

机器学习助力高效含能材料分子筛选与设计 ——基于主动学习的策略

杨 琳¹, 张晓龙², 王 鹤³, 周余伟^{3,4,*}, 滕波涛^{2,*}

(1. 山西财经大学 统计学院, 山西 太原 030006; 2. 天津科技大学化工与材料学院, 天津 300457;

3. 中国科学院山西煤炭化学研究所 煤炭高效低碳利用全国重点实验室, 山西 太原 030001;

4. 中科合成油技术股份有限公司, 北京 101407)

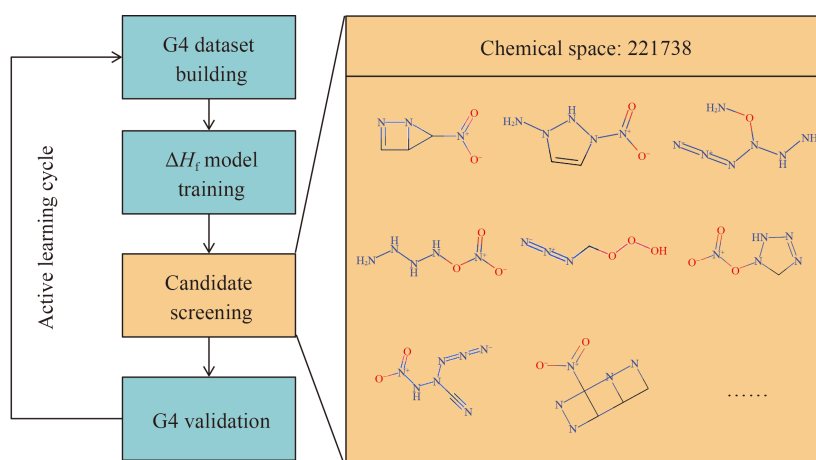
摘 要: 含能材料在军事和航天等领域应用广泛, 但其发现与合成主要依赖“试错法”, 严重制约了新型含能材料的研发与突破。本研究选取含能材料关键热力学性质——生成焓作为预测目标, 提出一种融合主动学习策略与 SMILES 分子特征表示的机器学习构效关系模型。基于 G4 高精度量子化学方法构建了包含 1447 种气相含能分子的数据集, 提取了 93 个 SMILES 有效描述符, 建立了基于线性模型的气相生成焓初步预测模型。对从 PubChem 数据库中获取的 221738 种潜在含能分子进行了系统预测。进一步引入主动学习策略,

对高误差样本迭代优化, 得到了泛化能力更好的预测模型, 该模型在典型含能分子上得到了验证。最终筛选出 20 个威力指数超过 2.0 倍 TNT 的候选分子, 其中, 大多数未见于现有已知含能材料库, 显示出本研究在高性能含能材料发现方面的潜力, 为新型含能材料的开发提供新策略与新途径。

关键词: 含能材料筛选; 主动学习; SMILES 特征; 生成焓; 爆炸热

中图分类号: O64

文献标识码: A



Machine learning-accelerated efficient screening and design of energetic material molecules ——A strategy based on active learning

YANG Lin¹, ZHANG Xiaolong², WANG He³, ZHOU Yuwei^{3,4,*}, TENG Botao^{2,*}

¹ School of Statistics, Shanxi University of Finance and Economics, Taiyuan 030006, China;

² College of Chemical Engineering and Materials Science, Tianjin University of Science and Technology, Tianjin 300457, China;

³ State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, China;

⁴ Synfuels China Technology Co., Ltd., Beijing 101407, China

Abstract: Energetic materials play a crucial role in military and aerospace applications. However, the discovery and synthesis of novel energetic compounds still largely rely on traditional trial-and-error approaches, which severely hinder the development of novel energetic

Received: 2025-10-13; Revised: 2025-11-10

Foundation items: Supported by National Natural Science Foundation of China (22372185, 22372120), Natural Science Foundation of Shanxi Province (202203021221219) and Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2023L164).

国家自然科学基金 (22372185, 22372120), 山西省自然科学研究面上项目 (202203021221219) 和山西省高等学校科技创新项目 (2023L164) 资助

*Corresponding author. E-mail: zhouyuwei@synfuelschina.com.cn, tbt@tust.edu.cn.

DOI: 10.3724/2097-213X.2025.JFCT.0030 CSTR: 32157.14.2097-213X.2025.JFCT.0176

materials. This study focuses on the prediction of a key thermodynamic property of energetic materials—heat of formation (HOF) and proposes a machine learning structure-property relationship model that integrates active learning strategies with SMILES-based molecular feature representation. A dataset containing 1447 gas-phase energetic molecules was constructed based on the high-accuracy G4 quantum chemical method, and 93 effective SMILES descriptors were extracted to establish a preliminary model for gas-phase HOF using linear model. Subsequently, the model was applied on the systematical prediction of 221738 potential energetic molecules retrieved from the PubChem database, enabling the screening of candidates with superior explosive performance. For samples with high prediction errors, an active learning strategy was implemented to iteratively refine the model parameters, significantly improving prediction accuracy. Validation on classical energetic molecules illustrated excellent predictive performance, highlighting the model's strong generalization capability. Finally, 20 candidate molecules with a TNT equivalent power index exceeding 2.0 were screened, most of which are new to the existing reservoir of known energetic materials. These results underscore the potential of this proposed approach in accelerating the discovery of high-performance energetic materials and offer a new strategy for the development of next-generation energetic compounds.

Key words: energetic materials screening; active learning; SMILES features; heat of formation; heat of explosion

含能材料是指能够储存大量化学能并在外界刺激下迅速释放能量的功能材料^[1-4], 例如 TNT (三硝基甲苯)、RDX(黑索金)等, 广泛应用于国防军工、航天推进、民用爆破等领域^[5-7]。由于在性能、安全性以及成本等方面的微小提升即可显著推动相关技术发展, 新型含能材料的发现与合成一直受到广泛关注。然而, 传统的含能材料研发模式主要依赖“试错法”, 即先基于化学直觉和经验设计分子结构, 再通过实验室合成及性能测试进行验证^[8-10]。该模式研发周期长、成本高且不确定性大, 严重限制了新型含能材料的研发效率。例如, CL-20(六硝基六氮杂异伍兹烷)^[11-12]这一里程碑式高能炸药从概念提出到最终投入使用历时近 20 年, 期间经历了多次失败。由此可见, 发展精准高效的筛选方法已成为突破含能材料研发瓶颈的关键途径。

密度、感度、爆速、生成焓和爆炸热等是评价含能材料性能的核心参数, 共同决定了其应用潜力与综合性能, 准确预测这些参数对于加速材料研发具有重要意义^[13-14]。其中, 生成焓(heats of formation, HOF)直接影响材料在爆轰或燃烧中释放的能量大小, 是计算爆速、爆压和爆炸热等重要性能指标的基础, 成为评估含能材料能量特性的关键依据。传统上, 含能材料生成焓的测定主要依赖实验方法, 如氧弹量热法^[15]、燃烧法^[16]等。然而, 这些方法操作复杂、耗时较长, 且对某些不稳定的高能化合物直接测量甚至存在安全风险, 因此, 不适用于大规模筛选。随着量子化学的发展, 理论计算逐渐成为含能材料生成焓性质获取的重要方法。例如, Byrd 等^[17]利用 G2 和 CBS-Q 方法计算了一系列硝基化合物和含能离子盐的生成焓, 计算结果与实验值吻合良好; Rice 及其合作者^[18]结合 B3LYP 泛函与 G3MP2 方法, 准确预测了五种四唑类含能化合物的生成焓。然而, 量子化学方法在处理大尺寸复杂分子体系时计算成本极

高, 难以支持高通量筛选的需求^[19]。

近十年来, 机器学习方法在分子设计及其性质预测中展现出强大潜力, 尤其在辅助药物发现领域取得显著成功^[20-22], 这为含能材料性质的快速预测提供了新的思路 and 手段^[23]。与传统理论计算不同, 机器学习依托数据驱动, 通过建立从输入(如分子结构)到输出(如材料性质)之间的构效关系, 实现对新材料性质的高效预测^[24], 从而大幅缩短研发周期。基于现有含能材料的性能数据, 多项研究已采用机器学习方法对其关键性能进行建模与筛选。例如, Nguyen 等^[25]基于 CSD 数据库中 10521 个分子数据集建立了高精度密度预测的机器学习模型, 均方根误差(RMSE)小于 0.05 g/cm³; Kang 等^[26]利用 41 种含能分子的爆炸热数据构建预测模型, 从 2732 种候选材料中筛选出 29 种爆炸威力指数大于 1.8 倍 TNT 当量的潜在化合物; 张炳儒等^[27]则借助机器学习实现了炸药撞击感度的智能识别。这些研究共同推动了含能材料性能预测的发展, 在此背景下, 生成焓作为决定材料能量性能的关键参数, 其准确预测已成为一个备受关注且极具挑战的前沿方向。

Elton 等^[28]较早尝试利用机器学习预测含能材料生成焓。他们构建了包含 109 种 CNOHF 类含能分子的数据集, 通过比较了不同特征化方法和机器学习模型, 最终模型的测试集平均绝对误差(MAE)为 17.1 kcal/mol, 决定系数 R^2 为 0.94。Chen 等^[29]提出了两种基于空间矩阵(体积占用矩阵和热贡献矩阵)表示的特征化方法, 在包含 451 种 CHNO 类有机含能化合物生成焓的数据集上训练模型, 测试集 MAE 达 14.056 kcal/mol。Xie 等^[30]提出了一种面向性质的自适应设计框架, 以 88 种已知含能材料化合物为训练集, 预测生成焓和爆炸热这两个关键性能指标。在线性核支持向量回归、E-state 指纹和 SOB 描述符的组合下取得 R^2 为 0.93 和 MAE 为 14.7 kcal/mol 的结果。上述研

究构建的数据集样本容量均小于 500, 为小样本条件下的生成焓机器学习预测提供了有益探索。

另一方面, Casey 等^[31]从 GDB 数据库中收集 26265 个分子构建 CSBB 数据集, 提出了一种基于三维卷积神经网络的新方法, 直接从分子的电子结构(包括电荷密度和静电势的空间点阵数据)中学习分子性质, 实现了包含生成焓在内的多个性质指标的同时预测。结果显示, 其生成焓预测的 RMSE 为 15 kcal/mol, R^2 为 0.979。2024 年, Zhang 等^[32]进一步采用无需人工设计描述符的消息传递神经网络模型, 实现了 RMSE 为 8.42 kcal/mol、 R^2 为 0.992 的生成焓预测。但是, 由于这些研究采用的 GDB 数据库中含能化合物数量较少, 因此, 相关模型在实际含能材料体系中的适用性仍需进一步验证。总体而言, 受限于含能材料生成焓可靠数据的匮乏, 机器学习研究在该领域的应用仍处于起步阶段, 尚待进一步成熟和完善。

针对当前含能材料生成焓可靠数据匮乏的现状, 本研究通过系统性的文献挖掘和数据库筛选, 结合高精度 G4 理论方法^[33]的计算, 构建了一个包含 C、H、O、N、F、Cl 元素的 1447 个分子的生成焓数据集, 为机器学习建模提供高质量的数据基础。然后, 论文基于化学 SMILES 表示, 采用 Mordred 特征化方法^[34]提取分子描述符, 并通过特征筛选策略获得精简有效的结构特征子集, 进一步构建了基于机器学习构效关系的生成焓预测模型。在此基础上, 本研究对从 PubChem 数据库中提取的 225543 种(经清洗后 221738 种)潜在含能材料分子进行了生成焓预测及爆炸热估算, 并筛选出爆炸威力指数较优的候选分子。在构建机器学习生成焓模型的过程中, 针对预测误差较大的样本, 采用主动学习策略将其重新纳入训练集, 并更新模型参数, 从而获得精度更高的生成焓预测模型。进一步地, 通过验证该模型在经典含能分子生成焓预测中的表现, 确认了其有效性及泛化能力。最终, 从候选分子库中筛选出 20 个威力指数超过 2.0 倍 TNT 的高能分子, 这为含能材料的快速筛选和研发提供了有价值的候选体系。本工作建立了一套计算高效、精度优良的机器学习生成焓预测方法, 为含能材料的理性设计和高效开发提供了新思路, 所构建的大规模生成焓数据集可为后续研究提供重要参考, 推动机器学习在含能材料领域的深入应用, 也为智能材料研发范式

的发展提供了典型范例和启发。

1 数据、特征及方法

1.1 含能材料生成焓数据集及筛选数据集的准备

为了训练和评估含能材料分子生成焓预测模型, 收集了一个包含 1447 个气相含能材料分子的数据集。该数据集中的分子主要由 C、H、N、O、F、Cl 等元素构成, 在结构与尺寸上均具有一定分布以增强数据的多样性和代表性——尺寸规模从最简单的双原子延伸至包含数十个原子的分子, 同时涵盖环状与链状等多种构型。在官能团方面, 数据集包含了硝基($-\text{NO}_2$)、氨基($-\text{NH}_2$)、羟基($-\text{OH}$)、硝酸酯($-\text{ONO}_2$)等多种典型含能基团。这种结构和组成的多样性有助于训练出稳健且泛化能力强的预测模型。总体而言, 该数据集规模适中, 包含上千个样本且便于 G4 方法计算处理, 为后续建模提供可靠的数据基础。图 1(a) 和 1(b) 分别展示了数据集中相对分子质量和原子数目的分布。相对分子质量集中分布于 50–200, 以 100–150 的中小分子数量最多; 原子数主要集中在 10–20, 其次是 5–10 和 20–25。此外, 从这些分子的 N 含量分布(见图 1(c))看出, 数据集中的含 N 量分布区间覆盖 0–1。因此, 以上分布特征表明, 本数据集以中等相对分子质量的含能分子为主, 同时兼顾了结构多样性。

由于实验测量的含能材料分子生成焓数据有限, 且实验测定易受样品纯度、稳定性、燃烧不完全等因素干扰, 导致不同来源数据的一致性较差, 本研究采用量子化学高精度 G4 方法对上述所有分子进行了生成焓计算。G4 方法是一种复合的量子化学计算方法, 通过几何优化、频率分析、单点能计算以及高阶能量校正等多层次计算相组合, 能够获得接近化学精度的热力学性质与能量值, 被广泛认为是可与实验测量媲美的“黄金标准”方法^[33]。支撑材料表 S1 展示了几种典型含能分子生成焓的实验值与 G4 计算值的对比, 两者的一致性确认了 G4 方法在本研究中的适用性。因此, 本研究利用 Gaussian 09 程序包^[35], 在 G4 理论水平下对所有含能分子进行了生成焓计算, 并通过频率分析确保分子构型均处于势能面最低点。生成焓的计算依据文献^[36]提供的原子化反应途径进行: 首先计算 0 K 下的生成焓, 再引入热校正项得到 298 K 下的最终目标气相生成焓值(ΔH_f)。

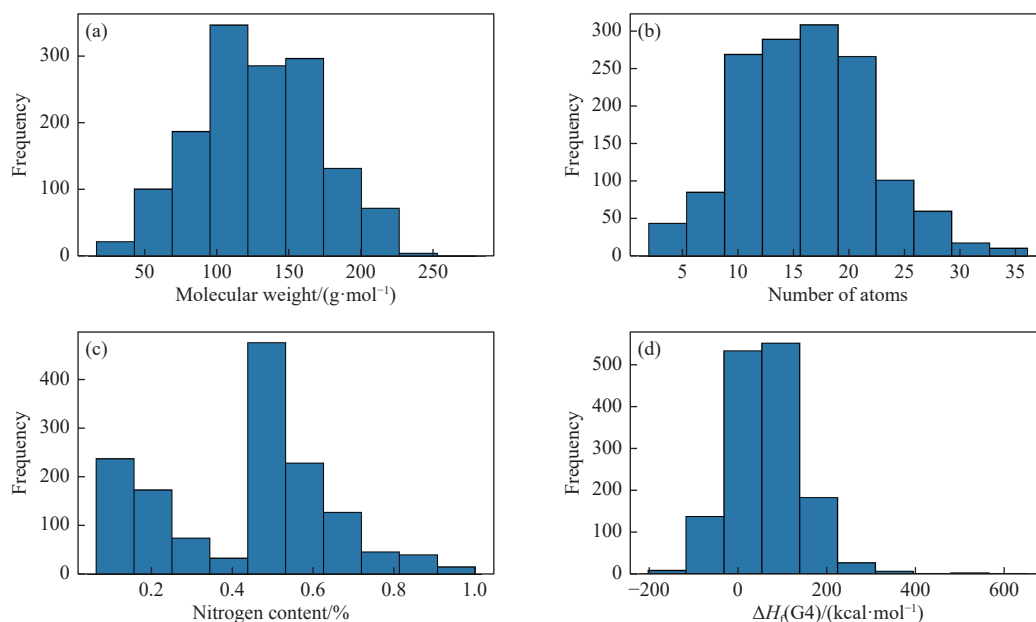


图1 含能材料分子集所含摩尔质量(中位数为 127.2 g/mol)、原子数(中位数为 16.0)、N 含量(中位数为 48.3%)和生成焓(中位数为 60.45 kcal/mol)分布

Fig. 1 Distribution diagram of molecular weight (median 127.2 g/mol), number of atoms (median 16.0), nitrogen content (median 48.3%) and heat of formation (median 60.45 kcal/mol) in energetic material molecules

由此,构建了一个包含 1447 个分子的高精度 G4 生成焓数据集,其 PubChem CID 和生成焓详见支撑材料表 S2,为后续机器学习建模提供了可靠的性质数据。图 1(d)展示了该数据集中生成焓的分布。结果表明,数据集中既包含高生成焓分子($\Delta H_f > 200$ kcal/mol),也有低生成焓分子($\Delta H_f < -200$ kcal/mol),而生成焓处于中间范围($-200 \sim 200$ kcal/mol)的分子数量最多。此类分子在能量和稳定性之间取得平衡,适用于训练具有宽生成焓范围预测能力的机器学习模型。

为了获得一个数量充足且种类丰富的含能材料分子候选集,从 PubChem 公开数据库^[37]上进行了检索和多级过滤,构建流程如图 2 所示。首先,基于含能材料分子常见的官能团特征,通过化合物名称过滤了包含硝基(nitro)、硝酸酯基(nitrate)、氰基(cyano)和叠氮基(azido)的化合物;然后,排除了带电体系(即形式电荷不为 0 的分子)和同位素体系(即同位素原子数不为 0 的分子);随后,进一步限定化合物仅由 C、H、O、N 四种化学元素组成;在此基础上,限定化合物的非氢重原子数目 ≤ 20 ;最终,过滤并去重后筛选得到了 225543 个候选分子(删除无效 Mordred 方法的 SMILES 特征值后得到 221738 个),由此构建得到了供筛选的含能材料分子数据集。

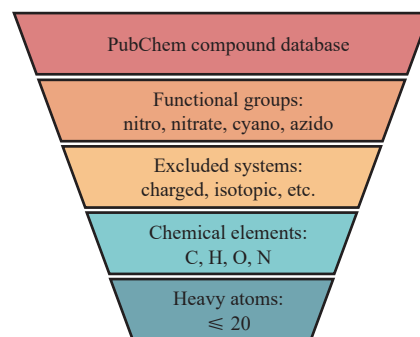


图2 从 PubChem 化合物数据库经过多级过滤得到候选含能分子的流程示意图

Fig. 2 Flowchart of obtaining candidate energetic molecules through multi-stage filtering from PubChem compound database

1.2 SMILES 特征和数据预处理

SMILES(Simplified Molecular Input Line Entry System)是上个世纪八十年代由 David Weininger 提出的一种基于 ASCII 字符串的分子结构表示方法^[38-39]。它使用由字母、数字和特殊符号组成的序列唯一地描述分子中的原子、化学键、支链和环结构等,具有良好的可传输性与可处理性,广泛应用于化学信息系统的数据存储和交换。此外,SMILES 可通过 RDKit^[40]、Mordred^[34]等工具进行分子描述符或特征提取和计算,以支持机器学习模型的构建;也可借助 RDKit、JSME^[41]等化学信息学软件转换为二维结构图像进行可视化。

在获取 PubChem 候选含能分子集后, 首先提取了各分子的 SMILES 序列, 并利用 Mordred 软件计算了所有可用的分子描述符, 包括宏观属性、拓扑指数、连通性指数以及诸如 HOMO/LUMO 能量和极化率等电子结构参数等。在初步得到的 1613 个原始特征中, 首先, 对计算失败和无效描述符(取值为 NaN)进行了删除; 剔除了常数列; 同时, 仅保留与生成焓相关性较大的特征剔除“冗余”特征(即特征之间 Pearson 相关系数大于 0.8 时); 此外, 剔除与生成焓 Pearson 相关系数小于 0.15 的特征。经过上述一系列的数据清洗操作, 最终获得一个包含 93 个特征的数据集。这些特征的具体描述详见支撑材料表 S3。图 3 展示了与生成焓具有最高 Pearson 相关性的前 30 个特征的相关系数热力图。可以看到, 这些特征与生成焓之间存在显著统计相关性, 表明通过 Mordred 提取

的特征能够有效反映生成焓的变化, 为构建高精度生成焓预测模型及含能材料筛选提供了可靠基础。基于高相关性特征的分析(图 3)表明, 生成焓由分子多尺度的结构与性质共同决定。其中, 氮原子的中心作用尤为突出, MID_N 的最高相关性, NdsN/NaasN 的贡献, 共同强调了氮元素的数量与键合形式是决定性因素。同时, nHRing/n5HRing 则标志着复杂环状骨架的重要性。除了这些直观的化学组成, 模型的预测能力更源于对深层物理化学因素的捕捉: AATS0d、BCUTdv-II 等电性特征反映了关键的微观电子效应, 而 Xch-6d、MIC5 等拓扑特征则量化了分子的三维拓扑复杂性与整体形状。该特征集合从组成、结构、电性与拓扑多个维度提供了互补信息, 共同奠定了模型卓越性能与高可解释性的物理基础。

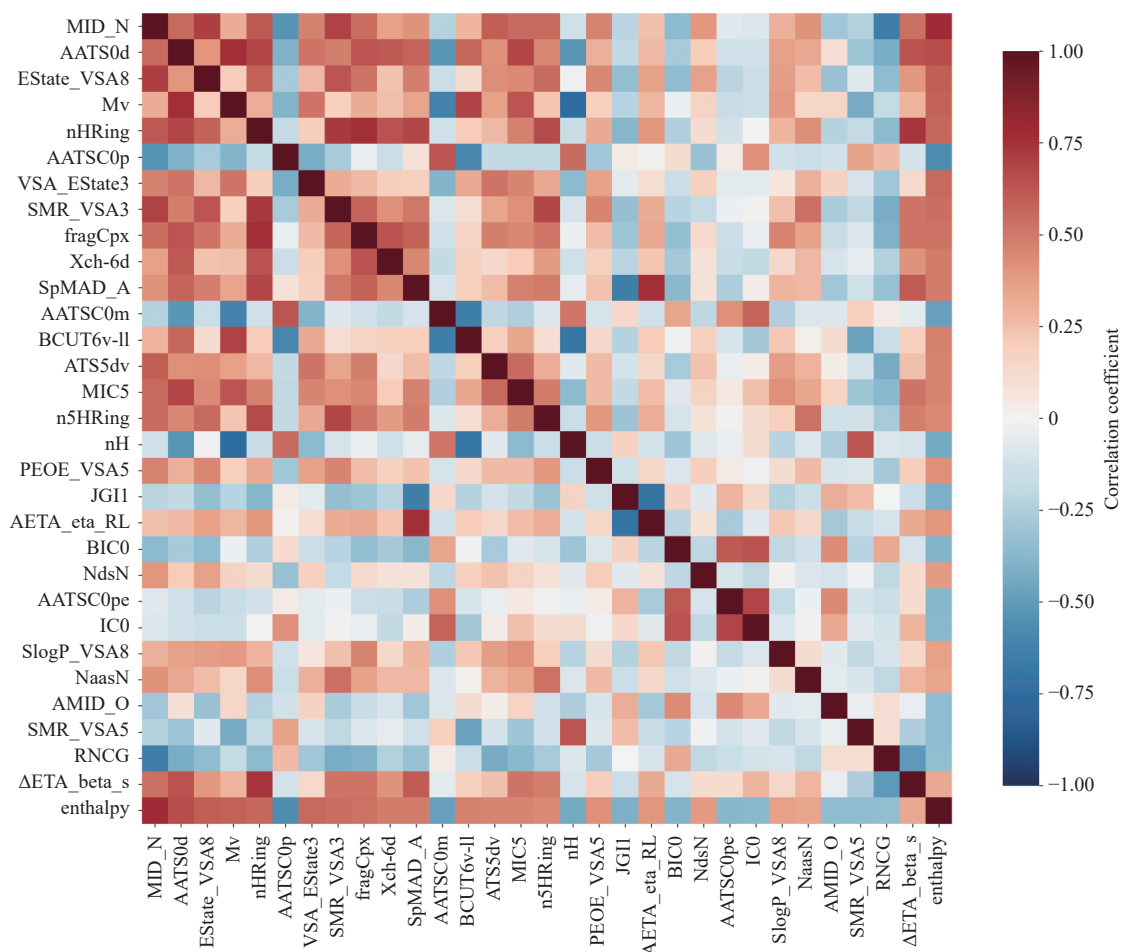


图 3 前 30 个特征的皮尔逊相关系数热力图

Fig. 3 Heat map of Pearson correlation coefficient for the top 30 features

1.3 生成焓预测模型方法及评估指标

在上述生成焓数据集构建以及特征筛选的基础上, 系统比较了六种经典的机器学习回归算法, 包括线性回归模型(Linear)、随机森林(RF)、K近邻(KNN)、梯度提升回归(GBR)、支持向量回归(SVR)以及神经网络(NN)。所有模型的训练均基于自主编写的 Python(版本 3.11)代码并借助 scikit-learn^[42](版本 1.6.1)完成。在模型训练阶段, 按 7:3 的比例使用预处理后的数据集划分为训练集和测试集。为了检验和评估模型结果, 选取平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)作为主要的评估指标。其中, MAE 表示的是预测值与真实值之间绝对误差的平均值; RMSE 表示预测值和真实值平方误差的离散程度; 决定系数 R^2 则是评价模型拟合优度的关键指标。MAE 和 RMSE 的值可以通过以下计算公式获得:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

式中: n 为样本的数量, y_i 和 $\hat{y}_i$ 分别是第 i 个样本的真实值和预测值。MAE 和 RMSE 的值越小, 意味着模型预测越准确。而决定系数 R^2 的计算公式如下:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \quad (3)$$

式中: SS_{res} 和 SS_{tot} 分别为残差平方与总平方和, 是反映真实值与预测值、真实值与平均值间的平方差异。 R^2 值为 0-1, 取值越接近 1 表示模型的拟合优度越好。

1.4 爆炸热及威力指数的计算方法

在一个给定的爆炸反应中, 热量和气体同时释放。因此, 为了获得性能优异的含能材料候选分子, 在生成焓预测的基础上进一步计算了 $\text{C}_a\text{H}_b\text{N}_c\text{O}_d$ 类含能材料的分解反应产生的气体体积(V_g)并估算了相应的爆炸热(ΔH_e)。在本工作中, 采用 Kistiakowsky 和 Wilson 发展的 K-W 规则^[43]来确定分解反应的产物, 具体规则如下: C 原子都转化为 CO; 如果有任何 O 残留, 则 H 氧化为 H_2O ; 如果仍有任何 O 残留, 则 CO 进一步氧化为 CO_2 ; 所有 N 原子转化为 N_2 。基于此, 即可计算出爆炸

产生的气体产物在标准温度和压强条件(298 K 和 1 atm)下的体积 V_g 。在标准工况条件下, 1 摩尔气体占 22.4 dm^3 的摩尔气体体积。爆炸热 ΔH_e 则可以通过反应物和产物的生成焓基于以下公式获得:

$$\Delta H_e = - \frac{\sum_{\text{products}} \Delta H_f - \sum_{\text{reactants}} \Delta H_f}{M} \quad (4)$$

式中: $\sum_{\text{products}} \Delta H_f$ 和 $\sum_{\text{reactants}} \Delta H_f$ 分别是产物和反应物的生成焓之和。 M 是反应物的摩尔质量。

爆炸威力的估计值 P_e 可以通过 V_g 和 ΔH_e 的乘积获得。基于此, 给定爆炸物的 P_e 值即可以和一个标准的爆炸物(这里采用 TNT 为标准物)的 P_e 值作对比, 从而得到 TNT 当量的威力指数 $P_{e(\text{TNT})}$:

$$P_{e(\text{TNT})} = \frac{\Delta H_e \times V_g}{\Delta H_e(\text{TNT}) \times V_g(\text{TNT})} \quad (5)$$

2 结果与讨论

2.1 机器学习模型评估

首先, 采用 Linear、RF、KNN、GBR、SVR 以及 NN 共六种常用的回归模型, 对构建的 1447 个分子的生成焓数据集进行了训练和测试。图 4 展示了各模型在训练集(红色)和测试集(蓝色)上的预测误差分布。从误差分布可以看出, 所有模型均表现出较为优异的预测性能, 其中, Linear、SVR 和 NN 的表现最佳, 其 MAE 分别为 8.83、8.53 和 9.58 kcal/mol, RMSE 分别为 14.74、14.67 和 14.55 kcal/mol, R^2 值均高于 0.96, 性能优于 Elton 等^[28]和 Xie 等^[30]所报道的含能材料生成焓预测模型。相比之下, KNN 模型预测性能相对最差, 其各项指标均显著低于其他模型, 且测试集 R^2 (0.84)明显低于训练集(0.95), 表明存在过拟合。这可能是由于 KNN 作为一种基于距离度量的学习算法, 对高维特征的非线性关系捕捉能力有限, 且易受噪声干扰。基于树模型的 RF 和 GBR 模型性能中等, 虽略逊于最优模型, 但仍能较好建立结构-性质映射关系, 其训练集与测试集性能的较大差距也提示可能存在一定过拟合。

综上所述, 基于 Mordred 方法提取的分子描述符特征, 结合 Linear、SVR 和 NN 等机器学习算法, 能够有效建立含能材料分子结构与生成焓之间复杂的构效关系, 形成预测性能优异的模型。

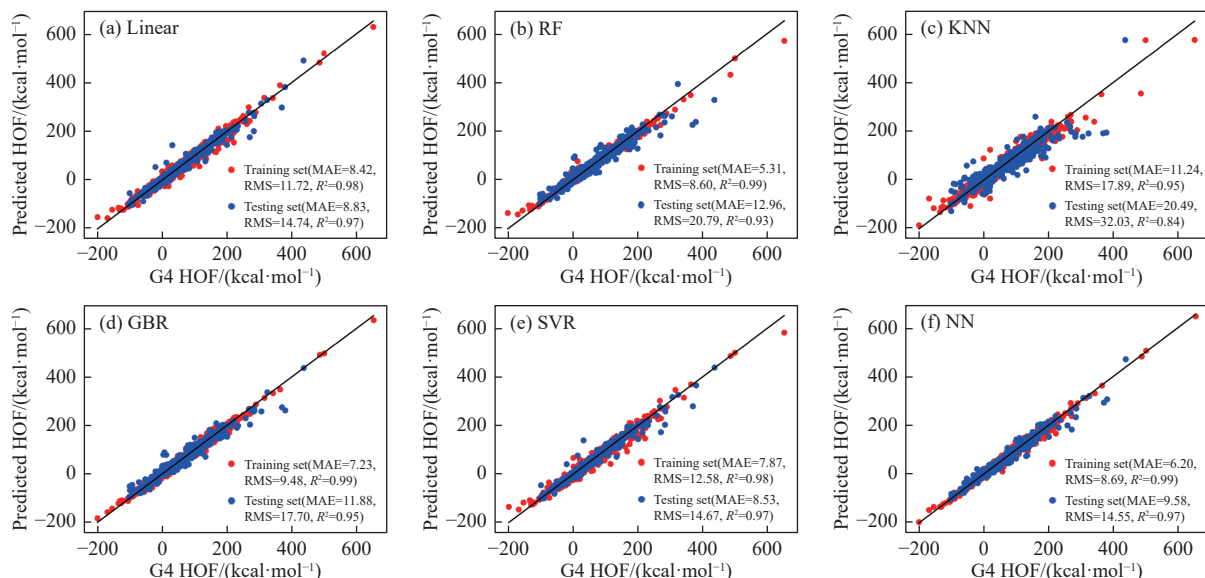


图4 六种机器学习模型训练和测试结果的误差相关图

Fig. 4 Error correlation plots of training set and testing set of the six machine-learning model

这也表明, 经筛选的描述符在关联生成焓性质方面具有强大潜力, 即使配合简单线性模型仍能保持较高预测精度。依据奥卡姆剃刀原理, 本研究最终选用兼具良好泛化能力与高可解释性且简洁的线性模型, 用于后续含能材料筛选任务。

2.2 主动学习策略构建生成焓预测模型

需要指出的是, 生成焓是指在标准条件(298 K, 1 atm)下, 由各元素最稳定单质生成 1 mol 化合物时所吸收或释放的能量, 是反映分子内部化学势能的关键热力学参数。然而, 含能材料在实际应用中的性能受多种因素共同影响, 仅凭生成焓难以全面评估其爆炸性能。相比之下, 爆炸热直接反映了单位质量炸药在绝热条件下爆炸并冷却至初始温度时所释放的热量, 是衡量含能材料做功能力的关键指标, 因而更适用于评价其综合性

能。为此, 本研究参考文献[26,44]中已有方法, 在生成焓预测模型基础上, 结合 K-W 规则及公式(4)和(5)进一步估算了候选分子的爆炸热及其相对于 TNT 的威力指数 $P_{e(TNT)}$ 。同样需要说明的是, 由于该方法的爆炸热计算仍采用 298 K 的气体生成焓, 其本质是理想气体条件下的理论值, 有别于实际物相条件的实验值。通过对从 PubChem 中提取的含能分子材料筛选数据集进行爆炸热和 $P_{e(TNT)}$ 的估算, 获得 $P_{e(TNT)}$ 大于 2.0 的潜在高能候选分子, 并对这些分子的生成焓进行了 G4 理论计算。与预测的生成焓对比发现, 预先构建的线性模型在候选的高能分子生成焓预测中表现有所下降, 其 MAE 和 RMSE 分别为 19.53 和 27.10 kcal/mol, 如图 5(a) 所示。

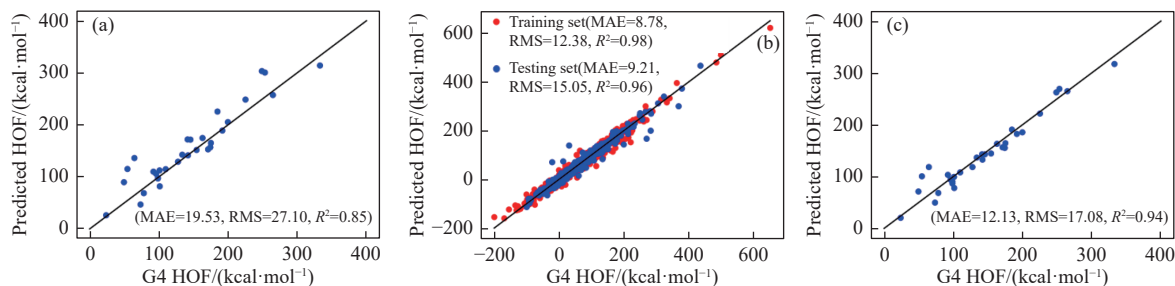


图5 (a)线性模型对候选分子的预测误差相关图; (b)主动学习后的线性模型训练和测试结果的误差相关图; (c)主动学习后的线性模型对候选分子的预测误差相关图

Fig. 5 Error correlation plots of (a) prediction errors for the candidate molecules using the linear model; (b) the training and testing results of the linear model after active learning; (c) prediction errors for the candidate molecules using the linear model after active learning

究其原因,可能是初始训练集与含能材料筛选集之间的数据分布差异导致的,这是一种常见的“领域适应”问题。为此,采用主动学习策略,将预测中绝对误差大于 10 kcal/mol 的 18 个候选样本加入训练集并重新线性训练模型。如图 5(b)所示,经过一次迭代的新模型在测试集上的表现仍旧保持良好的性能,MAE 和 RMSE 分别为 9.21 和 15.05 kcal/mol, R^2 为 0.963。同时,其对候选分子的预测精度显著提高,MAE 和 RMSE 分别降至 12.13 和 17.08 kcal/mol,而 R^2 从 0.85(图 5(a))提高至 0.94(图 5(c))。

为进一步验证模型的泛化能力与鲁棒性,选取了 TNT、RDX、TNAZ (1,3,3-三硝基氮杂环丁烷)和 FOX-7 (1,1-二氨基-2,2-二硝基乙烯)四种经典爆炸分子进行测试。其中,TNT 与 RDX 包含在初始数据集内,而 TNAZ 与 FOX-7 则为模型未见的新分子。预测结果显示,模型对这四种分子的生成焓预测均与参考值高度吻合。特别值得注意的是,对于未参与训练的 TNAZ 和 FOX-7,其预测值(分别为 35.46 与 2.69 kcal/mol)与真实值(22.01 与 -0.25 kcal/mol)非常接近。这表明,本工作的模型不仅对训练集分布内的数据具有良好拟合,对未知分子也展现出优异的预测精度。

综上所述,通过主动学习策略不断融合高误差样本并迭代再次训练,有效提升了生成焓预测模型对候选分子的泛化能力与鲁棒性,为含能材料的高通量筛选提供了更可靠的模型基础。这种“模型预测-识别不确定性-实验/高精度计算验证-模型更新”的主动学习迭代流程,构成一个高效的闭环优化系统,为解决含能材料生成焓预测模型中‘小数据’的普适性问题提供了一条可行路径。

2.3 高能材料的筛选

在验证生成焓预测模型的精确性和鲁棒性后,再次对全部筛选集中的含能分子进行了系统性的筛选。当设定威力指数 $P_{\text{e(TNT)}}$ > 1.5 时,共筛选出 420 种候选的含能分子;进一步提高阈值至 2.0 后,潜在候选集缩减至 20 个分子。表 1 列出了这 20 个分子的关键信息,包括分子式、PubChem

CID、2D 分子结构、生成焓、爆炸热、爆炸产生的气体体积和相对于 TNT 的威力指数等。其中,CID 编号为 129691886 的硝基四唑啉已在之前文献[26]中被报道为一种潜在爆炸物,而其余分子在现有含能材料研究报道中极少见到。

由表 1 可见,所有筛选出的分子爆炸热均高于 4000 kJ/kg,与典型含能材料(如 RDX: 5075.5 kJ/kg, FOX-7: 3881.13 kJ/kg 等)的性能特征相符。其中,(*E*)-叠氮基-[(*E*)-二氮烯基二氮烯基]二氮烯(分子式为 HN_9 , CID 编号为 59708004)的爆炸热最高,达 8722.84 kJ/kg。值得注意的是,该分子具有高达 99.2% 的含氮量,属于典型的多氮化合物,具有高氮低碳氢的特点,易于实现氧平衡,且爆炸产物以 N_2 为主,有望发展成为一类新型“绿色含能材料”。此外,多个其他分子(如 CID 编号为 148241938、154211030 等)的氮含量也超过 50%,且其威力指数普遍大于 2.5,显示出作为高能材料的潜力,值得结合其他研究进一步优化和筛选。另一方面,生成焓最低的是(CID 编号 129689476),其值(23.67 kcal/mol)接近 TNAZ,但其氧平衡约为 -22.42%,呈现为负氧平衡,化学性质可能更加稳定且机械感度可能也更低。若将威力指数阈值放宽至 1.9,则可筛选出生成焓更低的分子,例如二羟基(甲氧亚氨基)甲醇,其预测生成焓为 -5.10 kcal/mol(G4 计算值为 4.32 kcal/mol),低于 TNT(5.49 kcal/mol),意味着其合成路线可能更简单、安全性更高,更易于制备。该分子氧平衡约为 -17.38%,更接近零氧平衡,爆炸时能量利用率更高,具备替代 TNT 的潜力。此外,筛选结果中也包含氧平衡接近 0 的分子(CID 编号 148548035),但其生成焓较高,可能在实际合成中面临挑战。

值得注意的是,尽管上述筛选出的候选分子在理论上表现出优良性能,其实际应用仍需通过实验进一步验证和评估其相关物理化学性质。综上,基于主动学习策略训练并构建的生成焓预测模型,结合爆炸热与威力指数的评估,能够有效支撑多样化性能需求的含能材料分子筛选,为新型含能材料的设计与开发提供新思路 and 有力工具。

表 1 从候选集中筛选出的 20 个 $P_{\text{e(TNT)}} > 2.0$ 的候选分子
 Table 1 Screened out 20 candidates from the filter set with $P_{\text{e(TNT)}} > 2.0$

Formula	PubChem CID	2D molecular depiction	$\Delta H_f^a / (\text{kcal} \cdot \text{mol}^{-1})$	$\Delta H_c / (\text{kJ} \cdot \text{kg}^{-1})$	$V_g / (\text{dm}^3 \cdot \text{kg}^{-1})$	$P_{\text{e(TNT)}}$
$\text{CH}_6\text{N}_6\text{O}_3$	153952084		71.17 (49.35)	5945.75	1044.65	2.01
$\text{C}_2\text{H}_5\text{N}_5\text{O}_2$	57115395		108.05 (109.65)	5195.72	1196.08	2.02
$\text{H}_5\text{N}_7\text{O}$	148241938		155.42 (174.33)	7476.50	1128.58	2.74
$\text{H}_5\text{N}_5\text{O}_3$	87424472		68.53 (78.04)	7206.63	910.04	2.13
$\text{CH}_2\text{N}_2\text{O}$	172554591		49.77 (73.34)	5560.76	1157.82	2.09
$\text{CH}_3\text{N}_5\text{O}_3$	140363114		86.64 (98.65)	7354.06	841.68	2.01
CH_3N_{11}	140914248		317.28 (332.56)	7849.79	927.20	2.36
$\text{C}_4\text{H}_3\text{N}_5\text{O}_2$	129777279		262.75 (248.76)	8676.56	877.85	2.47
HN_9	59708004		264.92 (264.84)	8722.84	881.40	2.49
CHN_3O_2	129762552		78.20 (101.30)	8079.04	772.08	2.02
$\text{C}_2\text{H}_3\text{N}_{11}\text{O}_4$	148548036		269.21 (253.26)	7575.72	822.46	2.02
$\text{CH}_7\text{N}_7\text{O}_3$	129640395		99.42 (101.01)	6121.00	1085.32	2.15
$\text{CH}_3\text{N}_5\text{O}_2$	129691886		100.67 (54.55)	6626.80	956.71	2.06
CHN_3O	129813566		103.59 (91.92)	7713.16	945.96	2.37
$\text{CH}_5\text{N}_3\text{O}_3$	129689476		20.66 (23.67)	6361.31	1046.05	2.16
CH_3N_7	129767002		182.25 (191.63)	6743.12	990.41	2.17
$\text{H}_4\text{N}_6\text{O}$	154211030		136.78 (133.67)	7805.55	1076.17	2.72
$\text{CH}_2\text{N}_8\text{O}_3$	148548035		185.61 (199.70)	8123.16	772.06	2.03
CHN_5	154414430		156.59 (170.94)	7888.61	809.11	2.07
CH_2N_6	153744440		164.71 (175.14)	7027.17	913.64	2.08

^a: The values in brackets are calculated using G4 theory.

3 总结和展望

本工作以含能材料为研究对象, 提出了一种基于主动学习策略结合 SMILES 分子特征的构建机器学习生成焓构效关系模型的新方法, 实现了大规模含能材料分子的高效筛选。本研究首先基于 G4 高精度量子化学方法构建了包含 1447 种含能材料的生成焓数据集, 并利用 Mordred 特征化方法提取 SMILES 分子描述符, 经清洗与预处理后筛选出 93 个有效特征, 并构建了基于线性模型的生成焓初步预测模型; 进一步地, 针对从 PubChem 数据库中获取的 221738 种潜在含能材料分子开展了生成焓与爆炸热的系统预测, 筛选出爆炸威力较优的候选分子, 并对预测误差较高的样本采用主动学习策略进行迭代再训练, 不断优化模型参数, 从而显著提升了生成焓的预测精度。经验证, 该模型在已报道的经典含能分子上的预测表现良好, 展现出较强的泛化能力与可靠性。基于上述方法, 最终从候选分子库中筛选出 20 个威力指数超过 2.0 倍 TNT 的候选分子。这些分子涵盖多种结构类型, 既包括含氮量超过 50%、具有“绿色含能材料”潜力的多氮化合物, 也包含生成焓较低、易于合成且安全性良好的候选物, 显示出良好的应用前景。

相较于传统实验筛选方法的高成本与长周期, 以及现有生成焓预测模型普遍存在的样本有限、化学空间覆盖窄等问题, 本研究提出的主动学习和 SMILES 特征相融合的机器学习构效关系模型新策略, 在大规模含能材料生成焓与爆炸热预测中表现出较高的可靠性与泛化能力, 适用于高通量高效筛选。本研究仍存在一些局限性, 值得在未来工作中进一步探索。首先, 在筛选广度上, 未来可结合生成式模型突破现有化学空间的限制, 主动发掘高性能候选分子; 其次, 在性能维度上, 除能量特性外, 还需将感度、稳定性等关乎实际应用的指标纳入全局视角, 通过多任务学习框架实现综合性能的协同预测与优化。尽管存在上述局限, 本研究无疑为含能材料领域基础物性数据的预测提供了一种高效实用的新方法, 也为高性能含能材料的快速发现与设计提供了新思路。

利益冲突声明

作者声明没有利益冲突

参考文献

- [1] FRIED L E, MANAA M R, PAGORIA P F, et al. Design and synthesis of energetic materials[J]. Annu Rev Mater Res, 2001, 31: 291–321.

- [2] SHUKLA M K, BODDU V M, STEEVENS J A, et al. Energetic materials[M]. Switzerland: Springer International Publishing, 2017.
- [3] PAGORIA P F, LEE G S, MITCHELL A R, et al. A review of energetic materials synthesis[J]. Thermochim Acta, 2002, 384(1): 187–204.
- [4] BADGUJAR D M, TALAWAR M B, ASTHANA S N, et al. Advances in science and technology of modern energetic materials: An overview[J]. J Hazard Mater, 2008, 151(2): 289–305.
- [5] PEDREIRA S M, PINTO J R A, CAMPOS E A, et al. Methodologies for characterization of aerospace polymers/energetic materials - a short review[J]. J Aerosp Technol Man, 2016, 8(1): 18–25.
- [6] ZLOTIN S G, CHURAKOV A M, EGOROV M P, et al. Advanced energetic materials: Novel strategies and versatile applications[J]. Mendelev Commun, 2021, 31(6): 731–749.
- [7] SIKDER A K, SIKDER N. A review of advanced high performance, insensitive and thermally stable energetic materials emerging for military and space applications[J]. J Hazard Mater, 2004, 112(1): 1–15.
- [8] TSYSHEVSKY R, SMIRNOV A S, KUKLJA M M. Comprehensive end-to-end design of novel high energy density materials: III. Fused heterocyclic energetic compounds[J]. J Phys Chem C, 2019, 123(14): 8688–8698.
- [9] MA P, JIN Y T, WU P H, et al. Synthesis, molecular dynamic simulation, and density functional theory insight into the cocrystal explosive of 2, 4, 6-trinitrotoluene/1, 3, 5-trinitrobenzene[J]. Combust Explo Shock+, 2017, 53(5): 596–604.
- [10] TSYSHEVSKY R, PAGORIA P, ZHANG M, et al. Comprehensive end-to-end design of novel high energy density materials: I. Synthesis and characterization of oxadiazole based heterocycles[J]. J Phys Chem C, 2017, 121(43): 23853–23864.
- [11] VISWANATH D S, GHOSH T K, BODDU V M. Emerging Energetic Materials: Synthesis, Physicochemical, and Detonation Properties[M]. Dordrecht: Springer Netherlands, 2018.
- [12] NIELSEN A T, CHAFIN A P, CHRISTIAN S L, et al. Synthesis of polyazapolycyclic caged polynitramines[J]. Tetrahedron, 1998, 54(39): 11793–11812.
- [13] ZANG X, ZHOU X, BIAN H, et al. Prediction and construction of energetic materials based on machine learning methods[J]. Molecules, 2023, 28(1): 322.
- [14] ZEMAN S, JUNGOVÁ M. Sensitivity and performance of energetic materials[J]. Propell Explo Pyrot, 2016, 41(3): 426–451.
- [15] WASHBURN E W. Standard states for bomb calorimetry[J]. J Res Natl Bur Standards, 1933, 10: 525–558.
- [16] SINDITSKII V P, EGORSHEV V Y, LEVSHENKOV A I, et al. Ammonium nitrate: Combustion mechanism and the role of additives[J]. Propell Explo Pyrot, 2005, 30(4): 269–280.
- [17] BYRD E F C, RICE B M. Improved prediction of heats of formation of energetic materials using quantum mechanical calculations[J]. J Phys Chem A, 2006, 110(3): 1005–1013.
- [18] RICE B M, HARE J J, BYRD E F C. Accurate predictions of crystal densities using quantum mechanical molecular volumes[J]. J Phys Chem A, 2007, 111(42): 10874–10879.
- [19] FELLER D. The role of databases in support of computational chemistry calculations[J]. J Comput Chem, 1996, 17(13): 1571–1586.

- [20] VAMATHEVAN J, CLARK D, CZODROWSKI P, et al. Applications of machine learning in drug discovery and development[J]. Nat Rev Drug Discov, 2019, 18(6): 463–477.
- [21] DARA S, DHAMERCHERLA S, JADAV S S, et al. Machine learning in drug discovery: A review[J]. Artif Intell Rev, 2022, 55(3): 1947–1999.
- [22] CARRACEDO-REBOREDO P, LIÑARES-BLANCO J, RODRÍGUEZ-FERNÁNDEZ N, et al. A review on machine learning approaches and trends in drug discovery[J]. Comput Struc Biotech J, 2021, 19: 4538–4558.
- [23] BUTLER K T, DAVIES D W, CARTWRIGHT H, et al. Machine learning for molecular and materials science[J]. Nature, 2018, 559(7715): 547–555.
- [24] SCHLEDER G R, PADILHA A C M, ACOSTA C M, et al. From DFT to machine learning: Recent approaches to materials science—a review[J]. J Phys Mater, 2019, 2(3): 032001.
- [25] NGUYEN P, LOVELAND D, KIM J T, et al. Predicting energetics materials' crystalline density from chemical structure by machine learning[J]. J Chem Inf Model, 2021, 61(5): 2147–2158.
- [26] KANG P, LIU Z, ABOU-RACHID H, et al. Machine-learning assisted screening of energetic materials[J]. J Phys Chem A, 2020, 124(26): 5341–5351.
- [27] 张炳儒, 李明, 文玉史, 等. 基于音频信号的含能材料撞击感度机器学习识别[J]. 含能材料, 2025, 33(2): 136–147. (ZHANG Bingru, LI Ming, WEN Yushi, et al. Machine learning recognition of impact sensitivity of energetic materials based on acoustic signals[J]. Chin J Energ Mater (Hanneng Cailiao), 2025, 33(2): 136–147.)
- [28] ELTON D C, BOUKOUVALAS Z, BUTRICO M S, et al. Applying machine learning techniques to predict the properties of energetic materials[J]. Sci Rep, 2018, 8(1): 9059.
- [29] CHEN C, LIU D, DENG S, et al. Accurate machine learning models based on small dataset of energetic materials through spatial matrix featurization methods[J]. J Energy Chem, 2021, 63: 364–375.
- [30] XIE Y, LIU Y, HU R, et al. A property-oriented adaptive design framework for rapid discovery of energetic molecules based on small-scale labeled datasets[J]. RSC Adv, 2021, 11(41): 25764–25776.
- [31] CASEY A D, SON S F, BILIONIS I, et al. Prediction of energetic material properties from electronic structure using 3D convolutional neural networks[J]. J Chem Inf Model, 2020, 60(10): 4457–4473.
- [32] ZHANG D, CHU Q, CHEN D. Predicting the enthalpy of formation of energetic molecules via conventional machine learning and GNN[J]. Phys Chem Chem Phys, 2024, 26(8): 7029–7041.
- [33] CURTISS L A, REDFERN P C, RAGHAVACHARI K. Gaussian-4 theory[J]. J Chem Phys, 2007, 126(8): 084108.
- [34] MORIWAKI H, TIAN Y-S, KAWASHITA N, et al. Mordred: A molecular descriptor calculator[J]. J Cheminf, 2018, 10(1): 4.
- [35] FRISCH M J, TRUCKS G W, SCHLEGEL H B, et al. Gaussian 09, Revision E. 01[CP]. Gaussian, Inc: Wallingford, CT, 2013.
- [36] OCHTERSKI J W. Thermochemistry in Gaussian[Z]. Gaussian Inc., 2000, 1–19. <https://gaussian.com/thermo/>.
- [37] KIM S, CHEN J, CHENG T, et al. PubChem 2023 update[J]. Nucleic Acids Res, 2022, 51(D1): D1373–D1380.
- [38] WEININGER D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. J Chem Inf Comput Sci, 1988, 28(1): 31–36.
- [39] WEININGER D, WEININGER A, WEININGER J L. SMILES. 2. Algorithm for generation of unique SMILES notation[J]. J Chem Inf Comput Sci, 1989, 29(2): 97–101.
- [40] RDKit: Open-source cheminformatics[EB/OL]. <https://www.rdkit.org>.
- [41] BIENFAIT B, ERTL P. JSME: A free molecule editor in JavaScript[J]. J Cheminf, 2013, 5(1): 24.
- [42] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. J Mach Learn Res, 2011, 12: 2825–2830.
- [43] AKHAVAN J. The Chemistry of Explosives 4E[M]. London: Royal Society of Chemistry, 2022.
- [44] MUTHURAJAN H, SIVABALAN R, TALAWAR M B, et al. Prediction of heat of formation and related parameters of high energy materials[J]. J Hazard Mater, 2006, 133(1): 30–45.

作者简介:



杨琳(1989), 博士, 主要从事生存分析理论方法研究和深度学习、机器学习等统计方法的改进与应用, ylintj@126.com。



滕波涛(1974), 博士, 天津市特聘教授, 博士生导师, 主要从事理论指导实验的催化剂与吸附材料模拟、设计与开发、模型催化剂谱学的理论与实验研究, tbt@tust.edu.cn。



周余伟(1986), 博士, 高级工程师, 主要从事量子化学计算和人工智能/机器学习方法在化学、材料科学中的研究和应用工作, zhouyuwei@synfuelschina.com.cn。