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

基于分子指纹与量子化学描述符预测聚酰亚胺玻璃化转变温度的机器学习模型

詹森华，石彤非
（广东工业大学轻工化工学院，广州 510006）

摘要 基于聚酰亚胺重复单元获得了分子访问系统(MACCS)指纹图谱和9种量子化学密度泛函理论(DFT)描述符，构建了MACCS，DFT和两者集成的3类预测模型. 通过比较分析随机森林(RF)、支持向量回归(SVR)、极致梯度提升(XGB)和梯度提升回归(GBR)等4种机器学习算法共12个机器学习模型来预测聚酰亚胺的玻璃化转变温度，并提取关键特征信息. 结果表明，最优的玻璃化转变温度预测模型是XGB集成模型，其训练集和测试集的决定系数(R^2)分别为0.956和0.811，测试集的均方根误差(RMSE)和平均绝对误差(MAE)分别为25.41和20.20. 此外，集成MACCS指纹和DFT的模型均比单一模型的效果好. 建立的集成模型框架可为聚酰亚胺材料及聚合物材料结构的设计提供参考.

关键词 机器学习；量子化学；分子指纹；聚酰亚胺
中图分类号 O63; O641 **文献标志码** A doi：10.7503/cjcu20240556

Machine Learning Model for Predicting the Glass Transition Temperature of Polyimides Based on Molecular Fingerprints and Quantum Chemical Descriptors

ZHAN Senhua, SHI Tongfei*
(*School of Chemical Engineering and Light Industry*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract Combining machine learning and quantum chemistry methods to construct predictive models can facilitate the design and screening of polyimide material structures. In this study, Molecular ACCess System (MACCS) fingerprints and nine density functional theory (DFT) quantum chemical descriptors were obtained from polyimide repeating units to construct three types of predictive models: MACCS, DFT and their integrated models. Twelve machine learning models were developed using four algorithms——random forest (RF), support vector regression (SVR), extreme gradient boosting (XGB) and gradient boosting regression (GBR)——to predict the glass transition temperature of polyimides and extract key feature information. The results showed that the optimal predictive model for the glass transition temperature is the integrated XGBoost model, with coefficient of determination (R^2) values of 0.956 and 0.811 for the training and test sets, respectively. The root mean square error (RMSE) and mean absolute error (MAE) for the test set are 25.41 and 20.20, respectively. Furthermore, the integrated MACCS fingerprint and DFT models performed better than the individual models. The established integrated model framework provides new

收稿日期：2024-12-23. 网络首发日期：2025-01-15.
联系人简介：石彤非，男，博士，教授，主要从事高分子物理方面的研究. E-mail：tfshi@gdut.edu.cn
基金项目：国家自然科学基金(批准号：2247030172)资助.
Supported by the National Natural Science Foundation of China(No.2247030172).

insights for the structural design of polyimide materials and other polymer materials.

Keywords Machine learning; Quantum chemistry; Molecular fingerprint; Polyimide

聚酰亚胺(PI)具有优异的热稳定性、卓越的机械性能、出色的化学稳定性和高耐辐射性[1~5]，被广泛应用于航天航空[6]、微电子[7,8]、光伏电池[9]及燃料电池[10]等领域. 1955年，美国杜邦公司合成出一种芳香型聚酰亚胺并申请了首个关于聚酰亚胺的专利[11]. 1969年，Frazer等[12]通过均苯四甲酸二酐(PMDA)与氧二苯胺(ODA)合成出一种具有优异热稳定性的均苯型聚酰亚胺，实现了聚酰亚胺的商业化. 此后，科研人员对聚酰亚胺进行了深入研究.

通过二酐和二胺或二异氰酸酯缩合可获得聚酰亚胺(图1). 聚酰亚胺独特的分子结构，如氮五元环和六元环的共轭作用、形成分子内和分子间电荷转移络合物(CTC)等，使其具有高玻璃化转变温度(T_g)等优异的性能[13]. 高T_g聚酰亚胺因具有优异的耐高温性能，被广泛应用于航天器和卫星的耐高温复合材料、光电器件以及高温高压环境下的密封材料中，已成为现代工业和高科技领域中不可或缺的关键材料. 常见的光学聚合物薄膜在高温加工时会丧失其原有的光学性能和机械性能，这限制了其在光电工程中的应用，而高T_g聚酰亚胺即使在高温环境下仍能够保持良好的透光率. 此外，多孔聚酰亚胺因具有优异的稳定性，常被用作航天航空领域的轴承材料，在太空的极端恶劣环境中展现出卓越的性能[14]. 因此，高T_g聚酰亚胺的设计已成为材料科学领域研究的热点. 不同的二酐和二胺或二异氰酸酯组合可以获得不同功能的聚酰亚胺. 目前，经典的T_g测量方法依赖于实验测试，存在实验成本高及周期性长等问题. 密度泛函理论(DFT)和分子动力学(MD)模拟等仿真模拟方法可以在一定程度上预测T_g，但仍存在一定的局限性，十分依赖科研人员的实验经验和高度的科学直觉，尚不能完全解决这一难题[15].

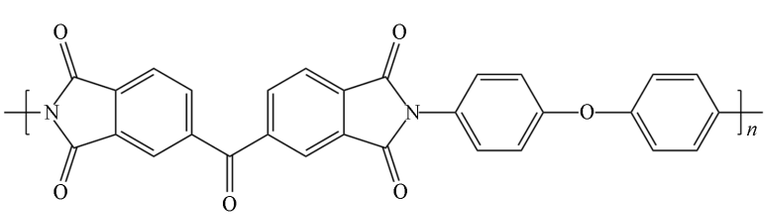


Fig. 1 Structure of polyimide(PI)

使用机器学习(ML)算法对分子结构进行设计和筛选可以加速开发新材料[16~19]. 与传统的实验方法和DFT与MD模拟相比，机器学习具有速度快和处理量多等优点[20,21]. 通过收集和构建数据库，基于机器学习算法可以对聚合物性质进行准确预测. 机器学习算法已经广泛应用于预测聚合物的介电[22]、热力学[23]、力学[24]、光学[25]及溶解性[26]等性质及临床诊断[27]等领域.

玻璃化转变作为一种二阶转变，是一种动力学现象，具有广泛的温度范围. Fox-Flory描述了T_g和分子量(M)之间的关系，其中T_g随着数均分子量的增加而升高[28,29]. 但这种关系不足以描述T_g和聚合物结构性能之间的关系，因此人们通过数学建模将分子结构和T_g关联起来. 例如，Alesadi等[30]将聚合物重复单元拆分为30多个结构和分子特征，包括侧链、芳香环及不同的原子等来预测T_g；Huang等[31]通过校正刚性和柔性基团对T_g的贡献值，修正了普适性基团贡献法，使其适用于预测聚酰亚胺体系，计算出的T_g与实验值有更好的相关性；Qiu等[32]收集了372个聚酰亚胺的T_g数据集，通过图神经网络(GNN)描述聚酰亚胺的结构和训练机器学习模型，表现出优秀的预测性能，测试集决定系数(R^2)高达0.83. 但他们所使用的模型和方法过于单一，只能得到T_g和分子结构之间的统计关系，缺少分子内部信息. 量子化学数据可提供分子的分子轨道、能级及电子密度分布等信息，有助于理解分子结构与T_g之间的关系. 通过分析这些电子结构特性，可以深入揭示聚合物的T_g如何受到分子内相互作用、链结构和分子刚性等因素的影响.

本文构建了1个包含287种组分的聚酰亚胺结构的数据库，提出了一种ML模型，该模型可以基于分子访问系统(MACCS)指纹和T_g相关的量子化学描述符(包括E_{HOMO}及E_{LUMO}等9种描述符)预测聚酰亚胺的T_g. 比较了4种机器学习模型预测结果的准确性，并通过网格优化预测模型的泛化能力，通过

SHAP方法对4种预测模型进行了解释.

1 材料与方法

1.1 数据集收集

本文使用的聚酰亚胺的分子结构和玻璃化转变温度来源于文献[33]. 共收集到372个T_g数据，通过量子化学计算筛选后得到287个有效数据点，分布如图2表示. 简化分子输入线性表示法(SMILES)[34]是一种简化的分子输入格式，通过使用文本字符来表示和描述化学结构，利用ASCII表示分子内的原子、连接性和手性信息，可以利用计算机进行各种机器学习任务[35]. 本文用重复单元的SMILES表示聚合物，用“*”来表示交界处. 在数据准备过程中，采用ChemDraw软件绘制聚酰亚胺的重复单元，然后转换为SMILES字符串，用来计算后续的分子指纹(MF)和量子化学描述符.

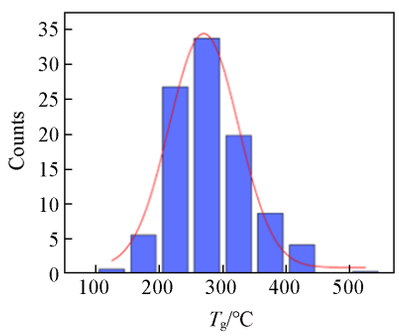


Fig. 2 T_g dataset for model building

1.2 分子指纹

在机器学习中，分子指纹是一种将分子的属性或特征拆解为数字表示的方法. 本文引入MF来表示聚酰亚胺重复单元的分子结构，这将为输入向量增加结构信息和物理化学信息. MACCS指纹图谱[36]将分子结构拆分为166个子结构，如芳香环及羰基等，由二进制表示是否存在子结构. 根据聚酰亚胺的SMILES，使用Python中的Rdkit库转化为166位MACCS指纹. 虽然MACCS分子指纹较短(只有166位)，存在无法描述更复杂的分子特征等局限性，但计算和存储效率较高，子结构特征容易解释. 因此，本文采用MACCS指纹图谱作为部分输入向量来增加分子结构信息.

1.3 量子化学描述符

采用量子化学(QC)方法可计算每个聚合物单体的化学性质，以增加预测模型中的化学性质描述符. 对于整个聚合物，难以用量子化学方法进行计算；对于二聚物和三聚物，尽管忽略了空间位阻效应，但计算结果与重复单元相比时间成本增加，因此，本文对聚酰亚胺重复单元进行计算来获得QC描述符. Zhang等[37]采用QC计算获得热能($E_{thermal}$)和总能量(E_{HF})作为特征值，通过高斯过程回归模型成功预测了聚丙烯酰胺的T_g，他们后续还增加了电荷和平均极化率等描述符，对聚甲基丙烯酸酯的T_g进行预测[38]. 但他们只描述了T_g与量子化学描述符之间的统计关系，而没有描述分子结构与T_g之间的关系.

玻璃化转变温度涉及分子结构的特性和分子间的相互作用，与多个量子化学描述符有关，反映分子结构的刚性、自由度、极性及热稳定性等特性. 选择合适的量子化学描述符才能准确描述聚合物T_g的影响机制. 最高占据分子轨道(HOMO)和最低占据分子轨道(LUMO)能级的分布决定了分子的电子转移能力和反应活性，可在一定程度上反映分子结构的热稳定性. 能隙(Gap)为HOMO和LUMO的能级差，对应分子从基态到激发态所需的能量. 能隙越大，意味着分子越难受热激发，从而表现为较高的热稳定性. 偶极矩(μ)对应分子间的相互作用，较高的电偶极矩通常意味着分子间存在较强的相互作用. 电子能量(E)可表示分子的稳定性. 零点能(Zero-point energy)、热能(Thermal Energy)、焓(Thermal Enthalpy)及自由能(Thermal free energy)均与分子内振动、旋转和平移能量有关，表示分子间的相互作用，与T_g密切相关. 这些特性均采用Gaussian程序包[39]，根据密度泛函理论进行QC计算. M06-2X泛函在大量数据库上进行了广泛测试，包括热化学、动力学和非共价相互作用数据集. 在需要中程相关能量和交换贡献的领域，其性能显著优于B3LYP等传统泛函. M06-2X泛函在计算有机化合物体系和能量计算精度方面有优势. 6-31G(d)基组是一种极化分裂价基组，在中等规模分子计算中可以很好地平衡计算效率和精度. 因此，本文采用M06-2X/6-31G(d)[40]水平进行理论计算，所有QC计算

的体系均为气相.

1.4 机器学习方法

采用随机森林(RF)[41]、支持向量回归(SVR)[42]、极致梯度提升(XGB)[43]和梯度提升回归(GBR)算法构建机器学习预测模型，对比了4种算法对模型预测精度的影响. 通过Python的Sklearn，Numpy和Pandas模块训练模型和进行数据预测.

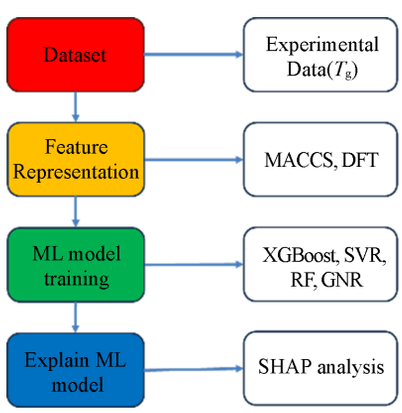
RF算法是一种集成学习方法，通过构建多个决策树分类器进行预测，将每棵树的结果结合起来(平均)，以生成最终预测结果，具有强鲁棒性和处理高维非线性的优点. SVR算法通过构建一个函数，使得多数数据点与实际值的误差在一个阈值内，通过核技巧处理问题. XGB算法是一种高效的梯度提升方法，核心是基于前一次误差不断迭代训练新的决策树，以获得最好的预测结果. XGB算法具有高性能和强鲁棒性的优点，内置正则化参数可防止过拟合. GBR算法也是一种集成方法，类似随机森林算法，通过多个决策树逐步提高模型预测精度.

为了比较不同训练模型的准确性，本文采用决定系数(R^2)、均方根偏差(RMSE)及平均绝对误差(MAE)来评估每个模型的性能和泛化能力. R^2可衡量模型对目标变量的解释能力，数值越接近1表示模型拟合效果越好；RMSE反映了预测值和真实值之间的偏差；MAE反映了预测值和真实值之间的平均绝对偏差.

2 结果与讨论

2.1 机器学习工作流程

本文的目的是建立一种集成分子结构和量子化学描述符的聚酰亚胺T_g预测模型，以指导高性能聚合物的分子设计. 本文的工作流程如Scheme 1所示：首先，收集了聚酰亚胺的分子结构和T_g实验值，并根据分子结构的SMILES提取了166位MACCS分子指纹以及高斯计算得到的9种量子化学描述符作为输入变量，同时将这两类特征进行集成，形成一种新的输入变量，将收集的T_g数据作为输出变量. 随后，为了探究分子指纹和量子化学描述符对模型预测性能的影响，分别以分子指纹、量子化学描述符及其集成特征进行预测，共构建了12个预测模型. 将数据集拆分为训练集和测试集，其中训练集由233个数据点(占比80%)组成，测试集由58个数据点(占比20%)组成. 通过R^2，RMSE和MAE评估每个模型的性能. 最后，利用SHAP值和相关性分析解释前10种重要子结构和量子化学描述符对T_g的影响，以便为聚酰亚胺的分子设计提供依据. 本文给4种模型的参数都设置了常见的参数范围，通过网格搜索等优化方法去找到每种最优模型的最优参数. 数据集、模型和参数配置的详细信息可在https://github.com/jidan-xm/MACCS-DFT-model.git网页上查看.



Scheme 1 Schematic illustration of the workflow

2.2 ML模型性能评估

通过评估RF，SVR，XGB和GBR等4种机器学习算法确定了预测聚酰亚胺T_g的最佳模型. 图3示出了4种集成机器学习模型训练集和测试集的准确性(蓝色表示训练集结果，红色表示测试集结果). 表1列出了4种机器学习共12个不同模型拟合度和性能，包括DFT模型，分子指纹模型和集成模型.

对于DFT模型，最优模型是XGB-DFT模型，其训练集和测试集的R^2分别为0.983和0.575，测试集的RMSE和MAE分别为38.07和27.90. 对于MACCS指纹图谱模型，最优的模型是XGB-MACCS模型，其训练集和测试集的R^2分别为0.814和0.785，测试集的RMSE和MAE分别为27.11和21.71. 整

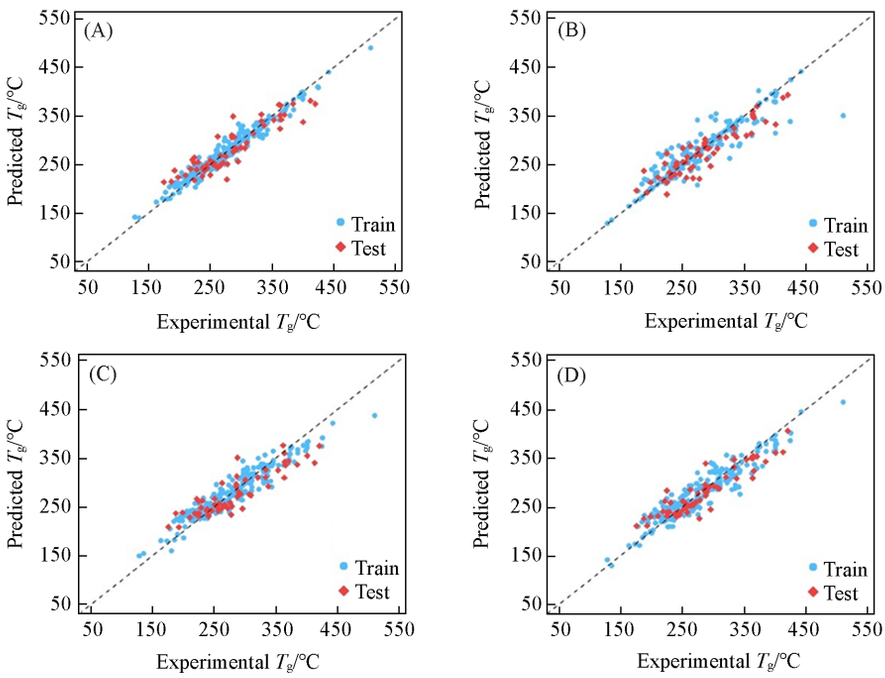


Fig. 3　Accuracy of training and test sets for four ensemble machine learning models

(A) XGB-Combined model; (B) SVR-Combined model; (C) RF-Combined model; (D) GBR-Combined model.

Table 1　Performance statistics of 12 predictive model

Model	Train R^2	Test R^2	RMSE	MAE
XGB-DFT	0.983	0.575	38.07	27.90
XGB-MACCS	0.814	0.785	27.11	21.71
XGB-Combined	0.956	0.811	25.41	20.20
SVR-DFT	0.497	0.512	40.79	30.73
SVR-MACCS	0.766	0.755	28.94	22.26
SVR-Combined	0.845	0.811	25.40	19.55
RF-DFT	0.813	0.529	40.11	28.65
RF-MACCS	0.793	0.751	29.13	23.03
RF-Combined	0.895	0.757	28.81	23.38
GBR-DFT	0.941	0.534	39.89	29.24
GBR-MACCS	0.833	0.731	30.28	24.63
GBR-Combined	0.882	0.822	24.64	18.86

体而言，与XGB-DFT模型相比，XGB-MACCS模型的性能有了明显提升，表明在机器学习模型中，分子结构的引入是必要的.

对于集成MACCS和DFT模型，最优的模型仍是XGB-Combined模型. 其训练集和测试集的R^2分别为0.956和0.811，测试集的RMSE和MAE分别为25.41和20.20. 其余集成模型也有较好的预测性能，其中SVR-Combined和GBR-Combined的训练集和测试集的R^2都达到了0.8以上. 与单独的DFT和MACCS模型相比，4个集成模型测试集的各个评估参数都有不同程度的提升. XGB模型采用逐步迭代的方法来提高模型的性能，因此在处理复杂特征关系中有更好的效果，在预测T_g方面具有更高的准确性和泛化能力. 总之，在比较分析4种预测聚酰亚胺T_g的机器学习模型中，分子指纹的效果大于DFT，而分子指纹和DFT的集成模型充分展示了集成模型具有更好的准确性和良好的泛化能力.

2.3　模型解释

在构建聚酰亚胺T_g预测模型后，需要找到一种方法来解释影响聚酰亚胺T_g的主要因素，从而指导聚酰亚胺材料的分子设计. Shapley Additive Explanations(SHAP)方法是现如今最好的机器学习模型解释工具之一，被广泛用于各种机器学习研究中. 基于博弈论中的Shapley值原理，可以计算每个输入变

量的贡献值. 本文通过SHAP方法对4种结合分子指纹的子结构和量子化学描述符的机器学习模型进行解释, 以获得每个分子指纹子结构和量子化学描述符对预测模型结果的贡献值. 图4示出了XGB-Combined model, SVR-Combined model, RF-Combined model和GBR-Combined model等4种集成模型中10个最重要的输入变量的SHAP值. y轴为特征变量, 越靠上说明其影响越大. x轴为SHAP值, 分布点越靠左说明特征对模型的负向影响越大; 分布点越靠右, 说明特征对模型的正向影响越大. 红色和蓝色表示特征取值的高低, 高特征值(红色)集中在右侧, 说明特征值越高, 越会正向影响预测结果; 反之, 蓝色集中在右侧则说明特征值越低越正向影响预测结果. 这一特征分布规律揭示了分子指纹和量子化学描述符对聚酰亚胺T_g的不同影响, 可为分子设计提供重要依据.

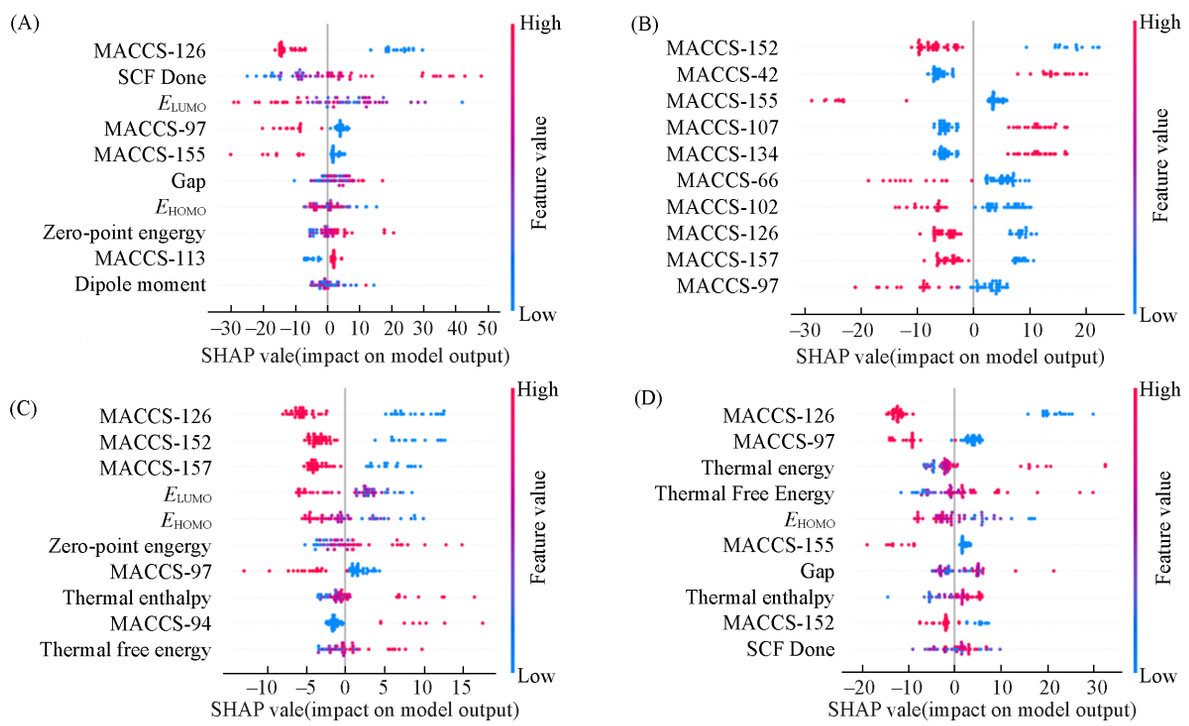


Fig. 4 SHAP values of the 10 most important input variable

(A) XGB-Combined model; (B) SVR-Combined model; (C) RF-Combined model; (D) GBR-Combined model.

对于性能最好的XGB-Combined模型, 分子指纹中影响最大的有子结构126, 97和155, 而量子化学描述符中影响最大的是电子能量、LUMO能级、能隙、HOMO能级、零点能和偶极矩. 由图4(A)可以看出, 子结构126, 97和155的特征值红色部分在左边, 说明这些子结构的特征值越大, 越会降低模型的预测结果. 由图5可见, 这些子结构分别对应芳香环、C—H键和单双键交替结构. C—H柔性结构增加会导致链段运动更容易, 从而降低聚酰亚胺的T_g; 芳香环和单双键交替结构等共轭结构增加也会导致链段间更容易滑动和堆叠, 从而降低聚酰亚胺的T_g. 可以看到, 能隙和零点能的增加则会提高T_g. 这是因为较大的能隙表示分子结构稳定, 难以发生链段运动; 而较高零点能意味着分子振动频率高, 刚性结构多, 从而导致T_g的升高. 此外, 子结构152出现在其余3种模型中, 为三取代碳原子. 子结构

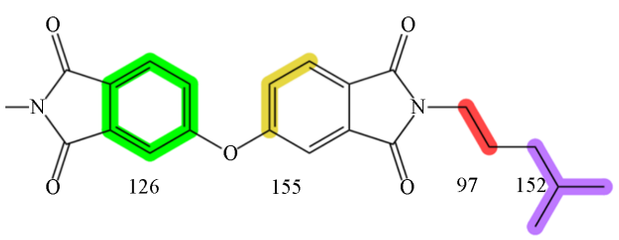


Fig. 5 Important substructures affecting the T_g model

152的增加意味着分子链的支化程度增大，这会降低分子堆叠效率，减少分子间作用力，从而导致T_g下降.

由图4(B)~(D)可以看出，对于RF-Combined模型和GBR-Combined模型，前10种最重要的输入变量既包括分子指纹，也包括量子化学描述符；而对于SVR-Combined模型，前10种最重要的输入变量均为分子指纹. 在4种模型的SHAP值图中，子结构126，97，155和能隙出现频率高，且对T_g影响较强，说明4种模型的结果具有较高的一致性. 在设计高T_g聚酰亚胺材料的时候，可以根据SHAP图的结果，适当增加刚性结构，同时减少C—H等柔性结构的成分，最后再通过量子化学计算能隙和零点能等简单的量子化学信息，进一步确认所设计的单体是否具有较高的T_g.

可以看出，在预测聚酰亚胺的T_g方面，基于分子指纹的模型优于基于DFT特征的模型. 其优势主要体现在分子指纹能够直接编码与T_g密切相关的化学功能基团和分子子结构信息. 聚酰亚胺分子中通常含有刚性的芳香环和酰亚胺基团，同时可能包含醚键和亚甲基等柔性链段. 分子结构中刚性链段的存在能够显著提高分子链的刚性，从而提升T_g；而柔性链段则通过影响分子间作用力和链段堆叠密度，对T_g也起到了重要作用. 分子指纹能够直接捕捉这些化学基团，从而有效表征分子结构对T_g的影响. 相比之下，DFT特征更注重单体内的电子结构(如LUMO和能隙等)，在对聚合物链段整体行为的描述上存在一定的局限性. 此外，MACCS等分子指纹提供的是高维特征，与机器学习模型的适配性更强，而低维的DFT特征更适合作为分子结构内在信息的补充，为模型预测提供额外的支持.

为探究10种影响因素的独立性，基于性能最优的XGBoost-Combined模型生成了前10种影响因素的相关性热图(图6)，红色表示相关性强，蓝色表示相关性弱. 从图6可以看出，大多数因素之间呈现弱相关性，说明它们是相对独立地影响着T_g，在分子设计过程中，可以单独考虑这些彼此之间是弱相关的结构. 此外，子结构113和126表现出较强的相关性，这是因为子结构113为碳碳双键结构，因此与芳香环126结构呈现强相关性；化学键强度、原子排列都会影响零点能和电子能量，因而呈现强相关性. 对于相关性弱的子结构，在分子设计中可以单独考虑他们的加入，从而获得更高的T_g.

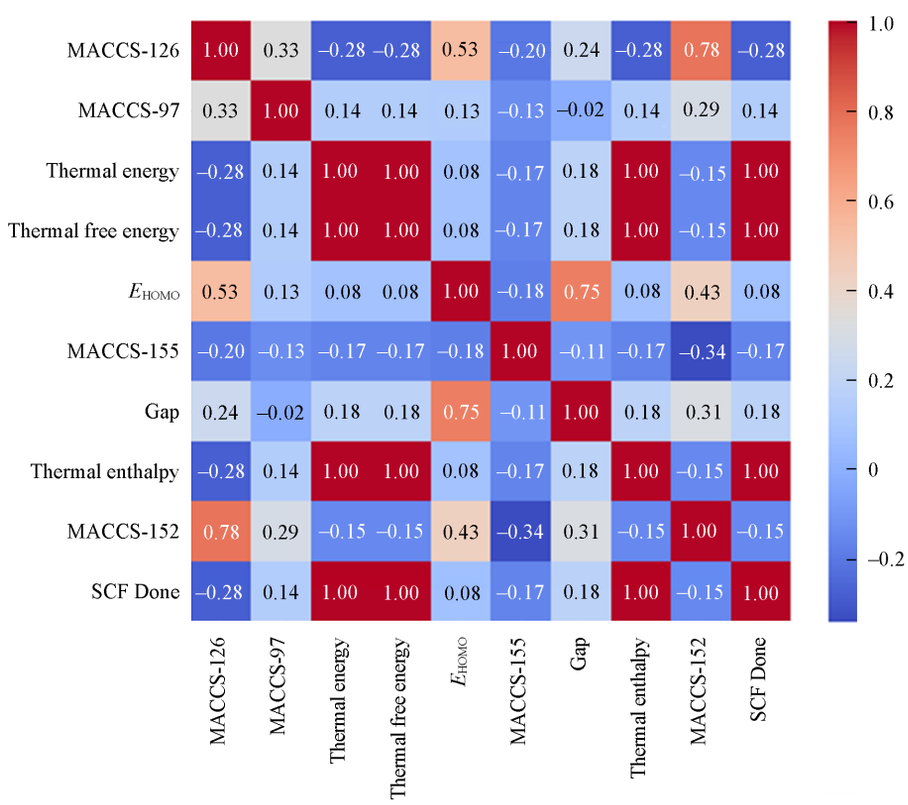


Fig. 6　Correlation analysis of the most important 10 substructures of the XGB-Combined model

基于上述SHAP分析解释，列出了4种不同结构的聚酰亚胺单体结构和其T_g（图7）. 在这4种结构中，PI-1比PI-2具有更高的T_g，而PI-2比PI-1有更多的芳香环结构和醚键基团，说明子结构126的增多对T_g有负贡献，符合SHAP值分析结果，但不能排除醚键对T_g的影响重大. PI-3比PI-4多了一个亚甲基（即子结构152），T_g下降，因为子结构152对T_g有负贡献. PI-3和PI-4显然比PI-1和PI-2增加了芳香环，且二酐和二胺的数量都有增加，但T_g有所下降，这表明增加刚性结构（芳香环）并不一定会使T_g升高，一方面要考虑其它基团的协同作用，另一方面也要考虑聚合后整体链段的刚性和运动性.

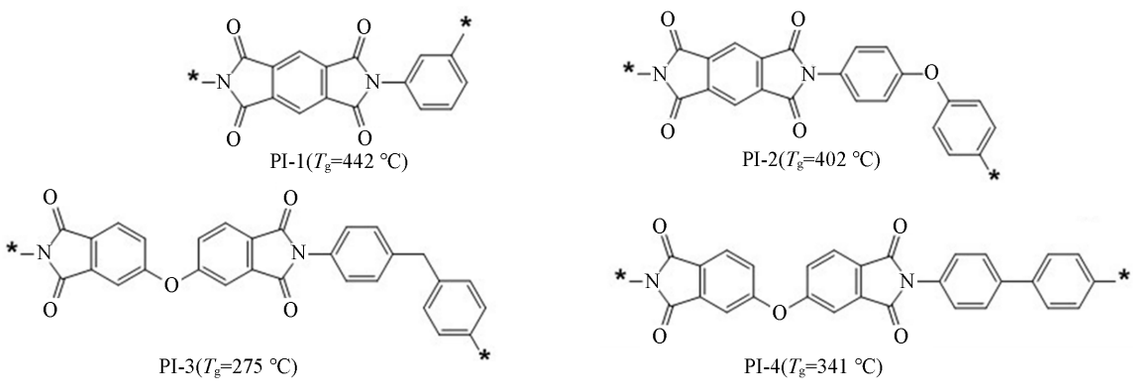


Fig. 7　Structures of PI-1, PI-2, PI-3 and PI-4

事实上，这些子结构和描述符对T_g的影响并非独立的，也并非不变的，每个子结构之间可能存在协同效应，同一个子结构可能在不同的聚合物中有截然相反的影响. 此外，不同子结构和描述符的组合对聚合物性能的影响也有显著差异. 因此，在分子设计中，应注重整合每个子结构及其协同作用对性能的影响，系统性地分析并优化综合贡献最大的子结构组合，以实现性能的最大化. 这种策略能够更精准地指导新型聚酰亚胺的设计和开发.

3　结　　论

为了加快聚酰亚胺材料的开发，本文通过整合文献数据、量子化学计算和机器学习方法，构建了一种通用的聚酰亚胺T_g预测模型. 该模型考虑了166位MACCS指纹图谱、9种量子化学描述符（E_{HOMO}，E_{LUMO}，Gap，μ，E，Zero-point energy，Thermal energy，Thermal enthalpy和Thermal free energy）以及两者的集成，比较了XGB，SVR，RF和GBR等4种模型的预测效果. 结果表明，除了RF-Combined模型外，XGB-Combined，SVR-Combined和GBR-Combined模型的测试集决定系数R^2均高于0.8，并且所有集成模型的R^2都比单一分子指纹输入和单一量子化学描述符输入模型的R^2高. 对于表现最好的XGB-Combined模型，训练集R^2达到了0.956，测试集R^2达到了0.811（比单一分子指纹模型高0.026，比单一量子化学描述符模型高0.236），表明集成模型有效结合了分子结构和量子化学信息，凸显了量子化学描述符在模型中的重要性. 利用SHAP值分析了重要子结构和量子化学信息的影响，从分子结构和量子化学的角度揭示了影响T_g的关键因素. 分析发现，子结构126，97，155和量子化学性质Gap出现频率高，且对T_g的预测值影响较大，表明对T_g影响因素的关键是芳香环、柔性结构以及分子结构稳定性，因此，可以在分子设计实验中特别关注这类结构的聚酰亚胺，从而提高聚酰亚胺的T_g. 综上所述，基于分子指纹和量子化学描述符相结合的机器学习方法不仅可显著提高单一输入的性能，还能降低实验试错成本. 通过解释分析模型，可以精准调控聚酰亚胺的分子设计，辅助开发各种高性能的聚酰亚胺材料. 在未来的工作中，我们将通过图神经网络（GNN）进一步获得更准确的聚酰亚胺的聚集结构与周期信息，还将考虑加工工艺参数与实际生产结合，开发聚合物高精度模型，以辅助新聚酰亚胺材料以及新聚合物材料的开发.

参　考　文　献

[1] Liaw D. J., Wang K. L., Huang Y. C., Lee K. R., Lai J. Y., *Prog. Polym. Sci.*, **2012**, *37*(7), 907—974

[2] Negi Y. S., Damkale S. R., Ansari S., *J. Macro. Sci.*, **2001**, *41*(1/2), 119—138
[3] Tao K., Qin F., Li Y., Zhang S., Han S., Liu G., Wang J., Shen J., Yang Z., Tang Y., Sun G., *High Perform. Polym.*, **2022**, *34*(9), 998—1008
[4] Ni L., Luo Y., Zhou C., Meng H., Wang G., Yan L., Liang M., Qiu S., Zhou S, Zou H., *Polymer*, **2024**, *298*, 126891
[5] Yang X., Geng W., Bi K., Mei L., Li Y., He J., Mu J., Hou X., Chou X., *Micromachines*, **2021**, *12*(1), 70—73
[6] Kimoto Y., Fujita T., Furuta N., Kitamura A., Suzuki H., *J. Spacecr. Rockets*, **2016**, *53*(6), 1028—1034
[7] Yi C., Li W., Shi S., He K., Ma P., Chen M., Yang C., *Sol. Energy*, **2020**, *195*, 340—354
[8] Wan B., Zheng M. S., Yang X., Dong X., Li Y., Mai Y. W., Chen G., Zha J. W., *Energy Environ. Mater.*, **2023**, *6*(1), e12427
[9] Zhou Z., Wang Y., Xu D., Zhang Y., *Sol. Energy Mater. Sol. Cells*, **2010**, *94*(12), 2042—2045
[10] Qiu X. S., Wu Q., Shi D. X., Zhang Y. Y., Chen K. C., Li H. S., *Chem. J. Chinese Universities*, **2022**, *43*(8), 20220140(仇心声, 吴芹, 史大昕, 张耀远, 陈康成, 黎汉生. 高等学校化学学报, 2022, 43(8), 20220140)
[11] Edwards W. M., Ivan M. R., *Polyimides of Pyromellitic Acid*, US 2710853A, 1955-06-14
[12] Frazer A. H., *J. Polymer Science Part A-1*: *Polymer Chemistry*, **1969**, *7*(8), 2464—2464
[13] Yang Y., Jung Y., Cho M. D., Lee S. G., Kwon S., *RSC Adv.*, **2015**, *5*(71), 57339—57345
[14] Zhang D., Wang C., Wang Q., Wang T., *Tribology International*, **2019**, *140*, 105728
[15] Zhou S. W., Yu C., Chen M., Shi C. Y., Gu R., Qu D. H., *Smart Molecules*, **2023**, *1*(2), e20220009
[16] Du Y., Jamasb A. R., Guo J., Fu T., Harris C., Wang Y., Blundell T. L., *Nat. Mach. Intell.*, **2024**, *6*(6), 589—604
[17] Sanchez-Lengeling B., Aspuru-Guzik A., *Science*, **2018**, *361*(6400), 360—365
[18] Kneiding H., Balcells D., *Chem. Sci.*, **2024**, *15*(38), 15522—15539
[19] Karthikeyan A., Priyakumar U. D., *J. Chem. Sci.*, **2021**, *134*(1), 1—20
[20] Prašnikar E., Ljubič M., Perdih A., Borišek J., *Artif. Intell. Rev.*, **2024**, *57*(4), 102—103
[21] Faber F. A., Hutchison L., Huang B., Gilmer J., Schoenholz S. S., Dahl G. E., von Lilienfeld O. A., *J. Chem. Theory Comput.*, **2017**, *13*(11), 5255—5264
[22] Yue D., Feng Y., Liu X. X., Yin J. H., Zhang W. C., Guo H., Lei Q. Q., *Adv. Sci.*, **2022**, *9*(17), 2105773
[23] Toland A., Tran H., Chen L., Li Y., Zhang C., Gutekunst W., Ramprasad R., *J. Phys. Chem. A*, **2023**, *127*(50), 10709—10716
[24] Zhao G., Xu T., Fu X., Zhao W., Wang L., Lin J., Du L., *Compos. Sci. Technol.*, **2024**, *248*, 110455
[25] Lightstone J. P., Chen L., Kim C., Batra R., Ramprasad R., *J. Appl. Phys.*, **2020**, *127*(21), 215105
[26] Liu T. L., Liu L. Y., Ding F., Li Y. Q., *Chin. J. Polym. Sci.*, **2022**, *40*(7), 834—842
[27] Hou Z. J., Li R. Q., Li J., Feng Y. N., Jin Q. Q., Sun J. H., Cao J., *Chem. J. Chinese Universities*, **2024**, *45*(9), 20240199(侯泽金, 李荣其, 李健, 冯怡宁, 靳茜茜, 孙俊红, 曹洁, 高等学校化学学报, **2024**, *45*(9), 20240199)
[28] Couchman P. R., *J. Appl. Phys.*, **1979**, *50*(10), 6043—6046
[29] Pásztor S., Becsei B., Szarka G., Thomann Y., Thomann R., Mühlhaupt R., Iván B., *Materials*, **2020**, *13*(21), 4822—4837
[30] Alesadi A., Cao Z., Li Z., Zhang S., Zhao H. Y., Gu X. D., Xia W. J., *Cell Rep. Phys. Sci.*, **2022**, *3*(6), 100911
[31] Huang C. C., Zhang B. Q., Liu C. Y., *Chem. J. Chinese Universities*, **2021**, *42*(8), 2617—2626(黄聪聪, 张宝庆, 刘琛阳. 高等学校化学学报, **2021**, *42*(8), 2617—2626)
[32] Qiu H., Qiu X., Dai X., Sun Z. Y., *J. Mater. Chem. C*, **2023**, *11*(8), 2930—2940
[33] Ding M. X., *Polyimide*: *Chemistry*, *Structure*, *and Relationship to Properties and Materials*, Chemical Industry Press, Beijing, **2006**, 221—504(丁孟贤. 聚酰亚胺: 化学, 结构与性能的关系及材料. 北京: 科学出版社, **2006**, 221—504)
[34] Weininger D., *J. Chem. Inf. Model.*, **1988**, *28*(1), 31—36
[35] Kwon J. H., Han J. Y., Kim M., Kim S. K., Lee D. K., Kim M. G., *Arch. Pharm. Res.*, **2024**, *47*(12), 914—923
[36] Rogers D., Hahn M., *J. Chem. Inf. Model*, **2010**, *50*(5), 742—754
[37] Zhang Y., Xu X., *Polymer Chemistry*, **2021**, *12*(6), 843—851
[38] Zhang Y., Xu X., *Mol. Cryst. Liq. Cryst.*, **2021**, *730*(1), 9—22
[39] Frisch M. J., Trucks G. W., Schlegel H. B., Scuseria G. E., Robb M. A., Cheeseman J. R., Scalmani G., Barone V., Mennucci B., Petersson G. A., Nakatsuji H., Caricato M., Li X., Hratchian H. P., Izmaylov A. F., Bloino J., Zheng G., Sonnenberg J. L., Hada M., Ehara M., Toyota K., Fukuda R., Hasegawa J., Ishida M., Nakajima T., Honda Y., Kitao O., Nakai H., Vreven T., Montgomery J. A. Jr., Peralta J. E., Ogliaro F., Bearpark M., Heyd J. J., Brothers E., Kudin K. N., Staroverov V. N., Kobayashi R., Normand J., Raghavachari K., Rendell A., Burant J. C., Iyengar S. S., Tomasi J., Cossi M., Rega N., Millam J. M., Klene M., Knox J. E., Cross J. B., Bakken V., Adamo C., Jaramillo J., Gomperts R., Stratmann R. E., Yazyev O., Austin A. J., Cammi R., Pomelli C., Ochterski J. W., Martin R. L., Morokuma K., Zakrzewski V. G., Voth G. A., Salvador P., Dannenberg J. J., Dapprich S., Daniels A. D., Farkas Ö., Foresman J. B., Ortiz J. V., Cioslowski J., Fox D. J., *Gaussian* 09, *Revision E*. 01, *Gaussian Inc.*, Wallingford CT, **2009**
[40] Zhao Y., Truhlar D. G., *Theoretical Chemistry Accounts*, **2008**, *120*(1), 215—241
[41] Breiman L., *Machine Learning*, **2001**, *45*(1), 5—32
[42] Smola A. J., Murata N., Schölkopf B., Müller K.R., *Asymptotically Optimal Choice of ε-Loss for Support Vector Machines*, *Proceedings of the ICANN 98*, Springer, London, **1998**
[43] Chen T., Guestrin C., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, **2016**

(Ed.: W, K, M)