大语言模型驱动的色谱工艺设计与优化平台：银杏叶提取物的案例应用

Zhilong Tang[1], Shaohua Wu[1], Xinyan Zhao[2], Yu Wang[1,3], Xingchu Gong[1,2,3,4,*]1. 浙江大学药学院药学信息学研究所，中国杭州 310058；2. 天津中医药大学中医药研究所，中国天津 301617；3. 浙江大学金华研究院，中国金华 321016；4. 浙江大学中药现代化国家重点实验室，中国杭州 310058* 通讯作者：gongxingchu@zju.edu.cn；电话：+86-571-88208427

摘要：色谱分离技术因其高效性，已广泛应用于制药、化工和食品行业。然而，传统依赖人工的色谱工艺开发面临诸多挑战，例如依赖专家经验、开发周期长且劳动强度大。本研究提出了 ChromR，这是一个由大语言模型（LLM）驱动的色谱工艺设计与优化平台。该平台整合了专为色谱领域训练的大语言模型 ChromLLM、多智能体系统以及自动化色谱实验装置。多智能体系统由四个智能体组成：领域知识问答、实验设计、实验执行和数据分析。ChromR 能够自动完成整个工作流程，包括初始工艺参数推荐、实验设计、自动化执行、数据分析和多目标优化。借助 ChromR，可有效降低对专家知识的依赖，同时显著减少人力投入和开发时间。本研究选取银杏叶提取物（EGBL）的色谱纯化作为

案例研究。ChromR 在一周内成功开发出满足多项目标的色谱工艺，包括馏分质量和生产效率，将开发时间缩短至传统范式所需时间的约七分之一。由此建立了一种智能化、自动化且具有普适性的新型色谱工艺开发范式。

关键词：色谱工艺；大语言模型（LLM）；大语言模型驱动实验室；工艺设计；多目标优化

1. 引言

色谱分离技术具有高效等优势，已广泛应用于化工、制药和食品行业的分离纯化过程[1-5]。色谱过程十分复杂，目标化合物的分离性能通常由多个参数共同决定，且参数间耦合性强，因此参数优化具有挑战性[6, 7]。在色谱工艺开发中，仅实现令人满意的分离是不够的，还必须考虑成本[8, 9]、绿色化学原则[10, 11]，和效率[12]。这些多目标要求使色谱工艺开发面临更多挑战。

传统的色谱工艺开发通常包括文献检索、固定相筛选[13, 14]，、单因素逐次试验[15, 16]，、实验设计研究[17, 18]，、基于模型的优化以及参数验证[19, 20]。即使对于经验丰富的研究人员而言，开发出令人满意的色谱工艺仍然耗时且劳动强度大。因此，有必要开发一种更高效的色谱工艺开发方法。

近年来，自驱动实验室[21-25]，已在化学[26-28]，生物学28]，和材料科学[29]，等研究领域崭露头角，并已被证明

是加速科学发现[30]的有效工具。自驱动实验室通过自动化实验生成高质量、高密度的实验数据。数据会传输至人工智能决策系统进行分析，该系统随后更新预测模型，并提出最有希望的下一步实验方案，形成闭环迭代，直至实现预先设定的研究目标[31]。贝叶斯优化（BO）常被用作人工智能决策系统中的实验规划算法[31, 32]。然而，贝叶斯优化对初始条件[33], 较为敏感，而现有的非信息性采样方法（例如随机采样或拉丁超立方抽样[34]）往往忽略文献中已有的先验知识，导致资源浪费。

自ChatGPT所代表的大语言模型发布以来，其在自然语言处理方面的出色表现促使科学家探索其在化学、生物学和材料科学等研究领域的应用潜力[35]，，例如文献数据挖掘[36-38]，、化学与材料发现[39, 40], 、蛋白质与基因序列设计[26, 41], 以及实验室自动化[42, 43]。大语言模型在海量语料库上训练后，能够学习数据中的复杂语言规则、语义和上下文关联，使其在接收到研究人员提供的复杂或非标准化输入后，能够基于先验知识进行合理推理并返回适当的响应[44]。来自各个领域的许多研究人员已经利用领域专用语料库，训练了专门服务于各自领域的大语言模型[45-47]。

目前，大语言模型在执行复杂数学计算方面仍然存在困难。然而，通过开发由大语言模型担任中央协调器、整合多种工具的智能体，可以解决这一局限[48, 49]。借助工具增强型大语言模型，可以根据需求通过预定义的API接口直接操作现实世界中的硬件，从而显著减少人工劳动并加速科学发现[50]。

本研究开发了一款由大语言模型驱动的色谱工艺设计与优化平台ChromR，旨在帮助研究人员加快色谱工艺开发

。银杏银杏叶提取物（EGBL）具有多种药理活性，包括改善认知功能[51],抗氧化作用[52], 和抗炎特性[53], 等。因此，本研究选取 EGBL 作为案例，展示 ChromR 的工作流程及其在色谱工艺设计与优化方面的表现。

2. 方法

2.1 ChromR概述

ChromR由两个部分组成：一个多智能体系统和一台自动化色谱实验装置。多智能体系统包含四个智能体，分别负责色谱领域知识问答、实验设计、实验执行和数据分析。在本研究中，我们首次训练了一个专门针对色谱领域的大语言模型（ChromLLM），并基于该模型开发了一个智能体（智能体A），用于初步色谱工艺设计。ChromR的详细架构如

图1。

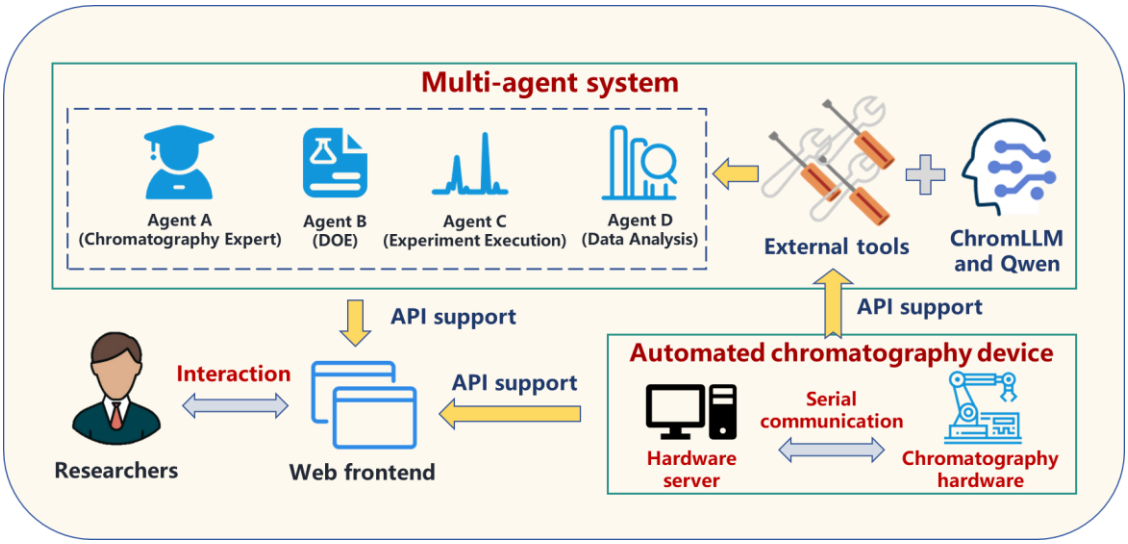


图1。 ChromR示意图。

2.2 ChromLLM的构建

2.2.1 数据收集

在传统的数据挖掘工作流程中，正则表达式是功能强大且广泛使用的工具。然而，这种方法耗费大量人力，并要求研究人员在编程、计算机科学和数据分析方面具备丰富的专业知识[54, 55]。在本研究中，为了最大限度地从文献中进行信息抽取，同时尽量减少人工操作，我们建立了专门用于文献数据挖掘的基于大语言模型的工作流程。

用于后预训练的数据属于未经标注的纯文本。无标签数据挖掘的具体流程如图2A所示。我们首先从数据库中检索并下载了大量与色谱相关的文献，随后使用支持最多 1000 万个 token 的 Qwen-Long 模型进行数据挖掘。在阿里云百炼平台（https://bailian.console.aliyun.com）上进行了批量推理。批量推理之前，我们运用提示工程优化 Qwen-Long 的响应，以确保提取数据的质量和准确性。推理结果中常包含不可见符号、个人信息、乱码和冗长的重复内容。因此，我们采用正则表达式和基于相似度的去重等数据清洗技术，以确保训练数据的质量。

用于获取微调标注数据的具体流程如图2B所示。我们首先选取了几篇具有代表性的文献，在优化提示工程后，使用通义千问-Qwen-Long模型进行批量推理，以获取监督微调数据。将数据拆分为问答对后，采用了类似的数据清洗流程。值得注意的是，数据清洗后，人类专家进一步审查并修订了数据，以确保其具有高质量并符合人类偏好。

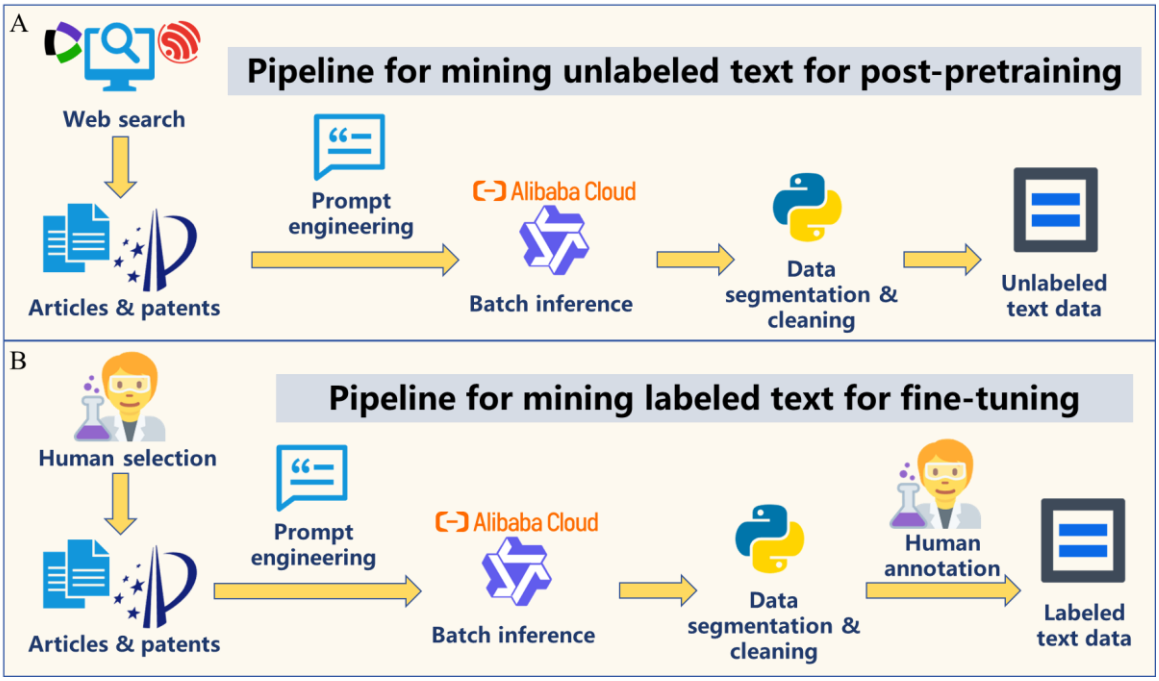


图2。 数据挖掘流程。（A）用于后预训练的无标签数据收集，预训练。（B）用于监督微调的有标签数据收集。

2.2.2 后预训练

我们使用第2.1.1节所述的流程，获得了一个总量超过7500万个词元的数据集。我们选用Qwen2.5-14B作为基础模型。模型训练在华为云（https://console.huaweicloud.com/modelarts）上进行。学习率设置为0.00002，训练轮数设置为1。训练损失曲线如图S1所示。由于损失已收敛至接近零，我们得出结论：模型已经充分学习了数据中蕴含的知识。

2.2.3 监督微调

经过人工检查，我们共获得 6,801 条适用于微调的数据，其中大多数属于多轮对话。本研究使用 Llama-Factory 0.9.2.dev0，在一台配备四块 NVIDIA GeForce RTX 3090 GPU 的计算机上，对经过后续预训练的大语言模型进行低秩适配（LoRA）微调。学习率设为 0.00005，训练轮数设为 1。监督微调过程中的损失曲线如图 S2 所示。损失收敛至约 0.8，表明模型已学习了数据中的大部分知识。

2.3 多智能体系统

该多智能体系统由四个智能体组成，分别负责色谱领域知识问答、实验设计、实验执行和数据分析。系统可以接收研究人员输入的提示和上传的文件，并根据研究人员的问题或需求，智能判断需要激活哪些智能体。随后，被激活的智能体会根据输入提示智能调用集成工具并生成响应。该系统通过提示工程，使用通义千问系列大语言模型[56]进行意图识别和规划。多智能体系统部署在 Dify 1.5.1 平台的本地环境中。

2.3.1 智能体A：领域知识问答

智能体A由三个主要组件构成：ChromLLM、知识库和文献检索模块。ChromLLM能够回答大多数专业问题。为提高回答的准确性，本研究建立了一个本地知识库，涵盖多项标准和法规，以支持检索增强生成（RAG）。由大语言模型驱动的文献检索模块可连接至包括 arXiv、PubMed 和 Scopus 在内的学术数据库。该模块能够根据研究人员的输入生成搜索查询，检索相关文献并返回最相关的数据，从而提升回答的先进性和时效性。接收三个组件的输出后，通义千问系列中的高性能大语言模型会综合这些信息，并向研究人员给出最终答案。

2.3.2 智能体B：实验设计

智能体B能够从用户提示中智能识别关键信息，并调用 Python执行器生成实验设计。智能体B能够生成常见的实验设计，包括确定性筛选设计（DSD）、Box-Behnken设计、中心复合设计等。

2.3.3 Agent C：实验执行

Agent C 解读研究人员的自然语言描述，生成硬件控制代码，并执行单项实验或一系列实验（例如完整的实验设计）。Agent C 集成了可直接操作物理硬件的 API接口，能够根据对研究人员需求的分析自动生成控制命令，并将其发送至硬件，从而加快研究构想在实验实践中的实现。

2.3.4 Agent D：数据分析

Agent D 集成了多种数据分析工具，包括逐步回归、设计空间计算[57], 以及基于 NSGA-II/III算法的多目标优化。在逐步回归过程中，它不仅会生成多项式模型，还会分析模型并提出改进建议。2.4 自动化色谱实验装置

尽管市面上有许多自动化色谱仪器，但其控制协议和操作方法大多属于专有技术，因此难以实现所有色谱单元与计算机控制器之间全面的双向通信[31, 58]。为与多智能体系统集成，并实现由大语言模型驱动的色谱过程设计与优化，我们开发了一套定制的自动化色谱实验装置，如图3所示（照片见图 S3）。整个系统由硬件服务器通过串行通信进行控制。两台蠕动泵 P1–P2（DIPump550-B146，上海科沫尔流体科技有限公司）分别安装在色谱柱的前后。阀门 V1–V9（DN8球阀，建凯强阀门经营部）由计算机控制，可灵活切换不同的色谱工艺阶段。该装置配备了三种在线传感器：pH计（LN-ISEP20L05）、电导率仪（LN-ISEP10L05）和氧化还原电位仪（LN-ISRC20L05），均购自上海蓝昌自动化科技有限公司；以及两套来自深圳市南方仪器科技有限公司的在线光谱仪，包括一台紫外光谱仪（紫外，SPEC-CDS350）

和一台近红外光谱仪（近红外，NIR-F210）。传感器和光谱仪数据存储在 MySQL 数据库中，用于判断再生和平衡阶段是否完成。此外，为保持色谱柱内液位稳定，在柱体上方安装了红外液位传感器（TOF-2m，成都中山科技有限公司）。泵 P2 的流速根据测得的液位由模糊控制器调节。系统控制源代码可在 https://github.com/tzl1125/ChromR 获取。

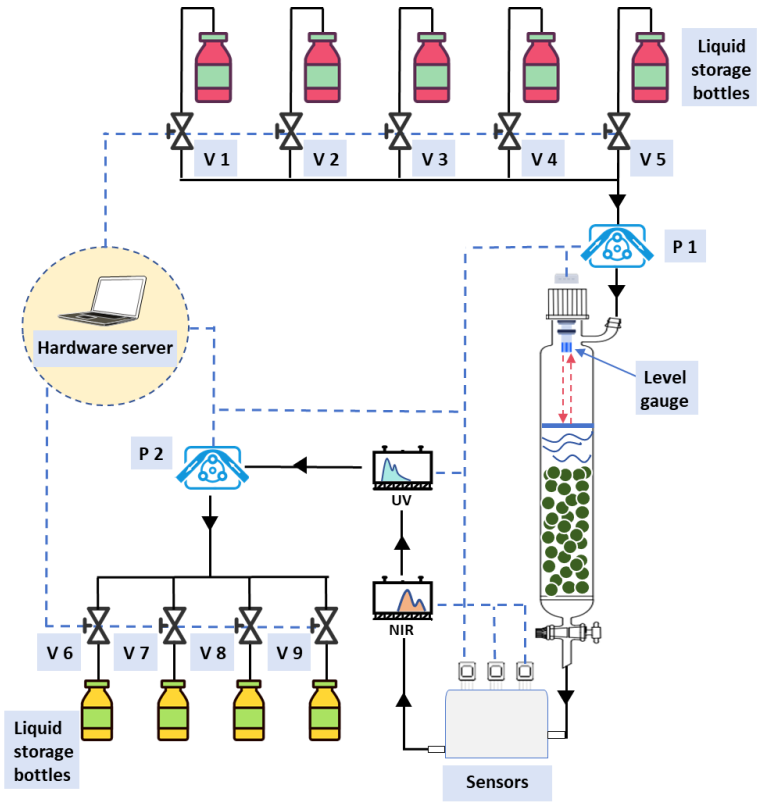


图3. 自动化色谱实验装置（实线表示流体路径，虚线表示串行通信）。

2.5 与 ChromR 的交互

为便于使用 ChromR，本研究还开发了一个用户友好的 Web 应用程序，如图 S4 所示。前端图形界面采用 Vue3.js 和 Node.js 框架开发。研究人员不仅可以通过前端界面与智能体交互以执行实验，还可以直接操作设备的各个单元。在后端，采用 FastAPI 框架

来管理多智能体系统和实验设备的各类任务，包括工具的 API 调用、传感器数据管理以及控制指令的传输。所有源代码均可在 https://github.com/tzl1125/ChromR 获取。2.6 ChromR 的应用

以 EGBL 的色谱纯化过程为例对 ChromR 进行了测试。在 EGBL 纯化过程中，大孔树脂充当分离介质。该过程包括加入水相沉淀溶液、用低浓度乙醇溶液进行清洗，以及用高浓度乙醇溶液洗脱最终馏分溶液。根据 2025 年版《中国药典》第一部，经柱色谱处理后，EGBL 中黄酮苷（FG）的含量不得低于 24.0%，萜类三内酯（TT）的含量不得低于 6.0%[59]。

优化过程中设定的工艺目标包括：在满足药典标准的同时，使 FG 和 TT 的纯度及生产率最大化。FG 和 TT 纯度的计算公式见公式（1），FG 和 TT 生产率的计算公式见公式（2）：

$$\text{Purity} = \frac{m_{target}}{m_{ts}} \times 100\% \quad (1)$$

$$\text{Productivity} = \frac{m_{tt}}{t} \quad (2)$$

其中，m_{target} 为每单位馏分体积中的目标物质质量（mg），m_{ts} 为每单位馏分体积中的总固体质量（mg），m_{tt} 为单个色谱纯化周期所得馏分中目标物质的总质量（mg），t 为一个色谱纯化过程所耗费的时间（h）。优化过程中未考虑平衡和再生步骤的持续时间。我们研究了六个工艺参数对这些目标的影响：进料流速、进料体积、洗涤流速、洗涤体积、洗脱流速和洗脱体积。平衡和再生流速固定为 3 柱体积/小时，而平衡和再生体积则根据传感器和光谱

信号确定。信号稳定后，系统自动切换至下一阶段。平衡溶剂为水，再生溶剂为 95%（v/v）乙醇。

与 ChromR 交互的主要步骤包括：(1) 为待研究的工艺参数推荐初始值，以及柱床长径比、树脂类型、洗涤溶剂和洗脱溶剂等其他必要参数；(2) 定义实验设计所需研究的工艺参数范围；(3) 根据实验设计表自动执行实验；(4) 分析实验数据、开发模型并优化工艺参数；(5) 验证最优工艺。

为更好地评估 ChromR 的性能，我们制备了13批银杏乙醇提取物，经水沉淀后作为进料溶液。选取浓度以及黄酮苷（FG）和萜类内酯（TT）的含量这四项作为材料属性。各批次的材料属性列于表 S1。图4展示了13批进料溶液的材料属性及其差异。其中10批用于工艺优化，其余三批——231201、250409和250401——留作验证优化后的工艺。详细实验方法见 S5节。

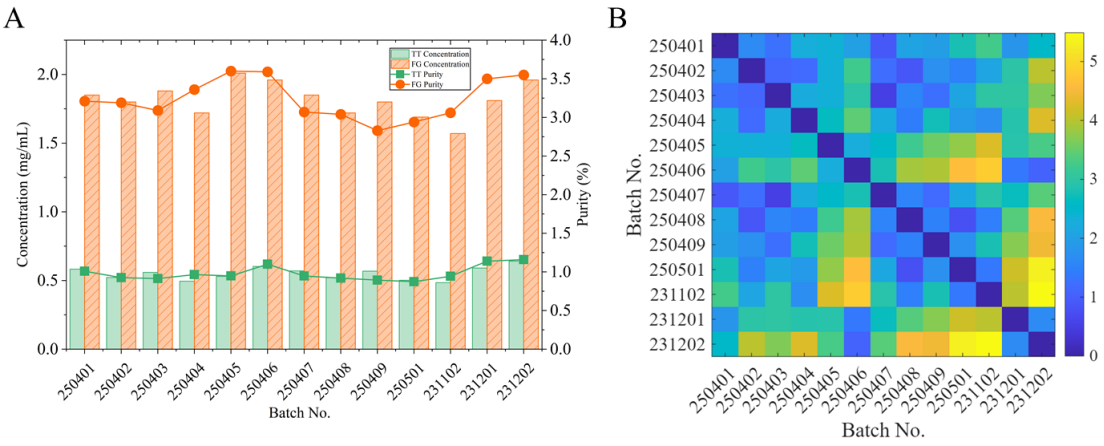


图4。各批次进料溶液的属性和差异。（A）物料attr13批进料溶液的属性。（B）进料溶液之间的欧几里得距离矩阵d 溶液批次。

3. 结果

3.1 初始色谱工艺条件

采用循序渐进的方法，让 ChromR 推荐初始工艺条件：首先确定树脂类型、洗涤溶剂和洗脱溶剂，然后再确定其他工艺参数。

通过提示工程，我们向 ChromR 说明了工艺的约束条件和优化目标，使其能够推荐合适的树脂类型以及洗涤溶剂和洗脱溶剂的乙醇浓度。详细的交互过程如图 S6 和图 S7 所示。

ChromR 成功检索到两篇相关文献[60, 61] 以及《 中国药典》中的相关规定。ChromR 首先推荐了 AB-8大孔吸附树脂，并基于树脂与目标化合物之间的化学性质分析给出了推荐理由。ChromR 还指出，文献中更倾向于使用 XDA-1树脂；然而，鉴于 AB-8 在成本和稳定性方面的优势，ChromR 最终选择了 AB-8树脂。该决策与已有报道的发现一致[62, 63],，尽管 ChromR 未检索到这些特定参考文献，这也展示了内置 ChromLLM 的强大推理能力。随后，基于“相似相溶”原理，ChromR 推荐使用 20%（v/v）乙醇水溶液作为洗涤溶剂，使用 75%（v/v）乙醇水溶液作为洗脱溶剂。

在大语言模型的训练阶段，我们专门从文献中收集了大量包含数值参数的色谱工艺条件，以增强 ChromLLM 准确推荐数值工艺参数的能力。同样，我们通过提示向 ChromR 提供已确定的工艺参数、进料溶液浓度和工艺目标，并要求其推荐其他具体工艺参数，包括柱床长径比，以及各色谱阶段的流速和持续时间。详细的问答过程见图 S8–S10。

尽管 ChromR 未检索到任何文献，但在 ChromLLM 和知识库的辅助下，它仍然给出了合理的建议。首先，综合考虑填料稳定性、流动均匀性和洗脱分辨率，ChromR 建议床层高径比为 12:1。随后，根据黄酮苷和萜类内酯在纯度和生产率方面的工艺目标，ChromR 推荐了以下参数：进料时间为 1.5 h，洗涤时间为 1 h，洗脱时间为 1 h，进料流速为 1 BV/h，洗涤流速为 2 BV/h，洗脱流速为 3 BV/h。

3.2 实验设计与执行

3.2.1 实验设计

在确定初始色谱工艺参数后，我们指示 ChromR 生成一份 DSD实验设计表，以进一步研究各项性能指标的响应面模型。我们将虚拟因子数设为 2，并增加了 3 次中心点实验。本研究选取了 10 批进料溶液用于工艺优化。ChromR 采用均匀随机分配法，将批号分配到实验设计表中。详细的交互过程见图 S11，ChromR 生成的 DSD实验设计表见表1。

表1。 实验设计表及结果。

No.	Feed flow X_1 (BV/小时)	Feed time X_2 (h)	Wash flow X_3 (BV/小时)	洗涤 时间 X4 (h)	洗脱 flow X_5 (BV/小时)	洗脱 time X6（小时）	Feed 溶液 批次 No.	TT 纯度 Y_1 (%)	TT 生产率 Y_2 (毫克/小时)	FG 纯度 Y_3 (%)	FG 生产率 Y_4 (毫克/小时)
1	1.0	2.0	2.5	1.5	3.5	1.5	250408	7.18	39.3	45.0	247
2	1.0	1.0	1.5	0.5	2.5	0.5	250403	1.14	31.5	6.65	184
3	1.5	1.5	1.5	0.5	3.5	0.5	231102	2.80	111	9.71	384
4	0.5	1.5	2.5	1.5	2.5	1.5	231202	7.58	23.3	30.3	93.0
5	1.5	2.0	2.0	0.5	2.5	1.5	250405	4.46	102	17.1	391
6	0.5	1.0	2.0	1.5	3.5	0.5	250405	6.87	16.0	31.5	73.1
7	1.5	2.0	2.5	1.0	2.5	0.5	250404	6.46	57.4	27.8	247
8	0.5	1.0	1.5	1.0	3.5	1.5	250402	4.47	16.4	21.6	79.2
9	1.5	1.0	2.5	1.5	3.0	0.5	250408	8.70	51.1	36.9	217
10	0.5	2.0	1.5	0.5	3.0	1.5	250406	1.79	29.7	8.14	134
11	1.5	2.0	1.5	1.5	3.5	1.0	250402	8.16	90.7	32.2	358

12	0.5	1.0	2.5	0.5	2.5	1.0	250407	3.62	23.4	14.4	93.4
13	1.5	1.0	2.5	0.5	3.5	1.5	250406	4.59	59.9	20.8	271
14	0.5	2.0	1.5	1.5	2.5	0.5	231202	4.62	18.8	21.8	89.0
15	1.5	1.0	1.5	1.5	2.5	1.5	250501	7.31	49.9	29.7	203
16	0.5	2.0	2.5	0.5	3.5	0.5	250403	4.20	38.4	17.8	162
17	1.0	1.5	2.0	1.0	3.0	1.0	250407	6.48	56.7	25.8	226
18	1.0	1.5	2.0	1.0	3.0	1.0	250501	7.06	57.4	28.4	231
19	1.0	1.5	2.0	1.0	3.0	1.0	250404	7.41	54.4	29.2	214
20	1.0	1.5	2.0	1.0	3.0	1.0	231102	6.97	65.9	24.3	230

3.2.2 实验执行

对于 ChromR 推荐的初始工艺条件，我们首先指示 ChromR 自动开展了 4 次操作条件相同的实验（实验 17、18、19 和 20）。结果见表1。尽管 4 批进料溶液的组成差异显著，但在 ChromR 推荐的初始工艺条件下获得的馏分已满足《中国药典》规定的萜类内酯和黄酮苷纯度要求。所有情况下，萜类内酯的生产率均超过 50 毫克/小时，黄酮苷的生产率均超过 200 毫克/小时。

随后，我们指示 ChromR 自动执行实验设计表中所有剩余的实验。对话界面如图S12–S14所示。得益于自动化，ChromR 可在夜间无人值守运行，在不到一周的时间内完成所有实验，显著降低了人工和时间成本。

3.3 工艺优化

3.3.1 模型开发

实验完成后，我们指示 ChromR 使用逐步回归（p=0.05）构建模型。具体对话内容见图S15和S16。ChromR 成功根据自动化实验运行所得数据，为每项性能指标建立了模型，等高线图见图5。各模型的偏回归系数列于表 S2，拟合结果见图S17。如图5所示，增加洗涤时间

和洗涤流速有助于提高萜类内酯和黄酮苷的纯度，而增加进料时间和进料流速则能提高萜类内酯和黄酮苷的生产率。

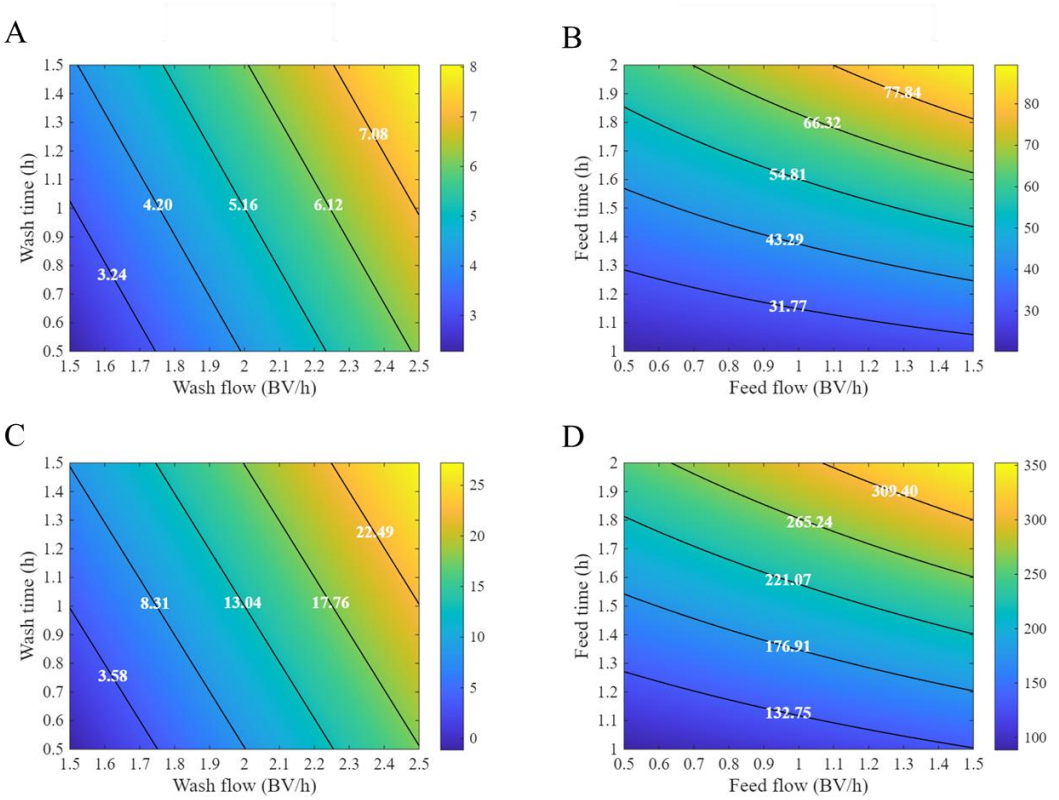


Fig. 5. 批次250401的等高线图。(A) TT纯度，X1 = 1.0 柱体积/小时，X5 = 3.0 柱体积/小时，X6 = 1.0 h。(B) TT生产率，X3 = 2.0 柱体积/小时，X4 = 1.0 h，X5 = 3.0 柱体积/小时，X6 = 1.0 h。(C) FG纯度，X1 = 1.0 柱体积/小时，X5 = 3.0 柱体积/小时，X6 = 1.0 h。(D) FG生产率，X3 =

2.0 BV/h, X_4 = 1.0 h, X_5 = 3.0 BV/h, X_6 = 1.0 h.

TT纯度和TT生产率的R^2 值均大于0.83，而FG纯度和FG生产率的R^2 值均超过0.93。这不仅表明所选工艺参数和材料属性能够解释大部分数据变异性，也在一定程度上体现了ChromR自动化实验执行的可靠性。此外，ChromR对所构建的模型进行了分析，并针对不同性能指标提出了合理的改进策略。关于提高TT和FG的纯度，ChromR建议增加洗涤时间和洗涤流速。这一建议是合理的：在洗涤阶段适当增加洗涤

体积可以有效去除高极性杂质，从而提高极性较低的目标化合物（TT和FG）的纯度。

3.3.2 多目标优化

在探索工艺参数空间并建立预测模型后，我们着手使 ChromR 能够基于所建立的模型执行多目标优化。在本研究中，我们选取了两批进料溶液——250401 和 250409，用于验证多目标优化。同样，我们通过简洁的提示词描述这些多目标优化问题，并将其交由 ChromR 自主优化。批次250401的提示词见图S18。

在不到1分钟内，ChromR 为每个多目标优化问题生成了五个不同的帕累托最优解，详见表 S3。我们从中为每个批次选取了一个最优溶液，由 ChromR 进行实验验证——250401 的溶液1和 250409 的溶液3。此外，我们还要求 ChromR 计算工艺参数和材料参数的设计空间。设计空间的定义条件为：TT 纯度必须超过6%，黄酮苷纯度必须超过24%，TT生产率必须超过50毫克/小时，黄酮苷生产率必须超过200毫克/小时。对于批次231201，我们还选取了设计空间之外的一个点，由 ChromR 进行实验验证。值得注意的是，所选的两个帕累托最优解均位于设计空间内。验证点的工艺参数列于表 S4，实验结果汇总于表2，设计空间验证结果见图6。

表2。 实验验证结果。

Feed	Y_1		Y_2		Y_3		Y_4		位置
溶液批次 No.	预测 cted 值	实验 精神 价值	预测 cted 价值	实验 精神 数值	预测 cted 数值	实验 性的 数值	预测 cted 值	实验 性 值	在 设计 空间
250401	6.80	7.81	96.5	98.0	29.4	32.8	380	412	内部
250409	8.52	8.28	88.6	89.2	37.5	33.8	355	364	内部

231201	2.47	1.98	38.0	43.4	8.54	7.99	157	175	外部

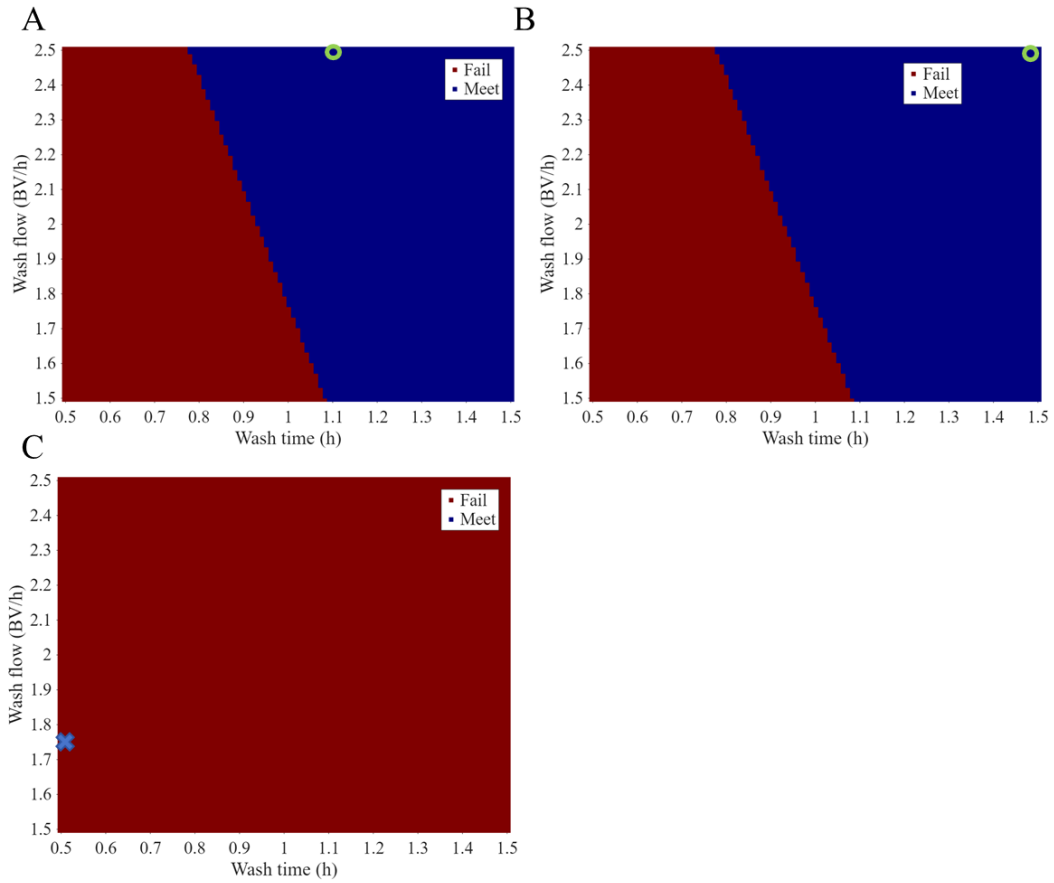


图6。 设计空间验证结果。(A) 批次250401，X1 = 1.50 柱体积/小时，X2 = 2.00 h，X5 = 3.50 柱体积/小时，X6 = 0.86 h。(B) 批次250409，X1 = 1.50 柱体积/小时，X2 = 2.00 h，X5 =3.50 柱体积/小时，X6 = 0.81 h。(C) 批次231201，X1 = 0.75 柱体积/小时，X2 = 1.00 h，X5 = 3.25 柱体积/小时，X6 = 0.50 h。符号× 表示设计空间外的点，而符号

○ 表示设计空间内的点。

帕累托最优解和设计空间均通过了验证。在优化后的工艺条件下，批次250401和250409均超过了预设的纯度和生产率目标，证实了设计空间的可靠性，并展现了 ChromR 在色谱过程优化方面的卓越表现。

4. 讨论

4.1 ChromR的创新与价值

ChromR首次将大语言模型应用于突破传统色谱工艺开发中的关键瓶颈，即对专家经验的依赖、劳动强度大以及开发周期长。在训练过程中，ChromLLM充分利用了大量经过文献验证的先验知识，使其能够基于这些知识自主推荐可行的初始工艺参数。此外，ChromLLM通过检索增强生成和文献检索模块得到增强，并由此构建智能体A，进一步提升了参数建议的合理性和先进性，显著减少了对专家判断的依赖以及所需的实验试验次数。其次，将多智能体系统与自动化色谱设备相结合，使ChromR能够自主执行实验并收集数据，大幅减少人工劳动并缩短整体开发时间。

传统的色谱工艺开发通常包括文献综述（1周）、树脂筛选（2周）、单因素实验（1周）、实验设计（2周），以及工艺优化与验证（1周），总计约需7周[61]。相比之下，基于 ChromR 的开发范式通过智能体A进行初始工艺设计，省去了前三个步骤，并通过自动化实验加快实验设计和工艺验证。以 EGBL 为例，我们仅用一周就获得了符合预设目标的工艺参数，展现出效率的大幅提升。

4.2 局限性

然而，本研究提出的 ChromR 仍有一些方面可以进一步改进。具体而言：（1）由于缺少在线混合器，无法连续改变流动相组成的梯度；（2）尚未实现自动样品制备和组分分析；（3）ChromR 目前仅限于多项式统计建模，无法建立机理模型或混合模型。

5. 结论

将大语言模型与实验室自动化技术相结合，构建了 ChromR——一个由大语言模型驱动的色谱过程优化平台。ChromR 是一个通用的色谱工艺开发平台，可有效应对色谱开发中由来已久的挑战，例如依赖专家经验、开发周期漫长以及操作耗费大量人力。通过利用 ChromLLM 从科学文献和监管指南中深度提取先验知识，ChromR 能够智能生成初始色谱参数。此外，将多智能体系统与自动化色谱实验装置相结合，大幅减少了人力投入，同时显著提高了实验效率和数据质量。在 EGBL 色谱过程优化案例中，ChromR 成功兼顾了馏分纯度和生产率要求，将工艺开发周期缩短至传统范式所需时间的 1/7。本研究为色谱过程优化范式的智能化转型提供了可复用的技术路径和工程实践参考。

CRediT 作者贡献声明

Zhilong Tang：概念构思、数据整理、正式分析、调查研究、方法论、验证、可视化、撰写初稿、审阅和编辑。Shaohua Wu：数据整理。Xinyan Zhao：审阅和编辑。Yu Wang：数据整理。Xingchu Gong：概念构思、经费获取、项目管理、资源提供、监督、审阅和编辑。

利益冲突声明

作者声明，他们不存在可能被认为会影响本文所报告研究工作的已知竞争性经济利益或个人关系。

致谢

本研究由国家重点研发计划（2024YFC3506900）资助。

作者感谢程一宇教授提出的宝贵建议。作者感谢李甜甜、冯宇轩和李金浩对数据标注工作的贡献。

附录 A。补充数据

以下是本文的补充数据：

补充数据.docx

数据可用性

本研究使用的所有代码均可在 GitHub 上公开获取：https://github.com/tzl1125/ChromR。

参考文献

[1] L. Zhao, G.H. Ma, 《复杂及超大型生物分子的色谱介质与纯化工艺：综述》, J. Chromatogr. A 1744 (2025) 13, 465721 https://doi.org/10.1016/j.chroma.2025.465721。[2] A.S. Rathore, D. Kumar, N. Kateja, 《生物药物色谱纯化的最新进展》, Biotechnol. Lett. 40 (2018) 895-905 https://doi.org/10.1007/s10529-018-2552-1。[3] G.A. Junter, L. Lebrun, 《用于病毒纯化和病毒清除的多糖基色谱吸附剂》, J. Pharm. Anal. 10 (2020) 291-312 https://doi.org/10.1016/j.jpha.2020.01.002。[4] K. Galyan, J. Reilly, 《药品纯化的绿色化学方法》, Curr. Opin. Green Sustain. Chem. 11 (2018) 76-80 https://doi.org/10.1016/j.cogsc.2018.04.018。[5] R. Amin, F. Alam, B.K. Dey, 等, 《多维色谱及其在食品、生物样品和毒素产品中的应用：全面综述》, Separations 9 (2022) 13, 326 https://doi.org/10.3390/separations9110326。[6] C.R. Bernau, M. Knödler, J. Emonts, 等, 《利用预测模型开发基于色谱的纯化工艺》, Frontiers in Bioengineering and Biotechnology 第10卷—2022年 (2022) https://doi.org/10.3389/fbioe.2022.1009102。[7] F. Michalopoulou, M.M. Papathanasiou, 《色谱分离工艺混合建模方法》, Digital Chemical Engineering 14 (2025) 100215 https://doi.org/10.1016/j.dche.2024.100215。[8] M.L. Yu, M.R. Yu, F. Qian, 《采用新型两阶段色谱工艺纯化质粒DNA》, J. Chromatogr. B 1250 (2025) 8, 124381 https://doi.org/10.1016/j.jchromb.2024.124381。[9] T. Matos, D. Hoying, A. Kristopeit, 等, 《大型病毒颗粒的连续多膜色谱》, J. Chromatogr. A 1705 (2023) 11, 464194 https://doi.org/10.1016/j.chroma.2023.464194。[10] M.H. Ardakani, C. Nosengo, S. Felletti, 等, 《结合多柱逆流色谱与绿色溶剂，提升藏红花中藏红花素-I的纯化效果》, Anal. Bioanal. Chem. (2024) 12 https://doi.org/10.1007/s00216-024-05228-6。[11] G. Compagnin, C. De Luca, C. Nosengo, 等, 《采用双柱循环色谱与绿色溶剂实现可持续的大麻素纯化》, Anal. Bioanal. Chem. 416 (2024) 4091-4099 https://doi.org/10.1007/s00216-024-05332-7。[12] Y.N. Sun, W.W. Chen, S.J. Yao, 等, 《抗体分离连续色谱的模型辅助工艺开发、表征与设计》, J. Chromatogr. A 1707 (2023) 9, 464302 https://doi.org/10.1016/j.chroma.2023.464302。

[13] N. Beeler, T. Hühn, S. Rohn 等, 《利用大孔吸附树脂从可可（可可树）水提取物中纯化黄酮类化合物》, Molecules 30 (2025) 19, 2336 https://doi.org/10.3390/molecules30112336。[14] S.Q. Lv, B.R. Li, Y.H. Li 等, 《从大豆卵磷脂中精制磷脂酰胆碱：纯化、分子种类鉴定及抗氧化活性评价》, LWT-Food Sci. Technol. 223 (2025) 11, 117704 https://doi.org/10.1016/j.lwt.2025.117704。[15] X. Yu, H.M. Chen, Z.G. Sha 等, 《通过大孔吸附树脂柱色谱高选择性分离和纯化林可霉素》, J. Chromatogr. A 1735 (2024) 8, 465282 https://doi.org/10.1016/j.chroma.2024.465282。[16] Z.W. Wang, X.J. Hou, J.X. Jin 等, 《从枳实中分离高纯度柚皮苷和新橙皮苷的绿色优化工艺开发》, Sep. Purif. Technol. 371 (2025) 12, 133198 https://doi.org/10.1016/j.seppur.2025.133198。[17] C. Zhao, J. Xu, Y. Liu 等, 《杠板归中黄酮类化合物的提取与纯化及其抗病毒和抗氧化活性》, Molecules 30 (2025) 26, 29 https://doi.org/10.3390/molecules30010029。[18] Y.J. Zhu, L. Meng, Y. Zhao 等, 《黄芪和炒牛蒡子提取过程中大孔吸附树脂对木脂素、黄酮类化合物和酚酸的吸附与解吸》, Chem. Pap. 79 (2025) 5125-5138 https://doi.org/10.1007/s11696-025-04111-7。[19] L. Yin, Y. Xu, Y. Qi 等, 《通过两步大孔树脂柱色谱从穿龙薯蓣中工业规模制备薯蓣皂苷的绿色高效工艺》, Chem. Eng. J. 165 (2010) 281-289 https://doi.org/10.1016/j.cej.2010.09.045。[20] D. Keulen, G. Geldhof, O.L. Bussy 等, 《加速纯化工艺开发的最新进展：聚焦疫苗的综述》, J. Chromatogr. A 1676 (2022) 463195 https://doi.org/10.1016/j.chroma.2022.463195。[21] M. Sim, M.G. Vakili, F. Strieth-Kalthoff 等, 《ChemOS 2.0：化学自驱动实验室的编排架构》, Matter 7 (2024) 20 https://doi.org/10.1016/j.matt.2024.04.022。[22] C. Scheurer, K. Reuter, 《人在回路在多相催化新兴自驱动实验室中的作用》, Nat. Catal. 8 (2025) 13-19 https://doi.org/10.1038/s41929-024-01275-5。[23] N. Mukhin, P. Jha, M. Abolhasani, 《流动化学在自驱动实验室中的作用》, Matter 8 (2025) 12, 102205 https://doi.org/10.1016/j.matt.2025.102205。[24] P.M. Pittaway, S.T. Knox, O.J. Cayre 等, 《用于乳液聚合的自驱动实验室》, Chem. Eng. J. 507 (2025) 14, 160700 https://doi.org/10.1016/j.cej.2025.160700。

[25] C.W. Coley, D.A. Thomas, J.A.M. Lummiss 等，《受人工智能规划指导的有机化合物流动合成机器人平台》，Science 365 (2019) eaax1566 https://doi.org/10.1126/science.aax1566。[26] N. Singh, S. Lane, T.H. Yu 等，《人工智能驱动的自主酶工程通用平台》，Nat. Commun. 16 (2025) 13, 5648 https://doi.org/10.1038/s41467-025-61209-y。[27] J.T. Rapp, B.J. Bremer, P.A. Romero，《自驱动实验室自主探索蛋白质适应度景观》，Nat. Chem. Eng. 1 (2024) 97-107 https://doi.org/10.1038/s44286-023-00002-4。[28] S. Putz, J. Doettling, T. Ballweg 等，《用于固相萃取工艺优化及核酸纯化应用的自驱动实验室》，Adv. Intell. Syst. 7 (2025) 16 https://doi.org/10.1002/aisy.202400564。[29] F. Delgado-Licona, A. Alsaiari, H. Dickerson 等，《流动驱动的数据强化加速自主无机材料发现》，Nat. Chem. Eng. 2 (2025) 436-446 https://doi.org/10.1038/s44286-025-00249-z。[30] R.B. Canty, J.A. Bennett, K.A. Brown 等，《借助自驱动实验室实现科学加速与可及性提升》，Nat. Commun. 16 (2025) 11, 3856 https://doi.org/10.1038/s41467-025-59231-1。[31] G. Tom, S.P. Schmid, S.G. Baird 等，《面向化学与材料科学的自驱动实验室》，Chem. Rev. 124 (2024) 9633-9732 https://doi.org/10.1021/acs.chemrev.4c00055。[32] X.L. Wang, Y.C. Jin, S. Schmitt 等，《贝叶斯优化最新进展》，ACM Comput. Surv. 55 (2023) 36, 287 https://doi.org/10.1145/3582078。[33] Y.Z. Gao, T.C. Yu, J.L. Li，《第六届机器学习优化与数据科学会议论文集——LOD年度会议》，2020年，意大利锡耶纳，2020年，第350-361页。[34] L. Greif, N. Hübschle, A. Kimmig 等，《贝叶斯优化中的结构化采样策略：在数学场景和现实场景中的评估》，J. Intell. Manuf. (2025) 31 https://doi.org/10.1007/s10845-025-02597-2。[35] A. Mirza, N. Alampara, S. Kunchapu 等，《评估大语言模型化学知识与推理能力的框架：与化学家的专业水平对比》，《自然·化学》17 (2025) 1027-1034 https://doi.org/10.1038/s41557-025-01815-x。[36] J. Dagdelen, A. Dunn, S. Lee 等，《利用大语言模型从科学文本中进行结构化信息抽取》，Nat. Commun. 15 (2024) 1418 https://doi.org/10.1038/s41467-024-45563-x。[37] Z. Zheng, F. Florit, B. Jin 等，《融合机器学习与大语言模型以推进电化学反应探索》，《应用化学国际版》64 (2025) e202418074 https://doi.org/10.1002/anie.202418074。

[38] L. Hu, Z.R. Zhou, G.Z. Jia，基于大语言模型和自动化机器学习的一次性自动化框架：加速多孔碳材料的设计与碳捕集优化，Sep. Purif. Technol. 376 (2025) 15, 133487 https://doi.org/10.1016/j.seppur.2025.133487。[39] H.W. Yang, J. Wang, R.M. Mo 等，智能方法：将 ChatGPT 集成到土壤镉生物炭固定化实验规划中，Sep. Purif. Technol. 352 (2025) 9, 128170 https://doi.org/10.1016/j.seppur.2024.128170。[40] X.F. Bai, H.T. Wang, L.H. Xie 等，用于金属有机框架材料多目标筛选的集成式人工智能系统，Sep. Purif. Technol. 376 (2025) 7, 133939 https://doi.org/10.1016/j.seppur.2025.133939。[41] E. Nguyen, M. Poli, M.G. Durrant 等，利用 Evo 进行从分子尺度到基因组尺度的序列建模与设计，Science 386 (2024) eado9336 https://doi.org/10.1126/science.ado9336。[42] Y.X. Ruan, C.Y. Lu, N. Xu 等，由大语言模型驱动的自动化端到端化学合成开发平台，Nat. Commun. 15 (2024) 16, 10160 https://doi.org/10.1038/s41467-024-54457-x。[43] D.A. Boiko, R. MacKnight, B. Kline 等，利用大语言模型开展自主化学研究，Nature 624 (2023) 570-578 https://doi.org/10.1038/s41586-023-06792-0。[44] X.M. Yan, Y. Xiao, Y.C. Jin，《生成式大语言模型解析》，IEEE Comput. Intell. Mag. 19 (2024) 45-46 https://doi.org/10.1109/mci.2024.3431454。[45] Y. Zhang, Y. Han, S. Chen 等，利用大语言模型加速有机化学合成，《自然·机器智能》7 (2025) 1010-1022 https://doi.org/10.1038/s42256-025-01066-y。[46] Y.Z. Zheng, H.Y. Koh, J.X. Ju 等，用于分子性质预测的科学发现大语言模型，《自然·机器智能》7 (2025) 23 https://doi.org/10.1038/s42256-025-00994-z。[47] K.M. Jablonka, P. Schwaller, A. Ortega-Guerrero 等，利用大语言模型开展预测化学研究，《自然·机器智能》6 (2024) 161-169 https://doi.org/10.1038/s42256-023-00788-1。[48] Y. Kang, J. Kim, ChatMOF：利用大语言模型预测和生成金属有机框架的人工智能系统，Nat. Commun. 15 (2024) 4705 https://doi.org/10.1038/s41467-024-48998-4。[49] A. M. Bran, S. Cox, O. Schilter 等，利用化学工具增强大语言模型，《自然·机器智能》6 (2024) 525-535 https://doi.org/10.1038/s42256-024-00832-8。[50] Z. Zheng, N. Rampal, T.J. Inizan 等，面向网状化学的大语言模型，《自然综述：材料》10 (2025) 369-381 https://doi.org/10.1038/s41578-025-00772-8。

[51] C. Zhu, J. Liu, J.X. Lin 等，探究银杏叶提取物对阿尔茨海默病认知功能的影响，CNS Neurosci. Ther. 30 (2024) 18, e14914 https://doi.org/10.1111/cns.14914。[52] F.F. Wang, S.H. Ye, Y. Ding 等，银杏叶多糖的结构与抗氧化活性研究，J. Mol. Struct. 1252 (2022) 10, 132185 https://doi.org/10.1016/j.molstruc.2021.132185。[53] L.H. Zhang, X.Y. Fang, J.H. Sun 等，银杏叶提取物典型功能成分的协同抗炎作用研究，Molecules 28 (2023) 13, 1377 https://doi.org/10.3390/molecules28031377。[54] J. Deng, G. Liu, R. Tang 等，一种用于从文献中收集材料数据的自动选择性 PDF 表格提取方法，Advances in Engineering Software 204 (2025) 103897 https://doi.org/10.1016/j.advengsoft.2025.103897。[55] A. Şahin, B.C. Kara, T. Dirsehan，LitOrganizer：实现科学文献综述数据提取与组织过程自动化，SoftwareX 30 (2025) 102198 https://doi.org/10.1016/j.softx.2025.102198。[56] J. Bai, S. Bai, Y. Chu 等，通义千问技术报告，ArXiv abs/2309.16609 (2023)。[57] X. Zhang, Z. Tang, B. Chen 等，考虑材料和工艺两方面变化的参数优化：以黄芩提取物为例，Separations 12 (2025) 165 https://doi.org/10.3390/separations12060165。[58] T.M. Dixon, J. Williams, M. Besenhard 等，贝叶斯优化指导的无操作员高效液相色谱自动化方法开发，Digital Discovery 3 (2024) 1591-1601 https://doi.org/10.1039/d4dd00062e。[59] N.P. Commission，《中华人民共和国药典》。编者，《中华人民共和国药典》，丛书《中华人民共和国药典》，中国医药科技出版社，北京，2025。[60] I. Lejri, I. Vukalović, A. Grimm 等，银杏提取物 EGb 761® 中的原花青素可改善体外生物能量学并促进神经突生长，Frontiers in Pharmacology 16 - 2025 (2025) https://doi.org/10.3389/fphar.2025.1495997。[61] S.J. Zhang, X.C. Gong, H.B. Qu，采用大孔树脂柱层析制备富含莽草酸的银杏叶提取物的有效且全面的优化策略，Phytochem. Anal. 35 (2024) 1428-1442 https://doi.org/10.1002/pca.3375。[62] M. Kang, X.J. Wang，采用液相色谱-质谱联用对银杏叶中生物活性黄酮进行大孔树脂纯化与定量分析，CyTA-J. Food 23 (2025) 13, 2517329 https://doi.org/10.1080/19476337.2025.2517329。[63] L.H. Zhang, T.T. Wu, W. Xiao 等，采用大孔树脂从银杏提取物中富集和纯化银杏总黄酮O-苷，并评价其体外抗炎活性，Molecules 23 (2018) 12, 1167 https://doi.org/10.3390/molecules23051167。

以下为补充信息（SI）：

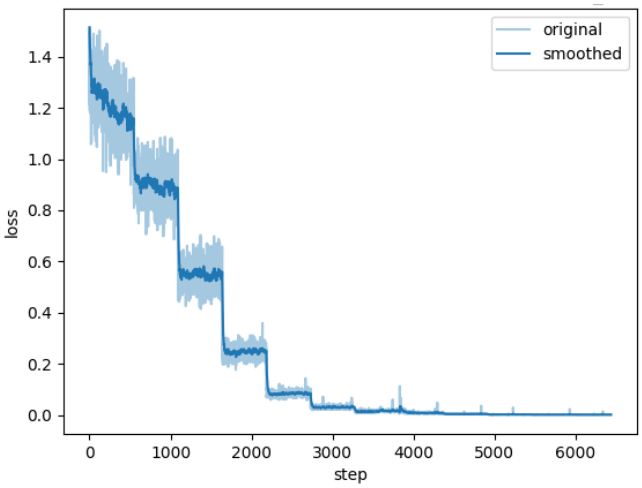
大语言模型驱动的色谱工艺设计与优化平台：银杏叶提取物的应用案例

Zhilong Tang[1], Shaohua Wu[1], Xinyan Zhao[2], Yu Wang[1,3], Xingchu Gong[1,2,3,4,*]1. 浙江大学药学院药学信息学研究所，杭州 310058，中国 2. 天津中医药大学中医药研究所，天津 301617，中国 3. 浙江大学金华研究院，金华 321016，中国 4. 中药现代化国家重点实验室，浙江大学，杭州 310058，中国

[*] 通讯作者：gongxingchu@zju.edu.cn；电话：+86-571-88208427

1. 后预训练

Qwen2.5-14B 后预训练的损失曲线如图S1所示。



图S1。后预训练的损失曲线。

2. 监督微调

大语言模型（LLM）后预训练后的监督微调损失曲线如图 S2 所示。

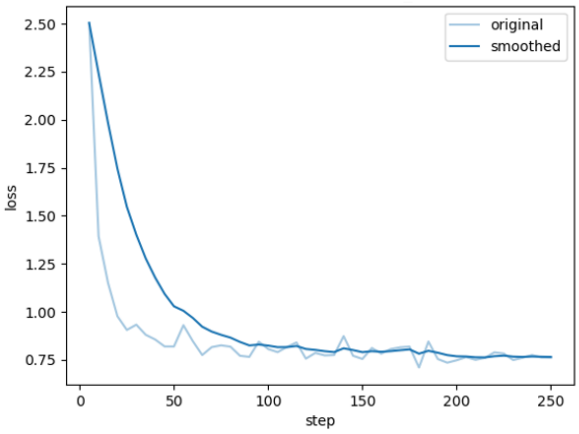


图 S2。监督微调的损失曲线。

3. 自动化色谱实验装置

自动化色谱实验装置的照片如图 S3 所示。



图 S3。自动化色谱实验装置。

4. Web 应用程序

图 S4 展示了 ChromR Web 应用程序前端的截图。该前端包括用于与多智能体系统交互的对话框、任务

执行栏、实验日志、实验记录、硬件控制系统、在线传感器数据和在线光谱数据。



图 S4. ChromR Web 应用程序前端截图。

5. 材料与方法

5.1 试剂与材料

槲皮素（批号：231227， ≥99 %）、山柰酚（批号：240225， ≥98 %）、异鼠李素（批号：240602， ≥99 %）、银杏内酯A（批号：200616， ≥99 %）、银杏内酯B（批号：191224，≥99 %）、银杏内酯C（批号：200729， ≥99 %）和白果内酯（批号：191221， ≥98 %）均购自上海威尔伯医药科技有限公司（中国上海）。所用甲酸（≥99 %，色谱级，ROE Scientific Inc.，美国纽瓦克）、甲醇（色谱级，安徽天地高纯溶剂有限公司，中国安徽）和盐酸（36–38%，分析纯，国药集团化学试剂有限公司，中国上海）。实验用水由超纯水系统（Milli-Q，Millipore，美国）制备。银杏叶醇提取物浓缩液由通化谷红药业有限公司（中国吉林）生产。

5.2 供料液制备

银杏叶乙醇提取水沉淀溶液用作进料溶液，制备方法如下：将银杏叶醇提浓缩物与水按 1:4（v/v）的比例混合，搅拌 30 分钟。随后在 4°C 下冷沉淀 12 小时，再进行过滤。采用此方法，将 13 批银杏叶醇提浓缩物制备成相应的进料溶液。通过黄酮苷（FG）和萜类三内酯（TT）的浓度与纯度表征材料属性。批号及材料属性列于表 S1。

表 S1. 进料溶液的批号和材料属性。

	TT		FG	
批次编号	浓度（毫克/毫升）	TT纯度 （%）	浓度（毫克/毫升）	FG纯度 （%）

250401	0.583	1.01	1.85	3.21
250402	0.522	0.926	1.80	3.19
250403	0.559	0.914	1.88	3.09
250404	0.495	0.966	1.72	3.36
250405	0.530	0.949	2.01	3.60
250406	0.602	1.10	1.96	3.59
250407	0.570	0.948	1.85	3.07
250408	0.520	0.920	1.72	3.04
250409	0.568	0.894	1.80	2.83
250501	0.502	0.874	1.69	2.94
231102	0.484	0.944	1.57	3.06
231201	0.591	1.14	1.81	3.50
231202	0.641	1.16	1.96	3.55

5.3 供试品溶液制备

供试品溶液的制备方法：取 1 mL银杏叶浓缩溶液，加入 1 mL 25% 盐酸和 4 mL 甲醇。将混合物置于 80 °C 水浴中回流加热 30 分钟，然后冷却至室温。加入甲醇，使总体积达到 10 mL。经 0.22 μm 滤膜过滤后，取续滤液作为供试品溶液。

混合对照品溶液的制备方法：分别准确称取适量槲皮素、山柰酚、异鼠李素、银杏内酯A、银杏内酯B、银杏内酯C和白果内酯对照品，置于10 mL容量瓶中。用甲醇溶液溶解并定容后，将容量瓶置于超声波清洗器中超声处理5分钟。精密吸取适量对照品储备液至容量瓶中，用甲醇稀释，配制成一系列不同浓度的对照

品溶液。经0.22 μm滤膜过滤后，取续滤液作为对照品储备液。5.4 分析条件

样品按照液相色谱-质谱法 [1]进行分析。质谱条件：单四极杆质谱仪检测器为 Agilent InfinitylabLC/MSDiQ（安捷伦科技，美国），配备电喷雾电离离子源，以负离子模式采集数据，监测槲皮素的 m/z 301、异鼠李素的 m/z 315、山柰酚的 m/z 285、白果内酯的 m/z 325、银杏内酯B的 m/z 423、银杏内酯C的 m/z 439，均以 $[M-H]^-$ 形式监测。银杏内酯A监测 m/z 453，形式为 $[M-H+COOH]^-$。雾化器压力为 50 psi，毛细管电压为 2000 V，干燥气温度为 325°C。白果内酯和银杏内酯A的碎裂电压为 90 eV，银杏内酯B和银杏内酯C为 100 eV，槲皮素、山柰酚和异鼠李素为 140 eV。

色谱条件：使用 Agilent 1290高效液相色谱系统（安捷伦科技，美国）和 Agilent InfinityLab Poroshell 120色谱柱（4.6 × 100 mm，2.7 μm，安捷伦科技，美国）进行分析。流动相由 A 相（水中含 0.1% 甲酸）和 B 相（甲醇中含 0.1% 甲酸）组成，流速为 0.66 mL/min，柱温维持在 30 °C。采用梯度洗脱，比例如下：0–15 min，0–35% B；15–30 min，35%–80% B；30–31 min，80%–95% B；31–41 min，95%–95% B。混合标准品和洗脱液的色谱图见图 S5。

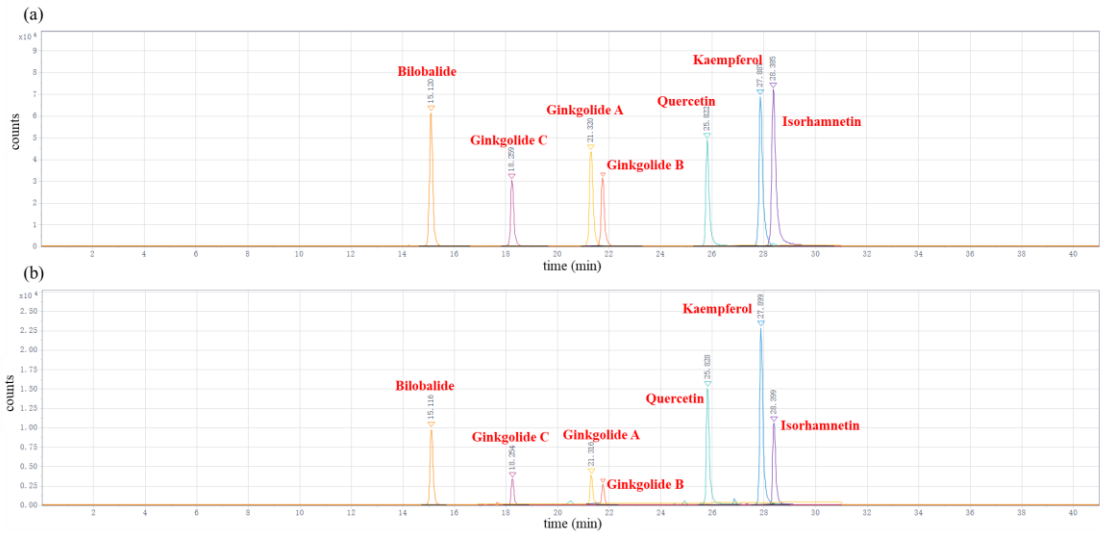


图 S5. 混合标准品（a）和洗脱液（b）的色谱图。

5.5 总固体测定

总固体测定方法参照《中华人民共和国药典》[2]。准确量取 4 mL 供试液，加入已恒重并预先称量的称量瓶中。随后将称量瓶放回鼓风干燥箱，在 105 °C 下干燥至少 3 小时。干燥后，将其转移至干燥器中静置 30 分钟，然后取出。称量称量瓶的总质量，并将其放回干燥箱继续干燥。重复此过程，直至连续两次总质量之差在 0.3 mg 以内，并记录最终质量。总固体的质量按最终质量减去空称量瓶的质量计算。

6. 银杏叶提取物色谱工艺的优化过程

以下为色谱工艺的优化过程，包括研究人员与 ChromR 之间的对话，以及 ChromR 提供的数据分析结果。

Conversation

Researcher:

In the column chromatography process for preparing Ginkgo biloba leaf extract, if—after sample loading and adsorption is complete—only one isocratic wash followed by one isocratic elution is performed, and the eluate is collected as the target product, then:

1. What type of macroporous resin should be selected to maximize the content of total flavonol glycosides and terpene lactones in the product?
2. What ethanol concentrations are recommended for the wash solution and the elution solution, respectively?

ChromR:

Agent A Execution Result:
In the column chromatography process for preparing Ginkgo biloba leaf extract, if only one isocratic wash and one isocratic elution are performed, and the eluate is collected as the target product, it is essential to holistically consider the type of macroporous resin and the elution conditions in order to maximize the content of **total flavonol glycosides** and **terpene lactones** (primarily including ginkgolides A, B, C, and bilobalide) in the final product.

☑ Recommended Protocol:

1. Macroporous Adsorption Resin Type: AB-8
- **Rationale**:
  - AB-8 is a nonpolar to weakly polar polystyrene-divinylbenzene copolymer macroporous resin, offering excellent adsorption capacity and selectivity for enriching and separating moderately polar compounds such as flavonoids and terpene lactones.
  - According to expert chromatographic models and practical process studies, AB-8 resin demonstrates high adsorption capacity and desorption efficiency for total flavonol glycosides and terpene lactones in Ginkgo biloba leaf extracts.
  - Although the Chinese Pharmacopoeia does not specify a particular resin type, it recommends the use of "macroporous adsorption resins" with sequential elution using water followed by ethanol gradients of varying concentrations. AB-8 is one of the standard resins widely adopted in industrial Ginkgo extraction processes.

> Additionally, literature [2] reports that XDA-1 resin also exhibits excellent performance under pH 2.0 conditions; however, AB-8 is generally preferred in routine production due to its lower cost, superior stability, and well-established application history.

---

2. Ethanol Concentration in Wash Solution: 20% (v/v)
- **Purpose**: To remove water-soluble impurities (e.g., sugars, certain organic acids, tannins, etc.) while retaining the target compounds on the resin.
- **Mechanism**: A 20% ethanol solution effectively elutes weakly adsorbed impurities, whereas total flavonol glycosides and terpene lactones remain largely retained on the resin at this concentration and are not significantly desorbed.
- **Basis**: Experimental studies indicate that ethanol concentrations below 30% have minimal impact on the target components while effectively enhancing the purity of the subsequent eluate [2].

Continuation

图 S6。ChromR 最初建议的工艺参数。

Conversation

Continuation

ChromR:

3. Ethanol Concentration in Elution Solution: 75% (v/v)
- **Purpose**: To maximize the elution of total flavonol glycosides and terpene lactones.
- **Mechanism**:
  - Total flavonol glycosides predominantly exist in glycosidic forms with moderate polarity; 75% ethanol provides an optimal polarity environment for efficient desorption.
  - Terpene lactones (e.g., ginkgolides A, B, C, and bilobalide), although relatively less polar, also achieve high recovery rates in 75% ethanol.
- **Experimental Evidence**:
  - Multiple studies have shown that 70–80% ethanol is the optimal range for eluting Ginkgo flavonoids and lactones, with 75% representing the best compromise to balance high yields of both compound classes [2].
  - If the ethanol concentration is too high (e.g., ≥90%), solubility of certain flavonoid glycosides may decrease, and more lipophilic impurities could co-elute; if too low, incomplete elution occurs.

☑ Summary of Final Recommended Parameters:

Parameter	Recommended Value	Notes
Macroporous Resin Type	**AB-8**	Nonpolar/weakly polar; suitable for enrichment of flavonoids and lactones
Ethanol Concentration in Wash Solution	**20%**	Removes impurities while retaining target compounds on the resin
Ethanol Concentration in Elution Solution	**75%**	Enables efficient co-elution of total flavonol glycosides and terpene lactones

---

Additional Notes:
- If selective enrichment of a specific compound class (e.g., prioritizing terpene lactones or flavonoids) is desired, consider fractionated collection or pH optimization (e.g., adjusting the sample loading solution to pH 2.0–3.0 to enhance adsorption) [2].
- Under the process constraint of "only one isocratic wash followed by one isocratic elution," the above parameters achieve maximal enrichment of target components while maintaining operational simplicity.

References

[1] Lejri, I., Vukalović, I., Grimm, A., & Eckert, A. (2025). Proanthocyanidins from Ginkgo extract EGb 761® improve bioenergetics and stimulate neurite outgrowth in vitro. *Frontiers in Pharmacology*, 16, 1495997. https://doi.org/10.3389/fphar.2025.1495997

[2] Zhang, S., Gong, X., & Qu, H. (2024). An effective and comprehensive optimization strategy for preparing Ginkgo biloba leaf extract enriched in shikimic acid by macroporous resin column chromatography. *Phytochemical Analysis*, 35(5), 1234–1245. https://doi.org/10.1002/pca.3375

> Note: Other literature sources were not directly relevant to resin selection or elution concentration optimization and are therefore not cited.

图 S7。图 S6 的续图。

Conversation

Researcher:

In the column chromatography purification process for Ginkgo biloba leaf extract, AB-8 macroporous resin is used as the adsorbent, and the sample loading solution is prepared at a concentration of 0.366 g of herbal material per mL. After sample loading and adsorption, an isocratic wash with 20% (v/v) ethanol followed by an isocratic elution with 75% (v/v) ethanol is performed, with the eluate collected as the target product.

To minimize processing time per cycle and maximize the contents of total flavonol glycosides and terpene lactones in the final product—while simultaneously reducing material waste—please address the following operational parameters:

1. **What should the bed height-to-diameter ratio (H/D) be?**
2. **What are the recommended volumes (in column volumes, BV) for sample loading, washing, and elution, respectively?**
3. **What flow rates (e.g., in BV/h or mL/min) are recommended for the loading, washing, and elution stages, respectively?**

ChromR:

Agent A Execution Result:

In the column chromatography purification process for *Ginkgo biloba* leaf extract, **AB-8 macroporous resin** is employed as the adsorbent. The objective is to **minimize material waste while achieving the shortest possible single-cycle processing time and maximizing the contents of total flavonol glycosides and terpene lactones in the final product**. Based on the preparation requirements specified in the *Chinese Pharmacopoeia* and optimization recommendations provided by a domain-specific large model for chromatography, the following key process parameters are recommended:

---

1. Bed Height-to-Diameter Ratio (H/D)

Recommended bed height-to-diameter ratio: 12:1 (e.g., 12 cm bed height / 1 cm inner column diameter).

This H/D ratio enhances column efficiency and separation selectivity, reduces axial dispersion while ensuring adequate mass transfer, and improves enrichment of flavonol glycosides and terpene lactones. It balances column packing stability, flow uniformity, and elution resolution—making it suitable for industrial scale-up.

> **Note**: An excessively low H/D ratio leads to incomplete separation, while an overly high ratio may cause excessive pressure drop and difficulties in flow control.

Continuation

图 S8. ChromR 推荐的其他工艺参数。

Conversation

Continuation

ChromR:

2. Loading, Washing, and Elution Volumes (expressed in Column Volumes, BV)

Step	Recommended Volume
Loading	**1.5 BV**
Washing	**2.0 BV** (20% ethanol)
Elution	**3.0 BV** (75% ethanol)

- **Loading at 1.5 BV**: At the given sample concentration (0.366 g herbal material/mL), this volume fully utilizes the resin's adsorption capacity without overloading, thereby preventing breakthrough of target compounds.
- **Washing with 2.0 BV of 20% ethanol**: Effectively removes water-soluble impurities (e.g., sugars, certain phenolic acids) while retaining flavonol glycosides and terpene lactones on the resin.
- **Elution with 3.0 BV of 75% ethanol**: Ensures complete desorption of strongly bound flavonol glycosides and terpene lactones, maximizing recovery while producing a concentrated eluate that facilitates downstream concentration.

---

3. Recommended Flow Rates (in BV/h)

Stage	Recommended Flow Rate
Loading	**1 BV/h**
Washing	**2 BV/h**
Elution	**3 BV/h**

- **Loading at 1 BV/h**: A slower flow rate allows sufficient diffusion of target compounds into the resin pores, improving adsorption efficiency and utilization of binding capacity.
- **Gradually increased flow rates during washing and elution**: Since washing primarily removes loosely bound impurities and 75% ethanol enables rapid mass transfer during elution, moderately higher flow rates do not compromise recovery but significantly reduce overall cycle time and improve throughput.

---

4. Expected Performance

Under the above conditions, model predictions indicate:

- **Total flavonol glycosides content: 24.79%**
- **Terpene lactones content: 6.42%**

These values meet the quality specifications outlined in the *Chinese Pharmacopoeia* for *Ginkgo biloba* leaf extract (total flavonol glycosides ≥24%; terpene lactones ≥6%), achieving an **optimal balance among product quality, production efficiency, and material utilization**.

Continuation

图 S9。图 S8 的续图。

Conversation

Continuation

ChromR:

5. Conclusion

To enable efficient, low-waste, and high-quality purification of *Ginkgo biloba* leaf extract, the following process parameters are recommended:

Parameter	Recommended Value
Bed height-to-diameter ratio	12:1 (H/D)
Loading volume	1.5 BV
Washing volume	2.0 BV (20% ethanol)
Elution volume	3.0 BV (75% ethanol)
Loading flow rate	1 BV/h
Washing flow rate	2 BV/h
Elution flow rate	3 BV/h

This protocol is designed based on the adsorption characteristics of macroporous resins and polarity differences among constituents. It aligns with the preparation principles of the *Chinese Pharmacopoeia* [1] and has been validated through model-based optimization, making it suitable for standardized production at both laboratory and pilot scales.

References

[1] Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People' s Republic of China (2020 Edition, Volume I)*: Vegetable Oils and Extracts – *Ginkgo biloba* Leaf Extract [S]. Beijing: China Medical & Pharmaceutical Press, 2020.

图 S10。图 S9 的续图。

Conversation

Researcher:

Run a DSD experimental design. The loading solution batches include ten lots: 231102, 231202, 250402, 250403, 250404, 250405, 250406, 250407, 250408, and 250501. The loading flow rate is set at 0.5–1.5 BV/h, with a loading time of 1–2 h. The wash flow rate is 1.5–2.5 BV/h, and the wash time is 0.5–1.5 h. The elution flow rate is 2.5–3.5 BV/h, and the elution time is 0.5–1.5 h. Two dummy factors are included, and three center point experiments are added.

ChromR:

Agent B Execution Result：

No.	Loading Flow Rate	Loading Time	Wash Flow Rate	Wash Time	Elution Flow Rate	Elution Time	Batch Number
1	1.0	2.0	2.5	1.5	3.5	1.5	250408
2	1.0	1.0	1.5	0.5	2.5	0.5	250403
3	1.5	1.5	1.5	0.5	3.5	0.5	231102
4	0.5	1.5	2.5	1.5	2.5	1.5	231202
5	1.5	2.0	2.0	0.5	2.5	1.5	250405
6	0.5	1.0	2.0	1.5	3.5	0.5	250405
7	1.5	2.0	2.5	1.0	2.5	0.5	250404
8	0.5	1.0	1.5	1.0	3.5	1.5	250402
9	1.5	1.0	2.5	1.5	3.0	0.5	250408
10	0.5	2.0	1.5	0.5	3.0	1.5	250406
11	1.5	2.0	1.5	1.5	3.5	1.0	250402
12	0.5	1.0	2.5	0.5	2.5	1.0	250407
13	1.5	1.0	2.5	0.5	3.5	1.5	250406
14	0.5	2.0	1.5	1.5	2.5	0.5	231202
15	1.5	1.0	1.5	1.5	2.5	1.5	250501
16	0.5	2.0	2.5	0.5	3.5	0.5	250403
17	1.0	1.5	2.0	1.0	3.0	1.0	250407
18	1.0	1.5	2.0	1.0	3.0	1.0	250501
19	1.0	1.5	2.0	1.0	3.0	1.0	250404
20	1.0	1.5	2.0	1.0	3.0	1.0	231102

图 S11. ChromR 提供的实验设计表。

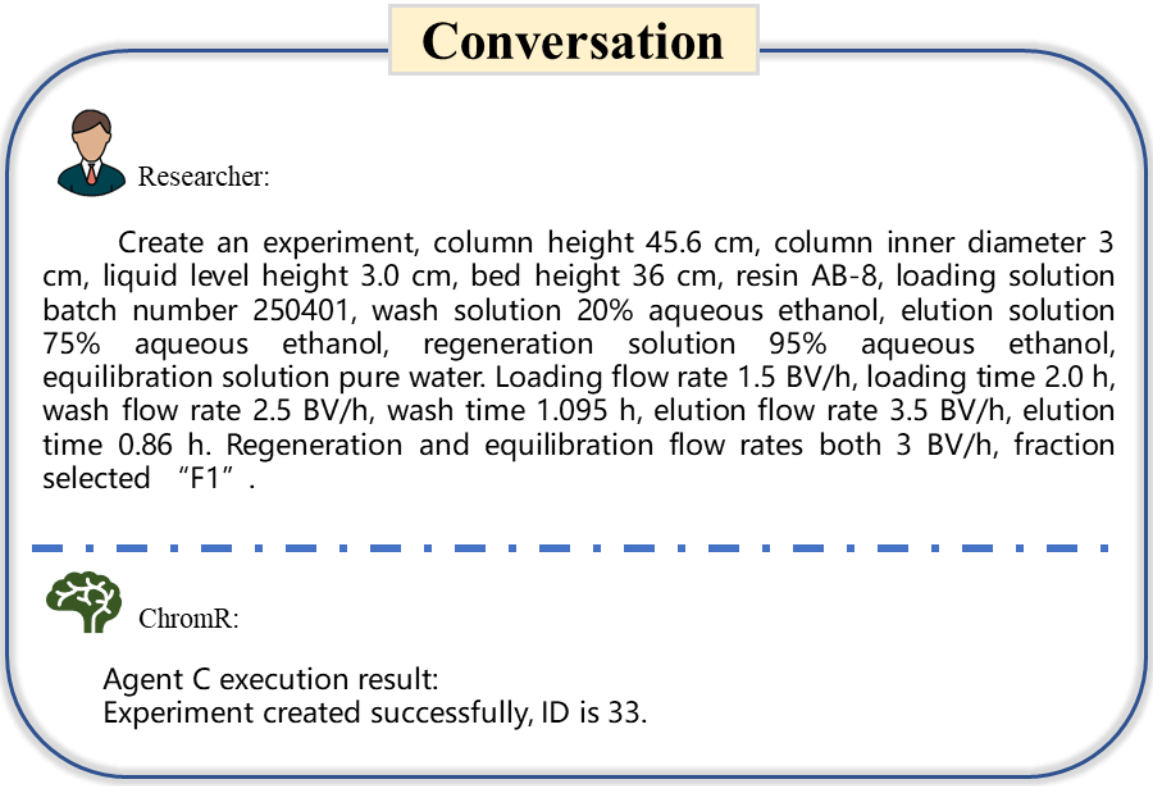


图 S12。在 ChromR 中通过提示创建实验。

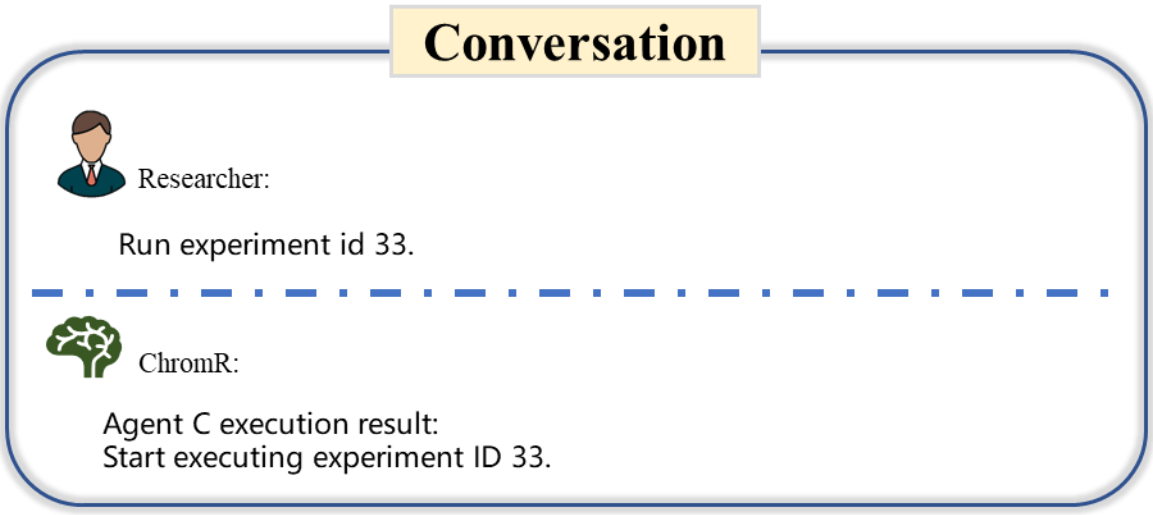


图 S13。在 ChromR 中通过提示运行指定实验。

Conversation

Researcher:

Create a DSD experimental design table, column height 45.6 cm, column inner diameter 3 cm, liquid level height 3.0 cm, bed height 36 cm, resin AB-8, wash solution 20% aqueous ethanol, elution solution 75% aqueous ethanol, regeneration solution 95% aqueous ethanol, equilibration solution pure water. Regeneration and equilibration flow rates both 3 BV/h, fraction selected "F1".

The experimental design table is as follows:

No.	Loading Flow Rate	Loading Time	Wash Flow Rate	Wash Time	Elution Flow Rate	Elution Time	Batch Number
1	1.0	2.0	2.5	1.5	3.5	1.5	250408
2	1.0	1.0	1.5	0.5	2.5	0.5	250403
3	1.5	1.5	1.5	0.5	3.5	0.5	231102
4	0.5	1.5	2.5	1.5	2.5	1.5	231202
5	1.5	2.0	2.0	0.5	2.5	1.5	250405
6	0.5	1.0	2.0	1.5	3.5	0.5	250405
7	1.5	2.0	2.5	1.0	2.5	0.5	250404
8	0.5	1.0	1.5	1.0	3.5	1.5	250402
9	1.5	1.0	2.5	1.5	3.0	0.5	250408
10	0.5	2.0	1.5	0.5	3.0	1.5	250406
11	1.5	2.0	1.5	1.5	3.5	1.0	250402
12	0.5	1.0	2.5	0.5	2.5	1.0	250407
13	1.5	1.0	2.5	0.5	3.5	1.5	250406
14	0.5	2.0	1.5	1.5	2.5	0.5	231202
15	1.5	1.0	1.5	1.5	2.5	1.5	250501
16	0.5	2.0	2.5	0.5	3.5	0.5	250403
17	1.0	1.5	2.0	1.0	3.0	1.0	250407
18	1.0	1.5	2.0	1.0	3.0	1.0	250501
19	1.0	1.5	2.0	1.0	3.0	1.0	250404
20	1.0	1.5	2.0	1.0	3.0	1.0	231102

ChromR:

Agent C execution result:

Experimental design table created successfully, experimental design table ID: 1, experiment IDs: [13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32].

图 S14. 通过 ChromR 中的提示创建实验设计表。

Conversation

Researcher:

Please perform stepwise regression analysis on the experimental results (with a p-value threshold of 0.05 for term selection) and provide predictive models for each response variable. Additionally, please identify the most influential terms in the models, interpret their practical significance, and suggest methods to improve the response variables.

Upload files: Experiment results | Parameter condition | Material condition

ChromR:

Agent D execution result is as follows:

Based on the stepwise regression analysis results, the predictive models for each response variable, the most influential terms, and methods to improve the responses are as follows:

1. Lactone Purity
Model:
lactone_purity = 3.1167 − 10.7425 × TT_concentration + 0.7026 × sample_flow + 3.9301 × washing_time + 1.8364 × washing_flow
Coefficient of Determination (R^2): 0.8538

Key Influential Terms:
- **washing_time (3.9301):** Has a significant positive effect on lactone purity.
- **washing_flow (1.8364):** Also has a significant positive effect on lactone purity.

Improvement Strategies:
- Increasing washing_time and washing_flow can effectively enhance lactone purity.

2. Flavonol Glycoside Purity
Model:
flavonol_glycoside_purity = 0.4002 − 50.0259 × TT_concentration + 18.7848 × washing_time + 9.5709 × washing_flow + 3.8743 × elution_time + 2.9674 × elution_flow
Coefficient of Determination (R^2): 0.9574

Key Influential Terms:
- **washing_time (18.7848):** Has a very strong positive impact on flavonol glycoside purity.
- **washing_flow (9.5709):** Also significantly improves flavonol glycoside purity.

Improvement Strategies:
- Increasing washing_time and washing_flow can effectively improve flavonol glycoside purity. Additionally, moderately increasing elution_time and elution_flow will also be beneficial.

Continuation

图 S15。通过 ChromR 进行逐步回归分析。41

Conversation

Continuation

ChromR:

3. Lactone Productivity
Model:
lactone_productivity = 32.8842 + 17.5758 × (sample_flow × elution_flow) − 2.3691 × sample_time − 23.4526 × sample_flow − 8.2481 × elution_flow − 5.7581 × washing_time − 9.4687 × (sample_flow × washing_time) + 20.6692 × (sample_time × sample_flow)
Coefficient of Determination (R^2): 0.8377

Key Influential Terms:
- **(sample_flow × elution_flow) (17.5758):** Positively affects lactone productivity.
- **sample_flow (−23.4526):** Has a significant negative effect on lactone productivity.

Improvement Strategies:
- Enhance the interaction between sample_flow and elution_flow (i.e., increase both simultaneously) to boost lactone productivity. However, sample_flow should be carefully controlled to avoid its strong negative impact.

4. Flavonol Glycoside Productivity
Model:
flavonol_glycoside_productivity = 47.4225 + 49.6461 × (sample_flow × elution_flow) + 13.6240 × sample_time − 18.6213 × sample_flow − 10.5198 × elution_flow − 22.3586 × washing_time − 26.3807 × (sample_flow × washing_time) + 58.9702 × (sample_time × sample_flow)
Coefficient of Determination (R^2): 0.9322

Key Influential Terms:
- **(sample_flow × elution_flow) (49.6461):** Has a significant positive effect on flavonol glycoside productivity.
- **(sample_time × sample_flow) (58.9702):** Also has a strong positive effect on flavonol glycoside productivity.

Improvement Strategies:
- Increase both the interaction of sample_flow with elution_flow and the interaction of sample_time with sample_flow to significantly enhance flavonol glycoside productivity.

Overall Evaluation:
- The models for **lactone purity** and **flavonol glycoside purity** show good fit, especially flavonol glycoside purity with an R^2 of 0.9574, indicating the model explains data variation well.
- The models for **lactone productivity** and **flavonol glycoside productivity** include interaction terms, but some coefficients may have relatively high p-values, suggesting these terms might not all be statistically significant. Further validation or model refinement is recommended.
- **washing_time** and **washing_flow** consistently show significant positive effects across multiple responses; thus, increasing these parameters is an effective strategy to improve product quality in practice.

图 S16。图 S11 的续图。

ChromR 使用二阶数学模型为每个指标建立了响应面模型，其定量模型见公式 S1。

$$Y = b_0 + \sum_{i=1}^{n} b_i X_{c,i} + \sum_{i=1}^{n} b_{ii} X_{c,i}^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} b_{ij} X_{c,i} X_{c,j} + \sum_{k=1}^{m} c_k Z_{c,k} \qquad (S1)$$

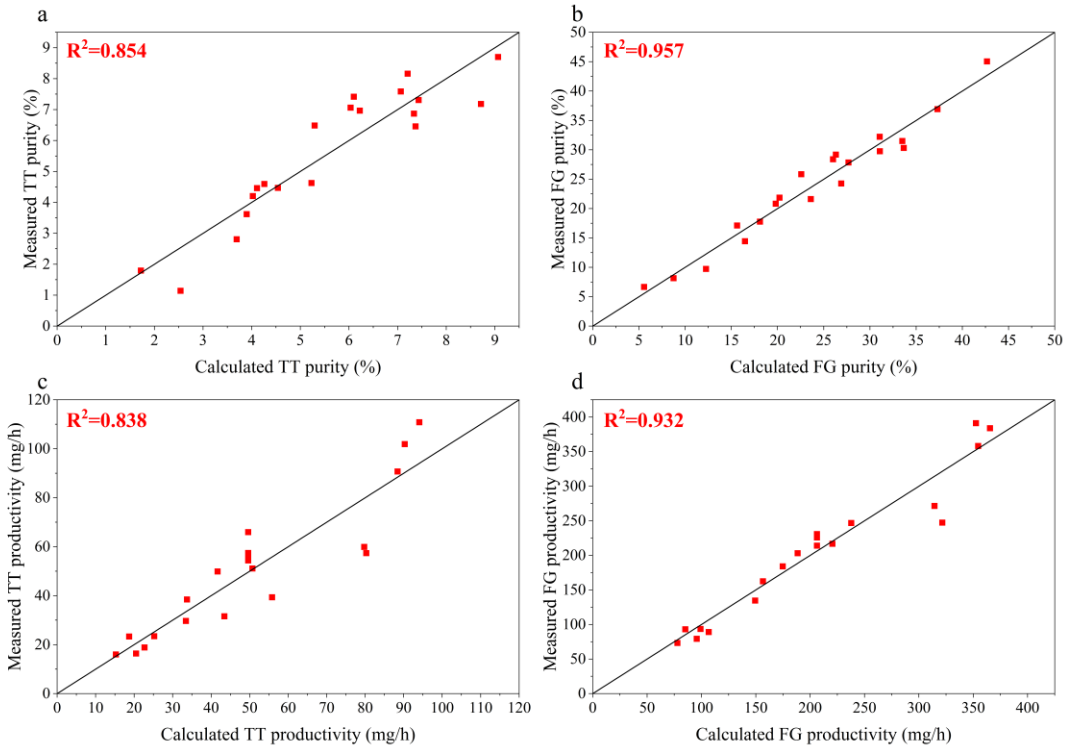
其中，n 和 m 分别为工艺参数和材料属性的数量；$X_{c,\ i}$、$X_{c,\ j}$ 为工艺参数；$Z_{c,\ k}$ 为材料属性；b_0 为常数项；b_i、b_{ii}、b_{ij} 和 c_k 为偏回归系数。ChromR 首先进行了逐步回归，以筛除不重要的项，随后进行了偏最小二乘拟合。各指标对应的模型见表 S2。工艺参数标记如下：进料流速（BV/小时）记为 X_1，，进料时间（小时）记为 X_2，，洗涤流速（BV/小时）记为 X_3，，洗涤时间（小时）记为 X_4，，洗脱流速（BV/小时）记为 X_5，，洗脱时间（小时）记为 X_6。材料属性标记如下：TT 浓度（mg/毫升）记为 Z_1，，TT 纯度（%）记为 Z_2，，FG浓度（mg/毫升）记为 Z_3，，FG纯度（%）记为 Z_4。响应指标标记如下：TT 纯度（%）记为 Y_1，，TT生产率（mg/小时）记为 Y_2，，FG纯度（%）记为 Y_3，，FG生产率（mg/小时）记为 Y_4。

表 S2。各

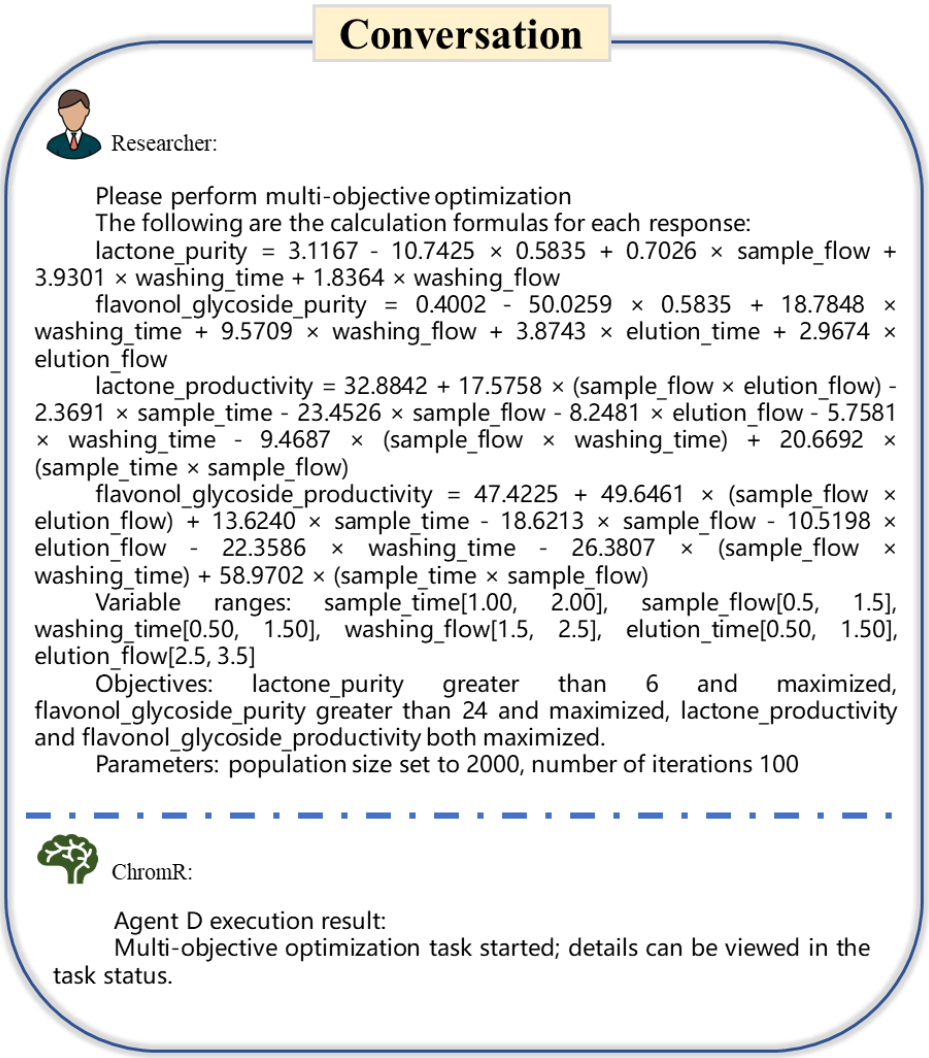
指标模型。

项	Y_1	Y_2	Y_3	Y_4
截距	3.1167	32.8842	0.4002	47.4225
Z_1	-10.7425		-50.0259	
X_1	0.7026	-23.4526		-18.6213
X_2		-2.3691		13.6240
X_3	1.8364		9.5709	
X_4	3.9301	-5.7581	18.7848	-22.3586
X_5		-8.2481		-10.5198
X_6			3.8743	
X1 × X5		17.5758		49.6461

术语	Y_1	Y_2	Y_3	Y_4
X1 × X4		-9.4687		-26.3807
X1 × X2		20.6692		58.9702
R^2	0.8538	0.8377	0.9574	0.9322



图S17. 各项指标的测量值和计算值。(a) 萜类内酯纯度。(b) 黄酮苷纯度。(c) 萜类内酯生产率。(d) 黄酮苷生产率。



图S18. ChromR多目标优化。

表S3. 进料溶液的帕累托最优解和预测指标值

250401和250409。

批次编号 feed 溶液	溶液 No.	X_1 (BV/小时)	X_2 (h)	X_3 (BV/小时)	X_4 (h)	X_5 (BV/小时)	X_6 (h)	Y_1 (%)	Y_3 (%)	Y_2 (毫克/小时)	Y_4 (毫克/小时)
250401	1	1.50	2.00	2.50	1.10	3.50	0.860	6.80	29.4	96.5	380
	2	1.50	2.00	2.50	1.33	3.50	0.501	7.74	32.5	91.7	365
	3	1.50	2.00	2.50	1.29	3.50	1.23	7.58	34.6	92.6	367

批次数量 feed 溶液	溶液 No.	X_1 (BV/小时)	X_2 (h)	X_3 (BV/小时)	X_4 (h)	X_5 (BV/小时)	X_6 (h)	Y_1 (%)	Y_3 (%)	Y_2 (毫克/小时)	Y_4 (毫克/小时)
	4	1.50	2.00	2.50	0.892	3.50	0.515	6.00	24.3	101	392
	5	1.50	2.00	2.50	1.40	3.50	1.33	7.99	36.9	90.4	361
250409	1	1.50	2.00	2.50	1.50	3.50	1.50	8.56	40.3	88.4	355
	2	1.50	2.00	2.50	1.31	3.50	1.50	7.83	36.8	92.2	366
	3	1.50	2.00	2.50	1.49	3.50	0.815	8.52	37.5	88.6	355
	4	1.50	2.00	2.50	1.45	3.50	0.501	8.38	35.6	89.4	357
	5	1.50	2.00	2.50	1.10	3.50	1.50	6.99	32.8	96.4	379

表 S4。验证实验的工艺参数。

进料批次编号 溶液	X1 (BV/小时)	X2 (小时)	X3 (BV/小时)	X4 (小时)	X5 (BV/小时)	X6 (小时)	设计空间中的位置 设计空间中的位置
250401	1.5	2	2.5	1.095	3.5	0.86	空间内部
250409	1.5	2	2.5	1.49	3.5	0.81	内部
231201	0.75	1	1.75	0.5	3.25	0.5	外部

参考资料

[1] J. Li、Y. Wang、J. Du 等，基于分析质量源于设计法定量分析银杏叶浓缩物中9种成分的高效液相色谱-质谱方法，《液相色谱与相关技术杂志》48（2025）484-493https://doi.org/10.1080/10826076.2025.2507957。

[2] N.P. Commission，《中华人民共和国药典》。编者，《中华人民共和国药典》图书，《中华人民共和国药典》丛书，中国医药科技出版社，北京，2025。