DOI:10.11918/202511031

基于主成分分析的神经网络用于新污染物环境影响预测

孙 叶[1,2],吴宝利[1,3],马 军[1,2],尤世界[1,2]

(1.哈尔滨工业大学 环境学院,哈尔滨 150090; 2.城乡水资源与水环境全国重点实验室(哈尔滨工业大学),哈尔滨 150090; 3.中国市政工程华北设计研究总院有限公司,天津 300381)

摘 要:新污染物具有结构多样性、环境持久性和生物毒性,对生态系统和人类健康构成潜在风险。生命周期评价(Life Cycle Assessment,LCA)是评估其环境影响的重要方法,但存在基础数据缺失、难以准确量化等问题。为此,提出基于主成分分析(Principal Component Analysis,PCA)优化输入特征的人工神经网络(Artificial Neural Network,ANN)模型,通过对大量分子描述符进行PCA降维,提取主要特征以减少数据冗余和维度负担,用于实现新污染物的环境影响预测。构建了4种不同特征保留比例的模型,即NOPCA(全部描述符)、PCA95(保留累计方差贡献率95%的主成分)、PCA85(保留累计方差贡献率85%的主成分)和PCA75(保留累计方差贡献率75%的主成分),评价并比较预测性能差异。结果表明:适度的特征提取可提升模型泛化能力,在全球变暖潜势(GWP)和人类毒性潜势(HTP)预测任务中,采用PCA85预处理描述符所构建的模型表现最佳,测试集R^2分别为0.64和0.73;在化石能源消耗潜势(FDP)和陆地酸化潜势(TAP)预测任务中,模型则在PCA95时达到最佳,测试集R^2分别为0.77和0.76。基于PCA特征提取的ANN模型可有效缓解LCA数据不全问题,为新污染物环境影响预测提供了高效、可靠的方法。

关键词:主成分分析;新污染物;人工神经网络;生命周期评价;分子描述符

中图分类号:X502 **文献标志码**:A **文章编号**:0367-6234(2026)06-0031-09

Neural network based on principal component analysis for predicting environmental impacts of emerging contaminants

SUN Ye[1,2], WU Baoli[1,3], MA Jun[1,2], YOU Shijie[1,2]

(1. School of Environment, Harbin Institute of Technology, Harbin 150090, China; 2. State Key Laboratory of Urban-rural Water Resource and Environment (Harbin Institute of Technology), Harbin 150090, China; 3. North China Municipal Engineering Design & Research Institute Co., Ltd., Tianjin 300381, China)

Abstract: Emerging contaminants have structural diversity, environmental persistence, and biological toxicity, which pose potential risks to ecosystems and human health. Life Cycle Assessment (LCA) is an important method for evaluating their environmental impacts, but it has problems such as a lack of basic data and difficulty in accurate quantification. To address these issues, this paper proposed an Artificial Neural Network (ANN) model based on Principal Component Analysis (PCA) to optimize input features. By performing PCA dimensionality reduction on a large number of molecular descriptors, main features were extracted to reduce data redundancy and dimensional burden for achieving the prediction of environmental impacts of emerging contaminants. Four models with different feature retention ratios were constructed, namely NOPCA (all descriptors), PCA95 (retaining principal components with 95% cumulative variance contribution rate), PCA85 (retaining principal components with 85% cumulative variance contribution rate), and PCA75 (retaining principal components with 75% cumulative variance contribution rate), and the differences in prediction performance were evaluated and compared. The results show that moderate feature extraction can improve the generalization ability of the model; in the prediction tasks of global warming potential (GWP) and human toxicity (HTP), the model constructed using PCA85-preprocessed descriptors performs the best, with test-set R^2 of 0.64 and 0.73, respectively; in the prediction tasks of fossil energy depletion (FDP) and terrestrial acidification (TAP), the model achieves optimal performance under PCA95, with test-set R^2 of 0.77 and 0.76, respectively. The ANN model based on PCA feature extraction can effectively mitigate the problem of incomplete LCA data and provides an efficient and reliable method for predicting the environmental impacts of emerging contaminants.

Keywords: Principal Component Analysis; emerging contaminant; Artificial Neural Network; life cycle assessment; molecular descriptor

收稿日期:2025-11-11;**录用日期**:2025-12-17;**网络首发日期**:2026-01-21
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20260120.1700.002
基金项目:国家自然科学基金(U25A20365,52400236,52325003)
作者简介:孙 叶(1991—),女,助理研究员;尤世界(1980—),男,教授,博士生导师
通信作者:尤世界,sjyou@hit.edu.cn

新污染物(emerging contaminants,ECs)是近年来环境科学与工程领域广泛关注的化学物质,主要包括药物、全氟化合物及其他难降解有机物等[1]。由于其结构复杂性和生物毒性,新污染物在环境介质中的广泛检出已对生态系统安全和人类健康构成潜在威胁[2]。科学评估其环境影响对新污染筛查、预防、评价和管控具有重要意义[3]。生命周期评价(Life Cycle Assessment,LCA)作为一种量化物质从原料获取、生产、使用到废弃全过程环境负荷的方法,已成为环境影响综合评价的重要手段[4-5]。然而,针对新污染物 LCA 研究面临数据匮乏的问题,尤其是在生命周期清单(LCI)阶段,缺乏基础实验数据与排放参数导致输入信息不足,从而影响评价结果的准确性与可比性[6]。

为解决 LCA 数据不全的问题,研究者提出了多种数据驱动方法,如线性回归、偏最小二乘回归(PLS)、随机森林(RF)及支持向量机(SVM)等[7-8]。其中,PLS 在处理多重共线性问题方面具有优势,SVM 在小样本和非线性问题中表现良好。然而,这类方法在面对高维分子描述符数据时仍存在一定局限性。PLS 主要通过线性潜变量提取实现降维,对复杂非线性关系的建模能力有限。SVM 对核函数选择与参数调优高度敏感,在高噪声、高维特征情境下模型稳定性与泛化能力易受影响[9-10]。人工神经网络(Artificial Neural Network,ANN)凭借其强大的非线性映射能力和自学习特征,在环境建模、污染物浓度预测和毒理效应评价等领域具有广泛应用。然而,高维输入特征易增加 ANN 模型过拟合风险,进而影响其在新污染物环境影响预测中的稳定性与泛化能力[11-13]。

主成分分析(Principal Component Analysis,PCA)可通过线性变换将原始高维数据空间转换为若干不相关的主成分,在保留主要信息的同时减少特征冗余和噪声干扰,是一种常用的特征降维方法[14]。PCA 在多变量分析和化学计量学等领域被广泛用于特征提取和降维处理。在分子描述符建模中,PCA 能够有效解决变量间的多重共线性问题,降低模型复杂度,提升计算效率,为后续机器学习模型的训练提供更具代表性的输入特征[15]。与 PLS 侧重于监督式潜变量提取不同,PCA 以无监督方式最大程度保留数据整体方差结构,避免了响应变量噪声对降维过程的干扰。因此,结合 PCA 与 ANN 建模可有效简化模型数据结构,提升预测稳定性与计算效率,为解决新污染物分子描述符高维特征与有限样本规模之间的失衡问题提供新思路和新方法。

针对新污染物环境影响预测数据不全与高维性问题,提出构建基于分子描述符的 ANN 模型,用于实现 LCA 清单数据预测,提升模型预测的准确性与可靠性。首先,利用 PCA 对新污染物分子描述符进行特征提取,构建不同特征保留比例(95%、85%、75%)的特征集;然后,基于各特征集分别训练 ANN 模型,并与未经 PCA 处理的模型(NOPCA)进行对比;最后,通过模型性能评估与特征重要性分析,探讨 PCA 在提升 ANN 预测稳定性方面的作用机制与应用潜力。

1 方 法

1.1 数据收集和预处理

以新污染物的生命周期评价(LCA)数据为研究对象,主要涉及全球变暖潜势(global warming potential,GWP)、人类毒性潜势(human toxicity potential,HTP)、化石能源消耗潜势(fossil depletion potential,FDP)和陆地酸化潜势(terrestrial acidification potential,TAP)4 个典型环境影响指标。数据来源于 LCA 商用数据库 Ecoinvent v3.8,收集了 205 组相对复杂的新污染物生命周期环境影响数据,其中,环境影响类型均为 ReCiPe Midpoint (H)。化学分子的结构信息通过 SMILES 字符串表示,利用 PaDEL-Descriptor 软件计算获得 1 444 个分子描述符[16],包括理化性质、电荷分布、分子片段等特征[17]。

鉴于新污染物生命周期清单(LCI)数据获取难度较大,本研究的样本规模相对有限,为防止模型过拟合,需对分子描述符数据进行预处理。利用 Z-score标准化法通过式(1)将数据进行标准化处理,即

$$Z=\frac{X-\mu}{\sigma} \tag{1}$$

式中:Z 为标准化特征,X 为标准化之前的原始特征,μ 为所有新污染物特征的平均值,σ 为特征标准差。

1.2 特征提取

为提升模型训练效率,采用主成分分析(PCA)对分子描述符进行降维处理[18]。PCA 通过正交变换将原始描述符转化为互不相关的主成分向量,在尽量保留总体方差信息的同时减少特征数量[19]。每个主成分的方差贡献率(explained variance ratio)反映其对总体信息的解释能力,前 k 个主成分的累计方差贡献率(cumulative explained variance ratio,R_{CEV})定义为

$$R_{\mathrm{CEV},k}=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{p}\lambda_j} \tag{2}$$

式中：λ_i 为第 i 个特征值，p 为原始特征总数。累计方差贡献率用于衡量前 k 个主成分保留的信息比例，当达到设定阈值时，可认为数据主要特征已得到充分保留。

对比 4 种包含不同比例分子描述符信息 PCA 处理的模型，分别为 NOPCA（不进行降维，使用全部描述符）、PCA95（保留累计方差贡献率 95% 的主成分）、PCA85（保留累计方差贡献率 85% 的主成分）和 PCA75（保留累计方差贡献率 75% 的主成分），处理后的数据集作为输入数据集进行模型训练。此外，为评估非线性降维方法在新污染物环境影响预测任务中的适用性，采用 t-SNE 与 UMAP 对分子描述符进行非线性降维，并将其作为模型输入数据进行对比。

1.3　模型构建

采用人工神经网络（ANN）建立分子描述符与生命周期评价（LCA）指标之间的非线性关系[20]。模型采用前馈全连接结构，包括一个输入层、若干隐藏层和一个输出层[21]。激活函数采用 ReLU，模型训练采用 Adam 优化算法，学习率设为 0.001，批量大小为 64，最大迭代数为 500。为防止过拟合，使用早停（Early Stopping）策略，当验证集损失在连续 20 轮内不再下降时终止训练。ANN 模型架构如图 1。

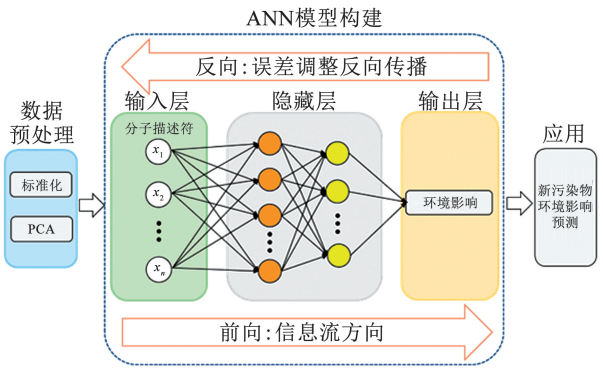


图 1　新污染物环境影响预测的 PCA-ANN 模型框架

Fig. 1　Framework of PCA-ANN model for predicting environmental impacts of emerging contaminants

模型在 Linux 系统下使用 Python 3.9，Google TensorFlow 框架开发。数据集按 9∶1 划分为训练集和测试集，训练样本用于参数学习，测试样本用于独立性能评估。训练阶段引入 10 折交叉验证（10-fold cross-validation），对训练数据进行多次划分与验证。各 PCA 特征集对应的 ANN 模型均采用相同的结构和训练参数，以确保不同特征降维处理下模型预测性能的可比性。为系统评估 PCA-ANN 模型的性能优势，引入随机森林（RF）、支持向量机（SVM）和多元线性回归（MLR）作为对比模型，所有模型均采用相同训练集与测试集划分方式。

1.4　模型评估

为定量评估不同 PCA 特征集下 ANN 模型的预测性能，选取决定系数（coefficient of determination，R^2）和均方根误差（root mean squared error，RMSE，E_{RMS}）作为评价指标。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_m)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$E_{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_m)^2}{n}} \tag{4}$$

式中：y_m 为样本真实值，y_i 为模型预测值，$\bar{y}$ 为样本均值，n 为样本数。

在本研究中，主要聚焦于网络结构参数（隐藏层数与神经元数）的系统优化，固定学习率、激活函数类型及权重初始化方式等其他超参数，以确保不同模型结果具有良好的可比性。

1.5　模型解释

为探究 PCA-ANN 模型的可解释性，采用 SHAP 方法对模型输入的主成分进行贡献度分析，量化各主成分对预测结果的影响。同时结合 PCA 的载荷矩阵，将重要主成分进一步映射回原始分子描述符空间，从而识别对环境影响预测贡献较大的关键特征。

为实现降维特征向原始分子描述符空间的可解释性映射，采用融合 SHAP 贡献度与 PCA 载荷矩阵的反向特征映射方法[22]。设 a_{jk} 表示第 j 个原始分子描述符在第 k 个主成分上的载荷系数，s_k 表示第 k 个主成分对应的 SHAP 重要性，K 为 PCA 降维后的特征数。首先，对主成分重要性进行归一化处理，得到权重系数，即

$$w_k = \frac{s_k}{\sum_{k=1}^{K} s_k} \tag{5}$$

随后，基于各描述符在关键主成分上的载荷强度，构建原始特征的综合重要性指标，即

$$I_j = \sum_{k=1}^{K} w_k \cdot a_{jk}^2 \tag{6}$$

式中：w_k 表征各关键主成分在模型中的相对贡献权重，a_{jk}^2 表示原始描述符在对应主成分中的结构贡献强度。通过对 I_j 进行排序，筛选出对模型预测贡献度较高的分子描述符，并进一步结合其结构属性进行类别归纳，从而实现由低维主成分空间向高维原始分子结构特征空间的定量映射与机理解释。

2 结果与讨论

2.1 主成分分析与特征提取

为减少分子描述符之间的多重共线性，提高 ANN 模型的计算效率与预测稳定性，对全部样本的描述符数据进行了 PCA 处理[18]。根据累计方差贡献，分别设置了 95%、85% 和 75% 3 个特征保留阈值，对应的模型分别记为 PCA95、PCA85 和 PCA75，同时保留未经 PCA 处理的 NOPCA 模型用于对照。4 种特征方案构建的 ANN 模型分别用于预测 GWP、HTP、FDP 和 TAP，其预测性能以 R^2（图 2）和 RMSE（表 1）进行比较。如图 2 所示，PCA 处理显著改善了模型的预测性能。NOPCA 模型在测试集上的 R^2 仅为约 0.52（GWP）、0.66（HTP）、0.63（FDP）和 0.63（TAP），表明原始高维特征中存在较强的噪声与冗余。经 PCA 降维优化后，模型性能均有提升，以 GWP 模型为例，PCA85 模型表现最优，测试集 R^2 提升到 0.64，较 NOPCA 模型提升了 23.08%。相比之下，PCA95 模型虽然保留了更多的方差信息，但由于部分冗余特征仍存在，模型复杂度较高，验证集性能略低于 PCA85。PCA75 模型因特征压缩过度丢失部分重要信息，预测精度有所下降。这表明当累计方差贡献率控制在 85% 左右时，模型在信息保留与泛化能力之间达到了最佳平衡。同样，对于 HTP、FDP 和 TAP 模型，分别在 PCA85、PCA95 和 PCA95 时性能最佳，测试集 R^2 分别为 0.73，0.77 和 0.76，这说明数据集的最佳特征保留阈值与数据本身密切相关，设置不同的保留阈值对模型结果有较大影响。

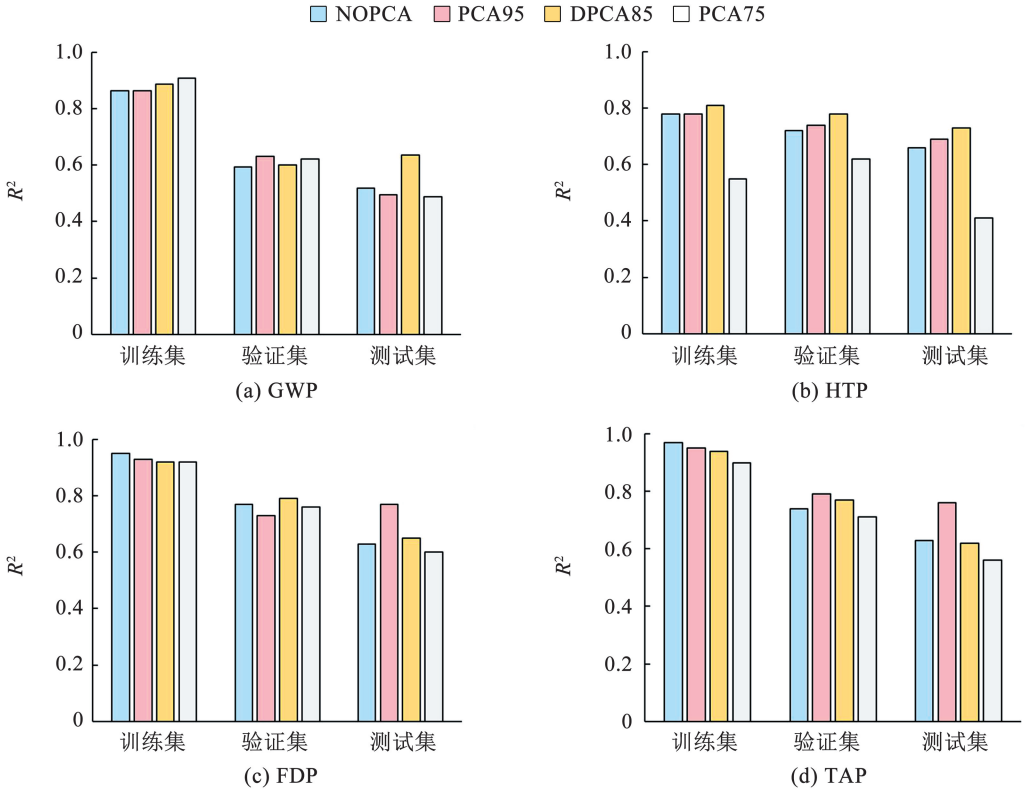


图 2 基于预处理后 NOPCA、PCA95、PCA85 和 PCA75 描述符构建的 ANN 模型性能（以 R^2 表示）

Fig. 2 Performance (R^2) of ANN models constructed based on preprocessed NOPCA, PCA95, PCA85, and PCA75 descriptors

表 1 基于 NOPCA、PCA95、PCA85 和 PCA75 数据降维处理后 ANN 模型的 RMSE

Tab. 1 RMSE of ANN models based on NOPCA, PCA95, PCA85, and PCA75 data dimensionality reduction

影响类别	RMSE			
	NOPCA	PCA95	PCA85	PCA75
GWP	1.329 2	1.384 0	1.213 1	1.374 5
HTP	1.487 4	1.440 8	1.418 8	1.559 7
FDP	1.190 0	0.815 1	0.995 5	1.288 0
TAP	0.007 4	0.006 8	0.007 3	0.007 8

以 GWP 模型为例，图 3 展示了 PCA85 处理后各主成分的信息贡献比例。结果显示，前 5 个主成分的累计方差贡献率已达 51.85%，前 10 个主成分可解释约 65.86% 的总体信息。这说明描述符数据中存在显著的相关性，主要信息集中在少数几个综合变量上。随着主成分序号增加，单个主成分的解释方差逐渐减小，当主成分数超过 28 时，新增主成分对总体信息的贡献率不足 0.60%，对模型性能的提升有限。因此，仅保留前 28 个主成分即可较完整

地描述化学分子的环境行为特征，实现有效降维。

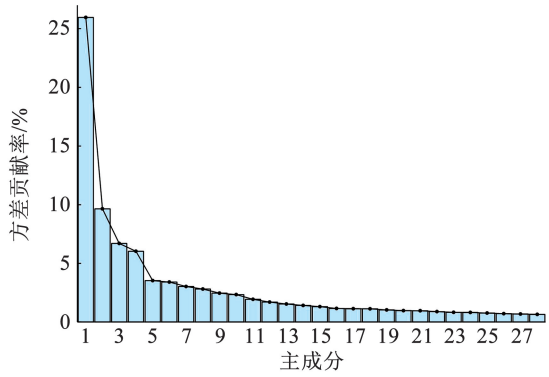


图 3　经过主成分分析（PCA85）预处理后 GWP ANN 模型中 28 个主成分的方差贡献率

Fig. 3　Variance contribution rate of 28 principal components in GWP ANN model after principal component analysis (PCA85) pretreatment

此外，为比较线性与非线性降维方法在新污染物环境影响预测任务中的适用性，采用 PCA、t-SNE 与 UMAP 将 1 444 维描述符统一降至 28 维，并在相同 ANN 结构与参数下进行性能对比。表 2 结果表明，与 PCA 方法相比，t-SNE 和 UMAP 在一定程度上能够捕捉特征空间中的非线性结构信息，但在本研究的小样本、高维分子描述符场景下，其预测性能并未优于 PCA 方法。因此，PCA 是本研究中更为稳健和实用的降维策略。

表 2　不同降维方法（PCA、t-SNE 和 UMAP）下模型在 GWP、HTP、FDP 和 TAP 上测试集 R^2 比较

Tab. 2　Comparison of test-set R^2 of models on GWP, HTP, FDP, and TAP under different dimensionality reduction methods (PCA, t-SNE, and UMAP)

模型	R^2			
	GWP	HTP	FDP	TAP
PCA + ANN	0.64	0.73	0.66	0.62
t-SNE + ANN	0.59	0.65	0.55	0.44
UMAP + ANN	0.61	0.69	0.63	0.60

2.2　模型建立优化

在完成特征提取后，为构建具有最优结构与性能的 ANN 模型，在 GWP 模型建立中，以 PCA85 预处理后的描述符作为输入变量，对模型结构参数进行了优化。本研究主要围绕网络结构参数（隐藏层数与神经元数）开展系统优化，并在此过程中固定学习率、激活函数类型等超参数，以保证不同模型之间具有良好的可比性。首先，为确定网络的最佳隐藏层数，设置隐藏层数为 1、2、3 和 4 的 4 种结构，保持每层神经元数一致（256 个），并在相同的学习率与批大小条件下进行训练。图 4（a）展示了不同隐藏层数下 GWP 模型的预测性能（以测试集 R^2 表示）。可以看出，随着隐藏层数的增加，模型的拟合能力在初期有所提升，训练集 R^2 由 1 层时的 0.89 提升到 4 层的 0.94。但当层数超过 1 后，测试集 R^2 没有提升反而下降，隐藏层数为 2 时，测试集 R^2 由一层时的 0.64 下降至 0.58，这表明过深的网络结构可能引入过拟合风险，增加训练复杂度与计算成本。因此，综合考虑模型精度与计算效率，单隐藏层结构为最优方案。

随后，为确定每层最优神经元数，在隐藏层固定为 1 层的条件下，将神经元数设置为 64、128、256、512 和 1 024 进行对比。结果如图 4（b）所示，随着神经元数的增加，模型的测试集 R^2 逐渐提升，当神经元数从 64 增加到 256，R^2 从 0.57 增加到 0.64；继续增加神经元数到 512 时，R^2 = 0.65，与 256 个神经元的模型性能没有明显提升。当神经元数增加到 1 024时，模型 R^2 反而降低至 0.62，这可能是由模型的过拟合导致。在一定范围内，增加模型的复杂度可以使训练的模型更加准确。但是，由于训练集数据数量的限制，增加模型的复杂度会导致过拟合的发生。因此，GWP 模型最终选取单隐藏层、256 个神经元作为 ANN 的最佳网络结构。

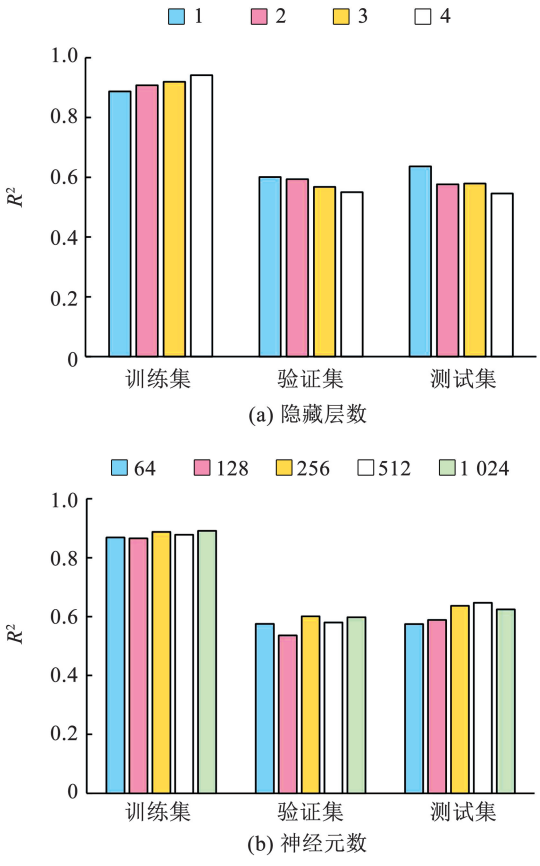


图 4　PCA85 描述符作为输入数据不同隐藏层数以及每层包含不同隐藏神经元数的 GWP 模型的预测性能（R^2）

Fig. 4　Prediction performance (R^2) of GWP models with different numbers of hidden layers and different numbers of hidden neurons in each layer using PCA85 descriptors as input data

最后,为确保模型在训练过程中达到收敛并避免过拟合,考察了不同迭代次数(epochs)对模型误差的影响。图 5 显示了 GWP 模型在不同迭代次数下的相对误差变化。可以看出,模型误差在初始阶段(前 50 次迭代)快速下降,模型的训练集和测试集的相对误差均迅速下降到 5% 以下,在约第 400 次迭代后趋于稳定,继续训练对性能提升作用有限。因此,采用早停策略,当验证集误差连续 20 轮无明显下降时停止训练,以防止过拟合[23]。

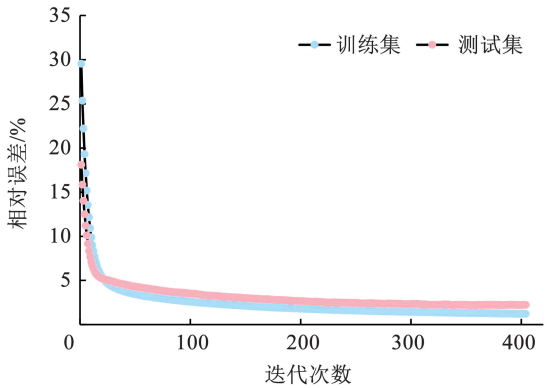


图 5　不同迭代次数下 GWP 模型的相对误差

Fig. 5　Relative error of GWP model under different numbers of iterations

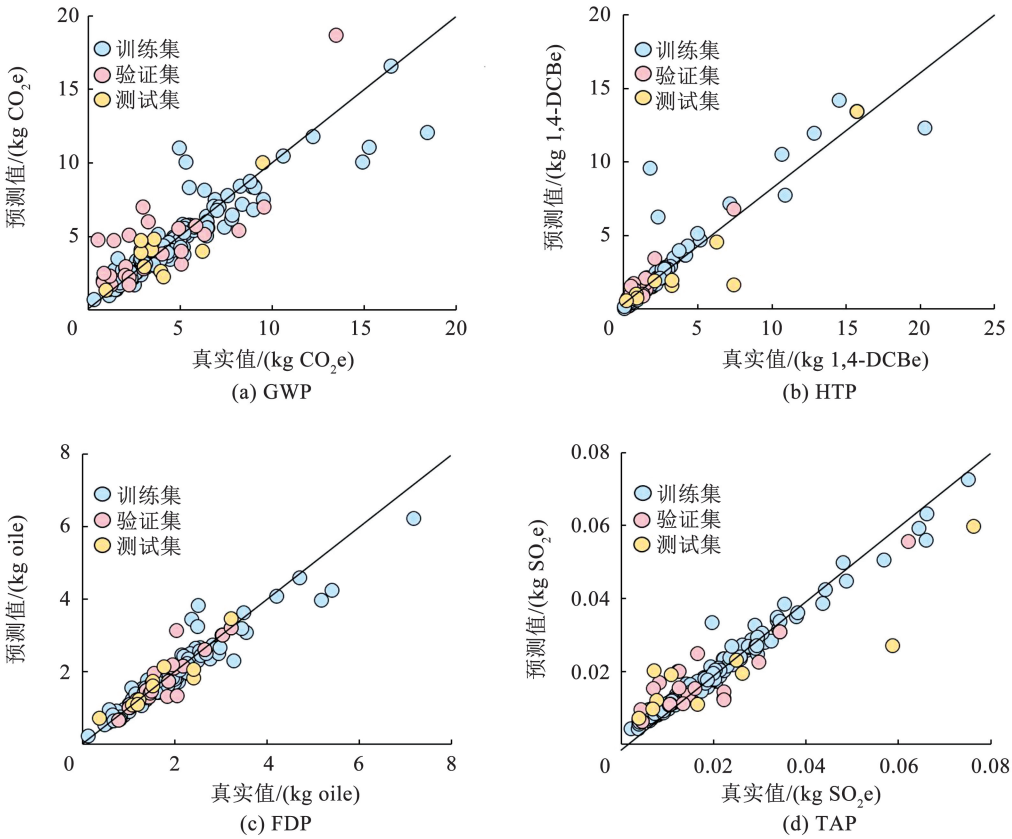
为进一步评估模型在不同环境影响类别中的适用性,分别构建并训练了针对 HTP、FDP 和 TAP 的 ANN 模型,各模型的最优结构参数见表 3。经 PCA 特征提取与参数优化后,HTP、FDP 和 TAP 模型均有较高的预测精度,测试集 R^2 分别为 0.73、0.77 和 0.76,性能与此前文献中报道的同类 ANN 模型相当[14],说明本研究所提出的建模策略具有较好的普适性与稳定性。

表 3　GWP、HTP、FDP 和 TAP 影响类别 ANN 模型的最优参数

Tab. 3　Optimal parameters of ANN models for impact categories of GWP, HTP, FDP, and TAP

影响类别	PCA 处理	隐藏层数	神经元数
GWP	PCA85	1	256
HTP	PCA85	1	512
FDP	PCA95	1	256
TAP	PCA95	1	256

图 6 展示了 GWP、HTP、FDP 和 TAP 4 个影响类别模型预测值与清单数据的对比。各模型的预测值均较为集中地分布在理想预测线附近,说明模型能够较好地捕捉输入描述符与环境影响指标之间的非



注:输入数据包括训练数据、验证数据和测试数据。

图 6　模型的清单数据和模型预测值

Fig. 6　Inventory data and predicted values of models

线性关系。训练集与验证集样本的拟合程度较高，表明模型在训练过程中收敛较好。测试集样本的预测结果同样表现出较好的线性相关性，进一步验证了模型的泛化能力。其中，HTP、FDP 和 TAP 模型的预测点分布最为集中，测试集 R^2 分别达到 0.73、0.77 和 0.76，说明其预测性能稳定且对未见样本具有较强的适应性。相比之下，GWP 模型的预测点略有分散，部分高值样本存在一定偏差，可能与 GWP 指标受能源利用方式、排放途径等多因素耦合影响有关，导致其特征空间复杂度较高[24-25]。结果表明，基于 PCA 优化的 ANN 模型能够实现对新污染物生命周期评价关键环境影响指标的可靠预测，为生命周期数据库中缺失数据的智能补全提供了可行的方法。需要指出的是，受限于数据库 LCI 数据的获取难度，本研究所采用的样本规模相对有限。尽管样本量有限可能在一定程度上影响复杂模型的稳定性，但交叉验证结果显示，模型在不同数据子集上的预测性能较为一致，表明该方法在当前数据条件下具有较好的稳健性。未来研究可通过扩大样本规模和整合多源数据，进一步提升模型的稳定性和泛化能力。

为评估 PCA-ANN 模型在新污染物环境影响预测中的优势，在最优 PCA 特征集条件下，构建了 MLR、SVM 和 RF 对比模型，对 4 类环境影响指标进行了比较，结果如表 4 所示。ANN 模型在 4 项指标上均表现出最佳拟合能力，其测试集 R^2 分别达到 0.64、0.73、0.77 和 0.76，整体优于传统 MLR 模型及经典机器学习模型（SVM、RF）。结果表明，高维分子描述符经 PCA 降维后，ANN 模型在捕捉环境效应关系方面具有更强的建模能力，进一步验证了 PCA-ANN 模型在新污染物环境影响预测任务中的有效性。

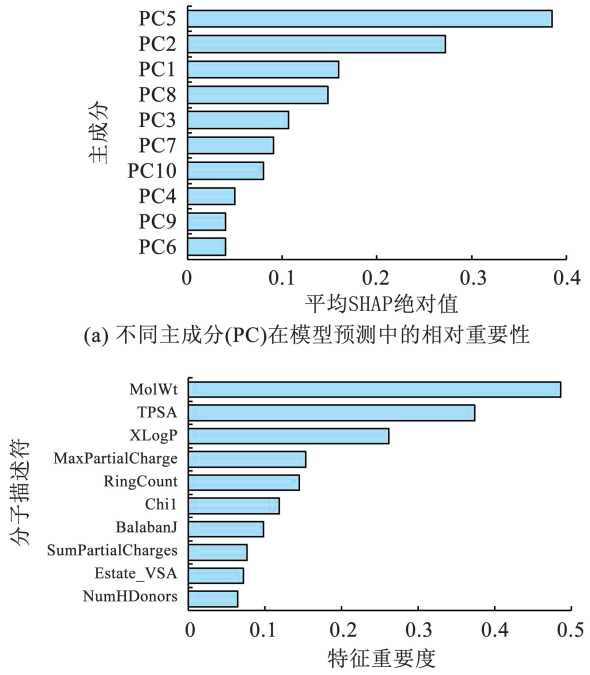
表 4　不同模型在 GWP、HTP、FDP 和 TAP 上测试集决定系数（R^2）比较

Tab. 4　Comparison of test-set coefficient of determination (R^2) of different models on GWP, HTP, FDP, and TAP

模型算法	R^2			
	GWP	HTP	FDP	TAP
MLR	0.48	0.55	0.51	0.49
SVM	0.59	0.65	0.59	0.64
RF	0.61	0.69	0.71	0.70
ANN	0.64	0.73	0.77	0.76

2.3　特征贡献分析

以 GWP 为例，SHAP 分析结果表明，不同主成分对模型预测的贡献存在显著差异（图 7(a)）。其中，PC5 的平均 SHAP 绝对值最高（0.385），表明其在模型决策过程中具有最强的解释能力；其次为 PC2（0.272）和 PC1（0.160），说明上述主成分在环境影响指标的预测中发挥了主导作用。相比之下，PC9、PC6 和 PC4 等主成分的 SHAP 贡献值较低，表明其对模型预测结果的贡献相对有限。同时，基于分子描述符特征重要度排名结果可知（图 7(b)），高贡献特征主要集中于分子极性参数（TPSA、XLogP、NumHDonors）、电荷分布指标（MaxPartialCharge、SumPartialCharges、EState_VSA）及拓扑结构相关描述符（MolWt、RingCount、Chi1、BalabanJ）。结果表明，PCA 降维后的特征空间中，各个主成分对模型预测贡献不同，这进一步验证了 PCA 在去除冗余信息和噪声特征方面的有效性，同时说明 ANN 模型能够基于关键主成分建立稳定的非线性关系。结合上述反向映射分析可知，这些高贡献主成分主要对应分子极性、电荷分布及拓扑结构等关键化学特征，进一步增强了模型预测结果的化学合理性与可解释性。



(a) 不同主成分(PC)在模型预测中的相对重要性

(b) 分子描述符在模型中的特征重要性排序

图 7　PCA-ANN 模型中主成分的 SHAP 重要性分布

Fig. 7　SHAP importance distribution of principal components in PCA-ANN model

2.4　新污染物环境影响预测验证

为验证基于 PCA-ANN 模型在 LCA 清单数据预测中的适用性，选取了 6 种具有代表性的新污染物（双酚 A、4-硝基苯酚、2,4-二氯苯酚、全氟化戊烷、六氟乙烷和十氟联苯醚）对 GWP 进行预测。上述新污染物涵盖了酚类、含氯芳香族化合物及含氟持久性有机物等多种结构类型，可用于评估模型在不

同分子复杂度与环境行为下的预测可靠性。如图 8 所示,模型预测结果与生命周期清单数据整体吻合良好。对于双酚 A,模型预测值为 4.42 kg CO_2e,相比真实值(4.12 kg CO_2e)相对误差为 7.41%;对于 4-硝基苯酚,模型预测值为 3.89 kg CO_2e,相比真实值(4.07 kg CO_2e)相对误差仅为 4.44%,这表明 ANN 模型能够准确识别芳香环及取代基对碳足迹的影响规律。经过 PCA 降维后的主成分仍充分保留了关键的结构信息,可有效支持模型的环境影响表征能力。相比之下,全氟化戊烷、六氟乙烷和十氟联苯醚等含氟化合物的预测值略高于清单数据,误差分别为 7.92%、10.37% 和 8.14%,其偏差可能源于含氟化合物在实际排放与分解过程中的高稳定性及生命周期数据库中该类物质数据的不确定性。以上结果证明了 PCA-ANN 模型在复杂新污染物环境影响预测中的可靠性与泛化性,为缺乏 LCA 清单数据的新污染物提供了一种可行的环境影响快速评估途径。

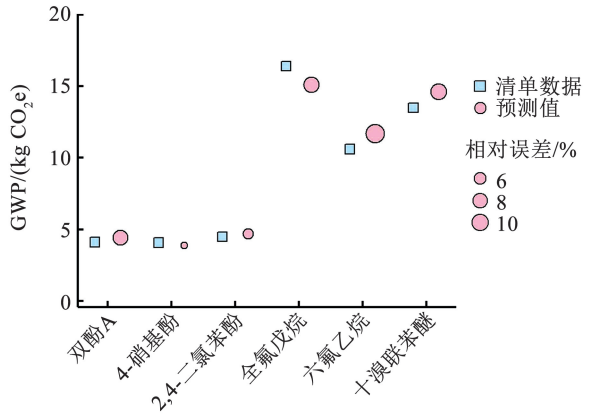


图 8 新污染物 GWP 模型预测结果与生命周期清单数据对比

Fig. 8 Comparison between predicted results of GWP model and life cycle inventory data for emerging contaminants

3 结 论

本研究构建了基于主成分分析(PCA)的人工神经网络(ANN)模型(PCA-ANN),用于实现新污染物环境影响的智能预测。主要结论如下:

1)通过 PCA 对分子描述符进行特征提取,显著降低了输入变量维度,改善了模型的训练效率与稳定性。以包含 85% 特征信息的 PCA85 方案为输入数据时,全球变暖潜势(GWP)和人类毒性潜势(HTP)ANN 模型在测试集上 R^2 分别达到 0.64 和 0.73,表明选择合理的 PCA 降维阈值有助于提升模型性能。

2)在化石能源消耗潜势(FDP)和陆地酸化潜势(TAP)预测任务中,模型在 PCA95 时达到最佳,测试集 R^2 分别为 0.77 和 0.76。本研究与现有研究结果相当,进一步说明合理设定降维阈值可在多类环境指标预测中取得相对较优的表现。

3)以双酚 A、4-硝基苯酚、2,4-二氯苯酚、全氟化戊烷、六氟乙烷和十氟联苯醚等新污染物为例,PCA-ANN 模型预测的 GWP 值与生命周期清单数据基本一致,预测误差大部分低于 10%,表明模型具有较好的泛化能力,有应用于实际生命周期数据库完善与构建的潜在价值。

综上,基于 PCA 优化的 ANN 模型可有效实现新污染物 LCA 清单数据预测,为无清单数据或新污染物的环境影响预测提供了一种可行方法。尽管样本规模相对有限可能对模型泛化能力产生一定影响,但交叉验证结果表明该模型在当前数据条件下仍具有较好的稳定性与适用性。未来随着相关数据的不断积累与完善,模型将持续迭代更新,并有望进一步融合非线性降维与注意力机制等先进方法,以提升对复杂分子体系的特征识别能力和环境影响预测精度。

参考文献

[1]时巧翠, 傅佳宇, 陈金媛, 等. 环境新污染物风险防范与化学品环境管理[J]. 化学试剂, 2022, 44(9): 1342
SHI Qiaocui, FU Jiayu, CHEN Jinyuan, et al. Risk prevention of emerging pollutants and environmental management of chemicals[J]. Chemical Reagents, 2022, 44(9): 1342. DOI: 10.13822/j.cnki.hxsj.2022.0333

[2]李芊, 张建, 陈信含, 等. 新污染物水生态环境风险及其防控形势[J]. 环境工程, 2024, 19: 107
LI Qian, ZHANG Jian, CHEN Xinhan, et al. Ecological risks of emerging contaminants in aquatic environments and their control strategies[J]. Environmental Engineering, 2024, 19: 107. DOI: 10.20087/j.cnki.1672-8114.2024.19.034

[3]胡洪营, 夏军, 陈卓, 等. 新污染物的水环境伦理问题及对策[J]. 环境工程, 2025, 43(1): 1
HU Hongying, XIA Jun, CHEN Zhuo, et al. Ethical issues and countermeasures regarding emerging contaminants in water environment[J]. Environmental Engineering, 2025, 43(1): 1. DOI: 10.13205/j.hjgc.202501001

[4]HOU Ping, CAI Jiarui, QU Shen, et al. Estimating missing unit process data in life cycle assessment using a similarity-based approach[J]. Environmental Science & Technology, 2018, 52(9): 5259. DOI: 10.1021/acs.est.7b05366

[5]ROMEIKO X X, LEE E K, SORUNMU Y, et al. Spatially and temporally explicit life cycle environmental impacts of soybean production in the U.S. Midwest[J]. Environmental Science & Technology, 2020, 54(8): 4758. DOI: 10.1021/acs.est.9b06874

[6]ZHU Xinzhe, HO Chihung, WANG Xiaonan. Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes[J]. ACS Sustainable Chemistry & Engineering, 2020, 8(30): 11141. DOI: 10.1021/acssuschemeng.

0c02211

[7] WERNET G, PAPADOKONSTANTAKIS S, HELLWEG S, et al. Bridging data gaps in environmental assessments: modeling impacts of fine and basic chemical production[J]. Green Chemistry, 2009, 11(11): 1826. DOI: 10.1039/b905558d

[8] LIU Xian, LU Dawei, ZHANG Aiqian, et al. Data-driven machine learning in environmental pollution: gains and problems [J]. Environmental Science & Technology, 2022, 56(4): 2124. DOI: 10.1021/acs.est.1c06157

[9] DOU Bozheng, ZHU Zailiang, MERKURJEV E, et al. Machine learning methods for small data challenges in molecular science[J]. Chemical Reviews, 2023, 123(13): 8736. DOI: 10.1021/acs.chemrev.3c00189

[10] ZHU Tengyi, CHEN Wenxuan, GU Yuanyuan, et al. Polyethylene-water partition coefficients for polychlorinated biphenyls: application of QSPR predictions models with experimental validation[J]. Water Research, 2021, 207: 117799. DOI: 10.1016/j.watres.2021.117799

[11] ALTMANN A, TOLOSI L, SANDER O, et al. Permutation importance: a corrected feature importance measure [J]. Bioinformatics, 2010, 26(10): 1340. DOI: 10.1093/bioinformatics/btq134

[12] WU Xindong, ZHU Xingquan, WU Gongqing, et al. Data mining with big data [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97. DOI: 10.1109/TKDE.2013.109

[13] CAI Jie, LUO Jiawei, WANG Shulin, et al. Feature selection in machine learning: a new perspective[J]. Neurocomputing, 2018, 300: 70. DOI: 10.1016/j.neucom.2017.11.077

[14] SONG Runsheng, KELLER A, SUH Sangwon. Rapid life-cycle impact screening using artificial neural networks[J]. Environmental Science & Technology, 2017, 51(18): 10777. DOI: 10.1021/acs.est.7b02862

[15] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014, 40(1): 16. DOI: 10.1016/j.compeleceng.2013.11.024

[16] YAP C W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints[J]. Journal of Computational Chemistry, 2011, 32(7): 1466. DOI: 10.1002/jcc.21707

[17] ZHANG Kai, ZHANG Huichun. Predicting solute descriptors for organic chemicals by a deep neural network (DNN) using basic chemical structures and a surrogate metric [J]. Environmental Science & Technology, 2022, 56(3): 2054. DOI: 10.1021/acs.est.1c05398

[18] 吕树彬, 杨婉琪, 李福生. 基于主成分分析和宽度学习系统的土壤铅镉重金属元素定量分析[J]. 光谱学与光谱分析, 2024, 44: 1852
LV Shubin, YANG Wanqi, LI Fusheng. Quantitative analysis of lead and cadmium heavy metal elements in soil based on principal component analysis and broad learning system[J]. Spectroscopy and Spectral Analysis, 2024, 44: 1852. DOI: 10.3964/j.issn.1000-0593(2024)07-1852-06

[19] 张雨辰, 赵军龙, 孙婧, 等. 主成分分析在低孔低渗储层孔隙度预测中的应用研究[J]. 河北地质大学学报, 2025, 48: 21
ZHANG Yuchen, ZHAO Junlong, SUN Jing, et al. Application of principal component analysis in porosity prediction of low porosity and low permeability reservoir [J]. Journal of Hebei Geo University, 2025, 48: 21. DOI: 10.13937/j.cnki.hbdzdxxb.2025.02.003

[20] KHOSHNEVISAN B, RAFIEE S, OMID M, et al. Prognostication of environmental indices in potato production using artificial neural networks [J]. Journal of Cleaner Production, 2013, 52: 402. DOI: 10.1016/j.jclepro.2013.03.028

[21] 周云鹏. 基于人工神经网络的二次供水系统余氯时间模型[J]. 净水技术, 2025, 44: 189
ZHOU Yunpeng. Time model of residual chlorine for secondary water supply system based on artificial neural network[J]. Water Purification Technology, 2025, 44: 189. DOI: 10.15890/j.cnki.jsjs.2025.08.024

[22] MYERS T, SONG Sejin, CHEN Yang, et al. Chronological age estimation from human microbiomes with Transformer-based robust principal component analysis[J]. Communications Biology, 2025, 8: 1159. DOI: 10.1038/s42003-025-08590-y

[23] ZHONG Shifa, ZHANG Kai, BAGHERI M, et al. Machine learning: new ideas and tools in environmental science and engineering [J]. Environmental Science & Technology, 2021, 55(19): 12741. DOI: 10.1021/acs.est.1c01339

[24] ZHAO Bu, SHUAI Chenyang, HOU Ping, et al. Estimation of unit process data for life cycle assessment using a decision tree-based approach [J]. Environmental Science & Technology, 2021, 55(12): 8439. DOI: 10.1021/acs.est.0c07484

[25] 杨科, 刘丽香, 韩永伟, 等. 不同水生植物对富营养化水体净化和温室气体排放的影响[J]. 环境污染与防治, 2024, 46: 618
YANG Ke, LIU Lixiang, HAN Yongwei, et al. Effects of different aquatic plants on eutrophic water purification and greenhouse gas emission[J]. Environmental Pollution and Control, 2024, 46: 618. DOI: 10.15985/j.cnki.1001-3865.2024.05.002

(编辑　刘　彤)