ML-QSPR 方法预测煤基液体的燃料性能

李文英[1,2,3]，王香玲[1]，范欢欢[1]，范鸿霞[1]，冯　杰[1,2,3]

(1. 太原理工大学 省部共建煤基能源清洁高效利用国家重点实验室, 山西 太原　030024; 2. 北京怀柔实验室, 北京　101400; 3. 怀柔实验室山西研究院, 山西 太原　030032)

摘　要: 煤基液体混合物如煤焦油、煤直接液化油的分子结构描述和性质预测是开发煤基液体产品高值化工艺和技术的重要基础。由于煤基液体主要由 C、H、O、N、S 元素构成数量庞杂、芳环结构各异的混合物，因此，使用 Python 中的 RDKit 工具包，利用简化分子线性输入规范 (Simplified Molecular Input Line Entry System，SMILES) 语言构建煤基液体中物质分子描述符，描述符包含样品元素信息、环数与环结构信息、原子数及分子量信息等共计 115 个分子描述符。对比人工信息提取方法，将所构建的分子描述符能够体现煤基液体分子结构碎片、分子量及原子个数信息等作为机器学习的特征输入变量，用于建立预测煤基液体的燃料性能的分子机器学习−定量结构性质关系方法 (ML-QSPR)，实现对燃料低位热值 (LHV)、液体密度 (ρ)、闪点 (FP)、十六烷值 (CN)4 个关键燃料性能参数的快速预测。模型验证分析表明 LHV、ρ、FP 模型的 R^2 分别为 0.996、0.988、0.987；CN 预测中加入混合物数据进行预测，R^2=0.959。与已公开报道的预测 LHV、ρ、FP、CN 性质方法对比，笔者提出 ML-QSPR 方法在预测 4 个关键燃料性能参数准确度方面有提升，在获取结果速度方面有显著优势。利用 ML-QSPR 模型预测得到的煤基液体制特种燃料性能参数数据库中的信息，分析增加不同族组分物质的碳原子数量时 4 个燃料性能参数的演变趋势，发现 LHV、ρ、FP、CN 四个燃料性能参数均受碳数 (n) 影响显著。由于 LHV 主要由 n 决定，不同族组分物质的 LHV 差距小；而不同族组分物质的 ρ、FP 和 CN 性质差距明显。此外，本研究训练好的模型可用于预测新的分子，为新型燃料分子设计提供参考；ML-QSPR 方法作为迁移学习模型可在今后用于煤基液体其他场景相关理化性质的分析。

关键词: 煤焦油; 煤直接液化油; 煤结构; 煤组成成分; 分子描述符

中图分类号: TQ53; V31; O21　　**文献标志码:** A　　**文章编号:** 0253−9993(2024)02−1098−13

Predicting the fuel performance of coal-based liquids using the ML-QSPR method

LI Wenying[1,2,3], WANG Xiangling[1], FAN Huanhuan[1], FAN Hongxia[1], FENG Jie[1,2,3]

(1. *State Key Laboratory of Clean and Efficient Coal Utilization*, *Taiyuan University of Technology*, *Taiyuan*　030024, *China*; 2. *Beijing Huairou Laboratory*, *Beijing*　101400, *China*; 3. *Shanxi Research Institute of Huairou Laboratory*, *Taiyuan*　030032, *China*)

Abstract: A comprehensive understanding of the composition and physicochemical properties of coal-based liquids, such as coal tar or coal direct liquefaction oil, is conducive to the rapid development of multi-purpose, high-performance and high-value-added products and the efficient use of oil properties. A full understanding of the composition of ideal components in the coal-based liquid mixtures and their physical and chemical properties is also the key to designing liquid fuels with some special properties. The authors use the RDKit toolkit in Python, a method based on the Simplified Molecular In-

收稿日期: 2023−12−17　**修回日期:** 2024−01−27　**责任编辑:** 钱小静　**DOI:** 10.13225/j.cnki.jccs.2023.1701

基金项目: 中国神华煤制油化工有限公司科技创新资助项目 (MZYHG-22-02)

作者简介: 李文英 (1968—), 女, 山西大同人, 教授。 Tel: 0351-6018453, E-mail: ying@tyut.edu.cn

引用格式: 李文英, 王香玲, 范欢欢, 等. ML-QSPR 方法预测煤基液体的燃料性能[J]. 煤炭学报, 2024, 49(2): 1098−1110.

LI Wenying, WANG Xiangling, FAN Huanhuan, et al. Predicting the fuel performance of coal-based liquids using the ML-QSPR method[J]. Journal of China Coal Society, 2024, 49(2): 1098−1110.

移动阅读

put Specification for Molecules (SMILES) language, to construct the molecular descriptors suitable for substances in the coal-based liquids. The constructed molecular descriptors are able to extract the required structural fragments for the molecules in the coal-based liquids, which are mainly composed of the elements C, H, O, N, and S and contain a large number of substances with polycyclic aromatic structures, so the constructed structural fragment descriptors are mainly considered from the perspective of the elemental and ring numbers of the polycyclic aromatic compounds. At the same time, the number of atoms and the molecular weight descriptors are added to the structural fragment descriptors, which the number of molecular descriptors is 115 in total. Compared with the traditional manual information extraction methods, the constructed molecular descriptors can quickly extract the information contained in a large number of molecules in the coal-based liquids. The structural fragments, molecular weights and atomic numbers of the coal-based liquid molecules obtained by the constructed molecular descriptors are used as input features in Machine Learning (ML) to establish a method of predicting the quantitative molecular structure-property relationship (ML-QSPR) of the coal-based liquids, which achieves the fast and accurate prediction of four properties, namely, the lower heating value (LHV), the density of the liquid (ρ), the flash point (FP) and the cetane number (CN). The model validation analysis shows that the model R^2 of LHV, ρ, and FP are 0.996, 0.988, and 0.987, respectively. The CN prediction is predicted by adding mixtures, and the R^2=0.959. The ML-QSPR method has been improved in terms of prediction accuracy compared to the methods in the literatures and has a significant advantage over the traditional experimental methods in terms of the speed of obtaining properties. Using the information in the property database obtained from the ML-QSPR predictions, the evolution of four combustion performance parameters of different groups of substances is investigated when the number of carbon atoms is increased, and all four properties are significantly affected by the carbon number (n). Comparison of the individual properties of substances of different families shows that the difference in the LHV of substances of different families is small, and the size of LHV is mainly determined by n. For ρ, FP and CN, the difference in the properties of substances of different families is obvious. The trained model can be used to predict new molecules for new fuel design. The ML-QSPR method is expected to be used as a transfer learning model for the property analysis of different coal-based liquids when being applied in other application scenarios.

Key words: coal tar; liquids from direct coal liquefaction; coal structure; coal composition; molecular descriptors

煤焦油、煤直接液化油等煤基液体是进一步深加工成高性能产品的优质原料。分析煤基液体的组成结构与性质有助于快速开发多用途产品、充分发挥油品特性, 开发适合的工艺。充分了解具有理想组分构成及其理化特性的煤基液体混合物, 能筛选出有前景的、具有理想燃料特性的专用燃料, 这也是燃料设计的关键步骤[1-5]。一些燃料性质的实验测定不仅耗时长而且费用高, 此外, 如测定十六烷值 (CN) 还需大量的测试燃料, 因此通过测试每一种燃料来进行燃料筛选不切实际[6], 迫切需要一种高效、精确的性质预测工具来快速估计大量燃料化合物的特性, 从而以低成本方式筛选燃料。

定量结构性质关系 (QSPR) 是由 HANSCH 等[7]于 1956 年开发的线性回归模型, 探索分子化学结构特征 (描述符) 与性质之间变化规律的影响[8], 它允许根据物质分子的化学结构预测其性质。机器学习 (ML) 能从已有的结构和性质数据中获取 2 者之间的规律, 建立 2 者间的联系, 快速预测出性质, 这极大地节省人力物力和成本, 因此, 本研究拟使用 ML 这一工具, 构建煤基液体的分子机器学习−定量结构性质关系 (ML-QSPR) 模型, 研究燃料的两大类理化特性——能量密度中的低位热值 (LHV)、液体密度 (ρ) 及可燃性中的闪点 (FP)、十六烷值 (CN)。构建分子描述符获取 ML 输入特征, 搭建 ML-QSPR 模型, 实现对 LHV、ρ、FP、CN 这 4 个燃料性能参数快速准确的预测。通过数学统计方法对获取的最佳模型进行检验, 成功搭建了燃料 4 个关键燃料性能参数的预测模型。

1　建模方法

煤基液体组成复杂, 将化合物直接作为 ML 输入特征不合理。ML 输入一般是固定大小, 这要求不同分子的向量大小要一致, 提取出的特征能够被计算机识别且便于分析运算, 因此从单一化合物入手, 构建单一化合物结构指纹, 以作为 ML 的输入特征。为验证所构建的分子描述符是否正确, 需要采用构建好的分子描述符作为 ML 输入特征, 进一步搭建 ML 模型进行相关性质预测。

用构建好的煤基液体通用描述符, 计算得到物质

中分子描述符特征向量,同时获取相应的 LHV、ρ、FP、CN 性质数据,用 ML 算法来映射结构特征和性质。

1.1 数据库

实验数据库的质量是任何典型数学模型的基石[9-11],文中 LHV、ρ、FP 和 CN 四个燃料性能参数的数据来源见文献[5, 12-22],其中 CN 数据包含二元、三元混合物。

煤基液体包括煤直接液化油、煤焦油及后续经过加工得到的石脑油、白油、柴油、航空煤油等。文献[23]对煤直接液化产物石脑油和粗汽油中含氧组分进行分析,主要有酚类、醇类和酮类物质,还定性检测出呋喃和羧酸;文献[24]表明煤直接液化油中存在含 N−,S−芳香族化合物。

课题组前期工作[25]通过 GC-MS 定性定量 77% 煤液化油轻质组分,它包含脂肪烃类与芳香烃类及含 O/N/S−芳香族化合物;文献[26]使用 GC×GC-MS 分析煤焦油,含有脂肪烃,其中烷烃包括脂肪烃、不饱和脂肪烃、环烷烃,芳香烃含有单环芳烃、双环芳烃、三环芳烃、四环芳烃,同时检测到含氧类化合物,包括酚类、酮类、醚类、脂类、醇类、醛类、呋喃等类物质。

结合文献报道与课题组前期对煤基液体结构组成分析,选用脂肪烃、芳香烃、含 N 化合物与 S 化合物作为煤基液体数据库中的数据。

数据库中存在无效数据需要对其进行清洗。首先,查找数据库中是否存在重复数据,对重复数据进行清洗。其次,对数据库中异常值 (数值为 0 或缺失) 进行筛选清除,最后得到所需数据库。

1.2 构建分子描述符

构建煤基液体中物质结构−性质预测的 ML-QSPR 模型,首先要对单分子信息进行解码,构建煤基液体通用的分子描述符,再将分子描述符数字化,成为 ML 能够识别学习的数据。以 RDKit 开源工具包中碎片匹配功能为基础,基于分子的 SMILES[27]语言编程构建一套对煤基液体中化合物适用的分子描述符。

1.2.1 结构分子描述符

煤基液体中的物质大都由 C、H、O、N、S 元素构成,SMILES 中一般省略 H,因而只构建与 C、O、N、S 有关的结构碎片分子描述符。此外,煤基液体中显著特征是包含较多稠环芳烃化合物,由此,单独构建与环有关的结构碎片描述符很有必要。元素与环的结构碎片信息具体如下:

(1) C、O、N、S 结构分子描述符。

煤基液体中各物质的结构由不同元素构成,其包含的原子个数和种类不同,物质性质必定存在差异。C、O 元素结构分子描述符的构建从 2 个方面进行考虑:一是结构的位置。结构是位于链上还是位于脂肪环上,若存在于环上,进一步分析环是脂肪环还是芳香环;二是该结构相接的结构类型及数量。相接的结构是链、脂肪环、芳香环中的哪一种或哪几种。

由于煤基液体中含 N、S 杂原子物质少,而 N、S 元素能构成的分子结构碎片多,因此对于 N、S 元素,简化其结构碎片,只考虑结构碎片的位置,构成的是环还是链。

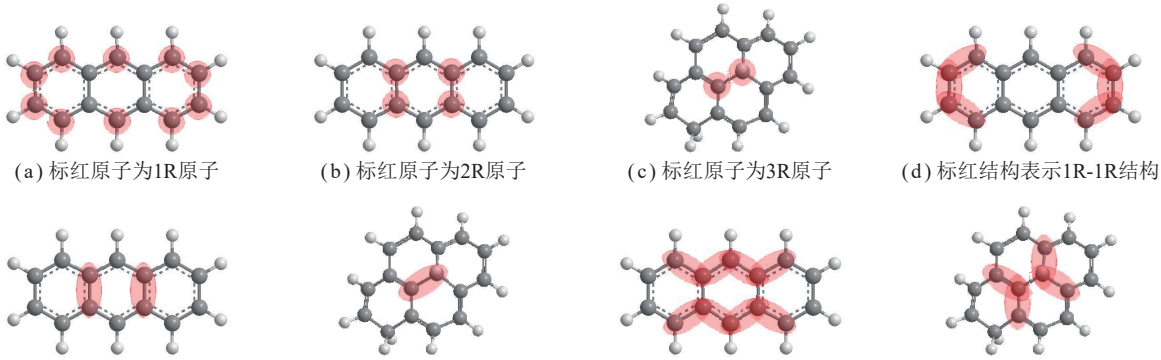
(2) 构建与环结构相关分子描述符。

该部分信息主要包括环与环连接信息、环自身结构信息及不同的大小环的数量信息。① 环上结构。构造的环上结构分子描述符从原子类型与键的种类进行考虑,具体如表 1 与图 1 所示。R 为除氢原子外环上任意原子;“-”代表任意键;1R、2R、3R 分别代表该原子为 1 个环共用、2 个环共用、3 个环共用。图 1(a)～(c) 为环上不同位置的原子个数信息,图 1(d)～(h) 环上的不同结构。构造的环结构碎片分子描述符一定程度上能够区分一部分环的同分异构体,如表 2 中蒽与菲的结构信息虽大部分一致,但 1R-2R 结构数量信息不同,这样,两者得以区分。② 环大小。煤基液体的数据库中构成单环与多环物质的单一环状结构大小不大于 7,因此这套分子描述符中构成 1 个环结构的最大原子个数为 7。③ 环连接方式及不同连接方式的环数量。煤基液体中含多元环,对多元环要考虑环与环间的连接方式和由不同连接方式连接环的个数。图 2(a)中联苯的 2 个苯环由非环上的单键连接,含有 2 个由脂肪键连接的六元环;图 2(b)中萘则是 2 个六元环环内相接。图 2(c)中 2−苯基萘既有脂肪键连接的六元环,又有通过环环内相接的情况。因此构建了能够区分环连接方式的分子结构描述符。

表 1 环上碎片结构信息

Table 1 Information on the structure of debris on the ring

项目	环碎片结构信息	蒽	菲	芘
环上不同位置的原子个数	1R原子数(图1(a))	10	10	10
	2R原子数(图1(b))	4	4	4
	3R原子数(图1(c))	0	0	2
环上的不同结构数量	1R-1R结构数量(图1(d))	6	7	6
	2R-2R结构数量(图1(e))	2	3	0
	3R-3R结构数量(图1(f))	0	0	1
	1R-2R结构数量(图1(g))	8	6	8
	2R-3R结构数量(图1(h))	0	0	4

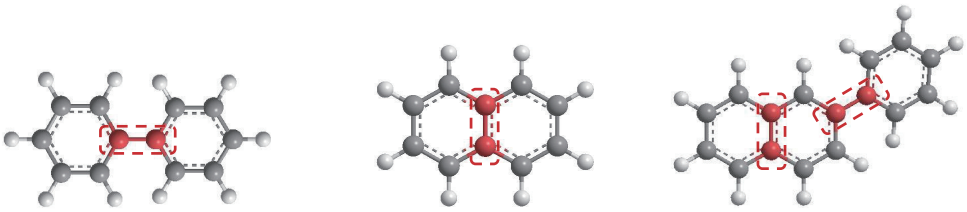


(a) 标红原子为1R原子　(b) 标红原子为2R原子　(c) 标红原子为3R原子　(d) 标红结构表示1R-1R结构

(e) 标红的结构表示2R-2R结构　(f) 标红的结构表示3R-3R结构　(g) 标红的结构表示1R-2R结构　(h) 标红结构表示2R-3R结构

图 1　环碎片结构信息

Fig.1　Illustration of the structure of debris on the ring



(a) 环与环由非环上的脂肪键连接　(b) 环与环由共有的键连接　(c) 包含(a)、(b)

图 2　环的不同连接方式

Fig.2　Different ways of connecting rings

表 2　C 和 O 元素碎片结构信息

Table 2　Information on the structure of the C and O elemental fragments

结构碎片	碎片划分方式
—C—	构成脂肪链；构成脂肪环；构成芳香环；连接脂肪环；连接芳香环；有两侧都连接环
—CH—	构成脂肪链；构成脂肪环；连接芳香环；连接脂肪环；连接芳香环；有两侧都连接环
—CH_2—	构成脂肪链；构成脂肪环；连接芳香环；连接脂肪环；有两侧都连接环
—CH_3	构成脂肪链；连接脂肪环；连接芳香环
C原子	脂肪环上C数；芳香环上C数
—C═C— (不考虑—C═C—OH)	构成脂肪链；构成脂肪环；连接脂肪环；连接芳香环；有两侧都连接环
—C≡C—	构成脂肪链；构成脂肪环；连接脂肪环；连接芳香环；有两侧都连接环
—COOH	构成脂肪链；连接脂肪环；连接芳香环
—COO—	构成脂肪链；构成脂肪环；C连接脂肪环；O连接脂肪环；C连接芳香环；O连接芳香环；有两侧都连接环
—OH	构成脂肪链；连接脂肪环；连接芳香环
—C═C—OH	构成脂肪链；构成脂肪环
—C(═O)H	构成脂肪链；脂肪环取代；芳环取代
—C(═O)—	构成脂肪链；构成脂肪环；构成芳香环；C连接脂肪环；O连接脂肪环；C连接芳香环；O连接芳香环；有两侧都连接环
—O—	构成脂肪链；构成脂肪环；构成芳香环；连接脂肪环；连接芳香环；有两侧都连接环

1.2.2　分子量和不同原子个数描述符的添加

WANG 等[28]将燃料特性与氢碳摩尔比 n(H/C) 和分子量 (M) 通过最小二乘法相关联, 表明 n(H/C) 和 M 的耦合适用于估计化合物的密度等性质。因此, 除构建分子结构碎片描述符外, 同时在分子描述符中增加分子量与不同原子个数的分子描述符, 这有益于后续 ML 训练出更好的性质预测模型, 分子量与不同原子个数的描述符来源于 RDKit 工具包本身。最终得到除分子量、不同原子个数描述符外的不同类型结构描述符, 共 115 个分子描述符, 详见表 2～4。

1.3　执行方式

分子中原子的追踪顺序不同使得同一分子可有

表 3 N 和 S 元素碎片结构信息

Table 3 Information on the structure of the N and S elemental fragments

结构碎片	碎片划分方式	结构碎片	碎片划分方式	结构碎片	碎片划分方式
—NO_2	构成脂肪链	—NH_2	构成脂肪链	—S—	构成脂肪链；构成环
N(═O)(O)O—	构成脂肪链	—C(═O)N—	构成脂肪链；构成环	—C═S	基团数量
—N═O	构成脂肪链	—C═N—	构成脂肪链；构成环	—S(═O)—	基团数量
—ONO_2	构成脂肪链	—C═C—N—	构成脂肪链；构成环	—S(═O)(═O)—	基团数量
—N—	构成脂肪链；构成环	—C≡N	基团数量	—SO_3H	基团数量
—NH—	构成脂肪链；构成环	—C═N—OH	基团数量	—NC═S	基团数量

表 4 环结构信息

Table 4 Ring structure information

考虑方式		结构碎片
环大小		三元环；四元环；五元环；六元环；七元环
环碎片结构		1R原子；2R原子；3R原子；1R-1R结构；2R-2R结构；3R-3R结构；2R-3R结构
环链接方式	仅存在1R	只有环通过脂肪键连接相接的环数
	仅存在1R和2R	同时有环与环本身相接及环与通过脂肪键连接相接情况，通过脂肪键连接的单环数；同时有环与环本身相接及环与通过脂肪键连接相接情况，环与环本身相接的环数；连接环的脂肪键数量；只有环与环本身相接及环与通过脂肪键连接相接情况
	同时存在1R、2R和3R	同时有环与环本身相接及环与通过脂肪键连接相接情况，通过脂肪键连接的单环数；同时有环与环本身相接及环与通过脂肪键连接相接情况，环与环本身相接的环数；连接环的脂肪键数量；只有环与环本身相接及环与通过脂肪键连接相接情况

多种 SMILES 表达式[29]，无论分子的 SMILES 为何种形式，通过本文所述方法拆解得到的分子碎片结果都一致，构建的分子描述符与分子是单向映射关系。此外，虽然构造的分子描述符适用于煤基液体中的物质，但这是根据本研究的实际情况出发的，对于其他研究，也可以根据需要编写代码，增减分子描述符。

提取分子描述符在得到分子信息前，只需获得各分子 CAS 号或 IUPAC 名称，通过开源程序 chemcell 自动获得各分子的 SMILES，煤基液体中物质种类繁杂，添加批处理指令，方便快捷。对于无法自动获得的分子 SMILES 信息，手动通过 CAS 号或 IUPAC 名称查找分子结构，利用 StoneMind 软件识别出 SMILES。将各分子 SMILES 信息输入程序中即可得到各分子描述符，保存至 EXCEL 表格以备 ML 模型调用。

1.4 ML-QSPR 模型开发

为验证构建的分子描述符正确性与可行性，选取经典的反向传播神经网络 (BPNN) 算法搭建 ML 单分子的结构−性质模型。BPNN 是经典的神经网络算法，其输入输出关系是一个高度非线性映射关系[30]，化学分子结构与性质间的关系也是高度非线性，因此使用 BPNN 算法作为 ML 算法。

ML-QSPR 模型开发步骤如下：

(1) 准备数据库。LHV、ρ、FP、CN 四个燃料性能参数的物质库及各物质对应的性质数据；

(2) 构建分子指纹。将数据库中的物质转化为 SMILES，成为计算机能识别的分子语言，基于 SMILES 语言构建分子描述符，构建的分子描述符包括结构分子描述符、分子量描述符、原子个数描述符；

(3) 计算分子描述符。使用构建的分子描述符对数据库中的分子进行拆解，得到后续用于 ML 的分子描述符库；

(4) 构建 ML-QSPR 模型。将步骤 (3) 中得到的分子描述符数据及其对应的性质数据提供给 ML 模型，对模型进行学习训练，最终得到参数性能最佳的 4 个性质预测模型。

图 3 为建立 ML-QSPR 模型步骤，上述构建的分子描述符为 ML 输入特征，通过 ML-QSPR 模型实现 LHV、ρ、FP、CN 性质预测。

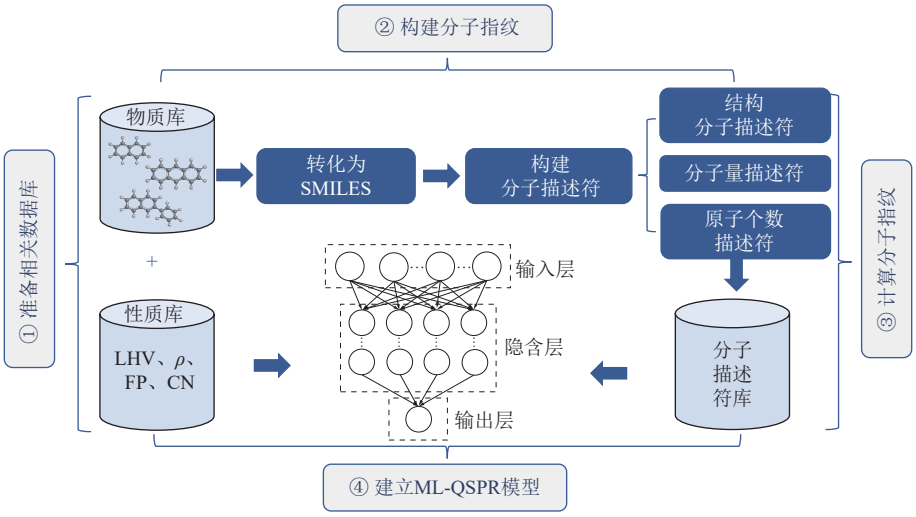


图 3　ML-QSPR 模型开发步骤

Fig.3　ML-QSPR model development steps

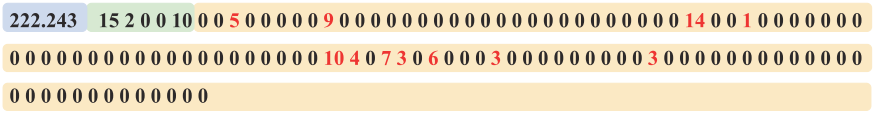
2　结果与讨论

2.1　分子描述符

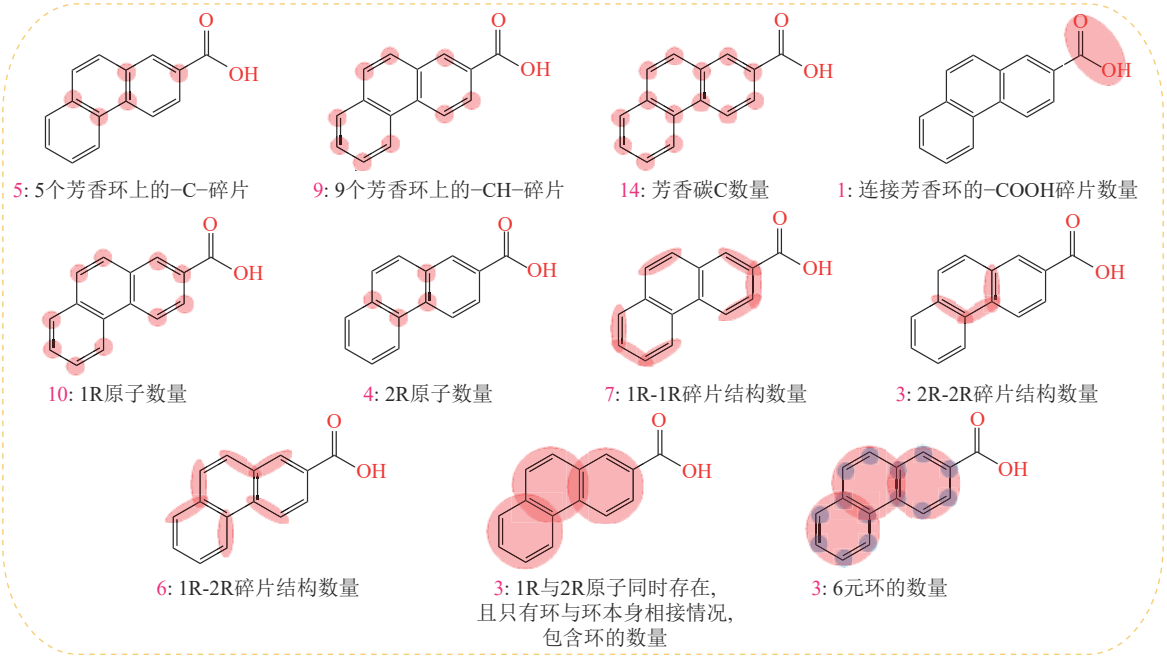
图 4 为以 2−菲甲酸为例进行拆解所获得的分子描述符结果, 图 5 为以 9−(4−联苯基) 咔唑进行拆解所获得的分子描述符结果。图 4(a)、5(a) 中蓝框的数为分子量, 绿框的数字分别代表 C、O、N、S、H 原子个数, 黄框为拆解的分子碎片结构, 分子结构中若包含该结构碎片, 便对该碎片进行计数, 否则对于不包含该碎片的结构数量记录为 0, 结构碎片除 0 外的数字含义如图 4(b)、5(b) 所示。

2.2　ML-QSPR 模型网络最优结构

数据集由随机划分算法按 7∶1.5∶1.5 分为训练集、验证集和测试集。BPNN 网络由输入层、隐含层及输出层组成, 网络最优结构见表 5, 输入节点/特征



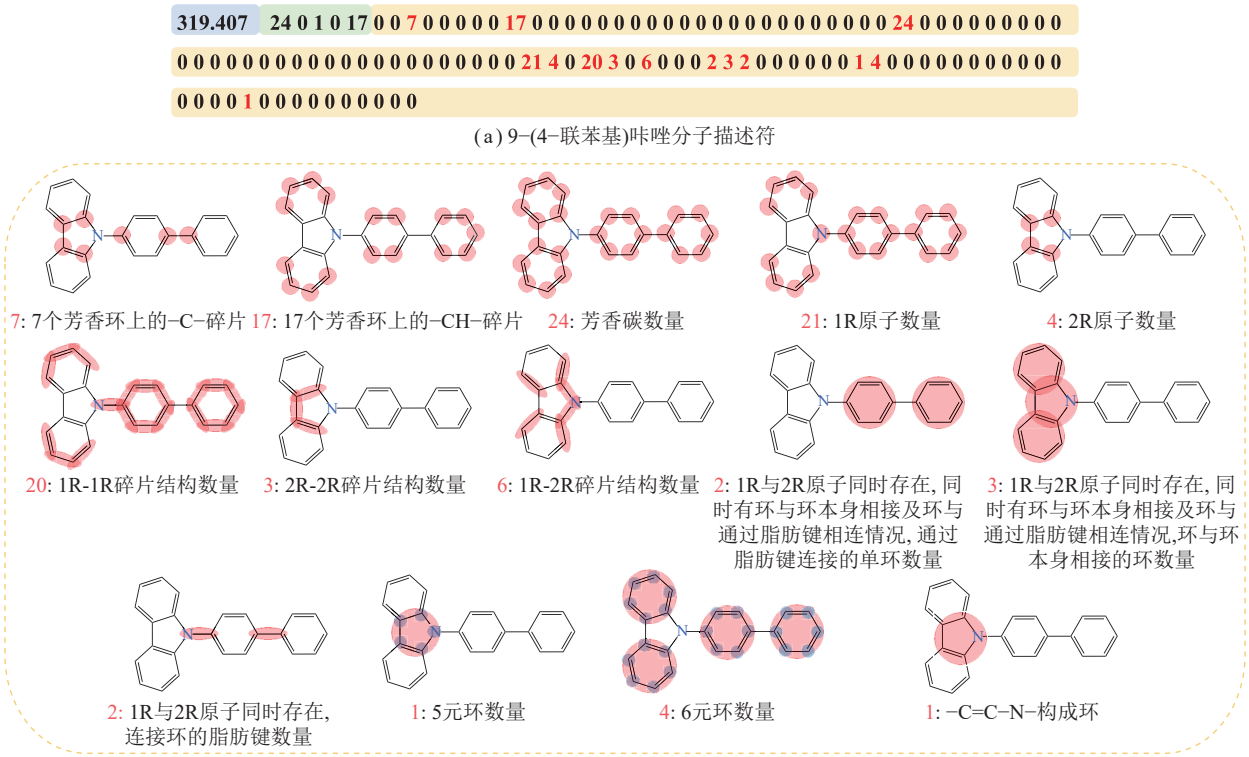
(a) 2−菲甲酸分子描述符



(b) 分子结构碎片中数字含义

图 4　以 2−菲甲酸为例得到的分子描述符信息

Fig.4　Molecular descriptor information obtained using 2-phenanthrenecarboxylic acid as an example



(a) 9−(4−联苯基)咔唑分子描述符

(b) 分子的结构碎片中数字含义

图 5 以 9−(4−联苯基) 咔唑为例得到的分子描述符信息

Fig.5 Molecular descriptor information obtained using 9-(4-biphenyl)carbazole as an example

为构建的分子描述符。Levenberg-Marquardt 训练算法一般情况下对中小型模型收敛速度快, 文中 4 个模型都为中小型模型, 故选用该训练算法。在输入层和输出层均使用线性函数作为激活函数, 隐含层中每一层使用双正切函数作为激活函数, 使得构造的神经网络能够解决非线性问题。

表 5 神经网络性质预测模型最优结构

Table 5 Optimal structure of neural network property prediction model

性质	输入层节点数	隐含层层数	各隐含层节点数	输出层节点数
LHV	115	3	30/10/5	1
ρ	115	3	20/10/5	1
FP	115	3	25/10/5	1
CN	115	3	25/10/5	1

2.3 ML-QSPR 模型性能

最佳模型的性能和精度通过 3 种类型的数学统计参数进行检验: 决定系数 (R^2)(式 (1))、均方根误差 (E_{RMS})(式 (2)) 和相对平均偏差 (D_{AAR})(式 (3))。

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

$$E_{RMS} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{2}$$

$$D_{AAR} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{3}$$

其中, m 为每个阶段 (训练、测试或整个数据集) 数据集中化合物的数量; y_i为 y 的实验值; $\hat{y}_i$为 y 预测值; $\bar{y}_i$为 y 的实验值平均值。最佳模型确定后, 通过该模型预测数据库各物质 LHV、ρ、FP、CN。表 6 为 ML-QSPR 模型性能结果。

图 6 为 4 个燃料性能参数 QSPR-ML 模型回归情况, 4 个燃料性能参数数据集总体 R^2>0.9; LHV、ρ、FP 的模型所有数据集 R^2 分别为 0.996、0.988、0.987; CN 预测中加入混合物数据进行预测, R^2=0.959, 各个性质模型实验值与预测值吻合较好。从各个性质模型 R^2 可看出, 4 个模型中性质的变异能通过回归关系被本文构建的分子描述符解释的比例高, 说明预测值与实验值拟合效果优良。也进一步表明 1.2 节针对煤基液体、基于化学结构构建的分子描述符为搭建的 ML-QSPR 模型提供了可靠输入特征, 能很好地与物质性质建立起关联, 因此构建的分子描述符能反映物质的真实性质。为进一步说明预测结果, 计算 LHV、ρ、FP、CN 预测误差分布如图 6 所示, 各子集的预测

表 6　模型不同数据集性能结果

Table 6　Model performance for different datasets

性质	数据类型	数据量	R^2	E_{RMS}	D_{AAR}
LHV/(kJ · mol^{-1})	训练集	960	0.995	186.410	0.022
	验证集	205	0.996	208.284	0.080
	测试集	208	0.996	172.808	0.016
ρ/(g · cm^{-3})	训练集	768	0.987	0.017	0.013
	验证集	165	0.991	0.013	0.011
	测试集	163	0.986	0.014	0.012
FP/K	训练集	740	0.994	8.453	0.018
	验证集	158	0.965	18.439	0.033
	测试集	158	0.976	17.067	0.031
CN	训练集	264	0.960	4.943	0.052
	验证集	56	0.945	5.220	0.052
	测试集	56	0.962	5.213	0.054

误差在零点两侧均匀分布，基本服从正态分布，说明该模型不存在系统性误差。同时对各个类别化合物的预测结果进行统计，见表 7～10。

对图 6 中各预测模型离散点进行分析，考虑原因有：① 离散点包含手性分子，如 CN 中 (7S,8S)−二甲基十四烷、(7R,8R)−二甲基十四烷、(7R,8S)−二甲基十四烷，模型对这些手性分子预测不准，是因未构建能区分手性分子的描述符；② 离散点还包含多种官能团且支链多的化合物，如 LHV 中的四醇四硝酸酯，FP 中乙烯基碳酸乙烯酯等物质，当物质结构复杂包含多种官能团且支链多的化合物时，模型可能对其预测不准确；③ 数据库中数据分布也会影响模型的预测准确性。如 LHV 离散点包括无环醚类物质四乙二醇二甲醚，四乙二醇二甲醚分子量为 194.227，该类数据中 92% 分子量均小于 190，分子量大于 190 分子数据少，这可能会使得模型由于缺少相关数据而不能很好拟合。

与以往预测研究进行比较，结果见表 11。本研究中使用的模型 ρ 的 R^2 略低于其他研究，但其余 3 种性质 R^2 均优于文献，说明构建的分子描述符与模型效果较好。

2.4　数据量对 ML-QSPR 模型的影响

以 LHV 为例，研究总数据量对模型的影响。首先，基于 1 373 个数据得到的 LHV 模型的最优参数，再随机试验 9 次，得到 10 个不同的模型，考查 10 个模型 3 个集合的 R^2、模型的稳定性及出现过拟合情况次数的概率。通过二分法对总数据量进行试验，将所有数据随机先划分成 2 份，使用其中一份数据量为

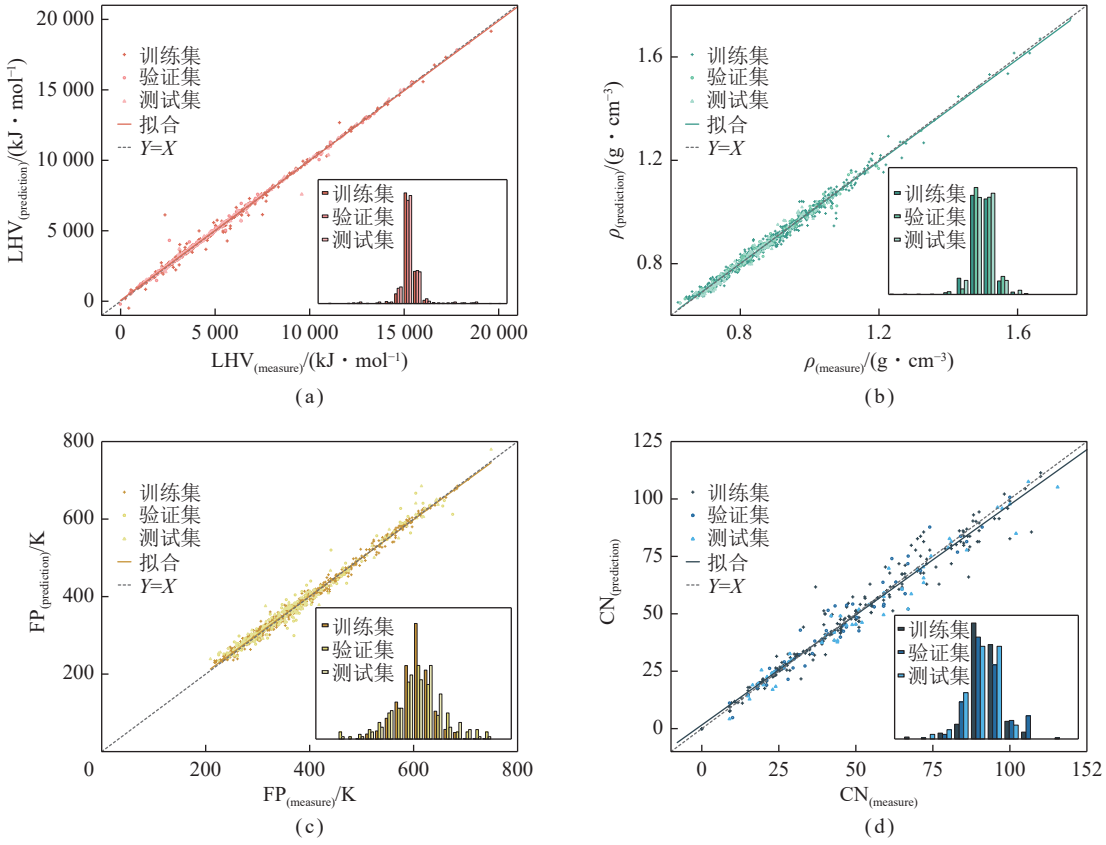


图 6　BPNN 模型回归

Fig.6　BPNN model regression plot

表 7 各个类型化合物 LHV 预测结果

Table 7 LHV prediction results for each type of compound

类别	小类	m	D_{AAR}	E_{RMS}/(kJ·mol^{-1})	R^2
脂肪烃	正构烷烃	31	0.012	342.393	0.998
	异构烷烃	70	0.006	34.439	0.999
	烯烃	107	0.008	58.613	0.999
	烷二烯	32	0.027	203.134	0.983
	炔烃	18	0.014	43.396	0.999
	环烷烃	98	0.009	149.348	0.999
	环烯	18	0.011	152.254	0.980
芳香烃	芳香族	168	0.014	207.608	0.995
含O化合物	醇类	121	0.019	177.202	0.995
	环醇	8	0.008	36.363	0.999
	醛类	34	0.009	40.769	0.999
	酮类	31	0.009	31.844	0.999
	环酮	7	0.021	129.041	0.995
	饱和酯	91	0.013	109.740	0.999
	不饱和酯	28	0.017	128.915	0.997
	无环醚	50	0.021	217.429	0.988
	呋喃类	6	0.010	31.110	0.999
	其他环醚	9	0.064	184.221	0.994
	碳酸酯	4	0.032	74.884	0.999
	羧酸	56	0.037	150.004	0.997
含N化合物		238	0.042	279.532	0.986
含S化合物		86	0.167	201.913	0.992
其他	多官能团	62	0.030	157.815	0.995

表 8 各个类型化合物 ρ 预测结果

Table 8 ρ prediction results for each type of compound

类别	小类	m	D_{AAR}	E_{RMS}/(g·cm^{-3})	R^2
脂肪烃	正构烷烃	25	0.007	0.007	0.997
	异构烷烃	75	0.015	0.013	0.993
	烯烃	96	0.01	0.009	0.997
	烷二烯	25	0.017	0.015	0.991
	炔烃	20	0.008	0.008	0.997
	环烷烃	71	0.009	0.011	0.988
	环烯	13	0.014	0.018	0.950
芳香烃	芳香族	146	0.008	0.013	0.981
含O化合物	醇类	117	0.016	0.017	0.988
	环醇	7	0.008	0.012	0.951
	醛类	33	0.015	0.016	0.975
	酮类	47	0.009	0.013	0.970
	环酮	3	0.007	0.001	0.954
	饱和酯	138	0.010	0.017	0.974
	不饱和酯	38	0.016	0.021	0.941
	无环醚	61	0.016	0.015	0.979
	呋喃类	12	0.006	0.008	0.985
	其他环醚	14	0.011	0.014	0.986
	碳酸酯	5	0.017	0.019	0.995
	羧酸	32	0.024	0.029	0.994
含N化合物		38	0.012	0.016	0.990
含S化合物		7	0.019	0.039	0.923
其他	多官能团	73	0.015	0.021	0.989

686 的数据集进行试验，同样随机试验 10 次；继续采用二分法分别随机取 343、172、86 个数据使用相同的模型参数各试验 10 次。统计上述获得的模型训练集、验证集、测试集 3 个集合的 R^2，见表 12，可以看出 50 次试验中训练集除了使用 86 个数据量训练，出现 1 次模型的训练集 R^2<0.95，其余训练集得到的 R^2 均高于 0.95，其中总数据量为 343、686、1 373 个时，R^2 均大于 0.99，说明对于训练集几种不同的总数据量训练效果都较好，增大总数据量有益于模型训练效果更好。但统计验证集与测试集 R^2<0.95 的情况，发现随着总数据量的下降，测试集与验证集的 R^2<0.95 的情况次数也随之增加，随机得到的模型更不稳定，也更易出现过拟合情况。当总数据量为 1 373 时，发现在 10 次随机训练中仍然出现一次测试集的 R^2<0.95，这说明数据量依旧需要继续增大，但相比于总数据量为 86、343、686，模型明显更稳定更不易过拟合。

此外，继续考察小类数据量对模型造成的影响。同样以 LHV 数据集为例，LHV 数据集中小类数据量最多的是含 N 化合物，因此选用含 N 化合物研究小类数据量大小对模型的影响。基于上文得到的最优参数，使用二分法逐步减少该类化合物的数量，统计模型在含 N 化合物减少得到的总数据集的 R^2 的情况，见表 13，当含 N 化合物数据量减少至 0 时，总数据量为 1 135 个，对总数据集的 R^2 影响不大，而当数据类别个数过小，可能使得模型对含 N 化合物预测效果差，如当含 N 化合物只剩 2 个时，含 N 化合物 R^2 只有 0.449。因此，LHV 选取的各类别化合物数据量对总数据集的 R^2 没有太大影响。

2.5 混合物性质预测

煤基液体产品为复杂的混合物，因此还需考虑混合物的性质与混合物中各组分化合物性质的数量关系，故在 CN 中对混合物性质的预测进行了初步探索。CN 数据库中包含部分二、三元混合物数据，在预测过程中考虑使用构建描述符按照各组分化合物比例关系相加得到混合物描述符进行性质预测。假设混合物包含组分 A、组分 B、…、组分 X，其具体算法如下：

表 9　各个类型化合物 FP 预测结果

Table 9　FP prediction results for each type of compound

类别	小类	m	D_{AAR}	E_{RMS}/K	R^2
脂肪烃	正构烷烃	28	0.012	7.317	0.992
	异构烷烃	64	0.017	6.596	0.995
	烯烃	67	0.024	7.917	0.994
	烷二烯	21	0.023	8.694	0.995
	炔烃	12	0.015	5.130	0.998
	环烷烃	36	0.026	13.233	0.979
	环烯	11	0.028	9.897	0.991
芳香烃	芳香族	128	0.021	15.391	0.918
含O化合物	醇类	123	0.026	11.941	0.950
	环醇	9	0.019	10.875	0.910
	醛类	37	0.020	8.964	0.986
	酮类	40	0.017	6.593	0.999
	环酮	2	0.059	27.629	0.583
	饱和酯	111	0.018	8.283	0.987
	不饱和酯	32	0.028	10.315	0.981
	无环醚	43	0.047	15.872	0.970
	呋喃类	7	0.028	8.473	0.994
	其他环醚	7	0.015	4.727	0.997
	碳酸酯	4	0.069	36.630	0.630
	羧酸	56	0.026	13.234	0.972
含N化合物		113	0.014	13.794	0.996
含S化合物		29	0.020	16.667	0.994
其他	多官能团	76	0.029	12.801	0.945

表 10　各个类型化合物 CN 预测结果

Table 10　CN prediction results for each type of compound

类别	小类	m	D_{AAR}	E_{RMS}	R^2
脂肪烃	烯烃	34	0.043	2.275	0.985
	链烷烃	49	0.133	4.759	0.966
芳香烃	芳烃	34	0.046	7.405	0.947
	萘类	24	0.053	2.739	0.980
含O化合物	醇类	17	0.048	3.051	0.964
	酯类	88	0.087	5.829	0.957
混合物	混合物	130	0.064	3.568	0.970

混合物第 i 个分子描述符 = 组分 A 第 i 个分子描述符×混合物中组分 A 占比 + 组分 B 第 i 个分子描述符×混合物中组分 B 占比 + … + 组分 X 第 i 个分子描述符×混合物中 组分 X 占比。

不同物质混合后 CN 的预测结果见表 14, 可以看出模型不能很好预测正癸烷−异辛烷和正十二烷−异辛烷混合物的 CN, 同样模型对正庚烷−正丁基环己烷和正庚烷−甲基环己烷预测效果不佳, 异辛烷与正庚烷物质混合又有良好的预测效果, 说明模型不能对所有的混合物进行完美拟合。一方面是各部分混合物数据量过少, 另一方面可能是由于当物质与不同化合

表 11　ML-QSPR 模型预测结果与文献比较

Table 11　Comparison of the model proposed in this paper with previous studies

性质	模型	方法	m	E_{RMS}	R^2
LHV/(kJ · mol^{-1})	本研究	ML-QSPR	1 373	187.878	0.996
	Frutiger[31]	MG GC	794	368.166	0.990
	Albahri[12]	ANN	586	946.731	0.994
ρ/(g · cm^{-3})	本研究	ML-QSPR	1 096	0.016	0.988
	Roubehie[11]	QSPR-ANN	222	0.006	0.995
	Saldana[32]	ANN	730	0.307	0.997
FP/K	本研究	ML-QSPR	1 056	12.022	0.987
	Satyanarayana[33]	Correlation (sp. gr. + NBP)	250		0.980
CN	本研究	ML-QSPR	376	5.026	0.959
	Saldana[20]	QSPR	625	6.300	0.959
	Saldana[20]	QSPR-GM	229	6.300	0.934

表 12　LHV 不同总数据量下 3 种数据集 $R^2 < 0.95$ 次数统计

Table 12　Statistics on the number of times $R^2 < 0.95$ for the three datasets with different total data volume

m	训练集	验证集	测试集
86	1	6	7
172	0	1	8
343	0	2	5
686	0	1	3
1 373	0	0	1

表 13　LHV 数据集中减少含 N 化合物数据个数对模型的影响

Table 13　LHV dataset impact of reducing the number of N-containing compounds on modeling

m	含N化合物数据量	总R^2	含N化合物R^2
1 135	0	0.988	0
1 136	1	0.984	0.994
1 137	2	0.996	0.449
1 139	4	0.989	0.999
1 144	9	0.995	0.974
1 150	15	0.992	0.999
1 165	30	0.996	0.999
1 195	60	0.994	0.999
1 254	119	0.995	0.997
1 373	238	0.996	0.986

物混合其性质差异大，致使模型无法成功对其进行预测。

表 14 混合物 CN 预测结果

Table 14 Mixture CN prediction

组成	m	R^2
正庚烷−异辛烷	12	0.990
正庚烷−异辛烷−甲苯	15	0.986
异辛烷−1−己烯	9	0.997
异辛烷−环己烷	9	0.999
异辛烷−四氢萘	9	0.998
正癸烷−甲苯	4	0.980
正癸烷−异辛烷	23	0.535
正癸烷−异辛烷−甲苯	2	0.909
正十六烷−异十六烷	5	0.951
正十二烷−异辛烷	5	0.765
正庚烷−正丁基环己烷	4	0.497
正庚烷−十氢萘	4	0.936
正庚烷−甲基环己烷	7	0.787
正庚烷−环戊烷	12	0.968
正十六烷−2,2,4,4,6,8,8−七甲基壬烷	10	0.999
合计	130	0.970

2.6 ML-QSPR 方法与实验方法对比

由表 15 可知，实际一次测量 4 个燃料性能参数需要耗费一定的时间与燃料样本。此外，实验人员在实验过程中较难获得纯净的实验样本；并且在测量有毒、挥发性、爆炸性或高反应活性物质的 LHV、ρ、FP 以及 CN 时存在挑战且耗时[34-35]。而本研究构建的 ML-QSPR 方法，仅需知道化合物的化学结构，便能对 LHV、ρ、FP 和 CN 这 4 个燃料性能参数进行估计，节约了时间与成本。

表 15 ML-QSPR 方法与实验方法对比

Table 15 Comparison of the ML-QSPR method with experimental

测量性质	需要样品的量		需要的时间	
	单次实验	ML-QSPR	单次实验	ML-QSPR
LHV	0.2～0.5 g[36]	0	>4 h[2]	<1 min
ρ	100 mL[37]	0		<1 min
FP	50～80 mL[38]	0		<1 min
CN	40～500 mL[17]	0	40 min[3]	<1 min

2.7 4 个燃料性能参数预测结果分析

化合物中原子在不同的化学环境中具有不同特性，例如含相同官能团但碳原子数不同的物质性质各异。因此，接下来对不同化学环境的化合物进行性质分析。研究 LHV、ρ、FP、CN 与分子结构的关系，为筛选适合于替代燃料应用的分子提供参考。

图 7 为当碳数 (n) 增加时不同的族化合物性质演化的趋势。需要说明的是图 6 是由模型预测得到的信息绘制，并且选取不同族组分的物质比较时，同一类物质的官能团或环，除异构烷烃与酯类物质外，都取代的是主链碳第 1 个或第 2 个碳原子。

(1) LHV: 如图 7(a) 所示，n 相同时，不同族组分的物质 LHV 差距小，LHV 主要由 n 影响，分子 n 增加，LHV 显著下降。

(2) ρ: 如图 7(b) 所示，正构烷烃、炔烃、非环醚、环烷烃、醛、醇类物质的 n 增加，ρ 增加，其中环烷烃类物质随着环烷烃环数增加明显，如环己烷预测 ρ 为 0.776 g/cm^3，环庚烷 ρ 为 0.814 g/cm^3，环上增加 1 个碳，ρ 增加 0.038 g/cm^3，而正构烷烃、炔烃、非环醚类增加量都小于 0.02 g/cm^3；芳香环、饱和酯、羧酸 ρ 随着 n 增加而减少。但随着 n 增加，除环烷烃外，各类物质的 ρ 变化趋势趋于缓和。

(3) FP: 如图 7(c) 所示，醇类、苯类、环己烷类、醛类、正构烷烃、异构烷烃、烯烃都随 n 增加而增加，FP 主要取决于分子中碳原子总数，这与文献[28]中提出的规律一致。这些小类中，n 相同时，醇类>醛类>六元芳香环>六元环烷烃>正构烷烃>甲基取代二位的异构烷烃>烯烃。相同碳数下，含氧类物质，FP 高于不含氧类物质。

(4) CN: 如图 7(d) 所示，分析醇类、芳香族、酯类、烯烃、烷烃 5 类物质，可以看出 CN 随着 n 增加而增加，且同碳数的情况下，烷烃的 CN 明显高于其余 3 类。

通过增加不同族组分的物质的碳原子数量得到的 4 个燃料性能参数演变趋势可以看出，4 个燃料性能参数都受 n 的影响明显。对比不同族组分的物质各个性质可知，不同族组分的物质 LHV 差距小，主要由 n 决定；而对于 ρ、FP 和 CN，不同族组分的物质性质差距明显。

3 结 论

(1) 构建了一套适合煤基液体物质的分子描述符，可用于提取煤基液体物质结构、分子量及原子个数的 115 种特征。建立分子机器学习−定量结构性质关系方法 (ML-QSPR)，实现对煤基液体的低位热值、密度、闪点和十六烷值 4 个燃料性能参数的快速准确预测。

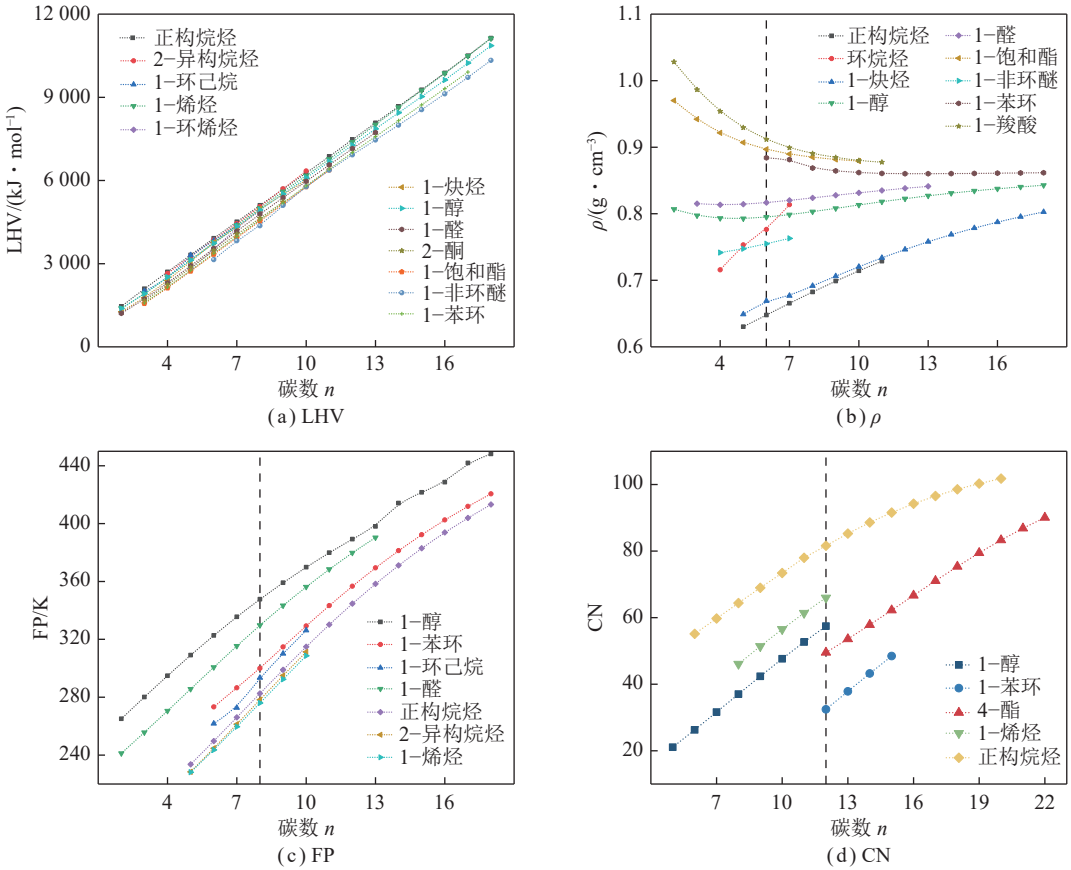


图 7　*n* 不同时各族组分物质性质演化趋势

Fig.7　Trends in the evolution of the properties of each group of substances with different *n*

(2) 采用 ML-QSPR 方法, 预测煤基液体 4 个燃料性能参数研究表明: 随着不同族组分的物质碳数增加, 4 个燃料性能参数都受碳数 *n* 的影响显著。不同族组分的物质低位热值差距小, 其主要由碳数决定; 而不同族组分物质的密度、闪点和十六烷值性质差距明显。

(3) 除了对文中所述的 4 个燃料性能参数能够进行较好的预测外, 本研究提出的分子描述符方法具有通用性, 不仅能够快速获取物质的分子特征, 而且能够根据需要增加或减少特征数量, 将训练好的模型用于预测新的物质分子。

参考文献(References):

[1] BITAM S, HAMADACHE M, HANINI S. Prediction of therapeutic potency of tacrine derivatives as BuChE inhibitors from quantitative structure−activity relationship modelling[J]. SAR and QSAR in Environmental Research, 2018, 29(3): 213−230.

[2] DAI Y M, ZHU Z P, CAO Z, et al. Prediction of boiling points of organic compounds by QSPR tools[J]. Journal of Molecular Graphics and Modelling, 2013, 44: 113−119.

[3] JIANG J C, DUAN W J, WEI Q, et al. Development of quantitative structure-property relationship (QSPR) models for predicting the thermal hazard of ionic liquids: A review of methods and models[J]. Journal of Molecular Liquids, 2020, 301: 112471.

[4] RAHAL S, HADIDI N, HAMADACHE M. In Silico prediction of Critical Micelle Concentration (CMC) of classic and extended anionic surfactants from their molecular structural descriptors[J]. Arabian Journal for Science and Engineering, 2020, 45(9): 7445−7454.

[5] ZUAS O, STYARINI D. A quantitative structure-property relationship (QSPR) evaluation of critical volume of unsaturated hydrocarbon alkenes and alkynes using simple connectivity indices[J]. Reaktor, 2009, 12: 260−267.

[6] LI R, HERREROS J M, TSOLAKIS A, et al. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types[J]. Fuel, 2021, 304: 121437.

[7] HANSCH C, FUJITA T. *p*-*σ*-*π* analysis. A method for the correlation of biological activity and chemical structure[J]. Journal of the American Chemical Society, 1964, 86(8): 1616−1626.

[8] 何婷. 基于机器学习结合定量构效关系 (QSPR) 的含能材料爆轰性能预估及筛选方法研究[D]. 西安: 西北大学, 2021.
HE Ting. Research on the prediction and screening method of energy-containing material blast performance based on machine learning combined with quantitative constitutive relationship (QSPR)[D]. Xi'an: Northwest University, 2021.

[9] HAMADACHE M, BENKORTBI O, HANINI S, et al. A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction[J]. Journal of Hazardous Materials, 2016, 303: 28−40.

[10] HAMADACHE M, HANINI S, BENKORTBI O, et al. Artificial neural network-based equation to predict the toxicity of herbicides on rats[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 154: 7−15.

[11] ROUBEHIE FISSA M, LAHIOUEL Y, KHAOUANE L, et al. QS-

PR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods[J]. Journal of Molecular Graphics and Modelling, 2019, 87: 109−120.

[12] ALBAHRI T A. Accurate prediction of the standard net heat of combustion from molecular structure[J]. Journal of Loss Prevention in The Process Industries, 2014, 32: 377−386.

[13] CRETON B, DARTIGUELONGUE C, DE BRUIN T, et al. Prediction of the cetane number of diesel compounds using the quantitative structure property relationship[J]. Energy & Fuels, 2010, 24(10): 5396−5403.

[14] DAHMEN M, MARQUARDT W. A novel group contribution method for the prediction of derived cetane number of oxygenated hydrocarbons[J]. Energy & Fuels, 2015, 29(19): 5781−5801.

[15] KARELSON M, PERKSON A. QSPR prediction of densities of organic liquids[J]. Computers & Chemistry, 1999, 23(1): 49−59.

[16] KUBIC W L, JENKINS R W, MOORE C M, et al. Artificial neural network based group contribution method for estimating cetane and octane numbers of hydrocarbons and oxygenated organic compounds[J]. Industrial & Engineering Chemistry Research, 2017, 56(42): 12236−12245.

[17] LI R Z, HERREROS J M, TSOLAKIS A, et al. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures[J]. Fuel, 2020, 280: 118589.

[18] PAN Y, JIANG J C, WANG R, et al. Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization[J]. Journal of Loss Prevention in the Process Industries, 2011, 24(1): 85−89.

[19] REN Y Y, ZHAO B W, CHANG Q, et al. QSPR modeling of nonionic surfactant cloud points: An update[J]. Journal of Colloid and Interface Science, 2011, 358(1): 202−207.

[20] SALDANA D A, STARCK L, MOUGIN P, et al. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods[J]. Energy & Fuels, 2011, 25(9): 3900−3908.

[21] WANG Y, CAO Y, WEI W, et al. A new method of estimating derived cetane number for hydrocarbon fuels[J]. Fuel, 2019, 241: 319−326.

[22] WON S H, DOOLEY S, VELOO P S, et al. The combustion properties of 2, 6, 10-trimethyl dodecane and a chemical functional group analysis[J]. Combustion and Flame, 2014, 161(3): 826−834.

[23] 武立俊, 皮中原, 王烨敏. 煤直接液化产物中含氧组分试验研究[J]. 中国煤炭, 2018, 44(1): 94−97.
WU Lijun, PI Zhongyuan, WANG Yemin. Experimental study of oxygenated components in coal direct liquefaction products[J]. China Coal, 2018, 44(1): 94−97.

[24] 李伟林, 石智杰, 张晓静, 等. 煤直接液化油中硫氮化合物的类型分布[J]. 洁净煤技术, 2015, 21(4): 55−57, 45.
LI Weilin, SHI Zhijie, ZHANG Xiaojing, et al. Type distribution of sulphur and nitrogen compounds in coal direct liquefaction oil[J]. Clean Coal Technology, 2015, 21(4): 55−57, 45.

[25] 慕海. 以物料衡算为约束的煤基粗油定性定量研究[D]. 太原: 太原理工大学, 2018.
MU Hai. Qualitative and quantitative research on coal-based crude oil with material accounting as a constraint [D]. Taiyuan: Taiyuan University of Technology, 2018.

[26] LI W Y, WANG W, MU H, et al. Analysis of light weight fractions of coal-based crude oil by gas chromatography combined with mass spectroscopy and flame ionization detection[J]. Fuel, 2019, 241: 392−401.

[27] TOROPOV A A, TOROPOVA A P, BENFENATI E. Additive SMILES-Based Carcinogenicity Models: Probabilistic Principles in the Search for Robust Predictions[J]. International Journal of Molecular Sciences, 2009, 10(7): 3106−3127.

[28] WANG X Y, JIA T H, PAN L, et al. Review on the relationship between liquid aerospace fuel composition and their physicochemical properties[J]. Transactions of Tianjin University, 2021, 27(2): 87−109.

[29] CHEN J H, TSENG Y J. Different molecular enumeration influences in deep learning: An example using aqueous solubility[J]. Briefings in Bioinformatics, 2020, 22(3): 1−13.

[30] 刘艳侠, 高新琛. BP 神经网络在材料领域中的应用 (综述)[J]. 辽宁大学学报 (自然科学版), 2007, 34(2): 116−119.
LIU Yanxia, GAO Xinchen. Application of BP neural networks in the field of materials (A review)[J]. Journal of Liaoning University (Natural Science Edition), 2007, 34(2): 116−119.

[31] FRUTIGER J, MARCARIE C, ABILDSKOV J, et al. A comprehensive methodology for development, parameter estimation, and uncertainty analysis of group contribution based property models—An application to the heat of combustion[J]. Journal of Chemical & Engineering Data, 2015, 61(1): 602−613.

[32] SALDANA D A, STARCK L, MOUGIN P, et al. Prediction of density and viscosity of biofuel compounds using machine learning methods[J]. Energy & Fuels, 2012, 26(4): 2416−2426.

[33] SATYANARAYANA K, RAO P G. Improved equation to estimate flash points of organic compounds[J]. Journal of Hazardous Materials, 1992, 32(1): 81−85.

[34] GHARAGHEIZI F, MIRKHANI S A, TOFANGCHI MAHYARI A-R. Prediction of standard enthalpy of combustion of pure compounds using a very accurate group-contribution-based method[J]. Energy & Fuels, 2011, 25(6): 2651−2654.

[35] KATRITZKY A R, STOYANOVA-SLAVOVA I B, DOBCHEV D A, et al. QSPR modeling of flash points: An update[J]. Journal of Molecular Graphics and Modelling, 2007, 26(2): 529−536.

[36] 蓝福燕, 许卫芹, 赵剑. 液体危险废物燃烧热值的测定氧弹量热法[J]. 皮革制作与环保科技, 2021, 2(19): 84−86.
LAN Fuyan, XU Weiqin, ZHAO Jian. Determination of calorific value of liquid hazardous waste by oxygen bomb calorimetry[J]. Leather Making and Environmental Protection Technology, 2021, 2(19): 84−86.

[37] 黄湘来, 赵珊红. 两种原油密度测量方法的比较试验[J]. 计量与测试技术, 2011, 38(5): 50−52.
HUANG Xianglai, ZHAO Shanhong. Comparative test of two crude oil density measurement methods[J]. Measurement and Testing Technology, 2011, 38(5): 50−52.

[38] 陈晓彤, 邹惠玲, 夏攀登. 石油产品闪点测定方法的探析[J]. 硅谷, 2009(22): 108−109.
CHEN Xiaotong, ZOU Huiling, XIA Pandeng. An analysis of methods for the determination of flash point of petroleum products[J]. Silicon Valley, 2009(22): 108−109.