大 学 化 学
Univ. Chem. **2025,** *40* (3), 171

•专题• doi: 10.12461/PKU.DXHX202406004 www.dxhx.pku.edu.cn

与机器学习相结合的计算材料学实验设计

周佳*，钟华英
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：设计了一个面向高年级本科生或者研究生的综合材料计算模拟实验。通过材料计算模拟和机器学习手段研究二维材料的禁带宽度。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，同时培养学生应用第一性原理和机器学习解决材料问题的能力。

关键词：二维材料、材料模拟、第一性原理、机器学习、禁带宽度
中图分类号：G64；O6

Experimental Design of Computational Materials Science Combined with Machine Learning

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

Abstract: This paper presents a comprehensive computational materials science experiment designed for senior undergraduate and graduate students. The band gaps of two-dimensional materials are investigated using materials simulation and machine learning techniques. Through this experiment, students will gain a foundational understanding of machine learning principles and workflows, while also developing their ability to apply first-principles calculations and machine learning to solve materials-related problems.

Key Words: Two-dimensional materials; Materials simulation; First-principles; Machine learning; Band gaps

二维材料是指电子仅可在两个维度的非纳米尺度(1–100 nm)上自由运动(平面运动)的材料。这些材料因为具备优异的电子迁移率、电导性和热导率等特性而备受关注[1–5]。对于二维材料来说，研究其带隙的大小及其影响因素是非常重要的[6]。目前实验测定二维材料的带隙仍然具有一定的难度，需要克服制备和表征方面的困难。因此，二维材料的带隙确定主要依赖于理论计算。其中，密度泛函理论是一种基于量子力学基本原理的计算方法，它可以直接计算材料的电子结构等物理化学性质，而无需引入任何经验参数[7–9]。

随着人工智能技术的不断发展，机器学习的能力不断提升，应用范围不断扩大。机器学习是人工智能的一个分支，它使计算机系统能够从数据中学习并改进性能。在机器学习中，模型是对现实世界现象的简化和抽象。模型的构建通常涉及选择一个算法和设置相应的参数。常见的机器学习算

收稿：2024-06-03；录用：2024-10-10；网络发表：2025-02-12
*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)

法包括线性回归、支持向量机、决策树、神经网络等。机器学习模型的训练过程包括使用训练数据集来调整模型的参数，以便模型能够准确地预测或分类数据。在训练完成后，通常会使用独立的测试数据集来评估模型的性能。机器学习在多个领域得到了广泛应用，如图像识别、自然语言处理等。机器学习也为二维材料的研究提供了新的契机。由于二维材料独特的电子性质，预测其能带结构对于深入了解材料和开发新的应用具有极其重要的意义。通过机器学习手段利用大量的实验和计算数据，可以自动挖掘其中隐含的模式和规律，进而精准预测新材料的能带结构。借助于机器学习技术，可以更加高效地发现和设计具有特定能带结构的二维材料，进一步推动材料科学和工程领域的进步。此外，机器学习还可以与密度泛函理论等计算方法进行结合，有效提高计算效率和准确性，更好地模拟和预测二维材料的能带结构[10,11]。因此，机器学习在二维材料能带结构上的应用不仅具有重要的实际应用价值，还展现了其独特的科学价值[12]。

当前，教学实践也开始尝试引入机器学习，取得了一定的教学效果[13–16]。本实验是化学创新研修实验课的一节，结合了当前热门的二维材料和机器学习，旨在引导学生应用最新的研究方法解决材料化学中的基本问题，为其今后开展相关科研打下坚实的基础。我们首先利用高度集成的二维材料数据库获取了包括石墨烯、氮化硼、二硫化钼及过渡金属二卤化物等在内的多种材料的详细晶胞结构信息。随后，运用材料计算模拟软件对所获取的二维材料的带隙进行了计算，从而更好地理解这些材料的电子性质。最后，将机器学习引入到实验中，通过使用大量计算数据进行训练，机器学习能够在这些数据中学习和提取隐藏的模式与规律，进而对二维材料的带隙进行分析、预测。

1 实验目的

(1) 通过实验，加深对计算材料学方法及原理的理解。
(2) 掌握Materials Studio建模的方法，以及应用CASTEP计算二维材料禁带宽度的方法。
(3) 掌握机器学习原理，以及Jupyter程序的使用方法。
(4) 掌握应用机器学习方法研究二维材料禁带宽度的方法。

2 实验原理

Materials Studio是一款广泛应用于材料科学、化学等领域的计算模拟软件，它提供了多种模块来进行各种类型的材料性质模拟[17]。其中，CASTEP模块是一种先进的量子力学计算程序，可以用于探索固态晶体和表面的特性，包括二维材料[18]。该模块采用了密度泛函理论方法，对电子的波函数进行求解，从而得到材料的能带结构和其他电子性质。在机器学习领域，Jupyter特别有用，因为它允许数据科学家和机器学习工程师交互式探索数据，进行探索性数据分析，并实时查看结果。Jupyter作为一种开发环境与Python结合紧密，可方便下载如Matplotlib、Seaborn和Plotly的Python包，方便创建图表和可视化。此外，也方便安装许多机器学习相关的Python包，诸如Scikit-learn、TensorFlow、PyTorch、Pandas和NumPy，进行大量的维度数组和矩阵运算。本实验通过使用Materials Studio中的CASTEP模块来计算二维材料的几何结构和能带结构，之后应用Jupyter中集成的机器学习算法，对二维材料的禁带宽度进行分析。

3 实验设备

Materials Studio软件，CASTEP，Jupyter，台式电脑(i7-9700，32G)。

4 实验步骤

(1) 利用网上材料数据库(Materials Project、Materials Cloud等)搜索材料结构，从数据库中再导出二维材料的结构文件；以Materials Cloud搜索氮化硼为例，在搜索框输入“BN”，单击搜索框下方出现的BN即可跳转BN结构的界面，单击“下载”图标，选择cif格式，即可下载BN.cif结构文件。

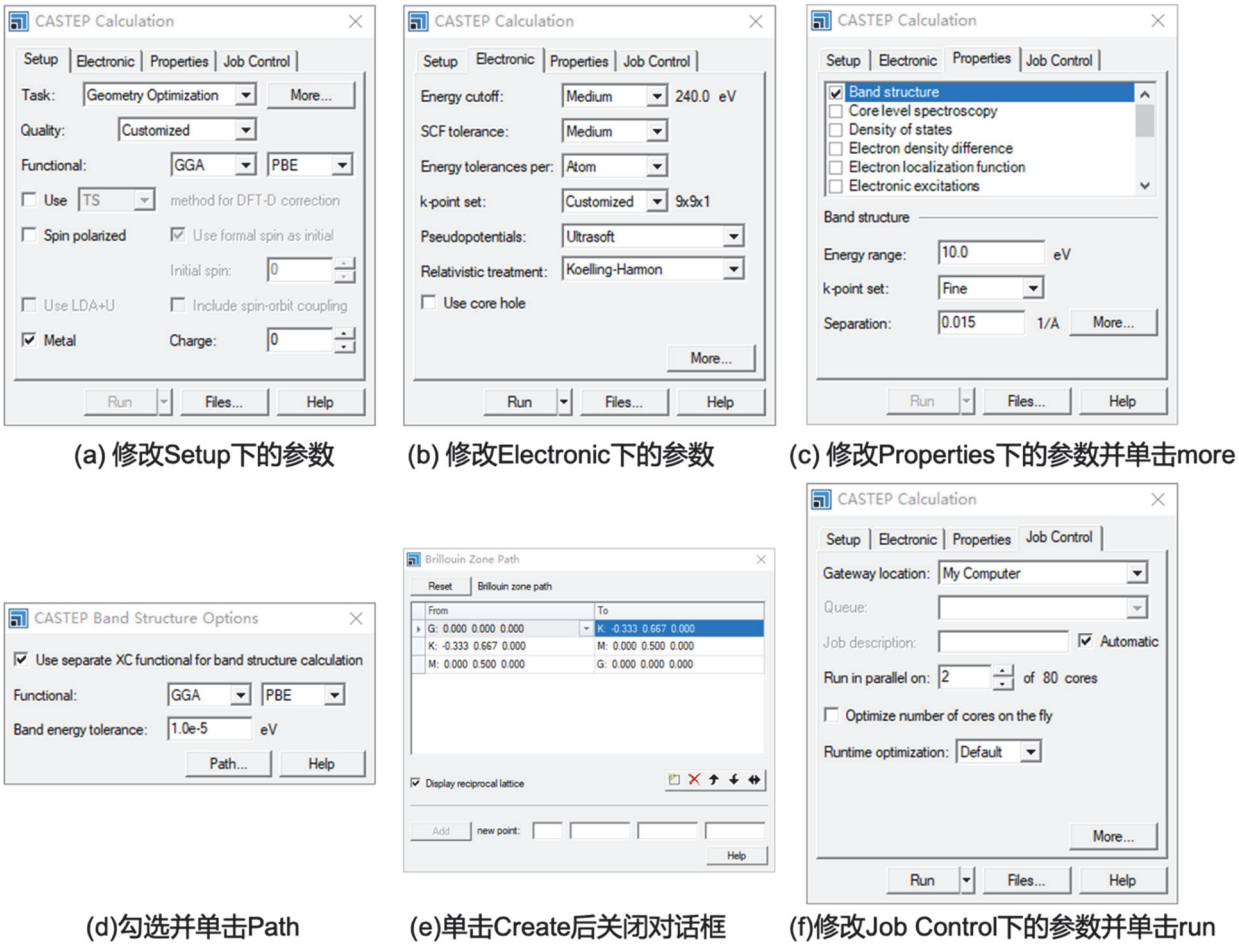
(2) 导入晶胞结构，计算能带结构

启动Materials Studio并创建一个新项目，打开New Project对话框，输入2D material作为项目名，单击OK按钮；在Project Explorer内，右击根目录从快捷菜单中选择New | Folder对话框输入2D；从菜单栏选择File | Import对话框，选择导入的氮化硼等二维材料结构文件。

在BN结构文件下，单击Modules工具栏上的CASTEP按钮，选择Calculation，或从菜单栏中选择Modules | CASTEP | Calculation，打开CASTEP Calculation对话框。选择Setup选项卡，把任务Task更改为几何优化Geometry Optimization，计算精度Quality设置为Fine，方法Functional选择GGA和PBE或者HSE06；单击几何优化计算任务Task后面的More…按钮，打开CASTEP Geometry Optimization对话框，勾选Optimize cell对话框，关闭对话框。选择Electronic选项卡，截断能Energy cutoff设置为Medium；SCF tolerance设置为Medium；K点大小设置单击More...按钮，选择K-point对话框，选中Custom grid parameters后，在Grid parameters中选择合适大小(如9 × 9 × 1)，关闭对话框；Pseudopotentials设置为Ultrasoft。选择Properties选项卡，勾选能带结构Band structure复选框，单击More...按钮，打开CASTEP Band Structure Options对话框；单击Path...按钮，打开布里渊区路径Brillouin Zone Path对话框。单击Create按钮，关闭两个对话框；单击Run按钮，关闭对话框如图1所示。

(3) 机器学习拟合禁带宽度

将计算所得的二维材料禁带宽度与网络数据库中的数据进行相互验证，确认数据库中数据的准确性。之后，将网上数据库中的二维材料数据导入到Jupyter平台，运行命令见图2。



(a) 修改Setup下的参数　(b) 修改Electronic下的参数　(c) 修改Properties下的参数并单击more

(d)勾选并单击Path　(e)单击Create后关闭对话框　(f)修改Job Control下的参数并单击run

图1　CASTEP计算操作步骤图示

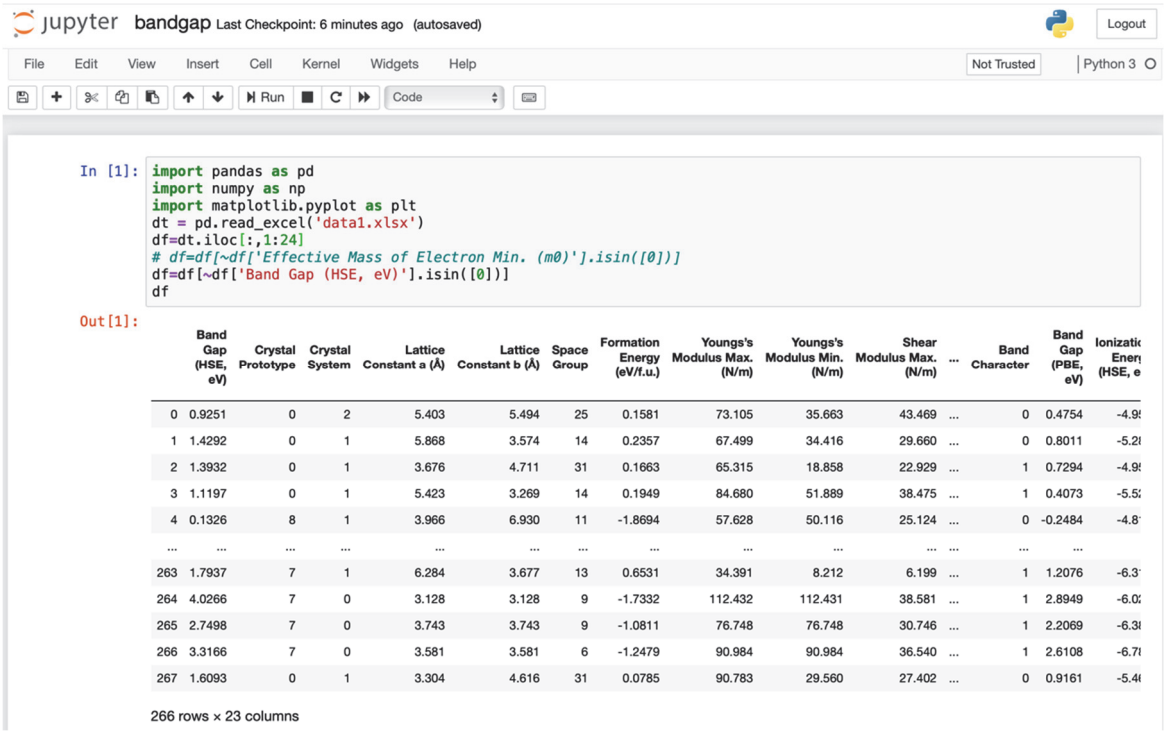


图2　二维材料数据(部分)

之后，应用各种不同的机器学习方法建立各种参数和HSE06禁带宽度之间的关系，包括线性回归、随机森林和神经网络等。图3展示的是在Jupyter中导入随机森林模型，并设置相关参数，进行数据训练、预测，并计算误差。

```
In [4]: from sklearn.ensemble import RandomForestRegressor
        from sklearn.model_selection import RandomizedSearchCV
        import matplotlib.pyplot as plt
        import sklearn.metrics as metrics

        # 引入随机森林回归模型
        rf = RandomForestRegressor(n_estimators=500,
                                   random_state=30,
                                   max_depth=12,
                                   oob_score=True)

        # 训练数据
        rf.fit(train_features, train_labels)
        # 使用模型预测数据
        predictions = rf.predict(test_features)
        plt.figure()
        plt.plot(np.arange(len(predictions)),predictions,'ko-',label="Predict value")
        plt.plot(np.arange(len(predictions)),test_labels,'ro-',label="True value")
        plt.title(f"RandomForest")
        plt.legend(loc="best")
        plt.show()
        # 计算误差值
        errors = abs(predictions - test_labels)
```

图3　随机森林模型参数

5　数据处理

(1) 六方氮化硼几何结构和能带结构

当结果文档被传输和下载后，将得到以下的几个文档：

BN.xsd：最终优化后的结构

BN.xtd：每个优化步骤后的结构组成的轨迹文件

BN.castep：包含优化信息的输出文本文件，包含有限的能带结构信息

BN.param：模拟中所使用的参数

更详细的信息包含在BN_Band Structure.castep文档中，单击Modules工具栏上的CASTEP按钮并选择Analysis，或从菜单栏里选择Modules | CASTEP | Analysis，打开CASTEP Analysis对话框：选择能带结构Band structure选项，单击View按钮，将创建一个包含能带结构的图文文档。

从输出文件中同样可查看六方氮化硼的的几何结构(图4a)。这里分别用PBE和HSE06两种方法计算六方氮化硼的能带结构(图4b、c)，其中PBE和HSE06计算出的带隙分别为4.673和5.561 eV。通常认为PBE低估了材料的带隙，而HSE06计算的带隙更加接近实际情况。

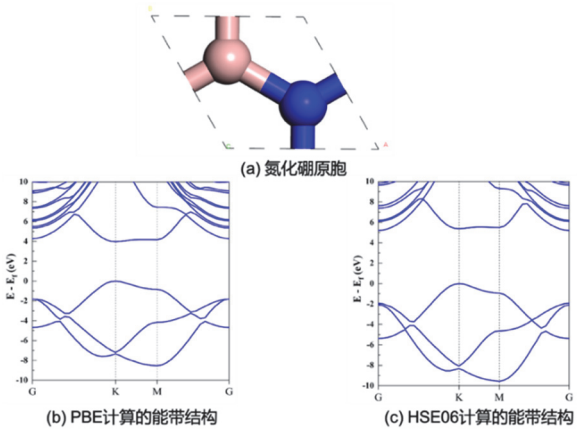


图4　六方氮化硼结构(a)和能带结构(b、c)

(2) 机器学习拟合HSE06带隙。

首先，我们计算出数据库中全体特征之间的皮尔森相关系数，如图5所示。皮尔森相关系数是一种用于度量两个变量之间线性相关程度的统计量。皮尔森相关系数的取值范围在−1到1之间，当其接近1时，表示两个变量之间存在强正相关，当其接近−1时，表示两个变量之间存在强负相关，当其接近0时，表示两个变量之间没有线性相关性。结果表明HSE06带隙值和PBE带隙值的相关程度最大。

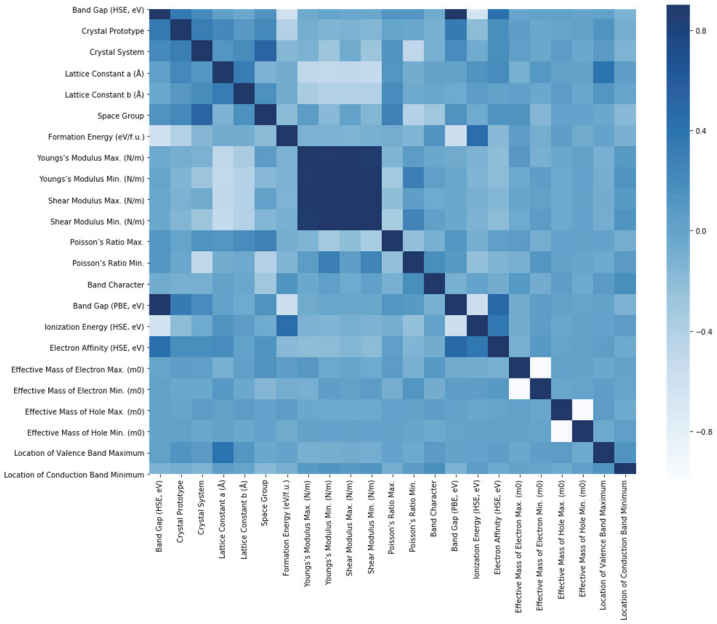


图5　全体特征皮尔森相关系数图

运行前述图3的随机森林模型，结果如图6所示。随机森林模型得分为0.9736。除了随机森林模型，还可以选择梯度提升、神经网络等机器学习方法，也可获得满意的预测结果。

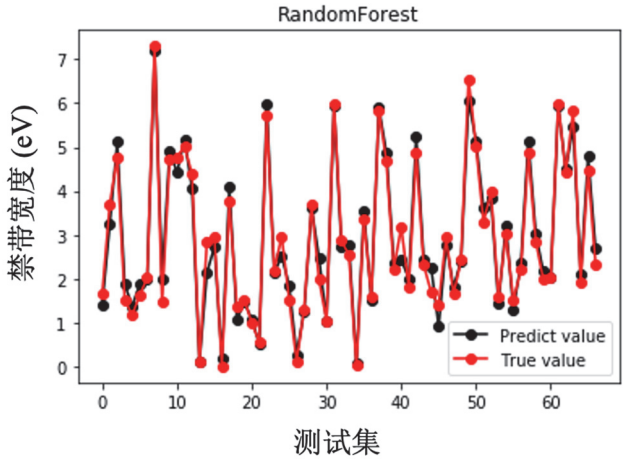


图6　随机森林模型运行测试集的结果

图7是模型预测结果与真实值的分布点图，清晰可见分布点基本在对角线附近，表明机器学习的结果令人满意。

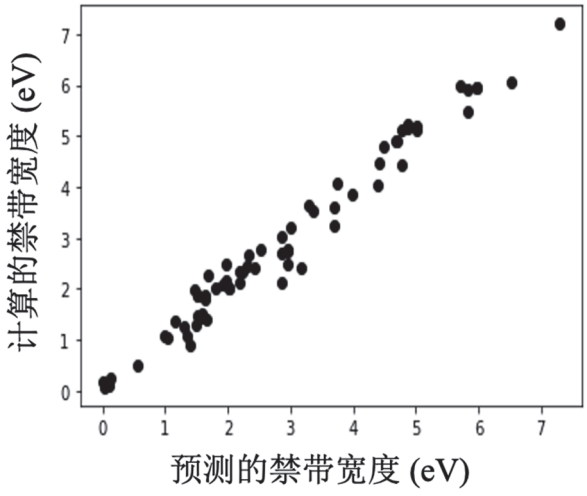


图7　模型预测结果与真实计算值的对比图

6　思考题

(1) 二维材料有哪些实际用途？计算二维材料的禁带宽度有哪些方法？

(2) 机器学习模型有哪些？请选择一种详细说明。

(3) 哪些机器学习模型适合预测二维材料禁带宽度？

7　实验教学安排

本实验主要面向大三下学期或大四上学期，已学习过结构化学相关知识的学生开设的探索性计算材料学实验。实验前会要求学生进行文献调研了解研究背景、二维材料发展历程和用途、机器学习原理等，并提出以下思考题：① 二维材料与传统材料相比有哪些优势？② 机器学习作为人工智

能的一个分支，它的优势主要体现在哪里？③ 目前机器学习存在的问题还有哪些？通过这些问题让学生提前理解实验内涵，好在听课时更容易进入状况。

实验用时4课时，具体为：简要介绍材料模拟和机器学习的相关知识，1课时；介绍 Materials Studio、CASTEP、Jupyter，讲解如何编制输入文件以及如何查看输出文件，1课时；上机计算并完成数据的计算与整理，2课时。课后学生独立完成实验报告和思考题。

本实验的难点在于为化学、材料背景的学生讲解机器学习的相关知识并学会初步使用机器学习模型解决材料化学相关问题。当前，我校学生大多具有使用python程序的经验，基于python的Jupyter机器学习环境搭建对于学生并非完全陌生，这里教学的重点在于机器学习模型的选择和相关参数的选择，教师会首先讲解实际案例，之后通过课堂上学生自行摸索，辅之教师现场指导，往往可以解决绝大部分学生所遇到的问题。最后通过学生完成实验报告的情况也会对课程的完成度有个明显的反馈。

8 结语

立足当前科研学术热点，设计与时俱进的创新型和综合型化学实验一致是我们教学研究的重点。学生通过本实验既可以学习二维材料的相关知识，也可以学习运用计算材料学和机器学习方法研究、分析材料相关性质。这将有助于激发学生的学习兴趣，锻炼学生解决问题的能力以及提高学生的科学素养。

参 考 文 献

[1] Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Zhang, Y.; Dubonos, S. V.; Grigorieva, I. V.; Firsov, A. A. *Science* **2004,** *306* (5696), 666.

[2] Pakdel, A.; Bando, Y.; Golberg, D. *Chem. Soc. Rev.* **2014,** *43* (3), 934.

[3] Chi, Z.; Zhao, J.; Zhang, Y.; Yu, H.; Yu, H. *Green Energy Environ.* **2022,** *7* (3), 372.

[4] Han, G. H.; Duong, D. L.; Keum, D. H.; Yun, S. J.; Lee, Y. H. *Chem. Rev.* **2018,** *118* (13), 6297.

[5] Miró, P.; Audiffred, M.; Heine, T. *Chem. Soc. Rev.* **2014,** *43* (18), 6537.

[6] Wang, Y.; Wang, L.; Zhang, X.; Liang, X.; Feng, Y.; Feng, W. *Nano Today* **2021,** *37*, 101059.

[7] 周公度, 段连运. 结构化学基础. 第5版 北京: 北京大学出版社, 2017: 128–129.

[8] Bickelhaupt, F. M.; Baerends, E. J. Kohn-Sham Density Functional Theory: Predicting and Understanding Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B. Eds.; John Wiley & Sons, Ltd.: New York, NY, USA, 2000; pp. 1–86.

[9] Wang, V.; Tang, G.; Liu, Y.-C.; Wang, R.-T.; Mizuseki, H.; Kawazoe, Y.; Nara, J.; Geng, W. T. *J. Phys. Chem. Lett.* **2022,** *13* (50), 11581.

[10] Hu, W.; Zhang, L. *Mater. Today Commun.* **2023,** *35*, 105841.

[11] Liu, H.; Xu, L.; Ma, Z.; Li, Z.; Li, H.; Zhang, Y.; Zhang, B.; Wang, L.-L. *Mater. Today Commun.* **2023,** *36*, 106578.

[12] 游洋, 杜婉, 李惟驹, 陈竞哲. 上海大学学报(自然科学版), **2020,** *26* (5), 824.

[13] 任更波, 李依然, 梅竣乔, 王美艳. 教育教学论坛, **2024,** No. 6, 132.

[14] 张乐飞, 罗勇, 杜博. 中国大学教学, **2023,** No. 5, 18.

[15] 朱红艳, 梁诗凯, 何富运, 黎海生, 夏海英. 广西物理, **2022,** *43* (2), 146.

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

[17] Materials Studio 2017; Accelrys Software Inc.: San Diego, CA, USA, 2017.

[18] Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. I. J.; Refson, K.; Payne, M. C. *Z. Kristallogr. Cryst. Mater.* **2005,** *220* (5–6), 567.