大 学 化 学
Univ. Chem. **2024,** *39* (3), 351

•化学实验•　　doi: 10.3866/PKU.DXHX202309060　　www.dxhx.pku.edu.cn

应用量子化学和机器学习构建水分子势能面
——介绍一个综合计算化学实验

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：当前机器学习在化学及其相关学科研究中的地位日渐提高，但是在本科化学实验中还未见涉及，现有的计算化学实验多侧重应用量子化学方法研究分子性质和化学反应。为了普及机器学习这一强大工具，本文将介绍一个综合计算化学实验，适用于高年级本科生或研究生。通过本实验，学生可以初步掌握机器学习的基本原理和操作流程，同时培养学生应用量子化学和机器学习解决化学问题的能力。

关键词：势能面；量子化学；机器学习；计算化学实验
中图分类号：G64；O6

Constructing Potential Energy Surface of Water Molecule by Quantum Chemistry and Machine Learning: Introduction to a Comprehensive Computational Chemistry Experiment

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

Abstract: The status of machine learning in chemistry and related disciplines is becoming increasingly prominent, but its application in undergraduate chemistry experiments has yet to be seen. Existing computational chemistry experiments mostly focus on the application of quantum chemistry methods to study molecular properties and chemical reactions. In order to popularize machine learning as a powerful tool, this article will introduce a comprehensive computational chemistry experiment suitable for senior undergraduate students or graduate students. Through this experiment, students can gain a preliminary understanding of the basic principles and operational procedures of machine learning, while also develop their ability to apply quantum chemistry and machine learning to solve chemical problems.

Key Words: Potential energy surface; Quantum chemistry; Machine learning; Computational chemistry experiment

在绝热近似下，势能面是化学体系在电子状态量子数守恒的情况下，粒子间相互作用势能随着原子间距的改变而改变所形成的空间曲面[1]。图1是一个简化了的用来描述化学反应的势能面，呈现为一个丘陵景观，包含山谷、山峰和山口。尽管大多数分子具有多个几何变量，但是这样的图形能够展示势能面大部分的关键特征。势能面的山谷代表了反应物、中间体和生成物。势能面中山谷的

收稿：2023-09-18；录用：2023-11-14；网络发表：2023-12-01
*通讯作者，Email: jiazhou@hit.edu.cn
基金资助：深圳市教育学会“十四五”规划2021年度教育科研一般课题(YB2021004)；哈尔滨工业大学(深圳)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)思政课程和课程思政专项课题(HITSZIP22017)

最小值代表平衡结构的位置。生成物谷和反应物谷之间的能量差则反映了反应的能量变化。分子在反应物和生成物平衡结构上振动运动可以用来计算零点能量，并进行计算焓差和自由能差所需的热修正。连接反应物谷和生成物谷的能量最低路径为反应路径。过渡态被定义为最低能量反应路径上的能量最高点，过渡态与反应物之间的能量差即为反应的能垒。过渡态在反应路径方向上具有最大能量，表示沿着连接反应物和生成物的方向上能量最高，而在与反应路径垂直的其他方向上，过渡态具有最小能量，因此过渡态也可以称为一阶鞍点。在图1中，可以将过渡态可视化为连接两个山谷的山口。

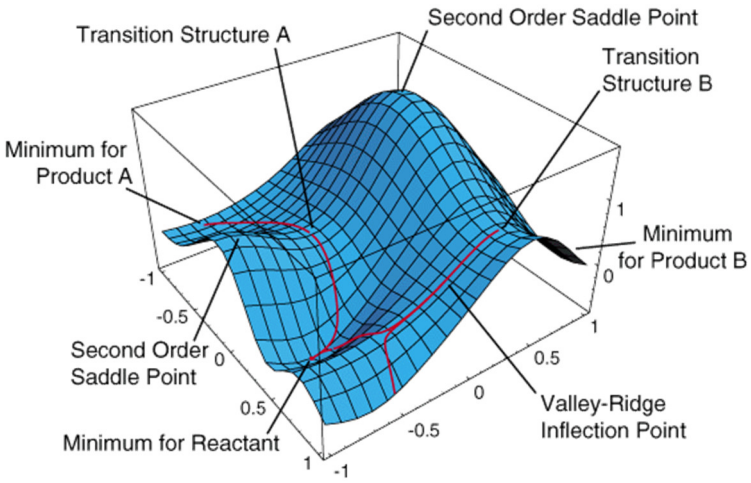

图1　势能面的示意图[2]

势能面在量子化学、分子光谱和分子力学模拟等领域中都具有重要地位。构建高质量的势能面是计算化学的主要目标和分子动力学模拟的起点，在化学研究中具有至关重要的作用。为此，厦门大学袁汝明等设计了一个面向高年级本科生的关于H_3反应势能面的计算化学实验[3]。类似地，国内其他大学也都有开设计算化学相关的实验[4–7]。对于由N个非线性原子组成的分子体系，势能面是一个在$3N-5$维空间中的超曲面，也就是说，它是由$3N-6$个坐标变量来描述的能量函数。通常，构建势能面的方法是通过求解Schrödinger方程来获得不同核构型下的电子态能量，这些能量值相当于势能面上的离散点。通常需要计算几千甚至几万个离散点。然后，通过采用适当的解析函数，该函数包含一些参数，并利用最小二乘法对计算得到的离散点进行拟合。通过确定这些参数，可以得到势能面的解析表达式。势能面的质量受多个因素影响，包括单点计算精度、单点布局和解析函数的适用性。特别是对于自由度较高的系统而言，要得到完整的势能面在当前的计算资源条件下几乎是不可能的任务[8]。近20年来，借助大数据拟合和精确的电子结构计算结果，机器学习技术已被证明可以用于构建高度准确的势能面[9–13]。机器学习是指从有限的观测数据中学习出具有一般性的规律，并将这些规律应用到未观测数据样本上的方法。机器学习任务本质上是根据特征判断性质，根据经验解决问题。基本流程是基于数据产生模型，利用模型预测输出，目标是让模型有较好的泛化能力。利用机器学习构建势能面当前正成为一种发展趋势。在这里，我们介绍一个结合量子计算和机器学习的计算化学综合实验。

1　实验目的

(1) 应用量子化学计算方法构建水分子的势能面，掌握势能面的含义，了解其应用。
(2) 学习和掌握Gaussian计算软件及其配套的可视化软件GaussView的基本使用方法。
(3) 学习使用机器学习方法构建水分子势能面，掌握Mathematica的使用方法。

2 实验原理

Gaussian软件[14]是一款广泛应用于量子化学计算的软件，目前最新版本为Gaussian 16。它基于量子力学的基本定律，能够预测在不同化学环境中化合物和反应的能量、分子结构、振动频率等性质。Gaussian 16包括多种计算方法，以满足不同类型化学计算的需求，其中一些常见的方法包括半经验方法(AM1、PM6等)，Hartree-Fock方法，密度泛函方法(B3LYP等)，后自洽场方法(MP2、CCSD、QCISD等)。需要注意的是，各种计算方法都基于一定的近似假设和简化，因此所描述的势能面会存在差异。GaussView是与Gaussian软件配套使用的图形界面软件。它提供了许多方便的功能，使用户可以更轻松地创建Gaussian的输入文件，并通过图形界面直观地运行Gaussian计算，而无需使用命令行指令。Mathematica [15]是一款集成了数值和符号计算引擎、图形系统、编程语言、文本系统以及与其他应用程序的高级连接功能强大的科学计算软件，它在许多领域内具有世界领先的地位。同时，它也被广泛应用于化学教学中[16]。Mathematica现已集成了机器学习框架，可以进行深度学习、神经网络、分类和聚类等任务。常见的机器学习算法包括线性回归、支持向量机、最近邻居、逻辑回归、决策树、随机森林、朴素贝叶斯等。其中，高斯过程回归是概率论和数理统计中随机过程的一种，是多元高斯分布的扩展，被应用于机器学习。神经网络又称人工神经网络，是机器学习的子集，同时也是深度学习算法的核心。其名称和结构均受到人脑的启发，可模仿生物神经元相互传递信号的方式。本实验中应用计算化学程序研究水分子势能面，之后应用机器学习算法，拟合势能面。

3 实验设备

Gaussian 16计算软件、GaussView 6软件、Mathematica 12软件、台式电脑(i7-9700，32G)。

4 实验步骤

(1) 应用GaussView构建水分子结构(图2)，键角设定为104.5° (此为实验值)。之后，创建Gaussian输入文件(表1)，扫描不同键长下水分子的能量。此处，两个O－H键保持一致。

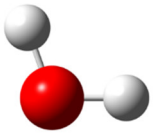

图2　水分子在GaussView中的结构显示

表1　应用Gaussian 16程序计算，输入文件和说明

输入文件	各部分的说明
# HF/6-31G(d,p) scan	任务类型，包括计算方法和基组
	空白行
water scan	计算说明
	空白行
0 1	电荷、自旋多重度
O	水分子内坐标
H　1　B1	
H　1　B1　2　A1	
	空白行
B1　0.5　31　0.05	键参数、初始键长、扫描步数、扫描步长
A1　104.5	键参数

(2) 放开两个O－H键，扫描不同O－H键长下水分子的能量，扫描变量分别为B1和B2 (表2)。

表2 Gaussian输入文件和说明

输入文件	各部分的说明
# HF/6-31G(d,p) scan	任务类型，包括计算方法和基组
	空白行
water double scan	计算说明
	空白行
0 1	电荷、自旋多重度
O	水分子内坐标
H 1 B1	
H 1 B2 2 A1	
	空白行
B1 0.5 31 0.05	键参数、初始键长、扫描步数、扫描步长
B2 0.5 31 0.05	键参数、初始键长、扫描步数、扫描步长
A1 104.5	键参数

(3) 利用Mathematica读取Gaussian计算数据，建立键长和分子能量的关系，应用机器学习方法，构建势能面。Mathematica读取单变量扫描的命令如下(Gaussian计算的输出文件为water_scan.log)。这里的目的是把扫描的水分子键长存在CoordScan表中，而把对应的水分子能量存储在EnergyScan表中，便于后续处理。

```
In[•]:= LogFile = "water_scan.log";

In[•]:= intxt = OpenRead[LogFile];
        Find[intxt, "Summary of the potential surface scan:"];
        Skip[intxt, Record, 2];
        Clear[info];
        info = Read[intxt, Table[{Number, Number, Number}, {i, 32}]];
        CoordScan = Table[info[[i, 2]], {i, 32}];
        EnergyScan = Table[info[[i, 3]], {i, 32}];
        dataset = MatrixForm[Table[{CoordScan[[i]], EnergyScan[[i]]}, {i, 32}]];
```

应用Mathematica自带的不同机器学习方法(高斯过程回归GaussianProcess、神经网络NeuralNetwork)训练数据的命令如下所示，目的是建立CoordScan到EnergyScan之间的关系。

```
In[•]:= pGPdataset = Predict[CoordScan → EnergyScan, Method → "GaussianProcess"]

Out[•]= PredictorFunction[ Input type: Numerical
                           Method: GaussianProcess ]

In[•]:= pNEdataset = Predict[CoordScan → EnergyScan, Method → "NeuralNetwork"]

Out[•]= PredictorFunction[ Input type: Numerical
                           Method: NeuralNetwork ]
```

在Mathematica中，有多个机器学习相关函数，包括Predict和Classify等，它们可以完全自动化的方式执行分类和预测任务，自动为给定的特定输入选择最佳方法。本课程主要使用的是Predict函数，它可以基于样本集训练预测器函数(如上面所示)，之后通过给定特征，预测新样本的值。此外，Predict

函数可通过更改Method后面的参数选择使用何种回归算法，诸如神经网络、随机森林、线性回归等等。

5 数据处理

(1) Gaussian计算所得的输出文件中会有如下信息，表示不同O—H键长(B1)下计算所得的水分子能量。

Summary of the potential surface scan:

N	B1	SCF
1	0.5000	−74.46288
2	0.5500	−74.98878
3	0.6000	−75.35297
4	0.6500	−75.60367
5	0.7000	−75.77384
6	0.7500	−75.88638
7	0.8000	−75.95752
8	0.8500	−75.99886
9	0.9000	−76.01880
10	0.9500	−76.02347
11	1.0000	−76.01731

用键长和能量作图，得到不同O—H键长下水分子的能量，如图3a所示。

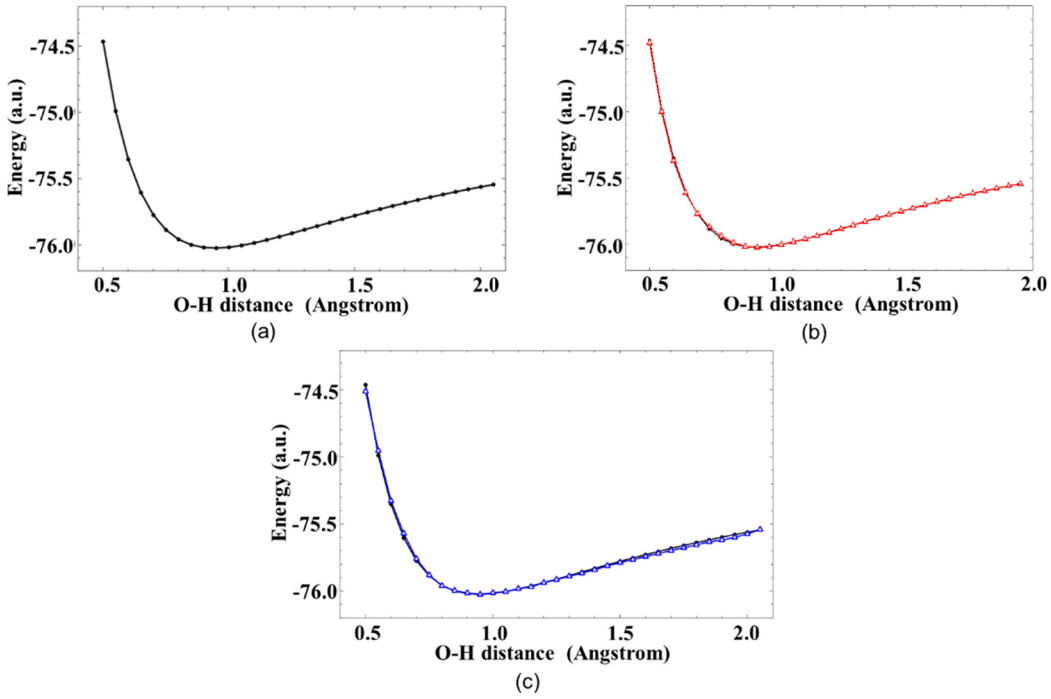

图3 (a) 水分子的计算势能曲线；(b) 水分子的机器学习势能曲线(高斯过程回归，红色)和计算势能曲线对比；(c) 水分子的机器学习势能曲线(神经网络，蓝色)和计算势能曲线对比

电子版为彩图，下同

量化计算的结果表明O—H键长为0.95 Å时分子的能量最低。

(2) 应用Mathematica将机器学习所得的势能面和量化计算的势能面做对比。Mathematica的命令如下所示：

```
In[0]:= Show[ListPlot[{Table[{CoordScan [[]], EnergyScan [[]]}, {i, 32}],
    Table[{CoordScan [[]], pGPdataset[CoordScan [[]]]}, {i, 32}]},
  PlotRange →{{0.4, 2.1}, {-76.2, -74.2}}, Axes →{False, True},
  AxesStyle →{None, {Black, Dashed}}, ImageSize →600, Frame →True,
  FrameTicks →True, Joined →True,
  FrameLabel →{"O-H distance (Angstrom)", "Energy (a.u.)"},
  LabelStyle →Directive[Bold, Italic, 15], PlotStyle →{Black, Red},
  PlotMarkers →{Automatic, "OpenMarkers"}]]
```

其中，pGPdataset就是前面训练得到的器函数。图3b是机器学习方法神经网络所得的势能曲线和量化计算所得的对比，红色为机器学习(高斯过程)所得的势能曲线，与量化计算值(黑色)非常接近。图3c是另外一个机器学习方法神经网络所得的势能曲线和量化计算所得的对比，结果同样表明机器学习的结果也和量化计算的结果非常吻合。

我们还可以应用机器学习所得的器函数，计算未知的键长下的能量，之后与量化计算做对比。Mathematica的命令为：pGPdataset[B1]。这里尝试B1 = 0.98 Å。机器学习预测值为76.025，而量化计算结果为−76.021 au，结果比较接近。

(3) 前面所计算的势能曲线可以用来描述水分子中两个O−H键对称伸缩振动的能量变化。当需要研究两个O−H键不对称伸缩振动的情况时，就要计算两个O−H键不同键长时水分子的能量。这样扫描所得的结果就是一个二维曲图。图4a是量化计算所得的势能曲面，图4b是机器学习(高斯过程回归)所得的势能曲面。

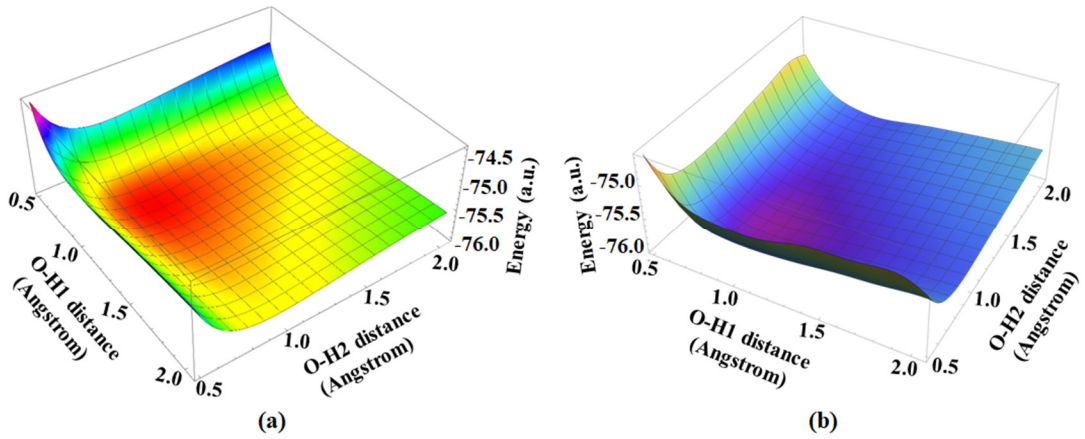

图4 (a) 水分子的计算势能曲面；(b) 水分子的机器学习势能曲面(高斯过程回归)

此外，还可以用如下的Mathematica命令计算机器学习所得的势能曲面与量化计算所得的势能曲面的偏差(图5)。

综上，我们可以发现机器学习能够根据一定数量的量化计算结果较好地预测水分子完整的势能面。

6 思考题

(1) 尝试其他量化计算方法构建水分子势能面，并以此作为机器学习的输入数据构建势能面。

提示：这里同学可以自行尝试应用DFT、MP2、CCSD等计算方法，重复前面的过程。

(2) 氢气分子的势能曲线之前常用Morse势拟合，请与机器学习所得进行对比。

提示：Morse势是一个解析的表达式，可以很容易计算不同氢原子距离下的势能。

(3) 机器学习的精度和哪些因素有关。如何提高机器学习的精度？

提示：课堂上我们只考虑键长这一描述符对能量的影响。但是并没有考虑诸如原子间距离变化中所引起的库仑相互作用的变化，如果以此库仑矩阵为描述符，机器学习的效果可以更好。

```
In[1]:= ListContourPlot[
  Table[{CoordScanX [[i]], CoordScanY [[i]],
    pGPdataset2D[{CoordScanXY [[i, 1]], CoordScanXY [[i, 2]]}] - EnergyScan2D [[i]]},
   {i, 1024}], Mesh →Automatic, ImageSize →600, PlotRange →All,
  InterpolationOrder →3, ColorFunction →"BlueGreenYellow",
  PlotLegends →Automatic]
```

Out[1]=

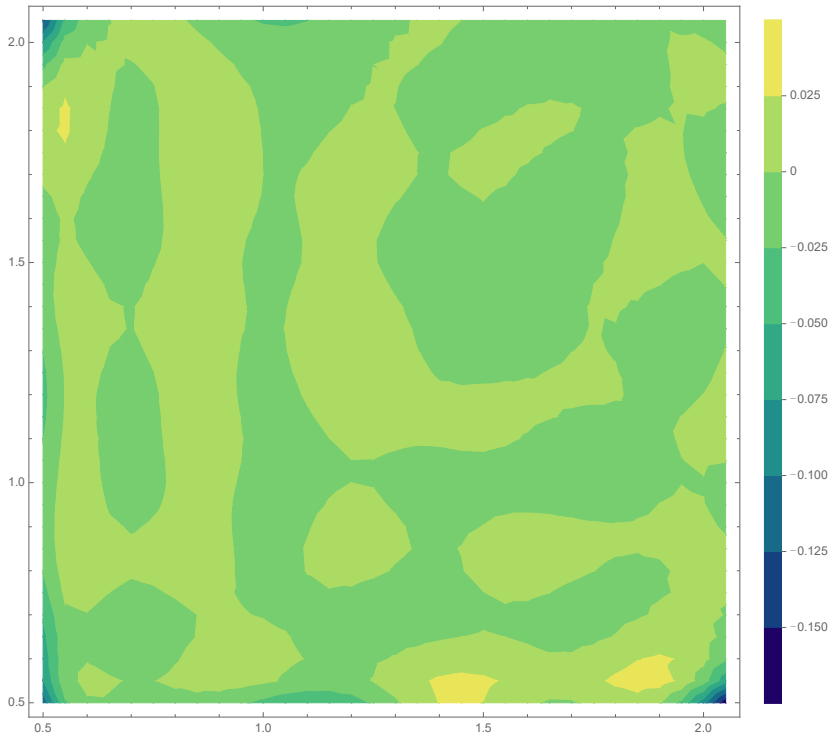

图5　水分子的机器学习势能和计算势能的偏差康托图

7　实验时间安排

这项实验旨在为已经掌握了物理化学课程中量子化学相关知识的学生设计，适用于大三下学期或大四上学期的学生。整个实验需要共计4个课时的时间进行。具体的课时安排如下：第一课时将简要介绍与量子化学计算和机器学习相关的知识；第二课时将详细介绍如何运用Gaussian 16、GaussView 6和Mathematica软件编写输入文件，并学习如何查看输出文件；第三和第四课时则主要用于实际的上机计算、数据处理和整理。在完成实验后，学生需要独立撰写实验报告并回答思考题。

8　结语

笔者在计算化学教学中，一直坚持与时俱进，努力把最新的学科进展融入到教学中，让学生接触到最新的学科发展方向，机器学习就是当前最热门的研究方向之一。我们既保证整体课程教学的连贯性，也逐年适度加入新知识，让这门课更有生命力，本次综合实验就是一个例证，在传统计算化学实验中加入了机器学习相关内容。通过本实验学生可以初步掌握结合量子化学和机器学习解决化学问题的基本原理及流程，了解机器学习的优势和应用前景，学会处理和分析计算所得数据。通过这个实验，学生将有机会结合量子化学和机器学习的理论知识，把它们应用到实际问题中，并通

过实践掌握数据处理和分析方法。这既拓宽了学生的学术视野，也为他们未来在化学领域的研究和工作提供了更广阔的可能性。

参 考 文 献

[1] 范康年, 周鸣飞. 物理化学. 第3版. 北京: 高等教育出版社, 2021: 851.

[2] Schlegel, H. *WIREs Comput. Mol. Sci*. **2011,** *1*, 790.

[3] 袁汝明, 傅钢摇, 韩国彬. 大学化学, **2011,** *26* (3), 47.

[4] 许秀芳. 大学化学, **2021,** *36* (2), 1912058.

[5] 陈国博, 李苓, 高先池. 实验技术与管理, **2021,** *38* (4), 231.

[6] 袁汝明, 傅强, 傅钢. 大学化学, **2020,** *35* (9), 1.

[7] 司聪慧, 刘金华, 卢启芳, 郭恩言. 实验室研究与探索, **2022,** *41* (2), 155.

[8] 苑世领, 张恒, 张冬菊. 分子模拟-理论与实验. 北京: 化学工业出版社, 2016: 202.

[9] Mills, A.; Goings, J.; Beck, D.; Yang, C.; Li, X. *J. Chem. Inf. Model*. **2022,** *62* (13), 3169.

[10] Eckhoff, M.; Reiher, M. *J. Chem. Theory Comput*. **2023,** *19* (12), 3509.

[11] Schmitz, G.; Godtliebsen, I.; Christiansen, O. *J. Chem. Phys*. **2019,** *150* (24), 244113.

[12] Goodlett, S.; Turney, J.; Schaefer, H. *J. Chem. Phys*. **2023,** *159* (4), 044111.

[13] Hughes, Z.; Thacker, J.; Wilson, A.; Popelier, P. *J. Chem. Theory Comput*. **2019,** *15* (1), 116.

[14] Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H.; *et al*. *Gaussian 16* Rev. C.01, Gaussian, Inc.: Wallingford, CT, USA, 2016.

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

[16] 周佳, 魏梦娇. 大学化学, **2022,** *37* (6), 2201033.