



孙连军,王宏畅,方婷,肖国政,侯晶晶,闫军,汪海,管旭东,王作平. 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(04):1-12.
SUN Lianjun, WANG Hongchang, FANG Ting, XIAO Guozheng, HOU Jingjing, YAN Jun, WANG Hai, GUAN Xudong, WANG Zuoping. Applications and advances of genomic selection in crop breeding[J]. *Journal of China Agricultural University*, 2026, 31(04): 1-12.
DOI: 10.11841/j.issn.1007-4333.2026.04.01

基因组选择技术在作物育种中的应用与进展

孙连军¹ 王宏畅^{1,2} 方婷¹ 肖国政¹ 侯晶晶¹ 闫军¹ 汪海¹
管旭东² 王作平²

(1. 中国农业大学 农学院, 北京 100193;
2. 三亚大北农创种基因科技有限公司, 海南 三亚 572024)

摘要 为全面揭示基因组选择(genomic selection, GS)技术在作物遗传育种中的现状和应用潜力,以“基因组选择”“作物育种”“深度学习”“统计模型”为关键词,在Web of Science、中国知网等数据库中检索相关文献共253篇,系统总结基因组选择技术的理论基础、方法演进、模型优化、应用实例等方面的进展和面临的挑战,重点分析采用多组学融合、神经网络等具体方法以提升基因组选择技术辅助育种的预测准确性。结果表明:基因组选择技术已从传统表型选择理论拓展至整合高通量基因组数据的现代育种实践,基因组选择模型亦由早期的以最佳线性无偏预测模型为(BLUP)核心的线性模型统计方法逐步演变为贝叶斯模型(Bayes model)、正则化方法、多性状预测、多环境反应规范模型和机器学习与深度学习模型,特别是在复杂性状预测中展现出显著优势。同时,大豆等重要作物已成为基因组选择辅助育种应用的典型代表,涵盖模型训练、训练群体优化及基因型-环境互作建模等方面。GS的突破有赖于多组学数据整合、环境-基因互作建模、可解释性算法、多模态深度学习框架,以及面向育种实践的大规模训练集构建等。可见,GS正逐步形成支撑新一代智能化育种体系的关键技术支柱,在耕地资源紧张与气候变化背景下,为实现作物遗传增益与可持续育种提供重要的支点。

关键词 基因组选择; BLUP; 贝叶斯方法; 机器学习

中图分类号 S-1

文章编号 1007-4333(2026)04-0001-12

文献标志码 A

Applications and advances of genomic selection in crop breeding

SUN Lianjun¹, WANG Hongchang^{1,2}, FANG Ting¹, XIAO Guozheng¹, HOU Jingjing¹,
YAN Jun¹, WANG Hai¹, GUAN Xudong², WANG Zuoping²

(1. College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China;

2. Sanya Dabei Agricultural Innovation Seed Gene Technology Co., Ltd., Sanya 572024, China)

Abstract To comprehensively elucidate the current status and application potential of genomic selection (GS) in crop genetic improvement and breeding, a total of 253 publications from Web of Science and CNKI were retrieved by using “genomic selection”, “crop breeding”, “deep learning” and “statistical models” as keywords. This study systematically summarized research progress in theoretical foundations, methodological evolution, model optimization, application cases, and outstanding challenges of GS, and focused on how specific strategies, such as multi-omics integration and neural-network-based approaches, can improve the predictive accuracy of GS-assisted breeding. The results indicate that: GS has expanded from conventional phenotype-

收稿日期: 2025-06-11; 接受日期: 2025-08-20

基金项目: 国家自然科学基金(32072089)

第一作者: 孙连军(ORCID:0000-0003-1541-6837), 教授, 主要从事大豆分子育种研究, E-mail: sunlj@cau.edu.cn

based selection theory to modern breeding practices that integrate high-throughput genomic data. Meanwhile, GS models have evolved from early linear statistical approaches centered on best linear unbiased prediction to Bayesian models, regularization methods, multi-trait prediction, genotype-by-environment (G×E) modeling frameworks, and machine-learning and deep-learning models, which have shown marked advantages for predicting complex traits. In addition, major crops such as soybean have become representative systems for applying GS in breeding, encompassing model training, optimizing of training populations, and modeling of G×E interaction. In the future, breakthroughs in GS will depend on multi-omics data integration, robust modeling of environment-genotype interactions, interpretable algorithms, multimodal deep-learning frameworks, and the construction of large-scale training datasets tailored to practical breeding programs. Overall, GS is becoming the key technological pillar of next-generation intelligent breeding systems and providing an important leverage point for achieving genetic gain and sustainable crop improvement under constraints of limited arable land and ongoing climate challenges.

Keywords genomic selection; BLUP; BAYES; machine learning

当前,全球农业可用耕地面积持续缩减,全球气温持续上升、极端天气事件日益频繁,对主要作物的稳产高产构成了严峻挑战^[1]。已有研究指出,到2050年全球粮食供给将呈下降趋势^[2]。因此,挖掘作物遗传潜力、提升优良品种选育效率,对于保障全球粮食安全至关重要。传统育种方法因周期长、效率低,难以满足快速增长的粮食需求。为突破这一瓶颈,基因组选择(genomic selection, GS)作为一种新兴育种技术应运而生。该方法通过在训练群体中整合全基因组标记与表型信息建立预测模型,直接估计个体的基因组育种值(genomic estimated breeding value, GEBV)。与传统方法相比,GS弱化了对单个主效基因(QTL)逐一解析的依赖,从而显著提升了早期选择效率并加快了育种进程^[3]。因此,在作物育种中开展基因组选择的应用研究,对提升复杂数量性状的早期选择效率及预测准确性、缩短育种周期并支撑主要作物稳产高产具有重要意义。

目前,GS已从理论层面的“方法学验证”逐步走向实际育种流程的应用实践。在水稻精英育种系中,已利用多环境信息与环境协变量开展多环境基因组预测以服务目标环境群(target population of environments, TPE)的材料筛选^[4]。在小麦中,复合群体的快速循环基因组选择(rapid-cycle recurrent GS, RCRGS)在多轮重组后获得了稳定的实际遗传增益,体现出缩短育种周期、提高单位时间遗传增益的潜力^[5];在玉米多环境试验中,基于统计与机器学习方法对缺乏实测数据的单交杂交组合的产量与花期进行预测,可为减少田间测试成本、优化杂交组合筛选提供了可操作路径^[6]。已有研究报道,在大豆中应用GS辅助育种,结果显示GS对提高籽

粒产量的改良效果可与表型选择的效果相当,并在一定选择强度下体现出优势^[7]。但GS在不同作物与育种体系中的“可复制落地”仍受关键环节制约:包括训练群体规模、代表性及其动态更新策略、多环境(G×E)和稀疏测试情景下预测稳健性,以及机器学习/深度学习在融合多组学和多模态数据时的真实增益边界及实现条件等^[8]。目前,针对上述核心制约因素的综合评价研究尚不多见。本研究围绕“统计模型-机器学习/深度学习-育种流程嵌入”的演进主线,系统梳理GS方法及其应用进展,并聚焦训练群体构建、深度模型训练及G×E建模等核心问题,旨在全面揭示基因组选择技术在作物遗传育种中的现状和应用潜力,以期作物遗传改良中GS规模化应用提供参考。

1 基因组选择的理论基础与方法演进

在同种作物种质材料中精准选出具有理想表型的个体,是育种者最为关注的核心目标。理想表型不仅能增强作物对环境胁迫的抗性,还能进一步提高产量^[9]。虽然合理的田间管理有助于作物的生长,但在特定环境条件下,作物的表型潜力主要由基因型决定。因此,在选育适应特定环境的优良品种过程中,环境因素、标准化的育种流程以及规范的测量手段同样发挥着至关重要的作用^[10]。Duvick^[11]提出的遗传增益理论指出,遗传育种的终极目标是能够预测并在杂交过程中获得最优的后代,从而实现产量的持续遗传增益。因此,表型预测成为育种流程中的关键环节。

传统育种依赖包括小区试验、区域试验及生产试验在内的多级田间验证体系,这些过程对人工表

型鉴定、环境条件和试验规模具有较高的要求,人工测量误差、环境噪声以及跨环境表型表现的不稳定性,往往限制了对基因型真实潜力的准确判断。此外,传统育种周期长、数据利用率有限,使得对品系的预测和筛选效率仍有待提升。因此,更为精准、高通量和数据驱动的表型预测方法已成为现代育种体系的重要发展方向。但要强调的是,GS并非脱离田间试验的“纯计算”方法,而是建立在育种体系长期积累的表型和基因型数据之上。

2001年,Meuwissen等^[12]提出基因组选择(genomic selection, GS)的概念。随着高通量测序技术的发展,借助全基因组分子标记进行标记辅助

选择可有效筛选出优良的种质资源。GS可被视为一种计算工具,能够从群体中筛选出最具竞争力的个体。目前,GS主要有两类方法:一类是以线性无偏预测理论(best linear unbiased prediction, BLUP)、贝叶斯(BAYES)及其变体为代表的统计模型,采用包含固定效应和随机效应的混合线性模型(如ssBLUP、GBLUP、BayesA、BayesB等),适用于加性遗传性状的预测^[13];另一类是机器学习模型,如随机森林(random forest)、支持向量机(supporting vector machine, SVM)及各类深度学习模型,这些方法可构建非线性的预测模型,更好地捕捉基因间复杂的非线性交互效应^[14](图1)。

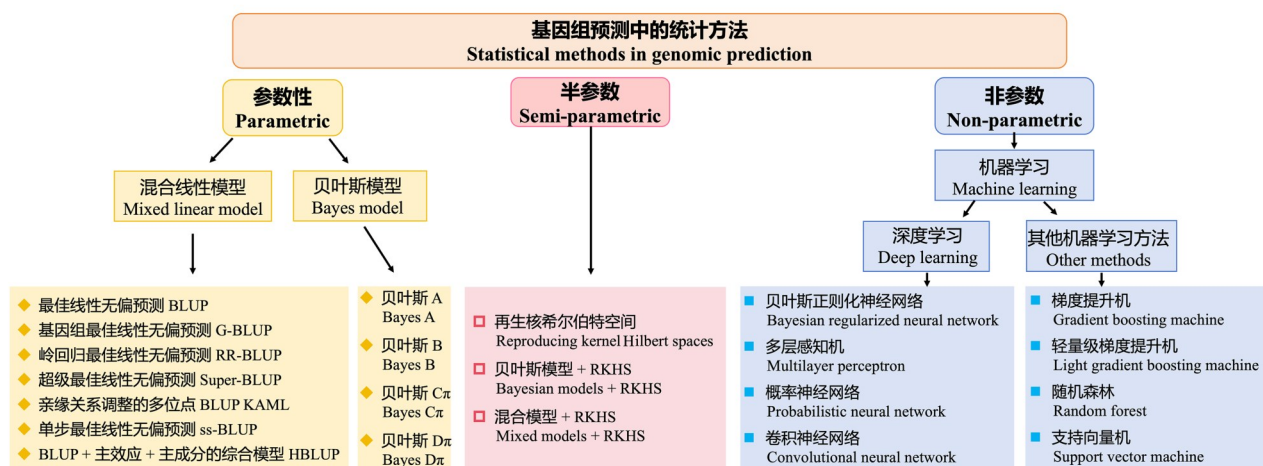


图1 基因组选择(GS)统计方法

Fig. 1 Statistical methods in genomic selection (GS)

伴随着深度学习技术的持续演进,计算机科学中的诸多先进算法被逐步引入基因组选择(GS)领域。在GS实际应用中,卷积神经网络(convolutional neural network, CNN)技术正发挥重要作用,通过引入增强型正则化层来减轻模型的过拟合问题,并采用更适用于作物基因组预测的激活函数,能够显著提升模型的预测性能。近年来,已有研究者开始整合包括基因组、转录组、蛋白质组和代谢组的多组学数据^[15],来提高GS的预测准确率。多源信息整合不仅提供了更为全面的生物学视角,有助于揭示控制性状形成的分子机制,还可能发掘出新的遗传标记。

2 分子标记在植物育种中的演进与作用

20世纪80年代以来,随着各类分子标记系统的不断发展,研究人员能获取的多态性标记数量大幅

增加^[16]。其中,最为广泛使用的高通量基因分型系统是单核苷酸多态性(SNP),它在数量性状位点(QTL)挖掘中应用广泛。目前已有覆盖12种作物的120余项研究,利用不同标记系统共识别出超过10 000个QTL,主要用于改良作物的经济性状^[17]。随着参考基因组的发布,针对特定作物的固相芯片技术也得到了快速发展。以大豆为例,随着‘Williams 82’参考基因组的公布,研究人员相继开发了Illumina SoySNP 50k及Affymetrix SoyaSNP 180k等高密度芯片。这些高通量分型工具大幅降低了全基因组分型的成本和难度,为GS在复杂性状预测中的规模化应用提供了坚实的数据支撑^[18]。

分子标记最初被整合在传统表型选择体系中,形成了标记辅助选择(marker-assisted selection, MAS)策略,通过选择与主要QTL相关的个体来提

高育种效率^[19]。然而,由于不同环境或遗传背景下难以稳定识别相同的QTL,该方法在改良复杂数量性状方面成效有限。为克服连锁分析的局限,引入了关联分析,利用非亲本群体和更高的重组率提升性状的精细定位能力,但在检测稀有变异时仍存在挑战^[20]。GS的出现革新了育种策略,其并不依赖少数显著标记,而是利用全基因组的分子标记数据,结合训练群体的基因型和表型信息,来预测测试群体中个体基因组的育种值(genomic estimated breeding value, GEBV)^[21]。

基因组选择(GS)在加快育种进程、降低田间表型鉴定成本方面展现出了极大的潜力,尤其在玉米等作物中,已有大量研究验证了其显著成效^[22]。然而,GS的广泛应用仍受到若干因素的制约,例如基因分型成本仍偏高、操作流程尚未标准化等问题,制约了其在实际育种工作中的推广。

与GS紧密相关的遗传预测(genetic prediction, GP)则更加强调对个体基因潜力的整体评估,涵盖了加性效应和非加性效应的预测能力。在具体的育种实践中,GS和GP的预测效果往往受多种因素影响,包括训练群体的数量和遗传多样性、目标性状的遗传力大小等。因此,如何优化这些参数以提升模型的准确性,仍是当前研究的重点。

尽管面临诸多挑战,GS和GP已在多个作物育种实践中获得了实际应用成效。国际玉米与小麦改良中心(Centro Internacional de Mejoramiento de Maiz y Trigo, CIMMYT)及国际干旱热带作物研究所(International Crops Research Institute for the Semi-Arid Tropics, ICRISAT)等机构通过多年实践证实,这些方法有助于加速育种材料的筛选过程,提升选育效率^[23]。分子标记辅助育种技术与基因组选择方法的结合,正在深刻改变传统育种的模式。尽管在数据处理和遗传多样性控制方面仍存在技术瓶颈,但从整体趋势来看,未来发展的潜力仍不可小觑。

3 统计模型在基因组选择中的应用

3.1 GS中的建模基础及策略

基因组选择(GS)准确性的实现依赖于高效且稳健的统计建模方法,以便在全基因组水平上快速而准确地估算个体的基因组育种值(GEBV)。GS是一种以全基因组分子标记为基础的辅助选择技

术,其核心做法在于,通过在全基因组范围内标记与目标性状表现值之间的关联,推断出各基因位点或染色体区段对性状的贡献,并将这些效应累加,从而获得个体遗传潜力的估计值。相较于传统的标记辅助选择(marker assisted selection, MAS),GS对多基因控制的复杂性状的预测具有显著优势。MAS常因所用标记数量有限,难以捕捉全部遗传变异或准确定位数量性状位点(QTL);而GS通过密集标记覆盖克服了这一限制,大大提升了对复杂性状遗传的解释能力。在实践中,提高GEBV的预测精度仍是关键技术目标之一,同时也要求对标记效应的估计更加精细和可信^[24]。

在深度学习和现代机器学习方法尚未普及之前,GS的研究主要依赖于传统的统计建模工具。模型的选择在很大程度上影响着预测结果的准确性和计算效率。目前被广泛应用的两类核心模型包括:以混合线性模型(如,BLUP系列)为代表的线性预测方法,以及基于贝叶斯理论框架构建的算法模型。这两大类方法各具优势,研究者通常会根据数据特征和性状遗传背景进行有针对性地选择。

3.2 基于BLUP ALPHABET的方法

从20世纪中期起,最早的BLUP就被广泛用于遗传评估。该理论于20世纪60年代由Charles Henderson提出,最初依赖系谱信息估算个体间的遗传方差与相关性,用于评估遗传值,在动物育种中发挥了重要作用^[25]。Henderson^[25]首次提出了混合线性模型(mixed linear model, MLM)方程,该模型将随机效应与固定效应整合进同一个方程中,解决了普通线性模型中最小二乘法过拟合的问题。MLM基于对数似然优化而非最小残差平方和,因此更具灵活性,尤其适用于自然语言处理等复杂模型任务。在20世纪70年代,Henderson又引入了可直接由系谱信息构建的逆相关矩阵,简化了遗传评估的计算过程。20世纪90年代,随着分子标记技术的兴起,Bernardo将标记间的遗传关系引入BLUP模型,标志着遗传评估方法从依赖系谱数据到依赖更精确的分子标记转变;Bernardo^[26]通过限制性片段长度多态性(restriction fragment length polymorphism, RFLP)标记构建玉米材料的全基因组关系矩阵,并使用BLUP预测未知杂交组合的产量,取得了较高的准确度。相比于系谱构建的遗传关系矩阵,基于分子标记构建的遗传关系矩阵更具

优势,因为后者能降低孟德尔随机分离误差、基因重组和等位基因偏分离等因素的影响。

2001年,Meuwissen等^[12]提出了一种新的BLUP模型,将所有标记效应视为独立、服从正态分布,均值为零、方差相等,并通过求解残差平方和最小值估计个体的遗传效应,这为gBLUP的建立奠定了理论基础。随后,Zhang等^[27]将标记构建的遗传矩阵引入MTDFREML模型。VanRaden等^[28]于2008年提出基于全基因组SNP的高效基因组关系矩阵(G矩阵)构建方法,显著提高了gBLUP的计算效率。2011年,岭回归(ridge regression)的引入扩展了gBLUP模型的应用^[29]。Misztal等^[30]则整合了系谱和标记信息,提出了单步BLUP(ssBLUP)^[31],可同时对基因型和非基因型个体进行遗传评估,成为育种项目中的主流方法。2019年,Wang等^[32]提出了SUPER BLUP和压缩BLUP(compressed BLUP),通过引入GWAS显著标记和对个体分组,在处理低遗传率性状及复杂群体结构方面提高了预测的准确性。此外,2020年Yin等^[33]提出了一种新的机器学习方法KAML(kinship-adjusted multiple loci mixed model),通过交叉验证、多重回归、网格搜索与二进制算法联合优化模型参数,不仅提高了预测准确率,还大幅降低了计算成本,预测性能超过传统的贝叶斯方法。2023年,Yin等^[34]又提出了HIBLUP模型,以往大多数工具需对矩阵反复求逆,而HIBLUP采用“HE+PCG”策略跳过大矩阵逆运算,显著提升了计算效率,有效解决了大规模基因组数据中方差组分估计计算复杂度过高的问题。

3.3 基于Bayesian ALPHABET的方法

在基因组学和数量遗传研究中,基于分子标记的亲缘关系分析(marker-based relationship analysis)是群体遗传结构和加性遗传方差分析的重要工具,可用于评估基因变异对目标性状表型表达的贡献^[28]。当分子标记广泛覆盖整个基因组时,亲缘矩阵可近似替代传统系谱关系,从而为遗传研究提供更全面的视角。

性状的总加性遗传效应可以被表示为各个标记效应之和。然而,面对“高维小样本”(large p , small n)问题,即参数数量远多于样本数的情况,传统遗传标记分析方法可能会导致模型过拟合而失效^[35]。为解决此问题,贝叶斯模型将遗传标记视为

随机效应,假设其效应服从以“0”为均值、方差未知的正态分布。Meuwissen等^[12]提出了基于贝叶斯理论的GS计算框架;这一框架不同于BLUP模型中假设所有标记具有相同的遗传变异,而是认为整个基因组中只有少部分SNP对性状具有显著作用,即可能与QTL相关联,大部分SNP是无效的。因此,贝叶斯模型中的标记效应方差被设定为服从某种先验分布,以引导模型自动识别重要标记。

在Bayes A模型中,假设每个标记效应都服从正态分布,其方差服从逆卡方分布,该分布由统一的超参控制。这一模型假设每个SNP都有一定效应,只是方差不同。Bayes A通常采用马尔可夫链蒙特卡罗方法(markov chain monte carlo, MCMC)中的吉布斯采样,以估算每个SNP的效应值及其他模型参数^[36]。若在Bayes A的基础上进一步假设只有少部分SNP具有非零效应,其余标记效应为0,这一变体即为Bayes B模型。Bayes B对SNP效应设置混合先验分布:部分标记效应为0,其余服从正态分布,方差仍为逆卡方分布。该模型在推理时会联合使用吉布斯采样和Metropolis-Hastings采样,以估计标记效应与其对应的方差。在贝叶斯B模型中, π 为关键参数,表示某一SNP效应为0的概率;当 $\pi=1$ 时,Bayes B即退化为Bayes A。

Bayes B特别适用于仅有少量遗传变异控制性状的情况,其预测精度通常优于Bayes A。但Gianola等^[37]研究指出,Bayes A和Bayes B的统计缺陷在于模型对先验参数设定的敏感性,SNP效应的先验分布仍可能对后验结果产生主导性影响,从而降低预测可信度。此外,Bayes A和Bayes B普遍假设 π 为已知固定值,实际模型中这种人为设定可能限制SNP效应的自由收缩程度,降低了模型的学习能力。为此Bayes C π 模型被提出,不再固定 π ,而是将其视为未知变量,为其设置先验分布,由模型通过数据计算出最优收缩比例。Bayes D π 模型则假设 π 和尺度参数 S 均为未知量,并设置(1,1)Beta分布先验,可以直接从后验分布中进行采样,从而提高参数估计的灵活性与鲁棒性。尽管Bayes系列模型在理论上具备强大的适应能力,但其主要挑战在于计算复杂度高,尤其是MCMC收敛过程需进行大量迭代,计算时间较长。目前已有多个R包(如BGLR)被开发以实现上述方法,极大便利了Bayes类方法在作物育种中的应用^[38]。

4 机器学习在基因组选择中的应用

4.1 机器学习助力GS的背景及价值

作为计算机科学的重要分支,机器学习(machine learning)的核心目标在于让计算系统能够通过算法自动识别数据中的潜在特征,并不断优化其预测性能,而无需依赖传统的显式编程过程^[39]。近年来,随着基因组测序技术的进步和成本的大幅下降,所能获取的基因组数据量迅速扩大,这也为机器学习在基因组选择(GS)领域的应用提供了更广阔的空间。与传统的统计方法相比,机器学习展现出独特的优势,它不仅具备处理大规模数据的能力,还能构建高效的预测模型以估计标记效应,从而覆盖复杂性状相关的遗传变异。相较于传统统计方法,机器学习不依赖于对数据分布或遗传模型的严格假设,能够以“黑箱模型”方式处理非线性与复杂的交互结构,在大群体样本表型预测中表现出较高的准确度^[40],并显示出良好的泛化能力和实际应用潜力。

4.2 监督学习及无监督学习

在机器学习的众多分支中,监督学习和无监督学习是最为基础的分类方式。其中,监督学习以已知的输入变量如单核苷酸多态性(single nucleotide polymorphism, SNP)和相应的输出变量(如表型)为基础,构建预测函数或回归模型,并通过验证集来评估模型的偏差和方差,从而优化预测性能^[18]。应用于基因组选择(GS)的常见监督学习算法包括随机森林(random forest, RF)、梯度提升机(gradient boosting machine, GTB)、再生核希尔伯特空间方法(reproducing kernel hilbert space, RKHS)以及支持向量机(support vector machine, SVM)等。这些方法的核心目标均在于有效控制模型的偏差和方差,从而增强泛化能力,提高对未见数据的预测准确率。

模型性能的优劣往往受多种因素影响,包括群体的遗传结构^[41]、性状的遗传构成(如加性效应和上位性效应)^[42]、所选算法的建模能力^[43],以及输入数据的质量和完整性等。在玉米、水稻和大豆等多种作物中,系统比较了3类线性模型(rrBLUP和Bayes A/B)和监督学习算法(如SVR、RF、GTB)之间的性能差异,并采用SVM、GBLUP和GBM型对水稻和小麦的表型进行预测,结果显示监督学习方

法在多个复杂性状的建模中具有较强优势,表现均优于传统统计模型^[44]。在处理特定抗病性状时,机器学习往往能提供不同于线性模型的视角。在针对大豆亚洲锈病的研究中,对比RR-BLUP、Bayesian LASSO和机器学习算法,发现Bayesian LASSO和RR-BLUP在预测准确度上略有优势(约0.52~0.53),但支持向量机(SVM)等方法也能达到与GBLUP相当的水平,这表明在缺乏显性主效基因的背景下,不同模型捕捉微效应点累加效应的能力存在差异,需根据性状遗传架构灵活选择^[45]。

无监督学习的目标是从输入变量中挖掘潜在结构,而无需依赖输出变量。当前常用的方法包括主成分分析(PCA)、聚类分析、自组织映射等。虽然它们不直接用于GS预测,但在揭示种群结构、遗传相关性和异质性方面具有重要作用^[46]。高级无监督方法如特征选择、层级聚类、密度峰估计等正在不断扩展GS的研究边界^[47]。

4.2.1 随机森林(random forest)

Breiman^[48]提出的随机森林是一种基于决策树的集成学习算法,适用于处理大量遗传标记数据。它通过构建多个独立的决策树并汇总其预测结果,大幅提高了预测准确性。每棵“树”的构建过程中随机选取部分特征用于节点分裂,有助于识别重要变量^[49]。变量重要性指标在做无基因组选择中一般用于分析各SNP对预测结果的贡献。

4.2.2 梯度提升机(gradient boosting machine)

GBM同样基于决策树,但采用迭代方式构建模型,每一步都试图修正前一模型的残差。每次节点分裂仅选取1个最优化的SNP,以最大化子节点纯度,所以它的模型更精细。虽然GBM的预测准确度高于RF,但其训练时间较长,且不支持并行计算。为了解决这一难点,Yan等^[50]开发了LightGBM,该模式具有极高的训练速度和内存效率,适合大规模结构化数据。LightGBM采用“叶子优先”决策树生长策略,可更有效地识别重要的SNP,同时提供SNP重要性分析工具,大大降低了基因分型成本。已有研究表明,LightGBM在预测准确率、模型稳定性和计算效率上均优于传统方法^[51]。

4.2.3 其他方法

SVM和RKHS均采用核技巧(kernel trick),通过在高维空间中进行线性运算以捕捉非线性关系。SVM不需要显式映射高维空间,而是通过核函数在

原始空间的计算点积来构建半参数非线性遗传模型。RKHS可通过高维核函数捕捉主效应、显性效应和上位性等高阶遗传作用^[52]。尽管RKHS不被视为典型的机器学习方法,但其在GS前期特征提取中可辅助提升后续模型效果。

4.3 集成学习(ensemble learning)

集成学习是通过组合多个学习模型来提升预测准确率的方法,可与监督或无监督学习结合使用。它适合处理复杂的基因组数据,能够有效整合多个模型的信息,从而增强对遗传变异与表型之间关系的预测能力。集成学习在GS中的典型应用包括:1)多模型集成:结合线性回归、SVM、RF等模型,利用模型多样性提升整体预测效果^[53];2)特征整合:将不同来源(如SNP、基因表达)的标记信息融合,获取更全面的遗传图谱^[54];3)数据整合:处理来自不同种群和环境条件的数据,提高模型在基因-环境互作预测中的鲁棒性^[55]。而典型方法包括:Bagging(套袋),如随机森林,通过对原始数据多次采样建立多个决策树,并取平均值(mean)或投票结果;Boosting(提升),如GBM,每个新模型致力于修正前一个模型的预测误差;堆叠(stack),利用多个基础模型的预测作为输入,再训练一个元模型进行最终预测。这些方法通过不同方式降低预测误差,提高模型准确性和稳定性,是GS中的重要工具。

5 深度学习在基因组选择中的应用

深度学习(deep learning, DL)作为机器学习的一个重要分支,适用于监督和无监督学习任务,其本质是通过多层神经网络结构自动提取输入数据中的特征信息。在DL模型中,多个隐藏层被逐层堆叠,每层包含多个神经元。每个神经元根据激活函数对前一层输入赋予权重后再传递,最终逐步推进至输出层。GS中常用的深度学习结构包括:前馈神经网络(feedforward neural network, FNN)、循环神经网络(recurrent neural network, RNN)和卷积神经网络(convolutional neural network, CNN)。其中FNN信息单向流动,层间无交互,更容易过拟合。CNN利用卷积操作将输入序列中关键区域集成用于决策,能有效提取局部特征。GS数据多为一维结构,因此常采用一维卷积(1D CNN)。考虑基因型数据多为0、1、-1的分类变量,通常先进行one-

hot编码处理后输入网络^[56]。

Gianola等^[37]将贝叶斯正则化神经网络(bayesian regularized neural network, BRNN)用于预测小麦产量,并与Bayesian岭回归模型对比,预测准确率提高了11%~18%。Pérez-Rodríguez等^[57]对比了浅层神经网络、RKHS方法和Bayes A/B模型,在12个环境中预测谷物产量和抽穗期,发现神经网络虽非在所有环境中最优,但整体预测能力优于线性模型。González-Camacho等^[58]基于玉米55 000个标记测试发现, RKHS和浅层神经网络优于Bayesian LASSO,在捕捉上位性方面表现更佳。随后,研究者广泛采用浅层全连接神经网络(含1~2个隐藏层)进行预测。在涉及小麦和玉米数据集的研究中,概率神经网络(probabilistic neural network, PNN)相比MLP模型可进一步提高谷物产量预测精度^[58]。

2018年, Ma等^[59]首次将CNN应用于GS, 提出DEEPGS模型。该模型结合卷积层、池化层和Dropout技术,考虑了近邻标记对表型的集成效应。粒长、千粒重、蛋白含量等6个性状的预测性能较RR-BLUP高2%~7%。与RR-BLUP、GBLUP和3种MLP模型比较, CNN预测性能最强相关系数(R^2)为0.742, 其他分别为0.737、0.731、0.409、0.363、0.428。此后, CNN在GS领域广泛应用。Jubair等^[10]将集成CNN模型用于预测6个小麦性状,发现其性能比非集成模型高20%~30%。

Liu等^[60]提出双通道CNN架构,输入后分别通过2个独立的CNN分支,输出后融合进入全连接网络进行预测。该结构引入残差连接,模型在大豆数据集中预测性能优于DeepGS、MLP、RRBLUP、Bayes A、Bayesian Lasso等方法。显著性分析显示,该模型更加关注与目标性状强相关的关键SNP位点。基于深度神经网络可视化解释方法,可从训练模型中识别具有生物学意义的标记,为优质种质资源筛选与基因-性状关系研究提供了新框架。除SNP信息外,也有研究者使用转录组或蛋白组信息作为输入^[51]。利用DNNGP模型将转录组数据用于玉米性状预测,在所有测试数据中预测准确率最高,超过GBLUP、LightGBM、SVR、DeepGS和DLGWAS等方法。

6 基因组选择育种平台的构建与应用

当前,无论是传统统计模型,还是基于机器学

习和深度学习的方法,均已在GS研究中构建起了较为成熟的理论体系和应用框架。然而在实际育种应用中,不同类型的数据特征常常对模型预测准确率产生显著影响,不同模型在实际场景中的表现亦存在差异^[61]。因此,构建一体化、自动化的数据处理平台,覆盖从原始数据读取到预测分析的完整流程,已成为现代育种工作中的关键技术方向之一^[62]。为满足这一需求,多个集成化平台相继开发并投入使用。Gao等^[63]提出的Crop-GPA是一款面向科研用户的在线平台,集成了文献挖掘、数据库管理和模型预测等功能,便于研究者从大量文献中快速识别潜在的基因-性状关系,并进行相关验证。随着人工智能技术的发展,现代智能育种平台已具备强大的数据整合能力,能够协同处理基因型与表型数据。Feng等^[64]通过集成多种经典机器学习与深度学习模型,显著提升了多作物育种的预测精度与效率。

在大豆育种中,Gao等^[65]发布了SoyDNGP,这是一个在线深度学习平台,允许用户上传VCF文件并调用平台内置的三维结构CNN模型进行GS预测分析。IP4GS是专为植物育种者设计的Web端GS分析平台,集成7种常用模型、11个评估指标及可视化模块,特别适合不具备生物信息学背景的用户进行快速GS分析^[66]。CropGS-Hub整合了多种作物的基因组和表型数据,内置多个GS预测和杂交设计算法,致力于支持跨物种的基因组研究,为大规模育种试验提供数据支撑^[67]。G2P容器平台是基于Singularity容器和R平台构建的,提供可复现的环境,集成16种GS模型和13种评估指标,实现了从数据预处理、模型构建到预测评估的完整流程^[68]。智能育种平台(smart breeding platform)面向高通量群体遗传学、表型组学和GS分析,支持数据管理、表型测量、基因组管理及案例复用,帮助育种者全面利用作物的遗传潜力^[69],见表1。

表1 育种平台功能对比
Table 1 Breeding platforms function comparison

模型 Model	主要功能 Main function	支持作物 Supporting crop	预测模型类型支持 Supported genomic prediction model types
Crop-GPA	集成文献挖掘、数据库管理、模型预测,支持基因-性状关联分析和验证	多作物	多种统计/ML模型
SoyDNGP	针对大豆分子育种设计,上传VCF后调用内置三维CNN模型进行预测分析	大豆	三维结构CNN(深度学习)
IP4GS	集成7种GS模型、11个评估指标及可视化模块,适合无生信背景用户快速分析	多作物	GBLUP、RR-BLUP、Bayes B
CropGS-Hub	跨物种数据整合,支持多模型预测与杂交设计,适合多作物联合分析	多作物	多模型集成
G2P 容器平台	基于Singularity + R,提供可复现分析环境,支持16种模型和13种评估指标	多作物	统计模型 + ML/DL 模型
Smart breeding platform	覆盖数据管理、群体遗传、GS分析与案例复用,全流程智能化	多作物	多种模型

这些平台由多模型集成,支持统计模型与深度学习模型的协同调用;操作简单,分析流程自动化,可减少人为操作,数据处理效率高,适合不具备生物信息学背景的用户;支持玉米、水稻、大豆、小麦等主粮作物。未来随着技术的不断发展,GS平台将向更深层

次的基因变异挖掘、复杂性状联合预测和性状选择方向演进,为作物育种提供更全面的决策支持。

7 总结与展望

基因组选择(GS)作为一种作物育种预测工具,

近年来在全球范围内受到广泛关注。它通过在测试群体中仅测定基因型、在训练群体中同时获取表型和基因型信息,从而在早期阶段实现候选个体的筛选。相比传统育种方法,GS成本更低、周期更短,能够大幅提高遗传增益。尤其对于大豆这类传统育种周期长达6~8年的自交作物,GS能够有效缩短世代间隔,加速优良变异的聚合。与依赖表型观察的传统杂交育种方式相比,GS能够直接利用基因信息识别优势等位基因,同时考虑基因间的互作和环境效应,揭示复杂性状的遗传结构,更精准评估基因潜力,有助于选育适应性更强的作物品种。

目前,GS技术正从传统线性模型向非线性机器学习模型、再向深度学习模型演进,模型复杂度和预测精度不断提升。然而,GS在实际应用中仍面临诸多挑战:1)训练深度学习模型仍是一项复杂任务,涉及大量超参数的调试,如网络层数、单层神经元个数、学习周期数、正则化方式、激活函数类型等,这些参数间存在高度耦合,需大量试验进行组合测试,训练难度和计算成本较高^[70]。2)GS模型的性能高度依赖训练群体的规模和代表性,群体过小则难以捕捉全面的遗传变异,群体过大则造成训练时间过长且模型过拟合的风险上升^[71]。此外,实际育种中难以构建长期稳定的训练集,育种项目需不断引入新种质资源,这限制了模型中长期积累有效基因型/表型数据的可能性。3)基因型 $\times$ 环境互作($G \times E$)是GS研究中的热点问题,它要求模型同时关注遗传变异和环境因素对作物表型的影响,尤其是在多环境/多性状的背景下,需要构建能同时处理遗传相关性和性状协方差的复杂模型。大豆对光周期和温度极为敏感,跨纬度种植时表现出强烈的环境适应性差异,这使得 $G \times E$ 效应尤为突出。Montesinos等^[72]研究显示,在未考虑 $G \times E$ 时,深度学习方法在9个数据集中有6个优于GBLUP模型,而在引入 $G \times E$ 项后,GBLUP在其中8个数据集中表现更佳,这说明深度学习在无需假设前提的情况下能更好地捕捉复杂结构,但 $G \times E$ 仍需精细建模,尤其当传统统计模型与深度学习结构进行融合时,通过引入协方差矩阵,环境主效应等方式可以显著提升 $G \times E$ 互作的解释力,但却需要引入更多的参数以确保模型的准确性。4)农业大数据往来源多样、结构复杂,包括基因型矩阵、多时空表型数据、环境变量、图像信息等。现有GS模型在处理这

些异构信息时,缺乏统一的数据融合机制,且对噪声、缺失值和批次效应较为敏感。此外,表型数据获取成本高、精度不一,进一步加剧了预测模型的误差传播。

GS的发展应优先关注以下3个方面:1)多组学数据融合,基因组预测以DNA序列为基础,但基因表达需经历转录和翻译过程,整合转录组和蛋白组数据将有助于揭示基因调控机制,提高模型预测准确率。2)多性状联合预测,单一性状的预测效率有限,联合多个目标性状进行筛选更符合现实育种需求,机器学习算法可实现特征融合、模型共享的理想育种策略。3)构建统一的数据标准和预处理框架,提高多源数据的一致性和可用性,从而确保其在农业大数据背景下的准确性。

参考文献 References

- [1] Azadi H, Ho P, Hasfiati L. Agricultural land conversion drivers: A comparison between less developed, developing and developed countries[J]. *Land Degradation and Development*, 2011, 22(6): 596-604
- [2] van Meijl H, Havlik P, Lotze-Campen H, Stehfest E, Witzke P, Domínguez I P, Bodirsky B L, van Dijk M, Doelman J, Fellmann T, Humpenöder F, Koopman J F L, Müller C, Popp A, Tabeau A, Valin H, van Zeist W J. Comparing impacts of climate change and mitigation on global agriculture by 2050[J/OL]. *Environmental Research Letters*, (2018-06-05). DOI:doi.org/10.1088/1748-9326/aabdc4
- [3] Alemu A, Åstrand J, Montesinos-López O A, Isidro Y Sánchez J, Fernández-González J, Tadesse W, Vetukuri R R, Carlsson A S, Ceplitis A, Crossa J, Ortiz R, Chawade A. Genomic selection in plant breeding: Key factors shaping two decades of progress[J]. *Molecular Plant*, 2024, 17(4): 552-578
- [4] Nguyen V H, Morantte R I Z, Lopena V, Verdeprado H, Murori R, Ndayiragije A, Katiyar S K, Islam M R, Juma R U, Flandez-Galvez H, Glaszmann J C, Cobb J N, Bartholomé J. Multi-environment genomic selection in rice elite breeding lines[J/OL]. *Rice*, (2023-02-07). DOI:doi.org/10.1186/s12284-023-00623-6
- [5] McGowan M, Wang J B, Dong H X, Liu X L, Jia Y, Wang X F, Iwata H, Li Y T, Lipka A E, Zhang Z W. Ideas in genomic selection with the potential to transform plant molecular breeding: A review[M]. In: Goldman I, eds. *Biology and Life Sciences Volume 45*, Hoboken: John Wiley & Sons, Inc, 2022
- [6] Barreto C A V, das Graças Dias K O, de Sousa I C, Azevedo C F, Nascimento A C C, Guimarães L J M, Guimarães C T, Pastina M M, Nascimento M. Genomic prediction in multi-environment trials in maize using statistical and machine learning methods[J/OL]. *Scientific Reports*, (2024-01-11). DOI:doi.org/10.1038/s41598-024-51792-3
- [7] Bandillo N B, Jarquin D, Posadas L G, Lorenz A J, Graef G L. Genomic selection performs as effectively as phenotypic selection for increasing seed yield in soybean[J/OL]. *The Plant Genome*, (2022-09-29). DOI:doi.org/10.1002/tpg2.20285
- [8] Montesinos-López O A, Chavira-Flores M, Kiasiantini, Crespo-Herrera

- L, Saint Piere C, Li H H, Fritsche-Neto R, Al-Nowibet K, Montesinos-López A, Crossa J. A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding[J/OL]. *Genetics*, (2024-11-05). DOI: doi.org/10.1002/tpg2.20285
- [9] Varshney R K. Exciting journey of 10 years from genomes to fields and markets: Some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut[J]. *Plant Science*, 2016, 242: 98-107
- [10] Jubair S, Domaratzi M. Crop genomic selection with deep learning and environmental data: A survey[J/OL]. *Frontiers in Artificial Intelligence*, (2023-01-10), DOI:doi.org/10.3389/frai.2022.1040295
- [11] Duvick D N. The contribution of breeding to yield advances in maize (*Zea mays* L)[J]. *Advances in Agronomy*, 2005, 86: 83-145
- [12] 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
- [13] Meher P K, Rustgi S, Kumar A. Performance of Bayesian and BLUP alphabets for genomic prediction: Analysis, comparison and results[J]. *Heredity*, 2022, 128(6): 519-530
- [14] Avasthi P, Mets D G, York R. Harnessing genotype-phenotype nonlinearity to accelerate biological prediction[J/OL]. *Arcadia Science*, (2023-09-22). DOI:doi.org/10.57844/arcadia-5953-995f
- [15] Moeiniazade S, Hu G P, Wang L Z, Schnable P S. Optimizing selection and mating in genomic selection with a look-ahead approach: An operations research framework[J]. *Genes Genomes Genetics*, 2019, 9(7): 2123-2133
- [16] Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de Los Campos G, Burguño J, González-Camacho J M, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, C, Gowda M, Roorkiwal M, Rutkoski J, Varshney R K. Genomic selection in plant breeding: Methods, models, and perspectives[J]. *Trends in Plant Science*, 2017, 22(11): 961-975
- [17] Biswas M K, Bagchi M, Biswas D, Harikrishna J A, Liu Y X, Li C Y, Sheng O, Mayer C, Yi G J, Deng G M. Genome-wide novel genic microsatellite marker resource development and validation for genetic diversity and population structure analysis of banana[J/OL]. *Genes*, (2020-12-09). DOI: doi.org/10.3390/genes11121479
- [18] Collard B C Y, Jahufer M Z Z, Brouwer J B, Pang E C K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts[J]. *Euphytica*, 2005, 142(1-2): 169-196
- [19] Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits[J]. *Genetics*, 1990, 124(3): 743-756
- [20] Grenier C, Cao T V, Ospina Y, Quintero C, Châtel M H, Tohme J, Courtois B, Ahmadi N. Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding[J/OL]. *PLoS One*, (2015-08-27). DOI: doi.org/10.1371/journal.pone.0136594
- [21] 郭亮虎, 逯腊虎, 张婷, 史晓芳, 袁凯, 王镇, 杨三维. 植物全基因组选择育种研究进展与前景[J]. *山西农业科学*, 2015, 43(11): 1558-1562
- Guo L H, Lu L H, Zhang T, Shi X F, Yuan K, Wang Z, Yang S W. Research progress and prospects of genome-wide selection breeding in plants[J]. *Journal of Shanxi Agricultural Sciences*, 2015, 43(11): 1558-1562 (in Chinese)
- [22] Santantonio N, Atanda S A, Beyene Y, Varshney R K, Olsen M, Jones E, Roorkiwal M, Gowda M, Bharadwaj C, Gaur P M, Zhang X C, Dreher K, Ayala-Hernández C, Crossa J, Pérez-Rodríguez P, Rathore A, Gao S Y, McCouch S, Robbins K R. Strategies for effective use of genomic information in crop breeding programs serving Africa and south Asia[J/OL]. *Frontiers in Plant Science*, (2020-03-27). DOI: doi.org/10.3389/fpls.2020.00353
- [23] Dreisigacker S, Crossa J, Pérez-Rodríguez P, Montesinos-López O, Rosyara U, Juliana P, Mondal S, Crespo-Herrera L, Govindan V, Singh R, Braun H. Implementation of genomic selection in the CIMMYT global wheat program, findings from the past 10 years[J/OL]. *Crop Breeding, Genetics and Genomics*, (2021-03-26) DOI: doi.org/10.20900/cbpg20210005
- [24] 王欣, 徐一亿, 徐扬, 徐辰武. 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, Xu C W. Research progress in genomic selection breeding technology for crops[J]. *Biotechnology Bulletin*, 2024, 40(3): 1-13 (in Chinese)
- [25] Henderson C R. Comparison of alternative sire evaluation methods[J]. *Journal of Animal Science*, 1975, 41(3): 760-770
- [26] Bernardo R. Prediction of maize single-cross performance using RFLPs and information from related hybrids[J]. *Crop Science*, 1994, 34(1): 20-25
- [27] Zhang Z, Todhunter R J, Buckler E S, Van Vleck L D. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood[J]. *Journal of Animal Science*, 2007, 85(4): 881-885
- [28] VanRaden P M. Efficient methods to compute genomic predictions[J]. *Journal of Dairy Science*, 2008, 91(11): 4414-4423
- [29] Endelman J B. Ridge regression and other kernels for genomic selection with R package rrBLUP[J]. *Plant Genome*, 2011, 4(3): 250-255
- [30] Misztal I, Lourenco D, Legarra A. Current status of genomic evaluation[J/OL]. *Journal of Animal Science*, (2020-04-04). DOI:doi.org/10.1093/jas/skaa101
- [31] Legarra A, Christensen O F, Aguilar I, Misztal I. Single Step, a general approach for genomic selection[J]. *Livestock Science*, 2014, 166: 54-65
- [32] Wang J B, Zhou Z K, Zhang Z, Li H, Liu D, Zhang Q, Bradbury P J, Buckler E S, Zhang Z W. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits[J]. *Heredity*, 2018, 121(6): 648-662
- [33] Yin L L, Zhang H H, Zhou X, Yuan X H, Zhao S H, Li X Y, Liu X L. KAML: Improving genomic prediction accuracy of complex traits using machine learning determined parameters[J/OL]. *Genome Biology*, (2020-06-17). DOI:doi.org/10.1186/s13059-020-02052-w
- [34] Yin L L, Zhang H H, Tang Z S, Yin D, Fu Y H, Yuan X H, Li X Y, Liu X L, Zhao S H. HIBLUP: An integration of statistical models on the BLUP framework for efficient genetic evaluation using big genomic data[J]. *Nucleic Acids Research*, 2023, 51(8): 3501-3512
- [35] Lin Z, Hayes B J, Daetwyler H D. Genomic selection in crops, trees and forages: A review[J]. *Crop and Pasture Science*, 2014, 65(11): 1177-1191
- [36] Hadfield J D. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package [J/OL]. *Journal of Statistical Software*, (2010-02-02). DOI: doi.org/10.18637/jss.v033.i02
- [37] Gianola D, Okut H, Weigel K A, Rosa G J. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat[J/OL]. *BMC Genetics*, (2011-10-07). DOI: doi.org/10.1186/1471-2156-

- 12-87
- [38] Belamkar V, Guttieri M J, Hussain W, Jarquín D, El-Basyoni I, Poland J, Lorenz A J, Baenziger P S. Genomic selection in preliminary yield trials in a winter wheat breeding program[J]. *Genes Genomes Genetics*, 2018, 8(8): 2735-2747
- [39] Samuel A L. Some studies in machine learning using the game of checkers[J]. *IBM Journal of Research and Development*, 1959, 3(3): 210-229
- [40] Liu X G, Wang H W, Wang H, Guo Z F, Xu X J, Liu J C, Wang S H, Li W X, Zou C, Prasanna B M, Olsen M S, Huang C L, Xu Y B. Factors affecting genomic selection revealed by empirical evidence in maize[J]. *Crop Journal*, 2018, 6(4): 341-352
- [41] Werner C R, Gaynor R C, Gorjanc G, Hickey J M, Kox T, Abbadi A, Leckband G, Snowdon R J, Stahl A. How population structure impacts genomic selection accuracy in cross-validation: Implications for practical breeding[J/OL]. *Frontiers in Plant Science*, (2020-12-16). DOI: doi.org/10.3389/fpls.2020.592977
- [42] Morgante F, Huang W, Maltecca C, MacKay T F C. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals[J]. *Heredity*, 2018, 120(6): 500-514
- [43] Farooq M, van Dijk A D J, Nijveen H, Mansoor S, de Ridder D. Genomic prediction in plants: Opportunities for ensemble machine learning based approaches[J/OL]. *F1000Research*, (2022-07-18). DOI: doi.org/10.12688/f1000research.122437.2
- [44] Grinberg N F, Orhobor O I, King R D. An evaluation of machine-learning for predicting phenotype: Studies in yeast, rice, and wheat[J]. *Machine Learning*, 2020, 109(2): 251-277
- [45] Xiong H Z, Chen Y L, Pan Y B, Wang J S, Lu W G, Shi A N. A genome-wide association study and genomic prediction for phakopsora pachyrhizi resistance in soybean[J/OL]. *Frontiers in Plant Science*, (2023-05-29) DOI: doi.org/10.3389/fpls.2023.1179357
- [46] Jannink J L, Lorenz A J, Iwata H. Genomic selection in plant breeding: From theory to practice[J]. *Briefings in Functional Genomics*, 2010, 9(2): 166-177
- [47] Xi J N, Yu Z H. Editorial: Unsupervised learning models for unlabeled genomic, transcriptomic & proteomic data[J/OL]. *Frontiers in Genetics*, (2021-11-11). DOI:doi.org/10.3389/fgene.2021.781698.
- [48] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5-32
- [49] Friedman J H. Greedy function approximation: A gradient boosting machine [J]. *Annals of Statistics*, 2001, 29(5): 1189-1232
- [50] Li B, Zhang N X, Wang Y G, George A W, Reverter A, Li Y T. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods[J/OL]. *Frontiers in Genetics*, (2018-07-04). DOI:doi.org/10.3389/fgene.2018.00237
- [51] Yan J, Xu Y T, Cheng Q, Jiang S Q, Wang Q, Xiao Y J, Ma C, Yan J B, Wang X F. LightGBM: Accelerated genomically designed crop breeding through ensemble learning[J/OL]. *Genome Biology*, (2021-09-20). DOI: doi.org/10.1186/s13059-021-02492-y
- [52] Morota G, Gianola D. Kernel-based whole-genome prediction of complex traits: A review[J/OL]. *Frontiers in Genetics*, (2014-10-16). DOI: doi.org/10.3389/fgene.2014.00363
- [53] Liang M, Chang T P, An B X, Duan X H, Du L L, Wang X Q, Miao J, Xu L Y, Gao X, Zhang L P, Li J Y, Gao H J. A stacking ensemble learning framework for genomic prediction[J/OL]. *Frontiers in Genetics*, (2021-03-04). DOI:doi.org/10.3389/fgene.2021.600040
- [54] Wang K L, Ali Abid M, Rasheed A, Crossa J, Hearne S, Li H H. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants[J/OL]. *Molecular Plant*, 2023, 16(1): 279-293
- [55] Wang P P, Lehti-Shiu M D, Lotreck S, Segura Abú K, Krysan P J, Shiu S H. Prediction of plant complex traits via integration of multi-omics data[J/OL]. *Nature Communications*, (2024-06-18). DOI: 10.1038/s41467-024-50701-6
- [56] Wang Z S, Li W C, Tang Z L. Enhancing the genomic prediction accuracy of swine agricultural economic traits using an expanded one-hot encoding in CNN models[J]. *Journal of Integrative Agriculture*, 2025, 24(9): 3574-3582
- [57] Pérez-Rodríguez P, Gianola D, González-Camacho J M, Crossa J, Manès Y, Dreisigacker S. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat [J]. *Genes Genomes Genetics*, 2012, 2(12): 1595-1605
- [58] González-Camacho J M, Crossa J, Pérez-Rodríguez P, Ornella L, Gianola D. Genome-enabled prediction using probabilistic neural network classifiers [J/OL]. *BMC Genomics*, (2024-04-09). DOI:doi.org/10.1186/s12864-016-2553-1
- [59] Ma W L, Qiu Z X, Song J, Cheng Q, Ma C. DeepGS: Predicting phenotypes from genotypes using deep learning[J/OL]. *bioRxiv*, (2017-12-17). DOI: doi.org/10.1101/241414
- [60] Liu Y, Wang D L, He F, Wang J X, Joshi T, Xu D. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean[J/OL]. *Frontiers in Genetics*, (2019-11-22). DOI: 10.3389/fgene.2019.01091
- [61] Azodi C B, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu S H. Benchmarking parametric and machine learning models for genomic prediction of complex traits[J]. *Genes Genomes Genetics*, 2019, 9(11): 3691-3702
- [62] Morales N, Ogbonna A C, Ellerbrock B J. Breedbase: A digital ecosystem for modern plant breeding[J/OL]. *Genes Genomes Genetics*, (2022-04-06). DOI: doi.org/10.1093/g3journal/jkac078
- [63] Gao Y J, Zhou Q, Luo J X, Xia C, Zhang Y H, Yue Z Y. Crop-GPA: An integrated platform of crop gene-phenotype associations[J/OL]. *npj Systems Biology and Applications*, (2024-02-12). DOI: doi.org/10.1038/s41540-024-00343-7
- [64] Feng W J, Gao P F, Wang X T. AI breeder: Genomic predictions for crop breeding [J]. *New Crops*, (2023-12-14). DOI: doi.org/10.1016/j.ncrops.2023.12.005
- [65] Gao P F, Zhao H N, Luo Z, Lin Y F, Feng W J, Li Y L, Kong F J, Li X, Fang C, Wang X T. SoyDNGP: A web-accessible deep learning framework for genomic prediction in soybean breeding [J/OL]. *Briefings in Bioinformatics*, (2023-11-01). DOI:doi.org/10.1093/bib/bbad349
- [66] Li T, Jiang S, Fu R, Wang X F, Cheng Q, Jiang S Q. IP4GS: Bringing genomic selection analysis to breeders[J/OL]. *Frontiers in Plant Science*, (2023-03-06). DOI:doi.org/10.3389/fpls.2023.1131493
- [67] Chen J X, Tan C, Zhu M, Zhang C Y, Wang Z H, Ni X M, Liu Y L, Wei T, Wei X F, Fang X D, Xu Y, Huang X H, Qiu J, Liu H. CropGS-Hub: A comprehensive database of genotype and phenotype resources for genomic

- prediction in major crops[J]. *Nucleic Acids Research*, 2024, 52(1): 1519-1529
- [68] Zeng S, Mao Z T, Ren Y J, Wang D L, Xu D, Joshi T. G2PDeep: A web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers[J]. *Nucleic Acids Research*, 2021, 49(1): 228-236
- [69] Li H H, Li X, Zhang P, Feng Y W, Mi J R, Gao S, Sheng L L, Ali M, Yang Z K, Li L, Fang W, Wang W S, Qian Q, Gu F, Zhou W B. Smart Breeding Platform: A web-based tool for high-throughput population genetics, phenomics, and genomic selection[J/OL]. *Molecular Plant*, (2024-03-05). DOI:doi.org/10.1016/j.molp.2024.03.002
- [70] Montesinos-López O A, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez C M, Montesinos-López A, Juliana P, Singh R. New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes[J]. *Genes Genomes Genetics*, 2019, 9(5): 1545-1556
- [71] Montesinos-López O, A, Montesinos-López A, Pérez-Rodríguez P, Barrón-López J A, Martini J W R, Fajardo-Flores S B, Gaytan-Lugo L S, Santana-Mancilla P C, Crossa J. A review of deep learning applications for genomic selection[J/OL]. *BMC Genomics*, (2021-01-06). DOI:doi.org/10.1186/s12864-020-07319-x
- [72] Montesinos-López A, Montesinos-López O A, Gianola D, Crossa J, Hernández-Suárez C M. Multi-environment genomic prediction of plant traits using deep learners with dense architecture[J/OL]. *Genes Genomes Genetics*, 2018, 8(12): 3813-3828

责任编辑: 吕晓梅



第一作者简介: 孙连军, 博士, 中国农业大学农学院教授, 博士生导师。主要研究方向: 大豆种子发育和产量形成的遗传基础, 开展优异种质资源挖掘和新品种创制及重要农艺性状形成机制解析。现担任中国农业机械化协会智慧农业发展分会委员, 主持国家重点研发计划课题、国家自然科学基金面上项目、部委战略咨询项目等6项。在 *Nature Genetics*、*Nature Plant*、*PNAS* 等学术期刊发表论文30余篇, 专利授权9项, 参编英文著作1部。