大 学 化 学
Univ. Chem. **2025,** *40* (9), 69

•专题• doi: 10.12461/PKU.DXHX202411067 www.dxhx.pku.edu.cn

高通量计算与机器学习相结合的综合计算化学实验设计与实践

周佳*
哈尔滨工业大学(深圳)理学院，城市水资源与水环境国家重点实验室，广东 深圳 518055

摘要：随着计算化学和人工智能技术的迅猛进步，它们在化学教育领域的应用愈发关键。本实验项目专门为化学高年级本科生及研究生搭建了一个综合性的实验平台，将计算化学与机器学习方法加以融合，对有机化合物的键解离能展开深入探究。课程内容包含量子化学计算方法的基础原理、实操技巧，还有机器学习模型的构建、训练以及验证等多个方面。学生将借由实际操作，学会运用先进的计算工具和算法来预测与分析化学键能，进而增进对化学反应机理的深度认知。课程的目标在于让学生不但能够熟练掌控数据处理和分析的技能，并且能够独立凭借这些技能从事化学问题的研究，为日后的科研工作或者跨学科领域的探索筑牢根基。

关键词：键解离能；计算化学；大数据；机器学习；SMILES
中图分类号：G64；O6

Design and Practice of a Comprehensive Computational Chemistry Experiment Based on High-Throughput Computation and Machine Learning

Jia Zhou *
State Key Laboratory of Urban Water Resource and Environment, School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, Guangdong Province, China.

Abstract: With the rapid advancements in computational chemistry and artificial intelligence technologies, their integration into chemical education has become increasingly vital. This experimental course is specifically tailored for senior undergraduate and graduate chemistry students, providing a comprehensive platform that merges computational chemistry with machine learning methodologies to explore the bond dissociation energies of organic compounds in depth. The curriculum covers fundamental principles and operational techniques of quantum chemical calculation methods, as well as the construction, training, and validation of machine learning models. Through hands-on experience, students will learn to utilize advanced computational tools and algorithms to predict and analyze chemical bond energies, thereby deepening their understanding of chemical reaction mechanisms. The objective of the course is to equip students with proficient data processing and analysis skills, empowering them to independently apply these skills to research chemical problems, thus establishing a strong foundation for future scientific endeavors or interdisciplinary explorations.

Key Words: Bond dissociation energy; Computational chemistry; Big data; Machine learning; SMILES

收稿：2024-11-19；录用：2025-01-02；网络发表：2025-03-26
*通讯作者，Email: jiazhou@hit.edu.cn
基金资助：哈尔滨工业大学深圳校区质量工程项目(高等教育教学改革项目) (HITSZERP22009)；哈尔滨工业大学深圳校区思政课程和课程思政专项课题(HITSZIP22017)

在当今的数字化时代，大数据与人工智能正以雷霆万钧之势改变着我们生活的世界。大数据所拥有的海量规模以及丰富多元的类型，为人工智能提供了充足的素材。凭借对海量数据的深度剖析和挖掘，人工智能能够持续学习并不断进化。二者的紧密融合，不但有力地推动了科学技术的进步，而且在各个领域引发了意义深远的变革。在化学领域，大数据与人工智能的应用正在引发革命性的转变[1–3]。大数据的驱动使得化学研究更为高效精确，人工智能的赋能促使化学创新更加便捷迅速。

与时代同步，培育适应新时代的人才始终是我们教育的重要目标之一。通过开设相关课程，让学生触及时代发展的前沿，无疑是一种行之有效的办法[4]。在此，我们设计了一个将高通量计算与机器学习相结合的全新综合计算化学实验。实验取材于最新的科研成果[1]，但通过适当修改使其更加适合化学高年级本科生及研究生在有限学时内学习掌握。本次实验改革将有益于化学专业的学生学习并掌握最新的机器学习方法，为日后的学习和研究筑牢根基。通过这些方法，学生们能够进行更高效的数据分析，拓展科学研究的边界，同时为未来可能的跨学科工作做好铺垫。

1 实验目的

(1) 了解分子信息学，掌握分子SMILES (Simplified molecular input line entry system，简化分子线性输入规)编码。

(2) 掌握ORCA计算软件[5]的使用方法。

(3) 了解机器学习方法，掌握Mathematica程序[6]的使用方法。

2 实验原理

键解离能(BDE)是评估分子稳定性和反应性的关键指标，通过对各种分子结构的BDE进行计算，学生能够深刻领会分子中化学键的本质以及其对分子性质的影响。在本次实验中，我们应用ORCA软件计算有机分子及其解离碎片的焓值，随后依据焓值来计算键解离能BDE。ORCA是一款由Frank Neese研究小组开发的计算化学软件，为学术机构提供免费版本以供学术使用。该软件涵盖了一系列的电子结构计算方法，其中包含密度泛函理论(DFT)、从头算(ab initio)方法，以及相关的后Hartree-Fock方法，例如MP2和耦合簇理论。

此外，Open Babel作为一款卓越的开源化学信息学工具[7]，在化学研究领域起着举足轻重的作用。它拥有强大的格式转换能力，无论是常见的分子结构描述格式，还是繁杂的化学文件类型，都能够实现便捷且精准的转换，极大程度地促进了不同化学软件之间的数据交流与共享。不仅如此，Open Babel还能够对分子施行一系列精细的操作，例如加氢、去氢等，并且可以精确计算分子的多种特性，诸如分子量、偶极矩等。它还在相似性搜索领域表现出众，能够切实有效地帮助研究人员筛选和对比分子。就连对于化学反应的处理，Open Babel也具备独特的方式和能力。总而言之，Open Babel凭借其丰富多样的功能和高效卓越的性能，为化学工作者给予了强有力的支撑，我们在本实验中应用Open Babel进行大规模化学数据格式转化，批量生成计算输入文件，再进行高通量计算。

Mathematica是Wolfram Research公司开发的一款集成了多种功能的科学计算软件，它不仅包含了强大的数值和符号计算引擎，还拥有高级的图形系统、编程语言、文本处理能力以及与其他应用程序的连接功能。Mathematica的这些特性使其在科学、工程、数学和计算领域都有着广泛的应用。在本教学实验中，Mathematica的机器学习框架被用于分析ORCA计算得到的键解离能数据，并构建模型以预测未知有机分子的BDE。学生将学习如何利用包括线性回归、支持向量机、决策树、随机森林、逻辑回归和朴素贝叶斯等机器学习算法，以及如何选择合适的算法和参数以最优化模型性能。

3 实验设备

KingDraw软件、Open Babel软件、ORCA软件、Mathematica软件、台式电脑。

4 实验步骤

具体的实验流程如图1所示。

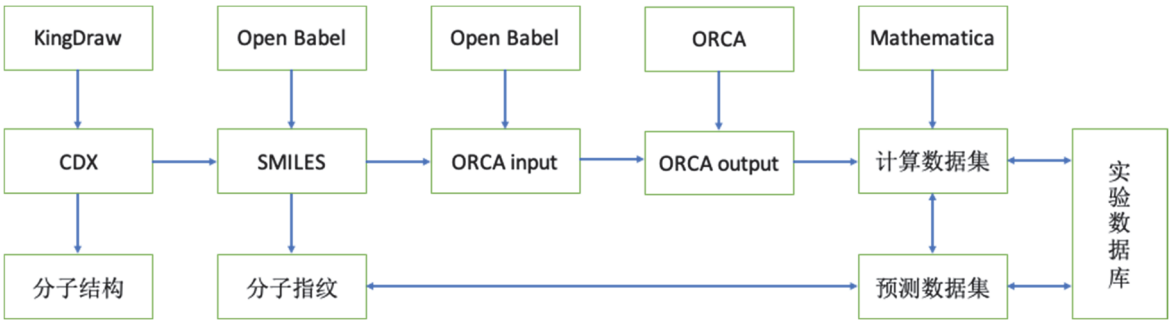


图1　实验流程图

(1) 应用KingDraw软件构建有机分子。

在KingDraw程序中画出大量(建议数量大于20)有机分子，最后保存成cdx文件。图2作为示意展示了一组从乙烷到正己烷的有机分子。

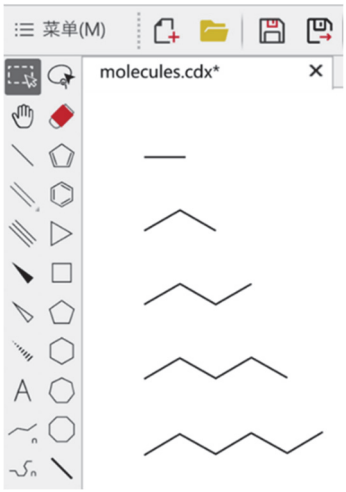


图2　应用KingDraw软件构建一系列有机分子

(2) 应用Open Babel软件生成SMILES编码以及ORCA计算的输入文件。

应用Open Babel软件，通过下述命令行将前面得到的cdx文件转为SMILES编码。

obabel molecules.cdx -O molecules.smi -m

同样应用Open Babel软件，通过下述命令行将SMILES文件转为一系列ORCA输入文件。

obabel molecules.smi -oorcainp -O molecules.inp -m --gen3D --minimize --ff uff

特别注意，可以通过文本编辑器自行修改ORCA输入文件，设置相应的计算方法。

注明：obabel为程序名。-O为输出参数，后接具体输出文件名，如molecules.smi。-m表示输出多个文件。-oorcainp表示输出文件的格式为ORCA的输入文件。--gen3D表示生成三维构型的分子坐标。--minimize表示进行能量最小化。--ff uff表示应用UFF分子力场进行能量最小化。

(3) 计算有机分子碳氢键和碳碳键均裂的键解离能。

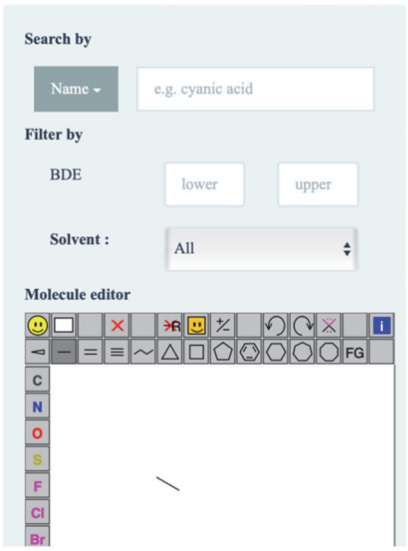
在命令行模式下输入：

<full orca binary folder path>/orca example.inp > example.out

进行ORCA计算。特别注意需要输入ORCA的绝对路径才能正常执行计算程序。

通过计算有机分子(例如C_2H_6)及其解离片段的焓值，计算键解离能。可以尝试不同的电子结构计算方法计算键解离能。

在iBond数据库(ibond.nankai.edu.cn)检索键解离能的实验值，搜索界面如图3所示。



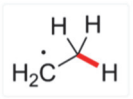
Structure	Solvent	BDE	Method	Ref.
CD_3CD_2-D	Gas	102.8±0.4	VLPP	89P
CH_3CH_2-H	Gas	100.5±0.5	FS	86B
	Gas	100.8±0.7	PIMS	88R
	Gas	101±0.4	PIMS	92S2
	Gas	100.5±0.3☆	VLPP	97D
	Gas	36.2±0.4☆	Der.	86P
	Gas	36.1±2.9	Der.	06G2
H_3C-CH_3	Gas	90.2±0.2	Der.	86P

图3　在iBond网站搜索乙烷的键解离能

(4) 应用机器学习方法预测键解离能。

图4是为本实验预先准备的有机分子键解离能数据集片段，第一列为有机分子的SMILES编码，第二和三列为相应化学键解离后片段的SMILES编码，第四列为计算得到的键解离能。

molecule	fragment1	fragment2	bde
C#C/C(C)=C/CNCC	[C]#C	C[C]=CCNCC	132.9189564
C#C/C(C)=C/CNCC	C#C/[C]=C/CNCC	[CH3]	90.54515614
C#C/C(C)=C/CNCC	[CH2]NCC	[CH]=C(C)C#C	92.5851879
C#C/C(C)=C/CNCC	C#C/C(C)=C/CN[CH2]	[CH3]	82.47350787
C#C/C=C/C(=O)O	[C]#C	[CH]=CC(=O)O	138.1536365
C#C/C=C/C(=O)O	O=[C]O	[CH]=CC#C	107.7853384
C#C/C=C/C=C/C=C/C	[C]#C	[CH]=C/C=C/C=C/C	139.4061444
C#C/C=C/C=C/C=C/C	[CH]=C/C=C/C	[CH]=CC#C	120.2106441
C#C/C=C/C=C/C=C/C	[CH]=C/C=C/C#C	[CH]=CC	118.0193827
C#C/C=C/C=C/C=C/C	[CH3]	[CH]=C/C=C/C=C/C#C	102.8261348
C#C/C=C/COCCCC	[C]#C	[CH]=CCOCCCC	138.8551915
C#C/C=C/COCCCC	[CH2]OCCCC	[CH]=CC#C	99.31020185
C#C/C=C/COCCCC	C#C/C=C/COC[CH2]	[CH2]C	88.55909015
C#C/C=C/COCCCC	C#C/C=C/COCC[CH2]	[CH3]	88.98893382
C#CC#CCCO	[C]#C	[C]#CCCO	164.3201342
C#CC#CCCO	C#CC#C[CH2]	[CH2]O	74.0529646
C#CC(=O)/C=C/C	C/C=C/[C]=O	[C]#C	113.1681106
C#CC(=O)/C=C/C	[CH3]	[CH]=CC(=O)C#C	102.7608738
C#CC(=O)NCCCCN	[C]#C	NCCCCN[C]=O	118.3707877
C#CC(=O)NCCCCN	C#CC(=O)N[CH2]	[CH2]CCN	84.16338961

图4　有机分子键解离能数据集片段

Mathematica可以识别分子的SMILES编码，并可以显示其三维结构，如图5所示，用来验证编码准确性。

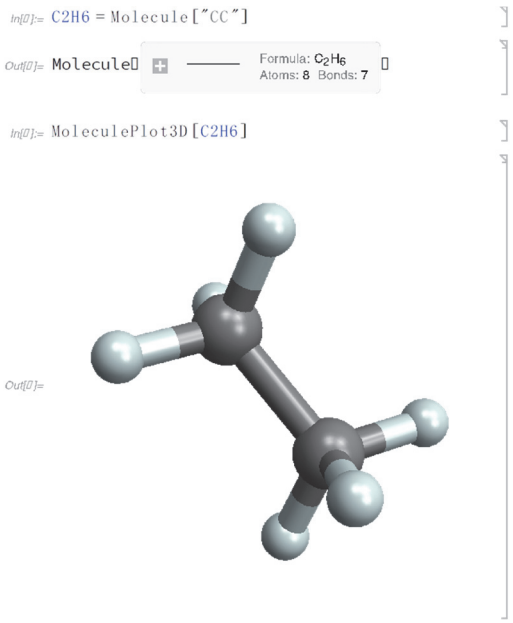


图5　Mathematica读取SMILES编码并显示分子三维结构

应用Mathematica读取上述键解离能的数据，应用Predict命令对数据进行机器学习，尝试不同机器学习算法。

```
In[]:= pGBdataset = Predict[BDInfo ▯ BDE, Method ▯ "GradientBoostedTrees"]
Out[]= PredictorFunction▯  Input type: {Text, Text, Text}
                           Method: GradientBoostedTrees
```

5　数据处理

(1) 优化乙烷分子C_2H_6及其解离片段结构，得到各自的焓。

$C_2H_6 \rightarrow 2CH_3$

$C_2H_6 \rightarrow C_2H_5 + H$

ORCA计算输出文件中的焓值信息如图6所示。

```
--------
ENTHALPY
--------

The enthalpy is H = U + kB*T
               kB is Boltzmann's constant
Total free energy                  ...    -79.62902567 Eh
Thermal Enthalpy correction        ...      0.00094421 Eh       0.59 kcal/mol
-------------------------------------------------------------------------------
Total Enthalpy                     ...    -79.62808146 Eh
```

图6　ORCA计算输出文件中的焓值部分

(2) 应用不同量子化学计算方法计算键解离能，例如HF/def2-SVP和B3LYP/def2-SVP。表1为相应计算所得的乙烷分子中碳氢键和碳碳键均裂的键解离能。将计算所得数据和实验值(见图3：BDE(C－C)；90.2 kcal·mol^{-1}；BDE(C－H)：100.5 kcal·mol^{-1}) (1 kcal·mol^{-1} = 4.19 kJ·mol^{-1})进行比较。

(3) 应用有机分子键解离能数据集，通过不同机器学习方法预测碳氢键和碳碳键均裂的键解离能。表2为不同机器学习方法所预测的乙烷分子中碳氢键和碳碳键均裂的键解离能。将机器学习预测所得数据和实验值进行比较。

表1　不同计算方法得到的键解离能及与实验值的相对偏差

in kcal·mol^{-1}	BDE(C－C)	相对偏差	BDE(C－H)	相对偏差
HF def2-SVP	60.0	33%	74.5	26%
B3LYP def2-SVP	88.2	2%	98.3	2%
……				

表2　应用不同机器学习方法得到的键解离能及与实验值的相对偏差

in kcal·mol^{-1}	BDE(C－C)	相对偏差	BDE(C－H)	相对偏差
Random Forest (随机森林)	93.4	4%	102.7	2%
Gradient Boosted Trees (梯度提升树)	93.0	3%	104.2	4%
……				

6　思考题

(1) 有机分子碳氢键和碳碳键均裂的键解离能和哪些因素有关?

(2) 计算其他有机分子中的碳氢键和碳碳键均裂的键解离能，并和实验值比较。

(3) 尝试其他机器学习算法预测键解离能，分析影响预测精度的因素？

7　实验教学组织与实施

(1) 教学设计和实施。

本实验是面向已学习过物理化学课程中量子化学相关知识的化学相关专业的高年级本科生开设的综合性实验。实验前会要求学生进行文献调研了解研究背景、有机分子键解离能的测定和实际用途、机器学习原理等，并提出以下思考题：

① 有机分子键解离能是如何实验测量的？

② 常见的机器学习方法有哪些，都有哪些特点？

③ 化学相关的数据库有哪些？国内的化学数据库有哪些？

通过这些问题让学生提前了解本实验内涵，更容易在课堂上完成实验计算操作。

实验共计用时4学时，具体为：简要介绍量子化学计算分子键能的相关知识，0.5学时；介绍ORCA、Avogadro、Mathematica、如何编制输入文件以及如何查看输出文件，1学时；上机计算并完成数据的计算与整理，2.5学时。课后学生独立完成实验报告和思考题。

(2) 考核方式。

本实验采取三段式评价方式，即实验前预习(20%)、上机实验(40%)和实验报告(40%)，见表3。

表3　实验各环节考核占比及标准

实验环节	考核标准	占比/%
实验前预习	预习报告	20
上机实验	计算程序操作	40
实验后总结	实验报告，数据处理和分析	40

(3) 教学实践情况。

本实验的难点在于为化学专业的学生讲解机器学习的相关知识，并让他们学会初步运用机器学习模型来解决化学相关问题。在前期的教学中，我们已经对Mathematica的基本运行模式进行过讲解，学生对Mathematica已不陌生。在此，教学的重点在于机器学习模型的选取以及相关参数的确定。教

师会先讲解实际案例，然后让学生在课堂上自行探索，同时辅以教师的现场指导。实践表明，这样的教学方式基本上能够解决绝大部分学生所遇到的问题。最后，通过检查学生的实验报告，能够对课程的学习情况获得清晰的反馈。实践表明，本次实验可以显著提高学生运用高通量计算获取大数据并进行处理的能力。部分学生反映，通过大数据结合人工智能的方式来解决化学问题，与以往的做实验或者进行计算模拟完全不同，体现了最新的科技发展趋向，他们对化学的未来也充满了期望。

(4) 思政教育。

在教学过程中，笔者尤为注重融入思政元素，以启迪学生的发散性思维，激发其学习兴趣，提升学习效率。在本次实验里，笔者将会阐述大数据以及人工智能不但正在改变我们的日常生活模式，而且也在变革我们的学习和科研方式。当下，我国正在大力推动与化学相关的大数据和人工智能的发展，已经获取了众多重大的进展和突破，例如中国科学技术大学研制出了初步实现智能化学范式的机器化学家。与此同时，全球范围内都在全力推进人工智能的发展，国家之间的竞争愈发激烈。我国若要在人工智能领域发挥引领作用，需要我们广大科技工作者贡献各自的聪明才智。化学作为传统学科，同样能够与人工智能相结合，从而焕发出崭新的活力。此外，国内工具软件的发展近年也取得了显著进步，本次实验中我们选用了国产的KingDraw软件，而非国外著名的ChemDraw。KingDraw软件的易用性非常好，我们也借此机会鼓励学生们多多使用国产软件。

8 结语

在计算化学实验教学过程中，笔者始终坚持将最新的学科进展融入实际教学当中，让学生能够接触到最新的学科发展动向，而大数据和人工智能正是当前最为热门的研究方向。笔者在传统的计算化学实验里增添了数据库和人工智能的相关内容，以使学生能够学会如何通过数据库查找相关数据，掌握运用计算化学和机器学习解决化学问题的基本原理与流程，了解人工智能的优势和应用前景。经由本实验，学生将获得契机，将计算化学、数据科学和人工智能的理论知识相结合，并将其应用于实际的有机化学键能研究中。本实验既拓展了学生的知识面，也为他们未来在化学等相关领域的研究和工作奠定了坚实的基础。

参 考 文 献

[1] St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. *Nat. Commun.* **2020,** *11* (1), 2328.

[2] Feng, C.; Sharman, E.; Ye, S.; Luo, Y.; Jiang, J. *Sci. China Chem.* **2019,** *62* (12), 1698.

[3] 张思源, 张志成, 李荣金. 大学化学, **2024,** *40* (1), 206.

[4] 周佳. 大学化学, **2024,** *39* (3), 351.

[5] Neese, F. *WIRES-Comp Mol Sci.* **2022,** *12* (5), e1606.

[6] Mathematica 12.0 ed.; Wolfram Research, Inc.: Champaign, IL, USA, 2019.

[7] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011,** *3* (1), 33.