



文献 CSTR:
32001.14.11-6035.csd.2025.0034.zh



文献 DOI:
10.11922/11-6035.csd.2025.0034.zh
数据 DOI:
10.57760/sciencedb.j00001.01295

文献分类: 物理学

收稿日期: 2025-03-15
录用日期: 2025-10-24
发表日期: 2025-12-29

适用于高效第一性原理计算的数值原子轨道数据

张文帅¹, 林霏泽², 陈默涵³, 何力新^{4*}

- 1. 中国科学技术大学, 公共实验中心, 网络信息中心, 科学数据中心, 合肥 230026
- 2. 合肥综合性国家科学中心人工智能研究院, 合肥 230026
- 3. 北京大学, 高能量密度物理数值模拟教育部重点实验室 应用物理与技术研究中心 工学院, 北京 100871
- 4. 中国科学技术大学, 量子信息实验室, 合肥 230026

摘要: 数值原子轨道是一种描述原子结构中电子波函数的高效基组, 特别适合大体系原子结构(数千个原子以上)中的电子结构模拟。由于该类数据的构建与测试流程较为复杂, 业内仍然缺乏兼具完善质量控制与广泛适用性的数据产品。为此, 本工作系统性地构建了覆盖 68 种元素的数值原子轨道数据集, 支持多种赝势、截断参数(半径/能量)及基组规模配置。在轨道数据构建过程中, 采用了分层优化策略, 先优化基组数量较少的低精度轨道, 再逐步提升到基组数量较多的高精度轨道, 从而满足不同效率与精度的第一性原理计算需求。此外, 为展示数据可靠性并为用户提供选择依据, 本数据集提供了全面的质量评估数据, 例如单晶态方程拟合数据和分子体系的键长-能量数据等。评估数据显示, 采用本数据集的原子轨道基组与使用标准平面波基组所计算的单晶体系能量的平均差异低至 0.66 meV/atom。本数据集已被 ABACUS (原子算筹) 第一性原理计算软件采用, 能大幅提升复杂体系的模拟效率与精度, 其在凝聚态物理、材料科学及化学模拟等领域具有应用价值。

关键词: 第一性原理计算; 原子轨道; 数值计算; 科学数据

数据库(集)基本信息简介

数据库(集)名称	适用大体系原子结构第一性原理计算的高效原子轨道数据
数据通信作者	何力新 (helx@ustc.edu.cn)
数据作者	张文帅、林霏泽、陈默涵、何力新
数据生成时间范围	2016–2024年
地理区域	全球
能量精度量级	1 meV/atom
数据量	1.2 GB (未经ZIP压缩前)
数据格式	*.orb, *.dat, *.txt, *.png, *.pdf
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01295
基金项目	国家自然科学基金(12105283); 中国科学院网络安全和信息化专项(CAS-WX2022SDC-SJ13); 国家重点研发计划(2016YFB0201202)。

* 论文通信作者
何力新: helx@ustc.edu.cn

数据库（集）组成

数据集共包含26648个数据文件（不含根目录的README.md及参考数据）。为提高传输效率，部分文件夹以压缩包形式发布，本统计描述仍基于原始文件结构。所有文件分为两类，一类为轨道数据（包括*.orb文件、*.txt文件以及构建流程的输入文件INPUT/Slab_INPUT），另一类为质量评估数据（位于以Test/Plot为前缀的文件夹中，包括*.dat、*.png、*.pdf文件）。前一类文件可按路径名字进行区分，分别对应3种赝势（根目录名）、不同优化算法（CGH/LRH）、不同基组规模配置（SZ/DZP/TZDP等）、不同能量截断（*Ry）、多种截断半径（*au）以及68种元素。后一类文件包括针对标准平面波基组或原子轨道基组的测试结果文件（分别位于以PW或LCAO为后缀的文件夹内）以及统计结果文件（位于Test_Summary文件夹内）。

引言

近年来，基于密度泛函理论（Density Functional Theory, DFT）^[1-2]的第一性原理计算方法成为研究凝聚态物理、材料科学、化学及生物学等领域复杂系统的有力工具。随着超级计算机的发展和数值算法的进步，使用第一性原理计算方法进行包含数千个原子的电子结构以及动力学性质的研究已经成为可能。在众多第一性原理计算算法中，传统常用的平面波（Plane Wave, PW）基组方法在处理大原子体系时具有扩展效率较低的问题。为提高第一性原理计算的扩展效率，使用数值原子轨道，并基于线性叠加的原子轨道方法（Linear Combination of Atomic Orbitals, LCAO），可以解决包含数千原子的大体系结构的模拟问题^[3]。该类基组具有紧凑性和局域性等特点，可以使核心热点计算部分更易并行扩展到支持数千 CPU 核心的计算资源^[4]。

特别是，结合赝势方法所获得的赝原子轨道^[5]，相比“全势全电子”原子轨道，减少了描述核心电子波函数的基组空间，进一步显著降低了热点数组与矩阵的大小，因此支持更高效的模拟计算^[6]。但是，目前的电子结构模拟仍然主要使用以平面波为代表的低效基组，使用原子轨道基组进行的模拟仍然不是主流。一个核心的原因是，业内缺乏稳定、完整且获得精确度验证的原子轨道数据以及使用该类基组的计算软件。例如，2016年 Kurt 等人在 Science 期刊发表的 DFT 计算方案评估论文中，研究了 25 种不同方案的计算精度差异，其中仅有一个方案（OpenMX-MBK2013）采用了赝原子轨道作为基组^[7]。并且，该方案的计算精度接近最低水平，所使用的原子轨道数据也仅支持一种赝势。

本工作聚焦于构建稳定、完整且高精度的赝原子轨道数据，以更好地支撑与推广基于原子轨道的高效电子结构模拟与计算。为提升稳定性，选择了多种参考系统（如不同键长的二聚体、三聚体）并进行了完善的质量测试；为保证完整性，支持了多种赝势、截断参数（半径/能量）及基组规模配置，并覆盖 68 种元素，以适用于广泛的材料类型；为同时提供高精度与高效率的原子轨道数据，采用了分层优化策略，先优化基组数量较少的低精度轨道，再逐步提升到基组数量较多的高精度轨道，以满足多样的计算需求。

区别于其他原子轨道数据，本次构建采用了两种优化算法，获得了两批数据产品。第一种是 Chen 等人在 2010 年提出的“可系统改进的”原子轨道基组的优化算法（后称 CGH 算法）^[8]。该算法不预先假设原子轨道的形状，而是通过最小化波函数的残差来优化径向函数，能够产生具有高可移植性的原子基组，并且可以通过增加基组大小或加入更高角动量的轨道来系统性地提升基组质量^[9]。第二种是 Lin 等人在 2021 年提出的新方案，通过将参考波函数的梯度纳入优化目标，可以构建出更精确的数值原子轨道基组（后称 LRH 算法）^[10]。前一种优化算法所生成的原子轨道经过了更长时间

的应用与检验，仍具有一定价值，并且可以作为基准，以评估第二种优化算法所生成的原子轨道的质量。

本工作不仅提供大量原子轨道数据，而且根据单晶物质状态方程等标准计算的结果，在筛除不合格原子轨道数据的基础上为留存的每个原子轨道数据提供了完善的质量评估结果数据，为基于该轨道基组数据集的第一性原理计算提供可信度与轨道选择依据。

本工作产生的数据已经通过 ABACUS 软件^[11-13]实现了重用价值，在针对多种复杂材料系统（例如从分子、固体、表面到缺陷等）的计算中表现出色，证明了该软件的广泛适用性和高精度，对材料模拟等多个领域的科学研究具有重要的支撑作用。

1 数据采集和处理方法

本文构建的数据集来源于第一性原理计算所得，通过第一性原理计算，可以不依赖实验观察数据获得真实高精度的物质性质数据。本文即采用这种采集途径，通过两种已发表的创新算法（CGH 与 LRH 算法），结合在迭代优化过程中择优选择的输入配置参数，创建了适合描述原子分子结构中电子波函数的基组数据。

以下介绍其中的算法核心以及优化的输入参数。

1.1 数据生成方法

原子轨道基组数据构建的核心原理是，先使用平面波基组算法，选择合适的赝势、能量截断等可调参数，计算出适当参考结构中的电子波函数，获得需要描述的希尔伯特空间，然后尝试不同位型与规模的局域轨道基矢，以少量基矢实现对目标电子波函数的精确表达。其中，最关键的是设计局域轨道的形状，通常只需要设计径向轨道形状即可。

CGH 与 LRH 算法均采用线性组合球贝塞尔函数的方式获得试探波函数^[8]，球贝塞尔函数具有零节点可调以及局域性两种优势，可以更好地描述轨道，该线性组合方式可表示为：

$$f_{\mu,l}(\mathbf{r}) = \begin{cases} \sum_q c_{\mu q} j_l(qr), & r < r_c \\ 0, & r \geq r_c \end{cases} \quad (1)$$

其中， $j_l(qr)$ 即为球贝塞尔函数， r_c 是设定的零点位置， q 被选择为使得 $j_l(qr_c) = 0$ ，下标 l 表示角动量量子数， $c_{\mu q}$ 是组合系数，下标 μ 是混合标签，被用于描述元素类型、原子序数、径向函数基组序数。

构建原子轨道数据的核心为优化组合系数 $c_{\mu q}$ ，因此需要评估两个尝试系数的优劣标准。在 CGH 算法中，采用波函数剩余残差 Spillage0 指标作为判定标准，其可以表达为：

$$\text{Spillage0} = \frac{1}{N_n} \sum_{n=1}^{N_n} || |\Psi_n\rangle - |\tilde{\Psi}_n\rangle ||^2, \quad (2)$$

$$\tilde{\Psi}_n = \sum_{\mu,l,m} a_{\mu l m n} f_{\mu,l}(\mathbf{r}) Y_{lm}(\mathbf{r}) \quad (3)$$

其中， $\tilde{\Psi}_n$ 是参考波函数 Ψ_n 通过投影到球贝塞尔函数与球谐函数空间的投影波函数， $a_{\mu l m n}$ 是投影系数。在 LRH 算法中，增加了一项针对波函数一阶导数的差异性数值，记为 Spillage1，可表达为：

$$\text{Spillage1} = \frac{1}{N_n} \sum_{n=1}^{N_n} [|| |\Psi_n\rangle - |\tilde{\Psi}_n\rangle ||^2 + \alpha || |\nabla \Psi_n\rangle - |\nabla \tilde{\Psi}_n\rangle ||^2], \quad (4)$$

其中 ∇ 表达一阶导数计算， α 是增加项的权重。

最终,原子轨道基组的优化问题变为最小化 Spillage0 或 Spillage1 的数值计算问题。因为 Spillage0 中没有对导数的任何约束,可能在生成过程中产生高动量的非物理轨道,在实际优化算法中,会加入对动能的计算优化,不显著影响 Spillage 的高动能轨道会被惩罚或消除。同时,考虑计算过程中需要不同精度与计算效率的轨道,更精简与更局域的轨道也会有计算速度较快的优势或适合的场景,同时也能满足特定计算的误差需求。因此有必要优化多种轨道,以满足不同的计算需求。

在本文的轨道数据生成工作中使用了分层优化策略:首先优化低精度与基矢数量较少的原子轨道作为 Level1。然后,在 Level1 优化结果的基础上,进一步增加轨道基矢,作为 Level2 轨道。后者优化过程中,可以仅对参考波函数减除 Level1 层轨道所获得的最佳拟合波函数后所得的残差进行最小化拟合,即固定(后文中称为 fix 参数) Level1 轨道,减少了 Level2 的待优化轨道空间,提升了优化速度。图 1 示意了本文中的原子轨道分层优化流程。

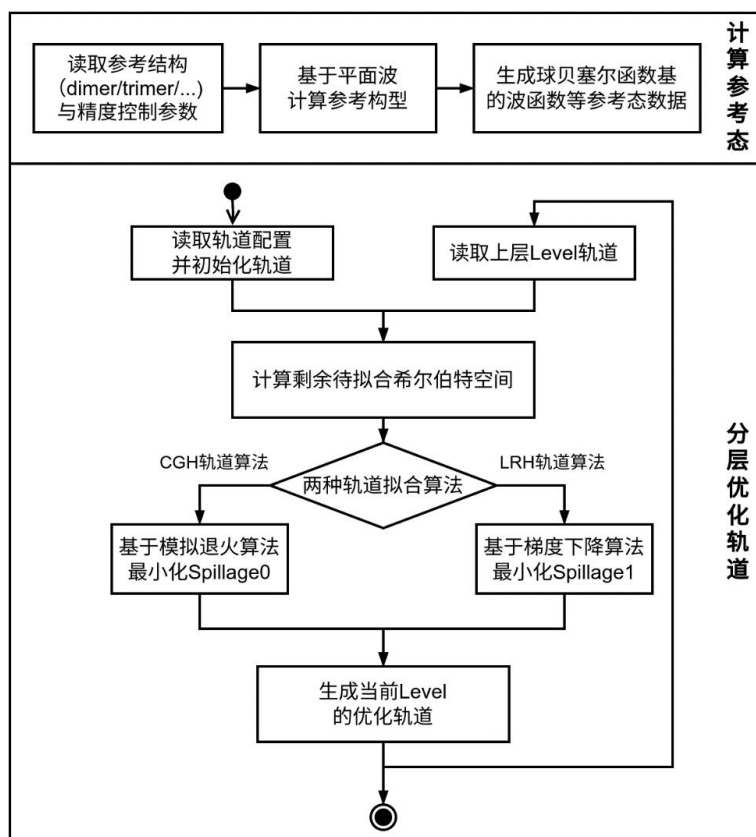


图 1 原子轨道基组分层优化流程示意图

Figure 1 Schematic diagram of the hierarchical optimization process for atomic orbital basis sets

1.2 数据制备步骤

本节讲述如何使用系统性可调的原子轨道生成器 SIAB^[14-15]来构建原子轨道数据,首先是准备依赖数据,以及输入文件来配置优化参数,其次是搭建计算环境与执行优化。

首先需要准备的数据是赝势,在进行平面波与基于原子轨道的第一性原理计算中均需要该输入数据。本工作中用到的赝势有三组,分别是 ONCV SG15 1.0 赝势^[16-17]以及 Dojo 赝势^[18]中的 NC SR 0.4 分支与 NC FR 0.4 分支。三组赝势均具有本文中涉及的 68 种元素结构计算所需的齐全数据,后面分别称为 SG、SR、FR 赝势,本文中均需要使用 UPF (Unified Pseudopotentials Format) 格式的数据^[19]。

其次,需要针对每个元素构建轨道生成器 SIAB 的输入文件。本工作中使用的 SIAB 程序有两个

版本，分别对应两种输入文件，第一个版本支持三种优化方法，分别是基于 Spillage0 指标进行模拟退火与梯度下降的两种优化方法以及基于 Spillage1 指标的梯度下降优化方法^[14]。该版本的输入文件的参数结构设计相对复杂，难以清晰地描述电子轨道配置的分层结构以及更简单地适配不同的构型，为帮助用户更容易地配置输入，开发了第二个版本^[15]，第二个版本仅支持最新的基于 Spillage1 的梯度下降优化方法。我们将第一个版本用于最原初的基于 CGH 算法的原子轨道基组构建中，将第二个版本用于更精确高效的 LRH 算法的基组构建中，两个版本均支持分层的原子轨道构建过程。

以下通过参数表详细说明两个输入文件的配置方法，表 1 中给出了第一个版本的 SIAB 程序的输入文件中所定义参数及其含义，本研究使用此表中的输入参数来运行 CGH 算法。

表 1 第一版 SIAB 程序的输入参数

Table 1 Input parameters of the first version of the SIAB program

序号	参数名	参数值描述
1	EXE_mpi	填写并行启动 MPIRUN 配置，例如 mpirun -np 8。
2	EXE_pw	填写 ABACUS 二进制程序路径或可执行命令，如 ABACUS.mpi。
3	EXE_orbital	填写支持 CGH 模拟退火轨道优化程序的路径或命令，如 SIA.exe。
4	targets	填写轨道优化的执行目录名，不存在时会自动生成，如 14_Si_100。
5	ref_bands	填写参考波函数的能带数。
6	nbands	填写用于计算的总能带数量。
7	Ecut	平面波计算时的能量截断（单位 Ry）。
8	Rcut	生成的原子轨道基组轨道半径截断值（单位 Bohr）。
9	Pseudo_dir	赝势文件所在目录。
10	Pseudo	赝势文件名。
11	sigma	平面波计算中使用的费米面附近高斯模糊展宽能量（单位 Ry），可选 0.01。
12	BLSTRU#	参数名中的#为数字，表达第#个参考构型，参数值为构型的键长列表。
13	SkipSTRU	针对每个构型，是否跳过基于该构型的轨道优化，适用于中断重启后的续算。
14	ListSTRU	构型的名称，例如 dimer、trimer 等，用于自动创建平面波计算的原子结构。
15	maxL	优化的多层级轨道中最大的角动量，用于创建可兼容复用的参考态。
16	Level	分层优化轨道时的层数。
17	level#	其中#为层数指标（从 1 开始），第一个值为本层轨道最大角动量量子数，后续值为每个角动量的轨道数，如当优化目标为 2s1p 轨道时，即为 1 2 1。
18	Start_tem_S	模拟退火优化 Spillage0 时的起始温度，例如 1.0e-3。
19	Start_tem_K	模拟退火降低动能项时的起始温度，例如 1.0e-1。
20	Step_S	优化 Spillage0 的随机步数，为保证终点足够稳定，可以设置高一些，如 50。
21	Step_K	降低动能项的随机步数，为更好地排除非物理的高动能成分，可设置为 30。
22	Delta_kappa	降低动能时对 Spillage0 变大的宽容度，可固定为 0.01。

当使用第二版 SIAB 原子轨道分层优化程序与 LRH 算法创建原子轨道数据时，输入文件中所需要配置的参数如表 2 所示。

表 2 第二版 SIAB 程序的输入参数

Table 2 Input parameters of the second version of the SIAB program

序号	参数名	参数值描述
1	EXE_env	Python 脚本启动系统 Shell 进程做平面波计算时，需要在 Shell 中重新加载环境所用到的执行命令。
2	EXE_mpi	填写并行启动 MPIRUN 配置，例如 mpirun -np 8。
3	EXE_pw	填写 ABACUS 二进制程序路径或可执行命令，如 ABACUS.mpi。
4	element	元素名，如 Ag。
5	Ecut、Rcut 等	Ecut、Rcut、Pseudo_dir、Pseudo、sigma 参数值与表 1 中解释相同。
6	STRU#	第#号参考构型的波函数计算中，所需要的构型名、能带数、最大角动量、自旋参数与键长，如 dimer 33 3 1 2.3 2.7 3.2。
7	Level#	第#层轨道优化时，需要使用的参考构型号（即 STRU#）、参考能带（RefBands）、是否固定输入的轨道（InputOrb）、轨道配置（OrbitalConf）。
8	RefBands	为 auto 时，自动拟合所有基态能带；为数字 n 时，表示对指定构型的所有键长参考实例都采用 n 作为参考波函数的能带数；当为列表时，需要为每个参考实例设定参考能带数，如有 3 个不同键长实例，可设置为 “[3,4,5]”。
9	InputOrb	可以为 none、fix、unfix，分别表示没有输入轨道、固定输入轨道、仅以输入的轨道基组作为初始试探基组的一部分，并继续进行优化。
10	Save#	分别给出需要保存的轨道优化层级以及保存名称，例如 Level2 DZP。

在创建原子轨道基组数据库的过程中，选择不同的赝势、截断半径、轨道配置以及参考态构型与键长等，均会导致所生成的轨道具有不同精度与迁移适应性。特别是，键长一般选择在平衡态附近，同时避免平面波计算的不收敛情况。当出现平面波计算不收敛时，仍会完成参考态输出并完成轨道优化，此时的轨道具有更大的质量风险，需要检查不收敛的具体原因，并仔细检验生成轨道的质量。

完成数据准备后，需要搭建计算环境与采用合适的资源完成计算优化。基础 ABACUS 软件的安装方法可见于官网^[13]的手册栏目给出的 PDF 文档或网页版文档中，主要需要 MPI 并行计算函数库与各种线性代数函数库。这里重点强调基于 LRH 算法进行原子轨道基组生成时需要安装 PyTorch 软件，利用该软件的高效梯度下降优化模块，可以大大加快构建速度。本文推荐使用测试有效的 v1.1.0 版本，其他版本有较大概率出现兼容问题。安装完成后，可以如两版软件原始发布网页^[14-15]中所示，分别运行其中提供的 Bash 与 Python 脚本即可。

2 数据样本描述

本节主要讲述生成的数据目录结构，给出数据的详细解释与用途。

2.1 核心数据文件目录结构

表 3 中给出了本文构建的核心数据：原子轨道基组数据，以及输入配置数据的文件目录结构，其中正体部分是文件（夹）示例，斜体字部分是解释。本文发布的数据的第一层目录中的

“Reference_Data”文件夹保存了质量评估环节需要引用的参考数据，这里不作详述。其他三个文件夹，分别是“Dojo-NC-FR-0.4”“Dojo-NC-SR-0.4”“SG15-1.0”，所述名字即为三组赝势的名字。

表 3 原子轨道基组数据文件与输入配置文件目录结构

Table 3 Directory structure of atomic orbital basis set files and input configuration files

Dojo-NC-FR-0.4							
	LRH__SZ-DZP-TZDP__E100Ry.zip						
Dojo-NC-SR-0.4							
	LRH__SZ-DZP-TZDP__E100Ry.zip						
SG15-1.0							
	LRH__SZ-DZP-TZDP__E100Ry.zip						
	CGH__MixDZP__E60Ry.zip						
	CGH__MixDZP-MixTZDP__E100Ry.zip						
	CGH__ZP-DZ-DZP-TZDP__E100Ry.zip						
		Orbitals_Complete_E100_DZP.json					
		Orbitals_Standard_E100_DZP.json					
		Orbitals_Dataset					
			Ag_DZP				
				Ag_gga_10au_100Ry_4s2p2d1f.orb			
				... Orbitals for other Rcuts			
				info			
					10 (注释：文件夹以 Rcut 数值命名)		
						ORBITAL_PLOTU.dat	
						ORBITAL_RESULTS.txt	
						SIAB_INPUT (注释：采用 LRH 算法时保存)	
				... Details for other Rcuts			
				ORBITAL_INPUT (注释：采用 CGH 算法时保存)			
Ag_ZP							
... Orbitals for other elements or configuration							

第二层目录名是产生原子轨道时使用的优化算法、基组规模配置以及计算参考态时使用的平面波的能量截断。如 Dojo-NC-FR-0.4 目录下有四个子目录，表示分别构建了 4 套原子轨道基组，其中 CGH__ZP-DZ-DZP-TZDP__E100Ry 表示使用 CHG 算法构建了 ZP、DZ、DZP、TZDP 四种不同规模的基组的轨道，并采用 100Ry 作为参考波函数的平面波能量截断。

在第三层目录中，包括 Orbitals_Dataset 文件夹以及多个 json 格式的文本文件，前者内含轨道基组数据，json 文件中写明了针对每个元素与对应基组类型所收录的轨道的 Rcut 列表，json 文件名中的“Complete”表示完整的 Rcut 列表，“Standard”表示基于后续质量评估结果所建议的标准 Rcut 值。

第四层目录 (Orbitals_Dataset 文件夹内) 包含针对每个元素与对应基组类型的轨道文件目录 (如

“Ag_DZP”），其中包括轨道文件以及 info 目录。一个轨道名示例，如“Ag_gga_10au_100Ry_4s2p2d1f.orb”，表示针对 Ag 元素、采用 GGA 泛函与 100Ry 的平面波能量截断制作参考波函数，所生成的具有 10 Bohr 轨道半径截断（Rcut）以及 4s2p2d1f 基组规模的原子轨道基组数据。info 目录中，给出了每个 Rcut 的输入与输出文件，采用 LRH 算法构建时，在每个 Rcut 文件夹中均保存了对应 Rcut 的输入文件“SIAB_INPUT”，方便使用者测试与修改；采用 CGH 算法时，仅在 info 目录内保存一个参考 Rcut 的输入文件“ORBITAL_INPUT”。同时，在所有 Rcut 文件夹中均保留了内层优化器的输入文件“INPUT”，其也是运行脚本输出文件，方便针对异常数据的交叉检验，一般情况下无需查看。

在每个 Rcut 文件夹中，保存有原始的基于球贝塞尔函数线性组合系数的原子轨道基组文件“ORBITAL_RESULTS.txt”，可以方便用户后续以此原子轨道作为输入初始值，进行 fix 或 unfix（涵义见表 2 中的 InputOrb 参数解释）的进一步优化。在部分 Rcut 文件中，还保留有 ORBITAL_PLOTU.dat 文件，其中的数据第一列为径向半径，后续列为各个数值的轨道，方便绘制轨道图形，所绘制的轨道图形例如图 2 所示。

在构建与发布以上数据时，因为其中有大量小文件不便传输，所以将部分第二层文件夹进行了压缩，以“.zip”文件的形式上传了数据，不影响整体的目录结构。

2.2 原子轨道基组数据结构

本小节描述数据集中最核心的原子轨道基组数据文件的数据结构，如表 4 所示。每个“.orb”后缀的轨道文件中，包含多个轨道径向函数基矢，对应多个角动量量子数 L 与多个主量子数 N，表中仅给出一个 L 或 N 作为示例。每个轨道径向函数基矢都有浮点数组表达，如当前示例中，数组长度为 801。

表 4 原子轨道基组数据文件中的数据结构示例

Table 4 Example of the data structure in an atomic orbital basis set

序号	数据结构键名	数据结构键值	解释
1	Element	Si	元素名
2	Energy Cutoff(Ry)	60	能量截断
3	Radius Cutoff(a.u.)	8	半径截断
4	Lmax	2	角量子数
5	Number of S orbital	2	S 角动量轨道数
6	Number of P orbital	2	P 角动量轨道数
7	Number of D orbital	1	D 角动量轨道数
6	Mesh	801	轨道径向函数数值网格数
7	dr	0.01	轨道径向函数数值网格间距
8	Type	0	类型占位，可用于元素识别
9	L	0	当前轨道径向函数基矢对应的角动量量子数
10	N	0	当前轨道径向函数基矢对应的主量子数
11	无键名符号	浮点数组向量	当前轨道径向函数基矢 $f(r)$ 的离散值

将所有的基矢进行绘图，即为后续 LCAO 计算中描述一个元素所需的所有径向自由度。图 2 给出了 Si 元素的一个原子轨道文件中所有径向函数的曲线示例，每条曲线即由表 4 中的第 11 项数据所绘制。

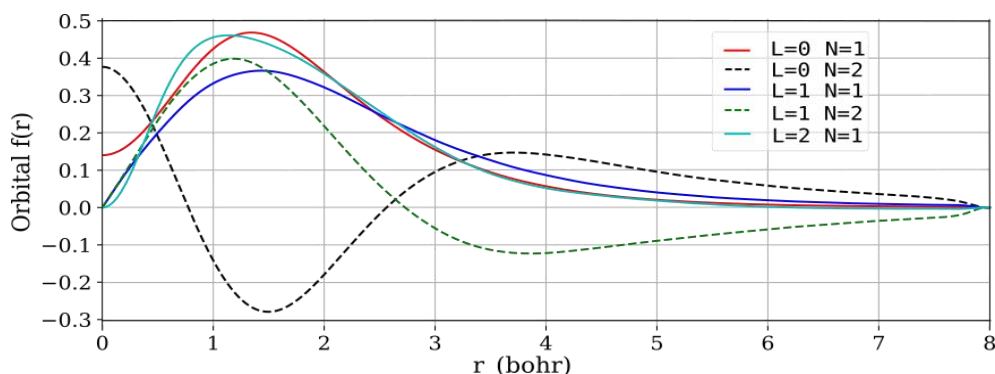


图 2 Si 元素 DZP 类型的原子轨道基组的径向分布

Figure 2 Radial distribution of DZP atomic orbitals for silicon

2.3 原子轨道基组大小的分类方式

本文构建的原子轨道基组的规模或大小可以分为 ZP、DZ、DZP、TZDP 等不同级别，并体现在前述基组数据目录结构的文件夹名称中，如“Ag_DZP”。本小节解释此分类的涵义。

以 Si 元素为例，其 SG15-1.0 赝势中定义了赝价电子层，对应的电子结构为 $1s1p$ ，该电子轨道分布被认为是 ζ 轨道（也称 Z 轨道）。比 Z 轨道中最高角动量的轨道高一个量子数的新轨道被称为极化轨道（即 P 轨道）。本文用 S 表示 Single，D 表示 Double，T 表示 Triple，所以，对于 SG15-1.0 赝势中的 Si 元素，SZ、DZ、ZP、DZP、TZDP 轨道的电子组态分别为 $1s1p$ 、 $2s2p$ 、 $1s1p1d$ 、 $2s2p1d$ 、 $3s3p2d$ 。

在前述目录结构中，出现有 MixDZP 标识，这是因为，有一部分元素的 Z 轨道中的最高角量子数轨道电子在里层，考虑，其被极化的程度可能更弱，无需进一步增加极化轨道，所以使用 DZ 配置代替了 DZP 配置，这部分元素有 Ga、Ge、In、Sn、Sb、Te、I、Xe、Ba、Hf、Ta、W、Tl、Pb、Bi。同时，在相应的 MixTZDP 轨道数据中，这些元素的轨道实际应该是 TZP 配置，即三层 Z 轨道加一层极化 P 轨道。我们将在本文后续的测试中，针对这部分元素，分析与展示 DZP 轨道与 DZ 轨道的精度差异，以及 TZDP 轨道与 TZP 轨道之间的精度差异。

对于部分元素的 TZDP 轨道，其 s 角动量的轨道数量过多，使得出现过拟合的现象，最后增加的 s 轨道常具有不平滑的形态，因此被舍弃，其不影响计算精度，也对计算速度影响极小，在本文中同样称为 TZDP 轨道。

在后续质量控制一节中，会展示针对每类基组的精度估算结果。在相关的数据文件中，还会针对每个元素，给出每个类别的轨道基组所对应的量子态层级结构。

3 数据质量控制和评估

本文提供的数据集中，特别增加了大量数据质量评估数据，以便用户能够更好地选择合适精度的轨道，对轨道的使用有很强的指导作用。为控制本轨道数据集的整体质量，本文中的原子轨道数据不仅基于 Spillage 目标进行优化，而且会基于本节中描述的质量评估指标进行数据筛选，剔除指

标数值不合格的数据，以构建更高质量门槛的原子轨道数据集。

3.1 质量评估数据文件的目录结构

质量数据的文件目录结构见表 5 所示，以下分别详述各个文件的意义。最基础的质量数据是优化目标 Spillage 的数值（包括前述 Spillage0 或 Spillage1），每个轨道的数据结果被汇总保存在 Test_Summary 文件夹中。例如文件“Spillages_DZP_orbE100.dat”中，包含有该文件的上层目录名所指示的赝势与优化算法下，基于 100Ry 平面波能量截断，所优化的每个元素的每个 Rcut 下的 Spillage。同时，在同名的“*.png”图片中，也绘制了相关数据。

表 5 原子轨道基组数据库质量评估数据

Table 5 Quality assessment of the atomic orbital basis set database

SG15-1.0	
	Test_VE_PW
	VE_WFEcut100Ry__pw_Ag.dat
	...
CGH__ZP-DZ-DZP-TZDP__E100Ry.zip	
	Plot_U
	10_Ne_DZP_orbE100_lcaoE100_PLOTU.pdf
	...
	Plot_VEBE
	10_Ne_DZP_orbE100_lcaoE100_VE.png
	10_Ne_DZP_orbE100_lcaoE100_BE.png
	...
Test_Summary	
	DeltaDFT_LCAO-PW_DZP_orbE100_lcaoE100.dat
	DeltaDFT_LCAO-WIEN2k_DZP_R10_orbE100_lcaoE100.dat
	DeltaDFT_WIEN2k-LCAO_DZP_orbE100_lcaoE100.dat
	Plot_DeltaDFT_WIEN2k-PW_E100.png
	EOS_Nmat_LCAO_DZP_Rstd_orbE100_lcaoE100.dat
	Spillages_DZP_orbE100.dat
	...
Test_VE_LCAO	
	VE_WFEcut100Ry__orb_Ag_gga_10au_100Ry_DZP.dat
	...

在目录“Plot_U”中，提供了每个轨道的曲线图，根据此图可以检查轨道的平滑度，确定其中是否有不合理的波动，以及 Rcut 截断是否过小，较小的 Rcut 会导致轨道在截断处的数值反常的急剧降低，影响原子轨道的精度。

在本节质量评估中，最关键的是被标识为 VE 的数据，其被保存在名字包含该标识的目录与文件中，这些文件保存了不同晶格基矢长度下的单晶晶格能量。例如，表 5 中所示“Test_VE_PW”文

文件夹包含每个元素与指定能量截断下的平面波（PW）计算所得的 VE 数据。文件“VE_WFEcut100Ry_pw_Ag.dat”有两列七行数据，第一列是平均到每个原子的晶格体积，第二列是相应晶格体积下的能量，七行数据分别对应单晶结构平衡态附近的 7 个体积点。基于 VE 数据计算获得的 DeltaDFT 数值被广泛用于验证使用不同 DFT 计算方法的计算程序之间的计算差异^[20]。所有 DeltaDFT 数据结果均被保存在 Test_Summary 文件夹中，文件名以“DeltaDFT”为前缀，该数据也被绘制于以“Plot_DeltaDFT”为前缀的图片文件中。

此外，基于 PW/LCAO 算法的多个 VE 数据的曲线图被保存在 Plot_VEBE 目录中。该目录中也存在名字以“BE.png”结尾的另一类图片，其中绘制了使用 PW/LCAO 方法所计算的分子体系的键长（Bond Length, BL）-能量关系曲线，所使用的分子体系为生成轨道时使用的分子参考结构。所有文件名称中，标识“orbE100”表示在生成轨道时使用了 100Ry 能量截断的平面波基组；标识“lcaoE100”表示在使用 LCAO 方法计算 VE/BE 曲线时，ABACUS 软件的输入文件“INPUT”中的 ecutwfc（波函数能量截断）参数值被设置为 100（单位是 Ry）。BE 图片文件中的 BL 坐标的单位为 Å。基于该 BE 数据，可以更好地选择合适的键长，以便对原子轨道数据做进一步优化，也可以用于筛选“平衡点”更精准的原子轨道。

3.2 质量评估与控制的核心指标 DeltaDFT

图 3 示例了使用不同软件或算法的代码 a 与 b 所计算的 VE 曲线，其中阴影部分表示差异的大小，即被称为 DeltaDFT。DeltaDFT 的具体表达式为：

$$\text{DeltaDFT}_i(a, b) = \sqrt{\int_{0.94V_{0,i}}^{1.06V_{0,i}} \frac{(E_{b,i}(V) - E_{a,i}(V))^2}{0.12V_{0,i}} dV} \quad (5)$$

其中， $V_{0,i}$ 表示第 i 个元素单质的平衡态体积，积分针对平衡态体积附近 6% 的变化值区间进行。本文中的平衡态结构点选择实验给出测量数据，对于极少数元素，如 Ba，在计算过程中不容易获得试验点附近的稳定极值，导致计算物态方程出错，6% 被扩大至 9%。本文轨道数据的构建与质量分析过程中，使用了三种不同算法或软件下的结果，包括基于 ABACUS 软件的 PW 或 LCAO 计算结果以及基于 WIEN2k 的全电子算法 FLAPW 的计算结果，因此对应有三对不同 DeltaDFT 结果，均保存在数据集中。

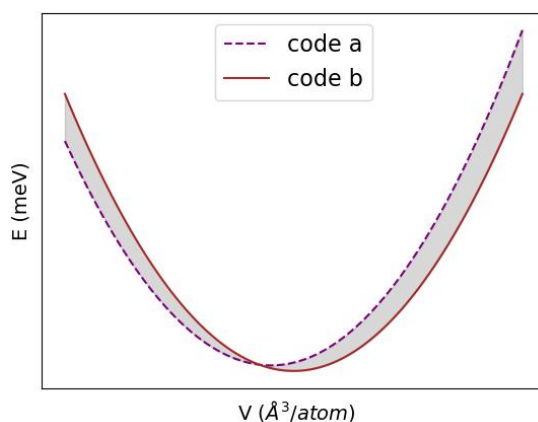


图 3 不同软件或算法所计算的 VE 曲线示例

Figure 3 Examples of VE curves calculated using different software or algorithms

所有的 DeltaDFT 计算结果被保存在 Test_Summary 文件夹中以“Delta”开头的文件中。LCAO 方法计算的 VE 数据被保存在各个“轨道类型”文件夹下的 Test_VE_LCAO 目录中, 该目录中还保存了名字前缀为“EOS_Nmat”的文本文件, 其第 2、3、4 列保存了平衡态体积 $V_{0,i}$ 数据、求解物态方程获得的体积模量数据 $B_{0,i}$ 与体积模量的压力导数 $B_{1,i}$ 数据, 三类数据的单位分别为 $\text{\AA}^3/\text{atom}$ 、GPa、无量纲。

3.3 评估轨道的多种属性变量对计算精度的影响

本小节根据 DeltaDFT 数据探讨与分析多种不同的赝势与轨道对 DFT 计算精度的影响, DFT 计算方法包括平面波 (PW) 方法、原子轨道线性叠加 (LCAO) 方法以及 WIEN2k 软件中采用的全势线性缀加平面波 (FLAPW) 方法^[21]。

WIEN2k 程序与 FLAPW 方法常被认为是 DFT 方法中最精确的参考方法, 因此, 可以作为基准, 来评估 PW/LCAO 方法的精确度。同时, LCAO 方法的原子轨道来源于逼近 PW 方法的参考结构量子态。因此, 通过对比 LCAO 与 PW 方法之间的 DeltaDFT 数值, 可以确定: 使用各个原子轨道基组, 相比参考波函数所使用的平面波基组, 会导致多大程度的计算偏差。本文的测试结果表明, 来源于局域原子轨道近似的平均能量偏差可以小至 0.66 meV/atom 。即便使用最精确的全电子方法的不同计算程序之间, 也常有如此大小的 DeltaDFT 差异^[7], 说明使用本文给出的原子轨道数据的 LCAO 计算, 可以与采用平面波基组的计算达到接近相同的物理分析与模拟效果。

(1) 评估不同规模轨道 (ZP、DZ、DZP、TZDP) 的精度与差异

图 4 展示了采用 SG15 赝势与 CGH 算法所构建的 ZP、DZ、DZP、TZDP 轨道的 DeltaDFT (LCAO, PW) 数值 (即基于相应轨道的 LCAO 计算结果与 PW 计算结果之间的差异), 图中横坐标 ME 表示所有元素的平均 DeltaDFT 数值。为简洁展示, 本图中仅绘制了 Rcut 为 10Bohr 的轨道的结果, 标准 Rcut 的结果具有相似性, 可以见于数据集中标识“Rstd”的 DeltaDFT 结果文件。可以看到, DZP 轨道整体上相比 ZP 与 DZ 轨道有明显的精度提升, 使得 DeltaDFT 数值基本降到了 10 meV/atom 以下。TZDP 轨道相比 DZP 轨道一般也有比较好的精度提升, 使得更多元素下 LCAO 的计算结果与 PW 计算的结果降到了 1 meV/atom 以下。对于多数元素下的 DFT 计算, 选择 DZP 轨道可以兼顾速度与精确性。

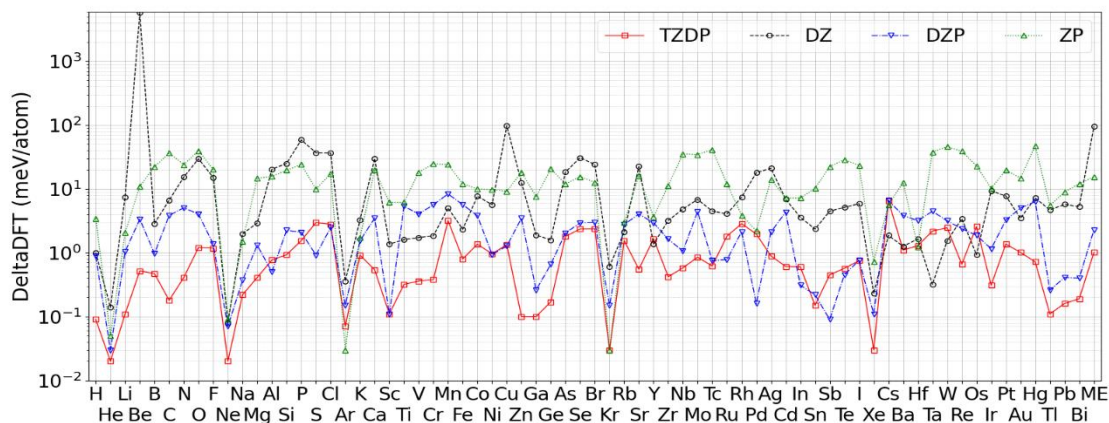


图 4 CGH 算法下 ZP、DZ、DZP、TZDP 轨道的 DeltaDFT(LCAO,PW)数值

Figure 4 DeltaDFT (LCAO, PW) values for ZP, DZ, DZP, TZDP orbitals under the CGH algorithm

(2) 评估使用不同平面波截断能量构建参考波函数对轨道精度的影响

以 MixDZP 轨道为例, 计算使用不同平面波能量截断的参考波函数所构建的轨道的精度差异, 结果如图 5 所示。图片显示, 该截断差异对精度影响不大, 绝大多数元素使用 60 Ry 截断即可达到结果收敛的效果。所有采用 60 Ry 截断的 DeltaDFT 结果均达到 10 meV/atom 以下, 已满足计算误差的基本需求。特别是, 少量采用 60 Ry 截断的结果反而更好, 说明这部分元素采用 100 Ry 截断会具有轻微过拟合的特征。图中的“ME”数据点显示, 两个截断能量下, 所有元素的 DeltaDFT(LCAO,PW) 的平均值几乎相同。因此, 综合评估, 在 60–100 Ry 范围内, 参考态能量截断对轨道质量的影响较小。

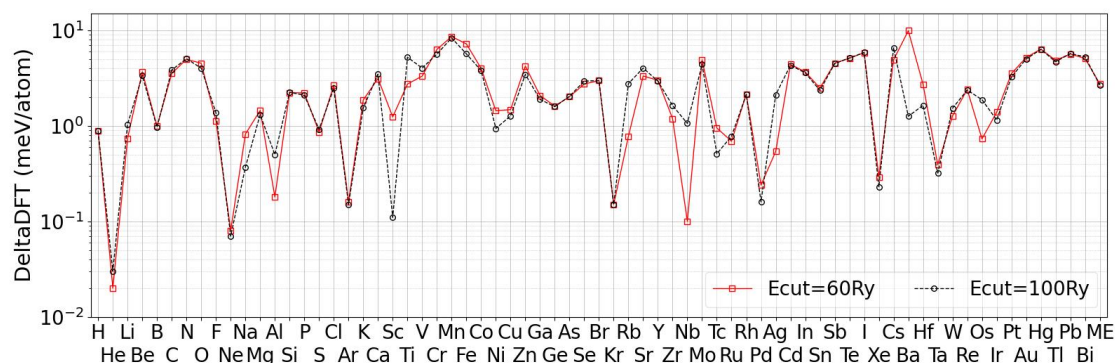


图 5 CGH 算法下使用不同平面波截断能量对 DeltaDFT(LCAO,PW)的影响

Figure 5 Effect of different plane-wave cutoff energies on DeltaDFT (LCAO,PW) under the CGH algorithm

(3) 内层满电子态的极化轨道对 LCAO 计算精度的影响

图 6 中绘制了 SG15 势下, Ga、Ge、In、Sn、Sb、Te、I、Xe、Ba、Hf、Ta、W、Tl、Pb、Bi 元素的 DZ、DZP、TZP、TZDP 轨道的 DeltaDFT(LCAO,PW) 差异, 该数值同样全部选择 Rcut 为 10Bohr 的轨道进行计算, 其精度更高, 可以更好地分辨出轨道规模差异的影响。这部分元素的 Z 轨道中的最高角量子数轨道被内层电子全满占据, 其被极化的程度可能更弱, 所以在本工作的生成初期减除了一层极化轨道 (P 轨道), 即使用了 DZ 轨道代替了 DZP 轨道, TZP 轨道代替了 TZDP 轨道, 以此创建了 MixDZP 与 MixTZDP 轨道数据集。根据表 6 所示, 在多数情况下 DZP 相比 DZ 具有明显的更小的 DeltaDFT, 同时 TZDP 相比 TZP 轨道的精度提升较小, 这也可以从平均值“ME 数据点”处得到明显结论。因此, 整体上, 建议至少保留一个 P 轨道。

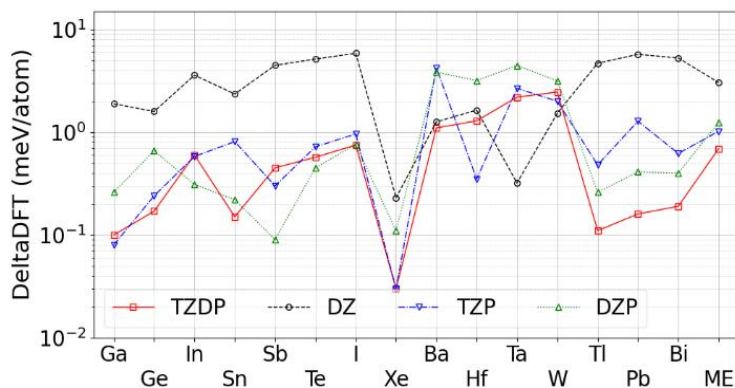


图 6 CGH 算法下内层满电子态的极化轨道的影响

Figure 6 Effect of polarization orbital on inner fully occupied states under the CGH algorithm

(4) 评估 LRH 轨道与 CGH 轨道的精度差异

图7展示了不同截断半径、不同基组大小的LRH或CGH算法优化的轨道的DeltaDFT(LCAO,PW)精度,所有轨道都基于SG15赝势来生成。图7中a、b、c三张图的整体DeltaDFT精度依次提升,对应的轨道,先由标准截断半径提升为10 Bohr截断半径,而后基组大小由DZP增加到TZDP。可以看到,LRH轨道相比CGH轨道的精度整体上有提升,但是在精度更低的图7a中表现更加明显。图7a中的LRH/CGH轨道平均精度值(ME)分别为1.68/3.55 (meV/atom);图7b中的LRH、CGH轨道ME值分别为1.12/2.27 (meV/atom);图7c中的LRH/CGH轨道ME值分别为0.66/1.01 (meV/atom)。

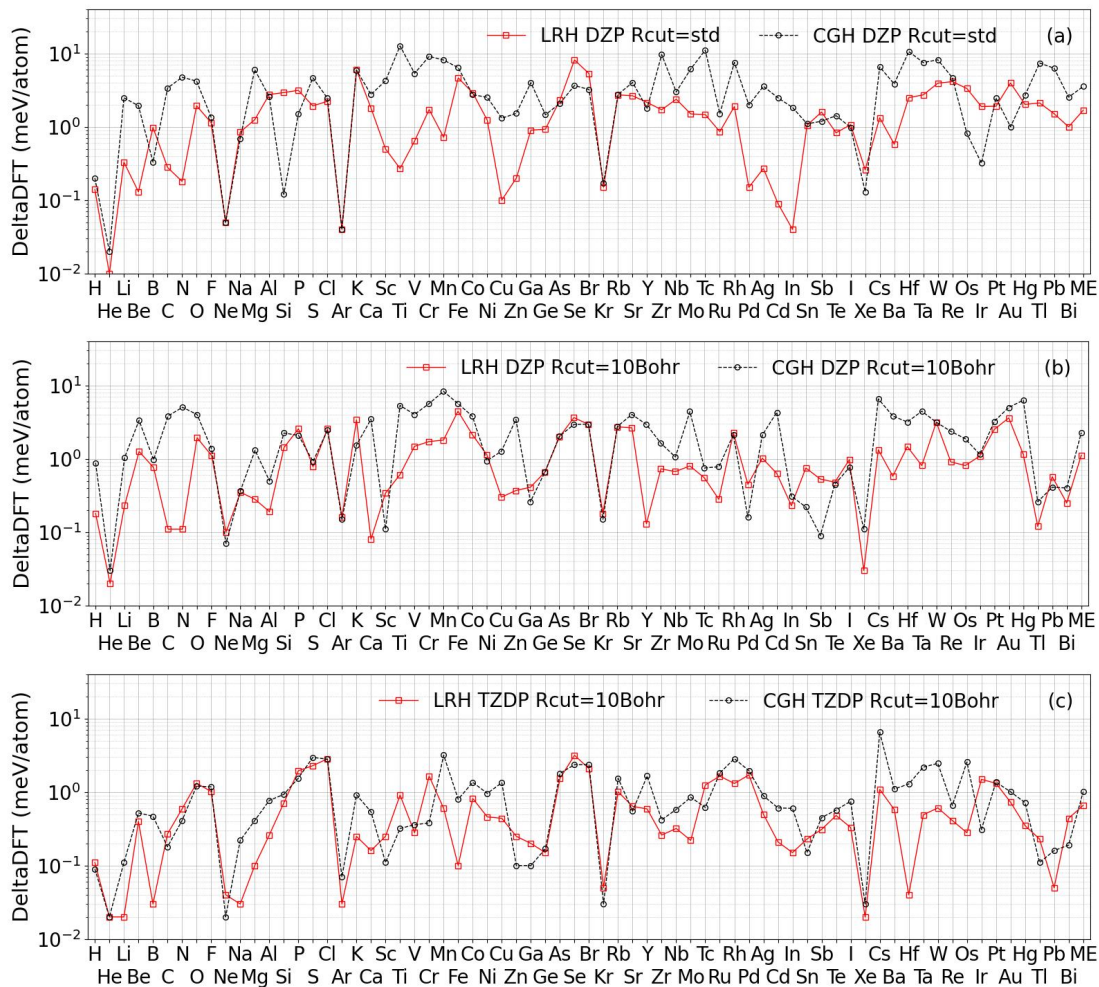


图7 基于不同截断半径、DZP或TZDP类型、LRH或CGH算法的轨道的DeltaDFT(LCAO,PW)

Figure 7 DeltaDFT (LCAO,PW) results for orbitals based on different Rcut, DZP or TZDP basis sets, and LRH or CGH algorithms

因此,LRH轨道相比CGH轨道在整体上有显著的提高,特别是对低截断半径的小基组轨道的优化提升更加明显,可以使DeltaDFT的平均值降低50%以上,其对高效率的ABACUS计算有更加明显的助益。

(5) 评估基于相同轨道但不同波函数能量截断的LCAO计算的精度差异

在使用ABACUS与原子轨道进行LCAO第一性原理模拟时,仍然需要指定波函数能量截断参数(ecutwfc),用于快速傅里叶变换。为测试原子轨道对不同截断能量的适应性,图8中展示了基

于 LRH 轨道并使用不同波函数能量截断进行计算的 DeltaDFT(LCAO, PW)值。在本文的数据库中, 也给出了基于标准 Rcut 轨道的评估结果, 但在图 8 中, 仅展示了 Rcut 为 10 Bohr 的计算结果, 以避免半径截断精度较低影响对波函数能量截断参数的评估。

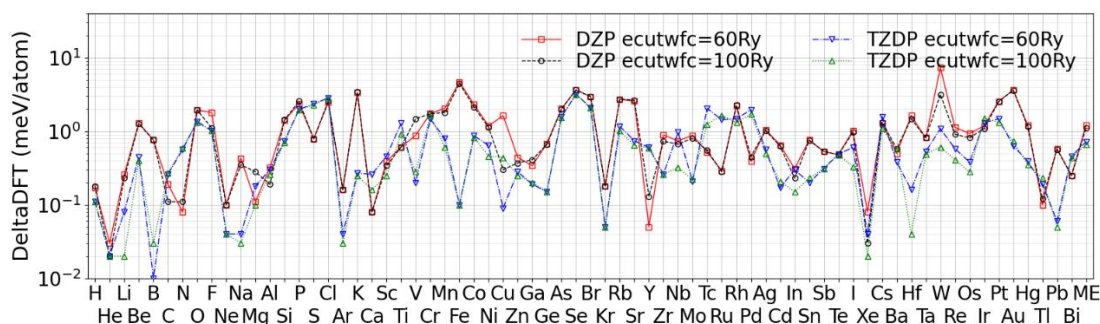


图 8 采用不同 $ecutwfc$ 截断进行 LCAO 计算时的 DeltaDFT (LCAO, PW)

Figure 8 DeltaDFT (LCAO, PW) results for LCAO calculations with different $ecutwfc$ cutoffs

图 8 中的计算所使用的轨道是基于 100 Ry 能量截断的平面波基组参考态与 LRH 优化算法所生成的 DZP、TZDP 轨道。对 DZP 轨道, 使用不同的 $ecutwfc$ 的 ABACUS 计算均给出了几乎相同的 DeltaDFT 数值, 所有元素的平均值 ME 由 60 Ry 时的 1.22 meV/atom 略微降低至 1.12 meV/atom。而对更精确的 TZDP 轨道, 依然变化不大, 但更多的情况下, 使用 $ecutwfc = 100$ Ry 可以给出稍好的结果, 相应的平均值 ME 从 0.72 meV/atom 降低至 0.66 meV/atom。如比较标准 Rcut 截断下的差异, 使用 TZDP 轨道时, $ecutwfc$ 的影响会更大一些。

(6) 评估基于三种不同赝势所构建的原子轨道的精度

图 9 中的计算均使用 10 Bohr 作为截断半径, 如需要标准截断半径下的结果, 可以在数据库中查询。图 9 显示, 三组赝势下基于更精确的 LRH 算法优化的 DZP/TZDP 轨道各有优劣, 从平均值“ME 数据点”的结果可知, 使用 DZP 轨道时, Dojo-NC-FR 赝势与 SG15 赝势的整体表现更优, 使用 TZDP 轨道时, 三者的平均精度相差无几。

本节仅基于 DeltaDFT 测试给出了若干有价值的评估结果, 并进行了简单的分析, 更多的数据, 如每个轨道对应的 Spillage 数值、具体的轨道形状、VE 曲线的形状与能量差异绝对值以及二原子或三原子参考结构下的 PW/LCAO 计算结果的对比等等, 可以见于数据库中, 以更全面地分析具体轨道的质量。在本文关联数据集的构建过程中, 针对 DZP 与 TZDP 类型的轨道, 设定了最小可接受的 DeltaDFT 指标数值为 30 meV/atom, 并逐个检查相关轨道的 VE 曲线的平滑性, 剔除了偏离抛物线曲线较多且具有震荡行为的对应原子轨道, 使轨道数据具有更好的可迁移性表现。

4 数据价值

本节展示基于本数据集中最精确的 DZP/TZDP 轨道的 ABACUS 计算在国内外其他相似计算中的精度水平, 以验证本数据的价值。本节的对比指标是基于各个原子轨道的 LCAO 计算与 WIEN2k 计算结果 (行业参考基准) 的 DeltaDFT (LCAO, WIEN2k) 数据。一个对比对象是行业中同样基于赝价电子轨道 (也称赝原子轨道) 的 OpenMX 计算^[22], 另一个对比对象是基于全电子轨道的 FPLO 计算^[23], 后者更精确但轨道数更多, 因此所需计算量更大。

DeltaDFT (LCAO, WIEN2k) 数据如图 10 所示, 全电子轨道明显优于其他赝价电子轨道。对比三

种廉价电子轨道的计算精度，DZP 原子轨道基组整体上较 OpenMX 给出的原子轨道基组计算结果精度更低，而 TZDP 原子轨道基组的精度更高。

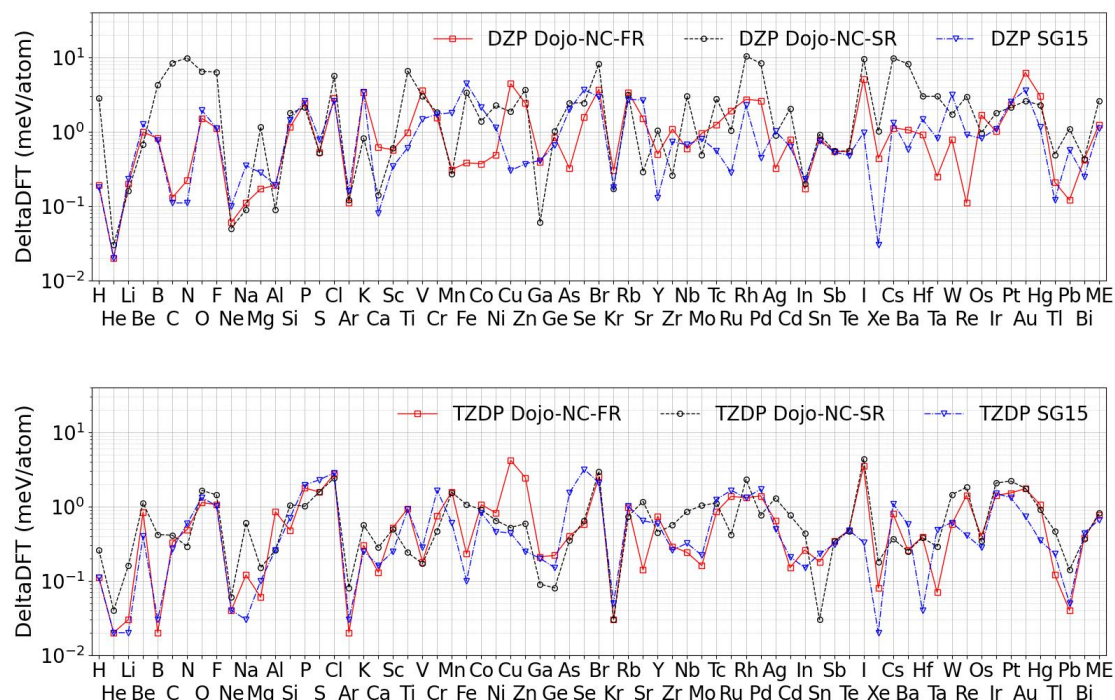


图9 三组赝势下的 DZP 与 TZDP 轨道的 DeltaDFT(LCAO,PW) 结果

Figure 9 DeltaDFT (LCAO, PW) results for DZP and TZDP orbitals under three pseudopotentials

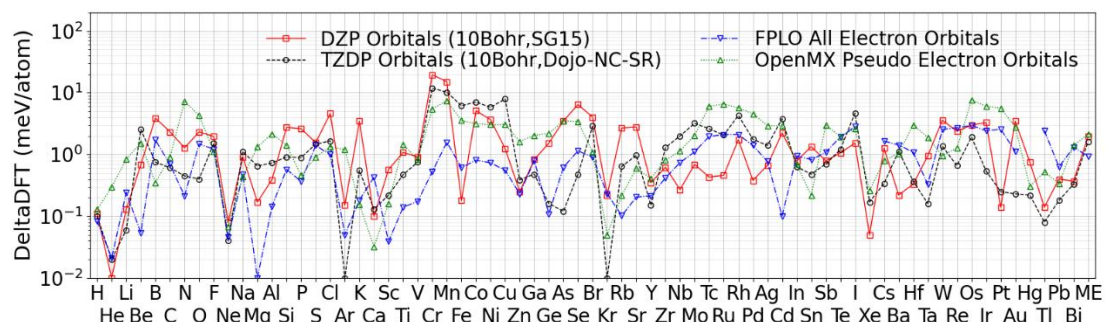


图10 基于本数据集的 LCAO 计算与行业其他 LCAO 计算的精算对比

Figure 10 Accuracy comparison of LCAO calculations based on this dataset and other LCAO calculations

表6中给出了基于本次发布的各个原子轨道基组数据的 LCAO 计算结果的 DeltaDFT 平均值，两列数据分别为 Rcut 取 10 Bohr 与标准半径 Rstd 时的结果，标准半径 Rstd 的具体数值可以见于数据集集中的 json 格式的文件中。

表6 基于原子轨道基组的 DeltaDFT(LCAO,WIEN2k) 平均值 (单位 meV/atom)

Table 6 Average DeltaDFT (LCAO-WIEN2k) values based on atomic orbital basis sets (Unit: meV/atom)

赝势, 轨道算法, 基组大小类型	DeltaDFT (Rcut = 10Bohr)	DeltaDFT (Rcut = Rstd)
Dojo-NC-FR, LRH, DZP	2.88	2.53
Dojo-NC-FR, LRH, TZDP	2.78	2.89

赝势, 轨道算法, 基组大小类型	DeltaDFT (Rcut = 10Bohr)	DeltaDFT (Rcut = Rstd)
Dojo-NC-SR, LRH, DZP	3.09	1.92
Dojo-NC-SR, LRH, TZDP	1.61	5.2
SG15, LRH, TZDP	1.72	2.5
SG15, CGH, MixDZP	2.85	3.69
SG15, CGH, MixTZDP	1.99	2.57
SG15, CGH, ZP	15.13	3.51
SG15, CGH, DZP	2.4	2.57
SG15, CGH, DZ	102.71	25.15
SG15, CGH, TZDP	1.91	16.69
SG15, CGH, DZP	1.95	1.91
LCAO High Orbitals from ATK ^[24]	1.8	(use Rcut from ATK)
Orbitals from OpenMX ^[24]	2.12	(use Rcut from OpenMX)
Orbitals from FPLO ^[24]	0.93	(use Rcut for FPLP)

在 DeltaDFT 数据汇总网站^[24]中, 基于赝价电子轨道的最好结果来自 ATK 软件提供的 LCAO High 轨道的 1.8 meV/atom (即表 6 中的倒数第三行结果)。本文给出的“Dojo-NC-SR, LRH, TZDP”“SG15, LRH, TZDP”两套轨道均获得了更精确的 DeltaDFT(LCAO, WIEN2k)结果, 分别为 1.61 meV/atom、1.72 meV/atom, 实现了当前已知的最高精度。因此, 本文给出的原子轨道基组数据, 不仅是 ABACUS 软件的必要计算基础, 具有很强的复用价值, 同时也为行业贡献了最高精度水平的数据, 其他使用数值原子轨道基组的第一性原理软件也可以复用本原子轨道数据 (一般需要进行读取格式兼容性开发)。

5 数据使用方法和建议

本文中给出的原子轨道数据适合关联生成其赝势来使用, 在交叉使用其他赝势时, 其精度与迁移适应能力不被保证。在使用本数据集时, 建议用户基于质量评估一节中提到的各个指标进行具体分析、具体选择。此外, 也建议用户针对自身的计算目标结构, 进行收敛性测试, 以选择最高效的原子轨道基组, 实现快速且精准的计算。

具体在 ABACUS 软件中使用本数据库数据的方法为: 先将希望使用的原子轨道文件放置在计算程序可访问的某目录下, 并将文件路径放置在“STRU”输入文件的“NUMERICAL_ORBITAL”文本行的下一行 (当原子轨道文件位于 ABACUS 程序的运行目录下时, 只需输入文件名), 每个元素的轨道路径需要依次逐行填写, 顺序与“ATOMIC_SPECIES”中的元素行保持一致, 如下文本所示:

```
ATOMIC_SPECIES
Mg  24.305      Mg_ONCV_PBE-1.0.upf
O   15.9994     O_ONCV_PBE-1.0.upf
NUMERICAL_ORBITAL
Mg_gga_9au_100Ry_4s2p1d.orb
O_gga_7au_100Ry_2s2p1d.orb
```

详细用法见 ABACUS 软件官网^[13]的“RESOURCES-MANUAL”栏目。

作者分工职责

张文帅（1987—），男，江苏徐州人，博士，高级工程师，研究方向为超算集群、数值计算与科学数据相关技术。主要承担工作：数据库的构建、流程实现与评测优化。

林霈泽（1992—），男，广东汕头人，博士，副研究员，研究方向为第一性原理计算方法的研究、开发与应用。主要承担工作：数据构建算法的设计与开发。

陈默涵（1985—），男，福建泉州人，博士，研究员，研究方向为密度泛函理论与分子动力学方法与应用。主要承担工作：数据构建算法的设计与开发。

何力新（1971—），男，湖南长沙人，博士，教授，研究方向为计算凝聚态物理与量子多体问题等。主要承担工作：数据库的构建算法设计与评测优化。

参考文献

- [1] KOHN W, SHAM L J. Self-consistent equations including exchange and correlation effects[J]. *Physical Review*, 1965, 140(4A): A1133 – A1138. DOI:10.1103/physrev.140.a1133.
- [2] KOHN W. Density functional and density matrix method scaling linearly with the number of atoms[J]. *Physical Review Letters*, 1996, 76(17): 3168 – 3171. DOI:10.1103/PhysRevLett.76.3168.
- [3] SÁNCHEZ-PORTAL D, ORDEJÓN P, ARTACHO E, et al. Density-functional method for very large systems with LCAO basis sets[J]. *International Journal of Quantum Chemistry*, 1997, 65(5): 453 – 461. DOI:10.1002/(SICI)1097-461X(1997)65:5<453::AID-QUA9>3.0.CO;2-V.
- [4] LIN L, CHEN M H, YANG C, et al. Accelerating atomic orbital-based electronic structure calculation *via* pole expansion and selected inversion[J]. *Journal of Physics: Condensed Matter*, 2013, 25(29): 295501. DOI:10.1088/0953-8984/25/29/295501.
- [5] BAKER J S, MIYAZAKI T, BOWLER D R. The pseudoatomic orbital basis: electronic accuracy and soft-mode distortions in ABO₃ perovskites[J]. *Electronic Structure*, 2020, 2(2): 025002. DOI:10.1088/2516-1075/ab950e.
- [6] TORRALBA A S, TODOROVIC M, BRÁZDOVÁ V, et al. Pseudo-atomic orbitals as basis sets for the[J]. *Journal of Physics: Condensed Matter*, 2008, 20(29): 294206. DOI:10.1088/0953-8984/20/29/294206.
- [7] LEJAEGHERE K, BIHLMAYER G, BJÖRKMAN T, et al. Reproducibility in density functional theory calculations of solids[J]. *Science*, 2016, 351(6280): aad3000. DOI:10.1126/science.aad3000.
- [8] CHEN M H, GUO G C, HE L X. Systematically improvable optimized atomic basis sets for *ab initio* calculations[J]. *Journal of Physics: Condensed Matter*, 2010, 22(44): 445501. DOI:10.1088/0953-8984/22/44/445501.
- [9] CHEN M H, GUO G C, HE L X. Electronic structure interpolation *via* atomic orbitals[J]. *Journal of Physics Condensed Matter*, 2011, 23(32): 325501. DOI:10.1088/0953-8984/23/32/325501.
- [10] LIN P Z, REN X G, HE L X. Strategy for constructing compact numerical atomic orbital basis sets by

- incorporating the gradients of reference wavefunctions[J]. *Physical Review B*, 2021, 103(23): 235131. DOI:10.1103/physrevb.103.235131.
- [11] LI P F, LIU X H, CHEN M H, et al. Large-scale *ab initio* simulations based on systematically improvable atomic basis[J]. *Computational Materials Science*, 2016, 112: 503 – 517. DOI:10.1016/j.commatsci.2015.07.004.
- [12] LIN P Z, REN X G, LIU X H, et al. Ab initio electronic structure calculations based on numerical atomic orbitals: Basic formalisms and recent progresses[J]. *WIREs Computational Molecular Science*, 2024, 14(1): e1687. DOI:10.1002/wcms.1687.
- [13] ABACUS official website [EB/OL]. (2016-12-19) [2024-12-31]. <https://abacus.ustc.edu.cn>.
- [14] First edition of the SIAB (Systematically Improvable Atomic orbital Basis generator) software [EB/OL]. (2016-12-19) [2024-12-31]. <https://github.com/abacusmodeling/abacus-develop/tree/develop/tools/SIAB>.
- [15] Second edition of the SIAB (Systematically Improvable Atomic orbital Basis generator) software [EB/OL]. (2022-08-31) [2024-12-31]. <https://github.com/abacusmodeling/ABACUS-orbitals/tree/main/SIAB>.
- [16] The SG15 Optimized Norm-Conserving Vanderbilt (ONCV) pseudopotentials v1.0 [EB/OL]. (2016-12-19) [2024-12-31]. http://www.quantum-simulation.org/potentials/sg15_oncv/.
- [17] SCHLIPF M, GYGI F. Optimization algorithm for the generation of ONCV pseudopotentials[J]. *Computer Physics Communications*, 2015, 196: 36 – 44. DOI:10.1016/j.cpc.2015.05.011.
- [18] Dojo pseudopotential official website [EB/OL]. (2016-12-19) [2024-12-31]. <https://www.pseudo-dojo.org/>.
- [19] Official documentation of Unified Pseudopotential Format (UPF) [EB/OL]. (2016-12-19) [2024-12-31]. <https://pseudopotentials.quantum-espresso.org/home/unified-pseudopotential-format>.
- [20] LEJAEGER K, VAN S, VAN O, et al. Error estimates for solid-state density-functional theory predictions: an overview by means of the ground-state elemental crystals[J]. *Critical Reviews in Solid State and Materials Sciences*, 2014, 39(1): 1 – 24. DOI:10.1080/10408436.2013.772503.
- [21] BLAHA P, SCHWARZ K, TRAN F, et al. WIEN2k: an APW+lo program for calculating the properties of solids[J]. *The Journal of Chemical Physics*, 2020, 152(7): 074101. DOI:10.1063/1.5143061.
- [22] OZAKI T. Variationally optimized atomic orbitals for large-scale electronic structures[J]. *Physical Review B*, 2003, 67(15): 155108. DOI:10.1103/physrevb.67.155108.
- [23] KOEPERNIK K, ESCHRIG H. Full-potential nonorthogonal local-orbital minimum-basis band-structure scheme[J]. *Physical Review B*, 1999, 59(3): 1743 – 1757. DOI:10.1103/physrevb.59.1743.
- [24] Comparing Solid State DFT Codes, Basis Sets and Potentials [EB/OL]. (2016-3-25) [2024-12-31]. <https://molmod.ugent.be/deltacodesdft>.

论文引用格式

张文帅, 林霏泽, 陈默涵, 等. 适用于高效第一性原理计算的数值原子轨道数据[J/OL]. *中国科学数据*, 2026, 11(1). (2025-12-29). DOI: 10.11922/11-6035.csd.2025.0034.zh.

数据引用格式

张文帅, 林霏泽, 陈默涵, 等. 适用于高效第一性原理计算的数值原子轨道数据[DS/OL]. V1. Science Data Bank, 2025. (2025-04-27). DOI: 10.57760/sciencedb.j00001.01295.

A dataset of high-efficiency atomic orbital data for first-principles calculations of large atomic systems

ZHANG Wenshuai¹, LIN Peize², CHEN Mohan³, HE Lixin^{4*}

1. Scientific Data Center, Network Information Center, Public Experimental Center, University of Science and Technology of China, Hefei 230026, P.R. China
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230026, P.R. China
3. HEDPS, CAPT, College of Engineering, Peking University, Beijing 100871, P.R. China
4. Laboratory of Quantum Information, University of Science and Technology of China, Hefei, 230026, Anhui, P.R. China

*Email: helx@ustc.edu.cn

Abstract: Numerical Atomic Orbitals (NAOs) are a highly efficient basis set for describing electron wave functions in atomic structures, particularly suitable for electronic structure simulations of large-scale atomic systems (thousands of atoms or more). Due to the complexity of constructing and testing such data, the research community still lacks data products that provide both comprehensive quality control and broad applicability. To address this, this work systematically constructed a NAO dataset covering 68 elements, supporting various pseudopotentials, cutoff parameters (radius/energy), and basis set size configurations. During the construction of the NAO dataset, a hierarchical optimization strategy was adopted: low-precision orbitals with fewer basis functions were optimized first, followed by a gradual progression to high-precision orbitals with more basis functions, thereby meeting the requirements of first-principles calculations at different levels of efficiency and accuracy. Moreover, to demonstrate the reliability of the data and provide a basis for user selection, this dataset includes comprehensive quality assessment data, such as single-crystal equations of state (EOS) fitting data and bond length-energy data for molecular systems. Assessment results show that the mean energy difference between calculations using this NAO dataset and standard plane-wave basis sets for single crystals is minimal, reaching as low as 0.66 meV/atom. This dataset has been adopted by the ABACUS first-principles computational software, significantly enhancing the efficiency and accuracy of complex system simulations. It has broad application value in fields such as condensed matter physics, materials science, and chemical simulations.

Keywords: first-principles calculations; Atomic orbitals; Numerical calculations; Scientific Data

Dataset Profile

Title	A dataset of high-efficiency atomic orbital data for first-principles calculations of large atomic systems
Datacorresponding author	HE Lixin (helx@ustc.edu.cn)
Data authors	ZHANG Wenshuai, LIN Peize, CHEN Mohan, HE Lixin
Time range	2016–2024
Geographical scope	Global (unspecified)
Magnitude of Energy Precision	1 meV/atom
Data volume	1.2 GB (before ZIP compression)
Data format	*.orb, *.dat, *.txt, *.png, *.pdf
Data service system	https://doi.org/10.57760/sciencedb.j00001.01295
Sources of funding	The National Natural Science Foundation of China (Grant No. 12105283); the CAS Special Fund for Network Security and Informatization (Grant No. CAS-WX2022SDC-SJ13); the National Key Research and Development Program (2016YFB0201202).
Dataset composition	The dataset contains 26,648 data files (excluding the root directory's README.md and referenced data). To improve transfer efficiency, some folders are distributed as compressed archives, though all descriptions in this table remain based on the original uncompressed file structure. All files are categorized into two types: orbital data (including *.orb files, *.txt files, and input files INPUT/SIAB_INPUT for the construction workflow) and quality assessment data (located in folders prefixed with Test/Plot, containing *.dat, *.png, and *.pdf files). The first category of files can be distinguished by their directory paths, corresponding to three types of pseudopotential (as named in root directories), different optimization algorithms (CGH/LRH), basis set size configurations (SZ/DZP/TZDP, etc.), energy cutoff parameters (*Ry), cutoff radii (*au), and 68 chemical elements. The second category includes test result files for standard plane-wave basis sets or atomic orbital basis sets (located in folders suffixed with PW or LCAO) , as well as statistical result files stored in the Test_Summary folder.