·专 论·

基于数据科学的中药膜过程研究的思考与实践

钟文蔚[1,2]，纪 超[3]，郭立玮[1,4*]，李 博[4]，徐雪松[4]

1. 广州中国科学院先进技术研究所，广东 广州 511458
2. 广州南沙资讯科技园博士后科研工作站，广东 广州 511458
3. 新南威尔士大学（宜兴）环境技术转移中心，江苏 宜兴 214200
4. 南京中医药大学，江苏 南京 210038

摘 要：“数据科学”是关于数据的科学，其宗旨为探索数据界奥秘的理论、方法和技术。鉴于中药膜分离过程是一个非线性系统，其工艺数据具有多变量、非线性、强噪声、自变量相关、非正态分布、非均匀分布等全部或部分特征，“数据科学”已成为中药膜过程复杂系统探索、技术创新的重要武器。“数据科学”引入中药膜科技领域的技术关键：（1）可精准表征中药膜过程传质的特征技术体系的建立；（2）“分子模拟”技术对膜传质过程的动态描述；（3）计算流体力学在膜领域的应用；（4）强大、先进的数据处理技术。并以课题组多年开展的基于数据科学的中药膜过程研究为实例进行阐述。

关键词：中药；膜过程；数据科学；绿色制造；分离机制

中图分类号：R283.6 **文献标志码：**A **文章编号：**0253 - 2670(2020)01 - 0001 - 08

DOI: 10.7501/j.issn.0253-2670.2020.01.001

Thoughts and practices on studies of membrane processes for Chinese materia medica based on data science

ZHONG Wen-wei[1,2], JI Chao[3], GUO Li-wei[1,4], LI Bo[4], XU Xue-song[4]

1. Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences, Guangzhou 511458, China
2. Guangzhou Nansha Information Technology Park Post-doctoral Scientific Research Station, Guangzhou 511458, China
3. UNSW Centre for Transformational Environmental Technologies (CTET), Yixing 214200, China
4. Nanjing University of Chinese Medicine, Nanjing 210038, China

Abstract: Data science is a form of data-oriented science, which serves as a set of theories, methodologies and technologies for data exploration and analysis. Considering that membrane separation process for Chinese materia medica (CMM) manufacturing is a non-linear system, the data obtained regarding its process can be either multivariate, non-linear, strong noise, non-normally distributed, or non-evenly distributed. Due to the special features of data science, research has shed lights on its application in the scientific exploration and technological innovation in the complex membrane processes based on CMM system. There are a few key aspects of data science that are worth mentioning, making it competent as an analytical tool for CMM manufacturing integrated membrane processes. They can be recognized as the following, such as (1) the ability to precisely describe the mass transfer properties of CMM integrated membrane processes; (2) the dynamic description of membrane mass transfer process by molecular simulation; (3) the application of computational fluid dynamics simulation in membrane technology, and (4) the powerful and advanced data processing technology. This paper also reviewed some real case studies encountered by the authors during the investigation of membrane technology for CMM manufacturing based on data science.

Key words: Chinese materia medica; membrane processes; data science; green manufacturing; separation mechanism

收稿日期：2019-11-17

基金项目：国家自然科学基金项目（30572374）；国家自然科学基金项目（30171161）；国家自然科学基金项目（30873449）；国家自然科学基金项目（81274096）；国家科技部“十五”科技攻关计划项目（2004BA721A42）；国家科技部“十一五”科技支撑计划项目（2006BAI09B07-03）；2017 中国工程院咨询项目（2017-XZ-08）；广州市科技项目（201904010054）

作者简介：钟文蔚，博士，助理研究员，主要研究方向为膜浓缩，中药制药工程。E-mail: ww.zhong@giat.ac.cn

*通信作者 郭立玮，南京中医药大学教授、广州中国科学院先进技术研究所客座研究员，从事以膜技术为主体的中药复方分离工程研究。E-mail: guoliwei815@126.com

近年来，随着信息技术的迅猛发展，以数据科学（date science）为核心的计算化学技术展现出广阔的前景。与此同时，计算机、激光、微生物及电子技术等被引进分离过程，成为现代分离科学的重要标志之一。融合计算机技术、网络技术、数学、化学及其相关学科的最新理论、集成多种关键软件的科技平台的出现，给面向制药分离工程的中药膜过程研究领域带来无限活力。

“数据科学”是关于数据的科学，其宗旨为探索数据界奥秘的理论、方法和技术，而将“信息”和“知识”转换为“数据”，是实施“数据驱动”策略，探索自然规律的关键技术[1-3]。就中药膜技术与“数据科学”的关系而言，随着膜分离技术在中药制药行业的广泛应用，迫切需要在膜过程中针对膜的污染程度进行即时分析和预测的综合分析系统，以便根据分析结果对中药体系进行相应的预处理，并制定适当的膜清洗方案。传统的分析方法主要基于统计学理论，单一使用回归分析、主成分分析等方法。但中药水提液是一个复杂系统，在膜工艺过程实验中采集到的关于中药水提液原液、提取液、膜分离过程等指标参数达 30 多个，这些表征数据具有多变量、非线性、强噪声、自变量相关、非正态分布、非均匀分布等全部或部分特征。“数据科学”研究领域的特征提取、遗传算法、神经网络、支持向量机等算法为上述复杂数据的分析和建模预测提供了新的技术手段。本文主要讨论“数据科学”在探索中药膜过程规律中的应用。

1 数据科学——中药膜过程复杂系统探索、技术创新的重要武器

1.1 中药膜过程的复杂系统特征

1.1.1 中药膜过程的数据结构组成 中药水提液是中药制药行业最普遍使用的物料。然而，中药水提液的膜过程却有着“谜”一般的表现。如理论上高分子物质不能透过孔径远小于其相对分子质量的膜，但实际的膜过程中却常常发现淀粉、果胶等几万、甚至十几万的相对分子质量的高分子物质出现在截留相对分子质量为 1 万，甚至 1 000 的超滤膜透过液中，严重限制了膜获取整体药效物质技术优势的发挥。又如，各中药水提液样品中，淀粉、果胶等“非药效共性高分子物质”均占很大比例（尤以淀粉、果胶等碳水化合物类为主），且是影响水提液的物理化学性质表征参数及导致膜通量衰减的主要因素[4]，因而起着“膜对抗作用”，成为导致膜污染的主要因素。而因为中药膜传质过程的化学物质高维多元，中药膜污染难于以常规数学模型进行预报、优化与监控，成为一种难以破解的“谜”[5]。

为了攻克中药水提液组成极其复杂，长期以来因缺乏密度、黏度、表面张力、导热系数、扩散系数等基本的物性数据，而造成的中药生产工艺设计难以与新型分离技术作用机制兼容的瓶颈，笔者课题组通过“分离科学”理论推导及以 200 多种具代表性的中药及其复方的膜过程所采集的上万个大样本的数据挖掘，构建了以下 3 类参数组成的中药膜过程数据集结构。

（**1**）与中药物料理化性质相关的数据

①中药水提液中“非药效共性高分子物质”的化学组成与含量：中药水提液中无一例外的均有大量构成各组织、器官细胞壁的成分及所贮藏的营养物质，如淀粉、果胶等“非药效共性高分子物质”。它们的热力学、动力学与电化学性质是影响膜过程的主要因素，因而可被视为膜对抗物质。因处方与提取工艺不同，各中药水提液体系中“非药效共性高分子物质”占有不同的比例，采用相对准确的化学分析方法测定它们的含量，可“定量”研究它们在不同膜过程中对膜结构与膜动力学参数的作用。

②物料“溶液环境”特征参数：根据分离科学一般理论和膜科学原理，物料体系的黏度、密度、浊度、电导、pH、粒径分布及不同溶质的浓度、化学势、淌度、平均相对分子质量（$\overline{M}$）、分子大小与形状等物理化学参数都可能对分离过程产生影响，它们共同构成了可科学地表征中药水提液对膜污染产生影响性质的集合。

③中药指标成分分子结构参数：基于药物分子本身的性质，常用的分子结构参数包括理化参数（如疏水性参数、沸点、熔点、NMR 谱、IR 谱等）；空间参数（包括二维、三维分子描述符）；电性参数（如 Hammett 电效应参数 σ）；量子化学参数（包括电荷参数及能量参数）等。

借助计算机化学技术，可通过中药药效成分的分子结构参数预测膜过程对目标物质超滤的透过/截留率，用以开展有关分离机制研究，为指导大规模实验以及生产实践提供科学依据。

④中药指标成分“溶液结构”特征参数[6]：中药物料体系中小分子与高分子物质在溶剂化过程中相互作用可形成“溶液结构”的微观结构，从而对膜过程产生影响。与“溶液结构”相关的参数也是

它们在溶剂化过程中相互作用而形成的，“溶液结构”表征参数主要有分子形态、粒径及其分布、结构构象、红外光谱等。此类数据可用于探索化学成分的空间结构与膜孔径的位阻作用及其机制。

（2）与膜材料微结构相关的数据：膜材料微结构可以膜孔径（d_m）和孔隙率（ε）等表征[7]。

①d_m：多孔膜中，孔的直径，评价膜分离功能的重要指标之一。

②ε：多孔膜中，孔隙的体积（$V_{孔}$）占膜的表观体积（$V_{膜表观}$）的百分数（$\varepsilon=V_{孔}/V_{膜表观}$）。

（3）与膜功能、膜传质过程及其机制相关的数据：与膜功能相关的数据主要有中药指标成分膜透过率、膜通量等。与膜传质过程及其机制密切相关的主要有膜阻力分布与膜污染度等膜过程特征表征参数。

①中药指标成分膜透过率：参照《中国药典》2015 年版技术要求，以 HPLC 等方法检测指标性成分及指纹图谱。

②膜通量：是膜分离过程的一个重要工艺运行参数，是指单位时间内通过单位膜面积上的流体量。

③膜阻力分布：在膜分离过程中溶剂或溶质透过速率的降低是由于膜的存在而引起的，则称为膜阻力，可细分为膜自身阻力、表面沉积阻力、膜堵塞阻力、浓差极化阻力等。

④膜污染度：表达膜过程中，污染物质在膜表面或膜孔内吸附、沉积造成膜孔径变小或堵塞，使膜产生透过流量与分离特性的不可逆变化现象的程度，膜污染度计算方法为初始纯水膜通量与被污染后稳定通量之差除以初始纯水膜通量的百分比。

1.1.2 中药膜过程复杂系统的特点 王永炎院士[8]指出：中医药研究所面临的是一个复杂巨系统。中药膜过程作为中医药复杂巨系统之一，具有“复杂数据”特征。

（1）具有“复杂数据”特征：中医药体系“复杂数据”具有多变量、变量相关、非均匀分布、非高斯分布等部分甚至全部特征，从而给构造模型、寻找规律造成很大的困难[9]。

（2）具有“适应性系统”的功能：复杂系统具有“适应性系统”的功能，这一系统是由许多平行发生作用的结构组成的网络。每一个复杂的适应性系统都具有多层次的组织，每一个层次的作用者对于更高层次来源的控制力并非集中而是分散。

（3）可通过模型模拟进行预测：复杂系统的预测还可以通过模型模拟进行研究。在已知的药味组成和临床有效的结果面前，认识复方的复杂系统，揭示中间的作用过程正是目前研究的目的。而所有复杂的适应性系统，都能建立其预测世界的模型[10]。

1.2 以数据科学解读中药膜过程是“中药制药”学科创新的要求

由于中药物料的特殊性等诸多因素的制约，目前中药制药工程理论研究和工艺技术的应用还处于粗放式的初级阶段，普遍存在提取工艺优化设计缺乏精准科学依据。如中药提取、精制过程中所涉及的流体力学过程、传热过程、传质过程的基本理论及工艺流程和生产装置至今尚处于“套用”相关领域学科知识的阶段。中药制药生产醇沉法、絮凝澄清法及大孔树脂吸附法等精制技术在安全性、有效性及技术经济指标等方面均不尽人意[11-12]。中药制药学科多年鲜见突破性进展。究其原因是多方面的，但中药制药领域的科学研究仍以“描述发现”为主，而未打响“机制探索”的攻坚战是重要因素之一。

1.2.1 现代膜科学技术的特点是引进以信息技术为代表的高新技术 信息技术在膜分离过程中的运用涉及到的热力学和传递性质、多相流、多组分传质、分离过程和设备的强化和优化设计等，如分子模拟大大提高了预测热力学平衡和传递性质的水平[13-14]；化工模拟软件的商品化、CAD（computer-aided design）和人工智能（artificial intelligence，AI）在化工中的广泛应用大大推动了分离过程和设备的优化设计和优化控制；信息技术和先进测试技术的高速发展为分离科学多层次、多尺度的研究提供了条件[15]。分离过程的研究已从宏观传递现象的研究深入到气泡、液滴群、微乳和界面现象等，加深了对分离过程中复杂传递现象的理解。功能齐全的计算流体力学（computational fluid dynamics，CFD）软件可以对分离设备内的流场进行精确地计算和描述[16]。实验研究和计算机模拟相结合成为分离技术研究开发和设计放大的主要途径。

如上所述，如此丰富的分离技术“信息”和“知识”都是以“数据”为符号或载体表达的；同时，这些以先进的仪器设备获取的“信息”和“知识”也只有以“数据”的形式，才可能被数据科学所接受，从而被加工为研究者所需要的新知识。

1.2.2 引进“数据科学”是中药膜过程由经验科学向量化科学过渡的必由之路 李静海院士[17]指出：使用多尺度的方法来描述微观、介观和宏观

上的物理变化是制药工程由经验科学向量化科学过渡的关键。

目前，制药工程已从传统总体性质的测量和关联，转向在分子和介观尺度上的观测和模拟。在微观层次上建立模型、模拟和定量分析，根据要求设计和生产产品，以实现从分子尺度到过程尺度的跨越。而“数据科学”则是实现这种跨越必不可少的利器。以“分离”为基本要素的中药制药生产过程，即是分子尺度的复杂药效组分的传递和再分布过程，反映在宏观过程尺度即是物料，如植物的药用组分在“场-流”条件下的能量交换或物质转运过程。因而，以分离目标为引领的中药复方分离过程研究是中药制药学科实现重大创新的突破口。

笔者课题组建立了 218 种中药单方、复方的 1 万多个膜过程数据挖掘基础上的“陶瓷膜精制中药的膜污染预报与防治系统”[18]，可对不同中药物料实现“表征参数检测-膜污染预报-提供优化治理方案”的个体化污染控制模式。该研究揭示了中药水提液在宏观与微观尺度膜分离过程的共性特征，使中药膜过程研究取得突破性进展；证明引进“数据科学”是中药膜过程由经验科学向量化科学过渡的必由之路。

2 “数据科学”引入中药膜科技领域的技术关键

为精准、动态描述膜传质过程，须采用实验研究与理论模型互补，宏观分析与微观表征并用，实现中药制药工程体系研究领域的多学科跨越。其中，数据科学引入中药膜科技领域的技术关键涉及以下内容。

2.1 可精准表征中药膜过程传质特征的技术体系的建立

中药膜传质过程精准表征技术体系的建立是中药膜科技研究进入“数据科学”领域的首要条件，也是“数据科学”研究结论科学、客观的保障。

2.1.1 表征方法的先进性与多样性及其优化组合、互相印证

（**1**）“溶液结构”研究手段的先进性与多样性：应用多种先进的形态研究方法（原子力显微镜、高分辨核磁共振等手段及计算机仿真技术），模拟不同尺度的小分子药效物质“溶液结构”特征；同时，应用多种先进的功能研究方法（储能模量、浊度、紫外/荧光光谱等）从不同侧面系统研究中药体系中大、小分子聚集体的分子内/分子间缔合行为；“形态研究”与“功能研究”方法的合理组合、优选及其研究结果的相互印证。

（**2**）膜结构参数计量、评价方法的可靠性与特异性：多种测定膜孔径、孔隙率以及膜孔密度等膜结构参数的方法（扫描电镜、渗透率法等）取长补短，所得电镜图片、数据经专业分析软件分析，优化建模，尽可能建立逼近真值的修正因子（模型）。

2.1.2 膜检测与诊断的先进技术 膜检测与诊断是评估膜通量降低和完整性问题原因的重要手段，而膜通量及其完整性与工艺终端产品成本及质量直接相关。澳大利亚新南威尔士大学（UNSW）联合国教科文组织膜技术研究中心（UNESCO Centre for Membrane Science and Technology）从“膜检测新视角”出发，研发出一整套膜诊断的先进技术，将液相色谱-有机碳分析仪（LCOCD），比表面积检测仪（BET）和场发射电子显微镜（FESEM）联用，鉴定膜对目标生物聚合物的去除效能，而传统检测技术对该生物聚合物无法量化。配合 Fujiwara 测试（确认膜表面是否被卤元素如氯或溴氧化），可判断卤化物对膜的损害与否。

2.1.3 实现“图状结构-性质参数”转换的定量构效关系（QSAR）手段 要进入“数据科学”领域，首先涉及到如何将药物分子的存在形态、空间结构特征转换为数值表征的问题。为此，需要引入定量构效关系（quantitative structure activity relationship，QSAR）技术手段。

化合物的性质取决于化合物的结构，即化合物的结构与其性质/活性具有相关性。如图 1 所示，化合物的结构是非数学量，要想建立某化合物结构与其性质/活性的相关性，则需要由结构图提取特征，并运用这些特征（作为变量）去构造数学模型，进而运用所构造的数学模型预测未知化合物在膜过程中的表现。

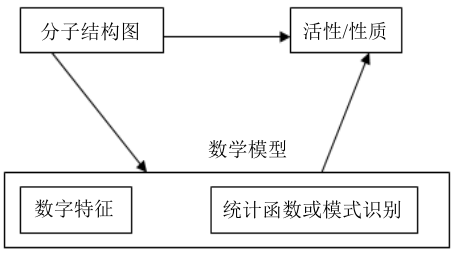


图 1 多元定量结构-活性/性质相关研究方法

Fig. 1 Multivariant research method for quantity analysis of structure-activity and characteristics

2.2 “分子模拟”技术对膜传质过程的动态描述

由于膜材料微孔体系的空间限制，其中流体的行为与性质难以通过实验观察和测定，而相关研究又具有重要理论意义，分子模拟技术正快速进入材料及化学工程等领域。该技术利用计算机以原子水平的分子模型来模拟分子的结构与行为、分子体系的各种物理化学性质，包括分子体系的动态行为，如氢键的缔合与解缔、吸附、扩散等[19-21]。本文涉及的中药“溶液结构”在膜过程中的动态表现及其对膜微结构的作用等研究内容，均可采纳目前探索的多尺度复杂现象的有效方法：实验工作先行，继以扫描探针显微镜技术佐证，最后用分子模拟技术研究机制并反馈进一步实验研究的方向和方法[22]。

2.3 计算流体力学（CFD）在膜领域的应用

2.3.1 CFD 预测流体运动规律 CFD 是通过数值方法求解流体力学控制方程，并以此预测流体运动规律的技术，具有成本低、速度快、资料完备、风险小等优点[23-26]。针对中药体系膜过程复杂的流场分布，引进 CFD 方法，以增加优化设计的可信度已成为制药分离技术领域的重要手段之一[27]。

2.3.2 膜器件的优化设计新手段 UNSW 联合国教科文组织膜技术研究中心拥有世界领先的膜组件设计及优化技术。其中包括：（1）利用数学模型及实验方法对不同膜组件的污染情况进行预测与优化，以最大程度地减少中试的成本和时间。（2）世界唯一的、使用 CFD 以及化工过程设计软件 Aspen Plus，设计节能高效的膜蒸馏组件和过程的技术。（3）中空纤维膜的强度及断裂原因分析技术。可通过优化膜丝自身的几何参数，如膜丝长度、松散度、最大位移、直径、注胶强度和方法，设计机械强度最优膜丝，延长膜丝的使用年限等。

2.4 强大、先进的数据处理技术

近年来，本课题组与上海大学陆文聪教授课题组合作，将建立在统计学习新理论基础之上的支持向量机（support vector machine，SVM）方法[28]应用于中药膜过程预测，通过调节 SVM 模型所选用的核函数及其参数以控制“过拟合”或“欠拟合”现象，从而较好地解决中医药体系复杂数据“建模结果好”而“预报结果不好”的问题，因而 SVM 有望成为中医药复杂体系数据挖掘和知识发现的新方法。

本课题组在“陶瓷膜精制中药的膜污染预报与防治系统”软件编制中涉及的算法超过 20 种。其中包括：最近邻（K-nearest neighbor，KNN）、主成分分析（PCA）、多重判别矢量（MDV）、白化变换（Sphere）、白化线性映射（Lmap）、球形映射（Lmap）、逆传播人工神经网络（back propagation-artificial neural network，BP-ANN）、最佳投影识别（optimal map recognition，OMR）、逐级投影（hierachical projection）、支持向量机回归、最佳投影回归（optimal projection regression）、核函数、支持向量机分类、超多面体（hyper-polyhedron）[29]等。

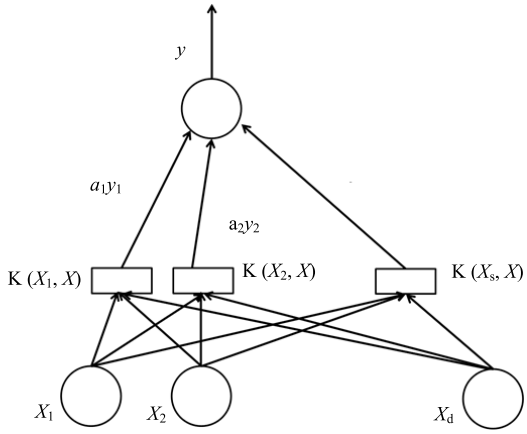
3 基于“数据科学”手段探索中药膜过程及其机制的研究实践

南京中医药大学中药膜科技团队在近 20 年来的研究中，比较深入、系统地开展了“基于数据科学的中药膜过程研究”。近年来，又通过与广州中国科学院先进技术研究所、江苏久吾高科技公司的“产学研”合作，不断深化该领域的研究。

3.1 建立在中药水提液理化性质表征技术基础上的膜过程优化研究

通过以支持向量分类算法为主的数据挖掘技术，研究中药水提液的理化数据（如黏度、密度、浊度、电导、pH、粒径分布等）及其中所含各种物质与膜通量之间的关系，从物理化学角度考察中药的膜分离过程，为科学地分离中药提供理论基础。利用支持向量网络对未知样本的类别属性（y）进行预报的示意图见图 2[30]。

初步研究结果表明，模式识别、SVM 等数据挖掘方法可以作为中药水提液复杂体系有效的数据处



输入向量 $X=(X_1, X_2, \cdots, X_n)$ 为第 j 个支持向量 X_j 与输入向量 X 的内积

Input vector $X=(X_1, X_2, \cdots, X_n)$ as the inner product of j^{th} supportive vector X_j and input vector X

图 2 SVM 模型预报未知样本类别图

Fig. 2 Unknown sample category graph predicted by support vector machine (SVM) model

理手段，并得到了适应 Al_2O_3 陶瓷微滤膜处理中药水提液的预报正确率高（或误报率低）、比较稳定的相关模型，但模型中目标变量和因变量之间的因果关系还有待进一步研究。

3.2 人工神经网络与支持向量机方法预测膜过程

以中药水提液膜过滤中得到的实验数据为对象，综合应用遗传算法、神经网络、支持向量机法对影响膜污染度的主要因素进行即时分析和预测。所构建的“基于特征提取的中药水提液膜分离预测系统”为实时预报系统，可根据膜分离前中药物料的物理化学参数、高分子物质含量及膜阻力分布数据等，实现不同数据源的信息处理和不同时效的膜污染预报，为不同中药体系实现“表征参数检测-膜污染预报-提供优化治理方案”模型下的个体化膜污染控制提供一种普适模式[31]。由表 1 可知，SVM、RBF（radial basis function）-ANN 都具有很好的泛化能力，BP-ANN 次之。

表 1 SVM、BP-ANN、RBF-ANN 3 种方法的比较

Table 1 Comparison of SVM, BP-ANN, and RBF-ANN

预测方法	相关系数	均方误差
SVM	0.968 5	0.000 3
BP-ANN	0.741 5	0.006 7
RBF-ANN	0.951 4	0.000 9

根据上述计算机化学方法建立的系统结构为 4 个部分：数据文件的获取和处理；特征因素筛选；预报模型优化和建立；预报结果输出，确定原液预处理方案和膜清洗方案。中药水提液膜分离预测系统结构见图 3。

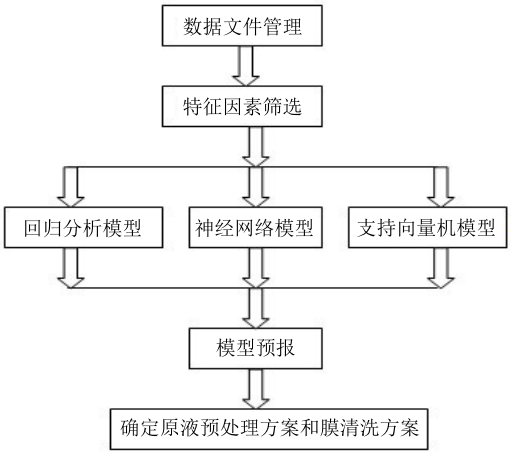


图 3 中药水提液膜分离预测系统结构

Fig. 3 Structure of membrane separation prediction system for water extract of CMM

该系统建立的支持向量机模型预测误差为 3.4%，较单一系统预测准确率高 1.1%，从而较好地解决了单一使用回归分析、主成分分析等方法预测误差大，难以有效进行膜污染预测及制定水提液预处理及膜清洗方案的工程难题。

3.3 SVM 算法用于中药挥发油含油水体超滤通量预测的研究

该研究选择 40 组数据进行模型参数的优化和训练，并对 10 组实验的稳定通量进行预测。同时，对 SVM 算法与 BP-ANN 算法的运行结果进行比较。结果表明，在该实验条件下 SVM 算法的预测能力显著强于 BP-ANN[32]。

应用设计好的算法对训练数据进行训练，MSE 达到 0.027 0，回归系数（r）为 0.850 1。采用该算法对测试数据进行预测。将实际值与预测值（包括 BP-ANN 预测值与 SVM 预测值）进行对比，结果见图 4。

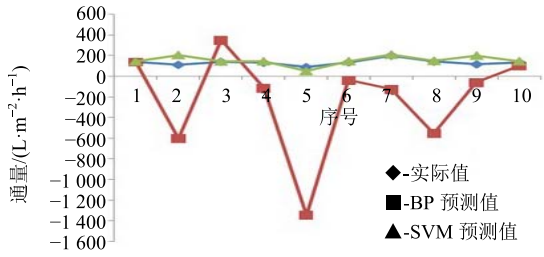


图 4 SVM 和 BP-ANN 算法的预测值与实际值的比较

Fig. 4 Comparison of predicted value and actual value based on SVM and BP-ANN algorithms

从实验结果看，SVM 算法以统计学习理论为基础，不涉及概率测度及大数定律等，可用于小样本的研究，它以训练误差作为优化问题的约束条件，以置信范围值最小化作为优化目标，故逼近能力和推广能力兼优，克服了神经网络方法在理论上的缺陷。由实验结果明显可见，其预测准确度较 BP-ANN 显著提高。

3.4 超滤膜对生物碱类等物质的透过/截留及其定量结构关系的研究

本课题组董洁等[33]先获取了生物碱类（小檗碱、巴马汀等）与环烯醚萜类（栀子苷、京尼平苷等）等 20 种中药成分在 5 种超滤膜（CA-1K、PS-1K、PES-1K、PS-3K、PES-3K，其中，CA 为醋酸纤维素膜、PS 为聚砜膜、PES 为聚醚砜膜，后缀 1K、3K 表示膜截留相对分子质量 1 000、3 000）过程中的透过率。再根

据膜科学理论，通过 Chemoffice 等软件计算和查阅资料得到可能影响膜透过率的 27 个结构参数，包括辛醇/水分配系数（Alog*P*）、偶极距（μ）、极化率（α）、相对分子质量（M_W）、摩尔体积（V_M）、表面积（*S*）等。最后采用偏最小二乘判别、SVM 和人工神经网络等作为建模方法，并结合线性相关、投票法、超多面体法删除等多种方法进行变量筛选，建立了上述生物碱类和环烯醚萜类物质共 8 种化合物的 5 种超滤膜透过率定量构效关系模型（表 2），建模过程使用 Master1.0 数据挖掘软件实现。

从选入模型的参数可以看出，在超滤膜分离化合物的过程中，影响透过率的因素主要包括化合物的自身性质（如得失电子能力、亲水/疏水性等）与膜性质的相互作用以及化合物的空间结构。

表 2 5 种超滤膜的构效关系模型

Table 2 Structure-activity relationship models of five ultrafiltration membranes

超滤膜	回归模型	建模方法	相关系数
CA-1K	Y=4.849 856 [CMR]−3.882 136 [LUMO]+11.424 632	SVM	0.933
PES-1K	Y=61.740−5.217 [LUMO]	PLS	0.989
PS-1K	Y=−40.606−22.670 [Alog$P$]−19.804 [LUMO]−87.665 [K&H_2] +131.533 [K&H_3]	PLS	0.996
PES-3K	Y=127.633−5.118 [ROG]−1.334 [L_X]	PLS	0.984
PS-3K	Y=103.634−0.917 [Alog$P$]−0.316 [S_{YZ}]	PLS	0.980

以青藤碱对表 2 中各模型的验证结果表明，上述各种超滤膜的构效关系模型有较好的预测能力，并对中药药效物质膜截留机制的阐述具有重要作用，可为指导大规模实验以及生产实践提供科学依据。

4 结语

基于数据科学的中药膜过程研究实践证明将数据科学引进中药制药学领域的重要性与迫切性。其意义和创新之处主要体现在以下 3 方面。

4.1 初步建立起计算机化学在中药制药学领域应用的基本模式

（1）大样本中医药体系的选择；（2）与中药制剂学或生物药剂学相关的技术参数表征体系的建立；（3）数据库设计与构建；（4）多种数据挖掘算法的筛选与相互印证；（5）知识发现——潜在规律的发现与验证。不但为中药及类似复杂体系的膜污染机制与防治提供了一种全新的研究模式，而且对探索中医药学新方法具有重要意义。

4.2 首建“中药水提液陶瓷膜污染基础数据库”

该数据库进一步丰富了中药工程学物性数据库，对制定和完善中药生产标准规范，提升中药工业整体工程技术水平具有重要启示作用。

4.3 首次提出对不同中药体系实现个体化膜污染控制模式的软件系统

该软件系统对拓宽计算机化学及膜科学在中药领域的应用，丰富计算机化学和膜科学的理论，探索化学计量学新方法具有重要意义。

参考文献

[1] 朝乐门. 数据科学 [M]. 北京: 清华大学出版社, 2016.

[2] Zhu Y Y, Zhong N, Xiong Y. *Data Explosion, Data Natrue and Dataology* [M]. Beijing: AMT-BI, 2009.

[3] 桑文锋. 数据驱动从方法到实践 [M]. 北京: 电子工业出版社, 2018.

[4] 郭立玮. 中药分离原理与技术 [M]. 北京: 人民卫生出版社, 2010.

[5] 郭立玮, 朱华旭. 基于膜过程的中药制药分离技术: 基础与应用 [M]. 上海: 科学出版社, 2019.

[6] 郭立玮, 陆 敏, 付廷明, 等. 基于中药复方小分子药效物质组“溶液结构”特征的膜分离技术优化原理与方法初探 [J]. 膜科学与技术, 2012, 32(1): 1-11.

[7] 徐南平, 李卫星, 邢卫红. 陶瓷膜工程设计: 从工艺到微结构 [J]. 膜科学与技术, 2006, 26(2): 1-5.

[8] 王永炎. 中医研究的三个重要趋势 [N]. 中国中医药报, 2005-03-04(06).

[9] 郭立玮, 付廷明, 李玲娟. 面向中药复杂体系的陶瓷膜污染机理研究思路与方法 [J]. 膜科学与技术, 2009, 29(1): 1-7.

[10] 王 阶, 王永炎. 复杂系统理论与中医方证研究 [J]. 中国中医药信息杂志, 2001, 8(9): 25-27.

[11] 郭立玮, 党建兵, 陈顺权, 等. 关于构建中药绿色制造理论与技术体系的思考和实践 [J]. 中草药, 2019, 50(8): 1745-1758.

[12] 丁 菲, 李除夕, 周 颖, 等. 基于“绿色设计”理念的中药制药膜分离工艺选择原则与方法 [J]. 中草药, 2019, 50(8): 1759-1767.

[13] O'Connell J, Neurock M. Trends in property estimation

for process and product design [A] // *Proceeding Coference on Foundations of Computer Aided Process Design* [C]. Breckenridge: AIChE Symp. Ser., 2000.

[14] Allen M P, Tildesley D J. *Computer Simulation of Liquids* [M]. New York: Oxford Univ Press, 1987.

[15] Delnoij E, Kuipers H A M, Swaaij W V, *et al*. Measurement of gas-liquid two-phase flow in bubble columns using ensemble correlation PIV [J]. *Chem Eng Sci*, 2000, 55(17): 3385.

[16] Baten J M, Krishna R. Modelling sieve tray hydraulics using computational fluid dynamics [J]. *Chem Eng J*, 2000, 77(3): 143.

[17] Li J H, Kwauk M. Exploring complex systems in chemical engineering—The multi-scale methoddology [J]. *Chem Sci*, 2003, 58(3/6): 521-535.

[18] 郭立玮, 李玲娟, 潘永兰, 等. 计算机软件著作权: 陶瓷膜精制中药的膜污染预报与防治系统 V1.0 [Z]. 中国: 软著登记第 0163739 号, 2009-09-04.

[19] 王艺峰, 程时远, 王世敏, 等. 高分子材料模拟中的分子力学法和力场 [J]. 高分子材料科学与工程, 2003, 19(1): 10-14.

[20] Sheetal J. Molecular modeling simulations to predict compatibility of poly(vinyl alcohol) and chitosan blends: A comparison with experiments [J]. *J Phys Chem B*, 2007, 111(10): 2431-2439.

[21] Jawalkar S S, Adoor S G, Sairam M, *et al*. Molecular modeling on the binary blend compatibility of poly(vinyl alcohol) and poly(methyl methacrylate): An atomistic simulation and thermodynamic approach [J]. *J Physical Chem B*, 2005, 109(32): 15611.

[22] 金万勤, 陆小华, 徐南平. 材料化学工程进展 [M]. 北京: 化学工业出版社, 2007.

[23] Wardeh S, Morvan H P. CFD simulations of flow and concentration polarization in spacer filled channels for application to water desalination [J]. *Chem Eng Res Des*, 2008, 86: 1107-1116.

[24] Rahimi M, Madaeni S S, Abolhasani M, *et al*. CFD and experimental studies of fouling of a microfiltration membrane [J]. *Chem Eng Proces*, 2009, 48(9): 1405-1413.

[25] Bacehin P, Espinasse B, Bessiere Y, *et al*. Numerical simulation of colloidal dispersion filtration: Description of critical flux and comparison with experimental results [J]. *Desalination*, 2006, 192(1/3): 74-81.

[26] 侯立安, 尹洪波. 计算流体力学在纳滤膜分离技术研究中应用 [J]. 膜科学与技术, 2011, 31(3): 5-10.

[27] 王福军. 计算流体动力学分析-CFD 软件原理与应用 [M]. 北京: 清华大学出版社, 2004.

[28] Vapnik V N. *Statistical Learning Theory* [M]. New York: Wiley-Interscience Publication, John Wiley & Sons. Inc., 1998.

[29] Liu X, Lu W C, Jin S L, *et al*. Support vector regression applied to materials optimization of sialon ceramics [J]. *Chemom Intell Lab Syst*, 2006, 82(1/2): 8-14.

[30] 郭立玮, 陆文聪, 董 洁, 等. 数据挖掘用于中药水提液膜过程优化的研究 [J]. 世界科学与技术—中医药现代化, 2005, 7(3): 42-47.

[31] 李玲娟, 郭立玮. 基于特征提取的中药水提液膜分离预测系统 [J]. 计算机工程与设计, 2010, 31(9): 2023-2026.

[32] 李玲娟, 洪 弘, 徐雪松, 等. 计算机化学及其在中药分离技术研究领域的应用进展 [J]. 中国中药杂志, 2011, 36(24): 3389-3396.

[33] 董 洁, 郭立玮. 截留相对分子质量 1 000 的超滤膜对生物碱和环烯醚萜类物质的透过率及其定量构效关系研究 [J]. 中国中药杂志, 2011, 36(2): 127-131.