



Correspondence

<https://doi.org/10.1631/jzus.B2600409>

A novel graph-text composite model for pesticide metabolic pathway prediction

Ruyue LI^{1,2*}, Fahui YUAN^{1*}, Mengcen WANG³, Yong HE¹, Yufei LIU¹✉

¹College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China

²College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

³State Key Laboratory of Rice Biology & Ministry of Agricultural and Rural Affairs Laboratory of Molecular Biology of Crop Pathogens and Insects, Zhejiang University, Hangzhou 310058, China

Agricultural chemicals are essential in modern agriculture for managing pests and diseases and increasing crop yields (Barone et al., 2024). However, the pervasive use of pesticides has raised primary concerns over potential risks to human health, as well as serious questions about environmental contamination and ecological hazards (Collinge et al., 2022). Phytotransformation of pesticides can influence their efficacy, the extent of plant uptake, and potential environmental exposure (Li et al., 2019). Therefore, elucidating the key metabolic pathways within the crop interleaf system is crucial for the design of safer and greener pesticides.

To address this, we introduce here a novel dual-modality feature-extraction mechanism, called the graph-text composite (GTC) model. Its primary contribution is the synergistic integration of molecular structure information from complementary perspectives, which overcomes the limitations of single-view representations. Evaluations on a multi-label dataset demonstrate that the GTC model achieves an accuracy of 96.05%, a precision of 82.26%, a recall of 78.67%, and an F1_score of 80.41%. These results outperform single-input models across all metrics. By integrating image and text features, the study provides a new approach to accurately predicting metabolic pathways in rice and related cereals.

The integration of artificial intelligence (AI) into crop breeding and field management represents a transformative shift in modern agriculture (Zhang et al., 2026). In recent years, AI has been transforming agriculture by enabling data-driven decisions, thereby enhancing efficiency, sustainability, and food security (Li et al., 2025). Prediction of compound metabolic pathways has evolved significantly through the application of AI technology. By converting Simplified Molecular Input Line Entry System (SMILES) strings into graph-structured representations, Baranwal et al. (2020) were able to use the graph convolutional network (GCN) to capture structural features that ultimately enabled accurate metabolic pathway prediction via a downstream classifier. Subsequent innovations, such as graph-attention networks, enhanced performance through dynamic feature weighting, yet remained constrained by the coarse taxonomic resolution of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Du et al., 2022).

Despite these advancements, critical gaps hinder broader application, particularly in the prediction of metabolic pathways for staple crops like rice. Current models often lack generalizability across diverse

✉ Yufei LIU, yufeiliu@zju.edu.cn

* The two authors contributed equally to this work.

Yufei LIU, <https://orcid.org/0000-0001-7027-8911>

Ruyue LI, <https://orcid.org/0009-0008-7899-6529>

Fahui YUAN, <https://orcid.org/0009-0005-8283-6905>

Received June 20, 2026; Revision accepted Oct. 13, 2026;

Crosschecked xxx. xx, 20xx

taxonomic groups and struggle with the rare or poorly annotated pathways prevalent in rice metabolism. To bridge this, we propose a framework utilizing a hybrid-architecture GCN trained on metabolomic data. This approach synergizes the global relational reasoning of GCN with the local feature-extraction capabilities of a convolutional neural network (CNN), enabling efficient knowledge transfer to plant-specific pathway-prediction tasks.

Specifically, the proposed GTC model comprises two specialized modules designed to capture molecular characteristics from complementary perspectives through two parallel modules: a GCN and a text-based CNN as shown in Fig. 1. The GCN module represents a molecule as a graph of atoms (nodes) and chemical bonds (edges). It leverages graph-transformer layers with a self-attention mechanism to aggregate information from neighboring atoms, thereby learning the topological features of the molecule. The graph transformer is configured with two layers, four attention heads, and a hidden embedding dimension of 50.

The CNN module processes the molecular SMILES string as textual data, utilizing molecular transformer embedding derived from a pretrained ChemBERTa model, followed by 1D convolutional and pooling operations to learn the string's sequential structural features. The feature vectors from both modules are then concatenated and integrated. The fused features are processed through a fully connected layer with dropout (rate = 0.5) to prevent overfitting. Finally, the output layer reduces the dimensionality of the integrated high-dimensional features to produce probability scores indicating the likelihood of the compound belonging to various metabolic pathways, accomplishing the multi-label prediction task.

To further ensure robust training and prevent overfitting beyond dropout, an L2 weight decay of $1e-4$ is applied. The model is trained using the Adam optimizer for 100 epochs with a batch size of 10. The initial learning rate is set to $1e-3$, which decays by a factor of 0.75 every 20 epochs. This dual-modality framework enables neural networks to synergistically integrate structural information through hierarchical feature abstraction. Through this dual-modality framework, the model achieves hierarchical abstraction and synergistic integration of molecule-structure information, thereby building a more comprehensive and discriminative molecular representation.

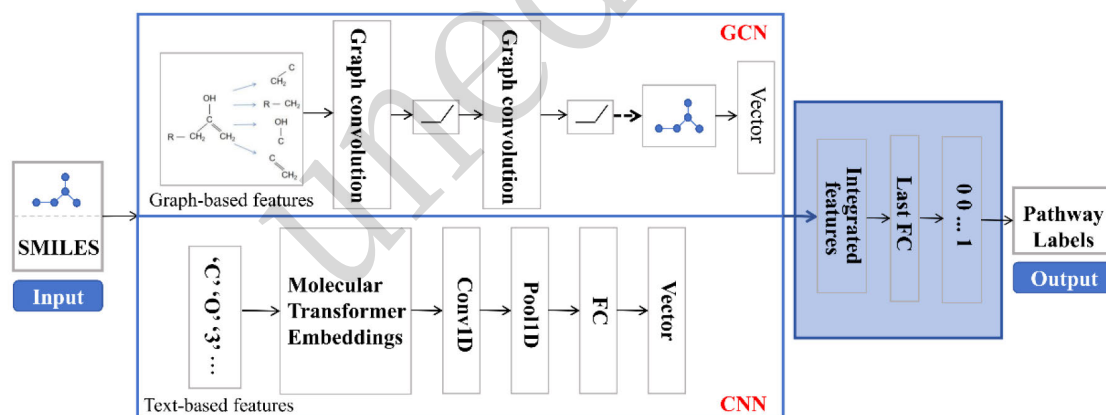


Fig. 1 Flowchart of the GTC model. SMILES: simplified molecular input line entry system. GCN: graph convolutional network. CNN: convolutional neural network. Conv1D: 1-dimensional convolutional layer. Pool1D: 1-dimensional pooling layer. FC: fully connected layer.

To evaluate the GTC model, we employed two complementary datasets (Dataset A and Dataset B) that reflect distinct prediction scenarios in plant metabolism. Dataset A (2,028 compounds, 18 pathway labels) was a multi-label dataset where a single compound could participate in multiple biosynthetic pathways; Dataset B (4,545 compounds, 11 primary pathway classes) was a single-label dataset, as each compound was assigned to only one core metabolic process. Datasets A and B were curated from the KEGG database, with pathway labels assigned directly from KEGG's internal compound-pathway mappings.

The use of two datasets addressed the inherent complexity of metabolic pathway prediction: Dataset A was

critical for secondary metabolism where pathways often overlap, while Dataset B was relevant to primary metabolism with mutually exclusive route assignments. We trained and evaluated the GTC model separately on each dataset to demonstrate its versatility across different biological realities, with each dataset partitioned into training, validation, and test sets at an 8:1:1 ratio. In addition, we employed a stratified sampling strategy and strict dataset segregation protocol to prevent overfitting and ensure the model's generalizability to unseen data. Based on these configurations, the GTC model successfully predicted the specific metabolic pathway affiliations for diverse chemical compounds. The specific prediction outcomes are detailed below.

First, we compared the new model against single-input baselines on Dataset A, such as GCN and CNN. As shown in Table 1, the GTC model achieved an accuracy of 96.05%, a precision of 82.26%, a recall of 78.67%, and an F1_score of 80.41%, outperforming both single-input models across all metrics. A two-tailed Student's t-test confirmed that the improvements were statistically significant ($p < 0.05$ for all 4 metrics compared to CNN and GCN). These results underscore that synergistically integrating graph-based topological features with sequential SMILES string features provides complementary information essential for resolving ambiguous multi-label pathway assignments.

Table 1 Performance of this model versus a single input model for metabolic pathway

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)
CNN	94.66	74.95	73.31	71.16
GCN	95.46	81.78	78.37	80.14
GTC	96.05	82.26	78.67	80.41

CNN: convolutional neural network. GCN: graph convolutional network. GTC: graph-text composite. The best results are highlighted in bold.

Next, we evaluated the GTC model on the single-label Dataset B to assess its performance across diverse primary metabolic classes. The model maintained strong and balanced results (Table 2). Notably, the consistently high precision and recall across all 11 categories indicated that our approach effectively minimized both false positives and false negatives, a critical attribute for reliable metabolic pathway annotation in real-world applications. In addition, we compared the classification accuracy of GTC, GCN, and CNN for predicting metabolic pathways of different compound categories, as shown in Fig.2. The results demonstrated that the GTC model consistently outperformed both single-input models across all tested compound classes.

Table 2 Performance of the GTC model in categorizing metabolic pathways of different kinds of compounds

Class	Accuracy (%)	Precision (%)	Recall (%)
Carbohydrate	98.06	88.95	83.31
Energy	95.46	89.78	95.37
Lipid	98.56	90.12	96.13
Nucleotide	97.10	91.28	90.49
Amino acid	94.64	93.25	93.24
Other Amino acid	95.99	90.61	89.54
Glycan	96.47	90.06	93.70
Cofactors/Vitamins	95.64	91.51	86.14
Terpenoids	97.36	91.41	89.65
Sec. metabolites	95.21	90.67	92.08
Xenobiotics	93.05	92.26	88.67

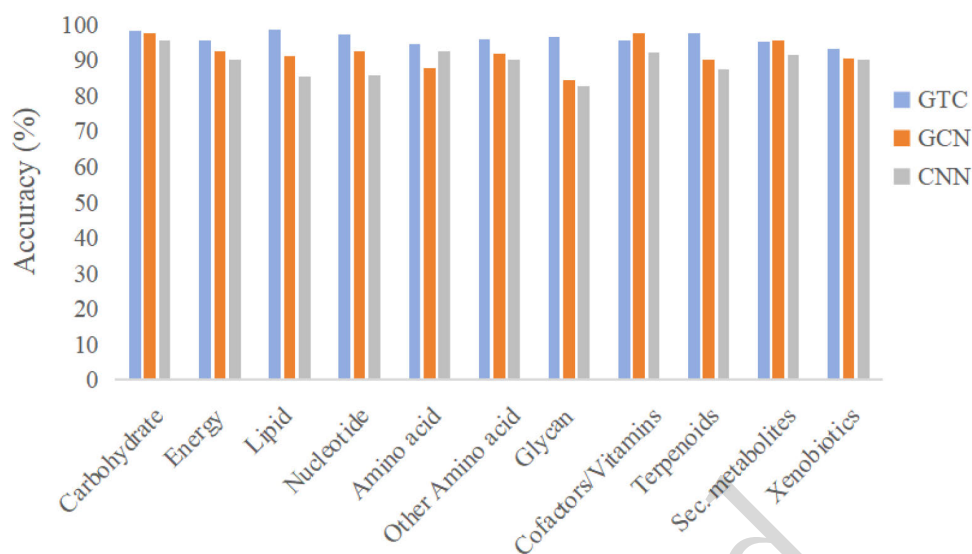


Fig.2 Accuracy of the three models in classifying metabolic pathways for different types of compounds. CNN: convolutional neural network. GCN: graph convolutional network. GTC: graph-text composite.

The above results showed that the GTC model exhibited consistent performance across both primary and secondary metabolic pathways, indicating its ability to capture deep chemical patterns that transcend pathway categories. This achievement provides new methodological support for metabolic research. In areas such as compound functional annotation, metabolic network reconstruction, and novel pathway discovery, the model could significantly reduce the cost and time required for experimental validation. In plant metabolic engineering, accurately predicting the metabolic fate of exogenous or newly synthesized compounds would substantially enhance the success rate of metabolic pathway design.

In addition, we conducted a controlled plant-metabolism study using rice seedlings to experimentally validate the predictive capability of the GTC model. The rice seedlings were cultivated for one month, transplanted, and divided into three groups, each with three replicates. Each group was sprayed with either distilled water, picoxystrobin ($C_{18}H_{16}F_3NO_4$, a strobilurin fungicide), or chlorantraniliprole ($C_{18}H_{14}BrCl_2N_5O_2$, an anthranilic diamide insecticide). Samples were collected 21 days post application and submitted to an ISO/IEC 17025-accredited facility for non-targeted analysis using ultra-performance liquid chromatography–quadrupole time-of-flight mass spectrometry (UPLC–QTOF-MS, mass error < 2 ppm). Putative metabolites were further purified by preparative high-performance liquid chromatography (Prep – HPLC, resolution ≥ 1.5 , purity > 98%).

We selected picoxystrobin and chlorantraniliprole as model compounds due to their contrasting chemical scaffolds and metabolic complexity. Their structures were input into the GTC model to predict plausible biotransformation pathways (Fig. 3). The predicted metabolites were then compared with experimentally observed structures elucidated by liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR) spectroscopy. The results showed that the experimentally determined metabolite structures were consistent with the model's predictions. This alignment preliminarily corroborated the GTC model's ability to infer metabolic pathways of xenobiotics in rice. Overall, these preliminary findings underscore the potential of the GTC model as a computational tool to decode complex xenobiotic metabolic pathways in rice.

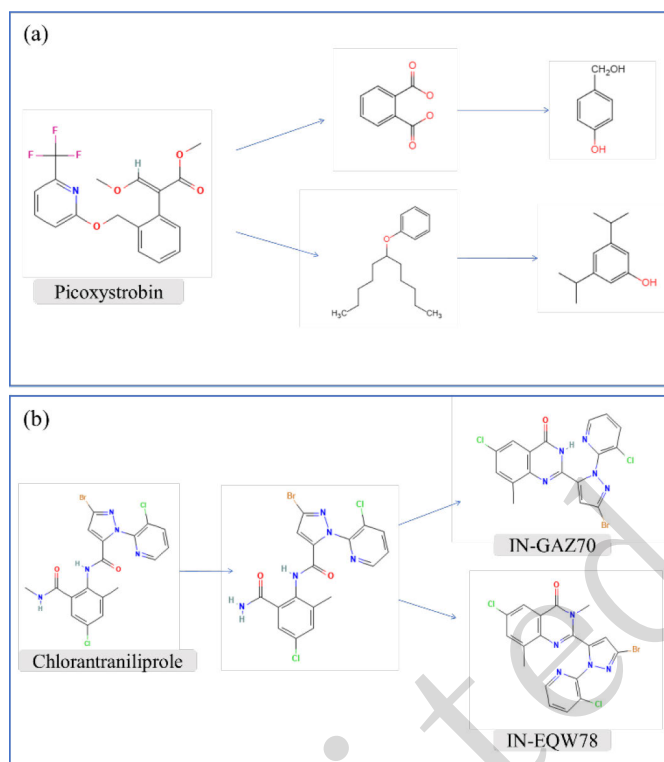


Fig. 3 Prediction of the Picoxystrobin and Chlorantraniliprole pathway by the proposed GTC model

In conclusion, our findings demonstrate that integrating molecular graph representations with sequential SMILES encoding significantly enhances the robustness of the prediction of pesticide metabolic pathways. The proposed GTC model achieves 96.05% accuracy by capturing both structural topology and symbolic patterns of pesticides, while this capability is notably lacking in unimodal approaches. The generalizability of this model suggests its potential to decipher metabolic behaviors of emerging or poorly characterized agrochemicals, even with limited data. While extensive validation is required to confirm broader physiological relevance across uncharacterized chemical spaces, this dual-modality framework provides a biologically informed computational strategy. Ultimately, by enabling accurate identification of metabolic products, this strategy can inform the design of safer and greener pesticides for sustainable crop protection.

Data availability statement

The data that support the findings of this study are available on GitHub at <https://github.com/pluto-46/GTC.git>.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 32572180], the Frontier Technologies R&D Program of Jiangsu [grant numbers BF2025312], the National Key Research and Development Program of China [grant numbers 2025YFD1600100].

Author contributions

Ruyue LI: Writing – original draft, Writing – review & editing, Methodology, Formal analysis, Conceptualization. Fahui YUAN: Writing – original draft, Writing – review & editing, Methodology, Formal analysis, Validation. Mengcen WANG: Writing – review & editing, Resources. Yong HE: Resources, Funding acquisition, Project administration. Yufei LIU: Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Compliance with ethics guidelines

Ruyue LI, Fahui YUAN, Mengcen WANG, Yong HE and Yufei LIU declare that they have no conflicts of interest. This article does not contain any studies with human or animal subjects performed by any of the authors.

Declaration on the use of generative AI tools

No generative AI tools were used in the preparation of this manuscript.

References

- Baranwal, M., Magner, A., Elvati, P., et al., 2020. Biological control of plant diseases. *Bioinformatics* 36, 2547–2553. <https://doi.org/10.1093/bioinformatics/btz954>
- Barone, G.D., Zhou, Y., Wang, H., et al., 2024. Implications of bacteria–bacteria interactions within the plant microbiota for plant health and productivity. *Journal of Zhejiang University-SCIENCE B* 25, 1039–1054. <https://doi.org/10.1631/jzus.B2300914>
- Collinge, D.B., Jensen, D.F., Rabiey, M., et al., 2022. Biological control of plant diseases – What has been achieved and what is the direction? *Plant Pathology* 71, 1024–1047. <https://doi.org/10.1111/ppa.13555>
- Du, B.-X., Zhao, P.-C., Zhu, B., et al., 2022. MLGL-MP: a Multi-Label Graph Learning framework enhanced by pathway interdependence for Metabolic Pathway prediction. *Bioinformatics* 38, i325–i332. <https://doi.org/10.1093/bioinformatics/btac222>
- Li, G., An, L., Yang, W., et al., 2025. Integrated biotechnological and AI innovations for crop improvement. *Nature* 643, 925–937. <https://doi.org/10.1038/s41586-025-09122-8>
- Li R., Dong F., Wu X., et al., 2019. Research progress in three-phase metabolic transformations of pesticides in plants mediated by enzymes. *Chinese Journal of Pesticide Science* 21, 799–814. <https://doi.org/10.16801/j.issn.1008-7303.2019.0105>
- Zhang, W., Ren, Y., Guo, Z., et al., 2026. Improved lightweight convolutional neural network models for the detection and evaluation of Fusarium head blight in wheat. *Journal of Zhejiang University-SCIENCE B* 27, 450–465. <https://doi.org/10.1631/jzus.B2500225>