第 54 卷第 3 期
2026 年 3 月

硅 酸 盐 学 报
JOURNAL OF THE CHINESE CERAMIC SOCIETY

Vol. 54，No. 3
March，2026

http://jccsoc.cbpt.cnki.net

DOI：10.14062/j.issn.0454-5648.20250728

AI 赋能的混凝土材料基因数据库智能设计及应用

刘楚瑶[1]，罗梓轩[1]，杨 柳[1]，胡金其[1]，刘赞群[2]，张增起[3]，杨文萃[4]，徐慧宁[4]，陈梦妮[1]

(1. 中南大学计算机学院，长沙 410075；2. 中南大学土木工程学院，长沙 410075；3. 北京科技大学冶金与生态工程学院，北京 100083；4. 哈尔滨工业大学交通科学与工程学院，哈尔滨 150090)

摘 要：针对混凝土数据多源、异构、多尺度的特点，本工作面向材料数据管理、基因智能分析到材料智能研发的典型任务，建立一套 AI 赋能的混凝土材料基因数据库智能设计框架，框架以标准化的混凝土材料基因关系型数据库与图数据库为数据存储基础，设计分层密度聚类算法提取对混凝土材料典型性能产生重要影响的基因组结构。利用机器学习正向设计混凝土材料典型性能预测方法，R^2 相较其他机器学习模型提升约 4%，实现了从特征选择、模型预测到贡献量化的可解释性推理路径，设计融合 ACO 与 NSGA-Ⅱ的、以性能目标驱动的混凝土材料配比逆向优化算法，突破传统配比经验依赖与单目标优化局限，通过全局搜索避免陷入局部最优解。系统化智能化框架旨在突破传统数据库在材料计算与智能分析方面的瓶颈，可为材料基因智能化研究提供借鉴。

关键词：混凝土；数据库智能设计；材料基因；材料性能；人工智能
中图分类号：TP181；TU528　　**文献标志码**：A　　**文章编号**：0454–5648(2026)03–0909–13
网络出版时间：2026–02–10

混凝土材料是全球用量仅小于水的材料[1]，由粗骨料、细骨料、水泥、矿物外加剂(如粉煤灰、矿粉、石灰石粉等)，化学外加剂(如减水剂、引气剂等)和水等原材料组成，其性能包括施工性能、力学性能和耐久性能。传统混凝土材料设计主要是基于鲍罗米公式的抗压强度设计方法，鲍罗米公式展示了水泥强度等级及用量、粗细骨料的种类/粒径分布特征及用量、矿物外加剂的品种及用量、化学外加剂的种类及用量、搅拌用水量等原材料与混凝土抗压强度之间的理论关系，上述过程需结合大量试验研究，并通过试错的方法来验证其使用性能是否满足工程要求，因此，其设计过程周期长，人力和材料成本高，研发效率低下。

2011 年，美国启动材料基因组计划(MGI)[2]，其主要研究思路是通过高通量试验表征技术得到大量有关材料组成基因及其对应的性能参数，建立材料基因组数据库平台，进而采用数学计算和理论方法分析，找出材料基因和性能之间的对应关系，从而预测满足某种功能的材料组成与配比，并通过试验合成该材料进行验证[3]。材料基因组计划可以大幅降低人工试验量，并指导改变基因来进行特定目标材料的设计。为实现这一变革性的重大工程，我国科学技术部于 2015 年设立材料基因工程关键技术与支撑平台重点专项，并在 2023 年的“十四五”材料领域设置 4 个重点专项，布局“材料基因工程技术应用”研究任务，旨在加速中国新材料的研发与应用，突破当前材料研发与产业应用过程中诸多瓶颈问题。

随着大数据与人工智能技术的快速发展，以机器学习为核心的人工智能技术为材料的研发提供了

收稿日期：2025–10–09。　修订日期：2025–10–30。
基金项目：国家重点研发计划(2022YFB2602602)。
第一作者：刘楚瑶(2002—)，女，硕士研究生。
通信作者：杨 柳 (1979—)，女，博士，教授；
刘赞群 (1975—)，男，博士，教授；
杨文萃 (1981—)，女，博士，副教授；
张增起 (1991—)，男，博士，教授。

Received date: 2025–10–09.　**Revised date:** 2025–10–30.
First author: LIU Chuyao (2002–), female, Master candidate.
E-mail: 244711047@csu.edu.cn
Correspondent author: YANG Liu (1979–), female, Ph.D., Professor.
LIU Zanqun (1975–), male, Ph.D., Professor.
YANG Wencui (1981–), female, Ph.D., Associate Professor.
ZHANG Zengqi (1991–), male, Ph.D., Professor.
E-mail: yangliu@csu.edu.cn; zanqun.liu@csu.edu.cn; yangwencui@hit.edu.cn; zhangzq@ustb.edu.cn

更加高效的方法，推动了材料研发由“经验+试错”向“数据驱动”研发模式的转变，更加容易建立材料“成分–结构–组织–性能”的构效关系，实现材料多层次、多维度目标量的有效预测，已在钢铁[4]、生物[5]、医药[6–7]、有机物[8]等材料研发中取得突破性进展，实现了基于特定目标的新材料研发。

采用机器学习[9]、神经网络[10]等人工智能技术研究混凝土配合比设计已在 20 世纪末和 21 世纪初开展[11] 。2003 年，刘国华等[12–13]率先将机器学习中的反向传播(BPNN)与蒙特卡罗(Monte-Carlo)随机试验法相结合，建立了普通混凝土配合比的优化设计模型。据文献[14]的不完全统计，至今约有 70 篇相关文献研究了人工智能在混凝土领域的单一目标优化或多目标优化方案，包括普通混凝土、超/高性能/高强混凝土、再生骨料混凝土、地聚物/碱激发混凝土、轻质骨料混凝土、其他类型混凝土等。分析上述文献可知，其收集的混凝土数据量有限，最少只有 10 组，最多为 10 000 组，其研究的变量(基因)主要为上述混凝土各种原材料组成的宏观质量比。然而，在材料研发模型开发中，数据质量直接影响构建模型的准确性，主要表现为数据误差会影响模型对数据规律的提取与学习，降低模型评估过程的可靠性，从而影响模型的泛化能力。对于不同来源的数据，其数据维度的差异会增加数据清洗、数据整合等工作难度，增加模型的复杂性与构建成本，降低研发效率。准确、高质量的数据库是实现模型开发的基础，不仅可以降低模型构建难度和减少优化工作量，而且可以提升模型预测精度与泛化能力[15]。因此，高质量数据库是混凝土材料人工智能研究的前提，更是材料基因工程实现的基础[16]。

材料的基因，可以是原子、分子、颗粒等不同粒度的基本单元，也可以是这些基本单元的组成结构特征(如孔隙结构等)。混凝土材料是一种多物相、多层次、非均质的颗粒堆积复合材料，随着混凝土科技的不断发展，各种矿物外加剂和化学外加剂等新型材料的应用，使得其组成与结构影响因素十分复杂，对于其基因的定义和选取标准很难取得统一；同时，构建混凝土材料数据库需要大量混凝土组成、结构和性能参数的高质量数据，而设计高通量试验由于其试验设计复杂导致获取大量数据十分困难，在实际生产过程中，很多商品混凝土搅拌站的数据并不能真实反映实际生产情况，但研究论文中的数据兼具真实性和可靠性。

为此，中南大学、哈尔滨工业大学、北京科技大学等高校开展联合研究，结合特征工程选定混凝土的 4 种典型性能，包括抗压强度、抗氯离子渗透性能、抗碳化性能及抗冻融破坏性能。从 2010—2025 年多国发表的 9000 余篇文献中，筛选出 1500 余篇满足以下条件的文献：研究主题明确涉及上述 4 类性能的定量分析，性能测试方法符合国际或国内标准，数据呈现形式为可直接提取的数值而非仅通过图表进行定性分析；同时过滤掉性能数据为模拟值、未说明测试方法或存在关键参数缺失的文献。论文所用代码的获取链接为 https://gitee.com/luo-zixuan/concert。

1 人工智能相关技术

本节介绍混凝土材料典型性能基因数据库智能设计中运用的主要人工智能技术，包括沙普利加性解释(SHAP)方法、聚类方法、非线性降维映射、随机森林回归器、蚁群算法、非支配排序遗传算法以及评估函数等。

1.1 SHAP 方法

SHAP(Shapley additive explanations)[17]是一种基于博弈论的模型可解释性方法，用于评估每个特征在机器学习模型中的重要性及其影响，其将每个特征视为一个参与者，通过计算该特征在所有可能的特征组合中的边际贡献，来量化特征在模型预测中的作用。

通过 SHAP 可直观表示每个特征对最终预测结果的贡献大小，以及其作用是正面还是负面。因此，SHAP 可视化呈现出特征对每个实例的影响，为研究人员提供数据分析与决策的可解释性依据。

1.2 聚类方法

聚类方法[18]是将无标签数据按相似性自动分组的无监督学习方法。K 均值(K-means)聚类和层次聚类是 2 种主要的聚类方法。

1) K-means 聚类法[19]的目标是将数据集划分为 k 个簇，使得每个数据点属于距离最近的簇中心。通过反复调整簇中心的位置，不断优化簇内的紧密程度，从而获得尽量紧凑、彼此分离的簇。后续对混凝土材料基因特征组的提取过程中，将改进 K-means 聚类方法，并结合加权欧氏距离和余弦相似度以提高聚类的准确性。

$$d_{\mathrm{m}}(x,y)=\alpha\cdot d_{\mathrm{w}}(x,y)+(1-\alpha)\cdot\left(1-\mathrm{sim}(x,y)\right) \quad (1)$$

式中：$d_{\mathrm{w}}(x,y)=\sqrt{\sum w_{\mathrm{i}}\left(x_{\mathrm{i}}-y_i\right)^2}$ 为加权欧氏距离；$\mathrm{sim}(x,y)$ 为余弦相似度；平衡系数 α 控制两者的贡

献度，可通过交叉验证选定。

2) 层次聚类法[20]是一种建立在距离度量基础上的聚类方法，以分层的方式逐步构建聚类结构。其通过递归合并相似的样本直至所有样本归为 1 个大类或达到预设聚类数，最终形成树状结构，可以捕捉混凝土材料基因特征的层级关系，直观基因特征之间的关联逻辑。

在混凝土材料的研究中，聚类方法不仅揭示了不同特征间的相互关系，还为进一步分析提供直观的可视化工具，使得特征重要性评估更为清晰。

1.3　非线性降维映射

非线性降维映射(UMAP)[21]是一种基于流形学习理论的非线性降维技术，旨在将高维数据映射到低维空间，同时保持数据的全局和局部结构。在降维过程中，UMAP 首先为高维数据中的每个点构建 1 个加权邻接图，并利用局部相似性来捕捉数据的整体结构，再通过优化目标函数将这种图映射到低维空间，以确保近邻关系尽可能被保留。最终，UMAP 将高维数据以二维或三维形式展现，方便进行样本间的相似性分析和特征聚类。

在混凝土材料的特征分析中，可以利用 UMAP 识别新的潜在特征组合，通过揭示复杂的特征间关系来准确分析混凝土材料基因组。

1.4　随机森林回归器

随机森林回归器(RFR)[22]是集成学习里的高性能回归模型，通过构建多棵决策树并对各棵树的预测结果进行集成，以完成回归任务。

在随机森林的构建过程中，每棵决策树均独立进行训练，且训练数据来自从原始数据中随机抽取的样本子集；同时，每棵树的每个分裂节点处，最优分裂特征的筛选范围被限定在随机选取的特征子集中。这种“样本抽样+特征选择”的双重随机性设计，能有效削弱单棵决策树对局部数据的过度依赖，进而显著降低模型的过拟合风险。

在混凝土典型性能预测任务中，RFR 可基于水、水灰比、粗细骨料等多维输入特征开展预测实验，通过多棵决策树的协同运算，对各决策树输出的性能预测值进行加权平均，形成最终的性能预测结果。这种集成学习机制不仅能高效处理混凝土性能与各输入特征之间的非线性耦合关系，还能显著削弱单一变量波动对预测结果的干扰，增强预测稳定性。

1.5　蚁群算法

蚁群算法(ACO)[23]是模拟蚂蚁觅食行为的仿生优化算法，通过信息素引导的群体协同决策来求解优化问题。蚂蚁觅食时会在路径上释放信息素，后续个体倾向选择信息素浓度高的路径，且会在该路径上进一步释放信息素以增强浓度，同时已释放的信息素会随时间自然衰减。通过这种累积与挥发的动态机制，蚁群的探索从无序逐步收敛于最优路径，形成具有自增强特性的正反馈循环。

ACO 全局搜索能力强，其群体协作可避免局部最优，对离散组合优化适配性好。在混凝土材料配比优化中，可将原材料组成、结构、配比用量映射为路径决策，以性能或成本为目标，通过迭代收敛得到多指标最优的配比方案，为材料工程智能配比研发提供技术支持。

1.6　非支配排序遗传算法

带精英策略的非支配排序遗传算法(NSGA-Ⅱ)[24]是基于进化理论的经典多目标优化算法，利用高效的非支配排序机制、精英保留策略及拥挤度计算方法，在多目标优化领域具备极强的实用性。

NSGA-Ⅱ通过模拟自然选择中的“选择–交叉–变异”遗传操作，实现多目标优化问题求解。NSGA-Ⅱ以构建初始候选群体为起点，对群体开展迭代更新：在每一轮迭代中，结合非支配排序机制与拥挤度计算方法筛选出优势个体，确保后续进化方向持续趋近多目标优化的最优前沿；同时融合不同优势个体在各目标维度的优势特征，生成继承双亲优良属性的新精英个体，提升子代群体的整体优化潜力；并通过微调部分基因参数，打破群体可能陷入的局部最优困境，为种群注入新的进化方向，进而有效避免算法出现早熟收敛问题。

在混凝土材料多目标配比优化场景中，将 NSGA-Ⅱ核心算法与蚁群算法结合起来，生成能够同时兼备混凝土材料强度、成本、耐久性等核心目标的帕累托(Pareto)最优配比集，满足实际工业场景对混凝土材料的多目标性能要求。

1.7　评估函数

为系统评估模型预测结果的准确性与可靠性，选取平均绝对误差(MAE)、均方误差(MSE)、决定系数(R^2)3 个评估指标，具体指标的定义与计算如下。

MAE 是预测值与实际值绝对误差的期望值，反映出预测误差的实际情况。

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|f_i-y_i\right| \tag{2}$$

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}\left(f_{\mathrm{i}}-y_{\mathrm{i}}\right)^2 \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(f_i - y_i)^2}{\sum_{i=1}^{m}(\bar{y} - y_i)^2} \tag{4}$$

式中：m为样本数量；MSE为预测值与实际值偏差平方的算术平均值，反映预测模型的精度；R^2反映回归模型对数据的拟合程度，取值范围为 0~1，其结果越接近 1，表示模型拟合效果越好；$\bar{y}$为实际值的平均值。

2 混凝土材料基因数据库智能设计框架

随着当前智能化水平的提高，研究者们开始借助人工智能在海量数据分析、复杂关系建模及多变量优化方面等优势来构建智能化材料数据库，进而推动材料领域向数字化、智能化的方向转型。人工智能赋能混凝土材料基因数据库设计，即通过机器学习、深度学习等人工智能方法挖掘混凝土材料成分、结构与性能之间的复杂关系，为高性能混凝土材料研发提供数据支撑。

本工作以“人工智能驱动混凝土材料智能设计”为核心目标，构建了 AI 赋能的混凝土材料基因数据库智能设计框架，如图 1 所示。该框架包括混凝土材料基因数据库设计、混凝土材料基因特征分析、混凝土材料智能设计 3 个部分。混凝土材料基因数据库设计遵循数据库标准化设计的 FAIR 原则与规范化设计的 BC 范式，搭建标准化的混凝土材料基因关系型数据库与图数据库，以实现混凝土全生命周期数据的规范存储与高效管理；混凝土材料基因特征分析对原始数据库中的数据进行预处理，并针对混凝土典型性能提取对应的基因特征组；混凝土材料智能设计包括材料正向设计与逆向优化设计 2 个方面，正向设计是通过混凝土材料的成分、结构与配比等特征预测材料性能，逆向优化设计是在约定混凝土材料目标性能的前提下，利用反馈迭代的方式优化材料的成分、结构与配比。

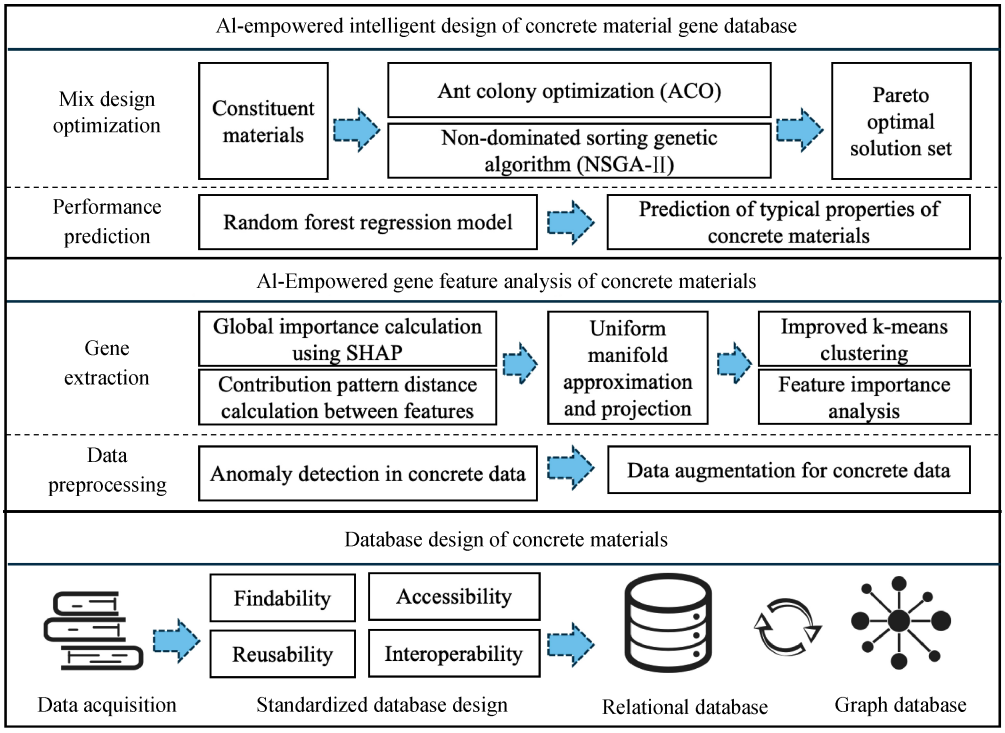


图 1 AI 赋能的混凝土材料基因数据库智能设计框架

Fig. 1 AI-Empowered Intelligent Design Framework for Concrete Material Gene Database

3 混凝土材料基因数据库设计

本节通过智能化技术构建了混凝土材料基因数据库(GMGD)，包括能够用于大规模数据存储管理的关系型数据库和建模基因特征关系的图数据库。

3.1 数据库设计思路

混凝土材料基因数据库的设计包括数据采集、标准化设计、混合存储架构 3 个核心环节。

数据采集是混凝土材料基因数据库构建的第一步。混凝土材料基因数据库中的原始数据是从筛选的 1500 余篇文献中采集得到，涵盖了混凝土抗压强度、混凝土抗氯离子渗透性能、混凝土抗碳化性能和混凝土抗冻性能等特征数据共计 5774 条，包括 130 000 多个特征数据，在已有研究对混凝土基

因宏观层次选定的基础上，扩展至水泥、矿物掺合料等化学组成，粗/细骨料粒径分布特征、粗细程度等细微观层面。混凝土材料原始数据集的数据特征描述如表 1 所示，上述数据包括材料成分、结构、配比参数、环境因素等信息以及抗压强度、抗碳化性、抗冻性等典型性能指标。其中，抗压强度数据表包括 72 个特征项、4083 条数据；抗氯离子渗透性能数据表包括 74 个特征项、2981 条数据；抗碳化性能数据表包括 34 个特征项、1587 条数据；抗冻性能数据表包括 99 个特征项、1206 条数据。

表 1　混凝土材料原始数据集的数据特征描述

Table 1　Data feature description of the concrete materials original dataset

Feature	Detailed description
Cement mineral composition	The main mineral composition of cement consists of CaO, SiO_2, Al_2O_3, MgO, and Fe_2O_3, unit %
Cement strength grade	Cement strength grade refers to the compressive strength achieved after 28 d of curing. China's silicate cement is divided into 4 grades and 8 types, with grades of 32.5, 42.5, 52.5, 62.5
Water content	Water usage, unit kg/m^3
Water-cement ratio	The ratio of water to cement mass in concrete
Water reducing agent dosage	Dosage of water reducing agent, unit kg/m^3
Water reducing agent's water Reduction rate	The ratio of water reduction after adding water reducing agent to water usage without water reducing agent, unit %
Aerating agent gas content	Air content in concrete after adding air-entraining agent, unit %
Air-entraining agent dosage	The dosage of air-entraining agent in concrete, unit kg/m^3
Crushed stone particle size	Divided into seven ranges: 5–10 mm, 5.0–12.5 mm, 5–16 mm, 5–20 mm, 5–30 mm, 10–20 mm, 20–30 mm
Crushed stone content	Crushed stone usage, unit kg/m^3
Natural sand types	Natural sand is a type of fine aggregate, formed under natural conditions, primarily consisting of river sand
Natural sand fineness modulus	An indicator that represents the degree of coarseness of natural sand, the larger the fineness modulus, the coarser the sand
Natural sand Bulk density	Bulk density refers to the mass per unit volume measured in the loose state, unit kg/m^3
Natural sand apparent density	Apparent density is the ratio of material mass to apparent volume, unit kg/m^3
Natural sand dosage	Natural sand usage, unit kg/m^3
Natural sand clay content	Clay content in natural sand, unit %
Artificial sand type	Artificial sand is formed by crushing rocks, also known as machine-made sand
Artificial sand fineness modulus	An indicator of the coarseness of artificial sand, with a larger fineness modulus indicating coarser sand
Artificial sand bulk density	The ratio of material mass to apparent volume, unit kg/m^3
Artificial sand powder content	Proportion of stone powder in manufactured sand, unit %
Artificial sand usage	The usage amount of manufactured sand, unit kg/m^3
Artificial sand ratio	Artificial sand's sand content ratio, unit %
Fly ash composition	Composed of seven indicators, including CaO, SiO_2, Al_2O_3, Fe_2O_3, MgO, SO_3, and loss, unit %
Fly ash class Ⅰ content	Ⅰ-grade fly ash usage, unit kg/m^3
Fly ash class Ⅱ content	Ⅱ-grade fly ash usage, unit kg/m^3
Fly ash content	Total fly ash usage, unit kg/m^3
Slag composition	Composed of seven indicators, including CaO, SiO_2, Al_2O_3, Fe_2O_3, MgO, SO_3, and loss, unit %
Slag content	Usage amount of slag, unit kg/m^3
Limestone powder composition	Composed of seven indicators, including CaO, SiO_2, Al_2O_3, Fe_2O_3, MgO, SO_3, and loss, unit %
Limestone powder specific Surface area	Limestone powder fineness index, unit m^2/g
Limestone powder dosage	Limestone powder usage, unit kg/m^3
Silica fume composition	Composed of seven indicators, including CaO, SiO_2, Al_2O_3, Fe_2O_3, MgO, SO_3, and loss, unit %
Silica fume specific surface area	Silica fume fineness index, unit m^2/g
Silica fume dosage	Silica fume usage, unit kg/m^3
Porosity	Ratio of total void space volume to total concrete volume, unit %
Slump	Degree of flow deformation under self-gravity during concrete construction

标准化设计是混凝土材料基因数据库实现开放、共享的基础。混凝土材料基因数据库设计遵循指导科学数据管理和治理的 FAIR 准则[25]，通过提升数据的可查找性(Findability)[26]、可访问性(Accessibility)[27]、互操作性(Interoperability)和可重用性(Reusability)，以促进数据的高效共享与长期价值挖掘。1)数据库的可查找性，GMGD 数据库结合标准化的元数据格式和唯一标识符按照对应数据来源加上序号的形式，确保数据集易于检索和引用，同时通过充分描述实验条件和数据来源，提升数据库的整体可查找性；2)数据库的可访问性，GMGD 数据库致力于设计一个开放获取的数据库，所有数据都可以在混凝土材料基因数据库平台上进行访问，确保混凝土材料数据通过合理的协议和访问控制，以有效促进数据的共享与使用；3)数据的互操作性，GMGD 数据库选择使用通用的标准格式 JSON，以便于与其他现有系统兼容，并促进跨平台的数据交换；4)数据的可重用性，GMGD 数据库参照国家标准计划《水泥基材料元数据》[16]，按照规范水泥基材料元数据的内容、格式和结构，确保混凝土材料数据符合规范，可应用于混凝土材料相关领域。因此，混凝土材料基因数据库设计总体遵循 FAIR 准则，旨在打破混凝土数据孤岛，促进数据的高效共享，充分挖掘混凝土数据的长期科学价值，最大限度发挥其在科研、工程等领域的作用。

混合存储架构是大规模数据智能化分析的重要保障。混凝土材料数据多样且复杂，单一数据库难以同时满足高效管理结构化数据与深度挖掘复杂关联关系的双重需求。因此，本工作针对混凝土材料特点，提出了关系型数据库与图数据库相结合的混凝土材料数据存储结构。混凝土材料关系型数据库以 MySQL 数据库为基础，用于管理结构化程度高、关系明确的混凝土数据，并利用其成熟的“表–关系”模型和结构化查询能力，实现混凝土材料的高效数据存储与检索。混凝土材料知识图谱以 Neo4j 图数据库为基础，用于探索混凝土材料特征之间的拓扑结构与耦合关系，以增强数据的结构表达能力，并以可视化图的方式实现对关联数据的高效分析。

3.2 混凝土材料基因关系型数据库设计

在混凝土材料基因关系型数据库设计过程中，E–R（实体–关系）模型设计和数据库表设计是 2 个核心工作。E–R 模型设计是描述物理世界中实体和实体之间关系的概念模型，数据库表设计进一步将概念模型转换为信息世界中的数据库结构。

针对混凝土材料多源、多粒度、多尺度特点，以及影响混凝土材料性能与成分、结构之间的关联关系，构建混凝土材料基因关系型数据库的E–R模型，如图 2 所示。其中，混凝土为核心实体，水、水泥、矿物、引气剂、粗细骨料、掺合料作为其组成成分形成多对多关系；部分组成成分还可以按类型细分，例如掺合料分为煤灰粉、矿渣、石灰粉、硅粉等；

不同的组成成分拥有各自的关键属性，例如碎石粒径包括 5~10 mm、5.0~12.5 mm、5~16 mm、5~20 mm、5~30 mm、10~20 mm、20~30 mm 等不同规格。另外，混凝土典型性能也作为关键实体进行描述，包括了力学性能、抗碳化性能、氯离子抗渗性能、抗冻融性能。因此，E–R 模型系统建模了混凝土的原材料成分、结构、配比以及性能之间的数据结构与数据关联。

混凝土材料基因关系型数据库设计是在 E–R 模型的基础上，将模型中的每个实体映射为数据库中一张独立表，并按数据类型、数据精度、约束范围等数据特性定义数据库的字段；依据关系型数据库标准化设计的BC范式(BCNF)优化数据库表结构，包括拆分表以消除关联依赖、关联多张表以处理多对多关系等；对数据库表设置表约束，包括设置唯一约束与非空约束、设置自增主键、关联表外键等；最后，统一规范混凝土材料的字段与类型，确保数据的完备性与兼容性，形成规范的大规模混凝土材料基因数据领域的结构化数据库设计。

基于上述结构化数据库设计，选择开源关系型数据库 MySQL 对混凝土材料基因数据进行数据存储与管理，以满足数据开放共享与智能化应用的需求。

3.3 混凝土材料基因图数据库设计

图数据库是以节点与边所形成的图结构来描述数据及数据间的复杂关系。因此，区别于混凝土材料基因关系型数据库的结构化管理方式，混凝土材料基因图数据库是针对混凝土材料性能与成分、结构、配比之间存在的复杂非线性关系，聚焦于材料基因要素间复杂关联的建模与挖掘。同时，基于图数据库建立的混凝土材料基因知识图谱，可以辅助构建混凝土材料基因领域知识体系，进一步支撑混凝土材料领域的智能化发展。

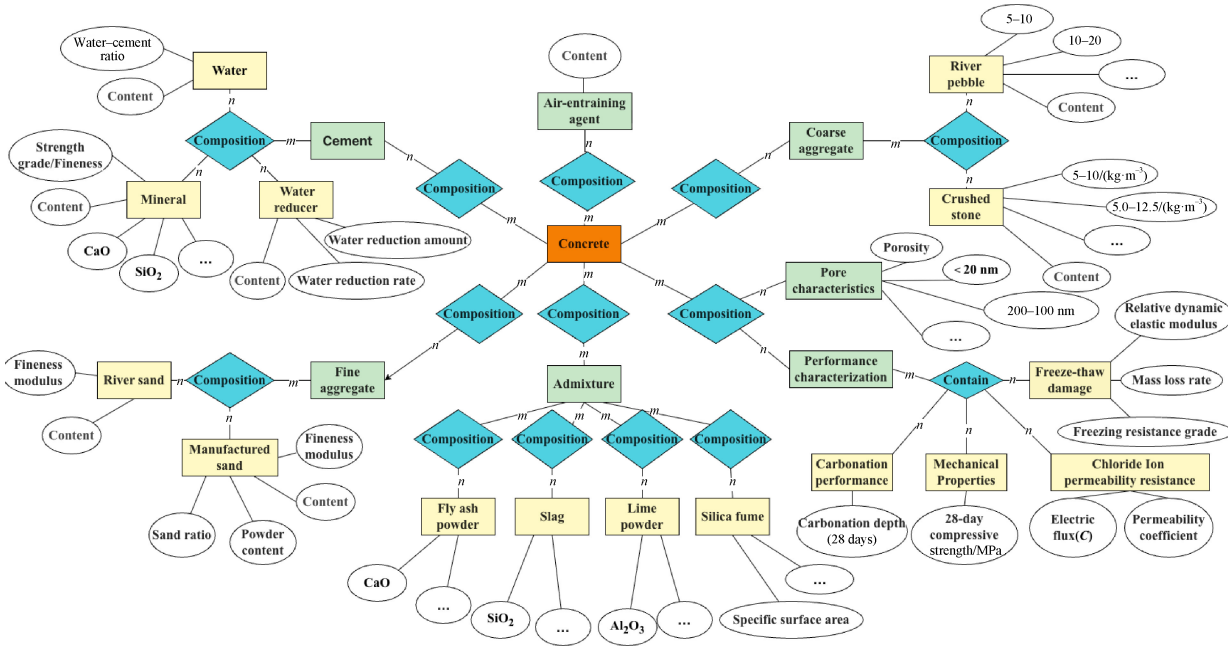


图 2　混凝土材料基因数据库 E–R 模型设计

Fig. 2　Design of E–R model for concrete material gene database

Neo4j 作为图数据库代表性工具，能够支持多格式导入及多语言应用程序编程接口(API)，与 MySQL 关系型数据库适配性强。因此，选用 Neo4j 作为图数据库工具，将混凝土材料关系型数据库映射为图数据库。将关系型数据库中的独立表直接转换为图数据库中的实体，并在图数据库中以节点的方式可视化，数据库表中原主键映射为图数据库中实体节点的唯一标识，而其他字段则映射为实体节点的关联属性；对于关系型数据库中表与表之间的关系，如“混凝土”与“水泥”之间的关系，映射为图数据库中带箭头的边。

基于混凝土材料图数据库中的数据构建出的混凝土材料基因知识图谱如图 3 所示，图中红色节点代表“混凝土”，绿色节点代表混凝土的组成成分、环境因素，橙色节点代表混凝土材料的典型性能，其他颜色的节点则是与之相连节点的关联属性。

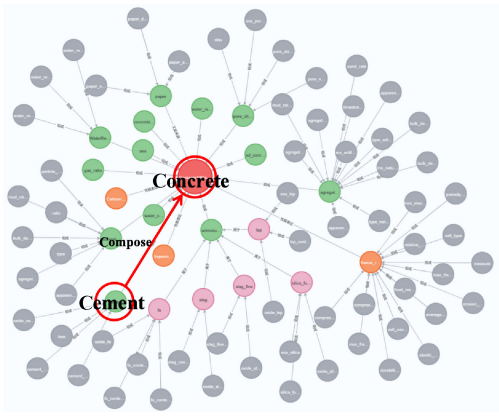


图 3　混凝土材料基因知识图谱

Fig. 3　Concrete material gene knowledge graph

4　人工智能赋能混凝土材料基因特征分析

在上述混凝土材料基因数据库的基础上，本工作提出基因重要性驱动分层密度聚类算法，实现对混凝土材料基因特征组的提取与分析。

4.1　混凝土数据预处理

混凝土数据预处理是为了保障数据的质量与特征的有效性。混凝土材料数据预处理的流程可以分为 2 个部分，异常检测和数据增强。首先对混凝土原始数据执行异常检测，精准识别出偏离正常分布的异常样本；再针对数据集中的空缺值开展定向填补，并通过独热编码、归一化等标准化技术，实现数据格式的统一与数据分布的平衡。最终得到混凝土增强数据，为开展混凝土材料基因特征分析与混凝土材料智能设计提供数据支撑。

4.2　混凝土材料基因特征组提取与分析

混凝土材料基因特征组是对混凝土典型性能具有直接且显著影响的多个特征的组合，然而在混凝土材料的配比设计与耐久性能优化中，研究者常面对数十甚至上百个配方因子(例如水胶比、矿物掺合料含量、砂率等)，它们既相互影响，又往往以群组形式共同作用，难以单独评估。因此，本工作

提出一种基因重要性驱动分层密度聚类算法，在 SHAP 可解释性分析方法的基础上，基于分层密度聚类算法构建基因组提取模型，实现基因集合的自适应识别，通过从普通特征中筛选对混凝土典型性能具有显著影响的核心关键的特征，并得到不同作用关系的基因特征组。图 4 是混凝土材料基因特征组提取与分析的基本流程。

1) 混凝土材料数据特征重要性分析。

对混凝土增强数据集使用 SHAP 方法计算每个特征的全局重要性，对 SHAP 值进行可视化分析，揭示特征对整体性能的重要贡献。将每个特征的全局重要性组合起来得到一个表明每个特征在混凝土典型性能评估中相对重要性的特征重要性向量。

2) 混凝土材料特征间贡献模式距离计算。

对特征重要性向量进行 Z 分数(Z-score)标准化，消除不同特征间量纲的差异，将所有特征的分布趋于标准正态分布。基于标准化后的向量，计算特征之间的欧氏距离矩阵。该矩阵反映了特征在模型贡献模式上的相似性。

3) 混凝土材料数据的非线性降维。

为减少冗余信息的干扰并降低计算复杂度，采用 UMAP 方法对计算得到的特征之间的欧氏距离向量矩阵进行非线性降维，从高维特征空间中提取出更为显著的低维特征表示。UMAP 旨在保留特征间的拓扑结构，通过最小化高维与低维表示之间的邻域分布差异，确保降维过程不会显著损失信息。这一步骤为后续的聚类分析奠定基础。

4) 混凝土材料数据的特征聚类。

在获得低维向量矩阵表示后，使用轮廓系数法选择最佳的聚类数，利用改进的 K-means 聚类算法，得到最终的特征聚类结果。改进的 K-means 聚类算法可以将混凝土特征分为若干组，以便识别其在对混凝土材料性能上影响的相似性和差异性。在此基础上，进一步使用层次聚类方法，可视化出各特征之间的结构关系和功能相似性。

5) 混凝土材料基因组特征分析。

利用得到的每个特征的全局特征重要性对所有特征进行排序，确定“主特征集”和“次特征集”。主特征集由显著高于中位数的特征组成，作为对典型性能具有关键影响的基因特征组，而次特征集则为补充性因子。利用简化的回归模型评估主特征集的预测能力，并进行增量测试来识别对特定性能指标有显著影响的辅助因子。

通过上述方法识别出对混凝土材料典型性能影响最大的主基因特征组，确立一阶核心基因单元，而其他特征则作为辅助因子。该混凝土材料基因特征组提取与分析方法构建了混凝土材料性能基因组结构模型，揭示各类特征在对不同混凝土材料典型性能影响过程中的系统性关联关系，为总体提升性能和优化配合比提供依据。

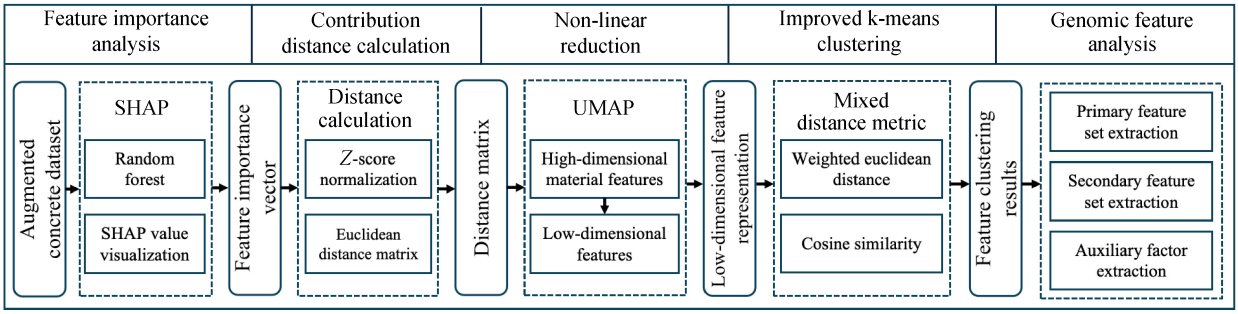


图 4 混凝土材料基因特征组提取与分析流程

Fig. 4 Extracting and analyzing the characteristic genome of concrete materials flowchart

5 人工智能赋能混凝土材料智能设计

本节围绕人工智能赋能混凝土材料智能设计展开分析，阐明混凝土典型性能预测分析与多目标配比优化的技术路径与实验分析过程。

5.1 混凝土材料典型性能预测分析

混凝土的典型性能直接关系到工程结构的安全性与使用寿命，因此，论文设计了一种基于随机森林与可解释性分析的性能预测模型，对混凝土典型性能实现精准预测，如图 5 所示。由图 5 可见，预测模型以混凝土材料的成分、结构和配比作为输入特征，采用随机森林算法构建高性能预测器，并通过交叉验证策略优化模型的泛化能力。同时进一步引入 SHAP 方法，定量评估各基因特征对目标性能

的贡献度，以提高模型的可解释性。

为验证模型准确率，论文针对混凝土的抗压强度、抗氯离子渗透性、抗碳化性及抗冻性 4 项典型性能开展预测实验。其中，抗压强度直接决定混凝土结构的承载能力，是混凝土结构设计中不可替代的首要依据；抗氯离子渗透性则直接影响混凝土在海洋、盐渍等恶劣环境下的耐久性与服役寿命，是混凝土防腐蚀设计的核心指标。因此，分别以 28 d 抗压强度、氯离子渗透深度为例，衡量上述 2 种混凝土材料典型性能的优劣，对二者开展可解释性分析，为混凝土性能优化设计提供切实的技术支撑。

1) 抗压强度性能

论文基于混凝土材料基因特征分析技术初步筛选出 11 个与 28 d 抗压强度紧密关联的关键基因特征，即氧化钙(CaO)、二氧化硅(SiO_2)、氧化镁(MgO)、氧化铁(Fe_2O_3)、水泥掺量、水掺量、水灰比、减水剂掺量、河砂掺量、孔隙率和煤灰(MH)总用量，以此构建了具有 1226 组有效样本的抗压强度数据集。最终模型在数据集上展现出优秀的准确率，其中，MSE 为 4.3636，MAE 仅为 34.2192，R^2 高达 0.9186；而在同类对比实验中，如极限梯度提升算法(XGBoost)与极端随机树算法(ExtraTree)的 R^2 分别为 0.8210 和 0.8679，且两者 MAE 和 MSE 均高于论文模型。

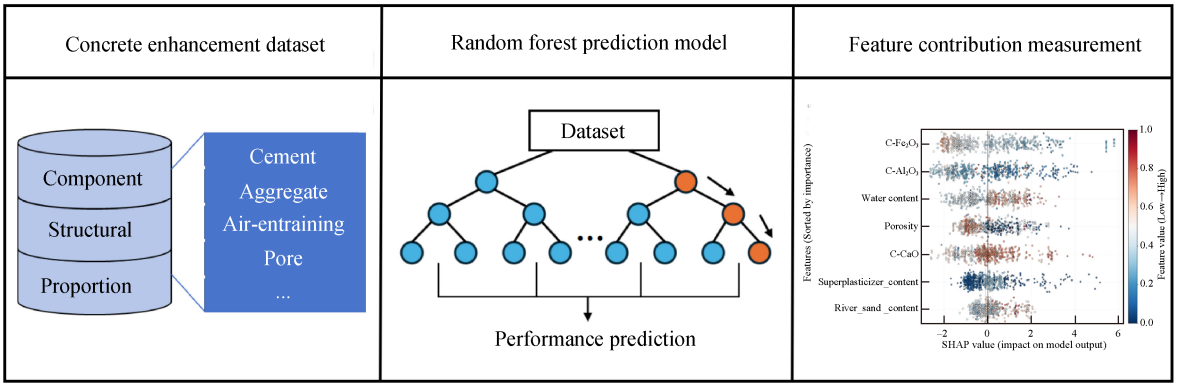
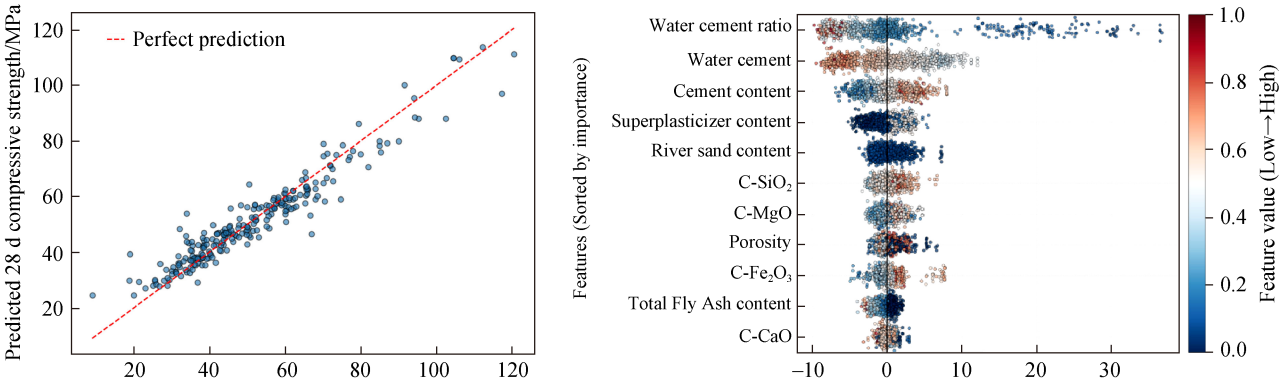


图 5　基于随机森林与可解释性分析的性能预测模型

Fig. 5　Performance prediction model based on random forests and interpretability analysis

混凝土抗压强度性能的预测分析结果如图 6 所示。从图 6a 预测拟合图可知，绝大多数样本点紧密分布于参考线两侧，说明模型拟合能力较强，预测误差小，预测准确率高。结合图 6b 对 28 d 抗压强度的特征贡献图可知，水灰比、水掺量及水泥掺量是影响抗压强度的关键特征。水灰比在特征值高的区间存在大量负 SHAP 值，表明高水灰比会显著降低混凝土抗压强度，而低水灰比通过调控水泥水化致密性，对抗压强度提升具有积极作用；同理水掺量高时 SHAP 值多为负向，体现了过量水分会增大硬化混凝土孔隙率，进而抑制抗压强度提升；水泥掺量在高值区间的 SHAP 值多呈正向，反映充足的水泥掺量为水化反应提供了充分的胶凝物质基础，对提升抗压强度贡献显著。



(a) 28 d compressive strength prediction fitting plot　　(b) 28 d compressive strength feature contribution plot

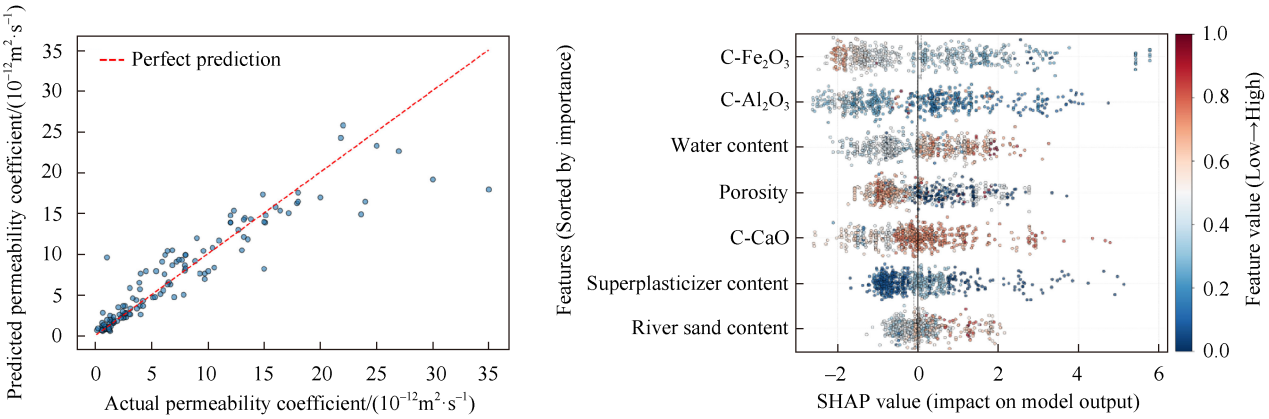
图 6　抗压强度性能预测分析

Fig. 6　Compressive strength performance prediction analysis

2) 抗氯离子渗透性能

论文基于混凝土材料基因特征分析技术初步筛选出 7 个与氯离子渗透深度紧密关联的关键基因特征，即 CaO、氧化铝(Al_2O_3)、Fe_2O_3、水掺量、减水剂掺量、河砂掺量、孔隙率作为输入，并构建了含 641 组有效样本的抗氯离子渗透性能数据集。最终预测模型表现出较高的准确率，其中 MSE 为 6.260 1，MAE 为 1.628 7，R^2 为 0.861 5，而 XGBoost 与 ExtraTree 的 R^2 分别为 0.815 1 和 0.833 5，显著低于所提模型。

抗氯离子渗透性能预测分析如图 7 所示。由图 7a 可见，大部分预测点集中在参考线附近，说明模型拟合能力较强，预测误差小，特别是在氯离子渗透深度低等至中等范围时表现最为稳定。个别样本在高氯离子渗透区间偏离较大，说明在极端工况下仍有提升空间。结合图 7b 中各个特征对氯离子渗透深度的特征贡献图可知，Fe_2O_3 和 Al_2O_3 在特征值高的区间存在较多负向 SHAP 值，表明当二者取高值时，能显著减小氯离子渗透深度，对提升抗氯离子渗透性能具有积极作用；水掺量和 CaO 在特征值高时 SHAP 值多为正值，其中水掺量的正向 SHAP 值分布更集中且幅度突出，说明水掺量取高值时会显著增大氯离子渗透深度导致性能降低；减水剂掺量在特征值低的区间的 SHAP 值多为负值，体现了低减水剂掺量对抗氯离子渗透性能的增强作用；孔隙率和河砂掺量的 SHAP 值正负分布相对均衡，表明其对抗氯离子渗透性能的影响同时具有促进与抑制双重效应。



(a) Chloride ion penetration depth prediction fitting plot　(b) Chloride ion penetration depth feature contribution plot

图 7　抗氯离子渗透性能预测分析

Fig. 7　Chloride ion penetration resistance prediction analysis

综上所述，该预测模型不仅能精准评估模型准确率，还能量化各基因特征对目标性能的贡献度，进而清晰揭示出特征与混凝土典型性能间的内在关系，为混凝土材料性能调优提供了明确的方向。

5.2　多目标驱动的混凝土配比优化

在成分复杂的混凝土材料体系中，配合比设计是决定其服役性能、经济成本与可持续性的核心要素之一。传统的配合比优化方法主要依赖试验经验与线性权衡思路，难以在性能达标、环境友好等多目标间实现高效协同优化。

针对上述局限，多目标优化算法已成为解决该问题的核心技术手段。当前主流方法涵盖粒子群优化算法、非支配排序遗传算法(NSGA-Ⅱ)及蚁群优化算法(ACO)等[28–29]。粒子群优化算法虽实现简单，但在复杂约束场景下易出现早熟收敛缺陷；NSGA-Ⅱ凭借快速非支配排序与拥挤度机制具备高效全局搜索能力，但其局部寻优精度易受种群多样性波动的制约；而 ACO 依托信息素累积-挥发机制，展现出精准挖掘局部最优解的特性，可有效弥补 NSGA-Ⅱ在局部搜索中的不足[30]。因此，本工作提出融合 ACO 和 NSGA-Ⅱ的混凝土配合比优化框架，构建“全局探索–局部开发”的协同增强效应，在保证解集全局多样性的同时提升局部寻优精度，以解决混凝土材料研发中多目标性能冲突问题(见图 8)。

混凝土配比优化框架以材料数据库、工程约束库和性能先验模型库为底层支撑，分别提供成分、设计规则与先验知识。在此基础上，通过实数编码将各成分映射为优化变量后，引入模糊理论与归一化权重构建多目标启发函数，以处理多性能指标的

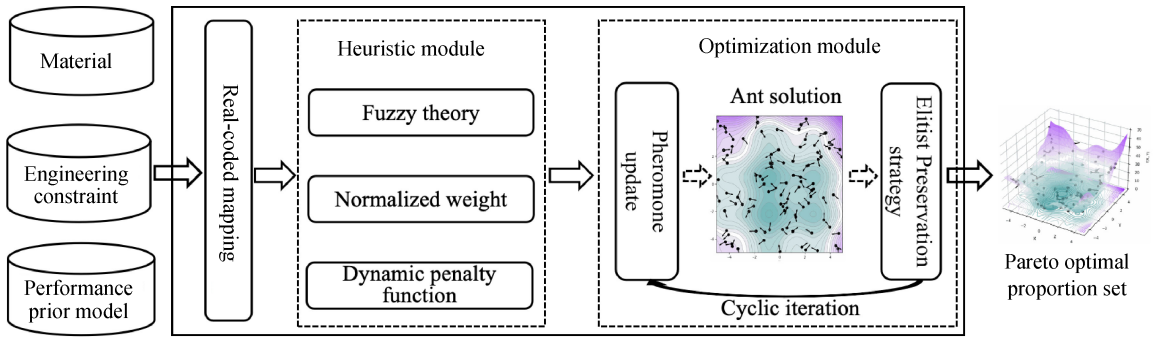


图 8　混凝土材料多目标配比优化框架

Fig. 8　Multi-objective mix design optimization framework for concrete materials

非线性协同问题。针对混凝土材料非线性及经验性约束，设计动态惩罚函数用以修正目标函数，提高对混凝土材料可行域的收敛能力。

核心多目标优化模块通过信息素更新机制，明确引导蚁群的搜索方向：在此过程中，蚁群负责构造候选配比方案，同时通过精英保留策略筛选非劣配比解；经过多轮循环迭代后，最终输出既能满足工程约束条件又能兼顾多目标性能需求的帕累托(Pareto)最优配比集，为混凝土材料配比设计提供丰富的优化方案。

考虑到抗碳化性能直接关联钢筋锈蚀风险与混凝土结构的耐久性，而抗压强度是混凝土最基础的力学性能表征，论文以“碳化深度最小化”与“28 d 抗压强度最大化”为双目标开展配比优化实验，获取了 Pareto 最优配比集，旨在探索不同配比对 2 种性能的影响规律。如图 9 所示的 Pareto 最优前沿图中，样本呈现出显著的性能权衡特征：当碳化深度控制在 0.5 mm 左右时，28 d 抗压强度可稳定处于 200～210 MPa 区间；而当优先保障 28 d 抗压强度(最高可达 260 MPa 以上)时，碳化深度则提升至 0.8 mm 左右。实验结果充分体现了不同工程场景下材料设计的灵活性，若工程更侧重力学承载需求，可选取 28 d 抗压强度高但碳化深度稍大的配比；若聚焦高碳化环境下的耐久需求，则可选用碳化深度小但强度适中的配比方案。

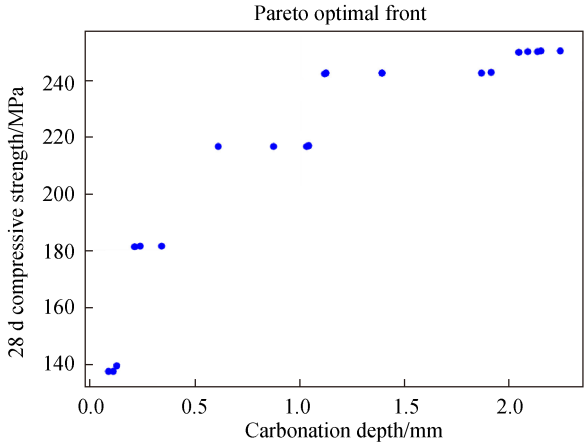


图 9　Pareto 最优前沿图

Fig. 9　Pareto optimal frontier chart

相较于传统混凝土配比设计方法，Pareto 最优配比集的优势方案更能有效实现力学强度与抗碳化性能的协同提升，为实际工程中混凝土材料的选型提供坚实的理论支撑。此外，通过约束配比参数的扰动区间、确保目标函数的物理合理性，可进一步使优化结果具备工程可实施性。与此同时，该配比优化框架还可拓展至混凝土抗氯离子渗透性、抗冻性等其他性能维度的并行优化中，为多性能需求下的混凝土配比设计提供更全面的解决方案。

6　结论与展望

论文研究所构建的 AI 赋能的混凝土材料基因数据库智能设计框架，利用机器学习、深度学习等人工智能前沿技术，不仅为混凝土材料的数据管理、基因智能分析、材料智能研发提供了一套可行的解决方案，而且对材料科学领域的数智化建设具有借鉴意义。在现有研究基础上，将结合混凝土材料数据相关的图像、文本等多模态特征，研究混凝土材料多模态表示学习方法，例如，利用 YOLO 模型进行扫描混凝土相关电子显微镜图像分析，引入大模型对混凝土相关文献文本进行处理，增强对混凝土材料基因特征表示，进一步探索混凝土材料微观结构与材料性能的潜在关系。

参考文献:

[1] LI Z Z, YOON J, ZHANG R, et al. Machine learning in concrete science: Applications, challenges, and best practices[J]. NPJ Comput

Mater, 2022, 8: 127.
[2] 李楠楠, 沈一笋, 臧亮, 等. 对比人类基因 探秘材料基因: 人类基因组计划对材料基因组计划的启发[J]. 中国材料进展, 2016, 35(2): 156.
LI Nannan, SHEN Yisun, ZANG Liang, et al. Mater China, 2016, 35(2): 156.
[3] 顾秉林, 汪洪, 向勇, 等. 材料基因组工程——推动材料创新的引擎[J]. 科技导报, 2015, 33(10): 3.
Gu, Binglin, Wang, Hong, Xiang, Yong, et al. Sci Technol Rev, 2015, 33(10): 3.
[4] SERLES P, YEO J, HACHÉ M, et al. Ultrahigh specific strength by Bayesian optimization of carbon nanolattices[J]. Adv Mater, 2025, 37(14): e2410651.
[5] AJMAL S, ATHAR HASHMI F, IMRAN I. Recent progress in development and applications of biomaterials[J]. Mater Today Proc, 2022, 62: 385–391.
[6] QI Y J, LIU C, et al. Deep learning for medical materials: Review and perspective[J]. ES Mater Manuf, 2021, 12: 17–28.
[7] SUWARDI A, WANG F K, XUE K, et al. Machine learning-driven biomaterials evolution[J]. Adv Mater, 2022, 34(1): e2102703.
[8] LI C H, BAO L Q, JI Y X, et al. Combining machine learning and metal–organic frameworks research: Novel modeling, performance prediction, and materials discovery[J]. Coord Chem Rev, 2024, 514: 215888.
[9] Zhou Z H. Machine learning[M]. Singapore: Springer nature, 2021.
[10] Gurney K. An introduction to neural networks[M]. Boca Raton, FL: CRC press, 2018.
[11] 杨鑫鑫. 基于机器学习的低收缩超高性能混凝土配合比优化设计[D]. 吉林: 东北电力大学, 2024.
YANG Xinxin. Optimization design of low shrinkage ultra-high performance concrete mixture ratio based on machine learning[D]. Jilin: Northeast Dianli University, 2024.
[12] 刘国华, 陈斌, 汪树玉, 等. 基于人工神经网络和 Monte-Carlo 法的混凝土配合比优化设计研究[J]. 水力发电学报, 2003, 22(4): 45–53.
LIU Guohua, CHEN Bin, WANG Shuyu, et al. J Hydroelectr Eng, 2003, 22(4): 45–53.
[13] 陈晓东, 陈斌, 刘国华. 基于BP ANN-GA 混合型算法的混凝土配合比优化设计研究[J]. 水力发电学报, 2007, 26(5): 59–63.
CHEN Xiaodong, CHEN Bin, LIU Guohua. J Hydroelectr Eng, 2007, 26(5): 59–63.
[14] 陈云, 付前旺. 机器学习和元启发式算法优化混凝土配合比研究进展[J]. 复合材料学报, 2024, 41(11): 5824–5851.
CHEN Yun, FU Qianwang. Acta Mater Compos Sin, 2024, 41(11): 5824–5851.
[15] 刘晓, 王思迈, 卢磊, 等. 机器学习预测混凝土材料耐久性的研究进展[J]. 硅酸盐学报, 2023, 51(8): 2062–2073.
LIU Xiao, WANG Simai, LU Lei, et al. J Chin Ceram Soc, 2023, 51(8): 2062–2073.
[16] 许成文, 叶家元, 高霖, 等. 熟料等水泥基材料数据库及其应用[J]. 硅酸盐学报, 2025, 53(5): 1328–1338.
XU Chengwen, YE Jiayuan, GAO Lin, et al. J Chin Ceram Soc, 2025, 53(5): 1328–1338.
[17] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Adv Neur Inf Process Syst, 2017, 30.
[18] Rokach L, Maimon O. Clustering methods[M].Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US, 2005: 321–352.
[19] AHMED M, SERAJ R, ISLAM S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.
[20] RAN X C, XI Y, LU Y G, et al. Comprehensive survey on hierarchical clustering algorithms and the recent developments[J]. Artif Intell Rev, 2023, 56(8): 8219–8264.
[21] HEALY J, MCINNES L. Uniform manifold approximation and projection[J]. Nat Rev Meth Primers, 2024, 4: 82.
[22] RIGATTI S J. Random forest[J]. J Insur Med, 47(1): 31–39.
[23] Fidanova S. Ant colony optimization[M].Ant colony optimization and applications. Cham: Springer International Publishing, 2021: 3–8.
[24] BAILEY E T, CALDAS L. Operative generative design using non-dominated sorting genetic algorithm II (NSGA-II)[J]. Autom Constr, 2023, 155: 105026.
[25] GOFAIR.FAIRPrinciples[EB/OL].[2020–07–21].https://www.go-fair.org/fair-principles/.
[26] RIOS R, CHILDS C, SMITH S, et al. Advancing cement-based materials design through data science approaches[J]. RILEM Tech Lett, 2021, 6: 140–149.
[27] AYLAS-PAREDES B K, HAN T H, NEITHALATH A, et al. Data driven design of ultra high performance concrete prospects and application[J]. Sci Rep, 2025, 15(1): 9248.
[28] SHARMA S, KUMAR V. A comprehensive review on multi-objective optimization techniques: Past, present and future[J]. Arch Comput Meth Eng, 2022, 29(7): 5605–5633.
[29] GIAGKIOZIS I, FLEMING P J. Methods for multi-objective optimization: An analysis[J]. Inf Sci, 2015, 293: 338–350.
[30] HUA Y C, LIU Q Q, HAO K R, et al. A survey of evolutionary algorithms for multi-objective optimization problems with irregular Pareto fronts[J]. IEEE/CAA J Autom Sin, 2021, 8(2): 303–318.

AI-Powered Intelligent Design of Concrete Material Gene Database

LIU Chuyao[1], *LUO Zixuan*[1], *YANG Liu*[1], *HU Jinqi*[1], *LIU Zanqun*[2], *ZHANG Zengqi*[3], *YANG Wencui*[4], *XU Huining*[4], *CHEN Mengni*[1]
(1. School of Computer Science, Central South University, Changsha 410075, China; 2. School of Civil Engineering, Central South University, Changsha 410075, China; 3. School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, China; 4. School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China)

Extended Abstract

Introduction Concrete as the second-most consumed material globally after water plays a vital role in modern construction and

infrastructure development due to its superior mechanical properties, longevity, and versatility. Conventional methodologies for concrete design predominantly rely on empirical approaches, often resulting in extended cycles for material development and suboptimal performance outcomes in practical applications. The advent of the Materials Genome Initiative (MGI), launched in the United States in 2011, highlights a need for data-driven approaches to material design. Concurrently, significant advancements in Artificial Intelligence (AI) have new avenues for enhancing material discovery and characterization. It is thus necessary to establish a robust and comprehensive database capable of accommodating the extensive, heterogeneous, and multi-scale data associated with various concrete properties and performance metrics. This study was to develop an innovative AI-powered intelligent design framework specifically tailored for a concrete material performance gene database. This framework could promise efficient data management and intelligent analytical capabilities and facilitate advanced material development and optimization.

Methods This framework system integrated the information on concrete material composition, mix parameters, environmental factors, *etc.*, and extracted the topological structure and coupling relationships between gene features to establish a concrete material knowledge graph based on conventional relational databases, realizing the visualization of gene feature relationships. This work designed a "gene importance-driven hierarchical density clustering" algorithm for quantitative analysis of concrete material gene features, extracting gene structure with different action relationships, *via* addressing the complex influencing factors and multi-scale coupling characteristics of concrete material performance. On this basis, this work also designed a closed-loop reasoning path from feature selection, model prediction, and contribution quantification for typical concrete material performance to reduce the cycle of traditional test methods, lower experimental costs, and achieve rapid evaluation and iterative optimization of concrete performance. This could provide a scientific theoretical basis for material selection *via* the performance prediction model as a core model and combining with multi-objective optimization algorithms. The framework could give a systematic technical framework and theoretical support for intelligent material research *via* implementing the whole-process design from concrete material data management to intelligent development, promoting the deep integration and innovative development of material science research and civil engineering decision-making.

Results and Discussion The implementation of the AI-enabled framework demonstrates substantial enhancements in the accuracy and reliability of concrete performance predictions. The algorithms effectively uncover intricate and complex relationships between various material constituents, and their resultant influence on the critical performance indicators, such as compressive strength, durability, and workability. The unique integration of multi-source data within the database facilitates enhances feature extraction and visual representation of critical metrics, effectively addressing the intricate coupling effects that often complicate conventional testing methodologies. Through rigorous testing, validation, and benchmarking against existing standards, the framework exhibits exceptionally high accuracy and reliability in predicting concrete properties. This robust performance ultimately supports the creation of optimized mix designs, propelling advancements in concrete technology and application.

Conclusions This study designed the AI-powered Intelligent Design of Concrete Material Gene Database, providing a systematic and technically sound foundation for concrete material research, and significantly enhancing the integration of material science principles and informed engineering decision-making. This study established a high-quality concrete gene database that could facilitate rapid evaluations and predictions *via* effectively leveraging advanced AI technologies. Furthermore, this study supported innovative advancements in concrete material design, paving a way for more efficient, sustainable practices in construction. Future research endeavors could broaden the system’s capabilities, focusing on cross-disciplinary applications and the extension of the database to include low-carbon, environmentally-friendly material options.

Keywords concrete; intelligent database design; material genome; material performance; artificial intelligence