Extracting Molecular Properties from Natural Language with Multimodal Contrastive Learning

Romain Lacombe [1] **Andrew Gaut** [1] **Jeff He** [1] **David Lüdeke** [1] **Kateryna Pistunova** [1]

Abstract

Deep learning in computational biochemistry has traditionally focused on molecular graphs neural representations; however, recent advances in language models highlight how much scientific knowledge is encoded in text. To bridge these two modalities, we investigate how molecular property information can be transferred from natural language to graph representations. We study property prediction performance gains after using contrastive learning to align neural graph representations with representations of textual descriptions of their characteristics. We implement neural relevance scoring strategies to improve text retrieval, introduce a novel chemically-valid molecular graph augmentation strategy inspired by organic reactions, and demonstrate improved performance on downstream *MoleculeNet* property classification tasks. We achieve a +4.26% AUROC gain versus models pre-trained on the graph modality alone, and a +1.54% gain compared to the recently proposed molecular graph/text contrastively trained *MoMu* model (Su et al., 2022).

1. Introduction

Deep molecular representation learning models have demonstrated significant potential for important tasks in computational biology and chemistry, such as predicting molecular properties or 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 work, we investigate whether learning molecular graph representations jointly with textual representations of the corresponding molecule improves those representations. Specifically, we improve on a recently proposed molecular multimodal model (*MoMu*) for contrastive joint text-graph representation learning (Su et al., 2022) by enhancing the relevance of natural language property descriptions to which we align neural molecular representations. We implement neural relevance based methods to improve text sampling, and introduce a novel, principled approach for chemically-valid graph augmentation which yields promising results.

We hope our improved multimodal pre-training strategy for property prediction, along with experimental results and identified avenues for future work, contribute to the development of more expressive models for computational biochemistry and molecular sciences.

2. Related Work

Molecular representation learning. Molecular representation learning has played a crucial role in recent computational biology advances. Traditional molecular representations, such as SMILES (Weininger, 1988) which represent molecules as strings of atoms, have limited capacity to capture complex molecular structures. Molecular graph representations (Duvenaud et al., 2015; Kearnes et al., 2016) have have been shown to better capture the structural and functional properties of molecules. Graph Convolutional Networks (GCNs) (Kipf & Welling, 2016), Graph Attention Networks (GATs) (Veličković et al., 2017), and Graph Isomorphism Networks (GINs) (Xu et al., 2018) are popular Graph Neural Network (GNNs) architectures which outperform traditional machine learning algorithms in various molecular prediction tasks (Gómez-Bombarelli et al., 2016; Gilmer et al., 2017; Wu et al., 2018b).

Language models in chemistry. The advent of large language models, such as GPT (Radford et al., 2018), BERT (Devlin et al., 2018), and T5 (Raffel et al., 2019), has transformed the field 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, retro-synthesis, and molecular property prediction (Kusner et al., 2017)—so much so that White (2023) concludes: “the future of chemistry is language.”

[1] Stanford University, Stanford, CA, United States. Correspondence to: Romain Lacombe <rlacombe@stanford.edu>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawaii, USA, 2023. Copyright 2023 by the author(s).

Contrastive learning Contrastive learning has been used to learn good pre-trained representations of data unsupervisedly (You et al., 2020; Wang et al., 2022). Contrastive learning can also be used in a multimodal setting to learn joint representations of the image and text modalities (Radford et al., 2021). Su et al. (2022) use a similar method to jointly train graph and text encoders and call their model *MoMu*. Our work builds on *MoMu* as our baseline model.

3. Methods

3.1. Foundation Model Paradigm: Pretrain & Finetune

We approach the task of extracting molecular properties from natural language through the lens of the foundation model paradigm, following a "pre-train and fine-tune" strategy (Bommasani et al., 2022) presented in figure 1, with (1) pre-trained text and graph encoder models, (2) aligned through contrastive learning, then (3) evaluated on downstream classification tasks:

- We use two previously pre-trained encoders: a bidirectional transformer (SciBERT) for text (Beltagy et al., 2019b), and a pre-trainGraph Isomorphism Network (GIN) (Xu et al., 2018) for graphs;
- We align their representations in a joint latent space through contrastive pre-training over graph-text pairs;
- We then fine-tune the graph encoder on a series of downstream molecular property prediction tasks, and evaluate the quality of our pre-training based on performance on these downstream tasks.

3.2. Multimodal Contrastive Pre-Training

3.2.1. Contrastive Learning Strategy

The core machine learning task in our approach is to learn aligned representations of pairs of molecular graphs and paragraphs of text in natural language 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 non-matching samples (negative pairs).

Building on the original *MoMu* implementation (Su et al., 2022), we use the following contrastive learning paradigm:

- Form of batch of N molecules $i \in [1, ..., N]$;
- Sample $2N$ relevant text fragments where $\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$ describe molecule i;
- From the original $\mathcal{G}_i$ graphs, form $2N$ graphs $\{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$ through deliberate graph augmentations;
- Update text and graph encoder through gradient descent on a loss designed to promote 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.

We implement the InfoNCE loss function (Oord et al., 2018), comprising of a term for graph pairs and a term for text-graph pairs (here the 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.

For our graph encoder, we use the GraphCL 80 pre-trained model (You et al., 2020), a 1.9 million parameters Graph Isomorphism Network (GIN) pre-trained through graph contrastive learning on *MoleculeNet* (Wu et al., 2018a).

3.3. Relevance-Based Sampling

3.3.1. Neural Text Relevance Scoring

The *MoMu* baseline retrieves text sequences by uniformly sampling two paragraphs associated with a molecule per epoch. As mentioned by the authors themselves, this approach assumes equal relevance of the retrieved paragraphs to the molecule's properties (Su et al., 2022).

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 SciBERT CLS token embeddings for (i) the paragraph and (ii) a query:

- **Mean similarity**: average embedding of the molecule name and its top 20 synonyms (see section 4.1)
- **Max similarity**: maximum similarity with any of the molecule name or its top 20 synonyms
- **Sentence similarity**: cosine similarity with a 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* $\{SYNONYM_1\}, \ldots, \{SYNONYM_i\}, \ldots,$ *or* $\{SYNONYM_N\}$*."*

We then apply **epsilon sampling** (Hewitt et al., 2022) to rank paragraphs by the cosine score and sample only from scores above a threshold, using the probability distribution (re-normalized over the strictly positive terms) and a temperature hyper-parameter to skew the sampling distribution towards the highest cosine score terms:

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

3.3.2. Chemically-Valid Principled Graph Augmentations

The baseline model is trained through the general contrastive learning strategy of modifying graphs by randomly dropping nodes and subgraphs, which had been shown to be effective for chemical tasks in the past (You et al., 2020). However, the resulting molecule graphs may not make physical sense.

Here, we introduce a graph augmentations inspired by biochemical reactions, which lead to chemically valid augmented graphs. Specifically, we implement augmentations $\{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$ to the molecular graph $\mathcal{G}_i$ which add or remove functional groups corresponding to the following methylation/de-methylation (replacing a hydrogen with a CH_3 group or vice versa), amination/de-amination reactions (replacing a hydrogen with an NH_3 or vice versa):

$$\mathrm{R{-}H} + \mathrm{CH_4} \rightleftharpoons \mathrm{R{-}CH_3} + \mathrm{H_2}$$

$$\mathrm{R{-}H} + \mathrm{NH_3} \rightleftharpoons \mathrm{R{-}NH_2} + \mathrm{H_2}$$

Notably, methylation and amination involve *adding* nodes to the molecular graph, instead of node dropping and random-walk subgraphs which *remove* nodes.

Each of these augmentations is performed on randomly selected nodes of the graph at batch sampling time (see appendix A.3). Crucially, we carefully control for chemical validity of the reactions, and update the molecular graph tensor to comply with fundamental chemical rules such a bond valences and implicit hydrogens count.

4. Experiments

4.1. Molecular Property Prediction

We measure the performance impact of our novel augmentations and pretraining strategy using the downstream task of molecular property prediction (Wu et al., 2018a). From pre-training, we obtain two encoders that embed molecular graph and text descriptions within the same joint latent space: $f_G : \mathcal{G} \to \mathbf{z}_\mathcal{G} \in \mathcal{Z}$ and $f_T : \mathcal{T} \to \mathbf{z}_\mathcal{T} \in \mathcal{Z}$. We fine-tune our graph encoder for classification tasks by adding a classifier MLP layer, which we adapt and fine-tune to each specific downstream task and dataset:

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

We pre-train on the molecular graph-text pairs dataset presented in figure 2, constructed in Su et al. (2022), which comprises of 15,613 graph-document pairs, with 37 million paragraphs or 47.5 gigabytes of text (∼3 megabytes per molecule) from scientific articles, presented in appendix A.1. We then evaluate our models by fine-tuning them on biochemical classification tasks from *MoleculeNet* (Wu et al., 2018a), a multi-faceted set of benchmark tasks and reference datasets. Specifically, we use 7 datasets from DeepChem and their associated classification tasks (BACE, BBBP, Clintox, MUV, SIDER, Tox21, and ToxCast), which are all detailed in appendix A.2.

We fine-tune and evaluate the graph encoder on seven MoleculeNet datasets: BACE (Subramanian et al., 2016), BBBP (Martins et al., 2012), Tox21, ToxCast (Richard et al., 2016), SIDER (Kuhn et al., 2016), ClinTox (Gayvert et al., 2016), and MUV (Rohrer & Baumann, 2009). We use Area Under Receiver-Operator Curve (AUROC) to measure performance and evaluate for three random seeds and report the mean and standard deviation in Table 1.

4.2. Experiments

We use the following baselines for experimentation:

- ***MoMu*:** the original model presented in Su et al. (2022) which samples text and graph augmentations with a uniform random distribution;
- **Naive text relevance:** as a non-neural control for the impact of relevance selection, we create a naive dataset with only sentences where molecule names and their synonyms appear explicitly;
- **Pruning:** to control for the impact of a smaller (thus potentially less noisy) data, we prune the dataset and keep only the first 256 characters of each paragraph;
- **Single modality pre-training:** as a baseline to measure performance gains from aligning graph representations with text, we also report performance of a GIN trained only on the graph modality.

We run the following experimental protocol and report results in table 1:

Experiment	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Graph only pre-training	70	65.8	74	63.4	57.3	58	71.8
Baseline (*MoMu*)	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
Mean cosine similarity (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
Max cosine similarity (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
Sentence cosine similarity (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
Principled 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

Table 1. Results of our experiments: AUROC classifier task performance for multiple random seeds for each *MoleculeNet* dataset, reported for each pre-training experiment and baseline model/dataset.

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`).

Cosine similarity retrieval: we ran experiments on the 3 cosine similarity query types, with hyper-parameters selected via an intrinsic evaluation based on hand-labeling of a small sub-set of text paragraphs, presented in appendix (table 2). For pre-trained each model for 30 epochs (2 hours each on an A100 GPU).

Chemically-relevant graph augmentations: 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 (2 hours on an A100 GPU).

4.3. Results

We report our experiment in table 1, where we observe a consistent improvement on baseline performance for 6 of the 7 *MoleculeNet* molecular property prediction datasets and associated classification tasks.

Overall, using our strategy, the AUROC performance metric for **molecular property prediction improves by an average of +4.26% across *MoleculeNet* classification tasks** compared to molecular representations trained on the graph modality alone. We found that `max` and `sentence` cosine similarity tend to outperform random draw most consistently, followed by `mean` cosine similarity, and that our principled graph augmentations markedly improved the results on the BBBP and SIDER datasets.

The performance increase with regards to *MoMu* and the baselines controlling for text pruning (relevance-based and length-based) are +1.54%, +1.59% and +1.49% respectively. Of note: the naive relevance strategy only improves performance by +0.06% vs. baseline, and pruning the paragraphs decreases performance by -0.05%. We conclude that it is the molecular property knowledge extracted from scientific papers that improves graph representations through the multimodal contrastive training process.

5. Conclusion

We demonstrated an improved strategy for multimodal contrastive learning of molecule representations from text corpora with principled augmentation and 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 with an average performance gain of +1.54%, and outperforms models trained on graphs only by +4.26%.

Our results provide strong evidence that natural language encodes key knowledge on the properties of molecules. Extracting this information effectively through a deliberate alignment of graph representation and text embeddings is a powerful approach to improve property prediction models, and holds clear promise for computational biology and molecular sciences.

Acknowledgements

The authors wish to thank Chris Manning and John Hewitt for their instrumental feedback on neural relevance scoring, natural language sentence retrieval, and intrinsic evaluation; as well as Jure Leskovec and Mohammad Aljubran, on molecular 2D graphs representation and augmentation. We also wish to acknowledge Su et al. (2022) for the original molecular multimodal contrastive learning implemen-

tation, You et al. (2020) for the pre-trained GraphCL GNN models, Beltagy et al. (2019a) for the pre-trained SciBERT model, the Allen Institute for the S2ORC dataset (Lo et al., 2020), and the National Institutes of Health for the *PubChem* database (Kim et al., 2022).

References

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

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

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Cohen, J. 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Duvenaud, D. K., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, 28:2224–2232, 2015.

Gayvert, K. M., Madhukar, N. S., and Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301, 2016.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR. org, 2017.

Gómez-Bombarelli, R., Duvenaud, D., Hernández-Lobato, J. M., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2016.

Hewitt, J., Manning, C. D., and Liang, P. Truncation sampling as language model desmoothing, 2022.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs, 2020. URL https://arxiv.org/abs/2005.00687.

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

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. 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.

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

Korolev, S., Tavakoli, M., and Lo, R. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1945–1954. PMLR, 2017.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 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.

Martins, I. F., Teixeira, A. L., Pinheiro, L., and Falcao, A. O. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

Napolitano, F., Candelieri, A., and Grandi, M. Molbert: Molecular representation learning with bert. *arXiv preprint arXiv:2102.01327*, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1801.06146*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., et al. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8):1225–1251, 2016.

Rohrer, S. G. and Baumann, K. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of chemical information and modeling*, 49(2):169–184, 2009.

Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Ji-Rong, W. A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language. 2022. URL https://arxiv.org/pdf/2209.05481.pdf.

Subramanian, G., Ramsundar, B., Pande, V., and Denny, R. A. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949, 2016.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Weininger, D. 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.

White, A. D. The future of chemistry is language. *Nature Reviews Chemistry*, 2023. doi: 10.1038/s41570-023-00502-0. URL https://doi.org/10.1038/s41570-023-00502-0. ISSN: 2397-3358.

Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018a. doi: 10.1039/C7SC02664A. URL http://dx.doi.org/10.1039/C7SC02664A.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018b.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

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

Zang, C. and Wang, F. 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

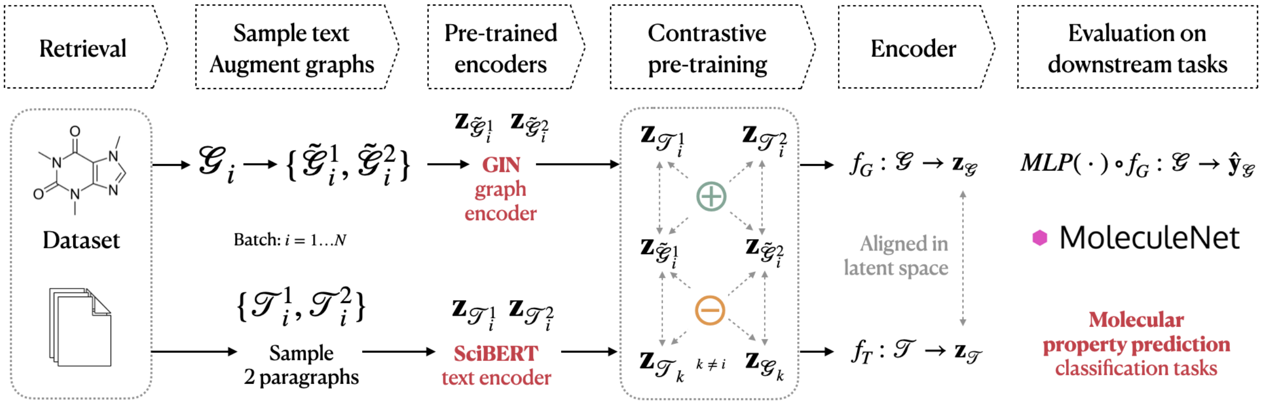
Retrieval
Sample text Augment graphs
Pre-trained encoders
Contrastive pre-training
Encoder
Evaluation on downstream tasks
Dataset
$\mathcal{G}_i \rightarrow \{\tilde{\mathcal{G}}_i^1, \tilde{\mathcal{G}}_i^2\}$
Batch: $i = 1 \ldots N$
$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$
GIN graph encoder
$\{\mathcal{T}_i^1, \mathcal{T}_i^2\}$
Sample 2 paragraphs
$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$
SciBERT text encoder
$\mathbf{z}_{\mathcal{T}_i^1}$ $\mathbf{z}_{\mathcal{T}_i^2}$
⊕
$\mathbf{z}_{\tilde{\mathcal{G}}_i^1}$ $\mathbf{z}_{\tilde{\mathcal{G}}_i^2}$
⊖
$\mathbf{z}_{\mathcal{T}_k}$ $k \neq i$ $\mathbf{z}_{\mathcal{G}_k}$
$f_G : \mathcal{G} \rightarrow \mathbf{z}_{\mathcal{G}}$
Aligned in latent space
$f_T : \mathcal{T} \rightarrow \mathbf{z}_{\mathcal{T}}$
$MLP(\cdot) \circ f_G : \mathcal{G} \rightarrow \hat{\mathbf{y}}_{\mathcal{G}}$
MoleculeNet
Molecular property prediction classification tasks

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

A.1. Pre-Training Dataset

We train on the molecular graph-text pairs dataset presented in figure 2, constructed following (Su et al., 2022) by retrieving scientific papers in the S2ORC (**?**) 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.

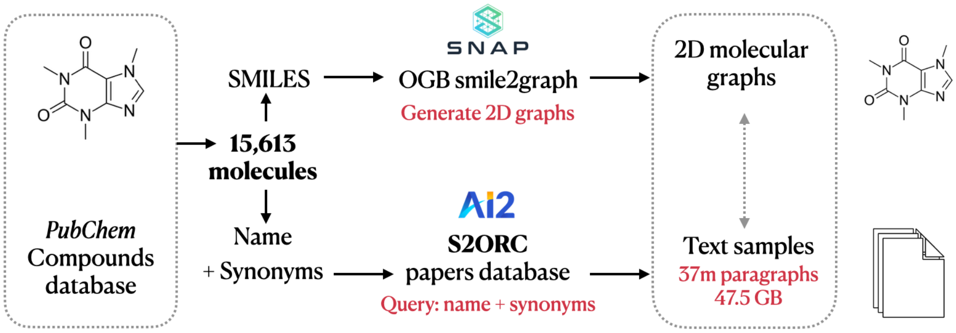

Figure 2. Joint molecular graph-text samples data set based on the *PubChem* and S2ORC database.

A.2. Downstream Molecular Property Prediction

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.

A.3. Chemically-Valid Principled Graph Augmentations

We implement algorithm 1:

Algorithm 1 Chemically-Valid Principled Graph Augmentations.
Example: methylation reaction, addition of a $-CH_3$ functional group to the molecular group.

```
Require: PyG graph tensor $x_i$, node features, edge features
  1. Randomly sample nodes that are C atoms with implicit hydrogen count $\geq 1$
  2. Add a new node to the graph for the additional functional group and update node features for valid covalence and implicit hydrogen numbers
  3. Add an edge to the molecule graph with a single bond feature to bind the additional functional group
  4. Decrease implicit hydrogen count for the original site to account for functional group addition
```

A.4. 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 2.

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**
$\mathcal{E}$-threshold	**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 2. Intrinsic evaluation for the selection of epsilon sampling hyper-parameters.

A.5. 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). As a next step, we could use the graph encoder and text encoder we trained to train generative models that help bridge these two modalities, with multi-model tasks such as:

- Molecular captioning: given a molecular graph, generating text that accurately describes the molecule and its properties;
- Molecular generation: given a text description of desiged properties in natural language, generate a graph for a molecule that exhibits such properties.

For that purpose, and following Su et al. (2022), we implemented MoFlow (Zang & Wang, 2020), a flow-based deep generative model, on our pre-trained graph encoders, to experiment with molecular generation from free text. This showed very promising results for zero-shot molecular generation (zero-shot since we did not fine-tune the flow model to match out encoder specifically). A logical avenue for future work could be to use our trained graph encoder as a teacher model to train our own flow or diffusion-based model and measure improvements in molecular generation capacity.