Nucleic Acids Research, 2025, **53**, D1404–D1414
https://doi.org/10.1093/nar/gkae1037
Advance access publication date: 18 November 2024
Database issue

HERB 2.0: an updated database integrating clinical and experimental evidence for traditional Chinese medicine

Kai Gao[1,2,3,†], Liu Liu[1,†], Shuangshuang Lei[1,†], Zhinong Li[2], Peipei Huo[2], Zhihao Wang[2], Lei Dong[1], Wenxin Deng[1], Dechao Bu[2], Xiaoxi Zeng[4], Chun Li[5], Yi Zhao[1,2,*], Wei Zhang[4,*], Wei Wang[5,*] and Yang Wu[2,*]

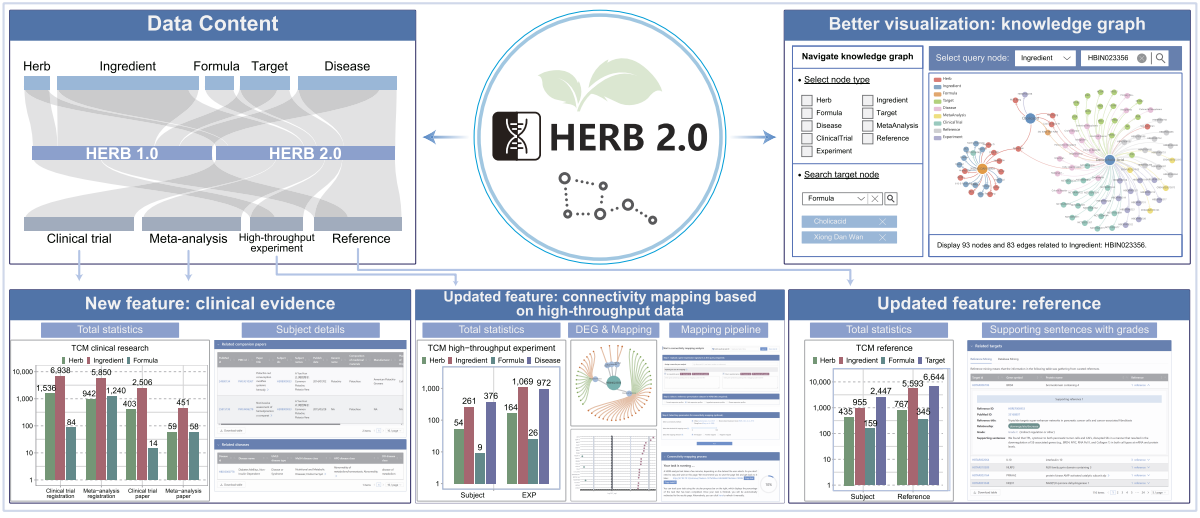
[1]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Chaoyang District, Beijing 100029, China
[2]Key Laboratory of Intelligent Information Processing, Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Respiratory and Critical Care Medicine, Ningbo No. 2 Hospital, Ningbo 315010, China
[4]West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu 610041, China
[5]State Key Laboratory of Traditional Chinese Medicine Syndrome, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

[*]To whom correspondence should be addressed. Tel: +86 10 6260 0879; Fax: +86 10 6260 0879; Email: wuyang@ict.ac.cn
Correspondence may also be addressed to Wei Wang, Tel: +86 20 3935 9999; Fax: +86 20 3935 9999; Email: wangwei26960@126.com
Correspondence may also be addressed to Wei Zhang, Tel: +86 28 8558 2944; Fax: +86 28 8558 2944; Email: weizhanghx@163.com
Correspondence may also be addressed to Yi Zhao. Tel: +86 10 6260 0879; Fax: +86 10 6260 0879; Email: biozy@ict.ac.cn
[†]The first three authors should be regarded as Joint First Authors.

Abstract

Clinical trials and meta-analyses are considered high-level medical evidence with solid credibility. However, such clinical evidence for traditional Chinese medicine (TCM) is scattered, requiring a unified entrance to navigate all available evaluations on TCM therapies under modern standards. Besides, novel experimental evidence has continuously accumulated for TCM since the publication of HERB 1.0. Therefore, we updated the HERB database to integrate four types of evidence for TCM: (i) we curated 8558 clinical trials and 8032 meta-analyses information for TCM and extracted clear clinical conclusions for 1941 clinical trials and 593 meta-analyses with companion supporting papers. (ii) we updated experimental evidence for TCM, increased the number of high-throughput experiments to 2231, and curated references to 6 644. We newly added high-throughput experiments for 376 diseases and evaluated all pairwise similarities among TCM herbs/ingredients/formulae, modern drugs and diseases. (iii) we provide an automatic analyzing interface for users to upload their gene expression profiles and map them to our curated datasets. (iv) we built knowledge graph representations of HERB entities and relationships to retrieve TCM knowledge better. In summary, HERB 2.0 represents rich data type, content, utilization, and visualization improvements to support TCM research and guide modern drug discovery. It is accessible through http://herb.ac.cn/v2 or http://47.92.70.12.

Graphical abstract



Received: September 21, 2024. **Revised:** October 15, 2024. **Editorial Decision:** October 16, 2024. **Accepted:** October 22, 2024

Introduction

Clinical trials provide efficient evaluations on the efficacy and safety of medical treatments, drugs or interventions, which have critical importance in evidence-based medicine and research (1–3). The most authoritative database documenting clinical trials, ClinicalTrials.gov, currently contains over 460 000 clinical studies submitted by researchers from more than 200 countries, reflecting the centrality of clinical research in drug discovery and therapeutic validation (4). As clinical trial data accumulates, systematic reviews and meta-analyses (called meta-analyses together hereafter) further aggregate and analyze the results of multiple independent clinical trials, offering quantitative conclusions on the reliability and precision of treatment effects (5,6). The most comprehensive database curating meta-analyses, PROSPERO, provides over 100 000 registrations from at least 118 countries (7).

Many clinical trials and meta-analyses are designed to explore traditional Chinese medicine (TCM), as TCM offers a rich resource for modern drug discovery and development. The study subjects include TCM herbs, herbal active ingredients and combinations of herbs as formulae (8–12). For example, breviscapine, a compound extracted from *Erigeron breviscapus*, has been validated through clinical trials for treating acute cerebral infarction (ACI) (13). Icaritin, a small molecule derived from *Epimedii Folium*, was approved in 2022 by the National Medical Products Administration (NMPA) to treat advanced hepatocellular carcinoma (14,15). These examples illustrate the practical application of clinical evidence in advancing TCM and modern medicine. However, such clinical evidence for TCM is scattered. A unified entrance is needed to navigate the complete landscape of TCM therapies evaluated under modern standards in one stop.

In HERB 1.0, we provide a comprehensive database containing solid experimental evidence for TCM by curating high-throughput experiments and published references. To this end, we organized the first pharmacotranscriptomics datasets of TCM and manually curated high-confident target and disease information from TCM-related references (16). We also conduct objective data-driven mapping among herbs, active ingredients and modern drugs to guide more effective therapeutics (16,17). Since the publication of HERB 1.0, these experimental evidence records have continuously accumulated. Accordingly, an effort to curate recent advances in this field is needed to update our understanding of the molecular mechanisms underlying the actions of TCM medicinal substances.

Therefore, in HERB 2.0, we comprehensively upgraded the database by integrating clinical and experimental evidence for TCM. We brought four improvements in data type, content, utilization and visualization to HERB. (i) For the first time, we carefully collected and curated 8558 clinical trials and 8032 meta-analysis information. (ii) Then we updated the original two types of experimental evidence, increasing the number of TCM-related experiments to 2231 and published references to 6644. (iii) We provide an analyzing interface for users to utilize our curated pharmacotranscriptomics datasets. (iv) We provide knowledge graph representations and visualizations of HERB entities and relationships alongside the relational databases for users. In summary, we provide a TCM-centric way to conveniently navigate 6892 herbs, 44 595 ingredients, 6743 newly added formulae, 15 515 gene targets, 30 170 diseases and their clinical and experimental evidence.

Materials and methods

Data updates

HERB 2.0 has five data components: herb, ingredient, newly added formula, gene target and disease. To ensure the data quality, we carefully refined the catalogs of herbs and ingredients in HERB 2.0 based on a range of authoritative databases to avoid redundancy (Supplementary Table S1). We then introduced a new data component: the formula, which consists of multiple herbs and represents the most common form used in TCM clinical practice (18). We compiled the TCM formulae based on the Pharmacopoeia of the People's Republic of China (2020 edition), authorized Chinese patent medicines issued by the China Food and Drug Administration (CFDA) and ancient classic formulae published by the National Administration of TCM in 2018 and 2022 (Table 1). Additionally, we cross-referenced these formulae with reputable sources such as the ETCM (19,20) and the ITCM database (21).

For targets, we updated the target list by adding new gene targets from the HIT 2.0 database (22) and our manual curation (described later). We standardized the target terminology using the Genbank database (23) and the GeneCards database (24). We cross-referenced these targets with the TTD database (25), a well-documented database about therapeutic targets for modern drugs. For disease, we updated the disease list by adding new diseases from the DisGeNET database (26) and our manual curation (described later). We standardized all disease terminologies using the DisGeNET database (26). We cross-referenced these diseases with the OMIM (27), HPO (28) and Disease Ontology (29) databases.

Clinical evidence for TCM medicinal substances

We curated clinical evidence for TCM herbs, ingredients and formulae. The first type of clinical evidence is clinical trial data collected from the *ClinicalTrials.gov* website (4), and the second type is systematic reviews and meta-analyses data from the *PROSPERO* database (7).

We searched for TCM-related clinical evidence in the ClinicalTrials.gov website and the PROSPERO database using the names and aliases of herbs, ingredients and formulae as keywords up to 1 January 2024. As a result, we obtained a considerable list of 102 151 clinical trials with particular research statuses including 'completed', 'recruiting', 'enrolling by invitation' and 'active but not recruiting', and 54 855 meta-analyses. However, most of the records in this list are false positives. Although the TCM medicinal substances appear in the intervention information, they are not *bona fide* study subjects. In most cases, they are just mentioned as the source of the actual study subject.

As a result, we made semantic judgments on the context of the intervention information to verify the relationships between TCM medicinal substances and their related clinical evidence. We first employed two cutting-edge large language models (LLMs), including *ChatGPT 3.5* (30) and *Gemini* (31). Then, we performed manual curation on the LLM-evaluated results, verified and standardized registration information such as disease information. We also evaluated the performance of LLMs using records with manual curation relationships (Supplementary Table S2). Finally, we obtained a list of 8558 clinical trials and 8032 meta-analyses.

Furthermore, we curated clinical conclusions for these clinical trials/meta-analyses by searching their companion supporting papers in the PubMed database. It is worth noting

Table 1. Overview of the data contents curated in HERB 2.0

Components	Amount in HERB 1.0	Amount in HERB 2.0	Data source
Herbs	7263	6892	De-redundancy based on multiple databases shown in Supplementary Table S1
Ingredients	49 258	44 595	De-redundancy based on multiple databases and methods shown in Supplementary Table S1
Formulae	/	6743	Novel component based on the pharmacopoeia of the People's Republic of China 2020, authorized Chinese patent medicines issued by the China Food and Drug Administration (CFDA), and ancient classic formulae released by the National Administration of TCM
Targets	12 933	15 515	Updated by reference curation and HIT 2.0, and cross-referenced with the GenBank, GeneCards, and TTD databases
Diseases	28 212	30 170	Updated by reference curation and cross-referenced with the DisGeNET, OMIM, HPO, Disease Ontology databases
Clinical trials	/	8558	Novel data for TCM curated from the ClinicalTrials.gov and PubMed databases
Meta-analyses	/	8032	Novel data for TCM curated from the PROSPERO and PubMed databases
High-throughput experiments	1037	2231	Updated by searching and re-analyzed the GEO database and by recruiting the CREEDS database
References	1966	6644	Updated by searching and manually curating from the PubMed database

that the herb/formula's therapeutic effect is highly influenced by its source and quality control method. Therefore, we extracted as many details about TCM subjects as possible in these papers. For herbs, we extracted information about their source, place of origin, processing method, *etc*. For formulae, we extracted information about their manufacturer, composition, quality evaluation method, *etc*.

High-throughput experiments for TCM medicinal substances and diseases

We updated high-throughput experiments for herbs, ingredients, and newly-added formulae from the GEO database (32). Moreover, we added high-throughput experiments for diseases from the CREEDS database (33) and the GEO database. These experiments include RNA-seq and microarray data in *Homo sapiens* and *Mus musculus* from 2001 to 2024. All data processing methods related to high-throughput experiments are the same as HERB 1.0 (Supplementary Figure S1A).

Firstly, all samples with unique GSM numbers were re-analyzed independently following the same unified pipeline as HERB 1.0 (16). For RNA-seq data, we used *Fastp* (34) to filter raw reads, *STAR* (35) to align the filtered reads to the genome, *featureCounts* (36) for expression quantification, and *Limma* (37) for expression normalization. We discarded samples with a mapping rate lower than 60% to ensure data quality. For microarray data, we extracted and normalized probe expression profiles from raw data using *Limma* or *Oligo* (38) for the Agilent or Affymetrix platform, and we used the normalized probe expression profiles directly for other platforms. Then, we transformed probe profiles into the normalized gene expression matrix for all array data using our custom script *probe2gene*, which embeds conversion files for 105 platforms and is downloadable on the HERB database.

Next, we conducted differential expression analysis (DEG) and functional enrichment analysis: (i) we manually dissected the relationships among samples (GSMs) in each GEO experiment (GSE) (Supplementary Figure S1B). We defined each HERB experiment (EXP) as a set of control and treatment samples in the same GSE related to a particular herb, ingredient, formula or disease. We required at least two biological replicates of control and treatment samples ($\geq$2 GSMs) in each EXP (Supplementary Figure S1C). (ii) we performed DEG analysis individually for each EXP using *Limma*. We selected genes with |log2(fold change)|$\geq$0.5 and $P \leq 0.05$ as differentially expressed genes. (iii) we performed GO and KEGG enrichment analysis for each EXP based on the DEG gene list using *clusterProfiler* (39). Enriched GO terms and KEGG pathways were selected when $P \leq 0.05$. (iv) we measured the EXP-level similarity of the DEG analysis by Pearson correlation coefficient (*PCC*) (Supplementary Figure S1D).

Then, we merged the analyzed results for each herb, ingredient, formula, or disease with multiple EXPs. We first transformed the initial two-tailed *P-value* for each gene in an EXP to *P/2* and *1-P/2* as two one-tailed *P-values* for up-regulation and down-regulation. Then, we merged two unified probabilities, *P-up* and *P-down*, for each gene across multiple EXPs using Fisher's method (40). Genes with only one significant P-value were retained in the DEG list, i.e. *P-up* < 0.05 or *P-down* < 0.05. We also required an average |log2(fold change)|$\geq$0.5 based on those EXPs with significant differences in the individual tests. Based on the merged DEG gene list, we performed additional GO and KEGG enrichment analyses for each herb, ingredient, formula and disease. Note that EXPs performed in *Homo sapiens* or *Mus musculus* were merged separately.

Data-driven connectivity mapping pipeline

Based on the re-analyzed gene expression profiles in HERB, we further evaluated all pairwise similarities among TCM medicinal substances, diseases and 2837 modern drugs with CMap gene expression profiles. Firstly, we performed connectivity mappings at the EXP level using the CMap (17,41) website (https://clue.io/query) when mapped with CMap purturbagens and the CMap custom script when mapped with others (Supplementary Figure S2). For each EXP about herbs, ingredients, and formulae, we mapped its DEG profile to that of CMap perturbagens and HERB-curated diseases, respectively. For each EXP about diseases, we mapped its DEG profile to that of CMap perturbagens and all HERB-curated TCM medicinal substances (Supplementary Figure S3A). We then merged the connectivity mapping results for each herb, ingredient, formula, and disease across their multiple EXPs fol-

lowing the CMap (17) method of maximum quantile statistic, which is the same as HERB 1.0 (Supplementary Figure S3B).

$$Summary\ \mathrm{score}_{s,d} = \begin{cases} Q_{\mathrm{high}}(score_{e,s,d}) \\ \quad \text{if} |Q_{\mathrm{high}}(score_{e,s,d})| \geq |Q_{\mathrm{low}}(score_{e,s,d})| \\ Q_{\mathrm{low}}(score_{e,s,d}) \quad otherwise \end{cases}$$

where the final *summary* $score_{s,d}$ represents the overall similarity measure between a query subject s in HERB and a DB instance d, and the $score_{e,s,d}$ indicates the individual similarity measure in each EXP e related to the query subject s and the DB instance d.

Furthermore, we provided a convenient analyzing interface for users to efficiently utilize our curated pharmacotranscriptomics datasets for drugs and diseases in a batch mode. Users can conduct connectivity mapping analysis using their uploaded gene expression profiles to that for TCM herbs, ingredients, formulae and diseases in the backend DBs. We provided three methods for users to select, including the classic CMap (17) method, the RGES (42) method adapted from CMap and the ZhangScore (43) method, to expand the utility of HERB and support researchers' customized requests.

Manually curated references in HERB 2.0

We updated curated references for TCM medicinal substances from the PubMed (44) database. For herbs and ingredients, we added novel related references since 2020. For newly added formulae, we searched all available PubMed references since 2000. Based on these references, we extracted target and disease information for herbs, ingredients and formulae using the Python natural language processing package '*Stanza*' (45), followed by manual curation.

After extraction, we provide the TCM–target and TCM–disease relationships and supporting sentences for these relationships. Moreover, we stratified the curated TCM targets into three grades according to their regulatory roles (Supplementary Figure S4). Grade A represents the highest reliability, with keywords about direct interactions like activation, inhibition, targeting, or binding; Grade B is related to indirect experimental evidence such as gene knockout, RNA interference, overexpression and affinity measurements; and Grade C covers other weak regulation evidence. We also retrieved supporting sentences and performed this stratification for curated references in HERB 1.0 to make them compatible with HERB 2.0.

Indirect associations among HERB components

After comprehensively upgrading data and evidence, we updated indirect associations among HERB components using Fisher's exact test followed by BH multiple test correction (46). We updated six types of indirect associations in total. For example, the indirect relationship between ingredient and disease can be obtained using the gene targets as the middle component (Supplementary Figure S5A), with reliable associations ($FDR < 0.01$) selected as the final statistical inference set. We also used this strategy to infer the herb–target and herb–disease relationships by using ingredients and targets as the middle components, respectively (Supplementary Figures S5B-C). Moreover, we inferred the formula–ingredient, formula–target and formula–disease relationships by using herbs, herbs, and gene targets as the middle components (Supplementary Figures S5D-F).

The knowledge graph representations of HERB data

In addition to relational databases, we provide knowledge graph representations and visualizations of HERB data to facilitate better retrieval of TCM knowledge. The knowledge graph consists of nine types of entities. The first five are basic components in HERB, including herbs, ingredients, formulae, targets and diseases. The other four are clinical and experimental evidence, including clinical trials, meta-analyses, high-throughput experiments and curated references (Supplementary Figures S6A).

We curated the relationships among these entities by merging multiple sources of evidence (Supplementary Figures S6B). To this end, we designed a scoring system to rank these merged relationships and provided weights for each relationship in the graph. This system has four rules: (i) The weights for built-in relationships are assigned as 1, as they are sufficiently reliable, for example, the formula-herb relationship. (ii) The weights for relationships recorded in authorized databases are assigned as 0.15–0.30, depending on single or multiple databases supported. (iii) The weights for relationships supported by clinical or experimental evidence curated in HERB are assigned as 0.10–0.15, depending on single or multiple instances of evidence supported. (iv) The weights for relationships inferred statistically are assigned as 0.05–0.10, depending on which significance level it reached. Taken together, the weights for built-in relationships are 1 according to the first rule, and the weights for other relationships range from 0 to 1 as they are the summation of the other three rules. These efforts allow us to prioritize the visualization of the most critical links in the knowledge graph.

We constructed the knowledge graph based on the Neo4j graph data platform (47). We used Neo4j's Python API for custom querying and the JavaScript-based framework Echarts (48) for graphical display. We provided the HERB-based knowledge graph on a stand-alone web page, which can be queried user-friendly using buttons or drop-down menus. Moreover, we integrated the knowledge graph with the HERB system, offering seamless cross-referencing between the graph and the detailed pages of five basic components and four evidence types.

Results

Data contents curated in HERB 2.0

The HERB 2.0 contains a comprehensive list of 6892 herbs, 44 595 ingredients, 6743 newly added formulae, 15 515 gene targets and 30 170 diseases (Table 1). We first refined the catalogs of herbs and ingredients to ensure high data quality. Then, we added a new data component, the formula, as it represents the most common form used in TCM clinical practice. Thirdly, we added new targets and diseases based on database mining and our manual curation. We also provided the statistics of all pairwise relationships among these five data components (Supplementary Table S3).

Subsequently, we systematically integrated clinical and experimental evidence for these five data components. For the first time, we carefully collected and curated 8558 TCM-related clinical trials and 8 032 meta-analyses information, enabling the HERB database to be guided by clinical evidence. We also updated the experimental evidence, including 2231 high-throughput experiments and 6 644 curated references. In this way, we aligned the TCM data in HERB 2.0

to a pyramid of evidence consisting of four types of high-quality data (Figure 1A). As a result, HERB-2.0 provides the first complete landscape of TCM therapies evaluated under modern standards and offers data-driven mapping results that can directly guide drug repositioning for TCM therapies and modern drugs. Compared with other TCM databases, HERB 2.0 collates the most comprehensive and high-quality evidence for TCM and provides more analytical functions (Supplementary Table S4).

Clinical evidence for TCM curated in HERB 2.0

For the first time, we curated a complete list of TCM-centric clinical evidence by searching for TCM-related records on the ClinicalTrials.gov website and the PROSPERO database. The initial list is extensive, including 102 151 clinical trials and 54 855 meta-analyses. However, most of these records are false positives that should be filtered. For example, in the 115 clinical trials related to '*ginseng*' satisfied time and research status cutoffs, 44 entries did not use the herb '*ginseng*' as their study subject (Supplementary Table S5). Although the word '*ginseng*' appears in the intervention information, the actual study subjects are often ginseng-related ingredients or formula, which we did not retain for the herb '*ginseng*'.

We used two cutting-edge LLMs, including *ChatGPT 3.5* and *Gemini*, to filter false positive records, followed by manual curation. To evaluate the performance of LLM models, we first manually determined 2438 clinical trials and 7884 meta-analyses whether each record should be retained or filtered. Then, we automatically checked these records by LLMs and used a cautious approach to filter records assessed as 'negative' by both models (Supplementary Figure S7A). We found that the total accuracy of LLMs in judging clinical evidence records is 51.3% (Supplementary Table S2). Among the wrong cases, only 1.4% were false negatives and 98.6% were false positives. As a result, we directly accepted the filtered set assessed as 'negative' by both LLMs and manually curated other records assessed as 'positive' by either LLM in order to reduce the workload. We should note that some false negatives generated by LLM judgments will be missing entries. We will gradually add these entries according to users' feedback, as the HERB database is widely used.

In total, HERB-2.0 provides curated information for 8558 clinical trials related to 249 herbs, 375 herbal ingredients and 59 formulae consisting of multiple herbs (Figure 1B), as well as 8032 registered meta-analyses covering 267 herbs, 399 ingredients, and 336 formulae (Figure 1C). These trials span interventions across 556 diseases, with registrations from 119 countries worldwide (Supplementary Figure S7B). Notably, TCM-related interventions cover all phases of clinical trials, with 12.5% in Phase 1, 19.6% in Phase 2, 12.6% in Phase 3 and 11.3% in Phase 4 (Supplementary Figure S7C). Around 3964 TCM-related clinical trials followed the strict randomized, double-blinded (or even more stringent, triple- or quadruple-blinded) principles (Supplementary Figures S7D-E). These meta-analyses span interventions across 485 diseases, with registrations from 81 countries (Supplementary Figure S7F). Moreover, we searched for 2903 and 563 companion supporting papers for 1 941 (22.7%) clinical trials and 593 (7.4%) meta-analyses. We curated clinical conclusions for these clinical trials/meta-analyses by reading their companion supporting papers. These high-quality clinical trials with clear conclusions involve 96 herbs, 165 ingredients and 13 formulae. These high-quality meta-analyses with clear conclusions involve 40 herbs, 129 ingredients and 61 formulae.

It is worth noting that the herb/formula's therapeutic effect is highly influenced by its source and quality control method. Therefore, we extracted and presented as many details about TCM subjects as possible in these papers. For herbs and formulae, we extracted information about their source, place of origin, manufacturer, composition, processing method, quality evaluation method, etc. (Supplementary Figure S8). We found that 83/28 herbs had related clinical trials/meta-analyses with detailed information like source, place of origin and processing methods. 13/46 formulae had related clinical trials/meta-analyses with detailed information like manufacturer, composition, and quality evaluation method. This holistic approach ensures that the contexts and outcomes of clinical evidence are thoroughly curated for researchers' convenience. We illustrate an example of clinical evidence for the herb, *Folium Camelliae Sinensis*, in Figure 2A–B.

Experimental evidence for TCM analyzed in HERB 2.0

In HERB 2.0, we substantially enhanced the experimental evidence for TCM. By integrating the explosion of this field in recent years, we increased the number of TCM-related high-throughput experiments from 1037 to 2231 (Table 1). We collected these data from 1206 GSEs (GSE) comprising 22 560 GEO samples (GSM). This amount of data is the most comprehensive among all TCM databases, with the number of sequenced samples (22 560) re-analyzed in HERB is much more extensive than that in the other database, ITCM (1488) (Supplementary Table S4).

We manually transformed these high-throughput data into 2 231 HERB experiments (EXP). Each EXP is related to a particular herb, ingredient, formula or disease and contains a set of control and treatment GEO samples, all performed within the same GSE (Supplementary Figure S1A). They cover 54 herbs, 261 ingredients, 9 formulae and 376 diseases (Figure 1D), with 77.9% of them consisting of microarray data and the remaining 22.1% as RNA-seq data (Supplementary Figure S1B). We re-analyzed the data for each HERB EXP by building automatic and stringent pipelines, the same as in HERB 1.0. Firstly, we required at least two biological replicates of control and treatment samples ($\geq$2 GSMs) in each EXP (Supplementary Figure S1C). The average number of biological replicates for control and treatment samples in HERB-EXPs were 5.7 and 6.4, respectively. Secondly, we discarded samples with insufficient mapping rates. Third, we normalized the gene expression profiles for both RNA-seq and microarray data to remove the batch effects. Finally, we evaluated the similarities of the gene expression levels among different EXPs by *PCC* for the same herb, ingredient, formula and disease. We found that 55.2% of EXP pairs have a *PCC* larger than 0.5 (Supplementary Figure S1D), reflecting that the independently collected data and our re-analyzing pipeline are re-producible. In summary, there were an average of 1631 DEGs upon EXP-level analysis and 2055 DEGs after merging the results from multiple EXPs for herbs, ingredients, formulae, and diseases. For GO terms, the average numbers were 863 and 632 for the EXP-level and merge-level results. For KEGG pathways, the two numbers were 39 and 32, respectively. These data statistics are coherent with that

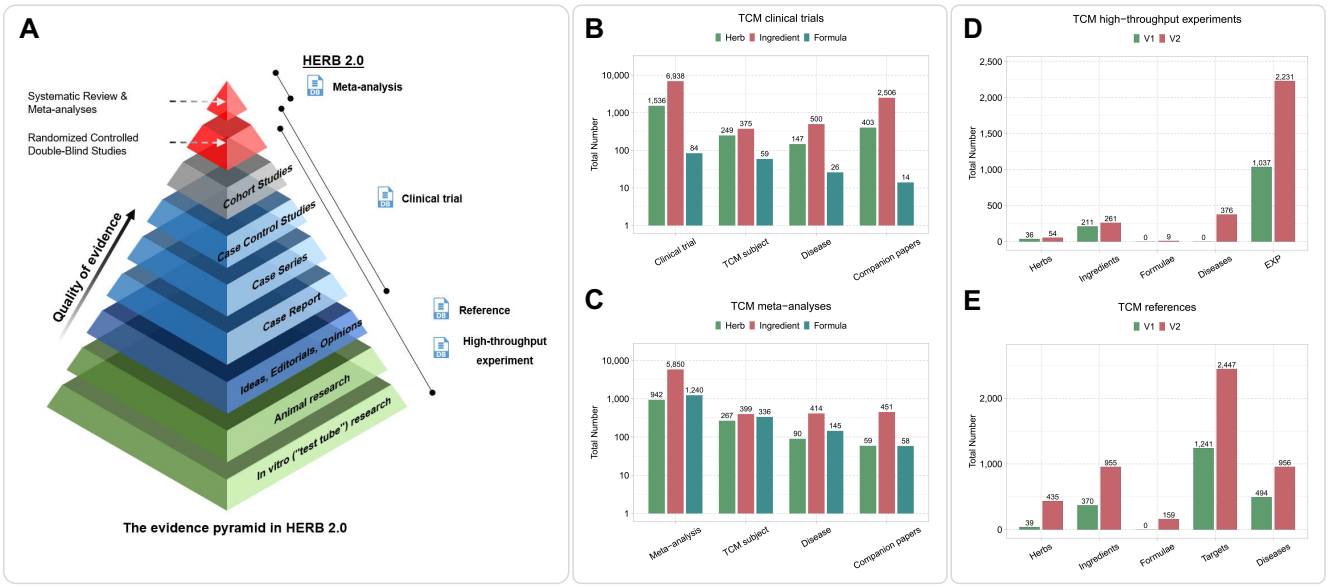


Figure 1. Illustration of clinical and experimental evidence for TCM. (**A**) The classical medical evidence pyramid model illustrates multiple levels of evidence credibility. We aligned four types of evidence in HERB 2.0 with this model. (**B**) Statistics of curated clinical trials for TCM in HERB 2.0. The number of clinical trials, TCM subjects, related diseases and companion supporting papers are shown in bar plots, with the three colors indicating herb, ingredient and formula. (**C**) Statistics of curated meta-analyses for TCM in HERB 2.0. The bar plots were arranged similar to Figure 1B. (**D**) Statistics of high-throughput experiments for TCM in HERB 2.0. The number of herbs, ingredients, and diseases with such experiments were shown in bar plots illustrating the difference between HERB 1.0 and 2.0. The number of HERB EXP, one for an HERB control-treatment comparison, was also shown. (**E**) Statistics of curated references for TCM in HERB 2.0. The differences between herbs, ingredients, formulae, curated targets and diseases were all shown.

in HERB 1.0. We illustrate an example of a high-throughput experiment for the herb, *Folium Camelliae Sinensis*, in Figure 2C.

We updated the curated references to 6644 (Table 1). Based on these references, we extracted 2447 targets and 956 diseases for 435 herbs, 955 ingredients and 159 formulae (Figure 1E). We provided detailed information about the supporting sentences for these TCM–target and TCM–disease relationships on the HERB website for the user's convenience. We also stratified the curated TCM targets into three grades according to their level of confidence (Supplementary Figure S4A), where Grade A is the most confident, and Grade C has the minor support. 3.4, 2.4 and 94.2% of TCM-target relationships are classified as Grade A, B and C, respectively (Supplementary Figure S4B). Besides, we carefully classified the TCM-related diseases into different classes according to the DisGeNET classification system. For example, 47.8% and 29.0% of TCM-related diseases are classified as disease or syndrome, neoplastic process, respectively (Supplementary Figure S4C). In summary, we increased the number of curated references and improved the data quality for this type of evidence. We illustrate an example of curated reference for the herb, *Folium Camelliae Sinensis*, in Figure 2D.

TCM-centric connectivity mapping pipeline in HERB 2.0

Utilizing the expanded high-throughput datasets, we extensively evaluated all pairwise similarities among herbs, ingredients, formulae, modern drugs and diseases. We measured these similarities using the CMap connectivity mapping score (17), which is a statistical framework to quantify the relationship between a query q and a DB EXP, r_i, based on the Kolmogorov–Smirnov enrichment statistic, and then summarize those scores across all EXPs for this DB subject r for robustness (Supplementary Figure S2). We performed EXP-level mapping first (Supplementary Figure S3A) and then merged them into the subject-level mapping results (Supplementary Figure S3B). We found that 19 herbs, 145 ingredients, 4 formulae and 286 diseases could be mapped to 2205 CMap compounds using a cutoff of absolute connectivity score above 80. 20 herbs, 183 ingredients and 2 formulae could be mapped to 132 diseases using a cutoff of absolute connectivity score above 80. The mapping results can directly inform drug repositioning efforts for TCM and modern therapeutics, significantly broadening the scope of HERB 2.0.

Our omics data-based CMap analysis can provide data-driven support for reusable TCM clinical practice experiences. For example, compound kushen injection (CKI) is a TCM formula approved by the NMPA in China and extracted through standardized Good Manufacturing Processes (49). CKI can effectively relieve cancer pain and bleeding and is widely used in the adjuvant treatment of many cancer types (50,51). In the CKI's CMap mapping results on the HERB 2.0 website, we found three cancer-related diseases out of the 10 top mapping hits, including malignant neoplasm of the pancreas, malignant neoplasm of the breast and cancer pain. We also conduct a reciprocal mapping analysis among different types of TCM medicinal substances by requiring mapping scores ≥ 75, and then annotate the results based on prior knowledge (Supplementary Table S6). We found that the HERB knowledge graph can validate 30.2% of these mapped pairs by requiring no more than one middle component between the query and target items. For example, in several studies, tomato and its main active ingredient, lycopene, have shown similar health benefits (52–54). Consistent with this, the tomato-to-lycopene and lycopene-to-tomato mapping scores are 77.0 and 79.8, respectively.

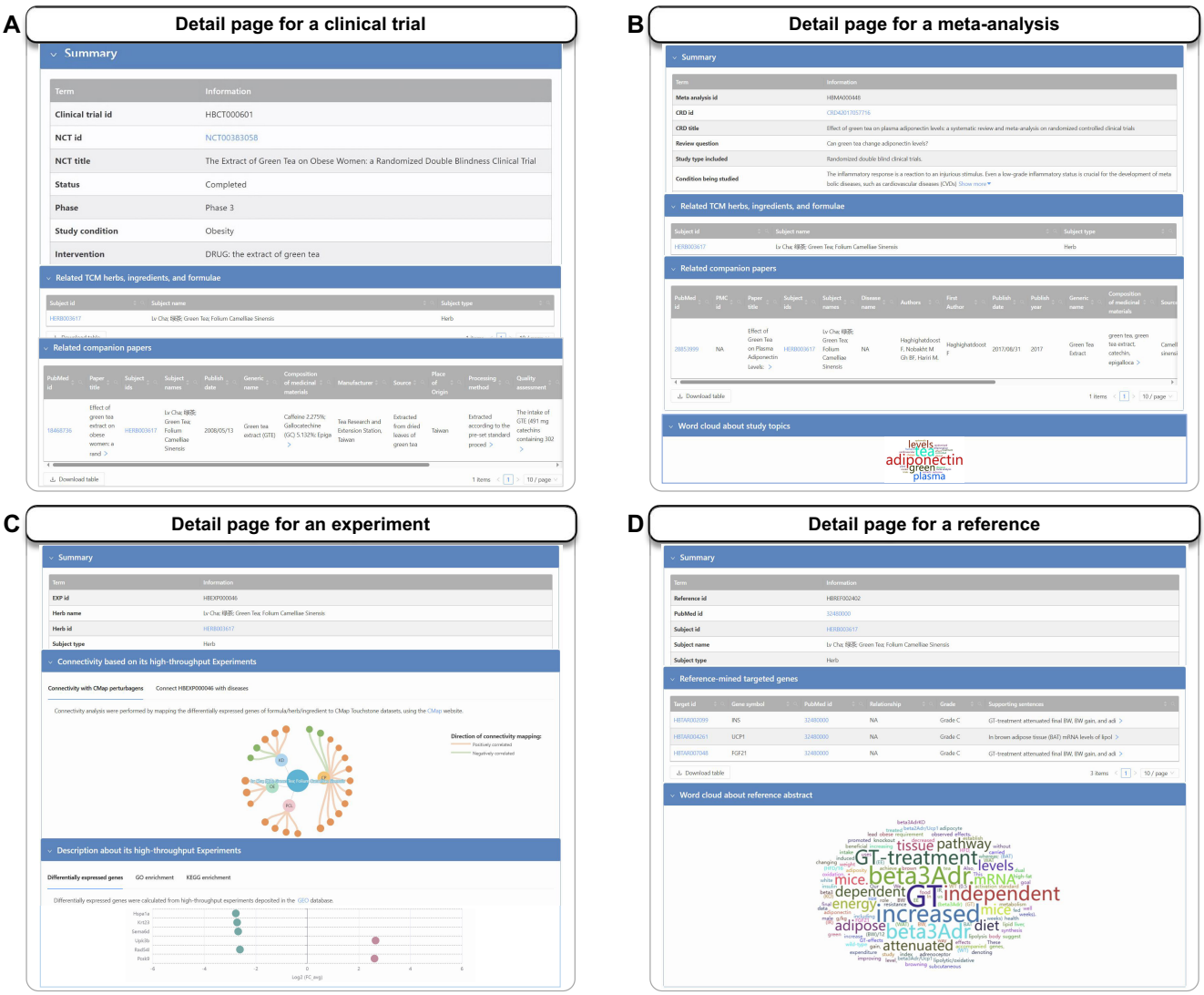


Figure 2. Demonstration of four types of evidence curated in HERB 2.0 for an herb, *Folium Camelliae Sinensis*. (**A**) Detail page for a clinical trial. (**B**) Detail page for a meta-analysis. (**C**) Detail page for a high-throughput experiment. (**D**) Detail page for a curated reference.

We visualized the connectivity mapping results on the HERB 2.0 website. The EXP-level mapping results are displayed on the data detail page for each HERB EXP, and the subject-level mapping results are displayed on the detail page for each herb, ingredient, formula and disease with high-throughput data. For example, on the detail page for the ingredient deoxycholic acid, its connectivity mapping result consists of two sections: one maps it with diseases, and the other maps it with CMap perturbagens (Figure 3A). Each section visualized the mapping results in a circled plot with the ingredient itself centered in the plot. The middle layer of the plot visualizes the classification of diseases or CMap perturbagens mapped to this ingredient. The third layer shows each disease category's mapped targets for this ingredient. The mapped targets are ordered by mapping score and colored by mapping direction. This allows the users to focus on the most robust mapping targets quickly.

Furthermore, we provide a convenient analyzing interface for users to efficiently utilize our curated pharmacotranscriptomics datasets for herbs, ingredients, formulae, and diseases. This interface is provided as an 'analyze' page on the HERB website (Figure 3B), which is similar to that on the CMap website (https://clue.io/) (17) but explicitly tailored for TCM. Users can map their gene expression profiles on this page against our curated datasets in three steps.

In step 1, users must prepare a gene expression signature, including upregulated and downregulated gene sets, following the same format as the example files. The signatures should be ranked gene lists, with the most critical genes at the top. The two gene lists are used as a query, q, for connectivity mapping analysis.

In step 2, users can select a reference perturbation dataset for connectivity mapping analysis. We have provided curated datasets for 22 herbs, 198 ingredients, 4 formulae and 308 diseases with human EXP datasets as backend DBs. For example, if the user chooses the ingredient dataset, the query gene lists, q, are mapped to each DB EXP, r_i, for every DB subject r (an ingredient in this case) in the dataset. Finally, a summary connectivity map score for the query and ingredient pair, $S_{q,\,r}$, was computed across all EXPs for this ingredient.

In step 3, users can adjust connectivity mapping methods and parameters. For example, users can choose from

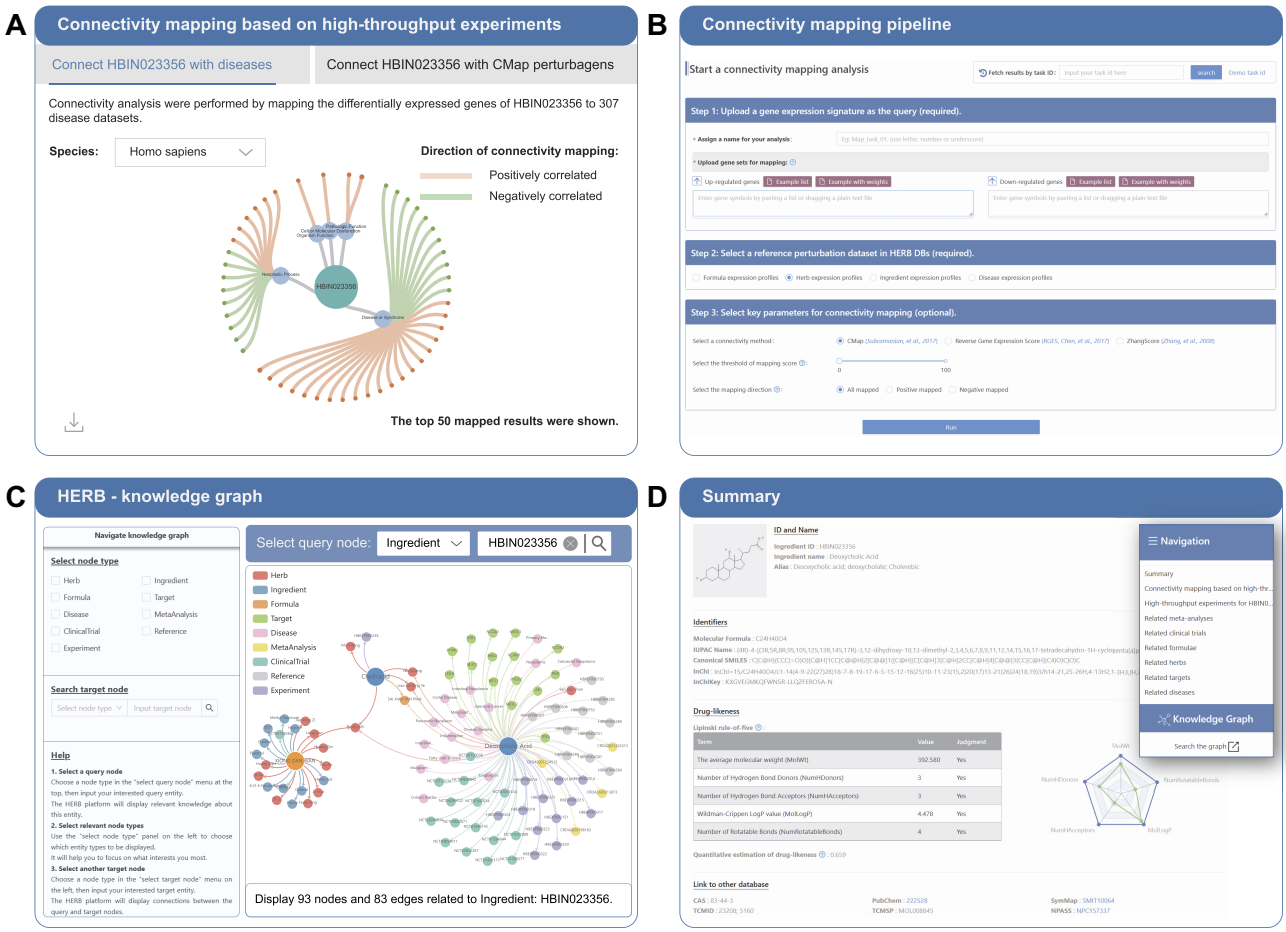


Figure 3. Illustration of new features in HERB website. (**A**) Connectivity mapping results for an ingredient on its detailed page are shown. Such subject-level connectivity mapping results were visualized in the details pages of all herbs, ingredients, formulae and diseases with high-throughput data. Besides, EXP-level connectivity mapping results were shown similarly on each data detail page. (**B**) The connectivity mapping pipeline for users to submit their gene expression profiles is shown. Titles represent three steps of using this pipeline. (**C**) Illustration of the stand-alone knowledge graph page in HERB 2.0. Users can query the knowledge graph friendly using buttons or drop-down menus. (**D**) A navigation bar for cross-referencing with the HERB knowledge graph is shown on the top-right of the detailed page for each data component and each evidence record.

the standard CMap (17), the adjusted RGES (42), or the ZhangScore (43) methods. The CMap method is the default (Supplementary Figure S2). Users can fine-tune the mapping threshold to control which parts of the mapping results to display. Users can also choose to map in similar, opposite or both directions depending on their research focus.

The automatic computation takes about 7.3 ± 0.6 min on average. We provide a circular progress bar when the task is running, which displays the percentage of the task that has been completed dynamically. After the mapping task is finished, the mapping results will be displayed similarly to the detailed pages for each search term in HERB (Figure 3A). This new feature enables the users to integrate their data with our curated datasets seamlessly. It expands the utility of HERB and supports researchers' customized requests for innovative discoveries in the TCM field.

Evidence-based TCM knowledge graph in HERB 2.0

We integrated all data components and all evidence in HERB into knowledge graph representations to better retrieve TCM knowledge. The HERB knowledge graph contains nine types of entities (Supplementary Figure S6A), including herbs, ingredients, formulae, gene targets, diseases, clinical trials, meta-analyses, high-throughput experiments and curated references, and the number of each entity type is the same as in Table 1. Theoretically, nine entity types may consist of 36 types of relationships. However, no relationships are established between the four types of evidence, and the component target is not connected to the two types of clinical evidence. As a result, the HERB knowledge graph contains 28 types of relationships, with the number for each relation type shown in Supplementary Figure S6B.

Among them, 18 types of relationships are built-in, which means the relationship is clear enough. For example, the relationship between the research subject (herb, ingredient, or formula) for a clinical trial is definite. The other ten types of relationships are not built-in but supported by different evidence with different strengths. For example, the relationship between an herb and a gene target may be curated in a reference or selected as a DEG in a high-throughput experiment. In such cases, we curated the relationships among these entities by merging multiple sources of evidence. We further designed a scoring system to prioritize the most critical relationships in the knowledge graph (Supplementary Figure S6B). As a result, all relationships in the HERB knowledge graph have weights

that range from 0 to 1. We further showed the distributions of the weights across all types of non-built-in relationships (Supplementary Figure S6C). We found that most (99.9%) of non-built-in relationships are supported by only one type of evidence, leaving another 3702 (0.1%) relationships supported by at least two. Among the relationships supported by a type of evidence, 51 195 (1.9%) have multiple instances of evidence.

We constructed the HERB knowledge graph using the *Neo4j* graph data platform (http://www.neo4j.org). We used Neo4j's Python API for custom querying and the JavaScript-based framework Echarts (48) for graphical display. As a result, we provided the HERB-based knowledge graph on a stand-alone 'knowledge graph' page (Figure 3C), which can be queried user-friendly using buttons or drop-down menus. Users can use the 'select query node' menu at the top of this page to input their query entity of interest. The HERB platform will then display relevant knowledge about this entity. Users can refine their search by using the 'select node type' panel on the top-left to choose which relevant entity types to display. Then, users can select another target node using the 'select target node' menu on the middle-left, and the HERB platform will display connections between the query and target nodes.

Moreover, we integrated the knowledge graph with the detailed page for each data component and each evidence record in the HERB database. For example, on the detailed page of the ingredient deoxycholic acid, we displayed a navigation bar on the right and highlighted the 'search the graph' button at its bottom (Figure 3D). This hyperlink will redirect the users to the ingredient's 'knowledge graph' page, enabling quick cross-linking between its knowledge graph page and its detailed page.

Discussion

The clinical practice of TCM across millennia provides valuable therapeutic candidates for modern drug discovery. As a result, researchers built several TCM databases to fully utilize these valuable TCM experiences and bridge the gap between TCM and modern technologies (Supplementary Table S4). However, the clinical trials and meta-analyses at the tip of the medical evidence pyramid have never been introduced. In HERB 2.0, we filled this gap by adding new solid links to clinical and experimental evidence for TCM, enabling the HERB database to be guided by a pyramid of evidence consisting of four types of high-quality data. HERB-2.0 provides the first complete landscape of TCM therapies evaluated under modern standards, evaluates all data-driven pairwise similarities among TCM herbs/ingredients/formulae, modern drugs, and diseases, and presents knowledge graph representations of all data components and all types of evidence for TCM for users' convenience. It represents rich improvements in data type, content, utilization, and visualization.

When upgrading HERB, we dealt with a vast amount of data, including 102 151 clinical trials and 54 855 meta-analyses for semantic judgments, 22 560 GEO samples with high-throughput data for re-analyzing, and 27 286 references for information extraction. Since the breakthrough of GPT-4 last year, AI-based LLMs have demonstrated impressive semantic understanding capabilities. A new technological revolution that combines LLMs with various vertical fields is rapidly unfolding. We thus utilized and evaluated cutting-edge LLMs, including ChatGPT 3.5 and Gemini, for data processing. According to our cautious pipeline, we found that the total accuracy of these LLMs in judging clinical evidence records is 51.3%, reflecting that current LLMs did not capture enough domain knowledge about TCM. If we want to elevate the depth and accuracy of LLMs in the TCM field, we need high-quality and TCM-specific datasets to fine-tune the intelligent models. Our HERB 2.0 platform goes as far as possible to curate evidence-based, high-quality, and comprehensive data that meet this need at the right time. We should note that there are false negatives in LLM judgement, and these records would be missing entries in HERB 2.0. We will continuously upgrade the HERB database according to users' feedback and integrate it with the latest artificial intelligence technologies to support TCM research and guide modern drug discovery.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Prof. Shunmin He, Lei Kong and Liang Sun for their kindly helps on high-performance computing. We thank Dr. Min Li and Fengzhen Chen for their helps on herb name curations. We thank 36 volunteers for their helps in manual curation (Supplementary Table S7).).

Funding

National Key R&D Program of China [2021YFC2500203]; National Natural Science Foundation of China [32070670, 92474204, 32341019]; Beijing Natural Science Foundation [L222007]; Ningbo major project for high-level medical and healthcare teams [2023030615, 2024020919]; Ningbo Science and Technology Innovation Yongjiang 2035 Project [2024Z229]; Major Project of Guangzhou National Laboratory [GZNL2023A03001]. Funding for open access charge: National Key R&D Program of China.

Conflict of interest statement

None declared.

References

1. Huang,Y., Xiong,W., Zhao,J., Li,W., Ma,L. and Wu,H. (2023) Early phase clinical trial played a critical role in the Food and Drug Administration-approved indications for targeted anticancer drugs: a cross-sectional study from 2012 to 2021. *J. Clin. Epidemiol.*, **157**, 74–82.
2. Mitra-Majumdar,M., Gunter,S.J., Kesselheim,A.S., Brown,B.L., Joyce,K.W., Ross,M., Pham,C., Avorn,J. and Darrow,J.J. (2022) Analysis of supportive evidence for US Food and Drug Administration approvals of novel drugs in 2020. *JAMA Netw. Open*, **5**, e2212454.
3. Yang,J., Yang,J. and Hu,Y.J. (2024) Characteristics of clinical trials of new oncology drugs approved in China. *Cancer*, **130**, 671–682.
4. Zarin,D.A., Tse,T., Williams,R.J., Califf,R.M. and Ide,N.C. (2011) The ClinicalTrials.gov results database–update and key issues. *N. Engl. J. Med.*, **364**, 852–860.
5. Siddaway,A.P., Wood,A.M. and Hedges,L.V. (2019) How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.*, **70**, 747–770.

6. Knoll,T., Omar,M.I., Maclennan,S., HernÃ¡ndez,V., Canfield,S., Yuan,Y., Bruins,M., Marconi,L., Van Poppel,H., N'Dow,J, *et al.*2018) Key steps in conducting systematic reviews for underpinning clinical practice guidelines: methodology of the European Association of Urology. *Eur. Urol.*, **73**, 290–300.
7. Schiavo,J.H. (2019) PROSPERO: an International Register of Systematic Review Protocols. *Med. Ref. Serv. Q.*, **38**, 171–180.
8. Hong,J.T., Lee,M.J., Yoon,S.J., Shin,S.P., Bang,C.S., Baik,G.H., Kim,D.J., Youn,G.S., Shin,M.J., Ham,Y.L., *et al.* (2021) Effect of Korea red ginseng on nonalcoholic fatty liver disease: an association of gut microbiota with liver function. *J. Ginseng. Res.*, **45**, 316–324.
9. Iturrino,J., Camilleri,M., Wong,B.S., Linker Nord,S.J., Burton,D. and Zinsmeister,A.R. (2013) Randomised clinical trial: the effects of daikenchuto, TU-100, on gastrointestinal and colonic transit, anorectal and bowel function in female patients with functional constipation. *Aliment. Pharmacol. Ther.*, **37**, 776–785.
10. Lu,C., Ke,L., Li,J., Zhao,H., Lu,T., Mentis,A.F.A., Wang,Y., Wang,Z., Polissiou,M.G., Tang,L., *et al.* (2021) Saffron (Crocus sativus L.) and health outcomes: a meta-research review of meta-analyses and an evidence mapping study. *Phytomedicine*, **91**, 153699.
11. Zhao,M.M., Lu,J., Li,S., Wang,H., Cao,X., Li,Q., Shi,T.T., Matsunaga,K., Chen,C., Huang,H., *et al.* (2021) Berberine is an insulin secretagogue targeting the KCNH6 potassium channel. *Nat. Commun.*, **12**, 5616.
12. Zhong,L.L.D., Cheng,C.W., Kun,W., Dai,L., Hu,D.D., Ning,Z.W., Xiao,H.T., Lin,C.Y., Zhao,L., Huang,T., *et al.* (2019) Efficacy of MaZiRenWan, a Chinese herbal medicine, in patients with functional constipation in a randomized controlled trial. *Clin. Gastroenterol. Hepatol.*, **17**, 1303–1310.
13. Lyu,J., Xie,Y., Sun,M. and Zhang,L. (2020) Clinical evidence and GRADE assessment for breviscapine injection (DengZhanHuaSu) in patients with acute cerebral infarction. *J. Ethnopharmacol.*, **262**, 113137.
14. Fan,Y., Li,S., Ding,X., Yue,J., Jiang,J., Zhao,H., Hao,R., Qiu,W., Liu,K., Li,Y., *et al.* (2019) First-in-class immune-modulating small molecule Icaritin in advanced hepatocellular carcinoma: preliminary results of safety, durable survival and immune biomarkers. *BMC Cancer*, **19**, 279.
15. Mo,D., Zhu,H., Wang,J., Hao,H., Guo,Y., Wang,J., Han,X., Zou,L., Li,Z., Yao,H., *et al.* (2021) Icaritin inhibits PD-L1 expression by targeting protein iκb kinase α. *Eur. J. Immunol.*, **51**, 978–988.
16. Fang,S., Dong,L., Liu,L., Guo,J., Zhao,L., Zhang,J., Bu,D., Liu,X., Huo,P., Cao,W., *et al.* (2021) HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. *Nucleic Acids. Res.*, **49**, D1197–D1206.
17. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K., *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
18. Li,F.S. and Weng,J.K. (2017) Demystifying traditional herbal medicine with modern approach. *Nat. Plants*, **3**, 17109.
19. Zhang,Y., Li,X., Shi,Y., Chen,T., Xu,Z., Wang,P., Yu,M., Chen,W., Li,B., Jing,Z., *et al.* (2023) ETCM v2.0: an update with comprehensive resource and rich annotations for traditional Chinese medicine. *Acta Pharm Sin B*, **13**, 2559–2571.
20. Xu,H.Y., Zhang,Y.Q., Liu,Z.M., Chen,T., Lv,C.Y., Tang,S.H., Zhang,X.B., Zhang,W., Li,Z.Y., Zhou,R.R., *et al.* (2019) ETCM: an encyclopaedia of traditional Chinese medicine. *Nucleic Acids. Res.*, **47**, D976–D982.
21. Tian,S., Zhang,J., Yuan,S., Wang,Q., Lv,C., Wang,J., Fang,J., Fu,L., Yang,J., Zu,X., *et al.* (2023) Exploring pharmacological active ingredients of traditional Chinese medicine by pharmacotranscriptomic map in ITCM. *Brief. Bioinform.*, **24**, bbad027.
22. Yan,D., Zheng,G., Wang,C., Chen,Z., Mao,T., Gao,J., Yan,Y., Chen,X., Ji,X., Yu,J., *et al.* (2022) HIT 2.0: an enhanced platform for Herbal Ingredients' Targets. *Nucleic Acids. Res.*, **50**, D1238–D1243.
23. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R., *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids. Res.*, **43**, D36–D42.
24. Fishilevich,S., Zimmerman,S., Kohn,A., Iny Stein,T., Olender,T., Kolker,E., Safran,M. and Lancet,D. (2016) Genic insights from integrated human proteomics in GeneCards.*Database (Oxford)*, **2016**, baw030.
25. Zhou,Y., Zhang,Y., Zhao,D., Yu,X., Shen,X., Zhou,Y., Wang,S., Qiu,Y., Chen,Y. and Zhu,F. (2024) TTD: therapeutic target database describing target druggability information. *Nucleic Acids. Res.*, **52**, D1465–D1477.
26. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids. Res.*, **48**, D845–D855.
27. Amberger,J.S., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2019) OMIM.Org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids. Res.*, **47**, D1038–D1043.
28. S.,K., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasilevsky,N.A., Danis,D., Balagura,G., Baynam,G., Brower,A.M, *et al.*2021) The Human phenotype ontology in 2021. *Nucleic Acids. Res.*, **49**, D1207–D1217.
29. Schriml,L.M., Munro,J.B., Schor,M., Olley,D., McCracken,C., Felix,V., Baron,J.A., Jackson,R., Bello,S.M., Bearer,C., *et al.* (2022) The Human Disease Ontology 2022 update. *Nucleic Acids. Res.*, **50**, D1255–D1261.
30. Brown,T.B., Mann,B., Ryder,N., Subbiah,M., Kaplan,J., Dhariwal,P., Neelakantan,A., Shyam,P., Sastry,G., Askell,A., *et al.* (2020) In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc., Vancouver, BC, Canada, pp. Article 159.
31. Team,G., Anil,R., Borgeaud,S., Wu,Y., Alayrac,J.-B., Yu,J., Soricut,R., Schalkwyk,J., Dai,A.M. and Hauth,A. (2023) Gemini: a family of highly capable multimodal models. arXiv doi: https://arxiv.org/abs/2312.11805, 19 December 2023, preprint: not peer reviewed.
32. Clough,E., Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., *et al.* (2024) NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids. Res.*, **52**, D138–D144.
33. Wang,Z., Monteiro,C.D., Jagodnik,K.M., Fernandez,N.F., Gundersen,G.W., Rouillard,A.D., Jenkins,S.L., Feldmann,A.S., Hu,K.S., McDermott,M.G., *et al.* (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.*, **7**, 12846.
34. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
35. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
36. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
37. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic. Acids. Res.*, **43**, e47.
38. Carvalho,B.S. and Irizarry,R.A. (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, **26**, 2363–2367.
39. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**, 100141.

40. Mosteller,F. and Fisher,R.A. (1948) Questions and answers. *The American Statistician*, **2**, 30–31.
41. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N., *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
42. Chen,B., Ma,L., Paik,H., Sirota,M., Wei,W., Chua,M.S., So,S. and Butte,A.J. (2017) Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat. Commun.*, **8**, 16022.
43. Zhang,S.D. and Gant,T.W. (2008) A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinf.*, **9**, 258.
44. Fiorini,N., Lipman,D.J. and Lu,Z. (2017) Towards PubMed 2.0. *eLife*, **6**, e28801.
45. Zhang,Y., Zhang,Y., Qi,P., Manning,C.D. and Langlotz,C.P. (2021) Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.*, **28**, 1892–1899.
46. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.
47. Miller,J.J. (2013) In: *Proceedings of the southern association for information systems conference*. Atlanta, GA, USA, Vol. **2324**, pp. 141–147.
48. Li,D., Mei,H., Shen,Y., Su,S., Zhang,W., Wang,J., Zu,M. and Chen,W. (2018) ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics*, **2**, 136–146.
49. Zhao,Z., Fan,H., Higgins,T., Qi,J., Haines,D., Trivett,A., Oppenheim,J.J., Wei,H., Li,J., Lin,H., *et al.* (2014) Fufang Kushen injection inhibits sarcoma growth and tumor-induced hyperalgesia via TRPV1 signaling pathways. *Cancer Lett.*, **355**, 232–241.
50. Yang,Y., Sun,M., Yao,W., Wang,F., Li,X., Wang,W., Li,J., Gao,Z., Qiu,L., You,R., *et al.* (2020) Compound kushen injection relieves tumor-associated macrophage-mediated immunosuppression through TNFR1 and sensitizes hepatocellular carcinoma to sorafenib. *J. Immunother. Cancer*, **8**, e000317.
51. Liu,X., Wu,Y., Zhang,Y., Bu,D., Wu,C., Lu,S., Huang,Z., Song,Y., Zhao,Y., Guo,F., *et al.* (2021) High throughput transcriptome data analysis and computational verification reveal immunotherapy biomarkers of compound kushen injection for treating triple-negative breast cancer. *Front. Oncol.*, **11**, 747300.
52. Li,N., Wu,X., Zhuang,W., Xia,L., Chen,Y., Wu,C., Rao,Z., Du,L., Zhao,R., Yi,M., *et al.* (2021) Tomato and lycopene and multiple health outcomes: umbrella review. *Food Chem.*, **343**, 128396.
53. Ratto,F., Franchini,F., Musicco,M., Caruso,G. and Di Santo,S.G. (2022) A narrative review on the potential of tomato and lycopene for the prevention of Alzheimer's disease and other dementias. *Crit. Rev. Food Sci. Nutr.*, **62**, 4970–4981.
54. Zhang,X., Zhou,Q., Qi,Y., Chen,X., Deng,J., Zhang,Y., Li,R. and Fan,J. (2024) The effect of tomato and lycopene on clinical characteristics and molecular markers of UV-induced skin deterioration: a systematic review and meta-analysis of intervention trials. *Crit. Rev. Food Sci. Nutr.*, **64**, 6198–6217.

Received: September 21, 2024. **Revised:** October 15, 2024. **Editorial Decision:** October 16, 2024. **Accepted:** October 22, 2024