高等学校化学学报
CHEMICAL JOURNAL OF CHINESE UNIVERSITIES
研究论文

基于机器学习方法的含能材料分解温度预测

郭莉莉[1]，户梦倩[1]，黄孟梅[1]，卢艳华[2]，
李　伟[2]，张庆友[1]，郭　翔[2]
（1. 河南大学河南省工业循环水处理工程研究中心，
河南省环境污染防治材料国际联合实验室，开封 475004；
2. 湖北航天化学技术研究所航天化学动力实验室，襄阳 441003）

摘要　提出一种新型复合描述符系统，构建了一个高性能的含能材料分解温度预测模型. 首先根据基团贡献理论提出了分子结构描述符，随后引入键解离能(BDE)作为关键补充参数，量化化学键强度对热稳定性的影响，最后通过RDKit软件生成了RDKit描述符，并将这3类描述符整合构建为一个多维特征集. 将该特征集分别提交给随机森林(RF)、支持向量机(SVM)及偏最小二乘(PLS)构建多个预测模型，并进行了系统性比较. 其中，使用随机森林构建的模型取得了最佳结果，其预测性能优于文献报道结果，表明所提出的复合描述符能够有效地捕捉影响分解温度的关键因素. 此外，借助沙普利加法解释(SHAP)可视化技术对最优模型进行了解析，为理解含能材料的热稳定性机理提供了有价值的数据洞察.

关键词　分解温度；分子基团描述符；定量构效关系；随机森林
中图分类号　O641；TJ55　　**文献标志码**　A　　doi：10.7503/cjcu20250297

Prediction of Decomposition Temperature of Energetic Materials Based on Machine Learning Methods

GUO Lili[1], HU Mengqian[1], HUANG Mengmei[1], LU Yanhua[2],
LI Wei[2], ZHANG Qingyou[1*], GUO Xiang[2*]
(*1. Henan Engineering Research Center of Industrial Circulating Water Treatment, Henan Joint International Research Laboratory of Environmental Pollution Control Materials, Henan University, Kaifeng* 475004, *China*;
2. Science and Technology on Aerospace Chemical Power Laboratory, Hubei Institute of Aerospace Chemotechnology, Xiangyang 441003, *China*)

Abstract　In this study, a novel composite descriptor system was suggested, and a high-performance prediction model for the decomposition temperature of energetic materials was constructed. First, molecular structure descriptors based on group contribution were proposed. Then, bond dissociation energy(BDE) was introduced as a key supplementary parameter to quantify the effect of bond strength on decomposition temperature. Finally, RDKit descriptors were generated using the RDKit software, ultimately integrating them into a multidimensional feature set. This feature set was submitted to random forest(RF), support vector machine(SVM), and partial least squares

收稿日期：2025-10-15. 网络首发日期：2025-12-14.
联系人简介：张庆友，男，博士，教授，主要从事化学信息学与化学计量学方面的研究. E-mail：zhqingyou@163.com
郭　翔，男，博士，研究员，主要从事固体推进剂和航天化学动力方面的研究. E-mail：guoxiang@casc42.cn
基金项目：中国航天化学动力实验室开放研究基金(批准号：120201B01)资助.
Supported by the Open Research Fund of the Science and Technology on Aerospace Chemical Power Laboratory, China(No.120201B01).

(PLS) individually to construct multiple prediction models and conduct a systematic comparison. Among them, the best results were obtained using the model built by RF. Its prediction performance was superior to the results reported in the literature, indicating that the proposed composite descriptors can effectively capture the key factors affecting the decomposition temperature. To further interpret the model and identify critical influencing factors, Shapley additive explanations(SHAP) visualization technology was employed to analyze the optimal model, thereby providing valuable data-driven insights into the thermal stability mechanisms of energetic materials.

Keywords Decomposition temperature; Molecular group descriptor; Quantitative structure-property relationship (QSPR); Random forest(RF)

含能材料是一类能够通过自身快速化学反应(通常是氧化还原反应)释放大量能量和气体产物的特殊物质,具有良好的热稳定性、爆速性能、机械敏感性等优异性能[1~4],而热分解温度是评估含能材料热稳定性的关键参数,也是直接影响其安全性、储存寿命和适用性的重要指标[5~7].

当前,含能材料的开发仍高度依赖试错式的合成-表征循环[8],实验上通常通过差示扫描量热法(DSC)[9]和热重分析法[10]等热分析技术测定化合物的分解温度.然而,一些含能材料本身具有高度危险性,且DSC测定需要复杂的且耗时的高质量样品制备[11],合成、纯化与高质量样品的制备过程不仅成本高昂、耗时费力,还伴随着显著的安全风险,这使得在实验研究前,利用理论方法对含能化合物的分解温度进行计算或预测变得尤为重要和必要.对未合成化合物的分解温度进行预测,还有助于摆脱文献束缚,拓展分子设计与筛选的思路.然而,热分解温度的预测仍面临重大挑战,因为其热分解机制涉及一系列复杂的物理化学过程(包括能量接收、传递、吸收与积累、晶格振动、化学键断裂、燃烧及爆轰等).这一机制与含能材料的多级结构、物性特征及激发过程密切相关.目前,研究者可通过量子力学方法和反应力场方法来探究初始分解机理、产物演变及分解过程[12,13].但由于需要在分子和晶体层面考虑大量物理化学参数,相关计算仍存在耗时冗长、计算成本高昂的问题[14],限制了其在结构多样性含能材料虚拟筛选中的应用.

随着计算机技术与化学信息学的发展,机器学习(ML)[15]方法为含能材料分解温度的预测提供了新途径.ML通过从实验和理论数据中提取规律,建立从分子/晶体特征描述符到分解温度的映射模型[16~18],旨在便捷、高效地评估含能化合物的热稳定性,加速材料的开发并降低成本.研究者已采用多种描述符构建预测模型,范围从简单的元素组成和分子骨架描述符,到复杂的拓扑结构、几何构型、量子化学描述符(如HOMO/LUMO能级)乃至自然键轨道电荷、静电势和核磁共振化学位移等[19].然而,现有预测模型的性能仍不尽理想,其主要涉及3个问题:(1)普遍面临着数据规模有限的挑战[14,16],这在一定程度上制约了模型的泛化能力与预测的可靠性;(2)当前的主流方法依赖Dragon或Materials Studio等工具的海量描述符筛选[20,21],描述符冗余同时缺乏可解释性,难以揭示影响热稳定性的内在机理;(3)现有模型对分解温度的预测能力相对较差[22,23].系统提升模型对分解温度的预测能力,已成为当前研究中的关键课题.

为了应对上述挑战,本文搜集到了一个相对较大的数据集,从分子结构本质与分解引发机理出发,自行开发设计与分子的分解温度相关性比较高的描述符.重点关注与热稳定性直接相关的结构特征(如共轭体系、分子对称性、环状结构等),并创新性地引入了键解离能(Bond dissociation energy, BDE)这一关键参数,直接量化化学键断裂的难易程度.旨在构建具有可解释性的预测模型,不仅提高了模型的预测能力,更有潜力辅助机理的研究和助力含能材料的分子设计.

1 数据集

数据来源于Wu等[21]的研究,数据集中共有1021个含能化合物,由C,H,O,N,F和Cl 6种元素组成.其分解温度分布于38~425 ℃之间,具体分布如图1所示.其中,温度值在130~270 ℃区间内出现频率最高.在建模过程中,为了保证模型的可靠性和准确性,将数据集随机划分为831个化合物的训练集和190个化合物的测试集.

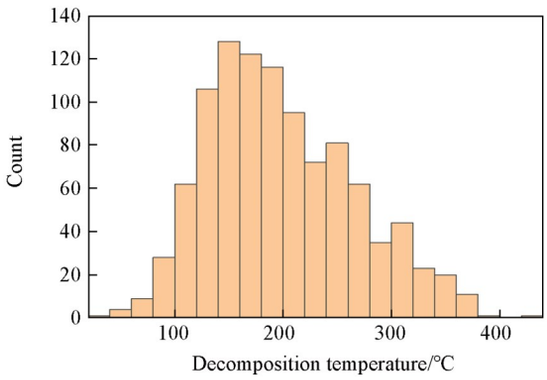


Fig. 1　Decomposition temperature distribution of the data set

2　方　　法

2.1　描述符

2.1.1　EGD描述符　分子分解温度与化学键断裂过程密切相关，因此在设计基团数量作为描述符时，首先基于化学键提取，即每个描述符至少对应一个由两个末端原子构成的化学键．在此基础上，为了更全面地捕捉官能团的整体效应与邻近基团的影响，描述符的提取范围被系统性地扩展至更大的原子组合．定义了两级描述符体系，一级描述符对应简单的基础化学键(如单键、双键等)，二级描述符由一级描述符与其直接相连的原子或基团构成，用于表征更完整的官能团环境，在构建二级描述符时，允许其包含的一级描述符中的化学键重复出现，是为了同时保留局部键信息与整体官能团信息．以图2中的三硝基丁酸为例，其中一级描述符GB1(C—OH)，GB2(C═O)和GB3(N═O)的数目分别为1，1和2；而二级描述符GB4(C—COOH)和GB5(—NO_2)的数目均为1．自主开发了相关的Java程序提取每个化合物中的一级和二级描述符，保证数据集中的每个化合物中的化学键都被提取到．基于数据集中分子结构的分析，共获得163个基团．考虑到分子体积通过影响范德华力和键能分布等因素与分解温度密切相关，进一步将分子中的每个原子近似为球体，采用范德华半径计算其体积，将每个基团的数量除以体积，共得到163个描述符．

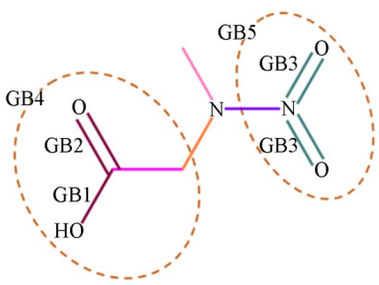


Fig. 2　Partial examples of first-level descriptors and second-level descriptors in trinitrobutyric acid

GB denotes group based on bonds.

除了化学键以外，化合物的组成也对分解温度产生影响．基于此，提取了各元素(C，H，O，N，F，Cl)的数量及非氢原子总数作为7个元素描述符．通过将得到的7个元素描述符与分子体积的比值作为新的描述符，构建了包含14个参数的描述符集(7个元素数量描述符和7个元素体积比描述符)．为了捕捉特定结构特征对分解温度的影响，额外提取了一系列分子结构描述符．这包括：(1) 反映电子离域与共振稳定性的共轭原子数描述符(C，N，O的共轭原子数及其总和)；(2) 表征结构对称性的分子对称性描述符(除对称轴或对称中心外的原子数与总原子数之比)；(3) 描述环状结构特征的小环数量及小环比描述符(三元环、四元环、五元环的数量及其与环上原子数的比例)；(4) 体现芳香稳定性的描述符；(5) 表征分子分支程度的维纳指数描述符．这些描述符同样进行了体积归一化处理，最终得到23个分子结构描述符．此外，还将分子体积本身作为1个描述符，共计24个描述符．

将以上得到的描述符统称为扩展基团描述符(Extended group descriptor，共201个描述符)，简称为EGD描述符．

2.1.2　键解离能描述符　键解离能描述符(BDE)[24]是指化学键发生均裂时所需的能量，是作为表征化学键强度的关键热力学参数．由于化合物分解的本质在于化学键的断裂过程，因此获取分子中各化学

键的强度信息至关重要. 采用自主开发的BDE预测模型[25], 对所有化合物中各化学键的BDE进行了计算. 以化合物2,4,6-三硝基甲苯为例, 其分子的结构式如图3所示, 计算了分子中所有化学键的BDE值, 并将其按照从小到大的顺序进行排列, 其部分数值列于表1.

从表1可以看出, 3个C—NO_2的BDE值是分子中最低的, 其中化学键10-7和2-3的环境相同, 相应的BDE值也相等. 由于前述2个化学键与甲基相邻, 空间位阻大, 因而两者的BDE值小于化学键5-13的BDE值. 同时文献[26]也指出, 大量含能化合物包含—NO_2作为爆炸基团, 与—NO_2(X-NO_2, X=C, N, O)相连的化学键在反应或分解

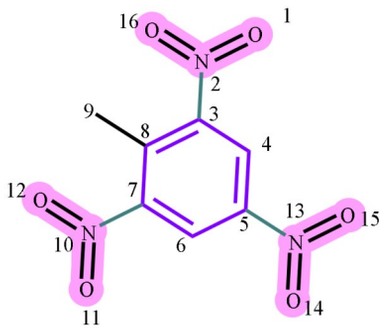


Fig. 3 Structure of 2,4,6-trinitrotoluene

Atom numbers: the same in Table 1 for bond dissociation energy analysis.

Table 1 Partial BDE values of 2,4,6-trinitrotoluene

BDE	Atom number 1	Atom number 2	Element 1	Element 2
58.92063	2	3	N	C
58.92063	10	7	N	C
59.81022	13	5	N	C
68.24258	3	8	C	C
69.09007	5	4	C	C
70.96597	7	6	C	C
79.13984	8	9	C	C
88.21220	1	2	O	N
88.21220	12	10	O	N
88.53529	15	13	O	N

过程中被认为是最快的键或触发键. 基于这一假设, X-NO_2的BDE通常被用作与化学键强度相关的热稳定性的关键指标. 如Harper等[26]通过分析证实, X—NO_2(X=N, C, O)键是三硝基甲苯(TNT), 黑索今(RDX)和季戊四醇四硝酸酯(PETN)等典型含能化合物中的关键触发键. 为了探究X—NO_2键的BDE在定量构效关系(Quantitative structure-property relationship, QSPR)中的潜在影响, 将X—NO_2键的最小BDE值作为描述符纳入建模分析. 如在三硝基甲苯中的3个C-NO_2基团, 在此选择C-NO_2中最小的BDE值, 即BDE(C-NO_2)描述符值为58.92063.

2.1.3 RDkit描述符 RDKit[27]是一款用于化学信息学的开源软件包. 使用RDKit计算和提取了分子的208个描述符, 主要包括反映分子的物化性质、分子的拓扑结构、电子性质及结构、几何形状、立体化学等性质的描述符. 最后, 删除了那些在整个数据集的所有分子中值都为零的描述符, 最终得到180个RDKit描述符.

2.2 机器学习方法

2.2.1 偏最小二乘法 偏最小二乘法(Partial least squares, PLS)[28]是一种用于多变量数据分析的统计方法, 它结合了主成分分析和多元回归的优点. PLS不仅能有效解决高维数据问题及多重共线性问题, 还能提高回归和预测模型的精度. 其中主成分[29]是PLS的一个重要属性, 主成分数量的选择直接影响建模结果后续数据分析的有效性. 较少的主成分数会导致较多信息的丢失, 降低模型的可解释性, 导致欠拟合, 使模型无法捕捉到数据中的所有信息. 而较多的主成分数可能会包含更多的噪声, 增加模型的复杂度, 同时数据过度拟合, 使得模型在新数据上的泛化能力下降. 对于所有PLS模型, 均根据训练集的10折交叉验证结果选择最佳主成分数. PLS在weka3.8.6的版本下实现.

2.2.2 随机森林 随机森林(Random forest, RF)[30]是一种基于树模型集成建模方法, 常用于解决分类判别或定量分析任务. 在回归问题中, 采用RF建模过程中的Bagging方法对回归的样本进行抽样, 并在每个回归树上随机选取一些变量进行拆分. 最后, RF模型的输出是所有树输出的简单平均值. RF

模型通常具有良好的分析性能，不易过拟合，可用于离散数据处理．RF在R-project[31]2.15.2版本中的随机森林库实现．

2.2.3　支持向量机　支持向量机(Support vector machine，SVM)[32]是一种强大的监督学习算法，被广泛应用于分类和回归问题．它基于“最大间隔分离”的原理，通过寻找一个最优的超平面来将不同类别的数据分开，并使得类别之间的间隔最大化．对于线性不可分的情况(即数据不能通过一个超平面线性分开)，SVM引入了核函数(如线性核函数、多项式核函数、径向基函数核等)来将数据映射到更高维的特征空间，从而实现非线性分类．SVM能有效避免过拟合，适合解决非线性、高纬度等问题．SVM在weka3.8.6的版本下实现，其中，训练集的评价采用了10折交叉验证．

3　结果与讨论

3.1　基于EGD+BDE描述符的预测结果

使用实验值与预测值的线性相关系数(Pearson correlation coefficient，R)值的平方、平均绝对误差(Mean absolute error，MAE)、均方根误差(Root mean square error，RMSE)来评估模型的性能．在一个模型中R^2越大，MAE和RMSE越小，说明模型越好．

3.1.1　RF的预测结果　采用EGD描述符作为分子特征输入，通过RF算法构建模型，预测结果列于表2．训练集的交叉验证结果为R^2=0.624，MAE=32.0 ℃，RMSE=42.1 ℃；测试集的预测结果为R^2=0.673，MAE=28.2℃，RMSE=38.5℃．其结果说明EGD描述符跟分子的分解温度具有一定的相关性．

Table 2　Modeling results based on EGD+BDE descriptors*

Machine learning	Descriptor {number of descriptors}	Cross-validation of the training set(R^2/MAE/RMSE)	Predication of the test set (R^2/MAE/RMSE)
RF	EGD{201}	0.624/32.0/42.1	0.673/28.2/38.5
RF	EGD+BDE(N-NO_2){202}	0.639/31.3/41.3	0.677/27.9/38.2
RF	EGD+BDE(C-NO_2){202}	0.630/31.8/41.8	0.695/27.6/37.4
RF	EGD+BDE(O-NO_2){202}	0.621/32.2/42.3	0.668/28.3/38.7
RF	EGD+BDE(X-NO_2){204}	0.637/31.3/41.4	0.704/26.9/36.8
SVM	EGD{201}	0.615/32.0/42.5	0.652/29.1/39.6
SVM	EGD+BDE(X-NO_2){204}	0.621/31.4/42.0	0.664/28.4/38.9
PLS	EGD{201}	0.549/36.2/46.2	0.594/33.4/42.8
PLS	EGD+BDE(X-NO_2){204}	0.558/35.6/45.3	0.609/32.4/41.8

* X-NO_2 represents N-NO_2, C-NO_2 and O-NO_2. The unit of MAE and RMSE: ℃.

考虑到化合物热分解过程主要涉及化学键断裂，因此使用本实验室开发的模型[25]计算了分子中含有的所有的X-NO_2(X=C，N，O)的BDE值，选取每个分子中X-NO_2中最小的BDE值作为指数提交给RF．在EGD描述符的基础上，当只有BDE(N-NO_2)作为补充描述符时，训练集的交叉验证结果为R^2=0.639，MAE=31.3 ℃，RMSE=41.3 ℃，测试的预测结果为R^2=0.677，MAE=27.9 ℃，RMSE=38.2 ℃．当BDE(C-NO_2)作为补充描述符时，训练集的交叉验证结果为R^2=0.630，MAE=31.8 ℃，RMSE=41.8 ℃，测试的预测结果为R^2=0.695，MAE=27.6 ℃，RMSE=37.4 ℃．当BDE(O-NO_2)作为补充描述符时，训练集的交叉验证结果为R^2=0.621，MAE=32.2 ℃，RMSE=42.3 ℃，测试集的预测结果为R^2=0.668，MAE=28.3 ℃，RMSE=38.7 ℃．结果显示，当BDE(N-NO_2)，BDE(C-NO_2)分别作为补充描述符构建模型时，训练集和测试集的结果都有所提高．

将X-NO_2(X=C，N，O)的BDE值作为补充描述符同时提交给模型，训练集的交叉验证结果为R^2=0.637，MAE=31.3 ℃，RMSE=41.4 ℃，测试集的预测结果为R^2=0.704，MAE=26.9 ℃，RMSE=36.8 ℃．相比于只有EGD描述符的模型，结果有了明显的提高，说明X-NO_2的BDE值确实影响化合物的分解温度，从结果中看出N-NO_2和C-NO_2对分解温度影响比较大，可能是这两个键更容易触发化合

物的分解.

3.1.2 SVM和PLS的建模结果 为了比较不同机器学习方法的建模结果，研究了分别采用SVM和PLS构建化合物的分解温度预测模型. 结果列于表2.

在SVM建模过程中，分别考察了EGD描述符及其与BDE描述符组合的建模效果. 结果显示，引入BDE描述符后，模型在训练集与测试集上性能均有一定改善，表明BDE对分解温度的预测具有辅助作用. 但是，SVM结果比RF模型差.

同时采用PLS方法构建模型，首先评估不同主成分数对应的模型性能，最终选定最优主成分数建立模型. 结果表明，引入BDE描述符后模型性能有所提升，进一步验证了该描述符的积极贡献. 值得注意的是，PLS模型性能明显比SVM模型差，这表明化合物的热分解温度与其分子结构之间并非简单的线性关系.

3.1.3 沙普利加法解释(SHAP)分析 为了进一步了解所提出的各个描述符对模型的贡献程度，对RF中变量重要性排名前20的变量进行了SHAP分析，用于量化每个特征对模型的影响. 图4(A)展示了基于SHAP值平均绝对值的特征重要性排序，条形长度直接反映各特征对预测结果的平均影响强度，从图中可以看出，共轭性能、BDE值对分解温度的影响比较大. 图4(B)是SHAP特征重要性摘要图，进一步揭示了特征值的具体影响方向. 分析结果表明，描述符*n*conjugation[C]/vanV(*n*conjugation[C]代表共轭碳原子的数量，vanV代表分子体积)重要性最高，其SHAP值显著为正，表明分子共轭性增强可提升分解温度，这可能源于π电子离域对热分解能量的分散作用. 此外，N-NO_2，C-NO_2的BDE值对模型的影响也比较大，说明BDE确实是影响化合物分解温度的重要因素. N-NO_2的BDE对模型最主要为负向影响，表明该键断裂可能是引发分解的关键步骤. C-NO_2的BDE对模型为正向影响，反映其较高的键能有助于维持分子稳定性. 这也间接说明N-NO_2的存在会降低化合物的分解温度，而C-NO_2的存在会增加化合物的分解温度.

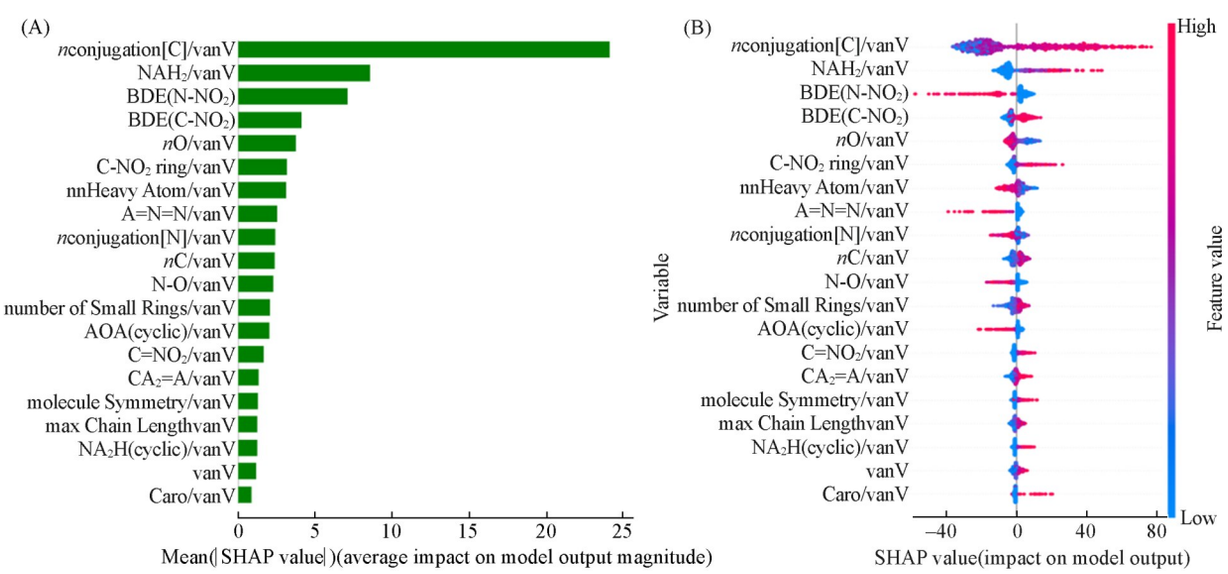


Fig. 4 Variable SHAP plots

(A) Feature importance diagrams for each variable; (B) summary diagram of feature importance

为了验证这一观点，对数据集中的1021个化合物分解温度与结构进行了筛选整理，其数值列于表3，对于分解温度(425~260 ℃)比较高的化合物，其结构中含有的N-NO_2基团的化合物比较少，其比例仅有5.08%，而含有C-NO_2基团的化合物却比较多，其比例为77.66%. 对于分解温度比较低(151~38 ℃)的化合物，其结构中含有N-NO_2的化合物占比42.61%，其结构中含有C-NO_2的化合物比例为58.45%. 这一分布规律表明：减少N-NO_2基团数量有助于提高化合物分解温度，与SHAP分析中N-NO_2呈负向影响，C-NO_2呈正向影响的结论一致.

Table 3 Decomposition temperature distribution of the compounds containing N-NO_2 and C-NO_2 in the data set

Decomposition temperature/℃	Number of compounds	Number of molecules containing N-NO_2(percentage)	Number of molecules containing C-NO_2(percentage)
425—260	197	10(5.08%)	153(77.66%)
260—151	540	143(26.48%)	329(60.93%)
151—38	284	121(42.61%)	166(58.45%)

3.2 基于RDKit描述符的预测结果

为了评估RDKit描述符在预测化合物分解温度中的有效性，分别采用RF，SVM，PLS构建了定量构效关系模型．由表4中数据可见，3种模型中RF表现出最优的预测能力．其训练集的交叉验证结果为R^2=0.636，MAE=31.3 ℃，RMSE=41.4 ℃；测试集的预测结果为R^2=0.708，MAE=27.1 ℃，RMSE=36.5 ℃．RF结果显著优于其它两种方法，SVM模型性能次之，而PLS模型则相对较差．该结果表明，在捕捉RDKit描述符(涵盖分子拓扑、电荷与电负性等信息)与分解温度之间复杂的非线性关系时，RF模型比SVM及PLS线性方法更具优势．

Table 4 Modeling results based on RDKit descriptors*

Machine learning	Descriptor {number of descriptor}	Cross-validation of the training set (R^2/MAE/RMSE)	Predication of the test set (R^2/MAE /RMSE)
RF	RDKit {180}	0.636/31.3/41.4	0.708/27.1/36.5
SVM	RDKit {180}	0.591/34.5/44.6	0.601/30.9/42.6
PLS	RDKit {180}	0.480/38.2/49.9	0.559/34.6/45.2

* X-NO_2 represents N-NO_2, C-NO_2 and O-NO_2. The unit of MAE and RMSE: ℃.

值得注意的是，RDKit描述符模型与EGD+BDE(X-NO_2)模型的预测结果相近．然而，从化学可解释性的角度看，EGD+BDE(X-NO_2)描述符直接源于特定的化学结构及其键解离能(BDE值)，能够更直观地揭示分子结构(如最弱键的强度)对热稳定性的影响机制，在平衡预测能力与模型可解释性方面展现出独特价值，对于含能材料的分子设计具有更明确的指导意义．

3.3 基于RDKit和EGD+BDE(X-NO_2)组合模型的预测结果

鉴于RDKit描述符同时包含分子结构信息(如原子数量)和电子特性参数(如原子电荷、电负性)，而本文提出的描述符主要基于分子结构和BDE衍生，二者结合可能产生互补效应．基于此，将RDKit描述符预测值与EGD+BDE(X-NO_2)描述符预测值的平均值作为最终的预测值构建模型，其结果列于表5．使用RF构建模型，训练集的交叉验证结果为R^2=0.657，MAE=30.5 ℃，RMSE=40.5 ℃，测试集的预测结果为R^2=0.721，MAE=26.1 ℃，RMSE=35.9 ℃．SVM的训练集的交叉验证结果为R^2=0.659，MAE=31.9 ℃，RMSE=41.8 ℃，测试集的预测结果为R^2=0.679，MAE=28.4 ℃，RMSE=39.2 ℃；PLS的训练集的交叉验证结果为R^2=0.567，MAE=34.8 ℃，RMSE=44.8 ℃，测试集的预测结果为R^2=0.623，MAE=31.4 ℃，RMSE=41.0 ℃．结果表明，RDKit描述符与EGD+BDE(X-NO_2)的组合建模策略优于单一特征建模方案，其中RF模型的综合性能表现尤为突出，其预测精度明显优于SVM和PLS模型．

Table 5 Predicted results of the combined model of RDKit and EGD+ BDE*

ML method	Descriptor	Cross-validation of the training set (R^2/MAE/RMSE)	Predication of the test set (R^2/MAE/RMSE)
RF	RDKit and EGD+BDE(X-NO_2)	0.657/30.5/40.5	0.721/26.1/35.9
SVM	RDKit and EGD+BDE(X-NO_2)	0.659/31.9/41.8	0.679/28.4/39.2
PLS	RDKit and EGD+BDE(X-NO_2)	0.567/34.8/44.8	0.623/31.4/41.0

* X-NO_2 represents N-NO_2, C-NO_2 and O-NO_2. The unit of MAE and RMSE: ℃.

3.4 基于RF的变量重要性选择变量子集的预测结果

在机器学习建模过程中，冗余特征的引入会显著增加运算负荷，削弱模型泛化能力，甚至引发过拟合风险. 为此，采用变量重要性分析策略开展变量筛选工作. 基于RF算法的变量重要性评分体系，分别对EGD+BDE(X-NO_2)描述符和RDKit描述符进行降序排列，进而遴选出具有最高表征能力的若干关键变量用于模型构建. 具体筛选结果列于表6. 经综合评估发现，采用EGD+BDE(X-NO_2)描述符组合的前170个最重要变量所建模型表现最优，而RDKit描述符则以前105个特征构建的模型性能最佳.

Table 6 Predicted results of variable subsets based on variable importance of RF*

Descriptor	Cross-validation of the training set (R^2/MAE/RMSE)	Predication of the test set (R^2/MAE/RMSE)
EGD+BDE(X-NO_2) top 170	0.642/31.1/41.2	0.700/27.3/37.1
EGD+BDE(X-NO_2) top 140	0.637/31.4/41.5	0.695/27.4/37.3
EGD+BDE(X-NO_2) top 110	0.640/31.2/41.3	0.696/27.4/37.3
RDKit top 155	0.634/31.3/41.5	0.709/26.9/36.5
RDKit top 130	0.635/31.2/41.5	0.712/26.9/36.3
RDKit top 105	0.639/31.3/41.3	0.711/26.9/36.4
RDKit top 80	0.636/31.3/41.5	0.716/26.6/36.1
Final model	0.663/30.4/40.3	0.720/26.3/36.0

* X-NO_2 represents N-NO_2, C-NO_2, and O-NO_2. The unit of MAE and RMSE: ℃.

最终模型整合了上述2个最优模型的预测值，取其平均值作为最终输出，其训练集的交叉验证结果为R^2=0.663，MAE=30.4 ℃，RMSE=40.3 ℃，测试集的预测结果为R^2=0.720，MAE=26.3 ℃，RMSE=36.0 ℃. 相较于全变量模型，变量筛选虽未显著提升模型预测性能，但有效缩减了描述符总量，降低了计算耗时，实现了模型效率与性能的良好平衡.

为了更直观地观察最终模型的预测能力，分别给出了训练集与测试集中实验分解温度同预测值的对比散点图[图5(A)和(B)]、预测绝对误差分布图[图5(C)和(D)]及绝对误差频率直方图[图5(E)和(F)]. 分析结果表明，模型在训练集和测试集上的预测绝对误差主要分布在50 ℃以内，仅少数样本的预测绝对误差超过100 ℃. 误差频率直方图进一步显示，误差值较小的样本占比较高，而误差值较大的化合物数量则比较少.

为了研究数据集划分对模型的影响，又进行了10次随机的数据集划分，并采用RF建立模型，所得的预测结果列于表7. 可以看出，训练集的结果相对比较稳定，其交叉验证的结果范围为：R^2=0.642~0.674，MAE=29.4~30.9 ℃，RMSE=39.1~41.2 ℃. 测试集预测结果的波动大一些，范围为：R^2=0.611~0.758，MAE=25.1~32.7 ℃，RMSE=33.7~42.2 ℃，这主要归因于测试集样本量有限，单次划分的偶然性影响被放大. 由于每次划分后训练集和测试集的样本不同，训练集和测试集的预测结果不适合直接用于模型性能的比较，因而采用整个数据集的MAE值评价模型，计算过程为

$$MAEwh=(MAEtr\times Ntr+MAEte\times Nte)/(Ntr+Nte) \quad (1)$$

式中：*MAEwh*代表整个数据集的MAE值；*MAEtr*代表训练集交叉验证的MAE值；*MAEte*代表测试集的MAE值；*Ntr*是训练集的化合物数量；*Nte*是测试集的化合物数量.

由式(1)所计算的10个随机模型和本文最终模型的结果列于表7. 可以看出，这些随机模型的整个数据集的MAE值在29.66~30.11 ℃之间，波动范围很小(最大差距仅为0.45 ℃)，证明了本模型具有较强的学习能力与拟合稳定性. 为了与采用相同数据集的文献结果[21]进行比较，将文献结果也列于表7. 由于文献中并未给出具体的训练集和测试集的划分，因而也采用整个数据集的MAE值与本文结果进行比较. 由表7可以看出，文献中整个数据集的MAE值为31.36 ℃，高于本文整体数据集的MAE值，即本文中的建模结果要优于文献中的结果.

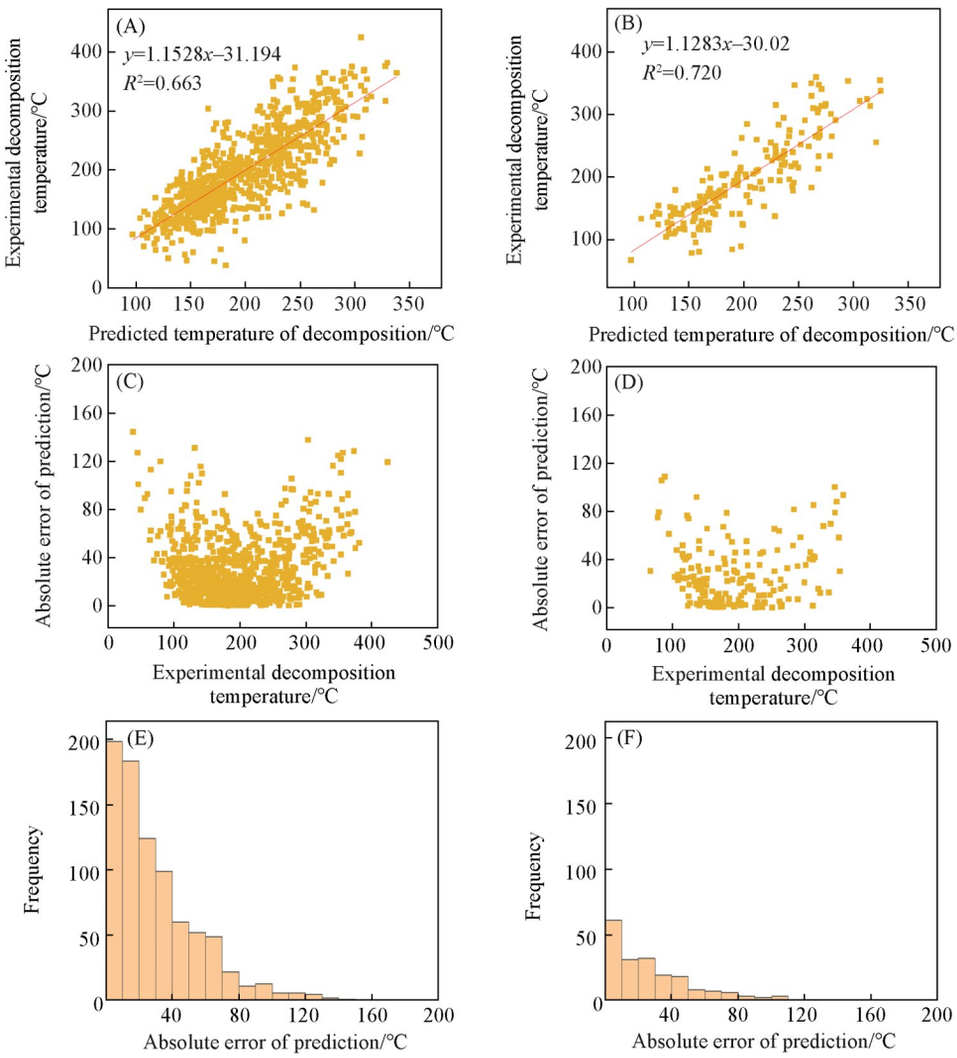


Fig. 5 Results of the best model's performance on the training and test sets

(A) Experimental *vs.* predicted decomposition temperatures for the training set; (B) experimental *vs.* predicted decomposition temperatures for the test set; (C) distribution of absolute prediction errors for the training set; (D) distribution of absolute prediction errors for the test set; (E) frequency distribution histogram of absolute errors for the training set; (F) frequency distribution histogram of absolute errors for the test set.

Table 7 Model results of the randomly partitioned data sets and in literature*

No.	Cross-validation of the training set (R^2/MAE/RMSE)	Predication of the test set (R^2/MAE /RMSE)	MAE of the whole data set
Final model	0.663/30.4/40.3	0.720/26.3/36.0	29.68
1	0.673/29.9/39.7	0.631/29.3/40.0	29.77
2	0.670/29.7/39.8	0.611/30.8/41.3	29.90
3	0.652/30.3/40.7	0.689/29.4/37.9	30.11
4	0.663/29.8/39.8	0.678/30.2/40.3	29.88
5	0.659/29.0/39.3	0.697/32.7/42.2	29.66
6	0.655/30.8/41.2	0.707/25.1/33.7	29.72
7	0.674/29.4/39.1	0.654/31.9/42.2	29.85
8	0.660/30.2/40.4	0.699/28.0/36.4	29.82
9	0.642/30.5/40.7	0.758/27.1/36.7	29.87
10	0.649/30.9/41.1	0.700/26.2/36.4	30.04
Literature results[21]	0.60/32.2/—	0.65/27.7/—	31.36

* X-NO_2 represents N-NO_2, C-NO_2, and O-NO_2. The unit of MAE and RMSE: ℃.

4 结　论

基于文献中搜集的1021个由碳、氢、氧、氮、氟和氯元素构成的含能化合物，创新性地构建了基于官能团特征的新型分子描述符体系，继而引入BDE作为补充描述符，同时使用RDKit软件包生成RDKit描述符．采用RF，SVM和PLS 3种机器学习算法构建预测模型．结果表明，纳入BDE描述符的模型较单纯使用EGD描述符的基准模型展现出显著的性能提升；而将EGD+BDE(X-NO_2)描述符、RDKit描述符进行多维特征融合所构建的模型表现尤为优异．值得注意的是，在所有建模方法中，RF算法的预测精度明显优于SVM和PLS模型．借助SHAP可视化技术对最优模型的关键影响因素进行解析，发现共轭性能、硝基官能团(N-NO_2，C-NO_2)的BDE参数对化合物分解温度具有显著影响．最后，通过多次随机的数据集划分，揭示了模型性能的波动范围，提供了对模型泛化能力更全面，更可靠的评估，更好地体现出了因特定划分方式的不同可能带来的偏差．

参　考　文　献

[1] Zlotin S. G., Churakov A. M., Egorov M. P., Fershtat L. L., Klenov M. S., Kuchurov I. V., Makhova N. N., Smirnov G. A., Tomilov Y. V., Tartakovsky V. A., *Mendeleev Commun.*, **2021**, *31*(6), 731—749
[2] Muravyev N. V., Fershtat L., Zhang Q., *Chem. Eng. J.*, **2024**, *486*, 150410
[3] Sabatini J., Oyler K., *Crystals*, **2015**, *6*(1), 5
[4] Yan Q. L., Zhao F. Q., Kuo K. K., Zhang X. H., Zeman S., DeLuca L. T., *Prog. Energy Combust. Sci.*, **2016**, *57*, 75—136
[5] Liu R., Liu J., Zhou Q., *Phys. Chem. Chem. Phys.*, **2024**, *26*(41), 26209—26221
[6] Lv M., Zhou P., Ma Y., Li W., Liu J., *J. Mol. Struct.*, **2022**, *1262*, 132955
[7] Bao G., Abe R. Y., Akutsu Y., *J. Therm. Anal. Calorim.*, **2020**, *143*(5), 3439—3445
[8] Liu Y., Zhao T., Ju W., Shi S., *J. Materiomics*, **2017**, *3*(3), 159—177
[9] Roduit B., Hartmann M., Folly P., Sarbach A., Brodard P., Baltensperger R., *J. Therm. Anal. Calorim.*, **2014**, *117*(3), 1017—1026
[10] Luo Z. H., Zhou J. J., Li H., Xia Y. H., Bai L. F., Yang H. J., *Energ. Mater. Front.*, **2023**, *4*(3), 125—133
[11] Abd-elghany M., Klapötke T. M., *Phys. Sci. Rev.*, **2018**, *3*(4), 20170103
[12] Hu J., Jin J. X., Hou X. J., Rao C. H., He Y., Wu K. J., *Ind. Eng. Chem. Res.*, **2025**, *64*(4), 2396—2405
[13] Gou Q., Liu J., Su H., Guo Y., Chen J., Zhao X., Pu X., *iScience*, **2024**, *27*(4), 109452
[14] Rein J., Meinhardt J. M., Hofstra Wahlman J. L., Sigman M. S., Lin S., *Angew. Chem. Int. Ed.*, **2023**, *62*(17), e202218213
[15] Wang B., Yi H., Xu K., Wang Q., *J. Therm. Anal. Calorim.*, **2016**, *128*(1), 399—406
[16] Pan Y., Zhang Y., Jiang J., Ding L., *J. Loss Prev. Process Ind.*, **2014**, *31*, 41—49
[17] Li X., Kong D., Luan Y., Guo L., Lu Y., Li W., Tang M., Zhang Q., Pang A., *Struct. Chem.*, **2024**, 35(5), 1375—1385
[18] Kong D., Luan Y., Zhao X., Lu Y., Li W., Zhang Q., Pang A., *Chemom. Intell. Lab. Syst.*, **2023**, *243*, 105021
[19] Adak P. K., Singh S. K., Singh J., Mahesh S., Jain M. K., Sagar P. A. K., Banerjee S., Khan A. S., *J. Mol. Model.*, **2022**, *28*, 400—404
[20] Zhang Z., Chen C., Cao Y., Wen L., He X., Liu Y., *Thermochim. Acta*, **2024**, *735*, 179717
[21] Wu J. N., Song S. W., Tian X. L., Wang Y., Qi X. J., *Energ. Mater. Front.*, **2023**, *4*(4), 254—261
[22] Song W., Chen F., Wang Y., Wang K., Yan M., Zhang Q., *J. Mater. Chem. A*, **2021**, *9*, 21723—21731
[23] Huang X., Li C., Tan K., Wen Y., Guo F., Li M., Huang Y., Sun C., Gozin M., Zhang L., *iScience*, **2021**, *24*(3), 102240
[24] Saju A., Gunasekera P. S., Morgante P., MacMillan S. N., Autschbach J., Lacy D. C., *J. Am. Chem. Soc.*, **2023**, *145*(24), 13384—13391
[25] Luan Y., Kong D. L, Guo L. L., Zhang Q. Y., Zhou Y. M., *Chem. J. Chinese Universities*, **2025**, *46*(3), 20240373(栾玥，孔丁羚，郭莉莉，张庆友，周艳梅. 高等学校化学学报，**2025**，*46*(3)，20240373)
[26] Harper L. K., Shoaf A. L., Bayse C. A., *ChemPhysChem*, **2015**, *16*(18), 3886—3892
[27] Scalfani V. F., Patel V. D., Fernandez A. M., *J. Cheminf.*, **2022**, *14*(1), 87
[28] Cook R. D., Forzani L., *J. Chemom.*, **2020**, *34*(10), e3294
[29] Camacho J., *Chemom. Intell. Lab. Syst.*, **2017**, *160*, 40—51
[30] Biau G., Scornet E., *Test*, **2016**, *25*(2), 197—227
[31] R Core Team. R: A Language and Environment for Statistical Computing[CP]. Vienna, Austria: R Foundation for Statistical Computing, 2012. https://www.R-project.org/
[32] Valkenborg D., Rousseau A. J., Geubbelmans M., Burzykowski T., *Am. J. Orthod. Dentofacial Orthop.*, **2023**, *164*(5), 754—757

(Ed.: F, K, S)