Generating Molecules from Natural Language with Multimodal Contrastive Pre-Training

Stanford CS224N Custom Project

Romain Lacombe
Stanford University
rlacombe@stanford.edu

Kateryna Pistunova
Stanford University
kpistunova@stanford.edu

David Lüdeke
Stanford University
dludeke@stanford.edu

Andrew Gaut
Stanford University (CS 224W)
agaut@stanford.edu

Jeff He
Stanford University (CS 224W)
jeff2024@stanford.edu

Abstract

Deep molecular neural representations have demonstrated significant potential in accelerating scientific research in the natural sciences. However, existing AI models typically focus on pure graph-based representations, or knowledge extraction from natural language, leaving a wide gap between these two modalities. In this paper, we explore what a foundation model for chemistry could look like, by introducing and evaluating improvements on a recently proposed molecular multimodal (MoMu) model for contrastive joint text-graph representation learning (Su et al. [2022]). By implementing neural relevance scoring strategies for retrieving molecule text descriptions, and augmenting molecular graphs via more chemically relevant transformations, we surpass MoMu's performance on most *MoleculeNet* molecular property predictions tasks. We explore zero-shot molecular generation based on the improved graph encoders we trained, propose avenues for future work, and highlight the ethics and safety implications of generative AI for molecules.

1 Introduction

Deep molecular representation learning models have demonstrated significant potential in accelerating important tasks in natural sciences, such as predicting molecular properties, or generating and screening candidates for drug discovery. However, existing AI models typically focus on either graph-based representations, or knowledge extraction from natural language, leaving a gap between these two modalities.

In this paper[1], we focus on improving the work presented in the paper "A Molecular Multimodal Foundation Model (MoMu) Associating Molecule Graphs with Natural Language"Su et al. [2022], which aims to bridge the gap between language-based and graph-based representations of molecules by introducing a molecular graph-text multimodal model trained through contrastive learning.

Our primary objective is to investigate how much molecular information natural language encodes by exploring the generation of graphs representing molecules based on textual descriptions of their desired properties. It is important to note at the outset that, while aligning graph and text representations is necessary in order to perform multimodal tasks such as molecule generation, it is an open question whether this also improves the performance of graph representation on other downstream tasks such as property prediction. We set out to explore this question in this paper.

[1] We are sharing the project across two classes—Prof. Chris Manning's Natural Language Processing with Deep Learning (CS224N) and Prof. Jure Leskovec's Machine Learning with Graphs (CS224W)—and are collaborating with Jeff He and Andrew Gaut (CS224W students).

Our contributions. We propose improvements to the model presented in the original paper aimed at enhancing the molecular representations it learns, as measured by their performance on downstream tasks. Specifically, we hypothesize that the text retrieval strategy employed for training the text encoder through contrastive learning is critical for the model's performance. We present neural relevance based methods to improve text sampling over uniform random draw, and implement novel approaches for chemically-relevant graph augmentation. We hope our improvements on the original model, along with experimental results and identified avenues for future work, contribute to the development of more expressive multimodal molecular models for the natural sciences.

2 Related Work

Molecular representation learning. Molecular representation learning plays a vital role in the study and analysis of chemical compounds. Traditional molecular representations, such as SMILES strings Weininger [1988] and InChI Heller et al. [2015], are linear notations that encode molecular structures into strings. However, they have limitations when it comes to modeling complex molecular properties and reactions. To overcome these shortcomings, several graph-based representations have been proposed, including molecular graphs Kearnes et al. [2016] and attributed molecular graphs Duvenaud et al. [2015]. These graph-based representations have been shown to better capture the structural and functional properties of molecules.

Graph-based methods. Graph-based methods have been widely adopted in cheminformatics to model molecules and predict their properties. Graph Convolutional Networks (GCNs) Kipf and Welling [2016], Graph Attention Networks (GATs) Veličković et al. [2017], and Message Passing Neural Networks (MPNNs) Gilmer et al. [2017] are popular graph neural network architectures used for molecular property prediction, drug discovery, and materials science. These methods have demonstrated their potential to outperform traditional machine learning algorithms in various molecular prediction tasks Wu et al. [2018a].

Language models in chemistry. The advent of deep learning-based language models, such as GPT Radford et al. [2018], BERT Devlin et al. [2018], and T5 Raffel et al. [2019], has revolutionized natural language processing. Researchers have started exploring their potential in cheminformatics, leading to the development of models such as ChemBERTa Korolev et al. [2020] and MolBERT Napolitano et al. [2021]. These models have shown promising results in tasks like reaction prediction, retrosynthesis, and molecular property prediction. Furthermore, generative models have been used to design novel molecules with desired properties Gómez-Bombarelli et al. [2016], Kusner et al. [2017].

Multimodal learning. Multimodal learning aims to integrate information from different modalities to improve overall model performance. Recent advances in this area include models such as CLIP Radford et al. [2021] and ALIGN Jia et al. [2021], which bridge the gap between vision and language tasks. MoMu Su et al. [2022] is a pioneering work in the field of molecular multimodal learning, aiming to combine graph-based and language-based representations of molecules for improved performance in various tasks.

3 Approach

3.1 Foundation Model Paradigm: Pretrain & Finetune

We approach the task of generating molecules from natural language through the lens of the foundation model paradigm, following the pretrain and finetune methodology (Liang et al. [2021]). Specifically, we proceed in three steps: pretraining a model, evaluating its performance by fine-tuning it on downstream classification tasks, and ultimately using it as the basis for a downstream generation task.

We train and evaluate joint text-graph representations for deep multi-modal generation as follows:

- We jointly pretrain a bidirectional transformer for text encoding (SciBERT) Beltagy et al. [2019b] and a Graph Isomorphism Network (GIN) Xu et al. [2018] for graph encoding through contrastive learning over a joint text-graph dataset;

- We then finetune the graph encoder on a series of molecular property prediction tasks, and evaluate the quality of our pretraining based on performance on these downstream tasks;
- Finally, we use embeddings encoded by our text encoder as the input to a separate flow-based deep graph generative model, to generate molecules from natural language.

3.2 Multimodal Contrastive Pre-Training

3.2.1 Contrastive Learning Strategy

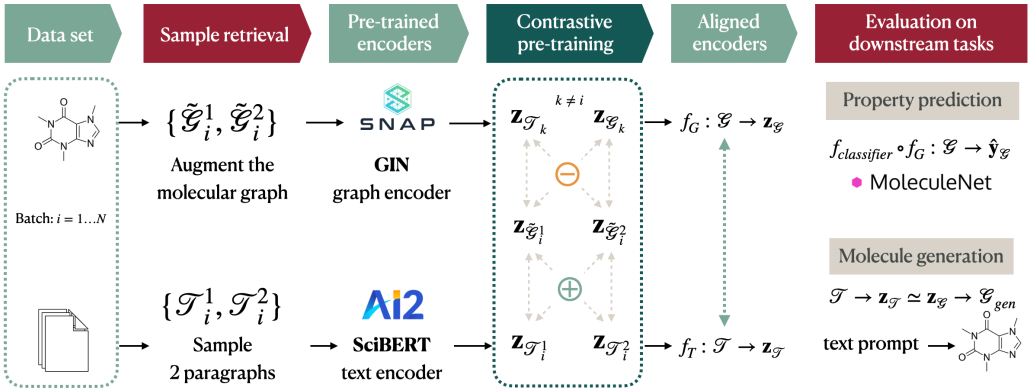

Data set
Sample retrieval
Pre-trained encoders
Contrastive pre-training
Aligned encoders
Evaluation on downstream tasks
Batch: $i = 1...N$
$\{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$
Augment the molecular graph
SNAP
GIN
graph encoder
$k \neq i$
$\mathbf{z}_{\mathcal{T}_k}$
$\mathbf{z}_{\mathcal{G}_k}$
$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$
$\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$
$\ominus$
$\oplus$
$\mathbf{z}_{\mathcal{T}_i^1}$
$\mathbf{z}_{\mathcal{T}_i^2}$
$f_G : \mathcal{G} \to \mathbf{z}_{\mathcal{G}}$
$f_T : \mathcal{T} \to \mathbf{z}_{\mathcal{T}}$
$\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$
Sample 2 paragraphs
Ai2
SciBERT
text encoder
Property prediction
$f_{classifier} \circ f_G : \mathcal{G} \to \hat{\mathbf{y}}_{\mathcal{G}}$
● MoleculeNet
Molecule generation
$\mathcal{T} \to \mathbf{z}_{\mathcal{T}} \simeq \mathbf{z}_{\mathcal{G}} \to \mathcal{G}_{gen}$
text prompt

Figure 1: Contrastive pre-training of joint representations of molecular graph-text. Our contribution focuses on improvements to the **retrieval strategy**, which we evaluate on **downstream tasks** .

The core machine learning task in our approach is to learn aligned representations of pairs of molecule graphs and text describing the properties of that molecule. We use the self-supervised learning technique of contrastive learning, based on a loss function which promotes smaller euclidian distances in the joint latent space between graph and text samples of the same data samples (positive pairs), and larger euclidian distances between different samples (negative pairs). Building on the original implementation of Su et al. [2022], we use the following contrastive learning paradigm:

- At train time, form a random batch of molecules $i \in [1, ..., N]$ molecules.
- From the original $\mathcal{G}_i$ graph, form $2N$ graphs $\{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$ through random augmentations.
- Randomly sample $2N$ text samples $\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$, each associated with molecule i.

The InfoNCE loss function (Oord et al. [2018]) promotes proximity between matching cross-modality $(\mathcal{T}_i, \tilde{\mathcal{G}}_i)$ and graph $(\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2)$ embedding pairs from the same molecule, and higher distance between non-matching pairs, by summing the following pair-wise losses (here, for a cross-modality pair):

$$\ell(\mathcal{T}_i, \tilde{\mathcal{G}}_i) = -\log \frac{\exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_i^{\mathcal{G}})/\tau\right)}{\sum_{j \neq i} \exp\left(\cos(\mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_j^{\mathcal{G}})/\tau\right)}$$

3.2.2 Pre-trained text and graph encoders

The goal of contrastive pre-training is to align the representations of matched text fragments and molecular 2D graphs in the same embeddings space. For efficiency purposes, we start with previously pre-trained models for both our text encoder and our graph encoder, which we present.

To optimize for extraction of information from fragment of scientific papers, we base our text encoder on SciBERT (Beltagy et al. [2019a]), a pre-trained language model based on BERT (Devlin et al. [2019]), trained on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks.

Graph Isomorphism Networks (GINs) are a class of Graph Neural Networks (GNNs) which are demonstrably the most expressive graph network. For our graph encoder, we use the GraphCL 80 GIN model from You et al. [2020], a 1.9 million parameters model pre-trained through graph contrastive learning on *MoleculeNet* Wu et al. [2018b].

3.3 Relevance-Based Sampling

3.3.1 Neural Text Relevance Scoring

The original paper retrieves text sequences by sampling two paragraphs per molecule with uniform sampling. This approach does not consider the relevance of the retrieved paragraphs to the molecule's properties or structure, as mentioned by the authors themselves.

To address this issue, we propose a neural text retrieval strategy informed by the relevance of each text segment for the molecule it describes. For each paragraph, we compute the cosine similarity between the paragraph and a natural language query. We then sample paragraphs according to the distribution of the cosine similarity scores, with a higher probability of selecting paragraphs with higher cosine similarity.

Specifically, we experiment with queries designed to ensure that the selected paragraphs are more likely to be relevant to the molecule's properties or structure:

- **mean similarity**: average embedding vector of name and top 20 synonyms of the molecule
- **max similarity**: maximum cosine score with any of the name or the top 20 synonyms
- **sentence similarity**: natural language query consisting of the following sentence:

> *"Molecular, chemical, electrochemical, physical, quantum mechanical, biochemical, biological, medical and physiological properties, characteristics, and applications of {$name$}, a compound also known as {$synonym_1$}, …, or {$synonym_n$}."*

We then apply **epsilon sampling** to rank paragraph by the cosine score and sample only from scores above a threshold, using the probability distribution (re-normalized over the strictly positive terms). We also introduce a temperature hyper-parameter to skew the sampling distribution towards the highest cosine score terms, and run experiments for several values:

$$\mathbb{P}(\text{paragraph}_{i\in[1..N]}) = \text{Softmax}\left(\frac{\cos(\mathbf{z}_{query}, \mathbf{z}_i)}{\text{Temp}}\right) \quad \text{if } \geq \frac{\epsilon}{N}$$

3.3.2 Chemically Relevant Graph Augmentations

The original model aim to improve on graph contrastive learning by adding random graph augmentations at sample time and add a intra-modality constrative loss terms to the loss function. Howeer, their augmentation strategy uses random node drops or random-walk subgraphs, which do not take into account the chemical constraints of the molecule graph such as bond valence.

Here, we introduce graph augmentations inspired by actual chemical reactions, in hope to improve molecule representations after training. Specifically, we implement augmentations corresponding to the following methylation/de-methylation and amination/de-amination reactions:

$$\text{R}-\text{H} + \text{CH}_4 \rightleftharpoons \text{R}-\text{CH}_3 + \text{H}_2 \qquad \text{R}-\text{H} + \text{NH}_3 \rightleftharpoons \text{R}-\text{NH}_2 + \text{H}_2$$

3.4 Molecular Property Prediction

Molecular generation is an open-ended tasks, and measuring performance is a challenge in the absence of access to a wet lab to synthesize and evaluate properties of generated molecules. Instead, as our measure of the quality of our graph representations, we use a set of downstream classification tasks aimed at predicting various properties of molecules based only on their molecular graph structure.

From pre-training, we obtain two encoders that embed molecular graph and text descriptions within the same joint latent space:

$$f_G : \mathcal{G} \to \mathbf{z}_{\mathcal{G}} \qquad f_T : \mathcal{T} \to \mathbf{z}_{\mathcal{T}}$$

We fine-tune our graph encoder for classification problems by adding a classifier layer, which we adapt and fine-tune to each specific downstream task and dataset:

$$\text{CLASSIFIER}(\cdot) \circ f_G : \mathcal{G} \to \hat{\mathbf{y}}_{\mathcal{G}}$$

3.5 Zero-Shot Molecular Generation from Natural Language

The ultimate goal of training aligned text-graph neural representations is to enable the flow of information from one modality to the other. Specifically, the authors of Su et al. [2022] introduce zero-shot molecular generation from natural language. Here, we leverage our text encoder's ability to encode a representation of text that corresponds to a molecular graph in the joint latent space.

In this task, we use MoFlow, a previously-trained flow-based deep molecular generative model that's design to generat chemically valid molecular graphs based on their latent representation (Zang and Wang [2020]). We feed embeddings generated by our text encoder to MoFlow and use it to generate candidate molecular graphs intended to match the natural language prompt:

$$\text{MoFlow}(\cdot) : \mathbf{z}_{\mathcal{G}} \to \mathcal{G}_{gen} \quad \Rightarrow \quad \text{MoFlow}(\cdot) \circ f_{\mathcal{T}} : \mathcal{T} \to \mathbf{z}_{\mathcal{T}} \simeq \mathbf{z}_{\mathcal{G}} \to \mathcal{G}_{gen}$$

Here, zero-shot refers to composing our text encoder with the flow-based model out of the box, without any fine-tuning. Training our own flow-based model was beyond the scope of this project. For future work, we could train a generative model to reverse our graph encoder, and generate molecules from prompts embedded by our text encoder.

4 Experiments

4.1 Data

4.1.1 Contrastive Pre-Training Data

We train on the molecular graph-text pairs dataset presented in figure 3, constructed in Su et al. [2022] by retrieving scientific papers in the S2ORC [Lo et al., 2020] database by using the name and synonyms of compounds from *PubChem* [Kim et al., 2022] as query, and transforming their SMILES intro a molecular graph using OGB smile2graph Hu et al. [2020].

The dataset comprises of 15,613 graph-document pairs, with 37 million paragraphs or 47.5 gigabytes of text ($\sim$3 megabytes per molecule). To make training tractable, the text beyond the first 500 paragraphs per molecule is left out.

Importantly, the molecule graph and text sequences datasets are only weakly correlated: text fragments are extracted form the original SO2RC database on the basis of the name of the molecule appearing in that paragraph, with no further controls for relevance.

Lastly, the dataset is highly bi-modal: out of 15,613 text-graph pairs, 8,700 samples have less than 50 paragraphs of text, and 2,967 molecules have $\geq$500 paragraphs. Our sampling strategies based on cosine similarity scores aim to counter this inherent imbalance, by training on most of the small text corpus for the sparsely described molecules, and only relevant text for richly described ones.

4.1.2 Downstream Task Evaluation Data

We evaluate our models by fine-tuning them on the relevant classification task of a series of chemical and biological datasets from *MoleculeNet* Wu et al. [2018b], a multi-faceted set of benchmark tasks and reference datasets. Specifically, we use the following datasets retrieved from DeepChem:

- BACE: classification of inhibitors of a human enzyme involved in Alzheimer, which, if blocked, may prevent build up of proteins in the brain associated with the disease.
- BBBP: classification for the prediction of blood-brain barrier penetration by small molecules.
- Clintox: classification of drugs approved/rejected by the FDA for toxicity.
- MUV: classification for virtual molecule screening built on *PubChem*.
- SIDER: classification of adverse side reactions of marketed drugs.
- Tox21: classification of toxicity measured by biological reactions and stress response.
- ToxCast: classification over 600 tasks linked to *in vitro* toxicology data.

4.2 Evaluation Metrics

We use downstream task Area Under Receiver-Operator Curve (AUROC) metric as our main performance metric to evaluate the quality of the representations learned through each experiment.

Specifically, we fine-tune our pre-trained models on 7 different classification tasks on corresponding *MoleculeNet* datasets. We fine-tune each model separately for 20 epochs for each of the data sets (15 epochs for MUV), and for 3 random seeds each (about 40 minutes for each model). We then compute the test set AUROC obtained on the epoch with the highest validation set during fine-tuning. Finally, we average this score over several random seeds to report more statistically robust results.

We use this max-validation test set AUROC performance metric averaged across downstreams as a measure of the quality of each pretrained model, and report it in the summary table in figure 4.

4.3 Experimental Details

The core hypothesis behind our work is that training on more relevant text should increase the quality of the graph representations we learn. We set out to test this hypothesis with the experiments below.

4.3.1 Baseline

As our baseline, we use the original model presented in Su et al. [2022] which uses the architecture introduced here, but samples text and graph augmentations with a uniform random distribution. We run a pre-training of this model on our full joint text-graph dataset, for 30 epochs with learning rate 0.001, which took about 2 hours on a Google Cloud NVIDIA A100 VM GPU.

We also seek to evaluate any gains in performance on downstream tasks from contrastive pre-training, and present the downstream tasks performance of the GraphCL 80 GIN with *no joint graph-text pre-training*, as reported by the authors of Su et al. [2022].

4.3.2 Naive Text Relevance

Relevance-based sentences pruning: to explore whether sampling more relevant text during pre-training improves the quality of learned representations, we pre-process each paragraph in the text corpus by leaving only the sentences of each paragraph which explicitly include the name of the molecule, or any of the top 20 synonyms retrieved from PubChem Kim et al. [2022]. This results in an $\sim$80% smaller text corpus, and we report results in the Experiments section showing that this approach does result in measurable gains in downstream performance.

Length-based paragraph pruning: for validation purposes, we control whether these gains may be simply due to training on a smaller, thus potentially less noisy dataset. We train and evaluate a model trained on the first 256 characters of each paragraph.

Performance gains from the first method are consistent across downstream property prediction tasks, compared to inconsistent results form the second one, which encourages us to seek to improve text relevance at sampling time.

4.3.3 Neural Text Relevance

Cosine similarity pre-processing: to speed up retrieval at train time, we pre-compute the cosine similarity scores for each paragraph in the dataset, with each of the query types in our experiments (`mean`, `max`, `sentence`). We computed embeddings and similarity score over the entire dataset and set of queries, for a total of about 6 hours on a Google Cloud VM NVIDIA A100 GPU.

Cosine similarity retrieval: based on the hyper-parameter search detailed above, we ran experiments on the 3 cosine similarity query types, with 3 hyper-parameters each. We conducted an intrinsic evaluation based on hand labeling of a small sub-set of text paragraphs, which we present in appendux (table 1). To maximize the retrieval F1 score implied by our intrinsic evaluation, we chose $\epsilon = 0.5$ and Temperature $= \{0.05, 0.1, 0.2\}$. For each model, we ran 30 epochs over the shortened dataset, for about 2h each.

Pre-training	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Before any pre-training (reported)	70 ±0	65.8 ±0	74 ±0	63.4 ±0	57.3 ±0	58 ±0	71.8 ±0
Baseline	70.31 ±3.67	68.04 ±1.67	74.6 ±0.68	63.27 ±0.53	59.39 ±0.51	61.09 ±1.1	75.66 ±0.55
Baseline (pruned)	71.14 ±1.93	67.86 ±2.1	74.77 ±0.37	62.71 ±1.3	59.31 ±0.72	61.17 ±1.39	75.18 ±1.06
Baseline (relevant)	72.13 ±0.47	68.73 ±2.21	74.85 ±0.3	62.47 ±0.66	60.05 ±0.7	59.99 ±1.73	74.47 ±0.95
Cosine similarity mean (best)	72.6 ±2.77	68.48 ±1.68	74.54 ±0.7	63.37 ±0.72	60.07 ±0.41	61.36 ±3.36	75.07 ±1.13
Cosine similarity max (best)	72.71 ±0.59	68.27 ±2.35	74.77 ±0.45	63.73 ±0.59	60.14 ±1.05	62.28 ±1.61	75.15 ±1.07
Cosine similarity sent (best)	72.05 ±0.52	68.11 ±2.5	74.94 ±0.79	63.6 ±0.29	59.84 ±0.24	61.47 ±2	74.61 ±0.27
Graph augmentation	71.45 ±2.24	69.23 ±0.93	74.31 ±0.36	62.61 ±0.49	61.33 ±0.69	58.97 ±2.22	75.03 ±1.52

4.3.4 Chemical Graph Augmentation Relevance

Lastly, we trained a comparison model trained on a uniform random text sampling strategy, but with chemically-relevant molecular graph augmentations, for a full 30 epochs run.

4.4 Results

We report the experiments we ran and their performance on downstream molecular property predictions tasks in table 4.4.

5 Analysis

5.1 Molecular Property Prediction

We improved on the original paper for 6 of the 7 MoleculeNet datasets on the molecular property prediction downstream task, as seen in the experimental results displayed in Table 4.4.

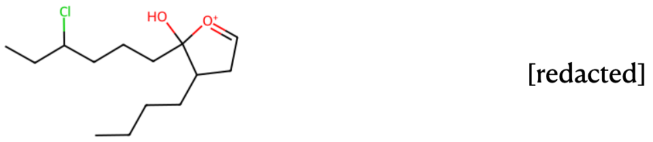
Specifically, `max` and `sent` cosine similarity in general tend to perform better than `mean` cosine similarity.

Lastly, we found that graph augmentations which happen during dataset retrieval markedly improved the results on two datasets: BBBP and SIDER. Our augmentation, which modifies the graph in a more chemically-relevant manner, outperformed all other neural text retrieval methods when run on uniforma draw.

We conclude from our experiments that improving relevance of sample retrieval improves the quality of the representations obtained through contrastive pre-training.

5.2 Molecular Generation

We experiment with molecular generation from natural language prompts using MoFlow and our text encoder trained through contrastive pre-training, and illustrate results in figure 2.



"This molecule has a hydroxyl group and a carbonyl group"

"This molecule [is very hazardous for health]"

Figure 2: Examples of molecular generation from a text prompt using MoFlow and our trained text encoder.

We observe that zero-shot molecular graph generation does generate plausible structures from natural language, and offers the exciting prospect of interacting with molecular structures through instructions in natural language.

An important observation is while molecules generated by MoFlow are designed to follow the laws of chemistry (valence, partial charges), zero-shot molecular generation can lead to unexpected results. A demonstration here for the molecule generated from the prompt *"this molecule has a hydroxyl group and a carbonyl group"*, which does include a hyroxyl group (-OH), but includes the carboxyl group inside of a furan ring, and includes a Chlorine atom.

This points to a potential future improvement, as we expect to generate more usual molecules if we trained our own flow-based generative model aligned with the embeddings generated by our text encoder.

6 Conclusion

In conclusion, we have demonstrated an improved strategy for multimodal contrastive learning of molecule representations from text corpora by incorporating neural relevance scoring at sampling time. Our approach outperforms the baseline model (MoMu) for the downstream task of molecular property prediction on most *MoleculeNet* datasets.

6.1 Future work

Evaluation of deep generative tasks in general, and molecular generation tasks in particular, is an open challenge in machine learning Yousefzadegan Hedin [2022]. An important next step would be to develop quantitative. We could both to ensure generated graphs are chemically relevant and to improve molecule fit with desired properties. An

Another potential improvement on . Training on flow-based generative model from the start was beyond the scope of the current project, but we could use our trained graph encoder as a teacher model to train our own flow

6.2 AI ethics and safety

As we continue to develop advanced foundation models for chemistry, it is essential to consider the ethical implications and safety concerns associated with the generation of molecules using AI. In this section, we discuss the potential risks associated with generative AI models for molecules, outline strategies for mitigating these risks, and propose guidelines for responsible development and deployment of these models.

6.2.1 Generation of Dangerous Molecules

The ability of AI models to generate novel molecules based on textual prompts can potentially lead to the creation of dangerous or harmful compounds. These could include toxic chemicals, environmental pollutants, or even molecules with potential applications in biological or chemical warfare. As illustrated in Figure 2, our model can potentially generate such molecules. To address this concern, we propose the following strategies:

- *Restricted access* to limit model access to authorized researchers and institutions.
- *Output filtering* to prevent the generation of dangerous molecules through safety filters.
- *Collaboration with regulatory bodies*, such as United States Environmental Protection Agency (US EPA), European Chemicals Agency (ECHA) or Food and Drug Administration (FDA), to ensure compliance with existing chemical safety regulations.
- *Education and awareness* to raise ethical considerations among researchers, institutions, and the public.

By following the above guidelines, we can ensure that the development and use of AI models in chemistry remain secure and ethical.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. 2019a. doi: 10.48550/ARXIV.1903.10676. URL https://arxiv.org/abs/1903.10676.

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

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL `https://doi.org/10.1177/001316446002000104`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

David K Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, 28: 2224–2232, 2015.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

Rafael Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2): 268–276, 2016.

Stephen R Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi, the iupac international chemical identifier. *Journal of Cheminformatics*, 7(1):1–34, 2015.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs, 2020. URL `https://arxiv.org/abs/2005.00687`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, James Hays, Pietro Perona, Jonathon Shlens, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 2021 Conference on Neural Information Processing Systems*, 2021.

Steven Kearnes, Brian Goldman, and Vijay Pande. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac956. URL `https://doi.org/10.1093/nar/gkac956`.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Sergey Korolev, Mohammadamin Tavakoli, and Ryan Lo. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1945–1954. PMLR, 2017.

Percy Liang, Christopher Potts, Frank Chen, John Hewitt, Daniel Jurafsky, and Noah A. Smith. Foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL `https://aclanthology.org/2020.acl-main.447`.

Flavio Napolitano, Antonio Candelieri, and Mattia Grandi. Molbert: Molecular representation learning with bert. *arXiv preprint arXiv:2102.01327*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL `https://arxiv.org/abs/1807.03748`.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1801.06146*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Bing Su, Dazhao Du, Zhao Yang, Youjie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Wen Ji-Rong. A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language. 2022. URL `https://arxiv.org/pdf/2209.05481.pdf`.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018a.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018b. doi: 10.1039/C7SC02664A. URL `http://dx.doi.org/10.1039/C7SC02664A`.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Sam Yousefzadegan Hedin. Evaluation of generative machine learning models: Judging the quality of generated data with the use of neural networks, 2022.

Chengxi Zang and Fei Wang. MoFlow: An invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. ACM, aug 2020. doi: 10.1145/3394486.3403104.

A Appendix

A.1 Dataset Construction

We train on the molecular graph-text pairs dataset presented in figure 3, constructed in Su et al. [2022] by retrieving scientific papers in the S2ORC [Lo et al., 2020] database by using the name and synonyms of compounds from *PubChem* [Kim et al., 2022] as query, and transforming their SMILES intro a molecular graph using OGB smile2graph Hu et al. [2020].

The dataset comprises of 15,613 graph-document pairs, with 37 million paragraphs or 47.5 gigabytes of text (∼3 megabytes per molecule). To make training tractable, the text beyond the first 500 paragraphs per molecule is left out.

We present the dataset construction process in figure 3.

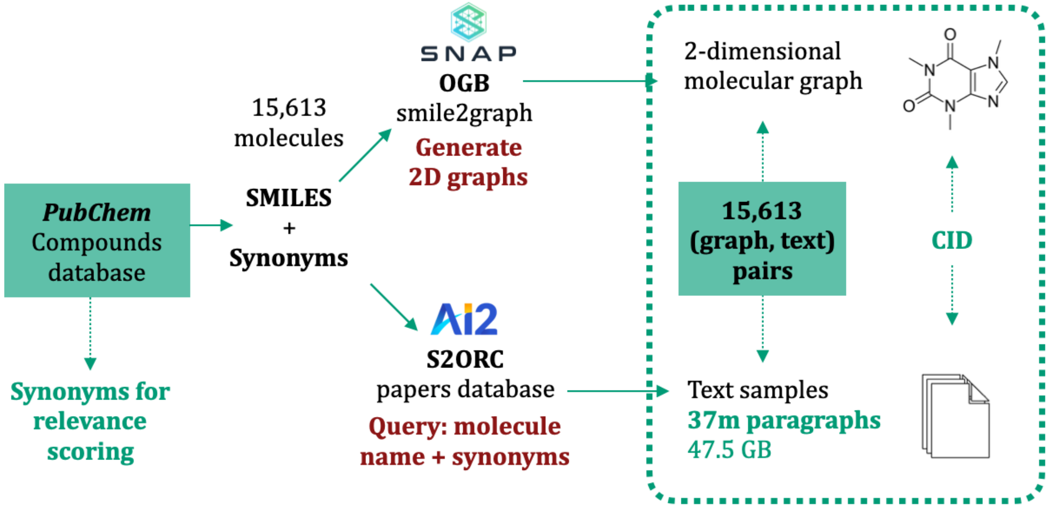


Figure 3: Joint molecular graph-text samples data set based on the PubChem and S2ORC database.

A.2 Intrinsic Evaluation for Hyper-parameter Search

To inform our search for the hyper-parameters with which to compute cosine similarity scores for sampling purposes, we ran an intrinsic evaluation of several potential retrieval methods and hyper-parameters.

We hand labeled each paragraph in a small subset of text samples, and used paragraphs which all labelers classified as relevant to the molecule as the ground truth for our retrieval problem.

We controlled for consistency between different human labelers by using Cohen's Kappa (Cohen [1960]). We report a score of 0.4874.

We varied the temperature and epsilon hyper-parameters and computed recall, precision and F1 score based on the ground truth from hand labeling. Results for the mean similarity query schema are reported in figure 1.

On the basis of these results, we chose to run our cosine similarity pre-training experiments with $\epsilon = 0.5$ and Temperature $= \{0.05, 0.1, 0.2\}$.

Temperature	**0.05**	**0.1**	**0.2**	**0.05**	**0.1**	**0.2**
Epsilon	**0.5**	**0.5**	**0.5**	**1**	**1**	**1**
Recall	0.5	0.7419	0.9354	0.3	0.4375	0.5
Precision	0.5172	0.5227	0.5178	0.5294	0.56	0.5172
F1 score	0.5085	0.6133	0.6667	0.383	0.4912	0.5084

Table 1: Intrinsic evaluation for the selection of epsilon sampling hyper-parameters.

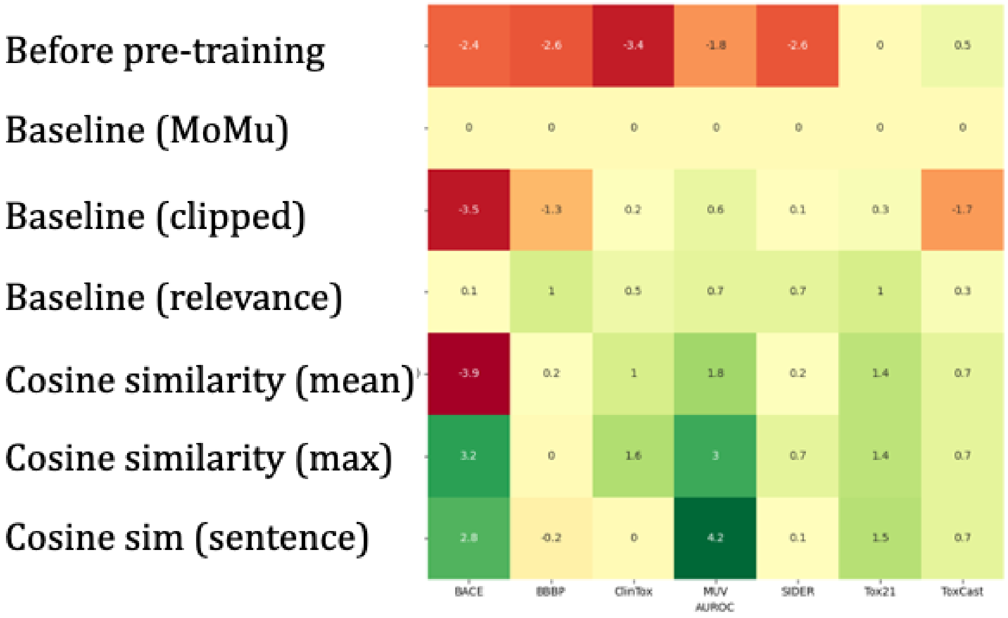


Figure 4: Visualization of the results of our experiments.