EDBench: Large-Scale Electron Density Data for Molecular Modeling

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose EDBench, a large-scale, high-quality dataset of electron density (ED) designed to advance learning-based research at the electronic scale. EDBench comprises 3,359,472 drug-like molecules with corresponding ED distributions and a comprehensive set of quantum chemical properties, including energy components, orbital energies, and multipole moments, thus providing a solid foundation for systematically investigating the role of ED in molecular modeling. We outline the AI tasks, data rationale, acceleration potential, and concrete pathway to data-creation with cost and scalability of EDBench.

1 AI task definition

EDBench is purpose-built to support three tightly-coupled scientific tasks that together push machine-learning models from the atomic scale to the electronic scale (The methods and significance of these tasks see Appendix A.1):

- **Quantum property prediction**: regress or classify ground-state quantum properties (energy components, orbital energies, multipole moments, open/closed-shell character) directly from the 3D electron density.
- **Cross-modal retrieval**: Retrieve a molecule's structure from its density and vice-versa, enabling electron-level virtual screening and inverse design.
- **ED generation**: Electron density (ED) prediction from molecular structures, aimed at approximating DFT-level density accuracy at significantly reduced computational cost.

2 Dataset rationale

The rapid integration of deep learning into molecular-dynamics (MD) simulations has established machine-learning force fields (MLFFs) as efficient and promising computational tools across physics, chemistry, biology, and materials science [1, 2, 3]. Nevertheless, prevailing MLFFs emphasize atom-level many-body interactions[4, 5], largely overlooking the pivotal role of microscopic electron distribution in governing interatomic forces[6, 7, 8]. Electron density (ED), as a fundamental physical quantity in quantum mechanics that describes the spatial distribution of electrons, offers a more fine-grained and physically grounded representation of molecular systems according to Hohenberg–Kohn theorem [9]. Explicit incorporation of ED into MLFFs is therefore expected to bridge the gap between microscopic electronic behavior and macroscopic force fields, enhancing both accuracy and generalizability.

Advancing MLFFs toward electron-level modeling confronts two principal challenges: (i) the absence of large-scale, high-quality ED datasets essential for pre-training and potentially paradigm-shifting architectures, and (ii) the lack of an ED-centric benchmark for systematically evaluating the feasi-

Submitted to AI for Science workshop (NeurIPS 2025).

bility and efficacy of ED-based frameworks. ED data can be acquired experimentally (e.g., X-ray diffraction)[10, 11, 12] or theoretically. Experimental routes are constrained by costly instrumentation and limited throughput, whereas theoretical approaches—predominantly density-functional theory (DFT)—are computationally demanding and resource-intensive, rendering large-scale ED curation arduous[13, 14]. Concurrently, the MLFF community remains in its infancy regarding effective ED representation learning, underscoring the urgency of establishing a comprehensive ED-based evaluation protocol to accelerate methodological progress. For more background, see Appendix A.2. To address this gap, We construct a more comprehensive large-scale dataset:

- **Scale**: 3.3 million drug-like molecules.
- **Type and resolution**: Cube files containing electron density (ED) data with a grid spacing of 0.4 Bohr, a padding of 4.0 Bohr, and a density fraction threshold of 0.85.
- **Molecule elements**: H,C,N,O,Ti,Ar,S,Se,He,Be,F,P,Si,Ca,Ga,Zn,Ge,Mg,B,Cl,As,Br.
- **Labels**: Electron density ρ, 6 energy components(DF-RKS Final Energy, Nuclear Repulsion Energy, One-Electron Energy, Two-Electron Energy, Exchange-Correlation Energy, Total Energy), 7 frontier orbital energies(HOMO-2, HOMO-1, HOMO-0, LUMO+0, LUMO+1, LUMO+2, LUMO+3), 4 multipole moments(3 Dipoles X, Y, Z, Magnitude), ED visualization.

3 Acceleration potential

EDBench supports the next-generation machine-learning model development for exploring a broader chemical space and designing molecules with target properties. It provides systematic and extensive quantum mechanical data as rich training samples, enabling more accurate and efficient predictive models for molecular property prediction, functional molecule design, and reaction pathway optimization, thereby accelerating new material discovery and drug development. We have evaluated several state-of-the-art deep learning models on the designed benchmark tasks, and the evaluation results show that learning from Edbench is not only feasible, but also achieves high accuracy (For more details about the experiment see Appendix A.3). Further impacts and visions are as follows:

- Replace or warm-start expensive DFT cycles in high-throughput screening.
- Enable joint geometric-electronic architectures that learn transferable chemical rules, improving $\mathrm{p}K_{\mathrm{a}}$, redox potential and binding-affinity prediction across chemical space.
- Provide "electronic fingerprints" for similarity search and retrosynthetic planning, accelerating lead-optimization cycles by weeks.
- Drug/catalyst design by specifying desired density features at active sites.

4 Data-creation pathway

- **Source and Engine**: 3.36 M molecules from PCQM4Mv2 processed with Psi4 1.7.
- **Functional and Basis-Set**: B3LYP hybrid functional; 6-31G** for molecules without S, 6-31+G** (diffuse functions) for S-containing molecules.
- **Spin Treatment**: Closed-shell (multiplicity = 1) $\rightarrow$ RHF reference; Open-shell (multiplicity > 1) $\rightarrow$ UHF reference.
- **Post-SCF cube files Generation**: Grid spacing: 0.4 Bohr; Isosurface defined at 0.85 density-fraction threshold..

Regarding data quality and reliability, refer to Appendix A.4.1.

5 Cost and scalability

Compute cost. All computations were carried out on a high-performance server equipped with 8 Intel(R) Xeon(R) Platinum 8270 CPUs, each with 26 physical cores and 2 threads per core, yielding a total of 416 logical cores. The total computational cost exceeded 205,000 core-hours, equivalent to approximately 23.4 years of single-core compute time.
Scalability. While the EDBench project has made significant progress in the scale and quality of ED data, surpassing existing datasets, there remains room to further enhancement. In future work, we plan to expand the dataset to include higher-level functionals and material-related molecules, and to develop advanced models tailored for ED, enabling EDBench to support a broader range of scientific applications in physical chemistry.

References

[1] Shiru Wu, Xiaowei Yang, Xun Zhao, Zhipu Li, Min Lu, Xiaoji Xie, and Jiaxu Yan. Applications and advances in machine learning force fields. *Journal of Chemical Information and Modeling*, 63(22):6972–6985, 2023.

[2] Adil Kabylda, Valentin Vassilev-Galindo, Stefan Chmiela, Igor Poltavsky, and Alexandre Tkatchenko. Efficient interatomic descriptors for accurate machine learning force fields of extended molecules. *nature communications*, 14(1):3562, 2023.

[3] J Thorben Frank, Oliver T Unke, Klaus-Robert Müller, and Stefan Chmiela. A euclidean transformer for fast and stable machine learned force fields. *Nature Communications*, 15(1):6539, 2024.

[4] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Zhiling Zheng, Chenru Duan, Zhi-Ming Ma, Omar Yaghi, Animashree Anandkumar, Christian Borgs, Jennifer Chayes, et al. Symmetry-informed geometric representation for molecules, proteins, and crystalline materials. *Advances in neural information processing systems*, 36:66084–66101, 2023.

[5] Tianze Zheng, Ailun Wang, Xu Han, Yu Xia, Xingyuan Xu, Jiawei Zhan, Yu Liu, Yang Chen, Zhi Wang, Xiaojie Wu, et al. Data-driven parametrization of molecular mechanics force fields for expansive chemical space coverage. *Chemical Science*, 2025.

[6] Norbert Schuch and Frank Verstraete. Computational complexity of interacting electrons and fundamental limitations of density functional theory. *Nature physics*, 5(10):732–735, 2009.

[7] Wenpei Gao, Christopher Addiego, Hui Wang, Xingxu Yan, Yusheng Hou, Dianxiang Ji, Colin Heikes, Yi Zhang, Linze Li, Huaixun Huyan, et al. Real-space charge-density imaging with sub-ångström resolution by four-dimensional electron microscopy. *Nature*, 575(7783):480–484, 2019.

[8] Ryan Pederson, Bhupalee Kalita, and Kieron Burke. Machine learning and density functional theory. *Nature Reviews Physics*, 4(6):357–358, 2022.

[9] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[10] Nicola Casati, Annette Kleppe, Andrew P Jephcoat, and Piero Macchi. Putting pressure on aromaticity along with in situ experimental electron density of a molecular crystal. *Nature communications*, 7(1):10901, 2016.

[11] Hidetaka Kasai, Kasper Tolborg, Mattia Sist, Jiawei Zhang, Venkatesha R Hathwar, Mette Ø Filsø, Simone Cenedese, Kunihisa Sugimoto, Jacob Overgaard, Eiji Nishibori, et al. X-ray electron density investigation of chemical bonding in van der waals materials. *Nature materials*, 17(3):249–252, 2018.

[12] Tim Gruene, Julian J Holstein, Guido H Clever, and Bernhard Keppler. Establishing electron diffraction in chemical crystallography. *Nature Reviews Chemistry*, 5(9):660–668, 2021.

[13] Libero J Bartolotti and Ken Flurchick. An introduction to density functional theory. *Reviews in computational chemistry*, pages 187–216, 1996.

[14] Michael G Medvedev, Ivan S Bushmarinov, Jianwei Sun, John P Perdew, and Konstantin A Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355(6320):49–52, 2017.

[15] Hongxin Xiang, Shuting Jin, Jun Xia, Man Zhou, Jianmin Wang, Li Zeng, and Xiangxiang Zeng. An image-enhanced molecular graph representation learning framework. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[16] Hongxin Xiang, Li Zeng, Linlin Hou, Kenli Li, Zhimin Fu, Yunguang Qiu, Ruth Nussinov, Jianying Hu, Michal Rosen-Zvi, Xiangxiang Zeng, et al. A molecular video-derived foundation model for scientific drug discovery. *Nature Communications*, 15(1):9696, 2024.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[18] Chuan Shi, Xiao Wang, and S Yu Philip. *Heterogeneous graph representation learning and applications*. Springer, 2022.

[19] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[20] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

[21] Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

[22] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[23] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

[24] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2022.

[25] Yusong Wang, Tong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1):313, 2024.

[26] Zun Wang, Guoqing Liu, Yichi Zhou, Tong Wang, and Bin Shao. Efficiently incorporating quintuple interactions into geometric deep learning force fields. *Advances in Neural Information Processing Systems*, 36:77043–77055, 2023.

[27] Prokop Hapala, Martin Švec, Oleksandr Stetsovych, Nadine J Van Der Heijden, Martin Ondráček, Joost Van Der Lit, Pingo Mutombo, Ingmar Swart, and Pavel Jelínek. Mapping the electrostatic force field of single molecules from high-resolution scanning probe images. *Nature communications*, 7(1):11560, 2016.

[28] Sérgio Filipe Sousa, Pedro Alexandrino Fernandes, and Maria Joao Ramos. General performance of density functionals. *The Journal of Physical Chemistry A*, 111(42):10439–10452, 2007.

[29] Schrödinger and E. An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, 28(6):1049–1070, 1926.

[30] J. M. Combes, P. Duclos, and R. Seiler. *The Born-Oppenheimer Approximation*, pages 185–213. Springer US, Boston, MA, 1981.

[31] Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, Incorporated, Dover Publications, Incorporated, Jul 02 1996.

[32] Frank Gaitan and Franco Nori. Density functional theory and quantum computation, 2009.

[33] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[34] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[35] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, 1988.

[36] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[37] Taoyong Cui, Chenyu Tang, Mao Su, Shufei Zhang, Yuqiang Li, Lei Bai, Yuhan Dong, Xingao Gong, and Wanli Ouyang. Geometry-enhanced pretraining on interatomic potentials. *Nature Machine Intelligence*, 6(4):428–436, 2024.

[38] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16):9759–9815, 2021.

[39] Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.

[40] Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34:13757–13769, 2021.

[41] Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

[42] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693, 2019.

[43] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

[44] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

[45] Yanan Zhang and Xiangzhi Bai. Geometry-augmented molecular representation learning for property prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2024.

[46] Yusong Wang, Shaoning Li, Tong Wang, Bin Shao, Nanning Zheng, and Tie-Yan Liu. Geometric transformer with interatomic positional encoding. *Advances in Neural Information Processing Systems*, 36:55981–55994, 2023.

[47] Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

[48] Xin Deng, WenYu Zhang, Qing Ding, and XinMing Zhang. Pointvector: A vector representation in point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9455–9465, June 2023.

[49] Shuofeng Sun, Yongming Rao, Jiwen Lu, and Haibin Yan. X-3d: Explicit 3d structure modeling for point cloud recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2024.

[50] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[51] Justin M Turney, Andrew C Simmonett, Robert M Parrish, Edward G Hohenstein, Francesco A Evangelista, Justin T Fermann, Benjamin J Mintz, Lori A Burns, Jeremiah J Wilke, Micah L Abrams, et al. Psi4: an open-source ab initio electronic structure program. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(4):556–565, 2012.

[52] Daniel GA Smith, Lori A Burns, Andrew C Simmonett, Robert M Parrish, Matthew C Schieber, Raimondas Galvelis, Peter Kraus, Holger Kruse, Roberto Di Remigio, Asem Alenaizan, et al. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of chemical physics*, 152(18), 2020.

[53] Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(273), 2022.

[54] Bartlomiej Fliszkiewicz and Marcin Sajdak. Fragments quantum descriptors in classification of bio-accumulative compounds. *JOURNAL OF MOLECULAR GRAPHICS & MODELLING*, 125, DEC 2023.

[55] T. Harimoto, M. Kikuchi, T. Suzuki, and et al. Diverse redox-mediated transformations to realize the para-quinoid, -bond, and ortho-diphenoquinoid forms. *Nature Communications*, 16(4088), 2025.

[56] Y. Yang, X. Zhang, L. P. Zhong, et al. Unusual kie and dynamics effects in the fe-catalyzed hetero-diels-alder reaction of unactivated aldehydes and dienes. *Nat. Commun.*, 11(1850), 2020.

[57] E. R. Atangana Nkene, R. A. Y. Kamsi, and H. Z. Adjia. Comparative study of physico-chemical properties of some molecules from khaya grandifoliola plant. *Scientific Reports*, 15(10430), 2025.

[58] Balazs Komjati, Akos Urai, Sandor Hosztafi, Jozsef Koekoesi, Benjamin Kovats, Jozsef Nagy, and Peter Horvath. Systematic study on the td-dft calculated electronic circular dichroism spectra of chiral aromatic nitro compounds: A comparison of b3lyp and cam-b3lyp. *SPECTROCHIMICA ACTA PART A-MOLECULAR AND BIOMOLECULAR SPECTROSCOPY*, 155:95–102, FEB 15 2016.

[59] Lutong Zhang, Jisen Li, Yifan Zhang, et al. White light-excited organic room-temperature phosphorescence for improved in vivo bioimaging. *Nat. Commun.*, 16(3970), 2025.

[60] Francisco E. Jorge, Sara S. Jorge, and Rogerio N. Suave. Electronic circular dichroism of chiral alkenes: B3lyp and cam-b3lyp calculations. *CHIRALITY*, 27(1):23–31, JAN 2015.

[61] R. Behjatmanesh-Ardakani, A. Ebady, and M. A. Karimi. Basis set effects on the prediction accuracy of relative acidity constants of primary and secondary amines. *COMPUTATIONAL AND THEORETICAL CHEMISTRY*, 966(1-3):54–61, JUN 2011.

[62] D Yamaki, T Ohsaku, H Nagao, and K Yamaguchi. Formulation of unrestricted and restricted hartree-fock-bogoliubov equations. *INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY*, 96(1):10–16, JAN 5 2004. 43rd Annual Sanibel Symposium, ST AUGUSTINE, FLORIDA, FEB 23-MAR 01, 2003.

[63] N. D. Woods, M. C. Payne, and P. J. Hasnip. Computing the self-consistent field in kohn-sham density functional theory. *JOURNAL OF PHYSICS-CONDENSED MATTER*, 31(45), NOV 13 2019.

[64] Pierre-Francois Loos, Jean-Louis Rivail, and Xavier Assfeld. Iterative stochastic subspace self-consistent field method. *JOURNAL OF MOLECULAR MODELING*, 23(6), JUN 2017.

A Appendix

A.1 Detailed introduction of AI tasks

A.1.1 Construction details of tasks

To comprehensively evaluate the capacity of the model to understand ED data, we define a suite of tasks based on both molecular structures (MS) and ED, focusing on three fundamental capabilities: prediction of quantum property, retrieval between MS and ED, and generation of ED based on MS . These tasks are constructed and conditionally sampled from the EDBench. To facilitate the development of ED-oriented machine learning methods within the community, we set the dataset size to a moderate scale of up to $n^{max} = 50,000$ molecules, with remaining data available for future research on pre-training strategies. We use scaffold split to divide the dataset into 80% training set, 10% validation set and 10% test set, which is an out-of-distribution split setting and is widely used to evaluate the generalization ability of the model [15, 16]. We summarize the statistics of the designed datasets in Table 1 . We next explain the construction details of these tasks.

Table 1: Statistical information of designed 6 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

Prediction of quantum property. To construct four task-specific datasets—ED5-EC (energy components), ED5-OE (orbital energies), ED5-MM (multipole moments), and ED5-OCS (open/closed-shell)—we design a structure- and label-balanced sampling strategy based on the full EDBench dataset (n molecules). We first extract 2D ECFP4 fingerprints ($fp^{2D} \in \mathbb{R}^{n\times 2048}$) and 3D USR descriptors ($fp^{3D} \in \mathbb{R}^{n\times 12}$) for each molecule, concatenate them, and apply k-means clustering ($k = 100$) to obtain structure clusters C^s. For the multi-dimensional labels y^{EP} (6D), y^{GE} (7D), and y^{MMR} (4D), we similarly apply k-means ($k = 100$) to produce clusters C^{EC}, C^{OE}, and C^{MM}, respectively; for y^{OCS} (binary), we use the original label. We then form sampling groups as (C^s, C^{EC}), (C^s, C^{OE}), (C^s, C^{MM}), and (C^s, y^{OCS}), and uniformly sample $m = \texttt{max}(n^{max}//n^{group}, 1)$ molecules from each group to construct the final datasets, ensuring diversity in both structure and property space.

Retrieval between MS and ED. Retrieval between MS and ED is a fundamental task. Retrieving molecular structures from ED (`ED` $\rightsquigarrow$ `MS`) enables electron-level virtual screening, while retrieving ED from structures (`MS` $\rightsquigarrow$ `ED`) supports electron-aware models—facilitating molecular representation learning, inverse design, and quantum-informed modeling. To construct the ED5-MER dataset for bidirectional retrieval between MS and ED, we group all molecules in EDBench by structure cluster C^s and uniformly sample m anchor (MS and ED) from each group. For each anchor, we sample $n^{neg} = 10$ negative samples: half from the same cluster (easy negatives) and half from different clusters (hard negatives). The final task involves identifying the correct ED (or MS) from a set of 11 candidates given an anchor MS (or ED).

Generation of ED based on MS. Generating ED from MS (`MS` $\rightarrow$ `ED`) is a highly valuable task, as it can significantly reduce the computational cost associated with DFT-based ED calculations. Since ED is inherently dependent on both molecular connectivity and 3D geometry, we ensure diversity in both structure and density by grouping molecules via C^s and uniformly sampling m MS-ED pairs from each group. In this task, the model is given an MS as input and is required to predict its ED.

A.1.2 Methods

To assess the model's understanding of electron density (ED), we design tailored learning paradigms for each task type (prediction, retrieval, and generation). For clarity, we formalize the molecular

structure (MS) with n atoms as $\mathcal{G} = (\mathcal{V}, \mathcal{Z}^{\mathcal{G}})$, where $\mathcal{V} = \{v_1, v_2, ..., v_n\} \in \mathbb{R}^{n\times 1}$ and $\mathcal{Z}^{\mathcal{G}} = \{z_1^{\mathcal{G}}, z_2^{\mathcal{G}}, ..., z_n^{\mathcal{G}}\} \in \mathbb{R}^{n\times 3}$ denote atomic types and their corresponding 3D coordinates, respectively. The ED data with m points is denoted as $\mathcal{P} = (\mathcal{Z}^{\mathcal{P}}, \mathcal{D})$, where $\mathcal{Z}^{\mathcal{P}} = \{z_1^{\mathcal{P}}, z_2^{\mathcal{P}}, ..., z_m^{\mathcal{P}}\} \in \mathbb{R}^{n\times 3}$ represents the ED coordinates and the corresponding density values $\mathcal{D} = \{d_1, d_2, .., , d_m\} \in \mathbb{R}^{n\times 1}$. We denote the MS encoder and ED encoder as $\texttt{Enc}_{\mathcal{G}}$ and $\texttt{Enc}_{\mathcal{P}}$, respectively, to extract latent representations from MS and ED.

For prediction tasks, we introduce an additional task-specific prediction head $\texttt{Enc}_t$, whose output dimension matches the number of target labels for each task. The learning paradigm is defined as follows: the ED encoder $\texttt{Enc}_{\mathcal{P}}$ first extracts features from $\mathcal{P}$, which are then passed through $\texttt{Enc}_t$ to generate task-specific predictions $\hat{y}$. This process can be formalized as:

$$\hat{y}^{\bullet} = \texttt{Enc}_t^{\bullet}(\texttt{Enc}_{\mathcal{P}}(\mathcal{P})) \tag{1}$$

where $\bullet$ denotes a specific task, such as EC, OE, MM, or OCS. Accordingly, on the ED5-EC, ED5-OE, ED5-MM, and ED5-OCS datasets, we compute the loss between $\hat{y}^{\bullet}$ and the corresponding ground truth $y^{\bullet}$ to optimize the model. Specifically, cross-entropy loss is used for classification tasks, while L2 loss is applied for regression tasks.

For retrieval tasks, we utilize $\texttt{Enc}_{\mathcal{G}}$ and $\texttt{Enc}_{\mathcal{P}}$ to extract latent representations $h_{\mathcal{G}}$ and $h_{\mathcal{P}}$ from the MS $\mathcal{G}$ and ED $\mathcal{P}$, respectively, which can be formalized as:

$$h_{\mathcal{G}} = \texttt{Enc}_{\mathcal{G}}(\mathcal{G}), \quad h_{\mathcal{P}} = \texttt{Enc}_{\mathcal{P}}(\mathcal{P}) \tag{2}$$

The models are trained with the InfoNCE loss [17], which pulls matched pairs closer in the embedding space while pushing apart mismatched ones. Formally, given a batch of n paired samples $\{(\mathcal{G}_i, \mathcal{P}_i)\}_{i=1}^{n}$, the loss for a positive pair $(\mathcal{G}_i, \mathcal{P}_i)$ is defined as:

$$\mathcal{L}_{\text{ret}} = -\log \frac{\exp(\text{sim}(h_{\mathcal{G}_i}, h_{\mathcal{P}_i})/\tau)}{\sum_{j=1}^{n} \exp(\text{sim}(h_{\mathcal{G}_i}, h_{\mathcal{P}_j})/\tau)} \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ denotes a similarity function (e.g., cosine similarity), and $\tau = 0.07$ is a temperature.

For the generation task, we construct a heterogeneous graph [18], defined as:

$$\mathcal{HG} = \{(\mathcal{V}, \mathcal{Z}^{\mathcal{G}}), (\mathcal{Z}^{\mathcal{P}}, \mathcal{D}), \mathcal{E}\} \tag{4}$$

where $\mathcal{HG}$ contains two types of nodes: atoms and electrons. To construct the edge set $\mathcal{E}$, we perform a k-nearest neighbor search ($k = 9$) for each node, retrieving the k closest nodes of the same type and k of the opposite type, which results in 18 edges per node, forming atom–atom, atom–electron, and electron–electron connections. Since the goal is to predict ED from MS, we mask all ED values to obtain the masked graph $\hat{\mathcal{HG}}$. We extend Equivariant Graph Neural Network (EGNN) [19], called HGEGNN, to support heterogeneous graph. In HGEGNN, we treat electrons as special atoms and apply the same EGNN operations as used for regular atoms. We then input $\hat{\mathcal{HG}}$ into an HGEGNN to extract node representations $h^{\mathcal{HG}}$, which are split into atomic features $h_{\mathcal{G}}^{\mathcal{HG}}$ and electronic features $h_{\mathcal{P}}^{\mathcal{HG}}$. Finally, we apply a prediction head $\texttt{Enc}_t^{\text{EDP}}$ to the electronic features to generate the masked density values:

$$h^{\mathcal{HG}} = \text{HGEGNN}(\hat{\mathcal{HG}}), \quad \hat{\mathcal{D}} = \texttt{Enc}_t^{\text{EDP}}(h_{\mathcal{P}}^{\mathcal{HG}}) \tag{5}$$

where $\hat{\mathcal{D}} \in \mathbb{R}^{n_{\mathcal{P}}\times 1}$ is the predicted ED. We minimize the discrepancy between $\hat{\mathcal{D}}$ and the ground-truth $\mathcal{D}$ by the following L2 loss:

$$\mathcal{L}_{\text{gen}} = \|\hat{\mathcal{D}} - \mathcal{D}\|_p, \quad p = 2 \tag{6}$$

A.1.3 Evaluations

In the prediction tasks, the predicted labels are obtained via Equation 1. Specifically, in ED5-EC, ED5-OE, and ED5-MM, we evaluate the prediction performance using MAE between the predicted and ground-truth values, i.e., $(\hat{y}_{\mathcal{P}}^{EC}, y_{\mathcal{P}}^{EC})$, $(\hat{y}_{\mathcal{P}}^{OE}, y_{\mathcal{P}}^{OE})$, and $(\hat{y}_{\mathcal{P}}^{MM}, y_{\mathcal{P}}^{MM})$. For ED5-OCS, we assess classification performance using accuracy, ROC-AUC, AUPR, and F1-score between the predicted logits $\hat{y}_{\mathcal{P}}^{OCS}$ and ground-truth labels $y_{\mathcal{P}}^{OCS}$. **In the retrieval task**, we evaluate the quality of the latent features $h_{\mathcal{G}}$ and $h_{\mathcal{P}}$ extracted via Equation 2. Specifically, in ED5-MER, given a molecular

feature $h_{\mathcal{G}_i}$ as the anchor, we retrieve from a set of ED features by computing cosine similarities and ranking the results; Top-k accuracy ($k = 1, 3, 5$) is used as the evaluation metric, where a hit is counted if the correct match appears in the top k results. Similarly, we perform retrieval in the opposite direction using $h_{\mathcal{P}_i}$ as the anchor and $h_{\mathcal{G}}$ as the candidate set. **In the generation task**, the predicted ED $\hat{\mathcal{D}}$ is obtained via Equation 5. We evaluate the generation performance using MAE, Pearson and Spearman's rank correlation coefficients between $\hat{\mathcal{D}}$ and the ground-truth $\mathcal{D}$ in ED5-EDP.

A.1.4 Significance of tasks

We define three core tasks that capture distinct yet complementary capabilities of modeling electron density (ED), each grounded in both scientific motivation and real-world utility:

- **Prediction of quantum property.** As ED fundamentally determines molecular quantum behavior, predicting properties such as total energy, dipole moment, and orbital energies from ED allows us to assess whether a model has captured the underlying physical principles linking electron distributions to quantum observables. Despite ED being typically computed via expensive DFT simulations, it encodes richer quantum information than molecular geometry alone. Accurate property prediction from ED thus serves as a proxy for model fidelity to quantum mechanics and offers a potential route to accelerate quantum property estimation in applications like drug discovery, catalysis, and materials design.
- **Retrieval between MS and ED.** Bidirectional retrieval between MS and ED enables molecule-level search in ED databases and supports structure inference from electronic environments. MS-to-ED retrieval facilitates functional site localization and electron distribution analysis, while ED-to-MS retrieval provides a foundation for inverse design driven by electronic requirements. This dual capability is essential for high-resolution virtual screening pipelines grounded in electronic behavior.
- **Generation of ED based on MS (Molecular Structure).** Learning to generate high-fidelity ED distributions directly from molecular structures bypasses the computational burden of DFT, making ED accessible to downstream tasks such as deep molecular dynamics, quantum-aware neural force fields, and reaction path modeling. This capability bridges the gap between computational efficiency and quantum-level accuracy, unlocking ED-driven learning for large-scale modeling scenarios.

A.1.5 Detailed statistics of 6 benchmarks

We provide a detailed statistical analysis of six benchmarks in the EDBench suite: ED5-EC, ED5-OE, ED5-MM, ED5-OCS, ED5-MER, and ED5-EDP. Figures 1, 2, and 3 illustrate the distributions of the number of atoms, the number of ED points at the threshold $\rho_\tau = 0$, and the per-molecule mean ED values at $\rho_\tau = 0$, respectively. As shown, the number of ED points significantly exceeds the number of atoms, which provides richer information for force field learning and related downstream tasks.

Furthermore, we report the distribution of ED point counts and mean ED values under a higher threshold $\rho_\tau = 0.05$ in Figures 4 and 5, respectively. By applying a larger threshold (e.g., $\rho_\tau = 0.05$), the overall ED point count is significantly reduced, which can lead to improved computational efficiency. This suggests that threshold tuning offers a practical way to control the data volume without severely compromising structural fidelity. In addition, Figure 5 reveals that increasing the ED threshold implicitly forces the model to focus more on high-density regions, which are typically more chemically informative and relevant for modeling interactions.

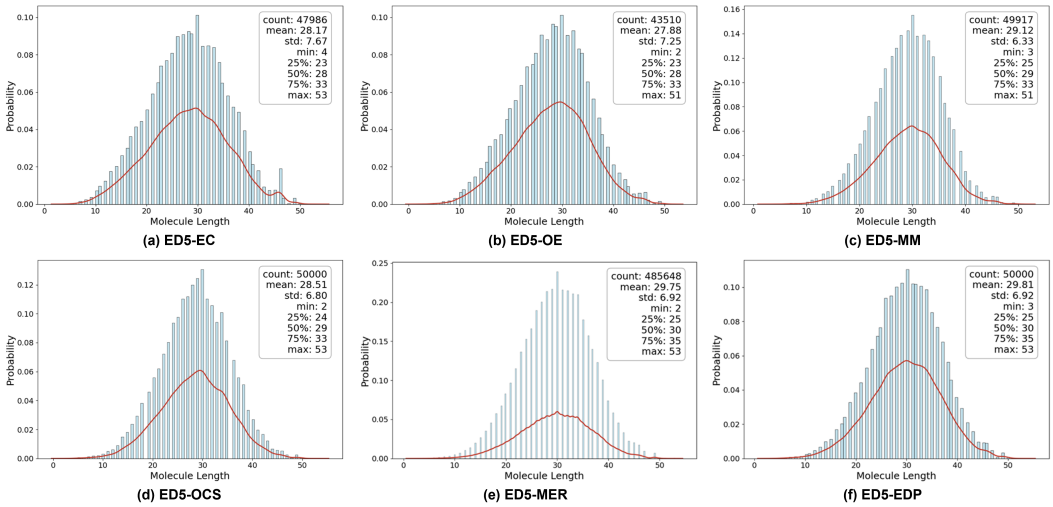

Figure 1: Distribution of the number of atoms in the 6 benchmark datasets.

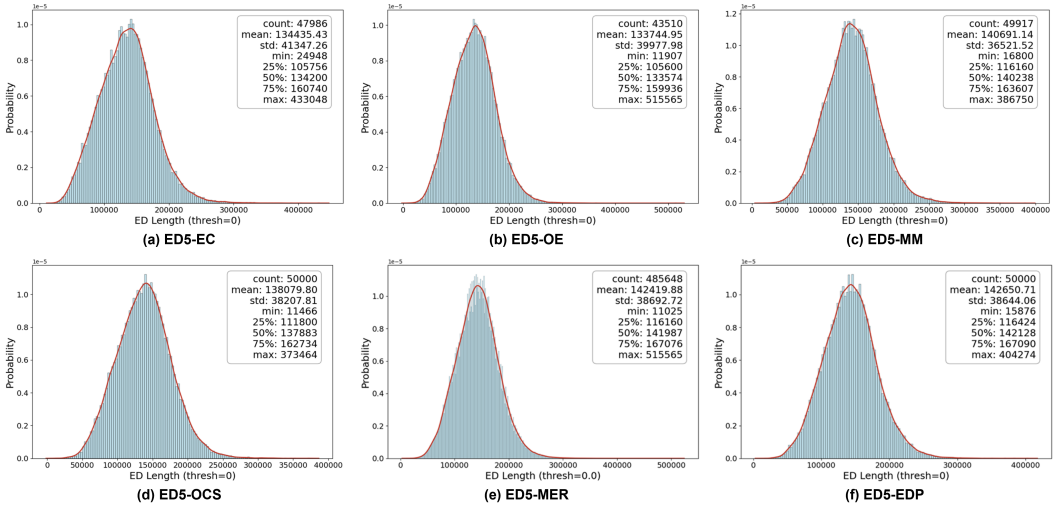

Figure 2: Distribution of the number of ED points in the 6 benchmark datasets with ED threshold $\rho_\tau = 0$.

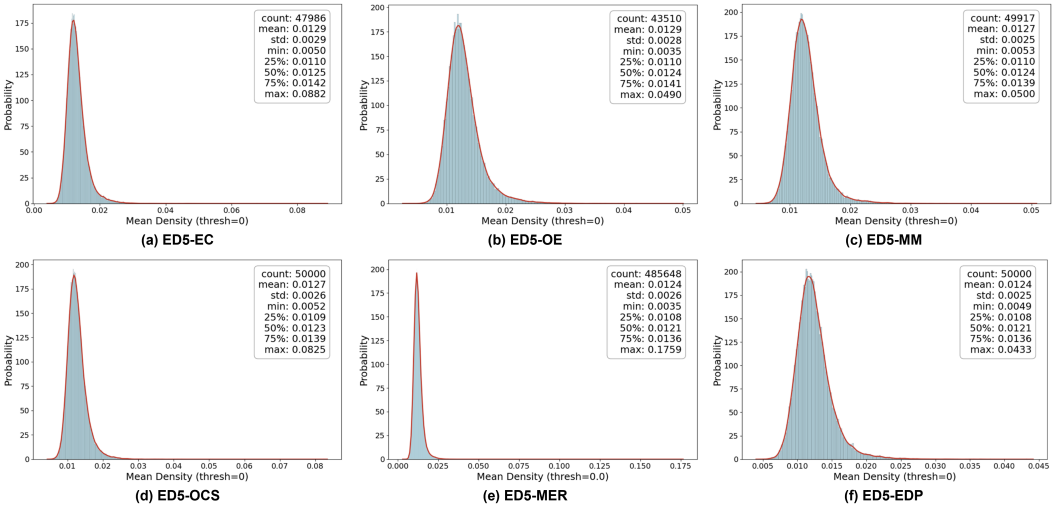

Figure 3: Distribution of per-molecule mean ED values in the 6 benchmark datasets with ED threshold $\rho_\tau = 0$.

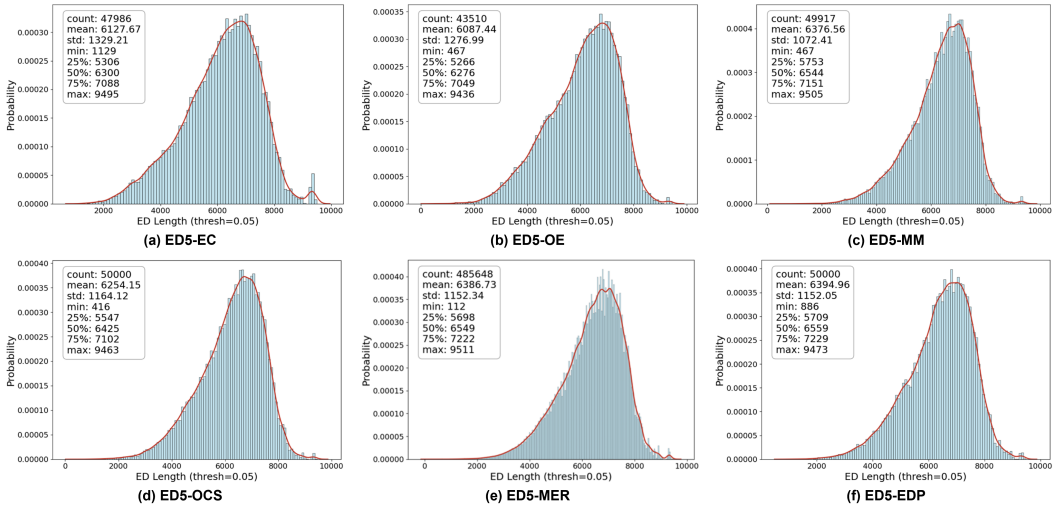

Figure 4: Distribution of the number of ED points in the 6 benchmark datasets with ED threshold $\rho_\tau = 0.05$.

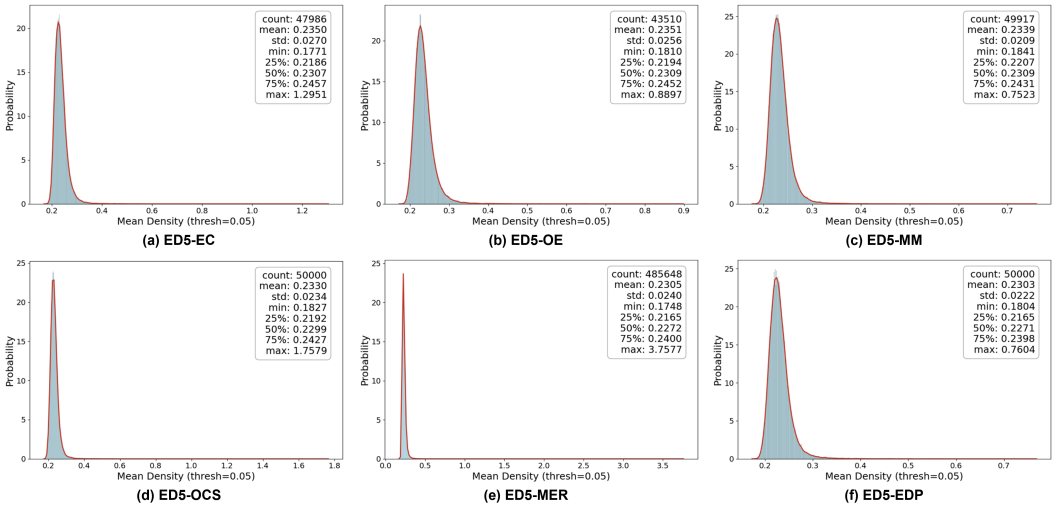

Figure 5: Distribution of per-molecule mean ED values in the 6 benchmark datasets with ED threshold $\rho_\tau = 0.05$.

A.2 Background

A.2.1 Introduction

With the widespread adoption of deep learning in molecular dynamics (MD) simulations, machine learning force fields (MLFFs) have become efficient and promising computational tools, significantly advancing research in physics, chemistry, biology, and materials science [1, 2, 3]. State-of-the-art MLFFs methods typically employ geometric deep learning to model atomic interactions within molecules, a strategy that has proven to be effective [5]. These models are generally built upon many-body interactions at the atomic level, including one-body (atomic attributes such as types and coordinates [19]), two-body (interatomic distances [20, 2]), three-body (bond angles [21, 22]), four-body (torsions [23, 24, 25] and improper torsions [5]), and five-body interactions [26].

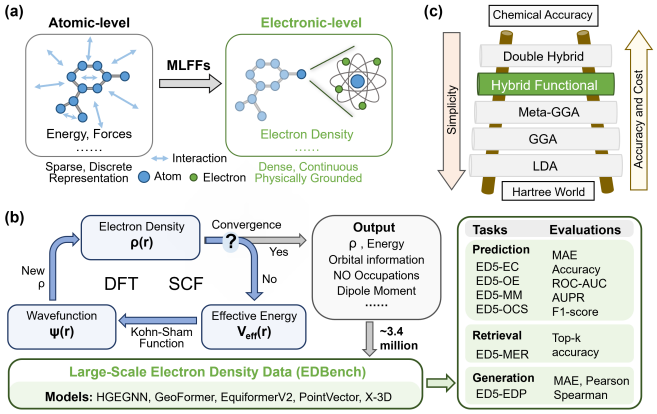

Figure 6: (a) Advancing MLFFs from atomic-level interactions—based on discrete atomistic representations—to electronic-level modeling using continuous ED, enabling richer and more physically grounded supervision; (b) Overview of the proposed EDBench dataset; (c) DFT method selection guided by Jacob's ladder to balance accuracy and computational cost.

Although existing MLFFs have demonstrated great potential in modeling molecular force fields (MFFs), they primarily focus on capturing coarse-grained, atom-level many-body interactions [4, 5], while often overlooking the critical role of microscopic electron distribution in understanding molecular interactions [6, 7, 8]. It is well known that the spatial distribution of electrons directly influences the interactions between atoms within a molecule, providing the most direct and fundamental information for interpreting molecular force fields [27]. Electron density (ED), as a fundamental physical quantity in quantum mechanics that describes the spatial distribution of electrons, offers a more fine-grained and physically grounded representation of molecular systems according to Hohenberg–Kohn (HK) theorem [9]. Therefore, explicitly incorporating ED into the modeling process holds promise for bridging the gap between microscopic electronic behavior and macroscopic force fields, further

improving both the accuracy and generalization of MLFFs. Therefore, as illustrated in Figure 6(a), the primary objective of this work is to advance current MLFFs beyond the atom-level learning paradigm toward electron-level modeling, enabling a more accurate and physically grounded description of molecular interactions.

However, advancing MLFFs toward an electron-level understanding faces two major challenges: (i) *the lack of large-scale, high-quality ED datasets*, which are essential for pretraining and could fundamentally reshape the paradigm of MLFFs modeling. (ii) *the absence of an ED-centric benchmark* to systematically explore the feasibility and effectiveness of ED-based modeling frameworks. Specifically, the acquisition of ED data can be categorized into two approaches: experimental methods (such as X-ray diffraction [10, 11], electron diffraction [12]) and theoretical calculation methods. Due to the reliance on expensive physical equipment, experimental methods inevitably limit data acquisition, making theoretical methods more popular. Theoretical calculations typically use density functional theory (DFT) [13, 14], the most common approach, to compute the ED of molecules. Although DFT does not depend on specialized observation equipment, its calculations are highly computationally intensive and time-consuming, making the acquisition of large-scale, high-quality ED datasets particularly difficult. In addition, the MLFFs community is still in the early stages of learning effective representations from ED, which makes the development of an ED-based evaluation protocol particularly important for the rapid advancement of ED representation learning.

To address the two key challenges outlined above, we introduce EDBench, a large-scale and high-fidelity dataset of ED, as shown in Figure 6(b). Following Jacob's ladder [28], as shown in Figure 6(c), we adopt higher-rung hybrid functionals as the underlying DFT methods to ensure the quality of the EDBench dataset.

A.2.2 Density functional theory (DFT)

The quantum mechanical description of many-electron systems is one of the core issues in modern physics and chemistry. Schrödinger equation [29] as the fundamental equation of quantum mechanics, is challenging to solve directly. Consequently, researchers introduced various wave function-based approximation methods to simplify the problem, such as, Born–Oppenheimer [30] and Hartree-Fock method [31]. Those methods scale with the number of electrons n as $\mathcal{O}(n^4)$ or more, its computational cost remains prohibitive for large polyatomic molecules. In contrast, Density Functional Theory (DFT) is more suitable for complex systems due to its lower computational cost ($\mathcal{O}(n^3)$) and incorporation of electron correlation effects [32]. The core concept of DFT is to use electron density (ED) as the fundamental variable instead of the wave function. The Hohenberg-Kohn theorem is the cornerstone of DFT [9], which states that the external potential field and the ground-state energy can be completely determined by ED. Thus, by solving for the ED distribution $\rho(r)$ that achieves the lowest energy, the properties of the stable system can be confirmed. The ED $\rho(r)$ can be expressed as:

$$\rho(\mathbf{r}) = \rho_\alpha(\mathbf{r}) + \rho_\beta(\mathbf{r}) \tag{7}$$

where $\rho_\alpha(\mathbf{r})$ and $\rho_\beta(\mathbf{r})$ are the density of α-spin electrons and β-spin electrons.

This concept is concretely realized in the Kohn-Sham equations, which transforms the polyelectron system with interactions into single-electron system without interaction, and adds interactions among electrons to exchange-correlation potential [33]. The Kohn-Sham equations is shown as:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{eff}}(r)\right]\psi_i(r) = \epsilon_i \psi_i(r) \tag{8}$$

where $\psi_i(r)$ and ϵ_i are, respectively, the wave function and energy of the i-th single-electron orbital, and $V_{\text{eff}}(r)$ is the effective single-electron potential energy. The basis of DFT is Hohenberg-Kohn theorem, and Kohn-Sham equation is the practical application form of DFT. In Kohn-Sham equation, $V_{\text{eff}}(r)$ is the effective single-electron potential energy, defined as

$$V_{\text{eff}}(r) = V_{\text{ext}}(r) + V_{\text{H}}(r) + V_{\text{xc}}(r) \tag{9}$$

The external potential $V_{\text{ext}}(r)$ is typically provided by the atomic nuclei. $V_{\text{H}}(r)$ is the Hartree potential, which is represented by the convolution of the ED with the Coulomb kernel. The exchange-correlation potential $V_{\text{xc}}(r)$ is the variational derivative of the exchange-correlation energy functional.

The solution of the Kohn-Sham equations is typically achieved through self-consistent field (SCF) iterations, as shown in the figure 6(b). Initially, a set of initial electron densities $\rho(\mathbf{r})$ is selected, and

the effective potential $V_{\text{eff}}(\mathbf{r})$ is calculated based on this initial guess. The Kohn-Sham equations are then solved to obtain new single-electron orbital wave functions $\psi_i(\mathbf{r})$ and energies ε_i, which are used to update the ED $\rho(\mathbf{r})$. This process is repeated until convergence is achieved, yielding the ED $\rho(\mathbf{r})$ and simultaneously stabilizing the total energy E.

In addition, to solve the equation, it is usually necessary to select the basis set, pseudopotential, and exchange correlation functional. The basic set includes plane wave method, numerical atomic orbital method, and augmented wave method. Norm-conserving pseudopotential (NCPP), ultrasoft pesudopotential (USPP) and projector augmented wave (PAW) are common pseudopotential methods. The exchange-correlation energy functional includes the Local Density Approximation (LDA) [34], the Generalized Gradient Approximation (GGA) [35], and hybrid functionals (such as B3LYP) [36]. In this paper, the exchange-correlation functional used is B3LYP, and the 6-31G**/+G** basis set is selected for combination. B3LYP integrates the advantages of the Hartree-Fock method and DFT. The 6-31G**/+G** basis set enhances computational accuracy by splitting the valence electron orbitals into two sets of basis functions and further incorporating diffuse functions. This combination achieves a great balance between precision and efficiency, making it more suitable.

A.2.3 Molecular geometry learning in quantum chemistry

Geometric Deep Learning (GDL) has become a dominant approach for modeling machine learning force fields (MLFFs), primarily focusing on atom-level information such as atomic attributes and interatomic interactions. Specifically, GDL models are built upon first-order atomic features, including atom types and 3D coordinates [19, 37]. To capture geometric relationships while preserving physical consistency, GDL methods incorporate symmetries such as rotational and translational invariance in 3D space [38, 39]. Consequently, a wide range of models have been developed with built-in invariance or equivariance to Euclidean group E(3) [19] or special Euclidean group SE(3) [40, 41], ensuring that predictions are physically meaningful. Given that atomic interactions—such as bonding, repulsion, and van der Waals forces—play a crucial role in molecular fields, modern GDL methods further incorporate second-order geometric features, including interatomic distances [42, 43], bond types [44], and spatial neighborhood structures [45]. To more precisely capture local structural features, some approaches even extend to higher-order geometric relations such as bond angles (three-body interactions) [21, 22] and torsional angles (four-body interactions) [23, 24, 25], thereby improving the expressiveness and accuracy of force field modeling. In contrast to prior works that focus primarily on atom-level representations, our proposed EDBench introduces a large-scale dataset of electronic density (ED), laying the foundation for extending molecular modeling from the atomic scale to the electronic scale. It also provides a new platform and evaluation benchmark for developing GDL methods tailored to electronic structure modeling.

A.3 Experiment and discussion

A.3.1 Experiment settings

Baseline. For comprehensiveness of the evaluation, we evaluate both molecular structure-based and electron density-based methods. Specifically, we selected several state-of-the-art baselines for evaluation on the proposed benchmark: (i) two geometric models based on molecular structure (MS): GeoFormer [46] and EquiformerV2 [47]; (ii) two point cloud models based on electron density (ED): PointVector [48] and X-3D [49]. GeoFormer and EquiformerV2 are Transformer-based architectures that use Interatomic Positional Encoding (IPE) and higher-degree tensors, respectively, to learn the interaction relationships between atoms. Unlike GeoFormer and EquiformerV2, which are specifically designed for molecules, PointVector and X-3D are the latest methods that focus on real-world point clouds. They are MLP(Multi-layer Perceptron)-based and explict structure-based architectures, respectively, offering excellent computational efficiency to handle large-scale point clouds.

Setup. The codes of all baselines are available from their GitHub repositories and we reproduce them on our benchmarks. We use the same experimental settings as these baselines. All datasets are split using a scaffold split [15] based on the out-of-division (OOD) scenario, which enables evaluating the generalization of the model. We repeat the experiments three times with different random seeds and report the means and standard variances on the test set. The test set results are selected according to the best validation set performance. Due to the excessive length of the ED vectors (Figure **??**(c)), we introduce a threshold ρ_τ to filter out electrons in regions with negligible density (all ED values below

ρ_τ are discarded). All models were trained using either NVIDIA A100 (80GB PCIe) or GeForce RTX 3090 (24GB) GPUs, depending on their memory requirements.

A.3.2 Details of computational efficiency

We conduct a computational efficiency analysis of all baseline models presented in this work, including molecular geometry-based methods—HGEGNN, EquiformerV2, and GeoFormer—and ED point cloud-based methods—PointVector and X-3D. As a first step, we report the parameter count of each model to assess their relative model capacities. The details are summarized in Table 2. We find that the model sizes of EquiformerV2 and GeoFormer are significantly larger than the other models.

Table 2: The number of parameters of different models. #Params represents the number of parameters of the model. M stands for Million.

	HGEGNN	Equiformerv2	GeoFormer	PointVector	X-3D
#Params (M)	0.574	27.9	9.5	1.5454	0.9476

Next, we report the GPU memory usage and training time for each model. Due to varying memory requirements across models, we had to use different GPU devices to accommodate specific models and avoid out-of-memory (OOM) issues. Tables 3 and 4 present the computational efficiency of PointVector and X-3D, respectively. As expected, both GPU memory consumption and training time increase consistently with the number of sampling points ξ.

Table 3: The computational efficiency of PointVector with different number of sampling points ξ on ED5-OE dataset with batch size of 32 and epoch of 100. Time refers to the total time spent on the entire training process.

ξ	GPU Memory (MiB)	Time (minutes)	GPU
512	4,425	$\sim$100	3090
1024	6,623	$\sim$150	3090
2048	11,453	$\sim$325	3090
4096	20,757	$\sim$433	a100-80gb-pcie
8192	38,083	$\sim$850	a100-80gb-pcie

Table 4: The computational efficiency of X-3D with different number of sampling points ξ on ED5-OE dataset with batch size of 32 and epoch of 100. Time refers to the total time spent on the entire training process.

ξ	GPU Memory (MiB)	Time (minutes)	GPU
512	3,431	$\sim$71	3090
1024	4,747	$\sim$88	3090
2048	7,951	$\sim$156	3090
4096	13,701	$\sim$305	3090
8192	21,351	$\sim$750	3090

Additionally, Table 5 shows the time efficiency of HGEGNN on the ED5-EDP dataset. A similar trend is observed: as the ED threshold ρ_τ decreases, the number of ED points increases, leading to higher memory usage and longer training times. These results collectively highlight the sensitivity of model efficiency to both the resolution of input data and the complexity of the architecture.

Table 5: The computational efficiency of HGEGNN with different ED threshold ρ_τ on ED5-EDP dataset. MiB/mol represents the total memory usage divided by the batch size.

ρ_τ	GPU Memory (MiB/mol)	Time (minutes/epoch)	GPU
0.1	2,153	$\sim$15	a100-80gb-pcie
0.15	907	$\sim$7.5	a100-80gb-pcie
0.2	616	$\sim$5	a100-80gb-pcie

A.3.3 Performance on prediction tasks

Table 6: 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**

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

	LUMO+1	HOMO-2 LUMO+2	HOMO-1 LUMO+3	HOMO-0	LUMO+0
PointVector	**1.73±0.01**	**1.68±0.01**	**1.92±0.01**	**3.08±0.05**	**2.86±0.05**
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			

Table 8: 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**

Tables 6, 7, 8, and 9 report the performance of recent models on the ED5-EC, ED5-OE, ED5-MM, and ED5-OCS datasets, respectively. We observe that X-3D consistently outperforms PointVector, achieving the best results on ED5-EC (Table 6), ED5-MM (Table 8), and ED5-OCS (Table 9). Notably, both X-3D and PointVector exhibit significantly stronger performance on orbital energy prediction (Table 7) than on energy component prediction (Table 6). This is likely due to the stronger locality of orbital energies, which are more directly linked to local ED patterns, allowing models to extract relevant features more effectively. In contrast, predicting energy components requires integrating over the entire ED, demanding the learning of more complex global interactions. These results further validate the effectiveness of using ED as a model input and demonstrate its utility in capturing physically meaningful patterns.

Table 9: 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

While X-3D and PointVector were not designed for ED data, their strong performance on our benchmarks underscores the potential of ED-based learning in quantum property prediction. We expect tailored models to further improve performance, advancing more accurate and efficient quantum modeling.

A.3.4 Performance on retrieval tasks

We first use GeoFormer and EquiformerV2 as molecular structure (MS) encoders, and PointVector and X-3D as electron density (ED) encoders. These encoders are combined pairwise—GeoFormer+PointVector, GeoFormer+X-3D, EquiFormer+PointVector, and EquiFormer+X-3D—to systematically evaluate cross-modal retrieval performance. Table 10 reports the Top-k accuracy on both `ED` $\rightarrow$ `MS` and `MS` $\rightarrow$ `ED` tasks. Results reveal substantial performance differences among combinations. For example, GeoFormer+PointVector achieves only 17.67% Top-1 accuracy, while GeoFormer+X-3D reaches 68.32%, yielding an absolute improvement of 50.65%. Similarly, EquiFormer+PointVector achieves just 10.24% Top-1 accuracy, whereas EquiFormer+X-3D reaches 78.71%—an absolute gain of 68.47%. These results highlight the critical importance of selecting

appropriate encoder architectures for effective cross-modal representation learning between MS and ED.

Table 10: The Top-k accuracy (%) on ED5-MER dataset. ED → MS represents using electron density (ED) to retrieve molecular structure (MS).

MS model	ED model	ED → MS			MS → ED		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
GeoFormer	PointVector	17.67±2.10	46.09±4.53	67.63±5.92	27.01±1.69	59.02±2.49	77.42±3.01
	X-3D	68.32±3.70	92.18±2.41	97.31±1.29	70.01±2.93	92.08±2.01	97.17±0.92
EquiformerV2	PointVector	10.24±1.28	32.47±2.69	53.42±2.67	22.18±0.64	54.61±2.89	76.83±2.90
	X-3D	**78.71±0.69**	**94.78±0.40**	**98.13±0.07**	**78.36±0.65**	**94.19±0.14**	**97.74±0.29**

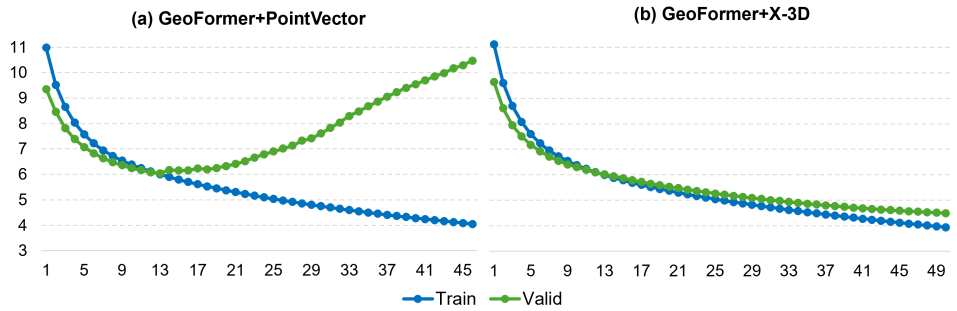

Figure 7: Comparative Learning Loss of GeoFormer+PointVector and GeoFormer+X-3D on ED5-MER training (Train) and validation (Valid) sets.

To further understand the performance gap, we closely analyzed the training logs of GeoFormer+PointVector and GeoFormer+X-3D. Figures 7(a) and 7(b) show their contrastive learning loss curves on the training and validation sets, respectively. While both combinations exhibit steadily decreasing training loss, GeoFormer+PointVector suffers from overfitting—as evidenced by its increasing validation loss despite continued improvement on the training set. In contrast, GeoFormer+X-3D maintains a consistently decreasing loss on both training and validation sets, explaining its significantly better retrieval performance.

Overall, the strong bidirectional retrieval performance of GeoFormer+X-3D and EquiFormerV2+X-3D demonstrates the feasibility of learning the complex mapping between MS and ED, providing a solid foundation for retrieval-based applications. For example, retrieving the most compatible MS given an ED can enable a novel perspective on high-throughput virtual screening—particularly valuable in scenarios where the ED is known but the MS is unknown or ambiguous. Conversely, retrieving approximate ED distributions from MS opens a promising direction for building structure-driven, density-aware models, potentially enhancing the physical faithfulness of downstream tasks such as molecular property prediction and reactivity analysis.

We employ {GeoFormer (G), EquiformerV2 (E)} and {PointVector (P), X-3D (X)} as the MS encoder $\texttt{Enc}_{\mathcal{G}}$ and ED encoder $\texttt{Enc}_{\mathcal{P}}$, respectively, in Equation 1. These components are paired to form four combinations: G-P, G-X, E-P, and E-X. Their retrieval performance is evaluated in Figure 8. The results show that combinations involving E (i.e., E-P and E-X) consistently outperform those involving G, highlighting the importance of selecting an appropriate encoder for retrieval tasks. Overall, the strong performance of E-P and E-X demonstrates their potential for ED-based virtual screening and MS-based electronic-level molecular understanding.

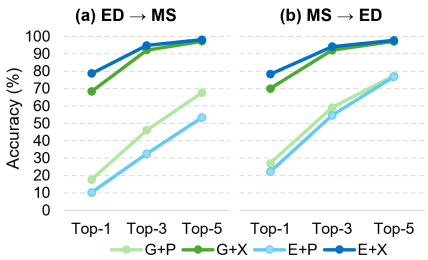

Figure 8: The retrieval performance on ED5-MER.

A.3.5 Performance on generation task

Table 11 presents the results of HGEGNN on the ED prediction task under different ED thresholds ρ_τ. We observe that, given the molecular structure (MS), the model can accurately predict ED values, achieving low MAE and high Pearson and Spearman correlations. This indicates that the deep learning method can significantly accelerate the generation of ED while reducing the computational cost associated with DFT calculations.

Table 11: The performance of HGEGNN on ED generation of ED5-EDP dataset. The unit of Time is second/mol.

	ρ_τ	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

Notably, the model performance improves with increasing ρ_τ, indicating it effectively captures high-ED regions. This aligns with chemical intuition, as high-density regions often correspond to chemically significant areas such as atomic cores and bonding regions, where the spatial patterns are more structured and consistent across molecules, making them easier for the model to learn.

A.3.6 Quality analysis of ED outputs from the generation task

To assess the quality of the ED data generated by HGEGNN, we employ models trained with three different random seeds, as described in Section A.3.5, to generate ED5-EC data with a density threshold of $\rho_\tau = 0.2$, denoted as G#1, G#2, and G#3. Figure 10 compares the average MAE performance of different data sources using the PointVector as baseline, where red values denote relative improvements compared to DFT-based data source. We observe that G#1, G#2, and G#3 consistently outperform the DFT-based data, indicating that HGEGNN generates high-quality ED. These demonstrate the potential of using predicted ED directly to enhance the model's understanding of MFFs.

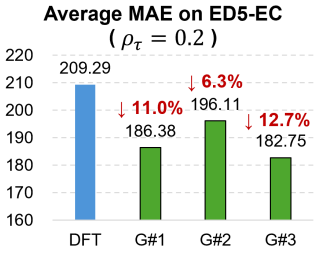

Figure 9: The average MAE of PointVector on ED5-EC generated by DFT and G#{1,2,3}.

To evaluate the quality of ED outputs in the generation task, we replace the DFT-based ED5-EC data (with a density threshold of $\rho_\tau = 0.2$) with new ED data generated by HGEGNN models trained on the original ED5-EC dataset. These new datasets are denoted as HGEGNN(2024), HGEGNN(2025), and HGEGNN(2026), where the numbers indicate different random seeds used during training. We then train PointVector—configured with the minimal ED length sampling rate—on each of these generated datasets. Detailed results are shown in Table 12. Compared to PointVector trained on the DFT-based ED5-EC, PointVector models trained on HGEGNN(2024)-, HGEGNN(2025)-, and HGEGNN(2026)-based ED5-EC all achieve superior performance. These findings support the feasibility of using deep learning models to accelerate DFT-level computations and suggest that the generated data is more learnable, thereby improving downstream model performance.

Table 12: MAE Performance of PointVector on DFT-based and HGEGNN-generated ED5-EC datasets with $\rho_\tau = 0.2$. 2024, 2025, 2026 represent seeds of training HGEGNN on original ED5-EC dataset.

	E1	E2	E3	E4	E5	E6	Mean
DFT	224.13±43.47	155.85±28.75	451.59±58.53	190.47±25.62	9.57±1.56	224.13±43.47	209.29
HGEGNN (2024)	195.48±2.77	137.69±11.48	408.40±10.61	172.98±7.30	**8.25±0.12**	195.48±2.77	186.38
HGEGNN (2025)	208.18±16.43	142.33±9.94	428.24±7.31	180.90±5.02	8.85±0.36	208.18±16.43	196.11
HGEGNN (2026)	**190.37±2.50**	**128.61±3.79**	**408.33±3.40**	**170.47±2.34**	8.35±0.04	**190.36±2.50**	**182.75**

A.3.7 Ablation study on thresholds and sampling points

Overview. Due to the substantial number of ED points and their direct influence on computational efficiency, it is crucial to study the effects of ED thresholds (ρ_τ) and sampling point counts (ξ) on model performance. Figure 10 shows the ablation results of PointVector under varying ρ_τ and ξ. We observe that performance does not improve proportionally with decreasing ρ_τ or increasing ξ, highlighting the importance of carefully selecting these hyperparameters to strike a balance between accuracy and computational cost.

Ablation Study on the ED Threshold ρ_τ. The ED threshold ρ_τ plays a critical role in representing electron density, as it governs the trade-off between model performance and computational efficiency.

In this ablation study, we evaluate the impact of ρ_τ using the PointVector model with a default number of sampled points $\xi = 2048$. However, when ρ_τ exceeds 0.05, the total number of ED points in some molecules falls below 2048, causing PointVector to fail due to insufficient input length. To minimize modifications to the original PointVector implementation, we set ξ to the minimum ED length across the dataset. Table 13 shows the MAE of PointVector on the ED5-EC dataset under various ED thresholds. The results indicate that the best performance is achieved at $\rho_\tau = 0.2$, with an average MAE of 209.29. This demonstrates that tuning ρ_τ can effectively balance accuracy and computational cost.

Table 13: Ablation study (ED5-EC dataset) of PointVector on ED threshold ρ_τ with MAE metric.

ρ_τ	ξ	E1	E2	E3	E4	E5	E6	Mean
0.05	2048	243.49±74.72	325.65±160.17	858.77±496.74	389.24±217.51	17.54±10.85	243.49±74.73	346.36
0.1	716	187.29±7.78	189.33±73.71	548.92±113.00	239.12±63.24	12.08±3.59	187.29±7.78	227.34
0.2	218	224.13±43.47	155.85±28.75	451.59±58.53	190.47±25.62	9.57±1.56	224.13±43.47	**209.29**
0.3	66	197.77±6.97	179.51±15.69	501.98±51.22	218.29±19.69	9.73±0.69	197.76±6.97	217.51
0.4	28	188.67±2.60	233.84±14.46	666.75±111.88	282.53±24.46	12.91±1.96	188.07±3.16	262.13

Ablation Study on the Number of Point Cloud Samples ξ. Point cloud-based methods (e.g., PointVector and X-3D) commonly adopt farthest point sampling (FPS) [50] to reduce the number of input points. Therefore, the number of sampled points, denoted as ξ, is a critical hyperparameter that directly affects both the model's capacity to capture spatial structures and its computational efficiency. A larger number of points allows the model to better represent the geometric details of ED, particularly in regions with ambiguous boundaries or sharp density gradients, facilitating the learning of fine-grained spatial features. However, increasing ξ also leads to higher memory consumption and longer training and inference times, especially when dealing with large-scale ED datasets. Therefore, choosing an appropriate number of points is essential to balance representational power and computational cost. To investigate this trade-off, we evaluate the performance of the point cloud-based PointVector model under different sample sizes $\xi = \{512, 1024, 2048, 4096, 8192\}$. Table 14 reports the results on the ED5-OE dataset. We observe that PointVector achieves the best performance when $\xi = 2048$, reaching an average MAE of 0.0248. Additionally, the model performance does not monotonically improve with increasing ξ. This may be attributed to the model's limited capacity—PointVector contains only 1.5454M parameters—which may constrain its ability to effectively leverage a large number of ED points. This observation highlights the need for more strong and ED-specialized architectures in future work.

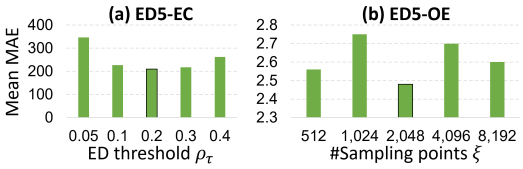

Figure 10: Ablation results of PointVector on (a) different ED thresholds ρ_τ and (b) different number of sampling points ξ.

Table 14: Ablation study (ED5-OE dataset) of PointVector on the number of sampling points ξ with MAE$\times$100 metric and $\rho_\tau = 0.05$.

ξ	HOMO-2	HOMO-1	HOMO-0	LUMO+0	LUMO+1	LUMO+2	LUMO+3	Mean
512	1.78±0.01	1.75±0.01	2.00±0.00	3.15±0.02	2.94±0.02	3.17±0.01	3.14±0.02	2.56
1024	1.79±0.01	1.74±1.98	3.19±2.99	3.19±0.02	2.99±0.02	3.19±0.02	3.14±0.01	2.75
2048	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	**2.48**
4096	1.87±0.09	1.76±0.07	2.01±0.02	3.40±0.07	3.21±0.12	3.38±0.19	3.29±0.13	2.70
8192	1.82±0.03	1.78±0.03	1.99±0.03	3.17±0.22	2.96±0.21	3.23±0.27	3.22±0.24	2.60

A.4 Details of the EDBench Database

A.4.1 Discussion on the Quality of the EDBench Database

To ensure the reliability and scientific utility of EDBench, we adopted a systematic and well-established protocol for electronic density (ED) calculation grounded in density functional theory (DFT) [32]. The entire workflow was designed to maximize both physical fidelity and computational robustness, while minimizing potential sources of error or bias.

First, all ED data were generated using Psi4 1.7, a widely used and validated open-source quantum chemistry package that supports high-accuracy ab initio and DFT calculations [51, 52, 53, 54].

We selected the B3LYP hybrid functional, a time-tested method known for its balance between computational efficiency and accuracy across a wide variety of molecules. Currently, B3LYP has been extensively applied in the domains of synthetic chemistry [55], molecular dynamics [56], phytochemistry [57], spectroscopy [58], medicine [53, 59] and physics [60]. This choice ensures that the resulting ED data reflect physically meaningful electron distributions rather than numerical artifacts.

Meanwhile, basis sets were systematically assigned based on molecular composition, employing 6-31G** for general cases and 6-31+G** for sulfur-containing molecules to capture diffuse electronic effects [61]. Compared with the basic 6-31G, 6-31G**/+G** provides the description of polarizable electron distribution and electron correlation effect by adding polarization function, which improves the processing ability of molecular polarization effect. This tailored approach enhances the accuracy of electron densities, particularly in chemically relevant regions such as lone pairs, -systems, or polarizable atoms.

In addition, the reference wavefunction was selected according to spin multiplicity, with restricted Hartree-Fock (RHF) applied to closed-shell systems and unrestricted Hartree-Fock (UHF) used for open-shell systems [62], in line with Hund's rule. This guarantees correct treatment of and spin components, reducing the risk of spin contamination and ensuring consistent modeling of open-shell species. To further control data quality, we enforced strict self-consistent field (SCF) convergence criteria before ED extraction [63, 64]. Electron density grids were then generated using a uniform grid spacing of 0.4 Bohr and a 4.0 Bohr padding, ensuring comprehensive spatial coverage without introducing undersampling or boundary artifacts. Additionally, a density fraction threshold of 0.85 was applied to focus on the physically relevant isosurface, filtering out low-density noise while preserving chemically meaningful features.

In sum, the quality of EDBench is supported by:

- A chemically sound and standardized computational protocol,
- Systematic and molecule-composition-based basis set selection,
- Accurate and consistent treatment of spin multiplicity,
- Rigorous convergence criteria and grid generation settings,
- Comprehensive spatial coverage and meaningful feature preservation.

These efforts collectively ensure that EDBench provides physically meaningful, reproducible, and high-resolution ED data at scale. We believe these safeguards sufficiently mitigate concerns of noise, bias, or low-quality samples, and position EDBench as a reliable benchmark for ED-aware machine learning research.

A.4.2 Example of ED visualization

Figure 11 illustrates the visualization of a molecule's electron density (ED) under varying threshold values ρ_τ. A higher ρ_τ retains only regions with a higher probability of electron presence. When $\rho_\tau = 0$, all possible electron positions are preserved, resulting in a dense, cuboid-like distribution. As ρ_τ increases, the number of ED points gradually decreases, and the molecular contour becomes more visually distinct.

It is worth noting that $\rho_\tau = 0$ leads to an overly dense ED representation, which poses challenges for both storage and computation. By tuning ρ_τ, we can achieve a balance between ED information retention and computational efficiency. In our experiments, we adopt $\rho_\tau \leq 0.2$ as a practical choice. As shown in Figure 11, this setting significantly reduces the number of ED points while preserving the essential ED structural features of the molecule.

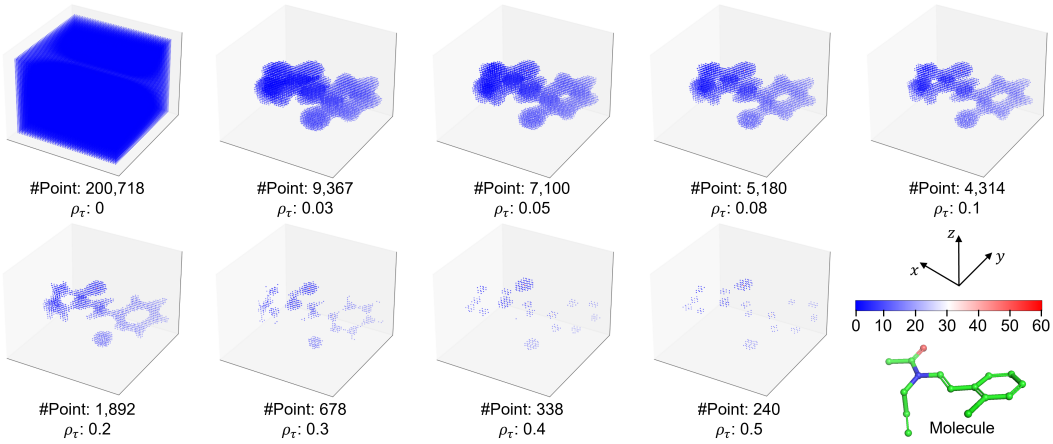

Figure 11: Example of ED visualization of a molecule with different thresholds ρ_τ. Point represents the number of ED points.