



OPEN ACCESS

EDITED AND REVIEWED BY
Jiedan Chen,
Chinese Academy of Agricultural
Sciences, China

*CORRESPONDENCE
Lijun Dou
✉ douljun777@163.com
Quan Zou
✉ zouquan@ncclab.net

RECEIVED 26 January 2026
REVISED 11 February 2026
ACCEPTED 16 February 2026
PUBLISHED 25 February 2026

CITATION
Sun S, Zou Q and Dou L (2026)
Editorial: Machine learning for
mining plant functional genes.
Front. Plant Sci. 17:1795967.
doi: 10.3389/fpls.2026.1795967

COPYRIGHT
© 2026 Sun, Zou and Dou. This is an
open-access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Editorial: Machine learning for mining plant functional genes

Shanwen Sun¹, Quan Zou^{2*} and Lijun Dou^{3*}

¹College of Life Science, Northeast Forestry University, Harbin, China, ²Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China,

³Cleveland Clinic Research, Cleveland Clinic, Cleveland, OH, United States

KEYWORDS

large language models, machine learning, multi-omics integration, plant functional genomics, regulatory network inference

Editorial on the Research Topic

Machine learning for mining plant functional genes

Plants in natural and agricultural environments are continuously challenged by diverse biotic and abiotic stresses, placing sustained pressure on crop productivity and global food security. The identification and functional characterization of plant genes underlying stress adaptation, development, and agronomic traits are therefore central to modern plant biology and precision breeding (Gaccione et al., 2025; Zhang et al., 2025). Although advances in high-throughput sequencing have generated vast genomic resources, experimentally validated functional annotations remain limited, leaving a substantial fraction of plant genes poorly characterized. A central challenge in contemporary plant genomics is thus to bridge the gap between rapidly expanding data volumes and biologically meaningful functional insights (Zhang et al., 2025). Recent progress in machine learning (ML) has created unprecedented opportunities to address this challenge. By integrating heterogeneous data types, spanning genome sequences, epigenomic marks, transcriptomic profiles, protein features, metabolite measurements, and regulatory interactions, ML-based approaches can model complex, nonlinear relationships that are difficult to resolve using conventional analytical frameworks (Sasse et al., 2024). Within this context, the Research Topic Machine Learning for Mining Plant Functional Genes brings together eight studies that collectively demonstrate how data-driven computational strategies are reshaping functional gene discovery, regulatory analysis, and trait dissection in plants.

A critical methodological shift in recent years is the emergence of foundation models (FMs) and large language models (LLMs) for biological sequence analysis. By conceptualizing DNA as a structured language, these models leverage large-scale pretraining to generate transferable representations that capture latent regulatory and functional features. Compared with task-specific models, FMs offer enhanced generalization, cross-species transferability, and scalability, making them particularly attractive for plant systems characterized by genomic complexity and limited functional annotations. Providing a systematic overview of this paradigm, Xu et al. present a mini-review synthesizing recent advances in foundation models for plant molecular biology. The review traces the evolution from general DNA language models to plant-specific tools and highlights key challenges unique to plant systems, including polyploidy, repetitive genomes, and sparse experimental annotations. By outlining future directions such as multimodal integration and computational efficiency, this work establishes a conceptual framework for

understanding how FMs are redefining computational plant biology and guiding next-generation model development.

Building on this FM paradigm, several contributions demonstrate how representation learning can be applied to concrete biological problems. Zhang et al. applied a DNABERT-2-based framework combined with gradient boosting to identify DNA N6-methyladenine modifications in rice, illustrating how foundation models can enhance epigenetic marker detection while mitigating data sparsity. This work exemplifies a broader shift toward pretraining-based strategies in plant genomics, with implications for cross-species prediction and regulatory annotation.

Extending LLM-based approaches to cis-regulatory element discovery, Pu et al. developed an enhancer identification framework that couples DNABERT-2 feature extraction with a support vector machine classifier. Beyond predictive performance, this study addresses the interpretability challenge inherent to deep learning by introducing a differential entropy-based analysis to monitor class separation during fine-tuning. The results provide theoretical insight into training dynamics and offer a principled strategy for model optimization, highlighting the growing emphasis on transparency and interpretability in genomic ML.

In parallel with advances in representation learning, architectural innovation is driving progress in modeling gene expression and regulatory complexity. Guo et al. proposed a hybrid framework that integrates Transformer-based global attention with state space models to efficiently capture both long-range dependencies and local regulatory motifs. Validated across multiple crop species, this approach demonstrates improved accuracy and generalization over conventional convolutional architectures, underscoring the importance of model design choices in decoding plant regulatory syntax.

Beyond sequence-centric modeling, graph-based learning emerges as a powerful strategy for uncovering regulatory interactions mediated by non-coding RNAs. Addressing the sparsity and noise characteristic of experimental interaction data, Liao et al. introduced an interpretable graph representation learning framework for predicting plant RNA–RNA interactions. By combining robustness-enhancing masking strategies with biologically interpretable decoding, this work advances the analysis of post-transcriptional regulatory networks, particularly in the context of stress responses.

Several studies focus directly on agriculturally relevant traits and stress adaptation. Qiao et al. developed a machine learning framework for identifying saline–alkali tolerance genes, explicitly addressing data imbalance and sequence divergence through cost-sensitive learning and evolutionary feature extraction. Importantly, their analysis links predictive features to known physiological mechanisms, illustrating how ML can generate biologically interpretable insights rather than purely statistical predictions. Complementing this gene-centric perspective, Saavedra et al. applied interpretable ML to whole-genome methylation data to resolve dormancy stages in sweet cherry. By integrating ensemble learning with feature attribution analysis, the study identifies epigenetic markers associated with developmental transitions and agronomic traits, demonstrating the potential of ML-derived biomarkers for crop management.

Finally, Ju et al. exemplify the value of multi-omics integration by combining genome-wide association analysis with transcriptomic and metabolomic data to dissect seed germination mechanisms in sorghum. By linking genetic variation to hormone signaling and metabolic flux, this systems-level approach moves beyond association toward mechanistic understanding, reinforcing the importance of integrative frameworks in functional gene discovery.

Taken together, the contributions in this Research Topic highlight the transformative role of machine learning and foundation models in plant functional genomics. By advancing representation learning, model architecture, interpretability, and multi-omics integration, these studies move the field beyond traditional sequence-based annotation toward predictive, mechanism-aware, and application-oriented frameworks. Continued synergy between computational innovation and experimental validation will be essential for translating these advances into resilient, high-yield crops capable of meeting future agricultural and environmental challenges.

Author contributions

SS: Writing – original draft, Writing – review & editing. QZ: Writing – review & editing. LD: Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by Natural Science Foundation of Heilongjiang Province (YQ2024F002); the National Natural Science Foundation of China (62273086); and China Postdoctoral Science Foundation (2024M760379).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Gaccione, L., Toppino, L., Bolger, M., Schmidt, M., Tassone, M. R., Sulli, M., et al. (2025). Graph-based pangenomes and pan-phenome provide a cornerstone for eggplant biology and breeding. *Nat. Commun.* 16, 9919. doi: 10.1038/s41467-025-64866-1
- Sasse, A., Chikina, M., and Mostafavi, S. (2024). Unlocking gene regulation with sequence-to-function models. *Nat. Methods* 21, 1374–1377. doi: 10.1038/s41592-024-02331-5
- Zhang, S., Chen, S., Fu, Z., Li, F., Chen, Q., Ma, J., et al. (2025). Integration of digital phenotyping, GWAS, and transcriptomic analysis revealed a key gene for bud size in tea plant (*Camellia sinensis*). *Horticulture Res.* 12, uhaf051. doi: 10.1093/hr/uhaf051
- Zhang, H., Liu, X., Liu, W., Wang, S., Li, Y., Xiang, W., et al. (2025). Deep learning-based annotation of plant abiotic stress resistance genes for crops. *Plant J.* 124, e70556. doi: 10.1111/tpj.70556