Geometry-text Multi-modal Foundation Model for Reactivity-oriented Molecule Editing

Haorui Li[1,2*] **Shengchao Liu**[3*] **Hongyu Guo**[4] **Anima Anandkumar**[1]
[1]Caltech [2]Huazhong University of Science and Technology [3]Indepedent
[4]National Research Council Canada [*] Equal contribution

Abstract

Recent breakthroughs in foundation models have revolutionized the science domain with their promising generalization performance to solve challenging open questions. In chemistry and biology, the textual data enriches comprehensive knowledge about the molecule's functionalities, thus serving as a complementary modality to the chemical structures. However, existing multi-modal foundation models mainly focus on 2D topology rather than 3D geometry. To handle this issue, we construct a large-scale 3D structure-text dataset with conformations calculated by semi-empirical quantum methods. Then we propose MoleculeSTM-3D, a geometry-text multi-modal foundation model to align the two modalities through contrastive learning. For downstream, we apply MoleculeSTM-3D to the reactivity-oriented molecule editing task. Our empirical results demonstrate that it achieves a 9.00% higher editing success rate and significantly reduces invalid molecule generation by 10.07% compared to baseline methods. These preliminary results reveal the potential of utilizing MoleculeSTM-3D to solve more challenging tasks.

1 Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized various scientific fields, including chemistry [1, 2, 3, 4], materials science [5, 6, 7], and biology [8, 9, 10, 4, 11]. AI's ability to process vast amounts of data and recognize complex patterns has opened up unprecedented opportunities for scientific discovery, ranging from molecule generation [12] to protein folding prediction [13, 14]. A crucial aspect of these breakthroughs is the application of foundation models (FMs) [15]—large-scale predictive or generative models trained on extensive datasets—which have played a transformative role across various scientific domains. These models are characterized by their capability to learn general-purpose representations that can be adapted to a wide range of downstream tasks with minimal fine-tuning [16, 17].

In the field of AI for science, research into multi-modal foundation models [18, 19] is growing, reflecting their potential to integrate multiple data modalities, enrich representations, and enhance task performance across a broad spectrum of downstream applications. These models aim to leverage the strengths of different data sources, such as textual information and structural data, to create more comprehensive and robust representations [10]. In the context of chemistry, multi-modal foundation models seek to integrate textual descriptions with molecular structures, to enhance molecular representations[20, 21, 22]. Textual descriptions, often detailing the chemical and pharmacological attributes of molecules, complement the structural data, which represent the spatial configuration and bonding patterns within molecules. This approach leverages the complementary nature of these two modalities, as textual data provide high-level insights about molecular characteristics, while structural data capture the molecular geometry and atomic interactions essential for understanding molecular behavior in chemical reactions. By combining these distinct yet related modalities, multi-modal

AI for New Drug Modalities Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

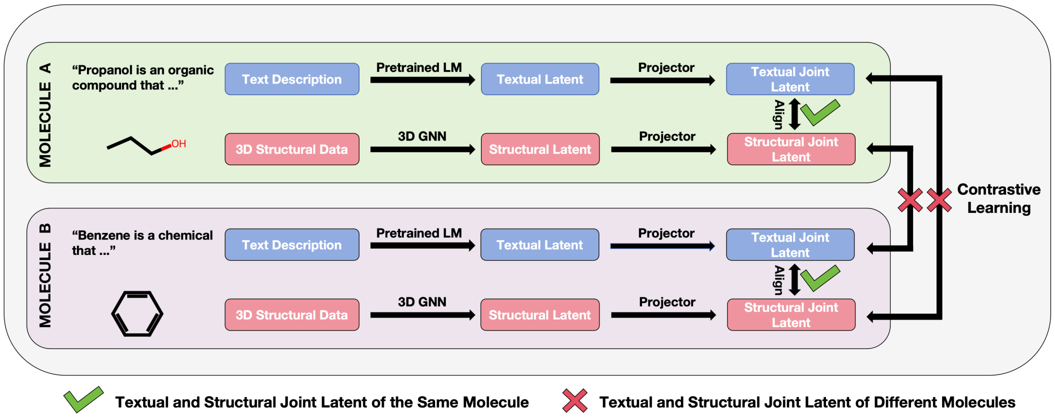

Figure 1: Pipeline of pretraining in MoleculeSTM-3D with two branches, the 3D structure (pink) and textual description (blue).

models are able to generate richer and more nuanced molecular representations, leading to improved performance in various downstream tasks.

One notable example of a multi-modal foundation model is MoleculeSTM [3], which maps structural and textual representations into a joint space using contrastive learning, reducing the representation distance between a molecule's chemical structure and its textual description while increasing the distance between different molecules. To demonstrate the advantages of introducing the language modality, the researchers designed two challenging downstream tasks: the structure-text retrieval task and the text-based molecule editing task, and they applied the pretrained MoleculeSTM in a zero-shot manner, achieving promising results. However, the model has limitations: for molecular structural data, it only focuses on 1D SMILES and 2D graphs, without considering the more information-rich 3D molecular graphs. Additionally, for molecule editing tasks that require a generative model, it uses MegaMolBart [23], a BART-based pretrained sequence-to-sequence model that can only process and generate SMILES as sequence data, significantly limiting its ability to generate more effective and robust molecules.

Recent studies have begun to explore the integration of 3D structural data with textual data to improve molecular representations [24, 25]. While this approach shows promise for creating richer and more comprehensive representations, these efforts are still in the early stages and face several significant challenges. One of the most critical obstacles is **the lack of benchmark:**

• **Dataset**. The shortage of large-scale, high-quality multi-modal datasets that combine detailed textual descriptions with accurate 3D structural data for molecules makes it difficult to train sophisticated multi-modal models capable of fully capturing the complexities of molecular structures alongside their textual attributes.

• **Evaluation**. The lack of innovative downstream tasks designed to specifically evaluate the effectiveness of 3D structural data further hampers progress. Most existing evaluation tasks are not designed to fully leverage the rich geometric information encoded in 3D molecular structures, making it difficult to accurately assess the impact of 3D data on model performance.

Our contributions. In response to the aforementioned limitations, we propose a **Geometry-text Multi-modal Foundation Model**, called **MoleculeSTM-3D**. In this work, we focus on aligning the latent representations of 3D structural data and textual descriptions of the same molecules to generate richer, more robust, and more effective representations for downstream tasks. Our main contributions are as follows:

• **3D Structure-Text Dataset Construction:** We construct a novel dataset containing approximately 163K molecules with 202K text-structure pairs. The textual data, sourced from PubChem, provides comprehensive descriptions of each molecule, including its physical, chemical, biological, and pharmacological properties. The structural data, extracted from the PubChemQC PM6 dataset, includes the constituent atoms and their corresponding 3D coordinates.

• **Alignment of 3D and Textual Modalities:** We process the 3D structural information and textual descriptions of the same molecules using one of state-of-the-art 3D GNNs and the cutting-edge pretrained language models, respectively, to obtain two distinct latent representations. We apply contrastive learning to align latent representations between 3D molecular structures and textual descriptions. This alignment enables our model to integrate both modalities effectively, enhancing molecular representations for various downstream applications.

• **Reactivity-oriented Molecule Editing Task:** To demonstrate the effectiveness of our joint latent representation, we propose a novel and challenging downstream task: reactivity-oriented molecule editing. This task is designed to modify molecules to increase their reactivity toward the specific reaction type which is described by a textual description. Additionally, we introduce benchmark metrics to rigorously evaluate the editing performance.

Using the learned alignment between 3D structures and textual representations, we apply our model to perform reactivity-oriented molecule editing. Empirically, our model demonstrates a significant improvement in performance, with an average increase of 9.00% in editing success rate and a 10.07% decrease in invalid molecule generation rate, showcasing the robustness and effectiveness of our multi-modal foundation model.

2 3D Structure-Text Dataset Construction

To combine textual and 3D structural latent representations of the same molecules and obtain more robust joint latent representations, we first need to collect both structural and textual data for the same molecules, forming 3D structure-text pairs.

2.1 Textual Data Component

For the textual data part of our dataset, we choose to extract it from PubChem Dataset [26], which is an online database managed by the National Center for Biotechnology Information (NCBI) of the United States and provides detailed chemical substances and biological activity data. This resource includes the structure, properties, biological activity and related information of millions of chemical substances, making it an ideal source for extracting textual information. Through the official API or website provided by PubChem, researchers can easily access detailed records of specific compounds, including chemical names, synonyms, molecular weights, and more. Among these records are comprehensive textual descriptions of molecules, which are especially valuable for constructing structure-text pairs.

We adopt the method of processing and extracting text from MoleculeSTM. First, we use one of the PubChem APIs, PUG View [27], to download the textual descriptions of molecules. Specifically, PUG View is a REST-style web service that provides information content that is not directly contained within the primary PubChem Substance, Compound, or BioAssay records. Its purpose is primarily to drive the PubChem database summary record web pages, but can also be used independently as a programmatic web service. PUG View is mainly designed to provide complete summary reports on individual PubChem records. It has many types of records, and here we focus on the annotations records, which contain textual descriptions of molecules. These textual descriptions are constantly updated and increased. As of the time we downloaded, there were a total of 422 pages of data. Each page is downloaded in JSON file format. In each JSON file, there is a key called "strings," and its corresponding value is a comprehensive description of a molecule, including its physical properties, chemical properties, and more. We use it as the textual description we need. In order to obtain standardized and uniform text data that is easy for the pretrained language model to load, we preprocess these text descriptions into a consistent format, and store them as JSON files indexed by PubChem ID (CID). It should be noted that each molecule may have more than one annotation. Therefore, there are a total of 333K molecules with relevant textual descriptions, and a total of 396K PubChem ID-text pairs. With this, we have preliminarily completed the establishment of the textual data component.

2.2 Structural Data Component

For the 3D structural data part of our dataset, it's important to note that commonly used molecular datasets do not meet our needs. We require 3D data for a broad range of molecules, and the volume required is substantial. However, well-known datasets that provide molecular 3D structures often have limitations in both the variety of molecules and the number of available entries. For instance, the

QM9 dataset [28] only includes 3D data for molecules with fewer than nine heavy atoms, covering just 134K molecules. Many of the molecules in our textual data fall outside this range, significantly limiting the dataset size and negatively impacting pretraining performance, which ultimately reduces the robustness of the latent representations. Therefore, we use the relatively new molecular dataset, PubChemQC PM6 [29], which optimizes molecular geometries and electronic properties calculated by the PM6 method for 94.0% of the 91.6 million molecules cataloged in PubChem Compounds retrieved on August 29, 2016.

There are a total of about 86M molecules with 3D structural data, and each molecule may have multiple sets of structural data corresponding to different states; besides the consistently present ground state (S_0), molecules may also exist in cationic, anionic, and the lowest triplet states (T_0). Here we only focus on the 3D structural data of the ground state S_0, because this is the state in which most molecules exist in nature, with the lowest energy and the most stable. All these molecular data are divided into 4860 compressed files in tar.xz format according to their PubChem ID (CID) and stored on a Sharepoint webpage. Since the PubChem IDs corresponding to the molecules in our text data are roughly evenly distributed across the entire range, we need to download and process nearly all of these compressed files. Since these compressed packages are all on the Sharepoint website, we choose to use a web crawler to download these compressed packages in a semi-automatic and semi-manual way. We decompress them one by one to retrieve the molecular data whose corresponding CIDs that also appear in our textual data part, and then extract the JSON files containing their structural data of the ground state S_0. It's important to note that these compressed files are extremely large, with the whole decompressed data amounting to several terabytes. Downloading them via crawling and decompressing all the compressed packages is both time-consuming and cumbersome. Since not all of the 333K molecules with textual descriptions have corresponding 3D structural data, we eventually identify and extract the 3D structural data of the ground state S_0 for 163,467 corresponding molecules, each stored in a separate JSON file. After removing the molecules from the text data component that do not have corresponding 3D structures, we finalize a 3D structure-text dataset consisting of 163,467 molecules, forming a total of 202,272 3D structure-text pairs, each with corresponding textual descriptions and 3D structural data of the ground state S_0.

3 3D Structure-Text Multi-modal Pretraining Framework

3.1 Text Modality Architecture

For the textual descriptions in our dataset, we require a language model that can understand and process the textual descriptions of molecules, encoding them into latent representations that capture each molecule's characteristics and properties for further use. Therefore, we chose two language models for experimentation: SciBERT and Llama2-7B.

SciBERT [30], based on the BERT architecture, is a powerful transformer-based model that captures contextual relationships in text by processing words bidirectionally, making it ideal for grasping nuanced and intricate textual meanings. Specifically, SciBERT is pretrained on a large corpus of scientific texts, including a random sample of 1.14 million papers from Semantic Scholar. This corpus consists of 18% computer science papers and 82% biomedical papers, using the full text rather than just abstracts. The average paper length is 154 sentences (2,769 tokens), resulting in a total corpus size of 3.17 billion tokens. Since 82% of the corpus is from the biomedical domain, we believe SciBERT is well-suited to understanding comprehensive molecular descriptions and encoding them into textual latent representations.

In addition, we experiment with Llama2-7B [31], a state-of-the-art large language model from Meta's Llama series. Llama2-7B is a generative transformer-based model with 7 billion parameters, designed for a wide range of natural language processing tasks. Unlike SciBERT, which is specifically tailored for scientific texts, Llama2-7B is a general-purpose model but exhibits remarkable performance across diverse text datasets due to its scale and the quality of its training data. By leveraging its vast capacity, Llama2-7B can also capture complex semantic relationships and offer robust latent representations for molecular descriptions.

3.2 3D Structure Modality Architecture

For the molecular 3D structural data, our goal is to obtain a robust latent representation that captures the rich 3D information of each molecule. Symmetry-informed geometric representations, which leverage physical principles (i.e., group theory for depicting symmetric particles) into spatial representations, have emerged as a promising approach. To process our 3D structural data, we need a 3D GNN model. Additionally, to effectively capture and handle the rotational and translational equivariance of molecules in 3D geometric space while maintaining physical consistency, we need an SE(3)-equivariant model capable of generating symmetry-informed geometric representations. To meet these requirements, we select a pretrained PaiNN [32, 33], one of the state-of-the-art 3D equivariant graph neural networks. PaiNN leverages the message-passing mechanism, which facilitates information propagation along the graph structure by updating node embeddings through neighborhood aggregation. This equivariant GNN simultaneously updates both invariant and equivariant features, making it suitable for practical tasks, such as molecular dynamics simulations, that require equivariant outputs.

3.3 Contrastive Learning Paradigm

Once we obtain both the structural and textual latent representations of the same molecule, our goal is to combine these two representations to create a more comprehensive and robust molecular latent representation. The textual latent representation can be viewed as domain knowledge that strengthens the structure latent representation. To achieve this, we employ contrastive learning, which works by reducing the representation distance between a molecule's chemical structure and textual description, while increasing the distance between different molecules. Through this approach, we map the 3D structural and textual latent representation spaces into a joint latent space, ensuring that the latent representation of the molecule contains both structural and textual information (Figure 1).

It is important to note that our 3D GNN model and language model operate independently, each generating its own latent space from the dataset. This independence allows us to utilize pretrained checkpoints for both models, enabling us to focus solely on pretraining through contrastive learning to efficiently map these two spaces into a joint space, thereby enhancing training efficiency and boosting the joint latent space's ability to capture complex relationships.

Specifically, we use two projectors to convert the 3D structure and textual latent representations into a space with the same dimensions. In this space, for the same molecule, we have both its structural and textual latent representations, referred to as a positive 3D structure-text pair. We then create a negative 3D structure-text pair by combining the structural and textual latent representations from different molecules. Then, we pretrain using EBM-NCE or InfoNCE, which align positive 3D structure-text pairs while contrasting them against negative pairs. The objectives for EBM-NCE [34] and InfoNCE [35] are:

$$\mathcal{L}_{\text{EBM-NCE}} = -\tfrac{1}{2}\Big(2 \cdot \mathbb{E}_{x_c,x_t}\big[\log \sigma(E(x_c,x_t))\big] + \mathbb{E}_{x_c,x_{t'}}\big[\log(1-\sigma(E(x_c,x_{c'}t)))\big] + \mathbb{E}_{x_{c'},x_t}\big[\log(1-\sigma(E(x_{c'},x_t)))\big]\Big), \quad (1)$$

$$\mathcal{L}_{\text{InfoNCE}} = -\tfrac{1}{2} \cdot \mathbb{E}_{x_c,x_t}\left[\log \frac{\exp(E(x_c,x_t))}{\exp(E(x_c,x_t)) + \sum\limits_{x_{t'}} \exp(E(x_c,x_{t'}))} + \log \frac{\exp(E(x_c,x_t))}{\exp(E(x_c,x_t)) + \sum\limits_{x_{c'}} \exp(E(x_{c'},x_t))}\right], \quad (2)$$

where σ is the sigmoid activation function, x_c and x_t form the structure-text pair for each molecule, and $x_{c'}$ and x_t are the negative samples randomly sampled from the noise distribution, which we use the empirical data distribution. $E(\cdot)$ is the energy function with a flexible formulation, and we use the dot product on the jointly learned space, i.e., $E(x_c, x_t) = \langle p_c \circ f_c(x_c), p_t \circ f_t(x_t) \rangle$, where $\circ$ is the function composition. Here, f_c represents the PaiNN model used to encode 3D structural data, and f_t denotes a pretrained language model used to encode textual descriptions. p_c and p_t are projectors designed to convert the 3D structure and textual latent representations into a space with the same dimensions.

4 Reactivity-oriented Molecule Editing Task

To evaluate the effectiveness and robustness of the joint latent representations which combine 3D structural and textual information of the same molecules, we design a reactivity-oriented molecule editing task. The goal of this task is to modify the chemical structure of molecules, such as by changing functional groups [36] or scaffold hopping [37], to increase their reactivity toward a specific type of reaction. In organic chemistry, although thousands of molecules can participate in various reactions, most reactions fall into a few primary types of reaction. The main types of reactions include nucleophilic substitution (S_N1, S_N2), electrophilic addition, elimination ($E1$, $E2$), polymerization, and condensation reactions. These reaction types have distinct characteristics and are strongly correlated with specific structural features of the reactants. Therefore, our aim is to determine whether, for a given molecule, specifying the desired type of reaction through a text description can result in an edited molecule with increased reactivity toward the expected type of reaction. This downstream task not only demonstrates whether the pretrained language model can correctly interpret the implicit characteristics of specific reaction types from the textual description, but also shows whether the joint latent representation, which integrates both 3D structural and textual data of the same molecule, can accurately align molecular properties with structural features to output valid and reasonable edited molecules. In essence, this downstream task serves as a strong validation of the superiority of the multi-modal latent representation that incorporates textual information.

To achieve this task, we need a generative model to output the edited molecules. The generative model has its own latent space, while we have the 3D structure-text joint latent space through pretraining. Based on the joint latent representation corresponding to the textual description of the desired reaction type, the ML editing methods can learn a semantically meaningful direction in the latent representation space of the generative model. The generative model then outputs edited molecules with the desired properties by moving along this direction. It's important to note that we need a 3D graph generative model capable of handling 3D molecules. We choose LDM-3DG [38], which is one of the state-of-the-art (SOTA) models. LDM-3DG proposes performing 3D graph diffusion in a latent space rather than the original space, ensuring that the latent space is low-dimensional but high-quality. This low-dimensional latent space is learned in a data-driven manner by pretraining a 3D graph autoencoder (AE). To differentiate it from the latent space obtained after diffusion, we refer to this as the AE-encoded latent space. LDM-3DG innovates the AE architecture by strategically decomposing topological and geometric features, ensuring that the symmetry constraints of permutation and SE(3) transformations are disentangled and properly addressed in separate but cascaded AE models. The diffusion generative model (DGM) is then trained in the resulting latent space to model distributions. After adding noise through the diffusion process, the latent space is further transformed, which we refer to as the diffusion latent space. In the context of LDM-3DG, this diffusion latent space serves as the latent space of the generative model, enabling the generation of 3D molecular structures with desired properties.

Specifically, the pipeline for this downstream task is divided into two parts (Figure 2). The first part is space alignment, where we need to align the latent representation of the same molecule in the generative model with its latent representation in MoleculeSTM-3D, thereby aligning the two different latent spaces. To achieve this, we use two adaptor modules: one maps the representation space from MoleculeSTM-3D to the generative model's space, and the other maps the generative model's space to MoleculeSTM-3D. The alignment is then optimized using the following objective functions:

$$\mathcal{L} = \left\| m_{g2f} \circ f_g(x_c) - p_c \circ f_c(x_c) \right\|^2 , \tag{3}$$

where $\circ$ is the function composition function, and m_{g2f} is the adaptor module optimized to align the two latent spaces. Here, f_g corresponds to the process in LDM-3DG that generates the molecule's diffusion latent representation via the autoencoder (AE) followed by the diffusion process.

The second part is latent optimization, where the adaptor module trained in the first part comes into play. For a given input molecule and a text description specifying the desired reaction type, the goal is to directly optimize the latent representation w of the generative model. This w represents the edited molecule's latent representation. Using the previously obtained adaptor module, we transform w into the joint latent representation in MoleculeSTM-3D. This way, the latent representations of the edited molecule and reaction type description are in the same latent space, and we optimize w

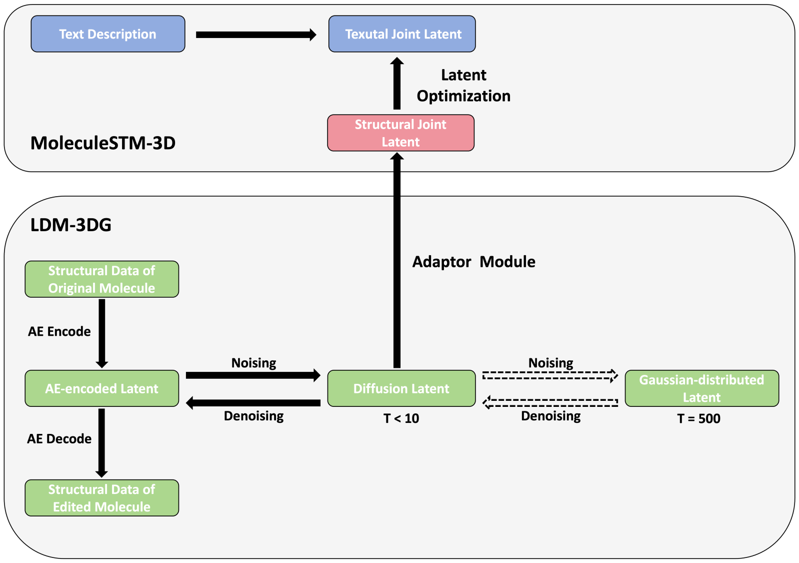

Figure 2: Pipeline of reactivity-oriented molecule editing task.

of the generative model to bring its corresponding representation in MoleculeSTM-3D as close as possible to the textual latent representation. At the same time, we ensure that w does not stray too far from the input molecule's latent representation of the generative model. The purpose of this is to modify the molecular structure to achieve the desired reactivity while preserving the molecule's core structure and fundamental characteristics, which is the essential requirement of molecule editing. The optimization of w can be expressed by the following formula:

$$w = \arg\min_{w \in \mathcal{W}} \big(-\mathcal{L}_{\text{cosine-sim}} \left(m_{g2f}(w), p_t \circ f_t(x_t) \right) + \lambda \cdot \mathcal{L}_{l_2} \left(w, f_g(x_{c,\text{in}}) \right) \big), \tag{4}$$

where $\mathcal{W}$ is the latent code space, $\mathcal{L}_{\text{cosine-sim}}$ is the cosine-similarity, and $\mathcal{L}_{l_2}$ is the l_2 distance, and λ is a coefficient to balance these two similarity terms.

Finally, after obtaining the optimal w, LDM-3DG uses this latent representation to generate the edited molecule through reverse diffusion and a 3D graph autoencoder (AE). It's important to note that during this editing process, both the MoleculeSTM-3D (f_c, p_c, f_t, p_t) and the LDM-3DG remain frozen.

The choice of the diffusion latent space in LDM-3DG is crucial. As a latent diffusion model, LDM-3DG is originally designed for de novo molecular generation in its paper. Consequently, the original implementation uses a Gaussian-distributed latent space, which is obtained by fully noising the AE-encoded latent space, as the diffusion latent space. However, in our scenario, the goal is not to generate entirely new molecules but to make slight modifications to existing molecules while preserving their core structural features to meet specific requirements.

To achieve this, we apply only mild diffusion to the AE-encoded latent space, ensuring that the diffusion latent representation retains the essential characteristics of the original molecule. Subsequently, inference (reverse denoising) on this representation effectively modifies the molecule while preserving its core structure, aligning seamlessly with our objectives for molecule editing. In contrast, if the AE-encoded latent space of the input molecule undergoes full Gaussian noising, as per the standard procedure in LDM-3DG, the resulting diffusion latent representation essentially becomes a random point in the Gaussian space, leading to the generation of a completely new molecule.

Experiments confirm this distinction: when using the parameters from the original LDM-3DG paper to noise the AE-encoded latent representation and then performing reverse denoising and AE decoding, the resulting molecules differ significantly from the original input molecules in structure. This outcome contradicts the fundamental intent of molecule editing. Conversely, by significantly reducing the number of timesteps and noise intensity during diffusion (e.g., timestep=3, noise=1e-4),

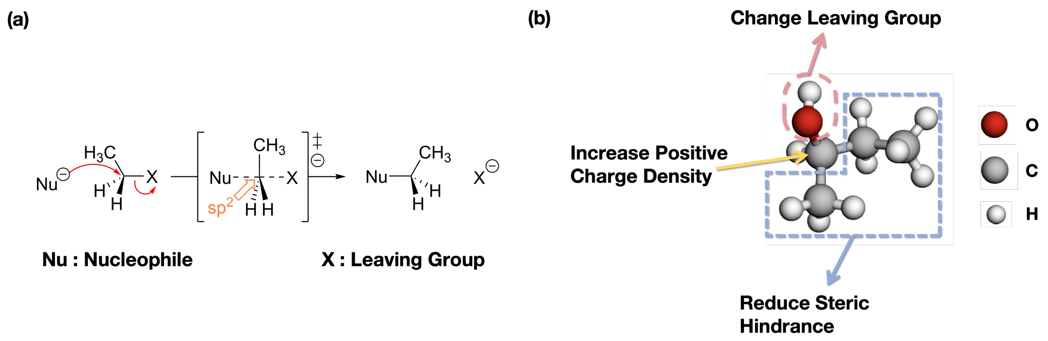

Figure 3: (a) Schematic diagram of the S_N2 reaction mechanism. (b) The three most important factors affecting the reactivity of the substrate in an S_N2 reaction.

the edited molecules maintain the essential structural features of the original molecules, thereby meeting the requirements of molecule editing effectively.

5 Experiment

In organic chemistry, the main types of reactions include bimolecular nucleophilic substitution S_N2, unimolecular nucleophilic substitution S_N1, electrophilic addition, bimolecular elimination $E2$, and unimolecular elimination $E1$. For our experiment, we choose the bimolecular nucleophilic substitution reaction S_N2 [39] as the target reaction type, as S_N2 reactions are among the most common and widely applied in fields such as drug synthesis and molecule discovery.

S_N2 reactions involve a single-step mechanism where the nucleophile attacks the carbon atom from the opposite side of the leaving group, resulting in the simultaneous bond formation and bond breaking (Figure 3 (a)). This reaction is characterized by its concerted mechanism, where the rate of the reaction depends on the concentration of both the nucleophile and the substrate (hence, "bimolecular").

In our task, we use textual descriptions of the key characteristics of S_N2 reactions to guide the editing of input molecules. The aim is to modify several critical factors affecting S_N2 reactivity, such as the leaving ability of the leaving group, the charge density on the carbon atom undergoing nucleophilic attack, and the steric hindrance surrounding that carbon atom (Figure 3 (b)).

Dataset. Alcohol molecules are a typical class of compounds that rarely undergo S_N2 reactions due to the presence of the hydroxyl group (-OH), which is a poor leaving group. Therefore, we select alcohol molecules to build a dataset for testing whether we can enhance their S_N2 reactivity. It is important to note that, to eliminate interference from other functional groups that may already be reactive in S_N2 reactions, we restrict the selected molecules to those containing only H, N, O and C atoms. We exclude molecules with halogens or ether bonds, as these functional groups are relatively better leaving groups. Additionally, we ensure that each selected molecule contains only one hydroxyl group. Following these criteria, we randomly sample 100 molecules that meet the requirements from the 3D structural data component of the 3D Structure-Text Dataset we build before.

Evaluation metrics. To evaluate whether the edited molecules are more likely to undergo an S_N2 reaction, we propose novel evaluation metrics that assess the success of the editing based on three key factors. A successful edit is defined as meeting the requirements of any one of these factors:

• **Leaving Group:** If the hydroxyl group in the original molecule is replaced by a better leaving group, such as an ether linkage or halogen, while the basic structure of the molecule remains intact, we consider the editing successful. A better leaving group increases the likelihood of the molecule undergoing an S_N2 reaction.

• **Charge Density on the Carbon Atom:** If the edited molecule remains an alcohol, we calculate the charge density of the carbon atom attached to the hydroxyl group. If the charge density on this reaction center carbon increases beyond a certain threshold compared to the original molecule, we also consider the editing successful. A higher, more positive charge on the carbon atom indicates

Table 1: Results of S_N2 reactivity-oriented molecule editing task on alcohol molecules.

	MoleculeSTM		MoleculeSTM-3D	
	SMILES	Graph	SciBERT	Llama2-7B
Editing Success Rate (%)	17.00	18.00	25.00	28.00
Invalid Molecule Generation Rate (%)	12.88	13.00	2.00	3.75

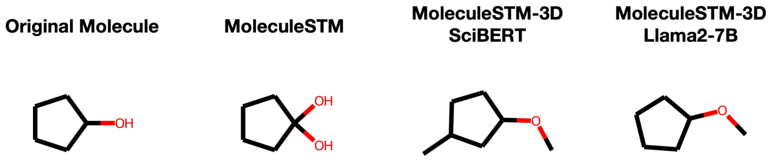

Figure 4: Different edited molecules generated by using different models.

that it is more electrophilic and more likely to be attacked by a nucleophile, thus facilitating the S_N2 reaction.

• **Steric Hindrance Around the Carbon Atom:** If the molecule remains an alcohol after editing, we also evaluate the number of atoms surrounding the carbon atom attached to the hydroxyl group. If the number of atoms within a certain range around this reaction center carbon decreases by more than a defined threshold, we deem the editing successful. Reduced steric hindrance around the carbon atom makes it easier for nucleophiles to approach and attack, increasing the likelihood of an S_N2 reaction.

Additionally, we apply some basic chemistry-based exclusions. If the edited molecule does not meet certain fundamental conditions required for an S_N2 reaction substrate, we consider this editing a failure. For example, if an alcohol is originally a primary alcohol, but the edited molecule becomes a tertiary alcohol, this is considered a failed editing. From a chemical standpoint, tertiary alcohols cannot undergo S_N2 reactions due to their high steric hindrance, making such a transformation incompatible with the S_N2 reaction mechanism.

Main Results. We test two versions of MoleculeSTM-3D: one using SciBERT as the language model, and the other using Llama2-7B. We compare their performance against MoleculeSTM, which is one of the state-of-the-art (SOTA) multi-modal foundation models, particularly well-suited for molecule editing tasks. We use alcohol molecules as input molecules and test by providing the characteristics of S_N2 reaction substrates as the textual description. The results are as follows (Table 1):

We observe that, regardless of whether SciBERT or Llama2-7B is used as the language model, MoleculeSTM-3D consistently outperforms MoleculeSTM, with an average increase of 9.00% in editing success rate and an average decrease of 10.07% in invalid molecule generation rate. We attribute this improvement to our use of 3D molecular structure data instead of 2D graph data or SMILES. The 3D data provides a richer structural context of the molecules, and our use of SE(3)-equivariant 3D GNNs allows for better processing of this 3D structural information, resulting in a more comprehensive molecular representation. This aligns well with the core premise of our research.

Furthermore, when comparing the SciBERT version to the Llama2-7B version within MoleculeSTM-3D, we find that Llama2-7B slightly outperforms SciBERT. This may be due to Llama2-7B's superior capability in understanding the relationship between the molecular structures and the likelihood of S_N2 reactions from the textual descriptions of S_N2 reaction characteristics.

Case One. To better illustrate the editing effects of different models, let's take cyclopentanol as an example (Figure 4):

• **MoleculeSTM** edits cyclopentanol into 1,1-cyclopentanediol. However, since both hydroxyl groups are attached to the same carbon atom, this molecule is highly unstable and prone to losing a molecule of water, forming cyclopentanone. This is considered a failed edit because it introduces instability.

• **MoleculeSTM-3D (SciBERT)** produces 1-methoxy-3-methylcyclohexane. In this case, the hydroxyl group is successfully converted into a methoxy group, which is a better leaving group for an S_N2 reaction. However, an additional methyl group is added elsewhere in the ring, which does not contribute to increasing S_N2 reactivity and is an unnecessary modification.

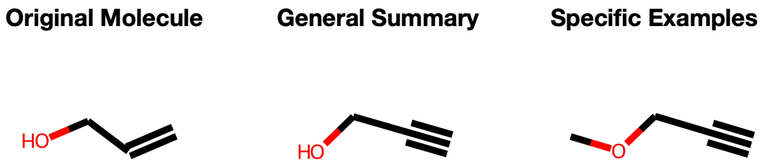

Figure 5: Different edited molecules generated by using different textual descriptions.

• **MoleculeSTM-3D (Llama2-7B)** results in methoxycyclopentane. This edit simply converts the hydroxyl group into a methoxy group, forming an ether bond, without making any other unnecessary modifications. This directly enhances S_N2 reactivity, making it a successful edit.

Case Two. We also discover that the textual content describing S_N2 reactions significantly influences the editing outcomes. We employ two different approaches to describe S_N2 reactions (Figure 5):

• **General Summary:** This description provides an overview of the characteristics that a good S_N2 substrate should possess. For example, using this type of description, the editing of allyl alcohol results in propargyl alcohol, where the formation of a carbon-carbon triple bond increases the positive charge density on the reaction carbon atom. This edit is considered effective.

• **Specific Examples:** This approach details specific molecules that are prone to S_N2 reactions, such as halogenated hydrocarbons, highlighting their characteristics and the reasons they facilitate S_N2 reactions. With this description, the molecule edited from allyl alcohol is propargyl methyl ether, which not only features a carbon-carbon triple bond to increase the positive charge density but also forms a better leaving group, methoxy, making it a more successful edit.

Our experiments, conducted on the alcohol dataset mentioned earlier, indicate that using the General Summary as the textual description yields 53 successful edits, whereas the Specific Examples approach results in 66 successful edits, an improvement of 24.53%. These results suggest that detailed, specific descriptions are more effective in achieving successful molecular edits.

6 Conclusion

In this work, to address the lack of large-scale 3D structure-text datasets and the absence of novel downstream tasks to effectively evaluate the performance of multi-modal foundation models using 3D data in the field of chemistry, we take two key steps. First, we construct a comprehensive 3D Structure-text dataset, and second, we design an innovative and challenging downstream task, the reactivity-oriented molecule editing task, along with corresponding evaluation metrics. We then utilize PaiNN and SciBERT/Llama2-7B, applying a contrastive learning paradigm, to build MoleculeSTM-3D, a multi-modal foundation model trained on our 3D structure-text dataset. Finally, we apply MoleculeSTM-3D to the reactivity-oriented molecule editing task. Our results show that it outperforms MoleculeSTM, achieving a 9.00% higher editing success rate and significantly reducing invalid molecule generation by 10.07%. Although our preliminary tests have primarily focused on S_N2 reactions, the promising results lead us to believe that MoleculeSTM-3D will not only perform well in reactivity-oriented molecule editing for more complex molecules and more difficult reaction types, but will also demonstrate effectiveness in other more challenging downstream tasks.

References

[1] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

[2] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[3] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

[4] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Zhiling Zheng, Ali H Alawadhi, Saumil Chheda, S Ephraim Neumann, Nakul Rampal, Shengchao Liu, Ha L Nguyen, Yen-hsu Lin, Zichao Rong, J Ilja Siepmann, et al. Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned gpt models. *Journal of the American Chemical Society*, 145(51):28284–28295, 2023.

[6] Junkil Park, Youhan Lee, and Jihan Kim. Multi-modal conditioning for metal-organic frameworks generation using 3d modeling techniques. *ChemRxiv*, 2024.

[7] Shengchao Liu, Divin Yan, Weitao Du, Zhuoxinran Li, Zhiling Zheng, Omar M. Yaghi, Christian Borgs, Hongyu Guo, Anima Anandkumar, and Jennifer T Chayes. A geometric foundation model for crystalline material discovery. 2024.

[8] Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

[9] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[10] Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023.

[11] Yanjing Li, Hannan Xu, Haiteng Zhao, Hongyu Guo, and Shengchao Liu. Chatpathway: Conversational large language models for biology pathway detection. In *NeurIPS GLFrontiers Workshop 2023 Oral*, 2023.

[12] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pages 4839–4848. PMLR, 2020.

[13] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[14] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[16] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[18] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

[19] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

[20] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.

[21] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021.

[22] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

[23] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

[24] Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Yanchen Luo, Sihang Li, Zhiyuan Liu, Jiancan Wu, Zhengyi Yang, Xiangnan He, Xiang Wang, and Qi Tian. Text-guided diffusion model for 3d molecule generation. *OpenReview*, 2024.

[26] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021.

[27] Sunghwan Kim, Paul A Thiessen, Tiejun Cheng, Jian Zhang, Asta Gindulyte, and Evan E Bolton. Pug-view: programmatic access to chemical annotations integrated in pubchem. *Journal of cheminformatics*, 11(1):56, 2019.

[28] 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.

[29] Maho Nakata, Tomomi Shimazaki, Masatomo Hashimoto, and Toshiyuki Maeda. Pubchemqc pm6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *Journal of Chemical Information and Modeling*, 60(12):5891–5899, 2020.

[30] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[32] 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.

[33] Shengchao Liu, Weitao Du, Zhiming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *ICML*, 2023.

[34] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[36] Peter Ertl, Eva Altmann, and Jeffrey M McKenna. The most common functional groups in bioactive molecules and how their popularity has evolved over time. *Journal of medicinal chemistry*, 63(15):8408–8418, 2020.

[37] Hans-Joachim Böhm, Alexander Flohr, and Martin Stahl. Scaffold hopping. *Drug discovery today: Technologies*, 1(3):217–224, 2004.

[38] Yuning You, Ruida Zhou, Jiwoong Park, Haotian Xu, Chao Tian, Zhangyang Wang, and Yang Shen. Latent 3d graph diffusion. In *The Twelfth International Conference on Learning Representations*, 2023.

[39] Francis A Carey, Richard J Sundberg, Francis A Carey, and Richard J Sundberg. Nucleophilic substitution. *Advanced Organic Chemistry: Part A: Structure and Mechanisms*, pages 389–472, 2007.