

LLM-H2G: biomedical semantic-enhanced hypergraph contrastive learning for herb--disease association prediction

Jun Zhang

Xinjiang Technical Institute of Physics & Chemistry

Hengchuang Yin

Xinjiang Technical Institute of Physics & Chemistry

Chao Wu

Xinjiang Normal University

Jinhao Yao

Xinjiang Technical Institute of Physics & Chemistry

Yue Yang

Xinjiang Technical Institute of Physics & Chemistry

Ziwen Cui

Xinjiang Technical Institute of Physics & Chemistry

Dongxu Li

Xinjiang Technical Institute of Physics & Chemistry

Pengwei Hu

Xinjiang Technical Institute of Physics & Chemistry

Lun Hu

`hulun@ms.xjb.ac.cn`

Xinjiang Technical Institute of Physics & Chemistry

Research Article

Keywords: herb–disease association prediction, hypergraph learning, biomedical language model, contrastive learning, traditional Chinese medicine

Posted Date: August 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10539263/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

LLM-H2G: biomedical semantic-enhanced hypergraph contrastive learning for herb–disease association prediction

Jun Zhang^{1,3}, Hengchuang Yin^{1*}, Chao Wu², Jinhao Yao^{1,3},
Yue Yang^{1,3}, Ziwen Cui^{1,3}, Dongxu Li^{1,3}, Pengwei Hu^{1,3},
Lun Hu^{1,3*}

¹Xinjiang Technical Institute of Physics and Chemistry, Chinese
Academy of Sciences, Urumqi, Xinjiang, China.

²College of Computer Science and Technology, Xinjiang Normal
University, Urumqi, Xinjiang, China.

³University of Chinese Academy of Sciences, Beijing, China.

*Corresponding author(s). E-mail(s): yinhengchuang@ms.xjb.ac.cn;
hulun@ms.xjb.ac.cn;

Abstract

Background: Herb–disease association prediction is central to computational traditional medicine, but existing graph and hypergraph methods mainly rely on observed topology and underuse biomedical textual semantics, especially in heterogeneous or sparse association networks. **Results:** We propose LLM-H2G, a biomedical semantic-enhanced hypergraph contrastive learning framework that integrates herbs, compounds, proteins and diseases through an enhanced heterogeneous hypergraph. Entity names are encoded as large language model-derived semantic representations, injected into compound and protein propagation, and aligned with herb and disease structural readouts through an InfoNCE objective. Across TCM-suite and Ethnobotany benchmarks, LLM-H2G consistently outperforms representative network embedding, graph neural network, graph contrastive learning and hypergraph baselines, with the largest performance gaps observed for sparsely connected herbs. Ablation and parameter analyses show

that both semantic enrichment and enhanced hypergraph construction are necessary for the observed gains. Case studies further show two complementary uses of the model: attribution and docking recover plausible compound–protein mediators for known associations, while literature-supported *de novo* predictions reveal candidate unannotated therapeutic paths. **Conclusions:** LLM-H2G provides an effective framework for integrating biomedical semantics with heterogeneous hypergraph structure and may support interpretable discovery of biologically plausible herb–disease associations. As a representative new path, the *Ginkgo biloba*–Radiation Injuries prediction is supported by a coherent radiation-related MeSH cluster and PubMed-indexed experimental evidence.

Keywords: herb–disease association prediction, hypergraph learning, biomedical language model, contrastive learning, traditional Chinese medicine

Plain language summary. Many herbal medicines are used for complex diseases, but it is difficult to identify which herb–disease links are plausible because herbs contain many compounds and may act through many proteins. We developed LLM-H2G, a computational framework that combines biomedical language-model information with hypergraph learning to predict herb–disease associations. In this study, we show that adding semantic information from biomedical entity names helps the model when known associations are sparse, a common situation in traditional medicine resources. We also show that the model can suggest molecular paths through compounds and proteins, making the predictions easier to interpret. Across two benchmark datasets, LLM-H2G outperformed representative graph, hypergraph, and contrastive-learning baselines. The case studies illustrate how this approach can recover known therapeutic links and highlight candidate associations that are supported by independent literature but missing from curated benchmarks. These results suggest that language-guided hypergraph learning can support more systematic prioritization of herbal therapeutic hypotheses.

1 Background

Traditional Chinese medicine (TCM) has been widely used for centuries to prevent and treat complex diseases. Unlike conventional drugs that often follow a single-compound, single-target paradigm, herbal medicines typically exert therapeutic effects through a multi-component, multi-target manner, in which multiple bioactive compounds jointly modulate diverse proteins and pathways. This systems-level view is consistent with network pharmacology and TCM systems-pharmacology resources [1, 2]. It is also reflected in recent HDA models that explicitly organize herbs, compounds, targets, and diseases into multi-level structures [3, 4]. This intrinsic complexity provides the pharmacological basis for the broad therapeutic potential of herbs, but it also makes the identification of reliable herb–disease associations (HDAs) highly challenging. A central difficulty lies in how to computationally model the cross-level biological relationships among herbs, compounds, proteins, and diseases in a unified and predictive framework [5, 6].

With the rapid development of network pharmacology and biomedical databases, computational approaches have become an important strategy for HDA discovery. TCM-specific databases provide herb, ingredient, symptom, and disease knowledge for large-scale computational modeling [2, 5, 6]. General biomedical resources further support compound, drug, disease, and protein relation curation [7–9]. Protein interaction resources complement these data by describing functional associations among targets [10]. Early network-based methods have demonstrated the utility of integrating heterogeneous biological associations for prediction [11]. Random-walk representation learning methods provide a foundation for embedding nodes from complex networks [12, 13]. Graph neural networks and attention-based variants further improve representation learning on graph-structured data [14–16]. Related biomedical association models use bipartite mutual information, meta-path reasoning, graph convolution,

or graph co-contrastive learning for drug–disease and heterogeneous network prediction [17–19]. However, many of these methods either operate on pairwise relations or focus on graph-level connectivity, making them insufficient for fully characterizing the multi-component, multi-target mechanisms of herbal medicine. In particular, the high-order group relationships naturally induced by “multiple compounds belonging to one herb” and “multiple proteins associated with one disease” are difficult to capture using ordinary graph formulations [20, 21].

To better model such high-order dependencies, hypergraph learning has recently emerged as a promising direction. Hypergraph neural networks have been used to represent higher-order biomedical associations beyond ordinary pairwise edges [20, 21], and recent HDA-specific models further exploit herb–compound and disease–target hypergraphs for TCM association prediction [4, 22]. The representative HGHDA framework [4] models herb–component and disease–protein relationships using a dual-channel hypergraph convolutional network, enabling herb and disease embeddings to be learned from high-order structural patterns. This design significantly advances HDA prediction beyond conventional pairwise graph modeling. Nevertheless, HGHDA mainly relies on similarity-derived component and protein features and performs prediction at the herb–disease level, without explicitly incorporating rich biomedical semantic information or modeling fine-grained molecular communication across the component–protein layer.

Subsequently, HDCTI [23] extends the hypergraph paradigm from macro-level HDA prediction to the more fine-grained task of compound–target interaction (CTI) prediction. By constructing herb–compound and disease–target hypergraphs and linking them through compound–target associations, HDCTI improves the ability to identify therapeutic targets of natural compounds. This task depends on reliable compound and drug records [7, 8], as well as disease and protein association resources [9, 10]. In addition, the integration of PageRank and multi-head attention

enhances global influence propagation and adaptive neighborhood aggregation [15]. However, despite its improved structural modeling, HDCTI is designed primarily for CTI prediction rather than direct HDA inference, and its feature learning still centers mainly on structural relationships, leaving the role of explicit biomedical semantics largely underexplored.

More recently, iCAM-Net [24] further improves interpretability by introducing a cross-channel attention mechanism and an auxiliary component–protein association task, enabling the model to reveal molecular interaction signals underlying predicted HDAs. Other TCM-oriented heterogeneous graph models also show the value of meta-path representations, attention mechanisms and molecular prioritization in complex herbal systems [25]. This represents an important step toward interpretable herb–disease prediction. However, existing methods, including iCAM-Net, still only partially exploit semantic information. Their semantic priors are mostly limited to molecular fingerprints or protein sequence embeddings, while the biomedical semantic knowledge carried by the names or textual descriptors of herbs, compounds, proteins, and diseases is not systematically modeled in a unified representation space. Moreover, the semantic view and the structural view are typically learned separately, without explicit alignment.

In practice, biomedical entities are inherently semantic objects. Herb and symptom names encode clinically meaningful TCM concepts [5, 6], while compound and disease vocabularies carry additional chemical and biomedical semantics [7, 9]. Pretrained biomedical language models such as PubMedBERT [26] and SapBERT [27] provide a powerful opportunity to extract such semantic priors from text. These semantic representations are highly complementary to hypergraph topology: while hypergraph structure describes *how entities are connected*, semantic embeddings provide clues about *what those entities biologically mean*. Therefore, a key unresolved problem is

how to effectively integrate semantic information with structural hypergraph learning for more robust and biologically meaningful HDA prediction [16, 20].

To address this issue, we propose LLM-H2G, a large language model-enhanced hypergraph contrastive learning framework for herb–disease association prediction. LLM-H2G constructs a four-entity heterogeneous system consisting of herbs, compounds, proteins, and diseases, together with three biologically grounded association matrices: herb–compound, compound–protein, and protein–disease. The model adopts a dual-encoder architecture. In the semantic branch, entity names are encoded by a frozen biomedical language model and projected into a unified latent space. In the structural branch, stacked residual hypergraph blocks model high-order herb–compound and disease–protein interactions, while an explicit compound–protein bridge enables information propagation between the two channels and captures the biological transmission path from herbs to diseases through molecular mediators.

Importantly, LLM-H2G introduces two complementary semantic injection strategies to match the roles of different entity types. For compounds and proteins, semantic embeddings are used to *replace* the zeroth-layer structural inputs, allowing the structural encoder to start from biologically informed node representations. For herbs and diseases, semantic embeddings are fused with the structural readouts by *addition*, yielding final herb and disease representations that combine semantic priors with multi-layer hypergraph aggregation. This differentiated “replace vs. add” design allows semantic information to be injected in a role-aware manner, rather than being naively fused at a single stage.

Furthermore, to explicitly bridge the semantic and structural spaces, we introduce an InfoNCE contrastive objective that aligns the semantic embeddings of herbs and diseases with their fused hypergraph representations [28]. Together with the supervised binary classification loss for HDA prediction, this contrastive learning strategy encourages semantic–structural consistency, improves representation robustness, and

helps the model better generalize under sparse and noisy association settings. Based on the final herb and disease embeddings, the model predicts HDA probabilities through an MLP decoder that jointly uses the herb representation, the disease representation and their pairwise interaction information.

In summary, this work presents a semantically enhanced hypergraph framework for herb–disease association prediction. Compared with previous hypergraph-based methods, LLM-H2G not only preserves the ability to capture high-order structural relations, but also explicitly introduces biomedical semantic priors, leverages an interpretable compound–protein bridge, and aligns semantic and structural representations through contrastive learning. By jointly modeling biological structure and biomedical semantics, LLM-H2G provides a more comprehensive and biologically informed solution for HDA prediction.

2 Methods

2.1 Problem definition

We formulate herb–disease association prediction as a link prediction task over a heterogeneous biomedical network with four entity types: herbs $\mathcal{H} = \{h_i\}_{i=1}^{N_H}$, compounds $\mathcal{C} = \{c_i\}_{i=1}^{N_C}$, proteins $\mathcal{P} = \{p_i\}_{i=1}^{N_P}$ and diseases $\mathcal{D} = \{d_i\}_{i=1}^{N_D}$, where N_H , N_C , N_P and N_D are the numbers of entities of each type. The observed herb–compound, compound–protein and protein–disease relations are represented by binary incidence matrices $\mathbf{B}^{HC}$, $\mathbf{B}^{CP}$ and $\mathbf{B}^{PD}$, respectively. Compounds and proteins have precomputed feature matrices $\mathbf{X}^C$ and $\mathbf{X}^P$, and all entities have names that are encoded by the semantic module (the Methodology section).

The associations to be predicted are collected in the herb–disease matrix $\mathbf{Y} \in \{0, 1\}^{N_H \times N_D}$, where $y_{ij} = 1$ indicates that herb h_i is known to be associated with disease d_j , and $y_{ij} = 0$ otherwise. Because only positive associations are curated while the remaining pairs are unverified rather than confirmed negatives, we adopt a

positive-unlabelled formulation [29]. We define the observed positive set $\Omega^+ = \{(i, j) \mid y_{ij} = 1\}$ and the unlabelled set $\Omega^0 = \{(i, j) \mid y_{ij} = 0\}$, and denote by $\Omega^* \supseteq \Omega^+$ the true but partly unobserved set of associations. During training, sampled negatives $\Omega^- \subseteq \Omega^0$ are drawn at a one-to-one ratio, i.e. $|\Omega^-|/|\Omega^+| = 1$; test negatives are held out and are never used during training.

Given these relations, features and entity names, the goal is to learn a scoring function $f(h_i, d_j) = \hat{y}_{ij} \in [0, 1]$ for each unobserved herb-disease pair. Because herbs and diseases are not directly connected by any input relation, the score must be inferred from the herb-compound-protein-disease chain, which motivates the heterogeneous hypergraph model in the Methodology section.

2.2 Model architecture

2.3 Overview

LLM-H2G is designed around a simple hypothesis: reliable HDA prediction requires both biomedical meaning and pharmacological connectivity. We therefore separate the model into three consecutive steps. First, entity names are mapped into a shared semantic space so that herbs, compounds, proteins and diseases can carry text-derived biomedical priors even when the observed network is sparse. Second, an enhanced heterogeneous hypergraph propagates evidence along the herb-compound-protein-disease chain, allowing herbs and diseases to be represented through their incident molecular mediators. Third, semantic and structural views are fused and contrastively aligned before a supervised decoder assigns confidence scores to candidate herb-disease pairs. This design keeps the biological transmission path explicit while using semantic information to stabilize prediction for unannotated or weakly connected entities.

LLM-H2G predicts herb-disease associations by jointly modelling biomedical semantics and heterogeneous hypergraph structure. The overall workflow is summarized in Fig 1. It couples two encoders: a *semantic encoder* that turns each entity

name into a biomedical language-model embedding, and a *structural encoder* that propagates information over an enhanced heterogeneous hypergraph in which herbs and diseases are treated as hyperedges over compound and protein nodes. The two sources are fused at two different granularities: at the node level the semantic embeddings *initialise* the compound and protein states, so that the propagated structural representations are already semantically informed, whereas at the hyperedge level the herb and disease semantic embeddings are *added* to the pooled structural representations. A contrastive objective further aligns the semantic and fused representations, and a multilayer perceptron scores each herb–disease pair [28].

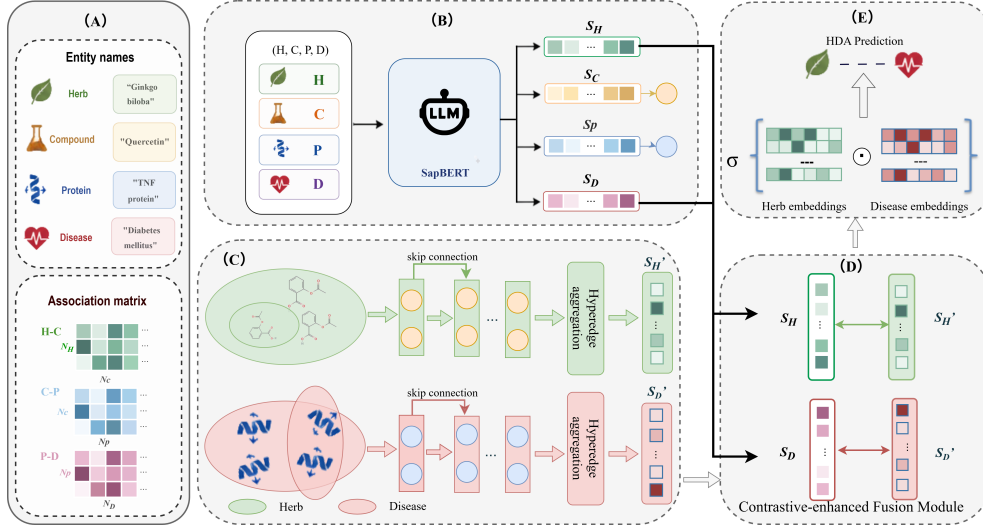


Fig. 1 Overall architecture of LLM-H2G for herb–disease association prediction. (A) Entity names and association matrices are collected for herbs (H), compounds (C), proteins (P) and diseases (D). (B) Entity names are encoded by the biomedical language-model module to obtain semantic representations S_H , S_C , S_P and S_D . (C) Enhanced heterogeneous hypergraph propagation aggregates compound- and protein-level structural information with skip connections, producing structural herb and disease representations S'_H and S'_D . (D) The contrastive-enhanced fusion module aligns and combines semantic and structural representations. (E) The final herb and disease embeddings are used for HDA prediction.

2.4 Notation

We reuse the entity sets, incidence matrices $\{\mathbf{B}^{HC}, \mathbf{B}^{CP}, \mathbf{B}^{PD}\}$, node feature matrices $\{\mathbf{X}^C, \mathbf{X}^P\}$ (with $d_C=64$) and label matrix $\mathbf{Y}$ defined in the Problem definition section. Throughout, d denotes the hidden dimension and L the number of propagation layers.

For any pair of entity types u and v , we write $\mathbf{B}^{uv}$ for the incidence matrix whose rows correspond to type u and columns to type v . The row-normalised propagation operator is

$$\begin{aligned}\hat{\mathbf{A}}^{uv} &= (\mathbf{D}^{uv})^{-1} \mathbf{B}^{uv}, \\ (\mathbf{D}^{uv})_{aa} &= \max\left(\sum_{b=1}^{|v|} (\mathbf{B}^{uv})_{ab}, 1\right),\end{aligned}\tag{1}$$

where the degree is clamped at 1 to avoid division by zero for isolated entities. This normalization follows the broader graph neural network practice of stabilizing message passing over sparse and irregular graph structures [14, 16].

2.5 Biomedical semantic encoder

For every entity we build a short text from its name and encode it with SapBERT, a PubMedBERT-based model specialised for biomedical entity representation; we denote this frozen language-model encoder by $\text{LLM}(\cdot)$ in the equations and Fig 1 [26, 27]. Using the [CLS] token as a pooled sentence embedding, we obtain a 768-dimensional vector for each entity,

$$\mathbf{s}_k^X = \text{LLM}(\text{name}(k)) \in \mathbb{R}^{768}, \quad X \in \{H, C, P, D\},\tag{2}$$

which is computed once offline and kept frozen. A type-specific linear layer projects these embeddings into the hidden space,

$$\mathbf{S}_k^X = \mathbf{s}_k^X \mathbf{W}_{\text{sem}}^X, \quad \mathbf{W}_{\text{sem}}^X \in \mathbb{R}^{768 \times d}, \quad (3)$$

yielding semantic matrices $\mathbf{S}^H, \mathbf{S}^C, \mathbf{S}^P, \mathbf{S}^D$. Only the projections $\mathbf{W}_{\text{sem}}^X$ are trainable.

2.6 Enhanced heterogeneous hypergraph encoder

The structural encoder maintains states for compound and protein nodes, while herbs and diseases act as hyperedges whose embeddings are read out at every layer. The node states are initialised by the semantic embeddings,

$$\mathbf{C}^{(0)} = \mathbf{S}^C, \quad \mathbf{P}^{(0)} = \mathbf{S}^P, \quad (4)$$

so that biomedical semantics enter the propagation from the first layer; when semantic embeddings are unavailable, LLM-H2G falls back to linear projections of the raw features $\mathbf{X}^C$ and $\mathbf{X}^P$. At layer $l \in \{1, \dots, L\}$, the herb and disease hyperedge embeddings are obtained by pooling their incident node states,

$$\mathbf{E}_H^{(l)} = \hat{\mathbf{A}}^{HC} \mathbf{C}^{(l-1)}, \quad \mathbf{E}_D^{(l)} = \hat{\mathbf{A}}^{DP} \mathbf{P}^{(l-1)}. \quad (5)$$

Messages then flow back to the nodes, augmented by a bidirectional compound–protein *bridge* that couples the two relation chains,

$$\begin{aligned} \mathbf{M}_C^{(l)} &= \hat{\mathbf{A}}^{CH} \mathbf{E}_H^{(l)} + \hat{\mathbf{A}}^{CP} \mathbf{P}^{(l-1)}, \\ \mathbf{M}_P^{(l)} &= \hat{\mathbf{A}}^{PD} \mathbf{E}_D^{(l)} + \hat{\mathbf{A}}^{PC} \mathbf{C}^{(l-1)}, \end{aligned} \quad (6)$$

where the second term of each line is the bridge contribution ($\hat{\mathbf{A}}^{CP}$ and $\hat{\mathbf{A}}^{PC}$ are built from $\mathbf{B}^{CP}$ and its transpose via Eq. (1)). Each node state is updated with a linear transformation, a LeakyReLU nonlinearity, a residual connection, dropout and row-wise ℓ_2 normalisation,

$$\begin{aligned}\mathbf{C}^{(l)} &= \text{Norm}\left(\text{Drop}\left(\text{LeakyReLU}(\mathbf{M}_C^{(l)}\mathbf{W}_C) + \mathbf{C}^{(l-1)}\right)\right), \\ \mathbf{P}^{(l)} &= \text{Norm}\left(\text{Drop}\left(\text{LeakyReLU}(\mathbf{M}_P^{(l)}\mathbf{W}_P) + \mathbf{P}^{(l-1)}\right)\right),\end{aligned}\tag{7}$$

with $\mathbf{W}_C, \mathbf{W}_P \in \mathbb{R}^{d \times d}$ and $\text{Norm}(\cdot)$ the row-wise ℓ_2 normalisation.

2.7 Semantic–structural fusion and readout

The final herb and disease representations aggregate the hyperedge embeddings across all L layers and add the corresponding semantic embeddings,

$$\mathbf{Z}^H = \sum_{l=1}^L \mathbf{E}_H^{(l)} + \mathbf{S}^H, \quad \mathbf{Z}^D = \sum_{l=1}^L \mathbf{E}_D^{(l)} + \mathbf{S}^D.\tag{8}$$

Semantics are thus injected at two granularities: implicitly, through the node-level initialisation in Eq. (4), and explicitly, through the hyperedge-level addition in Eq. (8).

2.8 Association prediction and training objective

For a herb–disease pair (h_i, d_j) , the herb representation $\mathbf{z}_i^H$ and disease representation $\mathbf{z}_j^D$ (the i -th and j -th rows of $\mathbf{Z}^H$ and $\mathbf{Z}^D$) are combined and scored by a two-layer perceptron,

$$\begin{aligned}\mathbf{f}_{ij} &= [\mathbf{z}_i^H \parallel \mathbf{z}_j^D \parallel \mathbf{z}_i^H \odot \mathbf{z}_j^D], \\ \hat{y}_{ij} &= \sigma(\mathbf{w}^\top \text{Drop}(\text{ReLU}(\mathbf{W}_p \mathbf{f}_{ij}))),\end{aligned}\tag{9}$$

where $\parallel$ denotes concatenation, $\odot$ the element-wise product, $\mathbf{W}_p \in \mathbb{R}^{d \times 3d}$, $\mathbf{w} \in \mathbb{R}^d$ and σ the sigmoid function. Over a mini-batch $\mathcal{B}$ of labelled pairs the supervised loss is the binary cross-entropy

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{B}|} \sum_{(i,j) \in \mathcal{B}} \left[y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}) \right]. \quad (10)$$

To align the two views of each entity, a contrastive term treats the semantic embedding and the fused representation of the same entity as a positive pair and all other in-batch entities of the same type as negatives [28]. Let $\mathcal{H}_{\mathcal{B}}$ be the distinct herbs appearing in $\mathcal{B}$; with ℓ_2 -normalised rows $\bar{\mathbf{s}}^H, \bar{\mathbf{z}}^H$ and temperature τ , the similarity logits are $\mathbf{Q}_{ab}^H = (\bar{\mathbf{s}}_{i_a}^H)^\top \bar{\mathbf{z}}_{i_b}^H / \tau$ for $i_a, i_b \in \mathcal{H}_{\mathcal{B}}$, and

$$\mathcal{L}_{\text{con}}^H = \frac{1}{2} \left[\text{CE}(\mathbf{Q}^H, \mathbf{y}^*) + \text{CE}(\mathbf{Q}^{H^\top}, \mathbf{y}^*) \right], \quad y_a^* = a, \quad (11)$$

where CE is the cross-entropy and each row is matched to its own column as the positive. The disease term $\mathcal{L}_{\text{con}}^D$ is defined analogously over the distinct diseases in $\mathcal{B}$, and the two are averaged, $\mathcal{L}_{\text{con}} = \frac{1}{2} (\mathcal{L}_{\text{con}}^H + \mathcal{L}_{\text{con}}^D)$. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{con}}, \quad (12)$$

with contrastive weight $\lambda \geq 0$ and temperature $\tau > 0$. Both are dataset-specific hyperparameters selected on the validation folds (the Results section); setting $\lambda=0$ recovers a purely supervised model built on the semantically initialised hypergraph encoder.

2.9 Implementation details and evaluation protocol

The semantic embeddings are produced by the biomedical language model described in the Biomedical semantic encoder subsection, using [CLS] pooling and a maximum token length of 25; they are precomputed once and cached. Sparse hypergraph propagation is carried out in full precision. Models are optimised with AdamW using a learning rate of 1.4×10^{-3} , weight decay 1.5×10^{-4} , batch size 4096 and dropout 0.1. On the primary TCM-suite benchmark we use $d=128$ and $L=7$; the Ethnobotany configuration uses $d=128$ and $L=4$. The contrastive weight λ and temperature τ are tuned per dataset as reported in the Results section. For every dataset, herb–disease pairs are split by five-fold stratified cross-validation with a fixed random seed of 2026 and a 1:1 positive-to-negative ratio, and all reported results are averaged over the five folds.

2.10 Attribution and molecular docking protocol

For molecular interpretation, we apply post hoc gradient attribution to the trained prediction score. For a selected herb–disease pair, the absolute gradient of $\hat{y}_{ij}$ with respect to each incident compound and protein embedding is accumulated across the five trained folds and averaged to obtain compound- and protein-level attribution scores. Candidate mechanistic paths are then ranked by combining the attributed compound and protein scores along the observed compound–protein bridge, so that a high-ranking path must be supported by both the herb–compound and protein–disease sides of the hypergraph.

Molecular docking is used only as an independent structural plausibility check for the highest-attribution compound–protein mediators. Docking is performed with AutoDock Vina. For the COX-2 case, coumarin (PubChem CID: 323) is docked to the PTGS2/COX-2 crystal structure PDB 3LN1; the geometric centre of the co-crystallised celecoxib ligand (residue name CEL) is used as the docking-grid centre. For the TP53 case, α -linolenic acid (PubChem CID: 5280934) is docked to PDB 2VUK;

the geometric centre of the co-crystallised PhiKan083 ligand (residue name P83) is used as the grid centre. The top-ranked Vina poses are inspected for binding-pocket placement and hydrogen-bond contacts, and the Vina affinity scores are reported as qualitative structural support. These docking results are not used as labels during training or evaluation.

3 Results

3.1 Benchmark datasets

We evaluate LLM-H2G on two benchmark datasets, TCM-suite and Ethnobotany, which provide complementary settings for herb-disease association prediction. As summarized in Table 1, TCM-suite contains 1,009 herbs, 1,193 compounds, 7,258 targets and 11,071 diseases, while Ethnobotany contains 19,653 herbs, 71,179 compounds, 359 targets and 4,888 diseases. The two datasets differ markedly in scale and entity composition, allowing us to assess the model under both structured TCM knowledge-graph conditions and broader ethnobotanical conditions. Such heterogeneity is consistent with the diverse entity coverage observed in TCM resources [2, 5, 6] and in general biomedical databases [7–9].

TCM-suite is used as the primary structured benchmark. It provides multi-type biomedical entities covering herbs, compounds, proteins and diseases, making it suitable for evaluating whether LLM-H2G can capture the multi-component and multi-target characteristics of traditional Chinese medicine [1–3]. Its relatively well-organized entity relations also allow the enhanced heterogeneous hypergraph construction to be directly evaluated.

Ethnobotany is included as a more challenging complementary benchmark. It contains a much larger herb and compound vocabulary but fewer targets, reflecting the broader and more heterogeneous nature of ethnobotanical records. In addition, its disease identifiers are less uniformly aligned across relation types, which makes pure

topology-based prediction more difficult. This setting is useful for examining whether large language model-derived semantic enrichment can improve robustness when graph structure is sparse, noisy or partially inconsistent [16, 29].

Together, TCM-suite and Ethnobotany enable a balanced evaluation of LLM-H2G. The former tests the model’s ability to exploit structured multi-relation topology, whereas the latter tests its robustness to heterogeneous entity scales and less regular biomedical annotation.

Table 1 Dataset statistics used in the experimental analysis.

Dataset	Herbs	Compounds	Targets	Diseases
TCM-suite	1009	1193	7258	11071
Ethnobotany	19653	71179	359	4888

3.2 Evaluation metrics

We evaluate each method with AUROC, AUPRC, F1-score, precision and recall. AUROC and AUPRC measure threshold-independent ranking performance, with AUPRC providing a more positive-class-focused view under class imbalance [30, 31]. Precision and recall assess the reliability and coverage of predicted herb–disease associations, and F1 summarizes their balance. Let TP, FP and FN denote true positives, false positives and false negatives, respectively:

$$\begin{aligned}
 \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
 \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
 \text{F1} &= 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.
 \end{aligned} \tag{13}$$

Together, these metrics evaluate LLM-H2G both as a classifier and as a prioritization model for high-confidence HDA discovery.

3.3 Baseline methods

We compare LLM-H2G with representative graph representation and biomedical association prediction methods. The selected baselines cover heterogeneous network embedding and bipartite mutual-information learning [11, 17], multi-instance and graph convolutional biomedical association prediction [18, 32], graph contrastive learning [19], hypergraph neural networks and hypergraph Transformers [21, 22], and TCM-specific compound-target or molecular prioritization models [23, 25].

- **HTINet** [11]: a symptom-related heterogeneous network model that uses node2vec embeddings and supervised classifiers for herb-target prediction.
- **BiGI** [17]: a bipartite graph embedding method that maximizes mutual information between local edge-level and global prototype representations.
- **MilGNet** [18]: a multi-instance heterogeneous graph neural network for drug-disease association prediction with attention-based multi-scale predictors.
- **LHGCE** [32]: an end-to-end heterogeneous graph convolutional encoder that learns drug and disease embeddings by biomedical knowledge-graph message passing.
- **SMGCL** [19]: a graph co-contrastive learning method based on drug-disease, drug-similarity and disease-similarity views.
- **HGNNLDA** [21]: a dual-channel hypergraph neural network that captures high-order connectivity through hypergraph convolution.
- **HGTHDA** [22]: a hypergraph Transformer for herb-disease association prediction using herb-compound and disease-target hypergraphs.
- **HDCTI** [23]: a TCM compound-target prediction model integrating hypergraph convolution, PageRank and multi-head attention.
- **MAHGCN** [25]: a metapath and heterogeneous graph convolution model for TCM small-molecule prioritization.

3.4 Baseline implementation

All methods are evaluated using the same five-fold splits, negative-sampling ratio and metrics defined in the Implementation details and evaluation protocol subsection. LLM-H2G is implemented in Python with PyTorch and trained on a Linux server equipped with an NVIDIA GeForce RTX 4090 GPU. Baseline methods are re-run or evaluated on the same benchmark datasets with identical train/test partitions whenever applicable, so that performance differences mainly reflect model design rather than data partitioning or hardware variation.

3.5 Performance comparison

We evaluated whether LLM-H2G improves herb–disease association prediction compared with representative graph-based and biomedical association prediction methods. All methods are evaluated under the fixed 5-fold H-D split. The reported values are the mean and standard deviation across the five cross-validation folds, where the mean reflects the overall predictive performance and the standard deviation reflects fold-level variability under different train-test partitions.

As shown in Table 2, LLM-H2G achieves the best performance on both datasets across all five evaluation metrics. On the TCM-suite dataset, LLM-H2G obtains 0.9970 AUROC, 0.9970 AUPRC, 0.9727 F1, 0.9713 Precision and 0.9742 Recall. Compared with the strongest baseline results, LLM-H2G improves AUROC by 0.0625, AUPRC by 0.0643, F1 by 0.1124, Precision by 0.1145 and Recall by 0.0984. These gains are substantial given that graph-based, hypergraph-based and TCM-oriented molecular prioritization baselines are already competitive on this structured benchmark [18, 23, 25, 32]. Nevertheless, their performance remains clearly below that of LLM-H2G, suggesting that topology alone is insufficient to fully characterize herb–disease associations.

Table 2 Performance comparison of LLM-H2G and baseline methods on the TCM-suite and Ethnobotany datasets. The best results are marked in bold, and the second-best results are underlined.

Dataset	Method	AUROC $\uparrow$	AUPRC $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
TCM-suite	HTINet	0.9242 $\pm$ 0.0006	0.9208 $\pm$ 0.0004	0.8463 $\pm$ 0.0025	0.8408 $\pm$ 0.0103	0.8522 $\pm$ 0.0151
	BiGI	0.9241 $\pm$ 0.0007	0.9203 $\pm$ 0.0004	0.8464 $\pm$ 0.0020	0.8389 $\pm$ 0.0119	0.8544 $\pm$ 0.0164
	MilGNet	0.9345 $\pm$ 0.0007	0.9327 $\pm$ 0.0006	0.8583 $\pm$ 0.0024	0.8543 $\pm$ 0.0062	0.8625 $\pm$ 0.0102
	LHGCE	0.9344 $\pm$ 0.0008	0.9326 $\pm$ 0.0005	0.8578 $\pm$ 0.0011	0.8568 $\pm$ 0.0017	0.8588 $\pm$ 0.0038
	SMGCL	0.8623 $\pm$ 0.0007	0.8643 $\pm$ 0.0007	0.7790 $\pm$ 0.0030	0.7744 $\pm$ 0.0087	0.7839 $\pm$ 0.0150
	HGNNLDA	0.9242 $\pm$ 0.0005	0.9206 $\pm$ 0.0003	0.8463 $\pm$ 0.0035	0.8406 $\pm$ 0.0100	0.8525 $\pm$ 0.0166
	HGTHDA	0.8623 $\pm$ 0.0008	0.8640 $\pm$ 0.0011	0.7781 $\pm$ 0.0025	0.7777 $\pm$ 0.0083	0.7787 $\pm$ 0.0130
	HDCTI	0.9343 $\pm$ 0.0008	0.9325 $\pm$ 0.0008	0.8603 $\pm$ 0.0010	0.8453 $\pm$ 0.0038	0.8758 $\pm$ 0.0044
	MAHGCN	0.9344 $\pm$ 0.0008	0.9327 $\pm$ 0.0010	0.8584 $\pm$ 0.0018	0.8538 $\pm$ 0.0059	0.8631 $\pm$ 0.0089
	LLM-H2G	0.9970 $\pm$ 0.0001	0.9970 $\pm$ 0.0001	0.9727 $\pm$ 0.0004	0.9713 $\pm$ 0.0049	0.9742 $\pm$ 0.0051
Ethnobotany	HTINet	0.6485 $\pm$ 0.0321	0.6451 $\pm$ 0.0343	0.5726 $\pm$ 0.1202	0.6480 $\pm$ 0.0267	0.5422 $\pm$ 0.1725
	BiGI	0.6431 $\pm$ 0.0214	0.6534 $\pm$ 0.0189	0.4675 $\pm$ 0.1201	0.7099 $\pm$ 0.0469	0.3700 $\pm$ 0.1499
	MilGNet	0.6379 $\pm$ 0.0486	0.6402 $\pm$ 0.0419	0.5541 $\pm$ 0.1668	0.6471 $\pm$ 0.0815	0.5578 $\pm$ 0.2581
	LHGCE	0.6210 $\pm$ 0.0526	0.6187 $\pm$ 0.0490	0.5431 $\pm$ 0.1953	0.6296 $\pm$ 0.0248	0.5376 $\pm$ 0.2416
	SMGCL	0.6488 $\pm$ 0.0407	0.6554 $\pm$ 0.0298	0.5476 $\pm$ 0.1800	0.6594 $\pm$ 0.0798	0.5415 $\pm$ 0.2554
	HGNNLDA	0.6500 $\pm$ 0.0396	0.6603 $\pm$ 0.0308	0.5341 $\pm$ 0.2103	0.6571 $\pm$ 0.0875	0.5412 $\pm$ 0.2843
	HGTHDA	0.6560 $\pm$ 0.0325	0.6611 $\pm$ 0.0245	0.5962 $\pm$ 0.0821	0.6199 $\pm$ 0.0717	0.6247 $\pm$ 0.2141
	HDCTI	0.6162 $\pm$ 0.0566	0.6267 $\pm$ 0.0446	0.5393 $\pm$ 0.1439	0.6426 $\pm$ 0.0339	0.5010 $\pm$ 0.1989
	MAHGCN	0.6065 $\pm$ 0.0532	0.6188 $\pm$ 0.0429	0.4076 $\pm$ 0.2413	0.6792 $\pm$ 0.1040	0.4000 $\pm$ 0.3496
	LLM-H2G	0.8524 $\pm$ 0.0039	0.8640 $\pm$ 0.0030	0.7628 $\pm$ 0.0238	0.7972 $\pm$ 0.0511	0.7396 $\pm$ 0.0800

The advantage of LLM-H2G is more pronounced on the Ethnobotany dataset. LLM-H2G achieves 0.8524 AUROC and 0.8640 AUPRC, whereas the best baseline results are 0.6560 AUROC and 0.6611 AUPRC. This corresponds to absolute improvements of 0.1964 and 0.2029, respectively. Similar improvements are observed for threshold-dependent metrics, with F1 increasing from 0.5962 to 0.7628, Precision from 0.7099 to 0.7972 and Recall from 0.6247 to 0.7396. The larger margin on Ethnobotany is especially informative because this dataset is broader and less uniformly aligned than TCM-suite. In such a setting, methods that mainly rely on graph topology or bipartite structural signals can be sensitive to sparse evidence [11, 17]. Meta-path, similarity-view and hypergraph propagation methods face a similar limitation when the observed network is incomplete or unevenly aligned [18, 19, 22]. By contrast, LLM-H2G incorporates large language model-derived biomedical semantic embeddings, which provide additional entity-level information beyond observed graph edges.

The comparison also reveals how different modeling assumptions affect performance. Network embedding methods such as HTINet and BiGI learn useful low-dimensional representations, but they are limited in capturing high-order biomedical dependencies [11, 17]. Graph convolutional and multi-instance methods such as MilGNet and LHGCE improve relational modeling, yet their representations are still primarily driven by observed graph structure [18, 32]. Hypergraph-based methods, including HGNNLDA, HGTHDA and HDCTI, are better aligned with the multi-component and multi-target nature of TCM, but they do not explicitly integrate biomedical language-model semantics into the representation learning process [21–23]. SMGCL introduces contrastive learning, but its contrastive views are constructed from graph similarity rather than from biomedical textual semantics [19]. These differences help explain why the baselines perform reasonably well on TCM-suite but degrade more noticeably on Ethnobotany.

Overall, the performance comparison demonstrates that the superiority of LLM-H2G comes from the joint modeling of biomedical semantics and heterogeneous hypergraph structure. The semantic component enriches entity representations with biomedical textual priors, while the enhanced hypergraph construction preserves multi-relation dependencies among herbs, compounds, proteins and diseases. The consistent improvements across AUROC, AUPRC, F1, Precision and Recall indicate that LLM-H2G improves both ranking quality and binary association prediction, making it more suitable for prioritizing biologically meaningful herb–disease associations.

3.6 Ablation study

To further investigate the contribution of each key design in LLM-H2G, we conduct an ablation study on both the TCM-suite and Ethnobotany datasets. The purpose of this experiment is to examine whether the performance gain of LLM-H2G mainly comes from biomedical semantic enrichment, enhanced heterogeneous hypergraph construction, or their joint effect. As shown in Fig 2 and Fig 3, we compare the full LLM-H2G model with two ablated variants: LLM-H2G w/o semantics and LLM-H2G w/o enhanced hypergraph. The former removes large language model-derived biomedical semantic embeddings while retaining the LLM-H2G propagation framework, whereas the latter removes the enhanced hypergraph construction and reverts to a simpler H-C/P-D dual-branch hypergraph structure.

The results first show that large language model-derived semantic representations contribute substantially to the predictive performance of LLM-H2G. On the TCM-suite dataset, removing the semantic encoder reduces AUPRC from 0.9970 to 0.9716 and F1 from 0.9727 to 0.8955. A more pronounced performance drop is observed on the Ethnobotany dataset, where AUPRC decreases from 0.8640 to 0.6895. These results indicate that large language model-derived entity embeddings provide complementary information beyond graph topology. This effect is particularly evident in Ethnobotany,

where the topological evidence is relatively sparse and heterogeneous; in such cases, semantic priors from biomedical text help stabilize entity representations and improve herb-disease association prediction.

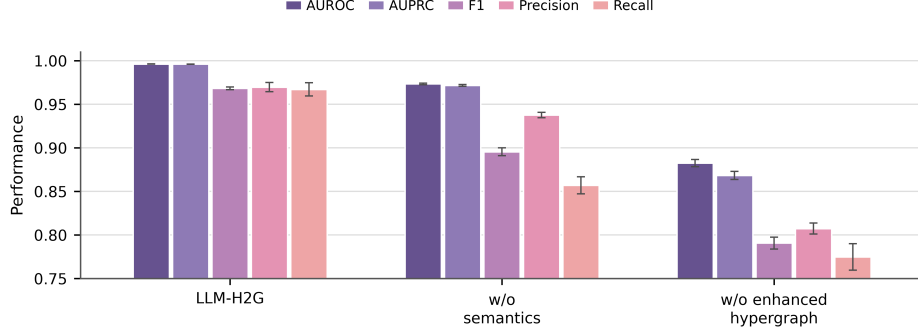


Fig. 2 Ablation analysis on the TCM-suite dataset. Bars indicate the mean performance over five folds, and error bars indicate the standard deviation.

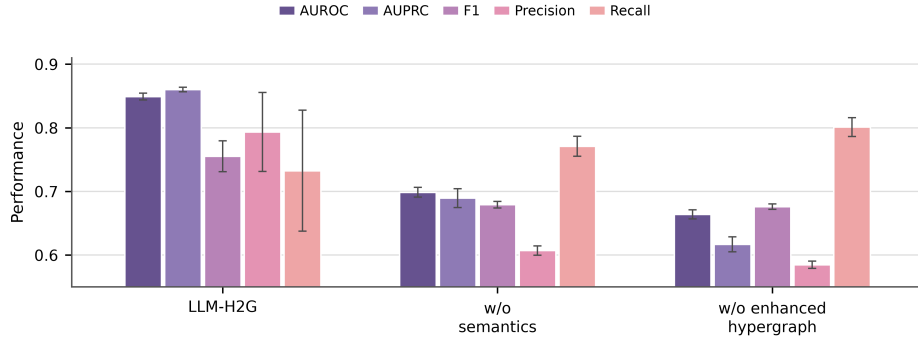


Fig. 3 Ablation analysis on the Ethnobotany dataset. Bars indicate the mean performance over five folds, and error bars indicate the standard deviation.

Removing the enhanced hypergraph construction leads to an even larger degradation. On TCM-suite, LLM-H2G w/o enhanced hypergraph achieves only 0.8825 AUROC, 0.8684 AUPRC and 0.7907 F1. On Ethnobotany, the corresponding AUROC and AUPRC further decrease to 0.6640 and 0.6169, respectively. This result suggests that the proposed hypergraph construction is not merely an auxiliary structural

component, but an important part of the representation learning framework. By integrating herb-compound, compound-protein and protein-disease relations, the enhanced hypergraph enables LLM-H2G to capture multi-relation dependencies that are consistent with the multi-component and multi-target characteristics of TCM. When this construction is removed, the model loses the ability to jointly exploit semantic and heterogeneous topological signals, resulting in substantially weaker ranking performance.

Overall, the ablation study demonstrates that both large language model-derived semantic enrichment and enhanced heterogeneous hypergraph construction are necessary for the strong performance of LLM-H2G. The full model consistently outperforms both ablated variants across the two datasets, supporting the importance of jointly modeling biomedical semantic information and multi-relation hypergraph structure for herb-disease association prediction.

3.7 Parameter analysis

We analyse three model-related hyperparameters that directly affect the proposed architecture: hidden dimension d , propagation depth L , and the contrastive weight-temperature pair (λ, τ) . Single-parameter scans show that $d=128$ provides sufficient representation capacity on both datasets. The best depth differs by dataset: TCM-suite benefits from deeper propagation ($L=7$), whereas Ethnobotany uses $L=4$, consistent with its sparser and less uniformly aligned topology.

We then examine the interaction between contrastive weight and temperature using a two-dimensional grid, with $\lambda \in \{0.002, 0.004, 0.006, 0.008, 0.010\}$ and $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Fig 4, TCM-suite remains strong across a broad region around $\lambda=0.004$, indicating that a moderate contrastive constraint is sufficient when the graph is relatively dense. Ethnobotany is more sensitive and reaches its best AUROC/AUPRC near $\lambda=0.008$ and $\tau=0.1$, suggesting that a stronger and sharper

semantic-structural alignment is useful under noisier and more heterogeneous graph conditions. These results support dataset-specific contrastive settings rather than a single fixed contrastive configuration.

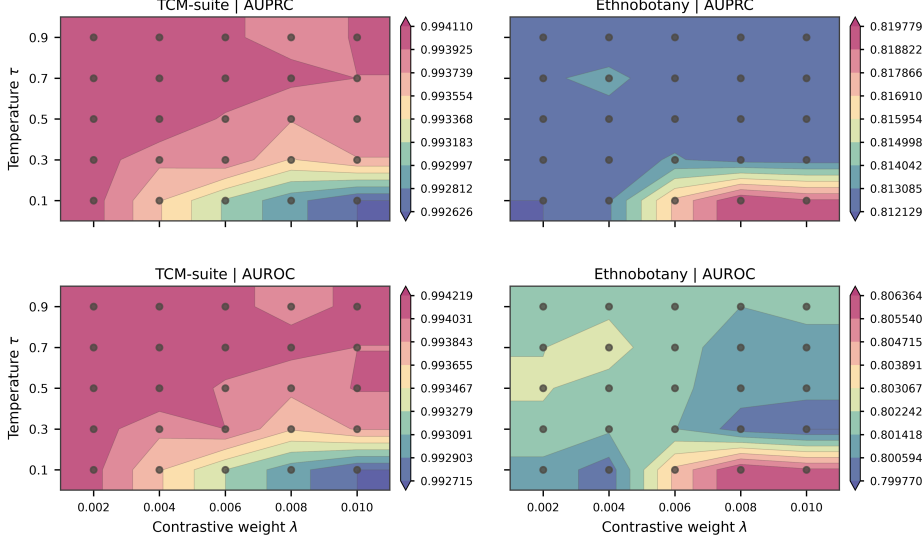


Fig. 4 Two-dimensional sensitivity analysis of the contrastive objective. Contours show single-fold screening performance for AUPRC and AUROC under different contrastive weights λ and temperatures τ . Dots mark the evaluated grid points.

3.8 Robustness to structural sparsity

A central motivation for injecting biomedical semantics is to compensate for sparse or missing topological evidence. To test this claim directly, we stratified the Ethnobotany test pairs by *herb degree*—the number of known disease associations of the herb—and re-evaluated the trained five-fold models within each degree bin. We compare LLM-H2G with HGHDA as a representative structure-driven hypergraph baseline without language-model semantics. Figs 5 and 6 report AUPRC and AUROC as a function of herb degree.

Semantic enrichment is most valuable exactly where structural evidence is scarcest. For herbs with a single known association (degree = 1), LLM-H2G attains 0.499

AUPRC and 0.747 AUROC, whereas HGHDA collapses to 0.152 AUPRC and 0.498 AUROC, i.e. essentially random ranking. As herb degree grows, the structural baseline improves and the gap narrows: the AUPRC gap decreases monotonically from +0.347 (degree 1) and +0.344 (degree 2–3) to +0.288 (degree 4–7) and +0.107 (degree ≥ 8). In AUROC, LLM-H2G remains stable at 0.75–0.79 across all bins, whereas HGHDA stays near 0.50 for low-degree herbs and reaches only 0.56 for the densest bin, confirming that structure-only evidence is insufficient for ranking associations of poorly connected herbs. These trends provide direct, quantitative evidence that language-model-derived semantics act as an effective prior that stabilises entity representations under structural sparsity, complementing the aggregate ablation results.

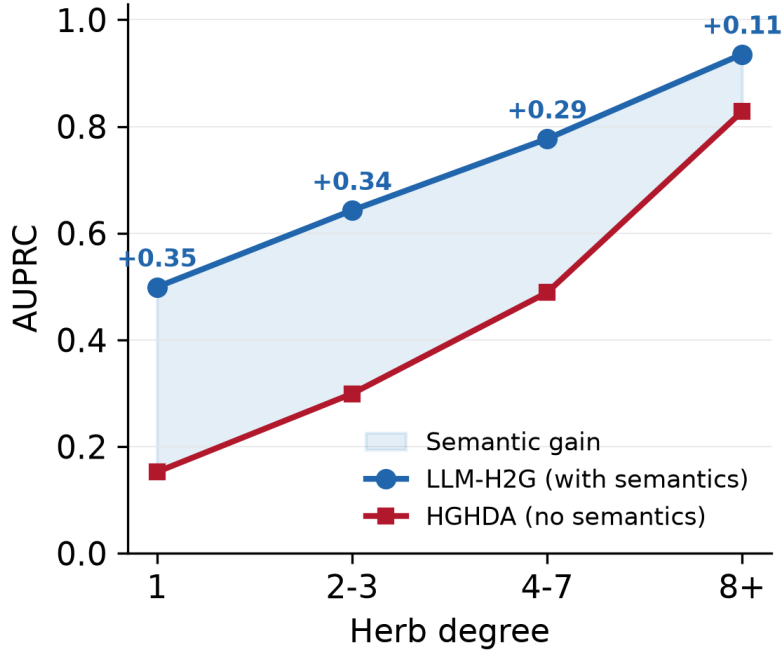


Fig. 5 AUPRC by herb degree on the Ethnobotany dataset. Test herb–disease pairs are grouped by herb degree (number of known disease associations per herb), and AUPRC is computed within each bin for LLM-H2G and the structure-driven hypergraph baseline HGHDA. Blue annotations report the performance gap (LLM-H2G – HGHDA), which is largest for the sparsest herbs and shrinks as more associations become available.

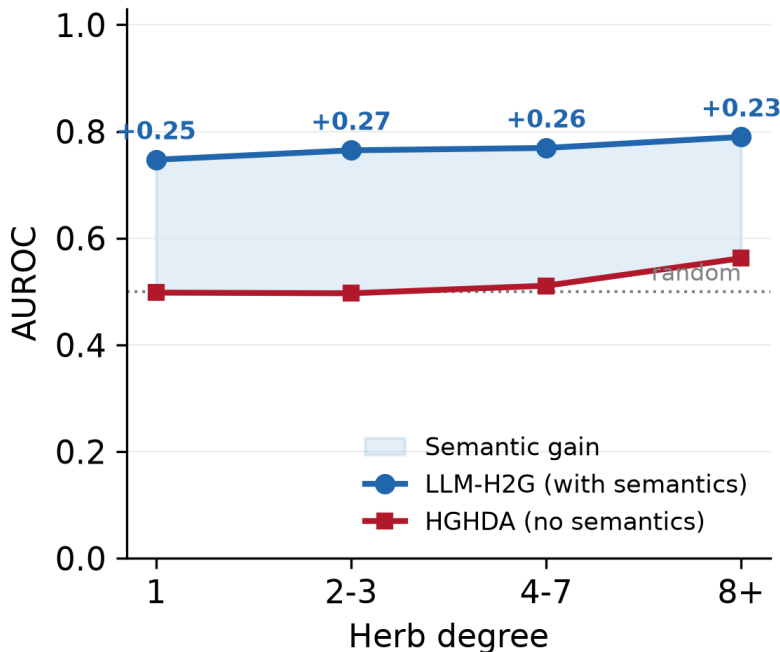


Fig. 6 AUROC by herb degree on the Ethnobotany dataset, computed within the same herb-degree bins as Fig 5. LLM-H2G stays well above the structure-driven hypergraph baseline HGHDA across all bins; HGHDA remains close to random ranking (dotted line, AUROC = 0.5) for low-degree herbs. Blue annotations report the performance gap (LLM-H2G – HGHDA).

3.9 Mechanistic interpretation of known associations

To test whether LLM-H2G can provide molecularly interpretable evidence for high-confidence herb–disease links, we first analysed two known positive associations from the TCM-suite benchmark. Following the protocol in the Attribution and molecular docking protocol subsection, we used gradient attribution over the compound and protein embeddings to identify the molecular mediators that contribute most strongly to each prediction. For the association between the spine of Chinese honeylocust and inflammatory disease, the attribution identified coumarin as the leading bridging compound and PTGS2 (cyclooxygenase-2, COX-2) as the dominant mediating protein, consistent with the central role of COX-2 in inflammation. For the association between common flax seed and colorectal adenocarcinoma, α -linolenic acid was highlighted

as the leading compound acting on the tumour suppressor TP53. In both cases, the mediating protein received a markedly larger attribution weight than any individual compound, indicating that the model concentrates its evidence on a specific molecular mediator rather than spreading it diffusely across all constituents of the herb.

To provide independent structural support for these attributed compound–protein mediators, we docked the top attributed compound of each case into its predicted protein (Fig. 7). Coumarin docked into PTGS2/COX-2 (PDB 3LN1) with a Vina affinity of -7.47 kcal/mol, and α -linolenic acid docked into TP53 (PDB 2VUK) with a Vina affinity of -5.12 kcal/mol. Both compounds adopt stable poses within the target binding pockets and form multiple hydrogen bonds (2.7 – 4.2 Å) with pocket residues—TYR371, HIS193, THR192 and ALA188 in panel A, and GLU224, SER227, CYS229 and THR231 in panel B. The docked geometries are consistent with the compound–protein relationships emphasised by the model, suggesting that the attributed paths correspond to plausible physical interactions rather than only statistical correlations.

3.10 Discovery of unannotated therapeutic paths

We next asked whether the model could identify plausible *new* paths that are absent from the curated positive set. Under the positive–unlabelled setting adopted throughout this work, high-scoring unannotated herb–disease pairs are best interpreted as candidate missing associations rather than as confirmed negatives. Using the final five-fold ensemble on the Ethnobotany dataset, we ranked all unannotated candidates and, for a panel of well-studied medicinal herbs, retrieved the highest-ranked disease predicted for each herb (Table 3).

Although none of the pairs in Table 3 is annotated in the benchmark, every one of them is independently corroborated by pharmacological literature, and the recovered associations span clinically distinct therapeutic areas: hepatoprotection of milk thistle (*Silybum marianum*) against drug-induced liver injury [33], ginseng (*Panax ginseng*)

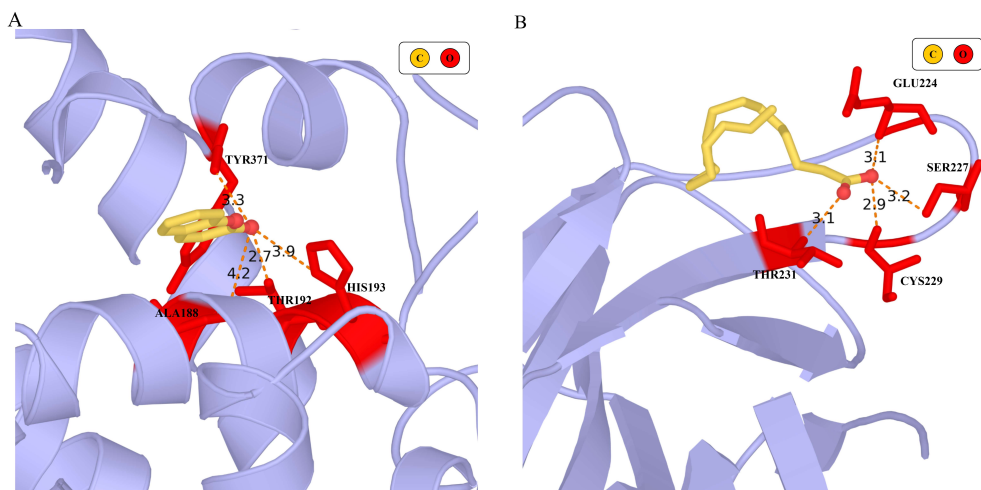


Fig. 7 Molecular docking of the top model-attributed compound into its predicted protein for two representative herb-disease predictions. (A) Coumarin (PubChem CID: 323), the leading bridging compound for the predicted spine-of-Chinese-honeylocust-inflammatory-disease association, docked into PTGS2/COX-2 (PDB 3LN1; Vina affinity, -7.47 kcal/mol). (B) α -Linolenic acid (PubChem CID: 5280934), the leading compound for the predicted common-flax-seed-colorectal-adenocarcinoma association, docked into TP53 (PDB 2VUK; Vina affinity, -5.12 kcal/mol). Ligand atoms are coloured by element (carbon in yellow, oxygen in red); dashed lines denote hydrogen bonds with the labelled binding-pocket residues, annotated with their distances in Å.

for erectile dysfunction [34], turmeric (*Curcuma longa*) for colorectal neoplasms [35], danshen (*Salvia miltiorrhiza*) for venous thromboembolism [36], *Scutellaria baicalensis* for lung cancer [37], ginger (*Zingiber officinale*) for nausea [38], ashwagandha (*Withania somnifera*) for anxiety [39], and green tea (*Camellia sinensis*) for prostate cancer [40]. The consistent agreement between these top-ranked predictions and documented pharmacology—across distinct herbs and unrelated diseases—indicates that LLM-H2G can recover biologically plausible associations that are merely missing from the benchmark labels. We present these as illustrative, literature-supported examples and do not claim exhaustive herb-level specificity.

As a focused example of a newly recovered therapeutic path, we further examined the predicted association between *Ginkgo biloba* (ncbitaxon:3311) and Radiation Injuries (MeSH:D011832), which ranked 13th among all unannotated Ethnobotany

Table 3 High-confidence *de novo* herb–disease predictions from LLM-H2G on the Ethnobotany dataset for well-studied medicinal herbs. All listed pairs are absent from the benchmark positive set (unannotated) yet are corroborated by independent pharmacological literature, indicating that top-ranked predictions can recover biologically plausible associations. Scores are the mean over the final five-fold ensemble.

Herb	Predicted disease (MeSH)	Score	Literature
<i>Silybum marianum</i>	Chemical- and Drug-Induced Liver Injury (D056486)	0.996	[33]
<i>Panax ginseng</i>	Erectile Dysfunction (D007172)	0.995	[34]
<i>Curcuma longa</i>	Colorectal Neoplasms (D015179)	0.992	[35]
<i>Salvia miltiorrhiza</i>	Venous Thromboembolism (D054556)	0.988	[36]
<i>Scutellaria baicalensis</i>	Lung Neoplasms (D008175)	0.944	[37]
<i>Zingiber officinale</i>	Nausea (D009325)	0.931	[38]
<i>Withania somnifera</i>	Anxiety Disorders (D001008)	0.902	[39]
<i>Camellia sinensis</i>	Prostatic Neoplasms (D011471)	0.832	[40]

candidates with a mean ensemble score of 0.998 (standard deviation 0.001). This prediction is not an isolated output: two closely related MeSH terms were also scored highly for *Ginkgo biloba*, namely Radiation Injuries, Experimental (rank 2, 0.999) and Abnormalities, Radiation-Induced (rank 10, 0.998), showing that LLM-H2G recovers a coherent radiation-injury concept cluster rather than a single spurious pair.

Independent PubMed-indexed studies support the plausibility of this prediction. *Ginkgo biloba* extract has been reported to protect against ionizing-radiation-induced oxidative organ damage in rats and to attenuate radiation-induced dermatitis through antioxidant-related effects [41, 42]. Additional animal studies showed protective effects of *Ginkgo biloba* supplementation or extract against radiation-induced oxidative injury in the lens and against technetium-99m-sestamibi-induced oxidative stress and lens damage [43, 44]. At the mechanistic level, EGb 761 and its constituents exhibited anticlastogenic activity in irradiated rats, while ginkgolide B was recently shown to alleviate radiation-induced brain injury by reducing oxidative stress and neuronal apoptosis through DCC/MST1-related signaling [45, 46]. Together, these findings indicate that the unannotated *Ginkgo biloba*–Radiation Injuries prediction is biologically plausible and may represent a missing association in the benchmark positive set.

4 Discussion

This study presents LLM-H2G for herb–disease association prediction. The model integrates large language model-derived entity semantics with an enhanced heterogeneous hypergraph over herbs, compounds, proteins and diseases, thereby capturing both textual biomedical meaning and high-order pharmacological structure. These two aspects are central to network pharmacology and TCM systems modeling [1, 2], and they are naturally compatible with hypergraph representation learning [20]. The contrastive alignment objective further encourages consistency between semantic and hypergraph-derived representations, improving robustness across datasets with different levels of sparsity and heterogeneity [28, 29].

Experiments on TCM-suite and Ethnobotany show that LLM-H2G consistently outperforms representative network embedding, graph neural network, graph contrastive learning and hypergraph baselines. The comparison covers established network embedding and graph neural network settings [13, 14], as well as graph contrastive and hypergraph learning settings [19, 20]. Ablation, parameter and sparsity analyses confirm that both biomedical semantic enrichment and enhanced hypergraph construction are necessary for the observed gains. The case studies further show that LLM-H2G can support both mechanistic interpretation of known associations and discovery of literature-supported unannotated therapeutic paths, exemplified by the coherent *Ginkgo biloba*–Radiation Injuries prediction. These results indicate that jointly modeling biomedical semantics and multi-relation hypergraph structure is a promising direction for interpretable and biologically meaningful herb–disease association discovery [3, 5].

5 Conclusions

LLM-H2G integrates large language model-derived biomedical semantics with enhanced heterogeneous hypergraph learning for herb–disease association prediction. Across two benchmark datasets, the model consistently improves ranking and threshold-based prediction over representative network embedding, graph neural network, graph contrastive learning and hypergraph baselines. The ablation, sparsity and case-study analyses further indicate that semantic enrichment and multi-relation hypergraph construction provide complementary benefits, especially under sparse topological evidence. These results support semantic–structural alignment as a practical strategy for interpretable and biologically meaningful herb–disease association discovery.

Supporting information

Supporting information, source code, processed data and additional files are available as described in the Availability of data and materials statement.

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The source code and processed data are available at <https://github.com/basehc/LLM-H2G>. All data needed to reproduce the analyses are included in the repository or can be obtained from the public source databases cited in the manuscript. Raw databases and pretrained language-model weights should be obtained from their original public sources.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the National Natural Science Foundation of China (62373348 and 62302495), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2023D01E15 and 2025D01A136), the Tianshan Talent Training Program (2023TSYCLJ0021), and the Xinjiang Leading Talent Recruitment Program (XJRC-2025-KJ-YJ-CXPT-103). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions. J.Z. and H.Y. conceived the study and designed the model. J.Z., W.C. and J.Y. performed the experiments and figure preparation. Y.Y., Z.C., D.L. and P.H. contributed data processing, literature validation and case analysis. H.Y. and L.H. supervised the study. All authors contributed to manuscript writing and approved the final draft.

Acknowledgements. The authors thank the contributors of the public biomedical databases and open-source software packages used in this study.

References

- [1] Hopkins, A.L.: Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4**(11), 682–690 (2008) <https://doi.org/10.1038/nchembio.118>
- [2] Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Li, P., Guo, Z., Tao, W., Yang, Y., *et al.*: TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *Journal of Cheminformatics* **6**, 13 (2014) <https://doi.org/10.1186/1758-2946-6-13>
- [3] Liu, Z., Guo, F., Wang, Y., Li, C., Zhang, X., Li, H., Diao, L., Gu, J., Wang, W., Li, D., *et al.*: BATMAN-TCM: a bioinformatics analysis tool for molecular mechanism of traditional chinese medicine. *Scientific Reports* **6**, 21146 (2016) <https://doi.org/10.1038/srep21146>

- [4] Hu, L., Zhang, M., Hu, P., Zhang, J., Niu, C., Lu, X., Jiang, X., Ma, Y.: Dual-channel hypergraph convolutional network for predicting herb–disease associations. *Briefings in Bioinformatics* **25**(2), 067 (2024) <https://doi.org/10.1093/bib/bbae067>
- [5] Wu, Y., Zhang, F., Yang, K., Fang, S., Bu, D., Li, H., Sun, L., Hu, H., Gao, K., Wang, W., *et al.*: SymMap: an integrative database of traditional chinese medicine enhanced by symptom mapping. *Nucleic Acids Research* **47**(D1), 1110–1117 (2018) <https://doi.org/10.1093/nar/gky1021>
- [6] Huang, L., Xie, D., Yu, Y., Liu, H., Shi, Y., Shi, T., Wen, C.: TCMID 2.0: a comprehensive resource for TCM. *Nucleic Acids Research* **46**(D1), 1117–1120 (2017) <https://doi.org/10.1093/nar/gkx1028>
- [7] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., *et al.*: PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**(D1), 1102–1109 (2018) <https://doi.org/10.1093/nar/gky1033>
- [8] Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.*: DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**(D1), 1074–1082 (2017) <https://doi.org/10.1093/nar/gkx1037>
- [9] Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I.: DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, 028 (2015) <https://doi.org/10.1093/database/bav028>
- [10] Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S.,

- Doncheva, N.T., Legeay, M., Fang, T., Bork, P., *et al.*: The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**(D1), 605–612 (2020) <https://doi.org/10.1093/nar/gkaa1074>
- [11] Wang, N., Li, P., Hu, X., Yang, K., Peng, Y., Zhu, Q., Zhang, R., Gao, Z., Xu, H., Liu, B., *et al.*: Herb target prediction based on representation learning of symptom related heterogeneous network. *Computational and Structural Biotechnology Journal* **17**, 282–290 (2019) <https://doi.org/10.1016/j.csbj.2019.02.002>
- [12] Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710 (2014). <https://doi.org/10.1145/2623330.2623732> . <https://doi.org/10.1145/2623330.2623732>
- [13] Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864 (2016). <https://doi.org/10.1145/2939672.2939754> . <https://doi.org/10.1145/2939672.2939754>
- [14] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations* (2017). <https://arxiv.org/abs/1609.02907>
- [15] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: *International Conference on Learning Representations* (2018). <https://arxiv.org/abs/1710.10903>
- [16] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning*

Systems **32**(1), 4–24 (2019) <https://doi.org/10.1109/TNNLS.2020.2978386>

- [17] Cao, J., Lin, X., Guo, S., Liu, L., Liu, T., Wang, B.: Bipartite graph embedding via mutual information maximization. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 635–643 (2021). <https://doi.org/10.1145/3437963.3441783>
- [18] Gu, Y., Zheng, S., Zhang, B., Kang, H., Li, J.: Milgnet: a multi-instance learning-based heterogeneous graph network for drug repositioning. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 430–437 (2022). <https://doi.org/10.1109/BIBM55620.2022.9995152> . IEEE
- [19] Gao, Z., Ma, H., Zhang, X., Wang, Y., Wu, Z.: Similarity measures-based graph co-contrastive learning for drug–disease association prediction. *Bioinformatics* **39**(6), 357 (2023) <https://doi.org/10.1093/bioinformatics/btad357>
- [20] Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3558–3565 (2019). <https://doi.org/10.1609/aaai.v33i01.33013558> . <https://doi.org/10.1609/aaai.v33i01.33013558>
- [21] Liu, D., Li, X., Zhang, L., Hu, X., Zhang, J., Liu, Z., Deng, L.: Hgnnlda: predicting lncrna-drug sensitivity associations via a dual channel hypergraph neural network. *IEEE/ACM transactions on computational biology and bioinformatics* **20**(6), 3547–3555 (2023) <https://doi.org/10.1109/TCBB.2023.3302468>
- [22] Gao, S., Qiao, Y., Zhang, J.: Hypergraph transformer for herb–disease association prediction. In: Proceedings of the 9th International Conference on Electronic Information Technology and Computer Engineering, pp. 1077–1081 (2025). <https://doi.org/10.1145/3766671.3766857>

- [23] Qiao, Y., Hu, L., Zhang, J., Hu, P., Luo, X.: Identifying novel therapeutic targets of natural compounds in traditional chinese medicine herbs with hypergraph representation learning. *Briefings in Bioinformatics* **26**(4), 399 (2025) <https://doi.org/10.1093/bib/bbaf399>
- [24] Zheng, D., Qin, C., Wang, Z., Li, Y., Liu, Y., Wang, Y., Kan, H., Hu, J.: iCAM-Net: Interpretable herb-disease association prediction via cross-channel attention and molecular interaction signals. *Phytomedicine* **148**, 157491 (2025) <https://doi.org/10.1016/j.phymed.2025.157491>
- [25] Sheng, N., Liu, J., Fang, W., Zheng, C., Ma, Y., Sun, Y., Chen, H.: Metapath2vec and attention-driven heterogeneous graph learning for prioritizing tcm-derived small molecules in gastric cancer. *Chinese Medicine* **21**(1), 158 (2026) <https://doi.org/10.1186/s13020-026-01431-y>
- [26] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* **3**(1), 1–23 (2022) <https://doi.org/10.1145/3458754>
- [27] Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238. Association for Computational Linguistics, Online (2021). <https://doi.org/10.18653/v1/2021.naacl-main.334> . <https://www.aclweb.org/anthology/2021.naacl-main.334>
- [28] Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018) [arXiv:1807.03748](https://arxiv.org/abs/1807.03748)

- [29] Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 213–220 (2008). <https://doi.org/10.1145/1401890.1401920> . <https://doi.org/10.1145/1401890.1401920>
- [30] Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006). <https://doi.org/10.1145/1143844.1143874> . <https://doi.org/10.1145/1143844.1143874>
- [31] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE **10**(3), 0118432 (2015) <https://doi.org/10.1371/journal.pone.0118432>
- [32] Ghasemi, S., Lakizadeh, A.: Drug-disease association prediction based on end-to-end multi-layer heterogeneous graph convolutional encoders. Informatics in Medicine Unlocked **37**, 101177 (2023) <https://doi.org/10.1016/j.imu.2023.101177>
- [33] Tighe, S.P., Akhtar, D., Iqbal, U., Ahmed, A.: Chronic liver disease and silymarin: A biochemical and clinical review. Journal of Clinical and Translational Hepatology **8**(4), 454–458 (2020) <https://doi.org/10.14218/JCTH.2020.00012>
- [34] Jang, D.-J., Lee, M.S., Shin, B.-C., Lee, Y.-J., Ernst, E.: Red ginseng for treating erectile dysfunction: a systematic review. British Journal of Clinical Pharmacology **66**(4), 444–450 (2008) <https://doi.org/10.1111/j.1365-2125.2008.03236.x>
- [35] Sharma, R.A., Euden, S.A., Platton, S.L., Cooke, D.N., Shafayat, A., Hewitt, H.R., Marczyklo, T.H., Morgan, B.: Phase I clinical trial of oral curcumin: biomarkers of systemic activity and compliance. Clinical Cancer Research **10**(20),

- 6847–6854 (2004) <https://doi.org/10.1158/1078-0432.CCR-04-0744>
- [36] Pei, G., Wang, L., Yang, C., Xu, L., Li, W., Cui, W., Li, H.: Study on the effect of wine processing on the anti-thrombotic activity of *Salvia miltiorrhiza*. *Journal of Ethnopharmacology*, 121325 (2026) <https://doi.org/10.1016/j.jep.2026.121325>
- [37] Gao, J., Morgan, W.A., Sánchez-Medina, A., Corcoran, O.: The ethanol extract of *Scutellaria baicalensis* and the active compounds induce cell cycle arrest and apoptosis including upregulation of p53 and bax in human lung cancer cells. *Toxicology and Applied Pharmacology* **254**(3), 221–228 (2011) <https://doi.org/10.1016/j.taap.2011.03.016>
- [38] Thomson, M., Corbin, R., Leung, L.: Effects of ginger for nausea and vomiting in early pregnancy: a meta-analysis. *Journal of the American Board of Family Medicine* **27**(1), 115–122 (2014) <https://doi.org/10.3122/jabfm.2014.01.130167>
- [39] Chandrasekhar, K., Kapoor, J., Anishetty, S.: A prospective, randomized double-blind, placebo-controlled study of safety and efficacy of a high-concentration full-spectrum extract of ashwagandha root in reducing stress and anxiety in adults. *Indian Journal of Psychological Medicine* **34**(3), 255–262 (2012) <https://doi.org/10.4103/0253-7176.106022>
- [40] Kumar, N.B., Pow-Sang, J., Egan, K.M., Spiess, P.E., Dickinson, S., Salup, R., Helal, M., McLarty, J.: Randomized, placebo-controlled trial of green tea catechins for prostate cancer prevention. *Cancer Prevention Research* **8**(10), 879–887 (2015) <https://doi.org/10.1158/1940-6207.CAPR-14-0324>
- [41] Sener, G., Kabasakal, L., Atasoy, B.M., Erzik, C., Velioglu-Ogunc, A., Cetinel, S.,

- Gedik, N., Yegen, B.C.: Ginkgo biloba extract protects against ionizing radiation-induced oxidative organ damage in rats. *Pharmacological Research* **53**(3), 241–252 (2006) <https://doi.org/10.1016/j.phrs.2005.11.006>
- [42] Yirmibesoglu, E., Karahacioglu, E., Kilic, D., Lortlar, N., Akbulut, G., Omeroglu, S.: The protective effects of Ginkgo biloba extract (EGb-761) on radiation-induced dermatitis: an experimental study. *Clinical and Experimental Dermatology* **37**(4), 387–394 (2012) <https://doi.org/10.1111/j.1365-2230.2011.04253.x>
- [43] Okumus, S., Taysi, S., Orkmez, M., Saricicek, E., Demir, E., Adli, M., Al, B.: The effects of oral Ginkgo biloba supplementation on radiation-induced oxidative injury in the lens of rat. *Pharmacognosy Magazine* **7**(26), 141–145 (2011) <https://doi.org/10.4103/0973-1296.80673>
- [44] Khedr, M.H., Shafaa, M.W., Abdel-Ghaffar, A., Saleh, A.: Radioprotective efficacy of Ginkgo biloba and Angelica archangelica extract against technetium-99m-sestamibi induced oxidative stress and lens injury in rats. *International Journal of Radiation Biology* **94**(1), 37–44 (2018) <https://doi.org/10.1080/09553002.2018.1407463>
- [45] Alaoui-Youssefi, A., Lamproglou, I., Drieu, K., Emerit, I.: Anticlastogenic effects of Ginkgo biloba extract (EGb 761) and some of its constituents in irradiated rats. *Mutation Research* **445**(1), 99–104 (1999) [https://doi.org/10.1016/s1383-5718\(99\)00139-4](https://doi.org/10.1016/s1383-5718(99)00139-4)
- [46] Yang, T., Du, X., Xu, L.: Radioprotective effect of Ginkgolide B on brain: the mediating role of DCC/MST1 signaling. *International Journal of Radiation Biology* **100**(3), 371–384 (2024) <https://doi.org/10.1080/09553002.2023.2281515>