

文章编号:1006-9941(2024)06-0573-11

机器学习辅助的[5,6]稠环含能化合物高通量设计

潘林虎,王睿辉,樊明仁,宋思维,王毅,张庆华

(西北工业大学航天学院, 陕西 西安 710072)

摘要: 与经验和计算指导的研发模式相比,机器学习辅助的含能分子高通量虚拟筛选技术,在分子设计效率及构效关系定量分析方面都展现出明显优势。鉴于富氮稠环含能化合物较好的能量-稳定平衡特性,研究利用机器学习辅助的高通量虚拟技术对[5,6]富氮稠环类含能分子的化学空间进行了探索研究,基于[5,6]全碳骨架,通过组合枚举和芳香性筛选得到142689个[5,6]稠环类化合物,同时采用核岭回归算法建立并优化了6个含能分子性能预测模型(密度,分解温度,爆速,爆压,撞感和生成焓),分析了稠环上的氮氧原子以及分子上官能团对含能化合物性能的影响。结果发现,所生成稠环化合物的构效关系与含能化合物能量与稳定性相关性的一般规律相符,验证了模型的合理性。以爆速和分解温度作为能量和热稳定性的标准,研究进而筛选获得了5个综合性质较为突出的分子,利用DFT等量子化学计算的结果与本研究模型预测结果符合良好,进一步验证了预测模型的精度。

关键词: 机器学习;高通量筛选;核岭回归;分子设计;[5,6]稠环含能化合物

中图分类号: TJ55

文献标志码: A

DOI: 10.11943/CJEM2024055

0 引言

含能化合物在武器装备和航天推进等领域应用广泛,其分子结构中所蕴含的化学能是毁伤和推进的主要能量来源^[1-2]。19世纪末以来,含能化合物的设计和合成主要依靠科研人员的经验试错,例如三硝基甲苯(TNT)、三氨基三硝基苯(TATB)、黑索今(RDX)、奥克托今(HMX)等性能优异的含能化合物,其发现和利用都离不开反复的实验探索^[3-5]。然而,由于含能化合物具有的亚稳态特性,在无法掌握化合物能量及安全特性的前提下,基于实验试错的传统含能化合物研发模式具有很高的危险性和偶然性。因此,如量子化学、分子动力学等的计算化学方法被广泛应用于指导含能

化合物的设计和评估。但由于耗时较长、计算成本昂贵和使用门槛高等问题,计算化学方法在满足高通量的含能分子设计及筛选需求方面仍有一定困难^[6-9]。

随着大数据时代的到来,含能材料基因工程、数据驱动的含能化合物研发等新概念被先后提出,为含能化合物设计、制备以及性能预测等相关研究提供了新的方向和动力^[9-11]。借助材料基因工程、机器学习和高通量筛选等先进理念及方法,近年来含能化合物研究取得了丰硕成果^[12-16]。例如,相比传统含能化合物设计,Wen等^[17]提出碎片组合设计和高通量设计的方法,得到2000个具有与CL-20相当的目标分子;Li等^[18]构建了由3个利用迁移学习关联的循环神经网络(RNN)模型,在有限数据集下生成7153个高能化合物。与此同时,机器学习也被用于分子的性能预测。例如,Li等^[19]利用随机森林(RF)算法成功预测含能分子的密度,决定系数(R^2)和均方根误差(RMSE)分别为0.9768,0.0578 $\text{g}\cdot\text{cm}^{-3}$;Phan等^[20]利用消息传递神经网络(MPNN)算法预测晶体密度, R^2 和RMSE分别为0.914,0.044 $\text{g}\cdot\text{cm}^{-3}$ 。基于机器学习和高通量设计,Liu等^[14, 21-22]搭建了EM-Studio平台,可快速实现分子三维结构的构建、爆轰性能预测等功能,其中的分子设计模块采用RNN算法生成新分子,利用定向消息

收稿日期: 2024-02-07; 修回日期: 2024-03-26

网络出版日期: 2024-06-05

基金项目: 国家自然科学基金(22205218, 22075259, 22175157)

作者简介: 潘林虎(1999-),男,硕士研究生,主要从事含能材料机器学习研究, e-mail: xinbopan@163.com

通信联系人: 宋思维(1992-),男,副教授,主要从事含能材料设计计算及合成研究。 e-mail: ssw_sv@nwpu.edu.cn

王毅(1988-),男,教授,主要从事含能材料设计及合成研究。 e-mail: ywang0521@nwpu.edu.cn

引用本文: 潘林虎,王睿辉,樊明仁,等. 机器学习辅助的[5,6]稠环含能化合物高通量设计[J]. 含能材料, 2024, 32(6):573-583.

PAN Lin-hu, WANG Rui-hui, FAN Ming-ren, et al. Machine Learning Assisted High-throughput Design of [5,6] Fused Ring Energetic Compounds[J]. *Chinese Journal of Energetic Materials (Hanneng Cailiao)*, 2024, 32(6):573-583.

传递神经网络架构(D-MPNN)算法构建爆轰性能预测模型,其模型的 R^2 均在0.9以上。上述研究表明,基于机器学习的含能化合物设计和性能预测,是实现含能化合物高效研发的途径之一^[23-26]。然而在含能化合物的设计与合成研究中,应秉持能量(高爆轰)与热稳定性(耐热)的双重优化原则,旨在实现含能化合物安全可靠、持久性能、环境友好性的全方位提升^[27-29]。

富氮稠环类含能化合物键能较高,具有更高的氮含量和生成焓,且整个骨架结构近似处于平面,有助于形成 π - π 相互作用、降低化合物机械感度,此外其爆轰产物中氮气含量高,更加绿色环保,因而近年来富氮稠环类含能化合物的设计及合成引起了含能材料领域的广泛兴趣^[30-31]。

基于此,本研究从[5,6]碳环骨架出发,进行含能分子高通量设计,生成得到了142689个[5,6]稠环化合物分子,利用核岭回归算法建立优化了6个含能分子性能预测模型,对生成分子的密度、分解温度、爆速、爆压、撞击感度和生成焓进行预测,以高能量($D_v > 9000 \text{ m} \cdot \text{s}^{-1}$)和高热稳定性($T_d > 200 \text{ }^\circ\text{C}$)为标准,筛选得到了5个综合性能优异的分子。研究可为含能分子数据库的构建奠定化学结构基础,同时也可含能化合物的结构设计提供参考。

1 机器学习辅助的高通量设计方法

1.1 数据收集和特征提取

数据量是含能化合物机器学习面临挑战之一。目

前含能化合物数据库的构建方法主要包括2种,一种是人工搜集相关文献数据,另一种是从公共数据库中提取数据。研究通过手动搜集文献,整理了一个含有1000多种含能分子的数据库,该数据库包含各类化合物,如脂肪族化合物和芳香族化合物等。数据库最终获得密度数据1487个,爆速1045个,分解温度1027个(分解温度包含初始分解温度和峰值),爆压1026个,生成焓961个,撞击感度1005个,具体数据集所包含结构及对应性能参数见支撑文件附录1。在此基础上,为保证数据质量,采用异常检测算法(Isolation Forest)进行数据预处理,去除明显结构和数据分布异常的样本。

准备好训练数据后,需选择合适的分子表示方法以构建分子结构特征。研究以简化分子线性输入规范(SMILES)作为特征化的起点,通过Open-Source Cheminformatics Software(RDKit)实现分子三维结构生成并采用MMFF94S、UFF力场对分子结构进行优化^[32-34],在此基础上构建了由电拓扑(E-state)指纹和自定义描述符组成的复合描述符集对分子进行表示。电拓扑指纹包括特定类型的原子数量和相应的电拓扑状态指数,自定义描述符则强化对含能特征、基团特征以及分子间弱相互作用特征的描述,具体描述符集见表1。

1.2 机器学习模型构建

含能化合物性能预测通常属于回归问题,常见的回归算法包括K最近邻(KNN)^[35]、岭回归(RR)^[36]、核岭回归(KRR)^[37]和神经网络(MLP)^[38]等。为选择最

表1 自定义分子描述符

Table 1 Customized molecular descriptors

abbreviation	description	abbreviation	description
nHbondA	number of hydrogen bond acceptor	nH	number of hydrogen atom
nHbondD	number of hydrogen bond donor	nC	number of carbon atom
nNH ₂	number of amino group	nN	number of nitrogen atom
nAHC	number of aromatic heterocycle	nO	number of oxygen atom
nACC	number of aromatic carbocycle	PBF	plane of best fit
nRbond	number of rotatable bond	TPSA	topological polar surface area
nR	number of ring	OB	oxygen balance
nNNO ₂	number of nitramine group	molecular weight	molecular weight
nCNO ₂	number of nitro group	PMI ₃	principal moments of inertia 3
nC(NO ₂) ₃	number of nitroform group	nCH ₃	number of methyl group
nC(NO ₂) ₂	number of dinitro group	nOCH ₃	number of methoxy group
nCNO ₂	number of nitro group	NPR1	normalized principal moments ratios 1
MinPartialCharge	minimum value of partial charge	NPR2	normalized principal moments ratios 2
MaxPartialCharge	maximum value of partial charge	MOLvolume	molecular volume

合适机器学习模型的回归算法,研究首先在部分数据集对上述算法的预测效果进行测试,得到了决定系数(R^2)和平均绝对误差(MAE)的数据结果,如图1所示,由图1可以看出,核岭回归(KRR)算法(红色折线)在决定系数(R^2)中表现最好,平均绝对误差(MAE)中,除了 T_d 误差较大,其余性能预测模型误差几乎相同。K最近邻(KNN),前馈神经网络(FNN),岭回归(RR)算法在模型训练中,虽然平均绝对误差(MAE)较小,但决定系数(R^2)不如KRR算法,因此,本研究统一采用核岭回归(KRR)算法训练密度、爆速、爆压、生成焓、撞击感度、分解温度6个含能分子性能预测模型。

为增加6个含能分子性能预测模型的拟合优度,研究基于贝叶斯优化^[39]进行超参数的选择。在训练模型的超参数中,正则化参数(Alpha)用于控制模型复杂度和泛化能力之间的平衡,在[0.0, 15.0]搜索范围进行均值采样。核函数从径向基函数(Rbf)^[40]和多项式函数(Poly)^[41]中选择。Gamma是核函数中内置参数,搜索范围设为[-10.0, 10.0]。Degree表示多项式核函数的阶数,搜索范围设为[0.0, 20.0]。Coef0是多项式核函数中内层函数的常数项,其均值和标准差分别为0.0和10.0。最终迭代产生的含能分子性能

预测模型的最佳超参数取值如表2所示。

表2 6个含能分子性能预测模型的最佳超参数取值

Table 2 Hyperparameter of six models

models	alpha	coef0	degree	gamma	kernel type
ρ _model	7.8330	1.3013	13.8973	0.0044	poly
D_v _model	10.9222	1.7014	9.5205	0.0032	poly
p _model	8.7090	1.2274	10.4000	0.0081	poly
$\Delta H_{f, \text{solid}}$ _model	0.0214	3.2477	10.7618	1.6233	poly
IS _model	5.4577	0.0845	5.0248	0.0309	poly
T_d _model	6.1071	1.2311	9.7314	0.0062	poly

Note: alpha is the regularization strength. coef0 is a constant term in the polynomial kernel function. Degree is the degree of the polynomial kernel function. Gamma is an inherent parameter of the kernel function. The kernel function is selected from radial basis function (Rbf) and polynomial function (Poly).

1.3 模型验证

研究将数据集分为80%的训练集和20%的测试集进行模型训练,采用五折交叉验证方法^[42]避免因数据集划分不合理而导致的模型偏差。为评估模型精度和泛化能力,选取了平均绝对误差(MAE)、平均绝对百分比误差(MAPE)和决定系数(R^2)作为模型性能评价指标,计算公式如式(1)~(3)所示:

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

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (2)$$

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

式中, n 为样本个数, y_i 为真实值, \hat{y}_i 为预测值, \bar{y} 为真实值的平均值。

表3列出了各性能预测模型的预测效果评价指标 R^2 、MAE和MAPE,由表3可知,模型对密度 ρ 、爆速 D_v 和爆压 p 的预测效果评价指标中, R^2 值较高(0.79~0.91)、MAE(0.041 g·cm⁻³, 240.3 m·s⁻¹, 2.194 GPa)和MAPE值(2.4%~8.0%)较低。这说明研究建立的模型对 ρ 、 D_v 和 p 具有非常好的预测能力^[9, 19],这主要归功于数据集和分子描述符的准确性。同时,在生成焓 $\Delta H_{f, \text{solid}}$ 预测方面, R^2 为0.96,MAPE值却为13.9%,这可能因为模型对 $\Delta H_{f, \text{solid}}$ 拟合优度很高,但由于 $\Delta H_{f, \text{solid}}$ 存在较小数值,在该处百分比误差易被放大。同时,为比较预测值和实际值之间的误差,验证本研究含能分子性能预测模型的稳定性,研究绘制了图2所

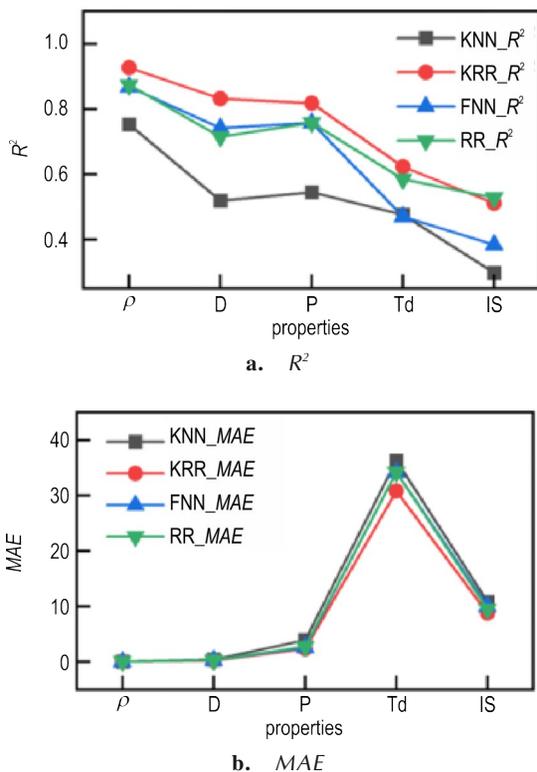
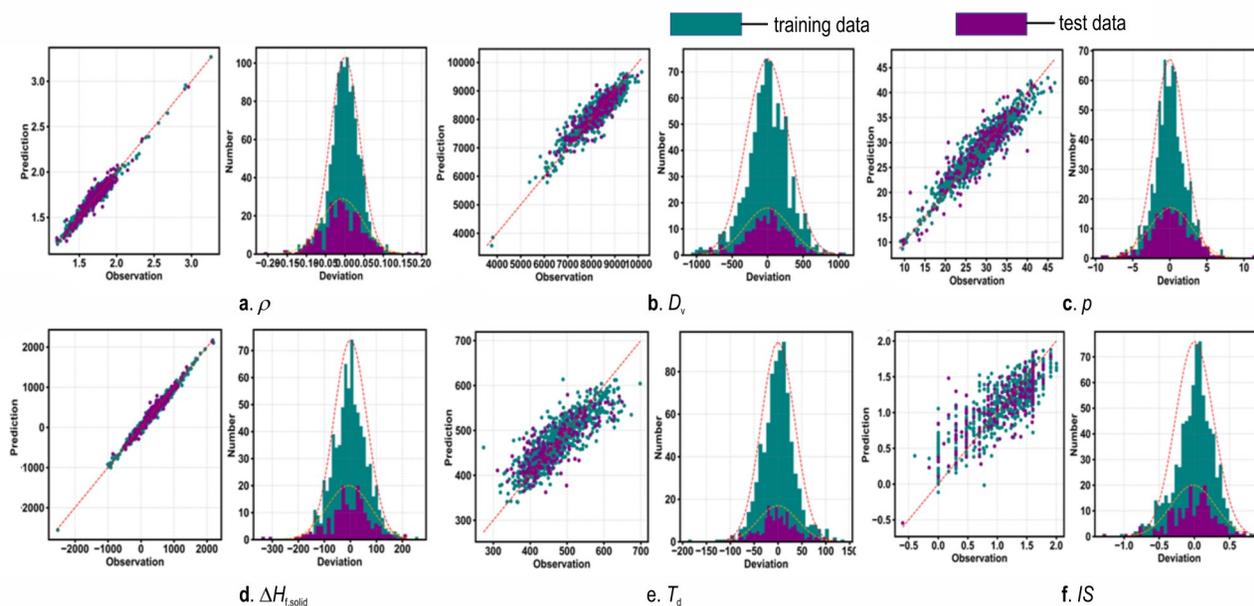


图1 不同回归算法预测效果比较

Fig.1 Comparison of the prediction effect for different regression algorithms

表3 6个含能分子性能预测模型预测效果评价的 R^2 , MAE , $MAPE$ 结果Table 3 R^2 , MAE and $MAPE$ results for six property prediction model

scores	ρ _model	D_v _model	p _model	$\Delta H_{f, \text{solid}}$ _model	T_d _model	IS _model
R^2	0.91	0.79	0.81	0.96	0.63	0.61
MAE	0.041 $\text{g} \cdot \text{cm}^{-3}$	240.3 $\text{m} \cdot \text{s}^{-1}$	2.194 GPa	62.94 $\text{kJ} \cdot \text{mol}^{-1}$	31.24 $^{\circ}\text{C}$	0.24 J
$MAPE / \%$	2.4	3.0	8.0	13.9	6.7	21.3

图2 ρ , D_v , p , $\Delta H_{f, \text{solid}}$, T_d 和 IS 模型预测的结果和文献中报道结果的奇偶图和偏差分布Fig.2 Odd-even diagrams and deviation distributions of density ρ , detonation velocity D_v , detonation pressure p , enthalpy of formation $\Delta H_{f, \text{solid}}$, decomposition temperature T_d and impact sensitivity IS

示的奇偶图进行评估,由图2可以发现,6种性质的模型预测值和实际测试值具有较好的一致性。从图2a中可知,所有点都紧密聚集在一条斜线上,表明模型对于含能化合物的 ρ 预测模型具有很高的精度;右侧子图为预测误差直方图,可知偏差(虚线)基本符合正态分布,表明模型预测结果的误差大多数情况下集中在均值附近,极端误差相对较少。其余图2b~2d也显示模型对 D_v , p , $\Delta H_{f, \text{solid}}$ 具有很好的稳定性和泛化能力。本研究中,分解温度 T_d (图2e)和撞击感度 IS (图2f)预测模型的精度相对较低(R^2 :0.61和0.63, MAE :0.24 J和31.24 $^{\circ}\text{C}$),分析认为主要原因在于含能化合物的稳定性不仅仅决定于分子结构,还受到分子间相互作用、晶体堆积、颗粒尺寸及形状等多尺度因素耦合的影响,因此仅通过分子描述符无法有效表示稳定性相关的结构特征。但总的来说,考虑到训练数据中分子数量和结构的多样性,可认为模型预测结果对评估分子稳定性仍具有一定参考价值^[43]。

为进一步评估本研究建立的含能分子性能预测模

型精度,将本研究建立的模型与部分文献预测模型进行对比研究^[9-10, 12, 18-19, 44-45],结果见表4。由表4可以看到,本研究建立的含能分子性能预测模型对 ρ 的预测效果评价指标($R^2=0.91$, $MAE=0.041$)上远优于Elton等^[9]的研究评价指标($R^2=0.74$, $MAE=0.06 \text{ g} \cdot \text{cm}^{-3}$),在其他性质预测方面(如 D_v , p , $\Delta H_{f, \text{solid}}$ 模型)也展现了较好的效果。Casey等^[10]的模型使用3D卷积神经网络预测含能材料的多重性质,模型精度很高,然而其网络特征的构建需要进行高成本的电子结构计算。此外,Chen^[44]、Hou^[12]的模型数据量较少,且预测的性能只针对含能化合物某一方面(如只关注爆轰性能不关注稳定性),无法对含能化合物的综合性质进行全面评估。可见,本研究建立的含能分子性能预测模型在性能评估的全面性、评估效率及精度方面均展现出了一定的优势。

1.4 含能化合物分子结构高通量设计

研究通过组合枚举方法^[11]进行含能分子结构设计,流程如图3所示。以[5,6]全碳分子作为母环骨架,母环上加入双键(以分子单键数一半为上限),得到

表4 本研究含能分子性能预测模型和已报道模型的对比^[9-10, 12, 18-19, 44-45]

Table 4 Comparison of energy molecule performance prediction models between this study and reported models

model	number of data	algorithms	featurization	R^2 (MAE)					
				ρ _model	D_v _model	ρ _model	$\Delta H_{f, \text{solid}}$ _model	T_d _model	IS _model
this study	1000	KRR	e-state fingerprint, custom descriptors	0.91 (0.041)	0.79 (240.3)	0.81 (2.194)	0.96 (62.94)	0.63 (31.24)	0.61 (0.24)
ELTON ^[9]	109	KRR, ridge, SVR, RF, KNN	custom descriptors rdkit, coulomb matrix	0.74 (0.06)	0.94 (71.41)	—	—	—	—
CASEY ^[10]	26265	3D CNNs	grid data for electron charge density and electrostatic potential	0.943 (0.011)	0.974 (96)	0.965 (0.584)	0.979 (47.09)	—	—
Li ^[18]	303	RNN	SMILES	—	0.9572 (80.1)	—	—	—	—
Li ^[19]	162	RF	topological index, geometric configuration, electrostatic coefficients, quantitative descriptors	0.9596 (0.0437)	—	0.9768 (0.0578)	—	—	—
Chen ^[44]	451	LASSO, KRR, BRR, SVR, RFR, KNN	VOM, HCM	(0.035)	—	—	(40.44)	—	—
Song ^[45]	1000	KRR	e-state fingerprint, custom descriptors	0.930 (0.042)	0.830 (240)	0.820 (2.379)	—	—	—
Hou ^[12]	436	LM	coulomb matrix	0.986 (0.026)	0.928 (345.6)	0.966 (1.493)	—	—	—

Note: KRR is Kernel Ridge Regression. SVR is Support Vector Regression. RF is Random Forest. KNN is K-Nearest Neighbors. 3D CNNs is 3-Dimensional Convolutional Neural Networks. RNN is Recurrent Neural Network. LASSO is Least Absolute Shrinkage and Selection Operator. BRR is Bayesian Ridge Regression. RFR is Random Forest Regression. LM is Levenberg-Marquardt. VOM is volume occupation spatial matrix. HCM is heat contribution spatial matrix.

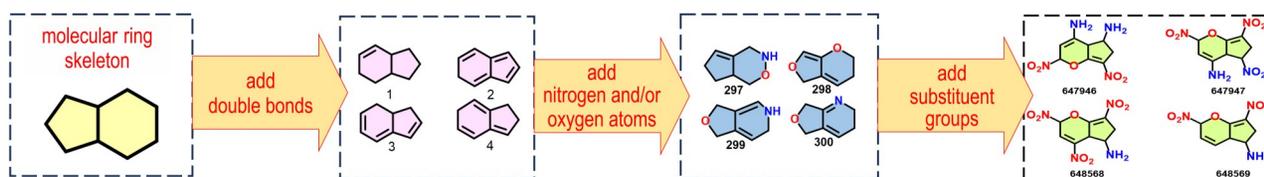


图3 含能分子设计流程图

Fig.3 Flow chart of energetic molecular design

48个[5,6]分子骨架。向分子骨架中引入氮氧元素能够提高含能化合物的密度及氧平衡,利于保证能量水平,但会降低骨架稳定性,减少可取代位点。因此,向全碳骨架中只引入1个氧原子或者1~8个氮原子,生成[5,6]氮氧分子骨架共28252个。然后根据分子取代位点随机加入氨基、硝基、叠氮基团,得到超过 10^7 个分子,并利用自制脚本对分子进行芳香性筛选,最终得到142689个具有共轭结构的化合物。

2 结果与讨论

2.1 生成分子性质预测及分析

2.1.1 稠环分子预测性质间的关系

为了探究所生成化合物各预测性质之间的关系,

研究采用三维散点图进行分析,结果如图4所示。由图4a可知,随着 ρ 的增加, D_v 、 ρ 也随之增加,当 D_v 大于 $9000 \text{ m} \cdot \text{s}^{-1}$ 时, ρ 大于 $1.7 \text{ g} \cdot \text{cm}^{-3}$, ρ 大于 31 GPa ,说明 ρ 、 D_v 、 ρ 三者之间存在很强的正相关性。图4b显示, IS 和 $\Delta H_{f, \text{solid}}$ 的关系类似于负指数曲线,即随着 $\Delta H_{f, \text{solid}}$ 的增加, IS 数值呈指数降低,与此同时, ρ 与 IS 呈负相关关系,且绝大多数高热稳定性的分子($T_d > 300 \text{ }^\circ\text{C}$)的 ρ 在 25 GPa 以下。然而,由图4c可知,分子的 IS 与 T_d 呈正指数曲线关系,而高密度($\rho > 1.8 \text{ g} \cdot \text{cm}^{-3}$)的分子主要集中在 ρ 大于 25 GPa 的范围内。由上述分析可以得知,含能化合物在能量与稳定性上存在天然的矛盾,而机械感度与热稳定存在一定的正相关性,这与目前对含能化合物构效关系的一般认识相符,一定程度上反应了机器学习模型的合理性。

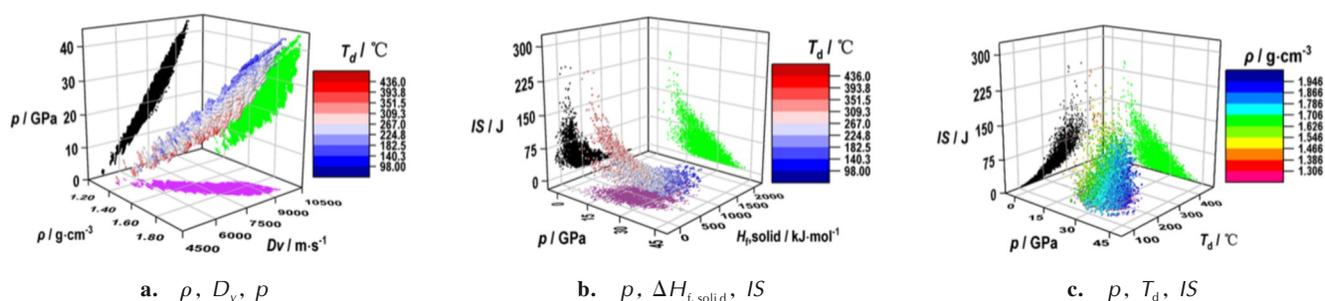


图4 生成的含能化合物各预测性质之间的关系

Fig.4 Relationship among the predicted properties of energetic compounds

2.1.2 稠环上氮原子数量对分子性能的影响

为探究稠环上氮原子数量对分子性能的影响,对生成含能化合物稠环上氮原子数量进行统计分析,结果如图5所示。如图5a所示,分子稠环上氮原子数量越多,高爆速的分子占比越高,例如,当分子稠环上没有氮原子时, D_v 大于8000 $\text{m}\cdot\text{s}^{-1}$ 的分子占没有氮原子的稠环分子总数的7.51%;当分子稠环上氮原子为4个时,占比增长到71.69%;当稠环上氮原子数量大于6个时,占比达到100%,所有分子的 D_v 均大于8000 $\text{m}\cdot\text{s}^{-1}$ 。同时还发现,当稠环上氮原子数量每增加一个,最低 D_v 增加470~700 $\text{m}\cdot\text{s}^{-1}$ (图5b),可能的原因是氮原子数量增加意味着更多高能C—N或N—N键的形成,对提升分子的生成焓是有利的,同时高氮含量可以保证分子分解时产生更多氮气产物,因此,增加氮原子数量有利于提升分子的 D_v 。然而由图5c可知,当稠环上氮原子数量过多时,其 T_d 也会随之降低,例如当稠环上氮原子数量在0~2个时, T_d 大于180 $^{\circ}\text{C}$ 的分子占比由14.90%增高到28.44%;但当氮原子数量继续增加时, T_d 大于180 $^{\circ}\text{C}$ 的分子占比逐渐降低。可能的原因是,稠环上氮原子增多时,由于氮原

子的电负性较高,使得周围电子云更稳定,从而增加分子活化能;但随着稠环上氮原子数量继续增加,分子极性增大,分子内部环境发生显著变化,降低了分子的活化能,从而影响分子的分解温度。因此,设计稠环类含能分子时,母环上氮原子数量对于实现化合物能量和稳定性的良好平衡十分关键。

2.1.3 稠环上氧原子数量对分子性能的影响

噁唑等含氧骨架能增加氧平衡,有利于提升能量水平。因此,为分析环上氧原子数量对分子性质的影响,研究通过小提琴图^[45]对分子的性质分布进行了统计分析,结果如图6所示。由图6a~6c可以发现,在分子 ρ 、 D_v 、 ρ 上,环上含有1个氧原子和不含有氧原子的分子差距不是很大,中位数的差值分别为0.01 $\text{g}\cdot\text{cm}^{-3}$, 5.1 $\text{m}\cdot\text{s}^{-1}$, 0.7 GPa。说明环上氧原子数量为0个或者1个对分子的能量性质影响不大,可能的原因是氧原子数量较少,对氧平衡及生成焓贡献不显著。然而,从图6d~6f可以发现,含有0或者1个氧原子稠环分子的IS中位数差值高达4.6 J,而 $\Delta H_{f,solid}$ 和 T_d 的中位数的差值也高达145.7 $\text{kJ}\cdot\text{mol}^{-1}$ 和29.4 $^{\circ}\text{C}$,可知氧原子数量对于IS、 $\Delta H_{f,solid}$ 、 T_d 的影响显著。可能的原因是氧原子的

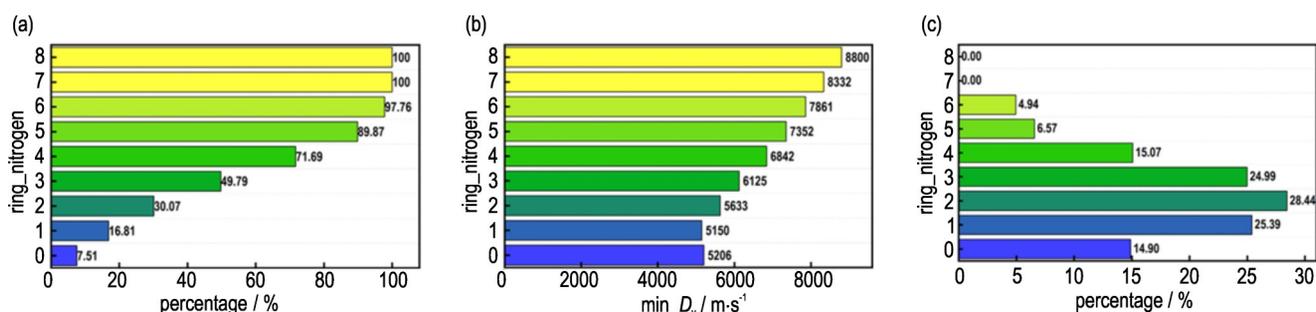


图5 (a) D_v 大于8000 $\text{m}\cdot\text{s}^{-1}$ 的稠环分子在此氮原子数量分子总数中的占比;(b)具有不同氮原子数量稠环分子的最低 D_v ;(c) T_d 大于180 $^{\circ}\text{C}$ 的稠环分子在此氮原子数量分子总数中的占比

Fig.5 (a) Proportion of fused ring molecules with D_v greater than 8000 $\text{m}\cdot\text{s}^{-1}$ in the total number of molecules with different number of nitrogen atoms; (b) minimum D_v of fused ring molecules with different number of nitrogen atoms; (c) proportion of fused ring molecules with T_d greater than 180 $^{\circ}\text{C}$ in the total number of molecules with different number of nitrogen atoms

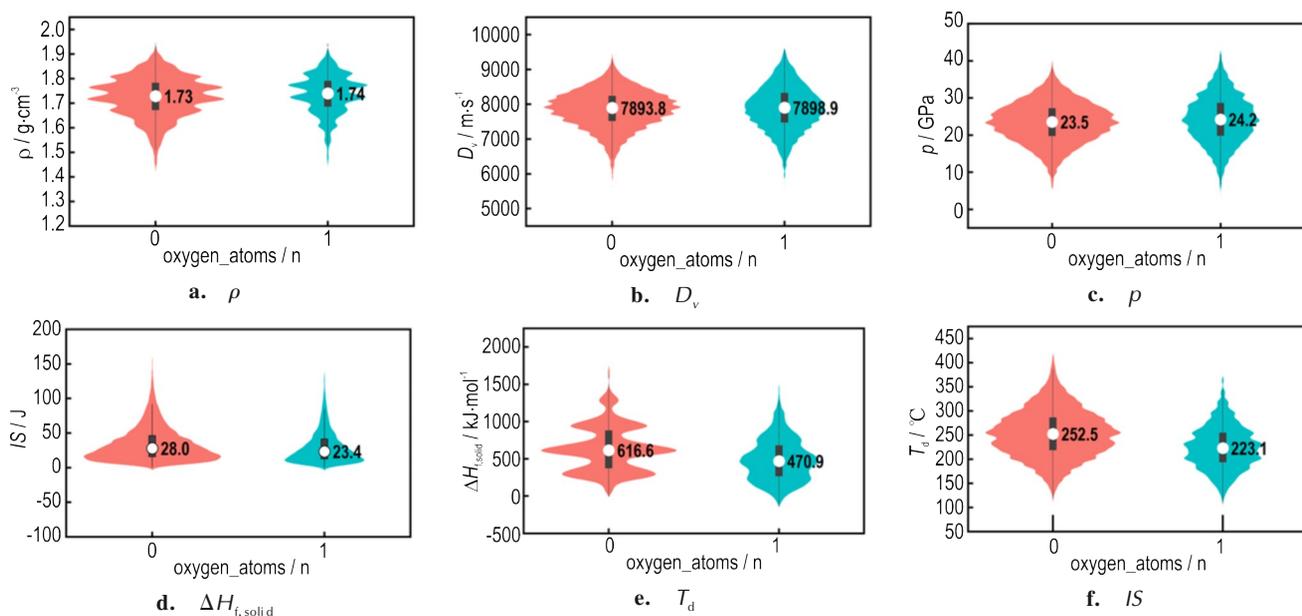


图6 [5,6]稠环骨架上含有0或者1个氧原子分子的 ρ 、 D_v 、 p 、 $\Delta H_{f,\text{solid}}$ 、 T_d 、 IS 的小提琴图(黑色矩形、白点、彩色范围分别代表四分位范围、中位数、95%置信区间和概率)

Fig.6 Violin plots of ρ , D_v , p , $\Delta H_{f,\text{solid}}$, T_d and IS for [5,6] heterocyclic molecules with 0 or 1 oxygen atom in the framework (where the black rectangle, white dot and color range represent the interquartile range, median, 95% confidence interval, and probability density, respectively)

出现造成分子的不对称以及极性的改变,同时降低分子的稳定性,使分子更容易在外界刺激下发生分解。

2.1.4 取代基种类和数量对分子性能的影响

研究从 D_v 、 p 和 T_d 三个方面,进一步分析了取代基(叠氮、硝基、氨基)数量对分子性质的影响,结果如图7所示。通过图7a和7b可以发现,随着叠氮和硝基的数量增加,高爆轰性能分子($D_v > 8000 \text{ m} \cdot \text{s}^{-1}$, $p > 19 \text{ GPa}$)占比更多;与之相反的是,随着氨基数量增加,高爆轰性能分子占比反而更少。分析原因,叠氮基团和硝基基团都能增加化合物的爆速,因为硝基基团有较多的氧原子,可以发生强烈的氧化反应,释放大能量,而叠氮基团的三键解离能极高,它们在分解时会迅速释放大气;而N—H键的能量不高,氨基分

解主要为氨气和水蒸气,对爆轰性能贡献很小。在图7c中,高热稳定性分子($T_d > 180 \text{ }^\circ\text{C}$)占比随氨基和硝基数量的增加而逐渐增加,而与叠氮数量呈负相关。这是因为氨基可以通过形成氢键增加分子内部的结构稳定性,而硝基提高了分子的氧化能力,增加分子的能量密度,并且其分解能较高,导致化合物整体分解所需能量增加;叠氮基团则是一种十分敏感且高能的官能团,其分解会产生大量的氮气,降低分子的稳定性。

2.2 分子筛选及性质对比

为保证分子兼具高能量(高爆速)与较高的热稳定性(高热分解温度),以 D_v 大于 $9000 \text{ m} \cdot \text{s}^{-1}$ 和 T_d 大于 $200 \text{ }^\circ\text{C}$ 作为筛选标准,得到681个分子,其中含氧骨架

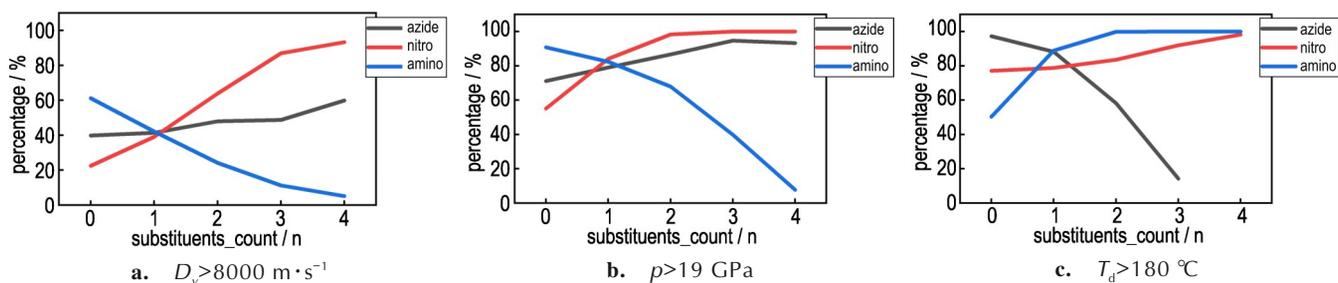


图7 随着取代基数量增加, D_v 大于 $8000 \text{ m} \cdot \text{s}^{-1}$ 、 p 大于 19 GPa 或 T_d 大于 $180 \text{ }^\circ\text{C}$ 的分子占比情况

Fig.7 Proportion of molecules with D_v greater than $8000 \text{ m} \cdot \text{s}^{-1}$ or p greater than 19 GPa or T_d greater than $180 \text{ }^\circ\text{C}$ with the increase of the number of substituents

分子(即环上有氧原子的分子)与碳氮杂环骨架分子比例约为1:2。然而研究发现含氧骨架类分子中的氧主要出现在六元环上,合成难度较大,难以成为潜在的候选分子。此外,筛选得到的分子中绝大多数都含有硝基基团,而叠氮基团和氨基基团比较少,主要是因为叠氮基团稳定性较差且密度较低造成的,而氨基基团对能量提升没有贡献,但会提高分子稳定性,增加分子的热分解温度。最终,在这681个分子中选取了5个能量性能较为突出的分子,并对这5个分子进行进一步量化和性质计算,以验证模型的精准度。

基于等键反应(Scheme 1),使用Gaussian09软件包在B3LYP/6-311++G(d,p)和MP2/6-311++G(d,p)理论水平下对化合物的气态标准生成焓进行计算^[46]。对于等键反应方程式中无法获得实验生成焓的反应物或产物,则基于原子化反应,采用G4(MP2)_6x方法计算得到目标分子的气态标准生成焓^[47]。根据式(5~6)计算得到分子的固体生成焓^[48-49],同时采用式(6)计算得到分子的晶体密度,并根据式(7~8)所示的Kamlet-Jacobos方程(K-J方程)计算分子的爆轰性质^[50]。

$$\Delta H_{f, \text{solid}} = \Delta H_{f, \text{gas}} - \Delta H_{\text{sub}} \quad (4)$$

$$\Delta H_{\text{sub}} = aA^2 + b\sqrt{17.505856\nu\sigma_{\text{tot}}^2} + c \quad (5)$$

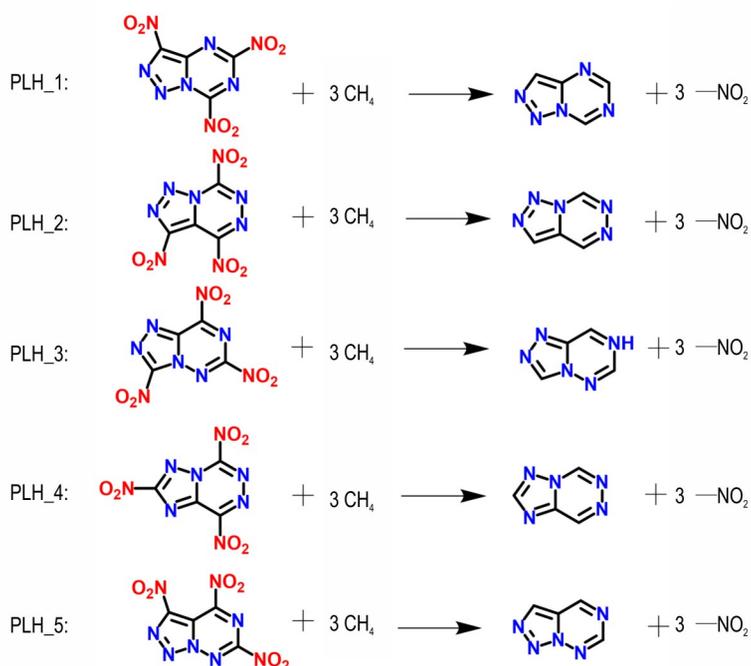
$$\rho = \alpha\left(\frac{M}{V_{0.001}}\right) + \beta \times 17.505856\nu\sigma_{\text{tot}}^2 + \gamma \quad (6)$$

$$D_v = (1.011 + 1.312\rho)(N\bar{M}^{0.5}Q^{0.5})^{0.5} \quad (7)$$

$$\rho = 1.558\rho_0^2 N\bar{M}^{0.5}Q^{0.5} \quad (8)$$

式中, $\Delta H_{f, \text{gas}}$ 是气体生成焓, $\text{kJ}\cdot\text{mol}^{-1}$; ΔH_{sub} 是升华焓, $\text{kJ}\cdot\text{mol}^{-1}$; $\Delta H_{f, \text{solid}}$ 是固体生成焓, $\text{kJ}\cdot\text{mol}^{-1}$; A 是分子表面积, \AA^2 ; 拟合的系数 $a=2.130$, $b=0.930$, $c=-17.844$ ^[49]; $\nu\sigma_{\text{tot}}^2$ 是分子表面总静电势的方差和电荷平衡度的乘积, $\text{kJ}^2\cdot\text{mol}^{-2}$; M 是分子质量, $\text{g}\cdot\text{mol}^{-1}$; $V_{0.001}$ 是分子的范德华体积, \AA^3 ; $\alpha=0.9183$, $\beta=0.0028$, $\gamma=0.0443$ 采用文献[48]拟合得到的参数; ρ 是爆压, GPa ; D_v 是爆速, $\text{m}\cdot\text{s}^{-1}$; N 是每克炸药产生的气体产物摩尔数; \bar{M} 是爆轰产物的平均摩尔质量, $\text{g}\cdot\text{mol}^{-1}$; Q 是爆热, $\text{kJ}\cdot\text{g}^{-1}$ 。

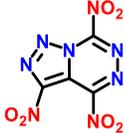
通过上述方法及公式,计算得到筛选出的5个分子的 ρ 、 $\Delta H_{f, \text{solid}}$ 、 D_v 和 ρ 性能结果,将其与本研究所得6个含能材料性能模型预测结果进行对比,列于表5。由表5可以发现,本研究模型所得预测结果与文献方法的计算结果几乎一致, ρ 的差值为 $0.02\sim 0.07\text{ g}\cdot\text{cm}^{-3}$, $\Delta H_{f, \text{solid}}$ 的差值为 $-35\sim 32\text{ kJ}\cdot\text{mol}^{-1}$, D_v 的差值在 $240\text{ m}\cdot\text{s}^{-1}$ 以内, ρ 的差值在 2.5 GPa 以内。可见,本研究建立的含能分子性能预测模型的有效性,表明本模型在预测含能分子的相关性质时具有较高的精度。值得一提的是,当计算分子 ρ 时,考虑静电势的影响会使计算值与预测值的差值增大。但总体而言,本研究采用的机器学习辅助虚拟筛选系统^[13]在新型含能分子设计筛选方面较传统的计算方法有显著优势,能够在保持较高精度的同时,显著降低计算所需的资源和时间消耗。



Scheme 1 Isodesmic reactions for computing the heat of formation

表5 含能化合物性能预测结果与和文献计算方法的所得结果对比

Table 5 Comparison between predicted properties obtained by this study machine learning and other reference method

ID	molecular	$\rho / \text{g} \cdot \text{cm}^{-3}$		$D_v / \text{m} \cdot \text{s}^{-1}$		p / GPa		$\Delta H_{f, \text{solid}} / \text{kJ} \cdot \text{mol}^{-1}$		$T_d / ^\circ\text{C}$	IS / J
		this study	other method	this study	other method	this study	other method	this study	other method		
PLH_1		1.90	1.88	9102	8912	38.6	36.1	445.3	431.0	227.1	15.5
PLH_2		1.90	1.86	9140	9044	38.5	37.2	483.7	519.1	208.6	12.4
PLH_3		1.90	1.83	9164	8941	38.8	36.4	463.1	438.9	210.3	14.8
PLH_4		1.90	1.84	9164	8954	38.8	36.5	469.5	487.9	213.8	14.6
PLH_5		1.90	1.85	9141	8992	38.5	36.7	463.7	498.7	207.3	13.2

3 结论

本研究展示了机器学习辅助的高通量筛选在[5,6]稠环含能分子设计及构效关系定量分析中的应用。通过对回归算法比较,选择KRR算法构建含能分子性能预测模型,结果表明所得预测模型对密度 ρ ,爆速 D_v ,爆压 p 具有高 R^2 (0.79~0.91),较低的MAE(0.041 $\text{g} \cdot \text{cm}^{-3}$, 240.3 $\text{m} \cdot \text{s}^{-1}$, 2.194 GPa)和MAPE值(2.4%~8.0%),对于撞感IS和分解温度 T_d 也有很好的预测能力($R^2 > 0.6$),未来的工作可能包括筛选合适的分子描述符、扩大数据集,对算法进行组合优化等方法,以提高对富氮稠环类化合物多重性质的预测能力。

其次,对稠环分子预测性质间的关系进行定量分析后,发现稠环上的O、N原子数量对分子的分解温度和稳定性具有重要的影响,叠氮和硝基基团对于稠环分子的爆轰性能起到积极的效果,而氨基和硝基基团会提高稠环分子的热稳定性。最后,通过性能综合筛选得到5个稠环分子并进行量子化学计算,结果表明

研究建立的模型对富氮稠环类含能化合物的多重性质预测方面具有较好的精度。

总之,本研究采用的机器学习辅助高通量虚拟筛选方法实现了特定体系含能分子的高通量设计,并可以在个人计算机上进行分子性质高效预测,对实验人员友好,有望大大加速分子的研发过程,推动新型含能化合物的发现。

参考文献:

- [1] PANG Wei-qiang, YETTER R A, DELUCA L T, et al. Boron-based composite energetic materials (B-CEMs): preparation, combustion and applications [J]. *Progress in Energy and Combustion Science*, 2022, 93: 101038.
- [2] XU Jing, SUN Jian, MA Xiao-xia, et al. Reactive antistatic additive modified copper(II) azide as a primary explosive with simultaneously enhanced stability and energy[J]. *Chemical Engineering Journal*, 2023, 471: 144440.
- [3] BAYAT Y, TAHERIPOUYA G, ZEYNALI V, et al. Methods and strategies to achieve hexanitrohexaazaisowurtzitane (HNIW or CL-20): A comprehensive overview [J]. *Journal of Energetic Materials*, 2023: 1-35.
- [4] YANG Xiu-rong, DANG Jia, ZHANG Chi, et al. Comparing the catalytic effect of metals for energetic materials: Machine

- learning prediction of adsorption energies on metals[J]. *Langmuir*, 2024, 40(1): 1087–1095.
- [5] WANG He, XU Ya-bei, WEN Ming-jie, et al. Kinetic modeling of CL-20 decomposition by a chemical reaction neural network[J]. *Journal of Analytical and Applied Pyrolysis*, 2023, 169: 105860.
- [6] CAO Yu-dong, ROMERO J, OLSON J P, et al. Quantum chemistry in the age of quantum computing[J]. *Chemical Reviews*, 2019, 119(19): 10856–10915.
- [7] SIFAIN A E, TADESSE L F, BJORGAARD J A, et al. Cooperative enhancement of the nonlinear optical response in conjugated energetic materials: A TD-DFT study [J]. *Journal of Chemical Physics*, 2017, 146(11): 114308.
- [8] VARANDAS A J C. Extrapolation in quantum chemistry: insights on energetics and reaction dynamics[J]. *Journal of Theoretical and Computational Chemistry*, 2020, 19(7): 2030001.
- [9] ELTON D C, BOUKOUVALAS Z, BUTRICO M S, et al. Applying machine learning techniques to predict the properties of energetic materials[J]. *Scientific Reports*, 2018, 8(1): 9012–9059.
- [10] CASEY A D, SON S F, BILIONIS I, et al. Prediction of energetic material properties from electronic structure using 3D convolutional neural networks [J]. *Journal of Chemical Information and Modeling*, 2020, 60(10): 4457–4473.
- [11] SONG Si-wei, CHEN Fang, WANG Yi, et al. Accelerating the discovery of energetic melt-castable materials by a high-throughput virtual screening and experimental approach [J]. *Journal of Materials Chemistry A, Materials for Energy and Sustainability*, 2021, 9(38): 21723–21731.
- [12] HOU Fang, MA Yi, HU Zheng, et al. Machine learning enabled quickly predicting of detonation properties of N-containing molecules for discovering new energetic materials [J]. *Advanced Theory and Simulations*, 2021, 4(6): 2100057.
- [13] SONG Si-wei, WANG Yi, CHEN Fang, et al. Machine learning-assisted high-throughput virtual screening for on-demand customization of advanced energetic materials[J]. *Engineering*, 2022, 10: 99–109.
- [14] LIU Jian, ZHAO Shi-cao, DUAN Bo-wen, et al. High-throughput design of energetic molecules [J]. *Journal of Materials Chemistry A, Materials for Energy and Sustainability*, 2023, 11(45): 25031–25044.
- [15] QIAN Wen, HUANG Jing, GUO Shi-tai, et al. Searching for the analogues of 1, 1-dinitro-2, 2-diamino ethylene (FOX-7) by high-throughput computation and machine learning [J]. *Firephyschem*, 2023, 3(4): 339–349.
- [16] WEN Lin-yuan, YU Tao, LAI Wei-peng, et al. Transferring the available fused cyclic scaffolds for high-throughput combinatorial design of highly energetic materials via database mining [J]. *Fuel*, 2022, 324: 124591.
- [17] WEN Lin-yuan, WANG Bo-zhou, YU Tao, et al. Accelerating the search of CHONF-containing highly energetic materials by combinatorial library design and high-throughput screening [J]. *Fuel*, 2022, 310: 122241.
- [18] LI Chuan, WANG Cheng-hui, SUN Ming, et al. Correlated RNN framework to quickly generate molecules with desired properties for energetic materials in the low data regime [J]. *Journal of Chemical Information and Modeling*, 2022, 62(20): 4873–4887.
- [19] LI Mao-gang, LAI Wei-peng, LI Rui-rui, et al. Novel random forest ensemble modeling strategy combined with quantitative structure-property relationship for density prediction of energetic materials [J]. *Acs Omega*, 2023, 8(2): 2752–2759.
- [20] NGUYEN P, LOVELAND D, KIM J T, et al. Predicting energetics materials crystalline density from chemical structure by machine learning [J]. *Journal of Chemical Information and Modeling*, 2021, 61(5): 2147–2158.
- [21] QIAN Wen, HUANG Jing, GUO Shi-tai, et al. Identifying the determining factors of detonation properties for linear nitroaliphatics with high-throughput computation and machine learning [J]. *Energetic Materials Frontiers*, 2023, doi: 10.1016/j.enmf.2023.05.002.
- [22] GUO Shi-tai, HUANG Jing, QIAN Wen, et al. Discovery of high energy and stable prismane derivatives by the high-throughput computation and machine learning combined strategy [J]. *Firephyschem*, 2024, 4(1): 55–62.
- [23] WANG Xian-shuang, HE Ya-ge, CAO Wen-li, et al. Fast explosive performance prediction via small-dose energetic materials based on time-resolved imaging combined with machine learning [J]. *Journal of Materials Chemistry A, Materials for Energy and Sustainability*, 2022, 10(24): 13114–13123.
- [24] WU Jun-nan, SONG Si-wei, TIAN Xiao-lan, et al. Machine learning-based prediction and interpretation of decomposition temperatures of energetic materials [J]. *Energetic Materials Frontiers*, 2023, 4(4): 254–261.
- [25] MAKAROV D M, FADEEVA Y A, SHMUKLER L E, et al. Machine learning models for phase transition and decomposition temperature of ionic liquids [J]. *Journal of Molecular Liquids*, 2022, 366: 120247.
- [26] YANG Chun-ming, CHEN Jie, WANG Run-wen, et al. Density prediction models for energetic compounds merely using molecular topology [J]. *Journal of Chemical Information and Modeling*, 2021, 61(6): 2582–2593.
- [27] ZHANG Jia-heng, MITCHELL L A, PARRISH D A, et al. Enforced layer-by-layer stacking of energetic salts towards high-performance insensitive energetic materials [J]. *Journal of the American Chemical Society*, 2015, 137(33): 10532–10535.
- [28] 张朝阳. 含能材料能量-安全性间矛盾及低感高能材料发展策略 [J]. *含能材料*, 2018, 26(1): 2–10.
ZHANG Chao-yang. On the energy & safety contradiction of energetic materials and the strategy for developing low-sensitive high-energetic materials [J]. *Chinese Journal of Energetic Materials (Hanneng Cailiao)*, 2018, 26(1): 2–10.
- [29] 张计传, 王振元, 王滨, 等. 富氮稠环含能化合物: 平衡能量与稳定性的新一代含能材料 [J]. *含能材料*, 2018, 26(11): 983–990.
ZHANG Ji-chuan, WANG Zhen-yuan, WANG Bin-shen, et al. Fused-ring nitrogen-rich heterocycles as energetic materials: Maintaining a fine balance between performance and stability [J]. *Chinese Journal of Energetic Materials (Hanneng Cailiao)*, 2018, 26(11): 983–990.
- [30] CUI Wen-hao, LIU Qi, YE Zhi-wen, et al. Design and synthesis of bistetrazole-based energetic salts bearing the nitrogen-rich fused ring [J]. *Organic Letters*, 2023, 25(30): 5661–5665.
- [31] 刘赛, 石伟, 王毅, 等. 富氮稠环类氮氧化物的研究进展 [J]. *含能材料*, 2021, 29(6): 567–578.
LIU Sai, SHI Wei, WANG Yi, et al. Research progress of nitrogen-rich fused-ring N-oxides [J]. *Chinese Journal of Energetic Materials (Hanneng Cailiao)*, 2021, 29(6): 567–578.
- [32] RAPPE A K, CASEWIT C J, COLWELL K S, et al. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations [J]. *Journal of the American Chemical*

- Society*, 1992, 114(25): 10024–10035.
- [33] HALGREN T A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 [J]. *Journal of Computational Chemistry*, 1996, 17(5–6): 490–519.
- [34] WEININGER, DAVID. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. *Journal of Chemical Information and Computer Sciences*, 1988, 28(1): 31–36.
- [35] DU Ming-jing, DING Shi-fei, JIA Hong-jie. Study on density peaks clustering based on k-nearest neighbors and principal component analysis [J]. *Knowledge-Based Systems*, 2016, 99: 135–145.
- [36] CARNEIRO T C, ROCHA P A C, CARVALHO P C M, et al. Ridge regression ensemble of machine learning models applied to solar and wind forecasting in brazil and spain [J]. *Applied Energy*, 2022, 314: 118936.
- [37] ZHANG Yu-chen, DUCHI J C, WAINWRIGHT M J. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates [J]. *The Journal of Machine Learning Research*, 2015, 16(1): 3299–3340.
- [38] VELO R, LÓPEZ P, MASEDA F. Wind speed estimation using multilayer perceptron [J]. *Energy Conversion and Management*, 2014, 81: 1–9.
- [39] GUO Zhen-dong, LIU Hai-tao, ONG Y S, et al. Generative multiform bayesian optimization [J]. *Ieee Transactions On Cybernetics*, 2023, 53(7): 4347–4360.
- [40] HUŠEK P. On monotonic radial basis function networks [J]. *Ieee Transactions On Cybernetics*, 2024, 54(2): 717–727.
- [41] ZHAO Yang, WEI Yong-yue, CHEN Feng. Forecasting of COVID-19: Transmission models and beyond [J]. *Journal of Thoracic Disease*, 2020, 12(5): 1762–1765.
- [42] BROWNE M W. Cross-validation methods [J]. *Journal of Mathematical Psychology*, 2000, 44(1): 108–132.
- [43] WANG Qing-sheng, WANG Jie-jia, LARRANAGA M D. Simple relationship for predicting onset temperatures of nitro compounds in thermal explosions [J]. *Journal of Thermal Analysis and Calorimetry*, 2013, 111(2): 1033–1037.
- [44] CHEN Chao, LIU Dang-yang, DENG Si-yan, et al. Accurate machine learning models based on small dataset of energetic materials through spatial matrix featurization methods [J]. *Journal of Energy Chemistry*, 2021, 63: 364–375.
- [45] WANG Yi, LIU Yu-ji, SONG Si-wei, et al. Accelerating the discovery of insensitive high-energy-density materials by a materials genome approach [J]. *Nature Communications*, 2018, 9(1): 2444.
- [46] FRISCH M J, TRUCKS G, SCHLEGEL H B, et al. Gaussian 09w, revision A. 02[CP]. 2009.
- [47] CHAN B, DENG JIA, RADOM L. G4 (MP2)-6X: a cost-effective improvement to G4(MP2) [J]. *Journal of Chemical Theory and Computation*, 2011, 7(1): 112–120.
- [48] RICE B M, BYRD E F. Evaluation of electrostatic descriptors for predicting crystalline density [J]. *Journal of Computational Chemistry*, 2013, 34(25): 2146–2151.
- [49] BYRD E F C, RICE B M. Improved prediction of heats of formation of energetic materials using quantum mechanical calculations [J]. *Journal of Physical Chemistry A*, 2006, 110(3): 1005–1013.
- [50] POLITZER P, MARTINEZ J, MURRAY J S, et al. An electrostatic interaction correction for improved crystal density prediction [J]. *Molecular Physics*, 2009, 107(19): 2095–2101.

Machine Learning Assisted High-throughput Design of [5,6] Fused Ring Energetic Compounds

PAN Lin-hu, WANG Rui-hui, FAN Ming-ren, SONG Si-wei, WANG Yi, ZHANG Qing-hua

(School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Compared with the research and development model guided by experience and calculations, machine learning-assisted high-throughput virtual screening technology for energetic molecules has shown obvious advantages in terms of molecular design efficiency and quantitative analysis of structure-activity relationships. In view of the fact that nitrogen-rich fused ring energetic compounds usually show better energy-stable balance properties, this study uses machine learning-assisted high-throughput virtual technology to conduct chemical space exploration of [5,6] nitrogen-rich fused ring energetic molecules. Based on the [5,6] all-carbon skeleton, this study obtained 142,689 [5,6] fused ring compounds through combined enumeration and aromatic screening. At the same time, a machine learning algorithm was used to establish and optimize an energetic molecular property prediction model (including density, decomposition temperature, detonation velocity, detonation pressure, impact sensitivity and enthalpy of formation). The effects of nitrogen and oxygen atoms on the fused ring and functional groups on the molecule on the performance of energetic compounds were analyzed. The research results show that the structure-activity relationship of the generated fused ring compounds is consistent with the general correlation between energy and stability of energetic compounds, verifying the rationality of the prediction model. Taking detonation velocity and decomposition temperature as the criteria for energy and thermal stability, five molecules with outstanding comprehensive properties were screened, and the quantum chemical calculation results were in good agreement with the machine learning prediction results, which further verified the accuracy of the prediction model.

Key words: machine learning; high-throughput screening; kernel ridge regression; molecular design; [5,6] fused ring energetic compounds

CLC number: TJ55

Document code: A

DOI: 10.11943/CJEM2024055

Grant support: National Natural Science Foundation of China (Nos. 22075259, 22175157, 22205218)

(责编: 卢学敏)