



药络智控

药络智控

——多组分均化解决方案

作品名称：“药络智控”——基于深度学习及网络
药理学的中药多组分智能均化系统

参赛成员：黄安东、张博扬、欧阳文锐、王嘉芃

参赛单位：广州医科大学

标题名称：以科技赋能中药产业，解决关键质量
控制难题

前言

中医药作为中华民族的瑰宝，在全球健康领域扮演着日益重要的角色。然而，在其现代化进程中，如何确保中药产品质量的稳定性和一致性，始终是一个亟待解决的瓶颈问题。中药具有多组分、多靶点的复杂特性，这既是其疗效的物质基础，也导致了不同批次间有效成分含量与整体药效的波动，成为长期困扰产业发展的关键质量控制难题。

在此背景下，将人工智能、大数据等前沿信息技术与传统中药产业深度融合，为破解这一难题提供了前所未有的机遇。本报告所呈现的“药络智控”——基于深度学习及网络药理学的中药多组分智能均化系统，正是应对这一挑战的创新性探索与实践。该系统旨在通过科技赋能，为中药生产过程中的质量控制提供智能化、可视化的解决方案。

本系统综合运用了机器学习、多目标优化算法（如 SLSQP 与 NSGA-II）及网络药理学知识，构建了一套完整的智能决策流程。它不仅能够对涵盖理化指标、指纹图谱及成本信息的多维数据进行深度分析与质量评分，更能在此基础上，智能求解兼顾质量最优与成本可控的最佳批次配比方案。为提升系统的可用性与透明度，我们依托 Streamlit 框架构建了高度可视化的交互界面，通过交互式仪表盘、统计图表等形式，将复杂的优化过程转变为一个透明、可交互的数据分析与决策支持平台，而非一个封闭的“黑箱”。

本研究的意义不仅在于提供了一个解决中药均化问题的实用工具，更在于探索了一条将现代数据科学与传统中药理论相结合的技术路径。“药络智控”系统旨在为中药产业的质量控制与现代化转型提供一套科学、高效、智能的解决方案，为推动中医药的继承创新与高质量发展贡献一份力量。

团队介绍



黄安东，“药络智控”项目负责人，广州医科大学麻醉学专业本科生。曾获国家奖学金、一等奖学金，获评广州医科大学“优秀共青团员”、优秀三好学生、优秀学生干部、广州青年马克思主义工程优秀学员。获竞赛奖项、大创项目多项，以共同作者发表高水平 SCI 论文 1 篇，国家软件著作权 1 项，在投 SCI 期刊 4 篇。



张博扬，“药络智控”项目核心成员，广州医科大学口腔医学专业本科生。曾获医学创新大赛国银奖，为国家级大创核心成员，发表 2 篇 SCI 论文。研究方向包括临床药物治疗和随机对照研究等。



欧阳文锐，“药络智控”项目核心成员，华南师范大学数据科学与大数据技术专业本科生。曾获第十九届挑战杯“揭榜挂帅”专项赛国家级二等奖和 2024 年全国大学生数学建模大赛省级一等奖，为省级大创项目负责人，以第一作者发表高水平 SCI 论文 1 篇，MSCE 2025 EI 会议 1 篇。



王嘉芃，“药络智控”项目核心成员，广州医科大学医学影像学专业本科生。曾获第十一届全国大学生统计建模大赛广东省二等奖，以独立第一作者在 Pharmaceutics 杂志上发表综述，以主要成员参加 2025 国家级大学生创新创业项目，以独立一作参加“中华医学会第二十四次超声医学学术会议”壁报交流等。



汪李炎，“药络智控”项目核心成员，江西中医药大学中西医临床医学专业本科生，作为主编出版 1 本中药科普类书籍，发表 1 篇论文。

目录

一、绪论	5
1.1 研究背景	5
1.1.1 中药产业的战略地位与现代化瓶颈	5
1.1.2 技术发展机遇：人工智能与大数据赋能中药现代化	6
1.2 国内外研究现状	7
1.2.1 国内研究现状	7
1.2.2 国外研究现状	8
1.3 研究内容	8
1.3.1 中药的多组分特性	8
1.3.2 中药功效物质的发现	9
1.3.3 确保中药质量一致性	11
1.3.4 均化效果及综合评价	11
1.4 研究意义	12
二、程序设计	13
2.1 系统整体架构	13
2.2 开发环境	14
2.3 数据扩充	15
2.4 数据统计	17
2.5 数据评分	17
2.6 优化求解 SLSQP	19
2.7 优化求解 NSGA-II	20
2.8 智能化决策支持与人机交互模块	22
2.9 程序搭建	23
三、实验与结果展示	24
3.1 数据集介绍	24
3.2 机器学习评分	25
3.3 批次质量分析仪表盘	28
3.4 成分含量深度分析统计图:	32
3.5 可视化引擎对均一化结果的优化	33
3.6 程序展示（程序使用方法界面介绍）	33
3.7 实验结果分析	39
3.7.1 SLSQP 实验结果图	40
3.7.2 NSGA-II 实验结果图	43
四、总结	45
五、参考文献	46

一、绪论

1.1 研究背景

中药作为中华民族数千年智慧的结晶，其独特的理论体系和在疾病预防、治疗及康复中的显著优势，使其在全球健康领域日益受到关注。尤其是在全球大健康理念和预防医学日益普及的背景下，中药的整体观念和个性化治疗方案使其在全球范围内获得了广泛认可。这种深厚的历史底蕴和不断拓展的国际影响力，奠定了中药在全球健康事业中的战略地位。与此同时，面对中药生产与应用中存在的质量均一性挑战，中药多组分均化技术应运而生。这一技术通过对多种活性成分的科学配比与均衡混合，能够有效降低批次间成分含量差异，显著提升制剂的标准化水平和质量一致性，不仅保障了药效的稳定性与可重复性，还为中药的国际化推广提供了关键技术支撑。

1.1.1 中药产业的战略地位与现代化瓶颈

当前，中药市场正经历一个快速发展的阶段。根据中康科技等机构的预测，到 2025 年，中国中医药市场规模有望达到 2.2 万亿元。这一增长源于多方面需求，包括民众日益提升的健康意识，以及对中药在疾病预防和康复中的信任。国际市场方面，中药的接受度也在不断提高。2024 年，我国中药类商品对外贸易总额已达到 83.9 亿美元，全球中药市场规模持续扩张，显示出中药产业强劲的发展趋势。



图 1.1 2009~2022 年国内中药材产供需评估（来源：中康科技中药大数据中心）

尽管中药具有显著的独特价值，但其现代化和国际化进程长期以来面临一个根本性挑战：中药制剂的质量均一性问题。中药制剂批间质量差异带来的影响是多方面的，不仅局限于生产环节，更对中药的临床应用、市场推广乃至国际地位产生深远制约。首先，批次间有效成分含量的波动直接导致药效的不可预测性，这可能严重影响治疗效果的稳定性与可重复性。更甚者，若批次间成分含量不足或杂质超标，可能引发潜在的安全性问题，直接影响药效稳定性与安全性。其次，缺乏批间质量的一致性，使得中药难以进行大规模、标准化的临床

试验，从而限制了其在现代医学体系中的推广和认可。这直接阻碍了中药循证医学的发展，使其难以被更广泛的医疗体系所接纳。最后，国际药品监管机构对药品的质量标准和可重复性有着极其严格的要求，例如良好生产规范和各国药典标准。中药固有的批间差异性使其难以满足这些国际标准，从而阻碍国际市场拓展。这直接限制了中药在全球范围内的注册和流通，削弱了其国际竞争力。

1.1.2 技术发展机遇：人工智能与大数据赋能中药现代化

面对中药质量均一性的挑战，中药多组分均化技术应运而生，并被视为一项具有深远战略意义的创新性解决方案。该技术通过精确优化多组分药材的混合比例，能够显著提升中药制剂的标准化水平，从而保证不同批次产品的质量一致性，进而确保药效的可靠性与可重复性。这种质量一致性的实现，是中药作为药品最核心属性的保障，也是其走向现代化、实现精益化生产的关键一步。

中药多组分均化技术在推动中药现代化方面具有重要前景，但目前仍处于发展初期，尚未形成一套成熟的标准化方案。这为技术创新提供了广阔空间，亟待突破性进展。近年来，随着中药现代化研究的深入，以及人工智能和大数据技术的迅速崛起，为该领域的软件开发奠定了前所未有的技术基础，并带来了巨大的创新机遇。这些前沿工具为解决中药固有的复杂性问题提供了全新途径。

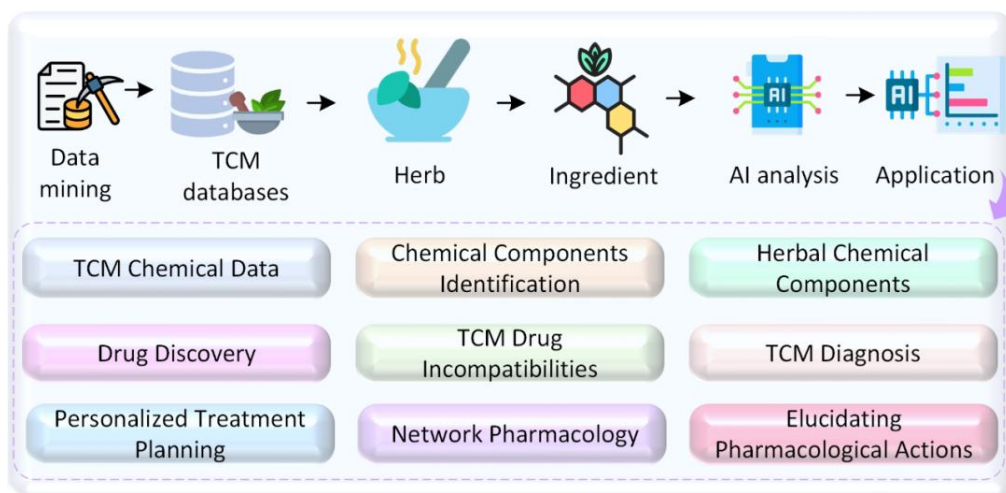


图 1.2 基于人工智能的中医药化学数据分析

（来源：北大陈语谦团队《AI empowering traditional Chinese medicine?》）

人工智能与大数据技术的深度融合，不仅仅是工具的简单叠加，更是推动中药研发与生产模式从传统的“经验试错”向“预测优化、智能决策”转变的关键驱动力。传统中药的研发和生产高度依赖经验判断和耗时耗力的试错过程。而本项目所采用的数据处理、机器学习评分、优化求解及 AI 智能助手等技术，从根本上改变了这一范式：机器学习评分实现了预测性质量评估，从被动控制转向主动预判；优化求解通过算法确定理想配比策略，最大限度减少人工干预并提高效率；AI 智能助手则通过提供智能、上下文感知的指导和分析，进一步增强了这一转变。这些技术共同促成了中药研发与制造流程的量化、预测和高度优化，对于加速创新、缩短开

发周期、确保产品质量一致性具有不可替代的作用。

1.2 国内外研究现状

随着中医药产业的快速发展和国际化进程的加速，中药质量控制技术正面临前所未有的机遇与挑战。由于中药成分的多样性和生产工艺的复杂性，传统质量控制模式已难以满足现代化生产和国际标准的要求。目前，国内外研究者正从不同角度探索解决中药质量一致性难题的创新路径。国内研究主要聚焦于以功效物质为核心的质量控制理念创新，积极探索数字孪生、人工智能等前沿技术在中药智能制造中的应用；而国外则更注重建立系统化、法规化的质量管理框架，强调“质量源于设计（Quality by Design, QbD）”的全过程控制策略。尽管研究路径存在差异，但均旨在通过技术创新和管理优化，实现中药产品质量的标准化和一致性。

1.2.1 国内研究现状

我国在实现中药质量均一化方面已作出多种尝试。操作上通过超临界流体萃取、微波萃取、色谱分离等多技术耦合，实现中药组分的高效提取与标准化分离。技术上建立标准组分库和多模式色谱系统，以系统化、标准化的方式获取中药化学物质基础，减少研究重复性，提升研究效率。

在科学性方面，均一化契合中药多成分协同作用和成分存在变异性的特点，能有效稳定药效相关成分含量，减少不同批次产品之间的疗效差异。Liu YS 等^[1]、Yang M 等^[2]均提出通过混合多个批次的提取物，使混合物和参考标准之间的含量相似度最大化的均一化方法。然而在工业生产中通常首选混合较少批次的提取物，以便于操作和在产品缺陷的情况下根据生产规范指南进行追溯。Bin-jun Yan 等（PMID: 24190450）^[3]提出使用最小批次的提取物以简化混合作，为间歇式混合工艺的设计提供了参考，初步实现在保证生产效率的同时减少中药质量。

为解决质量控制技术难题，阳长明等^[4]通过源头质量控制（如药材种植、工艺参数优化）和过程质量监控（如智能反馈调控）解决质量波动问题。陈霞等^[5]利用国家药品抽检和探索性研究优化质量标准，提升中药质量控制水平。通过非线性规划、遗传算法等数学模型优化多组分混合比例，可显著提高不同批次产品质量的一致性，减少成分波动。已有研究证明，绿茶提取物制剂通过多批次混合均一化和双定性双定量相似度混批勾兑技术可消除批次间化学成分差异^[6]。

将智能质量控制技术融入智能制造将为中药产品符合国际质量标准提供强有力的技术支撑。唐雪芳等^[7]对中药生产过程控制技术的综合分析表明，智能质量控制技术已初步应用于中药生产全过程；大数据、人工智能等技术将提升中药生产过程的质量控制水平，实现生产过程的精准调控和质量均一性提升。肖伟等^[8]研发基于数字孪生的智能生产设备和工业大脑，通过知识图谱、人工智能等技术优化生产流程，提升中药制造的智能化和质量稳定性。在质量控制技术的创新与标准化领域，李玲等^[9]建立符合中药特点的标准化质量评价体系，通过色谱、光谱、人工智能多技术耦合建立中药质量标准体系，推动中药质量控制的科学化和规范化。

我国当前研究聚焦于多技术集成、标准化平台构建及质量控制，但多组分均化技术的系统性研究仍需进一步深化，特别是在多组分均化工艺的标准化和大规模应用方面存在挑战。科技赋能通过智能制造、智能控制、标准化和政策支持，能系统性解决中药产业的关键质量控制难题，推动了中药产业的现代化和高质量发展。

1.2.2 国外研究现状

国际上，虽然药物理论体系不同，但针对植物药内在复杂性和多变性带来的质量挑战，采用了相似但体系化、法规化的策略。

欧洲药品管理局（EMA）和美国食品药品监督管理局（FDA）等监管机构，已经建立了法规驱动的质量管理体系，其特点在于并非仅在生产末端进行检验，而是将质量控制前置到研发和设计阶段。QbD 的理念要求在研发初期明确产品的关键质量属性，最终确定一个能稳定生产出合格产品的“设计空间”。这种方法将质量控制从“测试合格”的被动模式转变为“构建进去”的主动模式，提升了产品质量的稳定性和可预测性^[10]。

然而，国际监管框架的重点在于保证商业化生产的每一批产品都与通过临床试验证明安全有效的批次相同，这可能限制了对草药更优配比或更高效提取工艺的探索。其次，该体系高技术门槛和高昂成本不容忽视。全面实施质量全过程控制策略需要大量的先期投入、先进的仪器设备和专业的跨学科人才团队，这对于中小型企业构成了挑战。且该体系虽然承认其多成分协同作用的复杂性，但往往不过分深究其具体的物质基础和作用机理，这在一定程度上阻碍了对草药功效本质的深入理解。

国际研究的成果和框架为解决中药材功效物质发现与质量一致性控制的难题提供了灵感。FDA 在植物药指南中鼓励在化学标志物与疗效关联尚不明确时，使用生物测定来评价批次间的功能一致性。中国在“功效物质”研究中已经积累了关于药效评价的模型和方法，因此，应将这些方法标准化、规范化，并将其作为常规放行检测的一部分，与化学指纹图谱互为补充，从而确保混批中药产品不仅具有化学一致，而且功效一致。

1.3 研究内容

中药多组分均化技术，是依据《中药均一化研究技术指导原则（试行）》^[11]（以下简称《指导原则》），结合中药多组分、多靶点、整体发挥疗效的特性，为保证最终制剂批间质量稳定、解决中药关键质量控制难题而开展的一项系统性研究。其研究内容包括中药多组分特性的研究、中药功效物质的发现、中药质量一致性研究、均化效果总结及综合评价等。

1.3.1 中药的多组分特性

中药的多组分特性是中药多组分均化技术的物质基础。与现代西药通常是“单一成分、单一靶点”的模式不同，中药是一种复杂的天然混合物。一味中药里往往含有成百上千种化学成分，包括主要活性成分、次要活性成分以及微量成分。这些成分共同构成了一个有机的整体，通过不同的方式协同作用产生最终的治疗效果。传统中药的运用是在保证用药安全的前提下，依据中药的药性和配伍原则将药材或饮片制成方剂；随着现代组

分中药的研究不断深入，中药的运用比传统的配伍方式更加丰富、全面，其“多组分、多靶点”的特点在中医药治疗体系中得到了广泛的运用，但也为功效物质发现、中药质量控制、中药生产标准化以及确保疗效一致性带来了巨大挑战。

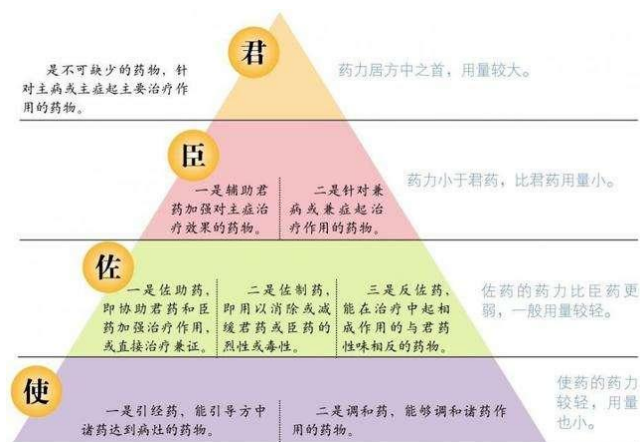


图 1.3 中医药理论中方剂的组方配药原则——君臣佐使理论

中药的多组分特性可以在实际运用中得到具体体现。在成分基础方面，一味中药是包含多种天然化学成分复杂体系。在作用模式方面，这些成分并非简单叠加，而是通过协同增效、相互制约（减毒）、多靶点治疗等方式有机地结合在一起。在疗效体现方面，最终产生的整体药效远比任何单一成分所能达到的效果更为全面和温和，完美契合了中医治疗疾病时强调的“平衡阴阳”和“整体观念”的哲学思想。

1.3.2 中药功效物质的发现

发现中药功效物质并择取有效和关键物质是中药多组分均化技术的实践途径。目前检测并发现中药功效物质的方法有化学分析方法、物理分析方法、指纹图谱与多组分定量分析等。

化学分析方法^[12]主要通过测定样品中特定化学成分的含量或整体化学轮廓评估其均一性。高效液相色谱法（HPLC）及超高效液相色谱法（UPLC）是评估中药均一性，特别是含量均匀度和化学指纹图谱一致性的核心技术；液相色谱-质谱联用技术（LC-MS）在中药复杂体系的成分鉴定和定量方面发挥着关键作用，它不仅可以提供组分的分子量信息，还能通过多级质谱提供结构信息，有助于区分同分异构体；紫外-可见分光光度法（UV-VIS）常用于测定具有紫外或可见吸收的组分含量或总成分（如总黄酮、总皂苷）的量，也可用于含量均匀度的初步评估；其他化学方法如薄层色谱法（TLC）、气相色谱法（GC）及其与质谱的联用（GC-MS）、毛细管电泳（CE）也可用于分析中药样品中化学成分的含量等。

物理分析方法通常是非破坏性的，能够快速提供样品整体或局部物理化学性质的信息，常用于过程监控和快速筛选。近红外光谱（NIR）技术因其快速、无损、可在线/原位测量的特点，在评估中药制剂（特别是固体制剂）混合均匀性、水分含量、粒径以及活性成分含量均匀度方面得到了广泛应用；拉曼光谱（Raman Spectroscopy）技术与 NIR 类似，也是一种振动光谱技术，能够提供样品的化学成分和分子结构信息；光谱成

像技术（如 NIR 成像、拉曼成像）将光谱信息与空间信息相结合，能够生成化学图像，直观显示不同组分在样品表面的分布情况，从而直接评估其空间均一性；热分析技术如差示扫描量热法（DSC）和热重分析法（TGA）可以提供关于样品热力学性质和组成的信息，有时可用于评估混合物的均一性或鉴别不同组分；机器视觉与数字图像处理技术：通过分析样品的外观（如颜色、纹理、斑点）来评价其均匀性，尤其适用于散剂、颗粒剂等的外观均匀度检查。结合色彩色差计等工具，可以将颜色等视觉特征数字化，进行定量评估。



图 1.4 SWIFT XS 拉曼光谱仪（来源：HORIBA 公司产品展示页）

指纹图谱技术^[13]是评价中药整体化学特征及其一致性的重要手段。它不追求对所有化学成分的绝对定量，而是通过色谱（HPLC、GC、UPLC 等）或光谱（NIR、IR 等）手段获得一个能够代表该中药样品特性的、相对完整的化学信息图谱。通过比较不同样品指纹图谱的相似度（如共有峰模式、峰面积比等），可以评估其批内和批间的均一性。多组分定量分析则更进一步，旨在同时测定中药中的多个活性或指标成分的含量。这比单指标成分测定更能反映中药的整体质量状况。例如，“一测多评”（QAMS）模式通过测定一个易得成分，利用其与多个其他成分间的定量关系，实现对多个成分的同时测定，为解决中药对照品缺乏的问题提供了有效途径。多元多维定量指纹图谱则结合了指纹图谱的整体性和多组分定量的精确性，从多个侧面表征中药的化学组成和含量特征，为质量控制提供了更全面的信息。

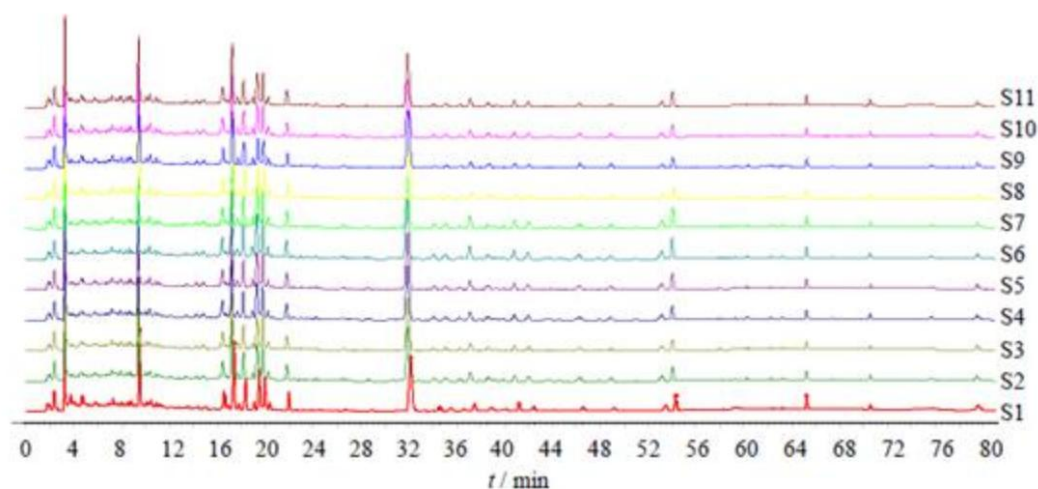


图 1.5 11 批不同产地麦冬 HPLC 指纹图谱

(来源: 陈文,张辉,崔小莹等,中药麦冬 HPLC 指纹图谱研究)

1.3.3 确保中药质量一致性

确保中药批间质量的一致性与稳定性是中药多组分均化技术的核心环节。中药研发和生产中所有均化工作的根本和首要目标是确保稳定的批间质量一致性。这意味着特定中药产品的每一批都应在多组分化学特征及其延伸的疗效和安全性方面,与先前和后续批次尽可能相同。在均化过程中,可以通过控制均化评分标准、设置饮片及中间体质量标准、实现含量均匀度达标等确保中药批间质量的一致性。

《指导原则》明确规定,如果均一化涉及使用如提取物之类的处方药味,其质量标准必须包括其制备工艺,强调了对中间体一致性的需求。饮片和提取物通常是生产中药最终剂型的直接投入物料。确保它们在化学特征和物理性质方面的均一性,是成品质量一致性的关键控制点。含量均匀度是一个关键的质量属性,直接反映了在单个剂量单位水平上均化工作的成功。它是大多数固体制剂和某些液体制剂的关键监管要求。

除了确保散装物料或少数几种标志性化合物的均匀分布外,先进的中药均化策略可以确保整体“物质基础”——即被认为是产生治疗效果的复杂化学成分混合物的一致性。这进一步延伸到确保各种有效成分或成分组之间“配伍”的持续一致性。不同成分间的协同或调节相互作用是中药药理学的一个标志,维持这些相互作用的一致性对于可靠的疗效至关重要。实现这种更高层次均一化的策略包括有效成分组合的筛选与鉴定、组分比例的优化等^[14]。

1.3.4 均化效果及综合评价

设计并实现均化全过程是中药多组分技术的最终目标。在中药领域,均化是一个综合性概念,其内涵远不止于通过简单的物理混合达到外观上的均匀。它涵盖了多组分体系在单个制剂批次内不同剂量单位、不同生产批次中质量的一致性。《指导原则》明确指出,“均一化研究应以制剂批间质量稳定为目标”,说明中药均化

并非仅仅追求产品外观的一致，而是要确保其复杂的化学指纹图谱，乃至其潜在疗效，在每一剂药物和每一个批次中都能得到可靠的重现。这为所采用的技术和策略设定了很高的标准。

QbD 的原则是实现中药生产中稳健均化的基础性指导思想。QbD 强调一种系统性的研发方法，始于预先设定的目标，并基于可靠的科学依据和质量风险管理（Quality Risk Management, QRM），注重对产品和工艺的理解及过程控制。QRM 是 QbD 不可或缺的组成部分，用于在产品生命周期的各个阶段主动识别、分析和减低可能导致不均一的潜在风险。同时，《指导原则》明确倡导 QbD 的理念。QbD 为从一开始就将均一性构建到中药产品和工艺中提供了一个主动的、科学驱动的、基于风险的框架，这与传统上严重依赖最终产品检验来推断质量的方法相比，是一个显著的进步。

在均化后的效果评估方面，检测并发现中药功效物质的方法如化学分析方法（如 HPLC、LC-MS）、物理分析方法（如 NIR、拉曼光谱）以及指纹图谱和多组分定量等整体表征手段的综合运用同样适用于评估均化后的中药批次，为全面评价中药的均一性提供了多样化的工具。同时，引入过程分析技术（PAT）^[15]，尤其是近红外光谱等在线监测技术的应用，可以使对生产过程进行实时监控和反馈控制成为可能，从而将质量保证从事后检验推向过程构建。



图 1.6 QbD 理念的螺旋式上升结构

1.4 研究意义

本项目“中药多组分智能均化软件”的研究与开发，在理论和实践上均具有较强的实用意义，对中药产业的现代化、智能化以及全球化发展具有重要的推动作用。该项目通过引入 LightGBM 等机器学习模型进行多维度质量评分，并结合多目标优化算法实现智能混批，突破了传统中药质量控制中依赖经验和单一指标的局限性。同时，项目旨在通过技术创新解决中药产业的核心质量控制难题，而非仅仅是应用现有成熟技术。在经济效益方面，该软件能够显著降低中药材的资源浪费，通过智能配比不同质量等级的药材批次，有效优化成本结构，提升中药用户的经济效益和市场竞争能力。在社会效益方面，通过确保中药制剂的质量均一性，该项目直接保障

了公众用药的安全性及药效的可靠性，增强了消费者对中药产品的信心，与国家“健康中国”战略高度契合。从国家战略和产业发展高度看，本项目将为中药产业的数字化转型和智能化升级提供关键支撑，引领整个产业迈向更高发展阶段，为中药走向世界开辟新的发展道路。

二、程序设计

2.1 系统整体架构

本研究构建的优化求解与分析系统，旨在针对中药多组分配方在质量均一性、成分达标率及成本控制等多目标优化需求，提供一套从数据准备、特征处理、建模评估到优化求解和结果可视化的完整技术方案。系统整体架构采用模块化设计思想，由数据预处理与扩充模块、特征分析与统计模块、机器学习评分模块、多目标优化求解模块、结果可视化与决策支持模块五个核心部分组成，各模块之间通过统一的数据接口与参数传递机制实现有机联动。

在数据预处理与扩充阶段，系统首先对原始中药批次检测数据进行清洗与格式化处理，去除缺失值和异常值，并统一度量单位与成分命名规范。针对样本数量有限的问题，引入 CTGAN（Conditional Tabular GAN）生成模型进行合成数据扩充，通过深度生成网络在保持原始数据统计分布特征的前提下，生成高质量的虚拟批次数据，以提升后续建模与优化的鲁棒性与泛化能力。

特征分析与统计模块利用集成分析对用户的批次数据进行成分分布、批次间差异以及指标相关性的多维统计分析，生成成分浓度分布直方图、批次用量比例图、成分达标率对比图等基础统计图表。这一过程不仅为后续机器学习模型提供数据洞察，也为优化目标的设定提供了量化依据。

机器学习评分模块基于前述处理后的数据，建立批次质量评分模型。该模块综合考虑甘草苷、甘草酸等核心成分的含量与波谱相似度指标，采用加权打分体系，将批次转化为可比较的质量评分值。模型训练过程中引入交叉验证方法，以确保评分结果的稳定性和可靠性。评分结果不仅作为优化算法的输入特征之一，还用于在可行解筛选阶段剔除低质量批次。

多目标优化求解模块是系统的核心计算引擎，支持 NSGA-II（Non-dominated Sorting Genetic Algorithm II）与 SLSQP（Sequential Least Squares Programming）两种优化策略的切换。NSGA-II 用于求解批次含量偏差最小化与指纹图谱相似度最大化等冲突目标之间的帕累托最优解集，并提供可视化的前沿分布图以辅助权衡分析；SLSQP 则用于在约束条件严格、目标函数连续可导的情况下进行精确优化，得到单一最优解。该模块可结合批次使用比例、成本约束、质量评分等多维度条件，输出符合实际生产要求的原料组合方案。

结果可视化与决策支持模块负责将优化结果以多种可视化形式呈现，包括帕累托前沿散点图、批次使用比例环形图、批次用量分布柱状图、成分达标率对比图、质量评分分布图、成本效益分析图以及库存使用率监控

图等。这些图表直观展示了优化过程中的关键数据特征与结果差异，为配方决策提供了多角度支撑。系统同时支持多方案雷达图对比功能，可在相似度、含量偏差、批次数量等指标维度下对不同优化方案进行综合性能评价。

整体而言，该系统实现了从数据、建模、优化、可视化到决策的闭环流程，能够为中药多组分配方的生产与质量管理提供科学、可量化的技术支持，并具备良好的可扩展性和适应性，可根据不同药材品种、质量指标和生产需求进行快速调整与部署。



图 2.1 系统整体架构图

2.2 开发环境

本项目的开发与运行在高度集成的数据科学与交互可视化环境中进行，旨在为中药多组分均化问题提供高效、稳定且可扩展的技术支持。在实际研发过程中，团队选择了 64 位 Windows 11 专业版作为主要操作系统平台，配备 AMD Ryzen 7 5800H 处理器与 NVIDIA GeForce RTX 4090 独立显卡。

在软件环境配置方面，项目依托 Anaconda 发行版进行 Python 数据科学与机器学习开发，核心版本为 Python 3.9。Anaconda 的虚拟环境隔离机制使得数据扩充（CTGAN）、统计分析（R 语言接口）与主程序开发可以在不同环境中独立运行，避免依赖冲突。数据处理与分析主要依赖 Pandas 和 NumPy 库，其中针对 NumPy 2.x 版本的兼容性问题，代码中加入了动态属性别名的前向兼容补丁，以确保在未来版本中依旧能够稳定运行。优化计算部分结合了 SciPy 提供的 SLSQP 算法与自研 NSGA-II 遗传算法，满足单目标与多目标均化策略的不同需求。机器学习评分模块使用 LightGBM 对药材批次进行多维度质量评分，评分结果直接作为优化求解的输入，提高了均化配比的科学性与准确性。

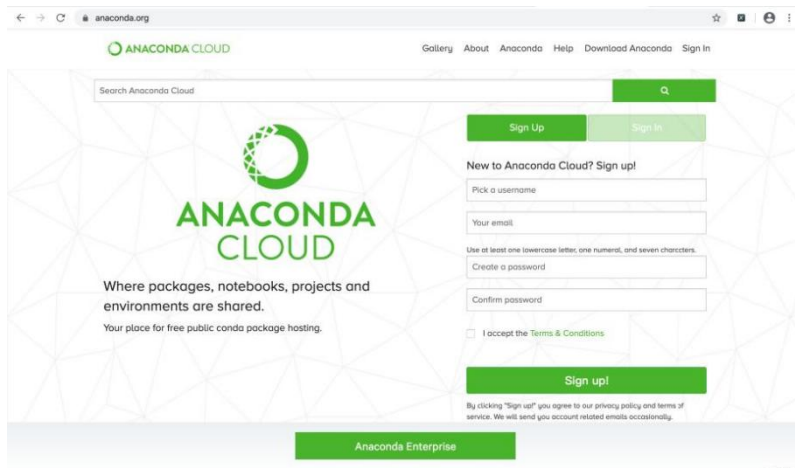


图 2.2 Anaconda 官网示意图

可视化与用户交互方面，系统基于 Streamlit 框架搭建 Web 应用前端，结合自定义 CSS 与 HTML 组件美化界面，实现了卡片式布局、动态步骤指示器、渐变按钮等增强设计，使得复杂的数据分析流程以直观、友好的方式呈现。Matplotlib 在绘制过程中通过自定义字体加载策略解决了跨平台中文显示的问题，Plotly 与 Seaborn 的引入为系统提供了交互式仪表板与高密度统计图形展示功能，从原始批次分析到优化结果解读均具备良好的可视化表现力。同时，系统集成了基于 Requests 库的 AI 智能助手，能够与外部大模型 API 通信，并在网络不可用时自动回退到本地规则式应答模式，为用户提供实时、上下文感知的操作指导与分析建议。

在实际开发过程中，团队使用 PyCharm 2023 专业版作为主要集成开发环境，结合 Git 进行版本管理与多人协作。为确保环境一致性与部署可移植性，项目在早期即建立了 environment.yml 文件与依赖锁定机制，方便不同开发节点和生产环境的快速搭建。整体而言，该开发环境的配置不仅保证了项目在性能、稳定性与交互体验上的综合需求，也为后续功能扩展与产业化落地奠定了坚实基础。

本系统在设计过程中，围绕中药多组分均一化和批次质量评价的核心需求，构建了集数据生成、统计分析、机器学习建模、优化求解和可视化交互于一体的综合性平台。系统结构呈现高度模块化特征，各模块既可独立运行，也可按既定顺序实现自动化串联，确保从原始数据导入到优化方案输出的全过程高效、稳定、可重复。具体包括数据扩充（CTGAN）、数据统计（R 语言集成）、数据评分、优化求解（NSGA-II）以及程序搭建五个主要部分。

2.3 数据扩充

在中药材批次数据采集过程中，受限于检测成本、样本获取难度及实验周期等多方面因素，原始数据规模往往不足，且在不同批次、不同产地之间存在显著分布差异。尤其是在高质量样本数量偏少的情况下，直接训练机器学习模型容易出现欠拟合或在测试集上性能下降的问题。为应对这一挑战，系统在数据扩充环节引入了 CTGAN 算法。这是一种专门面向表格型数据的生成对抗网络变体，能够有效建模混合型数据的复杂分布，并在保留变量间相关关系的同时生成具有统计一致性的合成样本。

CTGAN 的工作原理基于生成器和判别器之间的博弈过程。生成器根据输入的条件向量和噪声信号生成伪造样本，判别器则尝试判断样本的来源是真实数据还是生成数据。随着训练迭代的进行，生成器不断提升生成样本的真实性，而判别器则不断提高判别的准确度，最终达到纳什均衡，使得生成样本在分布特性上与真实数据高度接近。通过条件向量机制，系统可以在特定产地、年份或工艺条件下生成目标样本，从而在保证多样性的同时兼顾数据的针对性。扩充完成后，系统对合成样本与原始样本进行分布一致性检验，包括 K-S 检验、特征直方图对比和相关性矩阵分析，以确保扩充数据不会引入显著的分布偏差。这一过程有效缓解了数据稀缺与不平衡问题，为后续建模提供了坚实的数据基础。

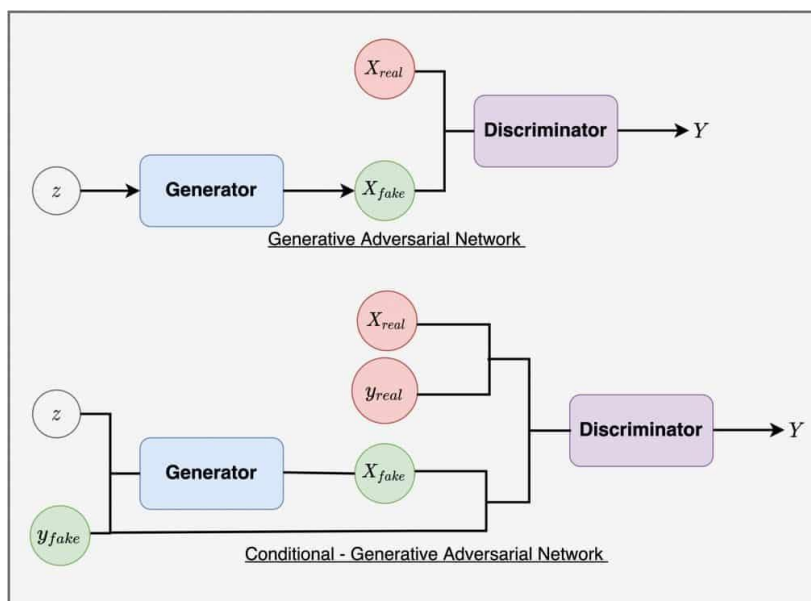
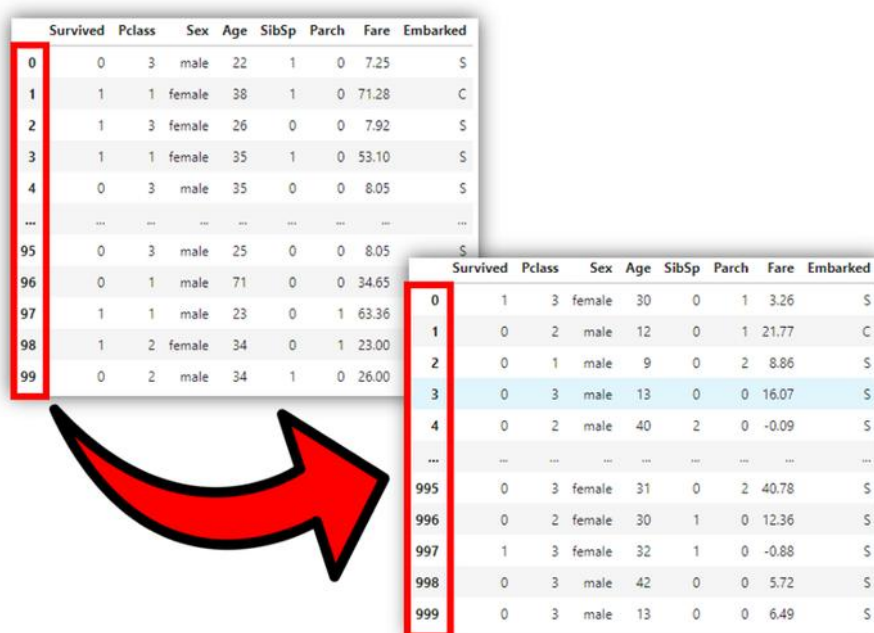


图 2.3 CTGAN 整体流程图示



	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22	1	0	7.25	S
1	1	1	female	38	1	0	71.28	C
2	1	3	female	26	0	0	7.92	S
3	1	1	female	35	1	0	53.10	S
4	0	3	male	35	0	0	8.05	S
...
95	0	3	male	25	0	0	8.05	S
96	0	1	male	71	0	0	34.65	S
97	1	1	male	23	0	1	63.36	S
98	1	2	female	34	0	1	23.00	S
99	0	2	male	34	1	0	26.00	S

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	3	female	30	0	1	3.26	S
1	0	2	male	12	0	1	21.77	C
2	0	1	male	9	0	2	8.86	S
3	0	3	male	13	0	0	16.07	S
4	0	2	male	40	2	0	-0.09	S
...
995	0	3	female	31	0	2	40.78	S
996	0	2	female	30	1	0	12.36	S
997	1	3	female	32	1	0	-0.88	S
998	0	3	male	42	0	0	5.72	S
999	0	3	male	13	0	0	6.49	S

图 2.4 CTGAN 扩展数据效果示意图

2.4 数据统计

在本研究中，统计分析的核心任务是对多批次原料及优化配方相关指标进行定量化描述、差异性比较与规律性挖掘，为后续的机器学习建模与多目标优化提供坚实的数据基础。统计原理的应用贯穿数据收集、预处理、特征分析及结果评价的全过程，其理论基础主要涵盖描述统计、推断统计、多变量分析和相关性分析等模块。

下文系本程序所用到的统计方法与意义：

2.4.1 在数据初步分析阶段，采用描述统计方法对原料批次的主要理化指标（如关键成分含量、质量评分、成本等）进行集中趋势（均值、中位数）与离散程度（方差、标准差、极差）计算。这一过程能够揭示数据总体分布特征及批次间差异的数量化表现，并为检测潜在异常值与数据不一致性提供参考。

2.4.2 在批次间的成分差异和质量差异比较中，引入推断统计原理，如假设检验（t 检验、方差分析）与置信区间估计，用于判断不同批次原料在特定指标上的差异是否具有统计学显著性。这一环节的分析结果，有助于确认哪些批次原料在关键质量特征上存在系统性差异，从而为批次选择与优化提供决策依据。

2.4.3 在涉及多个成分指标的综合分析中，本研究引入多变量统计方法包括主成分分析（PCA）和聚类分析，以降低数据维度、提取主要特征因子，并根据样本间的相似性进行分类分组。这一方法能够提高数据解释效率，并为优化模型中的相似度评价指标提供了数学支撑。

2.4.4 相关性与回归分析是本研究统计原理中的重要组成部分。通过计算皮尔逊相关系数或斯皮尔曼秩相关系数，可以定量化评估各成分指标之间的相关程度；通过多元回归分析，可以探讨成分含量、成本、质量评分等变量之间的线性或非线性关系。这一分析对于理解指标间的相互影响机制、构建优化目标函数具有重要意义。

2.4.5 在模型结果与优化方案的评估阶段，通过方差分析比较不同优化方案在关键指标上的波动程度，通过显著性检验判断优化效果是否超越原始方案的随机波动范围。这些统计方法可明确结论的科学性与可信度。

2.5 数据评分

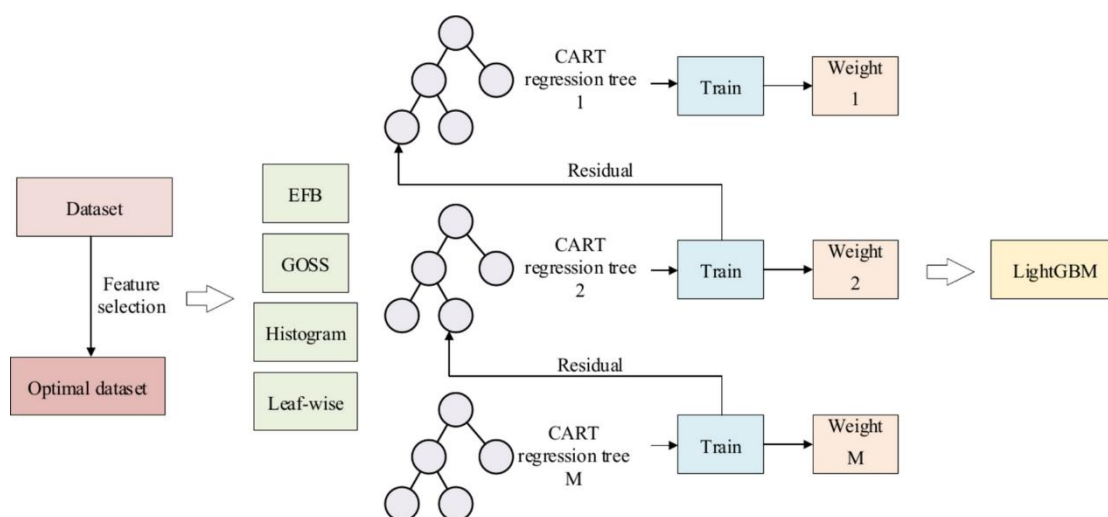
在完成了前期的数据清洗与统计分析后，系统进入了核心的质量评分阶段。本阶段旨在建立一个能够将多维化学检测指标、指纹图谱特征以及批次生产信息等综合转化为连续化质量分数的预测模型，从而为后续的优化求解模块提供可量化、可比较的目标值。传统的规则评分方法虽然在药典标准和专家经验的指导下具有较强的可解释性，但在捕捉成分之间复杂的非线性交互关系以及处理高维指纹图谱数据方面存在先天局限，难以满足实际生产中对评分精细化与差异化的需求。基于此，本研究在系统设计中引入了机器学习评分机制，将领域知识与数据驱动的建模方法有机结合，使评分既能继承专家知识体系，又能从数据中学习潜在模式，从而显著提升评价的科学性与适用性。

机器学习评分的构建过程采用“教师—学生”训练模式，其中教师模型为基于药典阈值与VIP权重计算的规

则评分体系，评分范围为 0 至 5，能够反映批次样本是否符合药典标准及其相对优劣。学生模型则基于梯度提升决策树框架 LightGBM 训练得到，其学习目标是教师模型的离散分数映射并细化为 1 至 10 的连续分值，以增加评分的分辨力。这种设计不仅保留了规则评分的可解释性，还使模型能够在训练过程中自主发现规则未明确刻画的特征关系，例如某些次要成分与整体质量之间的非线性关联。由此，机器学习评分在批次质量排序、差异化区分以及优化目标函数构建中发挥了重要作用。

在特征构建方面，系统对原始数据进行了多层次组织。首先，主要化学成分含量（如甘草苷、甘草酸等）统一以 mg/g 为单位，以消除量纲差异。其次，高维指纹图谱特征涵盖色谱峰高、峰面积及保留时间等信息，通过主成分分析提取关键信息以降低维度和噪声。此外，产地、采收年份、炮制方式等批次属性，以及原料成本与库存等经济性指标也被纳入模型。数据进入模型前，系统会自动完成数值变量归一化、类别变量编码、异常值处理等预处理步骤，并确保训练与推理阶段的一致性。对于检测成本较高导致的样本数量不足，系统引入了 CTGAN 数据扩充模块，通过条件生成方法增加样本多样性，缓解类别不平衡和高质量样本稀缺的问题，从而提升模型的泛化能力。

在模型训练环节，LightGBM 因其在表格型混合特征、非线性关系和高维稀疏数据处理方面的优势而被选为核心算法。训练过程采用严格的交叉验证策略，确保在有限样本条件下评估结果的稳定性与泛化能力。超参数选择结合网格搜索与贝叶斯优化方法，自动调整树深度、叶子节点数、学习率、子采样比例等关键参数，以在预测精度与计算效率之间取得平衡。模型性能通过均方根误差（RMSE）、平均绝对误差（MAE）、决定系数（ R^2 ）和 Spearman 等级相关系数等多项指标评估，结果表明模型在数值精度与排序一致性方面均达到了预期标准。训练完成后，模型被缓存至平台，确保在优化求解阶段调用评分函数时能够快速响应而无需重复加载或训练。



LightGBM 核心技术流程图

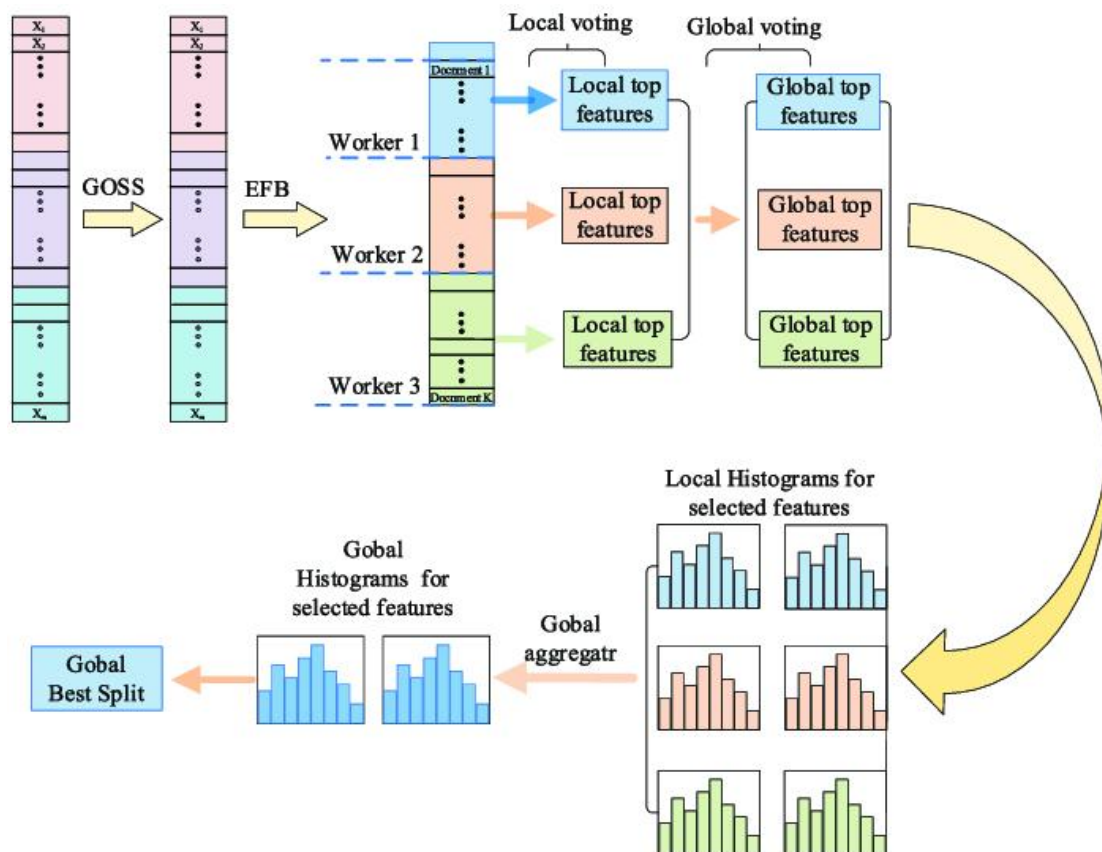


图 2.5 LightGBM 模型结构示意图

2.6 优化求解 SLSQP

在本研究中，为了最大化质量评分，并在质量与其他约束条件（如成本控制、成分比例限制等）之间找到最佳平衡，本研究采用了 SLSQP（Sequential Least Squares Quadratic Programming）单目标优化算法。SLSQP 是一种基于梯度的信息型优化方法，能够高效求解具有多个约束的非线性优化问题，特别适用于解决带有连续决策变量的优化任务。

SLSQP 的核心思想是通过二次规划近似来求解目标函数和约束条件的优化问题。在每一次迭代过程中，SLSQP 将目标函数进行二次近似，并对约束条件进行线性化，从而构造出一个可以有效求解的二次规划子问题。该子问题能够提供当前解的搜索方向，并通过线性搜索算法进行解的更新，最终得到一个符合约束的优化解。SLSQP 能够在保证计算效率的同时，保持较高的优化精度，因此特别适合处理像本研究中复杂的多批次原料优化问题。

在本研究中，优化的目标函数是机器学习评分，反映了中药配方中各批次原料组合的质量表现。目标函数计算的结果由质量评分模型给出，具体形式为最大化质量评分。为确保最终优化结果在实际生产中具有可操作性，我们在优化过程中设定了多个约束条件。主要的约束包括各批次原料的用量上下限，以保证生产过程的可行性，且总配方量需满足一定的标准。此外，还需满足成分含量的偏差不超过药典要求，并且优化后的配方在质量评分上的相似度必须达到预定的标准。

在 SLSQP 的求解过程中，首先通过经验或历史数据确定初始解。初始解需满足所有约束条件，并能够为后续的优化过程提供一个合理的起点。随后，通过对目标函数及约束条件进行梯度计算，SLSQP 在每一代迭代中通过二次规划构建一个优化子问题，并对当前解进行更新。每次迭代更新后，算法通过线搜索来确定更新步长，从而在不违反约束的前提下，逐步逼近最优解。最终，当目标函数的变化小于设定阈值，且所有约束条件都得到满足时，SLSQP 认为算法已经收敛，输出最优解。

SLSQP 算法的优势在于其计算效率和高精度。在优化过程中，SLSQP 能够快速收敛，适用于具有较多约束的连续变量优化问题。此外，SLSQP 能够灵活地处理非线性约束，特别是在中药配方优化过程中，能够准确处理复杂的质量评分与成分含量的非线性关系。因此，SLSQP 成为本研究中优化求解的核心算法之一。

在中药配方的优化应用中，SLSQP 算法能够根据机器学习评分模型和药典要求，对多批次原料的使用量进行优化。在具体应用中，SLSQP 通过精确调整每个批次的使用量，在保证质量评分的同时，还能够满足成分比例、成本控制等约束条件。例如，在甘草苷和甘草酸的配比优化问题中，SLSQP 能够根据给定的目标最大化质量评分，同时避免成分比例偏差过大，确保优化方案既符合药典要求，又具备较高的质量评分。通过这种方式，SLSQP 算法为中药配方的优化提供了可靠的解决方案，能够在实际生产中实现质量与成本的平衡。

总的来说，SLSQP 算法在本研究中表现出了较高的优化效果，不仅能够提供精确的最优解，还能够有效处理多种约束条件，确保优化结果的可行性。尽管在某些复杂的多峰问题中可能存在局限，但通过与其他优化算法的结合，SLSQP 仍然可以为中药配方的优化提供强大的技术支持。

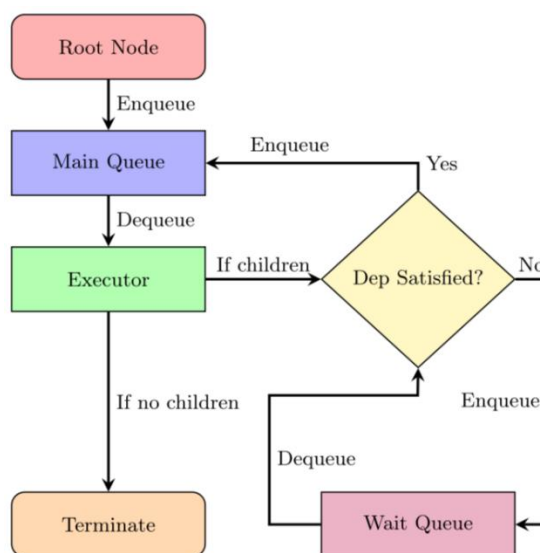


图 2.6 基本 SLSQP 流程图

2.7 优化求解 NSGA-II

在完成质量评分模型和初步的优化求解后，本研究进一步引入了 NSGA-II（Non-dominated Sorting Genetic Algorithm II）多目标优化算法，旨在实现质量评分最大化与成本控制、成分比例限制等多个优化目标之间的平

衡。NSGA-II 作为一种非支配排序的遗传算法，通过对多个目标的同时优化，能够提供一组 Pareto 最优解，帮助用户在不同目标之间做出合理的选择。与传统的单目标优化方法不同，NSGA-II 能够同时处理多个冲突的目标，并在一次运行中提供多个优化方案，满足实际生产中的复杂需求。

NSGA-II 算法的核心思想是通过非支配排序和拥挤距离计算，将解集划分为不同等级，并通过保持解的多样性来实现 Pareto 前沿的均匀分布。在 NSGA-II 中，每个个体代表一个潜在的生产配方或工艺参数组合，种群在经过多代的交叉与变异操作后，逐渐收敛到一个均匀分布的 Pareto 最优解集。这些最优解代表了不同目标之间的权衡，用户可以根据实际需求选择最适合的解决方案。

NSGA-II 的运行过程开始时通过随机生成初始种群，每个个体都代表一种原料配方的组合。初始种群的个体会根据其适应度进行非支配排序，优秀的个体会得到更高的选择概率。随后，种群通过交叉和变异操作生成新的个体，交叉操作模拟自然选择中的基因重组，变异操作则通过随机改变个体的基因来增加种群的多样性。在每一代的优化中，NSGA-II 通过非支配排序将种群中的个体进行排序，并通过拥挤距离计算确保 Pareto 前沿解的多样性，从而避免算法陷入某些局部解。

具体来说，NSGA-II 算法在优化过程中分为多个阶段：首先，种群通过非支配排序和拥挤距离计算进行初步评估；接着，通过选择、交叉和变异操作生成下一代个体；然后，新的个体与当前种群进行合并，再通过非支配排序和拥挤距离计算决定保留哪些个体，最后通过选择机制生成新一代种群。这个过程会持续进行，直到满足设定的迭代次数或收敛准则。

NSGA-II 的优势在于其能够处理多目标优化问题，特别适合那些具有多个相互冲突目标的复杂问题。在本研究中，NSGA-II 的引入，使得系统能够同时优化多个目标，如最大化质量评分、控制生产成本、保证成分比例和指纹图谱的相似度等。这种多目标优化方法能够帮助用户更全面地理解不同目标之间的权衡，提供多种可选的优化方案。与传统的单目标优化方法相比，NSGA-II 不仅能够提供全局最优解，还能生成多个 Pareto 最优解，为用户在实际生产中提供更多的灵活性。

在应用上，NSGA-II 能够通过对质量评分、原料成本、成分比例等目标的综合优化，为中药配方的设计提供合理的解决方案。例如，在优化甘草苷和甘草酸的配比问题时，NSGA-II 可以在保证质量评分的同时，根据成本限制和成分比例要求找到一组最优配方。这使得优化方案能够在满足质量要求的前提下，避免原料浪费，并有效控制生产成本。此外，NSGA-II 还能够提供多个优化方案，供用户根据实际生产情况选择最适合的方案。

综上所述，NSGA-II 在本研究中的应用展现了其强大的多目标优化能力。通过引入 NSGA-II，系统能够在多个冲突目标之间进行平衡，提供多种可行的优化方案，从而为中药配方的设计和优化提供了科学依据。NSGA-II 的引入，不仅提高了优化过程的灵活性和多样性，还确保了最终方案在质量、成本、成分比例等方面的综合最优，从而为实际生产中的决策提供了支持。

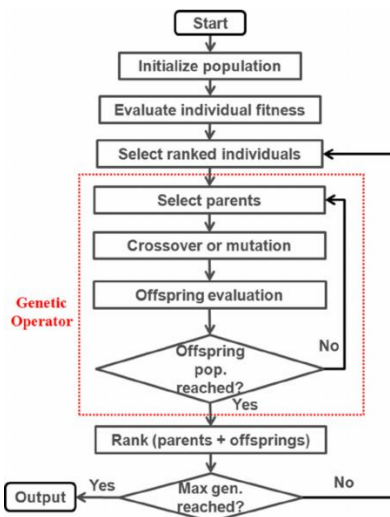


图 2.7 基本 NSGA-II 流程图

2.8 智能化决策支持与人机交互模块

为有效降低中药多组分配方优化这一复杂系统的认知负荷与操作门槛，本研究创新性地设计并集成了一套智能化决策支持与人机交互模块。该模块的核心目标是将系统从一个被动的计算工具，转变为一个能够与用户进行智能对话、提供主动式引导与专业性解读的“分析伙伴”。它旨在弥合深奥的数据科学算法与中药研发人员实际工作场景之间的鸿沟，确保用户即便不具备深厚的编程或算法背景，也能充分利用本系统的强大功能，做出科学、高效的决策。该模块的技术核心是一种创新的双模混合架构，它有机结合了大语言模型（Large Language Model, LLM）的深度语义理解与推理能力，以及本地规则式回退引擎的稳定性，在智能化水平与系统健壮性之间取得平衡。

(1) 主模式：基于动态上下文注入的远程 LLM 协同分析

在主运行模式下，系统通过 HTTPS 协议与我们进行 Prompt 训练的大语言模型服务进行 API 通信。为克服通用大模型缺乏领域知识与即时上下文的局限性，我们设计并实现了一套上下文注入式提示工程机制。该机制在用户发起交互请求时，后端逻辑自动、实时地从 Streamlit 的会话状态中抓取并整合一系列关键上下文信息。这些信息被结构化为一 段丰富的背景描述，与用户问题共同构成一个高信息密度的提示。注入的上下文信息具体包括：

数据概况：已上传数据集的摘要信息，如批次数、关键成分（如甘草苷、甘草酸）的均值与标准差、质量评分的分布区间等。

用户操作路径：用户在多步骤流程中的当前位置及已完成的关键操作（例如，已完成 CTGAN 数据扩充，已选择 NSGA-II 作为优化引擎）。

中间件与最终结果：当前界面上展示的核心数据，如特定批次的质量评分、优化算法输出的帕累托前沿解集的坐标、最终推荐配方的具体构成比例及预期成本等。

通过此机制，大语言模型能够精确感知用户所处的具体分析场景，从而提供高度相关且具有深度洞察的响应。例如，用户可提问“帕累托前沿右下角的点代表什么？”，系统不仅能解释其在多目标优化中“高成本、高质量”的普遍含义，更能结合当前数据，具体指出“该点代表的配方虽然成本较高，但其甘草酸含量和指纹图谱相似度均达到了最优水平，适合用于对质量有极致要求的特定产品线”。这种深度结合上下文的交互方式，极大地提升了分析结果的可解释性与决策支持的有效性。

(2) 备用模式：保障系统稳定性的本地规则式回退引擎

考虑到实际部署环境中网络连接可能存在的不稳定性或外部 API 服务潜在的延迟与中断风险，系统设计了完善的回退（Fallback）机制。当 API 调用因超时、网络错误或服务器无响应而失败时，系统将无缝、自动地切换至本地规则式推理引擎。该引擎本质上是一个预先构建的专家知识库，其中包含了一系列与程序状态严格绑定的“条件-响应”规则对。

2.9 程序搭建

系统整体程序的搭建依托 Streamlit 框架完成，利用其前后端一体化的特性，大幅降低了 Web 界面开发的复杂度。用户界面采用模块化卡片式布局，并结合渐变色按钮、步骤进度条与响应式设计，使整个分析流程清晰直观，用户可以按照提示逐步完成数据上传、扩充、统计、评分与优化操作。

为了确保用户在多步骤操作中的数据一致性，系统利用 Streamlit 的会话状态功能对关键变量进行持久化管理，防止页面刷新或模块切换导致的数据丢失。可视化部分除 Matplotlib 外，还引入了 Plotly 和 Seaborn 提供交互式仪表板和统计图形展示功能，显著提升了结果的可读性与分析深度。此外，系统集成了 AI 智能助手模块，能够在用户使用过程中提供上下文感知的实时解释与建议；在网络不可用或外部 API 失效的情况下，系统会自动切换至本地规则式推理引擎，以保障交互体验的连续性和稳定性。

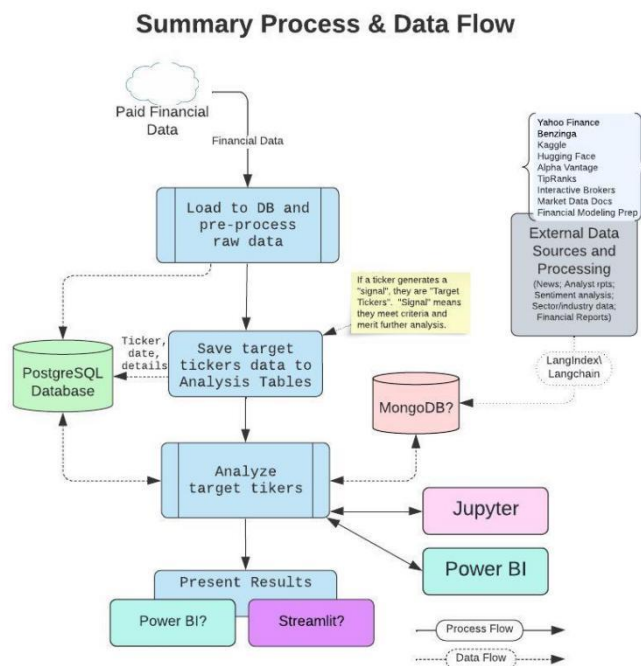


图 2.8 基于 Streamlit 框架的程序搭建架构及前后端交互逻辑

三、实验与结果展示

3.1 数据集介绍

本研究所使用的数据集来源于中药多组分均化生产过程中多个批次的检测记录，涵盖了理化指标、指纹图谱特征以及经济类信息等多方面内容，能够全面反映产品质量特征及生产条件。以中药材甘草为例，在理化指标部分，数据包含甘草苷、甘草酸等主要活性成分的含量^[16]，这些指标对药材药效和质量评估具有直接影响。指纹图谱特征通过高效液相色谱（HPLC）方法获得，以标准样本为参照计算相似度，并进一步分解为多个特征峰（F1-F11），用于更为细致的化学组成分析。经济类信息包括单位成本和库存数量，为后续在优化配比时实现质量与成本的综合平衡提供了约束条件。

在正式使用之前，数据集经过了完整的清洗与预处理。首先，对缺失值进行合理填补或删除，确保数据完整性；其次，结合箱线图与行业标准范围对异常值进行检测与剔除，排除因检测误差或其他偶然因素导致的极端数据；最后，将不同批次中存在差异的计量单位（如 mg/g 与 %）统一转换，以保证数据在后续分析与建模过程中的可比性与一致性。



图 3.1 数据集示意图

批次	相似度	严格甘草质量分数	甘草质量分数	异甘草质量分数	甘草素质量分数	异甘草素质量分数	甘草酸质量分数	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	库存
51	0.974	4.18	5.29	1.84	1.08	0.63	21.81	22.73	93.48	427.47	22.13	15.8	59.11	97.07	18.68	37.1	285.43	15.35	500
52	0.968	4.36	4.48	0.56	0.61	0.4	15.97	20.94	97.99	170.69	18.74	13.11	23.12	45.45	39.17	23.99	163.48	19.61	500
53	0.969	4.28	4.53	0.55	1.06	0.55	20.62	9.92	96.1	184.46	17.67	15.71	32.43	94.74	17.23	32.23	260.6	39	500
54	0.974	7.74	4.69	0.9	0.42	0.65	24.02	42.15	181.19	235.51	50.61	22.66	39.66	27.2	22.25	37.99	331.54	54.58	500
55	0.994	5.38	4.8	0.57	0.29	0.66	21.3	33.69	122.98	270.34	32.21	20.44	32.92	12.83	27.12	38.7	274.81	40.98	500
56	0.99	6.29	4.65	0.66	0.55	0.52	18.99	23.65	145.57	223.91	26.49	16.78	34.71	40.93	17.66	30.41	226.64	30.44	500
57	0.997	4.1	4.66	0.45	0.54	0.44	18.76	34.8	91.69	227.02	15.48	13.85	30.45	39.51	15.77	25.94	221.83	27.85	500
58	0.975	4.44	4.52	0.34	0.96	0.49	19.66	13.11	99.93	180.94	13.54	14.58	28.24	84.44	15.19	28.94	240.53	37.85	500
59	0.969	5.73	4.59	0.99	0.38	0.5	19.17	39.1	131.7	205.47	29.62	13.18	41.6	22.78	18.54	29.36	230.34	34.83	500

图 3.2 数据字段

由于原始数据的批次数量相对有限，直接用于模型训练可能会影响结果的稳定性与泛化能力。为此，引入了条件表格生成对抗网络（CTGAN）技术对数据进行扩充。该方法能够学习原始数据的分布特征，并在此基础上生成与真实样本在统计特性和相关结构上高度一致的合成数据。在扩充过程中，对生成结果进行了分布一致性检验和变量相关性分析，确保合成样本与原始数据在均值、方差及变量间相关性等方面保持一致，同时剔除扩充过程中产生的异常样本。

最终，原始数据与扩充数据合并形成了规模较原始提升约三倍的最终数据集，并按照八比二的比例划分为训练集与测试集。该数据集在样本规模、特征多样性和分布稳定性方面均能够满足后续的统计分析、评分预测与优化求解需求，为实现中药多组分均化目标提供了坚实的数据基础。

3.2 机器学习评分

为了进一步验证机器学习评分模型的有效性和预测精度，本研究通过一系列分析对比了模型预测值与标准分之间的关系，并对批次质量的分布、特征重要性及其对质量评分的影响进行了深入研究。首先，我们分析了模型预测值与标准分之间的相关性，以确定模型在整体拟合能力方面的表现。接着，针对机器学习评分的整体分布情况进行了可视化，考察了不同批次质量分数的分布特征，帮助我们理解样本间的质量差异。随后，我们对输入特征的重要性进行了排序，进一步揭示了哪些成分在质量评估中起到关键作用。最后，通过相关性热力图分析，进一步量化了各特征与机器学习评分之间的关系，为配方优化和生产工艺的调整提供了科学依据。以下将详细讨论这些分析结果，并展示它们在优化和决策过程中的实际应用价值。

为了验证模型预测值与标准分之间的关系，将二者进行对比分析，如图 3.3 所示。图中大部分样本点均分布在接近对角线的位置，呈现明显的正相关趋势，说明模型预测值与参考值之间具有较高一致性。多数批次样本的预测值与标准分差距较小，偏差控制在可接受范围内，表明模型在整体拟合能力方面表现优异。值得注意的是，个别高分段样本点出现了较大偏离，这与其化学成分比例及指纹图谱特征的特殊性有关，提示模型在少

数特殊样本上的泛化性能仍有进一步优化空间。

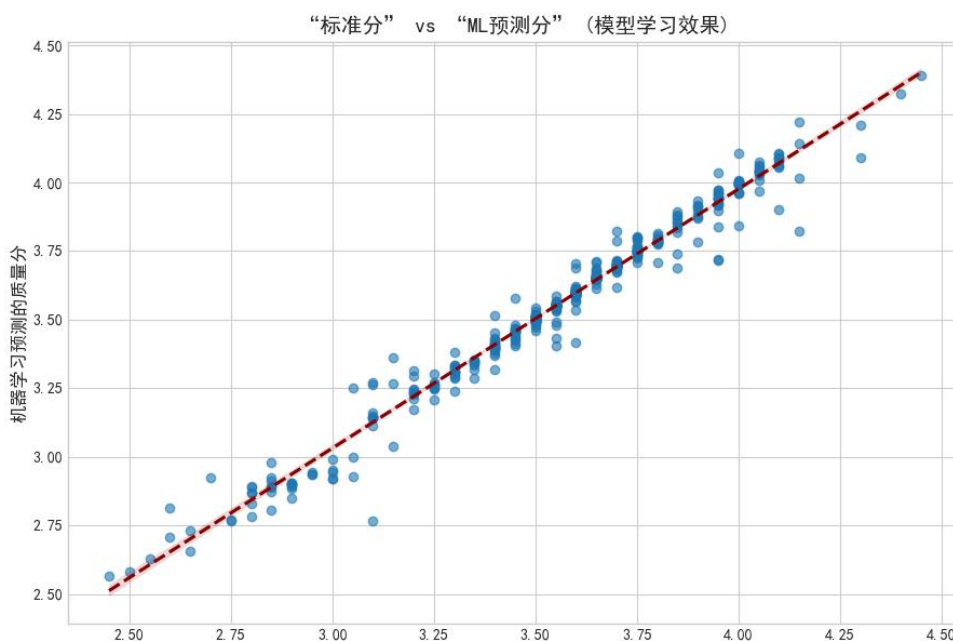


图 3.3 标准分与机器学习预测评分的对比关系

为考察整体评分分布特征，对所有批次的 ML 评分进行了直方图与核密度估计可视化，如图 3.4 所示。结果显示，预测得分主要集中在 3.0 至 3.8 之间，分布曲线呈单峰为主并带有轻微双峰特征，反映出批次间质量差异相对集中。尾部存在少量低分样本，需要在后续生产和优化中重点关注，而高分段样本虽然数量有限，但可能具有较高经济与品牌价值，因此也应在优化策略中予以保留和强化。

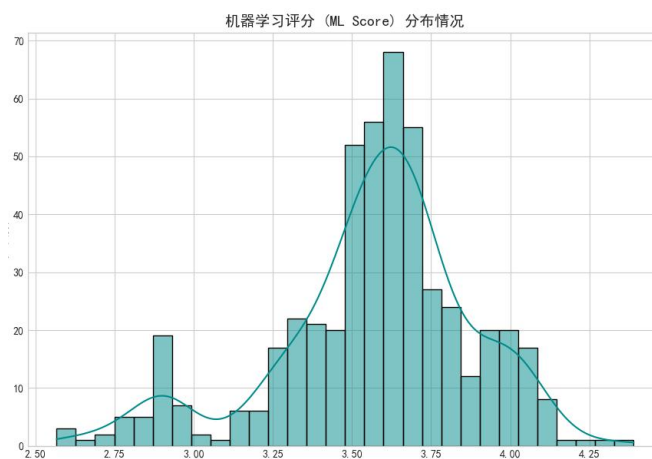


图 3.4 机器学习评分分布情况

在可解释性分析中，对输入特征的重要性进行了排序，如图 3.5 所示。“甘草酸 1”在模型中占据绝对主导地位，重要性显著高于其他特征，符合其在药理活性及质量标准中的核心地位。“相似度”指标排名第二，进一步凸显指纹图谱一致性在质量评估中的作用；“浸膏量”位居第三，对评分结果亦有明显贡献。其他特征的影响相对较小，但可能通过与关键特征的交互作用间接影响评分。

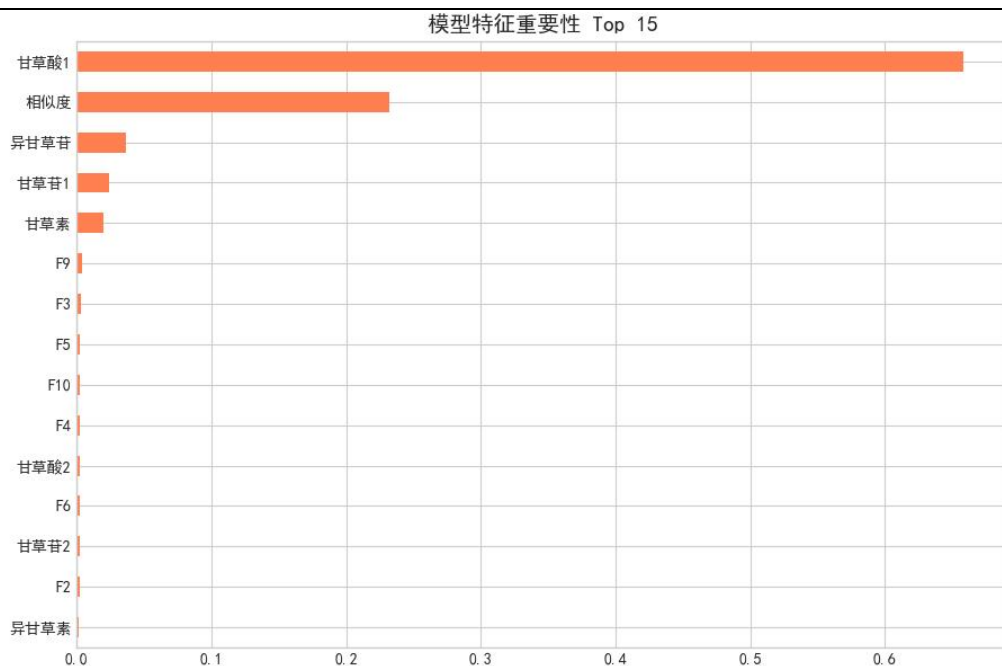


图 3.5 模型特征重要性 (Top 15)

最后，为更直观地理解各特征与质量评分之间的关系，绘制了相关性热力图（如图 3.6 所示）。从图中可以看出，“甘草酸 1”与 ML 评分的相关系数高达 0.67，呈显著正相关，意味着该成分含量的增加通常会带来评分的提升。“相似度”与评分的相关系数为 0.42，进一步验证了化学指纹图谱一致性在质量评估中的作用。其他成分与评分的直接相关性相对较弱，但这些指标可能通过与关键变量的交互作用间接影响评分结果。因此，在优化配方与生产工艺时，既要重点关注直接相关性高的指标，也应综合考虑变量间的相互作用，以实现评分结果的整体优化。



图 3.6 关键指标与机器学习评分的相关性热力图

综上所述，机器学习评分模型在预测精度、可解释性及应用价值方面均表现出色，不仅能够准确量化批次质量，还能揭示关键影响因素，为配方优化、生产过程控制及质量一致性管理提供可靠的技术支撑。更为重要

的是，机器学习评分在系统架构中与优化求解模块紧密衔接，为后续多目标优化提供了连续、稳定且可微的目标函数，使整个决策流程形成了从数据获取到优化输出的完整闭环。

3.3 批次质量分析仪表盘

该仪表盘在一个 2x3 的网格中同时展示了六个关键图表，涵盖了质量、成本、库存和成分等多个维度。首先通过质量分数的直方图给出一个全局视野，让用户对整体批次的质量水平和一致性有一个宏观的认识。随后，通过 Top 10 批次的条形图聚焦于表现最优的个体，为寻找“标杆”提供了明确的目标。接着，分析深入到质量的驱动因素层面：两个散点图中，“甘草苷 vs 甘草酸”图揭示了关键成分与质量的内在关联，而“成本 vs 质量”图则引入了经济效益视角，通过回归趋势线直观地评估“性价比”，帮助决策者在成本和质量之间做出权衡。最后，通过库存分布和指纹图谱相似度这两个直方图，将分析扩展到供应链和生产一致性的维度，评估原料的稳定性和充足性。下文将以甘草酸和甘草苷数据为例进行图表解读

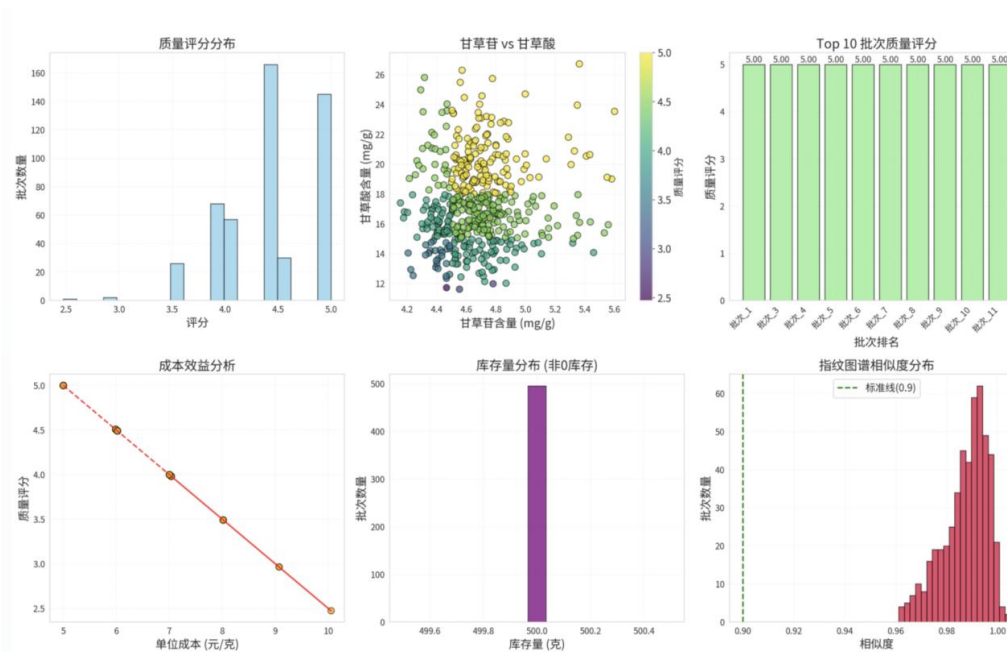


图 3.7 批次质量分析仪表盘

3.3.1 质量评分分布图直方图（Quality Score Distribution）

质量评分分布直方图帮助用户了解整体批次的质量水平和分布情况。质量评分从 2.5 到 5.0，分为五个区间：2.5-3.0, 3.0-3.5, 3.5-4.0, 4.0-4.5, 4.5-5.0。纵轴表示每个质量评分区间内的批次数。分布呈现出右偏（正偏）形态，即大部分批次的质量评分集中在较高的评分区间（4.0-5.0）。大部分批次的质量评分集中在 4.5-5.0 区间，表明整体质量水平较高。评分分布较为分散，从 2.5 到 5.0 都有分布，但主要集中在高分区间。分布的右偏形态表明大部分批次的质量评分较高，少数批次评分较低。在实际应用的质量控制中，这种分布可以帮助企业

了解其产品质量的总体水平，识别出需要改进的低评分批次。高评分批次较多，可能表明产品在市场上具有较高的竞争力。对于评分较低的批次，企业可以进一步分析原因，采取改进措施。

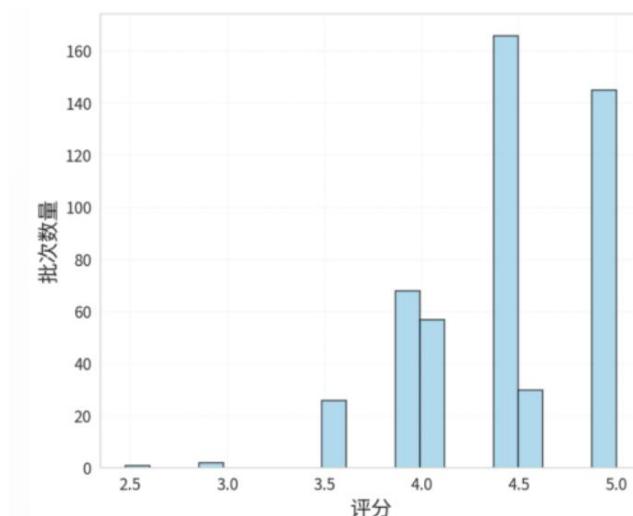


图 3.8 质量评分分布直方图

3.3.2 甘草苷和甘草酸含量质量评分散点图

核心指标相关性散点图揭示两个关键成分（如甘草酸和甘草苷）之间的关系，颜色深浅代表质量评分，有助于发现“高质高量”的批次。横轴表示甘草苷的含量，单位是毫克每克（mg/g）。纵轴表示甘草酸的含量，单位也是毫克每克（mg/g）。每个点代表一个批次，其位置由该批次的甘草酸和甘草酸含量决定。数据点在图中呈现出一定的分布模式，显示出两种成分含量之间的相关性。大多数数据点集中在甘草苷含量 4.5 到 5.5 mg/g，甘草酸含量 14 到 26 mg/g 的区域内。颜色深浅代表质量评分，可以帮助我们发现哪些批次具有较高的质量和成分含量。颜色从深紫色到黄色，表示质量评分从低到高。可以观察到，颜色较浅（黄色和浅绿色）的点主要集中在图的上部和右侧，表明这些批次的质量评分较高，同时甘草酸和甘草酸的含量也较高。颜色较深（深绿色和紫色）的点主要集中在图的下部和左侧，表明这些批次的质量评分较低，同时甘草酸和甘草酸的含量也较低。从相关性分析可以看出，甘草苷和甘草酸的含量之间存在一定的正相关关系，即甘草酸含量较高的批次，甘草酸含量也倾向于较高。质量评分与两种成分的含量之间也存在一定的正相关关系，即两种成分含量较高的批次，质量评分也倾向于较高。实际应用中，这种分析可以帮助企业识别出哪些批次的质量较高，从而进行优先选择或进一步分析。通过了解成分含量与质量评分之间的关系，企业可以优化生产过程，提高产品质量。高评分批次较多，可能表明产品在市场上具有较高的竞争力。

3.3.3 前 10 名批次质量评分图

直观展示最优批次，并精确标注其分数。横轴表示批次的排名，从 0 到 9。纵轴表示质量评分，范围从 0 到 5。每个条形的高度代表对应批次的质量评分。所有条形的高度相同，均为 5.00，表示这些批次的质量评分都是满分。所有前 10 个批次的质量评分都是 5.00，显示出这些批次在质量上的高度一致性和优越性。所有批

次的质量评分都集中在最高分，表明这些批次在质量控制或生产过程中表现出色。由于所有批次的评分相同，离散程度（如标准差）为零，显示出没有变异性。这种高度一致的高质量评分表明生产过程可能非常稳定，或者质量控制措施非常有效。虽然这些批次的质量评分相同，但可以进一步分析这些批次的成分含量、生产条件等，以了解是什么因素导致了这种高质量的一致性。建议持续监控这些批次的生产过程，以确保质量的持续稳定。可以考虑将这些批次作为基准，与其他批次进行比较，找出影响质量的关键因素。

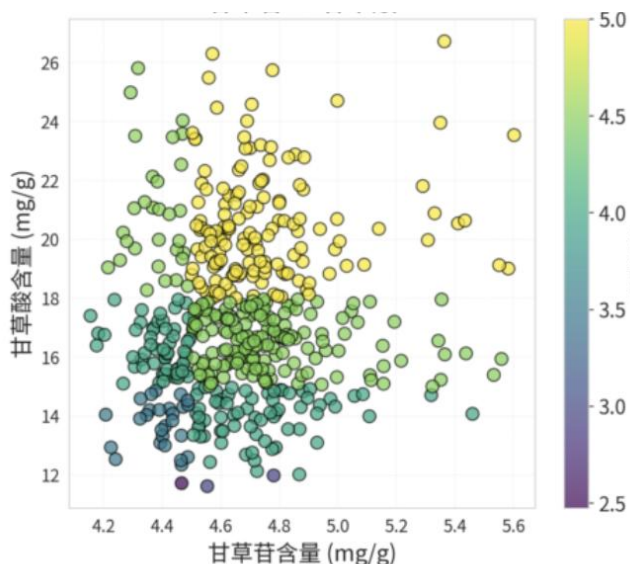


图 3.9 甘草苷和甘草酸含量质量评分分布图

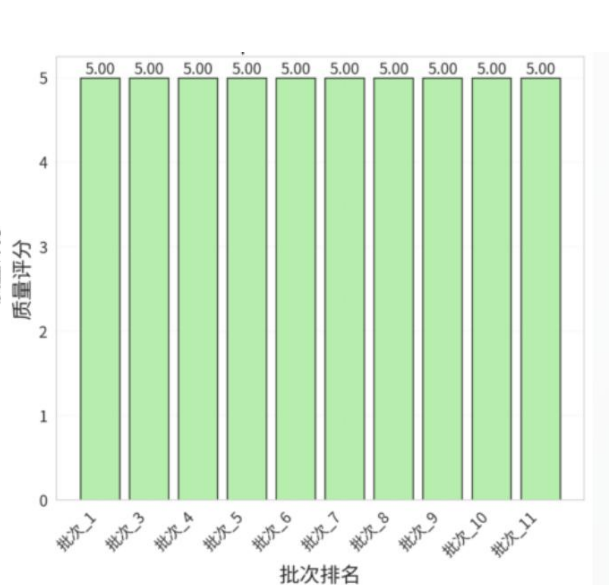


图 3.10 前 10 名批次质量评分图

3.3.4 成本效益回归分析

成本效益分析(散点图将成本与质量置于同一平面，并增加了线性回归趋势线，为用户决策提供了“性价比”的视角。横轴表示产品的单位成本，从 5 元/克到 10 元/克。纵轴表示产品的质量评分，从 2.5 到 5.0。每个点代表一个批次，其位置由该批次的单位成本和质量评分决定。数据点通过颜色或形状可能还表示其他变量（如批次类型），但在此图中未明确标出。图中的红色虚线是线性回归趋势线，它显示了单位成本与质量评分之间的线性关系。趋势线的斜率为负，表明随着单位成本的增加，质量评分呈下降趋势。从趋势线可以看出，单位成本与质量评分之间存在负相关关系。这意味着在所分析的数据集中，成本较低的产品往往具有较高的质量评分，而成本较高的产品则倾向于具有较低的质量评分。这种关系为用户提供了一个“性价比”的视角，即在成本较低的情况下，可能获得较高的质量评分，从而实现更好的性价比。企业可以根据这一分析结果调整成本结构，以实现更高的性价比。进一步分析建议考虑引入其他可能影响质量评分的变量，如生产方法、原材料质量等，进行多变量回归分析。库存与相似度分布直方图帮助用户评估原料的充足性和一致性。

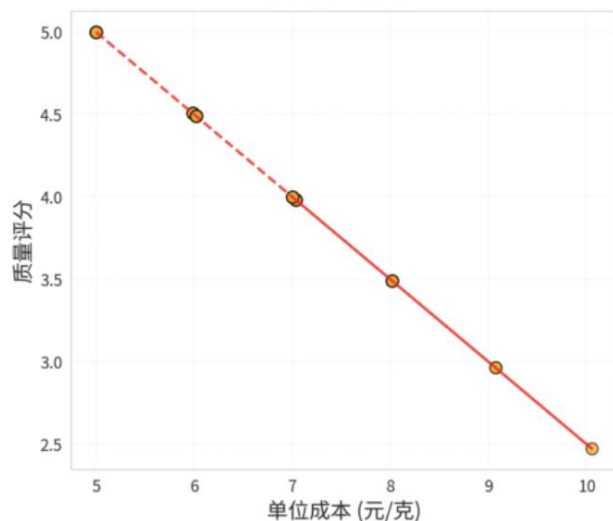


图 3.11 成本效益回归分析

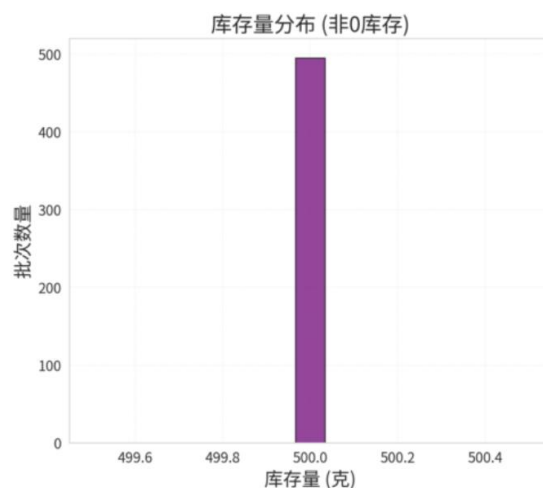


图 3.12 库存分布直方图

3.3.5 指纹图谱相似度分布

横轴表示指纹图谱的相似度分数，范围从 0.90 到 1.00。相似度分数越接近 1.00，表示两个样本的化学成分模式越相似。纵轴表示具有特定相似度分数的批次数量。直方图显示了不同相似度分数的批次数量分布。大多数批次的相似度分数集中在 0.98 到 1.00 之间，表明这些批次的化学成分模式非常相似。图中绿色虚线表示标准线，位于 0.90 处。表示一个阈值，用于区分可接受的相似度分数和不可接受的相似度分数。几乎所有的批次相似度分数都高于这个标准线，表明这些批次的相似度都达到了可接受的水平。大多数批次的相似度分数都非常高（接近 1.00），这表明这些批次的化学成分模式非常一致，质量控制良好。数据分布的集中趋势表明大多数批次的相似度分数都集中在高值区域，显示出高度的一致性。数据分布的离散程度较小，表明批次间的相似度分数差异不大，这进一步支持了高质量控制的结论。指纹图谱相似度分布图可以帮助企业确保其产品的化学成分一致性，从而保证产品质量。通过识别相似度分数较低的批次，企业可以进一步分析原因，优化生产过程。在药品和其他受监管的产品中，可以帮助企业满足监管要求，确保产品的一致性和可追溯性。

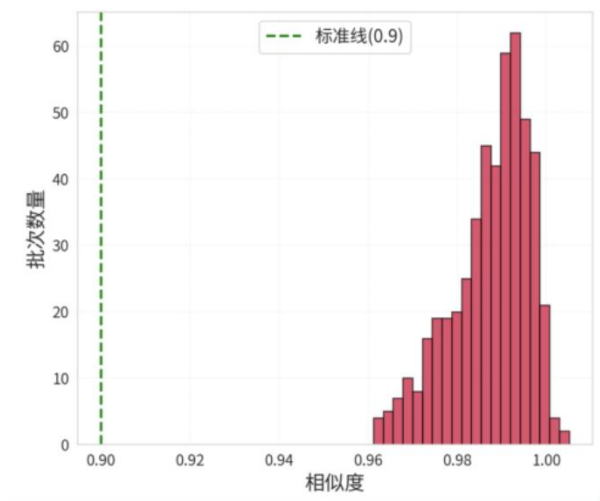


图 3.13 指纹图谱相似度分布

3.4 成分含量深度分析统计图:

3.4.1 箱线图和小提琴图

两种成分(Glycyrrhizin 和 Glycyrrhizic Acid)的箱线图和小提琴图展示了各成分含量的分布、中位数、四分位数和离群点，信息密度远高于简单的条形图。集中趋势比较中，中位数位置可以判断哪种成分整体含量更高，箱体的位置反映数据的集中区间。离散程度分析中，箱体高度(IQR，四分位距)显示中间 50%数据的离散程度，须线长度反映数据的整体波动范围。小提琴图的宽度变化展示不同含量值的概率密度。从分布形态中，箱线图中中位数与箱体的相对位置可初步判断偏态，小提琴图的形状能更精确显示分布是否对称、单峰或多峰。

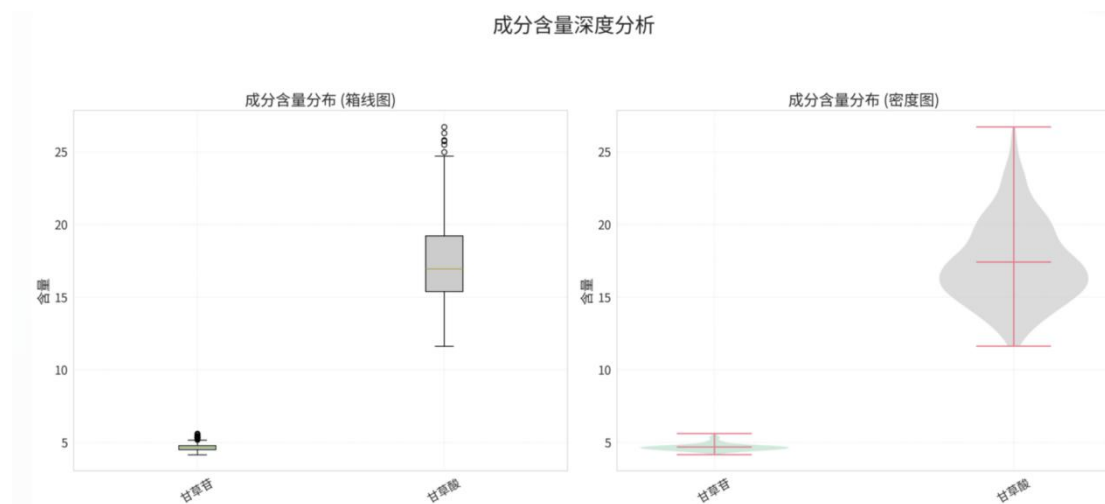


图 3.14 成分含量深度分析箱线图（左）和小提琴图（右）

3.4.2 相关性热力图

直观地展示了所有成分指标之间的两两相关性，并用数值标注了相关系数，有助于发现成分间的协同或拮抗关系。图中展示了两种成分：Glycyrrhizin（甘草苷）和 Glycyrrhizic Acid（甘草酸）的相关性系数。热力图使用颜色编码来表示相关性系数的大小。红色表示正相关，蓝色表示负相关，颜色越深，相关性越强。右侧双变量关系图中，横轴表示 Glycyrrhizin 的含量，纵轴表示 Glycyrrhizic Acid 的含量。每个点代表一个样本，点的颜色表示质量评分。颜色条显示了质量评分的范围，从 2.5 到 5.0，颜色从紫色到黄色。颜色的变化显示了质量评分的分布情况，颜色较深的点（紫色）表示质量评分较低，颜色较浅的点（黄色）表示质量评分较高。质量评分较高的样本（黄色点）分布在整个图中，没有集中在某个特定的区域。

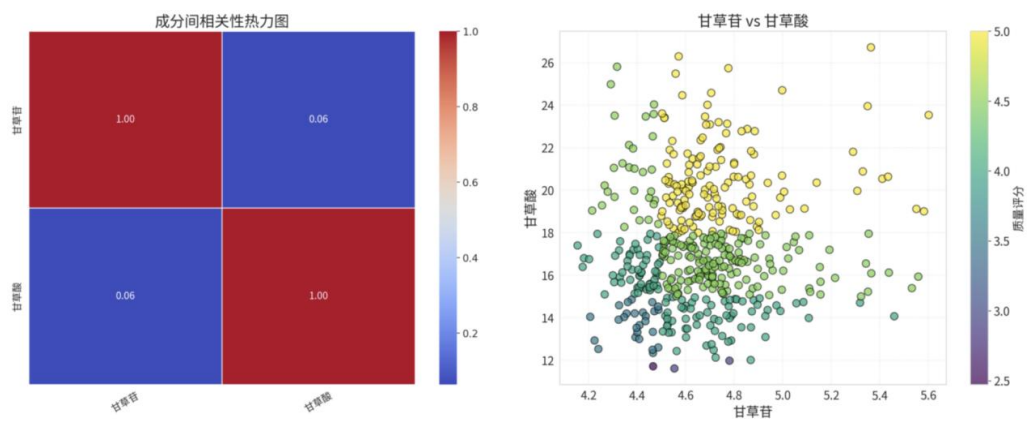


图 3.14 相关性热力图

这些仪表板的加入，使得软件不再是一个“黑箱”优化器，而是成为一个透明、可交互的数据分析平台，用户可以在充分理解数据特性的基础上，再进行后续优化决策。

3.5 可视化引擎对均一化结果的优化

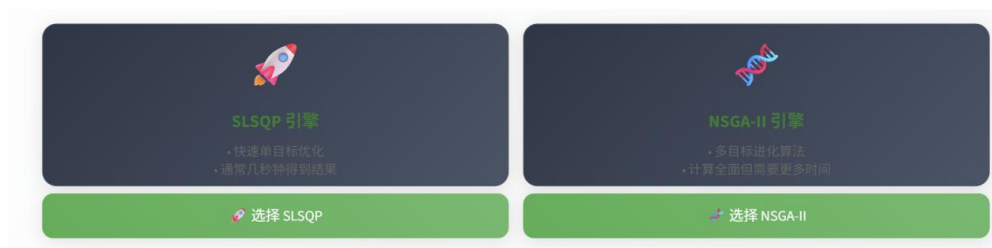


图 3.15 均一化引擎

3.6 程序展示（程序使用方法界面介绍）

步骤 1、上传前数据准备

在上传数据前，请确保数据文件采用 Excel (.xlsx) 或 CSV (.csv) 格式，并使用 UTF-8 编码。以甘草数据为例，建议的列与命名如下：批次标识可使用 Batch_ID，亦可不设专列而以行索引代替；指纹相似度为 0~1 的数值列；成本列建议使用自定义列；库存信息建议提供“库存量(克)”或“预设库存量”中的其一。单位规范方面，成分以 mg/g 计，相似度取值范围为 0~1，成本为元/克，库存以克为单位。各列并非强制全部提供，但缺失会限制相应功能，例如缺少成本列将无法进行成本分析。

批次	相似度	蔗糖	甘草苷质量分数	甘草苷质量分数	异甘草苷质量分数	甘草素质量分数	异甘草素质量分数	甘草酸质量分数	F1	F2	F3	F4	F5	F6
S1	0.974	4.18	5.29	1.84	1.08	0.63	21.81	22.73	93.48	427.47	22.13	15.8	59.11	
S2	0.988	4.36	4.48	0.58	0.61	0.4	15.97	20.64	97.98	170.69	18.74	13.11	33.12	
S3	0.969	4.28	4.53	0.55	1.06	0.55	20.62	9.92	96.1	184.46	17.67	15.71	32.42	
S4	0.974	7.74	4.69	0.9	0.42	0.65	24.02	42.15	181.19	235.51	50.61	22.66	39.66	
S5	0.994	5.38	4.8	0.57	0.29	0.66	21.3	33.59	122.98	270.34	32.21	20.44	32.92	
S6	0.99	6.29	4.65	0.66	0.55	0.52	18.99	23.65	145.57	222.91	26.49	16.78	34.71	
S7	0.997	4.1	4.66	0.45	0.54	0.44	18.76	34.8	91.69	227.02	15.48	13.85	30.45	
S8	0.975	4.44	4.52	0.34	0.96	0.49	19.66	13.11	99.93	180.54	13.54	14.58	28.24	
S9	0.989	5.73	4.59	0.99	0.38	0.5	19.17	39.1	131.7	205.47	29.82	13.18	41.6	
S10	0.991	4.94	4.72	1.19	1.02	0.51	19.84	29.77	112.33	245.08	29.85	18.27	45.6	
S11	0.988	3.2	4.82	1.14	0.38	0.51	18.82	40.96	69.55	276.9	15.87	9.79	44.57	
S12	0.979	4.18	5.03	1.08	0.84	0.49	19.1	20.69	93.6	344.59	18.37	13.31	43.33	
S13	0.987	5.48	5.35	1.84	0.99	0.73	23.96	26.87	125.5	446.63	35.13	20.44	58.98	
S14	0.973	3.9	5.09	1.09	0.86	0.51	19.14	21.54	86.75	362.97	17.05	12.72	43.61	
S15	0.994	4.79	4.69	0.74	0.46	0.52	20.91	40.06	108.62	237.58	28.23	12.86	36.48	

图 3.16 上传前数据准备

步骤 2、进入“智能优化”

应用首页提供四张功能卡片，分别引导不同的工作流。点击“智能优化”可进入完整的优化流程，包括数据上传、列映射、批次筛选、参数设置与求解执行；“快速计算”用于按目标产量与称量精度估算理论可用的最大批次数；“可视化分析”可一键生成总览六联图与成分分析四联图，便于快速洞察数据分布与相关性；“精准配比”提供精度等级说明与配比计算器，支持生产称量与误差控制。点击“智能优化”卡片后，会展开操作说明与“快速操作”按钮；继续点击“立即开始”，应用即进入上传流程。任意卡片在被点击时，其对应内容会在页面下方展开显示；如需收起，可点击右上角“✕ 关闭”。

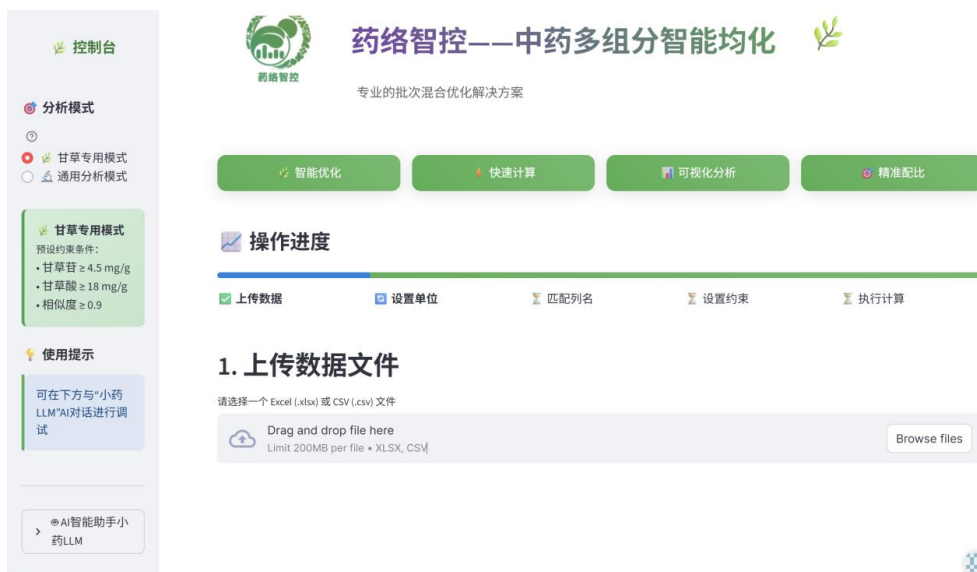


图 3.17 上传数据后进行智能优化

步骤 3、单位选择

系统在内部统一采用 毫克/克 (mg/g) 作为标准计量单位进行所有的质量评估与约束判断，若单位不一致，将直接导致优化失败或产生完全错误的配比方案。具体来说，系统能够智能识别并处理两种常见的数据录入格式：第一种是 百分比 (%) 格式，通常在表格中表现为小数形式，例如，单元格中的数值 0.025 即被系统解读为 2.5% 的含量；第二种是更为直接的毫克/克 (mg/g) 格式，即单元格中的数值（如 25）直接代表其 mg/g 的含量值。

2. 设置数据单位

请选择您上传文件中，核心成分含量所使用的单位。

单位选择

- ☐ 百分比 (%) - 例如 2.5% 表示为 0.025
- ☒ 毫克/克 (mg/g) - 例如 2.5% 表示为 25

确认单位并继续

图 3.18 单位选择

步骤 4、设置列映射

将列名映射到系统字段：甘草苷 → 映射为 **gg_g**；甘草酸 → 映射为 **ga_g**；相似度 → 映射为 **sim**；成本 → 映射为 **cost**（若无则使用默认 模拟成本 或忽略成本分析）；库存 → 建议映射 库存量 (克)，如缺失可退化为 预设库存量 或置空（优化会用大数兜底），确认映射后继续。

3. 匹配数据列

	批次	相似度	蔗糖甘草苷质量分数	甘草苷质量分数	异甘草苷质量分数	甘草素质量分数	异甘草素质量分数	甘草酸质量分数	F1	F2	F3
0	S1	0.974	4.18	5.29	1.84	1.08	0.63	21.81	22.73	93.48	427.47
1	S2	0.988	4.36	4.48	0.58	0.61	0.4	15.97	20.64	97.98	170.69
2	S3	0.969	4.28	4.53	0.55	1.06	0.55	20.62	9.92	96.1	184.46
3	S4	0.974	7.74	4.69	0.9	0.42	0.65	24.02	42.15	181.19	235.51
4	S5	0.994	5.38	4.8	0.57	0.29	0.66	21.3	33.59	122.98	270.34

核心指标 (必选)

甘草苷含量 (必选)

甘草苷质量分数

甘草酸含量 (必选)

甘草酸质量分数

相似度 (必选)

相似度

参考与管理指标 (可选)

异甘草素含量 (参考)

--

异甘草苷含量 (参考)

--

甘草素含量 (参考)

--

管理与成本信息 (可选)

单位成本列

--

② 批次编号列 (可选)

--

库存量列

库存

指纹图谱特征列 (F1, F2 ...)

F1 × F2 × F3 × F4 × F5 × F6 × F7 × F8 × F9 × F10 × F11 ×

确认列匹配并开始处理

图 3.19 设置列映射

步骤 5、批次筛选与实时预览

勾选/筛选将参与混合的批次（本演示选择 3 个批次）。预览区显示预期质量评分（选中批次平均（或混合）评分），预期总成本（需映射到单位成本列），库存充足率（库存>0 的批次占比），建议优先保留“质量评分高且成本中等、库存充足”的批次。



图 3.20 批次筛选与实时预览

步骤 6、设置优化参数与约束（侧边栏或参数面板）

如设置目标产量: 填写 1000.0 g; 关键成分最低标准（甘草）: 甘草苷 ≥ 4.5 mg/g, 甘草酸 ≥ 18.0 mg/g



图 3.21 设置优化参数

步骤 7、引擎选择:

SLSQP: 单目标快速求解, 输出唯一方案（本演示选此）。NSGA-II: 多目标求解, 输出帕累托解集（可作为进阶步骤）。选择批次数限制: 如需固定参与批次数, 可填 3。成本/偏差/相似度的权重或启用开关。精度等级用于称量指导（标准/精密/超精密）。默认启用库存约束, 任何批次用量不得超过库存。

4. 选择批次并执行优化



SLSQP 引擎

- 快速单目标优化
- 通常几秒钟得到结果

选择 SLSQP



NSGA-II 引擎

- 多目标进化算法
- 计算全面但需要更多时间

选择 NSGA-II

查看总数据分析仪表板

选择显示方式: ②

- ☒ 智能检测 (推荐)
- ☐ 仅英文标签
- ☐ 强制中文标签

生成数据分析报告

图 3.22 引擎选择

步骤 8: 执行优化 (SLSQP, 引擎演示)

按“开始优化”按钮；几秒内完成均一化过程。求解成功后显示配比结果表，即每个批次的“比例% / 用量(g) / 评分 / 成本”。结果解读要点，饼图用于显示每个批次在 1000 g 中的占比。量柱图用于显示各批次称量克数（总和应为 1000 g）。达标对比用于显示“标准线 vs 实际达成”（甘草苷、甘草酸皆应 \geq 标准）。成本-质量散点：标注被采用批次，辅助解释“性价比”选择。评分分布对比图用于显示所有候选 vs 最终采用批次的直方图对比。库存使用率用于显示各被采用批次的库存占用百分比（ $>80\%$ 提醒供应风险）。若失败，则自动检查“总量=1000 g”的可行性、库存是否不足、是否存在某成分无法达标、列映射是否错误。

已选择 SLSQP 引擎 - 单目标快速优化

优化参数设置

设置混合后产品总量 (克) ① - + ☒ 启用指纹图谱一致性约束 ②

最低指纹图谱相似度要求 ②

含量目标优化

☒ 启用含量目标引导 ②

目标甘草苷质量分数含量 (mg/g) ③ - + 目标甘草酸质量分数含量 (mg/g) ④ - +

✔ 约束指标达标情况

	指标名称	预期值	标准要求	是否达标
0	甘草苷质量分数	4.8969	≥ 4.5	✔
1	甘草酸质量分数	20.0011	≥ 18	✔
2	相似度	0.9765	≥ 0.9	✔

🎯 目标含量达成情况

	指标名称	实际值	目标值	偏差百分比
0	甘草苷质量分数	4.8969	5.0000	2.06%
1	甘草酸质量分数	20.0011	20.0000	0.01%

图 3.23 执行引擎优化

步骤 9、进阶 NSGA-II 多目标演示（可选）

适合需要“含量偏差最小化 + 相似度最大化 +（可选）成本”的多目标平衡。

常见参数（界面可能在侧栏或弹窗）：种群规模（如 80~200），迭代代数（如 200~800）

交叉/变异概率（0.7/0.1 参考），变异强度（小数，比如 0.05），固定参与批次数（如 3）

是否剔除极端解（remove_extremes），运行后可见帕累托前沿演化与实时指标，结束后第一前沿为最优解集散点。自动逐一查看候选方案的配比与达标性，选择“最合适”的可行解进入生产。

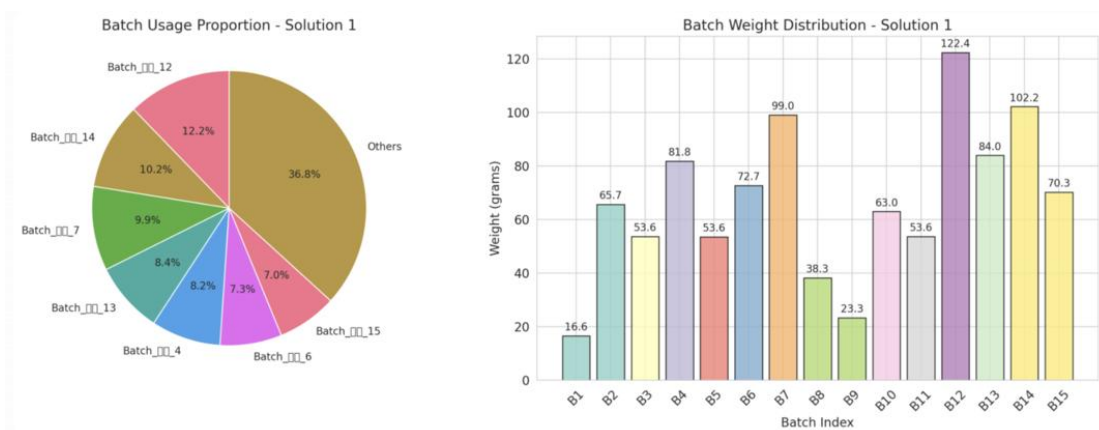


图 3.24 NSGA-II 多目标配比优化结果图

步骤 10、结果导出

打开“结果导出”区域（优化完成后显示）。导出 Excel 中包含处理好的原始数据、优化配方表（实际 >0.001 g 才会出现）、统计描述；文件名含时间戳；导出图表（PNG）包括重新生成并保存主分析图，适合报告插图。；生成 PDF 包含：AI 执行摘要（若配置 API Key）、推荐配方表等。AI 摘要（可选）需要在侧栏填入 github_api_key（无则跳过，PDF 仍可生成）。



图 3.25 结果导出

3.7 实验结果分析

在完成数据预处理、统计分析及机器学习质量评分模型的构建之后，本研究进入了配方优化求解阶段。该阶段旨在在既定的质量评价体系下，综合批次原料的化学成分、质量评分、成本信息及库存数据，生成满足不同生产目标的最优配伍方案。为实现这一目标，系统同时引入了 SLSQP（Sequential Least Squares Programming）单目标优化算法与 NSGA-II（Non-dominated Sorting Genetic Algorithm II）多目标进化优化算法，从而兼顾高效性与多样性，满足不同的决策需求。

在单目标优化情境下，SLSQP 以最大化机器学习评分为核心优化目标，通过快速迭代求解，在满足批次数量、用量范围及库存限制等约束条件的同时，锁定评分最高的原料组合。其特点是收敛速度快、计算效率高，适合需要即时生成可执行配方的场景。在多目标优化情境下，NSGA-II 则能够在多个相互冲突的优化指标之间寻找平衡点。本研究中，该方法主要用于在“加权含量偏差最小化”和“化学指纹相似度最大化”两个目标下，生成覆盖不同权衡关系的帕累托最优解集。算法通过种群进化与非支配排序，探索目标空间的多样性，使得生成的解在不同质量与一致性要求下均具备较高可行性与参考价值。

综合两种方法的优势，系统既能在需要快速出具方案时利用 SLSQP 获得单一高质量解，又能在需要权衡多目标的情况下通过 NSGA-II 生成丰富的候选解集。这样不仅满足了实际生产中的多样化需求，也提升了决策的科学性与灵活性。最终，优化结果通过可视化图表呈现，包括帕累托前沿分布、批次使用比例、用量分布、库存使用率及成本效益分析等，为用户提供了直观、量化的参考依据。

3.7.1 SLSQP 实验结果图

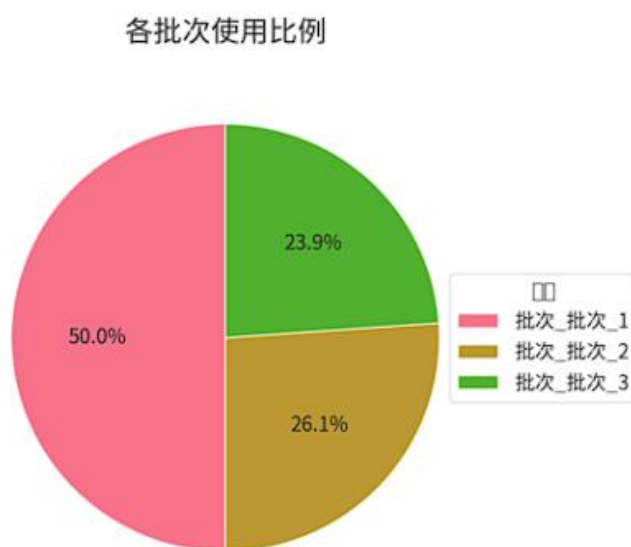


图 3.26 各批次使用比例分布

该图为饼图，用于展示不同批次在整体使用量中所占的比例分布情况。图中以三种颜色分别代表批次 1（粉色）、批次 2（棕色）和批次 3（绿色），并在各扇形区域内标注了对应的百分比数值。饼图的面积比例与数值成正比，便于直观比较各批次的使用占比。本示范结果中，批次 1 的扇形面积最大，占比达到 50.0%，表明其在整体使用量中占据主导地位；批次 2 的占比为 26.1%，位列第二；批次 3 的占比最小，仅为 23.9%。这种可视化方式有助于快速判断资源投入的集中程度以及各批次在配方或生产过程中的重要性。

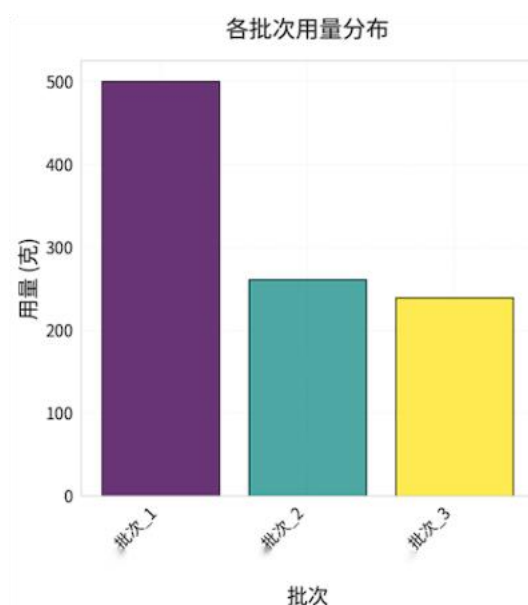


图 3.27 批次用量分布图

该图为柱状图，用于展示不同批次在实际生产或配方方案中对应的用量分布。横坐标为批次类别，包括批次 1、批次 2 和批次 3；纵坐标表示用量，单位为克（g）。每个柱形的高度直接对应该批次的实际使用量，便于直观比较不同批次在总用量中的占比情况。从图中可以看到，批次 1 的柱形最高，用量 500 克，在三者中最

多；批次 2 的用量约为 280 克，位列第二；批次 3 的用量最低，约为 240 克。该图形直观反映了不同批次的原料投放差异，有助于分析配方结构、评估批次间的投入均衡性以及制定后续的批次选择与优化策略。

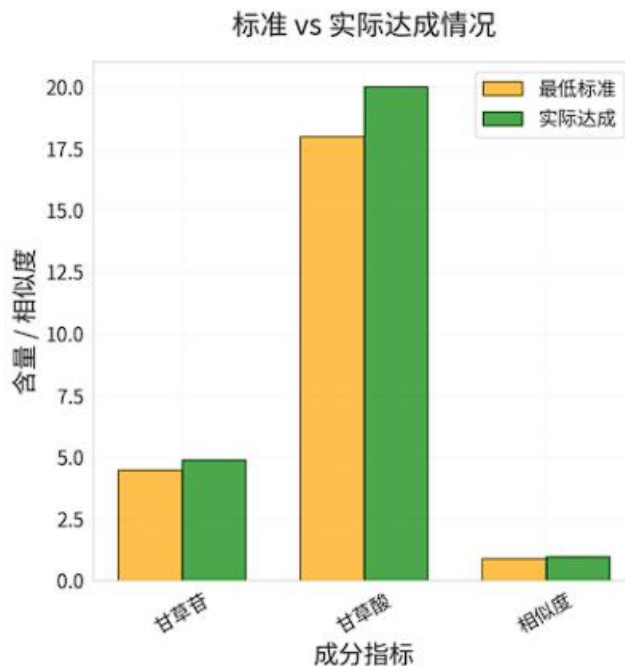


图 3.28 达标柱状图

该图为分组柱状图，用于比较不同成分指标在“最低标准”与“实际达成”两种情况下的数值表现。横坐标为成分指标，包括甘草苷、甘草酸以及相似度；纵坐标为对应的含量或相似度数值，具体单位取决于指标性质，其中含量以 mg/g 表示，相似度为无量纲比例。图中每个成分指标对应两根并列的柱形，橙色代表“最低标准”，绿色代表“实际达成”值。通过并列对比，可以直观评估实际生产批次在各个指标上是否满足或超越既定标准。例如，若绿色柱形高于橙色柱形，则表明该成分的实际达成值超出最低标准；若低于标准，则可能存在质量控制风险。此类图形有助于快速识别成分达标情况及潜在改进方向，为质量检测与优化调整提供直观依据。

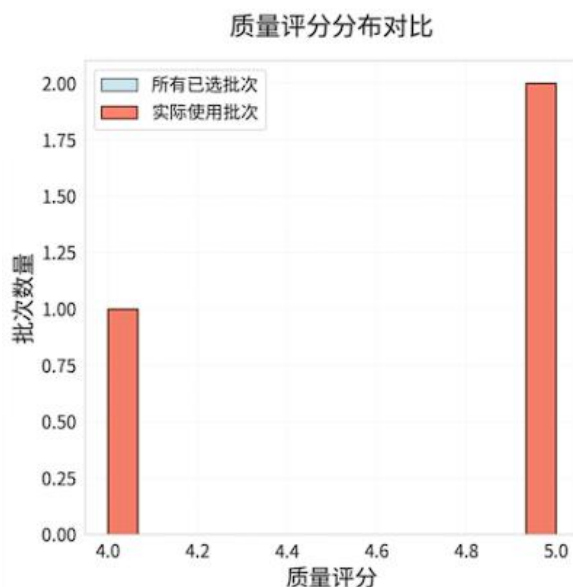


图 3.29 质量评分分布图

该图为分组柱状图，用于对比“所有已选批次”与“实际使用批次”在不同质量评分区间内的分布情况。横坐标为质量评分，纵坐标为对应评分区间内的批次数。蓝色柱形表示所有已选批次在各评分区间的数量分布，红色柱形表示最终实际投入使用批次的数量分布。通过该图，可以直观地观察不同评分等级的批次在选择与使用环节的差异情况。例如，如果红色柱形高度明显高于蓝色柱形，说明该评分等级的批次在实际使用中占比更高；反之则可能表明该评分等级的批次在使用环节中被淘汰或减少。此类可视化方式有助于评估质量评分与批次使用决策之间的关系，为优化原料选择策略和质量控制提供数据支撑。

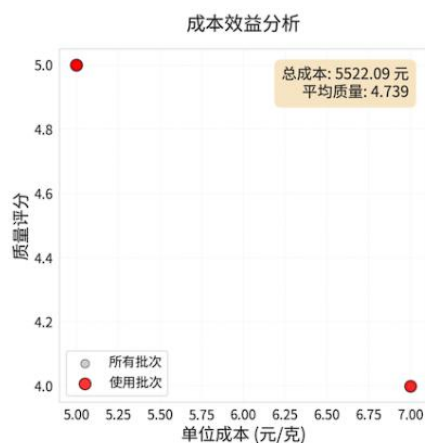


图 3.30 成本效益图

该图为散点图，展示了不同批次在单位成本与质量评分之间的关系，并结合了总体成本与平均质量的综合信息。横坐标表示单位成本（元/克），纵坐标表示质量评分，图中每一个点代表一个批次。灰色圆点表示所有批次的分布位置，红色圆点表示最终被实际选用的批次，便于区分筛选前后的差异。图右上方的标注框汇总了实际使用批次的总成本与平均质量评分，为整体性评估提供直观数据支撑。从图中可以观察到，不同批次在单位成本与质量评分上的分布情况及其相互关系，例如是否存在高评分伴随高成本，或者低成本批次质量表现优异的情况。此类可视化分析有助于综合评估成本与质量之间的权衡关系，为优化采购与生产决策提供量化依据。

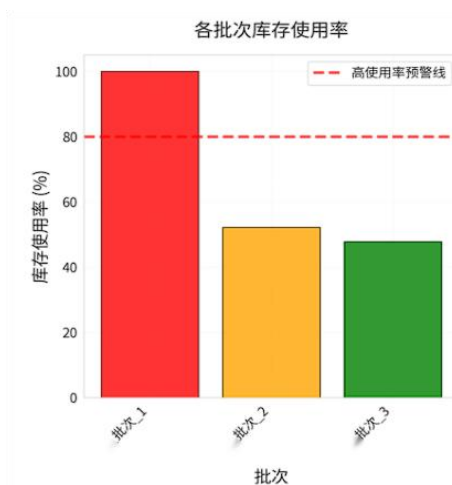


图 3.31 库存使用率分析图

该图为柱状图，用于展示各批次原料在库存中的使用率情况。横坐标表示不同批次的编号（批次 1、批次 2、批次 3），纵坐标表示库存使用率（%）。不同颜色的柱体对应不同批次，便于区分批次间的使用水平差异。为防止意外情况，我们设计了一条预警线，代表高使用率预警线（80%），用于标识超过该阈值的批次一次用量会过多消耗库存。从图中可见，批次 1 的使用率已达到 100%，明显超过预警值，而批次 2 与批次 3 的使用率则相对较低，分别位于 70% 左右和 50% 左右。该可视化结果能够直观反映库存消耗状况，为库存管理与采购补货决策提供参考依据。

3.7.2 NSGA-II 实验结果图

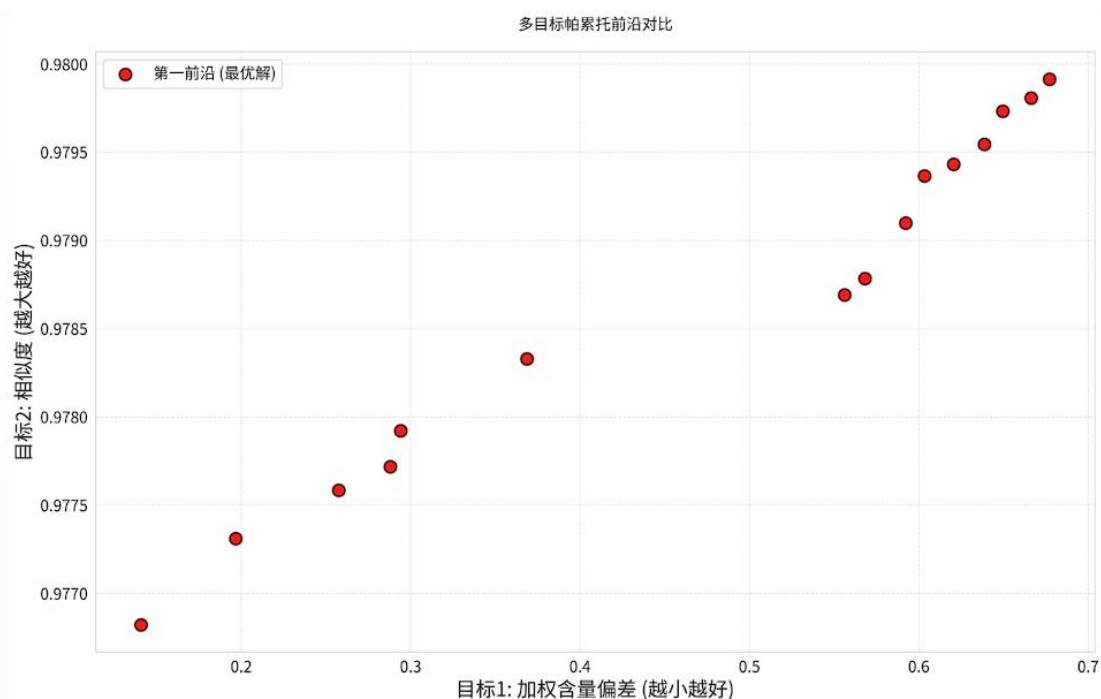


图 3.32 NSGA 帕累托图

该图展示了基于 NSGA-II 算法所得的多目标优化结果，反映了两个评价指标在优化过程中的分布关系。横坐标为“目标 1：加权含量偏差”，数值越小表示样品中各主要成分的含量越接近目标配方的设定值；纵坐标为“目标 2：相似度”，取值范围接近 1，数值越大表示样品的化学指纹图谱与参考标准的相似程度越高。图中的每一个红色点代表一次可行解，其在坐标平面上的位置由该解在两个目标上的评价价值决定。整体来看，这些点分布形成了从左下方向右上方延伸的趋势，呈现出典型的多目标优化结果的帕累托前沿特征，即在减少加权含量偏差与提高指纹图谱相似度之间形成了不同程度的权衡关系。该可视化后续为从众多可行解中筛选最优方案提供了直观依据。

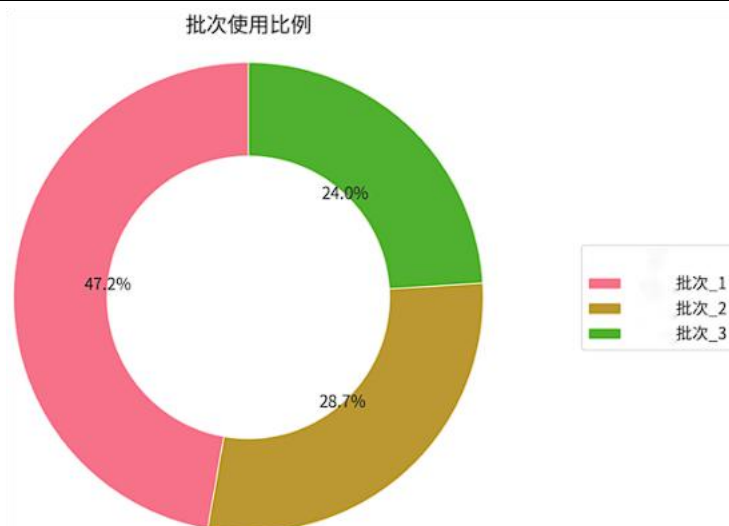


图 3.33 批次比例环状图

该图采用环形图形式展示了优化方案中不同批次原料的使用比例分布情况。图中不同颜色的扇区分别对应不同批次的原料，其中粉色代表“批次 1”、棕色代表“批次 2”、绿色代表“批次 3”，扇区面积大小与对应批次在最终配方中所占比例成正比。图上标注的百分比为各批次在总用量中的占比，例如“批次 1”占比 47.2%，表示该批次在优化配方中的使用量接近总量的一半；“批次 2”占比 28.7%；“批次 3”占比 24.0%。该可视化直观反映了不同批次原料在配方构成中的相对重要性，为分析原料来源的组合策略及其对质量和成本的影响提供了基础信息。

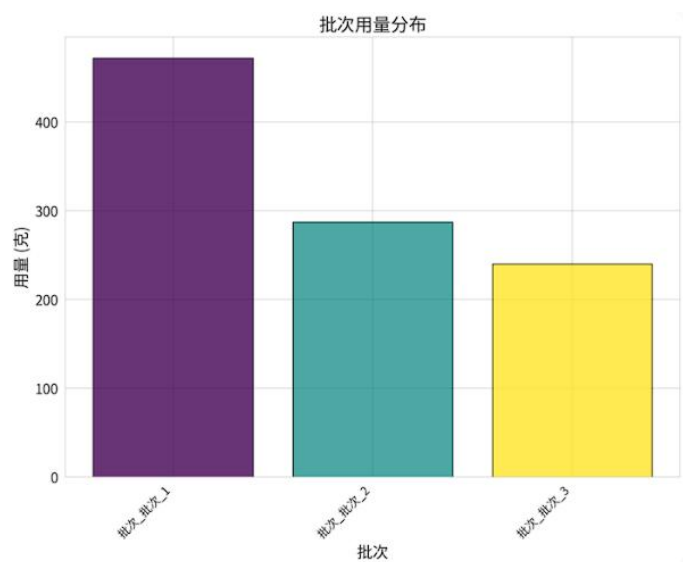


图 3.34 批次用量柱状图

该图以柱状图形式展示了优化方案中不同批次原料的实际使用量分布情况。横坐标为批次编号，包括“批次 1”“批次 2”和“批次 3”；纵坐标为对应批次的用量，单位为克（g）。柱子的高度与批次使用量成正比，从而直观体现了各批次在配方中投入的绝对数量。从图形可见，“批次 1”的柱高最高，用量超过 400 克，在所有批次中占据最大份额；“批次 2”用量约为 280 克；“批次 3”用量最低，约为 240 克。该可视化有助于分析配方构

成中各批次原料的实际投入规模，为后续评估其对质量评分和成本控制的贡献提供了基础数据支持。

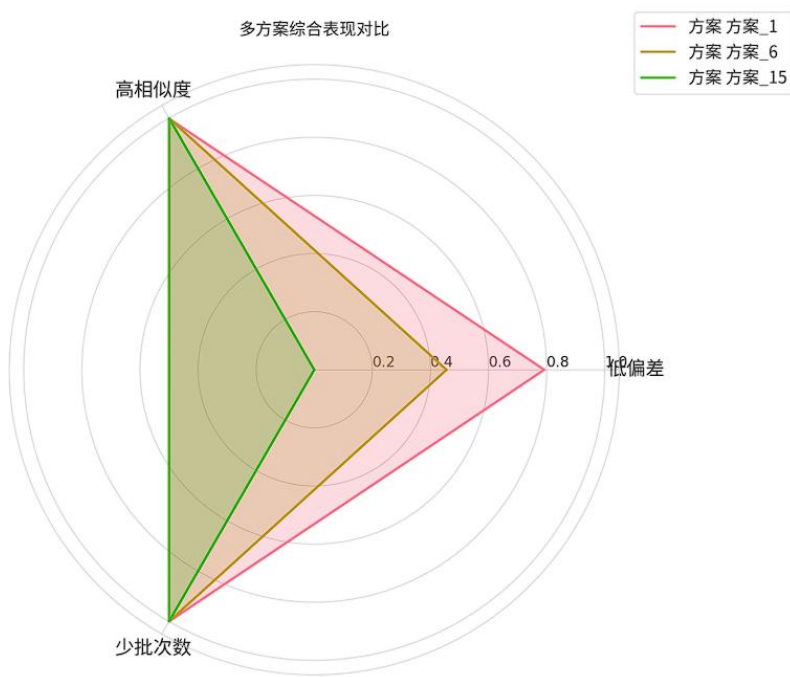


图 3.35 表现雷达图

该图为雷达图，用于展示多种优化方案在不同评价指标上的综合表现对比。图中选取了三个关键评价维度：相似度、配方含量偏差以及批次数量，分别位于雷达图的三个顶点。每条彩色折线代表一个具体方案，其中红色代表方案 1，棕色代表方案 6，绿色代表方案 15。折线的轮廓与填充面积直观反映了各方案在不同指标上的相对表现，数值越接近雷达图外圈，说明该方案在该指标上的表现越优。在“批次数量”指标上，方案 15 与方案 1 的表现相似且较高，而方案 6 略低；在“含量偏差”指标上，不同方案差异较为明显，方案 1 的数值最高，方案 15 最低，反映了在成分含量稳定性方面的差异。该图可为综合选取最优方案提供多维度参考，有助于用户在不同目标之间进行权衡。

四、总结

本系统的程序搭建以 Streamlit 框架为核心，构建了轻量化、可扩展的交互式用户界面，实现了数据扩充、统计分析、模型评分与优化求解等模块的无缝集成。通过前端浏览器与后端服务端的实时通信机制，用户可直观地完成数据上传、参数调整及结果查看，显著提升了操作的便捷性与结果的可视化程度。该架构不仅保证了系统运行的稳定性与交互响应速度，还为后续功能扩展和模块替换提供了良好的灵活性与兼容性，为多组分中药均化计算平台的落地应用奠定了坚实的技术基础。

五、参考文献

- [1] Liu Y, Cao M, Chen Y, Hu Y, Wang Y, Luo G. [Study of blending method for the extracts of herbal plants]. Se Pu. 2006 Mar;24(2):117-21. Chinese. PMID: 16830456IF: 1.0 Q4 .
- [2] Yang M, Zhou Y, Chen J, Yu M, Shi X, Gu X. [Application of genetic algorithm in blending technology for extractions of Cortex Fraxini]. Zhongguo Zhong Yao Za Zhi. 2009 Oct;34(20):2594-8. Chinese. PMID: 20069899.
- [3] Yan BJ, Qu HB. An approach to optimize the batch mixing process for improving the quality consistency of the products made from traditional Chinese medicines. J Zhejiang Univ Sci B. 2013 Nov;14(11):1041-8. doi: 10.1631/jzus.B1300059. PMID: 24190450; PMCID: PMC3829653.
- [4] 阳长明,陈霞,赵巍,等.基于源头控制的中药制剂质量研究[J].中草药,2021,52(02):321-326.
- [5] 阳长明,陈霞,马秀璟,等.从国家药品抽检探索性研究谈中药制剂质量控制[J].中草药,2023,54(01):1-7.
- [6] 钱忠直, 于江泳, 梁鑫淼. 中药生产过程中产品质量均一化要求的科学性、可操作性与风险评析[J]. 中国食品药品监管, 2025, 0(6): 4-11
- [7] 唐雪芳,齐飞宇,王团结,等.中药生产过程智能质量控制专利技术进展[J].中国中药杂志,2023,48(12):3190-3198
- [8] 肖伟,张欣,徐芳芳,等.以功效物质为质量目标的中药生产过程质量控制研究与实践[J].南京中医药大学学报,2024,40(10):1030-1038.
- [9] 李玲,张彤.现代科技赋能中药全产业链的传承与创新[J].中成药,2024,46(10):3540-3544.
- [10] ICH. (2009). Q8(R2): Pharmaceutical Development. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
- [11] 国家药监局药审中心,《中药均一化研究技术指导原则（试行）》(2020年11月)
- [12] 韩丽, 张芳, 张定堃, 等. 中药制剂含量均匀度的测定方法初探[J]. 中草药, 2014, 45(15): 2125-2131.
- [13] 江振作, 王跃飞. 基于“药材基原-物质基础-质量标志物-质控方法”层级递进的中药质量标准模式研究[J]. 中草药, 2016, 47(23): 4127-4133.
- [14] 陈聪聪, 秦雪梅, 杜冠华, 周玉枝. 中药有效组分“配伍艺术”的研究基础——组分配比研究策略[J]. 药学报, 2019, 54(5): 808-817.
- [15] 省盼盼, 罗苏秦, 尹利辉. 过程分析技术在药品生产过程中的应用[J]. 药物分析杂志, 2018, 38(5): 748-757.
- [16] 张育贵,宋沁洁,李国峰,等. 基于高效液相色谱指纹图谱、模糊物元法、多指标定量与化学计量学的市售甘草饮片质量评价[J]. 中华中医药杂志,2022,37(5):2584-2592.