, switching to the enterprise version of Neo4j or distributed deployment can also hold up, and the front end doesn't need to modify a single line of code. Functionally, we follow a modular approach. New features can be used as soon as they are installed, just like plugins. I tried adding a relationship prediction module before, and it only took a few lines of interface code to run it successfully. It won't be difficult to add path algorithms or time series analysis later. Intercommunication has also been taken into account. The system supports apis and SPARQL endpoints, and external platforms can directly retrieve data. Each node on the interface can jump to an external database. For example, clicking on a protein will lead to Uniprot. We also tested importing the data into Cytoscape and directly connecting it to NCBI for comparison, and it went quite smoothly. There are also loopholes left in the security aspect, making it convenient to connect to OAuth or manage permissions later. Overall, this system is flexible enough, scalable and can also keep pace with other platforms (Stear et al., 2023; Glen et al., 2025).

## 7 Challenges and Prospects

### 7.1 Data heterogeneity and dynamic update issues

Although the knowledge graph of molecular interactions has a broad prospect, it also has many troubles. First of all, the data is in a mess, with too many sources and too diverse formats. It includes both structured data and a bunch of literature texts and semi-finished products. Different experiments and database standards operate independently, resulting in varying degrees of reliability. When constructing a map, it is often necessary to manually remove duplicates and unify naming. For instance, in PPI, the same protein may be written with several alternative names, and there may also be duplicate names across species. Multi-omics integration is even more challenging, as data conflicts arise as soon as they are fused. In the future, we will have to rely on smarter algorithms to automatically align and reduce manual patching (Xiao et al., 2023). The second issue is the update problem. Biological research is advancing at an astonishing pace. New discoveries are made every day, and relationships that were valid yesterday may be overturned today. If the graph is not updated in a timely manner, it will lead to the wrong rhythm. Especially for the relationship related to drugs or diseases, if the old information is not revoked, the model analysis will be biased. To solve this problem, it is not only necessary to add new data, but also to be able to remove the old and retain the version. Perhaps a timestamp, confidence level or even an expiration date can be added to each relationship. In addition, automatic text mining is added to enable the system to discover on its own which knowledge should be updated (Hoyt et al., 2019). Over time, redundant data also needs to be cleared to prevent the graph from becoming bloated. Ultimately, for this system to always "survive", it must be able to adapt to changes and evolve on its own. Only in this way can it truly become a breathing warehouse of biomedical knowledge.

### 7.2 Model interpretability and knowledge uncertainty

The "black box" problem of models is particularly prominent in the biomedical field. Algorithms can produce results but cannot explain "why" clearly, which makes it difficult for researchers and doctors to be fully convinced. For instance, the model claims that proteins A and B interact with each other, but if no reason is given - even if it's just a simple statement like "They are both involved in the same pathway" - the conclusion would seem empty. For this reason, many studies have begun to attempt to make models "speak": find paths, draw subgraphs, and add



rules. Sometimes the model can be interpreted as "A is connected to B via X and Y", or the most critical local structures can be marked. Such results are more acceptable. The ideal system in the future should be able to provide reasoning basis like a human being, rather than merely throwing out a score (Rajabi et al., 2022). On the other hand, the issue of uncertainty is also tricky. Biological knowledge has never been black and white. Some interactions only hold true in specific cells, and some conclusions remain at the hypothesis stage. If a graph is described entirely by definite relationships, it will instead be distorted. A more realistic approach is to score each edge and indicate the confidence level. For instance, 0.8 represents high support, while 0.3 indicates insufficient evidence. In this way, when doctors see the prediction that "drug X may act on disease Y", they can also have a clear idea. One more point needs to be reminded: The incompleteness of the map does not mean "there is no such relationship", but just "it has not yet been discovered". It would be best for future systems to present such a grayscale, for instance, allowing users to see confidence intervals, literature differences, and even providing a "counterexample" button to help people understand the boundaries of knowledge (Bahaj et al., 2024). Science has always advanced in uncertainty, and knowledge graphs should also learn to recognize this.

### 7.3 Future development direction

The molecular interaction knowledge graph is just the starting point; there are simply too many directions for future development. For instance, multimodal fusion - most of the current graphs only handle structured data, which seems a bit "thin". If the molecular structure, cell images and even the content of the literature could all be integrated, the information would be much more three-dimensional. Imagine that after the protein-protein interaction network is combined with the three-dimensional structure, researchers can directly see where the interaction interface is. For instance, intelligent question answering. In the future, researchers may not need to go through databases. They just need to ask, "Which proteins are involved in insulin signaling and are related to Alzheimer's disease?" The system can answer and also provide the literature path. The scene where a doctor asks about the patient's condition and the system makes a diagnosis is not far from reality. The collaboration of knowledge graphs is also a trend - relying on a few database teams for updates is clearly insufficient. In the future, perhaps everyone will be able to upload new discoveries, allowing knowledge graphs to grow and roll like Wikipedia. Of course, to ensure quality, audits and standards are indispensable. There are also more cutting-edge directions such as cross-species and spatio-temporal maps, which not only allow for the observation of differences but also the tracking of dynamic changes in life processes (López et al., 2024). Even the graph can guide experiments in reverse, with machines proposing hypotheses, automatically verifying them, and then providing feedback for updates. By then, scientific research and Atlantis may have merged into one. In other words, this technology is just getting started, and its future is more wonderful than we can imagine now.

## 8 Conclusion

This research is essentially answering a question: How to organize the scattered information on molecular interactions into a knowledge system that can be understood by computers and directly used by researchers. Let's start with the most basic data. First, process the biological information from various sources, unify the names, relationships and formats, so that they can make sense on the same picture. Next comes the modeling process, which involves determining which nodes to count, which relationships to count, and what kind of structure to use to accommodate these elements. After the graph was set up, we also implemented an automated process to make the construction process more like an assembly line, allowing for repeated execution. The analysis part is more like mining: using embedding learning methods to convert the graph into vectors that machines can calculate, and then identifying key molecules and potential functional modules through metrics such as centrality and community. We chose protein-protein interactions as a case study, and the results were quite interesting. Besides reproducing the known relationships, we also unearthed some new interactions, and even had experimental support. Finally, we also developed a prototype system that can be both checked and viewed. Overall, this framework has successfully completed the process of "construction - analysis - verification - application", laying a solid foundation for more complex graph research in the future.

This research is not merely a matter of technology stacking. First, we brought the idea of knowledge graphs into the field of molecular interactions, attempting to use it to integrate biological data that "speak different languages"

from each other. Modeling, automated construction, and heterogeneous data fusion - these seemingly dull parts actually support the entire framework. We also wanted to see if it could truly serve scientific problems, such as identifying key proteins and recognizing functional modules. Therefore, topological analysis was introduced, treating the atlas as an experimental field in systems biology. The extent to which the model can explain has always been a pain point. We have added an "evidence path" in the reasoning stage, making the prediction no longer just about the result but also allowing us to see the process. On the other hand, we have also developed a practical system by ourselves, incorporating tens of thousands of protein interaction relationships, allowing researchers to directly query and analyze them. Even more surprisingly, through the prediction of the atlas, we really discovered and verified new protein-protein interactions, which indicates that it is not just a theoretical tool but can bring about new biological discoveries. It can be said that this work has opened up a gap in concept and built a step in practice, leaving room for more complex applications in the future, such as clinical diagnosis and drug design.

Although we have already built the prototype of the knowledge graph of molecular interactions, it is still far from being "complete". The next step is to expand the scale - not only to look at proteins, but also to incorporate aspects such as transcriptional regulation and metabolic pathways, making the map more like a true molecular panorama. Of course, this also means that the data is more complex and the relationships are more chaotic, and our methods need to be refined again. Algorithmically, we also aim to take another step towards explainability, such as introducing the causal reasoning approach, so that the model can not only predict but also clearly explain "why". Knowledge graphs are not static either. We plan to attempt modeling in the time dimension to see how the interaction network changes during the disease process. The system should also be more flexible, allowing for direct invocation in R or Python. It would be best if biologists could use it without learning new systems. We also plan to carry out more cooperation in the future and apply the atlas to real research, such as assisting in the analysis of rare disease mechanisms or the screening of drug targets. Ideally, researchers can not only use it but also feed their new discoveries back into the graph, allowing the system to "grow knowledge" on its own. Overall, our paths can be roughly divided into three: larger, smarter, and more practical. This is just the beginning. There is still a long way to go.

### Acknowledgments

I would like to express my heartfelt thanks to all the teachers who have provided guidance for this study.

### Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

### References

- Aloci D., Mariethoz J., Horlacher O., Bolleman J., Campbell M., and Lisacek F., 2015, Property graph vs RDF triple store: a comparison on glycan substructure search, *PLoS ONE*, 10(12): e0144578.  
<https://doi.org/10.1371/journal.pone.0144578>
- Bahaj A., and Ghogho M., 2024, A step towards quantifying, modelling and exploring uncertainty in biomedical knowledge graphs, *Computers in Biology and Medicine*, 184: 109355.  
<https://doi.org/10.1016/j.combiomed.2024.109355>
- Clancy R., Ilyas I., and Lin J., 2019, Scalable knowledge graph construction from text collections, In: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pp.39-46.  
<https://doi.org/10.18653/v1/d19-6607>
- Collura V., and Boissy G., 2007, From protein-protein complexes to interactomics, *Sub-cellular Biochemistry*, 43: 135-183.  
[https://doi.org/10.1007/978-1-4020-5943-8\\_8](https://doi.org/10.1007/978-1-4020-5943-8_8)
- Feng Z., Shen Z., Li H., and Li S., 2022, e-TSN: an interactive visual exploration platform for target-disease knowledge mapping from literature, *Briefings in Bioinformatics*, 23(6): bbac465.  
<https://doi.org/10.1093/bib/bbac465>
- Glen A., Deutsch E., and Ramsey S., 2025, PloverDB: a high-performance platform for serving biomedical knowledge graphs as standards-compliant web APIs, *Bioinformatics*, 2025: btaf380.  
<https://doi.org/10.1101/2025.03.09.642156>

- Habibi M., Weber L., Neves M., Wiegandt D., and Leser U., 2017, Deep learning with word embeddings improves biomedical named entity recognition, *Bioinformatics*, 33(14): i37-i48.  
<https://doi.org/10.1093/bioinformatics/btx228>
- Hoyt C., Domingo-Fernández D., Aldisi R., Xu L., Kolpeja K., Spalek S., Wollert E., Bachman J., Gyori B., Greene P., and Hofmann-Apitius M., 2019, Re-curation and rational enrichment of knowledge graphs in biological expression language, *Database*, 2019: baz068.  
<https://doi.org/10.1093/database/baz068>
- Hu Y., Oleshko S., Firmani S., Zhu Z., Cheng H., Ulmer M., Arnold M., Colomé-Tatché M., Tang J., Xhonneux S., and Marsico A., 2024, Path-based reasoning for biomedical knowledge graphs with BioPathNet, *bioRxiv*, 17: 599219.  
<https://doi.org/10.1101/2024.06.17.599219>
- Kim M., Park J., Bouhaddou M., Kim K., Rojc A., Modak M., Soucheray M., McGregor M., O'Leary P., Wolf D., Stevenson E., Foo T., Mitchell D., Herrington K., Muñoz D., Tutuncuoglu B., Chen K., Zheng F., Kreisberg J., Diolaiti M., Gordan J., Coppe J., Swaney D., Xia B., Van 't Veer L., Ashworth A., Ideker T., and Krogan N., 2021, A protein interaction landscape of breast cancer, *Science*, 374(6563): eabf3066.  
<https://doi.org/10.1126/science.abf3066>
- Kishan K., Cui F., and Haake A., 2020, Predicting biomedical interactions with higher-order graph convolutional networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2): 676-687.  
<https://doi.org/10.1109/TCBB.2021.3059415>
- Li N., Yang Z., Luo L., Wang L., Zhang Y., Lin H., and Wang J., 2020, KGHC: a knowledge graph for hepatocellular carcinoma, *BMC Medical Informatics and Decision Making*, 20(Suppl 3): 135.  
<https://doi.org/10.1186/s12911-020-1112-5>
- Liu Y., Hildebrandt M., Joblin M., Ringsquandl M., Raissouni R., and Tresp V., 2021, Neural multi-hop reasoning with logical rules on biomedical knowledge graphs, In: *European Semantic Web Conference*, Springer International Publishing, pp.375-391.  
[https://doi.org/10.1007/978-3-030-77385-4\\_22](https://doi.org/10.1007/978-3-030-77385-4_22)
- López V., Hoang T., Martinez-Galindo M., Fernández-Díaz R., Sbodio M., Ordonez-Hurtado R., Zayats M., Mulligan N., and Bettencourt-Silva J., 2024, Enhancing foundation models for scientific discovery via multimodal knowledge graph representations, *Journal of Web Semantics*, 84: 100845.  
<https://doi.org/10.1016/j.websem.2024.100845>
- Lu Y., Goi S., Zhao X., and Wang J., 2025, Biomedical knowledge graph: a survey of domains, tasks, and real-world applications, *arXiv Preprint*, 2501: 11632.  
<https://doi.org/10.48550/arxiv.2501.11632>
- MacLean F., 2021, Knowledge graphs and their applications in drug discovery, *Expert Opinion on Drug Discovery*, 16(9): 1057-1069.  
<https://doi.org/10.1080/17460441.2021.1910673>
- Mall R., Ullah E., Kunji K., D'Angelo F., Bensmail H., and Ceccarelli M., 2017, Differential community detection in paired biological networks, In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp.330-339.  
<https://doi.org/10.1145/3107411.3107418>
- Mavridis A., Tegos S., Anastasiou C., Papoutsoglou M., and Meditskos G., 2025, Large language models for intelligent RDF knowledge graph construction: results from medical ontology mapping, *Frontiers in Artificial Intelligence*, 8: 1546179.  
<https://doi.org/10.3389/frai.2025.1546179>
- Nicholson D., and Greene C., 2020, Constructing knowledge graphs and their biomedical applications, *Computational and Structural Biotechnology Journal*, 18: 1414-1428.  
<https://doi.org/10.1016/j.csbj.2020.05.017>
- Oughtred R., Rust J., Chang C., Breitkreutz B., Stark C., Willems A., Boucher L., Leung G., Kolas N., Zhang F., Dolma S., Coulombe-Huntington J., Chatr-Aryamontri A., Dolinski K., and Tyers M., 2020, The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions, *Protein Science*, 30: 187-200.  
<https://doi.org/10.1002/pro.3978>
- Peng J., Xu D., Lee R., Xu S., Zhou Y., and Wang K., 2022, Expediting knowledge acquisition by a web framework for knowledge graph exploration and visualization (KGEV): case studies on COVID-19 and Human Phenotype Ontology, *BMC Medical Informatics and Decision Making*, 22(Suppl 2): 147.  
<https://doi.org/10.1186/s12911-022-01848-z>
- Popik O., Saik O., Petrovskiy E., Sommer B., Hofestädt R., Lavrik I., and Ivanisenko V., 2014, Analysis of signaling networks distributed over intracellular compartments based on protein-protein interactions, *BMC Genomics*, 15(Suppl 12): S7.  
<https://doi.org/10.1186/1471-2164-15-S12-S7>
- Rajabi E., and Kafaie S., 2022, Knowledge graphs and explainable AI in healthcare, *Information*, 13(10): 459.  
<https://doi.org/10.3390/info13100459>
- Schulz S., and Jansen L., 2013, Formal ontologies in biomedical knowledge representation, *Yearbook of Medical Informatics*, 22(01): 132-146.  
<https://doi.org/10.1055/s-0038-1638845>
- Seoane L., 2024, Topological communities in complex networks, *arXiv Preprint*, 2409: 2317.
- Stear B., Mohseni Ahooyi T., Vasisht S., Simmons A., Beigel K., Callahan T., Silverstein J., and Taylor D., 2023, Petagraph: a large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data, *Scientific Data*, 11(1): 1338.  
<https://doi.org/10.1038/s41597-024-04070-w>
- Sun Z., Deng Z., Nie J., and Tang J., 2018, RotatE: knowledge graph embedding by relational rotation in complex space, *arXiv Preprint*, 1902: 10197.

- Sung M., Jeong M., Choi Y., Kim D., Lee J., and Kang, J., 2022, BERN2: an advanced neural biomedical named entity recognition and normalization tool, *Bioinformatics*, 38(20): 4837-4839.  
<https://doi.org/10.1093/bioinformatics/btac598>
- Taneja S., Callahan T., Paine M., Kane-Gill S., Kilicoglu H., Joachimiak M., and Boyce R., 2022, Developing a knowledge graph framework for pharmacokinetic natural product-drug interactions, *Journal of Biomedical Informatics*, 140: 104341.  
<https://doi.org/10.1016/j.jbi.2023.104341>
- Tomaszuk D., Angles R., and Thakkar H., 2020, PGO: describing property graphs in RDF, *IEEE Access*, 8: 118355-118369.  
<https://doi.org/10.1109/ACCESS.2020.3002018>
- Xiao Y., Steinecke D., Pelletier A., Bai Y., Ping P., and Wang W., 2023, Know2BIO: a comprehensive dual-view benchmark for evolving biomedical knowledge graphs, *arXiv Preprint*, 2310: 03221.  
<https://doi.org/10.48550/arxiv.2310.03221>
- Zhan X., Li H., Jin J., Ju X., Gao J., Chen X., Yuan F., Gu J., Xu D., and Ju G., 2024, Network pharmacology and experimental validation to explore the role and potential mechanism of Liuwei Dihuang Decoction in prostate cancer, *BMC Complementary Medicine and Therapies*, 24(1): 284.  
<https://doi.org/10.1186/s12906-024-04572-5>
- Zhou C., Cai C., Huang X., Wu S., Yu J., Wu J., Fang J., and Li G., 2024, TarKG: a comprehensive biomedical knowledge graph for target discovery, *Bioinformatics*, 40(10): btac598.  
<https://doi.org/10.1093/bioinformatics/btac598>



#### Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---