Graphormer-Integrated Interaction Network Construction and Action Prediction of Major Bupleurum Chinense Compounds and Targets

Jinjin Li[1]
[1]College of Pharmacy and Chemical Engineering, Shaanxi A&F Technology University, Yangling 712100, Shaanxi, China
Corresponding author: Jinjin Li (e-mail: yzyhljj@163.com)

ABSTRACT To address the low accuracy of multi-component-multi-target interaction prediction in Bupleurum chinense caused by complex molecular structures and limited target associations, this paper employs advanced graph models combined with biological network topology information. The same graph-based representation strategy is relevant to materials informatics, including the screening of molecular or polymer structures for electromagnetic sensing, dielectric response, and functional coating design. Specifically, RDKit is used to construct chemically defined molecular graphs, with atoms as nodes and covalent bonds as edges, and to encode multidimensional atomic features. These graphs are directly input into Graphormer. Graphormer then incorporates a fully connected attention mechanism that integrates topological distances and spatial geometric information from RDKit graphs to extract higher-order molecular representations. Node2Vec is used to embed the protein-protein interaction network in a low-dimensional space and identify potential target pathways. Finally, compound and protein embeddings are concatenated and fed into a two-channel neural network to predict interaction probabilities. With negative sampling for training, the method achieves stable average AUROC values of 0.91-0.93 and AUPRC values of 0.88-0.90. Even when only 10% of targets are visible, the F1-score reaches 0.683±0.021 and Coverage Rate@50 reaches 76.4%, indicating strong inductive reasoning under sparse knowledge conditions.

INDEX TERMS graphormer, molecular graph, protein interaction network, chemical sensing materials, electromagnetic material screening

I. INTRODUCTION

As a traditional Chinese herbal medicine used to soothe liver qi and relieve depression, Bupleurum chinense can be effective due to the synergistic regulation of multi-target networks by its various saponins [1,2]. However, these compounds are complex and their pathways of action are cryptic, making it challenging to reveal the whole pharmacological behavior of Bupleurum chinense through traditional pharmacological models [3]. Accurately characterizing the interactions between these compounds and their targets not only helps to explain the scientific essence of the "multi-component-multi-target" synergistic intervention idea in Traditional Chinese Medicine (TCM), but also offers a novel approach to the innovation of drug design. Similarly, this interaction characterization provides a deeper understanding of the synergistic effects between dyes and fabric treatments, which is critical for advancing material science and innovating new, functional products [4 -6]. Constructing high-precision predictive models is increasingly becoming the link connecting the chemical basis of traditional Chinese medicines and their biological functional responses.

When constructing compound-target association models of Bupleurum chinense, most methods encounter challenges in capturing the long-range electronic effects and spatial topological dependencies between atoms in the molecule [7,8]. When the target protein is situated at the margins of an existing knowledge graph, the signaling pathways connecting it to other molecules are destroyed and prediction gaps appear [9,10]. Furthermore, most joint modeling strategies statically splice the molecular representations with protein features and fail to dynamically integrate the semantic-level information. This makes them have low generalization performance when facing low-similarity targets and novel ligands [11,12]. All these problems aggregate and make it hard to infer high-precision interaction probabilities based on chemical structures.

This paper aims to build a high-precision target recognition system for major compounds in Bupleurum chinense. It attempts to solve the above problems to overcome the following limitations of existing methods:

(1) Existing methods focus on structural characterization and network inference of target compounds but are stuck in

the bottleneck.

(2) An end-to-end prediction framework combining graph morphology analyzer and PPI topology embedding is proposed.

(3) Compared with previous works, this paper introduces advanced graph attention mechanism into the field of active ingredient analysis in traditional Chinese medicine.

(4) The Rational Discovery Kit (RDKit) was used to accurately generate the target information of molecular map structure. A multidimensional feature system including atomic type, charge, hybridization state and aromaticity was designed to ensure the chemical rationality of the input information. The Graphormer model (A graph neural network capable of capturing long-range spatial and electron-dependent relationships within molecules) was employed to capture long-range intramolecular dependencies and enhance high-order characterization capabilities. Node2Vec (A network embedding method based on random walks is proposed for compressing the topology of protein interaction networks) was also used to perform nonlinear dimensionality reduction on the PPI network in the STRING database, completing missing functional association pathways and enhancing the completeness of the target space. On this basis, a dual-channel feedforward network was constructed to process compound and protein embedding vectors separately, dynamically merging them through learnable weights to output interaction probability scores. A frequency-weighted negative sampling strategy was introduced during the training phase to alleviate class imbalance and improve the model's sensitivity to low-frequency targets. The resulting "Bupleurum chinense compound-target" prediction system is not only highly interpretable but also extensible to the study of the mechanisms of action of other traditional Chinese medicine ingredients, providing a computable and iterative technical path for the modernization of traditional Chinese medicine.

II. COLLABORATIVE CONSTRUCTION OF BUPLEURUM CHINENSE COMPOUND MOLECULAR REPRESENTATION AND TARGET NETWORK

Figure 1 shows a target prediction model for Bupleurum chinense compounds that integrates Graphormer and protein interaction networks. The molecular processing pipeline uses RDKit to construct a molecular graph of radix stellaviaes, and the Graphormer encoder extracts a molecular representation including longrange atomic interactions. The protein processing pipeline uses Node2Vec to perform topological embedding of the PPI network. The dual-channel features undergo dynamic semantic alignment in the fusion module, and the interaction probability is output through a neural network optimized by negative sampling. This architecture enables cross-hierarchical modeling from the atomic scale to protein networks, providing a computational framework for analyzing the multi-component and multi-target mechanisms of action of Bupleurum chinense.

A. MOLECULAR GRAPH STRUCTURE ANALYSIS AND ATOMIC-LEVEL FEATURE ENCODING OF MAJOR BUPLEURUM CHINENSE COMPOUNDS

1) Atomic-Level Topological Modeling and Feature Space Construction of Molecular Graph Structures

In the structural analysis of the main active components of Bupleurum chinense, the SMILES strings of compounds such as radix stellaviaes D, A, and C were used as input, and the RDKit cheminformatics toolkit was used to perform automated molecular graph analysis. The process first converts the linear symbol sequence into graph-structured data with clear atomic connection relationships, where each non-hydrogen atom is regarded as a node in the graph, and covalent bonds are connected as edges to form an undirected graph G = (V, E), where V represents the set of atomic nodes and E is the set of chemical bond edges. On this basis, a twelve-dimensional physicochemical feature vector is extracted for each atomic node, including atomic number, number of valence electrons, degree of hybridization (sp3, sp2, sp, etc.), aromaticity, formal charge, hydrogen bonding capacity, conjugation state, atomicity, van der Waals radius, electronegativity correction, lone pair presence, and ring member properties, forming the initial node embedding.

In order to make the molecular geometry structure more realistic, a topological fingerprint extension mechanism based on Morgan algorithm is introduced. This mechanism finds local substructure patterns by setting neighborhood radius of 2 and adjusting initial values of interatomic bond length and angle. Then, this paper implements energy minimization optimization under the guidance of force field function (Universal force field for example). The atomic coordinate update equation is solved iteratively as follows:

$$x_i^{(t+1)} = x_i^{(t)} - \eta \nabla_{x_i} U(\{x_j\}) \tag{1}$$

x_i represents the spatial coordinates of the i-th atom, U is the total potential energy as a function of the system including bond stretching, angle bending, dihedral torsion and non-bonded interaction potential terms, and η is the learning rate parameter. The optimization process converges to a local energy minimum within the maximum number of iterations and returns a molecular structure with a certain degree of three-dimensional spatial orientation. The final three-dimensional coordinates are embedded into node positional attributes to facilitate the calculation of the relative distance matrix in the following graph attention mechanism and preserve certain spatial folding information. Finally, the entire pipeline achieves a certain mapping from a one-dimensional character sequence to a three-dimensional graph representation and provides the geometric basis for capturing the long-range spatial effect between the glycosyl side chains and steroid core in complex saponin backbones.

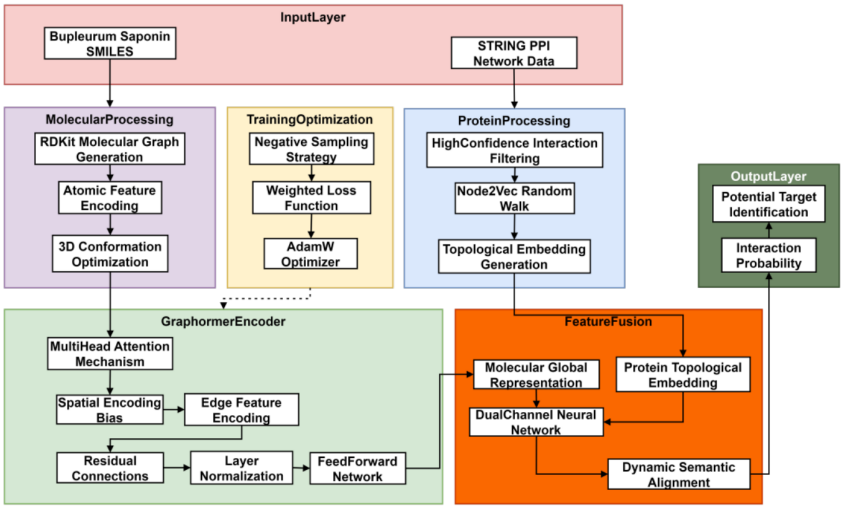

FIGURE 1. Target prediction model for Bupleurum chinense compounds

2) Atomic Environment Context-Aware Feature Enhancement Strategy

After constructing the basic graph structure, it is further aimed to implement contextual expansion of atomiclevel features to help the model better identify the local chemical microenvironment. This paper uses the atomic adjacency matrix to define adjacent relationships between nodes and constructs weighted adjacency tensor based on edge type (single bond, double bond, triple bond, or aromatic bond). This allows the graph structure to distinguish between electron density differences caused by different bond orders. In this framework, characteristic statistics of each atom in its k-th order neighborhood are calculated, including derived indices such as the average electronegativity of neighboring atoms, the proportion of aromatic rings, and the integral of positive charge density, and these are appended to the original feature vector to form an expanded high-dimensional input representation. In addition, a truncation threshold function based on the Euclidean distance between atomic pairs is introduced to define the effective interaction range:

$$d_{ij} = \|x_i - x_j\|_2 < d_{\text{cut}} \tag{2}$$

When the distance between two atoms is less than a preset cutoff value $d_{\text{cut}} = 5.0$ Å, they are considered to have a potential spatial proximity effect and are considered to have a potential functional association, even if they are not covalently linked. Such distant but spatially adjacent atom pairs are explicitly modeled in the subsequent attention mechanism, helping to capture non-covalent interactions across ring systems or glycosidic bonds.

B. GRAPHORMER-BASED LEARNING OF LONG-RANGE INTERACTION REPRESENTATIONS OF COMPOUNDS

1) Modeling Long-Range Interatomic Dependencies Based on a Global Attention Mechanism

Based on the molecular graph structure of radix stellaviae, a deep representation learning framework with graphormers at its core was constructed, focusing on addressing the information attenuation problem of traditional graph neural networks in capturing long-range atomic interactions. This model models molecules as fully connected graphs. Each atomic node not only communicates with its directly bonded neighbors but also establishes dynamic connections with all other atoms in the graph through self-attention to capture the global structural information. In practice, a multi-head attention structure is trained to perform weighted aggregation of semantic relevance on pairs of atom pairs. The main computational process is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B + E\right)V \tag{3}$$

Q, K and V represent the query, key and value matrices obtained by linear transformation from initial atomic embeddings, respectively. d_k is the dimension of attention head and used to normalize the dot product intensity. B is a position bias term based on shortest path distance between two atoms and used to encode the topological distance relationship. E comes from edge feature enhancement term and jointly encoded by chemical bond type and spatial distance function. The two strategies borrow geometric and chemical prior knowledge to encode topological distance relationship such that the attention weights can not only encode the semantic similarity between two atoms but also encode the potential transmission paths of electronic effects between atom-pairs not directly connected and spatial folding trends between directly connected atoms.

With this design, the model can effectively encode the coupling relationship of charge distribution between steroid nucleus and distal sugar group as well as dipole interaction

between ring system caused by their spatial proximity. This mechanism for polycyclic fused systems and branched oligosaccharide structures of radix stellaviaes avoids the receptive field limitation of local convolution operation and thus fully preserves the synergistic electronic effect between synergistic multiple higher-order functional groups. The term 'high-order' here specifically denotes long-range dependencies between atoms modeled through global attention, rather than higher-order structures in graph topology (e.g., clusters or ring motifs).

2) Deep Semantic Aggregation and Molecular-Level Compact Representation Generation

To further aggregate high-level abstract features, this paper uses a stacked encoder structure, with six layers of Graphormer encoder modules featuring parameter sharing gradually deepen the contextual information of atomic representation. After the update of atomic representation based on attention mechanism, each encoder layer is connected with a feedforward neural network submodule which has two layers of fully connected transformation and a Gaussian error linear unit (GELU) activation function internally, defined as follows:

$$h_i' = W_2 \cdot \mathrm{GELU}(W_1 \cdot h_i + b_1) + b_2 \tag{4}$$

h_i is the hidden state of the i-th atom at current layer, W_1 and W_2 are learnable weight matrices, b_1, b_2 are bias terms, and GELU is chosen as the activation function. GELU maintains nonlinear expressiveness while exhibiting smooth gradient property to promote optimal convergence. With the increase of layers, the atomic representation gradually accumulates multi-level structural information from the whole image and finally construct a deeply fused representation with local chemical information and global conformational information. In particular, the high-level encoding explicitly encodes the change of electron cloud density at the glycosidic bond site and the spatial clustering pattern in hydrophobic region, which makes the prediction more sensitive for potential protein-binding interface. After the encoding process of the final layer, this paper adopts hierarchical pooling strategy to get a fixed-dimensional vector representation of whole molecule. Firstly, it adopts graph-level sum pooling operation to aggregate all atomic latent states.

$$z = \sum_{i=1}^{n} h_i^{(L)} \tag{5}$$

$h_i^{(L)}$ represents the atomic latent vector output by the Lth layer, and z constitutes the final molecular-level embedding. This vector fully embodies the long-range electronic regulation and three-dimensional spatial organization of the molecule, serving as high-quality input for subsequent cross-modal fusion. The entire training process is completed in an end-to-end manner, without the need for manual descriptor design or step-by-step feature extraction, achieving automated dimensionality upgrade from the raw molecular graph to a high-level semantic representation.

In Bupleurum chinense compound representation learning, the core hyperparameter configurations for the Graphormer model are listed in Table 1. These include key settings like the depth of the network structure, the scale of the attention mechanism, the dimensionality of the feature representation, the control of the training process, and limits on input graph size. In order to effectively model the long-range atomic interactions of complex saponin molecules, these parameters were chosen to strike a balance between model expressiveness and computational efficiency. For cross-modal integration with protein embeddings, all parameters have been validated for end-to-end learning from molecular graphs to high-order semantic vectors.

TABLE 1. Core Configuration of the Graphormer Model

Parameter	Value	Description
Encoder Layers	6	Number of stacked Graphormer encoder blocks
Attention Heads	8	Number of parallel attention heads
Hidden Dimension	768	Dimension size of node and graph embeddings
Batch Size	32	Number of molecules per training batch
Learning Rate	5e-5	Initial learning rate for AdamW optimizer
Dropout Rate	0.1	Dropout probability in attention
Training Epochs	200	Total number of training epochs
Max Atom Count	100	Maximum number of atoms allowed in a molecule

The Graphormer model's hierarchical feature extraction mechanism for learning radix stellaviae molecular representations is depicted in Figure 2. With increasing network depth, both representation quality and the capacity to capture long-range interactions improve significantly, but the growth patterns vary, as depicted in Figure 2(a). The model places a high priority on developing fundamental structural cognition, as evidenced by the rapid improvement in representation quality in the first three layers before reaching saturation. Longdistance dependencies must be identified using a deeper semantic abstraction because long-distance capture capability increases with network depth. This hierarchical learning feature enables the model to take into account both the local chemical environment and the global spatial constraints. The evolution of attention distribution in Figure 2(b) further reveals its working mechanism: the shallow layer mainly focuses on local chemical bonds, while the deep layer significantly enhances the perception of medium- and long-range interactions. This dynamic adjustment stems from the need to transmit electronic effects within the molecule - there is spatial electronic coupling across multiple chemical bonds between the polycyclic skeleton of radix stellaviae and the sugar side chains.

C. PROTEIN INTERACTION NETWORK TOPOLOGY EMBEDDING AND POTENTIAL TARGET COMPLETION

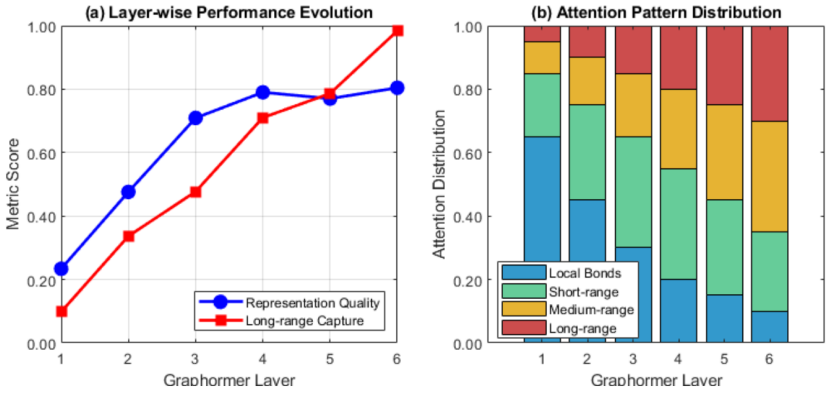

FIGURE 2. Hierarchical Feature Extraction Mechanism

1) Protein Network Construction and Low-Dimensional Topology Encoding Based on High-Confidence Interaction Data

This paper uses the STRING database, a collection of high-quality human protein-protein interactions verified experimentally and integrated with text mining, as a data source. It screened interaction pairs with a composite score of 0.9 or higher to construct a weighted undirected graph with biological credibility. This network uses proteins as nodes and physical associations, co-expression, or functional associations between them as edges. Edge weights are quantified by confidence scores provided by STRING, reflecting the strength of support for interaction events. The resulting interaction graph contains 18,432 protein nodes and 327,514 weighted edges, forming a sparse yet highly modular complex system. Based on this, the Node2Vec algorithm was used to perform nonlinear dimensionality reduction, mapping each protein into a 128-dimensional continuous vector space, effectively compressing the original topological structure while preserving its semantics. The core of Node2Vec is to design a second-order random walk strategy. By adjusting the return parameter and the inward exploration parameter, it balances the depth-first and breadth-first path sampling modes, thus taking into account both local neighborhood aggregation and global network connectivity. During the walk, the probability of the current node selecting the next-hop neighbor is determined by the conditional distribution:

$$P(v_{i+1} = t \mid v_i = s) = \frac{\pi_{st}}{Z} \tag{6}$$

π_{st} represents the unnormalized transfer weight from node s to t, which is affected by the shortest path distance and edge weight; Z is a normalization constant that ensures that the sum of probabilities is 1. By generating a large number of node sequences with fixed lengths, the propagation trajectory of information in the network is simulated, and then these sequences are trained using the Skip-gram model to maximize the contextual co-occurrence probability:

$$L = \sum_{u \in V} \log P(NS(u) \mid z_u) \tag{7}$$

$NS(u)$ represents the set of contextual neighbors of node u during a random walk, and z_u represents its corresponding embedding vector. The optimization process employs a negative sampling strategy to accelerate convergence, enhancing the ability to capture rare but critical connections while maintaining computational efficiency. The resulting embedding vector implicitly encodes functional similarities between proteins, pathway co-occurrence trends, and potential regulatory pathways. Even if there is no direct experimental evidence linking two proteins, as long as they occupy similar topological positions in the network, they can still be assigned highly similar representations, thus providing a mathematical basis for inferring missing links.

2) Identification and Set Expansion of Potentially Active Proteins Guided by Integrating Known Targets

After completing the full-structure protein embedding, a target-driven expansion mechanism was further introduced to infer candidate proteins that are highly adjacent to known targets in the existing PPI network topology and may participate in the same biological pathways. It should be emphasized that these candidate proteins are not newly discovered therapeutic targets, but rather verifiable hypotheses proposed based on network topological similarity, whose pharmacological relevance still requires further experimental validation. Based on the set of experimentally confirmed active ingredient targets, this served as the initial "seed" set. The geometric proximity principle in the embedding space was used to assess the functional association of other proteins with these targets. Specifically, the average cosine similarity between each non-seed protein and all known targets was calculated:

$$\mathrm{Sim}(p_i, T) = \frac{1}{|T|} \sum_{t_j \in T} \frac{z_{p_i} \cdot z_{t_j}}{\|z_{p_i}\| \|z_{t_j}\|} \tag{8}$$

z_{p_i} and z_{t_j} are the embedding vectors of protein p_i and known target t_j, respectively. T represents the set of known targets. This metric measures the overall proximity of a target protein to a real-world pharmacodynamic target in a low-dimensional semantic space. A higher value indicates a greater likelihood of participating in the same biological

process. A threshold of 0.75 was set, and proteins with similarity exceeding this threshold were selected as the initial candidate set.

Finally, this paper only keeps highly connected proteins top 10% ranked ones in centrality to exclude marginal proteins in the network from proposed prediction range. Finally, proteins with both high embedding similarity and strong topological influence are expanded into target set to construct an enhanced target pool covering a more complete signal transduction chain. This approach can not only improve the coverage of existing knowledge base, but also strengthen the ability of subsequent interaction prediction model to reveal implicit biological path.

Figure 3 shows three kinds of centrality indices for 12 representative proteins in a protein-protein interaction network, which can reflect network structure and potential biological property of the protein. Horizontal axis represents protein number (P1-P12), left vertical axis represents Degree Centrality which indicates the number of direct interaction of each protein. Right vertical axis represents Normalized Centrality which represents Closeness and Betweenness respectively and it indicates the efficiency of information transfer and the semantic role of protein in shortest paths between any two proteins in the network. The result shows that P5 and P6 are high in all three kinds of centrality, and their degrees are more than 10, and both of their closeness and betweenness are high. It means that P5 and P6 are not only extensively connected with other proteins, but also in a central position in the network, and they are key hubs in the information transfer. While P7-P9 are low in three kinds of centrality, it means that they are in the periphery of the network and their functional influence is limited.

D. JOINT REPRESENTATION OF COMPOUND-TARGET HETEROGENEOUS NETWORKS AND AFFINITY PREDICTION MODELING

1) Design of the Heterogeneous Feature Fusion Architecture and Dual-channel Semantic Alignment Mechanism

To simulate the interaction between Bupleurum chinense compounds and target proteins, this paper developed an integrated representation learning framework based on isomorphism graph data. The core of this framework lies in the structured integration of multimodal high-dimensional semantic information. By combining molecular vectors extracted through Graphormer with protein embedding vectors generated by Node2Vec, this paper constructed a unified feature representation format: specifically, the compound representation vector is concatenated with the protein embedding vector to form a joint feature tensor. The semicolon symbol represents the vector concatenation operation. This format preserves the independence of original semantic information across different modalities while providing independent information sources for subsequent nonlinear transformations.

Building on this foundation, this paper designed a dual-channel feedforward neural network architecture with independent processing branches for compound pathways and protein pathways. To prevent feature confusion, these branches are processed separately at the input layer. Each branch then progressively abstracts semantic patterns through three fully connected layers, with hidden layer dimensions sequentially set to 512, 256, and 128. This progressive compression bottleneck structure enables the model to capture the most discriminative interaction features during dimensionality reduction. The transformation operations across layers are composed of combinations of linear mappings and nonlinear activation functions:

$$h^{(l+1)} = \mathrm{GELU}\left(W^{(l)}h^{(l)} + b^{(l)}\right) \tag{9}$$

$W^{(l)}$ and $b^{(l)}$ represents the weight matrix and bias term of lth layer, and $h^{(l)}$ represents the activation output of current layer. GELU is smooth and has probabilistic interpretation, which could alleviate the jump of gradient and make the model more sensitive to weak signal. The dual-channel architecture allows the semantic information of compound and protein representations are refined in independent space, which reduces the suppression of information caused by the difference of feature scale and further reduces the suppression of information caused by the difference of feature scale.

2) Loss Optimization and Interaction Probability Inference Mechanism Driven by Negative Sampling

To reduce the risk of false negatives, negative samples were not randomly selected from the full protein space but were restricted to compound-protein pairs that had been experimentally validated as non-interacting through high-throughput experiments or showed no association records in multiple databases (STITCH, BindingDB, TCMSP). Additionally, to avoid introducing potential true positives, proteins with a topological distance ≤ 2 to known targets in the PPI network were excluded as negative sample candidates, thereby controlling false negative contamination at the source. Given the severe class imbalance in compound-target association data (where validated positive samples far outnumber unrecorded potential negative samples), this paper proposes a frequency-weighted negative sampling method to balance training sample distribution and enhance the model's ability to detect rare yet significant interactions. Specifically, during each training iteration, all genuine interaction pairs are designated as positive samples, while unmarked protein pairs from the global database are randomly selected as negatives. By maintaining a 1:3 positive-to-negative sample ratio, this approach effectively prevents negative samples from overwhelming the random selection of non-informative samples. Additionally, sampling weights are allocated based on proteins' occurrence frequency in known interaction networks, thereby reducing the probability of negative samples derived from high-frequency "hub proteins". This strategy avoids both overfitting common targets and excessive learning of marginal functional proteins. The model's output layer employs a sigmoid activation

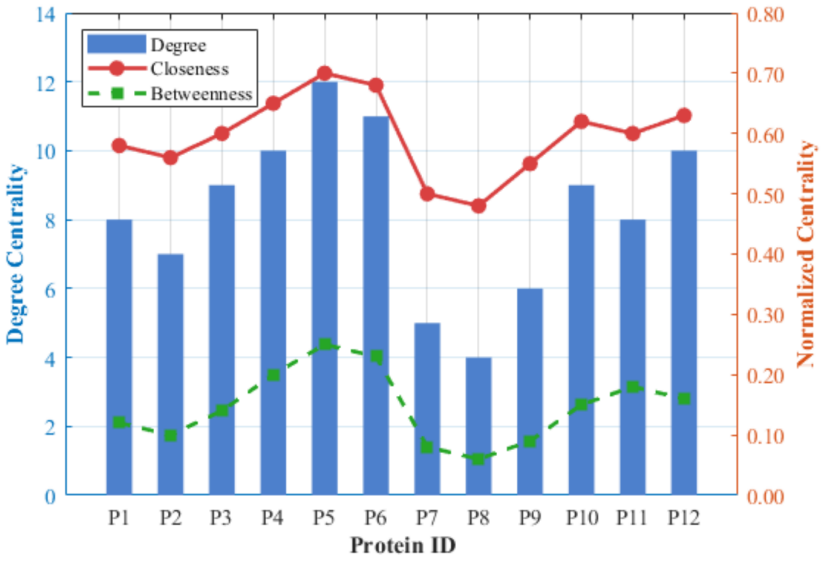

FIGURE 3. Comparative Centrality Indices of PPI Networks

function to scale neuron responses into the [0, 1] interval, providing interpretable estimates of target probabilities for physical or functional interactions between compounds and proteins:

$$\hat{y} = \sigma(f(c,p)) = \frac{1}{1+\exp(-f(c,p))} \tag{10}$$

f represents the composite mapping function of the entire two-channel network, and $\hat{y}$ is the prediction score. The training objective is defined as the weighted binary cross entropy loss function:

$$L = -\sum_{i=1}^{N} w_i \left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i) \right] \tag{11}$$

y_i is the true label of i-th sample (1 means interaction, 0 means not interaction), $\hat{y}_i$ is the corresponding predicted probability, and w_i is the sample weight and calculated according to class frequency to balance positive and negative sample contribution. In the training process, AdamW (Adam optimizer combined with weight decay) parameter update strategy is adopted, and gradient clipping and learning rate preheating strategy are combined to ensure the stability of model convergence.

III. COMPREHENSIVE COMPARISON OF BUPLEURUM CHINENSE COMPOUND-TARGET PREDICTION PERFORMANCE

A. EXPERIMENTAL DATA

The inclusion of Bupleurum compounds in this study was limited to saponins with clearly defined structures and target annotations in ChEMBL, PubChem, and TCMSP, totaling 48 compounds. Screening criteria included:

(1) a content $\geq 0.1\%$ (w/w) in Bupleurum root as verified by HPLC or MS;

(2) at least one reported pharmacological activity (e.g., anti-inflammatory, antidepressant); and

(3) the smiles structure could be successfully resolved by RDKit without stereochemical conflicts.

Known saponins that did not meet any of the above conditions (such as trace components like Bupleurum saponins F and G) were excluded. This study focused on the core active components of radix stellaviaes to construct a high-quality compound-target interaction dataset. Molecular structural information was obtained from the ChEMBL and PubChem databases, including standardized SMILES representations of major monomers such as radix stellaviae D, A, and C. Functional group correction and three-dimensional conformational optimization were performed using RDKit to ensure the chemical plausibility of the input structures. Target association data was integrated from STITCH, BindingDB, and TCMSP platforms. Low-confidence annotations were removed through literature mining and experimental validation. Ultimately, 654 high-confidence interaction relationships between 48 compounds and 217 targets were identified. Protein sequences and functional annotations were obtained from the UniProt database. Combined with physical and co-expression evidence from STRING with a confidence level of 0.9 or higher, a human-based PPI network was constructed for topological embedding learning.

1) Molecular Structure and Characterization of Radix Stellaviae

Based on the typical structure of radix stellaviae D, an adjacency matrix of its atomic connections was constructed, and the dynamic convergence of potential energy and geometric deviation during the iterative energy minimization process was simulated. Subsequently, multidimensional eigenvectors were generated using predefined atomic types and physicochemical rules, and microenvironmental indicators

for different atomic types were calculated using statistical methods. The entire process was implemented using MATLAB matrix operations, random number generation, and visualization functions, aiming to accurately replicate the molecular graph parsing and feature encoding process.

Figure 4 shows the whole computing procedure of radix stellaviae from topological structure to 3D conformation. Figure 4(a) shows the molecular topological connectivity based on RDKit analysis. The connectivity can be visualized as an adjacency matrix of 30 × 30. Nonzero elements (black squares) in the matrix are covalent bonds between atoms. It is evident that the matrix has a blocky structure. The red dashed box denotes the high-density connection region formed by the molecular steroid nucleus. The blue dashed box denotes the branched structure of the sugar chains attached to the nucleus, which presents a sparse connection region. The connecting bridges between the nucleus and the sugar chains (outside the wireframe) can be clearly seen. Therefore, the complex three-dimensional skeleton and branching topology of radix stellaviae molecule can be obtained. It provides the structural basis for the subsequent graph neural network modeling. The whole process of molecular conformation change under the guidance of the force field in the process of energy minimization is shown in Figure 4(b). The vertical axis on the left represents the total potential energy U (kcal/mol) of the system. It quickly decreases from an initial value of about 170 kcal/mol after 100 iterations and then slowly converges to about 20 kcal/mol with small oscillations. It is worth noting that the molecular structure is gradually relaxing to a local energy minimum. The right vertical axis represents the root mean square deviation (RMSD) of the bond angles, shown as red dashed line. It quickly decreases from an initial value of about 2.0 rad with the increase of iterations and then converges to about 0.3 rad. The coordinated convergence of the two indicators further verifies that the optimization effect is effective.

Figure 5 uses parallel coordinates to clearly show the 8-dimensional distribution of the 8 characteristic dimensions of key physicochemical characteristics of 36 atomic nodes in radix stellaviae molecule (choosing main 8 dimensions from 12 dimensions). The horizontal axis is the eight characteristic dimensions; the vertical axis is the normalized characteristic value (0-1). Dark blue represents the carbon atom in the steroid nucleus that is aromatic. Its characteristic trajectory is high in terms of aromaticity and ring membership, atomicity, hybrid orbit, but low in terms of atomic number, valence number, hydrogen bonding capacity, which is accurate for electron delocalization and structural rigidity of aromatic ring system. Black represents the oxygen atom in the nucleus that is aromatic. Its characteristic trajectory decreases significantly in the hybridization dimension, but remains at a moderate level in the atomicity dimension, reflecting the chemical behavior of saturated carbon chain. The red trajectory represents the oxygen atom of the sugar chain. The characteristic trajectory shows significantly low values in hybridization, aromaticity, formal charge, atomicity, and ring membership, while the atomic number remains high, and the valence electron count reaches a peak. It also reaches a peak in hydrogen bond capacity, reflecting the electronegativity of oxygen atoms as strong hydrogen bond acceptors. The orange trajectory represents the carbon atom of the sugar chain. The characteristic trajectory shows a moderate to low overall dimensionality, reflecting the aliphatic nature of the sugar chain.

2) Model Stability Assessment

This research adopts a fivefold cross-validation approach, assessing classification effectiveness through AUROC (Area Under the Receiver Operating Characteristic Curve) and AUPRC (Area Under the Precision-Recall Curve). While AUROC evaluates the model's capability to differentiate between positive and negative instances across a range of thresholds, AUPRC emphasizes the concentration of true positives among higher-ranked predictions. To ensure fair comparison under the same partitioning conditions, the study evaluates the proposed Graphormer against conventional methods such as Graphormer, GCN (Graph Convolutional Network), GAT (Graph Attention Network), DeepDTI (Deep Learning-based Drug-Target Interaction prediction), and TransDTI (Transformer for Drug-Target Interaction Prediction).

Figure 6 shows the comparative discriminative abilities of various approaches in identifying positive and negative compound-target interactions. Across five-fold cross-validation, the Graphormer method introduced in this research showed consistent stability, delivering average AUROC values between 0.91 and 0.93, with minimal variation. By combining long-range molecular structure dependencies with protein topology attributes, the method significantly bolstered classification robustness. In contrast, conventional graph neural network approaches, restricted by their local receptive field limitations, experienced notable performance drops during cross-validation. The subfigure highlights a concentration of positive samples within high-confidence prediction ranges. Proposed approach achieved superior performance, with average AUPRC values of 0.88 to 0.90, effectively pinpointing true interactions even amidst pronounced sample imbalance. The narrower error ranges underline the reduced cross-validation variance of proposed model, which can be credited to its architecture design enhancing generalization.

The AUROC (0.91–0.93) and AUPRC (0.88–0.90) reported in this study were both based on five-fold cross-validation, reflecting the model's generalization capability within the known compound-target space. Although no in vitro experiments or cross-species external validation have been conducted, the high AUPRC values (particularly under sparse positive sample conditions) indicate that the model demonstrates strong ranking and candidate screening capabilities for potential interactions, providing priority guidance for subsequent experimental validation.

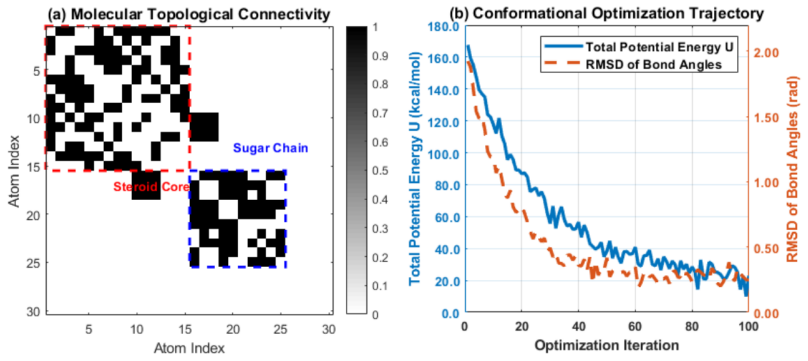

FIGURE 4. Molecular Graph Construction and 3D Conformation Optimization of Radix Stellaviae

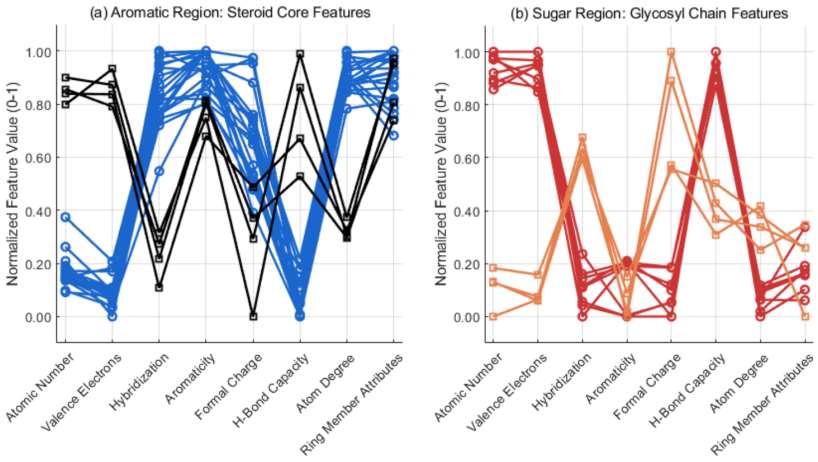

FIGURE 5. Parallel Coordinate Analysis of Atomic Physicochemical Characteristics

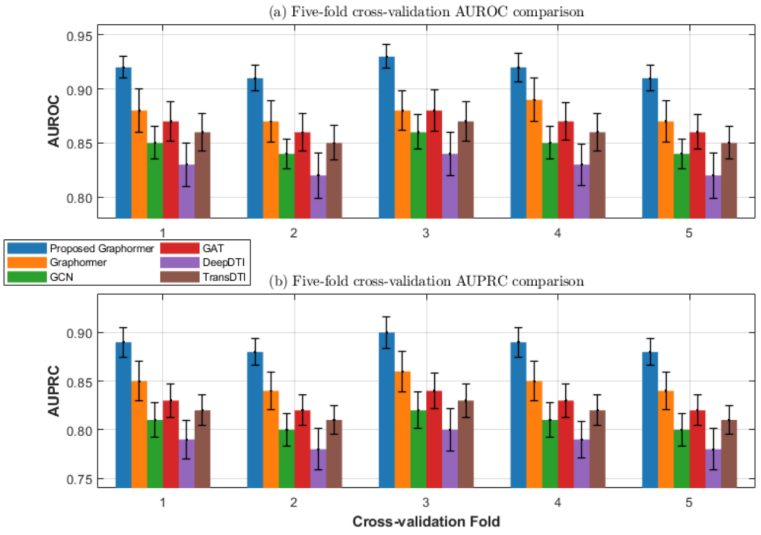

FIGURE 6. Comparison of AUROC and AUPRC

3) New Target Identification Ability

To simulate scenarios where target knowledge is sparse, only 10%, 20%, and 30% of known targets were retained for model training, respectively. The remaining targets were excluded from the training set. During the testing phase, the model's recall performance for unseen targets was evaluated. The F1-score measures the harmonic mean of precision and recall, and Coverage Rate@50 indicates the proportion of true interacting targets covered by the top 50 predictions. A horizontal comparison of the proposed Graphormer with tra-

ditional Graphormers, GCN, GAT, DeepDTI, and TransDTI under the same masking setting was conducted to examine its generalization ability in the absence of information.

Table 2 shows the ability of various models to identify unseen targets under varying conditions of target knowledge sparsity. As the proportion of available targets during training decreases, the performance of all methods decreases, reflecting the limiting effect of scarce labeled data on model learning. The proposed Graphormer framework maintains a relatively stable F1-score (0.683 ± 0.021) and high coverage (Coverage Rate@50 of 76.4%) even in an extremely low-resource scenario (only 10% of targets are visible).

TABLE 2. Analysis of New Target Identification Ability

Method	Shielding Ratio	F1-score (±Std)	Coverage Rate@50 (%)
Proposed Graphormer	90%	0.683 ± 0.021	76.4
	80%	0.712 ± 0.018	80.1
	70%	0.735 ± 0.015	83.6
	90%	0.621 ± 0.025	67.3
Traditional Graphormer	80%	0.649 ± 0.023	70.8
	70%	0.674 ± 0.020	74.2
	90%	0.543 ± 0.031	56.7
GCN	80%	0.572 ± 0.028	60.3
	70%	0.596 ± 0.026	63.9
GAT	90%	0.587 ± 0.027	61.5
	80%	0.613 ± 0.024	65.2
	70%	0.638 ± 0.022	68.7
	90%	0.521 ± 0.033	53.4
DeepDTI	80%	0.548 ± 0.030	57.1
	70%	0.573 ± 0.027	60.5
	90%	0.564 ± 0.029	58.2
TransDTI	80%	0.592 ± 0.026	62
	70%	0.617 ± 0.024	65.4

Note: Std is the abbreviation for standard deviation.

4) Recommendation Performance Verification

A 5-fold cross-validation approach was used to evaluate the model's target recommendation capabilities without any historical information about the molecule. This paper used Precision@10 and Recall@10 as evaluation metrics. The former measures the proportion of correct predictions among the top 10 targets, while the latter reflects the probability that a truly associated target was successfully ranked in the top 10. The proposed graphormer was compared with traditional graphormers, GCN, GAT, DeepDTI, and TransDTI on the same independent compound dataset, maintaining input consistency to eliminate bias.

Figure 7 shows the comparison of recommendation performance of each model on independent compound sets based on Precision@10 and Recall@10 metrics. The method achieves higher prediction consistency and cross-fold stability on two metrics. The average Precision@10 of all proposed models in this paper is in the range of 0.87-0.89, and the average Recall@10 is in the range of 0.79-0.81. The error range is small. It shows that the molecule-target co-characterization framework this paper constructed can generalize to unseen compounds and rank targets accurately based on drug similarity. Precision@10 result shows that proposed method has stronger discriminative ability in high confidence interval and can reduce the proportion of false positive recommendations. That is because Graphormer can spatially perceive the key pharmacophores in the complex structure of radix stellaviaes and preserve functional semantics in protein embedding space. Recall@10 result further shows that proposed method has more reliable ability to comprehensively capture true interacting targets in a limited prediction window. It shows that the model not only depends on a memory of known interactions but also infers potential pathways of action from topological proximity and functional similarity. While traditional architectures show large performance fluctuation on novel molecular scenarios due to the lack of modeling long-range atomic dependencies or inadequate biological networks prior guidance.

5) Model Robustness Testing in the Face of Structured Noise

To systematically evaluate the robustness of the model against molecular graph structural anomalies, edge perturbation experiments were designed. By randomly inserting artificial chemical bonds into saponin molecular graphs, this paper simulated topological deviations caused by conformational prediction errors or atomic type misjudgments. The perturbation levels were set at 5%,10%,15%, and 20%, with noise intensity increasing from mild to severe. After each perturbation, the training model parameters were retained, and the noisy molecular graph was reprocessed to output affinity predictions. Mean Square Error (MSE) was used to quantify the numerical deviation between predicted values and true labels, reflecting model stability. Meanwhile, Kendall's rank correlation coefficient was calculated to assess the stability of true rankings versus predicted rankings under noise interference, thereby evaluating the model's robustness in suppressing disturbances during critical target prioritization.

Table 3 illustrates the evolving trends in prediction stability and ranking consistency for each model when molecular graph structural noise is incrementally introduced. As the proportion of incorrect chemical bonds rises, all methods exhibit varying degrees of performance decline, underscoring the disruptive impact of abnormal topological connections on molecular representation learning. The Graphormer framework proposed in the study exhibits superior resistance to noise with regards to both MSE and Kendall's Tau, achieving values of 0.237 ± 0.013 and 0.681 ± 0.018, respectively, at a perturbation level of 20%. The prediction error increases gradually, and the rank correlation decreases smoothly, indicating that the global attention mechanism effectively mitigates the influence of local noise through long-range dependency modeling, preserving an accurate understanding of critical functional substructures. Traditional graph neural networks, which rely on local neighborhood aggregation, often experience feature confusion under accumulated noise, resulting in highly unstable outputs. Additionally, methods based on pure sequence processing or static attention fail to distinguish real connections from false ones effectively, thereby significantly impairing ranking performance.

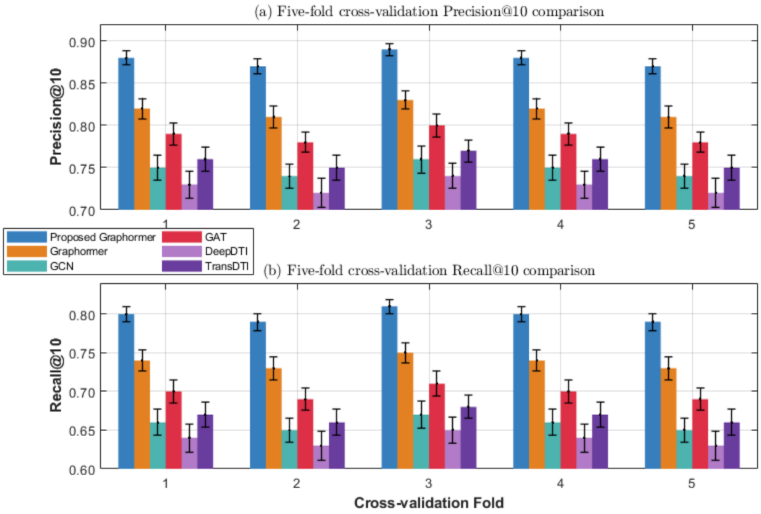

FIGURE 7. Comparison of Precision@10 and Recall@10

TABLE 3. Model Robustness Test

Method	Noise Level	MSE (±Std)	Kendall's Tau (±Std)
Proposed Graphormer	5%	0.183 ± 0.006	0.742 ± 0.012
	10%	0.196 ± 0.008	0.728 ± 0.014
	15%	0.214 ± 0.010	0.705 ± 0.016
	20%	0.237 ± 0.013	0.681 ± 0.018
Traditional Graphormer	5%	0.211 ± 0.009	0.698 ± 0.015
	10%	0.238 ± 0.012	0.673 ± 0.017
	15%	0.272 ± 0.016	0.641 ± 0.020
	20%	0.315 ± 0.021	0.602 ± 0.023
GCN	5%	0.264 ± 0.014	0.632 ± 0.019
	10%	0.301 ± 0.018	0.598 ± 0.022
	15%	0.347 ± 0.023	0.554 ± 0.025
	20%	0.402 ± 0.027	0.508 ± 0.028
GAT	5%	0.237 ± 0.011	0.665 ± 0.016
	10%	0.269 ± 0.014	0.638 ± 0.018
	15%	0.308 ± 0.017	0.607 ± 0.021
	20%	0.354 ± 0.020	0.573 ± 0.024
DeepDTI	5%	0.288 ± 0.016	0.601 ± 0.021
	10%	0.331 ± 0.019	0.567 ± 0.023
	15%	0.376 ± 0.022	0.529 ± 0.026
	20%	0.423 ± 0.025	0.491 ± 0.029
TransDTI	5%	0.252 ± 0.013	0.649 ± 0.018
	10%	0.291 ± 0.016	0.612 ± 0.020
	15%	0.335 ± 0.019	0.574 ± 0.022
	20%	0.382 ± 0.023	0.533 ± 0.025

IV. CONCLUSION

This study constructed a collaborative prediction framework that integrates graph transformers and protein interaction network topology embedding to achieve high-precision modeling of key component-target interactions in Bupleurum chinense. A parallel approach in the industry could leverage graph transformers and material interaction network topology embedding to model the complex, multivariable outcomes of novel fiber and chemical combinations, enhancing product development and quality control across manufacturing facilities. By introducing a global attention mechanism for molecular graphs, the long-range electronic effects and spatial folding properties of saponins are effectively captured. Combined with Node2Vec's low-dimensional semantic encoding of PPI networks, the identification of potential functional association pathways is improved, enhancing the integrity of the target space. A dual-channel neural network is used to dynamically integrate compound and protein representations, demonstrating superior generalization and stability compared to existing methods in five-fold cross-validation, new target identification, independent compound recommendation, and noise robustness testing. Experimental results show that the model can not only accurately predict known interaction relationships, but also reliably infer unknown targets in knowledge-sparse scenarios. This study provides an interpretable and scalable computational paradigm for the multi-component-multi-target system of Traditional Chinese Medicine, and offers a new technical path for the modernization of Traditional Chinese Medicine research. Furthermore, this paradigm is readily extendable to the industry, providing an interpretable and scalable method for modeling the complex interactions of multi-chemical-multi-fiber systems to optimize manufacturing processes and predict material durability. It should be noted that this model focuses on predicting static interactions between compounds and targets, without incorporating dynamic pharmacological dimensions such as temporal evolution, dose dependence, or component synergy. Therefore, its output reflects potential binding possibilities rather than the temporal trajectory of pathway activation or the strength of multi-component synergy. This limitation indicates that the model is more suitable for high-throughput preliminary

screening of candidate targets rather than replacing dynamic systems pharmacology modeling. Future work will integrate pharmacokinetic and pathway dynamics data to expand toward a time-aware, multi-component integrated model.

AUTHOR CONTRIBUTIONS

Jinjin Li designed, collected and analyzed the data, and drafted the manuscript. Jinjin Li conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jinjin Li participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Sui, W. J. Han, C. R. Zhu, and J. H. Wei, "Recent progress in saikosaponin biosynthesis in Bupleurum," Current Pharmaceutical Biotechnology, vol. 22, no. 3, pp. 329-340, 2021, doi: 10.2174/1389201021999200918101248.

[2] E. Hamzaoğlu, "A new variety of the Bupleurum (Apiaceae) from Türkiye," Anatolian Journal of Botany, vol. 7, no. 1, pp. 46-49, 2023, doi: 10.30616/ajb.1249583.

[3] A. V. Myagchilov, P. G. Gorovoi, and L. I. Sokolova, "Flavonoids of the Far Eastern species of the genus Bupleurum L," Russian Journal of Bioorganic Chemistry, vol. 50, no. 7, pp. 2884-2889, 2024, doi: 10.1134/S1068162024070227.

[4] S. Hou, R. Zhang, C. Zhang, L. Wang, H. Wang, and X. X. Wang, "Role of vermicompost and biochar in soil quality improvement by promoting Bupleurum falcatum L," nutrient absorption. Soil Use and Management, vol. 39, no. 4, pp. 1600-1617, 2023, doi: 10.1111/sum.12955.

[5] J. Ghosh and M. Bhaumik, "Typification of nine names of the genus Bupleurum (Apiaceae)," Feddes Repertorium, vol. 136, no. 2, pp. 114-127, 2025, doi: 10.1002/fedr.70010.

[6] D. Mu and Q. Ma, "A review of antidepressant effects and mechanisms of three common herbal medicines: Panax ginseng, Bupleurum chinense, and Gastrodia elata," CNS & Neurological Disorders - Drug Targets-CNS & Neurological Disorders, vol. 22, no. 8, pp. 1164-1175, 2023, doi: 10.2174/1871527322666221116164836.

[7] W. Han, J. Xu, H. Wan, L. Zhou, B. Wu, J. Gao, et al., "Overexpression of BcERF3 increases the biosynthesis of saikosaponins in Bupleurum chinense," FEBS Open Bio, vol. 12, no. 7, pp. 1344-1352, 2022, doi: 10.1002/2211-5463.13412.

[8] M. Gao, X. Huo, L. Lu, M. Liu, and G. Zhang, "Analysis of codon usage patterns in Bupleurum falcatum chloroplast genome," Chinese Herbal Medicines, vol. 15, no. 2, pp. 284-290, 2023, doi: 10.1016/j.chmed.2022.08.007.

[9] G. Zhang, H. Wang, J. Jiang, Q. Wang, B. Li, L. Shi, et al., "The complete chloroplast genome of Bupleurum marginatum var," stenophyllum (H. Wolff) Shan & Yin Li (Apiaceae), a new substitution for Chinese medicinal material, Bupleuri Radix (Chai hu). Mitochondrial DNA Part B, vol. 6, no. 2, pp. 441-443, 2021, doi: 10.1080/23802359.2020.1870899.

[10] Z. W. Li, J. T. Wu, Y. Bi, M. M. Li, A. Naseem, J. Pan, et al., "A new triterpenoid saponin and a new natural saponin from the roots of Bupleurum scorzonerifolium Willd," Natural Product Research, vol. 39, no. 15, pp. -4346, 2025, doi: 10.1080/14786419.2024.2338827.

[11] S. Khaled, M. Benahmed, S. Akkal, and H. Laouer, "Experimental and theoretical approaches to examine the effects of Bupleurum montanum extracts as corrosion inhibitors on carbon steel in an acidic solution," Chemical Papers, vol. 79, no. 3, pp. 1707-1737, 2025, doi: 10.1007/s11696-025-03884-1.

[12] X. Zou, Y. Sui, X. Y. Tang, R. Zhang, Q. Shu, T. Gong, et al., "Mechanism of Bupleurum scorzonerifolium and Paeonia lactiflora herbal pair against liver cancer: An exploration based on UPLC-Q-TOF-MS combined with network pharmacology," Zhongguo Zhongyao Zazhi (China Journal of Chinese Materia Medica), vol. 47, no. 13, pp. 3597-3608, 2022, doi: 10.19540/j.cnki.cjcmm.20220110.402.