文章编号:1006-9941(XXXX)XX-0001-13

机器学习辅助的烃类分子性质预测与高通量筛选燃料

侯 放[1],齐晓宁[4,5],刘睿宸[1],李 玲[6],王 莅[1,2,3],张香文[1,2,3],李国柱[1,2,3]

(1. 天津大学化工学院, 天津 300072; 2. 先进燃料与化学推进剂教育部重点实验室, 天津 300072; 3. 物质绿色创造与制造海河实验室, 天津 300192; 4. 中国科学院计算技术研究所, 北京 100190; 5. 中国科学院大学, 北京 100190; 6. 承德钒钛新材料有限公司, 河北 承德 067102)

摘 要: 通过数据收集、结构优化和量化计算,建立了碳数从1到50的2899个烃类分子"结构-多种性质"数据集,性质包含熔点(T_m)、沸点(T_b)、密度(ρ)、0 K下的内能(U_0)、298.15 K下的内能(U)、298.15 K下的焓(H)、298.15 K下的吉布斯自由能(G)。以表示分子结构的库伦矩阵作为模型输入,建立了决策树回归模型、交叉验证的最小绝对收缩和选择算子回归模型、交叉验证的岭回归模型、极限梯度提升回归模型4种不同的机器学习模型。通过比较不同模型预测性质的精度得出,极限梯度提升回归模型更适用于预测烃类分子的熔点、沸点、密度等通过实验测得的性质,交叉验证的岭回归模型更适用于预测烃类分子的内能、焓、吉布斯自由能等能量的通过理论计算得到的性质。同时,最优的机器学习组合模型可以准确预测相同碳数、不同种类和同分异构体烃类分子的性质。使用最优的机器学习模型计算了34种已通过实验合成的高密度碳氢燃料的密度,计算值与实验值的平均绝对误差为0.0290 g·cm^{-3}。进而,预测了开源数据库GDB-13C中的319,893个烃类分子的燃料性质,并高通量筛选出了37种低凝固点、高密度的新型碳氢燃料候选分子。采用基团贡献法和DFT方法进一步计算了筛选出的碳氢分子的关键燃料性质,这些新型分子与典型燃料JP-10和QC的质量热值和比冲相当。

关键词: 机器学习;烃类分子;高密度碳氢燃料;性质预测;高通量筛选

中图分类号: TJ5;O64;V31 **文献标志码:** A **DOI:**10.11943/CJEM2024276

0 引 言

烃作为有机化合物中的一类重要物质,广泛应用于有机化工、燃料、溶剂等众多领域。基于烃类分子构效关系的性质理论预测,对于开发新型碳氢燃料、推进剂等至关重要,有利缩短研发周期、降低开发成本,同时可以避免复杂的实验和潜在的危险,指导材料合成并与实验结果交叉验证[1-3]。机器学习(Machine Learning, ML),作为一种新兴的人工智能算法,在烃类化合物的性质预测方面广泛地应用起来[4-8]。Wang等[9]构建了k-近邻算法、支持向量机、随机森林、提升树4种机器学习模型,预测了55种烃类混合物的燃烧下限;4种模型中,随机森林模型在测试数据上表现出最低的预测误差[9]。Yang等[10]测定了由12种纯烃类及不同配比的69种柴油的密度值,用神经网络和线性回归等模型有效关联了烃类化合物的质量百分比和密度。Guo等[11]在结构-特性关系的基础上建立了神经网络模型,有效预测了349种烃类和含氧化合物的十六烷值,预测精度优于多元线性回归。Liu等[12]比较了图卷积网络、图注意力网络、图同构网络三种图神经网络预测对燃料化合物闪点的精度,其中,耦合了分子和原子特征的图同构网络具有最佳的预测精度,预测闪点的平均绝对误差为3.952 K。在前期的研究中,机器学习模型预测烃类分子的性质部分来源于实验值,部分来源于基团贡献法、量子化学等经验理论方法的计算值,但针对实验测得或理论计算得到的性质的模型预测特点的一般性规律总结较少。

机器学习对烃类性质的有效预测可用于辅助燃料

收稿日期: 2024-10-28; **修回日期:** 2024-11-19

网络出版日期: 2024-12-20

基金项目: 国家自然科学基金(U2341278, 22178248);国家重点研发计划项目(2023YFlB4103000)

作者简介: 侯放(1995-),男,博士研究生,主要从事机器学习辅助的燃料理论设计研究。e-mail:tju_houfang@163.com

通信联系人: 李国柱(1982-),男,天津大学英才教授,主要从事先进碳氢燃料设计、合成与应用研究。e-mail:gzli@tju.edu.cn

引用本文: 侯放,齐晓宁,刘睿宸,等. 机器学习辅助的烃类分子性质预测与高通量筛选燃料[J]. 含能材料, DOI:10.11943/CJEM2024276.

HOU Fang,QI Xiao-ning,LIU Rui-chen,et al. Machine Learning Assisted Property Prediction of Hydrocarbon Molecules and High Throughput Screening for Fuel [J]. *Chinese Journal of Energetic Materials(Hanneng Cailiao)*, DOI:10.11943/CJEM2024276.

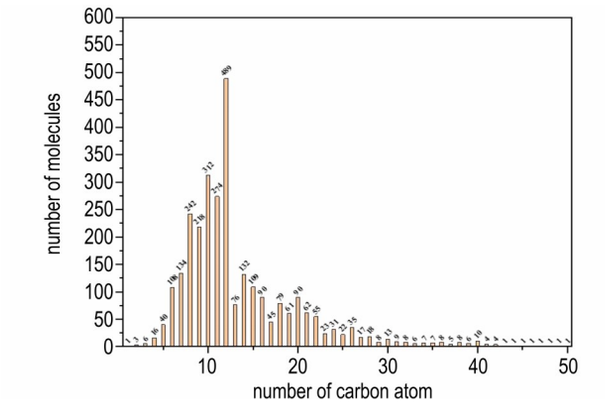
分子的筛选及设计，以加速了新型燃料的开发[12-14]。Li等[15-16]在准确预测多种燃料性质研究的基础上，提出了二级虚拟筛选特定性质燃料的方法，分别基于机器学习与定量结构-性能关系（ML-QSPR）模型及SI发动机的性能要求，最终筛选得到8种候选燃料分子。Zhou等[17]基于开源的ZINC20数据库，根据四环烷分子结构进行拓扑组装，得到了20种燃料分子结构；使用基团贡献法和密度泛函理论进一步评估分子的燃料特性，成功筛选出高体积热值、高比冲、低熔点的ZD-6分子结构。作者课题组也对机器学习辅助燃料性质预测进行了相关研究[18-21]，构建了多种神经网络模型准确预测了342种烃类子的密度、凝固点、热值等燃料性质，并成功筛选了一批性能突出的候选碳氢燃料分子。

基于以上背景，本研究通过构建以库伦矩阵为输入的多种机器学习模型，实现了对2899个烃类分子的3种实验测得的性质（熔点、沸点、密度）及4种理论计算得到的性质（0 K下的内能、298.15 K下的内能、298.15 K下的焓、298.15 K下的吉布斯自由能）较高精度地预测，比较得出了不同机器学习模型适用预测实验值或理论计算值等不同种类性质的一般性规律；最优的机器学习组合模型准确预测了细分数据集内的相同碳数、不同种类以及同分异构体烃类分子的多种性质，证明了模型具有较好的泛化能力；使用最优预测模型准确计算了34种已通过实验合成的高密度碳氢燃料的密度，与前期研究相比，模型基于燃料性质实验值训练得到，更具有实际应用价值；进一步对319，893个烃类分子的燃料性质进行预测，高通量筛选一批符合目标要求的候选碳氢燃料分子。

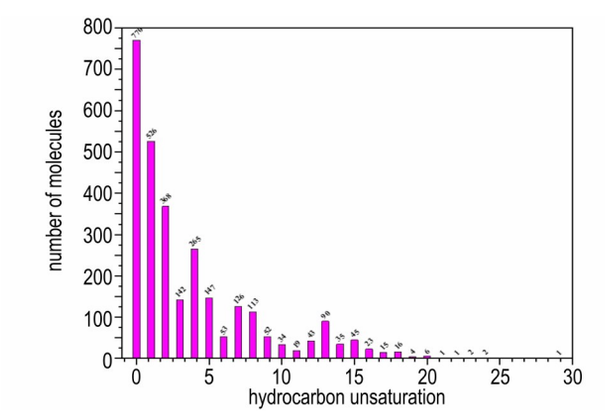
1 机器学习数据集和模型

1.1 数据集

从Yaws的《化合物性质手册》[22]选取碳原子数从1到50的全部烃类化合物分子，总计2899个。这2899个分子的碳原子数、氢不饱和度统计结果如图1所示。碳原子数为12的分子数目最多，共有489个，其次是碳原子数为10的分子，共有312个。氢不饱和度分布从0到29，其中有770个饱和烃类分子。上述分子结构包含了烷烃、烯烃、炔烃、链状烃、环状烃、芳香烃等多种碳氢化合物类型，足以涵盖基础烃类分子的所有结构特征。



a. carbon number



b. hydrogen unsaturation

图1 2899个烃类分子的碳数统及氢不饱和度统计

Fig. 1 (a) Statistical data of carbon number and hydrogen unsaturation for 2899 hydrocarbon molecules

Yaws《化合物性质手册》中提供了有机化合物的熔点（T_m）、沸点（T_b）和密度（ρ）的实验值。在上述提取的2899个烃类分子的性质数据库中，2520个分子具有熔点数据、2739个分子具有沸点数据、2899个分子具有密度数据。在DFT-B3LYP方法下，我们采用6-31G(d, p)基组通过高斯09软件进行了分子结构优化和频率计算。通过计算频率，所得优化结构均为分子势能面相对能量的最小值，且无虚频。基于结构优化，计算了2899个烃类分子的0 K下的内能（U_0）、298.15 K下的内能（U）、298.15 K下的焓（H）、298.15 K下的吉布斯自由能（G）。至此，建立了包含2899个烃类分子结构及其性质的数据集。具体分子结构和性质数值见支撑材料1。

为了进一步验证筛选出来的候选碳氢燃料分子的燃料性质，我们采用基团贡献法和DFT方法计算了烃类分子的质量热值（$NHOC$）和比冲（I_{sp}）。首先，使用基团贡献法[23]计算目标燃料分子在298 K下的标准蒸

发焓($\Delta_{vap}H^0$)；然后，配平等键反应方程，计算得到燃料分子的气相生成焓，气相生成焓与标准蒸发焓作差，进而得到其液相生成焓；最后，计算出燃料分子的燃烧热、质量热值和比冲性质。具体计算过程及公式如下。

(1)计算分子的标准蒸发焓

分子298 K下的标准蒸发焓$\Delta_{vap}H^0$(kJ·mol^{-1})计算公式见式(1)：

$$\Delta_{vap}H^0 - 11.733 = \sum_i n_i G_i \quad (1)$$

式中，n_i是基团类型i的数目，G_i是该基团对应该性质的贡献值。基团贡献法对应的参数详见支撑信息2中的附表1。

(2)配平等键反应方程

根据等键反应原则，方程两端反应物与生成物处于不同杂化状态的碳原子数量相等，方程两端反应物与生成物结合不同氢原子数(0，1，2，3)的C原子数量相等。反应方程式左端是目标燃料分子和乙烷，反应式右端是丙烷、异丁烷和新戊烷。根据等键反应原则和两端固定物质配平等键反应方程式。以四环庚烷(C_7H_8)为例，配平后的等键反应方程见式(Ⅰ)。

$$C_7H_8 + 10C_2H_6 = 1C_3H_8 + 6C_4H_{10} \quad (Ⅰ)$$

(3)计算分子的气相标准生成焓

结合等键反应方程式，计算方程式中各分子的298.15 K下的焓，从而计算等键反应的标准反应焓。下面以四环庚烷为例。通过计算，四环庚烷、乙烷、丙烷、异丁烷的298.15 K下的焓分别为：−271.323，−79.759，−119.046，−158.334 Hartree。乙烷、丙烷、异丁烷的标准生成焓分别为−84.76，−103.85，−134.52 kJ·mol^{-1}。通过差值计算得到目标燃料分子四环庚烷的气相标准生成焓。

(4)计算分子的液相标准生成焓

298K下的液相标准生成焓($\Delta_fH^0_{(l)}$)如式(2)：

$$\Delta_fH^0_{298(l)} = \Delta_fH^0_{298(g)} - \Delta_{vap}H^0_{298(l)} \quad (2)$$

式中，$\Delta_fH^0_{298(l)}$是物质的液相标准生成焓，$\Delta_fH^0_{298(g)}$是过程(3)中计算得到的气相标准生成焓，$\Delta_{vap}H^0_{298(l)}$是基团贡献法计算得到的标准蒸发焓。

(5)计算分子的标准燃烧热

298 K下的物质的标准燃烧热：

$$C_xH_{y(l)} + \left(x + \frac{y}{4}\right)O_{2(g)} \rightarrow xCO_{2(g)} + y2H_2O_{(g)} \quad (Ⅱ)$$

式中，目标化合物的$\Delta_fH^0_{298K}C_xH_{y(l)}$通过式(3)计算得出。二氧化碳和水蒸气的气相标准生成焓分别为$\Delta_fH^0_{298K}CO_{2(g)}$=−395.51 kJ·mol^{-1}，$\Delta_fH^0_{298K}H_2O_{(g)}$=−241.83 kJ·mol^{-1}。

(6)计算分子的质量热值

质量热值($NHOC$)：

$$NHCO = \Delta_fH^0_{298k}C_xH_{y(l)}/M \quad (3)$$

式中，M是分子的摩尔质量。

(7)计算分子的比冲

比冲(I_{sp})：

$$I_{sp} = (2 \times \eta NHOC)^{1/2}/g \quad (4)$$

式中，η是效率因子，多环烃类的η取0.556，g一般取9.8。

1.2 模型输入

本研究的机器学习使用了库仑矩阵[24-25](Coulomb matrix，CM)作为分子结构的数学表示。库伦矩阵包含分子结构和核电荷信息，用图2中所示的公式计算。其中，R_i是分子中原子的笛卡尔坐标，Z_i是原子的核电荷数，非对角位置代表不同原子i和j之间的库仑排斥，而对角位置的同一原子用多项式($0.5Z_i^{2.4}$)计算得到。首先将数据库中2899个烃类分子结构都转化成矩阵大小为200×200的二维库伦矩阵，其中，原子数不足200的分子的库伦矩阵通过填0

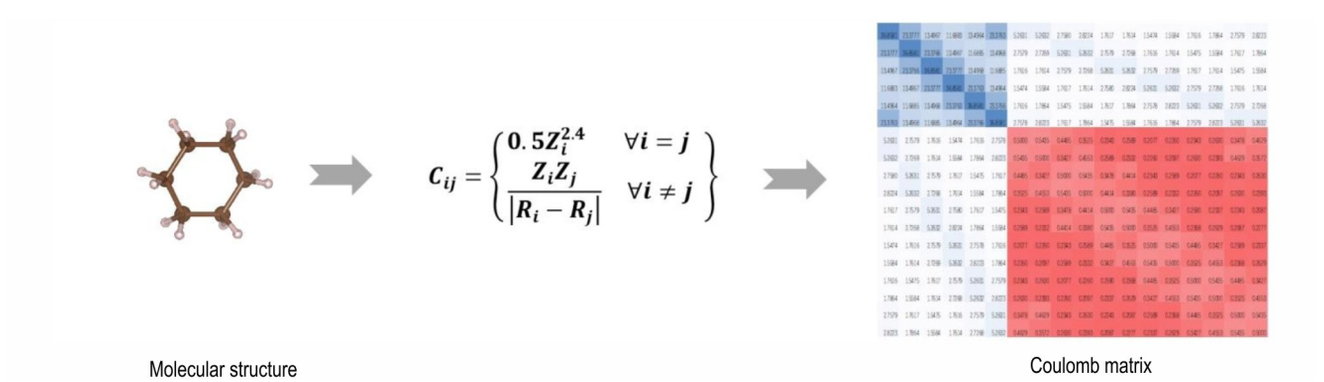


图2 分子结构的库伦矩阵表示

Fig.2 Coulomb matrix of molecular structures

扩展。最终，将获得的所有分子的二维库伦矩阵重构为一维矩阵特征值作为各个机器学习模型的输入。

1.3 模型建立及参数设置

研究构建了决策树回归模型（Decision Tree Regressor）、交叉验证的最小绝对收缩选择算子回归模型（Lasso CV）、交叉验证的岭回归模型（Ridge CV）、极限梯度提升回归模型（XGBoost Regressor）4种机器学习模型，用于预测2899个烃类分子的7种性质。其中，决策树回归模型通过scikit-learn机器学习库构建，以均方误差作为损失函数，每个分割节点通过“最佳”选择，树的最大深度为9。交叉验证的最小绝对收缩和选择算子回归模型通过scikit-learn机器学习库构建，采用5折交叉验证，沿正则化路径的正则化参数个数自动设置正则化参数，最大迭代次数为1000。交叉验证的岭回归模型通过scikit-learn机器学习库构建，正则化参数值的元祖为（0.1，1.0，10.0），采用5折交叉验证。极限梯度提升回归模型集成多个决策树，是一种基于提升方法增强策略的模型，训练时采用前向分布算法进行学习。模型中决策树之间是有先后顺序的，后一棵决策树的生成会考虑前一棵决策树的预测结果。同时，极限梯度提升回归模型对梯度提升模型进行了一系列优化，包括损失函数进行了二阶泰勒展开、目标函数加入正则项、支持并行和默认缺失值处理等。极限梯度提升回归模型的构建方法通过XGBoost开源库实现，针对数据库中不同的分子性质对模型进行分别训练，7种性质对应的模型参数见表1。

以上机器学习模型的训练集、验证集、测试集比例均分别设置为80%、10%、10%。通过训练优化，得到了具有最小损失函数的机器学习模型参数。通过计算预测值与真实值的决定系数（Coefficient of Determination，R^2）和平均绝对误差（absolute error，*MAE*）来评价模型预测性质的精度。两个评价指标的计算见式（5）和式（6）。其中，y_i代表真实值，$\widehat{y_i}$代表预测值，$Cov(y_i, \widehat{y_i})$代表两者的协方差，$Var\left[y_i\right]$代表真实值的方差，$Var[\widehat{y_i}]$代表预测值的方差。

$$R(y_i, \widehat{y_i}) = \frac{Cov(y_i, \widehat{y_i})}{\sqrt{Var\left[y_i\right]Var[\widehat{y_i}]}} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{6}$$

2 模型的验证

2.1 不同机器学习模型预测分子性质的精度比较

为了实现更准确地预测烃类性质，我们用决策树回归模型、交叉验证的最小绝对收缩选择算子模型、交叉验证的岭回归模型、极限梯度提升回归模型训练并预测了2899个烃类分子熔点、沸点、密度、0 K下的内能、298.15 K下的内能、298.15 K下的焓、298.15 K下的吉布斯自由能七种性质，计算了不同模型在训练集、验证集及测试集上的性质预测误差。通过比较不同模型预测不同性质在测试集上的误差，得到相应性质预测精度较高的模型。四种机器学习模型预测烃类7种性质在测试集上的平均绝对误差汇总结果见表2。

其中，极限梯度提升回归模型预测上述烃类分子的熔点、沸点、密度3种性质的精度最高。图3具体列出了极限梯度提升回归模型在训练集、验证集及测试集上预测烃类分子的熔点、沸点和密度3种性质的预测值与实验值的比较结果。从不同数据集上的性质预

表1 极限梯度提升回归模型的参数

Table 1 Parameters of XGBoost Regressor models

Model parameters	T_m / K	T_b / K	ρ / g·cm^{-3}	U_0 / Hartree	U / Hartree	H / Hartree	G / Hartree
learning_rate	0.2	0.03	0.03	0.1	0.1	0.1	0.1
n_estimators	400	900	950	900	900	900	900
max_depth	4	7	3	16	16	16	16
min_child_weight	4	4	6	4	4	4	4
subsample	0.8	0.7	0.9	0.9	0.9	0.9	0.9
colsample_bytree	0.7	0.9	0.8	0.6	0.6	0.6	0.6
gamma	0.1	0.1	0.1	0.1	0.1	0.1	0.1
reg_alpha	0.1	1	0.1	3	3	3	3
reg_lambda	0.05	0.1	1	1	1	1	1

Note: T_m, T_b, ρ, U_0, U, H and G respectively indicate melting point, boiling point, density, internal energy at 0 K, internal energy at 298.15 K, enthalpy at 298.15 K and Gibbs free energy at 298.15 K.

表 2 不同机器学习预测 2899 个烃类分子不同性质在测试集上的平均绝对误差(*MAEs*)

Table 2 *MAE*s of predicting different properties of 2899 hydrocarbon molecules via different machine learning methods on the test set

models	T_m / K	T_b / K	ρ / g·cm^{-3}	U_0 / Hartree	U / Hartree	H / Hartree	G / Hartree
Decision Tree Regressor	24.0925	15.2306	0.0266	0.7512	0.9946	1.0476	0.8138
Lasso CV	27.6617	11.8971	0.0303	4.3891	4.3885	3.2686	4.3900
Ridge CV	37.1347	10.6289	0.0277	0.1380	0.1375	0.1375	0.1390
XGBoost Regressor	19.9442	9.0034	0.0258	0.4964	0.5540	0.5540	0.5130

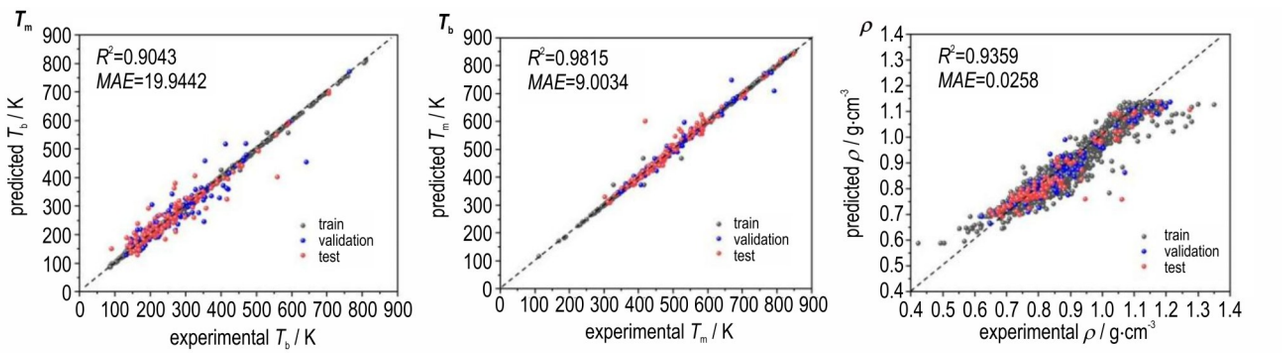


图 3 极限梯度提升回归模型预测不同烃类分子性质(T_m、T_b、ρ)的预测值与实验值比较结果(图中预测误差是模型在测试集上的误差)

Fig.3 Comparison of predicted and experimental values for different properties (T_m, T_b, ρ) by XGBoost Regressor models (the prediction error in the figure is on the test set)

测精度看,无论是在训练集、验证集还是在测试集上,模型均表现出了较高精度。这是由于数据库中烃类分子的熔点、沸点、密度性质取自 Yaws 的《化合物性质手册》,是通过多种来源获取到的实验值;而极限梯度提升回归模型使用的二阶导数有利于梯度下降得更快、更准,使用泰勒展开在不选定损失函数具体形式的情况下,仅依靠输入数据值即可进行叶子分裂优化计算,本质上是将损失函数的选取和模型参数优化分开。这种去耦合的方法增加了模型的适用性,能够在广泛的任务上取得良好的训练效果,从而更适用于处理不同数据源的真实数据,有效克服不同数据标准的干扰。因此,以库伦矩阵为输入的极限梯度提升回归模型后续被用于预测烃类分子的熔点、沸点和密度性质。

从预测误差结果可以看出,交叉验证的岭回归模型(Ridge CV)和极限梯度提升回归模型(XGBoost Regressor)两种模型都可以准确地预测烃类分子的 0 K下的内能(U_0)、298.15 K下的内能(U)、298.15 K下的焓(H)和 298.15 K下的吉布斯自由能(G)四种能量性质,且两种模型在训练集、验证集、测试集上均有着较高的决定系数,R^2均在 0.9999 以上。进一步对比两种模型预测四种能量性质的平均绝对误差,与极限梯度提升回归模型相比,交叉验证的岭回归模型预测四种能量性质的平均绝对误差较低,其在测试集上预测四种能量性质的平均绝对误差较极限梯度提升回归模型平均降低了 73.9%。这是由于上述四种能量性质均是通过 DFT 理论计算得到的,数据标准统一;而岭回归模型预测的优势就在于无偏估计。岭回归模型实质上是一种改良的最小二乘估计法,通过放弃最小二乘法的无偏性,以损失部分信息、降低精度为代价使回归系数更为符合实际,从而可以缓解多重共线问题,更适用于处理标准统一的数据。图 4 具体列出了交叉验证的岭回归模型在训练集、验证集及测试集上预测烃类分子的 0 K下的内能、298.15 K下的内能、298.15 K下的焓和 298.15 K下的吉布斯自由能四种能量性质的预测值与实验值的对比结果。

此外,其余模型预测不同性质的预测值与真实值比较结果分别见支撑材料 2 中的附图 1~4,模型在不同数据集上的预测误差汇总结果见附表 2~9。

2.2 细分数据集内的性质预测分析

将预测熔点、沸点、密度性质的极限梯度提升回归模型和预测 0 K下的内能、298.15 K下的内能、298.15 K下的焓和 298.15 K下的吉布斯自由能性质的交叉验证的岭回归模型作为最优机器学习组合模型,预测数据集中相同碳数以及不同种类烃类分子的

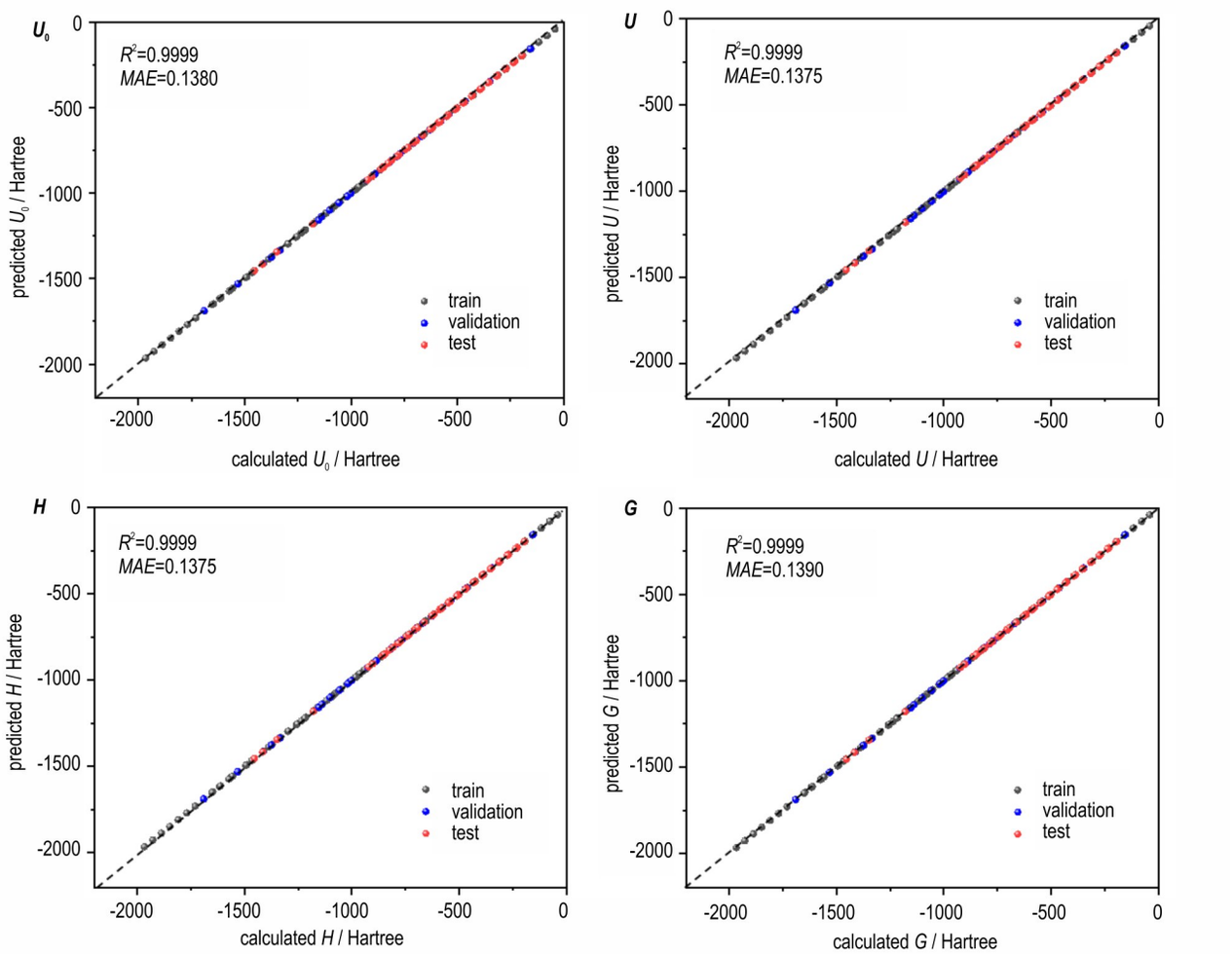


图 4 交叉验证的岭回归模型预测不同烃类分子性质（U_0、U、H、G）的预测值与理论计算值比较结果（图中预测误差是模型在测试集上的误差）

Fig.4 Comparison of predicted and experimental values for different properties (U_0, U, H, G) by Ridge CV models (the prediction error in the figure is on the test set)

多种性质。选取分子数目较多的同碳数分子进行验证，分别是C10（312个分子）、C11（274个分子）和C12（489个分子）。三种相同碳数分子的七种性质预测误差汇总结果见表4。链烷烃（770个分子）、环烷烃（311个分子）、烯烃（511个分子）、炔烃（124个分子）和芳香烃（690个分子）的七种性质预测误差汇总结果见表5。从表3和表4的预测结果可以看出，相同碳数以及不同种类烃类分子的性质预测精度与模型在测试集上的预测精度相近。这说明所得到的最优机器学习组合模型可以很好地应用到细分类数据集，具有较高的预测精度。

进一步将上述最优机器学习组合模型应用到同分异构体的性质预测。使用的库伦矩阵的分子结构表示方法可以准确描述分子结构，进而从机器学习模型的输入端实现精准区分同分异构体的不同分子构型。我们选取了数据集中分子数量前五的同分异构体烃分子

表 3 机器学习预测不同碳数分子（C10、C11和C12）的性质的平均绝对误差（*MAE*s）

Table 3 *MAE*s of predicting different properties of the fuel molecules with different carbon atoms (C10, C11 and C12) by machine learning

	T_m / K	T_b / K	ρ / g·cm^{-3}	U_0 / Hartree	U / Hartree	H / Hartree	G / Hartree
C10	5.6006	1.8174	0.0263	0.0779	0.0778	0.0778	0.0782
C11	4.1873	1.7865	0.0199	0.0568	0.0568	0.0568	0.0570
C12	3.9479	1.4238	0.0198	0.0332	0.0031	0.0331	0.0334

表 4 机器学习预测不同种类分子的性质的平均绝对误差(*MAEs*)

Table4 *MAEs* of predicting different properties of the fuel molecules with different types

	T_m / K	T_b / K	ρ / g·cm^{-3}	U_0 / Hartree	U / Hartree	H / Hartree	G / Hartree
Paraffin hydrocarbon	3.6359	1.4051	0.0203	0.0215	0.0214	0.0214	0.0217
Cycloalkanes	6.2300	2.0715	0.0273	0.0562	0.0562	0.0561	0.0562
Olefin	4.7784	2.1273	0.0285	0.0314	0.0314	0.0314	0.0316
Alkyne	5.5570	1.7478	0.0272	0.0491	0.0488	0.0488	0.0498
Aromatic hydrocarbon	14.4088	2.7554	0.0272	0.1741	0.1739	0.1739	0.1743

进行验证,分别是$C_{12}H_{26}$(355个分子)、$C_{11}H_{24}$(158个分子)、C_8H_{16}(131个分子)、C_9H_{18}(96个分子)和$C_{10}H_{22}$(75个分子)。如图5所示,上述五种同分异构体烃分子的熔点、沸点、密度、0 K下的内能、298.15 K下的内能、298.15 K下的焓、298.15 K下的吉布斯自由能7种性质预测误差均小于在最优机器学习组合模型在测试集上的预测误差。从预测误差比较结果可以看出,优化的机器学习模型已经能够比较准确地预测同分异构体的不同性质。

3 机器学习模型的性能预测研究

3.1 高密度碳氢燃料的密度预测

研究基于2899个烃类分子结构及其性质的数据库训练得到的多极限梯度提升回归模型,计算了34种已通过实验合成的高密度碳氢燃料的密度。其中,联环庚烷、联环己烷等10种碳氢燃料分子已经出现在机器学习模型训练数据集中。34种高密度碳氢燃料分子的结构见图6。密度是评估碳氢燃料综合性能的重

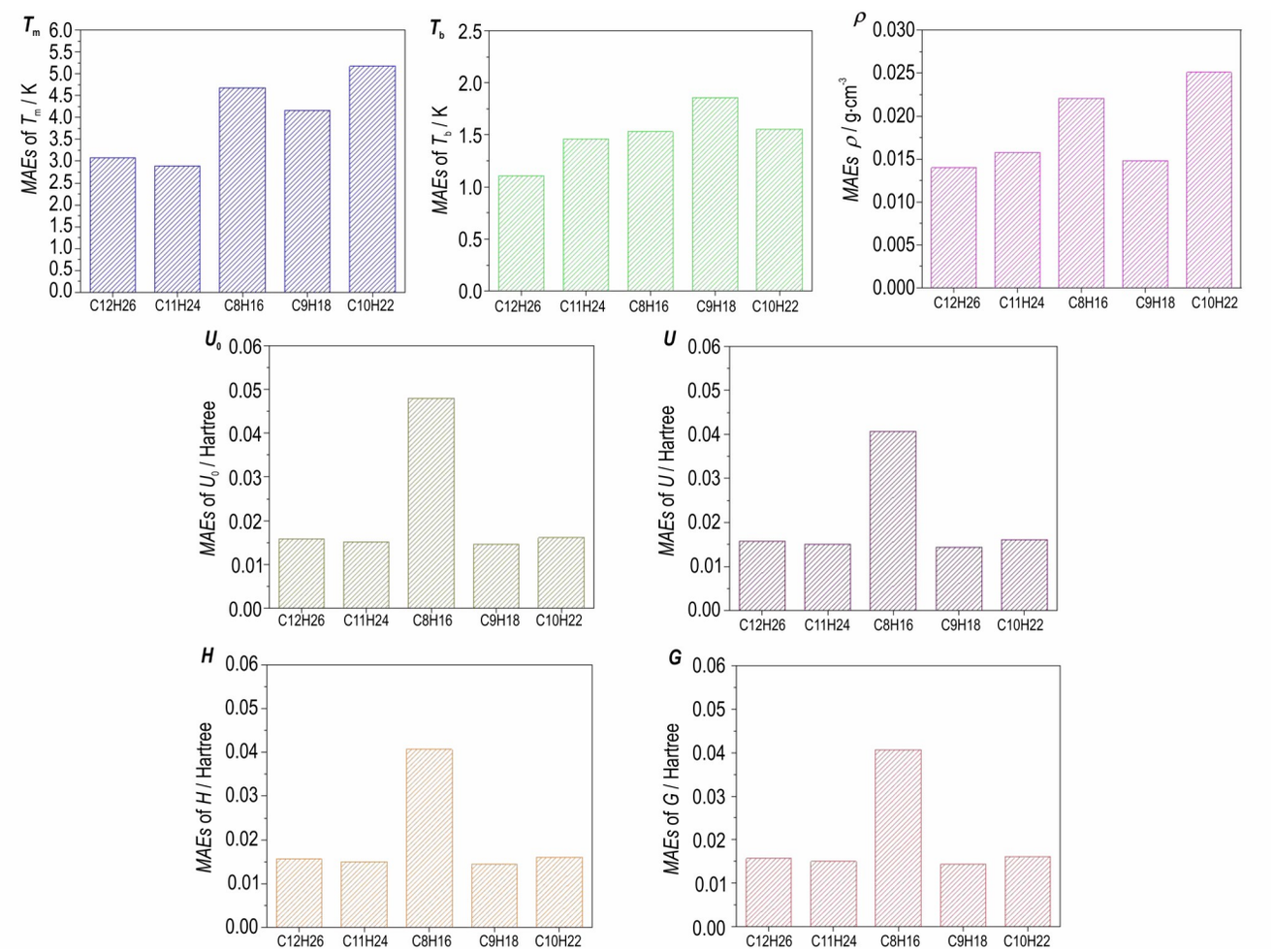


图 5 预测同分异构体的不同性质的误差结果比较

Fig.5 Comparison of predicting different properties for different isomers

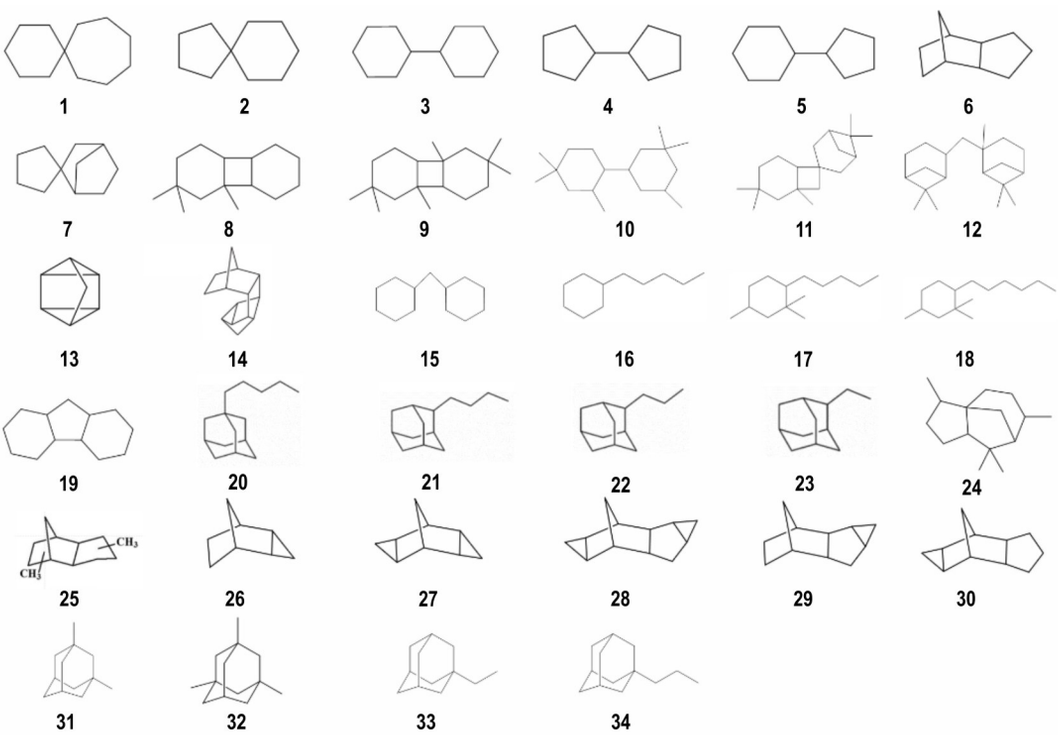


图6 34种碳氢燃料分子的结构(序号1,2,3,4,5,6,15,16,19,31分子已经出现在机器学习模型训练数据集)

Fig.6 Molecular structures of 34 hydrocarbon fuel molecules (No.1, 2, 3, 4, 5, 6, 15, 16, 19, 31 molecules appeared in the neural network model training dataset)

要指标之一，如液体碳氢燃料的密度与其体积热值呈正相关[26]。图6所示的这些燃料的密度值均已通过实验测得，可用于进一步评估机器学习模型的预测精度。34种碳氢燃料密度的机器学习模型的计算值和实验值的比较结果见图7。通过进一步计算，34种碳氢燃料密度的预测值和实验值之间的平均绝对误差为0.0290 g·cm^{-3}，除去模型训练集包含的10种碳氢燃料分子，训练数据集以外的24种碳氢燃料密度的计算值和实验值之间的平均绝对误差为0.0433 g·cm^{-3}。上述预测精度与极限梯度提升回归模型在测试集上的预测精度相近，证明了训练得到的机器学习模型较好地学习到了烃类分子的结构和性质信息，可以比较准确地扩展应用到训练数据集之外的烃类分子的性质预测，在烃类分子性质预测上具有较好的泛化能力。与前期工作相比，本研究的模型是基于2899个烃类分子的密度实验值训练得到的，预测结果更贴近于实验值，模型更具有实际应用价值。未来，设计优化的机器学习模型在高密度碳氢燃料、推进剂等烃类化合物实际应用领域具有重要应用价值。

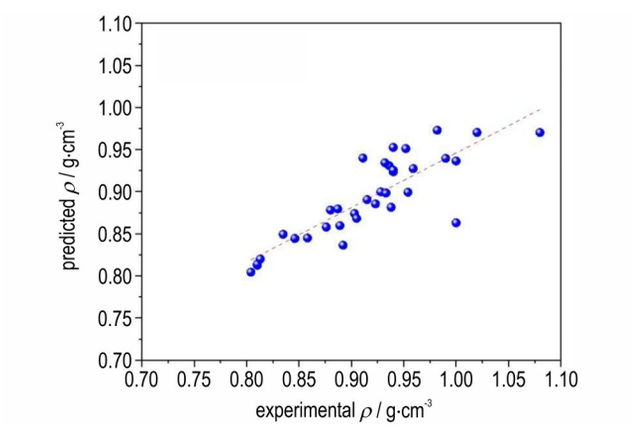


图7 34种碳氢燃料密度的预测值和实验值比较

Fig.7 Comparison of predicted and experimental values of density for 34 hydrocarbon fuel molecules

3.2 大体量分子结构-性质数据库构建及高通量筛选

2.1节中训练优化得到的机器学习模型，可以有效运用到训练集以外的烃类分子结构数据库上，快速地实现对其分子性质的预测。进而，可以构建包含大量分子的结构-性质的大体量数据库，便于后续用特定性质需求标准进行高通量筛选。GDB-13数据集包含977,468,314个有机小分子的结构，分子中包含若干C、H、N、O、S、Cl原子，分子中的总原子数均小于等于13。我们通过编写代码程序从GDB-13中筛选出饱和烃类分子，获得了319,893个仅包含碳和氢原子且没有任何不饱和键(双键和三键)的分子结构数据库，将其命名为GDB-13C。然后将GDB-13C数据库中的

319,893个分子结构的SMILE式转化为库伦矩阵特征值,输入到优化后的极限梯度提升回归模型中进行性质预测,快速得到了GDB-13C中所有分子的熔点、沸点、密度性质数据。GDB-13C中的319,893个烃类分子熔点、沸点、密度性质分布见图8,具体性质预测数据见支撑材料3。

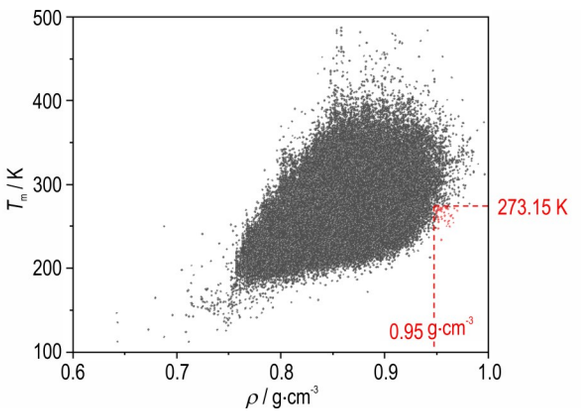


图8 319,893个分子的熔点、密度预测性质(筛选出的分子用红色表示)

Fig. 8 Predicted property values and screening criteria for melting points and densities of 319,893 molecules

基于上述得到的"分子结构-性质"数据库,可以根据特定燃料需求设定阈值,开展高通量筛选,获得所需要的分子。具体筛选过程是通过更改相关性质的阈值,将阈值设置为达到或超过某些性质最高值的某一百分比或某一特定值,从而获得具有不同性能的各种候选分子。为发现潜在的高密度液体碳氢燃料候选分子,将筛选标准设置为熔点低于273.15 K且密度高于0.95 g·cm^{-3},最终在319,893个分子中筛选出37种满足要求的分子结构。筛选出来的37种分子的数据分布情况见图8中红色数据点,分子的编号和结构见图9。对上述筛选出来的37个候选烃类燃料分子进行结构分析。37个分子中12个具有4个碳环(图9中的i-ii组),25个分子具有5个碳环(图9中的iii-vii组)。在这些分子结构中构成最多的碳环是五元环,五元环在91.2%分子中出现。此外,86.5%的分子具有三元环,54.1%、40.5%和5.4%的分子分别具有四元、六元和七元环。对两类分子(5个碳环和4个碳环结构)的环组成

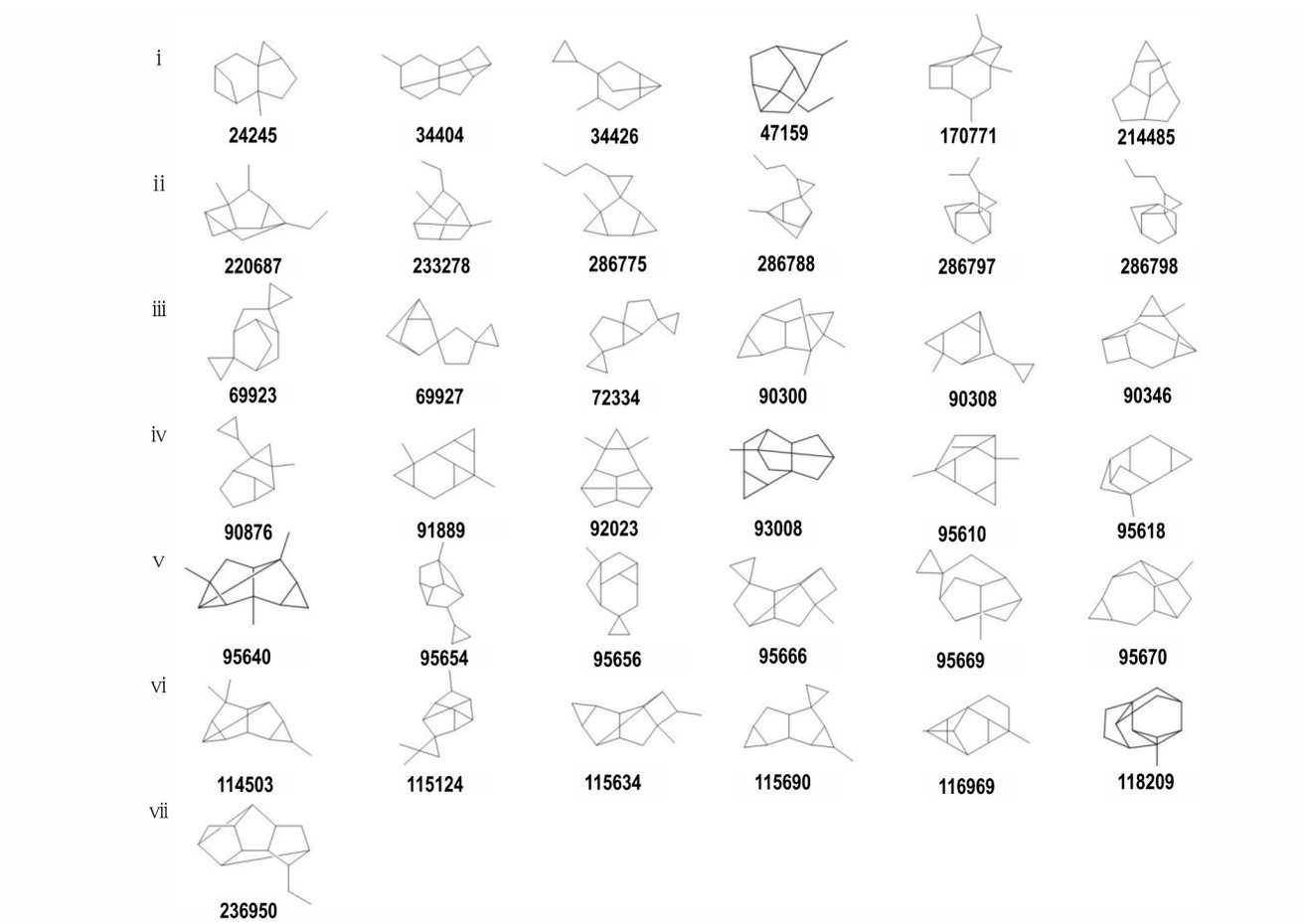


图9 筛选出来的37种烃类分子结构

Fig.9 Molecular structures of the as-screened 37 hydrocarbon molecules

也进行了具体分析，如图10所示。在4个碳环分子结构中，三元、四元、五元和六元环的含量分别为22.9%、29.2%、35.4%和12.5%。在5个碳环分子结构中，三元、四元、五元、六元和七元环的含量分别为36.8%、11.2%、40.8%、9.6%和1.6%。除了上述独特的空间结构外，筛选出来的新型候选燃料结构一个显著特征是三元环的存在，尤其是螺三元环，这很可能是设计新的碳氢燃料或改善现有燃料性能的通用设计手段。

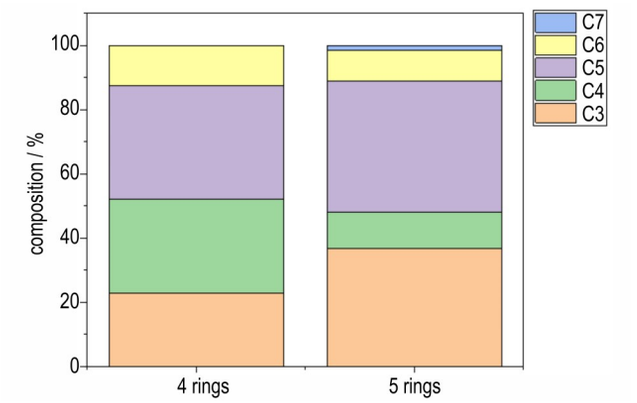


图10 37个烃类分子的环组成，C_x(x=3、4、5、6、7)是指x元碳环

Fig.10 Ring composition of the 37 hydrocarbon molecules C_x(x=3, 4, 5, 6, 7) means x-membered cyclic carbon ring

此外，通过基团贡献法和DFT方法计算得到上述37种烃类分子的质量热值(*NHOC*)和比冲(I_{sp})，并与传统燃料JP-10和QC的性质进行了比较，如表5所示。通过比较可以得出，筛选出来的分子在不同的燃料性质方面表现出各自的优势，可以指导后续实验合成新型高密度、低凝固点、高热值、高比冲的碳氢燃料。

4 结 论

研究以分子库伦矩阵作为输入，设计并优化了决策树回归模型、交叉验证的最小绝对收缩和选择算子回归模型、交叉验证的岭回归模型、极限梯度提升回归模型4种不同的机器学习模型，实现了对碳数从1到50的2899个烃类分子的熔点、沸点、密度、0 K下的内能、298.15 K下的内能、298.15 K下的焓、298.15 K下的吉布斯自由能7种性质的准确预测。比较不同性质的预测精度得出如下结论：

(1)极限梯度提升回归模型可以较高精度地预测烃类分子的熔点、沸点、密度等通过实验测得的性质，交叉验证的岭回归模型可以较高精度地预测烃类分子

表5 高通量筛选出来的37种烃类分子与典型燃料JP-10和QC的性质比较

Table 5 Calculated properties of 37 hydrocarbon molecules discovered by high-throughput screening and traditional fuels of JP-10 and QC

No.	T_m / K	ρ / g·cm^{-3}	*NHOC* / MJ kg^{-1}	I_{sp} / m·s^{-1}
24245	271.3	0.9593	42.56	340.3
34404	268.6	0.9698	44.68	348.7
34426	260.1	0.9655	42.55	340.2
47159	255.3	0.9518	45.55	352.0
69923	256.7	0.9579	42.13	340.2
69927	252.8	0.9609	42.10	340.1
72334	270.7	0.9610	42.41	341.4
90300	263.6	0.9642	43.99	347.6
90308	261.1	0.9511	42.38	341.2
90346	272.3	0.9585	46.31	356.7
90876	258.1	0.9516	44.50	349.6
91889	261.3	0.9544	44.93	351.3
92023	268.9	0.9513	43.39	345.3
93008	265.2	0.9615	42.96	343.5
95610	262.2	0.9633	46.22	354.1
95618	262.8	0.9578	42.80	342.9
95640	254.6	0.9660	42.83	340.8
95654	256.7	0.9619	42.40	341.3
95656	250.5	0.9594	43.54	345.8
95666	249.8	0.9625	41.94	339.4
95669	266.5	0.9573	42.10	340.1
95670	255.4	0.9609	44.20	348.4
114503	272.2	0.9538	41.94	339.4
115124	270.6	0.9549	42.42	341.4
115631	268.2	0.9500	42.36	341.1
115690	271.6	0.9668	43.67	346.4
116969	272.5	0.9553	43.28	344.8
118209	267.8	0.9546	45.18	352.3
170771	271.1	0.9537	44.46	347.3
214485	269.9	0.9523	42.07	337.8
220687	247.5	0.9515	42.33	338.8
233278	271.8	0.9512	42.41	339.2
236950	265.9	0.9500	41.80	338.9
286775	234.0	0.9544	44.49	347.4
286788	263.6	0.9502	42.55	339.7
286797	259.0	0.9550	42.65	340.1
286798	268.7	0.9536	42.63	340.1
JP-10	269.3	1.04	42.17	337.4
QC	230.0	1.09	43.02	347.4

的内能、焓、吉布斯自由能等能量的通过理论计算得到的性质。

(2)最优机器学习组合模型可以准确预测相同碳数、不同种类和同分异构体烃分子的性质,验证了模型在细分数据集上同样有效。

(3)将优化的极限梯度提升回归模型应用到训练数据集以外的分子结构库上,准确预测了34种实验已合成的高密度碳氢燃料的密度,平均绝对误差为0.0290 g·cm^{-3}。进而,使用优化的极限梯度提升回归模型预测了GDB-13C中319,893个烃类分子的熔点、沸点、密度性质预测,建立了大体量的燃料"分子结构-性质"数据库。通过设定燃料性质阈值,快速筛选出了37种低凝固点、高密度的烃类分子。这些筛选出的新型分子与传统碳氢燃料JP-10和QC质量热值和比冲相当,可以作为潜在的高能碳氢燃料候选者。

综上所述,基于基础烃类性质数据库训练优化的机器学习模型,可以有效地扩展应用到训练数据集以外的更多烃类分子,并高通量筛选特定性能的燃料分子,在碳氢燃料等领域具有重要应用价值。

参考文献:

[1] 余锐,刘显龙,史成香.高能碳氢燃料绿色合成技术研究进展[J].含能材料, 2022.
YU Rui, LIU Xian-long, SHI Cheng-xiang. Review on green synthesis of high-energy-density hydrocarbon fuel[J]. *Chinese Journal of Energetic Materials*(*Hanneng Cailiao*), 2022.

[2] 刘宁,史成香,潘伦.生物质替代石油原料合成高密度燃料的研究进展[J].燃料化学学报, 2021, 49(12): 1780-1790.
LIU Ning, SHI Cheng-xiang, PAN Lun. Progress on using biomass derivatives to replace petroleum for synthesis of high-density fuels[J]. *Journal of Fuel Chemistry and Technology*, 2021, 49(12): 1780-1790.

[3] ZHANG X, PAN L, WANG L. Review on synthesis and properties of high-energy-density liquid fuels: Hydrocarbons, nanofluids and energetic ionic liquids[J]. *Chemical Engineering Science*, 2018, 180: 95-125.

[4] NGUYEN P, LOVELAND D, KIN J T. Predicting energetics materials' crystalline density from chemical structure by machine learning[J].*J Chem Inf Model*, 2021,61(5):2147-2158.

[5] SUN X,ZHANG F,LIU J.Prediction of gasoline research octane number using multiple feature machine learning models[J]. *Fuel*, 2023, 333.

[6] PANWAR P, YANG Q, MARTINI A. Temperature-dependent density and viscosity prediction for hydrocarbons: Machine learning and molecular dynamics simulations[J]. *J Chem Inf Model*, 2023.

[7] JORGENSEN P B, SCHMIDT M N, WINTHER O. Deep generative models for molecular science[J]. *Mol Inform*, 2018, 37(1-2).

[8] PILANIA G, WANG C, JIANG X. Accelerating materials property predictions using machine learning[J]. *Sci Rep*, 2013,3: 2810.

[9] JIAO Z, YUAN S, ZHANG Z. Machine learning prediction of hydrocarbon mixture lower flammability limits using quantitative structure-property relationship models[J]. *Process Safety Progress*, 2019, 39(2).

[10] YANG HONG Z R, YEVGNIA BREKER. Neural network prediction of cetane number and density of diesel fuel form its chemical composition determined by LC and GC-MS[J]. *Fuel*, 2002, 81: 65-74.

[11] GUO Z, LIN K H, CHEN M.Predicting cetane numbers of hydrocarbons and oxygenates from highly accessible descriptors by using artificial neural networks [J]. *Fuel*, 2017, 207: 344-351.

[12] LIU J, GONG S, LI H. Molecular graph-based deep learning method for predicting multiple physical properties of alternative fuel components[J]. *Fuel*, 2022, 313.

[13] CLEMENS Hall B R, UWE BAUDER, MANFRED AIGNER. Comparison of probabilistic jet fuel property models for the fuel screening and design[J]. *Fuel*, 2023, 351.

[14] KUZHAGALIYEVA N, HORVATH S, WILLIAMS J. Artificial intelligence-driven design of fuel mixtures [J]. *Commun Chem*, 2022, 5(1): 111.

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

[16] LI R, HERREROS J M, TSOLAKIS A. Integrated machine learning-quantitative structure property relationship (ML-QSPR) and chemical kinetics for high throughput fuel screening toward internal combustion engine[J]. *Fuel*, 2022, 307.

[17] WEN L, SHAN S, LAI W. Accelerating the design of high-energy-density hydrocarbon fuels by learning from the data[J]. *Molecules*, 2023, 28(21).

[18] LI G, HU Z, HOU F. Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels[J]. *Fuel*, 2020, 265.

[19] LIU R, LIU R, LIU Y. Design of fuel molecules based on variational autoencoder[J]. *Fuel*, 2022, 316.

[20] LIU R, LIU Y, DUAN J. Ensemble learning directed classification and regression of hydrocarbon fuels[J]. *Fuel*, 2022, 324.

[21] LIU Y, LIU R, DUAN J. Deep generative fuel design in low data regimes via multi-objective imitation[J]. *Chemical Engineering Science*, 2023, 274.

[22] YAWS C L. Physical Properties-Organic Compounds[M]. The Yaws Handbook of Physical Properties for Hydrocarbons and Chemicals. 2015: 1-683.

[23] HUKKERIKAR A S, SARUP B, TEN KATE A. Group-contribution+(GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis[J]. *Fluid Phase Equilibria*, 2012, 321: 25-43.

[24] HOU F, WU Z, HU Z. Comparison study on the prediction of multiple molecular properties by various neural networks[J]. *J Phys Chem A*, 2018, 122(46): 9128-9134.

[25] HANSEN K, MONTAVON G, BIEGLER F. Assessment and validation of machine learning methods for predicting molecular atomization energies[J]. *J Chem Theory Comput*, 2013, 9(8): 3404-3419.

[26] 熊中强,米镇涛,张香文.合成高密度烃类燃料研究进展[J].化学进展, 2005, 17: 359-367.
XIONG Zhong-qiang, MI Zhen-tao, ZHANG Xiang-wen. Devlopment of synthesized high-densiyt hydrocarbon fuel[J]. *Progress in Chemmistry*, 2005, 17: 359-367.

Machine Learning Assisted Property Prediction of Hydrocarbon Molecules and High Throughput Screening for Fuel

HOU Fang[1], **QI Xiao-ning**[4,5], **LIU Rui-chen**[1], **LI Ling**[6], **WANG Li**[1,2,3], **ZHANG Xiang-wen**[1,2,3], **LI Guo-zhu**[1,2,3]

(1. *School of Chemical Engineering and Technology, Tianjin University, Tianjin* 300072, *China*; 2. *Key Laboratory for Advanced Fuel and Chemical Propellant of Ministry of Education, Tianjin* 300072, *China*; 3. *Haihe Laboratory of Green Creation and Manufacture of Matter, Tianjin* 300192, *China*; 4. *Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing* 100910, *China*; 5. *University of Chinese Academy of Sciences, Beijing* 100910, *China*; 6. *Chengde Vanadium Titanium New Material Co., Ltd., Chengde* 067102, *China*)

Abstract: A big database containing molecular structures and multiple properties of 2899 hydrocarbon molecules (the number of carbon atom is from 1 to 50), was constructed via data collection, structure optimization and quantum chemistry calculation. Seven properties were focused, including melting point (T_m), boiling point (T_b), density (ρ), internal energy at 0 K (U_0), internal energy at 298.15 K (U), enthalpy at 298.15 K (H) and Gibbs free energy at 298.15 K (G). Four different machine learning models were established, including Decision Tree Regressor, Lasso CV, Ridge CV and XGBoost Regressor, using coulomb matrix representing molecular structures as the input. In comparison, the XGBoost Regressor model is more suitable for regressing experimental melting point, boiling point and density of hydrocarbon molecules; Ridge CV model is more suitable for the prediction of four thermodynamic energy properties. In addition, the optimized machine learning combined model can accurately predict the properties of the hydrocarbon molecules with same carbon numbers, hydrocarbons with different types and hydrocarbon isomers. Furthermore, the densities of 34 high-density hydrocarbon fuels reported experimentally were calculated by the optimized machine learning model. The mean absolute error between the calculated values and the experimental values is only 0.0290 g cm^{-3}. Next, the fuel properties of 319,893 hydrocarbon molecules in GDB-13C were predicted by the machine learning model to establish a big database containing hydrocarbon structure and fuel properties. Based on high-throughput screening, 37 hydrocarbon fuel molecules with low freezing point and high density have been discovered. Through the proof-of-concept via group contribution method and DFT method, the net heat of combustion and specific impulse of the as-screened new molecules are similar to those of JP-10 and quadricyclane (QC).

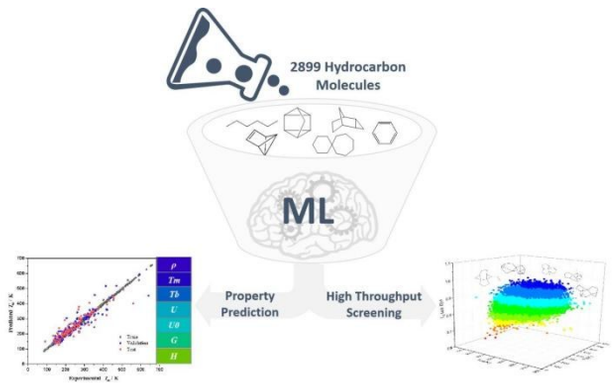
Key words: machine learning; hydrocarbon molecule; high-density hydrocarbon fuel; property prediction; high throughput screening

CLC number: TJ5;O64;V31 **Document code**: A **DOI**: 10.11943/CJEM2024276

Grant support: National Natural Science Foundation of China (No.22178248); National Key Research and Development Program Project (2023YFlB4103000)

(责编: 姜 梅)

图文摘要：



In this study, different machine learning models have been developed to accurately and quickly predict multiple properties of 2899 hydrocarbon molecules, and some general rules of machine learning models were obtained by comparing the prediction accuracy for different types of properties, such as experimental values and theoretical calculated values. The optimized machine learning model screens 319,893 hydrocarbon molecules using the key properties of fuel as the threshold values, and a group of highly potent hydrocarbon molecules as the next generation fuelswas identified. Additionally, this strategy can also be extended to discover other new molecules, e.g., lubricants, additives, and explosives.