本草基因编码天然多样性成分库 (GNDC) 在中医药研究中的应用

余志银[1,2], 陈 伟[1,2], 冷 梁[1,2], 孙 丹[1,2], 刘 灏[1,2], 宫瑞泽[1,2], 丛朝彤[1,2], 刘成成[1,2], 杨秀萍[3], 宾华超[1,2], 鲁 军[3], 张三印[1,2*], 宋 驰[1,2*]

(1. 成都中医药大学本草基因组学研究院, 四川 成都 611137; 2. 成都中医药大学中医药创新研究院, 四川 成都 611137; 3. 成都中医药大学药学院, 四川 成都 611137)

摘要: 天然成分作为物种适应环境与共进化的产物, 是药物发现的重要源泉。根据中心法则的遗传规律, 将天然成分分为基因直接编码的成分 (核酸和多肽) 和基因间接编码的成分 (初级代谢产物和次生代谢产物)。作者利用多组学和人工智能技术, 对全球八大权威药典中收载的 1 037 个物种的核基因组及细胞器基因组进行解析和注释, 并整合多维数据资源构建了全球首个亿量级本草基因编码的天然多样性成分库 (gene-encoded natural diversity components repository, GNDC, https://cbcb.cdutcm.edu.cn/gndc/)。本论文将系统介绍: ① GNDC 成分库的构建方法, 包括数据整合策略、注释流程及数据处理; ② 核心功能与数据规模: GNDC 目前收录超过 2.34 亿个基因直接或间接编码的天然成分 [包含四大专业子库: HerbalMDB (次级代谢产物): 232 万个小分子化合物; HerbalPDB (小肽): 2.29 亿条小肽; HerbalRDB (小 RNA): 238 万条小 RNA; HerbalCDB (碳水化合物): 26 万个碳水化合物]; ③ 探讨 GNDC 在中医药现代化研究中的应用。GNDC 提供了亿量级的天然成分, 将为药物发现提供前所未有的广阔“化学空间”, 有力驱动药物研发模式从“经验导向”向“大数据驱动”的转变, 为中医药现代化研究带来新的范式。

关键词: GNDC; 本草基因组学; 天然成分; 人工智能; 传统医药

中图分类号: R281 **文献标识码**: A **文章编号**: 0513-4870(2025)11-3408-10

Application of the gene-encoded natural diversity components repository (GNDC) in traditional Chinese medicine research

YU Zhi-yin[1,2], CHEN Wei[1,2], LENG Liang[1,2], SUN Dan[1,2], LIU Hao[1,2], GONG Rui-ze[1,2], CONG Zhao-tong[1,2], LIU Cheng-cheng[1,2], YANG Xiu-ping[3], BIN Hua-chao[1,2], LU Jun[3], ZHANG San-yin[1,2*], SONG Chi[1,2*]

(*1. Institute of Herbgenomics, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; 2. Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; 3. School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China*)

Abstract: Natural components are crucial sources for drug discovery. Based on the central dogma of molecular biology, we proposed a novel paradigm for classifying natural products by categorizing natural components into "direct gene-encoded components" (including nucleic acids and peptides) and "indirect gene-encoded components" (encompassing primary metabolites and secondary metabolites). Utilizing multi-omics and artificial intelligence technologies, we systematically analyzed the nuclear and organellar genomes of 1 037 medicinal species sourced from eight authoritative global pharmacopeias, integrating multidimensional data resources to establish the gene-encoded natural diversity components repository (GNDC). This paper comprehensively describes: (1) The

收稿日期: 2025-07-02; 修回日期: 2025-08-03.

基金项目: 四川省重大科技专项 (2024ZDZX0019).

*通讯作者 Tel: 86-28-82682984, E-mail: tcmzsy@cdutcm.edu.cn; songchi@cdutcm.edu.cn

DOI: 10.16438/j.0513-4870.2025-0749

construction methodology of GNDC, including data integration strategies, standardized annotation pipelines, and rigorous quality control; (2) Core features and data scale: GNDC is currently the world's largest repository of medicinal natural products, housing over 234 million gene-encoded (directly or indirectly) natural components. It encompassed four specialized sub-databases: HerbalMDB for 2.32 million secondary metabolites, HerbalPDB for 229 million small peptides, HerbalRDB for 2.38 million small RNAs, and HerbalCDB for 0.26 million carbohydrates; (3) Exploring the application prospects of GNDC in the modernization of traditional medicine research. GNDC will provide an unprecedented expansive "chemical space" for drug discovery. It will powerfully drive a paradigm shift in drug development from "experience-oriented" to "big data-driven" approaches, offering a transformative framework for traditional medicine research.

Key words: gene-encoded natural diversity components repository; herbgenomics; natural component; artificial intelligence; traditional medicine

1 前言

天然产物是生物体在进化过程中形成的产物, 是其适应环境、参与竞争及抵御外界威胁的重要媒介[1,2]。由于其显著的生物活性, 在人类医药史上始终占据不可替代的地位[3,4]。近年来, 核酸和多肽药物展现出了巨大的开发潜力, 核酸药物如反义寡核苷酸、小干扰RNA等, 能够精准地调控基因表达, 为治疗遗传性疾病、肿瘤等提供了全新的策略[5-7]。此外, 植物来源的miRNA可被哺乳动物吸收并调节基因表达, 为跨界调控开辟了新途径[8,9]。多肽药物因其独特的生物活性和相对较低的毒性, 在糖尿病、心血管疾病等治疗领域取得了显著进展[10]。然而, 传统的天然产物分类主要依据化学结构信息, 且多聚焦于次级代谢产物[11], 往往忽略了基因编码的分子 (如核酸和肽)。此外, 现有天然成分库的结构多样性受限于分离技术的选择性偏好, 多肽、核酸类成分常常被忽略。这使得重新认识天然产物的多样性, 构建覆盖生源合成全路径的动态化、系统性天然产物库已成为创新药物研发的迫切需求。

天然成分的合成演化过程遵循中心法则的遗传规律, 通过表达特定的基因, 生物体能够合成一系列结构多样的生物活性成分。基于此, 作者提出了基于中心法则的天然成分分类方法, 将天然成分分为基因直接编码的成分和基因间接编码的成分 (图1A)[12]。基因直接编码的成分是通过基因的转录和翻译直接指导合成的成分, 其结构由DNA序列决定, 这些分子的合成过程不涉及复杂的代谢途径, 主要包括核酸和多肽; 基因间接编码的成分是由基因编码的酶催化合成的化合物, 其生物合成过程涉及复杂的代谢途径, 通常包括初级代谢产物和次级代谢产物, 且可进一步基于生源合成途径进行分类, 如乙酸途径、莽草酸途径、甲羟戊酸途径 (MVA)、甲基赤藓醇磷酸途径 (MEP) 和氨基酸途径[13]。这一分类体系突破了传统化学结构分类的局限性, 将基因信息与代谢产物的化学多样性直接关联, 为系统性探究天然成分的生物功能, 遗传进化和生源合成路径提供了依据。

目前已有一些天然产物的数据库被报道[14], 例如次级代谢产物数据库COCONUT 2.0[15], 收录了近

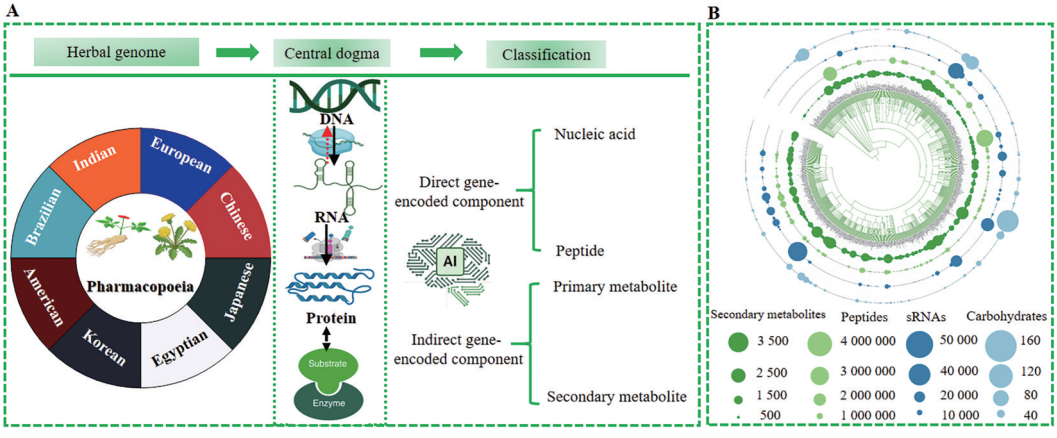


Figure 1 Classification of natural components based on the central dogma (A); phylogenetic tree of species in the eight pharmacopoeias and distribution of four types of components identified in each species (B). Some images were created using BioRender website. AI: Artificial intelligence

70万个成分, 但数据分类注释信息不足, 规模有限。此外, 植物小肽、小RNA和碳水化合物数据库也有报道, 但都聚焦于专一特征成分 (如NCBbook[16]、Bencao sRNA Atlas[17]和GlyTouCan[18]), 数据规模较小。传统药用植物蕴含大量活性成分, 但针对药用物种的专用数据库仍稀缺, 严重制约了其治疗潜力的挖掘。因此, 亟需构建一个整合多类天然成分的综合性数据库, 以支持特定物种分子多样性的系统性研究。

全球八部权威药典 (巴西、埃及、欧洲、印度、日本、韩国、中国及美国药典) 收录的药用植物具有明确疗效记载, 为此类研究提供了重要基础。近年来, 随着基因组学技术的迅猛发展, 大量药用植物基因组被解析, 为认识和挖掘天然成分的多样性奠定了基础, 并为基于基因组的药用物种代谢产物的挖掘提供了契机[19]。此外, 人工智能与多组学技术的深度融合, 正推动天然药物研究迈向“数据驱动”新范式, 为天然成分研究带来了革命性进展。如高通量基因组测序、代谢组学及机器学习能够实现天然成分的大规模鉴定与表征。基于深度学习的生物合成途径预测[20], 如BioNavi-NP [21], 可以实现部分天然成分的生物合成途径预测。通过构建结构多样性的成分库, 结合人工智能 (AI) 算法 (如深度神经网络、图卷积网络), 可高效评估化合物与靶点的结合亲和力及药代动力学 (ADMET) 特性, 并赋能百万级化合物的虚拟筛选[22]。基于此, 团队前期构建了全球药典基因组数据库[23] (GPGD)、传统药用植物基因组数据库[24] (TCMPG) 及千种本草基因组数据库[25] (1K-MPGD), 并成立了千种本草基因组计划联盟, 为天然成分的基因组学水平研究奠定了基础。

通过整合多组学技术和人工智能算法, 以中心法则为框架, 作者对全球八大权威药典中收载的1 037个物种的核基因组及细胞器基因组进行解析和注释, 并整合多维数据资源构建了全球首个亿量级草药基因编码的天然多样性成分库 (Gene-encoded natural diversity components repository, GNDC, https://cbcb.cdutcm.edu.cn/gndc/)[12]。当前, GNDC收录了超过2.34亿个基因直接或间接编码的天然成分, 包括四个专业的子数据库: HerbalMDB收录了232万个次级代谢产物, HerbalPDB收录了2.29亿个小肽, HerbalRDB收录了238万个小RNA, HerbalCDB收录了26万个碳水化合物。该数据库覆盖了目前发现的天然成分的类型, 包括核酸、小肽、次生代谢产物和碳水化合物, 并整合了多个成分功能预测分析模块。本论文将系统介绍GNDC成分库的构建方法、核心功能与数据规模, 重点探讨GNDC在中医药现代化研究中的应用前景, 如中药活性成分的快速发现与溯源、作用靶点预测、中药复方作用机制解析、中药质量评价与控制等。

2 构建方法

GNDC由四个子数据库组成, 包括HerbalMDB、HerbalPDB、HerbalRDB和HerbalCDB, 构建方法如下:

2.1 HerbalMDB 收录232万个次级代谢产物, 整合了来自多个数据库资源和八部全球药典物种的次级代谢产物信息。对每个获取的次级代谢产物进行了结构注释和分类, 还包含生物合成前体分析 (BioNavi-NP) [21]; 基于SMILES去除重复分子, 利用RDKit (v.2017.09.3) 对数据进行标准化处理和生成Morgan指纹, 并利用Tanimoto系数评估次级代谢产物之间的结构相似性; 使用NPClassifier[26]对化合物进行分类; 利用ADMET-AI[27]对化合物的吸收、分布、代谢、排泄和毒性等特性进行评估; 使用自主研发的基因表达扰动分析模型 (DeepICER) 评估次级代谢产物对细胞基因表达的扰动影响。所有注释流程的代码已公开发布在GitHub上 (https://github.com/GreatChenLab/GNDC/tree/main/SMDB)。

2.2 HerbalPDB 整合了来自1K-MPGD (http://www.herbgenome.com)、TCMPG (https://cbcb.cdutcm.edu.cn/TCMPG/) 和GPGD (http://www.gpgenome.com/) 三大数据库的基因组数据, 系统鉴定了2.297 7亿个小肽, 并对其进行了全面的功能注释与分析。具体方法为: 基因组数据使用EMBOSS (http://emboss.open-bio.org) 和ORFIPY (https://github.com/urmi-21/orfipy) 进行全基因组六框翻译, 生成所有可能的蛋白质编码序列, 并筛选出长度在5到75个氨基酸之间的序列, 进一步通过转录组数据和质谱数据进行验证; 采用HISAT2 (https://daehwankimlab.github.io/hisat2/) 将NCBI数据库中的转录组数据比对到对应的参考基因组上, 进一步提高预测的准确性。转录组数据使用StringTie (https://github.com/gpertea/stringtie) 进行处理, 利用TransDecoder (https://github.com/TransDecoder/TransDecoder) 预测开放阅读框 (ORFs), 获得潜在的编码序列; 使用mmseq2 (https://github.com/soedinglab/MMseqs2, release 17-b804f) 将转录组数据注释的小肽与基因组解析的小肽进行比对, 筛选出转录组支持的高置信度候选小肽, 进一步结合质谱数据, 提高预测小肽的可信度。利用NCBI的BLASTp (https://blast.ncbi.nlm.nih.gov/BlastAlign.cgi) 将实验验证的小肽序列与基因组及转录组预测数据进行比对, 保留具有显著匹配的肽段作为高置信度候选小肽。所有小肽注释流程的代码已公开发布在GitHub上 (https://github.com/GreatChenLab/GNDC/tree/main/PDB)。

小肽功能注释: 使用HMMSCAN (https://www.ebi.ac.uk/Tools/hmmer/search/ hmmscan) 对小肽KEGG通路进行注释, 识别与小肽相关的代谢及信号通路。通过BLASTp比对UniProt数据库, 结合序列同源性, 揭示其潜在生物学功能与分子互作机制。小肽理化性质评估使用Bio.SeqUtils (https://biopython.org/docs/1.75/api/Bio.SeqUtils.html) 中的ProtParam模块对小肽的理化性质 (如pH值、旋光度和等电点) 进行预测。除基础理化特征外, 还利用iAMPCN (https://github.com/joy50706/iAMPCN) 预测抗菌活性, 通过PepNet[28]评估抗炎活性, 并运用DeepB3P (https://github.com/GreatChenLab/DeepB3P) 预测血脑屏障通透性, 实现生物活性的多维功能评估。

2.3 HerbalRDB 收录了238万个小RNA (包括miRNA、tRNA、rRNA和snRNA)。RNA解析使用Rfam数据库 (v.14.10) 和INFERNAL软件 (v.1.1.4) 进行ncRNA预测。其中高置信度的预测结果, 进一步提取snRNA和miRNA。Rfam分析参数设置为: -cut_ga-rfam-nohmmonly-fmt 2; INFERNAL执行命令为: cmscan-cut_ga-nohmmonly-cpu16--tbloutout. tblout-fmt2-claninRfam.clanin-rfamRfam.cmgenome.fa。tRNA序列特征分析采用tRNAscan-SE (v.2.0.12), 参数: -E-jtRNA.gff-otRNA.result-ftRNA.struct。除此之外, 核糖体RNA (rRNA) 预测使用rRNAmmer (v.1.2), 参数设置: -S euk -m tsu,lsu,ssu以实现真核生物rRNA (5S/28S/18S) 的系统鉴定。

miRNA功能注释: 采用生物信息学分析流程对miRNA及其靶基因进行预测。首先从NCBI收集miRNA-seq数据, 并统一格式, 使用软件tBtools_JRE1.6.jar biocjava.sRNA.Tools.sRNAseqAdaperRemover进行接头序列去除, 随后通过对相同序列的读长进行合并和去冗余 (参数设置: --minLen 17)。基于基因组GFF文件注释信息, 利用gffread (v.0.12.7) 从基因组中提取相关序列。采用mmseq2软件 (release 17-b804f) (https://github.com/soedinglab/MMseqs2) 对miRNA-seq分析的前体miRNAs与基因组预测进行比对 (参数设置为"--search-type 3--threads 16"), 验证预测的前体miRNA的可靠性。通过RNAhybrid (v.2.1.2) 分析miRNA序列与靶标mRNA序列的互补配对, 以识别潜在的结合位点。同时, 通过计算miRNA与靶标的互作自由能, 评估靶基因的二级结构和稳定性, 优先选择具有高结合能、稳定性高和生物学相关性高的靶标进行分析。miRNA靶基因的功能注释: 对于编码RNA的靶标, 从HGNC数据库 (https://www.genenames.org/) 获取功能注释、基因特异性信息和GO注释; 对于非编码RNA靶标, 通过NCBI数据库获取基因信息与调控功能等信息。Seqkit软件用于识别和去除同一物种中输出的重复sRNA条目, 参数设置为"rmdup-s"。所有注释流程的代码已公开发布在GitHub上 (https://github.com/GreatChenLab/ GNDC/tree/main/RDB)。

2.4 HerbalCDB 收录了超过26万个碳水化合物, 数据主要来自GlyTouCan (https://glycosmos.org/download) 和CSDB (https://csdb.glycoscience.ru/database/) 数据库中的碳水化合物。通过PubChemPy (v.1.0.4) 工具, 依据GlyTouCan提供的PubChem CID号获取SMILES、InChIKey、分子式和分子量等化学信息。此外, 根据CSDB数据库提供的SMILES式, 利用PubChemPy (v.1.0.4) 补充对应的InChIKey、PubChem CID等信息。通过比对SMILES式和物种分类信息去除重复的碳水化合物。基于多糖的IUPAC命名, 利用GlycoWork (v.1.5.0) 与GlycoDraw工具构建糖的Glycan结构, 确保符合聚糖符号命名法 (SNFG) 规范。所有注释流程的代码已公开发布在GitHub上 (https://github.com/GreatChenLab/GNDC/tree/main/CDB)。

3 GNDC的组成和功能

3.1 GNDC组成 GNDC以全球八部药典收载的1 037个物种为基础, 依据中心法则对天然成分进行分类, 系统注释了药用物种的小RNA、小肽、次级代谢产物及碳水化合物。同时, 基于NCBI物种分类数据库构建了系统发育树, 实现了四类成分分布的可视化 (图1B), 揭示了天然多样性成分在生物分类谱系中的分布特征, 为理解其生物多样性提供了重要的视角和依据。基因编码天然多样性成分库包含四大特色数据库 (HerbalMDB、HerbalPDB、HerbalRDB和HerbalCDB), 收录了超过2.34亿个天然多样性成分, 并且具备多维信息注释体系, 为中医药研究提供了丰富的资源和工具。

HerbalMDB (本草代谢数据库) 收录了2 321 763个次级代谢产物, 其生物合成途径包含乙酸–丙二酸途径、甲戊二羟酸途径、甲基赤藓醇磷酸途径、莽草酸途径、氨基酸途径和复合途径[13]。该数据库提供生物合成前体分析、结构相似性分析、化合物分类、ADMET性质评估及基因表达谱预测等功能模块, 整合了基于这些功能分析的次级代谢产物信息。HerbalPDB (本草小肽数据库): 基于GPGD[23]、TCMPG[24]和1K-MPGD[25]三大数据库的基因组资源, 结合转录组和肽组学数据, 系统鉴定了229 772 278个小肽。除提供理化性质外, 数据库还涵盖抗菌活性、血脑屏障穿透性及抗炎等生物活性预测信息。HerbalRDB (本草小RNA数据库): 收录了2 382 463个小RNA, 涵盖miRNA、tRNA、rRNA

和 snRNA。该数据库整合了物种信息、GC 含量分析、miRNA 靶基因预测等功能模块。HerbalCDB (本草碳水化合物数据库) 存储了 264 564 个碳水化合物, 包括多糖、寡糖和单糖, 并提供物种来源、聚糖符号命名法 (SNFG) 等注释信息。

3.2 GNDC 功能介绍 GNDC 是一个综合性的数据库平台, 整合了四个专业子数据库: HerbalMDB、HerbalPDB、HerbalRDB 和 HerbalCDB, 为研究人员提供了丰富的天然产物数据和分析工具 (图 2)。HerbalMDB 专注于次级代谢产物, 提供数据浏览、查询和分析功能, 支持基于结构信息、理化性质、物种信息和分类信息 (包括生源合成路径) 的定制化搜索, 并整合了 NPClassifier、ADMET-AI 和 DeepICER 等工具, 用于结构注释和生物活性预测 (图 3)。HerbalPDB 收录了小肽信息, 支持基于物种名称、KEGG ID 和 UniProt ID 的高级搜索, 并提供小肽的基本信息 (如序列信息、物种信息、长度) 和功能注释 (如抗菌、抗炎活性、物理化学性质), 见图 4。HerbalRDB 专注于小 RNA, 提供数据浏览、查询和可视化功能, 支持用户通过物种名称和靶基因 ID 进行查询 (图 5), 并可链接到 HGNC 和 NCBI 数据库获取详细注释。此外, HerbalRDB 还预测了 RNA 的发夹结构。HerbalCDB 是一个碳水化合物数据库, 支持基于物种信息、IUPAC 名称、SMILES、CID、分子式和 InChIKey 等多种搜索方式, 并整合了 Glycowork、VirtualGlycome、RINGS、ISOGlyP、GRITS Toolbox 和 NMRDB 等分析工具, 用于糖的结构分析和 NMR 谱图预测 (图 6)。GNDC 丰富的天然成分和功能分析工具为天然产物的研究提供了强大的数据支持和分析能力, 助力药物发现和生物医学研究。

4 GNDC 在中医药研究中的应用

本草基因编码的天然多样性成分库基于全球八大权威药典收录的 1 037 个药用物种的多维数据, 运用自

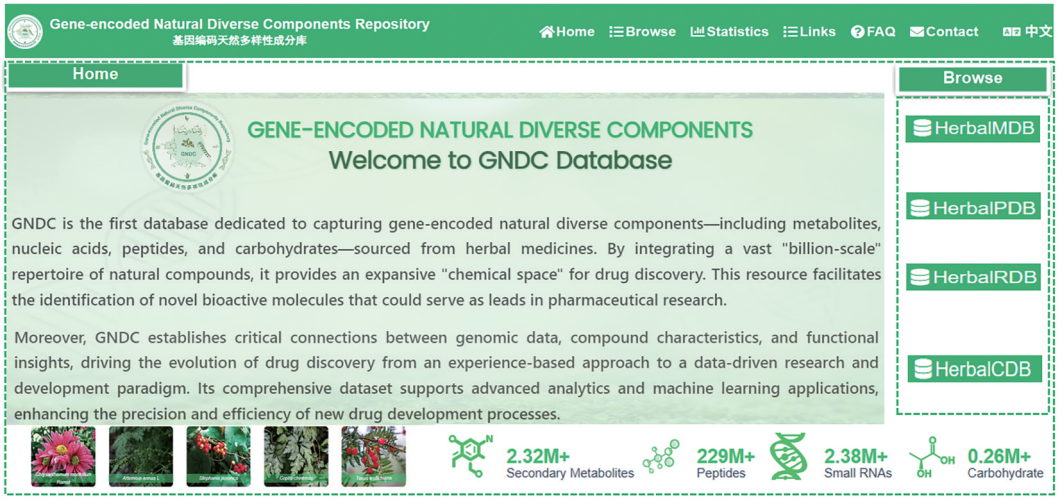


Figure 2 Screenshots of the web interfaces of gene-encoded natural diversity components repository (GNDC)

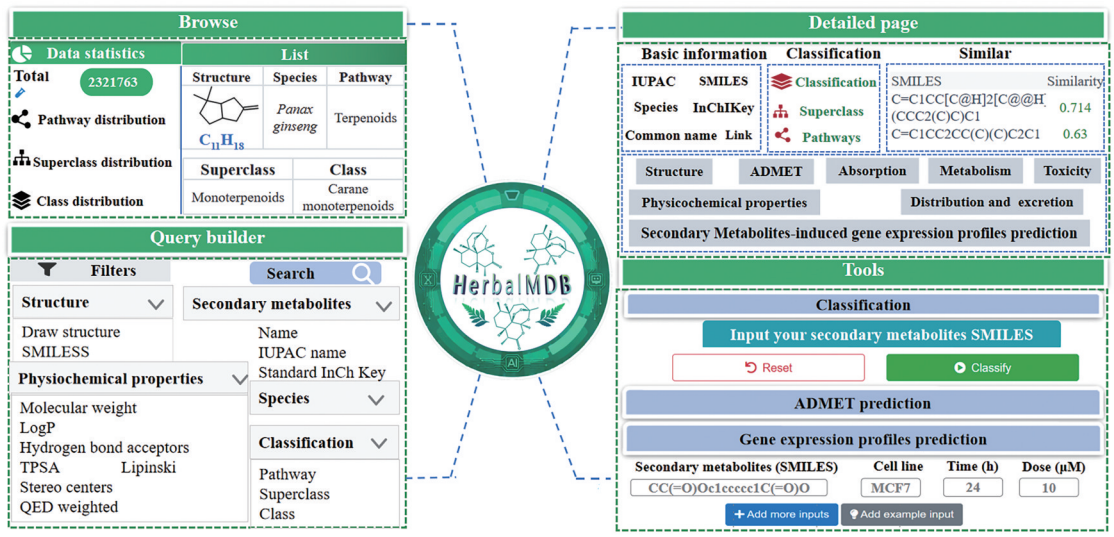


Figure 3 Structure and functionalities of the HerbalMDB database

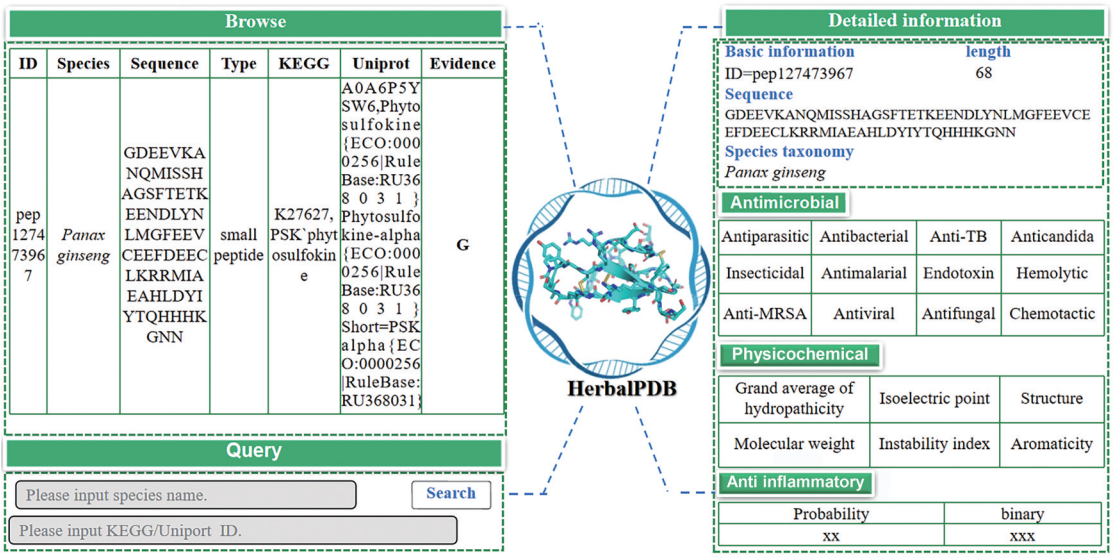


Figure 4 Structure and functionalities of the HerbalPDB database

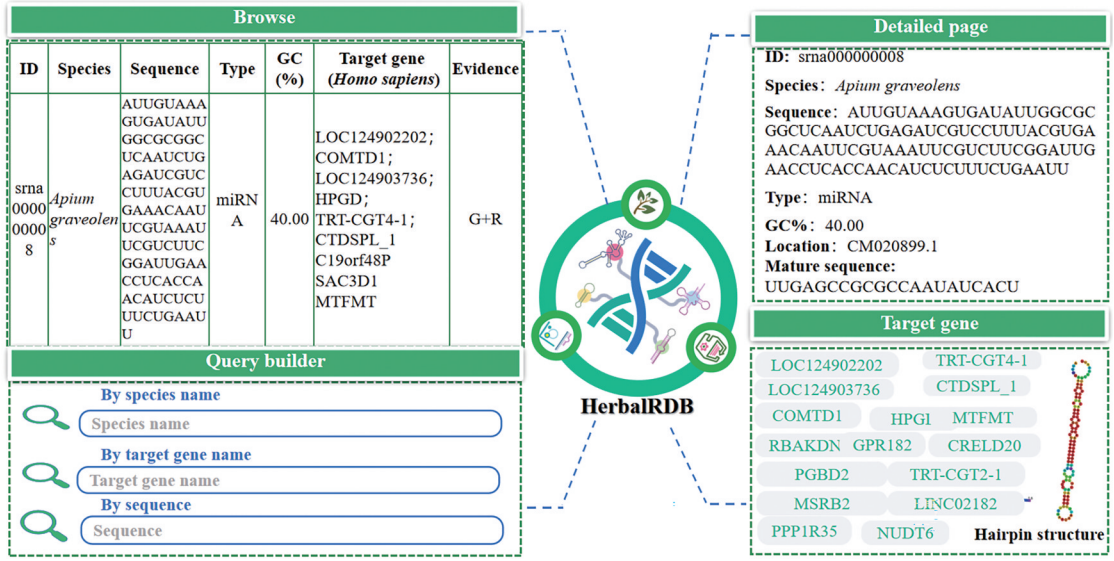


Figure 5 Structure and functionalities of the HerbalRDB database

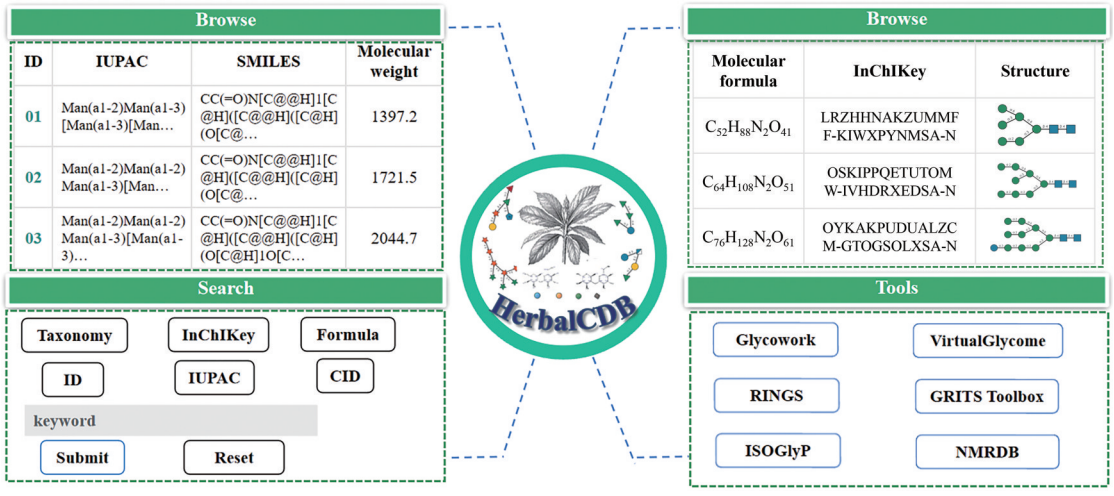


Figure 6 Structure and functionalities of the HerbalCDB database

主开发的算法结合多组学技术 (转录组、小RNA测序、肽质谱等) 挖掘天然多样性成分, 并整合人工智能算法 (自然语言处理、对抗生成网络、图卷积网络等) 实现成分识别、分类与功能注释。通过将多组学数据与AI技术深度融合, GNDC将为新药研发、合成生物学以及中药现代化研究等领域提供强有力的工具和平台。推动天然药物研究从传统经验模式向数据驱动模式转型。以下将详细探讨GNDC在中药新成分和新药发现、新靶点发现、中药质量评价与控制、大数据驱动的AI模型开发及中药药效物质的合成生物学等领域的应用潜力 (图7)。

4.1 中药新成分和新药发现 随着基因组学技术的迅猛发展, 研究人员可以深入剖析药用植物基因组, 获取丰富的生物信息[29], 尤其是GNDC基于基因组解析的小肽和小RNA研究模式, 将成为中药活性成分深入解析的范式。GNDC基于药用物种的基因组、转录组和代谢组等多组学数据, 解析获得了亿量级天然成分, 远远超过了传统天然成分发现的数量和种类。例如GNDC解析的小肽, 新增加的小肽数目接近98%是未报道的, 这些新增的独特成分也反映了GNDC与现有数据库相比的显著优势和价值。

庞大的天然成分与AI驱动的分析工具正在重塑新药发现的范式, 传统的药物发现过程通常依赖于经验性分离与活性测试, 这一过程不仅耗时耗力, 而且成功率较低, 常常导致资源的浪费与成本的增加。利用GNDC提供的亿量级天然成分, 研究人员可以借助虚拟筛选技术快速识别潜在的药物候选分子, 将极大提高药物筛选的效率。中药miRNA被发现能够参与跨界调控, 发挥治疗效果[17,30], 为代谢疾病和癌症等复杂疾病提供了新的治疗策略。基于中药物种的基因组数据, 可以从每个物种中解析出上万个小肽和miRNA, 这些将为中药新药开发提供创新分子。HerbalRDB收录了238万个小RNA, 并预测了miRNA的靶基因, 为中药核酸药物的原始创新提供了重要基础。此外, GNDC的HerbalMDB数据库集成了ADMET-AI预测模型, 能够评估化合物在药物代谢中的重要特性, 如化合物的吸收、分布、代谢、排泄和毒性。通过这些预测分析, 研究人员可以更好地了解中药药效成分的药物代谢动力学性质, 包括口服生物利用度、血脑屏障渗透率和CYP450酶抑制潜力等, 快速锁定低毒性、高生物利用度的分子, 从而高效筛选具有成药潜力的中药分子。

4.2 中药新靶点发现 GNDC整合了多个药物功能评价模型和亿量级的中药成分, 在中药新靶点发现方面的优势同样显著。通过结合海量的中药成分与药物靶点 (如膜蛋白受体和离子通道受体) 的AI虚拟筛选,

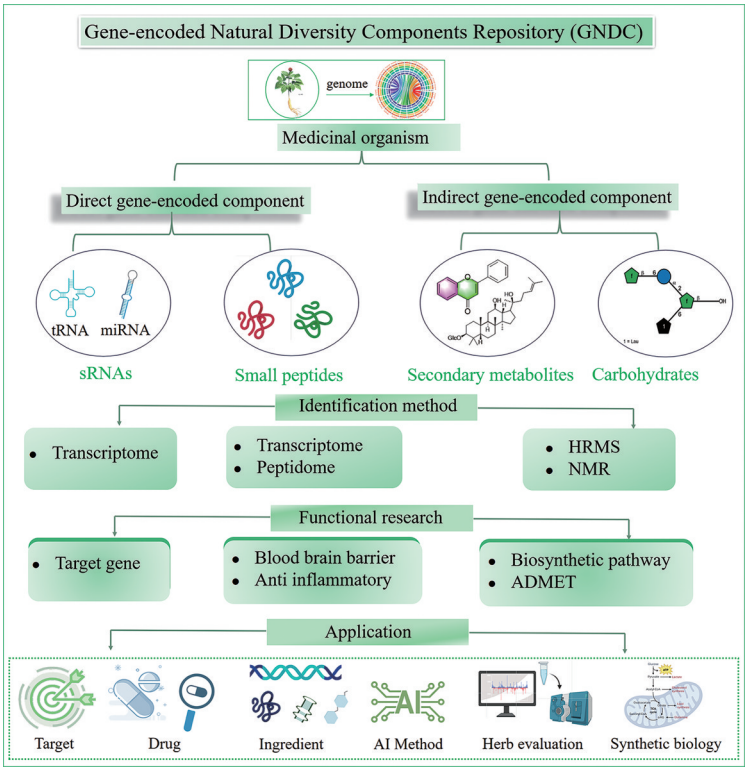


Figure 7 Schematic application of GNDC. Some images were created using BioRender website. HRMS: High-resolution mass spectrometry; NMR: Nuclear magnetic resonance

GNDC将突破传统药物靶点发现的瓶颈。例如GNDC整合了基因表达扰动分析模型—DeepICER, 该模型可评估次级代谢产物对细胞基因表达的扰动影响, 辅助发现新的药物靶点。此外, G蛋白偶联受体 (GPCR)是约40%临床药物的靶点[31], 前期团队已构建了人类泛基因组GPCR平台[32-35], 结合GNDC亿量级天然成分, 进行多维度的AI筛选, 加速中药靶标和新靶点发现。GNDC数据库中丰富的结构与功能信息的整合, 未来有望在孤儿受体、离子通道以及跨物种基因调控通路等领域的药物靶点发现中发挥重要作用。

4.3 中药质量评价 GNDC在传统医药的质量控制和评价方面也展现出巨大潜力。传统的药物质量控制主要集中在次生代谢产物, 难以全面反映药材的整体药效物质基础。中药材的药效是其所含的多种活性成分 (药效物质基础) 协同作用的结果, 单一或少数几类次生代谢物难以全面、精准地反映药材的整体品质、疗效及批次间一致性。GNDC基于药材的基因组、转录组和代谢组等多组学数据, 可以解析出四类成分 (小肽、小RNA、次级代谢产物和碳水化合物), 为药材质量控制和道地性研究提供了更为全面的指标。植物小肽在植物生长发育、胁迫响应及植物–微生物互作中发挥关键作用, 并在提升作物抗逆性、激活植物免疫及优化植物–微生物互作方面表现出巨大潜力[36]。例如, 稳定抗菌肽 (SAMPs) 既能破坏病原细胞膜, 又能激活植物的内源防御系统, 在防治柑橘黄龙病等重大病害中展现了显著效果[37]。GNDC解析的药用植物小肽, 将在中药材品质形成、道地性和自身病害防御中扮演重要角色。小RNA包括microRNA、siRNA等, 作为重要的基因表达调控因子, 参与多种生物过程, 其种类和丰度直接影响药材中其他活性成分的合成积累, 其本身也具有跨物种调控潜力, 是潜在的重要活性物质和质量控制指标。建立中药材DNA条形码与四类成分的指纹图谱, 实现药材的产地溯源与质量控制, 将为建立更为科学的中药材质量评价模型, 实现单一成分到多维指纹图谱的评价体系奠定基础, 推动传统经验式质控向数据驱动的精准质控转型。

4.4 中医药AI模型开发 GNDC的亿量级天然成分数据为生成式AI、深度学习模型及多组学关联分析提供了高质量数据集, 推动天然产物研究从“实验试错”转向“数据预测”。机器学习的有效性在于其对大量数据的依赖, GNDC正好提供了这样一个丰富的数据源。利用GNDC海量的天然成分信息, 研究人员可以构建和训练多种适合中医药的AI模型。例如通过深度的数据挖掘和应用人工智能算法, 研究人员能够挖掘出中药药效物质形成规律, 整合基因组信息和酶功能信息, 开发中药药效物质生物合成途径预测模型, 深入探究药效物质的生物合成机制。

4.5 中药药效物质的合成生物学研究 GNDC收录的基因编码的成分均可直接或间接关联到物种的基因组信息, 这一特性使得研究者能够深入解析中药药效物质的生物合成途径, 从而利用合成生物学技术实现异源合成。通过这种方式, 中药药效物质的生产方式正在经历一场根本性的革新, 逐步从传统的“植物提取”转向更高效的“细胞工厂”模式[38]。例如研究人员可以利用GNDC中的天然成分进行活性研究, 获得的候选目标活性分子, 可以通过合成生物学技术进行异源合成, 为新药研发提供原料基础。

5 讨论

中医药作为中华民族数千年医疗实践的结晶, 其最核心的根基在于历经数千年、跨越无数代人的持续性临床验证。从《神农本草经》、《本草纲目》等奠基性典籍的记载, 到历代医家对理法方药的不断丰富和完善, 中医药体系是在反复观察、实践、总结、修正中逐步构建, 积累了极其丰富的有效方剂。大量经典名方 (如桂枝汤、六味地黄丸、安宫牛黄丸等) 在防治常见病、多发病、慢性病乃至某些重大疾病方面, 至今仍在临床实践中展现出确切的疗效和独特的优势。但中医药研究也面临成分复杂、靶标和机制不清等挑战。

传统的天然产物数据库主要依赖于化学结构特征进行分类, 缺乏将基因组信息与天然多样性相结合的整合框架。GNDC首次提出了基于天然成分合成基因的分类方法, 并将次级代谢产物、小肽、小RNA和碳水化合物进行了系统整合。利用人工智能和多组学分析技术, 解析了超过2.34亿个天然成分, 为中药药效物质药物发现和作用机制解析提供了丰富的资源, 特别是基因直接编码的分子 (如小RNA和基因衍生的小肽)的纳入, 开辟了理解其调控作用和治疗潜力的新途径。此外, GNDC数据库中整合了多个功能分析模块, 例如NPClassifier、ADMET-AI和DeepICER等工具, 用于结构注释和生物活性预测。

GNDC也面临许多挑战, 首先, 中药次级代谢产物的生物合成通常涉及复杂的合成路径, 攻克这一核心科学难题, 需要借助基因组挖掘和合成生物学等多学科交叉研究, 开发构建智能化生物合成途径预测平台。该平台需整合基因组, 开发基于深度学习的酶注释算法和代谢网络模拟系统, 实现对其生物合成途径的精准预测, 以探索它们的化学多样性和药物开发潜力。此外, 小RNA (特别是调控RNA) 代表了有希望的治疗途径, 基因组和转录组测序已识别了大量这些分子, 但由于缺乏有效的功能验证方法, 它们的生物学功能

仍然难以捉摸。开发高通量的功能验证体系与基于机器学习的预测模型, 对于揭示其作用机制至关重要。未来, 随着基因组学的发展, GNDC将继续扩大其数据规模, 以纳入更广泛的天然成分, 包括来自海洋生物、微生物和药用动植物的成分, 实时更新数据, 确保基因组学、转录组和代谢组学数据的最新发现得到持续整合。

GNDC首次实现了基因组信息、成分信息与功能信息的整合, 弥补了当前天然产物数据库在结构系统性与功能解析方面的不足。作为首个基于基因组信息实现“遗传基础–化学多样性”全面映射的开放获取综合性数据库, 它还集成了先进AI分析工具, 支持基因编码机制解析、天然成分多样性挖掘及潜在药学应用评估。未来, GNDC将在中药新成分和新药发现、新靶点发现、中药质量评价与控制、大数据驱动的中医药AI模型开发及中药药效物质的作用机制解析等领域发挥巨大潜力, 为中医药的研究开辟新范式。

作者贡献: 余志银负责资料收集、论文撰写和修改; 陈伟、孙丹、宫瑞泽和刘灏负责四个数据库内容撰写和修改; 刘成成和杨秀萍参与图表绘制; 冷梁、丛朝彤、宾华超和鲁军负责论文修改完善; 宋驰和张三印确定论文选题、写作指导和论文审阅。

利益冲突: 本文所有作者声明不存在利益冲突关系。

References

[1] Barona-Gómez F, Chevrette MG, Hoskisson PA. On the evolution of natural product biosynthesis [J]. Adv Microb Physiol, 2023, 83: 309-349.

[2] Khursheed A, Rather MA, Jain V, et al. Plant based natural products as potential ecofriendly and safer biopesticides: a comprehensive overview of their advantages over conventional pesticides, limitations and regulatory aspects [J]. Microb Pathog, 2022, 173: 105854.

[3] Newman DJ. Natural products and drug discovery [J]. Natl Sci Rev, 2022, 9: nwac206.

[4] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019 [J]. J Nat Prod, 2020, 83: 770-803.

[5] Setten RL, Rossi JJ, Han SP. The current state and future directions of RNAi-based therapeutics [J]. Nat Rev Drug Discov, 2019, 18: 421-446.

[6] Yamada Y. Nucleic acid drugs – current status, issues, and expectations for exosomes [J]. Cancers, 2021, 13: 5002.

[7] Zhu Y, Zhu L, Wang X, et al. RNA-based therapeutics: an overview and prospectus [J]. Cell Death Dis, 2022, 13: 644.

[8] Qiao X, Huang F, Shi X, et al. Herbal small RNAs in patients with COVID-19 linked to reduced DEG expression [J]. Sci China Life Sci, 2023, 66: 1280-1289.

[9] Zhang L, Hou D, Chen X, et al. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA [J]. Cell Res, 2012, 22: 107-126.

[10] Muttenthaler M, King GF, Adams DJ, et al. Trends in peptide drug discovery [J]. Nat Rev Drug Discov, 2021, 20: 309-325.

[11] Reshi ZA, Ahmad W, Lukatkin AS, et al. From nature to lab: a review of secondary metabolite biosynthetic pathways, environmental influences, and *in vitro* approaches [J]. Metabolites, 2023, 13: 895.

[12] Chen W, Yu Z, Leng L, et al. Artificial intelligence-curated repository of gene-encoded natural diverse components from herbal medicines [J]. Innovation, 2025, 6: 101011.

[13] Gao L, Yu X, Lei X. Biosynthesis of natural products: exploring the secret of how nature produces the secondary metabolites [J]. Univ Chem (大学化学), 2019, 34: 45-53.

[14] Zeng T, Li J, Wu R. Natural product databases for drug discovery: features and applications [J]. Pharm Sci Adv, 2024, 2: 100050.

[15] Chandrasekhar V, Rajan K, Kanakam SRS, et al. COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database [J]. Nucleic Acids Res, 2024, 53: D634-D643.

[16] Sami A, Fu M, Yin H, et al. NCPbook: a comprehensive database of noncanonical peptides [J]. Plant Physiol, 2024, 196: 67-76.

[17] Cao YH, Lin YX, Sun N, et al. A comprehensive analysis of the Bencao (herbal) small RNA Atlas reveals novel RNA therapeutics for treating human diseases [J]. Sci China Life Sci, 2023, 66: 2380-2398.

[18] Aoki-Kinoshita K, Agravat S, Aoki NP, et al. GlyTouCan 1.0-the international glycan structure repository [J]. Nucleic Acids Res, 2016, 44: D1237-D1242.

[19] Cheng LT, Wang ZL, Zhu QH, et al. A long road ahead to reliable and complete medicinal plant genomes [J]. Nat Commun, 2025, 16: 2150.

[20] Xie X, Gui L, Qiao B, et al. Deep learning in template-free *de novo* biosynthetic pathway design of natural products [J]. Brief Bioinform, 2024, 25: bbae495.

[21] Zheng S, Zeng T, Li C, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP [J]. Nat Commun, 2022, 13: 3342.

[22] Gangwal A, Lavecchia A. Artificial intelligence in natural product drug discovery: current applications and future perspectives [J]. J Med Chem, 2025, 68: 3948-3969.

[23] Liao BS, Hu HY, Xiao SM, et al. Global Pharmacopoeia Genome Database is an integrated and mineable genomic database for traditional medicines derived from eight international pharmacopoeias [J]. Sci China Life Sci, 2022, 65: 809-817.

[24] Meng F, Tang Q, Chu T, et al. TCMPG: an integrative database

for traditional Chinese medicine plant genomes [J]. Hortic Res, 2022, 9: uhac060.

[25] Su X, Yang L, Wang D, et al. 1 K Medicinal Plant Genome Database: an integrated database combining genomes and metabolites of medicinal plants [J]. Hortic Res, 2022, 9: uhac075.

[26] Kim HW, Wang MX, Leber CA, et al. NPClassifier: a deep neural network-based structural classification tool for natural products [J]. J Nat Prod, 2021, 84: 2795-2807.

[27] Swanson K, Walther P, Leitz J, et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries [J]. Bioinformatics, 2024, 40: btae416.

[28] Han J, Kong T, Liu J. PepNet: an interpretable neural network for anti-inflammatory and antimicrobial peptides prediction using a pre-trained protein language model [J]. Commun Biol, 2024, 7: 1198.

[29] Xie LJ, Gong XJ, Yang K, et al. Technology-enabled great leap in deciphering plant genomes [J]. Nat Plants, 2024, 10: 551-566.

[30] Tian Y, Sun GB, Zhang WD. HerbRNomes: ushering in the post-genome era of modernizing traditional Chinese medicine research [J]. Acta Pharm Sin (药学学报), 2025, 60: 300-313.

[31] Lorente JS, Sokolov AV, Ferguson G, et al. GPCR drug discovery: new agents, targets and indications [J]. Nat Rev Drug Discov, 2025, 24: 458-479.

[32] Yang HT, Wang YF, Liu W, et al. Genome-wide pan-GPCR cell libraries accelerate drug discovery [J]. Acta Pharm Sin B, 2024, 14: 4296-4311.

[33] Bi Z, Li H, Liang Y, et al. Emerging paradigms for target discovery of traditional medicines: a genome-wide pan-GPCR perspective [J]. Innovation, 2025, 6: 100774.

[34] Zhang SY, Zhang T, Chen W, et al. A novel strategy for development and evaluation of traditional Chinese medicine based on genome-wide pan-receptor drug-target co-screening [J]. Acta Pharm Sin (药学学报), 2025, 60: 2375-2388.

[35] Zhang T, An WQ, You SJ, et al. G protein-coupled receptors and traditional Chinese medicine: new thinks for the development of traditional Chinese medicine [J]. Chin Med, 2024, 19: 92.

[36] Xiao F, Zhou H, Lin H. Decoding small peptides: regulators of plant growth and stress resilience [J]. J Integr Plant Biol, 2025, 67: 596-631.

[37] Huang CY, Araujo K, Sánchez JN, et al. A stable antimicrobial peptide with dual functions of treating and preventing citrus Huanglongbing [J]. Proc Natl Acad Sci U S A, 2021, 118: e2019628118.

[38] Seshadri K, Abad AND, Nagasawa KK, et al. Synthetic biology in natural product biosynthesis [J]. Chem Rev, 2025, 125: 3814-3931.