ENHANCING THE APPLICABILITY OF THE EQUIVARIANT DIFFUSION MODEL TO DIVERSE MOLECULE DATASETS

Jumin Lee, Chaeyoug Moon, Minju Na
Korea Advanced Institute of Science and Technology (KAIST)
{jmlee, cy.moon, ijudy96}@kaist.ac.kr

1 INTRODUCTION

Deep learning-based molecular generation holds great promise for various applications in physics, chemistry, drug discovery, materials design, and more. While 1D and 2D molecular generation have been successful, molecules exist in 3D physical space and possess geometric symmetries such as translation, rotation, and reflection. The 3D shape of a molecule is particularly important for real-world applications like structure-based drug design. Recent advancements in geometric deep learning have sparked research on representing molecules in 3D geometry, leading to progress in molecular design within 3D space.

Recently, diffusion-based generation models have been applied to the molecular generation problem with impressive performance. These approaches employ a strategy of corrupting data with diffuse noise and training a neural diffusion model to reverse the corruption process and generate meaningful data from the noise. Denoised diffusion-diffusion probability models (DDPMs) have shown state-of-the-art results in various tasks, including image generation from text and structure-based protein design, in domains such as computer vision and computational biology. Within the DDPM framework, methods have been proposed, such as the equivariable graph neural network(GNN) Hoogeboom et al. (2022), for generating 3D molecules.

EDMs generate molecular structures that exhibit equivariance, meaning their properties remain unchanged regardless of the molecule's orientation or position in 3D space. To achieve this, the model learns to denoise the diffusion process, considering both the continuous coordinates(spatial location of atoms) and the categorical atom types(carbon, hydrogen, oxygen, etc.). While diffusion methods have been proposed for datasets with these characteristics, there is a lack of evaluation on more complex datasets and a comparative analysis of different diffusion methods specifically applied to categorical data. In addition, using the EDM model for large and complex molecules requires a lot of computational resources. The EDM employs a one-hot vector to represent different atom types. As the molecule size increases, the length of the one-hot vector also grows. This can be one of the reasons for the increased computational cost. To improve this situation, it would be helpful to consider introducing another discrete diffusion method.

Therefore, in this project, we extend the **Equivariant Diffusion for Molecule Generation and Conditional Generation in 3D(EDMs)**. Our main contributions are:

1. Apply an alternative diffusion method for categorical data in conjunction with the existing model.
2. Assess their generalizability using a larger dataset of three-dimensional small molecules.
3. Extend our models to other types of datasets, such as dimers, dipeptides, solvated amino acids, and further broadening the scope of our research.

In this paper, we aimed to check the performance of EDM by applying it to a dataset containing larger molecules and various atoms in addition to the existing qm9 dataset, which had a small number of atoms (e.g., 1 to 9) and molecules (e.g., 1.3 million). Details of the applied datasets can be found in Table 1. In addition, considering that the diffusion model takes a lot of time in the model training process because it has a diffusion step, we tried to apply bit-diffusion to reduce the time and cost. In three molecule datasets (e.g., QM9, SPICE, and Molecule3D), the validity was over 79.9% and

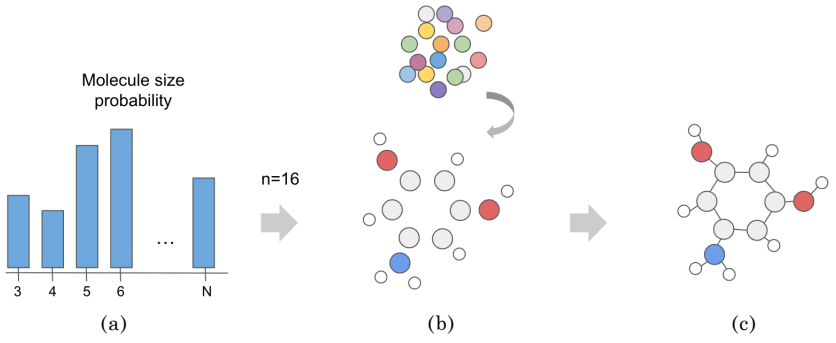

Figure 1: **Overall Equivariant Diffusion Model Framework** (a) Pre-generation stage involves sampling the desired number of atoms. (b) The Equivariant Diffusion Model then generates both the atom types and their corresponding coordinates. (c) Post-generation, bond types are predicted by analyzing the distances between atoms.

the uniqueness was over 98.5%. Our experiment code is available at `https://github.com/zoomin-lee/EDM_with_bit_diffusion`.

2 Related Work

2.1 Diffusion for Categorical Data

Diffusion models have been extensively studied in the context of continuous data. However, they often face challenges when handling discrete data, such as text and segmentation maps, due to the inherent discreteness that cannot be fully captured by continuous representations. To overcome this issue, researchers have developed discrete diffusion models Austin et al. (2021); Hoogeboom et al. (2021). These models update variable values at each time step using transition matrices that specify the probability of transitioning between different values.

More recently, there has been a shift towards studying continuous diffusion for discrete data, moving away from the use of transition matrices. This is because discrete diffusion models approximate the continuous situation using transition matrices, which can make accurate modeling difficult. Therefore, in the Enhanced Diffusion Models(EDM) Hoogeboom et al. (2022), discrete data is converted into a one-hot representation. Another approach, known as bit diffusion Chen et al. (2022), converts discrete data into a bit representation. These continuous representations of discrete data are then used to model the distribution using continuous diffusion methods, as they are treated as continuous values. In this project, we employ both one-hot and bit representations to model different diffusion processes for categorical data.

2.2 3D molecule generation

E-NF Satorras et al. (2021) and G-SchNet Gebauer et al. (2019). E-NF uses an equivariant EGNN-based regularization flow to transform arbitrary initial atomic positions into realistic molecular geometries. However, this model has only been validated for relatively small molecules, is very expensive to train, and produces molecules with low chemical validity rates. G-SchNet is an equivariate model that autoregressively generates 3D molecules, and has recently been extended to conditionally generate molecules with desired electronic properties and molecular fingerprints Gebauer et al. (2022). However, G-SchNet only generates atomic positions and does not generate molecular graphs, which provide chemical bonding information needed for many downstream applications.

3 Method

EDMs Hoogeboom et al. (2022) takes into account the equivariance principles concerning atom coordinates, encompassing rotation, translation, and reflection. Graph Neural Networks find exten-

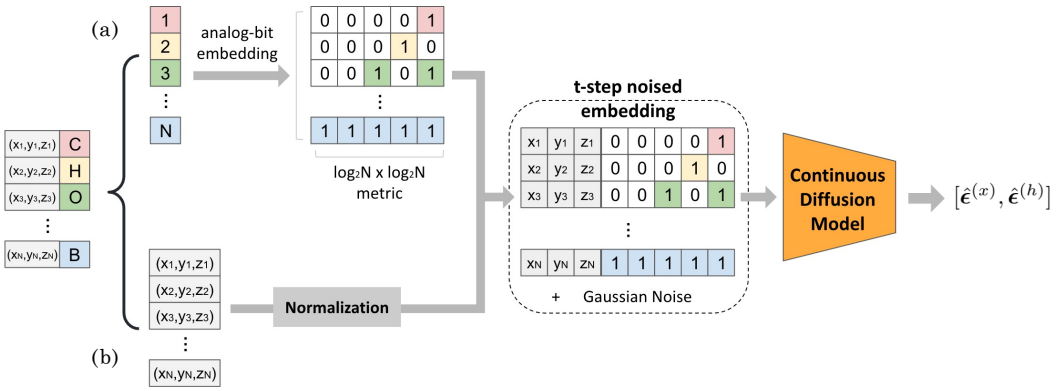

Figure 2: **Training Process of the Bit-EDM**. The following steps are undertaken for preprocessing: (a) The discrete atom types, represented by vector $\mathbf{h}$, are converted into analog bit embeddings. (b) The continuous coordinate values, denoted as $\mathbf{x}$, are normalized with respect to the origin. After completing the preprocessing stage, the equivariant noise process is applied to generate the t-step noised embedding. This perturbed embedding serves as the input for the continuous diffusion model.

sive employment in the representation of molecular structures, characterized by nodes and edges. Accordingly, GNN serves as the architectural backbone for the diffusion model within the context of EDM. EDM not only determines the coordinates but also incorporates the consideration of the molecular type within the diffusion process.

Consequently, the comprehensive generation framework can be outlined as in 1. Initially, the count of atoms to be encompassed within the molecule is established, paving the way for the diffusion model to undergo iterations over a predefined number of timesteps. These iterations facilitate the determination of both the molecular coordinates and its corresponding type. Subsequently, the inter-atomic distances are computed, effectively governing the establishment of edge connections within the molecular graph.

They define a set of points $(\mathbf{x}_i, \mathbf{h}_i)_{i=1,\ldots,M}$, where each node has associated to it a coordinate representation $\mathbf{x}_i \in \mathcal{R}^3$((b) in Fig. 2) and an atom types $\mathbf{h}_i$ ((a) in Fig. 2).

To generate t-step noised embedding in Fig. 2, we begin by normalizing the continuous coordinates and converting categorical atom types into either one-hot or bit representation. Subsequently, we apply the equivariant noise process $q(\mathbf{z}_t|\mathbf{x}, \mathbf{h})$ described in Eq. 1 to produce the t-step noised embedding.

$$q(\mathbf{z}_t|\mathbf{x}, \mathbf{h}) = q(\mathbf{z}_t^{(\mathbf{x})}|\mathbf{x}) \cdot q(\mathbf{z}_t^{(\mathbf{h})}|\mathbf{h}) = \mathcal{N}_x(\mathbf{z}_t^{(x)}|\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I}) \cdot \mathcal{N}_h(\mathbf{z}_t^{(\mathbf{h})}|\alpha_t \mathbf{h}, \sigma_t^2 \mathbf{I}) \tag{1}$$

where $\mathcal{N}_{xh}$ is a concise notation for the product of two distributions, one for the noised coordinates $\mathcal{N}_x$ and another for the noised atom types $\mathcal{N}$. $\alpha_t \in \mathcal{R}^+$ controls how much signal is retained and $\sigma \in \mathcal{R}^+$ controls how much noise is added.

In simpler terms, the process can be described as applying Gaussian noise to two features, as shown in the Eq. 2.

$$\mathbf{z}_t = \alpha_t[\mathbf{x}, \mathbf{h}] + \sigma_t[\epsilon^{(x)}, \epsilon^{(h)}] \tag{2}$$

where $[\cdot, \cdot]$ denote a concatenation.

So our network φ predicts $\hat{\epsilon} = [\hat{\epsilon}^{(x)}, \hat{\epsilon}^{(h)}]$. Using this prediction, we can obtain $[\hat{\mathbf{x}}, \hat{\mathbf{h}}] = \frac{\mathbf{z}_t}{\alpha_t} - \hat{\epsilon}_t \cdot \frac{\sigma_t}{\alpha_t}$. Therefore, our equivariant denoising process can be obtained as Eq. 3.

$$p(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}_{xh}(\mathbf{z}_{t-1}|\mu_{t\to t-1}([\hat{\mathbf{x}}, \hat{\mathbf{h}}], \mathbf{z}_t), \sigma_{t\to t-1}^2 \mathbf{I}) \tag{3}$$

In this project, we are working with a much larger dataset that includes a significantly higher number of molecules, atoms, and elements compared to the EDMs dataset, as shown in Table 1 1. Hence, we

Table 1: Dataset Information

Dataset	# Molecules	# Atoms	# Elements	Description
QM9	1.3 million	1∼9	5	Property descriptors & 3D atomic coordinates
Molecule3D	3.9 million	avg. 29.1	30	Ground-state 3D geometrics & Quantum properties
SPICE	1.1 million	2∼96	5∼15	small molecules, dimers, dipeptides, and solvated amino acids

investigate the equivariant noise process for categorical data $\mathcal{N}(\mathbf{z}_t^{(\mathbf{h})}|\alpha_t\mathbf{h}, \sigma_t^2\mathbf{I})$ in Equation 1 using both the one-hot representation $\mathbf{h}^{onehot}$ and the analog-bit representation $\mathbf{h}^{analog-bit}$ Chen et al. (2022).

The reason for considering both representations is that when using the one-hot representation, the embedding dimension after applying t-step noise increases based on the number of atom types in the dataset ($\mathbf{N}$). However, when using the analog-bit representation, the increase is only by $\log_2 \mathbf{N}$. Additionally, since there is no existing research analyzing the results of applying diffusion to both the one-hot representation and the analog-bit representation across multiple datasets, we aim to explore this aspect.

3.1 Dataset

In this project, we develop a model and evaluate its performance on three datasets. The details of the dataset are summarized in the Table 1.

The **QM9** dataset Ramakrishnan et al. (2014) contains molecular property descriptors and 3D atomic coordinates for 1.3 million small molecules. Each molecule in QM9 can contain up to nine heavy atoms, or 29 atoms if hydrogen is included.

Molecule3D Xu et al. (2021) is the first benchmark to systematically perform the task of ground-state 3D molecular shape prediction. The dataset contains information on more than 3.9 million molecules, including molecular graphs, ground-state 3D geometries, and various quantum properties. It is built upon the PubChemQC Nakata & Shimazaki (2017) dataset and curated to provide a more easily usable form for machine learning applications.

The SPICE Eastman et al. (2023) is a new quantum chemistry dataset for training potentials related to the simulation of drug-like small molecules interacting with proteins. It contains over 1.1 million conformations for a variety of small molecules, dimers, dipeptides, and solvated amino acids.

In this paper, we improve the algorithm by modifying the discrete diffusion method and conduct performance experiments on various datasets. We use the QM9 dataset used in the literature as a baseline, and expand the dataset to check the performance. Furthermore, we utilize more diverse datasets to test and validate that the model performs well for 3D coordinate generation.

4 Experiments

Metrics It uses the distance between pairs of atoms and the type of atom to predict the bond type (single, double, triple, or none). It then measures atomic stability (the percentage of atoms with the correct valence) and molecular stability (the percentage of molecules created in which all the atoms are stable).

Experiment Settings For the molecule generation, all models use 500 diffusion steps, EGNN with 3 layers, 128 features per layer, and SiLU activation, and are trained using Adam with a batch size of 256 and a learning rate of 10-4. We report the validity (measured by RDKit) and uniqueness of the generated compounds.

Table 2: Neg. log-likelihood, atom stability, molecule stability, validity and uniqueness over 1,000 molecules with standard deviation across 3 runs on QM9.

# Metrics	NLL	Atom stable (%)	Mol stable (%)	Valid (%)	Unique (%)
EDM	**12.4±0.48**	**98.05±0.00**	**86.53±0.01**	84.27±1.67	**100±0.00**
Bit-EDM (Ours)	589.9±16.4	1.3±0.00	0.0±0.0	**97.4±2.4**	1.0±0.7

Figure 3: Selection of samples generated by the denoising process of EDM (up) and Bit-EDM (down) on QM9.

4.1 QM9

In this experiment, we train Bit-EDM and EDM to unconditionally generate molecules with three-dimensional coordinates, 5 atom types (H, C, N, O, F), and integer values of atomic charges.

Results are reported in Table2. As shown in Table 2, EDM is able to generate a very high percentage of valid and unique molecules. While the original authors built their model with 9 layers and 3,000 training rounds, we built our model with 3 layers and 1,000 training rounds due to computational limitations, so we believe that our model is relatively under-trained, which may explain the positive NLL values for EDM and Bit-EDM in general. The validity of EDM is lower compared to Bit-EDM. This might be due to RDKit's correction process, where it adjusts valency by adding hydrogens to heavy atoms during calculation. We thought that using the bit-diffusion method instead of the traditional EDM method of one-hot encoding would reduce the time, which would be suitable for model training. However, in the actual model implementation stage, we had a problem that it took a little more time because we had to perform bit diffusion and then perform one-hot encoding again. We also expected that converting bits would reduce the amount of GPU memory required. However, we discovered that the size of the molecule has a much bigger impact compared to the representation of atom types. As a result, we were unable to achieve reduction in computational costs.

4.2 Conditional Molecule Generation - QM9

In this section, we aim to generate molecules that target several desired properties. Since the main goal of drug discovery is to generate molecules that are optimized for their properties, we train a conditional diffusion model in QM9 to generate according to the homo and lumo properties. Each property is described in more detail below. 1) HOMO: The highest occupied molecular orbital energy. 2) LUMO: The lowest unoccupied molecular orbital energy.

Experiment Settings In this conditional experiment, both EDM and Bit-EDM used EGNNs with 7 layers, 128 features per hidden layer, and SiLU activation. We used an Adam optimizer with a learning rate of 10-3 and a batch size of 512. We only modeled atom type (categorical) and position (continuous), and did not model atomic charge. All methods were trained for 1,000 epochs.

Results HOMO and LUMO values are measures of chemical reactivity and are important indicators in the field of molecule generation as well as drug discovery. We applied EDM and Bit-EDM to

Table 3: Mean Absolute Error for molecular property prediction on a QM9 subset, EDM and Bit-EDM generated samples.

Task	Homo	Lumo
EDM	**0.0544**	**0.0498**
Bit-EDM (Ours)	0.0571	0.0454

Table 4: Neg. log-likelihood, atom stability, molecule stability, validity and uniqueness over 500 molecules with standard deviation across 3 runs on SPICE.

# Metrics	NLL	Atom stable (%)	Mol stable (%)	Valid (%)	Unique (%)
EDM	**79.9±24.2**	**66.0±0.0**	0.0±0.0	**79.9±0.02**	**98.5±0.00**
Bit-EDM (Ours)	3855.5±83.8	28.75±0.02	**0.13±0.0**	65.33±0.29	5.50±0.05

derive MAE values for HOMO and LUMO values, and found that there was no difference between the results of Bit-EDM and the MAE values of EDM, as shown in Table 3.

4.3 SPICE

The SPICE Eastman et al. (2023) dataset contains 1.1 million molecules with sizes up to a maximum of 96. It contains dipeptides, solvated amino acids, dimers, monomers, ion pairs, and pubchem small moelcules. In this experiment, we train Bit-EDM and EDM to unconditionally generate molecules with three-dimensional coordinates, 15 atom types (H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I), and integer values of atomic charges.

Results The experimental results for the SPICE dataset are summarized in the Table4. Bit-EDM had lower atom stable and mol stable, which is too low to draw meaningful conclusions. We believe is due to the different size subsets of the SPICE dataset. For example, the SPICE dimers subset has 3,490 molecules with 2-34 atoms, and the solvated amino acids subset has 26 molecules with 79-96 atoms. As the number of atoms in each subset varies and the number of molecules varies, we believe that there were limitations in the molecule generation training process. Also, as with QM9, we found that the validity and uniqueness decreased in Bit-EDM compared to EDM.

Limitations In order to check validity and stability of the generated structures, we compute the distance between all pairs and use these distances to predict the existence of bonds and their orders. In addition, margins are defined in single, double, triple bonds m1, m2, m3 = 10, 5, 3 which were found empirically to describe the QM9 dataset well. However, for SPICE dataset molecules are much larger which introduces more atypical behavior.

4.4 MOLECULE3D

In this experiment, we train Bit-EDM and EDM to generate molecules using Molecule3D Xu et al. (2021) dataset.

Dataset Preparation The Molecule3D dataset contains 3.9 million molecules with the largest molecule composed of 51 heavy atoms. Since using large molecules with a high number of atoms requires a significant amount of GPU memory, we exclude the hydrogen atoms from the molecules. We randomly select one million molecules with sizes less than 30 from this dataset for experimentation. Then, the dataset contains 30 elements. The constructed dataset is divid into train, validation, and test sets with an 8:1:1 ratio, respectively.

Experiment Settings For the model architecture, we utilized EGNN with three layers and a hidden feature dimension of 128 for both EDM and Bit-EDM. The model was trained for a maximum of 500 epochs and updated for approximately 780,000 steps with a batch size of 512. The total number of diffusion steps was set to 500.

Results The experimental results for the Molecule3D dataset are summarized in the Table 5 above. The baseline model is the EDM model, and our model is called Bit-EDM. According to the table,

Table 5: Neg. log-likelihood, atom stability, molecule stability, validity and uniqueness over 1,000 molecules with standard deviation across 3 runs on Molecule3D.

# Metrics	NLL	Atom stable (%)	Mol stable (%)	Valid (%)	Unique (%)
EDM	45.9±1.8	**95.5±0.1**	**64.3±0.7**	**96.3±0.5**	**99.8±0.1**
Bit-EDM (Ours)	**-21.8±1.0**	81.6±0.1	7.7±0.6	28.1±0.5	60.7±1.7

Bit-EDM exhibits a lower negative log-likelihood value compared to EDM. However, EDM results in better performance in terms of other metrics such as validity, uniqueness, and stability.

5 Conclusion

This project aims to extend the **Equivariant Diffusion for Molecule Generation in 3D(EDMs)** Hoogeboom et al. (2022) by utilizing a larger and more diverse dataset that includes three-dimensional molecular structures such as small molecules, peptides, and amino acids. The objective is to assess the generalization capability of an alternative discrete diffusion method, which has not been employed in previous 3D molecule generation models.

We conduct experiments using three datasets: QM9 Ramakrishnan et al. (2014), SPICE Eastman et al. (2023), and Molecule3D Xu et al. (2021). EDM and our Bit-EDM generates the molecules in 3D structures and the generated molecules are evaluated with various metrics such as validity, uniqueness and stability. Based on the results, we demonstrate that the EDM model is well-suited for the Molecule3D dataset, while it is not as effective for the SPICE dataset. This could be because the SPICE dataset consists of more diverse molecules compared to the Molecule3D dataset. For example, in the case of the Molecule3D dataset, there is a size limitation, resulting in relatively small variations in the sizes of the molecules used.

Initially, our expectation was that the incorporation of Bit Diffusion would yield comparable results to the EDM model while reducing computational resources. However, our findings reveal that the utilization of Bit Diffusion actually leads to a performance decline. Our Bit-EDM model consistently generates molecules of lower quality compared to the existing model across all three experiments.

6 Future Works

In the project, both models generate only the atom type and 3D coordinates of the atoms. Consequently, in order to construct the complete molecular structure, it is essential to determine the bonds between the atoms. To do this, we calculate the distances between atoms and assign bonds based on these distances. In this approach, it is necessary to define distances based on both atom types and bond types. However, as the complexity of the molecule increases, the range of possible values becomes more diverse. This can result in potential errors in predicting the correct bonds. To address this issue, we propose utilizing the methods presented in this study Kim & Kim (2015). This particular method converts the 3D geometry into a molecular structure with atom connectivity. However, since our datasets contain some metal atoms, dataset filtering would be necessary. Nevertheless, incorporating this referenced method is anticipated to be beneficial in overcoming these challenges. Another limitation of this project is the bit-to-decimal conversion. During the conversion process, there is a possibility that the predicted decimal value exceeds the maximum allowable value. In such cases, we currently assign a random value from the range of possible integers to the decimal value. However, this approach can introduce errors and result in a degradation of performance. To address this issue in future research, it would be beneficial to explore alternative methods for handling this problem.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

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.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022.

Yeonjoon Kim and Woo Youn Kim. Universal structure conversion method for organic molecules: From atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society*, 36(7):1769–1777, 2015.

Maho Nakata and Tomomi Shimazaki. Pubchemqc project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57 (6):1300–1308, 2017.

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.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian B Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *arXiv preprint arXiv:2105.09016*, 2021.

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs. *arXiv preprint arXiv:2110.01717*, 2021.