UniTCM: A Computational Reverse Pharmacology Framework for Decoding Formula–disease Interactions in Traditional Chinese Medicine

Xiao Zheng[1], Yue Luo[3], Mengyue Hu[1,5], Wenpei Chen[4], Yichen Wang[1], Hong Li[4], Jianyong Zhang[3], Yurong Li[1], Jiahui Zhao[1,2], Yong Chen[1,6], Xuesong Liu[1,6], Su Zeng[1,5], Tengfei Xu[1,5*]

1 School of Pharmacy, Zhejiang University, Zhejiang University, Hangzhou 310058, China

2 Zhejiang Key Laboratory of Blood-Stasis-Toxin Syndrome, School of Basic Medical Sciences, Zhejiang Chinese Medical University, Hangzhou 310053, China

3 School of Pharmacy, Zunyi Medical University, Zunyi 563006, China

4 School of Pharmacy, Henan University, Kaifeng 475004, China

5 Research Center for Clinical Pharmacy, School of Pharmacy, Zhejiang University, Hangzhou 310058, China`

6 Jinhua Institute of Zhejiang University, Jinhua, Zhejiang, 321000, China.

*Corresponding author

E-mail address: tfxu@zju.edu.cn (T. Xu)

Abstract

Traditional Chinese Medicine (TCM) offers rich clinical wisdom for modern drug discovery. Reverse pharmacology provides an effective strategy to decode these complex remedies by moving from clinical efficacy to molecular mechanisms. However, applying this paradigm to multi-component formulas remains difficult due to fragmented research workflows and the lack of integrated platforms. To address this challenge, we developed UniTCM (https://unitcm.qfxulab.com), a unified computational reverse pharmacology framework. UniTCM integrates five functional modules covering terminology standardization, disease-formula mining, AI target prediction, multi-omics verification, and analytical toolkits. Using this framework, we investigated Wutou Tang (WTD) for the treatment of rheumatoid arthritis through combined target profiling, pharmacological assays, and metabolomics. This study successfully decoded the compatibility mechanism of WTD and identified the TNF-α/NF-κB/PTGS2 pathway as its core therapeutic axis. Overall, UniTCM provides a practical bedside-to-bench platform to accelerate TCM-inspired drug discovery.

Introduction

Traditional Chinese Medicine (TCM) is one of the most thoroughly documented and continuously practiced medical systems in human history. Over thousand years of clinical practice, it has generated a vast repository of therapeutic knowledge [1,2], and serves nearly a quarter of the global population in contemporary healthcare. Landmark discoveries such as artemisinin demonstrate that this traditional system remains a vital source for modern drug discovery [3,4]. However, the standard paradigm of modern drug development follows from single molecular targets to clinical indications. This approach is not well suited for complex TCM formulas, which characteristically involve multi-component, multi-target, and synergistic mechanisms. Consequently, this misalignment leaves the chemical basis, molecular mechanisms, and rational optimization of TCM formulas largely as a black box.

Although network pharmacology has advanced computational research in TCM [5,6], existing paradigms still fall short of fully explaining how these complex formulas work. Modern computational methods continue to struggle when translating TCM clinical knowledge into useful biological insights, primarily due to three challenges. First, classical TCM concepts are highly heterogeneous, and English translations for the same TCM medical terminologies vary widely across databases and literature. For example, "Sini Decoction" as a TCM formula with 4 herbs, appears as "Sini Decoction", "Sini Tang", "Frigid Extremities Decoction" or "Decoction for Resuscitation" across authoritative sources. This fragmentation leads to nearly a tenfold difference in PubMed search results (**Supplementary Table 1, 2**). Furthermore, core TCM concepts, such as syndromes (*ZhengHou*) and therapeutic principles, lack systematic mapping to modern molecular mechanisms. This gap hinders the translation of clinical experience into testable molecular hypotheses. Second, the TCM-related knowledge integration from various databases remains limited. This lag prevents researchers from fully leveraging up-to-date information. In addition, current platforms usually focus on isolated entities, such as herbs, active ingredients, targets, or omics data. They lack an end-to-end workflow that connects clinical records, molecular mechanisms, and multi-omics validation. Third, and most important, despite long-standing clinical validation, the mechanisms

underlying TCM remain elusive, which herein presents a key bottleneck to the internationalization and modernization of TCM.

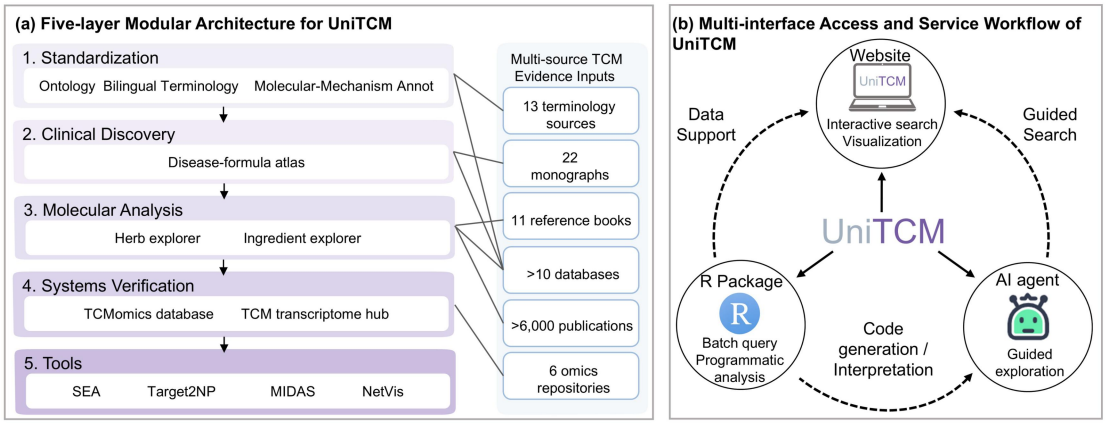
Reverse pharmacology offers a well-suited paradigm to address these bottlenecks [7,8]. Distinct from classical forward pharmacology, which begins with predefined molecular targets, reverse pharmacology starts from clinically or empirically validated therapeutic observations [9]. It works backward to identify active ingredients, molecular targets, and underlying mechanisms, which are then experimentally validated and optimized [10]. This bedside-to-bench approach is uniquely suited for TCM. However, traditional reverse pharmacology relies heavily on isolated case studies and expert intuition, making it difficult to scale up for hundreds of thousands of historical formula records. Although existing platforms such as TCMSP [11], ETCM [12], HERB [13], and SymMap [14] have compiled valuable TCM data, they still show clear limitations in providing an integrated, multi-layered analytical pipeline. Therefore, a platform that can transform unstructured clinical text into a structured, searchable, and mechanistically interpretable framework is urgently needed.

Here, we present UniTCM, a computational platform designed to execute bedside-to-bench reverse pharmacology. The framework overcomes existing analytical barriers through five core modules: (1) standardization to build a bilingual TCM ontology; (2) clinical discovery to mine disease-formula associations; (3) molecular analysis to predict targets and active ingredients using deep learning; (4) systems verification to conduct multi-omics validation; and (5) analytical toolkits to provide network visualization and API integration. This architecture establishes a closed loop connecting clinical phenotypes, molecular hypotheses, and experimental validation. Finally, we applied UniTCM to deconstruct a classical TCM formula, Wutou Tang, demonstrating a complete workflow from mechanistic discovery to prescription evaluation in rheumatoid arthritis treatment.

Results

Standardized Five-Layer Computational Architecture Enables Reproducible and Integrated Analysis of TCM

UniTCM is structured as a five-layer computational ecosystem, where each layer builds directly on the outputs of the previous one (**Fig. 1a**). The standardization layer addresses linguistic and conceptual barriers, establishes a unified TCM ontology, bilingual terminology corpus, and molecular-mechanism lexicon. The clinical discovery layer mines historical medical texts to construct a comprehensive disease-formula atlas. The molecular analysis layer reveals the material basis of TCM formulas by combining herb and ingredient profiling, ADMET evaluation, and target prediction. The systems verification layer provides multi-omics evidence to validate these predictions. Finally, the analytical tools layer provides interactive visualization, programmatic access, and AI-assisted querying. This layered design ensures that each analytical step rests on a standardized foundation. Consequently, research workflows remain fully reproducible and transparent. Moreover, UniTCM supports diverse user workflows by providing three flexible access modes: a web graphical interface, an R package, and an AI agent (**Fig. 1b**). For users seeking a graphical interface, the platform provides an interactive web application alongside an integrated chat assistant. For researchers requiring programmatic access and batch processing, it provides a standardized API and a dedicated R package. All three modes connect seamlessly to the UniTCM backend, enabling smooth querying, analysis, and data retrieval. Notably, UniTCM vastly expands on current databases by incorporating 28,728 herbs, 87,040 active ingredients, and 259,484 formulas (**Fig. 1c**). Beyond its large scale, UniTCM features a rich array of integrated analytical tools over existing platforms (**Fig. 1d**).



(c) Summary Statistics of the UniTCM Knowledgebase

632 Ontology Entities	52,275 Bilingual Terms
> 250,000 Formula Records	3204 ICD-11-standard diseases
28,728 Herbs	87,040 Active Ingredients
1,821 Multi-omics Datasets	1,037 Transcriptomic Studies

(d) Comparison of Core Data Between UniTCM and Other Established TCM Databases

Feature	TCMSP	SymMap	ETCM	BATMAN-TCM	TCMBank	HERB	**UniTCM**
Herbs	499	698	2,079	877	9,191	6,892	**28,728**
Ingredients	13,144**	25,975	38,298	7,862	61,965	44,595	**87,040**
Formulas / prescriptions	Approx. 800	Approx. 800	48,442	394	—	6,743	**259,484**

Figure 1. System Architecture, Access Workflows, and Data Ecosystem of the UniTCM Platform. ***(a)*** *Five-layer modular architecture and multi-source data inputs of UniTCM. The platform comprises five hierarchical layers: standardization; clinical discovery; molecular analysis; systems verification; and tools. This architecture integrates multi-source TCM evidence inputs, including terminology sources, monographs, reference books, databases, literature, and omics repositories.* ***(b)*** *Multi-interface access modes and cooperative service workflows of UniTCM. The platform establishes an interactive data ecosystem accessible via three core interfaces: the Website, the R Package, and AI agent. Dashed arrows illustrate the closed-loop synergy across data support, guided search, and code generation/interpretation.* ***(c)*** *Summary statistics of the UniTCM knowledgebase.* ***(d)*** *Comparison of core data capacities between UniTCM and other established TCM databases.*

Unified Ontology and Evidence-ranked Bilingual Corpus Enable Semantic Interoperability for TCM

Standardization remains a primary bottleneck in TCM modernization due to inconsistent conceptual and linguistic frameworks. To establish a rigorous analytical foundation, this module integrates three complementary features. It builds a hierarchical ontology to structure relationships among TCM entities while incorporating a comprehensive bilingual corpus to resolve translation ambiguities in international literature (**Fig. 2**). In addition, a molecular-mechanism lexicon maps classical clinical concepts directly to modern biological pathways.

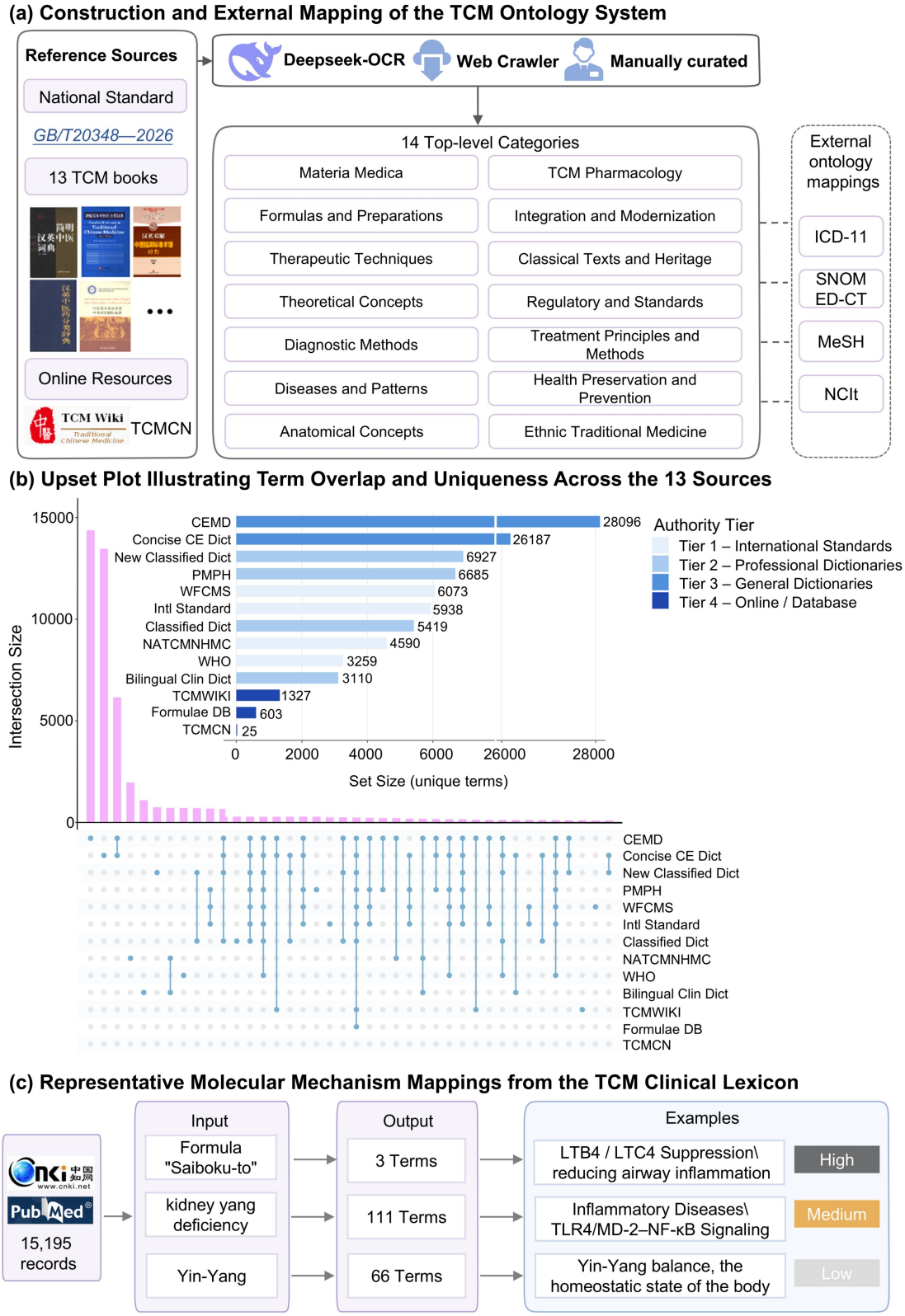


Figure 2. The Standardization Module Establishes the Digital Foundation of UniTCM. (a) Hierarchical tree map of the TCM ontology system showing 14 top-level categories and

representative subcategories. (b) Source composition of the bilingual TCM corpus*. Bar chart showing the number of terms contributed by each of the 13 authoritative sources, stratified by authority four tiers. Upset plot illustrating term overlap and uniqueness across the 13 sources, demonstrating the complementary coverage achieved through multi-source integration.* ***(c) Representative molecular mechanism mappings from the TCM clinical lexicon****. Schematic illustrating how three classical TCM concepts (Formula "Saiboku-to"; "kidney yang deficiency" and "Yin-Yang") are linked to their documented biomedical correlates (suppress the release of inflammatory mediators LTB4 and LTC4 from polymorphonuclear leukocytes, ID: 364; inhibiting the TLR4/MD-2-dependent NF-κB signaling pathway, ID: 103; and systemic metabolic balance, ID: 1), annotated by evidence type and source.*

Hierarchical TCM Ontology System Enables Semantic Interoperability with Mainstream Biomedical Knowledge Networks

To establish a structured semantic framework for TCM entities, we developed a four-level hierarchical ontology system (**Fig. 2a**). The ontology comprises 14 top-level categories spanning core domain concepts such as formulas, syndromes, meridians, and therapeutic principles, expanding into 487 subcategories and 632 standardized entities. Each entity is assigned a unique UniTCM Ontology Identifier (UOI) and cross-mapped to 29 external ontology resources, including ICD-11, SNOMED-CT, MeSH, and NCIt [15]. This comprehensive mapping bridges classical TCM terminology with mainstream biomedical knowledge systems.

High-density Bilingual Corpus Resolves Translation Inconsistencies Across TCM Literature

To resolve translation inconsistencies across multilingual TCM literature, we compiled a comprehensive Chinese-English terminology corpus (**Fig. 2b, Supplementary Table 3**). The corpus systematically consolidates terminological standards from 13 authoritative sources,

including international organizations, national pharmacopoeias, and domain-specific dictionaries. It integrates 52,275 standardized terms structured into four evidence tiers, making it nearly five times larger than the WHO International Standard Terminologies (2007) [16,17]. A key feature of this corpus is its evidence-based priority ranking system. When multiple English translations exist for a single Chinese term, candidate terms are ranked based on source authority, publication recency, and literature usage frequency. By prioritizing international standards over national pharmacopoeias, textbooks, and dictionaries, the system helps researchers select the most authoritative translation while retaining alternative terms for broader literature searches.

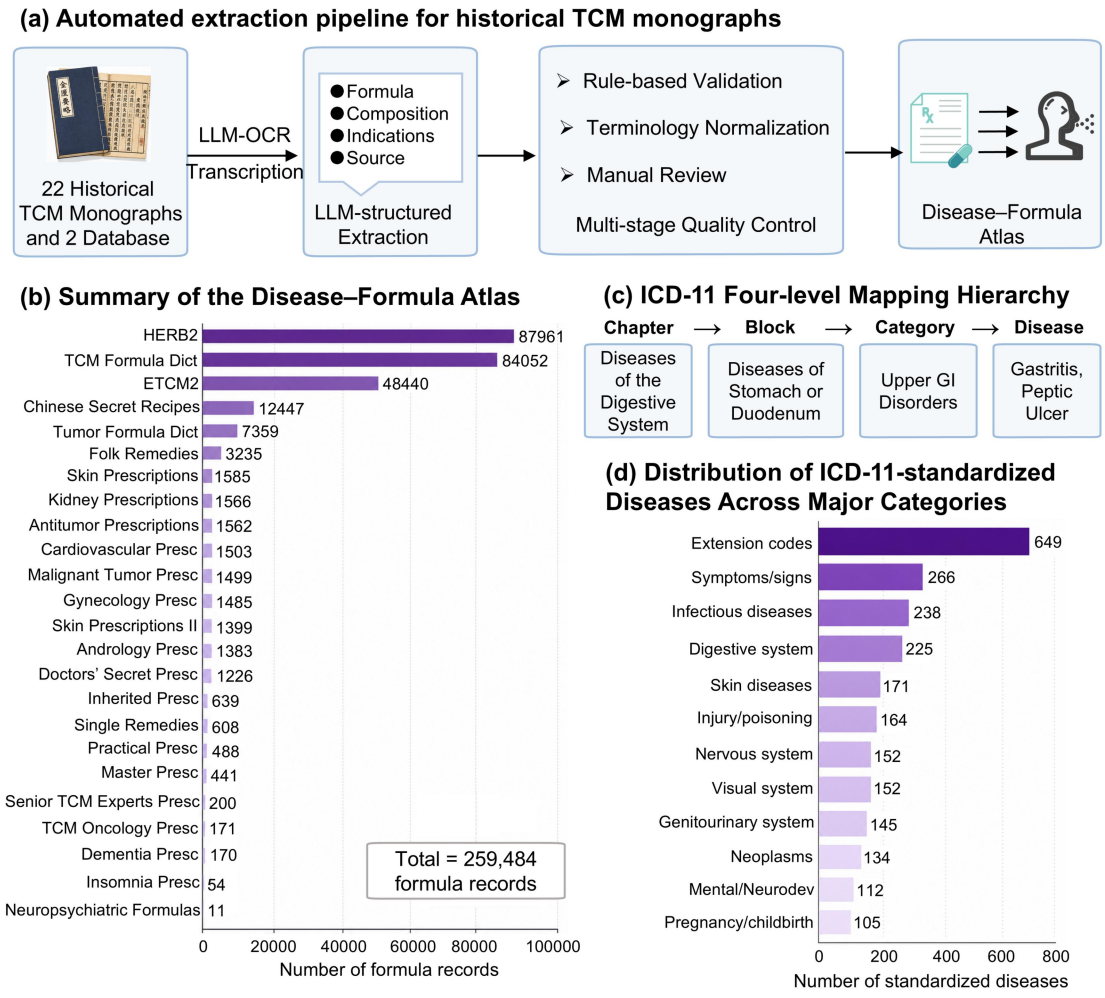
Molecular-mechanism Lexicon Bridges Classical TCM Concepts with Modern Biomedical Biology

Classical TCM terminology, such as "kidney yang deficiency", "liver qi stagnation", and "blood stasis", encodes centuries of clinical observations that effectively capture holistic, systemic pathological states. However, the lack of direct molecular alignment with modern biomedicine has long obscured the biological substrates of these concepts and hindered cross-disciplinary translation [18]. To address this challenge, we constructed a molecular-mechanism lexicon that systematically connects TCM clinical terms to their experimentally validated biological foundations (**Fig. 2c**). By mining 12,090 peer-reviewed articles from PubMed and CNKI, we established 15,195 structured knowledge records mapping TCM concepts to their underlying genes, proteins, pathways, and experimental evidence. For instance, "kidney *Yang* deficiency" is explicitly mapped to hypothalamic-pituitary-adrenal axis dysregulation, mitochondrial dysfunction, and the downregulation of steroidogenic enzymes (**Fig. 2c**). Each entry is rigorously annotated with study design, sample size, and publication source. By transforming qualitative clinical phenomena into structured, mechanistically interpretable networks, this lexicon provides a computable foundation for systematic target discovery and hypothesis-driven research in TCM modernization.

Deep-curated Disease–formula Atlas Bridges Historical TCM Practice with Modern Disease Classifications

TCM formulas represent the culmination of millennia of clinical practice, encoding complex therapeutic synergy and holistic treatment philosophies. However, extracting actionable insights from these historical remedies is severely constrained by severe data heterogeneity, archaic terminology, and historical measurement shifts. To unlock this rich clinical archive, we constructed a standardized disease-formula atlas through an integrated computational pipeline (**Fig. 3a**). This pipeline sequentially executes LLM-assisted historical text extraction, ICD-11 disease mapping, and dynasty-specific dosage normalization.

The historical text extraction workflow combines Optical Character Recognition (OCR) tailored for ancient scripts, LLM-based field parsing, and rigorous domain-expert curation. This strategy successfully unified 259,484 classical formulas mapped to 2,024 standardized disease entities across 20 primary sources, including the Great Dictionary of TCM Prescriptions and ETCM2 (**Fig. 3b, Supplementary Table 4**). By balancing computational scalability with expert validation, this framework enables the systematic digitization of rare historical monographs previously inaccessible to computational analysis.



(e) Historical Dosage Normalization for the Formula Guizhi Mahuang Geban Tang

Guizhi Mahuang Geban Tang
(桂枝麻黄各半汤)

Historical dosages across dynasties

Herb	Han: Shang Han Lun	Song: Sanyin Jiyi Bingzheng Fanglun	Qing:Shanghan Laisu Ji
Guizhi(桂枝)	1 *liang* 16 *zhu*	4 *qian* 1 *zi*	Guizhi decoction 3 pieces + Mahuang decoction 3 pieces
Shaoyao(芍药)	1 *liang*	0.5 *liang*	
Shengjiang(生姜	1 *liang*	Decocted with ginger and jujube	
Gancao(甘草)	1 *liang*	0.5 *liang*	
Mahuang(麻黄)	1 *liang*	0.5 *liang*	
Dazao(大枣)	4 pieces	Decocted with ginger and jujube	
Xingren(杏仁)	24 pieces	12 pieces	

UniTCM dose-unit normaliz-ation →

Herb	Original dose	Ratio	Clinical reference (g)	Dynasty	Note
Guizhi	1 *liang* 16 *zhu*	1.5	7.67	Qin/Han	Bark removed
Shaoyao	1 *liang*	1	4.6	Qin/Han	—
Shengjiang	1 *liang*	1	4.6	Qin/Han	Sliced
Gancao	1 *liang*	1	4.6	Qin/Han	Processed / honey-fried
Mahuang	1 *liang*	1	4.6	Qin/Han	Nodes removed
Dazao	4 pieces	—	4	Qin/Han	Split; counted by pieces, cannot be directly converted to grams
Xingren	24 pieces	—	2.4	Qin/Han	Soaked in hot water; cannot be directly converted to grams

Figure 3. The Clinical Discovery Module of the Disease–Formula Atlas. (a) Automated extraction pipeline for historical TCM monographs. *Schematic illustrating the workflow*

from source texts (22 historical TCM monographs and 2 databases) through OCR transcription, LLM-structured extraction, and multi-stage quality control, leading to the construction of the Disease–Formula Atlas. ***(b) Summary of the Disease–Formula Atlas.*** *Horizontal bar chart showing the number of formula records contributed by each major source, with a total of 259,484 formula records.* ***(c) ICD-11 Four-level Mapping Hierarchy.*** *Schematic representation of the four-level ICD-11-standardized disease classification mapping (Chapter -Block-Category-Diseases) using a representative path.* ***(d) Distribution of ICD-11-standardized diseases across major categories.*** *Horizontal bar chart illustrating the number of standardized diseases across major ICD-11 classification categories.* ***(e) Historical dosage normalization for the example formula "Guizhi Mahuang Geban Tang".*** *Workflow of the UniTCM dose-unit normalization, comparing historical dosages across multiple dynasties (Han, Song, and Qing) with their converted modern gram equivalents and clinical reference values (ID: 61184).*

To align classical TCM diagnoses with modern medical taxonomies, we implemented a semantic similarity algorithm grounded in ICD-11 (**Fig. 3c**). Using a biomedical language model fine-tuned on TCM corpora, the algorithm generates vector embeddings to rank candidates via cosine similarity against ICD-11 disease codes [19]. This process mapped the entire formula collection to 2,024 unique ICD-11 disease entities with a median mapping confidence of 0.89, covering 26 ICD-11 chapters with primary concentration in digestive, infectious, and ophthalmic disorders (**Fig. 3d**). This mapping bridges historical diagnostic concepts with global clinical taxonomies, facilitating cross-referencing with modern epidemiological data.

Furthermore, to enable precise quantitative pharmacological comparisons across two millennia of TCM practice, we established a dual-level dosage standardization pipeline (**Fig. 3e, Supplementary Table 5**). First, we calculated the relative percentage composition for each formula based on individual herb weights. Because weight ratios are invariant to historical unit changes, these normalized proportional profiles offer high-confidence,

cross-era comparability directly from raw classical texts. Second, to reconstruct absolute modern metric dosages, we curated dynasty-specific conversion factors based on source text dating to address dramatic historical variations in Chinese metrology, such as one *liang* shifting from 13.8 g in the Han Dynasty to 37.3 g in the Ming Dynasty [20]. By transforming fragmented historical prescriptions into a semantically aligned and mathematically normalized knowledge graph, this atlas provides a robust quantitative foundation that bridges classical empirical wisdom with modern systems biology.

Expanded Repository of TCM Ingredients and ADMET Profiles Connects Empirical Herbal Medicine with Modern Pharmacology

Deciphering the chemical material basis and key molecular targets of TCM formulas represents a crucial bridge connecting empirical herbal medicine with systems pharmacology. To achieve this, the molecular analysis module integrates a massive herb-ingredient repository, an ADMET profiling engine, and a dual-strategy target prediction framework.

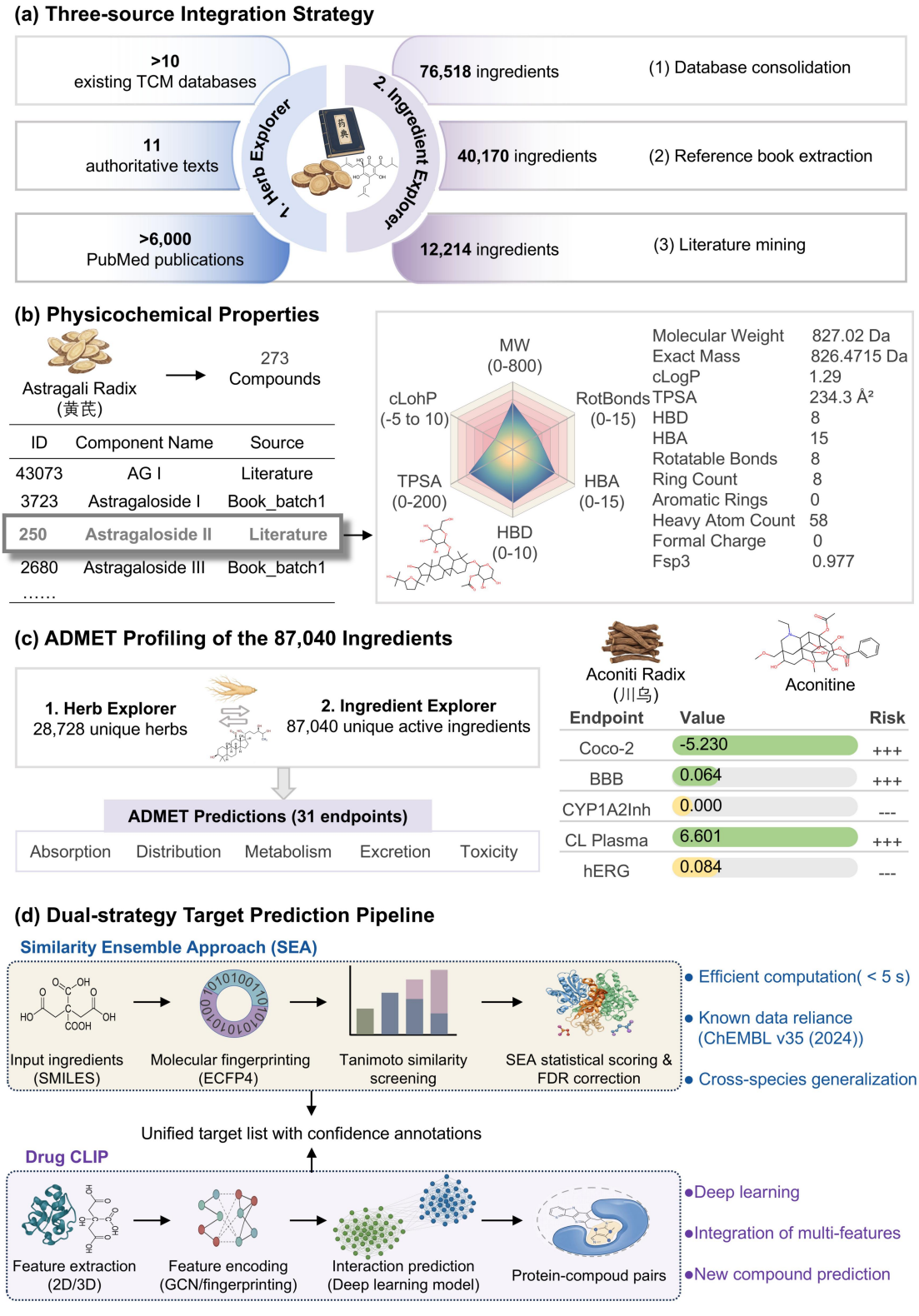


Figure 4. Molecular Analysis Module for Mapping Herbs to Their Targets. (a) Three-source integration strategy for the herb explorer and ingredient explorer databases.

Data are integrated from literature mining, reference book extraction, and database consolidation, yielding 28,728 herbs and 87,040 active ingredients after InChIKey/SMILES-based deduplication. ***(b) Calculation of basic physicochemical properties and the hierarchical matching strategy.*** *Left: Extraction of 273 compounds from a representative herb (Astragali Radix) and their profiling. Right: Radar chart and metrics showcasing the physicochemical properties calculated via RDKit (e.g., MW, cLogP, TPSA, HBD, HBA). The textual flowchart outlines the three-stage hierarchical matching strategy: InChIKey exact matching for structurally resolved compounds, canonical SMILES matching after RDKit standardization, and normalized chemical name matching for unresolved compounds.* ***(c) ADMET profiling of the 87,040 ingredients.*** *ADMET properties are predicted for all ingredients in Ingredient Explorer across 31 key endpoints (e.g., Caco-2 permeability, blood–brain barrier penetration, cytochrome P450 inhibition, clearance, and hERG inhibition). The right panel shows a representative risk profile for aconitine from Aconiti Radix.* ***(d) Dual-strategy target prediction pipeline.*** *Schematic illustrating the Similarity Ensemble Approach (SEA) (top; structural similarity and statistical enrichment against ChEMBL v35) and the DrugCLIP approach (bottom; contrastive learning and embedding similarity against PDB binding pockets), both converging on a unified target list with confidence annotations. Key characteristics of each method (e.g., computational efficiency and generalization ability) are highlighted on the right.*

Multi-source Data Integration Expands the TCM Herb Repository Beyond Current Benchmarks

Elucidating how TCM formulas function requires a comprehensive and rigorously annotated profile of their chemical constituents. However, existing databases frequently suffer from outdated records, unverified cross-citations, and sparse structural data [21]. To overcome these limitations, we engineered an herb-ingredient explorer by aggregating data across three complementary pipelines (**Fig. 4a**). First, by extracting structured information from 9,000 full-text PubMed articles using LLM-assisted mining with manual validation, we cataloged

12,214 ingredients across 4,005 herbs. Second, we deployed OCR and LLM tools to process 11 authoritative medical reference texts, including the Chinese Pharmacopoeia (2020 edition) and the Grand Dictionary of Chinese Materia Medica, recovering 40,170 ingredients from 16,718 herbs. Third, we harmonized data from over 10 public databases, including TCMSP [11], ETCM [12], HERB [13], and BATMAN-TCM [22]. After standardizing chemical structures and eliminating duplicates via InChIKey matching, the finalized database hosts 87,040 unique active ingredients mapped to 28,728 unique herbs. Each entry is enriched with taxonomic classifications, traditional properties, clinical indications, and external cross-references [23]. This repository expands herb coverage nearly 3.9-fold compared to HERB (7,263) and 57-fold compared to TCMSP (499).

Large-scale Chemical Integration and ADMET Profiling of Active Ingredients

We next evaluated the physicochemical and pharmacokinetic properties of these active ingredients. Each ingredient includes key physicochemical parameters from PubChem, including molecular weight, LogP, topological polar surface area, rotatable bonds, and hydrogen bond donors or acceptors (**Fig. 4b**). Furthermore, we integrated ADMETlab 3.0 [24] to predict 31 key pharmacokinetic and toxicological endpoints across absorption, distribution, metabolism, excretion, and toxicity. These endpoints cover critical clinical parameters such as Caco-2 permeability, blood-brain barrier penetration, CYP450 interactions, metabolic half-life, and hERG cardiotoxicity. This high-throughput profiling offers an immediate screen for drug-likeness and safety, allowing researchers to prioritize viable lead compounds early in the discovery pipeline (**Fig. 4c**).

Dual-strategy Prediction Pipeline Combining SEA and Deep Learning Reveals Ingredient-target Networks

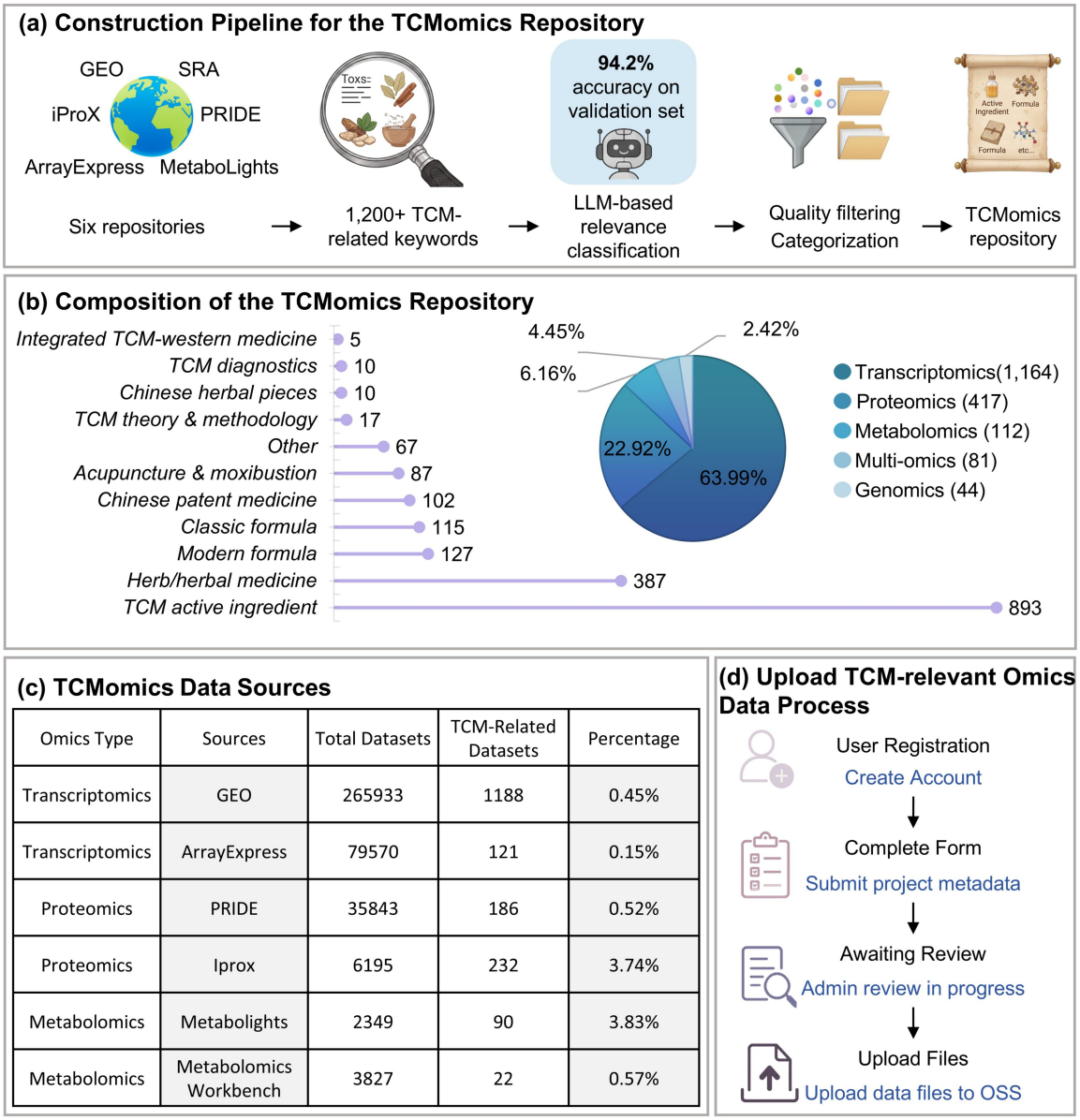
To systematically map ingredient-target interactions, we developed a complementary prediction pipeline uniting ligand-based similarity with deep-learning pocket representation (**Fig. 4d**). First, we calibrated the Similarity Ensemble Approach (SEA) against ChEMBL v35 human bioactivity data (pChEMBL $\geq$ 5, confidence $\geq$ 7), generating roughly 2 million

reference pairs. Using 2048-bit ECFP4 fingerprints, SEA identified potential targets based on statistically significant Tanimoto similarity scores. Second, we applied DrugCLIP, a contrastive learning model mapping small molecules and protein binding pockets into a shared latent space. We generated embeddings for 87,040 ingredients and roughly 25,000 PDB binding pockets, ranking targets by cosine similarity and retaining the top 100 per ingredient after bioactivity calibration. By pairing SEA with DrugCLIP, the pipeline achieves both reliability and discovery breadth. While SEA excels at detecting well-studied target families grounded in established chemical similarity, DrugCLIP captures novel compound-target pairs through learned structural representations.

In summary, by unifying a massive chemical repository, high-throughput ADMET profiling, and a dual-engine target prediction pipeline, this module transforms raw herbal data into an actionable chemical space. It enables systematic target mining and safety screening, providing a robust computational foundation for modern TCM drug discovery.

Curated Multi-omics Repository and Analytical Hub for Experimental Verification of TCM Mechanisms

Although omics studies in TCM have expanded rapidly, these datasets remain scattered across general-purpose repositories such as GEO [25], ArrayExpress [26], PRIDE [27], MetaboLights [28], and iProX [29]. Because these platforms are not tailored for TCM research, locating relevant datasets and performing cross-study meta-analyses remains difficult due to inconsistent metadata annotations. To address these limitations, the systems verification module of UniTCM integrates two complementary components: the TCMomics repository and the TCM transcriptome hub (**Fig. 5**). Together, they provide literature-derived experimental omics evidence to validate computational predictions and help elucidate the modern molecular mechanisms of TCM [30].



Omics Type	Sources	Total Datasets	TCM-Related Datasets	Percentage
Transcriptomics	GEO	265933	1188	0.45%
Transcriptomics	ArrayExpress	79570	121	0.15%
Proteomics	PRIDE	35843	186	0.52%
Proteomics	Iprox	6195	232	3.74%
Metabolomics	Metabolights	2349	90	3.83%
Metabolomics	Metabolomics Workbench	3827	22	0.57%

Figure 5. Architecture, Data Distribution, and Submission Workflow of the TCMomics Module. (a) Construction pipeline for the TCMomics repository. *Systematic searches across public repositories using 1,200+ TCM-related keywords, followed by LLM-based relevance classification, quality filtering, and categorization into 11 TCM intervention types.* ***(b) Composition of the TCMomics repository by intervention and omics type. (c) TCMomics data sources.*** *A detailed summary table mapping omics types (transcriptomics, proteomics, metabolomics) to their source repositories (GEO, ArrayExpress, PRIDE, iProX, MetaboLights, and Metabolomics Workbench), listing total public datasets, identified*

TCM-related datasets, and their respective percentages. ***(d) Upload TCM-relevant omics data process.*** *The four-step user submission workflow, spanning user registration (Create Account), complete form (Submit project metadata), awaiting review (Admin review in progress), and upload files.*

To build the repository, we systematically searched six major databases using over 1,200 Chinese and English terms covering herbs, formulas, and active ingredients. Following LLM-assisted relevance classification and manual curation, rigorous quality filtering removed records with incomplete metadata, missing raw files, or ambiguous designs (**Fig. 5a**). The resulting TCMomics repository contains 1,821 curated datasets, encompassing 1,352 transcriptomic (71.9%), 417 proteomic (22.2%), and 112 metabolomic (5.9%) studies (**Fig. 5b, 5c, Supplementary Table 6**). These datasets span ten intervention categories, primarily active ingredients (45.5%), herbs/extracts (26.4%), and formulas (15.4%). Each entry features standardized metadata detailing the entity name, intervention type, organism, disease model, tissue type, and original accession number. Additionally, TCMomics invites community-driven data deposition through a user-friendly submission pipeline, ensuring the platform evolves alongside the expanding TCM omics field (**Fig. 5d**).

To maximize data utility and lower computational barriers, we developed the TCM Transcriptome Hub, which provides standardized, pre-computed analysis pipelines across ingested studies (**Supplementary Fig. 1a**). Each raw dataset undergoes automated quality control, normalization, differential expression analysis, GO/KEGG/GSEA enrichment, PPI network reconstruction, immune cell infiltration profiling, and transcription factor activity inference (**Supplementary Table 7**). After processing multi-comparison studies into distinct analytical units, the Hub incorporates 1,037 studies covering 54 organisms and 451 TCM entities (**Supplementary Fig. 1b**). To make these findings accessible without requiring custom coding, outputs are interactively visualized through volcano plots, enrichment bubble charts, network graphs, and PCA plots. Furthermore, pre-computed transcriptomic signatures and therapeutic connectivity maps allow researchers to rapidly compare functional response

patterns across diverse interventions (**Supplementary Fig. 1c**).

Comprehensive Analytical Toolset and Multimodal Interfaces

The analytical tools layer extends UniTCM beyond its core data repositories by providing a suite of analytical services. PredictTarget-SEA enables target prediction for novel compounds by querying SMILES structures against ChEMBL v35, which contains about 2 million compounds and 5,000 human targets. Using the SEA algorithm, it generates ranked target lists with z-scores, p-values, and known active ligands (**Supplementary Fig. 2a**). To address the lack of confidence scoring in existing databases, Target2NP provides an evidence-graded resource for natural product-target interactions. It assigns evidence levels from 1 to 5 based on experimental stringency, enabling threshold-based filtering (**Supplementary Fig. 2b**). MIDAS evaluates gene-disease relationships by integrating about 5.75 million records across 11 databases into aggregated cross-database scores (**Supplementary Fig. 2c**).

For network pharmacology analysis, NetVis generates interactive graphs connecting herbs, ingredients, targets, and pathways, supporting node filtering, confidence-weighted edges, subnetwork extraction, and high-resolution figure exports (**Supplementary Fig. 2d**). Programmatic access is facilitated through a RESTful API and the unitcm R package, enabling batch queries, custom filtering, and bulk data downloads for reproducible bioinformatics workflows (**Supplementary Fig. 3a**). Finally, the UniTCM chatbot offers an AI-assisted natural language interface. It searches, cross-references, and synthesizes results across modules to streamline research workflows (**Supplementary Fig. 3b**).

Comparative Analysis with Existing TCM Platforms

To highlight the unique position of UniTCM, we systematically benchmarked it against six representative TCM databases across 12 key functional dimensions (**Supplementary Table 8**). While existing platforms primarily serve as static data repositories, UniTCM advances TCM modernization by establishing an end-to-end, closed-loop reverse pharmacology framework

within a single platform. Specifically, it integrates a dual-strategy target prediction engine with pre-computed multi-omics validation pipelines, directly bridging the long-standing divide between computational hypothesis generation and experimental proof. Beyond analytical integration, UniTCM empowers a versatile discovery ecosystem through a bilingual interface, a dedicated R package (unitcm), a RESTful API, and an AI-driven conversational agent. By unifying high-precision molecular inference, standardized omics verification, and seamless programmatic access, UniTCM transforms traditional medicine research from fragmented retrieval into an executable, reproducible, and translationally relevant discovery pipeline.

Deciphering the Combination Rationale of Wutou Tang in Treating Rheumatoid Arthritis via UniTCM

Rheumatoid arthritis (RA) is a chronic autoimmune disease whose current clinical management remains challenged by drug resistance, severe side effects, and heterogeneous patient responses [31–34]. Although Wutou Tang (WTD) exhibits marked efficacy in treating RA [35], computationally evaluating such traditional formulas is often hampered by non-standard terminology, unnormalized historical dosages, and disconnected target-prediction tools. To address these challenges, we developed a closed-loop reverse pharmacology framework within the UniTCM platform (**Fig. 6**), enabling automated formula decomposition and quantitative network analysis of WTD.

First, using the UniTCM bilingual corpus, we standardized WTD nomenclature across 13 TCM corpora and extracted dynasty-normalized dosages for its five component herbs (**Fig. 6a**). Subsequent ADMET filtering prioritized 1,203 bioavailable compounds from an initial pool of 2,137 constituents (**Fig. 6b**). Target prediction via the Target2NP module identified 501 unique WTD targets, 77.2% of which overlapped with RA pathogenic genes in the MIDAS database (**Supplementary Table 9, 10**). Mapping these herb-disease overlapping targets (**Fig. 6c**) and constructing a compound-target-disease PPI network (**Fig. 6d**)

highlighted key inflammatory hubs, including AKT1, TNF, and PTGS2. Pathway enrichment analysis further delineated distinct functional roles across the Jun-Chen-Zuo-Shi hierarchy (**Fig. 6e**): the Jun herb primarily regulated neuroactive ligand-receptor and cAMP signaling; the Chen herb dominated calcium and cGMP-PKG pathways; Zuo herbs modulated Th17 and TLR immune axes; and the Shi herb targeted TNF and IL-17 cascades. Crucially, negative network separation metrics (S_{AB}) across all four herb categories confirmed their synergistic convergence on the core RA network, with the Shi herb occupying the highest proportion of core disease targets as a network harmonizer (**Supplementary Table 11**). This entire analytical pipeline is fully executable via the companion unitcm R package.

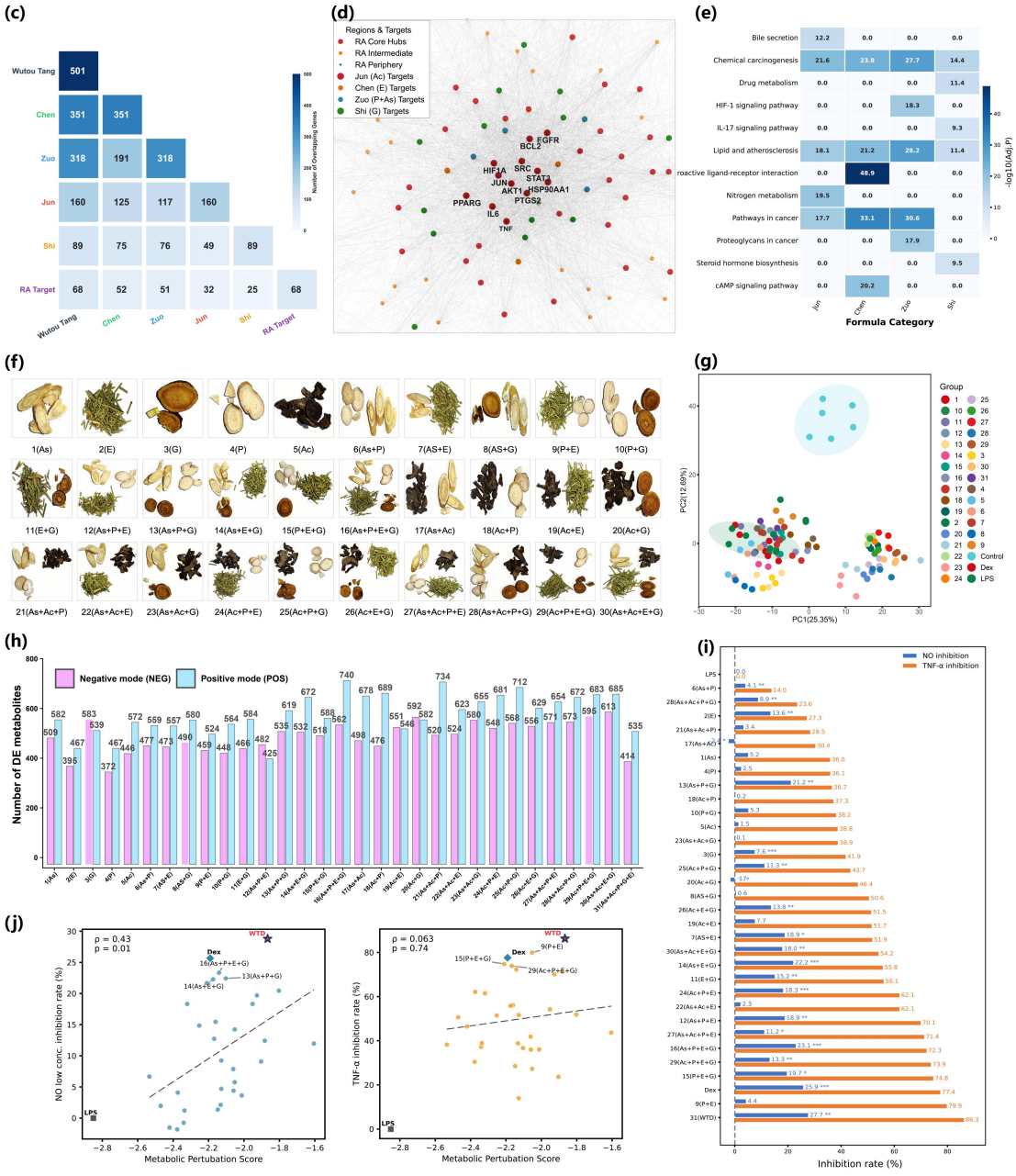
To experimentally validate these computational predictions through reverse pharmacology, we synthesized all 31 theoretical sub-formula combinations ($\sum_{k=1}^{5}\binom{5}{k} = 31$) of WTD (**Fig. 6f)** and conducted untargeted metabolomics profiling. LC-MS datasets were processed to resolve PCA trajectory separation (**Fig. 6g**) and differential metabolite profiles (**Fig. 6h, Supplementary Table 12**). Functional assays on LPS-stimulated RAW264.7 macrophages identified *Ephedrae Herba* as the primary anti-inflammatory driver, which exhibited potent synergy when combined with *Paeoniae Radix Alba* (**Fig. 6i**). To quantify global metabolic responses, we defined the Pathway Perturbation Score (PPS) (See **Methods**) based on altered metabolic pathways (**Supplementary Table 13**). Spearman correlation analysis between PPS and nitric oxide (NO) inhibition revealed that WTD alleviates systemic inflammation primarily by restoring metabolic homeostasis (**Fig. 6j, Supplementary Table 14**). Literature evidence retrieved from the platform "Terms Molecular Mechanisms" database substantially corroborated these predicted targets, pathway cascades, and metabolic shifts (**Supplementary Table 15**). Together, this integrated computational and experimental workflow successfully decodes the multi-target, multi-pathway compatibility mechanisms of WTD, establishing a generalizable paradigm for modernizing traditional complex formulas through reverse pharmacology.

(a)

Chinese	English	Sources	Authority
乌头汤	**Wutou Tang**	International Standard Chinese-English Basic Nomenclature of Traditional Chinese Medicine	Tier 1
	Wutou Tang:Aconite Main Root Decoction	World Federation of Chinese Medicine Associations (WFCMS)	Tier 1
	Aconite Main Tuber Decoction	Chinese-English-Medical-Dictionary	Tier 3

(b)

Herb_cn	Herb_en	Category (role)	Original dose	Dynasty	Ratio	Modern equivalent(g)	Raw Ingredients	After ADMET
ZhiChuanWu (制川乌)	Aconiti Radix Cocta	Jun/Sovereign	5 pieces	Han	2	6	335	206
MaHuang (麻黄)	Ephedrae Herba	Chen/Minister	3 *liang*	Han	3	9	744	410
HuangQi (黄芪)	Astragali Radix	Zuo/Assistant	3 *liang*	Han	3	9	484	523
ShengBaiShao (生白芍)	Paeoniae Radix Alba	Zuo/Assistant	3 *liang*	Han	3	9	388	
ZhiGanCao (炙甘草)	Glycyrrhizae Radix Praeparata	Shi/Courier	3 *liang*	Han	3	9	186	64



444

445 ***Figure 6. Deciphering the Anti-Inflammatory Mechanisms and Compatibility Principles of***

446 ***Wutou Tang Against Rheumatoid arthritis. (a) Comparison of English translations for***

447 ***Wutou Tang.*** *Wutou Tang recommended as the standardized nomenclature based on authority*

level and frequency. ***(b) Herb dosages of Wutou Tang.*** *Retrieved from the disease-formula atlas database (IDs: 16872, 104080) and the number of active ingredients per herb derived from UniTCM.* ***(c) Target overlap matrix among the five constituent herbs of Wutou Tang and rheumatoid arthritis (RA). (d) PPI network overlap between Wutou Tang herbs and RA targets. (e) KEGG pathway enrichment analysis of Wutou Tang targets. (f) Schematic of the 31 disassembled drug combinations derived from Wutou Tang. (g) PCA plot of untargeted LC-MS/MS metabolomics profiles across the 31 disassembled combinations (negative mode). (h) Differential metabolites identified among the 31 disassembled combinations. (i) Inhibitory effects of different combinations on LPS-induced NO and TNF-α production in macrophages. (j) Correlation between metabolic perturbation score and anti-inflammatory phenotypes.*** *As, Astragali Radix; E, Ephedrae Herba; G, Glycyrrhizae Radix; P, Paeoniae Radix Alba; Ac, Aconiti Radix.*

Discussion

The persistent decline in drug discovery productivity, which often termed Eroom's Law [36], has driven a clear shift from single-target therapies toward multi-target combination strategies [37–39]. While clinical experience with complex herbal formulations in TCM offers a uniquely validated foundation for multi-target therapeutics, modernizing these empirical remedies remains difficult. The main challenge lies in bridging the gap between holistic, multi-component herbal systems and modern molecular pharmacology [37,40]. Reverse pharmacology provides a compelling bedside-to-bench roadmap to address this challenge, yet its application has been constrained by data heterogeneity, unstructured historical literature, and fragmented prediction tools.

To overcome these barriers, we developed UniTCM, an integrated computational and experimental ecosystem designed to operationalize reverse pharmacology for TCM modernization. UniTCM systematically addresses key computational bottlenecks across multiple levels. At the terminological level, UniTCM resolves semantic heterogeneity across disparate historical and contemporary databases by establishing a unified TCM ontology and bilingual corpus. Its evidence-based ranking algorithm enables standardized, prioritized translation of complex traditional terminology. At the data and historical analysis level, UniTCM hosts a large-scale disease–formula atlas that exceeds existing databases in coverage by more than an order of magnitude. By digitizing rare historical monographs and implementing a dynasty-specific dosage normalization algorithm, UniTCM converts ancient measurement units into standardized metric doses, making it possible to perform quantitative pharmacological analyses across two millennia of clinical prescribing patterns. On the molecular analysis side, UniTCM integrates DrugCLIP and SEA within a dual-strategy target prediction pipeline. This approach substantially elevates target prediction precision over conventional methods, successfully transforming empirical clinical records into testable, mechanistically transparent molecular hypotheses.

We demonstrated the utility of UniTCM by deciphering the combination rationale of WTD in treating RA. By seamlessly coupling computational predictions with untargeted metabolomics

and macrophage assays, UniTCM revealed that WTD exerts its therapeutic effects through convergent modulation of core disease hubs, such as AKT1 and TNF, rather than isolated pathways. Notably, our platform-guided experimental workflow identified *Ephedrae Herba* as the primary anti-inflammatory driver, which acts synergistically with *Paeoniae Radix Alba* to maximize TNF-α suppression. Furthermore, the strong correlation observed between metabolic homeostasis restoration and immune suppression underscores how WTD achieves systemic efficacy through holistic metabolic and immune balance.

In conclusion, UniTCM offers a standardized, open-access, and end-to-end computational ecosystem for translating centuries of empirical clinical wisdom into actionable biological mechanisms. By bridging classical TCM principles with modern systems pharmacology, UniTCM establishes a transferable workflow for deciphering complex herbal formulas, accelerating evidence-based natural product discovery, and fostering the global integration of traditional and precision medicine.

Methods

TCM Ontology System Construction

We developed the TCM ontology system through an iterative process integrating expert consultation with a systematic literature review [15]. A panel of five domain experts, including three clinical practitioners, one pharmacognosist and one biomedical informatician defined the top-level categorical framework. The panel refined category boundaries through three rounds of delphi-style consensus building, resolving 23 contested categorization decisions using majority voting with documented rationale. Subcategories were constructed by extracting classification hierarchies from authoritative reference sources, yielding 487 subcategories and 632 total entities organized across four hierarchical levels. Each entity was assigned a unique UOI. We mapped these entities to 29 terms in external databases, including UMLS, ICD-11, GO, MeSH, and the NCBI Taxonomy, combining lexical matching with expert curation. The ontology is stored in web OWL format and is accessible through REST APIs.

Bilingual TCM Corpus Compilation

Terminology data were extracted from 13 authoritative sources organized into three hierarchy tiers. Tier 1 includes international standards, such as the WHO International Standard Terminologies on Traditional Medicine [41] and the World Federation of Chinese Medicine Societies (WFCMS) International Standard Chinese-English Basic Nomenclature of Chinese Medicine. Tier 2 covers national pharmacopoeias and government standards, including the Chinese Pharmacopoeia 2020 edition and the National Standard of the People's Republic of China, Basic Terminology of Traditional Chinese Medicine (GB/T 20348-2006). Tier 3 comprises academic reference works, including nine domain-specific dictionaries and terminology compilations covering herbal medicine, formulas, acupuncture, and clinical terminology. The raw extraction yielded 99,934 Chinese-English term pairs. Deduplication was performed in three sequential steps. First, we applied exact string matching following

unicode normalization and whitespace standardization. Second, we unified simplified and traditional Chinese characters using the OpenCC library [42]. Third, we conducted synonym clustering based on the Jaccard similarity of character n-grams with a threshold of 0.85, followed by manual adjudication of ambiguous clusters. This pipeline ultimately produced 52,275 unique standardized term entries.

Molecular Mechanisms Lexicon Construction

We searched PubMed and CNKI using structured queries combining TCM clinical terms (*e.g.*, "kidney yang deficiency") with molecular biology terms (*e.g.*, "mechanism","pathway", "biomarker", "molecular basis"). Searches were conducted for 856 unique TCM clinical concepts extracted from the ontology system. Retrieved articles (n = 12,090) were processed using an LLM-assisted extraction pipeline to identify molecular mechanisms, associated genes/proteins, implicated pathways, and experimental evidence types. Each extracted record was validated by domain experts, yielding 15,195 curated mechanism–concept associations [18].

Disease–Formula Atlas Construction

Physical monograph digitization. We processed 22 authoritative monographs through a three-stage pipeline. In the first stage, scanned PDF pages were processed using DeepSeek-OCR [43], a specialized model optimized for dense classical Chinese text featuring mixed font styles. The engine achieved an accuracy exceeding 97% at the character level on a manually verified benchmark of 500 pages. In the second stage, addressing structured extraction, the Gemini large language model (LLM) [44] was prompted with task-specific instructions to extract formula names, herbal compositions, dosages, indications, contraindications, source texts, and supplementary annotations from OCR output. Extraction templates were customized for each monograph to accommodate varying formatting conventions. In the third stage, dedicated to quality control, extracted records underwent three

559 sequential validation rounds. First, we conducted automated consistency checks, such as
560 verifying that dosage fields contained numeric values and that herb names matched entries in
561 the herb explorer. Second, a TCM pharmacologist performed a random-sample expert review
562 of 10% of records from each monograph, yielding a low error rate of 2%. Third, we
563 performed cross-reference validation against original scanned images for any flagged
564 discrepancies.

565 **Digital database integration.** Formula records were extracted from ETCM [12], comprising
566 1,124 entries, and HERB [13], comprising 7,301 entries, through their publicly accessible
567 data download interfaces. Additional formulas were collected from other online TCM formula
568 databases. All records were standardized using a unified schema including formula name,
569 herb composition, indications, dosages, source references, and dynasty of origin.

570

571 **Disease name standardization.** Traditional TCM disease names were mapped to ICD-11
572 codes [19] using a custom semantic similarity algorithm (**Fig. 3d**). For each traditional
573 disease name, we first generated a vector embedding using a biomedical language model
574 fine-tuned on TCM texts. Next, we computed cosine similarity against all ICD-11 term
575 embeddings. Finally, we extracted the top five candidate mappings for expert adjudication.
576 Mappings with a cosine similarity greater than 0.90 were accepted automatically, whereas
577 those with values between 0.70 and 0.90 underwent expert review. Mappings below 0.70 were
578 manually assigned or flagged as unmappable.

579 **Dosage normalization.** We compiled dynasty-specific weight and volume conversion factors
580 from three authoritative metrology references covering the Han, Tang, Song, Yuan, Ming, and
581 Qing dynasties, as well as the modern era. Dosages in original texts were parsed to identify
582 numeric values and unit terms. These values were then matched to the appropriate historical
583 period based on source text dating and converted into modern grams using corresponding
584 conversion factors. Percentage compositions were calculated as the weight of each herb in
585 grams divided by the total formula weight.

Herb Explorer Construction

Literature mining. We searched the PubMed database for articles published between January 2000 and March 2025 using queries combining TCM herb names, derived from Chinese Pharmacopoeia 2020 entries and common synonyms, with pharmacological terms. From 13,227 initial search results, greater than 6,000 publications containing full text or structured abstracts were selected for extraction. A LLM-assisted extraction pipeline identified herb-property associations, including taxonomic classification, medicinal properties, traditional indications, and modern pharmacological activities. Domain experts subsequently performed manual verification on a random 15% sample of extracted records.

Reference book extraction. Eleven authoritative reference books were digitized and structured using OCR and parsing pipelines. These sources included the Grand Dictionary of Chinese Materia Medica, the Chinese Pharmacopoeia 2020 edition (Volumes I and IV), the Encyclopedia of Chinese Materia Medica, and eight specialized reference works covering minority-ethnicity medicines, marine medicines, and Tibetan medicines.

Database consolidation. Herb data were integrated from greater than 10 established databases, including TCMSP [11], ETCM [12], HERB [13], BATMAN-TCM [22], TCMbank [45]. Following data extraction, all herb entries underwent rigorous normalization. We standardized Chinese herb names using the Chinese Pharmacopoeia as the primary authority, aligned Latin binomials against The Plant List and World Flora Online, and removed duplicates using normalized name matching and taxonomic identifier cross-referencing. This pipeline generated a final dataset of 28,728 unique herb entries.

Ingredient Explorer Construction and ADMET Profiling

Literature mining. From 13,227 PubMed articles identified through structured queries using the same search set as the herb explorer module, approximately 9,000 open-access full-text articles were retrieved. A LLM-assisted extraction pipeline identified herb-ingredient

associations. Each extracted ingredient was subsequently validated against PubChem, Chemical Entities of Biological Interest (ChEBI), and the LOTUS database using chemical names and structural matching, yielding 12,214 unique ingredients across 4,005 herbs from 6,005 publications.

Reference monograph extraction. Eleven authoritative reference books curated for the Herb Explorer were reprocessed using specialized extraction templates targeting chemical constituent sections. Ingredient names were normalized and cross-referenced against PubChem, yielding 40,170 unique ingredients across 16,718 herbs.

Database consolidation. Ingredient data integrated from greater than 10 TCM databases yielded 76,518 ingredients across 8,663 herbs.

Data merging and deduplication. The three source datasets were merged and deduplicated using a three-tiered matching strategy. First, we performed exact InChIKey matching for ingredients with resolved structures. Second, we conducted canonical SMILES matching following structure standardization with RDKit. Third, we applied normalized chemical name matching for ingredients lacking resolved structures. Herb-ingredient relationships were preserved and annotated with source provenance, generating a final consolidated dataset containing 87,040 unique active ingredients.

ADMET profiling. All ingredients with valid SMILES representations were submitted to ADMETlab 3.0 [24] for the prediction of 31 absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints. Key physicochemical properties, including molecular weight, log P, hydrogen bond donors and acceptors, topological polar surface area, rotatable bonds, and compliance with Lipinski's Rule of Five, were retrieved from PubChem using the PubChemPy library.

Target Prediction for Active Ingredients

Similarity Ensemble Approach (SEA)-based prediction. High-confidence bioactivity data were extracted from ChEMBL version 35 [46] using specific filtration criteria: human target

organisms, binding or functional assay types, pChEMBL values ≥ 5 (corresponding to IC50, EC50, or Kd ≤ 10 μM), and confidence scores ≥ 7. This pipeline yielded approximately two million compound-target pairs spanning 3,200 unique human protein targets. For each TCM ingredient with a valid SMILES representation, comprising 71,832 compounds, Tanimoto coefficients were computed against all ligand sets using extended-connectivity fingerprints with a diameter of 4 (ECFP4, 2048-bit, radius 2) generated via RDKit [47]. The SEA algorithm [48] was subsequently applied with an E-value threshold of 10^{-5} to identify statistically significant target predictions.

DrugCLIP-based prediction. The pre-trained DrugCLIP model [49] generated fixed-dimensional embeddings for all 87,040 ingredient molecular graphs and approximately 25,000 protein binding pocket structures extracted from the PDB [50]. We computed cosine similarity between each ingredient embedding and all pocket embeddings, retaining the top 100 targets ranked by cosine similarity as candidate predictions for each ingredient. To calibrate the prediction threshold, we evaluated model performance on a held-out benchmark set of 5,000 known compound-target pairs from ChEMBL, selecting the similarity cutoff that maximized the F1 score.

TCMomics Repository Construction

Data collection. We conducted systematic searches across six public repositories, including GEO [25], ArrayExpress [26], PRIDE [27], MetaboLights [28], iProX [29], and Sequence Read Archive (SRA). Search queries comprised greater than 1,200 TCM-related keywords, including individual herb names specified in Chinese pinyin and Latin binomials, formula names, and active ingredient names. Searches were executed in both Chinese and English to maximize coverage.

Relevance classification. Metadata from retrieved datasets were evaluated for TCM relevance through a two-stage process. First, we applied a fine-tuned LLM classifier, which achieved an accuracy of 94.2% on a manually labeled validation set of 500 datasets. Second,

667 we performed manual curation on all positively classified datasets alongside a random 20%
668 sample of negatively classified datasets to correct potential misclassifications.

669 **Quality filtering.** Datasets were excluded if they met any of the following criteria: lacking
670 raw data or processed expression matrices; containing insufficient biological replicates (n<2
671 per group for transcriptomics); utilizing non-standard or proprietary platforms without
672 accessible probe annotations; or displaying ambiguous experimental designs that precluded
673 definitive group assignment.

674 **Categorization.** Retained datasets were classified into 10 TCM categories based on
675 intervention type, including active ingredients, herbs and extracts, formulas, processed
676 products, acupuncture, moxibustion, external applications, dietary therapy, exercise therapy,
677 and other interventions. The final repository contains 1,821 curated datasets, comprising
678 1,352 transcriptomic, 417 proteomic, and 112 metabolomic profiles.

679

680 ## TCM Transcriptome Hub Analytical Pipeline

681 **Study definition.** Transcriptomic datasets were expanded into individual analytical units
682 based on experimental comparisons, where a single study was defined as a direct comparison
683 between one treatment group and one control group. Consequently, multi-arm experimental
684 datasets were decomposed into multiple independent contrast pairs. Following quality control
685 filtering, which excluded datasets with fewer than two biological replicates per group (n < 2),
686 non-human, non-mouse, or non-rat organisms lacking adequate gene annotations, and
687 corrupted data files, 1,037 studies were retained for downstream analysis.

688 **Normalization and quality assessment.** For microarray data, we applied robust multi-array
689 average (RMA) normalization combined with quantile normalization. For RNA sequencing
690 (RNA-seq) data, counts per million (CPM) normalization was performed using trimmed mean
691 of M-values (TMM) adjustment. Data quality was comprehensively evaluated through
692 sample-level principal component analysis (PCA), density distribution plots of expression
693 values, and inter-sample correlation heatmaps.

Differential expression analysis. Microarray datasets were analyzed using the limma package [51] with empirical Bayes moderated t-statistics. RNA-seq count matrices were analyzed using DESeq2 [52] with Wald tests. Differentially expressed genes (DEGs) were identified based on a threshold of |log2FC| >1 and a Benjamini-Hochberg adjusted p-value <0.05.

Pathway enrichment. Gene Ontology (GO) enrichment analysis, covering Biological Process, Molecular Function, and Cellular Component branches, alongside Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, were performed using clusterProfiler [53] with Benjamini-Hochberg correction. Gene Set Enrichment Analysis (GSEA) was executed using the fgsea package [54] with 10,000 permutations. Reactome pathway analysis was conducted using ReactomePA [55]. All enrichment procedures applied a significance threshold of adjusted p-value <0.05.

Network analysis. The identified DEGs were mapped to the STRING protein-protein interaction (PPI) database version 12.0 [56]. Networks were filtered using a combined interaction score threshold of 0.7, representing high-confidence interactions. Hub genes were identified using the cytoHubba algorithm [57], applying degree, betweenness centrality, and maximal clique centrality (MCC) as topological ranking metrics. Network visualization was performed using Cytoscape version 3.10 [58].

Immune cell infiltration. Immune cell proportions were inferred from normalized expression matrices using CIBERSORT [59] for human datasets, assessing 22 immune cell types, and ImmuCellAI [60] for mouse datasets, assessing 24 immune cell types.

Transcription factor prediction. Transcription factor activity was inferred from DEG lists using the TRRUST [61] and DoRothEA [62] databases within the decoupleR framework [63], implementing both univariate linear models and multivariate regression approaches.

TCM Transcriptomic Connectivity Map Analysis. We developed a Connectivity Map (CMAP)-based framework to identify potential TCM interventions that counteract disease transcriptomic signatures (Supplementary Methods). A reference matrix of $\log_2$FC values was

constructed from differential expression profiles of curated TCM datasets after gene symbol normalization and feature filtering. Given a user-supplied disease query, the framework calculates normalized connectivity scores (cs ∈ [-1,1]) using three complementary algorithms: vectorized cosine similarity, Spearman rank correlation, and weighted Kolmogorov-Smirnov enrichment statistics. A negative score denotes signature reversal, indicating putative therapeutic potential. Candidate interventions are subsequently evaluated for statistical significance, adjusted for multiple testing via the Benjamini–Hochberg false discovery rate (FDR) procedure, and ranked to prioritize top therapeutic leads.

Deciphering the Combination Rationale of Wutou Tang

Network pharmacology analysis. We conducted a network pharmacology analysis of WTD using the UniTCM platform. Five constituent herbs of WTD were matched to their corresponding UniTCM identifiers: ZhiChuanWu (ID 11151), MaHuang (ID 8076), HuangQi (ID 7055), ShengBaiShao (ID 4972, manually mapped to BaiShao), and ZhiGanCao (ID 11156). Associations between herbs and active ingredients were retrieved from the UniTCM herb explorer module, with chemical structures annotated through PubChem API queries to acquire PubChem Compound Identifiers (CIDs) and (SMILES strings. ADMET filtering was performed using the ingredient explorer module under rigorous criteria, which required a Lipinski violation count ≤ 2, human intestinal absorption ≥ 0.1, Ames mutagenicity score ≤ 0.7, and drug-induced liver injury probability ≤ 0.7.

Multi-layer target prediction integrated three evidence tiers. Level 1 experimental data extracted from the UniTCM Target2NP tool, level 2 SEA predictions, and level 3 DrugCLIP deep learning predictions. Targets validated by at least two prediction tiers were retained as integrated target sets. PPI networks were constructed using the STRING database version 12.0 with a minimum combined confidence score threshold of 0.4, and hub genes were identified based on degree centrality. GO biological processes, KEGG human pathways version 2021, and WikiPathways human database version 2021 enrichment analyses were conducted using Enrichr. RA disease-associated genes were cross-validated against the

Medical Information Database for Analysis and Synthesis (MIDAS), containing 7,201 known RA-associated genes. Jaccard similarity coefficients were calculated across target sets corresponding to each Jun-Chen-Zuo-Shi hierarchical category pair. Finally, a comprehensive herb-compound-target-disease network comprising 2,315 nodes was constructed for visualization in Cytoscape.

Topological and network separation metrics. To evaluate topological relationships, the Barabási network separation metric (S_{AB}) was computed between each herbal target set (B) and the RA disease module (A) within Network A [64]:

$$S_{AB} = \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \quad (1)$$

where d_{AB} represents the mean shortest-path distance between the RA disease module A (comprising 382 genes and 6,554 interactions from STRING with a combined score ≥ 0.4) and each herbal category target set B. A negative score ($S_{AB}<0$) indicates network overlap, whereas a positive score ($S_{AB}>0$) reflects topological separation. Single-source shortest-path lengths were computed via a parallel breadth-first search over the combined interactome, comprising 499 nodes and 19,200 edges, executing across eight parallel workers. A hyperparameter sweep was performed over the top-N genes ranked by degree centrality (N=10,15,20,…,50,60,80,…,200), and the optimal elbow point was determined using the Kneedle algorithm.

To eliminate potential selection bias inherent to degree centrality, random sampling validation was performed with 30 iterations per N value to construct 95% confidence intervals. Furthermore, the RA PPI network was partitioned into three topological regions based on degree percentiles: the core region (top 20% degree, ≥ 51 edges, 77 hub genes), the intermediate region (40% to 80% degree, 156 genes), and the peripheral region (bottom 40% degree, 149 genes). Per-node topological metrics, including betweenness centrality, minimum and mean shortest-path distances to the top 20 RA core hubs, and the core-to-periphery distance ratio, were calculated. Mann-Whitney U tests were employed to compare RA degree percentile distributions among different herbal categories.

Formula preparation and combinatorial design. WTD consisted of ZhiChuanWu (*Aconiti Radix Cocta*), MaHuang (*Ephedrae Herba*), HuangQi (*Astragali Radix*), ShengBaiShao (*Paeoniae Radix Alba*), and ZhiGanCao (*Glycyrrhizae Radix Praeparata*). All 31 possible herbal combinations (2^5-1) were prepared using a systematic combinatorial decomposition design. Each formulation was extracted twice by reflux in 10 volumes of distilled water for 1 hour per cycle, filtered, concentrated under reduced pressure, and lyophilized. Final crude drug concentrations were standardized to 0.1 g/mL for ZhiChuanWu and 0.15 g/mL for all other constituent herbs.

In vitro functional assays. RAW264.7 cells were cultured in DMEM with 10% FBS at 37 °C (5% CO_2). For cell viability, cells (5×10^3 cells/well, 96-well) were treated with WTD and single herbs (0.02–25 mg/mL) for 24 h using CCK-8 assay, yielding IC_{50} values of 2.70 mg/mL (WTD), 4.84 mg/mL (HuangQi), 15.41 mg/mL (ZhiChuanWu), 8.62 mg/mL (BaiShao), 1.12 mg/mL (MaHuang), and 25.17 mg/mL (ZhiGanCao). Treatments (0.5 and 1.0 mg/mL ZhiChuanWu equivalents) were selected to maintain viability > 80%. For anti-inflammatory evaluation, cells were stimulated with LPS (1 μg/mL) with herbal extracts or dexamethasone (2 μg/mL, positive control) for 24 h. Supernatants were collected for NO (Griess assay) and TNF-α/IL-1β quantification (ELISA). Protein levels of TNF-α and IL-1β were validated by Western blot. Group comparisons used one-way ANOVA with Tukey's post-hoc test ($P < 0.05$).

Untargeted metabolomics and multi-modal data integration. RAW264.7 cells (2×10^5 cells/well, 6-well) were treated with LPS and herbal extracts for 24 h. Intracellular metabolites were extracted with methanol:acetonitrile:water (4:4:2, v/v/v). Supernatants were lyophilized, reconstituted with internal standards, and analyzed by UPLC-Q-TOF-MS (ESI± modes), while protein pellets were quantified for BCA normalization. Raw .mzML files were processed using tidyMass [65]. Data pre-processing included quality filtering, minimum value imputation, and SVR normalization based on QC samples. PCA (prcomp), empirical Bayes moderated t-tests (limma), and OPLS-DA (ropls, validated by 7-fold CV and 1,000 permutations) were performed. Differential metabolites (VIP > 1.0, |FC| > 1.2, $P < 0.05$) were

annotated via MetID against HMDB (MS^1 15 ppm, MS^2 25 ppm tolerance) and mapped to KEGG pathways. Pathway p-values from both modes were integrated via Fisher's combined probability test (Supplementary Methods).

Integration of computational predictions and metabolic perturbations. To bridge computational predictions with biological evidence, the 31 decomposed formula groups were mapped to the Jun-Chen-Zuo-Shi hierarchical classification scheme and assigned functional category profiles (**Supplementary Table 16**). Pharmacological outcomes, specifically NO and TNF-α inhibition rates, were analyzed as a function of category composition, enabling direct alignment between computational metrics (SAB, core-periphery targeting patterns, and Jaccard similarity indices) and experimental observations. Furthermore, to quantify global metabolic alterations, the Pathway Perturbation Score (PPS) was defined as the mean log10-transformed Fisher-combined p-value across all detected KEGG pathways:

$$\mathrm{PPS} = \frac{1}{N}\sum_{i=1}^{N} \log_{10}(p_i) \tag{2}$$

where N represents the total number of detected pathways and pi denotes the integrated p-value for pathway i. Biologically, p-values approaching 1.0 indicate minimal pathway disruption relative to controls, whereas extremely small p-values reflect severe perturbation. Functional classification using curated KEGG keywords identified seven anti-inflammatory repair pathways, including tryptophan, pyruvate, and tricarboxylic acid (TCA) cycle metabolism, which aligns with the metabolic coverage inherent to untargeted metabolomics.

The relationship between PPS metrics and anti-inflammatory phenotypes was assessed using a two-sided Spearman rank correlation coefficient (ρ), with 95% confidence intervals estimated through 5,000 bootstrap resampling iterations. To evaluate sensitivity and robustness, a pre-specified parameter grid search evaluated 4,992 correlation scenarios across four p-value types, six top-N pathway cutoffs, four significance thresholds, eight score variants, and four phenotypic outcomes. To prevent potential overfitting, all scoring functions were constrained by pre-defined biological rationales, grid parameters were restricted to

828 physiologically plausible ranges, and primary correlation patterns were cross-validated across
829 all evaluated anti-inflammatory phenotypes.

830

831 Data Availability

832 All data presented in this study are freely accessible through the UniTCM platform at
833 https://unitcm.qfxulab.com. Comprehensive API documentation is provided at
834 https://unitcm.qfxulab.com/api/v1/docs, and companion R package is hosted at
835 https://github.com/zx122ty/UniTCM_R_Package. In addition, the raw metabolomics datasets
836 for the 31 Wutou Tang combinations can be downloaded directly from TCMomics at
837 https://unitcm.qfxulab.com/#/tcmomics/browse/tcmomics001841.

838 Code Availability

839 The web design and backend algorithm source code is available at
840 https://zenodo.org/records/21711676. Analysis scripts for the Wutou Tang case study are
841 included in the UniTCM R package at https://github.com/zx122ty/UniTCM_R_Package.
842 Other analysis, such as KEGG pathway analysis, data processing and transcriptome analytical
843 pipelines rely on standard open-source bioinformatics tools, including limma, DESeq2,
844 clusterProfiler, fgsea, STRING, and RDKit, without requiring proprietary software.

References

[1] Tang, J. L., Liu, B. Y., & Ma, K. W. (2008). Traditional Chinese medicine. The Lancet, 372(9654), 1938-1940.

[2] World Health Organization. (2019). WHO global report on traditional and complementary medicine 2019. World Health Organization.

[3] Tu, Y. (2011). The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. Nature Medicine, 17(10), 1217-1220.

[4] Tu, Y. (2016). Artemisinin - A gift from traditional Chinese medicine to the world (Nobel Lecture). Angewandte Chemie International Edition, 55(35), 10210-10226.

[5] Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology, 4(11), 682-690.

[6] Li, S., & Zhang, B. (2013). Traditional Chinese medicine network pharmacology: Theory, methodology and application. Chinese Journal of Natural Medicines, 11(2), 110-120.

[7] Vaidya, A. D. B. (2014). Reverse pharmacology: A paradigm shift for drug discovery and development. Current Research in Drug Discovery, 1(2), 39-44.

[8] Patwardhan, B. (2008). Reverse pharmacology and systems approaches for drug discovery and development. Current Bioactive Compounds, 4(4), 201-212.

[9] Vaidya, A. D. B. (2010). Reverse pharmacology - a transdisciplinary path from bedside to bench. Journal of Ayurveda and Integrative Medicine, 1(1), 22-25.

[10] Raut, A. A., Vaidya, A. D. B., & Amonkar, A. J. (2012). Reverse pharmacology for developing herbal therapeutics. In S. C. Mandal et al. (Eds.), Natural Products and Drug Discovery (pp. 317-342). Elsevier.

[11] Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., ... & Wang, Y. (2014). TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. Journal of Cheminformatics, 6, 13.

[12] Xu, H. Y., Zhang, Y. Q., Liu, Z. M., Chen, T., Lv, C. Y., Tang, S. H., ... & Yang, H. J. (2019). ETCM: an encyclopaedia of traditional Chinese medicine. Nucleic Acids Research, 47(D1), D976-D982.

[13] Fang, S., Dong, L., Liu, L., Guo, J., Zhao, L., Zhang, J., ... & Chen, X. (2021). HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. Nucleic Acids Research, 49(D1), D1197-D1206.

[14] Wu, Y., Zhang, F., Yang, K., Fang, S., Bu, D., Li, H., ... & Chen, X. (2019). SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Research, 47(D1), D1110-D1117.

[15] Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research, 49(D1), D325-D334.

[16] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11), 1251-1255.

[17] Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics, 16(6), 1069-1080.

[18] Zhao, L., Wu, H., Qiu, M., Sun, W., Yang, Y., Chen, X., ... & Li, J. (2016). Kidney-yang deficiency syndrome: a multi-dimensional characterization based on systems biology. Chinese Journal of Integrative Medicine, 22(6), 410-418.

[19] World Health Organization. (2019). International Classification of Diseases 11th Revision (ICD-11). WHO.

[20] Qiu, G. (2005). Weights and Measures in Ancient China. Science Press.

[21] Zhang, R. Z., Yu, S. J., Bai, H., & Ning, K. (2022). Traditional Chinese medicine: an updated overview on the current state of research and clinical applications. Phytomedicine, 95, 153890.

[22] Liu, Z., Guo, F., Wang, Y., Li, C., Zhang, X., Li, H., ... & Chen, X. (2016). BATMAN-TCM: a bioinformatics analysis tool for molecular mechanism of traditional Chinese medicine. Scientific Reports, 6, 21146.

[23] Chinese Pharmacopoeia Commission. (2020). Pharmacopoeia of the People's Republic of China, Volume I. China Medical Science Press.

[24] Fu, L., Shi, S., Yi, J., Wang, N., He, Y., Wu, Z., ... & Cao, D. (2024). ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Research, 52(W1), W422-W431.

[25] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets - update. Nucleic Acids Research, 41(D1), D991-D995.

[26] Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., ... & Brazma, A. (2007). ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucleic Acids Research, 35(Database issue), D747-D750.

[27] Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., ... & Vizcaino, J. A. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research, 50(D1), D543-D552.

[28] Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., & O'Donovan, C. (2020). MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Research, 48(D1), D440-D444.

[29] Chen, T., Ma, J., Liu, Y., Chen, Z., Xiao, N., Lu, Y., ... & Zhu, Y. (2022). iProX in 2021: connecting proteomics data sharing with big data. Nucleic Acids Research, 50(D1), D1522-D1527.

[30] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights, 14, 1177932219899057.

[31] Almutairi, K. B., Nossent, J. C., Preen, D. B., Keen, H. I., & Inderjeeth, C. A. (2021). The global prevalence of rheumatoid arthritis: A meta-analysis based on a systematic review. Rheumatology International, 41(5), 863-877.

[32] GBD 2021 Rheumatoid Arthritis Collaborators. (2023). Global, regional, and national burden of rheumatoid arthritis, 1990-2020, and projections to 2050: A systematic analysis of the Global Burden of Disease Study 2021. The Lancet Rheumatology, 5(10), e594-e610.

[33] Nagy, G., Roodenrijs, N. M., Welsing, P. M., Kedves, N., Hamar, A., van der Heijde, D., ... & van Laar, J. M. (2021). EULAR points to consider for the management of difficult-to-treat rheumatoid arthritis. Annals of the Rheumatic Diseases, 80(1), 31-35.

[34] Smolen, J. S., Aletaha, D., & McInnes, I. B. (2016). Rheumatoid arthritis. The Lancet, 388(10055), 2023-2038.

[35] Moon, J. H., Sung, W. S., Choi, S. K., Kim, J. H., Suh, J. W., Kim, J. H., ... & Kim, E. J. (2022). The effectiveness and safety of Wu tou decoction on rheumatoid arthritis: A protocol for systematic review and/or meta-analysis. Medicine, 101(11), e29105.

[36] Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Nature Reviews Drug Discovery, 11(3), 191-200.

[37] Patwardhan, B., & Vaidya, A. D. (2010). Natural products drug discovery: Accelerating the clinical candidates through reverse pharmacology. Current Science, 98(6), 780-790.

[38] Mokhtari, R. B., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., & Yeger, H. (2017). Combination therapy in combating cancer. Oncotarget, 8(23), 38022-38043.

[39] Mok, T. S., Wu, Y. L., Thongprasert, S., Yang, C. H., Chu, D. T., Saijo, N., ... & Fukuoka, M. (2009). Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. The New England Journal of Medicine, 361(10), 947-957.

[40] Willcox, M. L., Graz, B., Falquet, J., Diakite, C., Giani, S., & Diallo, D. (2011). A reverse pharmacology approach for developing an anti-malarial phytomedicine. Malaria Journal, 10(Suppl 1), Article S8.

[41] World Health Organization. (2007). WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region. WHO Western Pacific Region.

[42] Open Chinese Convert (OpenCC) Project. (2023). OpenCC: Open Chinese Convert. GitHub.

[43] Wei, H., Sun, Y., & Li, Y. (2025). Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

[44] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

[45] Lv, Q., Chen, G., He, H., Yang, Z., Zhao, L., Zhang, K., & Chen, C. Y. (2023). TCMBank-the largest TCM database provides deep learning-based Chinese-Western medicine exclusion prediction. Signal transduction and targeted therapy, 8(1), 127.

[46] Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., ... & Leach, A. R. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research, 52(D1), D1180-D1192.

[47] Landrum, G. (2024). RDKit: Open-Source Cheminformatics Software. GitHub.

[48] Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. Nature Biotechnology, 25(2), 197-206.

[49] Jia, Y., Chen, Y., Yan, G., Jiang, T., Zheng, S., & Zhang, L. (2026). Deep contrastive learning enables genome-wide virtual screening. Science, 391, eads9530.

[50] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... &

973 Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242.

974 [51] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015).
975 limma powers differential expression analyses for RNA-sequencing and microarray studies.
976 Nucleic Acids Research, 43(7), e47.

977 [52] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
978 dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

979 [53] Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for
980 comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology,
981 16(5), 284-287.

982 [54] Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A.
983 (2021). Fast gene set enrichment analysis. bioRxiv, 060012.

984 [55] Yu, G., & He, Q. Y. (2016). ReactomePA: an R/Bioconductor package for reactome
985 pathway analysis and visualization. Molecular BioSystems, 12(2), 477-479.

986 [56] Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., ... &
987 von Mering, C. (2023). The STRING database in 2023: protein-protein association networks
988 and functional enrichment analyses for any observed set of proteins. Nucleic Acids Research,
989 51(D1), D638-D646.

990 [57] Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2014).
991 cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC
992 Systems Biology, 8(Suppl 4), S11.

993 [58] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T.
994 (2003). Cytoscape: a software environment for integrated models of biomolecular interaction
995 networks. Genome Research, 13(11), 2498-2504.

996 [59] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., ... & Alizadeh,
997 A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature
998 Methods, 12(5), 453-457.

[60] Miao, Y. R., Zhang, Q., Lei, Q., Luo, M., Xie, G. Y., Wang, H., & Guo, A. Y. (2020). ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application to cancer immunotherapy. Advanced Science, 7(7), 1902880.

[61] Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., ... & Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research, 46(D1), D380-D386.

[62] Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research, 29(8), 1363-1375.

[63] Badia-i-Mompel, P., Velez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Muller-Dott, S., ... & Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances, 2(1), vbac016.

[64] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabasi, A. L. (2015). Uncovering disease-disease relationships through the incomplete interactome. Science, 347(6224), 1257601.

[65] Wang, X., Liu, Y., Jiang, C., Huang, Z., Yan, H., Wong, S. H., Johnson, C. H., Zhang, J., Ge, Y., Zhang, F., Zhang, J., Lai, R., Gao, P., Zhang, X., & Shen, X. (2026). TidyMass2: advancing LC-MS untargeted metabolomics through metabolite origin inference and metabolic feature-based functional module analysis. Nature Communications, 17, 68464.

Notice & Disclaimer

The content provided by this database is intended exclusively for academic research and educational exchange. Any direct or indirect commercial use is strictly prohibited. If you believe that any publicly accessible data or content on this platform constitutes an infringement of rights, please notify us via the email address above. Verified infringing material will be removed or amended accordingly.

The information on formula, medicinal materials, chemical ingredients, and targets contained in this database is provided solely for scientific research reference. It does not replace the clinical diagnosis, treatment plan, or prescription guidance of a licensed medical professional. Under no circumstances should users rely upon the data herein for self-medication or treatment adjustment. Always consult a qualified healthcare provider at a certified medical institution regarding any health concerns.

Acknowledgements

This work was conducted, in part, using the resources of the Information Technology Center and the State Key Laboratory of Computer-Aided Design & Computer Graphics at Zhejiang University. We thank Cheng Ma from the Core Facilities, Zhejiang University School of Medicine for his technical support. This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LMS25H280002