# EDBench: Large-Scale Electron Density Data for Molecular Modeling


Hongxin Xiang[1,5], Ke Li[2], Mingquan Liu[1], Zhixiang Cheng[1], Bin Yao[3], Wenjie Du[4], Jun Xia[5,6], Li Zeng[1], Xin Jin[7†], Xiangxiang Zeng[1†]

[1]College of Computer Science and Electronic Engineering, Hunan University [2]East China Normal University [3]College of Materials Science and Engineering, Hunan University [4]University of Science and Technology of China [5]The Hong Kong University of Science and Technology (Guangzhou) [6]The Hong Kong University of Science and Technology [7]Eastern Institute of Technology


## Motivation

**Critical Gap in Existing Molecular Machine Learning Force Fields (MLFFs)**

- MLFFs **focus on learning many-body interactions at the atomic level**, including one-body (atomic attributes), two-body (interatomic distances), three-body (bond angle), four-body (torsions and improper torsions), and five-body interactions.
- They largely **ignore the pivotal role of Electron Density (ED)**, which is the fundamental quantity determining all ground-state properties according to the Hohenberg-Kohn theorem.

**Key Challenge To Advance MLFFs Toward An Electron-Level Understanding**

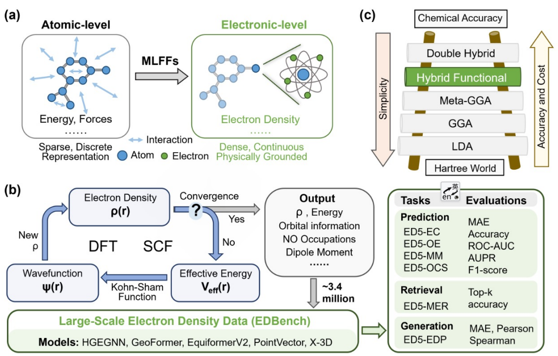
- ***The lack of large-scale, high-quality ED datasets***, which are essential for pretraining and could fundamentally reshape the paradigm of MLFFs modeling.
- ***The absence of an ED-centric benchmark*** to systematically explore the feasibility and effectiveness of ED-based modeling frameworks.

Comparison of various databases in quantum chemistry

| Category | Ours | Classical quantum chemistry databases and extensions | | | | | Molecular dynamics | | Pharmaceutical | | Material | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Datasets | EDBench | QM7-X | QM9 | QH9 | QM9-VASP | PubChem QC | WS22 | MD17 | QMugs | $\nabla^2$DFT | QMMD | ECD | MP |
| Source | PCQM4Mv2 | GDB-13 | GDB-17 | QM9 | QM9 | PubChem | - | - | CHEME | MOSES | ICSD | Magten | ICSD |
| Molecules | 3,359,472 | 7K | 134K | 130K | 134K | 85M | 10 | 8 | 665K | 1.9M | 1.2M | 140K | 577K |
| Element Count | 22 | 6 | 5 | 5 | 5 | 12 | 4 | 4 | 11 | 8 | - | - | - |
| Calc Method (Basis Set/ XC-Function) | B3LYP, 6-31G**/+G** | PBE0+ MBD | B3LYP, 6-31G (2df,P) +G4MP2 | B3LYP, Def2-SVP | PBE | B3LYP/6-31G* //PM6 | PBE0/ 6-31G* | PBE+ vdW-TS | GFN2-x TB+ωB97 X-D/def 2-SVP | ΩB97X-D/Def2-SVP | PAW-PBE | PBE/HSE 06,GGA+ U/Monkhorst-Pack | PBE,GGA /GGA+U, SCAN/R2 SCAN |
| Electron density $\rho$ | ✓-CUBE | × | × | × | ✓-CHGCAR | × | × | × | × | × | × | ✓-CHGCAR | ✓(122K)-CHGCAR |
| Total Energy | ✓ | ✓ | × | × | × | ✓ | × | ✓ | ✓ | ✓ | ✓ | × | ✓ |
| NO Occupation | ✓ | × | × | × | × | × | × | × | × | × | × | × | × |
| Canonical Orbit | ✓ | ✓ | ✓ | × | × | ✓ | ✓ | × | ✓ | ✓ | × | × | × |
| Dipole Moment | ✓ | ✓ | ✓ | × | × | ✓ | ✓ | × | ✓ | ✓ | × | × | × |
| E-Structure | ✓ | × | × | × | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| Software/Tool | Psi4 | FHI-aims | Gaussian | PySCF | VASP | GAMESS | Gaussian | ORCA/ FHI-aims | Psi4 | Psi4 | VASP | VASP | VASP |

## Overview of EDBench

- We introduce EDBench, **a large-scale, high-quality dataset of Electron Density for 3.3 million molecules**, built upon PCQM4Mv2.
- We design a **comprehensive benchmark suite** with ED-centric tasks (prediction, retrieval, generation) to rigorously evaluate model capabilities.
- We show that learning-based methods can **calculate ED with comparable precision** while **reducing the computational cost** relative to traditional DFT.




The overview of EDBench

## The Designed Benchmark

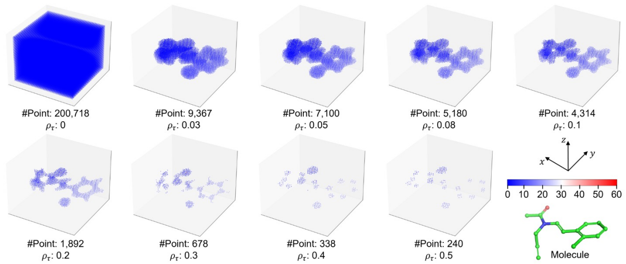
We define a suite of tasks based on both molecular structures (MS) and ED, focusing on three fundamental capabilities:

- **Prediction** of quantum property: ED5-EC, ED5-OE, ED5-MM, ED5-OCS
- **Retrieval** between MS and ED: ED5-MER
- **Generation** of ED based on MS: ED5-EDP

Statistical information of 6 designed benchmarks with a scaffold split

| Datasets | #Mol | #Train/#Valid/#Test | #Task | Task type | Task desc |
|---|---|---|---|---|---|
| ED5-EC | 47,986 | 38,388/4,799/4,799 | 6 | Regression | 6 energy components (DF-RKS Final Energy [E1], Nuclear Repulsion Energy [E2], One-Electron Energy [E3], Two-Electron Energy [E4], DFT Exchange-Correlation Energy [E5], Total Energy [E6]) |
| ED5-OE | 43,510 | 34,808/4,351/4,351 | 7 | Regression | 7 orbital energies (HOMO-2, HOMO-1, HOMO-0, LUMO+0, LUMO+1, LUMO+2, LUMO+3) |
| ED5-MM | 49,917 | 39,933/4,992/4,992 | 4 | Regression | 4 multipole moment (3 Dipoles {X, Y, Z}, Magnitude) |
| ED5-OCS | 50,000 | 40,000/5,000/5,000 | 1 | Classification | open-/closed-shell classification |
| ED5-MER | 50,000 | 40,000/5,000/5,000 | 2 | Retrieval | cross-modal retrieval between molecular structures and ED |
| ED5-EDP | 50,000 | 40,000/5,000/5,000 | 1 | Generation | ED prediction from molecular structures |

## Visualization of ED




## Experiments

### Results of EDBench on the Prediction of Quantum Property

- **ED5-EC**

The MAE performance on 6 energies from the ED5-EC dataset with $\rho_\tau$ = 0.05

| | E1 | E2 | E3 | E4 | E5 | E6 |
|---|---|---|---|---|---|---|
| PointVector | 243.49±74.72 | 325.65±160.17 | 858.77±496.74 | 389.24±217.51 | 17.54±10.85 | 243.49±74.73 |
| X-3D | **190.77±1.98** | **109.21±2.82** | **369.88±1.34** | **150.05±0.27** | **8.13±0.51** | **190.77±1.98** |

- **ED5-OE**

The performance of MAE×100 on 7 orbital energies of the ED5-OE with $\rho_\tau$ = 0.05

| | HOMO-2 | HOMO-1 | HOMO-0 | LUMO+0 | LUMO+1 | LUMO+2 | LUMO+3 |
|---|---|---|---|---|---|---|---|
| PointVector | **1.73±0.01** | **1.68±0.01** | **1.92±0.01** | **3.08±0.05** | **2.86±0.05** | **3.05±0.02** | **3.01±0.02** |
| X-3D | 1.75±0.02 | 1.72±0.02 | 1.98±0.00 | 3.21±0.01 | 3.02±0.02 | 3.25±0.04 | 3.20±0.03 |

- **ED5-MM**

The MAE performance on multipole moments from the ED5-MM dataset with $\rho_\tau$ = 0.05

| | Dipole X | Dipole Y | Dipole Z | Magnitude |
|---|---|---|---|---|
| PointVector | 0.9123±0.0203 | 0.9605±0.0053 | 0.754±0.0068 | 0.7397±0.0467 |
| X-3D | **0.8818±0.0010** | **0.9427±0.0008** | **0.7416±.0.0023** | **0.6820±0.0005** |

- **ED5-OCS**

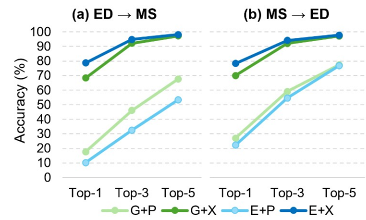
The performance (%) of open/closed-shell prediction on the ED5-OCS dataset with $\rho_\tau$ = 0.05

| | Accuracy | ROC-AUC | AUPR | F1-Score |
|---|---|---|---|---|
| PointVector | 55.57±2.14 | 55.97±5.17 | 57.62±3.91 | **66.96±2.08** |
| X-3D | **57.65±0.18** | **60.48±0.38** | **61.54±0.31** | 61.41±1.02 |

Overall, results **validate the effectiveness of using ED as a model input** and demonstrate its **utility in capturing physically meaningful patterns**.

### Results of EDBench on Retrieval Tasks

- **ED5-MER**




The retrieval performance on ED5-MER

The strong performance of E+P and E+X demonstrates their **potential for ED-based virtual screening and MS-based electronic-level molecular understanding**.

### Results of EDBench on Generation Task

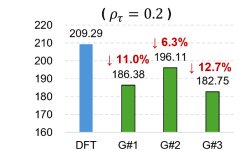
- **ED5-EDP**

The performance of HGEGNN on the ED generation of the ED5-EDP dataset. The unit of Time is second/mol.

| | $\rho_\tau$ | MAE | Pearson (%) | Spearman (%) | Time |
|---|---|---|---|---|---|
| HGEGNN | 0.1 | 0.3362±0.2900 | 81.0±8.1 | 56.4±13.7 | 0.024 |
| | 0.15 | 0.0463±0.0157 | 98.0±6.3 | 87.0±2.7 | 0.015 |
| | 0.2 | 0.0448±0.0133 | 99.2±0.8 | 91.0±9.1 | 0.013 |
| DFT | - | - | - | - | 245.8 |

The results show that the HGEGNN (i) **accurately and efficiently predicts ED**, (ii) **successfully captures key chemical features** in high-density regions, and (iii) **offers a powerful alternative to costly DFT calculations**.

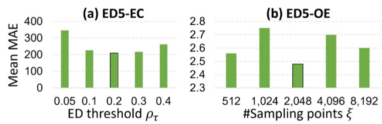
### Quality analysis of ED outputs from the generation task




The average MAE of PointVector on ED5-EC generated by DFT and G#{1,2,3}

Model-generated ED produces superior downstream performance than DFT-calculated ED. This demonstrates that **HGEGNN can create high-quality, machine-learning-friendly ED data** for advancing molecular force field models.

### Ablation study on thresholds and sampling points




Ablation results of PointVector on (a) different ED thresholds $\rho_\tau$ and (b) different numbers of sampling points $\xi$

Lower Threshold or More Points ≠ Better Performance. Striking a balance between accuracy and cost through hyperparameter tuning is key.

### References


[1] Xiang H, Li K, Liu M, Cheng Z, et al. EDBench: Large-Scale Electron Density Data for Molecular Modeling[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[2] Xiang H, Xia J, Jin X, et al. Electron density-enhanced molecular geometry learning[C]//Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. 2025: 7840-7848.