GeomCLIP: Contrastive Geometry-Text Pre-training for Molecules

Teng Xiao[†], Chao Cui[‡], Huaisheng Zhu[†], Vasant G. Honavar[†]
[†]*Artificial Intelligence Research Laboratory*, The Pennsylvania State University
[‡]*Tsinghua Shenzhen International Graduate School*, Tsinghua University
{tengxiao, hvz5312, vhonavar}@psu.edu chaocui01@gmail.com

Abstract—**Pretraining molecular representations is crucial for drug and material discovery. Recent methods focus on learning representations from geometric structures, effectively capturing 3D position information. Yet, they overlook the rich information in biomedical texts, which detail molecules' properties and substructures. With this in mind, we set up a data collection effort for 200K pairs of ground-state geometric structures and biomedical texts, resulting in a `PubChem3D` dataset. Based on this dataset, we propose the `GeomCLIP` framework to enhance for multi-modal representation learning from molecular structures and biomedical text. During pre-training, we design two types of tasks, i.e., multimodal representation alignment and unimodal denoising pretraining, to align the 3D geometric encoder with textual information and, at the same time, preserve its original representation power. Experimental results show the effectiveness of `GeomCLIP` in various tasks such as molecular property prediction, zero-shot text-molecule retrieval, and 3D molecule captioning. Our code and collected dataset are available at https://github.com/xiaocui3737/GeomCLIP.**

Index Terms—**molecule conformation, CLIP, molecule description, geometric pretraining**

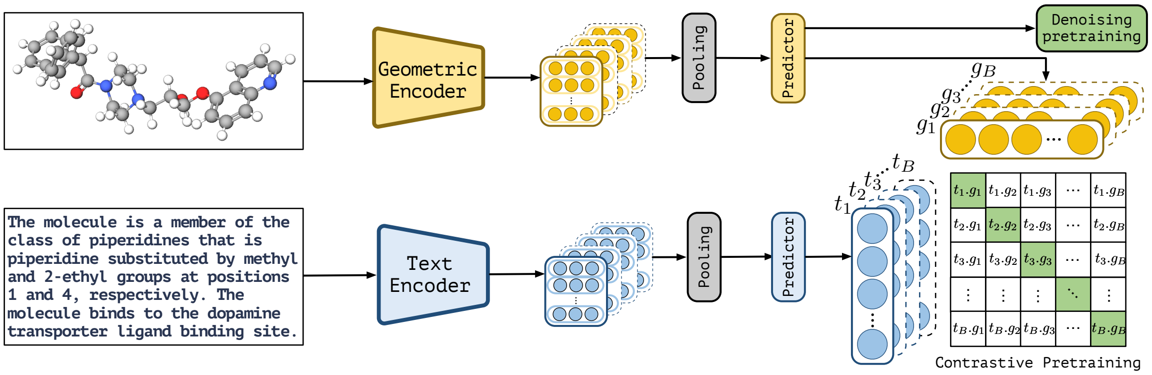
I. INTRODUCTION

The problem of learning useful image [35, 42], video [27], and audio [11] representations by incorporating text supervision [35] has been extensively studied in the literature. Because of the practical importance of generating molecular structures from textual descriptions of molecular characteristics, substructures, etc., there is a growing interest in multimodal representation learning from molecular structures and biomedical text [6, 8, 23, 26, 40, 53].

Existing works for multimodal learning on molecules and texts operate in two ways: (1) Sequence-based methods model molecules as 1D sequences, such as SMILES [5, 6, 25, 33]; (2) Graph-based methods aim to capture the 2D structures in molecules [4, 7, 26, 45, 51, 54, 55, 60]. These methods, however, do not investigate the effect of 3D geometric structures, which largely determine the physical and chemical properties of molecules [21, 22]. The learning of geometric representation for molecules is critical in various applications for quantum chemistry [10], protein structure prediction [38], materials science [37], and drug discovery [44]. Thus, a promising direction is to pretrain molecular representations based on 3D geometric structures and text descriptions, which is the main focus of this paper. Concurrent to our work, Tang *et al.* and Li *et al.* [18, 46] leverage 3D information to aid in multimodal learning. However, they rely directly on RDKit [17] to generate approximate 3D geometries from SMILES for evaluation and training, which are not ground-state geometries, are ambiguous, and can introduce significant noise [56, 57]. In addition, their heavy emphasis on cross-modal alignment causes them to overlook the unimodal information of 3D geometric structures, as demonstrated in our experiments.

In this work, we pioneer the creation of a high-quality dataset called `PubChem3D`, comprising `203,257` pairs of ground-state geometric structures and biomedical texts. While multimodal datasets exist in other domains [11, 35], a conspicuous gap remains when it comes to 3D molecule-text. This gap stems from two main challenges: (i) acquiring ground-state geometric structures is costly due to the time-intensive nature of methods such as density functional theory (DFT) [32]. (ii) Annotating the text of molecules is expensive given the need for depth of expert knowledge. To address these challenges, we describe a data collection effort, manually collecting molecule-text pairs from various licensed sources (see § III-A for details). To further enhance the representation learning of molecules through text, we propose a simple `GeomCLIP` framework, inspired by CLIP [35] to perform multimodal pretraining of 3D geometry structures and textual descriptions of molecules based on collected `PubChem3D`. As shown in Figure 1, `GeomCLIP` comprises two encoders, each tailored to learn 3D geometry or text representations of molecules. The geometry and text encoders are aligned via a task-agnostic joint contrastive objective to predict correct pairings within a batch of (geometry, text) pairs. `GeomCLIP` also employs a denoising objective to maintain the original effectiveness of the geometric encoder in capturing the 3D positions of molecules. Extensive experiments show the effectiveness of `GeomCLIP` in various downstream tasks such as molecular property prediction, text-molecule retrieval, and molecule captioning.

Our key contributions include: (1) We study a novel problem of learning multi-modal molecular representations, integrating 3D geometry with textual data. (2) We meticulously curate and develop a dataset combining text with ground-state 3D geometry of molecules. Building upon our dataset, we propose a straightforward yet highly effective approach, `GeomCLIP` for enhancing 3D geometry representations of molecules. (3) Extensive experiments show that our `GeomCLIP` yields improved performance on various downstream tasks such as 3D molecular property prediction, text-to-molecule (and back)



The molecule is a member of the class of piperidines that is piperidine substituted by methyl and 2-ethyl groups at positions 1 and 4, respectively. The molecule binds to the dopamine transporter ligand binding site.

$t_1 \cdot g_1$	$t_1 \cdot g_2$	$t_1 \cdot g_3$	$\cdots$	$t_1 \cdot g_B$
$t_2 \cdot g_1$	$t_2 \cdot g_2$	$t_2 \cdot g_3$	$\cdots$	$t_2 \cdot g_B$
$t_3 \cdot g_1$	$t_3 \cdot g_2$	$t_3 \cdot g_3$	$\cdots$	$t_3 \cdot g_B$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_B \cdot g_1$	$t_B \cdot g_2$	$t_B \cdot g_3$	$\cdots$	$t_B \cdot g_B$

Fig. 1. Framework of GeomCLIP: A dual-encoder pre-training scheme for 3D geometric molecules and their text observations.

retrieval, and 3D molecule captioning.

II. Related Work

A. Molecule Representation Learning

Molecular representation learning plays a crucial role in fields such as drug discovery [41] and material design [34]. In particular, self-supervised pretraining approaches applied to molecular data have exhibited promising performance without the need for labels. Some works focus on 1D SMILES strings [16] and 2D molecular graphs [47], leveraging sequence-based or graph-based pretraining methods to learn molecular representations effectively. However, they overlook the 3D geometric structure of molecules, which is vital since the physical and chemical properties of the molecule are significantly influenced by its 3D geometry [15, 39]. Thus, recent research focuses on learning representations from 3D geometric graphs through self-supervised methods [9, 24, 52]. Different from these studies, which neglect the semantic text information of molecules. Our work aims to enhance geometry representation of molecules by using textual descriptions.

B. Text-Molecule Multi-modal Learning

It has been broadly studied how to learn better image [35], video [27], and audio [11] representations by incorporating text supervision [35]. Since natural language enables nuanced expression of molecular characteristics, substructures, and biomedical understanding, multi-modal representation learning on molecule and biomedical text has recently attracted considerable attention [7, 20, 23, 26, 45]. Existing works for multi-modal learning on molecules in two ways: (1) Sequence-based methods model molecules on 1D sequences [6, 25, 33]; (2) Graph-based methods seek to capture 2D structures in molecules [7, 26, 29, 45, 49, 50]. However, these two lines of work investigate the effect of 3D geometric structures less. Concurrent to our work, [18, 46] leverage 3D information to help multi-modal learning. However, the 3D positions serve only as auxiliary information and cannot be used for supervision during alignment. Furthermore, they rely directly on RDKit [17] to generate 3D geometries from SMILES, which can be inaccurate and introduce significant noise [56, 57]. Consequently, their performance is suboptimal, as demonstrated in our experiments. In contrast, our work pioneers the creation of a dataset PubChem3D, comprising pairs of ground-state geometric structures and biomedical texts, and enhances the geometric representation learning of molecules through text.

III. Methods

A. Dataset Construction–PubChem3D

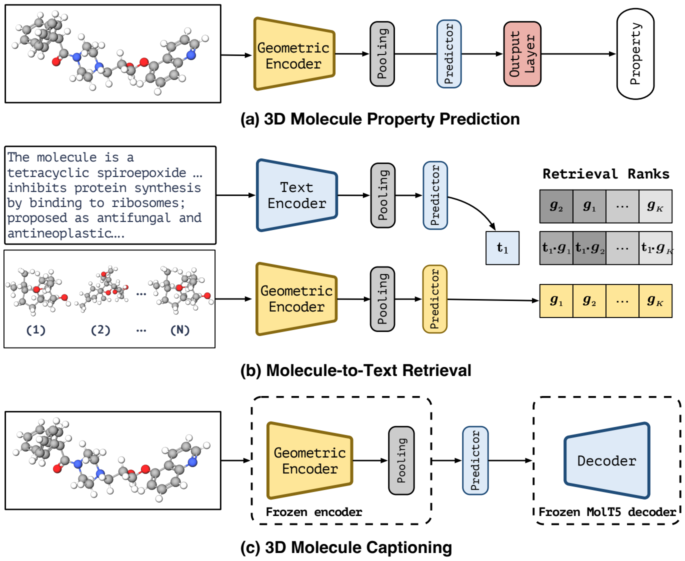
3D Geometry collection. To build our dataset, we consider two large databases containing ground-state geometries obtained by DFT computations. The first is PubChemQC [30], which contains over 3.9 million molecules, including their molecular graphs and ground-state 3D geometries. Furthermore, we also consider GEOM [1], a database of high-quality geometries for 430,000 molecules. Each molecule in GEOM has multiple geometric structures. As the top-10 conformers are sufficient to cover most conformers with an equilibrium state, we sample the top-10 geometry structures for each molecule with the highest possibility and lowest energy. We merge these two databases and extract the IUPAC International Chemical Identifier (InChI) [12] for each molecule.

Text annotation collection. Based on the 3D geometries collected, we aim to gather their text annotations. Manual annotation of molecules is costly due to its complexity. Hence, we turn to PubChem [13], a freely accessible and essential resource for chemical research, for comprehensive and authoritative molecular text annotations. PubChem contains text descriptions for many molecules, submitted by various research institutions. Using the InChI as a unique identifier, we retrieve text descriptions from PubChem for each geometry structure collected. The text annotations include descriptions related to properties or 3D geometry information. For instance, in molecule with CID 444795: “The molecule is a retinoic acid in which all four exocyclic double bonds have E- (trans-) geometry”, and (2) In molecule with CID 5375200:” The molecule is an abscisic acid in which the two acyclic double bonds both have trans-geometry”.

Final dataset. The final dataset PubChem3D contains 203,257 geometry-text pairs, in which 70,981 and 132,276 ground-state geometries come from PubChemQC and GEOM, respectively. We counted the types of molecular descriptions in the collected dataset, such as toxicity, solubility, and color in Table I, which shows the diversity of texts.

TABLE I
DATA VOLUME, AVERAGE NUMBER OF ATOMS, AND WORD COUNTS OF SOURCE DATASETS

Data Source	Quantity	Average Heavy Atoms	Average Words
PubChemQC	70981	13.79	68.54
GEOM-Drug	132276	25.58	31.81
GEOM-QM9	133885	8.80	—
PubChem3D	203257	21.46	44.64



(a) 3D Molecule Property Prediction

(b) Molecule-to-Text Retrieval

(c) 3D Molecule Captioning

Fig. 2. GeomCLIP can perform different downstream tasks: (a) molecular property prediction, where GeomCLIP is fine-tuned to predict properties of molecules. (b) Pretrained geometric and text encoders can perform zero-shot molecule-text retrieval without any annotations. (c) Molecule captioning, where we integrate GeomCLIP's aligned molecule representation with the MolT5 pretrained text decoder through optimizing predictor.

B. Joint Modeling of Geometries and Texts

Modality Alignment Task. Our work introduces a simple geometry-text pretraining approach, GeomCLIP, which enables advanced cross-modal representation. The alignment objective is inspired by the fact that both the semantic text representation and the 3D geometric representation of the same molecule should be as close to one another as possible. We align the embedding of corresponding geometry-text pairs while distancing other pairs in the same batch:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{B}|} \sum_{(g_i, t_i) \in \mathcal{B}} \log \text{NCE}(\boldsymbol{g}_i, \boldsymbol{t}_i) + \log \text{NCE}(\boldsymbol{t}_i, \boldsymbol{g}_i), \quad (1)$$

Where $\mathcal{B}$ represents the batch of geometry-text pairs, $\text{NCE}(\boldsymbol{g}_i, \boldsymbol{t}_i)$ and $\text{NCE}(\boldsymbol{t}_i, \boldsymbol{g}_i)$ denote the contrastive losses for geometry-to-text and text-to-geometry similarities, respectively. $\boldsymbol{g}_i$ and $\boldsymbol{t}_i$ are the representations from the geometric encoder f_θ and the text encoder f_ϕ, detailed in Appendix III-C, respectively. The geometry-to-text contrastive loss, $\text{NCE}(\boldsymbol{g}_i, \boldsymbol{t}_i)$, describes the likelihood of correctly ranking the molecules given its text.

$$\text{NCE}(\boldsymbol{g}_i, \boldsymbol{t}_i) = \log \frac{\exp(\cos(\boldsymbol{g}_i, \boldsymbol{t}_i)/\tau)}{\sum_{j=1}^{|\mathcal{B}|} \exp(\cos(\boldsymbol{g}_i, \boldsymbol{t}_j)/\tau)}, \quad (2)$$

where τ is temperature and $\boldsymbol{t}_i$ represents positive text embeddings that overlap with 3D geometry embedding. $\boldsymbol{t}_j$ is negative text embedding implicitly formed by other text embeddings in the batch. A symmetric equivalent, text-geometry contrastive loss $\text{NCE}(\boldsymbol{t}_i, \boldsymbol{g}_i)$ can be similarly calculated.

Denoising Pretaining Task. To preserve the unimodal information of 3D molecules when injecting the cross-modality information from biomedical texts, we incorporate a denoising pretraining task on geometric encoder that enables the model to effectively capture 3D structural information during alignment. Specifically, given an input 3D molecule G, we perturb it by adding i.i.d. Gaussian noise to its atomic positions $\boldsymbol{p}_i$ and masking atoms, resulting in a noisy version of the molecule:

$$\tilde{G} = \{(\tilde{\boldsymbol{p}}_1, \boldsymbol{x}_1) \ldots, (\tilde{\boldsymbol{p}}_N, \boldsymbol{x}_N)\}, \quad (3)$$
$$\text{where } \tilde{\boldsymbol{p}}_i = \boldsymbol{p}_i + \sigma \boldsymbol{\epsilon}_i \text{ and } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I),$$

where noise $\sigma = 1$. We also randomly mask 15% of the atoms in the molecule [59]. We denote the raw masked molecule and the noised complementary molecule by $G[m]$ and $\tilde{G}[1-m]$, respectively, where m is the binary index of the masked atoms. The task is formulated as a denoising autoencoder to predict noised coordinates and types of masked atoms:

$$\mathcal{L}_{\text{denoising}} = \sum_{i=1}^{|\mathcal{D}|} \| f_\psi(\tilde{\boldsymbol{g}}_i) - G_i[m] \|^2, \quad (4)$$

$\tilde{\boldsymbol{g}}_i = f_\theta(\tilde{G}_i[1-m])$ is the representation of noised complementary molecule, and f_ψ is the decoder proposed in [59].

Overall Objective: To promote representation alignment and maintain the capacity to capture geometric information, GeomCLIP simultaneously minimizes two losses:

$$\mathcal{L}_{\text{GeomCLIP}} = \mathcal{L}_{\text{con}} + \alpha \mathcal{L}_{\text{mask}}, \quad (5)$$

where α is the loss weighting hyperparameter.

C. Model Architecture

3D Molecular Encoder. The 3D molecular encoder f_θ in GeomCLIP draws inspiration from recent advancements in transformer-based models for encoding geometry information, as demonstrated in works [28, 59]. In this work, we employ Uni-Mol[1] [59] as our 3D molecular encoder, which is a transformer-based model with two inputs, atom types and atom 3D coordinates. The atom representation is initialized from atom types, by the embedding layer, and the representation for each pair of atoms is initialized using invariant spatial positional encoding from 3D coordinates. Then, the representations of atoms and atom pairs communicate with each other in the self-attention module. Formally, Uni-Mol f_{geom} performs 3D encoding steps to obtain sequential atomic representations:

$$[\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N] = f_{\text{geom}}(G), \quad (6)$$
$$\overline{\boldsymbol{z}} = \text{pooling}([\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]), \quad (7)$$

[1] https://github.com/dptech-corp/Uni-Mol/tree/main/unimol

TABLE II
RESULTS ON 12 QUANTUM MECHANICS PREDICTION TASKS FROM QM9 AND THE BEST RESULTS ARE MARKED IN **BOLD**.

Model	$\alpha \downarrow$	$\nabla\mathcal{E} \downarrow$	$\mathcal{E}_{\text{HOMO}} \downarrow$	$\mathcal{E}_{\text{LUMO}} \downarrow$	$\mu \downarrow$	$C_v \downarrow$	$G \downarrow$	$H \downarrow$	$R^2 \downarrow$	$U \downarrow$	$U_0 \downarrow$	ZPVE $\downarrow$
3D InfoMax	0.057	42.09	25.90	21.60	0.028	0.030	13.73	13.62	0.141	13.81	13.30	1.670
GraphMVP	0.056	41.99	25.75	21.58	0.027	0.029	13.43	13.31	0.136	13.03	13.07	1.609
MoleculeSDE	0.054	41.77	25.74	21.41	0.026	0.028	13.07	12.05	0.151	12.54	12.04	1.587
MoleculeJAE	0.056	42.73	25.95	21.55	0.027	0.029	11.22	10.70	0.141	10.81	10.70	1.559
3D-MoLM	0.055	42.53	24.76	21.39	0.023	0.026	12.51	11.55	0.135	10.78	11.22	1.468
Uni-Mol	0.051	41.01	23.31	20.75	**0.016**	0.023	9.52	8.73	0.128	9.77	9.65	1.345
GeomCLIP	**0.048**	**39.52**	**22.78**	**19.61**	**0.016**	**0.022**	**8.61**	**7.49**	**0.118**	**8.27**	**8.19**	**1.209**

TABLE III
MOLECULE-TEXT RETRIEVAL PERFORMANCES (%). † DENOTES METHOD WHICH IS ALSO PRETAINED ON OUR PUBCHEM3D.

	Molecule2Text		Text2Molecule	
Model	Acc	R@20	Acc	R@20
KV-PLM	35.12	78.91	36.33	74.71
MoMu	37.43	79.71	37.95	75.36
Text2Mol	38.27	79.52	38.96	77.27
MoleculeSTM	40.58	80.33	42.61	78.39
MolCA	47.71	82.47	42.71	80.73
MolCA†	49.96	83.61	44.35	81.71
3D-MoLM	50.92	81.34	43.15	80.25
3D-MoLM†	51.52	83.95	45.33	82.18
GeomCLIP	**53.50**	**85.53**	**48.71**	**83.50**

TABLE IV
MOLECULE CAPTIONING RESULTS ON PUBCHEM3D DATASET. FOR ALL METHODS, WE UTILIZE THE DECODER OF MOLT5-LARGE.

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
MolT5	26.71	18.34	35.25	16.91	25.37
MoMu	27.81	19.61	35.92	17.33	26.51
MolCA	28.02	20.68	35.73	18.06	27.70
3D-MoLM	28.71	22.19	36.83	19.86	28.75
GeomCLIP	**31.18**	**22.97**	**38.05**	**23.65**	**32.05**

where $\boldsymbol{z}_i$ corresponds to the representation of the i-th atom and we conduct a mean pooling operation to get the global representation of the molecule. We then use a projection MLP layer to map from the encoder's representation to the multimodal embedding space: $\boldsymbol{g} = f_{\text{proj}}(\overline{\boldsymbol{z}})$. The predictor customizes molecular representation for distinct pretraining tasks while sharing the previous backbone encoder, and acts as a natural task-specific adapter. The whole process can also be expressed as $\boldsymbol{g} = f_\theta(G)$ signifying the outcome of the encoding operation within the input 3D geometric graph.

Biomedical Text Encoder. The text encoder's foundation is rooted in the recent advances of transformer-based models to encode scientific textual descriptions into latent spaces. To inject potentially useful scientific knowledge from the literature into the text encoder, we initialize it with the Sci-BERT's checkpoint[2] [2] at denoted as $f_{\text{text}}(T)$, which is a transformer-based encoder pretrained on scientific publications. We utilize the pooling representation of the `[CLS]` token in Sci-BERT as whole text representations: $\boldsymbol{t}_{\text{CLS}} = f_{\text{text}}(T)[\text{CLS}]$. Similar to molecule, we also introduce an additional MLP predictor to map the text representation to multimodal space: The encoding process can also be expressed as $\boldsymbol{t} = f_\phi(T)$.

IV. RESULTS AND DISCUSSION

We evaluate GeomCLIP on three downstream tasks: property prediction, text-molecule retrieval, and molecule captioning. Figure 2 illustrates how GeomCLIP perform these tasks.

A. Implementation Details

The pretraining experiments are conducted on four NVIDIA A100 GPUs and downstream experiments are conducted on a single NVIDIA A100 GPU. For all methods, the batch size is 64 per GPU and the gradients are accumulated for 4 steps before updating, and the representation dimension is set to 512. The learning rate is set to 0.0001 with a warm-up for the first 1,000 steps and a linear decay for the remaining steps. We use Adam [14] optimizer for optimization, and the weight decay is set to 0.05. The parameters for Uni-mol are directly borrowed from [59]. A small grid search is used to select the best hyperparameter for all methods. For our GeomCLIP, we set temperature $\tau = 0.1$ and search α from {0.2, 0.4, 0.6, 0.8, 1.0}. We select the best configuration of hyper-parameters based using the validation set.

B. molecular property Prediction

We adopt the popular dataset: QM9 [36], which is a dataset of 134K molecules consisting of 9 heavy atoms, and the 12 tasks are related to quantum properties. We follow the official splits [48] and take 110K for training, 10K for validation, and 11K for testing. The metric is the mean absolute error (MAE).

Baselines. We consider the following baselines: 3D InfoMax [43], GraphMVP [24], MoleculeSDE [21], MoleculeJAE [3], Uni-Mol [59] and 3D-MoLM [18].

Results. As shown in Table II, GeomCLIP consistently outperforms all baselines on 12 tasks. This highlights the importance of 3D information for molecular property prediction and confirms GeomCLIP's superior capability in learning 3D molecular representations. Notably, GeomCLIP shows significant improvements over Uni-Mol and 3D-MoLM, demonstrating the advantages of integrating ground-state 3D geometrics with their textual descriptions for property prediction.

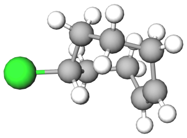
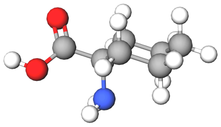
C. Zero-shot Molecule-Text Retrieval

With molecule-text aligned representation space, GeomCLIP allows for the retrieval of molecules using

[2] https://huggingface.co/allenai/scibert_scivocab_uncased

TABLE V

MORE CASES OF MOLECULE CAPTIONING AND CORRECTLY HIGHLIGHTED TEXTS ARE SHOWN IN RED. WE OBSERVE THAT GeomCLIP IS ABLE TO RECOGNIZE THE GENERAL CLASS OF MOLECULE IT IS ANALYZING AND IDENTIFY ITS FUNCTIONAL RELATIONS.

Molecule	Ground Truth	MolT5	GeomCLIP
	The molecule is a pyrimidone that is thymine in which the hydrogen at position 6 is substituted by a 1,3-dihydroxyisobutyl group. It is functionally related to a thymine.	The molecule is a pyrimidone that is thymine in which the hydrogen is replaced by a hydroxy group at the 5-position. It is functionally related to a uracil.	The molecule is a pyrimidone that is thymine in which the hydrogen at position 4 is replaced by a 1,3-dihydroxyacetone group. It is functionally related to a thymine.
	The molecule is an alpha-amino acid that is cyclohexanecarboxylic acid substituted by an amino group at position 1. It is functionally related to a cyclohexanecarboxylic acid.	The molecule is a non-proteinogenic alpha-amino acid that is serine in which the alcoholic hydroxy group has been formally oxidised to the corresponding formyl group. It is a non-proteinogenic alpha-amino acid and an alanine derivative.	The molecule is a non-proteinogenic alpha-amino acid that is that is cyclohexanecarboxylic acid substituted by an amino group at position 1. It is a non-proteinogenic alpha-amino acid and an alanine derivative. It is functionally related to a cyclohexanecarboxylic acid.

texts or vice versa. We evaluate GeomCLIP's performance in molecule-text retrieval on PubChem3D. We randomly select two subsets of 1, 500 pairs each for validation and testing. We measure retrieval performance using Accuracy and Recall@20 across the entire test set.

Baseline. We compare GeomCLIP with recent baselines: MoleculeSTM [23], MoMu [45], KV-PLM [58], Text2Mol [7], MolCA [26], and 3D-MoLM [18].

Results. The results in Table III reveal that GeomCLIP significantly outperforms existing baselines, including 1D SMILES-text models (e.g., KV-PLM) and 2D graph-text models (e.g., MoleculeSTM and MolCA). This underscores the advantage of incorporating 3D geometry in aligning the semantic spaces of molecules and texts. Furthermore, GeomCLIP exceeds the performance of 3D-MoLM, showcasing its ability to extract molecular features closely related to textual descriptions. The superior performance of GeomCLIP can be partially credited to our curated PubChem3D dataset, consisting of high-quality ground-state geometries. This is evidenced by 3D-MoLM† improved performance when retrained on PubChem3D, compared to it using unreal 3D geometries generated by RdKit.

D. Molecule Captioning

Essentially, GeomCLIP transforms 3D molecule presentation into underlying text space, thereby enabling it to perform a molecule-to-text generation task. Inspired by image captioning, which integrates CLIP's pre-trained image embeddings with GPT-2 pre-trained text generation model through a learnable mapping network. We adopt a similar strategy as shown in Figure 2, to facilitate integration with the MolT5's [6] pre-trained molecule-to-text decoder, we optimize our predictor f_θ with next-token prediction loss. This approach lets us leverage the existing pre-trained decoder without the need for training from scratch. We use BLEU [31] and ROUGE [19] as evaluation metrics.

Baselines. We compare with MolT5 [6], MoMu [45], Molca [26] and 3D-MoLM [18].

Results. Table IV shows that our GeomCLIP consistently outperforms the baselines by a large margin. Specifically, it achieves improvements of up to 2.47 and 3.79 points compared to 3D-MoLM in BLEU-2 and ROUGE-2, respectively. This showcases the effectiveness of 3D molecule-text alignment training in connecting 3D molecular representations with the input space of language models. Table V shows several molecule captioning examples of different models' outputs. We can observe that using our GeomCLIP as encoder leads to more accurate descriptions of molecule structures compared to baseline MolT5.

V. CONCLUSION

GeomCLIP introduces a novel pre-training strategy that effectively combines 3D geometric information of molecules with textual descriptions, addressing the issue that corresponding texts are usually overlooked in current molecular representation learning methods. Also, by leveraging a contrastive learning approach and a denoising pre-training strategy, GeomCLIP-induced geometric encoder has been verified to be effective across multiple downstream applications, including property prediction, molecule-text retrieval, and molecule captioning. The creation of the new PubChem3D dataset which aligns geometric representations with their diverse descriptions enriches the resources that researchers can use, thus promoting future study.

ACKNOWLEDGEMENTS

The work of Teng Xiao, Huaisheng Zhu was supported in part by graduate research assistantships funded by grants from the National Science Foundation (2226025, 2020243) to Vasant G. Honavar and from the National Center for Advancing Translational Sciences, and the National Institutes of Health (UL1 TR002014) to the Pennsylvania State University.

REFERENCES

[1] S. Axelrod and R. Gomez-Bombarelli, "Geom, energy-annotated molecular conformations for property prediction and molecular generation," *Scientific Data*, p. 185, 2022.

[2] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *EMNLP-IJCNLP*, 2019, pp. 3615–3620.

[3] J. Chen, X. Zhang, Z.-M. Ma, S. Liu *et al.*, "Molecule joint auto-encoding: Trajectory pretraining with 2d and 3d diffusion," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[4] Z. Chen, T. Xiao, and K. Kuang, "Ba-gnn: On learning bias-aware graph neural network," in *ICDE*, 2022, pp. 3012–3024.

[5] D. Christofidellis, G. Giannone, J. Born, O. Winther, T. Laino, and M. Manica, "Unifying molecular and textual representations via multi-task language modelling," in *ICML*. PMLR, 2023, pp. 6140–6157.

[6] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji, "Translation between molecules and natural language," in *EMNLP*, 2022, pp. 375–413.

[7] C. Edwards, C. Zhai, and H. Ji, "Text2mol: Cross-modal molecule retrieval with natural language queries," in *EMNLP*, 2021, pp. 595–607.

[8] Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen, "Mol-instructions: A large-scale biomolecular instruction dataset for large language models," *arXiv preprint arXiv:2306.08018*, 2023.

[9] S. Feng, Y. Ni, Y. Lan, Z.-M. Ma, and W.-Y. Ma, "Fractional denoising for 3d molecular pre-training," in *ICML*, 2023, pp. 9938–9961.

[10] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *ICML*, 2017, pp. 1263–1272.

[11] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP*, 2022, pp. 976–980.

[12] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "Inchi, the iupac international chemical identifier," *Journal of cheminformatics*, pp. 1–34, 2015.

[13] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, "Pubchem in 2021: new data content and improved web interfaces," *NAR*, pp. D1388–D1395, 2021.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd IICLR*, Y. Bengio and Y. LeCun, Eds., 2015.

[15] P. I. Koukos, L. C. Xue, and A. M. Bonvin, "Protein–ligand pose and affinity prediction: Lessons from d3r grand challenge 3," *Journal of computer-aided molecular design*, vol. 33, pp. 83–91, 2019.

[16] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (selfies): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, p. 045024, 2020.

[17] G. Landrum *et al.*, "Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling," *Greg Landrum*, vol. 8, p. 31, 2013.

[18] S. Li, Z. Liu, Y. Luo, X. Wang, X. He, K. Kawaguchi, T.-S. Chua, and Q. Tian, "Towards 3d molecule-text interpretation in language models," *arXiv preprint arXiv:2401.13923*, 2024.

[19] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[20] P. Liu, Y. Ren, J. Tao, and Z. Ren, "Git-mol: A multi-modal large language model for molecular science with graph, image, and text," *Computers in Biology and Medicine*, p. 108073, 2024.

[21] S. Liu, W. Du, Z.-M. Ma, H. Guo, and J. Tang, "A group symmetric stochastic differential equation model for molecule multi-modal pretraining," in *ICML*, 2023, pp. 21 497–21 526.

[22] S. Liu, H. Guo, and J. Tang, "Molecular geometry pretraining with se (3)-invariant denoising distance matching," in *ICLR*, 2022.

[23] S. Liu, W. Nie, C. Wang, J. Lu, Z. Qiao, L. Liu, J. Tang, C. Xiao, and A. Anandkumar, "Multi-modal molecule structure–text model for text-based retrieval and editing," *Nature Machine Intelligence*, pp. 1447–1457, 2023.

[24] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, "Pre-training molecular graph representation with 3d geometry," in *ICLR*, 2021.

[25] Z. Liu, W. Zhang, Y. Xia, L. Wu, S. Xie, T. Qin, M. Zhang, and T.-Y. Liu, "MolXPT: Wrapping molecules with text for generative pre-training," in *ACL(Short Papers)*, 2023, pp. 1606–1616.

[26] Z. Liu, S. Li, Y. Luo, H. Fei, Y. Cao, K. Kawaguchi, X. Wang, and T.-S. Chua, "Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter," in *EMNLP*, 2023.

[27] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[28] S. Luo, T. Chen, Y. Xu, S. Zheng, T.-Y. Liu, L. Wang, and D. He, "One transformer can understand both 2d & 3d molecular data," in *ICLR*, 2022.

[29] Y. Luo, K. Yang, M. Hong, X. Liu, and Z. Nie, "Molfm: A multimodal molecular foundation model," *arXiv preprint arXiv:2307.09484*, 2023.

[30] M. Nakata and T. Shimazaki, "Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry," *JCIM*, pp. 1300–1308, 2017.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[32] R. G. Parr, S. R. Gadre, and L. J. Bartolotti, "Local density functional theory of atoms and molecules," *PNAS*, vol. 76, no. 6, pp. 2522–2526, 1979.

[33] Q. Pei, W. Zhang, J. Zhu, K. Wu, K. Gao, L. Wu, Y. Xia, and R. Yan, "Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations," in *Proceedings of the 2023 Conference*

on Empirical Methods in Natural Language Processing, 2023, pp. 1102–1123.

[34] R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao *et al.*, "Data-driven strategies for accelerated materials design," *Accounts of Chemical Research*, pp. 849–860, 2021.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[36] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific data*, pp. 1–7, 2014.

[37] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, p. 83, 2019.

[38] K. Schütt, O. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra," in *ICML*. PMLR, 2021, pp. 9377–9388.

[39] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet–a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, 2018.

[40] P. Seidl, A. Vall, S. Hochreiter, and G. Klambauer, "Enhancing activity prediction models in drug discovery with the ability to understand human language," in *ICML*, 2023, pp. 30 458–30 490.

[41] J. Shen and C. A. Nicolaou, "Molecular property prediction: recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, pp. 29–36, 2019.

[42] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022, pp. 15 638–15 650.

[43] H. Stärk, D. Beaini, G. Corso, P. Tossou, C. Dallago, S. Günnemann, and P. Liò, "3d infomax improves gnns for molecular property prediction," in *ICML*, 2022, pp. 20 479–20 502.

[44] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, pp. 688–702, 2020.

[45] B. Su, D. Du, Z. Yang, Y. Zhou, J. Li, A. Rao, H. Sun, Z. Lu, and J.-R. Wen, "A molecular multimodal foundation model associating molecule graphs with natural language," *arXiv preprint arXiv:2209.05481*, 2022.

[46] X. Tang, A. Tran, J. Tan, and M. B. Gerstein, "Mollm: A unified language model to integrate biomedical text with 2d and 3d molecular representations," *bioRxiv*, pp. 2023–11, 2023.

[47] Y. Wang, R. Magar, C. Liang, and A. Barati Farimani, "Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast," *JCIM*, vol. 62, no. 11, pp. 2713–2725, 2022.

[48] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chemical science*, pp. 513–530, 2018.

[49] T. Xiao, Z. Chen, Z. Guo, Z. Zhuang, and S. Wang, "Decoupled self-supervised learning for graphs," *Advances in Neural Information Processing Systems*, pp. 620–634, 2022.

[50] T. Xiao, Z. Chen, D. Wang, and S. Wang, "Learning how to propagate messages in graph neural networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1894–1903.

[51] T. Xiao, C. Cui, H. Zhu, and V. G. Honavar, "Molbind: Multimodal alignment of language, molecules, and proteins," *arXiv preprint arXiv:2403.08167*, 2024.

[52] T. Xiao and D. Wang, "A general offline reinforcement learning framework for interactive recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 4512–4520.

[53] T. Xiao, Y. Yuan, H. Zhu, M. Li, and V. G. Honavar, "Cal-dpo: Calibrated direct preference optimization for language model alignment," *Advances in Neural Information Processing Systems*, 2024.

[54] T. Xiao, H. Zhu, Z. Chen, and S. Wang, "Simple and asymmetric graph contrastive learning without augmentations," *Advances in Neural Information Processing Systems*, 2024.

[55] T. Xiao, H. Zhu, Z. Zhang, Z. Guo, C. C. Aggarwal, S. Wang, and V. G. Honavar, "Efficient contrastive learning for fast and accurate inference on graphs," in *Forty-first International Conference on Machine Learning*, 2024.

[56] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang, "Geodiff: A geometric diffusion model for molecular conformation generation," in *ICLR*, 2021.

[57] Z. Xu, Y. Luo, X. Zhang, X. Xu, Y. Xie, M. Liu, K. Dickerson, C. Deng, M. Nakata, and S. Ji, "Molecule3d: A benchmark for predicting 3d geometries from molecular graphs," *arXiv preprint arXiv:2110.01717*, 2021.

[58] Z. Zeng, Y. Yao, Z. Liu, and M. Sun, "A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals," *Nature communications*, p. 862, 2022.

[59] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, and G. Ke, "Uni-mol: A universal 3d molecular representation learning framework," in *ICLR*, 2022.

[60] H. Zhu, T. Xiao, and V. G. Honavar, "3m-diffusion: Latent multi-modal diffusion for text-guided generation of molecular graphs," *arXiv preprint arXiv:2403.07179*, 2024.