Enhancing Molecular Expressiveness through Multi-View Representations

Indra Priyadarsini
IBM Research - Tokyo
indra.ipd@ibm.com

Seiji Takeda
IBM Research - Tokyo
seijitkd@jp.ibm.com

Lisa Hamada
IBM Research - Tokyo
lisa.hamada@ibm.com

Hajime Shinohara
IBM Research - Tokyo
hajime.shinohara1@ibm.com

Abstract

In the field of foundation models for materials science and chemistry, the quality of molecular representations plays a critical role in the success of machine learning models in downstream tasks. Transformer-based molecular representation models have shown great potential in these areas by generating high-quality latent representations of molecules. However these representations often fail to capture the full complexity of molecular structures, leading to suboptimal performance in predictive tasks. In this work we propose a simple yet novel multi-view representation method to improve the expressiveness of the molecular latent representations. We provide preliminary analysis of the proposed method which shows promising improvements compared to the conventional method, suggesting that the multi-view approach improves the quality of the latent representations.

1 Introduction

Large-scale molecular representation methods are shown to be useful in various material science applications, such as virtual screening, drug discovery, chemical modeling, material design, and molecular dynamics simulations. With the progress in deep learning, numerous models have been developed to derive representations directly from molecular structures. Recently, transformer-based molecular representations have gained prominence in material informatics, offering significant potential for advancements in drug discovery, materials science, and related fields. Recent works (1; 2; 3; 4; 5) have demonstrated the capability of transformer models in capturing complex relationships and patterns within molecular data with the help of attention mechanisms. Most of these works are based on SMILES (Simplified Molecular Input Line Entry System) (6). However, one of the drawbacks of SMILES is that it does not guarantee syntactic and semantic validity of the molecule (7), thus leading to a possibility of learning invalid representations. SELFIES (SELF-referencing Embedded Strings) is another molecular string representation that was introduced by (7) to overcome the drawbacks of SMILES. Furthermore, in addition to achieving high accuracy predictions of molecular properties, a key objective within computational material informatics is to devise novel and functional molecules. But most existing transformer models for material informatics are encoder-only models, which are not capable of generating new molecules.

In this paper, we introduce SELF-BART, a transformer-based model capable of capturing intricate molecular relationships and interactions. Unlike most existing works that utilize encoder-only models, we propose an encoder-decoder model based on BART (Bidirectional and Auto-Regressive Transformers) (8). This model not only efficiently learns molecular representations but is also capable of auto-regressively generating new molecules from these representations. This capability is

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

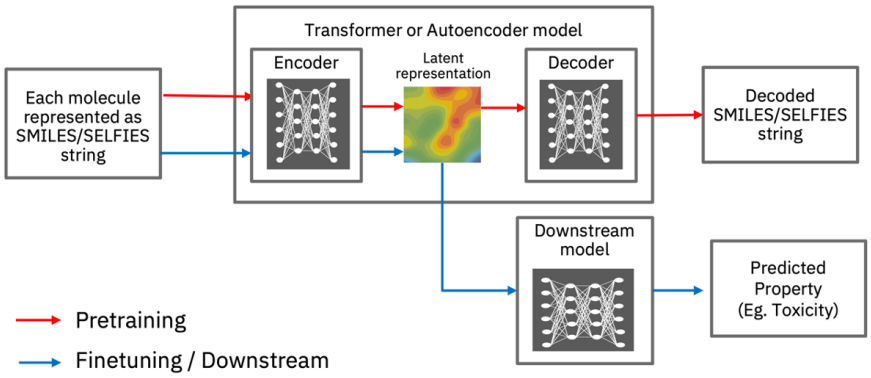

Figure 1: General architecture and flow of training

particularly impactful for novel molecule design and generation, facilitating efficient and effective analysis and manipulation of molecular data.

2 Background and Motivation

While foundation models have shown great promise in materials science and chemistry, they face a significant challenge: the limited availability of large, diverse datasets, especially in the downstream tasks. In contrast to the vast text corpora used to train large language models (LLMs), datasets in materials science and chemistry often contain only a few hundred samples. This data scarcity hampers the ability to train models that generalize well to unseen molecular structures, especially for tasks requiring high-quality latent representations.

One common approach to addressing this issue is SMILES enumeration (9), a data augmentation technique that generates multiple valid SMILES representations for the same molecule. The same enumeration can be extended to SELFIES strings too. Figure 2 illustrates an example of a molecule represented by several different SMILES/SELFIES strings. While this method increases the dataset sample size, it does not necessarily enhance the quality or expressiveness of the latent space learned by the model. Simply adding more samples might improve training performance, but it does not guarantee that the learned representations effectively captures the molecular properties.

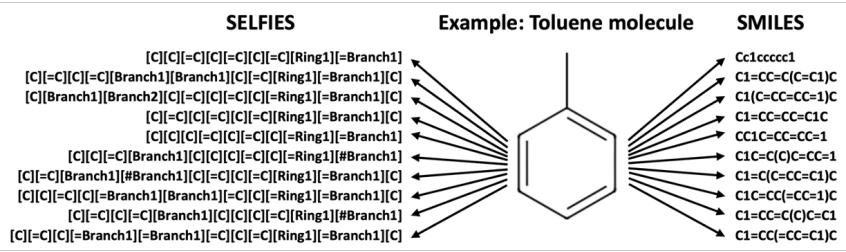

Figure 2: Example of SMILES/SELFIES enumeration where a single molecule can be represented in multiple forms

This study is driven by the need to understand how the latent representations of enumerated strings relate, given that they represent the same underlying molecule. To explore this, we conducted a preliminary analysis by selecting 10 random molecules from the MoleculeNet (10) BACE dataset and generated 100 alternative SMILES strings for each. We then extracted the latent representations using a transformer encoder-decoder model and visualized them using t-SNE visualization (11). As shown in Figure 3, clear clusters emerged, where each cluster corresponds to a molecule and its alternative representations. The latent representations of the enumerated SMILES/SELFIES form a cloud, indicating that these alternate forms cluster together and can be treated as different views of the same molecule, each conveying a different aspect of the same underlying molecule.

Building on these observations, we propose a novel framework called Multi-View Representation (MVR) to enhance the expressiveness of molecular representations. The proposed method is detailed in the following section.

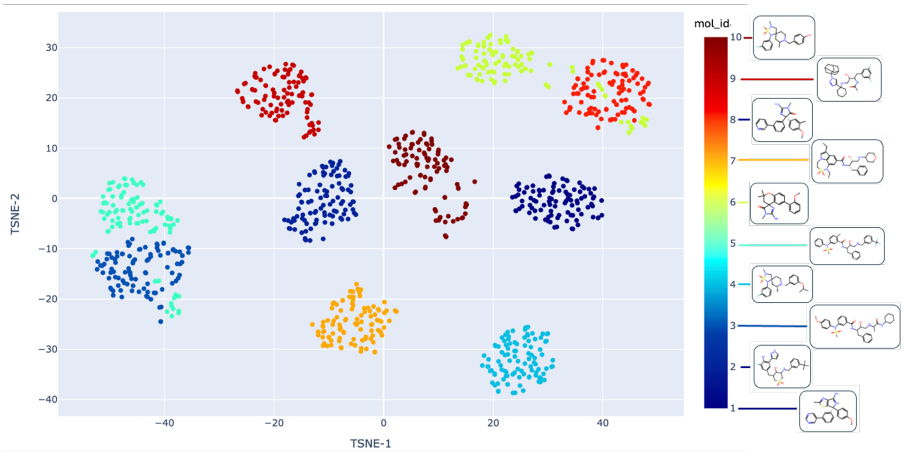

Figure 3: T-SNE plot of the latent representation 10 different molecules and their enumerated forms

3 Proposed Method

In this paper, we introduce a Multi-View Representation (MVR) framework aimed at enhancing the expressiveness of latent representations in molecular modeling. The core idea is to generate multiple latent representations for the same molecule, each capturing distinct features or "views" of the molecule. By systematically selecting and combining these features, we create a more comprehensive and enriched latent vector representation. This approach is expected to improve the quality of the latent space representation, consequently improving the performance in downstream tasks such as molecular property prediction. The schematic of the proposed MVR framework is illustrated in Figure 4. The proposed framework operates through three main steps:

- Generating multiple string representations: Obtain k different SMILES or SELFIES strings for the same molecule, including canonical and non-canonical variants. These alternate string representations provide different "views" of the molecule's structure.
- Extracting Latent Representations: For each generated string, we use a pretrained model (Eg. transformer-based encoder) to obtain its latent representation. Each latent representation is hypothesized to capture different aspects or "views" of the molecule's structure and properties.
- Selecting and Combining Latent Representations: To create an enriched representation, a greedy selection process is used to identify the most informative latent vectors. These selected vectors are concatenated to form a unified, comprehensive latent representation that leverages the diversity of the alternate views.

The final enriched feature vector is fed into a downstream model to make predictions. By leveraging multiple views of the molecule, this approach is expected to enhance molecular modeling by capturing a broader spectrum of molecular features from different latent views, ultimately improving performance in various cheminformatics tasks.

4 Results and Discussions

To evaluate the effectiveness of the proposed multi-view representation (MVR), we conducted experiments on four classification tasks from the MoleculeNet dataset. The proposed MVR method can be applied to SMILES and SELFIES based string representations. Thus to demonstrate this, we use Molformer (3), a molecular representation model trained on SMILES and SEFIES-BART (12), a

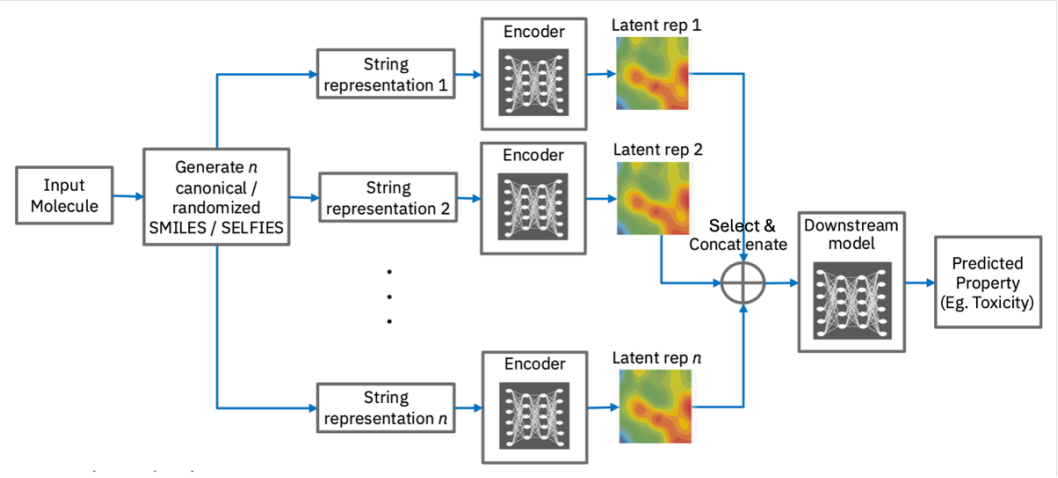

Figure 4: Proposed Multi-View Representation Framework

molecular representation model trained on SELFIES, for extracting the latent representations. For each molecule in the datasets used for evaluation, we generated 4 alternate SMILES/SELFIES string, one of which is the canonical set. Thus $k = 5$ including the original dataset. The latent representation of the molecules for each set is extracted using the encoders of Molformer and SELFIES-BART, respectively. These representations are used as input features in the downstream XGBoost model (13). The metric used for the evaluation is ROC-AUC score. The extracted latent representations are concatenated in combinations of k=2,3,4,5 and a greedy selection method is applied to select the best combinations to form the new enriched latent representation as detailed in Section 3. The corresponding results are reported in Table 1. The results of the original and alternate sets As seen from the results, the proposed MVR method shows significant increase in ROC-AUC scores. While further tuning of the downstream model could enhance performance, this was not pursued, as the focus of this study was on demonstrating the gains from the proposed method. Also, the number of alternate representations (k) is a hyperparameter, and future work may explore more efficient methods for optimizing it.

	MolFormer - SMILES				**BART - SELFIES**			
	BACE	**BBBP**	**ClinTox**	**HIV**	**BACE**	**BBBP**	**ClinTox**	**HIV**
Original	82.78	93.95	97.57	72.88	83.84	92.17	89.70	73.43
Canonical	84.43	91.01	86.39	72.35	86.52	87.06	84.44	71.50
Non-canonical (set 1)	78.87	89.81	75.62	72.69	70.78	83.38	76.80	71.09
Non-canonical (set 2)	70.27	88.34	69.23	72.21	77.48	84.57	81.12	69.72
Non-canonical (set 3)	73.68	83.29	81.07	68.99	76.96	84.71	79.23	70.26
MVR (k=2)	84.12	95.09	97.87	75.42	87.77	93.77	90.95	74.21
MVR (k=3)	**84.74**	95.19	98.28	**76.62**	84.39	93.83	**97.51**	**77.54**
MVR (k=4)	84.17	95.25	**98.52**	74.28	**87.81**	**94.22**	85.27	74.03
MVR (k=5)	80.59	**95.78**	97.40	70.01	86.13	93.31	86.33	70.33

Table 1: Performance comparison of the proposed method on MoleculeNet tasks.

5 Conclusion

In this paper, we presented a novel multi-view representation approach designed to enhance the expressiveness of molecular representations in transformer-based models. The core idea is to obtain different latent representations for the same molecule and systematically select features to form a more enriched latent vector. This can be regarded as capturing different view of the same molecule. Through a series of experiments on benchmark datasets, we demonstrated that incorporating multi-view representations can lead to better performance in various downstream property prediction tasks. This work opens several avenues for further exploration and investigate the potential of multi-view representations

References

[1] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: large-scale self-supervised pre-training for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[2] V. Bagal, R. Aggarwal, P. Vinod, and U. D. Priyakumar, "Molgpt: molecular generation using a transformer-decoder model," *Journal of Chemical Information and Modeling*, vol. 62, no. 9, pp. 2064–2076, 2021.

[3] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[4] G. Chilingaryan, H. Tamoyan, A. Tevosyan, N. Babayan, L. Khondkaryan, K. Hambardzumyan, Z. Navoyan, H. Khachatrian, and A. Aghajanyan, "Bartsmiles: Generative masked language models for molecular representations," *arXiv preprint arXiv:2211.16349*, 2022.

[5] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan, "Selformer: molecular representation learning via selfies language models," *Machine Learning: Science and Technology*, vol. 4, no. 2, p. 025035, 2023.

[6] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[7] 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*, vol. 1, no. 4, p. 045024, 2020.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[9] E. J. Bjerrum, "Smiles enumeration as data augmentation for neural network modeling of molecules," *arXiv preprint arXiv:1703.07076*, 2017.

[10] 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*, vol. 9, no. 2, pp. 513–530, 2018.

[11] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[12] I. Priyadarsini, V. Sharma, S. Takeda, A. Kishimoto, L. Hamada, and H. Shinohara, "Improving performance prediction of electrolyte formulations with transformer-based molecular representation model," *arXiv preprint arXiv:2406.19792*, 2024.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.