北京林业大学学报
JOURNAL OF BEIJING FORESTRY UNIVERSITY

深度学习在林木全基因组选择中的应用与挑战：从高维特征到多目标优化

韩富川 高暝 赵耘霄 陈益存 汪阳东

Deep learning for genomic selection in forest trees: applications and challenges from high-dimensional features to multi-objective optimization

Han Fuchuan, Gao Ming, Zhao Yunxiao, Chen Yicun, Wang Yangdong

引用本文:
韩富川, 高暝, 赵耘霄, 陈益存, 汪阳东. 深度学习在林木全基因组选择中的应用与挑战：从高维特征到多目标优化[J]. 北京林业大学学报, 2026, 48(6):9–20. doi: 10.12171/j.1000–1522.20260208
Han Fuchuan, Gao Ming, Zhao Yunxiao, Chen Yicun, Wang Yangdong. Deep learning for genomic selection in forest trees: applications and challenges from high–dimensional features to multi–objective optimization[J]. Journal of Beijing Forestry University, 2026, 48(6):9–20. doi: 10.12171/j.1000–1522.20260208

在线阅读 View online: https://doi.org/10.12171/j.1000–1522.20260208

您可能感兴趣的其他文章

Articles you may be interested in

关于我国林木育种向智能分子设计育种发展的思考

Thoughts on the development of forest tree breeding towards intelligent molecular design breeding in China

北京林业大学学报, 2024, 46(3): 1–7. https://doi.org/10.12171/j.1000–1522.20230338

论林木常规育种与非常规育种及其关系

On conventional and unconventional tree breeding and their relationships

北京林业大学学报, 2023, 45(6): 1–7. https://doi.org/10.12171/j.1000–1522.20230042

基于深度学习和上采样的杉木单木枯损模型构建

High–precision individual tree mortality prediction for Chinese fir using deep learning and oversampling

北京林业大学学报, 2025, 47(12): 39–47. https://doi.org/10.12171/j.1000–1522.20250103

面向钻蛀振动实时侦听的深度学习模型压缩

Deep learning model compression for real–time listening of boring vibration

北京林业大学学报, 2021, 43(6): 92–100. https://doi.org/10.12171/j.1000–1522.20200100

基于深度学习的树木根系探地雷达多目标参数反演识别

Deep learning–based inverse identification of multi–target parameters for tree rooting ground–penetrating radar

北京林业大学学报, 2024, 46(4): 103–114. https://doi.org/10.12171/j.1000–1522.20230259

基于特征选择和机器学习的森林蓄积量估算

Estimating forest stock volume based on feature selection and machine learning

北京林业大学学报, 2025, 47(4): 155–167. https://doi.org/10.12171/j.1000–1522.20240328

DOI:10.12171/j.1000−1522.20260208

深度学习在林木全基因组选择中的应用与挑战：从高维特征到多目标优化

韩富川[1,2]　高　暝[1]　赵耘霄[1,2]　陈益存[1,2]　汪阳东[1,2]
（1. 中国林业科学研究院亚热带林业研究所，浙江 杭州 311400；2. 中国林业科学研究院林木遗传育种国家重点实验室，北京 100091）

摘要：林木育种具有周期长、杂合度高以及受基因型与环境互作影响显著等特点，传统全基因组选择（GS）模型在处理高维小样本、复杂非加性效应时存在局限。深度学习凭借多层神经网络的非线性建模能力，为解析基因型与表型之间的复杂关系提供了新途径。该技术在作物 GS 中已广泛应用，并衍生出多种融合注意力机制和轻量化设计的模型。此外，多组学与环境数据的联合建模，以及面向多性状协同优化的选择指数构建，正逐渐成为研究热点。然而，针对林木育种中长周期环境互作、高杂合基因组背景及多目标选择需求，深度学习技术的适配性与整合框架尚缺乏系统梳理。本文综述了深度学习在林木 GS 中的应用，主要内容包括：（1）基因组输入特征由单核苷酸多态性向 k-mer 及图泛基因组节点类型的演变及其计算挑战；（2）多组学与环境数据的建模方法；（3）主流深度学习模型的结构特点、适用场景及超参数调优策略；（4）从单性状基因组估计育种值预测到多性状选择指数的理论演进。本文还分析了当前面临的数据稀疏、模型过拟合和可解释性不足等问题，指出迁移学习、半监督学习以及融合生物学先验的机制建模是潜在的解决方向。最后，展望了高通量表型技术与深度学习的进一步融合，提出构建集多组学数据管理、自动化分析流程和育种决策支持于一体的智慧育种平台，以推动林木育种向全基因组智能设计转变。

关键词：林木育种；全基因组选择；深度学习；多组学整合；选择指数；智慧育种
中图分类号：S722；Q37　　**文献标志码：**A　　**文章编号：**1000−1522(2026)06−0009−12
引文格式：韩富川，高暝，赵耘霄，等. 深度学习在林木全基因组选择中的应用与挑战：从高维特征到多目标优化［J］. 北京林业大学学报，2026，48(6)：9−20. Han Fuchuan, Gao Ming, Zhao Yunxiao, et al. Deep learning for genomic selection in forest trees: applications and challenges from high-dimensional features to multi-objective optimization［J］. Journal of Beijing Forestry University, 2026, 48(6): 9−20.

Deep learning for genomic selection in forest trees: applications and challenges from high-dimensional features to multi-objective optimization

Han Fuchuan[1,2]　Gao Ming[1]　Zhao Yunxiao[1,2]　Chen Yicun[1,2]　Wang Yangdong[1,2]
(1. Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou 311400, Zhejiang, China;
2. State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: Forest tree breeding is characterized by long cycles, high heterozygosity, and significant genotype-by-environment interactions. Traditional genomic selection (GS) models face limitations when dealing with high-dimensional small-sample data and complex non-additive effects. Deep learning, with its nonlinear modeling capability based on multi-layer neural networks, offers new approaches for dissecting the complex relationships between genotype and phenotype. This technology has been widely applied in crop GS, and

收稿日期：2026−04−29　**修回日期：**2026−06−11
基金项目：中国林科院中央级公益性科研院所基本科研业务费专项（CAFYBB2022ZA002），国家自然科学基金项目（32572095）。
第一作者：韩富川，博士生。主要研究方向：林木遗传育种研究。Email：hanfc@caf.ac.cn　地址：311400 浙江省杭州市富阳区大桥路 73 号中国林业科学研究院亚热带林业研究所。
责任作者：陈益存，研究员，博士生导师。主要研究方向：林木遗传育种研究。Email：chenyc@caf.ac.cn　地址：同上。汪阳东，研究员，博士生导师。主要研究方向：林木遗传育种研究。Email：wangyangdong@caf.ac.cn　地址：同上。
本刊网址：https://j.bjfu.edu.cn；https://journal.bjfu.edu.cn

various models incorporating attention mechanisms and lightweight designs have been developed. In addition, joint modeling of multi-omics and environmental data, as well as the construction of selection indices for multi-trait synergistic optimization, are emerging as research hotspots. However, for forest tree breeding—characterized by long-term environmental interactions, highly heterozygous genomic backgrounds, and multi-objective selection demands—the adaptability of deep learning techniques and their integrative frameworks have not been systematically reviewed. This paper reviews the applications of deep learning in forest tree GS. The main contents include: (1) the evolution of genomic input features from single nucleotide polymorphisms to k-mer and graphical pan-genome node types, along with associated computational challenges; (2) modeling methods for multi-omics and environmental data; (3) structural characteristics, applicable scenarios, and hyperparameter tuning strategies of mainstream deep learning models; (4) the theoretical progression from single-trait genomic estimated breeding value prediction to multi-trait selection indices. This review also analyzes current major challenges, including data sparsity, model overfitting, and insufficient interpretability, and identifies transfer learning, semi-supervised learning, and mechanism-guided modeling incorporating biological priors as potential solutions. Finally, we envision the further integration of high-throughput phenotyping with deep learning and propose the construction of a smart breeding platform that integrates multi-omics data management, automated analysis pipelines, and breeding decision support, thereby facilitating the transformation of forest tree breeding toward genome-wide intelligent design.

Key words: forest tree breeding; genomic selection; deep learning; multi-omics integration; selection index; intelligent breeding

林木作为可再生自然资源，在维持生态平衡和支撑林产工业方面具有重要的战略价值[1]。面对全球气候变化引发的环境胁迫与病虫害频发，定向培育高产、优质且抗逆性强的林木新品种，已成为保障生态安全与产业可持续发展的迫切需求[2]。然而，林木普遍具有世代间隔长、树体高大及童期漫长等特征。此外，林木基因组通常较大且杂合度高，自然群体内连锁不平衡（linkage disequilibrium，LD）衰减迅速，导致遗传变异解析相对困难[3]。林木的主要经济性状（如生长速率、木材理化特性）多属复杂的数量性状，受微效多基因联合调控，并在漫长的发育周期中受到基因型与环境互作及潜在上位性效应的影响[4]。基于表型或谱系的选择策略对低遗传力复杂性状的选择效率较低，难以满足快速定向育种的需求。

为突破林木漫长育种周期的限制，分子标记辅助选择（marker-assisted selection，MAS）曾用于早期筛选，即通过定位与目标性状相关的数量性状位点（quantitative trait locus，QTL）进行辅助选择[5]。但在林木育种实践中，控制复杂性状的微效位点在不同作图群体和环境中难以重复验证，导致 MAS 难以有效推动林木遗传改良。严格的显著性检验标准往往会遗漏大量具有较小加性效应的基因座，使得依赖少数标记的传统分析方法难以准确解析林木复杂表型的遗传基础。针对这一技术瓶颈，全基因组选择（genomic selection，GS）作为一种极具潜力的育种策略被提出[6]。GS 摒弃了预先定位 QTL 的步骤，直接利用覆盖全基因组的高密度遗传标记对复杂数量性状进行预测[7]。通过在幼苗期估算个体的基因组估计育种值（genomic estimated breeding value，GEBV），GS 能够更充分地利用微效多基因的遗传信息，从而提高遗传增益并加速育种进程[8−9]。目前，GS 技术体系已较为成熟，在牛（*Bos taurus*）[10]、猪（*Sus scrofa*）[11−14]等畜禽以及大豆（*Glycine max*）[15−16]、玉米（*Zea mays*）[17−19]、水稻（*Oryza sativa*）[20]、花生（*Arachis hypogaea*）[21]等农作物育种中取得了显著进展。近年来，GS 技术逐渐应用于林木育种领域，已在桉树（*Eucalyptus* spp.）[22]、火炬松（*Pinus taeda*）[23]、挪威云杉（*Picea abies*）[24]、美洲黑杨（*Populus deltoides*）[25−26]、马尾松（*Pinus massoniana*）[27]等主要用材树种[28]，以及油棕（*Elaeis guineensis*）[29]、山茶（*Camellia japonica*）[30]、山苍子（*Litsea cubeba*）[31]等经济树种中成功构建了育种模型。研究表明，桉树木材密度和纤维素含量等性状的预测准确度可达 0.55 ~ 0.88[22]，火炬松生长性状的预测精度多介于 0.20 ~ 0.50 之间[23]。在针叶树中，挪威云杉的跨世代独立验证研究证实了 GS 对木材密度等高遗传力性状的预测可行性[32]。在美洲黑杨的 GS 研究中，模型对胸径的预测精度最高达 0.73[26]。此外，在山苍子的 GS 研究[31]中，基于深度学习模型对萜类化合物的预测准确性最高达到 0.78。

传统基于显式加性效应假设的线性统计模型（如 GBLUP 及贝叶斯系列算法），在林木育种中面临 3 个瓶颈：（1）林木基因组杂合度高、LD 衰减快，在标记数远超样本量的“高维小样本”条件下，难以准确捕捉显性和上位性等非加性效应[33]；（2）传统模型将环境简化为地点等分类变量，无法有效利用长时间序列的气候与土壤数据，难以刻画多年生林木跨数十年的基因型与环境动态互作；（3）生长、材性、抗逆等多性状之间存在遗传负相关，且经济权重常呈非线性，线性选择指数难以实现多目标协同改良。

深度学习[34−35]凭借非线性映射、卷积架构的局部特征提取和注意力机制的异质数据对齐能力，为解决上述问题提供了重要方法。本文系统综述了深度学习在林木 GS 中的研究进展，主要围绕以下 4 个方面展开（图 1）：（1）特征表征层，分析输入数据从单核苷酸多态性（single nucleotide polymorphism，SNP）向 k-mer 及图泛基因组节点演变的趋势与计算挑战；（2）模型层，探讨时序网络与多分支架构如何将静态基因组与动态环境、多组学数据进行联合建模；（3）验证层，评述 CNN、Transformer 等模型在林木小样本条件下的超参数调优与泛化能力局限；（4）育种决策层，阐述如何将多性状 GEBV 转化为线性基因组选择指数（linear genomic selection index，LGSI）、二次选择指数（quadratic genomic selection index，QGSI）及最优贡献选择（optimal contribution selection，OCS），以平衡生长、材质与抗逆的协同改良。最后，本文分析了模型稳健性与可解释性面临的挑战，旨在为构建林木全基因组智能设计育种平台提供方法学参考。

1　文献检索与筛选方法

本文以 Web of Science、PubMed、Google Scholar 及中国知网为主要检索数据库，检索时间范围为 2015 年 1 月至 2026 年 3 月。英文检索关键词包括“forest tree genomic selection”“deep learning genomic prediction”“multi-omics genomic prediction”“k-mer genomic prediction”“graph pangenome in plant breeding”“genomic selection index”及“optimal contribution selection in forest trees”；中文检索关键词包括“林木全基因组选择”“深度学习基因组预测”“多组学整合育种”及“基因组选择指数”。初检共获得文献 564 篇，经去重后剩余 428 篇。经阅读标题和摘要，排除与主题无关的文献 226 篇；进一步阅读全文后，最终纳入文献 81 篇。

文献纳入标准如下：（1）以林木或多年生木本植物为对象的全基因组选择实证研究；（2）以农作物或畜禽为对象，且所提方法对林木 GS 具有可迁移参考价值的研究；（3）涉及深度学习、多组学整合、k-

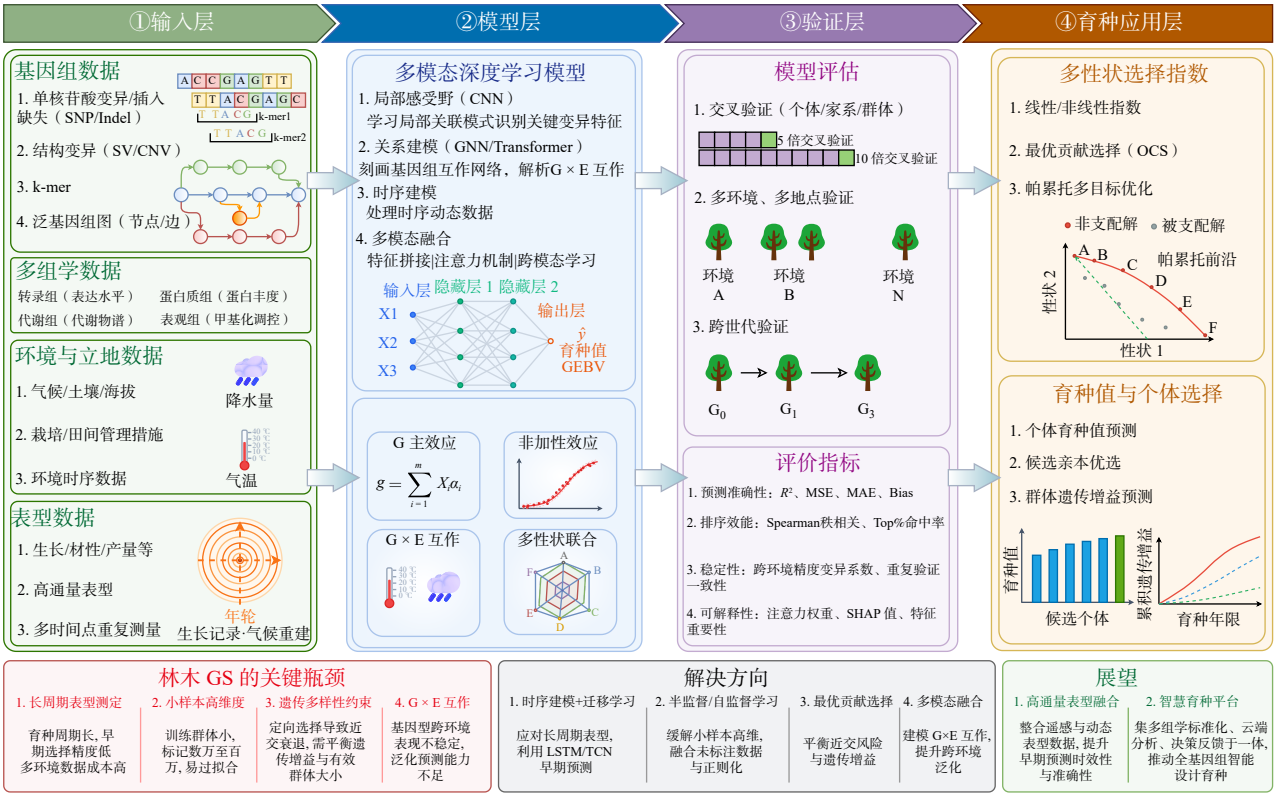


图 1　基于深度学习的林木全基因组选择多模态融合框架

Fig. 1　A deep learning-based multimodal fusion framework for genomic selection in forest trees

mer 特征或图泛基因组在全基因组选择中应用的研究；(4)关于多性状选择指数及多目标优化方法的研究；(5)同行评审论文、正式发表的综述，以及2024—2026 年间的代表性预印本。鉴于深度学习在林木 GS 中的直接实证研究目前仍较为有限，本文纳入了部分已在作物中验证的模型(如 DNNGP、GPformer、Cropformer 等)，以期为林木 GS 提供可借鉴的技术路线。

2 基因组特征的类型

基因型数据的类型和质量是影响 GS 模型预测准确性的重要因素。随着长读长测序技术的发展和林木泛基因组研究的深入，GS 模型的输入特征已从依赖单一参考基因组的 SNP，逐步扩展至结构变异(structural variation，SV)以及基于图泛基因组节点信息的序列特征。

2.1 基于参考基因组的 SNP 标记及其局限

长期以来，基于单一参考序列比对的 SNP 标记一直是林木 GS 模型的主要输入特征。获取高质量 SNP 数据，需要将短读长测序片段准确比对到参考基因组上。然而，林木通常具有较高的遗传杂合性和复杂的群体结构，个体间存在大量大片段基因组重排、插入和缺失[36]。传统分析方法在比对过程中，通常将群体中差异较大的非参考序列作为未映射读段予以剔除。这种由参考基因组偏差导致的多态性信息丢失，致使常规 SNP 矩阵无法全面反映群体的遗传多样性[37−38]。尤其对于基因组高度杂合且结构变异丰富的林木，单一线性参考基因组难以完整表征群体水平的泛基因组序列多样性，这已成为限制复杂数量性状预测精度的潜在因素[39−40]。

2.2 基于 k-mer 的无参策略

针对常规 SNP 序列比对固有的参考系偏差，基于 k-mer 的无参分析策略在复杂变异表征方面展现出显著优势。k-mer 算法将原始测序读段直接切割为固定长度的核苷酸序列，无需依赖线性参考基因组比对，从而无偏地保留群体特有的非参考序列、大型结构变异以及微小插入/缺失信息[37,39,41−42]。在 GS 的特征提取阶段，海量 k-mer 序列通常被转化为存在/缺失二元矩阵作为模型输入。该编码方式侧重于表征特定序列片段在不同个体基因组中的有无，对捕获存在/缺失变异和非同源重组片段尤为敏感。

尽管 k-mer 二元矩阵极大地丰富了变异维度，但其特征空间通常高达数千万至亿级。若将此类高维矩阵直接用于模型训练，不仅面临计算瓶颈，更易引发严重过拟合。在降维策略上，除利用主成分分析或深度自编码器将高维稀疏矩阵压缩为低维连续向量外，基于表型关联的特征筛选提供了另一种思路。例如，通过全基因组关联分析(genome-wide association study，GWAS)预先计算各 k-mer 与目标性状的关联显著性，将 k-mer 划分为显著关联集与全基因组背景集，再借助双路径卷积神经网络分别提取两类信号。这一策略能有效压缩输入维度降低噪声干扰，但需注意，GWAS 预筛更适用于加性效应主导的性状；对于非加性效应贡献较高的性状，过于严格的显著性阈值可能导致有互作价值的微效 k-mer 丢失，从而削弱网络捕捉上位性效应的能力。研究表明，引入 k-mer 特征的深度预测框架在农作物(如预测小麦品质、玉米花期)中已被证实能够达到或略优于常规 GBLUP 模型的预测精度[41−42]。目前，在杂合度较高的林木中，基于 k-mer 的 GS 研究仍相对有限，但将更丰富的序列变异整合到预测模型中，是提高复杂表型预测准确性的可行方向。

2.3 图泛基因组节点：GS 模型的图结构输入特征

k-mer 策略丰富了变异信息，但其丢失了序列的物理位置信息。近年来，杨属(*Populus*)、橡胶树(*Hevea brasiliensis*)等[43−44]林木的图泛基因组相继发表，进一步弥补了这一缺陷。图泛基因组将群体遗传多态性构建为由节点(代表非冗余、变长核苷酸片段)和连线(物理相邻关系)组成的拓扑网络，能够在无偏比对个体的同时，保留变异的序列拓扑关系和相对于参考基因组的位置信息[38]。目前，直接将图结构应用于全基因组预测的研究较少，但其在全基因组关联分析中的前沿探索为预测模型提供了重要的方法学参考。现有方法主要通过将重测序读段比对至泛基因组图，提取多态性节点的存在或缺失状态以构建群体基因型矩阵。这种节点特征矩阵展现出显著的数据表征优势：相较于无参 k-mer 方法中单一变异产生大量冗余序列的局限，以多态性节点为单位的特征表示有效降低了信号冗余，在提升变异检测敏感性的同时实现了更优的假阳性控制；此外，由于图泛基因组在构建时保留了节点与线性参考坐标的映射关系，多态性节点可直接对应至参考基因组上的物理位置，不仅增强了下游功能解析的可解释性，还有效缓解了单一直线参考在处理高杂合复杂基因组时的比对偏差。

尽管节点的存在或缺失状态显著优化了变异表征，但将其应用于林木 GS 仍面临基础资源与计算适配的双重挑战。受限于林木基因组的高杂合度、庞大基因组及丰富的重复序列，获取大规模高质量单倍型组装并构建泛基因组图的成本极高，导致目前可用的林木图泛基因组资源十分匮乏。此外，即使在现有图结构中，大型插入片段内部嵌套的小型

气泡（bubble）等复杂变异，也易干扰真实结构变异的精准提取。图节点矩阵的维度通常极高且数据高度稀疏，常规线性统计模型难以直接处理。在此背景下，深度学习方法为发挥图泛基因组的结构优势提供了潜在途径。图神经网络等模型不仅能处理高维多态性节点特征，还能直接利用节点间的连线拓扑信息，在隐空间中学习变异位点间的非线性互作组合。鉴于林木育种群体规模普遍较小且家系结构复杂，直接引入全量图节点易加剧网络过拟合风险。因此，随着林木泛基因组资源的逐步丰富，预筛选高质量多态性节点并结合深度学习模型挖掘非加性遗传效应，将是该领域具有前瞻性且亟待深入探究的技术方向。

3　多组学与环境数据在全基因组选择中的整合

传统 GS 模型通常直接建立 DNA 变异与表型间的关联，默认遗传变异可线性解释表型差异。然而，林木重要性状多受多基因控制及复杂调控过程影响。单一基因组数据难以反映表观修饰、基因表达变化及代谢调控等信息[45]。从 DNA 到表型的表达过程中存在多层级非线性调控，致使传统模型难以捕获上位性等复杂遗传效应，从而限制了预测精度[46]。

3.1　多组学整合与多模态融合策略

整合转录组、代谢组和蛋白组等中间分子表型，有助于从调控层面解析 DNA 变异与复杂性状之间的非线性关系[1,45,47]。然而，在林木育种中，多组学数据整合面临由林木生物学特性引发的时空异质性及样本匹配难题。

与一年生作物不同，林木具有显著的组织特异性和发育阶段差异。受取样可行性及成本限制，转录组或代谢组数据多采集于苗期叶片或幼嫩形成层，而目标经济性状（如材积、木材密度和纤维长度）通常需在数年至数十年后测定。这种组学测定时间与目标性状形成期的时间错位，导致早期分子表型与成树期性状之间的生物学关联不完整。同一基因型在不同发育阶段的表达模式可能存在显著差异，苗期获得的组学信号在成树期未必具有相同的调控权重。此外，多地点、多年份的组学采样会引入环境响应差异与批次效应；若缺乏标准化校正流程，模型可能将非遗传变异误判为遗传效应。

林木多组学研究还面临样本配对困难与数据缺失问题。受限于群体规模和测定成本，不同组学数据往往来自不完全重叠的个体集合，非配对样本现象较为常见[48]。这进一步增加了联合建模的难度，尤其在高维小样本条件下，配对样本的稀缺可能制约模型对组学间互作关系的学习。

在融合策略方面，直接拼接多组学矩阵的早期融合在林木小样本条件下易引发维度灾难与过拟合。后期融合策略先分别建立单组学模型再集成结果，虽稳健性较好，但可能忽略组学间的生物学互作。基于深度学习的模型级融合通过共享隐层表示或注意力机制联合学习多源特征，允许网络为不同组学数据动态分配权重，可在一定程度上缓解单一发育阶段采样带来的偏差[48]。然而，这类方法对配对样本和训练数据量的要求更高，在林木中的应用仍处于探索阶段。

总体而言，不同融合策略在林木 GS 中的适用性存在明显差异。对于当前样本规模有限的林木育种群体，信息增益并不必然转化为预测性能的稳定提升；如何在模型复杂度与泛化能力之间取得平衡，仍是林木多组学 GS 研究的核心问题。未来研究需发展适用于小样本、非配对数据和长期动态表型场景的多模态融合方法，并加强对模型可解释性与生物学一致性的验证。

3.2　环境数据整合与 G × E 时序建模

林木复杂性状的形成是基因型与长期环境互作共同作用的结果。与一年生作物相比，林木在长达数十年的生长过程中持续暴露于不同立地条件、气候波动和经营措施影响之下，因此其遗传效应往往具有显著的时间动态性与环境依赖性[46,49–50]。如何在复杂环境背景下实现稳定的全基因组预测，是林木 GS 面临的核心挑战之一。

近年来，循环神经网络（RNN）、长短期记忆网络（long short-term memory，LSTM）和 Transformer 等时序深度学习模型逐渐被应用于 G × E 建模[48]。这类模型能够整合温度、降水、光照及土壤等环境协变量，从连续时间序列中提取环境动态变化特征，从而提升复杂环境条件下的预测能力[51]。然而，当前多数研究仍主要依赖同一群体内部的交叉验证，其预测性能往往建立在训练群体与测试群体高度相关的基础之上。模型在林木育种中的实际价值，取决于其能否在跨地点、跨年份和跨世代的独立验证中维持稳定的预测能力。由于林木性状表达具有明显年龄依赖性，苗期性状与成熟期经济性状之间通常存在年龄相关性衰减现象，即随着年龄间隔增加，早期表型与成树期目标性状之间的遗传相关性可能逐渐降低。因此，即便模型在幼龄阶段获得较高预测精度，其长期育种价值仍需通过成树期及后续轮伐周期中的独立验证加以确认。此外，林木群体通常具有半同胞或全同胞家系结构，且有效群体大小相对有限，导致群体内 LD 结构在不同世代间可能发

生明显变化。当模型容量较大而训练样本不足时，模型可能过度拟合训练集的家系结构，而非捕捉个体的因果遗传变异，导致在独立群体中的预测精度明显下降。在此情况下，当模型应用于遗传背景差异较大的独立群体、后续世代或不同生态区时，其预测精度往往会显著下降。

除遗传背景差异外，长期环境异质性也进一步加剧了模型的泛化难度。林木长期生长于不同的立地与气候梯度下，同一基因型在不同生态区域中可能表现出差异化的遗传效应。因此，林木 GS 模型不仅需要在统计意义上实现跨环境泛化，更需具备跨立地迁移能力，以满足实际区域化育种部署需求。在评估模型性能时，不应仅关注随机交叉验证结果，还应综合考量训练群体与候选群体之间的遗传距离、环境差异、LD 衰减程度，以及表型测定年龄与目标选择年龄之间的关系。

未来研究需进一步加强跨环境、跨地点和跨世代的独立验证，并结合环境协变量、群体结构校正、迁移学习及领域自适应等方法，提高模型在长期育种场景中的稳健性与可迁移性。同时，应更加关注模型是否真正学习到稳定的生物学关联，而非仅仅拟合训练数据中的群体结构特征，从而推动深度学习模型由“高拟合预测”向“可泛化育种决策工具”转变。

4 面向全基因组选择的深度学习模型结构设计

随着林木基因组数据维度的增加和多源信息的引入，GS 的计算复杂度显著上升。传统线性模型在高维数据场景下面临局限，而深度学习通过灵活的网络结构，为解析复杂的基因型–表型关系提供了新途径。

目前，GBLUP 和贝叶斯方法仍是林木 GS 的主流工具。此类模型通过估计方差组分和设定先验分布实现参数收缩，能较好地估计加性遗传方差，且在样本量有限时表现较为稳定。然而，线性模型对遗传效应的描述存在两方面局限。(1)上位性和显性效应通常被归入残差项。林木生长、抗逆等性状中非加性效应占比较高，若直接在模型中引入交互项，参数数量将随标记数呈指数增加，现有群体规模难以支持有效求解。(2)当标记密度远大于样本量时，线性模型对微效多基因位点的效应区分能力有限，且高维标记矩阵中的共线性问题进一步降低了估计精度。

4.1 主流深度网络架构

多层感知机(MLP)是 GS 中较早应用的深度网络结构[52]。该网络以全基因组标记向量为输入，经全连接层与非线性激活函数变换后输出预测值。全连接结构能够拟合任意非线性映射，在不预先设定交互项具体形式的情况下捕捉全基因组范围内的上位性效应。对于受基因互作影响明显的性状，MLP 具备自动学习特征组合的潜力。因此，MLP 更适用于标记数量中等、群体规模相对有限的常规 SNP 数据集，可作为林木 GS 的基础非线性预测框架。然而，由于网络为全连接结构，对于超高维 k-mer 或泛基因组特征，其参数规模增长较快，应用受到一定限制(表 1)。

一维卷积神经网络(1D-CNN)利用局部感受野沿染色体坐标扫描标记序列，能够对相邻变异进行分组处理[31,34,53]。该算法能够有效利用由连锁不平衡(LD)导致的相邻位点非独立分离特性。此外，卷积核的参数共享机制减少了模型参数量，有助于提取局部单倍型组合特征。研究表明，对于受连续基因组区域调控的数量性状，1D-CNN 的预测性能常优于仅依赖单个位点独立效应的模型。然而，卷积核尺寸限制了单次运算覆盖的物理距离，导致跨较长物理区间或不同染色体之间的基因互作难以被直接建模。鉴于其能有效利用局部 LD 结构，CNN 特别适用于高密度 SNP 矩阵、单倍型区块及 k-mer 等具有局部序列特征的数据，在当前林木 GS 研究中具有较高的应用潜力。

Transformer 算法依靠自注意力机制计算序列内

表 1 不同深度学习架构在林木 GS 中的潜在应用场景

Tab. 1 Potential application scenarios of different deep learning architectures in forest tree GS

网络架构	主要优势	主要局限	潜在应用场景
MLP	可拟合复杂非线性关系	参数量大，难利用位置信息	中等规模SNP数据
CNN	利用局部LD和单倍型结构	难建模远距离位点互作	高密度SNP、k-mer
Transformer	捕获长程依赖关系	计算开销较大	多组学整合、G × E建模
LSTM	适合时序信息建模	依赖连续观测数据	生长轨迹、环境序列
CNN-LSTM	联合建模空间与时间特征	模型复杂度较高	多时期表型与G × E分析
CNN-Transformer	兼顾局部与全局特征	对样本和算力要求较高	复杂数量性状预测

任意两个变异位点之间的关联权重，不需要依赖位点在染色体上的邻近关系[54−55]。在 GS 建模时，自注意力可用于捕捉相距较远的位点间的上位性作用。在多组学数据整合场景下，交叉注意力模块能够对齐标记效应与环境、转录表达等变量间的依赖关系。但自注意力的计算复杂度与输入序列长度的平方成正比。直接将大量的变异位点作为序列输入，会极大消耗计算资源并增加显存占用，因此通常需要在建模前对标记进行筛选或降维。因此，Transformer 更适用于需要建模远距离位点互作、多组学融合或环境协变量整合的复杂预测任务。然而，其对训练样本规模和计算资源要求较高，在当前多数林木育种群体中的应用仍受到一定限制。

长短期记忆网络（LSTM）等循环结构适用于处理具有时间序列特征的观测数据。林木生长周期长，树高、胸径等性状的观测值随时间动态变化。LSTM 模型可接收早期观测序列、历史气象数据与固定基因型信息，学习性状随时间推移和环境波动轨迹。该结构为利用早期记录预测后期性状，或在跨年度环境下评估基因型表现提供了方法支持。但 LSTM 的应用依赖连续多年的表型记录，对于仅具备单次测定数据的育种项目，并不适用此类时序建模方法。对于包含多年连续观测数据、环境时间序列或生长轨迹记录的林木育种群体，LSTM 能够提供比静态模型更丰富的动态信息表征。然而，目前大多数林木试验尚缺乏长期连续观测数据，其应用范围仍相对有限。

4.2　面向基因组选择的深度学习模型

随着 GS 中标记维度提升与多源数据融合需求的增加，深度学习模型的设计已从直接套用基础架构，转向针对基因组数据特征的结构优化。目前，已有多种模型在卷积神经网络或 Transformer 基础上进行了专门改进（表 2）。

DeepGS 采用轻量化的一维卷积结构，依次包含卷积层、采样层、3 个 Dropout 层及两个全连接层[56]。DNNGP 则加深了网络层数，设置 3 个卷积层、一个批归一化层、两个 Dropout 层与一个全连接层，并支持多组学特征的联合输入[34]。SoyDNGP 面向大豆育种任务设计，其网络由 12 个卷积层和一个全连接层组成，并在起始与末尾卷积层后嵌入坐标注意力机制，以强化对标记位置信息的利用[57]。GPformer 基于 Transformer 编码器构建，依靠自注意力机制捕获全基因组范围内位点间的依赖关系，并引入知识引导模块融合 GWAS 结果等先验知识[58]。DeepCCR 将一维卷积网络与双向长短期记忆网络进行组合，在处理基因型空间结构的同时，将种植环境信息纳入联合建模，提升了跨环境预测的稳健性[59]。Cropformer 进一步集成了卷积操作与多头自注意力，在多个物种的独立测试集中维持了相对一致的预测精度[54]。DPCFormer 则在卷积与多头注意力的混合结构基础上，增设了基于最大信息系数的特征筛选模块，在输入维度压缩与模型表征能力之间取得平衡[55]。针对复杂的基因型–环境互作与群体特殊结构，近期模型展现出更深度的架构融合。GEFormer 构建了多模态学习框架，利用全维动态卷积提取环境时间序列特征，并通过跨模态门控机制实现基因型与环境特征的深度融合[60]。MeNet 是一种混合效应深度神经网络，通过对比学习分支提取样本间遗

表 2　面向全基因组选择的代表性深度学习模型汇总

Tab. 2　Summary of representative deep learning models for genomic selection

模型名称	核心架构	多组学	G × E	计算成本	林木适用性	参考文献/PMID
DeepGS	1D CNN + FC	否	否	较低	作物证据为主	PMID: 30101399[56]
DNNGP	CNN + BN + FC	是	否	中等	已有林木实证	PMID: 36366781[34]
SoyDNGP	CNN + CA	否	否	较高	作物证据为主	PMID: 37824739[57]
GPformer	CNN + Transformer + KGM	否	否	高	作物证据为主	PMID: 38058185[58]
DeepCCR	1D CNN + BiLSTM + FC	否	是	中等	作物证据为主	PMID: 38805625[59]
Cropformer	CNN + MHSA	否	否	较高	作物证据为主	PMID: 39690739[54]
DPCformer	Res-CNN + MHSA	否	否	高	预印本证据	arXiv:2510.08662[55]
GEFormer	gMLP + ODConv + CrossGatedMLP	否	是	高	作物证据为主	PMID: 39881541[60]
MeNet	RepGeno + VE + CIFF	否	是	较高	作物证据为主	PMID: 41267404[61]
PKDP	Dual-path CNN + FC	否	否	中等	已有林木实证	PMID: 41229065[31]

注：1D CNN.一维卷积神经网络；FC.全连接层；BN.批量归一化；CA.坐标注意力机制；KGM.知识引导模块；BiLSTM.双向长短期记忆网络；MHSA.多头自注意力机制；Res-CNN.残差卷积神经网络；gMLP.门控多层感知机；ODConv.全维动态卷积；CrossGatedMLP.跨模态交叉门控层；RepGeno.基于对比学习的基因型表征模块；VE.变异嵌入模块；CIFF.交叉信息特征融合单元；Dual-path CNN.双路径卷积神经网络。DPCFormer 为预印本状态（2025年发布于arXiv）。

传亲缘关系，并利用残差网络分支捕捉变异的非线性累积效应[61]。

针对林木基因组高杂合度、群体规模有限等特性，已有研究者开展了模型结构的针对性调整。例如，PKDP 模型采用双路径卷积架构，分别对 GWAS 筛选出的显著位点与全基因组背景 SNP 进行独立特征提取，旨在样本量受限条件下提升关键信号的贡献权重[31]。

目前多数深度学习 GS 模型主要在作物群体中开发与验证，其训练数据通常具有样本量较大、世代周期短、环境条件相对可控等特点。相比之下，林木 GS 具有明显不同的数据结构特征，包括群体规模有限、家系结构明显、长期环境暴露以及高杂合基因组背景。因此，作物模型在林木中的迁移应用仍面临若干限制。第一，深度网络通常依赖较大的训练样本规模，而林木育种群体规模往往仅为数百至数千个体，复杂模型容易出现过拟合。第二，林木群体中较强的亲缘结构可能使模型学习到家系关系而非稳定的标记−性状关联，从而降低跨群体预测能力。第三，林木性状具有长期动态变化特征，多数作物模型缺乏对多年生发育过程和环境时序效应的建模能力。第四，高杂合基因组及丰富的结构变异增加了标记编码和序列表示的复杂性。因此，当前作物领域的深度学习模型更适合作为林木 GS 的方法参考，而非可直接迁移的通用方案。未来需结合林木群体结构、长期环境数据和复杂基因组特征，对模型结构进行针对性调整。

4.3 超参数调优

深度学习模型的预测性能高度依赖于超参数的选择，包括隐藏层数量、神经元个数、激活函数类型、学习率、Dropout 率以及训练轮次等。为避免信息泄露并确保最终模型的泛化能力，通常将数据划分为训练集、调优集与测试集。其中，训练集用于拟合模型参数；调优集用于超参数选择与模型结构调整，即模型在训练集学习后，依据其在调优集上的表现更新超参数；测试集则用于评估模型对未知数据的预测性能[52]。当样本量有限时，可采用嵌套交叉验证替代上述固定划分：外层循环负责测试集轮换，内层循环完成训练集与调优集的划分及超参数搜索。针对林木 GS 场景样本量有限的特点，建议采用 k 折交叉验证结合早停机制来控制过拟合。在正则化策略上，宜优先采用小批量训练与较高 Dropout 率配置，以增强模型泛化能力。此外，应适当收窄超参数搜索空间，避免因数据量不足导致调优过程引入额外方差。实际操作中，可通过网格搜索或随机搜索遍历超参数空间，并以验证集上的均方误差（MSE）或平均绝对百分比误差（MAPE）作为优化目标。

4.4 深度学习的应用挑战

在 GS 预测任务中，深度学习并非普遍优于 GBLUP。其精度优势主要体现在训练样本量较大或目标性状上位性方差占比较高的场景，此时深度网络的预测误差相对较低。反之，当性状遗传力较高且主要由加性效应控制时，GBLUP 凭借卓越的抗过拟合能力往往表现更优[62]。因此，深度学习应定位为特定复杂场景下的补充方法，而非传统线性模型的通用替代方案。

林木高维特征及小样本训练集导致的模型不稳定性是当前突出的瓶颈之一。由于林木育种群体规模通常较小，面对百万级特征输入，网络极易过拟合训练集中的噪音[63]。尽管引入了 Dropout、早期停止及数据增强等正则化策略，深度模型的验证集方差通常仍高于线性模型。此外，超参数的微小扰动可能导致 GEBV 排名出现明显波动。模型解释性与生物学意义的缺失也是主要挑战：育种家不仅需要高准确度的 GEBV，还需解析驱动表型变异的生物学通路。尽管目前可借助 SHAP 或积分梯度等特征归因算法进行事后解释，但受限于林木功能基因组学基础相对薄弱，从非线性权重中准确追溯关键基因并验证其功能仍较为困难[64]。

最后，跨群体与跨环境泛化能力不足限制了 GS 模型的推广应用。深度模型极易对训练数据的分布特性产生过拟合。当测试群体与训练群体存在显著的数据分布差异时，其预测准确度往往大幅衰减。如何结合迁移学习与领域自适应技术，提升模型在复杂林木群体结构与强 G × E 条件下的跨群体泛化能力，是未来亟待解决的关键技术问题[50,65]。

5 从全基因组预测到多目标优化决策

在利用 GS 模型获得个体 GEBV 后，如何将预测结果转化为兼顾多性状改良、长期遗传增益与群体健康的育种决策，是 GS 技术实现实际应用的关键环节。对于林木而言，育种目标通常涉及生长量、木材品质、抗逆性、环境适应性以及遗传多样性等多个维度，因此林木遗传改良本质上属于多目标优化问题。随着深度学习模型逐渐应用于复杂性状预测，如何将单性状或多性状预测结果进一步整合至多目标选择体系，已成为智能育种研究的重要方向。

5.1 多性状 GEBV 与选择指数的整合

林木重要经济性状之间广泛存在由基因多效性或 LD 引起的遗传相关性，且部分性状间常表现为遗传负相关[66−67]。例如，生长速率提高往往伴随着木材密度下降，仅依据单一性状进行定向选择，容易

导致其他关键性状退化。因此，无论是基于传统统计模型，还是基于深度学习的单性状预测模型或多任务联合预测模型，最终均需要将多个目标性状的GEBV 进行综合评价，以服务于实际育种决策[68]。

基因组选择指数是连接 GEBV 预测与育种决策的基础框架[69−70]。其中，LGSI 通过对不同性状 GEBV 赋予经济权重并进行线性加权，实现对候选个体综合遗传价值的评价[71−72]。

$$I = \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{g}}$$

式中：I 表示综合选择指数，$\boldsymbol{w}$表示性状经济权重向量，$\hat{\boldsymbol{g}}$表示由 GS 模型输出的多性状 GEBV 向量。该方法能够在苗期对无表型个体进行早期综合评价，目前仍是林木多性状选择的主要方法。例如，在桉树和火炬松等育种项目中，常依据材积、生长量与木材密度等性状的重要性设定权重，以筛选综合表现较优的候选无性系或亲本材料[73]。

然而，LGSI 建立在线性可加假设基础之上，难以充分反映复杂性状之间的非线性关系。为此，研究者进一步提出了 QGSI，通过引入二次项描述性状间的非线性经济效应与非加性遗传互作[74]。

$$I = \boldsymbol{w}_1^{\mathrm{T}}\hat{\boldsymbol{g}} + \hat{\boldsymbol{g}}^{\mathrm{T}}\boldsymbol{W}_2\hat{\boldsymbol{g}}$$

式中：$\boldsymbol{w}_1$为线性权重向量，$\boldsymbol{W}_2$为反映性状间非线性经济价值及非加性效应的对称权重矩阵。尽管该方法在部分作物研究中表现出较高的理论选择响应，但其在林木中的应用仍受到一定限制。一方面，林木训练群体规模通常较小，而二次参数数量随性状数快速增加，易导致参数估计不稳定；另一方面，长期环境异质性使性状间协方差结构随时间和立地条件变化，从而降低非线性指数的稳定性与泛化能力。

5.2　基于帕累托最优的多目标选择

传统选择指数需要预先设定固定经济权重，然而实际林木育种目标常随造林区域、经营模式及市场需求动态调整。因此，多目标优化方法逐渐被引入林木 GS 决策体系。其中，以帕累托最优（pareto optimality）为基础的多目标进化算法受到广泛关注。在该框架下，生长量、木材密度、抗逆性和适应性等多个性状 GEBV 可同时作为优化目标，而不再被预先压缩为单一综合指标。以非支配排序遗传算法-II（NSGA-II）为代表的方法，通过迭代搜索获得一组互不支配的候选解，即帕累托前沿（pareto front）[75−76]。该结果可为育种者提供不同目标权衡下的候选个体组合（图 1）。例如，在特定立地条件下可优先选择生长量较高的材料；而在干旱区域，则可适当提高抗逆性与材性权重。相较于固定权重指数，帕累托优化提升了林木多目标决策的灵活性，更适用于复杂环境下的区域化育种部署。

5.3　遗传多样性约束与最优贡献选择

在长期林木改良过程中，持续的高强度定向选择可能导致群体近交水平上升和有效群体大小下降，从而影响长期遗传增益与群体适应能力[77]。因此，在多目标育种决策中，仅关注短期遗传增益是不够的，还需兼顾遗传多样性维持与近交控制。最优贡献选择（OCS）是目前重要的群体优化策略之一[78]。该方法在最大化群体综合遗传增益的同时，对候选个体间的基因组亲缘关系施加约束，以限制近交积累速度并维持群体遗传多样性。近年来，OCS 逐渐与 GS 和多目标优化方法相结合，用于平衡短期选择响应与长期育种可持续性。在云杉和杨树等林木的模拟研究中，结合多性状基因组估计育种值（GEBV）、帕累托优化与 OCS 的方法，已显示出在控制近交水平的同时维持长期遗传增益的潜力[78−80]。

总体而言，林木 GS 正由单性状预测逐步发展为多目标综合决策体系。未来的发展重点不仅在于提高预测精度，更在于构建能够兼顾复杂性状协同改良、环境适应性和长期群体健康的智能育种决策框架。

6　当前面临的挑战与未来展望

深度学习在林木 GS 中的应用前景，取决于能否在以下场景中展现传统方法难以替代的价值：一是多性状协同预测，尤其是性状间存在遗传负相关或非线性经济权重的情形；二是跨环境 G × E 建模，需整合长期气候时序数据以提升不同立地条件下的预测稳定性。在上述场景中，深度学习的非线性映射和多源数据融合能力具有明显优势。

然而，当前林木 GS 面临四方面关键瓶颈。第一，高质量表型数据的获取效率远滞后于基因分型能力。林木世代周期长、表型测定成本高，大规模训练群体难以建立，高维小样本问题普遍存在[3,9]。第二，模型的跨群体泛化能力不足。林木育种群体多由半同胞或全同胞家系构成，训练集与候选集之间的亲缘关系、LD 衰减模式和环境背景存在差异，导致模型在独立验证中精度显著下降[81]。第三，多组学与环境数据的整合仍面临较大挑战。由于不同组织、发育阶段和环境条件下的数据难以完全匹配，多组学数据容易受到批次效应、时间尺度差异及数据缺失等因素影响，尚未形成成熟的标准化融合框架[45]。第四，林木群体中丰富的复等位变异尚未在 GS 模型中得到充分利用。林木多为异交物种，基因组高度杂合，单个位点常存在多个等位基因。然而，现有 GS 方法多基于双等位 SNP 标记，难以完整表征复

等位结构及其连锁信息，可能降低 GEBV 的预测精度，尤其在显性或上位性效应显著的情况下。

未来研究应优先推进以下工作：一是建立涵盖多地点、多年份的林木 GS 独立验证群体，为模型在跨世代、跨环境条件下的泛化能力提供严格评估基准；二是将基因功能注释和调控网络等生物学先验融入模型结构，在降低过拟合风险的同时增强可解释性；三是构建集多组学数据管理、动态预测和育种决策支持于一体的智能化平台，将多性状 GEBV 预测与选择指数、最优贡献选择等方法衔接，实现从预测到决策的闭环。

参 考 文 献

[1] Feng J J, Dan X M, Cui Y K, et al. Integrating evolutionary genomics of forest trees to inform future tree breeding amid rapid climate change [J]. Plant Communications, 2024, 5(10): 101044.

[2] Borthakur D, Busov V, Cao X H, et al. Current status and trends in forest genomics [J]. Forestry Research, 2022, 2: 11.

[3] 杜庆章, 战鹏宇, 李鹏, 等. 基因组选择研究进展及其在林木中的发展趋势 [J]. 北京林业大学学报, 2020, 42(11): 1−8.
Du Q Z, Zhan P Y, Li P, et al. Advances in genomic selection and its development trend in forest [J]. Journal of Beijing Forestry University, 2020, 42(11): 1−8.

[4] Sharma U, Sankhyan H P, Kumari A, et al. Genomic selection: a revolutionary approach for forest tree improvement in the wake of climate change [J]. Euphytica, 2023, 220(1): 9.

[5] Muranty H, Jorge V, Bastien C, et al. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops [J]. Tree Genetics & Genomes, 2014, 10(6): 1491−1510.

[6] Meuwissen T H, Hayes B J, Goddard M E. Prediction of total genetic value using genome-wide dense marker maps [J]. Genetics, 2001, 157(4): 1819−1829.

[7] 刘策, 孟焕文, 程智慧. 植物全基因组选择育种技术原理与研究进展 [J]. 分子植物育种, 2020, 18(16): 5335−5342.
Liu C, Meng H W, Cheng Z H. Plant genome-wide selection breeding technical principle and research progress [J]. Molecular Plant Breeding, 2020, 18(16): 5335−5342.

[8] 张苗苗, 王军辉, 卢楠, 等. 林木全基因组选择研究现状和应用 [J]. 世界林业研究, 2021, 34(4): 26−32.
Zhang M M, Wang J H, Lu N, et al. Research progress and application of whole genome selection in forest tree breeding [J]. World Forestry Research, 2021, 34(4): 26−32.

[9] 朱崤, 黄敏仁. 基因组选择在林木遗传育种研究中的进展与展望 [J]. 林业科学, 2020, 56(11): 176−186.
Zhu S, Huang M R. Recent advances and prospect of the genomic selection in forest genetics and tree breeding [J]. Scientia Silvae Sinicae, 2020, 56(11): 176−186.

[10] 贾宏霞, 刘在霞, 周乐, 等. 基因组选择在肉牛中的研究进展 [J]. 畜牧兽医学报, 2024, 55(9): 3757−3768.
Jia H X, Liu Z X, Zhou L , et al. Research progress of genomic selection in beef cattle [J]. Acta Veterinaria et Zootechnica Sinica, 2024, 55(9): 3757−3768.

[11] 谈成, 边成, 杨达, 等. 基因组选择技术在农业动物育种中的应用 [J]. 遗传, 2017, 39(11): 1033−1045.
Tan C, Bian C, Yang D, et al. Application of genomic selection in farm animal breeding [J]. Hereditas, 2017, 39(11): 1033−1045.

[12] 倪世恒, 王子轶, 谢鑫峰, 等. 基因组选择技术在畜禽育种中的应用研究进展 [J]. 中国畜牧杂志, 2024, 60(6): 95−101.
Ni S H, Wang Z Y, Xie X F, et al. Advances in the application of genomic selection technology in livestock breeding [J]. Chinese Journal of Animal Science, 2024, 60(6): 95−101.

[13] 陈皖强, 樊艳凤, 唐修君, 等. 基因组选择模型及其在家禽育种中的应用 [J]. 中国家禽, 2023, 45(5): 100−106.
Chen W Q, Fan Y F, Tang X J, et al. Genome selection model and its application in poultry breeding [J]. China Poultry, 2023, 45(5): 100−106.

[14] 李棉燕, 王立贤, 赵福平. 机器学习在动物基因组选择中的研究进展 [J]. 中国农业科学, 2023, 56(18): 3682−3692.
Li M Y, Wang L X, Zhao F P. Research progress on machine learning for genomic selection in animals [J]. Scientia Agricultura Sinica, 2023, 56(18): 3682−3692.

[15] 李广, 魏天娇, 孟凡凡, 等. 大豆基因组选择与基因组选配育种的现状与展望 [J]. 吉林农业大学学报, 2025, 47(5): 761−771.
Li G, Wei T J, Meng F F, et al. Current situation and prospects of soybean genomic selection and genomic mating breeding [J]. Journal of Jilin Agricultural University, 2025, 47(5): 761−771.

[16] 齐新捧, 于洋, 张奇, 等. 全基因组选择技术在大豆育种中的应用与展望 [J]. 大豆科学, 2025, 44(6): 129−138.
Qi X P, Yu Y, Zhang Q, et al. Genomic selection in soybean breeding: progress and prospects [J]. Soybean Science, 2025, 44(6): 129−138.

[17] 张敖, 关媛, 张学才, 等. 玉米全基因组选择育种研究进展 [J]. 上海农业学报, 2023, 39(5): 13−18.
Zhang A, Guan Y, Zhang X C, et al. Research progress in genomic selection breeding of maize [J]. Acta Agriculturae Shanghai, 2023, 39(5): 13−18.

[18] 王欣, 徐一亿, 徐扬, 等. 作物全基因组选择育种技术研究进展 [J]. 生物技术通报, 2024, 40(3): 1−13.
Wang X, Xu Y Y, Xu Y, et al. Research progress in genomic selection breeding technology for crops [J]. Biotechnology Bulletin, 2024, 40(3): 1−13.

[19] 曹士亮, 张学才, 张建国, 等. 全基因组选择技术在玉米育种中的应用 [J]. 玉米科学, 2025, 33(6): 1−9.
Cao S L, Zhang X C, Zhang J G, et al. Application of genomic selection in maize breeding [J]. Journal of Maize Sciences, 2025, 33(6): 1−9.

[20] 孙连军, 王宏畅, 方婷, 等. 基因组选择技术在作物育种中的应用与进展 [J]. 中国农业大学学报, 2026, 31(4): 1−12.
Sun L J, Wang H C, Fang T, et al. Applications and advances of genomic selection in crop breeding [J]. Journal of China Agricultural University, 2026, 31(4): 1−12.

[21] 郭敏杰, 邓丽, 苗建利, 等. 基于全基因组选择的高产花生选育方法 [J]. 中国油料作物学报, 2024, 46(3): 697−702.
Guo M J, Deng L, Miao J L, et al. A method of peanut breeding of large pod and high yield based on genomic selection [J]. Chinese

Journal of Oil Crop Sciences, 2024, 46(3): 697−702.

[22] Resende M D V, Resende M F R, Sansaloni C P, et al. Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees [J]. New Phytologist, 2012, 194(1): 116−128.

[23] Resende M R, Muñoz P, Resende M V, et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.) [J]. Genetics, 2012, 190(4): 1503−1510.

[24] Chen Z Q, Klingberg A, Hallingbäck H R, et al. Preselection of QTL markers enhances accuracy of genomic selection in Norway spruce [J]. BMC Genomics, 2023, 24(1): 147.

[25] Guo C C, Yin T M, Wu H T, et al. Genomic selection with GWAS-identified QTL markers enhances prediction accuracy for quantitative traits in poplar (*Populus deltoides*) [J]. Communications Biology, 2025, 8(1): 1242.

[26] Zhou X, Zhang L, Zhang M, et al. Genomic selection for growth and wood properties in multi-generation hybrid populations of *Populus deltoides* [J]. Horticulture Research, 2025, 12(9): uhaf165.

[27] 郭臣臣, 李旗, 李思缘, 等. 杨树杂交子代苗期动态生长性状的全基因组选择 [J]. 林业科学, 2026, 62(1): 32−41.
Guo C C, Li Q, Li S Y, et al. Genomic selection for dynamic growth traits during the seedling stage of poplar hybrid population [J]. Scientia Silvae Sinicae, 2026, 62(1): 32−41.

[28] Zhu J, Liu Q, Diao S, et al. Development of a 101.6K liquid-phased probe for GWAS and genomic selection in pine wilt disease-resistance breeding in masson pine [J]. The Plant Genome, 2025, 18(1): e70005.

[29] Wong C K, Bernardo R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations [J]. Theoretical and Applied Genetics, 2008, 116(6): 815−824.

[30] Li J, Luo Y, Zhang R, et al. Decoding hybrid origins and genetic architecture of leaf traits variation in camellia via high-density 21K SNP array for genomic prediction [J]. Horticulture Research, 2025, 12(11): uhaf221.

[31] Han F, Gao M, Zhao Y, et al. Improving genomic selection accuracy using a dual-path convolutional neural network framework: a terpenoid case study [J]. New Phytologist, 2026, 249(2): 961−974.

[32] Hayatgheibi H, Hallingbäck H R, Gezan S A, et al. Cross-generational genomic prediction of Norway spruce (*Picea abies*) wood properties: an evaluation using independent validation [J]. BMC Genomics, 2025, 26(1): 680.

[33] Varona L, Legarra A, Toro M A, et al. Non-additive effects in genomic selection [J]. Frontiers in Genetics, 2018, 9: 78.

[34] Wang K, Abid M A, Rasheed A, et al. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants [J]. Molecular Plant, 2023, 16(1): 279−293.

[35] Montesinos-Lopez A, Crespo-Herrera L, Dreisigacker S, et al. Deep learning methods improve genomic prediction of wheat breeding [J]. Frontiers in Plant Science, 2024, 15: 1324090.

[36] Zhang T, Sun X, Li J, et al. Application and development prospect of genomic selection breeding in coniferous trees [J]. Planta, 2025, 262(6): 133.

[37] Zhang X, Chen S, Liu X, et al. A k-mer-based GWAS approach empowering gene mining in polyploids [J/OL]. Preprint at Research Square, 2025 [2026−05−01]. https://doi.org/10.21203/rs.3.rs-7347406/v1.

[38] Zhang Y, Wang Y, Wu T, et al. NodeGWAS: leveraging graph pangenomes for sensitive and accurate association analysis in diverse diploid and polyploid species[J]. Plant Communications, 2026: 101835.

[39] Lemay M A, De Ronne M, Bélanger R, et al. k-mer-based GWAS enhances the discovery of causal variants and candidate genes in soybean [J]. The Plant Genome, 2023, 16(4): e20374.

[40] Roberts M, Josephs E B. Previously unmeasured genetic diversity explains part of lewontin's paradox in a k-mer-based meta-analysis of 112 plant species[J/OL] Preprint at Research Square, 2024 [2026−05−01]. doi:10.1101/2024.05.17.594778.

[41] Zhang Z, Liu D, Li B, et al. A k-mer-based pangenome approach for cataloging seed-storage-protein genes in wheat to facilitate genotype-to-phenotype prediction and improvement of end-use quality [J]. Molecular Plant, 2024, 17(7): 1038−1053.

[42] He C, Washburn J D, Schleif N, et al. Trait association and prediction through integrative k-mer analysis [J]. The Plant Journal, 2024, 120(2): 833−850.

[43] Shi T, Zhang X, Hou Y, et al. The super-pangenome of populus unveils genomic facets for its adaptation and diversification in widespread forest trees [J]. Molecular Plant, 2024, 17(5): 725−746.

[44] Fang Y, Xiao X, Lin J, et al. Pan-genome and phylogenomic analyses highlight hevea species delineation and rubber trait evolution [J]. Nature Communications, 2024, 15(1): 7232.

[45] Montesinos-López O A, Montesinos-López A, Mosqueda-González B A, et al. Genomic prediction powered by multi-omics data [J]. Frontiers in Genetics, 2025, 16: 636438.

[46] Robinson H, Robles-Zazueta C A, Voss-Fels K P. Accelerating perennial crop improvement via multi-omics-based predictive breeding [J]. The Plant Genome, 2025, 18(4): e70058.

[47] Amin A, Zaman W, Park S. Harnessing multi-omics and predictive modeling for climate-resilient crop breeding: from genomes to fields [J]. Genes, 2025, 16(7): 809.

[48] Tsega A, Mullualem D. Machine learning for multi-omics data integration in crop improvement: a systematic review [J]. BMC Bioinformatics, 2026, 27(1): 81.

[49] Liu X, Wang M, Qin J, et al. GbyE: an integrated tool for genome widely association study and genome selection based on genetic by environmental interaction [J]. BMC Genomics, 2024, 25(1): 386.

[50] Brault C, Conley E J, Read A C, et al. Improving genomic prediction for plant disease using environmental covariates [J]. Plant Methods, 2025, 21(1): 114.

[51] Thingujam D, Gouli S, Cooray S P, et al. Climate-resilient crops: integrating AI, multi-omics, and advanced phenotyping to address global agricultural and societal challenges [J]. Plants, 2025,

14(17): 2699.

[52] Montesinos-López O A, Montesinos-López A, Pérez-Rodríguez P, et al. A review of deep learning applications for genomic selection [J]. BMC Genomics, 2021, 22(1): 19.

[53] Cecil R M, Sugden L A. On convolutional neural networks for selection inference: revealing the effect of preprocessing on model learning and the capacity to discover novel patterns [J]. PLoS Computational Biology, 2023, 19(11): e1010979.

[54] Wang H, Yan S, Wang W, et al. Cropformer: an interpretable deep learning framework for crop genomic prediction [J]. Plant Communications, 2025, 6(3): 101223.

[55] Deng P, Liu K, Zhou M, et al. DPCformer: an interpretable deep learning model for genomic prediction in crops[J/OL]. Preprint at Research Square, 2025 [2026−05−01]. https://doi.org/10.48550/arXiv.2510.08662.

[56] Ma W, Qiu Z, Song J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes [J]. Planta, 2018, 248(5): 1307−1318.

[57] Gao P, Zhao H, Luo Z, et al. SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding [J]. Briefings in Bioinformatics, 2023, 24(6): bbad349.

[58] Wu C, Zhang Y, Ying Z, et al. A transformer-based genomic prediction method fused with knowledge-guided module [J]. Briefings in Bioinformatics, 2024, 25(1): bbad438.

[59] Ma X, Wang H, Wu S, et al. DeepCCR: large-scale genomics-based deep learning method for improving rice breeding [J]. Plant Biotechnology Journal, 2024, 22(10): 2691.

[60] Yao Z, Yao M, Wang C, et al. GEFormer: a genotype-environment interaction-based genomic prediction method that integrates the gating multilayer perceptron and linear attention mechanisms [J]. Molecular Plant, 2025, 18(3): 527−549.

[61] Li Y, Ren S, Li J, et al. MeNet: a mixed-effect deep neural network for multi-environment genomic prediction of agronomic traits [J]. Plant Communications, 2026, 7(3): 101620.

[62] You F, Zheng C, Daniel J J Z, et al. MultiGS: a comprehensive and user-friendly genomic prediction platform integrating statistical, machine learning, and deep learning models for breeders[J/OL]. Preprint at Research Square, 2026 [2026−05−01]. https://doi.org/10.64898/2026.01.02.697306.

[63] Swain M K, Kamila N K, Jena L, et al. Hybrid deep learning framework for accurate classification of high dimensional genomic data [J]. Scientific Reports, 2026, 16(1): 5919.

[64] Novielli P, Romano D, Pavan S, et al. Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: a case study with a dataset from an almond germplasm collection [J]. Frontiers in Plant Science, 2024, 15: 1434229.

[65] Montesinos-López O A, Solís-Covarrubias A E, Hernández-Suarez C M, et al. A transfer learning method for enhanced genomic prediction [J]. Discover Plants, 2025, 2(1): 278.

[66] Villanueva B, Kennedy B W. Index versus tandem selection after repeated generations of selection [J]. Theoretical and Applied Genetics, 1993, 85(6−7): 706−712.

[67] Yasuda Y, Iki T, Takashima Y, et al. Inheritance of growth ring components and the possibility of early selection for higher wood density in japanese cedar (cryptomeria japonica D. Don) [J]. Annals of Forest Science, 2024, 81(1): 5.

[68] Moeinizade S, Kusmec A, Hu G, et al. Multi-trait genomic selection methods for crop improvement [J]. Genetics, 2020, 215(4): 931−945.

[69] Hazel L N. The genetic basis for constructing selection indexes [J]. Genetics, 1943, 28(6): 476−490.

[70] Cerón-Rojas J J, Crossa J. The statistical theory of linear selection indices from phenotypic to genomic selection [J]. Crop Science, 2022, 62(2): 537−563.

[71] Sinha D, Maurya A K, Abdi G, et al. Integrated genomic selection for accelerating breeding programs of climate-smart cereals [J]. Genes, 2023, 14(7): 1484.

[72] Ceron-Rojas J J, Crossa J, Arief V N, et al. A genomic selection index applied to simulated and real data [J]. G3: Genes, Genomes, Genetics, 2015, 5(10): 2155−2164.

[73] Duarte D, Jurcic E J, Dutour J, et al. Genomic selection in forest trees comes to life: unraveling its potential in an advanced four-generation eucalyptus grandis population [J]. Frontiers in Plant Science, 2024, 15: 1462285.

[74] Jesús Cerón-Rojas J, Montesinos-López O A, Montesinos-López A, et al. Nonlinear genomic selection index accelerates multi-trait crop improvement [J]. Nature Communications, 2026, 17(1): 1991.

[75] Castro I, Salas-González R, Fidalgo B, et al. Optimising forest management using multi-objective genetic algorithms [J]. Sustainability, 2024, 16(23): 10655.

[76] Vijai P. A hybrid multi-objective optimization approach with NSGA-II for feature selection [J]. Decision Analytics Journal, 2025, 14: 100550.

[77] Hamilton M G. Optimal contribution selection in highly fecund species with overlapping generations [J]. Journal of Heredity, 2020, 111(7): 646−651.

[78] El-Kassaby Y A, Cappa E P, Chen C, et al. Efficient genomics-based 'end-to-end' selective tree breeding framework [J]. Heredity, 2024, 132(2): 98−105.

[79] Araujo M J, Bush D, Tambarussi E V. Quantifying genetic and genotypic gain gaps in eucalyptus: the hidden cost of ignoring inbreeding and dominance [J]. Heredity, 2025, 134(9): 542−557.

[80] Gebregiwergis G T, Sørensen A C, Henryon M, et al. Controlling coancestry and thereby future inbreeding by optimum-contribution selection using alternative genomic-relationship matrices [J]. Frontiers in Genetics, 2020, 11: 345.

[81] Thistlethwaite F R, Gamal El-Dien O, Ratcliffe B, et al. Linkage disequilibrium vs. pedigree: genomic selection prediction accuracy in conifer species [J]. PLoS One, 2020, 15(6): e0232201.

（责任编辑 赵田芸 范 娟
责任编委 康向阳）