TCMM: A Unified Database for Traditional Chinese Medicine Modernization and Therapeutic Innovations

Zhixiang Ren[b], Yiming Ren[b], Zeting Li[b] and Huan Xu[a,c,*]

[a]School of Public Health, Anhui University of Science and Technology, Hefei, 231131, Anhui Province, China
[b]Peng Cheng Laboratory, Shenzhen, 518055, Guangdong Province, China
[c]Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism and Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai, 200237, China

ARTICLE INFO

Keywords:
Traditional Chinese Medicine
Database
Knowledge Graph
Deep Learning
Graph Neural Network

ABSTRACT

Mining the potential of traditional Chinese medicine (TCM) in treating modern diseases requires a profound understanding of its action mechanism and a comprehensive knowledge system that seamlessly bridges modern medical insights with traditional theories. However, existing databases for modernizing TCM are plagued by varying degrees of information loss, which impede the multidimensional dissection of pharmacological effects. To address this challenge, we introduce traditional Chinese medicine modernization (TCMM), the currently largest modernized TCM database that integrates pioneering intelligent pipelines. By aligning high-quality TCM and Western medicine data, TCMM boasts the most extensive TCM modernization knowledge, including 20 types of modernized TCM concepts such as prescription, ingredient, target and 46 biological relations among them, totaling 3,447,023 records. We demonstrate the efficacy and reliability of TCMM with two features, prescription generation and knowledge discovery, the outcomes show consistency with biological experimental results. A publicly available web interface is at `https://www.tcmm.net.cn/`.

1. Introduction

TCM is an empirical science based on thousands of years of clinical experience and continues to play a crucial role in disease diagnosis and treatment[1]. In recent years, a wealth of studies[2, 3, 4, 5] have demonstrated that TCM and Western medicine share the same theoretical foundation at the molecular level. Specifically, compounds can treat diseases by regulating the efficacy of targets. Despite this, current information on the chemical components, and metabolism mechanisms of TCM is still lacking, making the relation between drug components and pharmacological effects unclear, hindering the targeted TCM diagnosis and treatment. Therefore, in the modernization process of TCM, it is urgent to establish a comprehensive and highly reliable TCM modernization database, which would facilitate the identification of active ingredient within herb and elucidate the mechanism of action with biological targets.

Recently, numerous efforts[6, 7, 8, 9] have been made to construct TCM databases that incorporate information about herb, ingredient, and target to support related research. However, these works often overlook the critical concepts of "symptom" and "prescription" inherent in the theoretical system of TCM. SymMap[10] incorporates symptom information, which is beneficial for phenotype-based drug discovery research. LTM-TCM[11] and CPMCP[12] introduce the concept of prescription, which helps Chinese medicine practitioners understand the compatibility of herbs in prescriptions. TCMBANK[13] integrates a vast amount of data on herb, ingredient, disease, and target entities but neglects important concepts like prescription and symptom in the theoretical system of TCM. Despite these advancements, current integrative works on TCM and Western medicine have only partially involved modern medical concepts like disease, target, and ingredient. This simplifies the principles of ingredient-based disease treatment and ignores essential information such as anatomy, and pathway, which is not conducive to understanding the complete pharmacological effects. Additionally, in the aforementioned databases, relations are binary, while in the biological process, relations between entities are complex and diverse. Different relations may represent the alleviation or aggravation of diseases, such as ingredient-downregulate/upregulate-gene.

Knowledge discovery is a hot topic in TCM, encompassing various challenging tasks such as digging into the molecular mechanisms of herbs and discovering the treatment patterns of TCM prescriptions. Although some researchers have used network pharmacology to study the action mechanism of prescriptions[14, 15, 16] and explore herb-symptom correlations[17], these efforts have focused on analyzing specific prescription or disease without discovering a universal method for explaining the mechanism. With advancements in artificial intelligence (AI), some methods combining TCM data and deep learning (DL) have been developed for tasks such as prescription-target interaction[18] and herb-symptom correlation[19]. However, due to the lack of detailed information, the pharmacological principle of TCM remains largely unexplored, leaving a gap in the knowledge discovery of modernized TCM.

Symptom-based prescription plays a crucial role in the treatment process of TCM, and the combination of prescription generation with AI methods has become a trend in recent years, providing convenience for patients

*Corresponding author at: School of Public Health, Anhui University of Science and Technology
E-mail addresses: `xuhuan@ecust.edu.cn`

Database	Prescription	Herb	Ingredient	Target	Disease	MM Symptom	TCM Symptom	Syndrome	Pathway	Total
TCM-ID [30]	7443	2751	7375	2756	4111	/	/	/	/	24658
CHEM-TCM [8]	/	240	7000	78	/	/	/	/	/	7318
TCM@TAIWAN [7]	/	453	20000	/	/	/	/	/	/	20453
TCMSP [6]	/	502	13729	3339	867	/	/	/	/	18437
TM-MC 2.0 [9]	5075	635	34061	13991	27997	/	/	/	/	81759
HerDing [31]	/	**19476**	6655	16762	11394	/	/	/	/	54287
TCMID 2.0 [32]	46929	8159	43413	17603	4633	/	/	/	/	65649
ETCM [33]	3959	402	7284	2266	4323	/	/	/	/	18234
SymMap v2 [10]	/	698	25975	20965	14086	1148	**2285**	**233**	/	65390
HITv2 [34]	/	1250	1237	2208	/	/	/	/	/	4695
HERB [35]	/	7263	49258	12933	28212	/	/	/	/	97666
CPMCP [12]	2125	1560	27928	20965	14434	1148	2285	/	/	70445
LTM-TCM [11]	**48126**	9122	34967	13109	/	/	1928	/	/	107252
TCMBank [13]	/	9191	61965	15179	**32529**	/	/	/	/	118864
TCMM	48043	8932	**69816**	**76449**	22365	**17079**	1900	146	**3704**	**248434**

Table 1
TCM Database Comparison provides a quantitative comparison of the data scale of TCMM with existing TCM databases. To ensure fairness, we only count the entity types present in existing databases. The values are based on the most recent data from the relevant databases. The results show that TCMM contains the most comprehensive entity type and the largest amount of data.

and doctors. Recently, [20, 21] proposed the topic model for prescription generation that incorporates background knowledge such as herbal compatibility, aiming to facilitate the creation of prescriptions and integrate TCM theories into the generation process. To generate prescriptions that align more closely with TCM concepts, [22, 23, 24] focus on the use of seq2seq structures, by viewing TCM prescription generation as a sequence process task and decode the latent representation of symptoms into herb sequences. Additionally, GNN can effectively capture structural and semantic information between entities and is widely applied to prescription generation tasks. Many studies[25, 26, 27, 28, 29] have utilized GNN to capture high-order correlations among herbs, symptoms, and prescriptions, recommending high-quality herb collections based on background knowledge. However, existing methods overlooked internal prescription information such as compatibility principles and inherent properties of medicinal materials, while also lacking the use of mechanistic information like pathways and targets. nn In this work, we introduce TCMM, a unified database for TCM modernization that integrates 6 high-quality databases of TCM and Western medicine. According to Table 1, TCMM is the currently largest non-commercial database for TCM modernization, consisting of 20 typical entities such as prescription, ingredient, target and 46 relations among them, aiming to cover the theoretical systems of TCM and Western medicine as comprehensively as possible. A web-based interface is provided for users to explore relations among herb, ingredient, target, and related pathway or disease. Based on the constructed database, a prescription generation pipeline with pre-train method is proposed, which creates highly credible prescriptions, thereby proving the high performance of the database. This pipeline is integrated into the website to support user-customized prescription generation. Furthermore, in the absence of direct data, we pioneering use a multi-hop reasoning method to achieve TCM knowledge discovery. This includes 2 challenging tasks, prescription repositioning and symptom-related target prediction. The results highly match experimental conclusions in case studies, effectively supporting TCM modernization.

2. Materials and methods

2.1. Collection and processing of TCMM

Existing TCM databases exhibit substantial differences in terms of processing methods and data sources, necessitating the establishment of independent alignment strategies for different types of entities and relationships. The processing methods for the main entities and relations within TCMM are described in detail as follows.

For entities, **Medicinal Material**, sourced from CPMCP and TCMBank, combines information from the "2015 Chinese Pharmacopoeia", literature and multiple databases, such as TCMID, TCM-ID, SymMap, TCMSP, HERB, and TCM@TAIWAN. During data processing, records that share the same HERB ID or SymMap ID are merged. In instances where the attribute values differed, we combine the corresponding names separated by a semicolon. Subsequently, we extract attributes such as the "Four Qi and Five Flavors" and "Meridian Theory," which encompassed 13 kinds of Tropism, 4 categories of Toxicity, 13 varieties of Flavor, and 12 types of Medicinal Properties.

Prescription comes from the TCMID and CPMCP, with the former derived from books and articles, and the latter's from patent medicines and ancient prescriptions. We merge 2,140 prescriptions from CPMCP with 46,929

prescriptions from TCMID, resulting in 48,034 records after removing duplicates. Due to multiple names for some herbs in the prescriptions, we capture 4,947 alternative herb names from `https://zhongyibaike.com` and align the prescription information, in order to improve the recall rate of herb identification.

Ingredient comes from CPMCP and TCMBank. During the integration process, since CAS ID, PubChem ID, and InChI Key cannot uniquely determine the molecule, we merge data based on the same SymMap ID in TCMBank and CPMCP but remove records in TCMBank pointing to multiple SymMap IDs. For example, TCMBANKIN057994 points to SMIT00245 and SMIT1456. In this way, 69,816 ingredient records are ultimately obtained.

Target data is from CPMCP, TCMBANK, PharMeBINet, and PrimeKG databases. NCBI ID is used to integrate data from CPMCP, PrimeKG, and PharMeBINet, and supplement PrimeKG data with gene symbols. Furthermore, to better understand the role of targets in disease processes, we extract molecular function, biological process, and cellular component information for targets from PrimeKG and PharMeBINet and align them using GO ID.

Symptom and syndrome data comes from multiple databases. To explore the correlation between modern medicine (MM) and TCM diseases, MM symptoms and TCM symptoms are obtained separately. TCM symptoms come from CPMCP. Due to some overlapping semantics in TCM Symptoms and Syndromes in the source database, we utilize a large language model(LLM) to merge information from the source database and combine it with manual verification to improve accuracy. Specifically, the semantic descriptions of TCM Symptoms are first generated using prompt engineering combined with GPT methods[36, 37] and then convert the semantic descriptions into embedding representations using Sentence-BERT[38]. To reduce computation, symptoms are grouped according to locus and property and only calculate similarity within groups. Through manual verification, 0.98 is selected as the similarity threshold, and symptoms with a similarity greater than the threshold are merged into one record, resulting in 1,900 TCM Symptoms. To understand the cause of symptoms, TCMM also extracts Locus information from the CPMCP database as an attribute of TCM Symptoms. Syndrome information comes from the SymMap database, and 146 records are obtained after merging semantic information. In addition, 17,079 MM Symptom records are obtained by aligning information using shared MeSH IDs from PharMeBINet and CPMCP.

Disease comes from CPMCP, TCMBANK, PharMeBINet, and PrimeKG databases. We merge the data using keywords such as name, MONDO ID, OMIM ID, MeSH ID, Orphanet ID, and UMLS ID, resulting in 48,233 Disease records. Since a large number of phenotype information in TCMBANK, such as "Body Height" and "Cardiovascular Problem," are categorized as diseases, we only retain Disease information from the PharMeBINet and PrimeKG databases in this study while preserving their database IDs in CPMCP and TCMBank.

For relations, **Prescription - Medicinal Material**: TCMM, by means of deep parsing of prescription descriptions and a combination of LLM with rule matching, extracts 321,913 relations between prescription and herbs from 48,034 prescriptions, of which 264,728 relations contained dosage information. Specifically, we employ ChatGLM to parse the medicinal materials and dosage information in the prescription descriptions. However, as the model is not trained with similar data, the accuracy is only 57%. To achieve better results, we manually annotate a dataset for fine-tuning ChatGLM and randomly extract 200 prediction results for manual evaluation, achieving an accuracy of 96% and a recall rate of 76%. For incorrectly predicted prescriptions, further predictions are made using rule matching, achieving an accuracy rate of 98% under the same evaluation method. In addition, we only retain 17 commonly used weight/volume units in herbal prescriptions and convert the ancient or imprecise units of herbal dosages to g or ml for the modernization of TCM. The dosage unit conversion method is shown in supplementary table S1.

Prescription-Syndrome/Symptom: The relation data of prescription-symptom partially originates directly from the CPMCP database, totaling 50,813 records, but this only covers part of the prescriptions. To further supplement information for other prescriptions, TCMM extracts symptom and syndrome information from the indication and treatment attributes of prescriptions in the TCMID and CPMCP databases. Specifically, we use the already obtained symptom and syndrome entities as keywords, match them with indication descriptions, and ultimately obtain an additional 60,342 prescription-symptom relations and 13,736 prescription-syndrome relations.

Ingredient-Target: Within TCMM, the Ingredient-Gene relation is refined into four different types: associate, bind, downregulate, and upregulate. The data for bind, downregulate and upregulate relations come from the PharMeBINet database, while the associate relation is obtained by integrating information from PharMeBINet, SymMap, CPMCP, and TCMBank databases. To better align with practical application scenarios, we establish a "mutual exclusion" rule for relations. Specifically, for the same ingredient-target group, downregulate and upregulate relations cannot coexist, and associates cannot coexist with the other three relations. For instance, in the PharMeBINet database, the same ingredient-target group was simultaneously marked as downregulate and upregulate. To resolve this conflict, we adopt the more broadly defined associate relation as a replacement.

Target-Target: TCMM contains three types of target-target relations, namely associate, regulate, and covary. The information for regulate and covary comes from the PharMeBINet database, while associate information comes from the PharMeBINet and PrimeKG databases. In TCMM, the associate relation is mutually exclusive with the other two relations, meaning that for the same target-target group, the associate will not coexist with either regulate or covary.

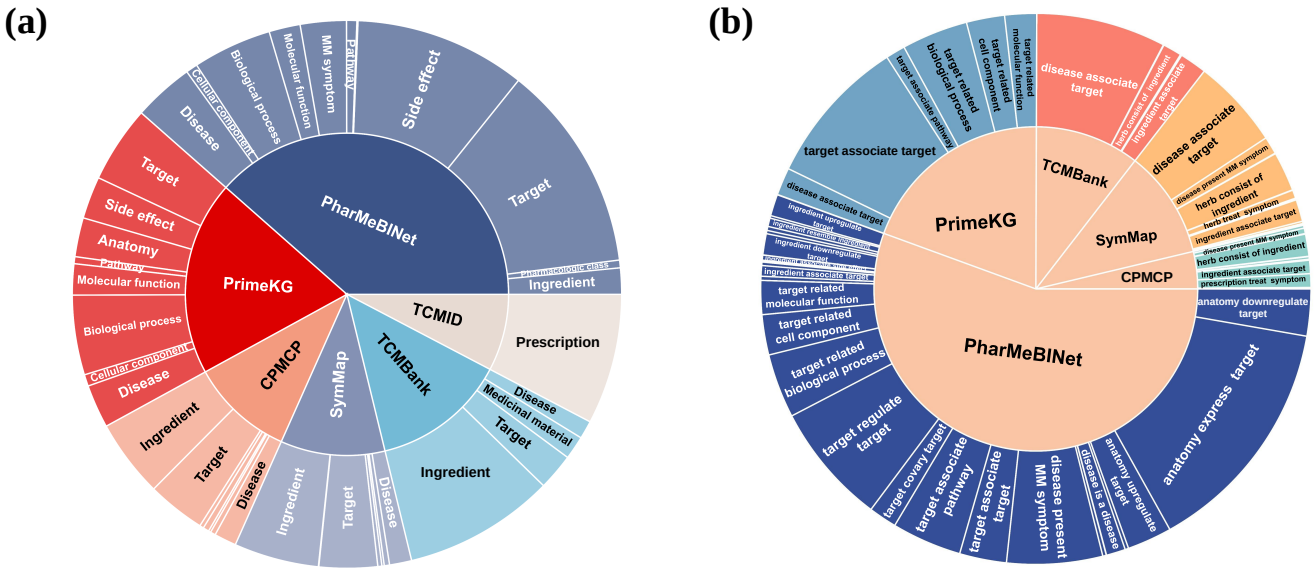

Figure 1: Data Source of TCMM Database provides a visual representation of the data sources of entities and relations in TCMM. The size of each area represents the quantity of entities or relations from that source. **(a), (b) Data Source of TCMM Entities and Relations** integrates Western medicine information from PharMeBINet, PrimeKG and integrates TCM information from TCMBank, CPMCP, TCMID and SymMap. Specific rules and LLM are used for information combination.

Disease-Disease: TCMM integrates resemble information from the PharMeBINet database. In this database, these two relations are deemed mutually exclusive. Therefore, for situations in the PharMeBINet database where two relations coexist, only the 'is a' information is retained.

Ingredient-Disease: The PharMeBINet database provides TCMM with induce, contraindicate, and treat relations for the ingredient-disease relation. In this database, since treat is semantically mutually exclusive with induce and contraindicate, we discard data from PharMeBINet where coexistence situations exist.

2.2. Data source and statistics of TCMM

TCMM is sourced from the 6 largest TCM knowledge bases, PrimeKG[39], PharMeBINet[40], TCMBank[13], SymMap[10], TCMID[32] and CPMCP[12], which results in a more detailed relations and a wider variety of entity types. In this work, the TCM or Western medical concepts, such as prescription, ingredient and target, are treated as entities. Additionally, a relation describes the biological correlation between entities. In this way, TCMM encompasses 20 types of entities and 46 kinds of relations, amounting to a total of 3,447,023 records, forming the most comprehensive TCM modernization database currently available. Data source of entities and relations is depicted in Figure 1, which integrates Western medicine information from PharMeBINet and PrimeKG, and integrates TCM information from TCMBank, CPMCP, and SymMap. Figure 2 presents detailed information and statistics of the database, which contains a total of 248434 entities and 3447023 triplets. The integration is beneficial for drug researchers to understand the pharmacological effects of TCM from a modern medical perspective. Additionally, it enhances the potential of the knowledge graph (KG), such as enabling it to address challenging tasks like prescription repositioning for modern diseases.

2.3. Knowledge augmentation based TCM prescription generation

In this study, we implement a TCM prescription generation pipeline upon the work proposed by [42],which grounds in the Encoder-Decoder framework shown in Figure 3. By combining it with the Diverse Beam Search[43], the model is proven to be capable of producing more diverse prescriptions. Furthermore, we extract relevant information from the database to construct a KG and incorporate Chinese and Western medical knowledge into the model via a graph pre-train task. Experimental results demonstrate that database knowledge enhances metrics and leads to more reasonable prescriptions.

2.3.1. Data processing

Prescription data containing dosage information is selected as the dataset to facilitate the construction of a sequence generation model from symptom sequence to herb sequence. To unify the order of sequence data, symptom sequences are sorted according to symptom ID in the database. Meanwhile, according to the principle of Jun-Chen-Zuo-Shi compatibility, the medicinal materials for treating the main symptoms usually have a larger dosage, so the herb sequences are sorted according to the weight proportion of the herbs in the prescription. The dataset contains a total of 20,911 prescriptions, which are randomly divided into training, validation, and test sets at a ratio of 8:1:1.

To clarify the relation between herbs and symptoms, the model's initial parameters employ embeddings pre-trained on

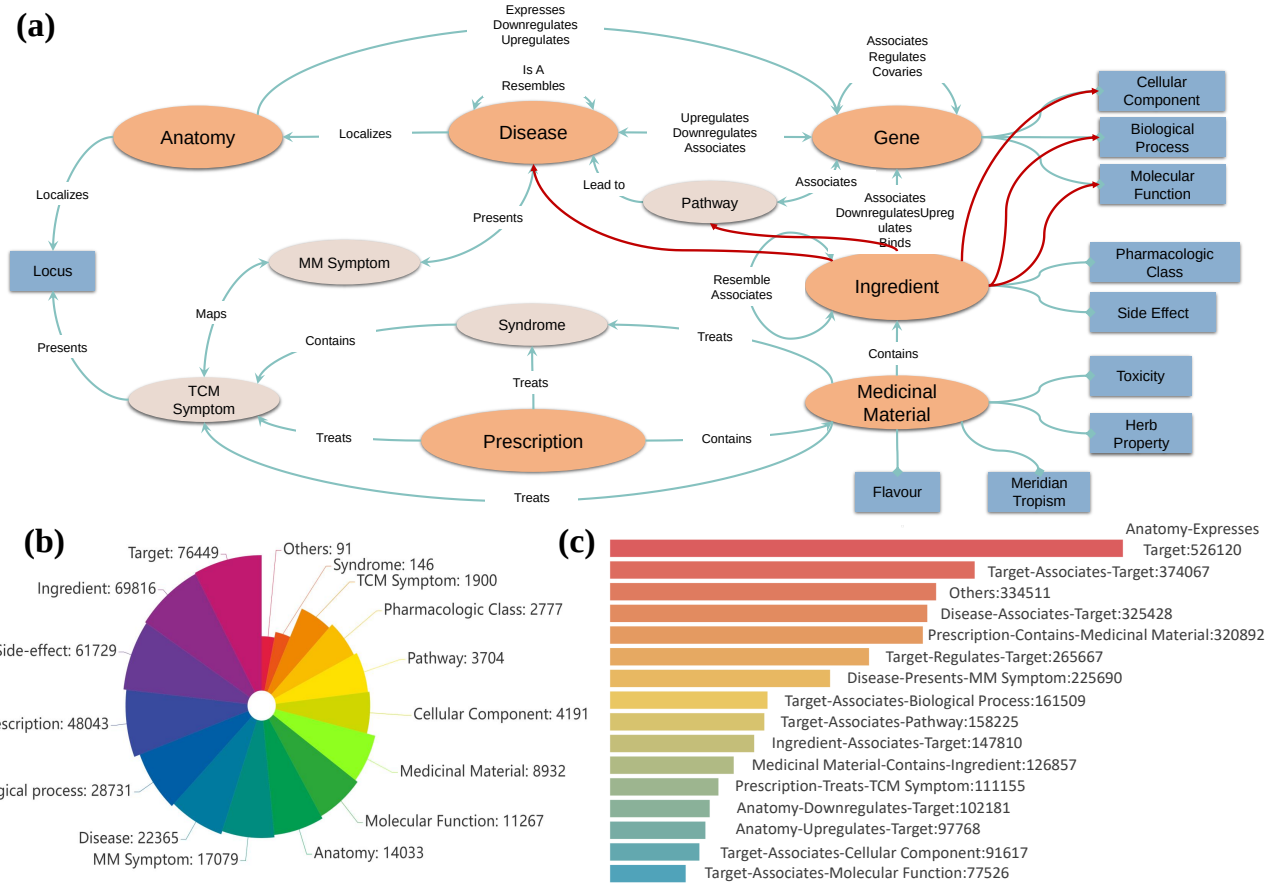

Figure 2: Overview of TCMM Database.(a) Details of TCMM Database displays the complete 20 entities and 46 relations in the database, while the blue bars represent the attribute entities. **(b) Entity Distribution in TCMM** provides a quantitative overview of the entity types in TCMM. TCM attribute entities are a small minority and are therefore grouped together as "others". **(c) Relation Distribution in TCMM** illustrates the quantity of each relation type within the TCMM. 31 types of relations, due to their low frequency, have been collectively categorized as 'others'

a KG. Expert knowledge directs the extraction of prescription generation-related information from the TCMM database, constructing a KG encompassing entities such as herb, prescription, ingredient, target, TCM symptom and MM symptom. The KG structure is depicted in supplementary figure S1.

2.3.2. Model description

To capture the long-term dependency in symptom sequences, the GRU[44] is utilized as the Encoder to transform variable-length symptom sequences $(s_1, ..., s_M)$ into hidden state sequences $(h_1, ..., h_M)$. Additionally, the bidirectional GRU is employed to adapt to the weak order characteristics of symptom sequences, capturing symptom information before and after the current moment. The hidden state $h_t \in R^{2*hs}$ at time t consists of the forward state $h_f \in R^{hs}$ and the reverse state $h_r \in R^{hs}$, with the calculation process as following:

$$\begin{aligned} h_t &= h_f \oplus h_r \\ h_f &= F_f\left(h_{t-1}, E_{st}\right) \\ h_r &= F_r(h'_{M-t}, E_{st}) \end{aligned} \tag{1}$$

In equation 1, E_{st} represents the embedding of s_t, while F_f and F_r are the forward GRU and reverse GRU networks, respectively. h'_{M-t} is the reverse symptom sequence.

The attention mechanism is utilized to enable the decoder to better model long sequence information of symptoms. Specifically, based on the previous state s_{j-1} and attention A_{j-1} of the symptom, the weight of the symptom hidden state at the current moment α_{tj} is measured, yielding the context vector c_j for the current moment of the decoder. The function a is a soft alignment function used to compute the correlation between s and h.

$$\begin{aligned} \mathrm{c_j} &= \sum_{\mathrm{t=1}}^{\mathrm{T}} \alpha_{\mathrm{tj}} \mathrm{h_t} \\ \alpha_{\mathrm{tj}} &= \mathrm{softmax}\left(\mathrm{a}\left(\mathrm{s_{j-1}}, \mathrm{h_t}, \mathrm{A_{j-1}}\right)\right) \end{aligned} \tag{2}$$

The Decoder utilizes the GRU to decode the hidden state of symptoms into a variable-length herbal sequence incrementally. Since the prescription generation task aims to generate a non-repetitive herbal sequence, a Cover layer is introduced to make the model aware of generated tokens, thereby generating a more diverse and reasonable herbal

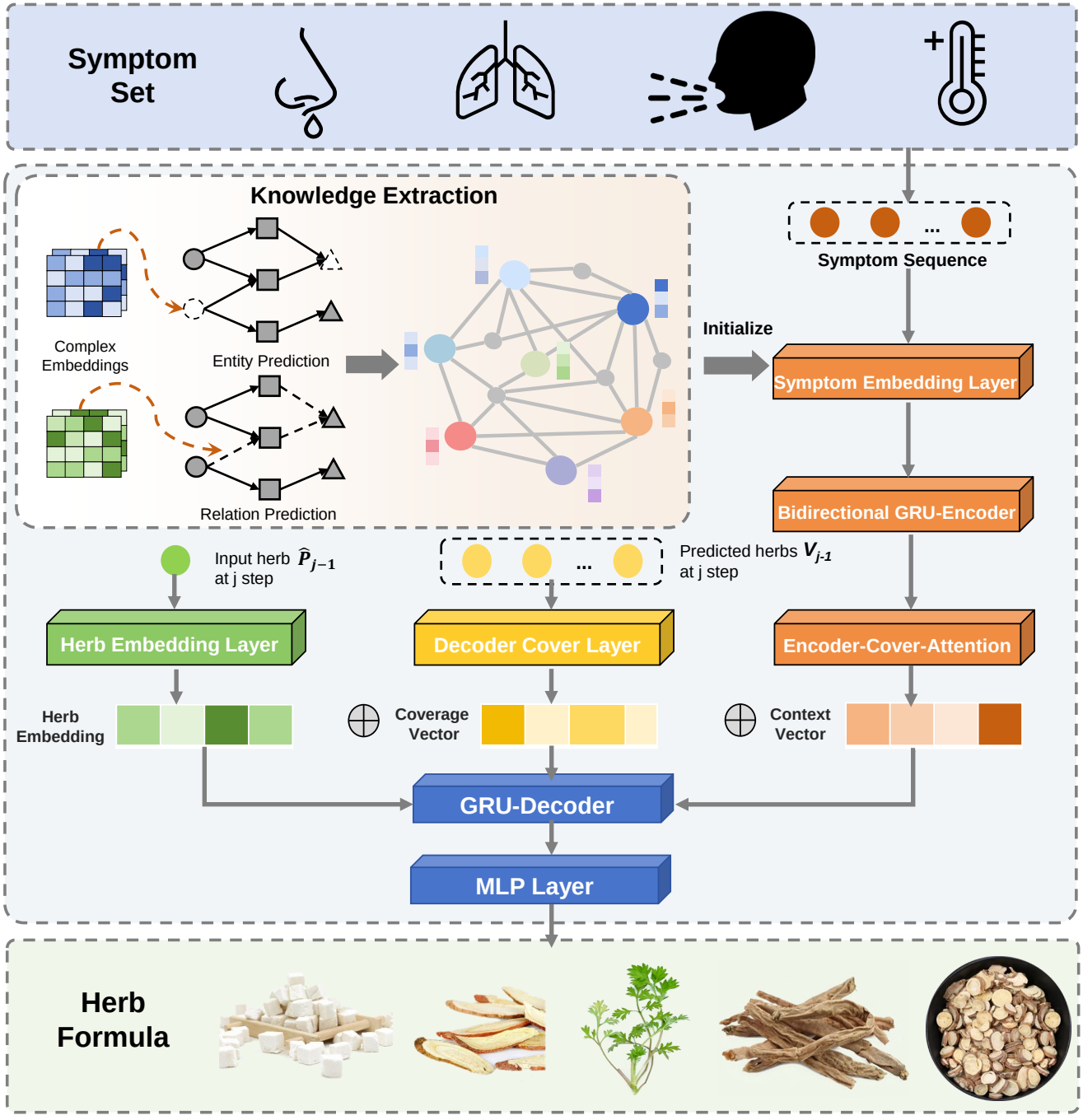

Figure 3: TCM Prescription Generation Model. Knowledge Extraction employs the ComplEX[41] model to treat 'knowledge graph completion' as a pre-training task, facilitating the integration of TCMM knowledge. The model first initializes the entities and relations in the knowledge graph to the complex space to acquire complex embeddings. Then, it conducts collaborative model training on two tasks: entity completion and relationship completion. The pre-trained entity embedding is then used as the initial parameter of symptom embedding layer. **Prescription Generation** model adopts a seq2seq architecture, mapping the input symptom set into a herb formula. Bidirectional GRU is utilized as the Encoder to map the pre-trained symptom embeddings into a context vector. The Decoder then progressively decodes the context vector into the herb ID P_j. In order to prevent the generation of duplicate herbs, a multi-hot vector V_j is introduced to annotate the predicted herbs.

formula. The predicted herbs are transformed into a multi-hot vector representation V_{j-1}, which is subsequently projected onto the hidden space CV_{j-1} via the Cover layer. This, in conjunction with the context vector c_j, the previous state s_{j-1}, and the herb embedding in the previous timestep E_hj, is collectively used to update the current state s_j of the Decoder.

The MLP then decodes s_j into the herb ID p_j.

$$\begin{aligned} p_j &= \mathrm{softmax}\left(\mathrm{W}_{\mathrm{out}}^{\mathrm{T}}\, \mathrm{s_j} + \mathrm{b_{out}}\right) \\ \mathrm{s_j} &= \mathrm{F_{Decoder}}\left(\mathrm{s_{j-1}}, \mathrm{Concat}\left[\mathrm{c_j}; E_{hj}; \mathrm{cv_{j-1}}\right]\right) \\ \mathrm{cv_{j-1}} &= \tanh\left(\mathrm{W}_{\mathrm{cover}}^{\mathrm{T}} \mathrm{V_{j-1}} + \mathrm{b_{cover}}\right) \end{aligned} \tag{3}$$

Pre-trained embedding is introduced as the initial parameter for our model to enhance the performance of downstream tasks. In order to capture the semantics within the complex graphical structure, we design a KGC task, which aims to explore high-order correlation and encode the intricate information in heterogeneous graph. ComplEX [41] is utilized as pre-trained model since its outstanding performance in the KGC task. According to [41], in order to more fully integrate the knowledge within the graph, both entity prediction and relation prediction are jointly used as training tasks.

$$\arg\max \sum_{\langle u,r,v\rangle \in \mathcal{G}} \log p(u \mid r, v) + \log p(v \mid u, r) + \lambda \log p(r \mid u, v) \tag{4}$$

As the equation 4 shown, the objective of ComplEX is to maximize the predictive probability of the combined optimization equation, encompassing three tasks: predicting the head entity u, the tail entity v, and the relation r.

Cross entropy is used as the loss function of the model. However, the original cross entropy function requires that the target sequence have a strict order in order to measure the loss of the predicted sequence. Since the herb sequence in the prescription generation task does not demand a strict order, this work employs the smoothed target probability y'_j to relax the penalty for incorrect prediction of herb position.

$$\begin{aligned} \mathbf{y'_j} &= \left(\mathbf{y_j} + \mathrm{g/T}\right)/2 \\ \mathrm{loss} &= -\sum_{\mathrm{j}} \sum_{\mathrm{vs}} \mathbf{y'_j} \log\left(\hat{\mathbf{p}}_{\mathbf{j}}\right) \end{aligned} \tag{5}$$

$\hat{\mathbf{p}}_{\mathbf{j}}$ is the predicted probability that the herb appears at the jth position of the prescription. y_j is the one-hot encoding of the herb at the jth position of the real prescription sequence. g is the multi-hot encoding of the herbs in the real prescription; if the herb appears in the real prescription, the element at the corresponding position of the g is 1, otherwise it is 0. T is the length of the real prescription and vs represents the vocabulary size.

To support the online inference, it is essential to ensure that the result is globally optimal and has a lower time complexity. Diverse Beam Search[43] strikes a balance between greedy search and exhaustive traversal, dividing the search space into G Beam Search groups, retaining the B sequences with the highest probability as candidates in each iteration, and introducing diversity penalty to enhance sequence diversity.

2.3.3. Experimental study

To validate the enhancement resulting from the incorporation of different types of modernized TCM knowledge into the prescription generation model, we conduct two comparative experiments.

The "Basic" model initializes embeddings arbitrarily without incorporating additional knowledge. The CPMCP database, which has the richest variety of relationship types among TCM databases, is utilized for comparison. The 'CPMCP-based' model is pre-trained using relation information from CPMCP. The 'TCMM-based' model is pre-trained with a KG (Supplementary Figure S1) extracted from TCMM, which contains a more diverse types of relations compared to CPMCP.

Regarding hyperparameters, the maximum epoch is set to 100 and an early stopping strategy is employed. The batch size is set to 128, the embedding dimension is 600, and the hidden dimension is set to 300. The Adam optimizer is chosen for optimization, with a learning rate of 0.001. Precision, Recall, and F1 score are utilized as evaluation metrics, supplementary table S2 shows the results.

The experimental results show that the introduction of additional TCM knowledge will significantly improve the model effect. Moreover, the model based on TCMM knowledge is better than the 'CPMCP-based' model in terms of Precision and F1 score for prescription generation, which means that a more comprehensive modernized TCM knowledge can help identify more symptomatic herbs, thus validating the potential of the TCMM database.

The evaluation metrics are based on a statistical analysis of actual prescriptions extracted from the test dataset.

2.4. Multi-hop reasoning based TCM knowledge discovery

Knowledge graph reasoning can infer new knowledge or query answers based on existing knowledge and has been widely applied in biomedical tasks such as drug-target interaction(DTI) and drug-drug interaction(DDI). However, the lack of direct relational data in the current TCM field makes it challenging for traditional knowledge graph reasoning methods to solve complex queries. Multihop reasoning, as a knowledge graph reasoning task, is primarily utilized to answer complex first-order logic (FOL) queries involving logical operations such as existential quantifier ($\exists$), conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$). Therefore, in this work, we introduce a multi-hop reasoning based pipeline in Figure 8 to transform knowledge discovery into complex logical queries, discovering new knowledge through specific meta-paths.

2.4.1. Data processing

To facilitate comparison with biological experiments, this work primarily concentrates on two tasks: prescription-disease and symptom-target. We extract relations from the database and construct a KG with task-relevant information specified based on expert knowledge. In contrast to the prescription generation task, which focuses on TCM-related attributes such as tropism and flavour, we retain the attribute entities of Western medicine, including biological process and pharmacological class, in order to investigate the

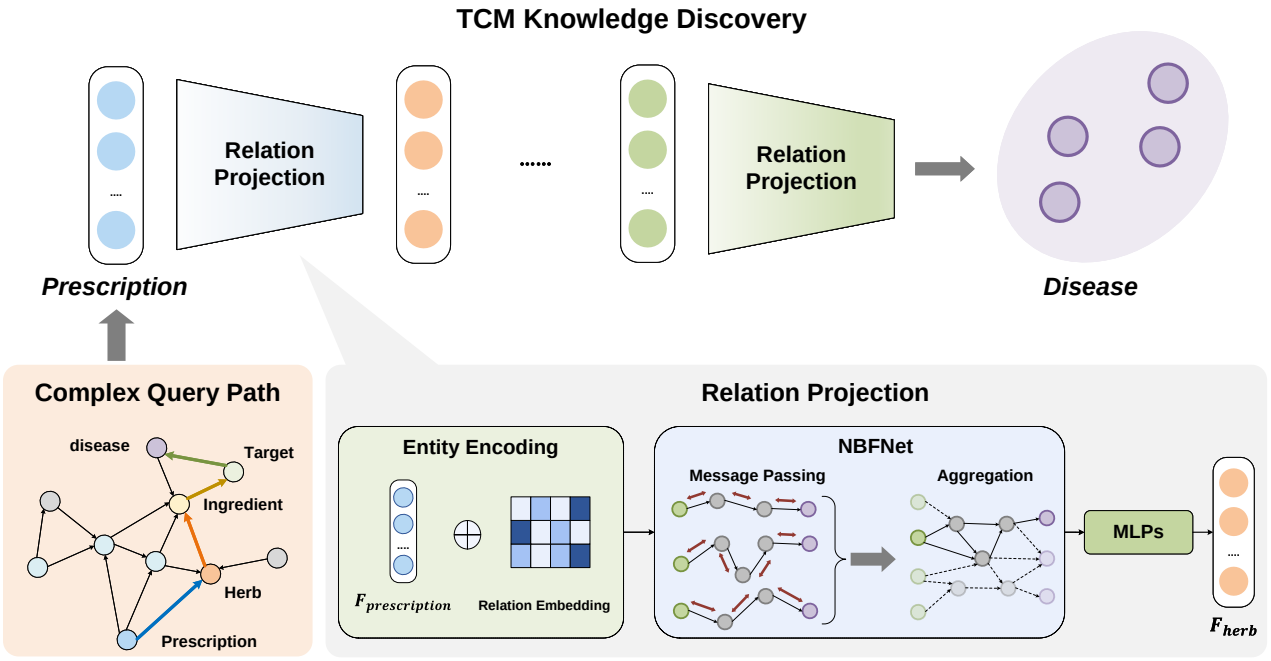

Figure 4: TCM Knowledge Discovery Model is employed to represent relations between entities within predefined complex query paths. Here, take prescription repositioning as an example, which utilizes the **complex query path** $prescription \longrightarrow herb \longrightarrow ingredient \longrightarrow target \longrightarrow disease$. **TCM Knowledge Discovery** takes the fuzzy set of the source entity $F_{prescription}$ as input and iteratively performs relation projection operations to obtain the answer set $A_{disease}$. **Relation Projection** initially merges the fuzzy set of the head entity in the corresponding triplet with the relation embedding, then utilizes the merged information as input into the NBFNet for message passing and aggregation. Finally, the representation is mapped to the tail entity's fuzzy set through MLP layers and sigmoid.

correlation between TCM entities and Western medicine entities. Detailed information can be found in supplementary figure S2. We generate training data based on the method stated in [45], and introduce certain constraints, that filter out multiple self-related loops between entities and prohibit the appearance of bidirectional relations in the same path, to ensure the rationality of the path. In addition, logical operations, such as disjunction ($\vee$), and negation ($\neg$), are not considered based on the path features of downstream tasks. Furthermore, since the downstream tasks involve 4p and 5p queries, 4p and 5p data is utilized for training to enhance performance. In order to balance the performance of each query, the quantity of 4p and 5p is limited to one-tenth for other queries because the amount of answers for these tasks is significantly higher than others. The dataset is shown in supplementary table S3.

2.4.2. Model description

GNN-QE[46], as the current state-of-the-art in the complex logical query task, is chosen to complete this task. This model predicts the fuzzy set of answer entities given the head entity and relation, where the elements in the set are all entities on the KG and are represented by a probability for the confidence of the tail entity.

GNN-QE initially transforms an FOL query into an expression of fundamental operations, enabling answer retrieval via expression execution. For example, FOL queries are decomposed into the expression6.

$P_r(x)$ represents the tail entity fuzzy set measured with the relation r and the fuzzy set x of the head entity. $\{prescription\}$ represents the entity sets related to the prescriptions.

Complex logical query on the KG attempts to predict the fuzzy set of the tail entity given the fuzzy set of the head entity x and the relation r. However, traditional KGC methods based on GNN concentrate on the relation projection between single entities, which is difficult to extend to large-scale fuzzy set prediction of the high time complexity. Therefore, NBFNet[47] is utilized to model relation projection in this study. Based on the generalized Bellman-Ford algorithm for single-source problems on graphs, NBFNet has been shown to perform well in the GNN-QE framework due to its lower complexity.

Following the NBFNet, we utilize the function to incorporate the probabilities in the head entity fuzzy set h_v into the relation embedding, which together serves as the entity's initial representation. The entity representation is then input into a multi-layer GNN, integrating the structural information in the graph through message passing and message aggregation. The output of the final layer is then passed to the MLP layer, which predicts the fuzzy set of the tail entity using the sigmoid function. Based on the multi-hop relations in the path, multiple iterations are performed

$$\textbf{FOL query:}\quad q = c : \exists a, b : Consist\ of\ Herb(prescription, a) \wedge \text{Consist of Ingredient}(a, b) \wedge \text{Associate Target}(b, c)$$
$$\textbf{Expression:}\quad \mathcal{P}_{\text{Associate Target}}\left(\mathcal{P}_{\text{Consist of Ingredient}}\left(\mathcal{P}_{\text{Consist of Herb}}(\{prescription\})\right)\right) \tag{6}$$

to ultimately predict the fuzzy set of answers. In equation 7, F represents the function that integrates the tail entity fuzzy set $P_r(x)$ of the previous hop and relation embedding. $\mathcal{E}(v)$ is the triplets of the training KG.

$$\begin{aligned} h_v^{(t-1)} &= \mathrm{F}(P_r(x)^{(t-1)}, r) \\ h_v^{(t)} &= \text{Aggregate}\left(\text{Message}\left(h_v^{(t-1)}, \mathcal{E}(v)\right)\right) \\ P_r(x)^{(t)} &= sigmoid\left(MLP\left(h_v{}^{(t)}\right)\right) \end{aligned} \tag{7}$$

According to [47], we choose binary cross entropy loss to train our model. Ans represents the answer set of the multi-hop query, $\mathcal{V}$ is the set of all entities and y_i represents the probability of i in the final answer fuzzy set.

$$\begin{aligned} \text{BCE Loss} = &-\frac{1}{|\text{Ans}|}\sum_{i\in\text{Ans}}\log y_i \\ &-\frac{1}{|\mathcal{V}-\text{Ans}|}\sum_{i\in\mathcal{V}-\text{Ans}}\log\left(1-\hat{y}_i\right) \end{aligned} \tag{8}$$

2.4.3. Experimental study

Prior work[48, 49, 45, 46] of multi-hop reasoning mainly concentrated on evaluating the performance of 1p,2p,3p queries. However, this study focuses on long-path reasoning. To verify the impact of introducing 4p and 5p data to the model, we conduct an ablation study. $GNN{-}QE_{org}$ refers to the basic model, with the dataset presented in supplementary table S3 and $GNN - QE_{short}$ removes the 4p and 5p data from trainset. To compare model performance, both models employ the same hyperparameters, which are shown in supplementary table S4.

The performance is measured by mean reciprocal rank (MRR). According to the results present in supplementary table S5, the use of long path data demonstrates an enhancement in the performance for queries involving 3p, 4p, and 5p, while it will decrease the performance on short path queries. However, since TCM knowledge discovery requires the integration of intermediate information, it is necessary to have high performance in long-path reasoning.

3. Results

3.1. Modernized TCM research via web interface

TCMM presents a user-friendly website shown in Figure 5, enabling users to effortlessly access comprehensive information and relations among various entities in both TCM and modern medicine. To enhance user experience, the website focuses on showcasing the nine most frequently used entities: prescription, medicinal material, ingredient, pathway, gene, disease, TCM symptom, MM symptom, and syndrome.

The homepage of TCMM is equipped with a multi-entity search function, a concise description of the database, and a navigation bar filled with diverse functions. The search function is designed to support fuzzy search, allowing users to perform searches in Chinese, Pinyin, English, or using Alias names.

The Browse section is a collection of detailed information and statistical results pertaining to the entities. Users can click on the ID within the details to view the relations of the entity. The Relation section is split into three parts: detail, network, and relationships. The Detail primarily shows the entity's attributes, while the network and relation part displays the relation types associated with the entity in the form of a knowledge graph and a table, respectively.

A key function of TCMM is Rx Gen, which is used for online prescription generation. Users can freely combine any of the predefined 1402 symptoms to perform online reasoning. The website is programmed to automatically filter and display the top three prescriptions, each containing detailed information about the medicinal material, thereby guiding the diagnosis and treatment in TCM.

3.2. Case studies

3.2.1. TCM prescription generation

Beyond comparing model performance, it is necessary to validate the feasibility of generating prescriptions based on different types of modernized TCM knowledge. To validate the efficacy of the generated prescriptions, we randomly select various samples from the test set and compare the prediction results of the three models with ground truth.

In Table 2, herbs that appear in actual prescriptions are highlighted in green, herbs that do not appear in actual prescriptions but are symptomatic are highlighted in blue, and herbs not related to symptom are highlighted in red. Following the design of the model performance comparison experiment, 'Basic' represents a model without any additional knowledge incorporated. 'CPMCP-based' signifies a model that includes relation information from CPMCP database. The 'TCMM-based' model integrates relevant knowledge extracted from TCMM. According to Table 2, the pre-train models have a greater tendency to predict symptomatic herbs than the model with random initialization, even if these herbs are absent from actual prescriptions. Meanwhile, the model trained with the TCMM KG is able to identify a greater amount of symptomatic herbs and fewer irrelevant herbs than the model trained with the CPMCP relations. This further validates the effectiveness of the abundant entity types and detailed relation types present in TCMM for prescription generation task.

Symptom	Ground Truth	Basic	CPMCP-based	TCMM-based
laryngitis	licorice, coptis chinensis, platycodon grandiflorus, belamcanda chinensis, arctium lappa	licorice, platycodon grandiflorus, ginseng, pericarpium citri reticulatae	licorice, belamcanda chinensis, cimicifugae Rhizoma, mirabilite	licorice, belamcanda chinensis, cimicifugae Rhizoma, rhinoceros horn, mirabilite
belching	licorice, poria cocos, semen trichosanthis, pinellia ternata, atractylodes macrocephala, scutellaria baicalensis, coptis chinensis, amomum villosum, pericarpium citri reticulatae, citri reticulatae pericarpium viride, nutgrass galingale rhizome	licorice, atractylodes macrocephala, pericarpium citri reticulatae, radix paeoniae alba, ginseng, angelica sinensis, ligusticum sinense	licorice, atractylodes macrocephala, scutellaria baicalensis, coptis chinensis, cyperus rotundus, radix paeoniae alba, angelica sinensis	licorice, poria cocos, atractylodes macrocephala, scutellaria baicalensis, pericarpium citri reticulatae, atractylodes lancea
heavy limbs	licorice, poria cocos, aurantii fructus, cinnamon, common ginger, platycodon grandiflorus, ginseng, Papaya, notopterygii rhizoma et radix, ligusticum sinense, atractylodes lancea, peucedani radix, aconiti lateralis radix praeparata	common ginger, ginseng, aconiti lateralis radix praeparata, atractylodes macrocephala, pericarpium citri reticulatae	licorice, poria cocos, ginseng, cinnamomum cassia atractylodes macrocephala, paeonia lactiflora, angelica sinensis, pinellia ternata, scutellaria baicalensis	poria cocos, common ginger, ginseng, aconiti lateralis radix praeparata atractylodes macrocephala, pinellia ternata
muscle atrophy	licorice, rehmannia, concha haliotidis, sesami nigrum semen, ophiopogon japonicus, pleuropterus multiflorus, dendrobii caulis, radix paeoniae alba, uncariae ramulus cumuncis, polygonati odorati rhizoma	rehmannia, rhizoma dioscoreae, cortex moutan, cornus officinalis, alisma orientale, ophiopogon japonicus	licorice, rehmannia, radix paeoniae alba, wolfiporia extensa, schisandra chinensis, cinnamomum cassia, astragalus membranaceus, atractylodes macrocephala, angelica sinensis, ginseng, achyranthes bidentata, ligusticum sinense	licorice, rehmannia, radix paeoniae alba, poria cocos, schisandra chinensis, cinnamon, astragalus membranaceus, atractylodes macrocephala, angelica sinensis, ginseng
prickly heat	licorice, rheum palmatum, poria cocos, trichosanthis radix, scutellaria baicalensis, coptis chinensis, Chinese mosla herb, angelica sinensis, artemisia annua, dandelion	angelica sinensis, rehmannia, astragalus membranaceus	coptis chinensis, cortex phellodendri	scutellaria baicalensis, coptis chinensis, rehmannia, cortex phellodendri, gardenia jasminoides ellis

Table 2
Case Study of Prescription Generation is used to validate the efficacy of the generated prescriptions. Herbs used in actual prescriptions are marked in green, while herbs not in the prescriptions but symptomatically relevant are in blue, and those unrelated to symptoms are in red. Besides, 'Basic' represents a model without any additional knowledge incorporated. 'CPMCP-based' signifies a model that includes relation information from CPMCP database. The 'TCMM-based' model integrates relevant knowledge extracted from TCMM.

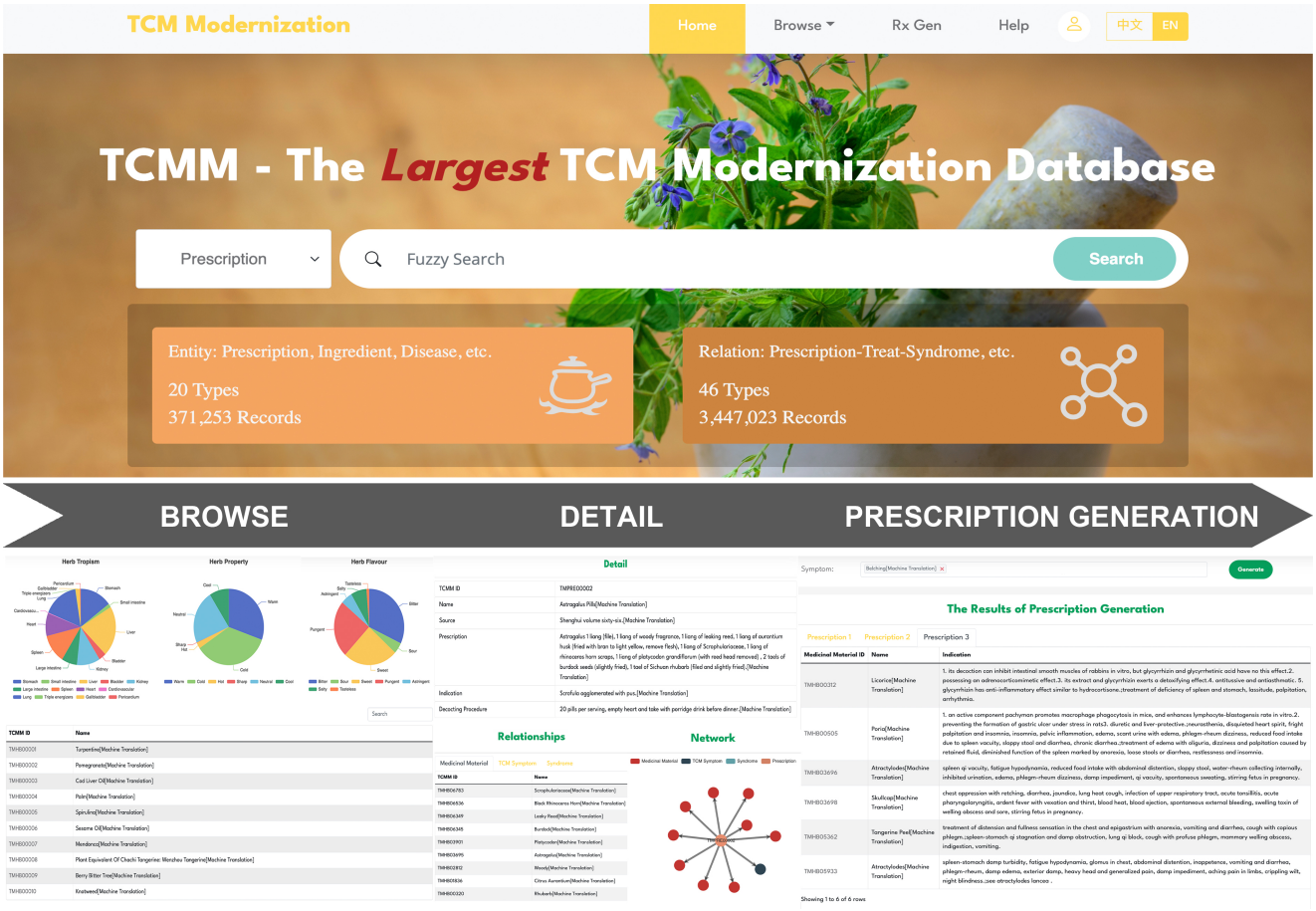

Figure 5: Overview of Web Interface. Home Page contains a navigation bar and a multi-entity search function. **Browse** shows the detailed information and statistical results of each type of entity. **Detail** displays the entity's attributes and the associated relation types in the format of network and table. **Rx Gen** integrates the function of customized prescription generation. Users can get the top-3 prescriptions by entering the combination of the predefined 1402 symptoms.

3.2.2. TCM knowledge discovery

We evaluate the effectiveness of multi-hop reasoning in knowledge discovery for TCM modernization with a series of cases. To discover new insights, we only consider entities outside the training KG as the answer set. Meanwhile, for comparing the model outputs with actual experimental results, we choose prescription repositioning and symptom-related target prediction as downstream tasks, which have extensive experimental cases.

Table 3 shows the case study of prescription repositioning. In order to guarantee the validity of the path data, we choose $prescription \xrightarrow{\text{consist of}} herb \xrightarrow{\text{consist of}} ingredient \xrightarrow{\text{associate}} target \xrightarrow{\text{associate}} disease$ as the inference path based on the expert knowledge. We focus on five prescriptions with primarily concentrated components and select the top five answer entities based on their scores for analysis. The results demonstrate that the model's predictions of high-scoring outputs are highly consistent with the biological experimental results. Additionally, some modern medical insights distinct from TCM knowledge have been discovered. For instance, Salvia miltiorrhiza is primarily used in TCM to treat heart disease, but its active ingredient, Tanshinone IIA, has recently been shown to be effective in cancer treatment[50]. Another example is Panax notoginseng, whose main function is to promote blood circulation, and is frequently used to treat traumatic injuries. Its ingredient notoginsenoside R1 has been proven to be effective against pressure overload-induced cardiac hypertrophy, and the model associates it with aortic stenosis, the related symptom of pressure overload-induced cardiac hypertrophy. Therefore, TCMM can be utilized to predict the potential biological activities of natural products in various herbs, contributing to the development of natural medicinal chemistry and accelerating the discovery of investigational new drugs.

Inflammatory response as a common symptom has been extensively studied, and the understanding of its gene regulation mechanism significantly benefits modern medical diagnosis and treatment. Therefore, for the case study of symptom-related target prediction, we adopt $TCM\,symptom \xrightarrow{\text{map}} MM\,symptom \xrightarrow{\text{presented by}} disease \xrightarrow{\text{associate}} target \xrightarrow{\text{associate}} target$ as the reasoning path and mainly analyze symptoms related to inflammation. The results in

Prescription	Disease	Publication
Galculus Bovis Capsules	sarcomatoid mesothelioma 0.31 ampulla of vater adenocarcinoma 0.31 stomach carcinoma in situ 0.3 ecthyma 0.3 linitis plastica 0.3	[51]
Danshen Granules	salivary gland neoplasm 0.42 intestinal neoplasm 0.41 lung adenocarcinoma 0.41 neurotic depression 0.41 breast fibrocystic disease 0.4	[50]
Patrinia Soup	cutaneous adenocystic carcinoma 0.3 splenic neoplasm 0.3 lung adenoid cystic carcinoma 0.29 acanthoma 0.29 pleomorphic lipoma 0.29	[52]
Panax Notoginseng Cream	diarrheal disease 0.35 aortic stenosis 0.32 spindle cell neoplasm 0.32 sebaceous gland disease 0.31 dental abscess 0.31 dentin dysplasia type II 0.3	[53]
Chuanbei Powder	functional gastric disease 0.35 dentin dysplasia type II 0.3 fish eye disease 0.3 tarsal tunnel syndrome 0.29 deficiency anemia 0.28	[54]

Table 3
Case Study of Prescription Repositioning shows the matching degree between the model's inference results and the experimental results of Prescription Repositioning task. If an item is marked in red, it indicates the consistency with the experimental results. 'Publication' shows the source of the experimental results.

Symptom	Target	Publication
Rheumatoid Arthritis	GABARAPL2 0.34 IQCJ-SCHIP1 0.34 TMEM33 0.33 SORT1 0.33 SH3GLB1 0.33	[55]
Pneumonia	USP50 0.29 STX12 0.29 NDUFA6 0.29 APEX2 0.28 VAMP4 0.28	[56]
Tracheitis	CREB3 0.4 PSMC5 0.4 TFAP2C 0.4 FRMD1 0.39 SNTB2 0.39	[57]

Table 4
Case Study of Symptom Related Target Prediction shows the matching degree between the model's inference results and the experimental results of Symptom Related Target Prediction task. If an item is marked in red, it indicates the consistency with the experimental results. 'Publication' shows the source of the experimental results.

Table 4 show that the model can effectively predict symptom-related regulatory genes. For example, GABARAPL2, as a specific protein, is preferentially recognized by autoantibodies from early rheumatoid arthritis patients[55]. This kind of prediction may provide valuable information for TCM researchers to identify the underlying molecular basis of patient symptoms.

4. Discussion

Based on information regarding TCM prescription, symptom, target, and ingredient, numerous TCM databases have been developed using various data sources. Unfortunately, current efforts to clarify the pharmacological mechanisms of TCM are hindered by the absence of correspondence between TCM and Western medical knowledge. To address these issues, we conduct a study that integrates existing TCM and Western medical databases, refining the relationships between the two fields through a standardized approach.

In order to enhance data quality, a rigorous screening and verification process is implemented for database entities. Specifically, we employ a data alignment process that filters entities based on specific rules, such as ID, name, description, and other relevant information. We also merge duplicate entities within the existing database to enrich entity information. For example, aliases are used to consolidate identical medicinal materials, and LLM is applied to merge semantically similar symptoms. These measures ensure that the database encompasses the majority of entities and relations in TCM and Western medicine, greatly enhancing its practicality in discovering modernized TCM knowledge. Moreover, the potential of the TCMM is validated from two perspectives. By integrating modernized TCM knowledge, the TCM prescription generation method is strengthened, resulting in prescriptions that are not only highly consistent with classical TCM prescriptions but also include new effective components to enhance the therapeutic effects of the original prescriptions. The field of TCM knowledge discovery is limited by the lack of relational data, leaving a gap in the use of computational methods to mine TCM knowledge. Existing methods typically rely on network pharmacology, with a singular and limited focus on specific prescriptions and diseases. The introduction of the TCMM database effectively bridges the gap between TCM and Western medical research, making AI-based, generalized TCM knowledge discovery possible. This study selects prescription repositioning and symptom-related target identification tasks with abundant cases for result validation. Experiments demonstrate that novel pathways consistent with the results of biological experiments can be discovered using TCMM knowledge, further proving the immense potential of TCMM in TCM knowledge discovery.

However, despite our experiments demonstrating the potential of TCMM knowledge, there are still some limitations. First, the database requires further improvements, particularly in the information of prescription. TCMM employs a hybrid approach that combines rules and LLM to extract relations among prescription, medicinal material, and symptom from text. Although the model achieves high accuracy in parsing prescriptions, extensive manual verification is still necessary to standardize the results, which will be addressed in future updates. Moreover, while TCMM data has bridged the knowlegde between TCM and modern medicine, enabling the discovery of modernized TCM knowledge. However, long-path logical queries introduce a large number of random variables, unavoidably reduces the accuracy of the results. Additionally, the training data has not been completely validated by biological experiments, containing noise. Therefore, future research will continue to integrate the latest studies, improve the authenticity of the data, such as target and disease-related information, supplementing graph data to shorten inference paths and enhance the confidence of new knowledge. TCMM will be regularly updated to ensure its ongoing relevance and performance, accelerating the progress of modernized TCM research.

5. Conclusion

In conclusion, the TCMM database represents a groundbreaking advancement in TCM research, offering the largest collection of modernized TCM data and deep learning integration. This resource enhances our understanding of TCM, promotes drug development, and aids clinical applications, setting a new standard for the fusion of traditional wisdom and modern science.

CRediT authorship contribution statement

Zhixiang Ren: Conceptualization, Formal analysis, Supervision, Funding acquisition, Writing review editing. **Yiming Ren:** Methodology, Model training, Writing-original draft. **Zeting Li:** Methodology, Data curation, Model training. **Huan Xu:** Conceptualization, Supervision, Funding acquisition, Writing-review editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Main data and related functions are publicly available through a web interface at `https://www.tcmm.net.cn/`. To obtain the complete data, please contact the corresponding author.

Funding

This work is funded by National Natural Science Foundation of China [grant No. 42177417]. This work is also funded by Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission).

Acknowledgements

The study is supported by the Peng Cheng Laboratory and Peng Cheng Cloud-Brain.

Supplementary Materials

Supplementary material to this article can be found online.

References

[1] Felix Cheung. Tcm: made in china. *Nature*, 480(7378):S82–S83, 2011.

[2] Youyou Tu. Artemisinin—a gift from traditional chinese medicine to the world (nobel lecture). *Angewandte Chemie International Edition*, 55(35):10210–10226, 2016.

[3] Jin-Fang Chen, Shi-Wei Wu, Zi-Man Shi, Yan-Jie Qu, Min-Rui Ding, and Bing Hu. Exploring the components and mechanism of solanum nigrum l. for colon cancer treatment based on network pharmacology and molecular docking. *Frontiers in Oncology*, 13:1111799, 2023.

[4] Xiusong Tang, Jing Chen, Peiqi Ou, Jingqin Chen, Shaohang Lan, Jiajing Luo, Yehao Luo, Yuzhi Shang, and Gang Fang. Chinese herbal compound prescription for endometriosis: a protocol for systematic review and meta-analysis. *Medicine*, 99(42), 2020.

[5] Tsukasa Fueki, Koichiro Tanaka, Kunihiko Obara, Ryudo Kawahara, Takao Namiki, and Toshiaki Makino. The acrid raphides in tuberous root of pinellia ternata have lipophilic character and are specifically denatured by ginger extract. *Journal of natural medicines*, 74:722–731, 2020.

[6] Jinlong Ru, Peng Li, Jinan Wang, Wei Zhou, Bohui Li, Chao Huang, Pidong Li, Zihu Guo, Weiyang Tao, Yinfeng Yang, et al. Tcmsp: a database of systems pharmacology for drug discovery from herbal medicines. *Journal of cheminformatics*, 6:1–6, 2014.

[7] Calvin Yu-Chian Chen. Tcm database@ taiwan: the world's largest traditional chinese medicine database for drug screening in silico. *PloS one*, 6(1):e15939, 2011.

[8] Thomas M Ehrman, David J Barlow, and Peter J Hylands. Phytochemical databases of chinese herbal constituents and bioactive plant compounds with known target specificities. *Journal of chemical information and modeling*, 47(2):254–263, 2007.

[9] Sang-Kyun Kim, SeJin Nam, Hyunchul Jang, Anna Kim, and Jeong-Ju Lee. Tm-mc: a database of medicinal materials and chemical compounds in northeast asian traditional medicine. *BMC Complementary and Alternative Medicine*, 15(1):1–8, 2015.

[10] Yang Wu, Feilong Zhang, Kuo Yang, Shuangsang Fang, Dechao Bu, Hui Li, Liang Sun, Hairuo Hu, Kuo Gao, Wei Wang, et al. Symmap: an integrative database of traditional chinese medicine enhanced by symptom mapping. *Nucleic acids research*, 47(D1):D1110–D1117, 2019.

[11] Xu Li, Jing Ren, Wen Zhang, Zhiming Zhang, Jinchao Yu, Jiawei Wu, He Sun, Shuiping Zhou, Kaijing Yan, Xijun Yan, et al. Ltm-tcm: A comprehensive database for the linking of traditional chinese medicine with modern medicine at molecular and phenotypic levels. *Pharmacological Research*, 178:106185, 2022.

[12] Chang Sun, Jipeng Huang, Rong Tang, Minglei Li, Haili Yuan, Yuxiang Wang, Jin-Mao Wei, and Jian Liu. Cpmcp: a database of chinese patent medicine and compound prescription. *Database*, 2022:baac073, 2022.

[13] Qiujie Lv, Guanxing Chen, Haohuai He, Ziduo Yang, Lu Zhao, Kang Zhang, and Calvin Yu-Chian Chen. Tcmbank-the largest tcm database provides deep learning-based chinese-western medicine exclusion prediction. *Signal Transduction and Targeted Therapy*, 8(1):127, 2023.

[14] Huali Zuo, Qianru Zhang, Shibing Su, Qilong Chen, Fengqing Yang, and Yuanjia Hu. A network pharmacology-based approach to analyse potential targets of traditional herbal formulas: an example of yu ping feng decoction. *Scientific Reports*, 8(1):11418, 2018.

[15] Zihao Wang, Hui-Heng Lin, Kegang Linghu, Run-Yue Huang, Guangyao Li, Huali Zuo, Hua Yu, Ging Chan, and Yuanjia Hu. Novel compound-target interactions prediction for the herbal formula hua-yu-qiang-shen-tong-bi-fang. *Chemical and Pharmaceutical Bulletin*, 67(8):778–785, 2019.

[16] Daiyan Zhang, Yun Zhang, Yan Gao, Xingyun Chai, Rongbiao Pi, Ging Chan, and Yuanjia Hu. Translating traditional herbal formulas into modern drugs: a network-based analysis of xiaoyao decoction. *Chinese medicine*, 15:1–11, 2020.

[17] Kuo Yang, Runshun Zhang, Liyun He, Yubing Li, Wenwen Liu, Changhe Yu, Yanhong Zhang, Xinlong Li, Yan Liu, Weiming Xu, et al. Multistage analysis method for detection of effective herb prescription from clinical data. *Frontiers of medicine*, 12:206–217, 2018.

[18] Zhuo Gong, Naixin Zhang, and Jieyue He. Kgrn: Knowledge graph relational path network for target prediction of tcm prescriptions. In *Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part III 17*, pages 148–161. Springer, 2021.

[19] Chunyang Ruan, Jiangang Ma, Ye Wang, Yanchun Zhang, Yun Yang, and S Kraus. Discovering regularities from traditional chinese medicine prescriptions via bipartite embedding model. In *IJCAI*, pages 3346–3352, 2019.

[20] Xinyu Wang, Ying Zhang, Xiaoling Wang, and Jin Chen. A knowledge graph enhanced topic modeling approach for herb recommendation. In *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part I 24*, pages 709–724. Springer, 2019.

[21] Liang Yao, Yin Zhang, Baogang Wei, Wenjin Zhang, and Zhe Jin. A topic modeling approach for traditional chinese medicine prescriptions. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1007–1021, 2018.

[22] Zeyuan Wang, Josiah Poon, and Simon Poon. Tcm translator: A sequence generation approach for prescribing herbal medicines. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2474–2480. IEEE, 2019.

[23] Chanjuan Li, Dayiheng Liu, Kexin Yang, Xiaoming Huang, and Jiancheng Lv. Herb-know: knowledge enhanced prescription generation for traditional chinese medicine. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1560–1567. IEEE, 2020.

[24] Zhi Liu, Zeyu Zheng, Xiwang Guo, Liang Qi, Jun Gui, Dianzheng Fu, Qingfeng Yao, and Luyao Jin. Attentiveherb: a novel method for traditional medicine prescription generation. *IEEE Access*, 7:139069–139085, 2019.

[25] Yuanyuan Jin, Wei Zhang, Xiangnan He, Xinyu Wang, and Xiaoling Wang. Syndrome-aware herb recommendation with multi-graph convolution network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 145–156. IEEE, 2020.

[26] Wuai Zhou, Kuo Yang, Jianyang Zeng, Xinxing Lai, Xin Wang, Chaofan Ji, Yan Li, Peng Zhang, and Shao Li. Fordnet: recommending traditional chinese medicine formula via deep neural network integrating phenotype and molecule. *Pharmacological research*, 173:105752, 2021.

[27] Wen Zhao, Weikai Lu, Zuoyong Li, Haoyi Fan, Zhaoyang Yang, Xuejuan Lin, Candong Li, et al. Tcm herbal prescription recommendation model based on multi-graph convolutional network. *Journal of Ethnopharmacology*, 297:115109, 2022.

[28] Yuanyuan Jin, Wendi Ji, Wei Zhang, Xiangnan He, Xinyu Wang, and Xiaoling Wang. A kg-enhanced multi-graph neural network for attentive herb recommendation. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(5):2560–2571, 2021.

[29] Yuanyuan Jin, Wendi Ji, Yao Shi, Xiaoling Wang, and Xiaochun Yang. Meta-path guided graph attention network for explainable herb recommendation. *Health Information Science and Systems*, 11(1):5, 2023.

[30] X Chen, H Zhou, YB Liu, JF Wang, H Li, CY Ung, LY Han, ZW Cao, and YZ Chen. Database of traditional chinese medicine and its application to studies of mechanism and to prescription validation.

British journal of pharmacology, 149(8):1092–1103, 2006.

[31] Wonjun Choi, Chan-Hun Choi, Young Ran Kim, Seon-Jong Kim, Chang-Su Na, and Hyunju Lee. Herding: herb recommendation system to treat diseases using genes and chemicals. *Database*, 2016:baw011, 2016.

[32] Lin Huang, Duoli Xie, Yiran Yu, Huanlong Liu, Yan Shi, Tieliu Shi, and Chengping Wen. Tcmid 2.0: a comprehensive resource for tcm. *Nucleic acids research*, 46(D1):D1117–D1120, 2018.

[33] Hai-Yu Xu, Yan-Qiong Zhang, Zhen-Ming Liu, Tong Chen, Chuan-Yu Lv, Shi-Huan Tang, Xiao-Bo Zhang, Wei Zhang, Zhi-Yong Li, Rong-Rong Zhou, et al. Etcm: an encyclopaedia of traditional chinese medicine. *Nucleic acids research*, 47(D1):D976–D982, 2019.

[34] Deyu Yan, Genhui Zheng, Caicui Wang, Zikun Chen, Tiantian Mao, Jian Gao, Yu Yan, Xiangyi Chen, Xuejie Ji, Jinyu Yu, et al. Hit 2.0: an enhanced platform for herbal ingredients' targets. *Nucleic acids research*, 50(D1):D1238–D1243, 2022.

[35] ShuangSang Fang, Lei Dong, Liu Liu, JinCheng Guo, LianHe Zhao, JiaYuan Zhang, DeChao Bu, XinKui Liu, PeiPei Huo, WanChen Cao, et al. Herb: a high-throughput experiment-and reference-guided database of traditional chinese medicine. *Nucleic acids research*, 49(D1):D1197–D1206, 2021.

[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[38] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[39] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[40] Cassandra Königs, Marcel Friedrichs, and Theresa Dietrich. The heterogeneous pharmacological medical biochemical network pharmebinet. *Scientific Data*, 9(1):393, 2022.

[41] Yihong Chen, Pasquale Minervini, Sebastian Riedel, and Pontus Stenetorp. Relation prediction as an auxiliary training objective for improving multi-relational graph representations. *arXiv preprint arXiv:2110.02834*, 2021.

[42] Wei Li and Zheng Yang. Exploration on generating traditional chinese medicine prescriptions from symptoms with an end-to-end approach. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 486–498. Springer, 2019.

[43] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[44] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[45] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. *Advances in Neural Information Processing Systems*, 33:19716–19726, 2020.

[46] Zhaocheng Zhu, Mikhail Galkin, Zuobai Zhang, and Jian Tang. Neural-symbolic models for logical queries on knowledge graphs. In *International Conference on Machine Learning*, pages 27454–27478. PMLR, 2022.

[47] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, 34:29476–29490, 2021.

[48] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

[49] Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *arXiv preprint arXiv:2002.05969*, 2020.

[50] Rui Guo, Lan Li, Jing Su, Sheng Li, Sophia Esi Duncan, Zhihao Liu, and Guanwei Fan. Pharmacological activity and mechanism of tanshinone iia in related diseases. *Drug Design, Development and Therapy*, pages 4735–4748, 2020.

[51] Zhen Zhang, Puhua Zeng, Wenhui Gao, Ruoxia Wu, Tianhao Deng, Siqin Chen, and Xuefei Tian. Exploration of the potential mechanism of calculus bovis in treatment of primary liver cancer by network pharmacology. *Combinatorial Chemistry & High Throughput Screening*, 24(1):129–138, 2021.

[52] Xiao-chen Li, Shuai Wang, Xin-xin Yang, Tian-jiao Li, Jia-xing Gu, Lin Zhao, Yong-rui Bao, and Xian-sheng Meng. Patrinia villosa treat colorectal cancer by activating pi3k/akt signaling pathway. *Journal of Ethnopharmacology*, 309:116264, 2023.

[53] Yuan-Yuan Chen, Quan Li, Chun-Shui Pan, Li Yan, Jing-Yu Fan, Ke He, Kai Sun, Yu-Ying Liu, Qing-Fang Chen, Yan Bai, et al. Qishenyiqi pills, a compound in chinese medicine, protects against pressure overload-induced cardiac hypertrophy through a multi-component and multi-target mode. *Scientific reports*, 5(1):11802, 2015.

[54] Qianqian Tang, Yunfei Wang, Lanjing Ma, Meiling Ding, Tingyu Li, Yongzhan Nie, and Zhengyi Gu. Peiminine serves as an adriamycin chemosensitizer in gastric cancer by modulating the egfr/fak pathway. *Oncology Reports*, 39(3):1299–1305, 2018.

[55] Caroline Charpin, Fanny Arnoux, Marielle Martin, Eric Toussirot, Nathalie Lambert, Nathalie Balandraud, Daniel Wendling, Elisabeth Diot, Jean Roudier, and Isabelle Auger. New autoantibodies in early rheumatoid arthritis. *Arthritis research & therapy*, 15(4):1–9, 2013.

[56] Jae Young Lee, Dongyeob Seo, Jiyeon You, Sehee Chung, Jin Seok Park, Ji-Hyung Lee, Su Myung Jung, Youn Sook Lee, and Seok Hee Park. The deubiquitinating enzyme, ubiquitin-specific peptidase 50, regulates inflammasome activation by targeting the asc adaptor protein. *FEBS letters*, 591(3):479–490, 2017.

[57] Luciana Sampieri, Pablo Di Giusto, and Cecilia Alvarez. Creb3 transcription factors: Er-golgi stress transducers as hubs for cellular homeostasis. *Frontiers in cell and developmental biology*, 7:123, 2019.