Supplementary Information

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

Contents

Supplementary Figures

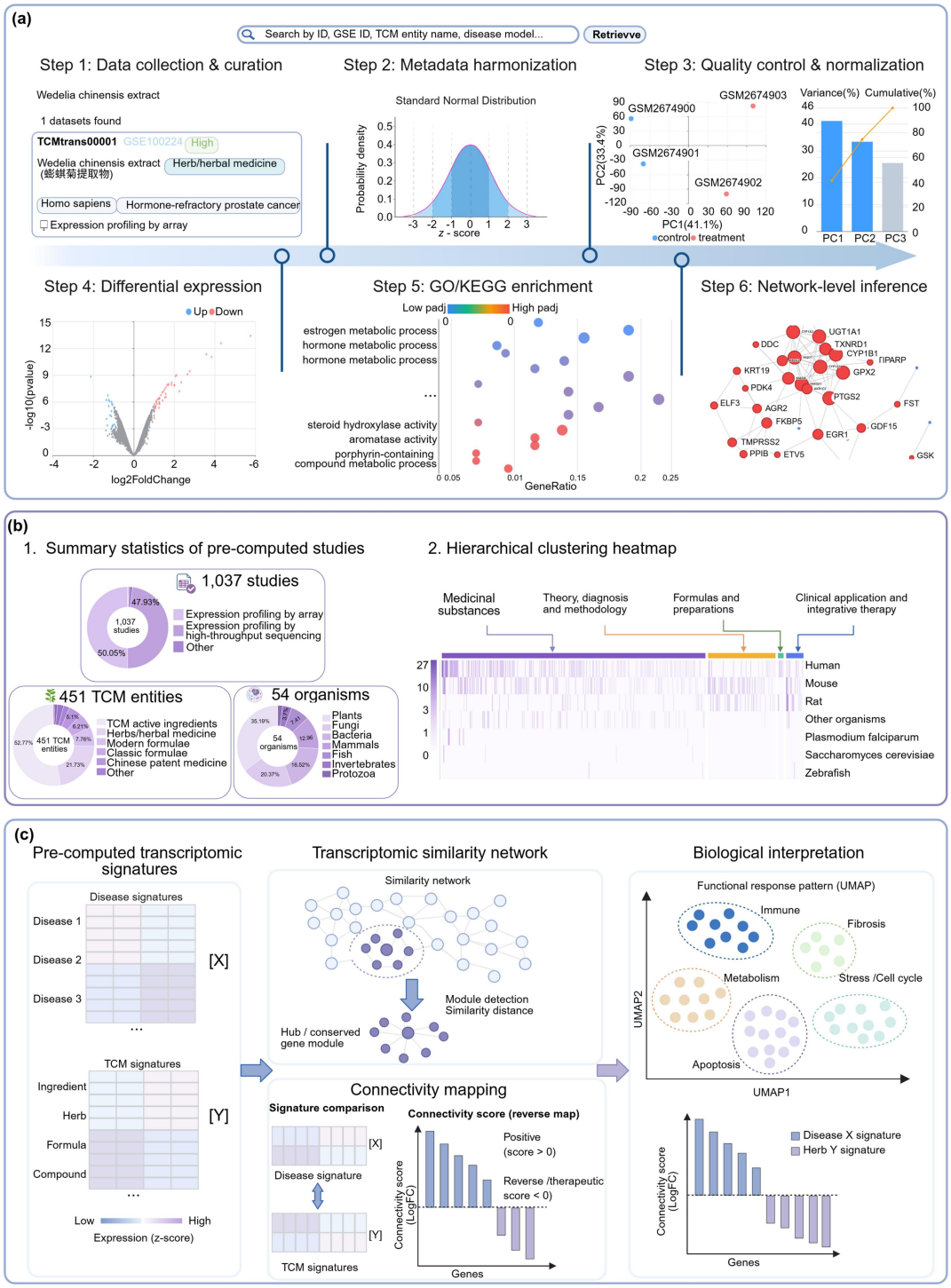

Figure S1. System architecture and analytical workflows of the TCM Transcriptome Hub.

(a) *Standardized bioinformatics analytical pipeline. The flowchart outlines the streamlined workflow from data collection and curation (Step 1), metadata harmonization (Step 2), and*

quality control & normalization (Step 3), through to differential expression analysis (Step 4), GO/KEGG pathway enrichment (Step 5), and network-level inference (Step 6), with representative visualization outputs highlighted for each stage. ***(b)*** *Standardized re-analysis across 1,037 studies, detailing the summary statistics of curated datasets (including TCM entities and organisms) and a comprehensive hierarchical clustering heatmap stratified by clinical applications and traditional therapeutic theories.* ***(c)*** *Integrative downstream analysis, illustrating the application of pre-computed transcriptomic signatures, transcriptomic similarity networks, therapeutic connectivity mapping, and biological interpretation via functional response patterns.*

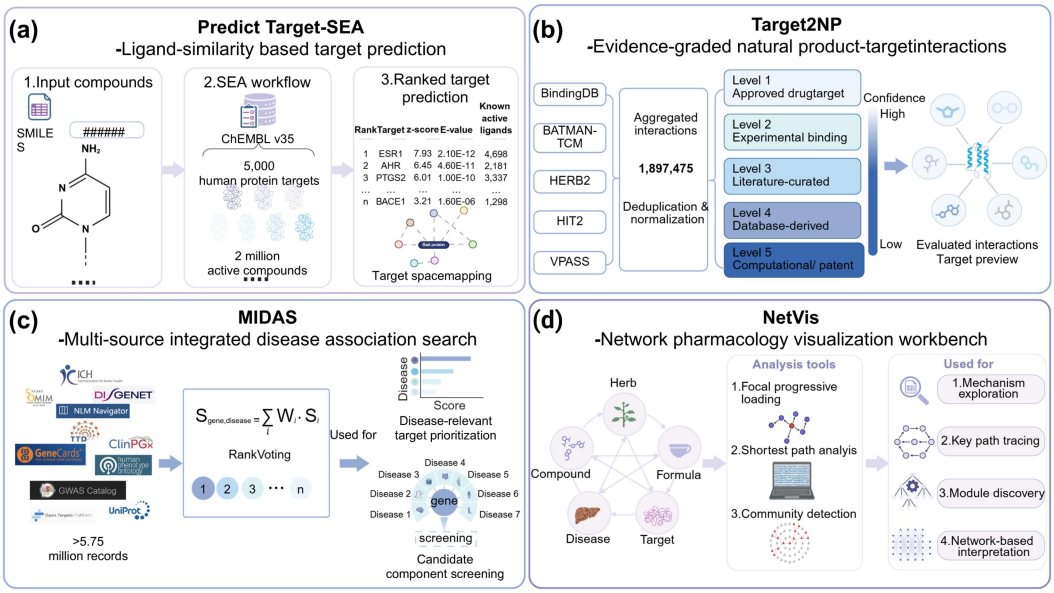

Figure S2. Overview of the analytical tools integrated into the UniTCM platform. (a) *Predict Target-SEA workflow for ligand-similarity-based target prediction. User-submitted compounds in SMILES format are screened against the ChEMBL v35 reference library (~2*

million active compounds mapped to ~5,000 human protein targets) using the Similarity Ensemble Approach (SEA) algorithm. The output features ranked target predictions accompanied by z-scores, E-values, and known active ligands, alongside a target space mapping visualization. ***(b)*** *Target2NP module for evidence-graded natural product-target interactions. Known interactions (1,897,475 pairs) are aggregated from five core databases (BindingDB, BATMAN-TCM, HERB, HIT2, and VPASS) followed by deduplication and normalization. Interactions are categorized into a five-tier confidence scale ranging from Level 1 (approved drug-target) to Level 5 (computational/patent).* ***(c)*** *MIDAS (Multi-source Integrated Disease Association Search) framework for gene-disease association mining. This tool aggregates over 5.75 million records from 11 authoritative databases. Candidate components and disease-relevant targets are prioritized via the Rank Voting algorithm () and visualized interactively. (d) NetVis network pharmacology visualization workbench. It enables interactive exploration of the five-layer TCM network (Disease-Formula-Herb-Compound-Target). Built-in analysis tools include focal progressive loading, shortest path analysis, and community detection, which facilitate mechanism exploration, key path tracing, module discovery, and network-based interpretation.*

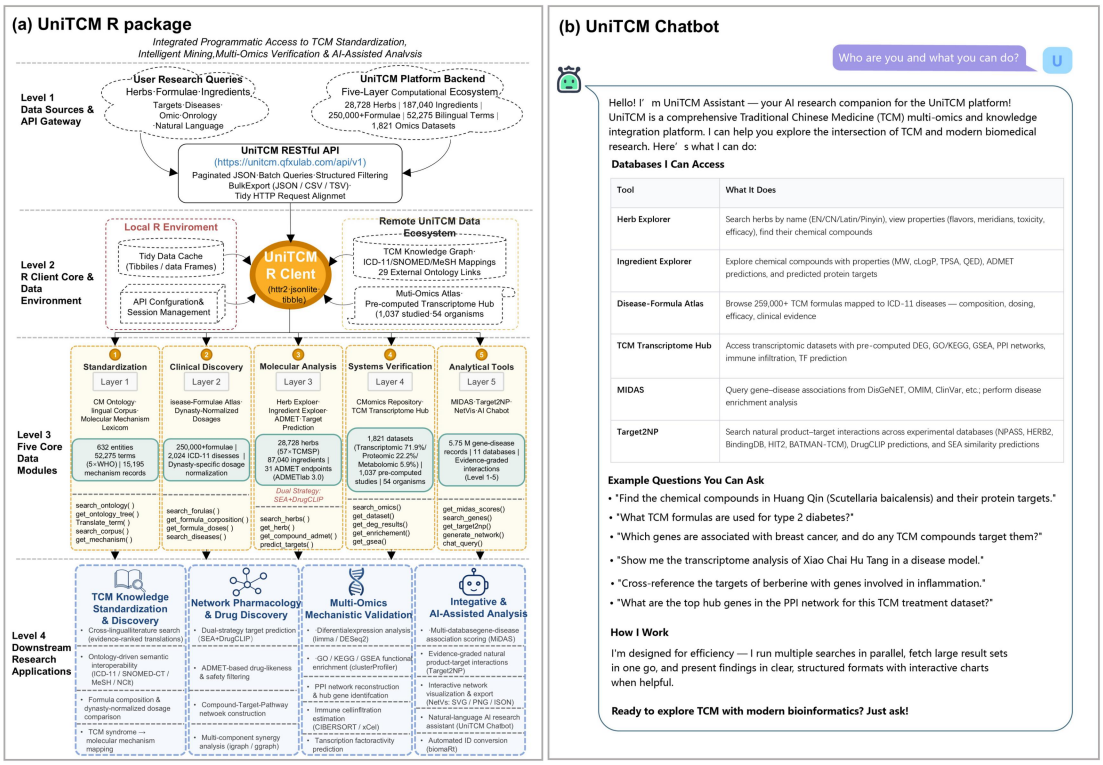

Figure S3. Programmatic and conversational access to UniTCM. (a) *UniTCM R package.* ***(b)*** *UniTCM Chatbot. A natural language interface designed for conversational queries across platform databases. It enables automated cross-referencing, data synthesis, and multi-step reasoning across the core functional modules, including Standardization, Clinical Discovery, Molecular Analysis, Systems Verification, and Analytical Tools.*

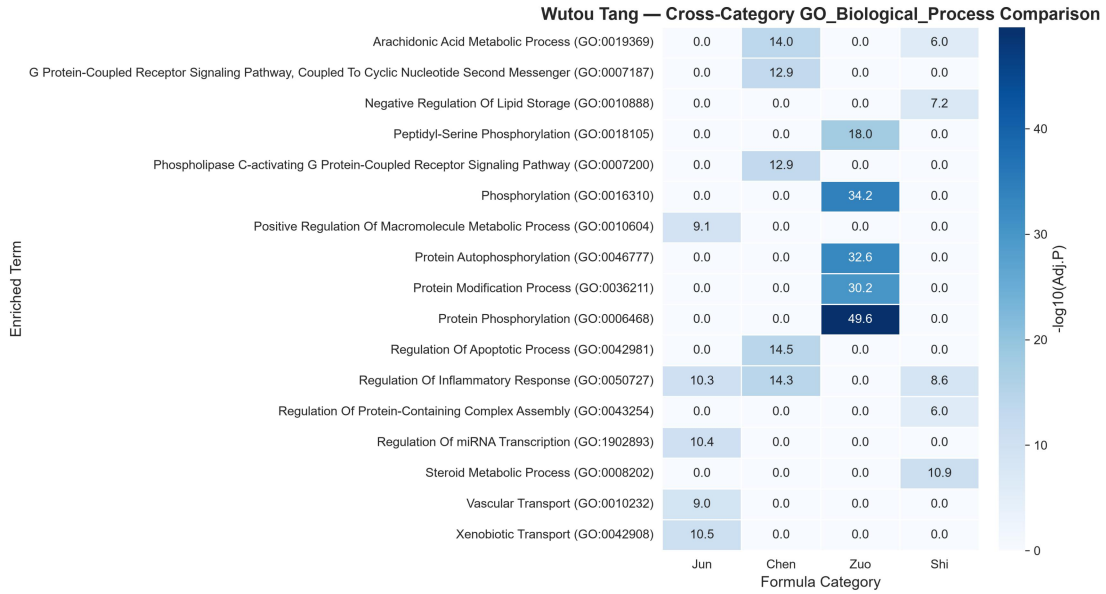

Figure S4. Per-category GO enrichment analysis of Wutou Tang.

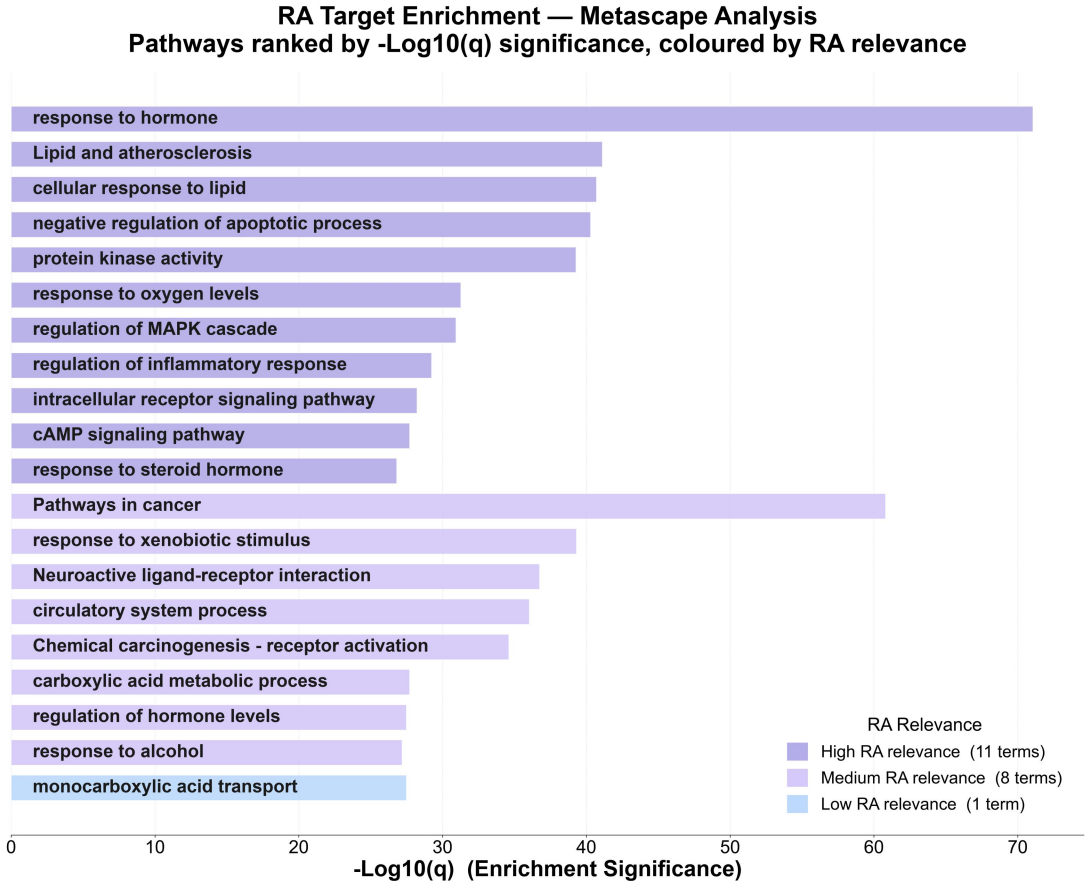

Figure S5. GO and KEGG analysis results of Rheumatoid arthritis (RA) overlapping targets.

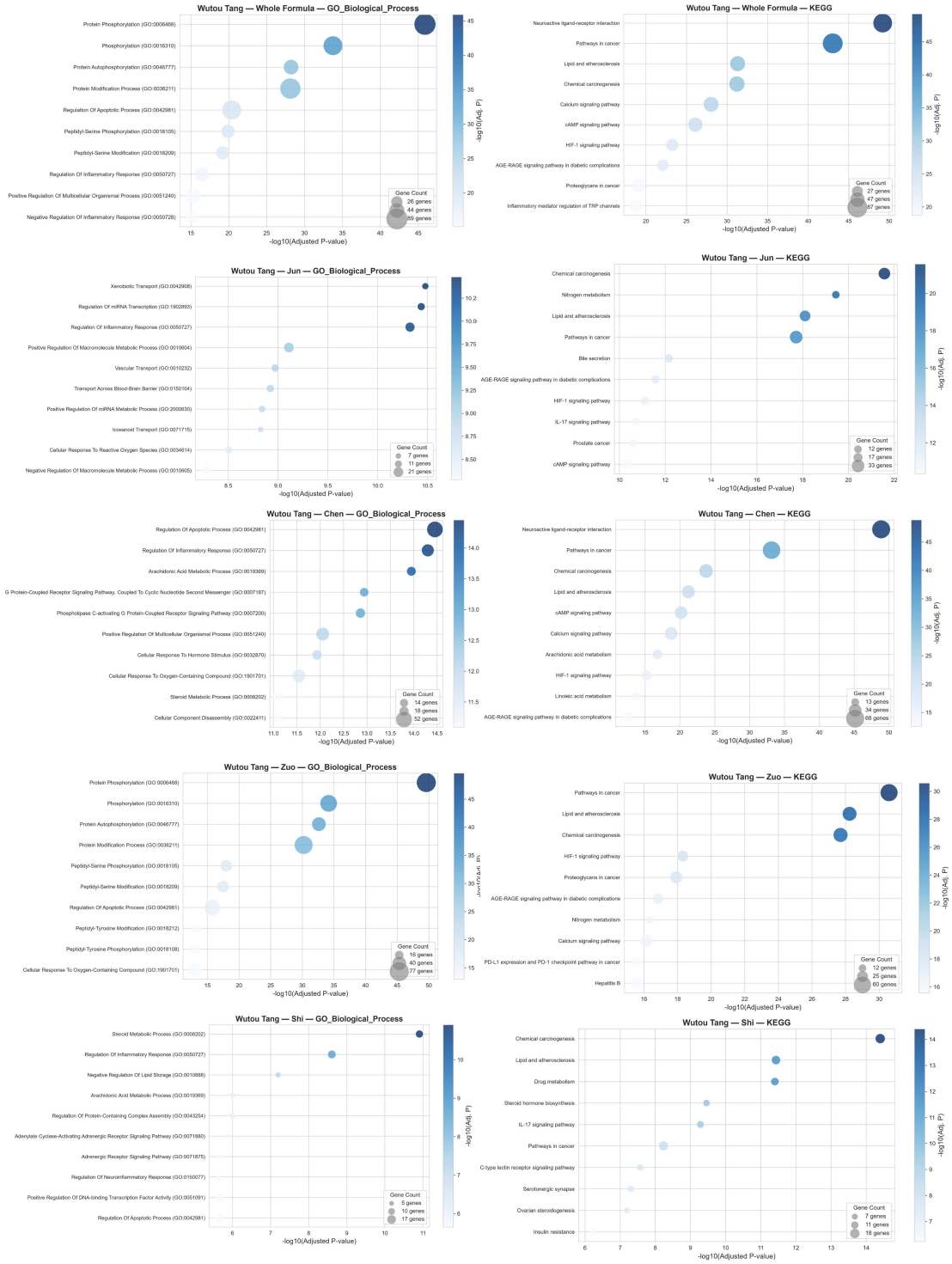

Figure S6. Per-category KEGG enrichment analysis of Wutou Tang.

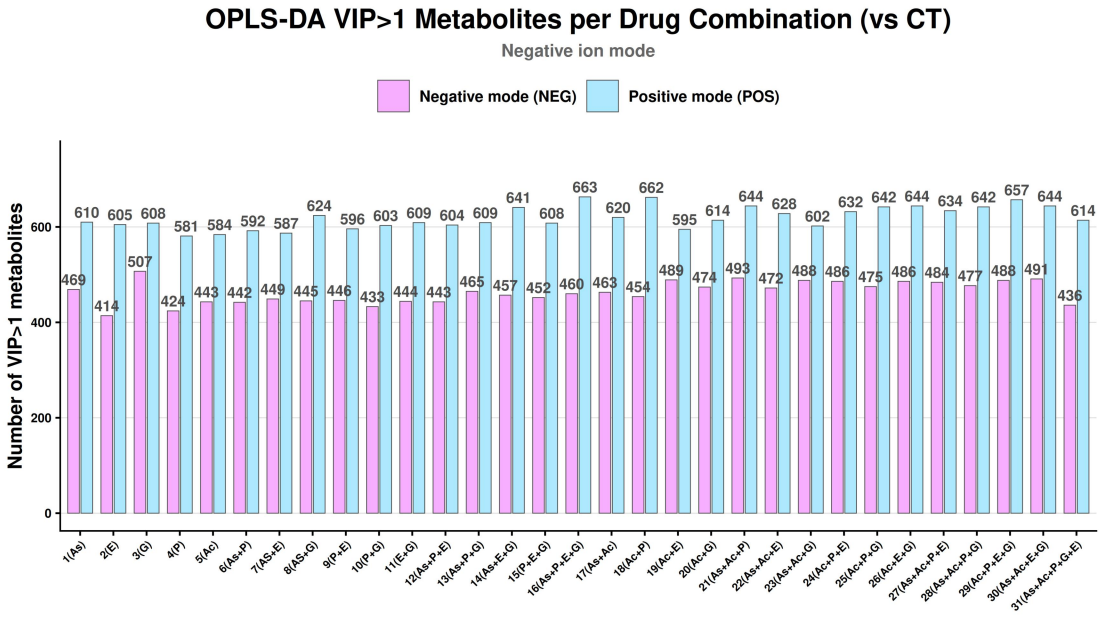

Figure S7. OPLS-DA analysis identify the metabolites with VIP >1.

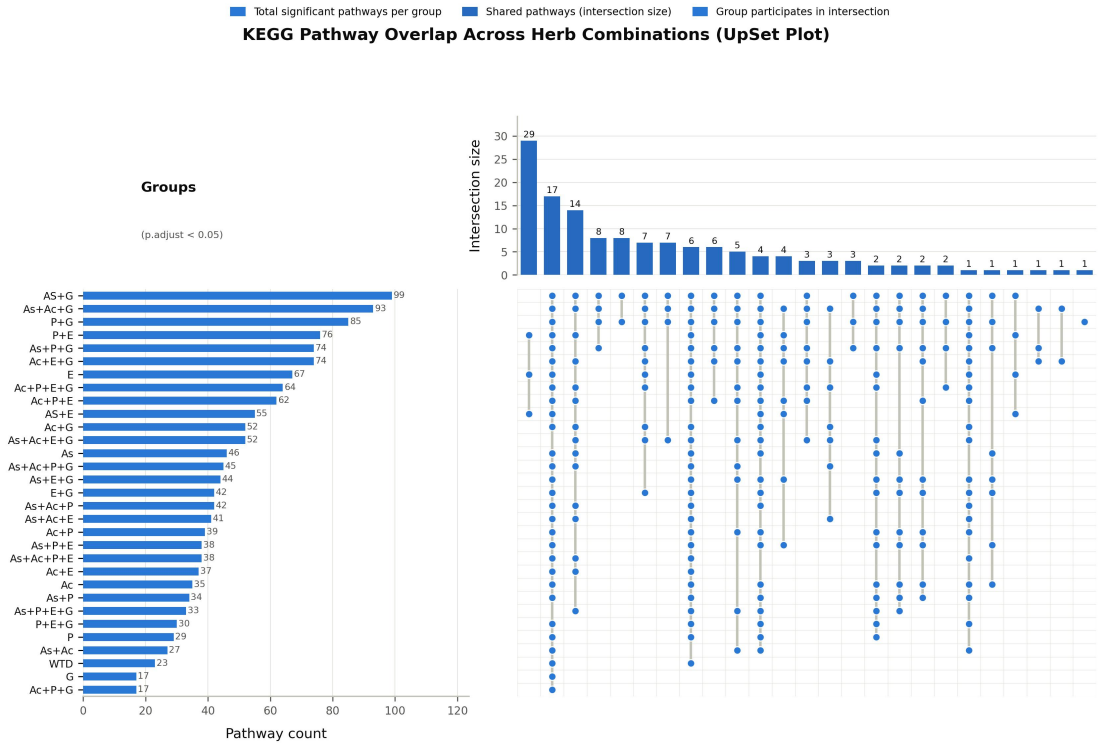

Figure S8. UpSet plot illustrating the overlap and distribution of significantly altered metabolic pathways (p-adjusted < 0.05) across combinations.

Supplementary Methods

Deciphering the Combination Rationale of Wutou Tang

Detailed Metabolomics Protocol

Cell Culture, Treatment, and Sample Preparation

RAW264.7 murine macrophages were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) at 37 °C in a humidified atmosphere containing 5% CO_2. For the anti-inflammatory activity evaluation, cells were seeded in 96-well plates at a density of 5×10^3 cells per well (100 μL/well) and incubated for 24 h to allow full adherence. Cell viability was determined by the CCK-8 assay following treatment with gradient concentrations of LPS or dexamethasone (DEX). For formula evaluation, cells were treated with LPS (1 μg/mL) together with different herbal formula extracts (containing 0.5 mg/mL or 1.0 mg/mL ZhiChuanWu equivalent) or DEX (2 μg/mL, positive control). After intervention, cell supernatants were collected to measure NO release (via Griess assay) and inflammatory cytokine levels (TNF-α and IL-1β, via ELISA kits).

For untargeted metabolomics, RAW264.7 cells in the exponential growth phase were seeded in 6-well plates at a density of 2×10^5 cells per well (2 mL system) and allowed to adhere completely. Cells were treated with LPS (1 μg/mL) and different herbal formula extracts for 24 h. After discarding the supernatant, cells were rinsed twice with pre-cooled PBS and scraped after adding an appropriate volume of pre-cooled extraction solvent (methanol: acetonitrile: water = 4:4:2, v/v/v). The cell suspension was centrifuged at 12,000 rpm for 10 min at 4 °C. The supernatant was collected and lyophilized to dryness, while the remaining

protein pellet at the bottom was measured using the BCA Protein Assay Kit for subsequent data normalization. The dried samples were reconstituted in an appropriate volume of extraction solvent containing internal standards prior to high-resolution mass spectrometry analysis.

Western Blot Analysis

Protein expression levels of TNF-α and IL-1β in RAW264.7 cells under specified treatments were evaluated by Western blot. Briefly, total cellular proteins were extracted, quantified, separated by SDS-PAGE, and transferred onto membranes. After primary and secondary antibody incubations, protein bands were visualized using enhanced chemiluminescence (ECL). Densitometric analysis of grayscale values was performed for semi-quantitative evaluation of protein expression differences between groups.

UPLC-Q-TOF-MS Analysis

Chromatographic separation was performed on an ACQUITY UPLC I-Class system (Waters Corporation, Milford, MA, USA) equipped with an ACQUITY UPLC BEH C_{18} column (2.1 × 100 mm, 1.7 μm particle size; Waters). The column temperature was maintained at 40 °C, and the autosampler tray was kept at 4 °C. The mobile phase consisted of (A) 0.1% formic acid in water (v/v) and (B) 0.1% formic acid in acetonitrile (v/v). The gradient program was set as follows: 0–2 min, 5% B; 2–8 min, 5–60% B; 8–14 min, 60–95% B; 14–17 min, 95% B; 17–17.1 min, 95–5% B; 17.1–20 min, 5% B. The flow rate was 0.3 mL/min, and the injection volume was 2 μL.

Mass spectrometric detection was carried out on a Xevo G2-XS Q-TOF mass spectrometer (Waters Corporation, Manchester, UK) equipped with an electrospray ionization (ESI) source in both positive (ESI+) and negative (ESI-) ion modes. Source parameters were: capillary voltage, 3.0 kV (ESI+) / 2.5 kV (ESI-); sampling cone voltage, 40 V; source offset, 80 V; source temperature, 120 °C; desolvation temperature, 450 °C; cone gas flow, 50 L/h; desolvation gas flow, 800 L/h. Full-scan spectra were acquired over m/z 50–1200 Da with a scan time of 0.2 s. Leucine-enkephalin was continuously infused via LockSpray for real-time mass correction. Quality control (QC) samples were injected periodically throughout the analytical sequence to monitor system stability.

Data Processing and Statistical Analysis

Raw LC-MS data files (.mzML) were processed using the tidyMass framework. Feature extraction and peak alignment were conducted via the massprocesser package (version 1.0.0) [1]. Prior to statistical analysis, raw feature tables underwent a rigorous quality control and filtering pipeline: (1) background noise features were excluded if the mean signal in biological samples was less than the configured fold-change threshold relative to the median signal in blank samples; (2) features exceeding the maximum missing value threshold per group were removed; and (3) feature stability was evaluated using Quality Control (QC) samples, excluding features with a relative standard deviation (RSD) exceeding the predefined threshold. Missing values were imputed using the minimum value method (impute_mv, method = "minimum"). To correct for signal drift across analytical runs, intensity data were normalized using Support Vector Regression (SVR) based on QC samples (normalize_data, method = "svr") [2].

Statistical analyses were conducted in R. Principal Component Analysis (PCA) was performed on auto-scaled (unit-variance) data using the prcomp function to visualize global metabolic profiles and evaluate sample clustering. Pairwise differential analysis was performed using empirical Bayes moderated t-tests via the limma package, combined with Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA) using the ropls package to calculate Variable Importance in Projection (VIP) scores [3]. Model validity for OPLS-DA was validated using 7-fold cross-validation (R^2Y and Q^2 metrics) and 1,000-iteration permutation testing (valid if the permutation-derived Q^2 intercept was < 0.05). Differential metabolites were identified based on VIP > 1.0, | Fold Change| >1.2, and p-value < 0.05 (or false discovery rate, FDR < 0.05, adjusted by the Benjamini-Hochberg method) [4,5,6].

TCM Transcriptome Hub Connectivity Map (CMAP) Analysis

Overview

To identify TCM interventions — herbs, bioactive ingredients, and formulae — whose transcriptomic perturbations counteract a user-supplied disease signature, a Connectivity Map (CMAP) module was implemented following the conceptual framework of Lamb et al [7]. Given a query expression signature, the module quantifies the transcriptional connectivity between the query and each reference perturbation profile in the TCM transcriptome compendium and returns a ranked list of candidate interventions. Three complementary scoring algorithms are provided: (i) a vectorised cosine similarity, (ii) a Spearman rank correlation, and (iii) the classical weighted Kolmogorov-Smirnov (KS) enrichment statistic. All three produce a normalised connectivity score $cs \in [-1,1]$, for which a negative score

denotes reversal of the query signature (a putative therapeutic or inhibitory effect) and a positive score denotes mimicry (a concordant or synergistic mechanism).

Reference Perturbation Database Construction

Reference perturbation profiles were derived from the differential-expression analyses of all curated datasets in the TCM Transcriptome Hub. For each dataset, the per-gene $\log_2$ fold-change (log2FC, treatment versus control) was extracted from its differential-expression results. Gene identifiers were harmonised to upper-case official HGNC gene symbols; where multiple probes or transcripts mapped to the same symbol within a dataset, the first occurrence was retained.

To suppress sparsely measured features and control memory footprint, only genes detected in at least τ datasets were retained (default τ=50; user tunable). The retained data were assembled into a single gene-by-dataset log2FC matrix:

$$\mathbf{M} \in \mathbb{R}^{G \times D}, M_{g,d} = \log_2\text{FC of gene } g \text{ in dataset } d$$

where G is the number of retained genes and D the number of reference datasets. Entries corresponding to genes not quantified in a given dataset were set to 0. The matrix was cast to single-precision floating point and persisted in columnar apache parquet format together with a companion metadata table (mapping each dataset identifier to its TCM entity name, botanical classification, organism, tissue/cell line, disease model, and GEO accession). At query time the matrix is memory-resident and cached to enable low-latency, whole-compendium scoring.

Query Signature Construction

The user submits a query signature as a list of gene-value pairs (gene symbol and its associated $\log_2$FC), supplied either as pasted text or an uploaded delimited file (comma-, tab-, or whitespace-separated; UTF-8 and GBK encodings supported). Header and comment lines are discarded, gene symbols are normalised to upper case, and non-numeric records are ignored.

Let the parsed query be the mapping $\{(g, x_g)\}$. The query is first aligned to the reference gene universe $\mathcal{G} = \{g: g \in \boldsymbol{M}\}$, yielding the matched signature:

$$\mathcal{Q} = \{\, g \in \mathcal{G}: (g, x_g) \in \text{query}\,\}, n = |\mathcal{Q}| \quad (1)$$

Only the n matched genes contribute to scoring; the number of submitted genes and the number successfully matched are both reported to the user as quality-control diagnostics.

For the correlation-based methods (cosine, Spearman), the full aligned vector $\boldsymbol{u} = (x_g)_{g \in \mathcal{Q}}$ is used directly. For the KS method, a directional tag set is additionally extracted: genes are ranked by $\log_2$FC, and the top N up-regulated ($x_g > 0$) and top N down-regulated ($x_g < 0$) genes are taken as the up-tag set N and down-tag set D, respectively (default N=150; user tunable). When fewer than N genes are available in either direction, all available genes are used.

Connectivity Score Algorithms

Cosine similarity

For each reference dataset d, let $\boldsymbol{v}^{(d)} = (M_{g,d})_{g \in \mathcal{Q}}$ be the reference vector restricted to the matched genes. The connectivity score is the cosine of the angle between the query and reference vectors:

$$cs_d^{\cos} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}^{(d)}}{\|\boldsymbol{u}\|_2 \left\|\boldsymbol{v}^{(d)}\right\|_2} = \frac{\sum_{g \in Q} x_g \, M_{g,d}}{\sqrt{\sum_{g \in Q} x_g^2} \sqrt{\sum_{g \in Q} M_{g,d}^2}} \tag{2}$$

This computation is fully vectorised across all D datasets as a single matrix-vector product, making it the default method for interactive use. An approximate two-sided p-value is obtained by treating $cs_d^{\cos}$ as a product-moment correlation coefficient and applying the Fisher variance-stabilising z-transformation:

$$z_d = \operatorname{arctanh}(cs_d^{\cos})\sqrt{n-3}, p_d = 2(1 - \Phi(|z_d|)) \tag{3}$$

where Φ is the standard normal cumulative distribution function and the degrees-of-freedom term is floored at unity for very small signatures.

Spearman rank correlation

To provide a scoring scheme that is robust to outliers and to monotone but non-linear relationships, the Spearman rank correlation is computed between the query vector and each reference vector $\boldsymbol{v}^{(d)}$:

$$cs_d^{\rho} = \rho(\mathbf{u}, \mathbf{v}^{(d)}) = 1 - \frac{6 \sum_{i=1}^{n} \delta_i^2}{n\left(n^2 - 1\right)} \tag{4}$$

where δ_i is the difference between the rank of gene i in the query and its rank in reference profile d. Reference (or query) vectors with zero variance are assigned $cs_d^{\rho} = 0$ and $p_d = 1$. The associated p-value is obtained from the asymptotic t-approximation of the rank-correlation null distribution.

Weighted Kolmogorov - Smirnov connectivity score

The third method reproduces the enrichment-based connectivity score of the original Connectivity Map. For each reference dataset d, all genes are ordered by descending $\log_2$FC, so that rank 1 corresponds to the most strongly up-regulated gene. Given a directional tag set S (either $\mathcal{U}$ or $\mathcal{D}$) whose members occur at sorted rank positions $V(1) \leq V(2) \leq \cdots \leq V(t)$ in this ordering—where counts only tags present in the profile—the running-sum KS enrichment statistic is defined by the two extreme deviations.

$$a = \max_{1 \leq j \leq t} \left[\frac{j}{t} - \frac{V(j)}{G} \right], b = \max_{1 \leq j \leq t} \left[\frac{V(j)}{G} - \frac{j-1}{t} \right] \quad (5)$$

$$\mathrm{ES}(S, d) = \begin{cases} a, & a > b, \\ -b, & \text{otherwise} \end{cases} \quad (6)$$

A positive ES indicates that the tag set is concentrated towards the top (up-regulated pole) of the reference ranking, and a negative ES that it is concentrated towards the bottom (down-regulated pole). Writing $\mathrm{ES}_{\mathrm{up}} = \mathrm{ES}(\mathcal{U}, d)$ and $\mathrm{ES}_{\mathrm{down}} = \mathrm{ES}(\mathcal{D}, d)$, the connectivity score combines the two directions:

$$cs_d^{\mathrm{KS}} = \begin{cases} \dfrac{\mathrm{ES}_{\mathrm{up}} - \mathrm{ES}_{\mathrm{down}}}{2}, & \mathrm{sgn}(\mathrm{ES}_{\mathrm{up}}) \neq \mathrm{sgn}(\mathrm{ES}_{\mathrm{down}}) \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Requiring the two enrichment scores to have opposite signs enforces directional coherence: the score is non-zero only when the up- and down-tag sets are enriched at opposite poles of the reference profile, as expected under a genuine mimicry or reversal relationship. When the up-tags are enriched at the top and the down-tags at the bottom of a reference profile, cs_d^{KS} is positive (mimicry); the mirror configuration yields a negative score (reversal). By construction $cs_d^{\mathrm{KS}} \in [-1, 1]$.

Statistical Significance and Ranking

For the KS method, statistical significance is assessed empirically against the null distribution generated by the connectivity scores of the entire reference compendium. For each dataset, a two-sided empirical p-value is computed as the proportion of reference profiles that achieve a connectivity magnitude at least as extreme as the observed value, incorporating additive smoothing constants to prevent p-values of exactly zero.

Irrespective of the scoring method, the resulting per-dataset p-values are corrected for multiple hypothesis testing across all dataset comparisons using the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Once ordered by significance, these p-values are systematically adjusted to account for false positives.

Candidate interventions are ultimately ranked in ascending order by their connectivity score, ensuring that the candidates with the strongest potential to reverse the disease signature—those with the most negative scores—appear first. Each output record is annotated with its connectivity score, raw p-value, and FDR-adjusted p-value, alongside the corresponding TCM entity metadata. Finally, the results are visualized through an interactive lollipop ranking plot and a connectivity-score versus volcano plot.

Implementation

The CMAP module is implemented in Python 3.12 using NumPy and pandas for vectorised linear-algebra operations, SciPy for rank correlation, and statsmodels for Benjamini - Hochberg FDR control. The reference matrix is stored and loaded via Apache Parquet and

cached in memory to support whole-compendium scoring within a single request. All connectivity scores are bounded to [-1,1] and reported to six decimal places.

Tunable constants. Default parameter values are given above and may be adjusted for a given study: the minimum-dataset detection threshold (τ=50), the number of directional tag genes per pole for the KS method (N=150), and the choice of scoring algorithm (cosine, Spearman, KS).

Supplementary Results

KEGG Pathway Enrichment Analysis of Wutou Tang Case Study

Background

Wutou Tang (WTD) is a TCM formula consisting of five herbal components: Astragali Radix (As, Astragalus), *Ephedrae Herba* (E, Ephedra), *Glycyrrhizae Radix* (G, Licorice), *Paeoniae Radix Alba* (P, White Peony), and *Aconiti Radix* (Ac, Aconite). To investigate the compatibility rules within this formula, 31 herb combination groups were systematically designed by adding or removing individual herbs. Untargeted metabolomics was performed to profile the metabolic perturbations induced by each combination. This report analyzes the KEGG pathway enrichment results derived from the differential metabolites of each combination versus the blank control (CT) group.

Results

Overall Enrichment Profile

Across the 31 evaluated herb combinations, a total of 138 unique KEGG pathways were significantly enriched (adjusted $p < 0.05$). These enriched pathways were categorized into four major functional classes (**Table S1**). Disease-related pathways constituted the vast majority of enriched processes, accounting for 1,298 instances (77.6%), followed by metabolic pathways (261 instances, 15.6%), drug action pathways (112 instances, 6.7%), and protein-related pathways (2 instances, 0.1%). The marked dominance of disease-associated pathways indicates that the metabolic perturbations induced by Wutou Tang and its

constituent combinations predominantly target networks linked to systemic metabolic disorders, reinforcing the comprehensive metabolic regulatory role of the formulation.

Table S1. Classification and Distribution of Enriched KEGG Pathways Across 31 Herb Combinations

Pathway Class	Count	Proportion
Disease; primary pathway	1,298	77.6%
Metabolic; primary pathway	261	15.6%
Drug Action; primary pathway	112	6.7%
Protein; primary pathway	2	0.1%

Core Pathways Consistently Enriched Across All Combinations

Among the evaluated drug combinations, 30 pathways showed significant enrichment in at least half of the pairs (16 out of 31), with 18 pathways perturbed in nearly all combinations (29 out of 31, 94%). These broadly affected pathways point to the primary metabolic processes influenced by the WTD formula.

Branched-chain amino acid (BCAA) degradation and amino acid metabolism pathways represented the most broadly affected category, reaching high enrichment frequencies across 29 of the 31 combinations. Specific metabolic disruptions within this system included valine, leucine, and isoleucine degradation, alongside closely related metabolic conditions such as maple syrup urine disease, isovaleric acidemia or isovaleric aciduria, 3-methylcrotonyl-CoA carboxylase deficiency type I, 3-hydroxy-3-methylglutaryl-CoA lyase deficiency, 3-hydroxyisobutyric aciduria or dehydrogenase deficiency, 2-methyl-3-hydroxybutyryl-CoA

dehydrogenase deficiency, methylmalonate semialdehyde dehydrogenase deficiency, isobutyryl-CoA dehydrogenase deficiency, beta-ketothiolase deficiency, 3-methylglutaconic aciduria (types I, III, and IV), methylmalonic aciduria, and propionic acidemia.

Glutamate- and GABA-related pathways also demonstrated wide involvement, being enriched across 27 of the 31 combinations (87%). Key pathways in this class included glutamate metabolism, D- and L-2-hydroxyglutaric aciduria, 4-hydroxybutyric aciduria or succinic semialdehyde dehydrogenase deficiency, hyperinsulinism-hyperammonemia syndrome, homocarnosinosis, and succinic semialdehyde dehydrogenase deficiency.

In addition, fatty acid metabolism pathways were affected in 24 of the 31 combinations (77%). These pathways comprised malonic aciduria, malonyl-CoA decarboxylase deficiency, propanoate metabolism, and methylmalonic aciduria related to cobalamin metabolic disorders.

The extensive enrichment of BCAA degradation and glutamate metabolism across nearly all herb combinations suggests that these metabolic processes serve as primary multi-target pathways for WTD, supporting its traditional application in treating conditions marked by metabolic dysregulation and neuro-inflammatory responses.

Pathway Enrichment Variation Across Drug Combinations

The number of significantly enriched pathways varied substantially across the 31 evaluated drug combinations, ranging from 17 to 121 pathways (**Table S2**). Among these combinations, the two-herb pairing of *Astragalus* and Licorice (AS+G, Drug 8) exhibited the highest degree

of pathway perturbation, enriching 121 pathways. High enrichment was also observed in multi-herb combinations, including *Astragalus* + Aconite + Licorice (As+Ac+G, Drug 23; 110 pathways), Peony + Licorice (P+G, Drug 10; 107 pathways), and *Astragalus* + Peony + Licorice (As+P+G, Drug 13; 95 pathways). Conversely, single-herb Licorice (G, Drug 3) and the tri-herb combination Aconite + Peony + Licorice (Ac+P+G, Drug 25) showed the lowest enrichment, each altering only 17 pathways. Remarkably, the complete formula, Wutou Tang (WTD, Drug 31), significantly perturbed a relatively modest total of 23 pathways.

These results reveal several distinct biological insights into the formula interactions. Although single-herb Licorice (Drug 3) altered only 17 pathways, its combination with *Astragalus* (Drug 8) yielded a more than sevenfold increase in pathway perturbation (121 pathways), demonstrating a marked synergistic effect between these two herbs. Notably, the maximal pathway disruption observed with AS+G far exceeded that of the full formula WTD (23 pathways). This shift suggests that while *Astragalus* and Licorice together trigger broad metabolic responses, the inclusion of additional herbal components within the full formula modulates, refines, and restricts these broad perturbations. Rather than causing widespread system disruption, the complete WTD decoction achieves a highly focused and balanced metabolic output, reflecting the traditional Jun-Chen-Zuo-Shi (sovereign-minister-assistant-courier) compatibility strategy, in which individual components harmoniously moderate one another to optimize overall efficacy.

Table S2. Representative Herbal Combinations and Their Corresponding Number of Enriched Pathways.

Drug ID	Meaning	Enriched Pathways
8	AS+G (Astragalus + Licorice)	**121**
23	As+Ac+G (Astragalus + Aconite + Licorice)	**110**
10	P+G (Peony + Licorice)	**107**
13	As+P+G (Astragalus + Peony + Licorice)	95
3	G (Licorice alone)	**17**
25	Ac+P+G (Aconite + Peony + Licorice)	**17**
31	WTD (full formula)	23

Effect of Aconite (Ac) Addition

The incorporation of Aconite into various herbal combinations altered the overall pathway enrichment profiles through distinct modes of regulation. In certain contexts, the addition of Aconite broadened pathway coverage. For instance, comparing the two-herb pair *Astragalus* + Peony (As+P, Drug 6; 34 pathways) with the three-herb combination *Astragalus* + Aconite + Peony (As+Ac+P, Drug 21; 42 pathways) revealed 14 newly enriched pathways upon Aconite addition. These newly targeted processes were heavily enriched in purine and xanthine metabolism, encompassing conditions such as adenine phosphoribosyltransferase deficiency, adenosine deaminase deficiency, general purine metabolism, and Lesch-Nyhan syndrome. Similarly, adding Aconite to the three-herb mixture *Astragalus* + Peony + *Ephedra* (As+P+E, Drug 12; 38 pathways) to yield the four-herb combination *Astragalus* + Aconite + Peony + *Ephedra* (As+Ac+P+E, Drug 27; 38 pathways) introduced 9 newly affected pathways, despite the total pathway count remaining constant.

Beyond simple expansion, Aconite demonstrated a notable modulatory effect when integrated into the complete formulation. Transitioning from the four-herb combination *Astragalus* +

Peony + *Ephedra* + Licorice (As+P+E+G, Drug 16; 33 pathways) to the full WTD formula containing Aconite (Drug 31; 23 pathways) resulted in the resolution of 10 pathways alongside the emergence of 17 distinct pathways. These newly targeted processes were predominantly involved in branched-chain amino acid and propionate metabolism. Overall, introducing Aconite to the four-herb base (As+P+E+G) effectively reoriented the systemic metabolic response from broad purine and lipid metabolism toward more focused amino acid degradation, highlighting the functional role of Aconite in directing and refining the metabolic spectrum of the Wutou Tang formula.

Drug-Drug Similarity Analysis

Jaccard similarity analysis based on shared enriched pathways delineated the hierarchical relationships among the 31 evaluated herb combinations. The highest pathway profile similarities ($J > 0.95$) were observed between specific pairs, notably single-herb Licorice (G, Drug 3) and the tri-herb combination Aconite + Peony + Licorice (Ac+P+G, Drug 25), which displayed identical pathway enrichment profiles ($J=1.0$). High profile similarities were also observed between Aconite + *Ephedra* (Ac+E, Drug 19) and *Astragalus* + Aconite + Peony + *Ephedra* (As+Ac+P+E, Drug 27; $J=0.974$), single-herb Aconite (Ac, Drug 5) and *Astragalus* + Peony (As+P, Drug 6; $J=0.971$), as well as single-herb Peony (P, Drug 4) and Peony + *Ephedra* + Licorice (P+E+G, Drug 15; $J=0.964$).

Regarding the complete formulation, WTD exhibited the highest similarity to combinations incorporating Peony (P) and Licorice (G), including single-herb Peony ($J=10.79$), Peony + *Ephedra* + Licorice ($J=10.77$), single-herb Licorice ($J=10.74$), and Aconite + Peony +

Licorice (J=10.74). These strong correlations suggest that Peony and Licorice exert primary influences in shaping the global metabolic response of WTD. In contrast, WTD showed minimal similarity to four-herb combinations lacking either Peony or Aconite, such as *Astragalus* + Aconite + *Ephedra* + Licorice (As+Ac+E+G; J=0.09), *Astragalus* + Peony + *Ephedra* + Licorice (As+P+E+G; J=0.12), and the two-herb pair *Astragalus* + *Ephedra* (As+E; J=0.28). This pronounced divergence indicates that the complete five-herb formulation generates a distinct metabolic signature that cannot be replicated by incomplete sub-combinations.

Mode-Specific Detection

Analyzing pathway distribution across ionization modes revealed that 22 pathways were identified simultaneously in both positive and negative ion modes. Meanwhile, 86 pathways were detected exclusively in positive ion mode, and 30 pathways were captured solely in negative ion mode. The marked predominance of positive-mode detection indicates that the majority of differential metabolites driving these enriched pathways undergo more efficient ionization under positive mode conditions.

Conclusion

This study systematically delineates the metabolic perturbation patterns of WTD and its various herbal sub-combinations, providing metabolomic evidence for traditional herb compatibility principles. WTD and its constituents consistently target branched-chain amino acid (BCAA) degradation and glutamate/GABA metabolic pathways across diverse herb

combinations. These robust perturbations indicate that BCAA and glutamate/GABA pathways represent the primary metabolic nodes regulated by WTD, directly aligning with its traditional clinical applications in managing pain, chronic inflammation, and neurological conditions.

Herbal interaction analysis demonstrates both marked synergy and sophisticated metabolic moderation within the formula. The sharp expansion in pathway enrichment from single-herb Licorice (17 pathways) to the Astragalus–Licorice pair (121 pathways) illustrates a potent synergistic interaction between these two herbs. Conversely, the complete five-herb WTD formula alters fewer pathways (23 pathways) than many of its smaller sub-combinations. This constrained, highly targeted response indicates that the complete formulation achieves homeostatic metabolic regulation through mutual moderation among individual component herbs, rather than inducing non-specific, widespread perturbation.

Within this regulatory network, individual herbs contribute distinct functional roles. The addition of Aconite specifically redirects the metabolic profile toward amino acid degradation while dampening broad-spectrum perturbations in purine and lipid metabolism, providing a clear mechanistic rationale for its traditional designation as a "minister" (Chen) herb.

Meanwhile, Jaccard similarity profiling reveals that WTD shares its highest pathway homology with combinations containing Peony and Licorice, establishing these two herbs as primary drivers of the core metabolic response. Overall, the hierarchical clustering of herb combinations shows that compositional similarity governs metabolic alignment, yet the full WTD formulation generates a distinct metabolic signature with low similarity to partial sub-combinations. These findings demonstrate that the complete WTD decoction operates as

an integrated system, generating emergent therapeutic properties that transcend the simple sum of its individual herbal parts.

Supplementary Notes

Detailed description of the evidence-based priority ranking system for bilingual translation.

To systematically select recommended English translations when multiple variants exist for a single TCM term, an evidence-based priority ranking algorithm was established. The algorithm assigns a cumulative score to each candidate translation based on three distinct parameters:

Authority Tier Scoring: Each source is categorized and assigned a base score reflecting its authoritative weight, where Tier 1 sources (international standards) receive 3 points, Tier 2 sources (national standards and official dictionaries) receive 2 points, and Tier 3 sources (databases and specialized glossaries) receive 1 point.

Cross-Source Frequency Scoring: For each unique English translation, the count of independent sources referencing that specific variant is added directly to its base score.

Biomedical Consistency Bonus: Candidate translations incorporating standard biomedical nomenclature—such as species-level Latin binomials for botanical drugs or IUPAC-compliant chemical names for active compounds—receive an additional bonus of 0.5 points.

The final ranking is determined by arranging total scores in descending order, with the top-ranked translation designated as the primary recommended term. In cases of identical scores, ties are resolved by prioritizing the candidate originating from the higher authority tier.

As a illustrative application, the ranking algorithm evaluated candidate translations for Sini

Tang (四逆汤). The translation Frigid Extremities Decoction achieved a total score of 8 points, derived from a Tier 1 source (WHO, 3 points), a Tier 2 source (Wiseman, 2 points), and a cross-source frequency bonus across three independent databases (3 points), establishing it as the primary recommended term. In contrast, alternative candidate translations received lower overall scores: Counterflow Cold Decoction scored 4 points (WFCMS Tier 1, 3 points; 1 source, 1 point), Sini Decoction scored 3 points (Chinese Pharmacopoeia Tier 2, 2 points; 1 source, 1 point), and Decoction for Resuscitation scored 2 points (HERB Tier 3, 1 point; 1 source, 1 point).

Community feedback and data update mechanism.

UniTCM implements a structured community contribution system to ensure sustainable data quality and growth:

Error reporting. Users can flag potential errors in any data entry (term translation, formula composition, ingredient structure, target prediction) through an in-page "Report Issue" button. Each report requires a brief description and optional supporting evidence.

Data contribution. Registered users can submit new data through standardized contribution templates available for each module (new terms for the Bilingual Corpus, new formulas for the Disease-Formula Atlas, new herb-ingredient relationships, etc.). Submissions undergo automated format validation followed by expert review.

Review and integration. Reported issues and new contributions are reviewed by the curation team on a monthly cycle. Accepted corrections are applied to the database with a versioned changelog, and contributors are credited in the platform's acknowledgment registry.

Versioning. The UniTCM database follows semantic versioning (MAJOR.MINOR.PATCH).

Major versions reflect substantial data expansion or architectural changes; minor versions reflect monthly curated updates; patches reflect individual error corrections. A complete version history is maintained and accessible through the API.

Data Availability

All data presented in this study are freely accessible through the UniTCM platform at https://unitcm.qfxulab.com. API documentation is available at https://unitcm.qfxulab.com/api/v1/docs. Companion packages are available at https://github.com/zx122ty/UniTCM_R_Package. The network pharmacology analysis was performed using the UniTCM platform, publicly accessible at https://unitcm.qfxulab.com. All herb-compound associations were retrieved from TCMKD and TCMbank databases integrated within UniTCM. Compound structures were annotated via the PubChem API (https://pubchem.ncbi.nlm.nih.gov). PPI networks were constructed using STRING v12.0 (https://string-db.org). Pathway enrichment analyses utilized Enrichr (https://maayanlab.cloud/Enrichr) with KEGG 2021 Human, GO Biological Process 2023, and WikiPathway 2021 Human gene set libraries. RA disease gene validation was performed against the MIDAS database (Metabase Integrative Disease Association Set). Metabolomics data were processed with tidymass2 [8].

References

[1] Smith CA, Want EJ, O'Maille G, et al. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. 2006;78(3):779-787.

[2] Wishart DS, Guo AC, Oler E, et al. HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 2022;50(D1):D622-D631.

[3] Thevenot EA, Roux A, Xu Y, et al. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. J. Proteome Res. 2015;14(8):3322-3335.

[4] Pang Z, Zhou G, Chong J, et al. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Res. 2024;52(W1):W398-W406.

[5] Sumner LW, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics. 2007;3(3):211-221.

[6] Xia J, Wishart DS. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr. Protoc. Bioinform. 2011;34:14.10.1-14.10.48.

[7] Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929-1935.

[8] Wang X, Liu Y, Jiang C, et al. TidyMass2: advancing LC-MS untargeted metabolomics through metabolite origin inference and metabolic feature-based functional module analysis.

Nat. Commun. 2026;17(1):1755.