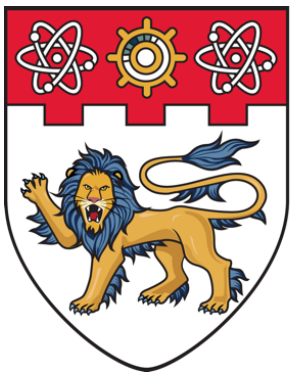


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

DEVELOPMENT AND APPLICATION OF A NOVEL NETWORK PHARMACOLOGY AND REVERSE DOCKING METHOD FOR NATURAL PRODUCT DISCOVERY: A CASE STUDY WITH GASTRODIA ELATA (TIAN MA)

SUN AO

SCHOOL OF BIOLOGICAL SCIENCES

2024

DEVELOPMENT AND APPLICATION OF A NOVEL NETWORK PHARMACOLOGY AND REVERSE DOCKING METHOD FOR NATURAL PRODUCT DISCOVERY: A CASE STUDY WITH GASTRODIA ELATA (TIAN MA)

SUN AO

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Master of Science

2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

1/16/2024 Sun Ao

. . . . . .

Date [Input Name Here & Sign above]

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Su I-Hsin

[Supervisor]

[Input Date Here]

Jan 22, 2024

Date

Mu Yuguang

[Co-Supervisor]

Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

*(A) This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

1/16/2024	Sun Ao
. . . .	. . . . .
Date	[Input Name Here & Sign Above]

Content

Abstract

Gastrodia elata, commonly referred to as 'Tian Ma' in China, is a prominent Traditional Chinese Medicine (TCM) utilized in clinical treatments for neurological disorders, hypertension, and other ailments. Despite extensive research on this TCM, numerous studies have primarily focused on its traditional or well-established ingredients, diseases, and pathways. Leveraging advancements in computational techniques, our research endeavors to uncover potential 'Molecule-protein-pathway-disease' axes, aiming to bridge the gap between traditional knowledge and modern drug discovery. Our methodology commences with the integration of TCM compound databases and literature reviews to predict Tian Ma's potential efficacious molecules. With our curated list, we employ AlphaFold2 for Reverse Docking—an AI-assisted visual screening—to pinpoint potential target proteins for each molecule. Subsequent integrative analyses involve GO and KEGG functional annotations, complemented by Disease Association Analysis using DisGeNET and Human Phenotype Ontology (HPO) databases. Utilizing Cytoscape software, we visualize the intricate network, striving to identify the potential 'Molecule-protein-pathway-disease' axes. Upon identification of promising axes, we engage in a preliminary 'dry lab' verification using Molecular Dynamics via Gromacs to ascertain the stability of the compounds. Positive outcomes in this phase will guide our evaluation on the viability of advancing to 'wet lab' pharmacological verification, ensuring a comprehensive and methodical approach to harnessing the therapeutic potential of 'Tian Ma'. Furthermore, I also developed a python-based software for quick literature review with NLP and machine learning applied.

Introduction

Gastrodia elata: The Versatile Tian Ma

Gastrodia elata, commonly known as Tian Ma, is derived from the dried tuber of the orchid plant Gastrodia elata Bl. According to Pharmacopoeia, the harvesting should occur from around 'Lidong' (the start of winter, one of China's 24 solar terms) until just before the Qingming festival in the subsequent year. Following harvest, it's imperative to clean the tuber immediately, steam it thoroughly, and then dry it at a low temperature, then cut in slices and ready for used [1].

As one of the most widely used TCM, this product is renowned for its efficacy in treating neurological disorders. Traditionally, it has been a remedy for conditions such as convulsions, vertigo, paralysis, and epilepsy. Its effectiveness against ailments like tetanus, asthma, and immune dysfunctions further emphasizes its versatility [2]. Additionally, its documented role in managing dizziness and certain emotional symptoms solidifies its reputation in TCM [3].

Historically prescribed for headaches and dizziness, it hints at its potential in preserving neuronal functions [4]. Contemporary research has shed light on its active constituent, Gastrodin, which exhibits anti-seizure and neuroprotective effects. This compound augments the expression of the GABAA receptor, suggesting a pathway for creating new antiepileptic drugs inspired by TCM [5].

Apart from neurological benefits, Gastrodia elata has showcased profound advantages for gastrointestinal health. Research underscores its potential in repairing damaged gastric glands, boosting gastric acid secretion, and effectively reducing stomach inflammation, emphasizing its therapeutic promise for various gastric issues [6].

Gastrodia elata's therapeutic spectrum also encompasses mental health. Studies have pinpointed its antidepressant-like effects, particularly in models of chronic social defeat stress, positioning it as a potential natural solution for mood disorders [Yu-En Lin et al., 2018]. An exciting avenue of Gastrodia elata's application lies in neurodegenerative diseases. Research suggests that its long-term administration might offer therapeutic avenues for conditions like Alzheimer's disease, presenting a beacon of hope for natural interventions against such challenging disorders [7].

Additionally, Tian Ma has demonstrated potential in mitigating the inflammatory response associated with rheumatoid arthritis. Specifically, it attenuates inflammation in

rheumatoid arthritis fibroblast-like synoviocytes by inhibiting the NF-κB pathway, underscoring its therapeutic promise for this autoimmune disorder. [8]

Current Gaps

Despite the extensive studies on Gastrodia elata and its primary active component, Gastrodin, most research has focused on traditional 'single molecule and single target' approaches. This narrow focus limits our understanding of the broader therapeutic potential and the underlying mechanisms involving multiple targets and pathways. There is a significant gap in employing holistic, multi-target methodologies to explore the full potential of Gastrodia elata in modern medical research. This gap presents an opportunity for innovative research approaches that integrate traditional knowledge with contemporary scientific techniques.

Objective and Significance of This Work

Our research aims to bridge this gap by leveraging advanced bioinformatics tools and databases to develop a more efficient and high-throughput virtual screening process. This approach will not only enhance our understanding of Gastrodia elata's therapeutic effects but also set another approach for integrating TCM with modern drug discovery techniques. The significance of this work lies in its potential to uncover new therapeutic targets and applications for Gastrodia elata, thereby contributing to both TCM and modern pharmacology.

Reverse docking

Reverse docking is a commonly utilized computational technique in modern drug research[33]. In the early stages of drug development, it is employed to screen a broad range of protein target databases for target small molecular compounds, predicting potential active molecules. Traditional molecular docking addresses the question, "Which small molecule can bind to this protein?" In essence, it uses a known protein structure as the "receptor" and "docks" various small molecules to its binding site to assess their adaptability and potential binding strength[9]. This process subsequently filters potential small molecules for a specific target for further research validation. For instance, molecular dynamics software like GROMACS is used to rigorously verify the stability of its binding. Once confirmed, subsequent molecular biology experiments are conducted for validation. In contrast, Reverse Docking addresses the question, "Which proteins might this drug bind to?" Simply put, its approach is opposite to traditional docking. It uses a known small molecule as the "ligand" and docks it to multiple protein structures to determine which proteins it might bind to. This method is often used for the functional rediscovery of existing drugs. Beyond pharmacological research, it is also frequently applied in toxicological studies, such as predicting potential side effects of

existing drugs[10]. Thanks to advancements in modern computational technology, in our research, we have refined these methods to accommodate a broader scope and higher throughput of virtual drug screening studies.

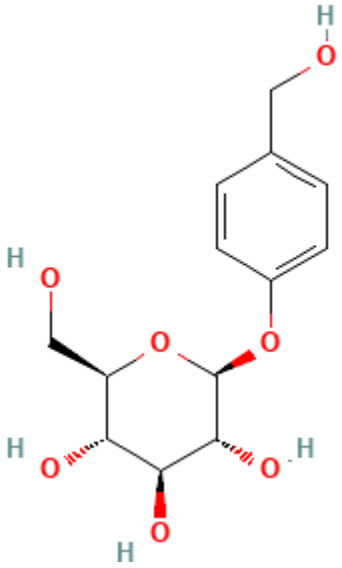
Network Pharmacology and its important role Traditional Chinese Medicine(TCM) Research

Historically, drug discovery was focused on maximizing the likelihood of a drug binding to a single protein receptor, making it the sole target of the drug. As time progressed and with technological advancements, the concept of systems biology became widely accepted, emphasizing a holistic approach to understanding biological processes. It became evident that drug discovery should not be a meticulous process targeting a single molecule and its pathway. For instance, Sildenafil citrate, initially developed by Pfizer as a new-generation antihypertensive and angina drug, showed minimal effects in these areas during clinical trials. However, it significantly induced male erections, leading to its primary current use, becoming a classic example of drug repositioning[11]. Therefore, drug development should be a holistic, multi-target, multi-pathway process. We need an efficient and high throughput means to screen at the beginning of the research. With the advent of bioinformatics and big data, network pharmacology emerged. Network pharmacology is an interdisciplinary field that studies the interactions between multi-target drugs, multi-component drugs, and biological networks. It combines computer science, systems biology, and pharmacology, aiming to provide a comprehensive view of the organism, revealing how drugs function at molecular, cellular, and organismal levels. Its foundational process can be summarized as data collection, network construction and analysis, drug screening, and experimental validation. It offers a fresh perspective for drug discovery and research, especially in TCM research. Due to the multi-component, multi-target nature of TCM, traditional single-target drug research methods struggle to fully unveil its mechanisms. Network pharmacology, as an efficient tool, can save significant time in preliminary screenings and provide potential research directions.

Our Works

In conclusion, research on TCM, traditional medicines, and natural products is on the rise. Our research aims to leverage existing bioinformatics tools and databases to develop a more efficient and high-throughput virtual screening process to aid TCM research. We observed that Gastrodia elata, widely used in traditional Chinese hospitals with evident efficacy, has been extensively studied. However, most of these studies are traditional 'single molecule and single target' research, lacking a holistic approach. Therefore, we hope to test Gastrodia elata while developing our research methods.

A



B

Figure 1: (A) Chemical structure of Gastrodin, one of the 15 molecules we used in this research, figure source: https://pubchem.ncbi.nlm.nih.gov/compound/Gastrodin; (B) UDP-glucuronosyltransferase 1A7 protein 3D structure from AlphaFold2, one of the potential target proteins of Gastrodin, figure source: https://www.alphafold.ebi.ac.uk/entry/Q9HAW7.

Developing a New Tool for Literature Review

In the realm of biomedical research, efficient literature retrieval and analysis are key steps in acquiring up-to-date scientific information and trends. Extensive literature retrieval and searching can guide our research, as well as ensure innovativeness and avoid duplication with previous studies. However, with the ever-increasing volume of scientific publications, traditional manual retrieval methods have become increasingly time-consuming and inefficient. Therefore, the development of a tool capable of rapidly and accurately retrieving and analyzing a vast array of literature has become particularly important. This study introduces a comprehensive literature retrieval and analysis software developed in Python, designed to address this issue.

This software integrates the Entrez interface of the Bio package, linear regression models, k-means and k-medoids algorithms, the Web of Science Impact Factor database, and Natural Language Processing (NLP) technologies. It automates various tasks, including retrieving literature from PubMed and MeSH databases, predicting research

trends, analyzing journal impact factors, matching keywords, and visualizing data. These features significantly enhance the efficiency of literature retrieval and analysis, aiding researchers in faster acquisition and processing of relevant information, thereby accelerating the pace of scientific research.

In this paper, we will also detail the design philosophy, functional implementation, and application cases of the software, demonstrating its practical value and potential applications in biomedical literature retrieval and analysis.

Method and Materials

List of Active Ingredients Tian Ma(Gastrodia elata)

The list of active compounds of Tian Ma was collated by our collaborators from public TCM databases such as TCMSP and BATMAN-TCM[13-14]. A subsequent literature review was conducted to filter out the 15 compounds with the most potential research value for our subsequent studies (for the complete list and the 15 potential active molecules, refer to Supplementary Data 1 and Supplementary Data 2). Due to the complexity and lengthiness of the IUPAC nomenclature, we utilized their sequence numbers from the compound list as indices for subsequent analyses. Thus, the fifteen compounds are identified as 9, 10, 15, 24, 36, 39, 42, 46, 47, 48, 49, 54, 55, 121, and 123.

Reverse Docking

For reverse docking, we employed the human protein database from AlphaFold2[15-16]. The docking software utilized was qvina2.1[17], an improved version tailored for high-throughput molecular docking based on AutoDock Vina. PointSite was employed for the identification of protein ligand binding atoms[18].

Network Pharmacology

We utilized three databases for our analysis: KEGG (Kyoto Encyclopedia of Genes and Genomes) [19-21], GO database (Gene Ontology database) [22-23] , and DisGeNeT[24]. The KEGG and GO databases were employed to perform functional enrichment and mapping of the protein list derived from reverse docking. Meanwhile, the DisGeNeT database was used for disease association analysis of the list.

Data analysis and Visualization

All data were saved in the CSV file format. Data processing was conducted using Python 3.10 and involved the invocation of specific Python packages. The packages utilized for data handling included: numpy, csv, pandas, sqlite3, scipy, and statsmodels. For mapping and functional enrichment, the packages employed were: bioservices, requests, goatools, and mygene. Data visualization was achieved using matplotlib and seaborn. All analysis workflow scripts were written by the author and will be uploaded to GitHub. For correlation analysis, we utilized the Linux GUI of Cytoscape to facilitate large-scale data processing.

For my software, these packages have been applied: NumPy, pandas, spacy, scispacy, argparse, Bio, re, sys, datetime, sklearn.

Systerm Information

The protein database used for reverse docking was deployed on our local server. Both the reverse docking and all data processing were conducted on our laboratory server running Ubuntu 22.04.2 LTS (GNU/Linux 5.15.0-76-generic x86_64). In terms of hardware, for the preparation of the protein database, developing software, I used our most powerful workstation whose processor is an AMD Ryzen™ Threadripper™ PRO 5975WX, and GPUs are dual NVIDIA RTX™ A6000. For data processing and the drawing of network pharmacology relationship diagrams, we utilized two Intel(R) Xeon(R) Silver 4116 CPUs @ 2.10GHz and two NVIDIA GTX 1080ti GPUs.

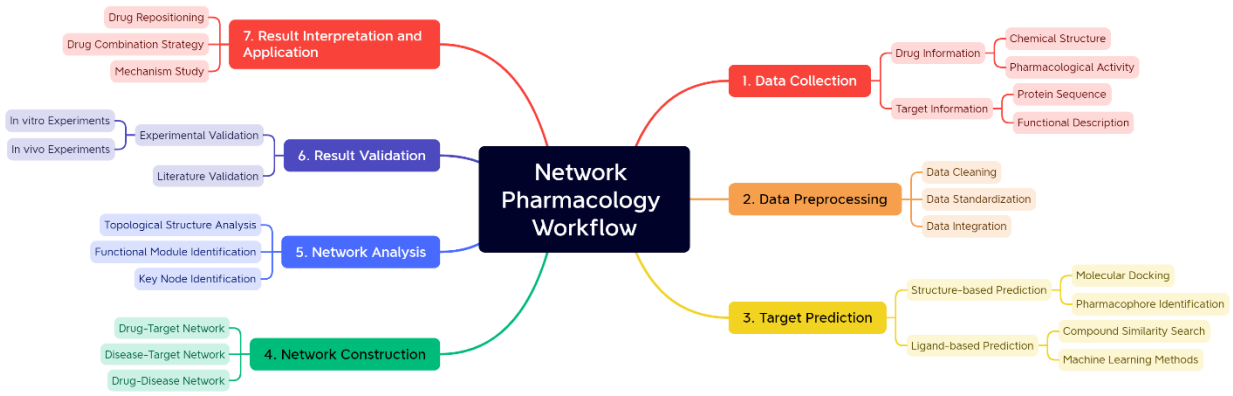


Figure 2: General workflow of Network Pharmacology

Results

Identification of Potential Protein Targets for Active Components of Gastrodia elata through Reverse Docking

Initially, our collaborators utilized the public Traditional Chinese Medicine Systems Pharmacology Database (TCMSP) to identify 154 molecular components from the traditional Chinese herb, Gastrodia elata. While each molecule was associated with a specific molecular structure, a few lacked names in the IUPAC nomenclature. Following a preliminary screening, they selected 15 molecules of significant research interest and handed them over to our team for further investigation. Given the extensive length of the IUPAC names, we opted to use the index numbers from the complete list as identifiers for these molecules, specifically: 9, 10, 15, 24, 36, 39, 42, 46, 47, 48, 49, 54, 55, 121, and 123.

To ensure comprehensive and accurate reverse docking, we turned to the human protein database of AlphaFold2(AF2) as our source of protein receptors. Benefiting from advancements in deep learning, AlphaFold2 offers one of the most exhaustive human protein databases to date, even encompassing proteins whose structures are yet to be definitively determined. Although AF2 provides an API access method, our objective to virtually screen the entire human protein database made individual API calls impractical. To achieve efficient parallel processing and facilitate repeated use, we decided to deploy a local protein database. We first downloaded the complete AF2 human protein database to our local server and employed the PointSite tool to predict potential binding sites for each protein. PointSite, a novel tool, specializes in accurately identifying ligand-binding atoms in proteins and can even discern binding sites on unbound proteins. After the PointSite prediction, the derived information was integrated with the original protein data for direct access by molecular docking software. This approach not only minimized manual intervention and potential errors but also leveraged our local computational power for parallel docking tasks, significantly saving time.

For molecular docking, we chose QuickVina 2.1 (qvina2.1), a tool previously developed by our lab as an optimized and accelerated version of AutoDock Vina. In tests on the core set of 195 protein-ligand complexes from the 2014 version of PDBbind, QuickVina 2.1 achieved a speed-up of up to 20.49 times compared to Vina. Moreover, qvina2.1's scoring mechanism, an enhancement over AutoDock Vina's, incorporates a new first-order consistency check heuristic to bolster significance testing, thereby improving reliability in identifying the lowest energy docking conformations.

Leveraging the capabilities of qvina2.1 and our newly established database, we scripted a high-throughput molecular docking screening for the 15 potential active molecules. Following the docking, based on qvina2.1's scoring of the docking results, we identified a total of 13,587 potential human protein targets. Subsequent analyses were based on these 13,587 potential proteins, and they all were stored in a UniProt IDs format combine with molecular indexes(supplement data 3).

Data Verification and Purification

Initially, we verified the accuracy of our data, ensuring that all genes in the protein list originated from human sources. To achieve this, we employed NCBI BLAST+ [25], loaded its refseq_protein database locally, and mapped the UniProt IDs to confirm that they all represent proteins expressed by human genes. Subsequently, we scripted in Python and utilized the 'mygene' package [26] to convert all UniProt IDs to NCBI's Entrez Gene IDs. This conversion served two primary purposes: on one hand is to re-verify the accuracy of our protein list by excluding proteins that did not match any Entrez Gene ID, on the other hands is to prepare for potential subsequent analyses. This is because although databases like KEGG and GO support queries using UniPort IDs, some of the other database may not really support using UniProt ID for searching. The Entrez Gene ID is a universally accepted format supported by most bioinformatics databases. After mapping, the Entrez Gene IDs were stored alongside the data from supplement data 3 in a new CSV file (supplement data 4).

Gene Ontology (GO) Functional Enrichment and Mapping

After data verification, our initial step was to associate the data with the GO database for enrichment analysis, which was divided into two main phases. In the first step, we conducted a functional enrichment analysis of the entire protein list. In the second step, we utilized the GO database to map each UniProt ID, annotating all GO terms associated with a particular UniProt ID. For the initial phase, we downloaded the GO ontology file, ‘go-basic.obo’, and the GO human gene association file, ‘goa_human.gaf’, from the official GO database website using the wget function on a Linux server. We scripted in Python and employed a specialized Python library for GO enrichment analysis called GOATOOLS. Specifically, we separated the protein list corresponding to each molecule and conducted enrichment analysis using parallel processing. This package calculates the enrichment P-value for each GO term post-enrichment and can directly apply the Benjamini-Hochberg method to the P-values for False Discovery Rate (FDR) correction, generating a column of P_fdr_bh values [27].

P-value measures the probability of obtaining test results at least as extreme as the observed results, under the null hypothesis. However, in multiple testing scenarios, the

P-value can be misleading due to the increased chance of Type I errors (false positives). The Benjamini-Hochberg method adjusts the P-values to control the FDR, which is the expected proportion of false positives among the rejected hypotheses. Therefore, FDR is a more accurate reflection of the true positive rate in the context of multiple hypothesis testing [34].

Given the large sample size of our analysis, there's a potential for numerous false positives, necessitating multiple hypothesis testing corrections for P-values. Based on this value, we filtered the enrichment results corresponding to each molecule's list, selecting the top 20 data based on P_fdr_bh values (supplement data 5) and merged them into a new list. Subsequent data visualization was also filtered based on P_fdr_bh values, with the results presented as shown in Figure 3A.

(A)

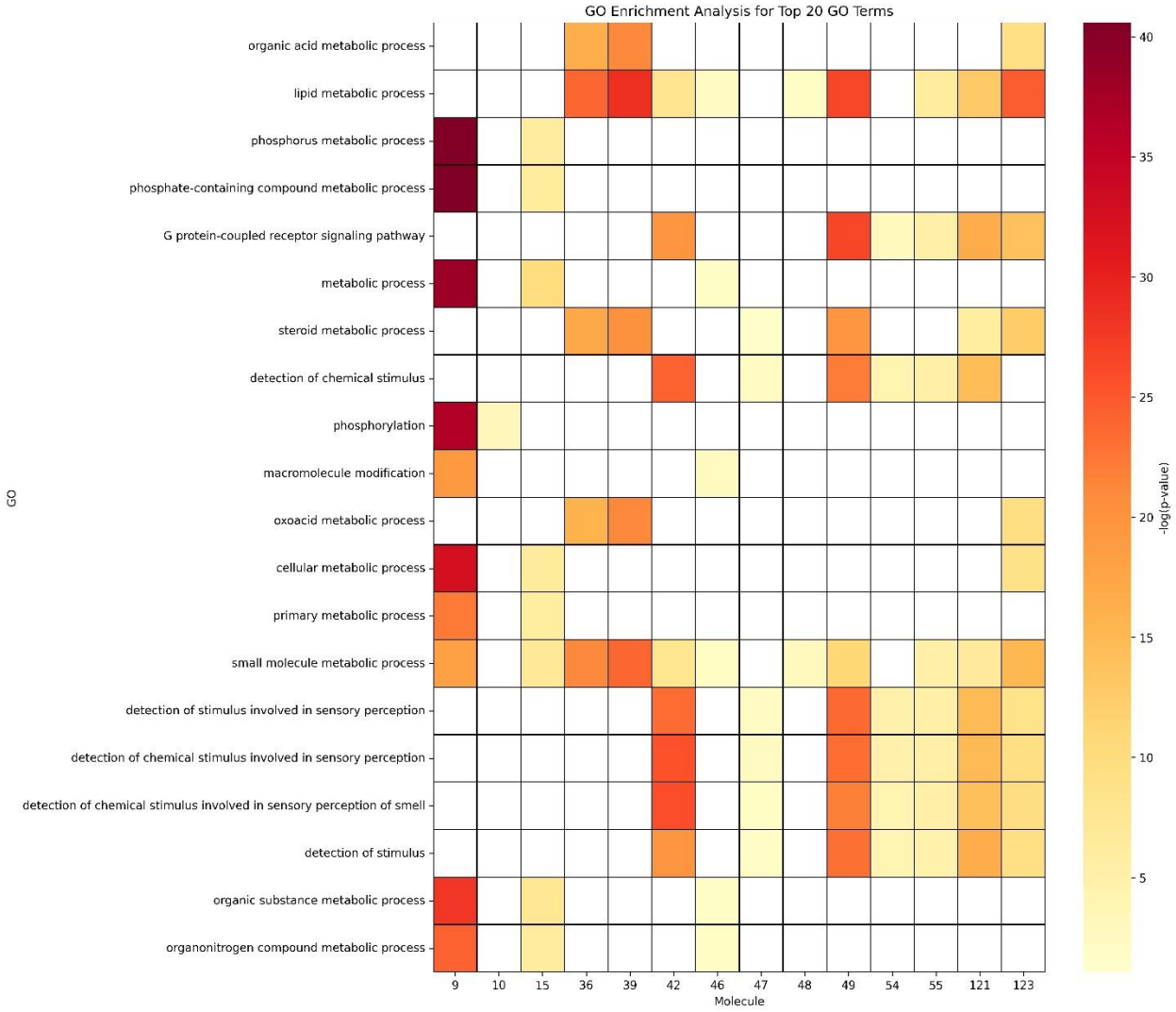
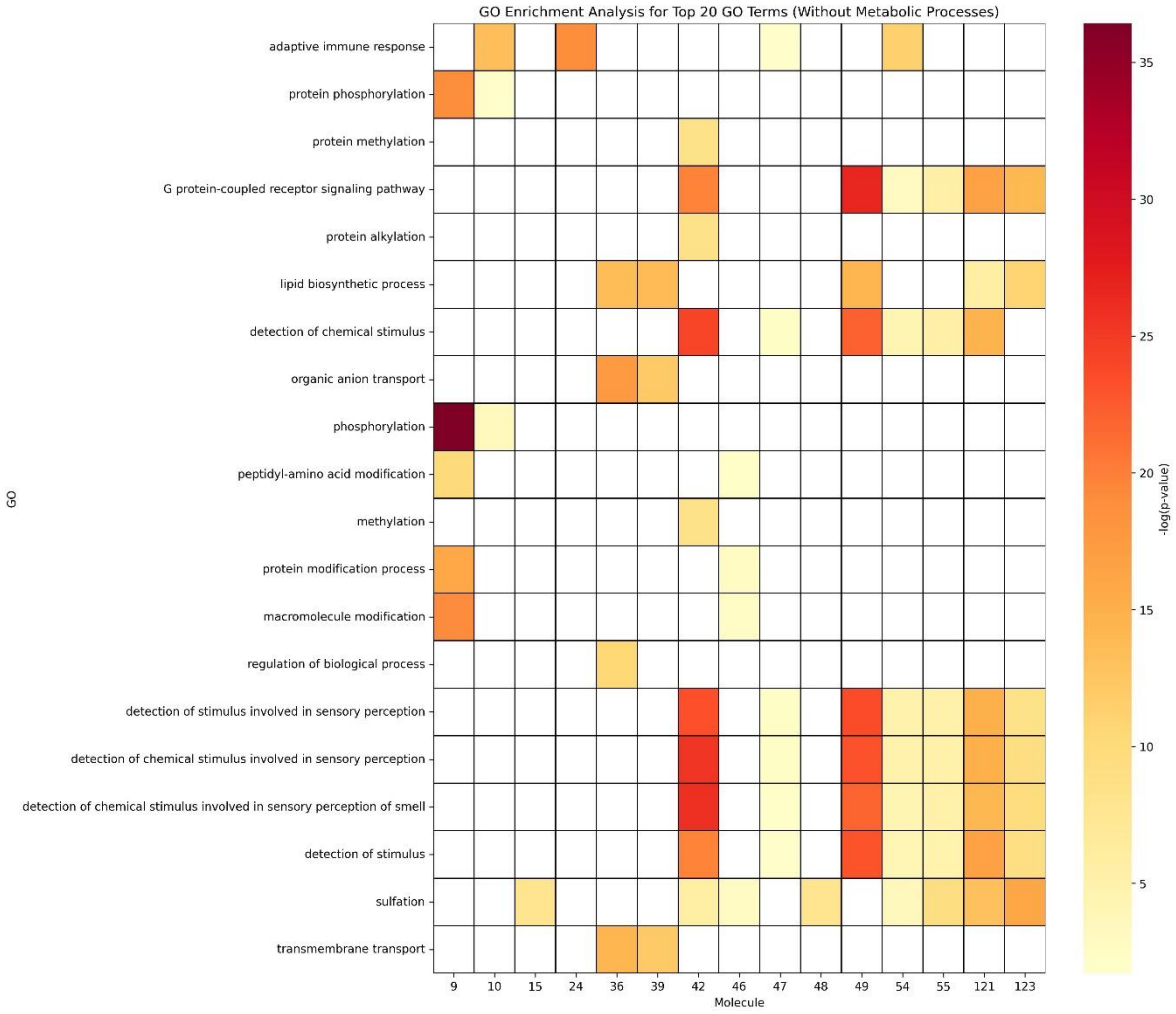


Figure 3: (A) GO enrichment analysis visualization for the top 20 enriched functions; (B) GO enrichment analysis visualization for the list that exclude 'metabolic process' related functions and visualized top 20 enriched pathway. The

visualization was based on python package numpy, pandas, matplotlib and seaborn.

(B)



From Figure 3A, we observe that a majority of proteins are significantly enriched in pathways associated with metabolic processes, such as "phosphate-containing compound metabolic process," "phosphorus metabolic process," "metabolic process," "cellular metabolic process," "lipid metabolic process," and "organic substance metabolic process." This indicates that these 15 molecules might profoundly influence the body's energy production, utilization, and overall metabolic equilibrium. Other pathways with notable enrichment include 'Phosphorylation,' 'G Protein-Coupled Receptor Signaling,' and 'Sensory Perception related processes.' The presence of 'Phosphorylation' suggests that Tian Ma plays a crucial role in cellular signaling and modulation. Phosphorylation, a vital biochemical process, is central to cellular signal transduction. It can activate or deactivate many protein functions, thereby affecting a range of cellular biological processes, from the cell cycle to cell growth, survival, and death [27]. 'G Protein-Coupled Receptor Signaling' is an essential pathway for transmitting external signals into the cell. When combined with 'Sensory Perception

related processes,' its significant enrichment suggests that these 15 compounds might influence various physiological responses, from sensory perception to hormone regulation. The pronounced enrichment of 'Sensory Perception Related Processes,' especially those associated with detecting external stimuli, resonates with Tian Ma's traditional use in Chinese medicine for conditions like vertigo and headaches. This suggests a potential influence of Tian Ma on the sensory system, possibly offering relief from stress responses triggered by external stimuli. Additionally, the recurring appearance of functions related to 'Lipid Metabolism' is intriguing. This suggests that Tian Ma might influence lipid synthesis, breakdown, or transport, a facet seldom highlighted in prior research. Dysregulation of lipid metabolism is linked to various diseases. For instance, disorders in lipid metabolism can lead to cardiovascular diseases, atherosclerosis, and coronary heart disease. Elevated levels of cholesterol and triglycerides are primary risk factors for cardiovascular diseases. Furthermore, lipid metabolic disorders are also associated with type 2 diabetes, fatty liver disease, and certain cancer types [29-30]. Considering Tian Ma's potential role in lipid metabolism, it may offer therapeutic potential for diseases related to lipid metabolic imbalances. The 'organic anion transport' also stands out, potentially related to metabolism, but this aspect has never been addressed in previous Tian Ma research.

However, we noticed that the top 20 pathways are predominantly associated with the 'metabolic process.' Over-enrichment in metabolic-related processes might cause us to overlook some specific functions or disease-related processes. Given our goal of identifying potential disease treatment targets, we are more focused on which functions or pathways are most relevant to potential disease treatments. Therefore, based on the merged data, we excluded processes related to 'metabolic process' and, using the same methodology, re-selected the top twenty processes, as shown in Figure 3B. In addition to the previously mentioned processes, this figure reveals more details, such as 'transmembrane transport,' which correlates with the previously enriched 'G protein-coupled receptor signaling pathway.' While the initial figure primarily highlighted 'macromolecule modification,' after excluding metabolic-related processes, we observed other processes such as 'protein modification,' 'methylation,' and 'protein alkylation, methylation, and phosphorylation.' These findings open avenues for potential future research and exploration, suggesting that these molecules may play roles in epigenetics, post-transcriptional regulatory mechanisms, and translation regulation. In conclusion, all the findings above can not only generally validate the traditional clinical uses of Tian Ma but also pave the way for its potential in addressing contemporary health challenges.

For the subsequent phase, we continued to employ GOATOOLS and the same database. However, the distinction lies in the methodology: we engaged in a mapping process rather than the prior enrichment. To elucidate, while enrichment associates proteins

with potential GO terms, mapping inversely associates GO terms with individual proteins. This approach is designed to distinctly delineate the signaling pathways pertinent to each protein, laying the groundwork for upcoming network pharmacology integrating analyses. So, we just making slight adjustments to our initial Python script to achieve mapping analysis and the resultant data was cataloged in supplement data 6.

KEGG Enrichment Result

Following the functional enrichment using GO, we sought to obtain more detailed pathway-related data to facilitate the subsequent construction of a comprehensive network pharmacology integrating analysis. For this purpose, we employed the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for further data mining. Similarly, our study was divided into two parts, akin to the GO analysis: one part focused on KEGG enrichment analysis, and the other on KEGG pathway mapping. The operational procedures during this process mirrored those in the corresponding sections of the GO analysis and are thus not reiterated here. However, due to the absence of root permissions, I was unable to set up the KEGG database locally. Therefore, we utilized another Python library, 'Bioservices', for the analysis. Bioservices offers a KEGG feature that allows remote access to the KEGG database. Yet, Bioservices does not offer P_fdr_bh correction functionality; it only displays the p-values post-enrichment. Consequently, we employed another Python library, 'statsmodels', which offers a function called ‘multitest’ to correct p-values to P_fdr_bh. Based on these packages, we scripted our Python analysis. The results of the first enrichment step are stored in Supplement Data 7, while the results of the second mapping step are in Supplement Data 8. Visualization tools for both sections were achieved using Python's pandas, numpy, matplotlib, and seaborn packages. The results of the pathway enrichment from the first part are shown in Figure 4. In this section, unlike the GO enrichment analysis, we did not exclude metabolic-related pathways. Instead, we extracted the top 15 from the protein list mapped for each molecule, integrated them into a comprehensive list, and visualized them directly. Upon visualizing the results of the KEGG enrichment, it is evident that the pathways with the highest enrichment are the 'Metabolic Pathway' and 'olfactory transduction'. However, discerning other data beyond these two pathways becomes challenging. Therefore, we employed a median-based method to simply filter out the 10 most significantly enriched signaling pathways for analysis. The median is a robust statistic, and from a statistical standpoint, it is insensitive to outliers. Given our large dataset, and the possibility that a signaling pathway might be enriched in multiple protein lists corresponding to various molecules, and that some pathways have extremely low p-values, we considered that these extreme p-values might influence the mean. Hence, we opted for the median-based approach. The ten most significantly enriched pathways, ranked from highest to lowest enrichment, are: Metabolic pathways,

Olfactory transduction, Pathways in cancer, Pathways of neurodegeneration - multiple diseases, Axon guidance, Neuroactive ligand-receptor interaction, Amyotrophic lateral sclerosis, Drug metabolism - cytochrome P450, Alzheimer's disease, and the Akt signaling pathway.

The significant enrichment of "Metabolic pathways", "Pathways in cancer", and "Pathways of neurodegeneration - multiple diseases" indicates that the proteins in our dataset play crucial roles in basic cellular metabolic activities, the onset and progression of cancer, and various neurodegenerative diseases. Additionally, the enrichment of pathways related to neural development and signal transduction, such as "Axon guidance" and "Neuroactive ligand-receptor interaction", underscores the importance of neural functions in these proteins. When combined with the results of the previous GO functional enrichment, a clear trend emerges. Apart from the metabolic-related pathways, a significant number of proteins in our dataset are associated with neural development, neural function, neurodegenerative diseases, and tumors. This suggests a strong association of our protein dataset with the nervous system, especially processes related to neural development and neurodegenerative diseases. Moreover, considering the highly enriched PI3K-Akt signaling pathway in KEGG, the highly enriched Adaptive Immunology Pathway in GO, and the mention of the pathway in cancer in both databases, we can also infer their relevance to tumors and immunity. This provides a general direction for our subsequent analyses.

For the second step, we employed a process similar to the one used in the GO analysis, with slight script modifications, to produce a KEGG pathway mapping file for each UniProt ID (Supplement Data 8).

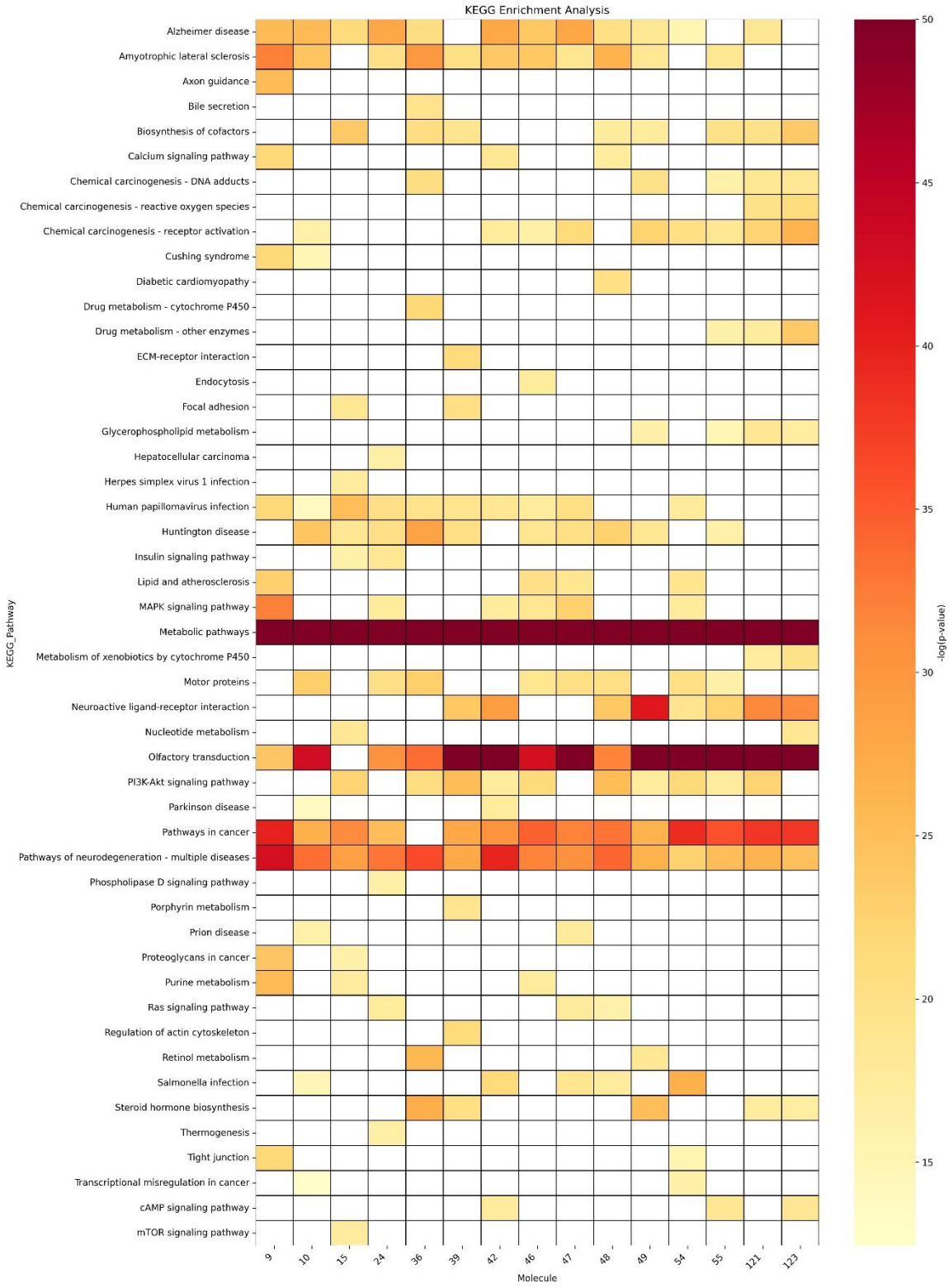


Figure 4: Visualization of the selected KEGG data via heatmap

Disease-related Analysis via DisGeNET Database

Subsequently, we employed our dataset to query the DisGeNET database. DisGeNET is a multifunctional research database primarily designed for exploring the molecular mechanisms of human diseases, characterizing disease genes, hypothesizing drug actions, and validating predicted disease genes [31]. DisGeNET offers two modes of association analysis: API and a local Structured Query Language database (SQL database). As an initial step, we intended to directly map all 13,587 proteins from our dataset. To facilitate rapid analysis, similar to our approach with the GO database, we attempted to deploy the DisGeNET SQL database locally. Firstly, we downloaded the DisGeNET SQLite 2020 - v7.0, the current release data source provided by GO, to our local server using the Linux wget tool and loaded it using the sqlite3 SQL database management tool. Subsequently, we scripted in Python utilizing its built-in sqlite3 module to access the database and perform the mapping. However, the local database did not support direct querying using UniProt ID. Hence, we resorted to using the previously mapped Entrez Gene IDs for our search, initially converting them to geneNID using the 'diseaseAttributes' in the database. To retrieve specific disease names, we needed to invoke the 'diseaseName' module in the database, where geneNID is the only supported format for querying, explaining our prior conversion of Entrez Gene IDs to geneNID.

To our surprise, none of the 13,587 proteins were enriched. Considering the possibility that the local database version might be outdated or not loaded correctly, we first verified its accessibility using SQL and confirmed it contained the data we sought. Following this, we rewrote our script, utilizing the requests package in Python to access the API link provided by DisGeNET and remapped. The outcome remained the same with all proteins off-target.This suggests that our approach to using the DisGeNET database might have been flawed.

Directly using proteins to search in the database like DisGeNET might not be appropriate for virtual screening. As a result, we decided to adopt a comprehensive approach for the selection of disease-related targets. Rather than relying solely on protein data from the databases, we opted for a more holistic strategy of manually collecting target data associated with diseases of our interest. These targets were derived from multiple sources, including the results of differential analysis from Genome-Wide Association Studies (GWAS), as well as disease target databases such as OMIM, DrugBank, GeneCard, and CTD. Additionally, targets mentioned in authoritative scientific literature were also considered. After collecting this data, it is then used as a reference dataset and cross-referenced with our protein dataset.

In future work, we plan to refine this section, with the specific improvements and more details discussed in the 'Future Work' section.

WangCai: A quick literature review software

During this study, we developed a comprehensive literature retrieval and analysis tool based on Python, aimed at optimizing the literature search and analysis process in the field of biomedical research. This software utilizes the Entrez interface provided by the Bio package to batch retrieve PubMed literature and downloads key information including titles, abstracts, PubMed IDs, journal names, publication dates, keywords, full-text links, and article types, saving them as CSV format files. This process not only enhances the efficiency of literature collection but also lays the foundation for subsequent analysis.

The software's second feature is trend prediction for the downloaded literature, using linear regression models to forecast the publishing trends of specific keywords. The accuracy of this function depends on the adequacy of the number of literatures and their broad time distribution. However, considering the limitations of linear regression in certain scenarios, we are exploring other machine learning methods to improve the accuracy and robustness of predictions, with a current focus on employing time series analysis methods as a substitute for linear regression.
The third feature focuses on analyzing the journals in which the articles are published. Initially, we manually downloaded the Web of Science Impact Factor database (2023 version). The software can annotate the literature with impact factors based on this database and add this information to the previously downloaded CSV files. Users can filter high-impact articles by entering the desired impact factor and generate new CSV files. Additionally, users can opt to visualize the analysis results, displaying the journals most receptive to the user's input keywords, thus providing intuitive guidance for journal selection.

To facilitate non-Linux users' access to the MeSH database, the software's fourth feature includes rapid downloading and decoding of the MeSH database. The fifth feature, one of the core functionalities, uses the MeSH database to match literature titles and abstracts, identifying keywords (such as diseases) within articles and adding MeSH term annotations to each literature entry in the newly generated CSV file. This feature plays a significant role in enhancing the efficiency and accuracy of keyword identification. Another core functionality of the software is the integration of Natural Language Processing (NLP), incorporating five pre-trained scispacy models: en_core_sci_md, en_ner_jnlpba_md, en_ner_bc5cdr_md, en_ner_bionlp13cg_md, and en_ner_craft_md[32]. These models identify keywords in abstracts and titles from various aspects such as biology, chemistry, pharmacology, and signal pathways. For instance, the en_ner_jnlpba_md model focuses on identifying entities in biomedical texts,

such as proteins and cell types, while the en_ner_bc5cdr_md model specializes in naming entity recognition of chemicals and diseases, suitable for clinical medicine and biomedical fields.

Finally, the software offers visualization features. First, it can visualize the results from the fifth and sixth steps, filtering MeSH terms or NLP keywords most relevant to the keywords. This functionality significantly reduces the workload of literature screening for users, enhancing efficiency compared to traditional individual review methods. Secondly, the software uses sklearn's k-means or k-medoids algorithms (with k-medoids being an option when fewer literatures are collected) for clustering based on two data sources: 1. Annotations from NLP and MeSH; 2. The abstracts of the literature. Considering the

convenience of individual users, the software also has a streamlined version packaged with pyinstaller, which can be run directly. Currently, we are optimizing and testing the software and preparing to develop APIs and a GUI interface in the future, to broaden the software's audience.

Overall, this software integrates multiple advanced technologies and tools, not only improving the efficiency and accuracy of literature retrieval but also providing biomedical researchers with a new, more efficient method of literature analysis through the application of data analysis and NLP technology.

Discussion

In this Research, we aimed to establish a comprehensive workflow for traditional Chinese medicine data mining, using Gastrodia elata (Tian Ma) as a test case, to elucidate potential molecular mechanisms and pathways associated with our dataset. We employed a combination of reverse docking, network pharmacology, and enrichment analysis, hoping to offer a more holistic perspective for traditional Chinese medicine research.
Reverse docking provided us with a unique approach to predict potential protein targets within our dataset. Despite its computational demands and inherent limitations, this method offers a rapid and efficient means to identify potential molecular interactions, significantly reducing the time researchers would spend on target selection based on literature.

Building on the results from reverse docking, network pharmacology further deepened our understanding of the predicted targets, offering a holistic view of potential interactions and pathways. This approach perceives diseases not merely as a result of a single dysregulated pathway but as a complex interplay of multiple pathways and interactions. Consequently, we aspire to establish a 'Molecule – Target Protein – Pathway – Function - Disease' axis to guide the research and development of traditional Chinese medicines.

In our test with Tian Ma, we utilized the KEGG, GO, and DisGeNET databases for enrichment analysis to determine specific pathways and processes that our proteins might be involved in. The KEGG analysis highlighted pathways related to metabolism, cancer, and neurodegenerative diseases, underscoring the multifunctional and multi- target nature of this traditional medicine. On the other hand, the GO analysis offered a broader perspective, emphasizing biological processes, molecular functions, and cellular components. These findings suggest a strong association of our protein dataset with the nervous system.

While our endeavors with the DisGeNET database did not yield the anticipated results, it did shed light on the limitations of bioinformatics tools. This serves as a reminder that, powerful as algorithms might be, there's a pressing need for data preprocessing and algorithmic optimization.

Future Work

Our future endeavors will primarily focus on optimizing methods and algorithms. Firstly, although we have established a novel database of proteins and their binding sites through reverse docking, the subsequent network pharmacology research process for Tian Ma still employs the classic network pharmacology workflow. Our exploration has revealed its limitations, particularly that the traditional process might not be suitable for handling such large datasets. In response, we plan to develop our own comprehensive workflow, starting from data preprocessing to association analysis, tailored to accommodate large-scale data.

Furthermore, while we have achieved enrichment results from molecules to proteins and then to KEGG and GO databases for our test dataset, a more comprehensive relationship network might require protein interaction data and disease association data within the dataset for network pharmacology. Due to time constraints, protein association analysis has not been conducted. Moreover, given the preliminary results from the DisGeNet enrichment analysis, we intend to adopt a different approach for disease association data. We will revisit the efficacy and primary diseases associated with Tian Ma as mentioned in traditional Chinese medical texts, correlate them with modern medical diseases, and, upon identifying specific diseases, combine them with existing KEGG and GO enrichment results. Subsequently, we will filter out related disregulated genes from multiple databases such as DisGeNet, HPO, and CTD, identify their intersections, extract the corresponding proteins from the database, and conduct association analysis with our protein list to identify potential targets and diseases. This might present a more rational approach.

Moreover, directly using proteins to search in the database like DisGeNeT might not be appropriate for virtual screening. As a result, we currently adopted a comprehensive approach for the selection of disease-related targets. Rather than relying solely on protein data from the databases, we opted for a more holistic strategy of manually collecting target data associated with diseases of our interest. These targets were derived from multiple sources, including the results of differential analysis from Genome-Wide Association Studies (GWAS), as well as disease target databases such as OMIM, DrugBank, Gene Card, and CTD. Additionally, targets mentioned in authoritative scientific literature were also considered.

To ensure the accuracy and relevance of the selected targets, we employed a method of cross-validation, selecting only those targets that were mentioned across GWAS analysis, databases, and literature. Subsequently, protein-protein interaction (PPI) analysis was conducted using the String database on potential targets identified through reverse

docking techniques. This step aimed to reveal the interactions between targets and explore their potential roles in the mechanism of action of drugs.
Ultimately, our goal is to construct a comprehensive drug-target-signal pathway-disease network. This network will not only aid in understanding how drugs exert their effects through specific targets and pathways but also provide deeper insights into the mechanisms of diseases. This systematic analysis approach is of significant value in the research of drug discovery and disease treatment strategies, offering theoretical foundation and practical guidance for precision medicine.

This integrated multi-source data approach, combining techniques from genomics, pharmacology, and bioinformatics, provides a comprehensive and in-depth framework for the identification and validation of disease targets. I'm currently optimizing and testing this approach. I believe through this method, we can more accurately pinpoint key targets associated with diseases, laying a solid scientific foundation for subsequent drug design and disease treatment.

If I'm fortunate enough to discover one or more 'Molecule – Target Protein – Pathway – Function - Disease' axes, our preliminary consideration for validation methods includes: firstly, using open-source databases, for instance, gene expression validation where we would download transcriptome datasets of healthy individuals and patients from the GEO database and then conduct differential analysis to verify the accuracy of our
predicted target genes and related diseases. Secondly, if the initial validation results are positive, we might attempt to use molecular dynamics software like Gromacs to verify the stability of the binding between small molecules and receptor proteins. Thirdly, once the results from the above validations are promising, we would forward the tested axes to our collaborative laboratories for molecular biology experimental validation.

Finally, in my research, almost every step of the script was written by myself, lacking automation. Therefore, future collaborations will focus on algorithm optimization to achieve a more robust automated process. Moreover, given the multi-target, multi-
functional, and complex nature of traditional Chinese medicine, we plan to explore the application of deep learning models to enhance the prediction accuracy of molecular mechanisms and potential therapeutic effects of traditional Chinese medicine.
Considering the advantages of deep learning in handling complex data patterns, capturing non-linear relationships, and automatic feature engineering, we believe it can offer us a more in-depth and accurate capture. Especially against the backdrop of the complex components of traditional Chinese medicine and their interactions, the attributes of deep learning might bring unexpected benefits to our research.

Conclusion

In this research, we embarked on a comprehensive exploration of traditional Chinese medicine data mining, using Tian Ma as a representative example. At the same time, I developed a literature review software by myself to ensure a more efficient retrieval and analysis. Through reverse docking, we identified potential protein targets, and the subsequent network pharmacology analysis provided insights into the intricate molecular mechanisms and pathways associated with our dataset. While our initial results from databases like KEGG, GO, and DisGeNET have been enlightening, they also highlighted the limitations of traditional workflows, especially when handling vast datasets. Moving forward, our primary focus will be on refining our methods and algorithms. We envision a future where our research is not only more automated but also enhanced by the integration of deep learning models, capitalizing on their ability to capture complex patterns and nonlinear relationships. This holistic approach, combining computational techniques with traditional Chinese medicine, promises to unveil deeper insights into the multifaceted nature of herbal treatments and their potential therapeutic effects.

Reference

[1] 国家药典委员会.中国药典[M].北京：中国医药科技出版社，2020:1088

[2] Tang, C., Wang, L., Liu, X., Cheng, M., Qu, Y., & Xiao, H. (2015). Comparative pharmacokinetics of gastrodin in rats after intragastric administration of free gastrodin, parishin and Gastrodia elata extract. Journal of ethnopharmacology, 176, 49–54. https://doi.org/10.1016/j.jep.2015.10.007

[3] Lin, Y. E., Chou, S. T., Lin, S. H., Lu, K. H., Panyod, S., Lai, Y. S., Ho, C. T., & Sheen, L. Y. (2018). Antidepressant-like effects of water extract of Gastrodia elata Blume on neurotrophic regulation in a chronic social defeat stress model. Journal of ethnopharmacology, 215, 132–139. https://doi.org/10.1016/j.jep.2017.12.044

[4] Tsai, C., Huang, C., Lin, Y., Lee, Y., Yang, Y., & Huang, N. (2011). The neuroprotective effects of an extract of Gastrodia elata.. Journal of ethnopharmacology, 138 1, 119-25 . https://doi.org/10.1016/j.jep.2011.08.064.

[5] Yang, C., Chiu, S., Liu, P., Wu, S., Lai, M., & Huang, C. (2020). Gastrodin alleviates seizure severity and neuronal excitotoxicities in the rat lithium-pilocarpine model of temporal lobe epilepsy via enhancing GABAergic transmission.. *Journal of ethnopharmacology*, 113751 . https://doi.org/10.1016/j.jep.2020.113751.

[6] Chen, C., Fu, Y., Li, M., Ruan, L., Xu, H., Chen, J., Zhao, W., Meng, H., Xing, Y., Hong, W., & Wang, J. (2019). Nuclear magnetic resonance-based metabolomics approach to evaluate preventive and therapeutic effects of Gastrodia elata Blume on chronic atrophic gastritis. *Journal of Pharmaceutical and Biomedical Analysis*, 164, 231–240. https://doi.org/10.1016/j.jpba.2018.10.035.

[7] Huang, G., Zhao, T., Muna, S., Jin, H., Park, J., Jo, K., Lee, B., Chae, S., Kim, S., Park, S., Park, E., Choi, E., & Chung, Y. (2013). Therapeutic potential of Gastrodia elata Blume for the treatment of Alzheimer's disease☆. *Neural Regeneration Research*, 8, 1061 - 1070. https://doi.org/10.3969/j.issn.1673-5374.2013.12.001.

[8] Li, Y., Wang, L. M., Xu, J. Z., Tian, K., Gu, C. X., & Li, Z. F. (2017). Gastrodia elata attenuates inflammatory response by inhibiting the NF-κB pathway in rheumatoid arthritis fibroblast-like synoviocytes. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 85, 177–181. https://doi.org/10.1016/j.biopha.2016.11.136

[9] Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

[10] Lee, M., Kim, D. Large-scale reverse docking profiles and their applications. BMC Bioinformatics 13 (Suppl 17), S6 (2012). https://doi.org/10.1186/1471-2105-13-S17-S6

[11] Ghofrani, H. A., Osterloh, I. H., & Grimminger, F. (2006). Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature reviews. Drug discovery*, *5*(8), 689–702. https://doi.org/10.1038/nrd2030

[12] Hopkins A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, *4*(11), 682–690. https://doi.org/10.1038/nchembio.118

[13] Jinlong Ru; Peng Li; Jinan Wang; Wei Zhou; Bohui Li; Chao Huang; Pidong Li; Zihu Guo; Weiyang Tao; Yinfeng Yang; Xue Xu; Yan Li; Yonghua Wang; Ling Yang. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminformatics. 2014 Apr 16;6(1):13.

[14] Liu, Z., Guo, F., Wang, Y. *et al.* BATMAN-TCM: a Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine. *Sci Rep* **6**, 21146 (2016). https://doi.org/10.1038/srep21146

[15] umper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[16] Mihaly Varadi et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D439–D444, https://doi.org/10.1093/nar/gkab1061

[17] Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio- Temporal Integration" Nafisa M. Hassan, Amr A. Alhossary, Yuguang Mu & Chee-Keong Kwoh. Nature Scientific Reports 7(1) (2017). DOI:10.1038/s41598-017-15571-7

[18] Yan, X., Lu, Y., Li, Z., Wei, Q., Gao, X., Wang, S., Wu, S., & Cui, S. (2022). PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms. Journal of chemical information and modeling, 62(11), 2835–2845. https://doi.org/10.1021/acs.jcim.1c01512

[19] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). [pubmed] [doi]

[20] Kanehisa, M; Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28, 1947-1951 (2019) [pubmed] [doi]

[21] Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. and Ishiguro-Watanabe, M.; KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 51, D587-D592 (2023). [pubmed] [doi]

[22] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25 (1):25-9. DOI: 10.1038/75556

[23] The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May 4;224 (1):iyad031. DOI: 10.1093/genetics/iyad031

[24] Piñero et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020 Jan 8;48(D1):D845-D855. DOI: 10.1093/nar/gkz1021

[25] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

[26] Wu C, MacLeod I, Su AI (2013) BioGPS and MyGene.info: organizing online, gene- centric information. Nucleic Acids Res 41(Database issue):D561-5. doi:10.1093/nar/gks1114

[27] Liu, L., Michowski, W., Kolodziejczyk, A. *et al.* The cell cycle in stem cell proliferation, pluripotency and differentiation. *Nat Cell Biol* **21**, 1060–1067 (2019). https://doi.org/10.1038/s41556-019-0384-4

[28] "G Protein-Coupled Receptor - an overview | ScienceDirect Topics". Progress in Molecular Biology and Translational Science. Retrieved 17 January 2023 – via ScienceDirect. Link to original publication

[29] Yoon, H., Shaw, J. L., Haigis, M. C., & Greka, A. (2021). Lipid metabolism in sickness and in health: Emerging regulators of lipotoxicity. *Molecular cell*, *81*(18), 3708–3730. https://doi.org/10.1016/j.molcel.2021.08.027

[30] Snaebjornsson, M. T., Janaki-Raman, S., & Schulze, A. (2020). Greasing the Wheels of the Cancer Machine: The Role of Lipid Metabolism in Cancer. *Cell metabolism*, *31*(1), 62–76. https://doi.org/10.1016/j.cmet.2019.11.010

[31] DisGeNet official website: https://www.disgenet.org/

[32] allenai/scispacy: A full spaCy pipeline and models for scientific/biomedical documents. (github.com)

[33] Chen, Y. Z., & Zhi, D. G. (2001). Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins, 43(2), 217–226. https://doi.org/10.1002/1097-0134(20010501)43:2<217::aid-prot1032>3.0.co;2-g

[34] An investigation of the false discovery rate and the misinterpretation of p-values" (https://doi.org/10.1098/rsos.14 0216).

Acknowledgement

First and foremost, I would like to express my profound gratitude to Prof. Mu Yuguang and Prof. Su I-Hsin for their invaluable support and guidance throughout this research. I am also indebted to our collaborators who generously provided us with the efficient molecular compounds of Tian Ma.

I extend my sincere thanks to all individuals who have assisted me in various capacities during this journey. Additionally, I am grateful to several organizations that have been instrumental in equipping me with computational skills and knowledge in data science. These include Github, Coursera, OpenAI, DeepLearning.ai, MIT Open Courseware, and Stanford Online. A special mention goes to Prof. Gilbert Strang from MIT for his enlightening courses on linear algebra and 'Learning from Data', Prof Zhang Min from Tsinghua University for her course in Machine Learning Algorithms and ChatGPT, assistance in debug and answering puzzle mathematical questions, which have been pivotal in my academic growth.