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

基于随机森林的有机小分子的化学键解离能预测

栾　玥, 孔丁羚, 郭莉莉, 张庆友, 周艳梅
(河南大学化学与分子科学学院, 河南省工业水循环利用工程技术研究中心, 开封 475004)

摘要　从iBonD有机物键能数据库中手动收集1208个含C, H, O, N和S原子的有机分子, 并记录相应的化学键解离能实验值. 提出了化学键类型描述符、杂原子描述符和支化度描述符, 并与此前提出的原子类型描述符结合, 从而更全面地描述目标化学键的周边环境. 采用随机森林建立键解离能的预测模型, 结果表明目标化学键周围的原子类型和化学键类型的描述符组合建模得到的预测结果最佳, 在没有量子化学辅助的情况下得到了较好的预测结果. 与已报道的预测结果进行比较发现, 本文结果优于文献中的相应结果. 此外, 还设计了一个应用域算法来初步判断预测结果的质量, 重新随机划分训练集和测试集来验证模型的稳定性, 与零模型比较来判断模型的可行性.

关键词　键解离能; 随机森林; iBonD; 原子类型; 化学键类型
中图分类号　O657　　**文献标志码**　A　　doi: 10.7503/cjcu20240373

Prediction of Chemical Bond Dissociation Energies of Small Organic Molecules Based on Random Forest

LUAN Yue, KONG Dingling, GUO Lili, ZHANG Qingyou*, ZHOU Yanmei*
(*Henan Engineering Research Center of Industrial Circulating Water Treatment*, *College of Chemistry and Molecular Sciences*, *Henan University*, *Kaifeng* 475004, *China*)

Abstract　1208 organic molecules containing C, H, O, N, and S were manually collected from the iBonD organic bond energy database, and the corresponding experimental bond dissociation energy values were recorded. Chemical bond type descriptors, heteroatomic count descriptors, and branch descriptors were proposed and combined with previously suggested atomic type descriptors to provide a more comprehensive description of the surrounding environment of the target chemical bond. The prediction models for bond dissociation energy were constructed using random forest, and the results show that the combination of the descriptors of atomic types and chemical bond types around the target chemical bond achieves the best prediction results, and satisfactory results were obtained without quantum chemistry assistance. Compared with the results in published literature, the predicted results herein are better than the corresponding results in the literature. In addition, an algorithm on the application domain was designed to assess the quality of prediction results in advance, and the training set and the test set were randomly re-partitioned to verify the stability of the model, as well as the feasibility of the model was evaluated by comparing it with a zero model.

Keywords　Bond dissociation energy; Random forest; iBonD; Atom type; Bond type

收稿日期: 2024-07-30. 网络首发日期: 2024-09-12.
联系人简介: 张庆友, 男, 博士, 教授, 主要从事化学信息学方面的研究. E-mail: qingyou@vip.henu.edu.cn
周艳梅, 女, 博士, 教授, 主要从事生物质基功能材料等方面的研究. E-mail: zhouym@henu.edu.cn
基金项目: 国家自然科学基金(批准号: 22278112)资助.
Supported by the National Natural Science Foundation of China(No.22278112).

化学键的形成和断裂是大部分化学转化的核心过程，而键解离能(Bond dissociation energy，BDE)作为表征化学键强度的重要物理量，对于理解化学反应机理、设计化学反应以及探索物质的内在属性具有重要意义[1~3]. 因此，对键解离能的研究日益增多，而机器学习方法因其计算速度快和成本低等优点也被应用于对键解离能的预测[4~6]. 如Meng等[7]通过机器学习方法预测了各种具有饱和氮原子的杂环化合物中C—N键的BDE；Gou等[8]基于文献检索和量子化学计算得到了一个含能材料的键解离能数据集，在此基础上构建了一个预测键解离能的机器学习模型；Raza等[9]使用多种机器学习算法(包括随机森林，前馈神经网络、最小绝对值收缩和选择算子回归)预测了C—F键的BDE；Yu等[10]使用机器学习建模预测羰基的BDE，其模型基于相关位置的键长和键角作为描述符，准确预测了C═O键的BDE. 然而，与大多数研究相似，其用于预测化学键解离能的数据来自量子化学计算，而非实验测量[11~13]. 由于量子化学计算本身存在一定误差，这影响了预测结果的准确性.

前文[14]摘录了有机小分子中C—H键的实验BDE值，通过量子化学计算的BDE值辅助构建随机森林(Random Forest，RF)模型，得到了较好的实验BDE值预测结果. 在此基础上，本文升级了结构特征描述符，并将该方法扩展到更复杂的不含氢化学键的解离能预测，在未采用量化计算BDE值辅助的情况下，分别采用随机森林、偏最小二乘和支持向量机建立了有机小分子中不含氢化学键的实验BDE值预测模型.

1 计算方法

1.1 数据集

BDE的数据来源于iBonD有机物键能数据库[15,16]. 从该数据库中摘录了1208个含C，H，O，N和S原子的有机小分子，每个分子中只有一个化学键(不含R—H)的实验BDE. 为了确保评估结果的可靠性，按照4∶1的比例将数据集随机划分为训练集和测试集，即966个化合物用作训练集，242个化合物用作测试集.

1.2 机器学习方法

1.2.1 随机森林 采用随机森林(由R程序4.2.2版本中的随机森林库实现)建立了机器学习模型. 它是多棵回归树的集成，每棵树都独立建模和预测，最终输出结果是所有回归树预测结果的平均值[17]. 由于随机森林结合了多棵回归树的预测结果，因此通常比单棵回归树有更高的预测准确性[18]. 它能够有效处理具有多变量和样本的数据集，且不需要对变量进行降维，因而广泛应用于各种机器学习任务中. 此外，随机森林采用袋外交叉验证结果评价训练集，通过随机选择样本和变量来避免出现过拟合[19]. 本研究采用1000棵树建模，其它参数均为默认值. Proximity是随机森林内置的一个功能，用于计算两个样本之间的相似度，其定义为两个样本落入同一个端点的次数除以树的总数. 可见，这种相似度计算值同时依赖于随机森林和描述符.

1.2.2 偏最小二乘 偏最小二乘(Partial least squares，PLS)是一种传统的多元统计分析方法，是主成分分析和多元线性回归的结合，因其实用性和通用性而成为常用的建模技术[20,21]. 本文中PLS算法在Weka软件(3.8.6版本)中实现[22].

1.2.3 支持向量机 支持向量机(Support vector machine，SVM)是一种监督的学习算法，它可以用于分类或回归问题，SVM通过引入核函数来处理线性不可分的数据，通过将数据映射到高维空间中，使得在原始空间中线性不可分的数据在新的空间中变得线性可分[23]. 本文中SVM在Weka软件(3.8.6版本)中实现.

1.3 描述符

在定量构效关系研究中，描述符的提取至关重要[24,25]. 化学键的解离能与其周边环境的杂原子、空间位阻、原子类型和化学键类型以及所属环信息等因素相关[12,26,27]，因此本研究提取了相应的结构特征描述符，用于预测化学键的解离能.

1.3.1 杂原子数描述符 在本数据集中，含有O，N和S 3个杂原子(非碳和氢的非金属原子). 如果

一个化学键的周边环境中存在杂原子，通常会影响化学键的解离能，而且杂原子离该化学键越近对解离能值的影响越大. 由此，基于层的概念，定义每一层各杂原子的数量作为描述符：第一层仅含中心原子，第二层与中心原子拓扑距离为1，…，第n层与中心原子拓扑距离为n−1. 由此，杂原子数(Heteroatomic count，HC)描述符的组成可以简化为图1所示的3n维描述符，其中n为层数.

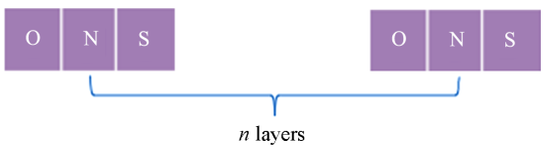

Fig. 1　Diagram of composition of the heteroatomic count descriptors

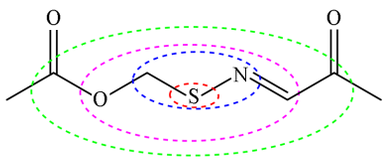

Fig. 2　Diagram of layers around a center atom

以图2中的分子示例HC描述符的生成如下：分子的第一层为中心原子(红色椭圆标识的一个S原子)，即第三个描述符为1；第二层(蓝色椭圆)包含一个氮原子，即第五个描述符为1；第三层(粉色椭圆)包含一个氧原子，即第七个描述符为1；第四层(绿色椭圆)没有杂原子；第五层包含两个氧原子，即第13个描述符为2. 详细信息可参见表1.

Table 1　Heteroatomic Count descriptors of the molecule in Fig.2

Descriptor number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Layer number	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Corresponding heteroatom	O	N	S	O	N	S	O	N	S	O	N	S	O	N	S
HC descriptors	0	0	1	0	1	0	1	0	0	0	0	0	2	0	0

1.3.2　*支化度描述符*　空间位阻是影响键解离能的另一个重要因素，与一个化学键周围每一层的非氢原子数量密切相关，由此，提取每一层的非氢原子数量作为支化度(Branch)描述符，其中第一个描述符为与中心原子拓扑距离为1的非氢原子数量，并依此类推. 由于不考虑中心原子，n层的支化度描述符是n−1维. 以图2为例，以红色圆圈内的硫原子为中心原子，第一层的蓝色圆圈由1个碳原子和1个氮原子组成，即第一个描述符为2；依此类推，第二个描述符为2；第三个描述符为2；第四个描述符为4.

1.3.3　*扩展原子类型描述符*　与本课题组[14]之前提出的用于预测C—H键的原子类型描述符类似，本文也采用扩展原子类型(Modified atomic type，MAT)描述符来代表原子. MAT描述符是由原子类型(Atomic type，AT)描述符扩展得到的. AT是基于原子类型和原子间的拓扑距离所衍生的指数，每个非氢原子都是通过中心原子周围的各层原子类型的数目来描述. 根据本数据集中化合物的结构特征定义了32种原子类型，并按元素(O，N，S，C)、邻接非H原子数量和H原子数量进行排序(详情见本文支持信息). 不同的是，前文[14]是对含氢化学键的解离能进行预测，忽略了氢原子的描述，即在生成MAT时仅以化学键的非氢原子为中心，生成该原子的MAT描述符. 而本文研究对象是不含氢化学键的解离能，1个化学键的2个非氢原子均需生成MAT描述符，由此产生了2个MAT描述符需要排序的问题. 为此，定义了化学键的2个原子的MAT描述符的优先级判断规则：对于由原子A和原子B组成的化学键，描述符按照杂原子、支化度和MAT描述符由大到小进行排序.

由此，每一层均由原子的32个原子类型得到32个描述符，扩展到n层，每个中心原子将生成32n个描述符. 第1到32维描述符代表第一层(仅含中心原子)，所对应原子类型的描述符为1；第33到64维描述符代表第二层(与中心原子拓扑距离为1的原子)；描述符为该层各类型的原子数，依此类推. 每层32维，则五层共计160维.

在32种原子类型中，除芳香性间接暗示环结构外，未引入环信息. 然而，原子周围环的数量和大小是影响BDE的重要因素. 因此，在AT描述符的基础上，增加了20个新的描述符用来表示中心原子所在的三元环到二十二元环的数量，称为MAT. 同样，MAT描述符的组成可以简化为图3所示的32n+20维的描述符.

以图2为例，分子第一层(红色圆圈)的原子类型为—S—；第二层(蓝色圆圈)的原子类型为═N—和>CH_2；第三层(粉色圆圈)的原子类型为—O—和═CH—；第四层的原子类型为>C═；第五层的原子类型为—CH_3和═O.

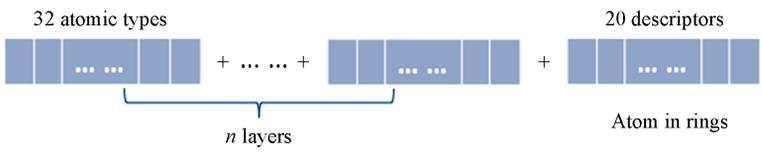

Fig. 3　Diagram of composition of atomic type descriptors

1.3.4　扩展化学键类型描述符　此前进行含氢化学键解离能预测时，仅描述了原子类型(MAT指数)．为了能够体现化学键信息，本文提出了化学键类型描述符．根据数据集中化合物的结构特点提出了28个化学键类型(Bond type，BT)，并将各类型化学键的数量作为描述符(见本文支持信息)．与原子描述符类同，化学键描述符是基于化学键类型和中心化学键间的拓扑距离(称为层数)所衍生的指数，即每个键生成28n个描述符．同样，n是化学键周围的层数．此外，在BT描述符的基础上，增加了20个新的描述符用来表示中心化学键所在的三元环到二十二元环的数量，由此所得到的描述符称为扩展化学键类型(Modified bond type，MBT)描述符．MBT描述符的组成可以简化为图4所示的28n+20维描述符．

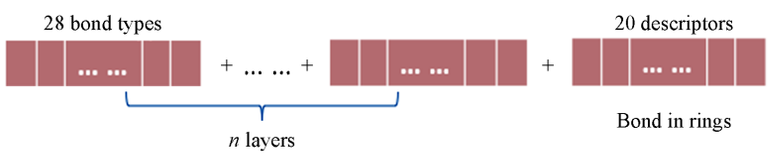

Fig. 4　Diagram of composition of bond type descriptors

以图5为例，该分子第一层(红色圆圈)有1个C—O键，第二层(蓝色圆圈)为2个C—C键和1个N—O键，第三层(粉色圆圈)为2个C—O键和2个N═O键，第四层(绿色圆圈)为2个N—O键，第五层为4个N═O键，以这些键的数量作为该分子的MBT描述符．

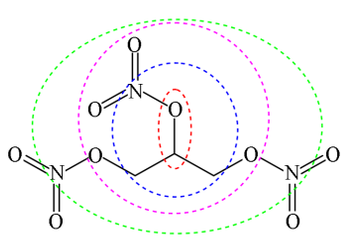

Fig. 5　Diagram of layers around a center bond

1.3.5　HAB描述符　将HC描述符、MAT描述符和MBT描述符组合成的描述符称为HAB描述符，可以简化为图6所示的2×(3n+32n+20)+28n+20=98n+60维描述符．

Fig. 6　Composition diagram of the HAB descriptor

为了实现描述符的自动生成，采用Java开发了前述程序．

2　结果与讨论

2.1　基于随机森林的BDE预测

采用随机森林，基于上述提取的描述符建立QSPR模型，并通过训练集的交叉验证和测试集的预测结果来比较模型的性能．

2.1.1　基于HC和Branch描述符的BDE预测　首先基于HC(3n×2)和Branch(n−1)×2描述符，使用RF构建模型来预测数据集中化学键的BDE．当n=5时，描述符的数量为38．将训练集提交给随机森林，使

用OOB交叉验证对训练集进行评估，然后将测试集提交给训练的模型进行预测，结果见表2.

Table 2　BDE prediction results based on different descriptors

Descriptors/Number of descriptors	Training set(cross-validation) R^2/MAE/RMSE/(kcal·mol^{-1})*	Test set R^2/MAE/RMSE/(kcal·mol^{-1})
HC+Branch/38	0.6860/7.68/11.79	0.7904/8.89/13.03
MAT/360	0.8675/5.82/9.72	0.8755/5.99/10.18
HC+Branch+MAT/398	0.8874/5.45/8.96	0.8709/6.34/10.27
MBT/160	0.8719/5.73/9.54	0.8699/6.22/10.31
HC+Branch+MBT/198	0.8891/5.34/8.88	0.8915/5.86/9.48
MAT+MBT/520	0.8894/5.22/8.88	0.8935/5.38/9.35
HAB/550	0.8977/4.96/8.53	0.8980/5.28/9.17
Branch+HAB/558	0.8963/5.13/8.60	0.8920/5.63/9.41

* 1 kcal≈4185. 85 J.

由表2可知，训练集的预测结果为R^2=0. 6860，MAE=7. 68 kcal/mol，RMSE=11. 79 kcal/mol；对于测试集，R^2=0. 7904，MAE=8. 89 kcal/mol，RMSE=13. 03 kcal/mol. 可见，模型对BDE的预测能力较差，这是因为HC和Branch描述符无法全面捕捉化合物中结构特征的细节，因而需要引入更多基于原子和化学键的结构描述符来提高模型的预测能力.

2.1.2　基于MAT描述符的BDE预测　基于MAT描述符，采用RF构建模型预测数据集中化合物的BDE. 当n=5时，MAT描述符的数量为360. 预测结果如表2所示，对于训练集，R^2=0. 8675，MAE=5. 82 kcal/mol，RMSE=9. 72 kcal/mol；对于测试集，R^2=0. 8755，MAE=5. 99 kcal/mol，RMSE=10. 18 kcal/mol. 此结果显著优于2. 1. 1节预测结果，这是因为加入MAT描述符后，补充了对中心原子周围各层的原子类型的描述，这说明原子类型是影响BDE的一个重要因素.

2.1.3　基于MBT描述符的BDE预测　为了对数据集中化合物的特征进行更充分的描述，采用化学键描述符(MBT)来建立模型. 当n=5时，MBT描述符的数量为160. 预测结果如表2所示，对于训练集，R^2=0. 8719，MAE=5. 73 kcal/mol，RMSE=9. 54 kcal/mol；对于测试集，R^2=0. 8699，MAE=6. 22 kcal/mol，RMSE=10. 31 kcal/mol. 训练集的预测结果比基于MAT描述符的建模结果好，但测试集的预测结果却略差. 这说明化学键类型同样是影响BDE的重要因素之一. 同时，由于MAT反映了原子的结构特征，而MBT反映了化学键的特征，双方各有侧重，但是都缺乏全面性，这也暗示二者的结合可能更好地反映目标化学键的周边环境.

2.1.4　基于HC，MAT和MBT组合描述符的BDE预测　对基于HC，Branch，MAT和MBT组合描述符的模型也进行了考察. 预测结果如表2所示，使用HAB描述符建模的结果最佳，训练集的预测结果为R^2=0. 8977，MAE=4. 96 kcal/mol，RMSE=8. 53 kcal/mol；测试集的预测结果为R^2=0. 8980，MAE=5. 28 kcal/mol，RMSE=9. 17 kcal/mol. 这说明HC描述符，MAT描述符和MBT描述符结合能对化合物的特征进行更完整的描述，获得更好的预测结果.

为了进一步观察使用HAB描述符建模得到的预测结果，预测值与实验值的散点图以及关于实验值与预测值误差的散点图如图7所示.

图7(A)和(B)结果显示，训练集和测试集的实验值与预测值之间呈现较高的相关性，相关系数R^2≈0. 90. 在训练集和测试集中，大部分数据的误差都在0 kcal/mol附近，并且只有小部分>20 kcal/mol，训练集中有两个数据的误差>80 kcal/mol，测试集中只有一个数据的误差>80 kcal/mol. 这一结果说明模型对于绝大部分化合物的预测结果是令人满意的.

为了进一步验证模型，基于本实验的训练集建立了零模型. 对训练集进行三折交叉验证，得到的预测结果如下：对于训练集，MAE=38. 08 kcal/mol，RMSE=54. 74 kcal/mol；对于测试集，MAE=36. 53 kcal/mol，RMSE=51. 64 kcal/mol. 如表2所示，实验中预测结果均明显优于零模型，说明所建立的模型具备可行性.

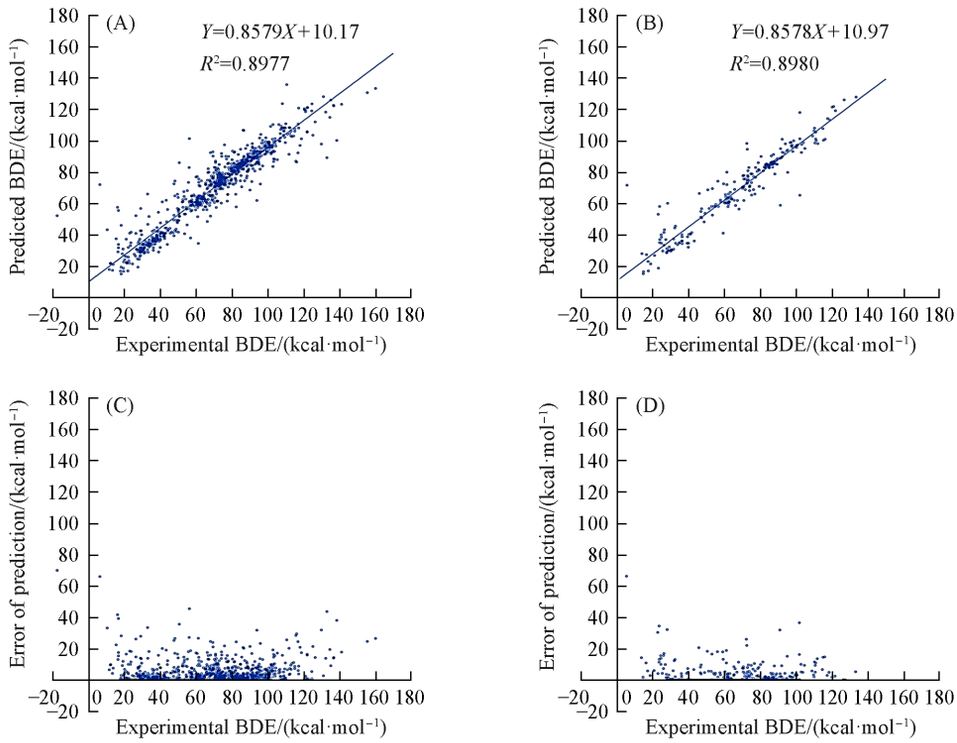

Fig. 7　Graphs of prediction results modeled using HAB descriptors

(A) Scatter plot of experimental and predicted values of the training set; (B) scatter plots of experimental and predicted values for the test set; (C) scatter plot of the error between experimental and predicted values of the training set; (D) scatter plot of the error between the experimental and predicted values of the test set.

2.2　描述符层数对模型的影响

MAT和MBT描述符都依赖于层数，因此中心原子描述符的层数有可能影响预测能力. 分别基于中心原子周围非氢原子达到6和7层时的MAT+MBT，HAB和HAB+Branch描述符建模，与5层描述符的建模结果进行了比较. 其中，每个模型均是在删除全零描述符后使用随机森林建立，结果如表3所示. 表3结果表明，采用HAB描述符进行建模得到的结果最佳，且使用5层、6层和7层HAB描述符获得的预测结果非常接近. 在训练集中，使用5层描述符时的MAE最佳，为4. 95 kcal/mol；而使用六层描述符时，达到了最佳的R^2和RMSE值；对测试集而言，使用5层描述符时训练集的R^2，MAE，RMSE都优于6层和7层. 虽然层数越多越能更好地反映连接环境，但是预测结果并未随着层数的增加而改善. 这可能是因为远离中心的环境对BDE影响小，而层数增加后变量的数量随之提高，降低了模型的预测能力，在两个因素共同影响下，未能提高预测结果. 在结果相近的情况下，通常采用更少的描述符，因此本研究采用5层.

Table 3　Prediction results of different description layers BDE based on RF

Descriptors/Number of descriptors/Number of layers	Training set(cross-validation) R^2/MAE/RMSE/(kcal·mol^{-1})	Test set R^2/MAE/RMSE/(kcal·mol^{-1})
MAT+MBT/352/5	0.8886/5.24/8.91	0.8919/5.42/9.43
MAT+MBT/403/6	0.8895/5.24/8.91	0.8903/5.41/9.48
MAT+MBT/449/7	0.8890/5.26/8.88	0.8891/5.50/9.55
HAB/379/5	0.8988/4.95/8.48	0.8977/5.31/9.17
HAB/434/6	0.8992/4.97/8.47	0.8956/5.34/9.28
HAB/484/7	0.8987/4.97/8.49	0.8957/5.33/9.26
HAB+Branch/387/5	0.8973/5.08/8.56	0.8930/5.57/9.36
HAB+Branch/444/6	0.8955/5.19/8.63	0.8905/5.65/9.45
HAB+Branch/496/7	0.8977/5.16/8.55	0.8918/5.63/9.42

2.3 应用域

应用域(Application domain, AD)是化学空间中的一个理论区域, 由构建模型的描述符以及化合物的响应数值所定义[28~30]. 引入应用域有助于增强机器学习构建的QSPR模型的实用性. 因为从机器学习角度来说, 应用域的设定可以防止因为待测试样本的结构特征与建模样本区别太大而导致的预测偏差过大的情况[31]. 从另一个角度, 若通过合适的应用域在评估前排除不适合的化合物样本, 同样有助于提高模型的准确性[32].

基于随机森林的Proximity功能计算一个被递交到模型的样本与训练集中所有化学键的相似度, 并将样本与训练集中最相似化学键的相似度作为它与训练集的相似度. 如果这个样本与训练集中的一个化学键的描述符完全相同, 则在每棵树上二者均会落入同一个端点, 即二者的相似度为1; 若两个样本从未落入同一个端点则相似度为0. 基于此, 将根据测试集中的样本与训练集的相似度来统计BDE预测值的误差, 以此定义模型的应用域.

实验分别计算了使用5层和6层HAB描述符建模的应用域, 计算结果列于表4. 当相似度>0.9时, 误差分别为2.77和2.44 kcal/mol; 相似度在0.9~0.7之间的误差为3.65和4.75 kcal/mol, 均优于测试集整体的MAE. 当相似度在0.5~0.3之间和0.3~0之间时, 误差明显差于测试集整体的结果. 可以看出, 整体趋势上, 与训练集的相似度越低, 测试集的预测结果越差. 基于此可以判断, 相似度>0.9的结果为优秀预测结果, 相似度<0.3的结果为不可靠的预测结果. 可见, 可以预先通过应用域初步评估模型对未知样本的预测能力, 从而预先排除可能预测偏差大的化学键, 提高模型的预测能力.

Table 4 Error under different proximities

Proximity	> 0.9	0.9—0.7	0.7—0.5	0.5—0.3	0.3—0
(5 layers)MAE(Number of compounds)	2.77(46)	3.65(49)	5.31(78)	5.81(53)	8.83(16)
(6 layers)MAE(Number of compounds)	2.44(28)	4.75(60)	4.00(82)	6.00(56)	7.52(16)

2.4 基于其他机器学习方法的BDE预测

分别采用偏最小二乘和支持向量机进行建模, 用于预测BDE实验值.

2.4.1 基于偏最小二乘的BDE预测 基于上述提出的HAB描述符建立PLS模型, 采用训练集的三折交叉验证考察了PLS中主成分数对BDE预测的影响, 预测结果列于表S3(见本文支持信息). 由表S3可见, 训练集的R^2在0.8250~0.8573 kcal/mol之间, 其MAE在6.49~7.56 kcal/mol之间, RMSE在10.16~11.15 kcal/mol之间; 测试集的R^2在0.8143~0.8403 kcal/mol之间, MAE在6.80~8.17 kcal/mol之间, RMSE在11.15 kcal/mol和12.22 kcal/mol之间. 这些预测结果均差于基于HAB描述符使用RF建模的结果, 暗示BDE值与描述符之间并不具备良好的线性加和性.

2.4.2 基于支持向量机的BDE预测 基于HAB描述符建立了SVM模型, 其中核函数采用nu-svr(regression), 评估性能后最佳的参数组合为C=700, Gamma=0.03. 基于此模型, 训练集的三折交叉验证结果为R^2=0.8738, MAE=5.39 kcal/mol, RMSE=9.46 kcal/mol; 测试集的预测结果为R^2=0.8781, MAE=5.35 kcal/mol, RMSE=9.73 kcal/mol. 结果表明, SVM也是一种较好的预测BDE值的机器学习方法, 所得结果好于PLS. 但是与RF的预测结果相比, 使用SVM建模的训练集交叉验证和测试集的预测结果均略差, 这可能是由于随机森林能更好地体现化学键之间的结构相似性所致.

2.5 基于文献验证模型

为了与文献[33]结果进行比较, 使用长整型高选择性拓扑指数[34,35]找出该文献测试集中与本文相同的化学键, 均为C—C键. 把这46个数据作为本研究中C—C键的测试集, 剩余的C—C键作为训练集, 得到的测试集预测结果如下: R^2=0.8522, MAE=5.47 kcal/mol, RMSE=7.62 kcal/mol. 而这些化合物在文献[33]中的预测结果为R^2=0.8176, MAE=6.93 kcal/mol, RMSE=8.62 kcal/mol. 可见, 本文预测结果优于文献结果, 表明了本模型的可行性.

2.6 模型稳定性的验证

为了验证模型的稳定性, 重新随机划分了训练集和测试集. 同样按照上文4∶1的比例将数据集分

为966个化合物的训练集和242个化合物的测试集. 基于5层HAB描述符使用RF建立模型, 得到的预测结果如下: 训练集的预测结果为R^2=0.9009, MAE=4.92 kcal/mol, RMSE=8.48 kcal/mol; 测试集的预测结果为R^2=0.8837, MAE=5.60 kcal/mol, RMSE=9.47 kcal/mol. 与原模型相比训练集结果略有提高, 测试集结果略有下降, 说明重新随机划分为训练集和测试集对模型的影响不大, 证明了模型的稳定性. 在此基础上, 进一步考察所建立模型对不同键型的BDE预测能力, 结果如表5所示.

Table 5 Results of BDE for different bonds based on RF

Chemical bond type (Number of compounds)	R^2	Training set(cross-validation) MAE/RMSE/(kcal·mol^{-1})	Test set MAE/RMSE/(kcal·mol^{-1})
O—O(93)	0.6398/0.4617	1.95/3.47	2.00/4.24
O—N(127)	0.7211/0.8494	5.16/9.96	6.45/8.74
O—S(23)	0.2616/0.9193	11.32/13.80	13.55/16.37
O—C(216)	0.8946/0.7837	5.08/7.68	6.03/12.50
N—N(91)	0.6582/0.6442	3.78/7.39	5.28/8.70
N—S(14)	0.3065/0.2564	5.46/9.22	4.66/4.91
N—C(142)	0.8498/0.8426	5.72/8.33	6.57/9.26
S—S(16)	0.2072/0.6208	11.03/13.60	12.72/13.56
S—C(90)	0.7891/0.6365	4.22/7.48	5.09/9.96
C—C(396)	0.7623/0.7979	4.97/8.98	5.07/8.09

由表5可知, C—C键训练集的预测结果为MAE=4.97 kcal/mol; O—C键训练集的预测结果为MAE=5.08 kcal/mol, 接近整体数据集的预测结果. 这是因为C—C键和O—C键在总数据集中占比最多, 分别是32.78%和17.88%, 因此对总体预测结果影响最大. 在所有键型中, O—O键的误差最小, 训练集的MAE=1.95 kcal/mol; N—N键和S—C键预测结果也优于整体预测结果, 说明该模型对这3种键型的BDE的预测能力较好. 而O—S键, N—S键和S—S键的误差最大, 这是因为它们的样本数量太低, 分别是23, 14和16个, 导致建模时能够学习到的结构特征太少, 因此预测能力较差. 未来若能有更多含S键型的数据发表, 将有助于改善模型对含硫键的预测能力.

3 结　　论

作为键解离反应的固有性质, BDE对于理解许多化学过程至关重要, 如药物代谢、生物燃料燃烧和污染物降解等. 对于分析化学、环境科学和药物开发等各个领域的研究人员来说, 基于BDE数据可以为某些应用设计更可靠、更安全的反应途径和更合理的分子结构. 本文基于MAT描述符、HC描述符、MBT描述符、Branch描述符以及HAB描述符, 使用RF建立模型预测了化合物的BDE. 结果表明, 使用HAB描述符得到的预测结果最好. 通过与文献报道的模型比较, 发现本文预测结果优于文献预测结果. 此外, 通过相似度设定应用域, 有助于初步评估模型对未知化合物的预测能力. 还采用PLS和SVM建立了模型, SVM建模结果好于PLS, 但差于RF的建模结果. 研究结果也说明该方法有潜力应用于含其它元素(如含氯和溴)的化学键, 若未来能有更多数量含S化合物的BDE数据发表, 将有助于提高模型的预测能力.

支持信息见http://www.cjcu.jlu.edu.cn/CN/10.7503/cjcu20240373.

参　考　文　献

[1] Liu Y., Li Y., Yang Q., Yang J., Zhang L., Luo S., *Chin. J. Chem.*, **2024**, *42*(17), 1967—1974
[2] Wang P., Gong S., Mo Y., *J. Chem. Phys.*, **2024**, *160*(16), 164302
[3] Nicolaides A., Tomioka H., *J. Phys. Org. Chem.*, **2024**, *37*(6), e4606
[4] Nakajima M., Nemoto T., *Sci. Rep.*, **2021**, *11*(1), 20207
[5] Wen M., Blau S. M., Spotte-Smith E. W. C., Dwaraknath S., Persson K. A., *Chem. Sci.*, **2020**, *12*(5), 1858—1868
[6] S. V S. S., Kim Y., Kim S., St. John P. C., Paton R. S., *Digit Discov.*, **2023**, *2*(6), 1900—1910

[7] Meng Q., Wang R., Shao H., Wang Y., Wen X., Yao C., Qiao J., *J. Phys. Chem. Lett.*, **2024**, *15*(16), 4422—4429
[8] Gou Q., Liu J., Su H., Guo Y., Chen J., Zhao X., Pu X., *iScience*, **2024**, *27*(4), 109452
[9] Raza A., Bardhan S., Xu L., Yamijala S., Lian C., Kwon H., Wong B., *Environ. Sci. Tech. Lett.*, **2019**, *6*(10), 624—629
[10] Yu H., Wang Y., Wang X., Zhang J., Ye S., Huang Y., Luo Y., Sharman E., Chen S., Jiang J., *J. Phys. Chem. A*, **2020**, *124*(19), 3844—3850
[11] Bao J., Welch B. K., Ulusoy I. S., Zhang X., Xu X., Wilson A. K., Truhlar D. G., *J. Phys. Chem. A*, **2020**, *124*(47), 9757—9770
[12] Qu X., Latino D. A., Aires-de-Sousa J., *J. Cheminform.*, **2013**, *5*(1), 34
[13] Feng C., Sharman E., Ye S., Luo Y., Jiang J., *Sci. China Chem.*, **2019**, *62*(12), 1698—1703
[14] Li W., Luan Y., Zhang Q., Aires-de-Sousa J., *Mol. Inform.*, **2023**, *42*(1), e2200193
[15] An H., Liu X., Cai W., Shao X., *J. Chem. Inf. Model*, **2024**, *64*(14), 5480—5491
[16] Liu J., He X., Xiong Y., Nie F., Zhang C., *Def. Technol.*, **2023**, *22*, 144—155
[17] Mantero A., Ishwaran H., *Stat. Anal. Data Min.*, **2021**, *14*(2), 144—167
[18] Scornet E., *J. Multivar. Anal.*, **2016**, *146*, 72—83
[19] Wesolowski B. C., *J. Educ. Meas.*, **2019**, *56*(3), 610—625
[20] Bian X., Li S., Fan M., Guo Y., Chang N., Wang J., *Anal. Methods*, **2016**, *8*(23), 4674—4679
[21] Kong D., Luan Y., Zhao X., Lu Y., Li W., Zhang Q., Pang A., *Chemometr. Intell. Lab.*, **2023**, *243*, 105021
[22] Frank E., Hall M., Trigg L., Holmes G., Witten I. H., *Bioinformatics*, **2004**, *20*(15), 2479—2481
[23] Shao X., Bian X., Liu J., Zhang M., Cai W., *Anal. Methods*, **2010**, *2*(11), 1662—1666
[24] Kaneko H., *ACS Omega*, **2023**, *8*(24), 21781—21786
[25] Li X., Luan Y., Lu Y., Li W., Ma L., Zhang Q., Pang A., *Chem. Res. Chinese Universities*, **2022**, *39*(2), 296—304
[26] Marque S., *J. Org. Chem.*, **2003**, *68*(20), 7582—7590
[27] Huang C., Zhao Y., Roy I., Cai L., Pitsch H., Leonhard K., *Combust Flame*, **2022**, *242*, 112211
[28] Gramatica P., *Qsar. Comb. Sci.*, **2007**, *26*(5), 694—701
[29] Netzeva T. I., Gallegos Saliner A., Worth A. P., *Environ. Toxicol. Chem.*, **2006**, *25*(5), 1223—1230
[30] Wang Z., Chen J., Hong H., *Chem. Res. Toxicol.*, **2020**, *33*(6), 1382—1388
[31] Roy K., Kar S., Ambure P., *Chemometr. Intell. Lab.*, **2015**, *145*, 22—29
[32] Wang Z., Chen J., Hong H., *Environ. Sci. Technol.*, **2021**, *55*(10), 6857—6866
[33] St John P. C., Guan Y., Kim Y., Kim S., Paton R. S., *Nat. Commun.*, **2020**, *11*(1), 2328
[34] Luan Y., Li X., Kong D., Li W., Li W., Zhang Q., Pang A., *J. Mol. Graph. Model*, **2024**, *129*, 108752
[35] Wu T., Chen M. Y., Xiao K. X., Zhou Y. M., Zhang Q. Y., *Chem. J. Chinese Universities*, **2019**, *40*(6), 1158—1163(吴婷, 陈梦瑶, 肖凯霞, 周艳梅, 张庆友. 高等学校化学学报, **2019**, *40*(6), 1158—1163)

(Ed.: N, K)