Electron Density-enhanced Molecular Geometry Learning

Hongxin Xiang[1], **Jun Xia**[2], **Xin Jin**[3], **Wenjie Du**[4], **Li Zeng**[1] and **Xiangxiang Zeng**[1,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]School of Engineering, Westlake University, Hangzhou, China
[3]Eastern Institute of Technology, Ningbo, China
[4]University of Science and Technology of China, Hefei, China

Abstract

Electron density (ED), which describes the probability distribution of electrons in space, is crucial for accurately understanding the energy and force distribution in molecular force fields (MFF). Existing machine learning force fields (MLFF) focus on mining appropriate physical quantities from the atom-level conformation to enhance the molecular geometry representation while ignoring the unique information from microscopic electrons. In this work, we propose an efficient **E**lectronic **D**ensity representation framework to enhance molecular **G**eometric learning (called EDG), which leverages images rendered from ED to boost molecular geometric representations in MLFF. Specifically, we construct a novel image-based ED representation, which consists of 2 million 6-view images with RGB-D channels, and design an ED representation learning model, called ImageED, to learn ED-related knowledge from these images. We further propose an efficient ED-aware teacher and introduce a cross-modal distillation strategy to transfer knowledge from the image-based teacher to the geometry-based students. Extensive experiments on QM9 and rMD17 demonstrate that EDG can be directly integrated into existing geometry-based models and significantly improves the capabilities of these models (e.g., SchNet, EGNN, SphereNet, ViSNet) for geometry representation learning in MLFF with a maximum average performance increase of 33.7%. Code and appendix are available at https://github.com/HongxinXiang/EDG

1 Introduction

Machine learning force fields (MLFF) is a computationally efficient and low-cost method for learning the interactions between atoms in molecular systems, bringing revolutionary advances to molecular dynamic simulations (MD) in many fields such as physics, chemistry and biology, and materials science [Chmiela *et al.*, 2017; Xiang *et al.*, 2024b; Wang *et al.*, 2024b]. Recent MLFF methods use geometric deep learning that represent atoms in molecular systems as nodes in a geometric graph and take into account physical symmetries have been shown to be effective in learning molecular force fields (MFF) [Liu *et al.*, 2022; Wang *et al.*, 2024a]. However, previous studies focused on mining physical quantities at the atomic level (such as coordinates, multi-body interactions, etc.) [Batzner *et al.*, 2022; Liao and Smidt, 2023; Wang *et al.*, 2024a], ignoring information at the electronic level.

Electron density (ED) is a core quantum mechanical property of the distribution of electrons within a molecule and is crucial for accurately predicting the quantum chemical properties of MFF [Sunshine *et al.*, 2023; Skogh *et al.*, 2024]. The application of ED faces two major challenges:

Challenge 1: High computational complexity of ED. Unlike the number of atoms in a molecule, which is usually on the scale of hundreds, the ED relies on a continuous spatial distribution and as the resolution increases, the number of data points may reach millions or even tens of millions (See Appendix A for details). As shown in Figure 1(a), there are two direct ways to represent ED: point cloud [Guo *et al.*, 2020] and voxel [Gong *et al.*, 2023]. As shown in Figure 1(b), we empirically show the limitations of point clouds and voxels as ED representations in energy prediction in force fields, efficiency of GPU memory and training (See Appendix B for details). In particular, point clouds and voxels are directly related to the resolution of the ED, so the computational efficiency will decrease as the resolution increases. The limitations above motivate us to propose a novel multi-view RGB-D image (the right subfigure of Figure 1(a)) for accurate and efficient ED representation [Xiang *et al.*, 2024a], which is independent of ED resolution and compresses the continuous ED signal in space into pixels. Compared with point clouds and voxels, the proposed images improve energy prediction ability, GPU memory efficiency, and training efficiency by 38.4%, 42.1%, and 4.8%, respectively.

Challenge 2: Expensive ED acquisition. The acquisition of ED data mainly relies on two types of technologies: experimental measurements and theoretical computations. Experimental measurements, such as X-ray crystallography [Nienaber *et al.*, 2000] or neutron diffraction [Goncharenko and Loubeyre, 2005], require high-precision instruments, sophisticated experimental setups, substantial

*Corresponding author (xzeng@hnu.edu.cn)

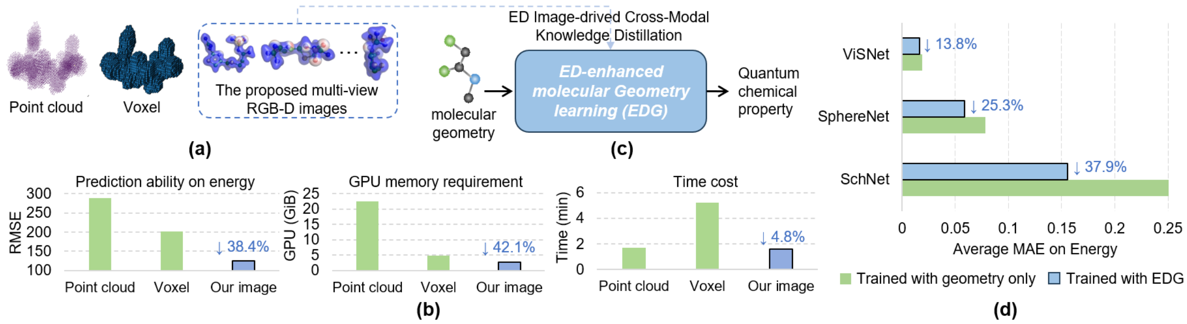

Figure 1: (a) ED representation methods. (b) RMSE in energy prediction, GPU memory, and training time cost of different ED representations with the same experimental setting. (c) The proposed EDG framework, which uses multi-view ED images with RGB-D channels to enhance molecular geometry learning. (d) Average MAE performance on 10 energy prediction datasets of rMD17 by using or not using the EDG.

time and technical expertise, making them resource-intensive. Theoretical computations, such as density functional theory (DFT) [Kohn and Sham, 1996], require a lot of computational time to obtain high-quality ED [Hegde and Bowen, 2017; Lee and Kim, 2024], which relies on high-performance computing clusters, resulting in significant cost. Given the limitations above and to improve data efficiency, as shown in Figure 1(c), we transform the ED-enhanced geometry learning problem into teaching excellent students (geometry) using an ED-aware teacher (image). Specifically, we first use DFT to obtain 2 million high-quality ED data and use them to train an ED representation learning model (called ImageED). Subsequently, we transfer the knowledge in ImageED to an ED-aware teacher that takes images without ED information as input, and use an ED-aware teacher to distill excellent geometry students (called EDG). This scheme allows us to obtain ED data only once and no more ED data is needed at any other time, which greatly improves the computational efficiency. As shown in Figure 1(d), student models (SchNet [Schütt *et al.*, 2017], SphereNet [Liu *et al.*, 2022] and ViSNet [Wang *et al.*, 2024a]) equipped with EDG achieve significant performance improvements.

We summarize the main contributions as follows:

- To the best of our knowledge, we are the first to exploit ED images to enhance molecular geometry learning.
- We propose an efficient multi-view ED images with RGB-D channels and design an ED representation learning method, called ImageED, to automatically extract ED-related features from images.
- We propose an ED-enhanced molecular geometry representation learning framework, called EDG, which equips with ED-aware teacher to improve the performance of a large number of geometry models.
- We show that our method achieves significantly better performance on 12 datasets from QM9 and 10 datasets from rMD17 and can substantially improve the performance of existing geometry representation models.

2 Related Work

Molecular Geometry Representation Learning. Geometry deep learning, which studies the interactions between atoms in molecular systems, is the key to the success of machine learning force fields (MLFF) [Liu *et al.*, 2022; Wang *et al.*, 2024a]. Recently, the main approaches have been to incorporate physical constraints such as roto-translational invariance of the geometry into the model architecture [Zaidi *et al.*, 2023; Wang *et al.*, 2024b], making the output features of the model invariant to the roto-translation of the molecule. Equivariant neural network (ENN) [Satorras *et al.*, 2021] is the most representative one and has been greatly developed in geometric representation learning. A simple way to achieve rotational-translational invariance is to construct invariant features based on the geometric conformation of the molecule, such as inter-atomic distances [Fuchs *et al.*, 2020], angles [Liu *et al.*, 2022], molecular descriptors [Todeschini and Consonni, 2009], etc. Besides these, there are many ENNs designed for invariance, such as models for modeling inter-atomic interactions [Schütt *et al.*, 2017; Satorras *et al.*, 2021; Liao and Smidt, 2023] and models for modeling multi-body interactions [Wang *et al.*, 2024b; Wang *et al.*, 2024a]. Our approach is agnostic to the model architecture, which can enhance any geometric representation learning model from a novel electronic perspective.

Electron Density Representation Learning. Existing electron density (ED) representation methods can be mainly divided into two categories: point cloud-based and voxel-based methods. The former considers all density values in the ED as a collection of points. For example, PointNet [Qi *et al.*, 2017] is used to classify the symmetry of inorganic compounds [Kim *et al.*, 2024]. The latter treats each point in the ED as a voxel. For example, 3D convolutional neural network (CNN) [Liu *et al.*, 2015] is used for the prediction of molecular exchange energy [Gong *et al.*, 2023] and discovery of guests of host molecules [Parrilla-Gutiérrez *et al.*, 2024], respectively. 3D-UNet [Çiçek *et al.*, 2016] is used to segment reactive sites in molecules and classify substances [Singh *et al.*, 2024]. Different from previous methods, we propose a novel multi-view RGB-D image to represent ED and design a representation learning method to extract features.

3 Our Method

3.1 Preliminaries

Background. Electron density (ED) is a key bridge to understand the prediction of energy and forces in molecular force fields (MFF). ED is not only the core output of quantum mechanical calculations, but also provides a solid theoretical basis for constructing high-precision MFF and understanding complex interactions between molecules. This demonstrates the significance of the proposed method to introduce ED into the learning framework of geometry representation. We provide more background details in the Appendix C.

Notation and Problem Formulation. The molecular geometry and the corresponding ground-truth labels with t prediction tasks are $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n} \in \mathbb{R}^k$ respectively, where $\mathcal{V} \in \mathbb{R}^{n_i^v \times (3+d_i^v)}$ (n_i^v, 3, d_i^v are the number of atoms, the coordinates of the atoms, and the feature dimensions of the atoms) and $\mathcal{E} \in \mathbb{R}^{n_i^v \times n_i^v \times d_i^e}$ (d_i^e is the feature dimensions of bonds). The corresponding multi-view ED image and structural image are $\mathcal{U} \in \mathbb{R}^{V \times 4 \times H \times W}$ and $\mathcal{S} \in \mathbb{R}^{V \times 3 \times H \times W}$ respectively, where V, H, W represent the number of views, the height, and the width of images, respectively. 3 and 4 represent RGB channels and RGB-D channels, respectively. This paper mainly defines three problems: (1) Pre-train a masked autoencoder (MAE) [He *et al.*, 2022] architecture consisting of an ED encoder f_{EDE} and an ED decoder f_{EDD} to learn useful representations $\mathcal{F}^{\mathcal{U}} \in \mathbb{R}^{d^U}$ (d^U is the dimension of features) from the ED image $\mathcal{U}$; (2) Pre-train an ED-aware teacher f_S and ED predictor f_{EDP} so that it can complete the mapping from structural images $\mathcal{S}$ to structural features $\mathcal{F}^{\mathcal{S}} \in \mathbb{R}^{d^S}$ and then to ED-related features $\mathcal{F}^{\mathcal{S}\to\mathcal{U}} \in \mathbb{R}^{d^U}$; (3) Distill a strong geometry student f_G using ED-aware teacher, ED predictor and mapper f_M. Among them, the geometry student f_G takes the molecular geometry as input and extracts the corresponding geometry features $\mathcal{F}^{\mathcal{G}} \in \mathbb{R}^{d^G}$ and the mapper is used to convert the geometry features into features recognized by ED predictor $\mathcal{F}^{\mathcal{G}\to\mathcal{S}} \in \mathbb{R}^{d^S}$ for distillation.

3.2 Overview of the Method

Here, we propose the **E**lectron **D**ensity-enhnaced molecular **G**eometry representation learning framework (called EDG). The overview of EDG is illustrated in Figure 2, which is divided into 4 main modules: (a) Given 2 million molecular conformations, necessary DFT data are generated and further processed together with the conformations into multi-view ED images with RGB-D channels (Section 3.3); (b) ImageED receives the 2 million multi-view ED image as input and utilizes two pre-training tasks to learn ED-related knowledge (Section 3.4); (c) We design an ED-aware teacher, which is optimized by minimizing the difference between the predicted ED features from the structural images and the true ED features from the ImageED on 2 million molecules (Section 3.5). (d) The ED-aware teacher is used to distill a strong geometry student by using ED predictor and mapper (Section 3.6). We summarize the main processed in Appendix D.

3.3 Generation of RGB-D Electron Density Images

As shown in Figure 2(a), we first obtain 2 million molecular conformations from PCQM4Mv2 [Hu *et al.*, 2021] and use density functional theory (DFT) with the basis set of 6-31G**/+G** and the exchange-correlation functional of B3LYP to generate DFT data for these molecular conformations [Sud, 2016]. For each molecule, the DFT data includes an electrostatic potential (ESP) file and an ED file stored in the form of a three-dimensional grid. Next, we describe the main details of ED image generation. The structural loader uses the command *load* {*conformation file*} in PyMol [DeLano and others, 2002] to load the structural information from the molecular conformation file. ED loader uses the commands *load* {*ED file*}, *ED; load* {*ESP file*}, *ESP; ramp_new legend, ESP, [-0.08, 0, 0.08], [red, white, blue]; isosurface surface, ED, 0.05; set surface_color, legend, surface* in PyMol to load the ED information, which uses a red-white-blue distribution range to describe the electrostatic potential distribution of ED and red, white, and blue represent positive, neutral, and negative regions, respectively. Finally, multi-view joint render uses commands *set transparency, 0.4; turn* {*axis*}, {*angle*}; *png* {*path*}, *width*={*width*}, *height*={*height*} in PyMol to render the ED information and structural information into multi-view RGB-D ED images $\mathcal{U} \in \mathbb{R}^{6\times4\times224\times224}$. Specifically, $\{axis\}$ and $\{angle\}$ represent the rotation angle degrees along axis and we set $(\{axis\}, \{angle\})$ to $(x, 0), (x, 180), (x, 90), (x, -90), (y, 90), (y, -90)$, which means generating images from 6 different views. $\{width\} = 224$ and $\{height\} = 224$ represent the width and height of the rendered image and $\{path\}$ represents the path where the image is saved. We describe more details of ED image rendering in Appendix E.

3.4 ED Representation Learning with ImageED

The proposed ED images have two properties: (1) RGB and D channels use color to represent the distribution of ED and depth to represent the spatial layout, respectively, which indicates that each pixel has a clear physical meaning; (2) the distribution of ED is continuous, which means that ED can be predicted by the context. Therefore, we propose a novel ED representation learning framework (called ImageED) with mask prediction and restoration prediction tasks to learn pixel-level local and contextual ED information from 2 million ED images. ImageED is an encoder-decoder architecture, which is built following the ViT-Base/16 [Dosovitskiy *et al.*, 2020] of MAE [He *et al.*, 2022]. For a given batch (n molecules) of ED images $u \in \mathbb{R}^{n\times c\times V\times H\times W}$ (where $V, c, H, W = 6, 4, 224, 224$), we first use a view-agnostic patch embed layer with a patch size of $n_p = 16$ to transform u into a pile of multi-view tokens mt.

$$mt_{i,j} = u[:, :, n_p i : n_p(i+1), n_p j : n_p(j+1)] \quad (1)$$

We assume $n_t = H/n_p = W/n_p$ and $mt = \{mt_{i,j} | i, j \in \{0, 1, ..., n_t\}\} \in \mathbb{R}^{n\times V\times n_t^2\times(n_p^2\times c)}$. Next, we add the positional embeddings and expand along the view to get tokens $t = \in \mathbb{R}^{n\times(V\times n_t^2)\times(n_p^2\times c)}$. We shuffle the order of tokens and randomly mask 25% of the tokens to obtain the masked tokens $t^m \in \mathbb{R}^{n\times(0.25\times V\times n_t^2)\times(n_p^2\times c)}$ and unmasked tokens $t^u \in \mathbb{R}^{n\times(0.75\times V\times n_t^2)\times(n_p^2\times c)}$ respectively. We input the t^u into the ED encoder f_{EDE} to get the encoded tokens:

$$h^u = f_{EDE}(t^u), h^u \in \mathbb{R}^{n\times(0.75\times V\times n_t^2)\times d_{\mathcal{U}}} \quad (2)$$

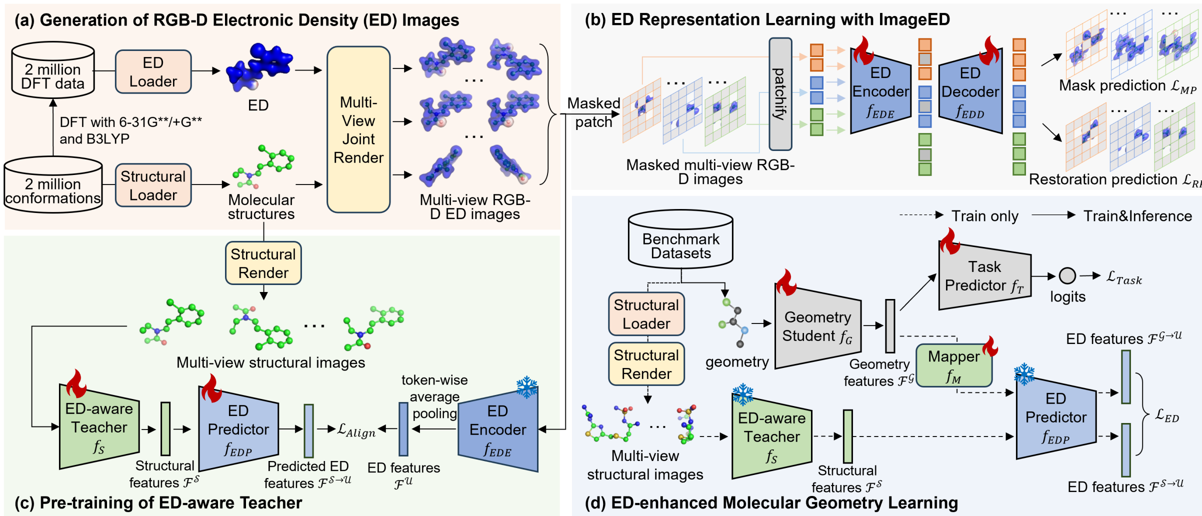

Figure 2: Overview of the proposed EDG framework. **(a)** Multi-view ED images with RGB-D channels are generated based on 2 million molecular conformers and DFT data, which contains the structural and ED information of the molecule. **(b)** ImageED with masked autoencoder (MAE) architecture and 2 pretext tasks ($\mathcal{L}_{MP}$ and $\mathcal{L}_{RP}$) is pretrained to extract ED-related features from multi-view ED images in (a). **(c)** The ED-aware teacher accepts the structural images as input and learns to transform it into ED features $\mathcal{F}^{\mathcal{S}\to\mathcal{U}}$ using ED predictor and ED encoder in (b). **(d)** In downstream tasks, the ED-aware teacher and ED predictor from (c) are frozen to enhance the geometry student. Note that there is no need for explicit involvement of electron density here because the ED teacher has the ability to extract ED-related information.

where $d_{\mathcal{U}}$ represents the feature dimension of h^u. Afterwards, we use additional mask tokens $\hat{h}^m$ that is initialized by 0 to the encoded tokens to obtain all encoded tokens h:

$$h = \pi(h^u \parallel \hat{h}^m) + pos \in \mathbb{R}^{n\times(V\times n_t^2)\times(n_p^2\times c)} \tag{3}$$

where π represents arrange them in the order of the images and pos represents the positional embeddings. The final predicted masked tokens $\hat{t}^m$ and unmasked tokens $\hat{t}^u$ can be obtained by input h into the ED Decoder f_{EDD}.

In order to optimize the f_{EDE} and f_{EDD} in ImageED, we define a mask prediction task $\mathcal{L}_{MP}$:

$$\mathcal{L}_{MP} = \frac{1}{n}\sum_{i=1}^{n} sim(t_i^u, \hat{t}_i^u) \tag{4}$$

where $sim()$ represents the Euclidean distance. However, we find that using only mask prediction tasks will limit the understanding of local features. We further introduce the restoration prediction task $\mathcal{L}_{RP}$:

$$\mathcal{L}_{RP} = \frac{1}{n}\sum_{i=1}^{n} sim(t_i^m, \hat{t}_i^m) \tag{5}$$

Finally, the overall loss function of ImageED is formulated:

$$\mathcal{L}_{ImageED} = \lambda_{MP}\mathcal{L}_{MP} + \lambda_{RP}\mathcal{L}_{RP} \tag{6}$$

where λ_{MP} and λ_{RP} is the balance coefficient and we set them to 1. After pre-training ImageED with 2 million molecules, we use the f_{EDE} to extract ED features.

3.5 Pretraining of ED-aware Teacher

Since the generation of ED data requires a lot of computing resources, it is resource-intensive and impractical to calculate ED data for every downstream task. Therefore, we hope to use an easily accessible intermediary to replace ED to generate ED features. We assume that the molecular conformation, ED data generated by DFT, and ED features are s, x, h, respectively. The acquisition of ED features follows this path: $s \xrightarrow{\text{DFT}} x \xrightarrow{\text{ImageED}} h$, which shows that when s is known, x and h can be obtained. Therefore, according to the probability chain rule, the joint probability distribution $p(h, x|s)$ can be decomposed into: $p(h, x|s) = p(h|x, s) \cdot p(x|s)$. According to the Markov hypothesis [Markov, 1960], we can get: $p(h|x, s) = p(h|x)$. In order to simplify $p(h, x|s)$ to a probability distribution that is independent of x, we further marginalize x (integrating with respect to x) and get:

$$p(h|s) = \int p(h|x) \cdot p(x|s)dx \tag{7}$$

Therefore, we can approximate $p(h|x) \cdot p(x|s)$ by directly learning a $p(h|s)$. Here, we propose an ED-aware teacher $f_{\mathcal{S}}$ and ED predictor f_{EDP} as $p(h|s)$ to learn the mapping of ED features directly from molecular conformations. We choose multi-view structural images as the input of the $f_{\mathcal{S}}$ (See Appendix F for specific reasons). Specifically, structural render uses command template *turn* $\{axis\}$, $\{angle\}$; *png* $\{path\}$ in PyMol to render molecular structures from 2 million conformations into multi-view images $\mathcal{S}$. Considering computational efficiency, we generate 4 views here and $(\{axis\}, \{angle\})$ is set to $(x, 0), (x, 180), (y, 180), (z, 180)$. $f_{\mathcal{S}}$ uses ResNet18 [He *et al.*, 2016] with a view-wise average pooling and f_{EDP} is a Multilayer Perceptron (MLP) with Linear Layer→Softplus Activator→Linear Layer. Given a batch (n molecules) of structural images s from $\mathcal{S}$ and ED images u from $\mathcal{U}$, we obtain structural features $\mathcal{F}^{\mathcal{S}}$ and ED features $\mathcal{F}^{\mathcal{S}\to\mathcal{U}}$:

$$\mathcal{F}^{\mathcal{S}\to\mathcal{U}} = f_{EDP}(\mathcal{F}^{\mathcal{S}}); \mathcal{F}^{\mathcal{S}} = f_{\mathcal{S}}(s) \tag{8}$$

Next, we freeze the ED encoder that accepts ED images u as input and use a token-wise average pooling to convert the token features output by the ED encoder into ED features $\mathcal{F}^{\mathcal{U}}$. Finally, we take $\mathcal{F}^{\mathcal{U}}$ as the ground-truth and train the ED-aware teacher and ED predictor to learn the mapping from structural features to ED features on 2 million molecules. The loss function $\mathcal{L}_{align}$ is defined as:

$$\mathcal{L}_{align} = L1(\mathcal{F}^{\mathcal{S}\to\mathcal{U}}, \mathcal{F}^{\mathcal{U}}) \tag{9}$$

where $L1$ represent L1 distance. With the ED-aware teacher, the costly ED image can be replaced by a cheaper structural image, significantly reducing DFT-related costs.

3.6 ED-enhanced Molecular Geometry Learning

In the training stage of downstream tasks, we first convert the molecular conformations in the dataset into geometric data and multi-view structural images using structural loader and structural render. Subsequently, geometry data and images are input into the geometry student f_G and frozen ED-aware teacher to extract features $\mathcal{F}^{\mathcal{G}}$ and $\mathcal{F}^{\mathcal{S}}$, respectively. Please note that f_G can be any geometry-based model, such as SchNet [Schütt *et al.*, 2017], EGNN [Satorras *et al.*, 2021], etc. Next, a mapper f_M is used to map the geometry features into the structural space to obtain $f_M(\mathcal{F}^{\mathcal{G}})$. Subsequently, the frozen ED predictor accepts $f_M(\mathcal{F}^{\mathcal{G}})$ and $\mathcal{F}^{\mathcal{S}}$ as input and obtains predicted ED features:

$$\mathcal{F}^{\mathcal{G}\to\mathcal{U}} = f_{EDP}(f_M(\mathcal{F}^{\mathcal{G}})); \mathcal{F}^{\mathcal{S}\to\mathcal{U}} = f_{EDP}(\mathcal{F}^{\mathcal{S}}) \tag{10}$$

To distill the ED knowledge from the teacher model to the student model, we define a consistency loss $\mathcal{L}_{ED}$:

$$\mathcal{L}_{ED} = SL1(\mathcal{F}^{\mathcal{G}\to\mathcal{U}}, \mathcal{F}^{\mathcal{S}\to\mathcal{U}}) \tag{11}$$

where $SL1$ represents smooth L1 distance [Girshick, 2015]. In order to obtain task-related labels, we define a task predictor f_T, which accepts geometry features $\mathcal{F}^{\mathcal{G}}$ to output task-related logits $\hat{y} = f_T(\mathcal{F}^{\mathcal{G}})$. The task-related loss function is defined as:

$$\mathcal{L}_{Task} = L1(\hat{y}, y) \tag{12}$$

The final loss of EDG is formulated:

$$L_{EDG} = \mathcal{L}_{Task} + \lambda\mathcal{L}_{ED}, \tag{13}$$

where λ is the balanced coefficient. In the inference phase, the prediction result is obtained by sequentially inputting the geometry data into the student network f_G and the task predictor f_T. Therefore, images are only involved in the training of the model and are not needed during inference, which further improve the efficiency.

4 Experiments and Results

4.1 Experimental Settings

Datasets and evaluation protocol. To pre-train ImageED, the ED-aware teacher, and the ED predictor, we select the first 2 millions unlabeled molecular conformations and their DFT-computed ED data from the EDBench database [Xiang *et al.*, 2025], generated by Psi4 software [Turney *et al.*, 2012] with a grid spacing of 0.4. In evaluation stage, we select 12 widely used tasks related to quantum mechanic properties from QM9 [Ramakrishnan *et al.*, 2014] and 10 common tasks related to energy/force from revised MD17 (rMD17) [Christensen and Von Lilienfeld, 2020]. It is worth noting that for the force prediction, we first predict the molecular energy and use the gradient of each node position as the force, that is, *force=-torch.autograd.grad(outputs=energy, inputs=positions)* in PyTorch [Paszke *et al.*, 2019]. The dataset split follows Geom3D [Liu *et al.*, 2024], i.e., using 110K for training, 10K for validation, and 11K for testing in QM9 and 950 for training, 50 for validation, and 1000 for testing in rMD17. We use mean absolute error (MAE) as evaluation metric.

Baselines. To verify the effectiveness of EDG, we select many geometry-based models with different architectures, such as SchNet [Schütt *et al.*, 2017], EGNN [Satorras *et al.*, 2021], Equiformer [Liao and Smidt, 2023], SphereNet [Liu *et al.*, 2022], ViSNet [Wang *et al.*, 2024a], as geometry students to verify the generalizability of EDG. Following [Liu *et al.*, 2022; Wang *et al.*, 2024a; Liu *et al.*, 2024], we ensure that each baseline is fully trained. For example, SchNet, EGNN, and SphereNet are trained for 1,000 epochs with a learning rate of 5e-4; Equiformer is trained for 300 epochs with a learning rate of 5e-4; and ViSNet is trained for 3,000 epochs with a learning rate of 0.0002. The batch size of SchNet, EGNN, SphereNet, and Equiformer is set to 128 in QM9 and 1 in rMD17, the batch size of ViSNet is set to 4 in rMD17.

Implementation details. The encoder and decoder of ImageED are built based on ViT-Base/16. In pre-training of ImageED on 2 million ED molecules, we use a learning rate of 1.5e-4, a batch size of 64, a mask ratio of 0.25, λ_{MP} and λ_{RP} of 1 for 20 epochs on 8 GeForce RTX 4090 (See Appendix G for more details). In pre-training of ED-aware teacher on 2 million molecules, we divide 2% as validation set and the rest as training set. We use a learning rate of 5e-3 and a batch size of 128 to train the ED-aware teacher and ED predictor for about 280k steps (See Appendix H for more details). In distillation stage of EDG, we select hyper-parameters λ from 1e-4 and 5e-4 to 1.0 with a 10x increasing in steps. Following [Wang *et al.*, 2024a; Liu *et al.*, 2024], we run the experiments with exactly the same parameter settings as the baselines and report test scores corresponding to the best validation performance. The mapper and task predictor consists of a simple linear layer.

4.2 Main Results

We first evaluate the performance of EDG on the 12 quantum properties from QM9 with 4 baselines (SchNet, EGNN, Equiformer, SphereNet) and Table 1 shows the main results. We find baselines equipped with EDG achieve the best performance. We observe that regardless of the architecture, the baselines after equipping EDG achieve consistent performance improvement with a relative performance increase ranging from 2.2% to 6.4% in average MAE performance. Except for the property U in Equiformer, all other performance are enhanced.

Model	$\alpha\downarrow$ α_0^3	$\nabla\mathcal{E}\downarrow$ meV	$\mathcal{E}_{\text{HOMO}}\downarrow$ meV	$\mathcal{E}_{\text{LUMO}}\downarrow$ meV	$\mu\downarrow$ D	$C_v\downarrow$ $\frac{cal}{mol\cdot K}$	$G\downarrow$ meV	$H\downarrow$ meV	$R^2\downarrow$ α_0^2	$U\downarrow$ meV	$U_0\downarrow$ meV	ZPVE $\downarrow$ meV
SchNet	0.07021	50.829	31.952	26.168	0.03013	0.03228	14.678	14.090	0.13455	14.142	13.915	1.714
EDG-SchNet	0.06866	49.778	31.884	25.972	0.02980	0.03162	14.022	13.841	0.12458	13.794	13.826	1.688
Δ	↑2.2%	↑2.1%	↑0.2%	↑0.7%	↑1.1%	↑2.0%	↑4.5%	↑1.8%	↑7.4%	↑2.5%	↑0.6%	↑1.5%
EGNN	0.06474	49.493	29.865	24.696	0.02981	0.03125	11.057	10.596	0.07494	11.013	10.150	1.519
EDG-EGNN	0.06147	46.979	28.319	24.283	0.02655	0.03078	10.708	10.298	**0.07225**	9.985	10.012	1.498
Δ	↑5.1%	↑5.1%	↑5.2%	↑1.7%	↑10.9%	↑1.5%	↑3.2%	↑2.8%	↑3.6%	↑9.3%	↑1.4%	↑1.4%
Equiformer	0.06762	46.308	26.017	23.681	0.02074	0.02733	18.439	16.453	0.45828	15.339	23.928	1.537
EDG-Equiformer	0.06476	45.813	25.492	23.266	**0.01985**	0.02642	15.976	14.451	0.43947	15.466	16.517	1.529
Δ	↑4.2%	↑1.1%	↑2.0%	↑1.8%	↑4.3%	↑3.3%	↑13.4%	↑12.2%	↑4.1%	↓0.8%	↑31.0%	↑0.5%
SphereNet	0.04670	40.129	22.007	19.435	0.02689	0.02437	7.875	7.199	0.25821	6.999	6.641	1.253
EDG-SphereNet	**0.04592**	**39.694**	**21.842**	**19.014**	0.02648	**0.02376**	**7.769**	**6.283**	0.24935	**6.502**	**6.101**	**1.206**
Δ	↑1.7%	↑1.1%	↑0.7%	↑2.2%	↑1.5%	↑2.5%	↑1.3%	↑12.7%	↑3.4%	↑7.1%	↑8.1%	↑3.8%

Table 1: The mean absolute error (MAE) performance of different methods on 12 quantum mechanics prediction tasks in QM9. Δ represents the relative improvement percentage calculated by $(1 - \frac{w/EDG}{w/oEDG}) \times 100$.

Model	Aspirin ↓	Azobenzene ↓	Benzene ↓	Ethanol ↓	Malona. ↓	Naphth. ↓	Paracetamol ↓	Salicylic ↓	Toluene ↓	Uracil ↓
SchNet	0.73909	0.39678	0.02052	0.12516	0.16142	0.21158	0.37097	0.19078	0.20797	0.07872
EDG-SchNet	0.35525	0.33441	0.01711	0.06061	0.11181	0.07338	0.28303	0.15697	0.08780	0.07433
Δ	↑51.9%	↑15.7%	↑16.6%	↑51.6%	↑30.7%	↑65.3%	↑23.7%	↑17.7%	↑57.8%	↑5.6%
SphereNet	0.18091	0.09794	0.00647	0.03784	0.06005	0.03823	0.10425	0.14119	0.03452	0.08088
EDG-SphereNet	0.13622	0.06788	0.00413	0.03575	0.05659	0.02753	0.09934	0.09569	0.02413	0.03683
Δ	↑24.7%	↑30.7%	↑36.2%	↑5.5%	↑5.8%	↑28.0%	↑4.7%	↑32.2%	↑30.1%	↑54.5%
ViSNet	0.05547	0.02081	0.00627	0.01095	0.01517	0.01313	0.02700	0.01966	0.01089	0.01238
EDG-ViSNet	**0.04650**	**0.01838**	**0.00616**	**0.00990**	**0.01395**	**0.01178**	**0.02491**	**0.01906**	**0.00998**	**0.01188**
Δ	↑16.2%	↑11.7%	↑1.8%	↑9.6%	↑8.0%	↑10.2%	↑7.8%	↑3.0%	↑8.3%	↑4.0%

Table 2: The MAE performance of different methods on 10 energy ($\frac{kcal}{mol}$) prediction tasks in rMD17. Malona. and Naphth. represents Malonaldehyde, Naphthalene, respectively. Δ represents the relative improvement percentage calculated by $(1 - \frac{w/EDG}{w/oEDG}) \times 100$.

In order to verify the effectiveness of EDG in more tasks, we further evaluate on 10 energy/force prediction tasks from rMD17 with 3 baselines (SchNet, SphereNet, ViSNet). Table 2 and Table 3 show the prediction performance on energy and force respectively. We find the same conclusion as on the QM9 benchmark, that is, EDG improves the performance of all baselines with average MAE performance improvements ranging from 8.1% to 33.7% on energy and 1.5% to 5.3% on force. It is worth noting that EDG has a larger improvement on energy prediction than on force. This is because ED can capture the global energy distribution, while force, as an energy gradient, depends on local atomic interactions, making the improvement brought by ED is not as obvious as energy. In any case, the performance improvements on energy and force prove the effectiveness of EDG. In addition, we also visualize absolute value of the difference between the predicted energy y_{pred} and the ground-truth y_{true} for all trajectories in the test set by showing the absolute value of the difference between them in Figure 3, which shows that EDG outperforms the baselines in energy prediction for almost all trajectories.

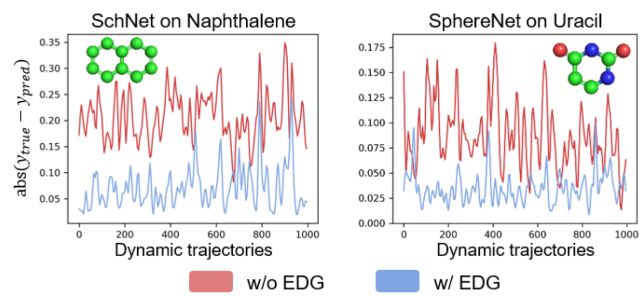

Figure 3: The visualization of SchNet on Naphthalene task and SphereNet on Uracil task. The y axis represents the absolute value of the difference between y_{pred} and y_{true} on the test set.

4.3 Hyperparameters Analysis

λ in Formula 13 is a parameter used to control the strength of distilling knowledge from the ED-aware teacher into the geometry students and a larger value will force the student to learn more knowledge from the teacher. Figure 4 shows the line figures of the performance of EGNN and ViSNet with different $\mathcal{L}_{ED}$ on QM9 and rMD17, respectively. Overall, we find that EDG can improve the performance of baselines to varying degrees with $\mathcal{L}_{ED}$. For example, on the aspirin task, the performance gain of EDG fluctuates from 8.2% to 16.2% with the adjustment of $\mathcal{L}_{ED}$. In addition, we also find several patterns: on the α and U tasks, as $\mathcal{L}_{ED}$ increases, the performance decreases overall; on the $\nabla\mathcal{E}$, $\mathcal{E}_{\text{LUMO}}$, aspirin, and malonaldehyde tasks, the performance curve changes with $\mathcal{L}_{ED}$ in a U-shaped manner. These findings suggest that by tuning the appropriate $\mathcal{L}_{ED}$, EDG can better enhance the baselines.

4.4 Results of ED images on Energy-related Tasks

Here, we describe the advantages of the proposed ED image on energy-related tasks. We sample 10,000 molecules from 2

Model	Aspirin ↓	Azobenzene ↓	Benzene ↓	Ethanol ↓	Malona. ↓	Naphth. ↓	Paracetamol ↓	Salicylic ↓	Toluene ↓	Uracil ↓
SchNet	1.04245	0.90082	0.18569	0.38519	0.65536	0.39851	0.82544	0.77487	0.48322	0.51399
EDG-SchNet	1.04910	0.91692	0.17113	0.37901	0.64678	0.39509	0.83296	0.74308	0.47790	0.50783
Δ	↓0.6%	↓1.8%	↑7.8%	↑1.6%	↑1.3%	↑0.9%	↓0.9%	↑4.1%	↑1.1%	↑1.2%
SphereNet	0.39134	0.21776	0.02151	0.19432	0.29278	0.11141	0.32265	0.28692	0.10978	0.27702
EDG-SphereNet	0.38598	0.21665	0.02101	0.18741	0.28456	0.11102	0.32023	0.28299	0.10804	0.17179
Δ	↑1.4%	↑0.5%	↑2.3%	↑3.6%	↑2.8%	↑0.3%	↑0.8%	↑1.4%	↑1.6%	↑38.0%
ViSNet	0.15164	0.05729	0.00656	0.05688	0.09275	0.02808	**0.10488**	0.08348	0.02980	0.05252
EDG-ViSNet	**0.14996**	**0.05691**	**0.00647**	**0.05558**	**0.08992**	**0.02798**	0.10599	**0.08107**	**0.02780**	**0.05100**
Δ	↑1.1%	↑0.7%	↑1.3%	↑2.3%	↑3.0%	↑0.3%	↓1.1%	↑2.9%	↑6.7%	↑2.9%

Table 3: The MAE performance of different methods on 10 force ($\frac{kcal}{mol \cdot Å}$) prediction tasks in rMD17. Malona. and Naphth. represents Malonaldehyde, Naphthalene, respectively. Δ represents the relative improvement percentage calculated by $(1 - \frac{w/EDG}{w/oEDG}) \times 100$.

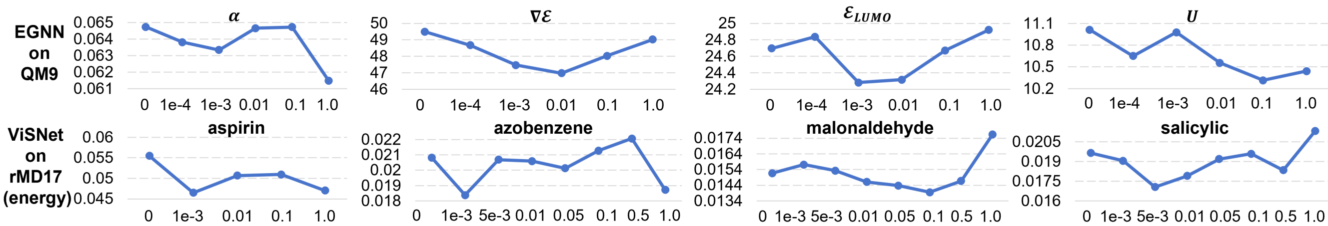

Figure 4: Performance of EDG with different $\mathcal{L}_{ED}$. The x axis and y axis represent the value of $\mathcal{L}_{ED}$ and the corresponding MAE performance, respectively. $\mathcal{L}_{ED} = 0$ means that EDG is not used.

million DFT data and predict the energy of the molecular system given the ED information. We use exactly the same experimental settings and hyperparameters and randomly split the dataset into training/validation/test sets with an 8:1:1 ratio for evaluation. More settings see Appendix B. For each ED representation, we select the corresponding popular encoder to extract features, such as point cloud-based PointNet [Qi *et al.*, 2017], voxel-based ResNet3D [Hara *et al.*, 2018] and image-based ResNet18 [He *et al.*, 2016]. As shown in Table 4, we find the proposed ED iamge achieves the best performance on 6 energy-related tasks with a relative performance gain ranging from 7.8% to 71.6%, which demonstrates the effectiveness of image as a representation of ED and that 2D images are easier to learn compared to 3D representations.

Models	E1	E2	E3	E4	E5	E6
Point	275.1	168.6	557.7	244.8	14.5	288.9
Voxel	121.1	313.9	947.8	271.2	7.9	202.0
Image	**111.6**	**47.9**	**349.8**	**85.6**	**4.4**	**124.5**
Δ	↑7.8%	↑71.6%	↑37.3%	↑65.0%	↑44.5%	↑38.4%

Table 4: RMSE (Root Mean Squared Error) performance of different ED representations on 6 energy prediction tasks. Point (point cloud), voxel, and image use PointNet, ResNet3D, and ResNet18 as encoders, respectively. E1-E6 represent DF-RKS Final Energy, Nuclear Repulsion Energy, One-Electron Energy, Two-Electron Energy, DFT Exchange-Correlation Energy, and Total Energy, respectively. Δ represents the relative performance gain of the image compared to the best other results.

4.5 Visualization of ImageED

As shown in Figure 5, we find that ImageED can generate ED images well compared original images, which indicates that ImageED can learn ED-related knowledge well. In addition, we find that simply applying the masked prediction task (ImageED w/o $\mathcal{L}_{RP}$) will limit the understanding of ImageED in local pixels, which shows the importance of restoration prediction task in ImageED. We show more examples in the Appendix I.

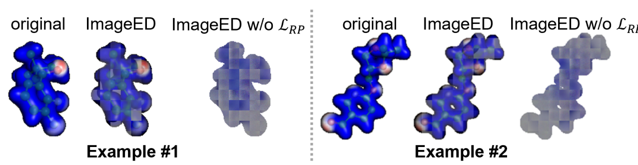

Figure 5: Several examples of ImageED output visualizations.

5 Conclusion

In this work, we propose a novel **ED**-enhanced molecular **G**eometry representation learning framework (called EDG), which is the first attempt to exploit ED images to improve the performance of geometry-based methods. We propose an efficient ED representation learning, called ImageED, to extract ED knowledge from images and further transfer the knowledge in ImageED to an ED-aware teacher to save the cost of DFT. By exploiting ED-aware teacher, EDG can significantly improve the performance of geometry-based methods without any architectural modifications on a large number of quantum chemical benchmarks. In addition, we experimentally show that using ED images can more accurately predict energy-related prediction tasks while saving memory and computational costs, enabling direct use of ED images in broader tasks like drug discovery and materials science.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant nos. U22A2037, 62425204, 62122025, 62450002, 62432011), Grants of Ningbo 2023CX050011.

References

[Batzner *et al.*, 2022] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[Chmiela *et al.*, 2017] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[Christensen and Von Lilienfeld, 2020] Anders S Christensen and O Anatole Von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4):045018, 2020.

[Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[DeLano and others, 2002] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[Fuchs *et al.*, 2020] 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.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[Goncharenko and Loubeyre, 2005] Igor Goncharenko and Paul Loubeyre. Neutron and x-ray diffraction study of the broken symmetry phase transition in solid deuterium. *Nature*, 435(7046):1206–1209, 2005.

[Gong *et al.*, 2023] Weiyi Gong, Tao Sun, Hexin Bai, Peng Chu, Anoj Aryal, Jie Yu, Haibin Ling, John P Perdew, Qimin Yan, et al. Incorporation of density scaling constraint in density functional design via contrastive representation learning. *Digital Discovery*, 2(5):1404–1413, 2023.

[Guo *et al.*, 2020] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

[Hara *et al.*, 2018] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Hegde and Bowen, 2017] Ganesh Hegde and R Chris Bowen. Machine-learned approximations to density functional theory hamiltonians. *Scientific reports*, 7(1):42669, 2017.

[Hu *et al.*, 2021] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *NeurIPS*, 34, 2021.

[Kim *et al.*, 2024] Seonghwan Kim, Byung Do Lee, Min Young Cho, Myoungho Pyo, Young-Kook Lee, Woon Bae Park, and Kee-Sun Sohn. Deep learning for symmetry classification using sparse 3d electron density data for inorganic compounds. *npj Computational Materials*, 10(1):211, 2024.

[Kohn and Sham, 1996] Walter Kohn and L Sham. Density functional theory. In *Conference Proceedings-Italian Physical Society*, volume 49, pages 561–572. Editrice Compositori, 1996.

[Lee and Kim, 2024] Ryong-Gyu Lee and Yong-Hoon Kim. Convolutional network learning of self-consistent electron density via grid-projected atomic fingerprints. *npj Computational Materials*, 10(1):248, 2024.

[Liao and Smidt, 2023] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023.

[Liu *et al.*, 2015] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 806–814, 2015.

[Liu *et al.*, 2022] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *International Conference on Learning Representations*, 2022.

[Liu *et al.*, 2024] 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. *NeurIPS*, 36, 2024.

[Markov, 1960] Andrei Andreyevich Markov. The theory of algorithms. *Am. Math. Soc. Transl.*, 15:1–14, 1960.

[Nienaber *et al.*, 2000] Vicki L Nienaber, Paul L Richardson, Vered Klighofer, Jennifer J Bouska, Vincent L Giranda, and Jonathan Greer. Discovering novel ligands for macromolecules using x-ray crystallographic screening. *Nature biotechnology*, 18(10):1105–1108, 2000.

[Parrilla-Gutiérrez *et al.*, 2024] Juan M Parrilla-Gutiérrez, Jarosław M Granda, Jean-François Ayme, Michał D Bajczyk, Liam Wilbraham, and Leroy Cronin. Electron density-based gpt for optimization and suggestion of host–guest binders. *Nature computational science*, 4(3):200–209, 2024.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

[Qi *et al.*, 2017] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[Ramakrishnan *et al.*, 2014] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[Satorras *et al.*, 2021] 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.

[Schütt *et al.*, 2017] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[Singh *et al.*, 2024] Satnam Singh, Gina Zeh, Jessica Freiherr, Thilo Bauer, Isik Türkmen, and Andreas T Grasskamp. Classification of substances by health hazard using deep neural networks and molecular electron densities. *Journal of Cheminformatics*, 16(1):45, 2024.

[Skogh *et al.*, 2024] Mårten Skogh, Werner Dobrautz, Phalgun Lolur, Christopher Warren, Janka Biznárová, Amr Osman, Giovanna Tancredi, Jonas Bylander, and Martin Rahm. The electron density: a fidelity witness for quantum computation. *Chemical Science*, 15(6):2257–2265, 2024.

[Sud, 2016] Manish Sud. Mayachemtools: an open source package for computational drug discovery. *Journal of chemical information and modeling*, 56(12):2292–2297, 2016.

[Sunshine *et al.*, 2023] Ethan M Sunshine, Muhammed Shuaibi, Zachary W Ulissi, and John R Kitchin. Chemical properties from graph neural network-predicted electron densities. *The Journal of Physical Chemistry C*, 127(48):23459–23466, 2023.

[Todeschini and Consonni, 2009] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*. John Wiley & Sons, 2009.

[Turney *et al.*, 2012] 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.

[Wang *et al.*, 2024a] 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.

[Wang *et al.*, 2024b] 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, 2024.

[Xiang *et al.*, 2024a] 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*, 2024.

[Xiang *et al.*, 2024b] 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.

[Xiang *et al.*, 2025] Hongxin Xiang, Ke Li, Mingquan Liu, Zhixiang Cheng, Bin Yao, Wenjie Du, Jun Xia, Li Zeng, Xin Jin, and Xiangxiang Zeng. Edbench: Large-scale electron density data for molecular modeling. *arXiv preprint arXiv:2505.09262*, 2025.

[Zaidi *et al.*, 2023] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2023.