基于机器学习算法的杏贝止咳颗粒脆碎度预测建模研究

赵媛媛[1, 2]，刘乐乐[1, 2]，张永超[1, 3]，张　欣[1, 3*]，王振中[1, 2, 3*]

1. 中药制药过程控制与智能制造技术全国重点实验室（江苏康缘药业股份有限公司/南京中医药大学），江苏 南京 211112
2. 南京中医药大学 康缘中药学院，江苏 南京 210023
3. 江苏康缘药业股份有限公司，江苏 连云港 222001

摘　要：目的　通过预测杏贝止咳颗粒（Xingbei Zhike Granules，XZG）颗粒脆碎度，结合最优算法，识别中间体的关键物料属性（critical material attributes，CMAs）。**方法**　以 6 个中间体物料的共 60 个物性参数作为输入，颗粒脆碎度作为输出，使用偏最小二乘回归算法（partial least squares，PLS）、决策树算法（classification and regression tree，CART）、广义路径追踪算法（generalized path seeker，GPS）、多元自适应回归样条算法（multivariate adaptive regression splines，MARS）、随机森林算法（random forest，RF）和树网随机梯度提升算法（TreeNet）共 6 种机器学习算法建立预测模型，根据模型拟合效果与预测误差确定最佳算法，筛选中间体的 CMAs。**结果**　基于 GPS 算法建立的预测模型表现最佳，可准确预测出 XZG 的颗粒脆碎度。其训练集决定系数（R^2_c）为 0.981，测试集决定系数（R^2_p）为 0.966，训练集均方根误差（root mean square error of calibration，RMSEC）为 0.976，测试集均方根误差（root mean square error of prediction，RMSEP）为 1.304，平均相对预测误差（average relative prediction error，ARPE）为 4.72%，低于 5%。共筛选出 6 个中间体的 15 个 CMAs。**结论**　基于中间体物性参数构建的预测模型，为干法制粒的颗粒脆碎度预测提供了一个新的思路；可以筛选出影响干法制粒颗粒脆碎度的关键物料属性，有助于提升产品质量。

关键词：杏贝止咳颗粒；颗粒脆碎度；干法制粒；物性参数；预测模型；广义路径追踪算法；关键物料属性

中图分类号：R283.6　　**文献标志码：**A　　**文章编号：**0253 - 2670(2025)23 - 8524 - 10

DOI: 10.7501/j.issn.0253-2670.2025.23.006

Modeling and prediction of friability for Xingbei Zhike Granules based on machine learning algorithms

ZHAO Yuanyuan[1, 2], LIU Lele[1, 2], ZHANG Yongchao[1, 3], ZHANG Xin[1, 3], WANG Zhenzhong[1, 2, 3]

1. State Key Laboratory of Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture (Jiangsu Kanion Pharmaceutical Co., Ltd., & Nanjing University of Chinese Medicine), Nanjing 211112, China
2. Kanion School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing 210023, China
3. Jiangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

Abstract: Objective　To identify critical material attributes (CMAs) of intermediates by predicting the granule friability of Xingbei Zhike Granules (XZG, 杏贝止咳颗粒) combined with the optimal algorithm. **Methods**　Sixty physical parameters of six intermediate materials were selected as input variables, and granule friability was chosen as the output variable. Six machine learning algorithms, including partial least squares (PLS), classification and regression tree (CART), generalized path seeker (GPS), multivariate adaptive regression splines (MARS), random forest (RF), and TreeNet, were employed to establish predictive models. The optimal algorithm was selected based on model fitting performance and prediction error to screen CMAs from the intermediates. **Results**　The prediction model constructed with the GPS algorithm exhibited the best performance for accurately predicting the granule friability of XZG. The model yielded a coefficient of determination for calibration (R^2_c) of 0.981, a coefficient of determination for prediction (R^2_p) of 0.966,

收稿日期：2025-07-22

基金项目：国家工信部产业基础再造和制造业高质量发展专项（TC2308068）

作者简介：赵媛媛，女，硕士研究生，研究方向为药物制剂与产品研发。E-mail: 3186466296@qq.com

***通信作者：**王振中，研究员，硕士生导师，研究方向为中药新药创制与过程控制研究。E-mail: kyyywzz@163.com

张　欣，博士，研究方向为中药制药过程新技术。E-mail: zxtcm@126.com

a root mean square error of calibration (RMSEC) of 0.976, a root mean square error of prediction (RMSEP) of 1.304, and an average relative prediction error of 4.72%, which is lower than 5%. A total of 15 CMAs from six intermediate materials were identified. **Conclusion** The predictive model based on intermediate material attributes offers a novel approach for predicting granule friability in dry granulation processes. It effectively identifies critical material attributes affecting granule friability, thereby contributing to improved product quality.

Key words: Xingbei Zhike Granules; granule friability; dry granulation; physical properties; predictive model; generalized path seeker; critical material attributes

杏贝止咳颗粒（Xingbei Zhike Granules，XZG）处方由麻黄、苦杏仁、甘草、浙贝母、桔梗等 9 味药材组成，其制备工艺包括提取浓缩、喷雾干燥及干法制粒等环节。中药颗粒剂作为中医药现代化的重要成果，具有便于服用、吸收快、生物利用度高、质量可控等优势，在临床治疗中得到广泛应用。

干法制粒工艺广泛应用于中药颗粒剂的制备过程，其制备的颗粒质量直接影响后续制剂的成型性、稳定性以及临床疗效。在干法制粒过程中，脆碎度是影响颗粒物理稳定性的重要质量属性[1]，是衡量颗粒抗破碎能力的重要指标[2]。当颗粒内部结合力低于临界值时，脆碎度显著升高，颗粒易在储存或运输过程中因机械应力发生破碎，导致溶散时限延长、装量差异增大等问题，直接影响产品的物理性能、生产效率和合规性。因此，准确预测颗粒脆碎度，对于保障中药颗粒剂的质量稳定性具有重要意义。

影响中药制剂脆碎度的因素有很多，如原辅料性质、制备工艺、环境温湿度等[3]。控制中药制剂脆碎度可以从处方因素和工艺因素 2 个角度入手。现有改良集中于制剂成型环节工艺改善，例如通过干法制粒替代湿法制粒以减少水分对结构破坏的风险[4]，或引入塑性辅料（如羟丙甲纤维素）增强颗粒抗破碎能力，未从生产中间体物料角度分析影响颗粒脆碎度的关键物料属性及相关控制限。已有研究证实，通过对不同批次中药中间体物料属性的量化分析，可有效揭示物料特性与颗粒质量间的关联特征。王晴等[5]以桂枝茯苓胶囊（Guizhi Fuling Capsules，GFC）制剂成型过程原料、中间体粉末和胶囊成品为研究对象，建立了 GFC 内容物的吸湿性预测模型；陈琪等[6]同样以 GFC 生产过程中的 5 种中间体为研究对象，基于不同算法建立 GFC 内容物吸湿性预测模型；陶振等[7]以不同辅料配比制成的混合粉为研究对象，比较 5 种算法对 XZG 溶解性的模型预测效果，确定最优算法和影响颗粒溶解性的中间体关键物料属性（critical material attributes，CMAs）。

本研究基于 XZG 多批次生产数据，运用偏最小二乘回归算法（partial least squares，PLS）、决策树算法（classification and regression tree，CART）、广义路径追踪算法（generalized path seeker，GPS）、多元自适应回归样条算法（multivariate adaptive regression splines，MARS）、随机森林算法（random forest，RF）、树网随机梯度提升算法（TreeNet）多种机器学习算法构建干法制粒后颗粒脆碎度预测模型，通过对不同算法模型的性能进行比较和分析，筛选出最优的预测模型。同时，利用变量重要性分析方法，识别出对颗粒脆碎度影响显著的 CMAs，为中药颗粒剂智能化质量控制提供方法学支持，推动中药生产过程质量控制从经验驱动向数据驱动转变，提高中药颗粒剂的质量稳定性和生产效率。

1 仪器与材料

1.1 仪器

BT-1001 型粉体特性测试仪、Bettersize2600 型激光粒度分布仪，丹东百特仪器有限公司；DHG-9145A 型电热鼓风干燥箱、LHS-250HC-II 型恒温恒湿箱，上海一恒科学仪器有限公司；SD20 型 pH 计，梅特勒-托利多仪器（上海）有限公司；DDS-307A 型电导率仪，上海雷磁创益仪器仪表有限公司；DC0506N 型恒温浴槽、NDJ-8S 型数字旋转粘度计、QBZY-1 型全自动表面张力仪，上海方瑞仪器有限公司；WYA-2W 型阿贝折射仪，上海仪电物理光学仪器有限公司；CJY-300E 型脆碎度仪，上海黄海药检仪器有限公司。

1.2 材料

XZG 生产过程中的 5 种中间体，包括水提浸膏（ST）、醇提浸膏（CT）、混合浸膏（HHJ）、喷干粉（PG）、混合粉（HHF）及 XZG 共 69 批，采集样本为 2024 年 4 月至 11 月期间生产，涵盖编号 240516 至 241214 的颗粒批次，所有物料均由江苏康缘药业股份有限公司供应。

2 方法与结果

2.1 中间体物料属性测定方法

参考文献方法[7-11]对物料的休止角（α）、崩溃角（β）、差角（CJ）、平板角（γ）、松装密度（D_a）、振实密度（D_c）、分散度（Dis）、吸湿性（H）、水分（HR）、卡尔指数（CI）、豪斯纳比（IH）、孔隙率（I_e）、相对均齐度（Iθ）、粒径（D_{10}、D_{50}、D_{90}）、分布范围（span）、宽度（width）、粒径＜50 μm 百分比（Pf）等粉体物理属性进行检测。参考文献方法[12-13]对物料的固含量（SC）、密度（ρ）、酸碱值（pH）、折光率（n）、黏度（μ）、表面张力（γ）、电导率（σ）等浸膏物理属性进行检测。

2.1.1 细颗粒占比（FPP） 采用手动筛分法，称取 20 g 样品颗粒，过 80 目标准筛，过筛时保持水平状态，左右往返，边筛动边拍打 3 min。取过筛的细颗粒，精密称定质量，计算其所占比例 FPP。

2.1.2 脆碎度 精密称取样品颗粒 10 g，置于脆碎度测定仪转鼓内，另加入 200 颗直径为 4 mm 的玻璃珠，设置转速为 25 r/min，连续转动 4 min 后取出，过 24 目筛，筛分后称定，通过 24 目筛的颗粒质量[14-17]。

脆碎度＝(供试品初质量－过筛后质量)/供试品初质量

2.2 物理性质测定结果

2.2.1 中间体物理性质测定结果 按照“2.1”项下方法测定 XZG 及其中间体物理性质，结果见表 1。由表 1 可知，多数中间体物理性质指标最大值与最

表 1 中间体物理性质测定结果

Table 1 Determination results of intermediate physical properties

物理性质	单位	最大值	最小值	平均值	物理性质	单位	最大值	最小值	平均值
ST-sc	%	28.21	19.56	24.25	PG-H	%	36.95	27.01	31.07
ST-ρ	g·cm^{-3}	1.13	0.80	0.98	PG-D_{10}	μm	13.70	6.49	10.23
ST-pH		6.16	4.26	5.35	PG-D_{50}	μm	60.85	20.01	35.23
ST-n	N·cm^{-1}	1.38	1.37	1.37	PG-D_{90}	μm	82.53	56.36	70.13
ST-μ	cp	371.47	22.61	79.87	PG-Pf	%	92.48	67.52	80.23
ST-γ	mN·m^{-1}	47.47	33.77	45.76	PG-span		2.16	1.55	1.85
ST-σ	μS·cm^{-1}	14.29	7.62	11.22	PG-Iθ		1.51×10^{-3}	1.02×10^{-4}	5.17×10^{-4}
CT-sc	%	28.22	16.80	23.66	PG-width		70.79	37.09	55.23
CT-ρ	g·cm^{-3}	1.13	0.99	1.07	HHF-α		53.23	42.94	49.10
CT-pH		7.73	2.88	4.66	HHF-β		50.10	35.73	42.92
CT-n	N·cm^{-1}	1.38	1.36	1.37	HHF-CJ		15.53	1.00	6.18
CT-γ	mN·m^{-1}	45.57	25.90	32.93	HHF-γ		82.58	42.09	71.82
CT-σ	μS·cm^{-1}	14.29	4.55	8.13	HHF-D_a	g·cm^{-3}	0.41	0.23	0.34
HHJ-sc	%	25.45	16.26	21.69	HHF-D_c	g·cm^{-3}	0.71	0.44	0.62
HHJ-ρ	g·cm^{-3}	1.29	0.86	1.05	HHF-IH		2.05	1.45	1.82
HHJ-pH		5.18	2.82	4.54	HHF-CI	%	0.51	0.31	0.45
HHJ-n	N·cm^{-1}	1.38	1.36	1.36	HHF-I_e		2.07	0.77	1.33
HHJ-μ	cp	117.41	11.38	34.98	HHF-HR	%	7.26	4.04	5.53
HHJ-γ	mN·m^{-1}	45.57	25.90	39.83	HHF-H	%	35.73	25.19	28.53
HHJ-σ	μS·cm^{-1}	14.29	4.55	9.42	HHF-D_{10}	μm	8.63	5.08	7.45
PG-α		55.57	42.16	50.23	HHF-D_{50}	μm	26.98	16.59	22.44
PG-β		53.94	40.98	47.67	HHF-D_{90}	μm	69.77	48.36	59.06
PG-CJ		10.51	0.57	4.84	HHF-Pf	%	90.89	77.34	84.21
PG-γ		85.61	45.74	70.92	HHF-span		2.61	1.97	2.30
PG-D_a	g·cm^{-3}	0.31	0.19	0.26	HHF-Iθ		1.53×10^{-3}	1.04×10^{-4}	5.85×10^{-4}
PG-D_c	g·cm^{-3}	0.64	0.39	0.51	HHF-width		61.42	41.24	51.61
PG-IH		2.39	1.59	2.03	HHF-Dis		33.13	8.43	19.95
PG-CI	%	0.58	0.41	0.50	XZG-FPP	%	9.90	0.51	3.09
PG-I_e		2.77	1.57	2.15	XZG-H	%	26.76	18.99	22.44
PG-HR	%	7.56	4.26	5.84	XZG-HR	%	6.30	3.40	5.30

小值有一定差距，说明多数物理性质指标存在一定的批间波动。

2.2.2 XZG 脆碎度测定结果 根据“2.1”项下方法测定颗粒脆碎度分布情况见图 1。可知收集样本的脆碎度范围为 10.17%～41.58%，数据波动较大。

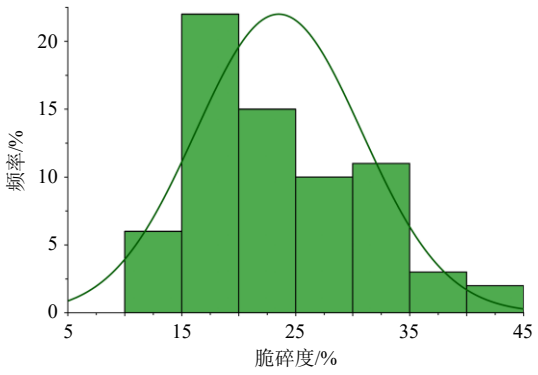

图 1 XZG 脆碎度分布图

Fig. 1 Distribution diagram of friability of XZG

2.3 样本集的划分

将 69 批样本按 4∶1 比例随机划分为训练集与测试集，其中，训练集 55 批，测试集 14 批。

2.4 建模原理与方法

2.4.1 PLS PLS 通过多轮循环计算，实现变量投影与参数优化，最终建立起 Y 与 X 的线性关系模型（$Y=bX+f$）。其核心机制在于通过线性组合各阶段的投影参数，逐步优化得到回归系数 b 和误差项 f。该方法基于成分提取原理实现变量降维，区别于传统降维方法之处在于其优先保留对响应变量解释力强的协变量。PLS 采用分阶段逼近策略，首轮计算获取主趋势框架后，通过逐次迭代修正局部参数，运用动态优化策略提升模型拟合精度[6]。

2.4.2 CART CART 算法基于属性特征空间训练数据，通过二分递归分割生成与目标变量高度适配的判别规则。该算法结合高效运算与智能化数据解析能力，具有以下技术优势：（1）数据驱动特性使其无需预设变量转换条件，通过排序信息直接构建模型，有效规避异常值干扰；（2）采用替代变量分裂机制实现缺失值自动填补，提升模型鲁棒性；（3）决策路径可视化输出支持多维数据关联性挖掘，为复杂数据集提供可解释性建模方案[6,18-19]。

2.4.3 MARS MARS 算法是一种非参数回归分析方法，该算法的核心优势在于能够量化评估各基函数对模型的贡献度，从而精确解析预测变量的独立作用及交互影响。在模型选择方面，MARS 采用广义交叉验证（generalized cross-validation，GCV）准则，以 GCV 值最小化为标准确定最优模型结构。该方法具有较高的计算效率，通过自适应分段回归构建显式的多维非线性模型，基于数据驱动的方式建立自变量与因变量之间的映射关系[20-22]。

2.4.4 GPS GPS 算法采用多元正态回归框架构建高精度模型，通过引入广谱弹性系数生成多组候选模型。该算法采用逐步优化策略，初始模型不含预测变量，在迭代过程中依次添加新变量或调整现有变量系数，形成不同步数的路径模型集合，并基于智能筛选机制确定最优解。该方法在计算效率和变量覆盖度上显著优化了正则化回归过程，其核心优势体现在：（1）能高效处理高维小样本数据；（2）对强相关性预测变量具有良好适应性。与传统回归方法相比，GPS 模型在预测性能和稳定性方面均展现出明显优势[23-24]。

2.4.5 RF RF 算法以集成学习为核心框架，它通过构建多棵决策树来进行决策，最终将多个模型的输出结果进行投票或平均，从而得到一个更加稳定和准确的预测结果。RF 通过以下步骤进行构建。

（1）随机采样：在训练阶段，将不同钻井参数和井眼轨迹参数作为输入变量，将钻速作为输出变量。RF 采用自助采样法，从原始训练数据集中有放回地随机抽取多个子集。每个子集包含与原始数据集相同数量的样本，且允许单个样本在子集中重复出现，用于独立训练一棵决策树。

（2）特征随机选择：每棵决策树的构建过程中，RF 不仅在样本上进行随机抽样，还在每个节点分裂时随机选择一部分特征进行分裂，而不是使用所有的特征，这降低了树与树之间的相关性，提高了模型的泛化能力。

（3）构建多棵决策树：对于每个从数据中抽取的子集，RF 都会训练一棵决策树。每棵树的训练过程中，树的深度可以控制，通过递归分裂节点，直到节点纯度达到某个阈值或树的最大深度。

（4）集成预测：在预测阶段，RF 对每棵树的预测结果进行集成，将所有决策树的输出结果求得平均值，作为最终预测结果[25]。

2.4.6 TreeNet TreeNet 是一种高效、容错且预测精度卓越的新型建模工具。它仅需简单的数据预处理，能智能处理异常数据，自动适应缺失值，并通过全面的自检机制确保模型在新数据上的稳定性。与传统提升方法不同，TreeNet 采用分阶段函数逼近技术，逐步优化前一阶段的残差，而非直接建模

目标变量。此外，它重点优化决策边界附近的样本分类，大幅提升预测准确性。基于 TreeNet 构建的模型本质上是一种评分卡模型，可为每条合规样本输出 0～1 的概率评分，评分越高，事件发生可能性越大[26]。

2.5　数据分析方法

采用 SPM 8.3 软件（美国 Salford Systems 公司）中不同算法构建颗粒脆碎度预测模型。

2.6　模型性能评价

2.6.1　PLS 模型评价　采用模型解释率（R_X^2）、预测决定系数（R_Y^2）和交叉验证决定系数（Q^2）来评估模型拟合效果，这些指标越接近 1，表示 PLS 模型的拟合和预测能力越优。当 R_Y^2 和 Q^2 均大于 0.5 时，模型性能良好。

2.6.2　机器学习算法模型评价　模型性能通过训练集决定系数（R^2_c）、测试集决定系数（R^2_p）、训练集均方根误差（root mean square error of calibration，RMSEC）、预测集均方根误差（root mean square error of prediction，RMSEP）、训练集平均绝对百分比误差（mean absolute percent arror of calibration，MAPEC）、预测集平均绝对百分比误差（mean absolute percent arror of prediction，MAPEP）、平均绝对偏差（mean absolute deviation，MAD）等指标进行评估。R^2 反映模型的拟合优度，MAPE 衡量预测误差的相对大小，RMSE、MAD 是衡量模型预测值与实际值之间差异的度量。其中，R^2 值趋近于 1，MAPE 值趋近于 0，说明模型拟合效果越好；而 RMSE、MAD 值越小，则表明模型的预测精度越高[27-28]。

$$R^2=1-\sum_{i=1}^{n}(y-y_i)^2/\sum_{i=1}^{n}(y-\bar{y})^2 \quad (1)$$

$$\mathrm{RMSE}=[1/n\sum_{i=1}^{n}(y-y_i)^2]^{1/2} \quad (2)$$

$$\mathrm{MAPE}=\sum_{i=1}^{n}|(y-y_i)/y|/n \quad (3)$$

n 为训练集或测试集的样本数，$i\in[1, n]$，y 为参考值，为预测值，$\bar{y}$ 为所有样品参考值的平均值

2.6.3　模型预测精度评价　根据测试集中平均相对预测误差（average relative prediction error，ARPE）来评价不同算法对模型预测精度的影响，ARPE 越小，模型预测精度越高，反之越低。

相对预测误差＝|预测值－参考值|/参考值[6]

2.7　CMAs 筛选方法

建模过程中，根据变量重要性投影（variable importance projection，VIP），将各参数按照 VIP 值从小到大依次剔除并逐一建模，直至筛选出最佳模型。此时，筛选出最佳模型对应的自变量参数即 CMAs[7]。

2.8　多算法模型的建立

2.8.1　PLS 建模结果　以颗粒脆碎度（Y）为因变量，60 个物性参数为自变量（X），建立的 PLS 预测模型的 Q^2 为 0.343，R^2_X 和 R^2_Y 分别为 0.302、0.621，$Q^2<0.5$ 且 $R^2_X<Q^2$，说明模型中存在较多的冗余信息，导致该模型预测性能较差。根据变量投影重要性（variable importance for the projection，VIP），将各参数按照 VIP 值从小到大依次剔除并逐一建模，直至筛选出最佳模型。建模结果见表 2、3。当变量数为 19 时，Q^2 达到最大值 0.531，R^2_X 和 R^2_Y 为 0.601、0.616，VIP 如图 2 所示。所建模型 R^2_X、R^2_Y、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.616、0.587、4.497、4.708、0.160、0.197、3.933，模型的拟合效果与预测精度达到最高，ARPE 为 19.74%。

2.8.2　CART 算法建模结果　树的节点对 CART 算法影响较大，开始建模前，可以通过 SPM 软件中

表 2　PLS 建模结果

Table 2　PLS modeling results

模型	自变量个数	潜变量数	R^2_X	R^2_Y	Q^2
模型 1	60	2	0.302	0.621	0.343
模型 2	19	2	0.601	0.616	0.531

表 3　不同变量数下 PLS 模型性能与预测能力比较

Table 3　Comparison of PLS model performance and predictive ability with different numbers of variables

自变量个数	训练集			测试集			MAD	ARPE/%
	R^2	RMSE	MAPE	R^2	RMSE	MAPE		
60	0.644	4.373	0.159	0.611	4.720	0.211	4.029	21.08
19	0.616	4.497	0.160	0.587	4.708	0.197	3.933	19.74

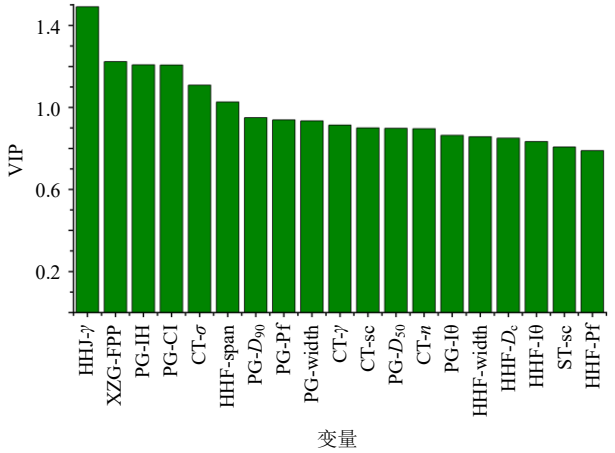

图 2　PLS 模型自变量 VIP 分布

Fig. 2　Distribution of independent variable VIP in PLS model

AUTOMATE 自动建模选项进行筛选，表 4 为不同节点数对应相对误差值，结果为当节点数为 8 时模型性能最好。60 个参数得到的最初模型性能指标：R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.696、0.567、3.507、4.689、0.127、0.127、2.731。依据变量重要性从小到大剔除变量并逐一建模结果见表 5，由表 5 可知当变量数为 4 时，模型性能最佳。所建模型 R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.890、0.786、2.113、3.298、0.076、0.125、2.604，模型的拟合效果与预测精度达到最高，ARPE 为 12.48%。

表 4　不同节点数对应相对误差值

Table 4　Relative error values corresponding to different numbers of nodes

结点数	相对误差	结点数	相对误差	结点数	相对误差
2	0.600	5	0.644	8	0.433
3	0.553	6	0.729	10	0.511
4	0.806	7	0.584		

表 5　不同变量个数建立的 CART 模型的性能和预测性能比较

Table 5　Comparison of CART model performance and prediction performance established with different numbers of variables

自变量	训练集			测试集			MAD	ARPE/
个数	R^2	RMSE	MAPE	R^2	RMSE	MAPE		%
60	0.696	3.507	0.127	0.567	4.689	0.127	2.731	12.68
44	0.661	3.702	0.137	0.697	3.923	0.134	2.731	13.40
28	0.434	4.785	0.171	0.621	4.389	0.171	2.974	17.06
15	0.536	4.335	0.157	0.649	4.219	0.161	2.974	16.08
4	0.890	2.113	0.076	0.786	3.298	0.125	2.604	12.48

2.8.3　GPS 算法建模结果　均方误差（mean squared error，MSE）是评估测试集模型性能的关键指标，其数值由算法内部优化过程产生的弹性系数确定。GPS 算法最优模型的选择标准是使测试样本的 MSE 达到最小值。如图 3 所示，MSE 随预测变量数量的变化呈现明显规律性变化，当变量数为 28 时，模型取得最小的 MSE 值，此时达到最佳预测效果。选择此模型，在其基础上依据变量重要性从小到大剔除变量并逐一建模，建模结果见表 6。由表可知，当变量数为 15 时模型效果最优，此时模型 R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.981、0.966、0.976、1.304、0.033、0.0475、0.987，ARPE 为 4.72%。

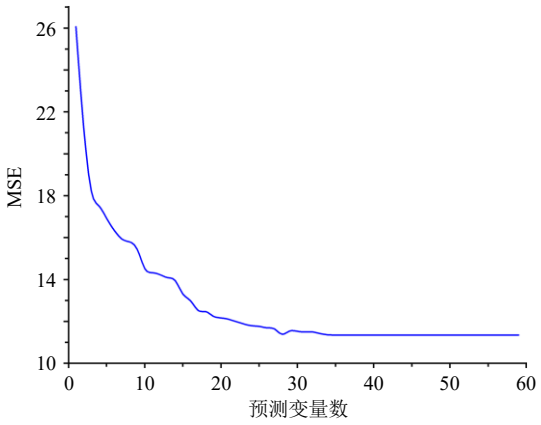

图 3　MSE 与预测变量数相关图

Fig. 3　Correlation diagram between MSE and the number of predictor variables

表 6　不同变量数下 GPS 模型性能与预测能力比较

Table 6　Comparison of GPS model performance and predictive ability with different numbers of variables

自变量	训练集			测试集			MAD	ARPE/
个数	R^2	RMSE	MAPE	R^2	RMSE	MAPE		%
60	0.967	1.297	0.046	0.857	2.687	0.091	1.874	9.11
28	0.977	1.084	0.037	0.926	1.918	0.065	1.337	6.49
15	0.981	0.976	0.033	0.966	1.304	0.047	0.987	4.72
11	0.970	1.229	0.042	0.952	1.560	0.054	1.165	5.41
8	0.929	1.879	0.075	0.906	2.174	0.086	1.709	8.58

2.8.4　MARS 算法建模结果　基函数的数量影响最优模型的选择，可以通过 GCV 性能图确定最优基函数，从而获得最佳模型尺寸[6]。由表 7 可知，GCV 值随基函数的变化规律，在实际建模中，应不断调试预设的基函数数量，以选择最优基函数并建立最佳模型。由表 7 可知，GCV 值最小时对应的基函数的个数为 5，以 5 个基函数建立模型，得到最初模型性能指标：R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.596、0.675、4.046、4.062、0.151、0.164、3.441。依据 VIP 值从小到大剔除变量并逐一建模，建模结果见表 8，随着剔除变量数的增加，所建模型稳定性与预测精度并未得到较大改善，表明 MARS 模型所含冗余信息较多。

表 7　不同基函数个数对应 GCV 值

Table 7　GCV values corresponding to different numbers of basis functions

基函数	GCV 值	基函数	GCV 值	基函数	GCV 值
1	34.919	5	32.563	9	54.114
2	33.911	6	32.973	10	68.339
3	33.128	7	36.873		
4	32.658	8	44.073		

表 8　不同变量个数下 MARS 模型性能与预测能力比较

Table 8　Comparison of MARS model performance and predictive ability with different numbers of variables

自变量个数	训练集			测试集			MAD	ARPE/%
	R^2	RMSE	MAPE	R^2	RMSE	MAPE		
60	0.596	4.046	0.151	0.675	4.062	0.164	3.441	16.36
41	0.552	4.258	0.164	0.600	4.511	0.173	3.570	17.34
22	0.584	4.106	0.158	0.583	4.606	0.163	3.566	16.32
8	0.600	4.023	0.150	0.678	4.046	0.162	3.421	16.16
4	0.599	4.031	0.150	0.679	4.037	0.161	3.394	16.08

2.8.5　RF 算法建模结果　树的数量影响 RF 模型性能高低，MSE 随树的数目变化趋势如图 4 所示，当树的数目为 200 时，模型的 MSE 最低，所建模型性能最好。初始模型的 R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.260、0.493、5.472、5.076、0.210、0.186、3.738。依据变量重要性从小到大剔除变量并重新建立模型结果见表 9，由表可知当变量数为 8 时，模型性能最佳。所建模型 R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.545、0.563、4.291、4.712、0.154、0.143、3.211，模型的拟合效果与预测精度均较高，ARPE 为 14.33%。

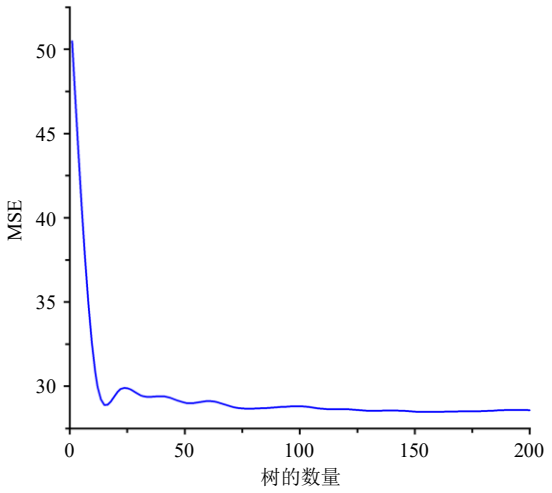

图 4　MSE 随树的数目变化趋势

Fig. 4　Trend of MSE with the number of trees

表 9　不同变量个数下的 RF 模型性能与预测能力比较

Table 9　Comparison of RF model performance and predictive ability with different numbers of variables

自变量个数	训练集			测试集			MAD	ARPE/%
	R^2	RMSE	MAPE	R^2	RMSE	MAPE		
60	0.260	5.472	0.210	0.493	5.076	0.186	3.738	18.64
40	0.318	5.253	0.201	0.538	4.846	0.174	3.688	17.45
20	0.413	4.875	0.184	0.594	4.544	0.160	3.438	16.00
12	0.475	4.609	0.169	0.609	4.458	0.144	3.098	14.40
8	0.545	4.291	0.154	0.563	4.712	0.143	3.211	14.33

2.8.6　TreeNet 算法建模结果　树的数量影响 TreeNet 模型性能高低，图 5 为测试集 R^2 随树数目变化趋势，当树数目为 200 时，模型 R^2 最大，所建模型性能最好。初始模型的 R^2_c、R^2_p、RMSEC、RMSEP、MAPEC、MAPEP、MAD 分别为 0.837、0.569、2.570、4.682、0.094、0.169、2.530。依据变量重要性从小到大剔除变量并重新建立模型结果见表 10，随着剔除变量数的增加，所建模型稳定性与预测精度并未得到较大改善，训练集与测试集 R^2 差值过大，模型过拟合[7]。

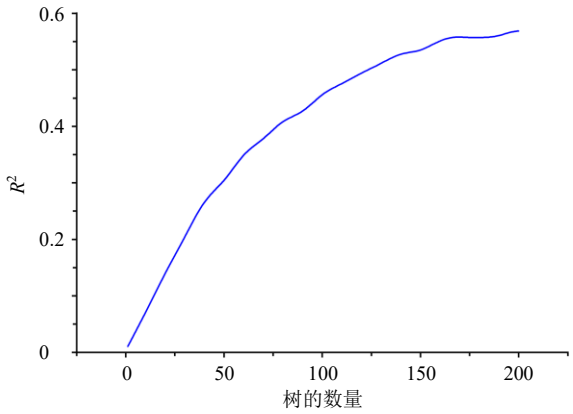

图 5　R^2 随树的数目变化趋势

Fig. 5　Trend of R^2 with varying number of trees

表 10　不同变量个数下 TreeNet 模型性能与预测能力比较

Table 10　Comparison of TreeNet model performance and predictive ability with different numbers of variables

自变量个数	训练集			测试集			MAD	ARPE/%
	R^2	RMSE	MAPE	R^2	RMSE	MAPE		
60	0.837	2.570	0.094	0.569	4.682	0.169	2.530	16.89
42	0.834	2.590	0.094	0.555	4.753	0.172	3.613	17.24
23	0.828	2.638	0.095	0.527	4.901	0.179	3.758	17.88
11	0.817	2.720	0.100	0.534	4.865	0.173	3.682	17.32
6	0.784	2.956	0.108	0.593	4.545	0.160	3.315	16.00

2.9　不同算法模型性能和预测精度比较结果

不同算法建立的最优模型性能和预测精度比较结果见表 11，可以看出 GPS 算法的最优预测模型在训练集和测试集上均获得了更高的决定系数，且各项误差指标均较低，表明 GPS 算法更适用于本数据建模。模型训练集和测试集的决定系数分别为 0.981、0.966，MAPE 分别为 0.033、0.047，RMSE 分别为 0.976、1.304，MAD 为 0.987，ARPE 为 4.72%＜5%，说明模型预测准确度较高。对该模型预测值与实际

表 11 不同算法下的模型性能与预测精度比较

Table 11 Comparison of model performance and prediction accuracy under different algorithms

算法	自变量个数	训练集			测试集			MAD	ARPE/%
		R^2	RMSE	MAPE	R^2	RMSE	MAPE		
PLS	19	0.616	4.497	0.160	0.587	4.708	0.197	3.933	19.74
CART	4	0.890	2.113	0.076	0.786	3.298	0.125	2.604	12.48
GPS	15	0.981	0.976	0.033	0.966	1.304	0.047	0.987	4.72
MARS	4	0.599	4.031	0.150	0.679	4.037	0.161	3.394	16.08
RF	8	0.545	4.291	0.154	0.563	4.712	0.143	3.211	14.33
TreeNet	6	0.784	2.956	0.108	0.593	4.545	0.160	3.315	16.00

值进行配对 t 检验，结果显示，$P=0.257>0.05$，认为该模型预测值与实际值间无显著性差异，说明模型具有较好的预测能力。

2.10 XZG 干法制粒后颗粒脆碎度 CMAs 的筛选

采用 GPS 算法建立的模型最佳，其包含 15 个变量，使用方差膨胀因子（variance inflation factor，VIF）评价 15 个参数的共线性程度，一般 VIF 值＜10 表明共线性较弱，VIF＞10 表明强共线性，评价结果见表 12，由表可知模型自变量之间独立性良好，因此可将这 15 个变量定义为影响颗粒脆碎度的关键物料属性[7]。

利用 SPM8.3 软件依据变量重要性对所建模型中 15 个变量进行打分，打分结果见表 13。得分反映了变量与目标变量的相关性或重要性，得分越高，表示该变量对目标变量的解释能力越强。由表可知，XZG-FPP 与 CT-sc 的得分均大于 90，表明它们对颗粒脆碎度具有最强的解释能力，为决定物料脆碎特性的关键因素；此外，PG-CI、HHJ-μ、HHJ-γ 等 5 个变量的得分介于 50～90，属于次关键影响因素，在颗粒脆碎度的形成机制中起到重要的协同作用；其余 8 个变量的得分均低于 50，对脆碎度的影响相对较弱，可视为辅助性因素。

表 12 模型自变量共线性诊断结果

Table 12 Results of collinearity diagnosis of independent variables in the model

变量	VIF 值	变量	VIF 值	变量	VIF 值
XZG-FPP	1.597	ST-pH	1.329	HHF-γ	1.360
CT-sc	2.414	HHF-I_e	1.723	PG-α	2.307
PG-CI	1.927	PG-β	2.111	HHF-Dis	2.427
HHJ-μ	1.896	ST-ρ	1.460	PG-HR	1.594
HHJ-γ	2.493	ST-sc	2.586	ST-γ	1.404

表 13 GPS 模型变量重要性评分结果

Table 13 Variable importance scores of GPS model

变量	得分	变量	得分	变量	得分
XZG-FPP	100.00	ST-pH	53.70	HHF-γ	29.82
CT-sc	96.88	HHF-I_e	52.49	PG-α	24.64
PG-CI	68.45	PG-β	39.60	HHF-Dis	23.43
HHJ-μ	68.37	ST-ρ	34.72	PG-HR	21.87
HHJ-γ	65.51	ST-sc	32.58	ST-γ	16.10

3 讨论

本研究通过对比 6 种机器学习算法对 XZG 干法制粒后颗粒脆碎度进行建模，结果表明 GPS 算法预测效果最优。推测原因如下：（1）数据层面：颗粒脆碎度通常由物料属性的复杂非线性交互共同决定，变量间存在高阶交互及多重共线性，例如差角（CJ）由休止角（α）与崩溃角（β）计算得到。传统线性模型（PLS）难以捕捉非线性和交互关系的复杂性，树模型（CART/RF）虽然具有较强的非线性表达能力，但在应对极端或高阶非线性关系以及小样本数据集时，会出现欠拟合或过拟合问题。GPS 通过符号回归直接挖掘关键变量组合，能够有效识别关键的非线性关系和变量交互，对小规模高维数据具有强噪声鲁棒性。（2）模型层面：GPS 通过进化机制自动搜索变量之间的数学表达式，能够有效捕捉高阶多项式、指数和对数等非线性关系。与其他模型相比，GPS 在非线性拟合和可解释性之间取得了良好的平衡。PLS 受到线性假设的限制，树模型的规则往往冗长且难以解析连续交互，MARS 则依赖于人工定义的基函数，而神经网络需要大量数据且缺乏透明性。GPS 兼顾模型精度与解释性的特点，适用于制药行业对工艺机理透明化和工程实用性的要求，成为处理复杂物料属性建模问题的理想选择。

本研究筛选出 XZG-FPP、CT-sc、PG-CI、HHJ-μ、HHJ-γ 等 15 个变量为影响 XZG 颗粒脆碎度的 CMAs，表明 XZG 颗粒脆碎度受多重因素影响，涉及多个中间体，需综合调控。根据变量重要性评分结果可知，细颗粒占比（XZG-FPP）和浸膏固含量（CT-sc、ST-sc）是调控颗粒脆碎度的核心因素。过高的细颗粒占比会导致颗粒大小分布不均，形成局部应力集中点，从而降低颗粒的整体强度，增加脆碎风险。可通过建立多级筛分系统、动态工艺调控和粘合剂协同作用实现其范围精准调控。浸膏固含量直接影响颗粒骨架致密性，低固含量易形成空心结构，导致颗粒易碎，需通过浓缩工艺控制固含量至合理阈值。

此外，卡尔指数（PG-CI）、黏度（HHJ-μ）与表面张力（HHJ-γ、ST-γ）属于次关键影响因素，卡尔指数反映粉体的压缩性，控制在合理范围内有助于增强颗粒的整体结构稳定性，减少因过大弹性变形导致的脆碎倾向。浸膏黏度影响颗粒结合力，高黏度易形成致密结构，增强颗粒抗破碎能力。浸膏 pH 则通过改变浸膏黏度等理化性质间接影响颗粒强度。浸膏表面张力低促进润湿性，减少颗粒内部缺陷，反之易聚集不均。在辅助变量方面，浸膏密度（ST-ρ）作为影响颗粒内部孔隙率和骨架均匀性的参数，与颗粒机械强度具有一定关联性。同样，中间体粉末的含水量（PG-HR）也会间接影响颗粒的成型性和后续稳定性，水分的适度控制对于避免团聚或结构松散至关重要。

虽然粉体的流动性参数（PG-β、HHF-γ、PG-α、HHF-Dis）和堆积性参数（HHF-I_e）分值较低，但可通过协同调控，避免在干法制粒和压片过程中因流动性不足或堆积不良而引起的颗粒应力集中，从而有效减少脆碎现象的发生。实际生产中应优先建立细颗粒占比、浸膏固含量及粉体压缩性的动态监控体系，通过工艺参数联动调整实现强度与脆碎度的最优平衡，确保产品质量稳定性。

中药制剂质量控制的提升关键在于积累大量反映中药生产质量传递规律的基础数据。通过数据挖掘和建模，明确并量化制剂过程中的质量传递因果关系，从中获取信息并建立模型，最终利用这些模型解决或预防质量问题[5,29]。本研究基于中间体物性参数构建预测模型，不仅为预测和控制干法制粒颗粒脆碎度提供了崭新的视角和可行路径，也为进一步提升制粒工艺的稳定性和产品质量的一致性奠定了坚实基础，最终为 XZG 质量控制提供了初步的智能化解决方案。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 魏瑞霞, 王璟璐, 冯中, 等. 柴银颗粒干法制粒的工艺优化 [J]. 中国医药工业杂志, 2023, 54(7): 1088-1094.

[2] 贾栓柱, 杜仕国, 甄建伟, 等. 金属基高热剂湿法制粒技术 [J]. 兵器装备工程学报, 2019, 40(4): 152-154.

[3] 王芳, 金曼, 朱传祥. 影响片剂脆碎度的因素分析 [J]. 齐鲁药事, 2012, 31(10): 613-614.

[4] 贾晓伟, 吴丹丹, 单长智, 等. 干法制粒技术在中药制剂中的应用探究 [J]. 中国设备工程, 2021(10): 197-198.

[5] 王晴, 徐冰, 王芬, 等. 桂枝茯苓胶囊内容物吸湿性预测建模研究 [J]. 中国中药杂志, 2020, 45(2): 242-249.

[6] 陈琪, 徐芳芳, 张欣, 等. 基于不同算法对桂枝茯苓胶囊内容物吸湿性预测建模研究 [J]. 中草药, 2021, 52(11): 3216-3223.

[7] 陶振, 洪韵, 安双凤, 等. 基于不同算法对杏贝止咳颗粒中间体物料属性与颗粒溶解性的相关性研究 [J]. 中草药, 2024, 55(22): 7644-7652.

[8] 汪盛华, 闫明, 徐芳芳, 等. 基于物料粉体学性质对 5 种制剂品种颗粒溶化性的相关性研究 [J]. 中草药, 2022, 53(22): 7082-7090.

[9] 秦春娟, 闫明, 王振中, 等. 基于多品种水-醇双提物的中药粉体性质影响颗粒吸湿性的研究 [J]. 中草药, 2023, 54(4): 1120-1126.

[10] 丁涵, 徐忠坤, 王振中, 等. 基于 AHP-CRITIC 混合加权法和 Box-Behnken 设计-响应面法优化羌芩颗粒成型工艺及其物理指纹图谱研究 [J]. 中草药, 2024, 55(3): 787-797.

[11] 戴胜云. 中药直接压片处方智能设计方法研究 [D]. 北京: 北京中医药大学, 2019.

[12] 童枫, 徐芳芳, 闫逸伦, 等. 热毒宁注射液金银花和青蒿（金青）萃取过程中固形物含量近红外光谱在线监测模型的建立及萃取终点判断研究 [J]. 中草药, 2024, 55(19): 6555-6565.

[13] 杨婉, 邹海英, 邱智东, 等. 基于物理指纹图谱与多指标成分定量测定构建升陷汤标准煎液质量评价方法 [J]. 中草药, 2023, 54(6): 1804-1813.

[14] 王满, 汪露露, 陈雪晴, 等. 星点设计-效应面法优化复方肠泰颗粒成型工艺 [J]. 中成药, 2017, 39(2): 420-423.

[15] 胡文军, 吴艳萍, 潘卫三, 等. 栀灵复方颗粒的成型工艺优化及其对小鼠自主活动的影响 [J]. 中国新药杂志, 2020, 29(6): 710-714.

[16] Inghelbrecht S, Remon J P. Roller compaction and tableting of microcrystalline cellulose/drug mixtures [J]. International journal of pharmaceutics, 1998, 161(2): 215-224.

[17] Remon J P, Schwartz J B. Effect of raw materlals and processing on the quality of granules prepared from microcrystalline cellulose-lactose mixtures [J]. *Drug Dev Ind Pharm*, 1987, 13(1): 1-14.

[18] 朱梦琦, 朱合华, 王昕, 等. 基于集成 CART 算法的 TBM 掘进参数与围岩等级预测 [J]. 岩石力学与工程学报, 2020, 39(9): 1860-1871.

[19] Ghiasi M M, Zendehboudi S, Mohsenipour A A. Decision tree-based diagnosis of coronary artery disease: CART model [J]. *Comput Meth Prog Bio*, 2020, 192 (prepublish): 105400.

[20] 雷建勤, 陈新, 洪鎏, 等. 基于 MARS 与 AIC 融合的能

源碳排放预测方法研究 [J]. 能源与环保, 2024, 46(12): 185-189+200.

[21] 张艺潇, 赵忠国, 郑江华. 利用 MARS 估算不同气象要素组合下的参考作物蒸散量 [J]. 武汉大学学报 (信息科学版), 2022, 47(5): 789-798.

[22] Zhang W G. *MARS Applications in Geotechnical Engineering Systems* [M]. Singapore: Springer, 2019.

[23] 张永超, 徐芳芳, 李执栋, 等. 基于广义路径追踪算法建立桂枝茯苓胶囊和天舒胶囊中间体水分的近红外光谱通用定量模型 [J]. 中草药, 2023, 54(22): 7436-7344.

[24] Friedman H J. Fast sparse regression and classification [J]. *Int J Forecast*, 2012, 28(3): 722-738.

[25] 张伟国, 蒋昆, 宋宇, 等. 基于机器学习和贝叶斯优化的大位移井钻井提速方法 [J]. 石油钻探技术, 2025, 53(2): 38-45.

[26] 周松元, 罗渭, 马伟东, 等. 基于 TreeNet 算法的煤与瓦斯突出预测模型构建研究 [J]. 矿业研究与开发, 2022, 42(5): 190-196.

[27] 刘元铭, 王振华, 王涛, 等. 热轧带钢出口凸度数据驱动建模及智能化预测分析 [J]. 中国机械工程, 2020, 31(22): 2728-2733.

[28] 郑伟达, 张惠然, 胡红青, 等. 基于不同机器学习算法的钙钛矿材料性能预测 [J]. 中国有色金属学报, 2019, 29(4): 803-809.

[29] 徐冰, 史新元, 罗赣, 等. 中药工业大数据关键技术与应用 [J]. 中国中药杂志, 2020, 45(2): 221-232.

[责任编辑 郑礼胜]