SELF-BART : A Transformer-based Molecular Representation Model using SELFIES

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

Emilio Vital Brazil
IBM Research - Brazil
evital@br.ibm.com

Eduardo Soares
IBM Research - Brazil
eduardo.soares@ibm.com

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

Abstract

Large-scale molecular representation methods have revolutionized applications in material science, such as drug discovery, chemical modeling, and material design. With the rise of transformers, models now learn representations directly from molecular structures. In this study, we develop an encoder-decoder model based on BART that is capable of leaning molecular representations and generate new molecules. Trained on SELFIES, a robust molecular string representation, our model outperforms existing baselines in downstream tasks, demonstrating its potential in efficient and effective molecular data analysis and manipulation.

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

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

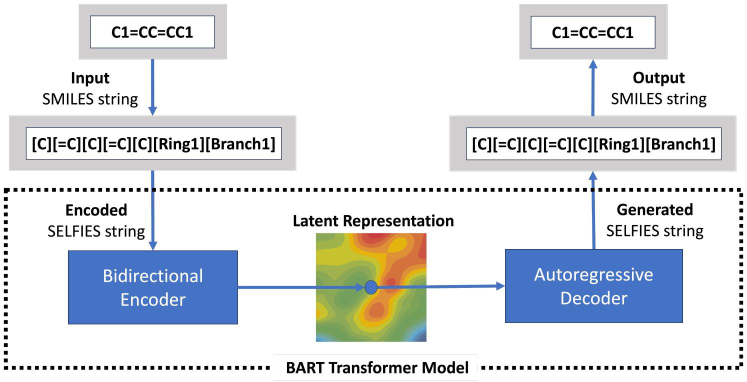

Figure 1: Model architecture

2 Model

The proposed SELF-BART model is an encoder-decoder architecture derived from the BART (Bidirectional Auto-Regressive Transformer) model (8). The encoder processes the sequence of input token bidirectionally and the decoder generates the sequence autoregressively. The SELF-BART model is trained using SELFIES as it provides a more concise and interpretable representation, making it suitable for machine learning applications where compactness and generalization are important (7). During pre-training the model is trained with a denoising objective function. The model is trained using the ZINC-22 (9) and PubChem (10) datasets. The dataset consists of molecules represented in SMILES notation. We convert these SMILES strings to SELFIES using the SELFIES API (7). In SELFIES each atom or bond is represented by symbols enclosed in **[ ]**, which are then tokenized using a word level tokenization scheme where each symbol or bond in **[ ]** is treated as a word. Further 15% of the tokens are randomly masked and the model is trained using a denoising objective where the model learns to predict the next token in the original sequence, conditioned on both the corrupted sequence and the already decoded part of the original sequence. The objective function is given as,

$$\mathcal{L}_{\text{denoise}} = -\sum_{t=1}^{T} \log P(Y_t | Y_{<t}, X_{\text{corrupt}}; \theta)$$

where, Y_t is the t-th token in the original sequence Y, $Y_{<t}$ represents the tokens preceding t in the target sequence, X_{corrupt} is the corrupted input sequence, θ are the model parameters, and $P(Y_t|Y_{<t}, X_{\text{corrupt}}; \theta)$ is the probability predicted by the model for token Y_t, conditioned on the corrupted input and the previously generated tokens. Figure 1 illustrates the pre-training model architecture. We hypothesize that the encoder-decoder structure of the SELF-BART model, combined with the denoising objective, provides better molecular representations. Moreover, training on SELFIES instead of SMILES ensures that the encoder output represents only valid molecules, enhancing the robustness of the molecular representations which are used for downstream tasks such as property prediction.

Dataset	**Description**	**#Samples**	**Metric**
BACE	Binary labels on β-secretase 1 (BACE1) binding properties	1,513	ROC-AUC
ClinTox	Binary labels on clinical toxicity data on FDA-approved drugs	1478	ROC-AUC
BBBP	Binary labels on blood–brain barrier permeability	2,039	ROC-AUC
HIV	Binary labels on the ability to inhibit HIV replication	41,127	ROC-AUC
SIDER	Drug side effect classification for 27 types of adverse effects	1,427	ROC-AUC
Tox21	Qualitative toxicity measurements on 12 targets	7,831	ROC-AUC
Esol	Water solubility prediction of small molecules	1,128	RMSE
Lipophilicity	Prediction of octanol-water partition coefficient (logD)	4,200	RMSE
Freesolv	Hydration free energy of small molecules in water	642	RMSE

Table 1: Description of the benchmark datasets used in the evaluation of the proposed model.

Model	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
RF (3)	71.4	71.3	78.1	86.7	68.4	76.9
SVM (3)	72.9	66.9	79.2	86.2	68.2	81.8
MGCN (12)	85.0	63.4	73.8	73.4	55.2	70.7
D-MPNN (13)	71.2	90.5	75.0	85.3	63.2	68.9
DimeNet (14)	-	76.0	-	-	61.5	78.0
Hu, et al. (15)	70.8	78.9	80.2	85.9	65.2	78.7
N-Gram (16)	91.2	85.5	83.0	87.6	63.2	76.9
MolCLR (17)	73.6	93.2	80.6	89.0	68.0	79.8
GraphMVP (18)	72.4	77.5	77.0	81.2	63.9	74.4
GeomGCL (18)	-	91.9	-	-	64.8	85.0
GEM (19)	72.4	90.1	80.6	85.6	67.2	78.1
ChemBerta (1)	64.3	73.3	62.2	79.9	-	-
ChemBerta2 (20)	71.94	90.7	-	85.1	-	-
Galatica 30B (21)	59.6	82.2	75.9	72.7	61.3	68.5
Galatica 120B (21)	66.1	82.6	74.5	61.7	63.2	68.9
Uni-Mol (22)	72.9	91.9	80.8	85.7	65.9	79.6
SELFormer (5)	90.2	-	68.1	83.2	**74.5**	65.3
MoLFormer-XL (3)	93.7	94.8	82.2	88.2	69.0	**84.7**
SELF-BART	**95.2**	**96.9**	**83.0**	**89.3**	65.0	76.5

Table 2: Results of the evaluation on classification tasks of MoleculeNet benchmark datasets

3 Results and Discussions

To evaluate the effectiveness of our proposed model on both molecular property prediction tasks and molecule generation tasks. For the molecule property predition tasks, we conducted evaluations using a comprehensive set of 9 distinct benchmark datasets sourced from MoleculeNet (11). The details of the benchmarks used are illustrated in Table 1. We evaluate 6 datasets for the classification task and 3 datasets for regression tasks. To ensure a robust and unbiased assessment, we maintained consistency with the MoleculeNet benchmark by adopting identical train/validation/test splits for all tasks (11). We compare the performance of the proposed SELF-BART model with various graph-based and text-based models. The SELF-BART model used in the evaluations is a 354M parameter model trained on 1B samples drawn from a combination of ZINC and PubChem datasets with a vocabulary of 3160 tokens. Futhermore, for the molecule generation tasks we conduct a preliminary analysis of the SELF-BART model and compare its results with existing molecular generative models.

3.1 Molecular Property Prediction Tasks

We evaluated the SELF-BART models on nine benchmark from MoleculeNet (11). These tasks include four binary classification tasks using BACE, ClinTox, BBBP and HIV datasets, two multi-label classification task using SIDER and Tox21 datasets, and three regression tasks using the esol, freesolv and lipophilicity datasets. For the evaluation, we used molecular embeddings generated by the SELF-BART models as input features. We use XGBoost (23) as the downstream task model and Optuna (24) for hyperparameter tuning. The results corresponding to the optimal hyperparameters are reported. The performance is measured using the ROC-AUC and RMSE metrics. Table 2 presents the performance of the SELF-BART models compared to other molecular graph-based, geomentry-based models and molecular string-based models. ChemBERTa, Galatica, Uni-Mol and MolFormer are trained on SMILES representations, while SELFormer and the proposed SELF-BART model are trained on SELFIES representations. As shown in Table 2, the SELF-BART model outperforms the other models in four out of six tasks. We also evaluate the performance of the models on 3 regression task, the results of which are presented in Table 3. The SELF-BART model outperforms the other models in two out of three tasks. The improved performance of SELF-BART can be attributed to encoder-decoder architecture of model being trained on SELFIES, which ensures that the learned representations correspond to valid molecules. This approach substantially improves the robustness and quality of the molecular representations. Although both SELFormer and the proposed SELF-BART model are trained on SELFIES, SELF-BART demonstrates superior performance. This enhancement is primarily due to SELF-BART's encoder-decoder architecture combined with a denoising objective, in contrast to SELFormer's encoder-only architecture. This design choice significantly improves the robustness and quality of the molecular representations.

Model	ESOL	FreeSolv	Lipophilicity
D-MPNN(13)	1.050	2.082	0.683
Hu et al.(15)	1.220	2.830	0.740
MGCN(12)	1.270	3.350	1.110
GEM(19)	0.798	1.877	0.660
SchNet(25)	1.050	3.220	0.910
KPGT(26)	0.803	2.121	**0.600**
GraphMVP-C(18)	1.029	-	0.681
GCN(27)	1.430	2.870	0.850
GIN(28)	1.450	2.760	0.850
MolCLR(17)	1.110	2.200	0.650
ChemBERTa-2(20)	-	-	0.986
MolFormer(3)	0.755	2.022	0.840
SELFformer(5)	0.682	2.797	0.735
SELF-BART	**0.454**	**1.397**	0.771

Table 3: Results of the evaluation on regression tasks of MoleculeNet benchmark datasets

3.2 Molecule Generation Task

The SELF-BART model is an encoder-decoder architecture, making it not only capable of providing robust molecular representations but also adept at generating molecules. In this section, we analyze the SELF-BART model's performance in non-conditioned molecular generation. Given the infinitely large and unexplored chemical space, it is crucial for a molecular generative model to understand molecular grammar and rules, ensuring the generation of novel and valid molecules. As a preliminary analysis, we evaluate the SELF-BART model's ability to generate molecules. For this purpose, we use the decoder, initializing it with the begin of sentence `<bos>` token to generate 10,000 molecules. This evaluation helps us understand the model's proficiency in producing diverse and valid molecular structures. The metrics we use in this analysis are validity, uniqueness, novelty and internal diversity. The metric scores are presented in Table 4. The metrics for CharRNN, VAE, AAE, LatentGAN, JT-VAE and MolGPT are values reported from (2) trained on MOSES dataset, while SELF-BART was trained on 1B samples from ZINC-22 and PubChem. From the results, we can observe that the SELF-BART model is equally performant in generating unique, valid, and novel molecules with the high internal diversity, thus confirming its effectiveness in generating molecules of varying structures and quality compared to similar baseline methods.

Models	Validity	unique@10K	Novelty	IntDiv$_1$	IntDiv$_2$
CharRNN	0.975	0.999	0.842	0.856	0.85
VAE	0.977	0.998	0.695	0.856	0.85
AAE	0.937	0.997	0.793	0.856	0.85
LatentGAN	0.897	0.997	0.949	0.857	0.85
JT-VAE	1.0	0.999	0.914	0.855	0.849
MolGPT	0.994	1.0	0.797	0.857	0.851
SELF-BART	0.998	0.999	1.0	0.918	0.908

Table 4: Comparison of different models based on various metrics used in evaluating molecular generative models.

4 Conclusion

This paper presents SELF-BART, an encoder-decoder transformer model designed to effectively learn representations of the chemical space. By training on SELFIES strings, SELF-BART ensures the validity of the molecules during pre-training, which enhances the robustness of its molecular representations. The model's effectiveness is demonstrated through performance evaluations on benchmark classification and regression tasks from MoleculeNet. The SELF-BART model achieved state-of-the-art results in most tasks. Although the primary focus is on molecular representation for downstream tasks, we provided an initial exploration of the model's ability to generate molecules without conditioning. The preliminary analysis showed that the model was capable of generating valid and novel molecules with good structural diversity. Future work will investigate the model's generative capabilities further, including conditioned molecular generation, and examine its performance with scaling and conditioned generative modeling.

References

[1] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: large-scale self-supervised pretraining 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] B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz, and J. J. Irwin, "Zinc-22 a free multi-billion-scale database of tangible compounds for ligand discovery," *Journal of chemical information and modeling*, vol. 63, no. 4, pp. 1166–1176, 2023.

[10] S. Kim, J. Chen, A. Gindulyte, J. He, S. He, B. A. Shoemaker, P. A. Thiessen, E. E. Bolton, G. Fu, L. Han, *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2016.

[11] 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.

[12] C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, and L. He, "Molecular property prediction: A multilevel quantum interactions modeling perspective," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 1052–1060, 2019.

[13] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, *et al.*, "Analyzing learned molecular representations for property prediction," *Journal of chemical information and modeling*, vol. 59, no. 8, pp. 3370–3388, 2019.

[14] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," *arXiv preprint arXiv:2003.03123*, 2020.

[15] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[16] S. Liu, M. F. Demirel, and Y. Liang, "N-gram graph: Simple unsupervised representation for graphs, with applications to molecules," *Advances in neural information processing systems*, vol. 32, 2019.

[17] Y. Wang, J. Wang, Z. Cao, and A. Barati Farimani, "Molecular contrastive learning of representations via graph neural networks," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 279–287, 2022.

[18] S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, and J. Tang, "Pre-training molecular graph representation with 3d geometry," *arXiv preprint arXiv:2110.07728*, 2021.

[19] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, and H. Wang, "Geometry-enhanced molecular representation learning for property prediction," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, 2022.

[20] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta-2: Towards chemical foundation models," *arXiv preprint arXiv:2209.01712*, 2022.

[21] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

[22] 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," 2023.

[23] 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.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[25] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions," *Advances in neural information processing systems*, vol. 30, 2017.

[26] H. Li, D. Zhao, and J. Zeng, "Kpgt: knowledge-guided pre-training of graph transformer for molecular property prediction," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 857–867, 2022.

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

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