Leveraging Implicit Representations of Atomistic Foundation Model for Generation and Optimization of Molecules

Yubo Hou[*a], **Keyu Wu**[*a], **Ji Wei Yoon**[a]

[a] *Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #21-01, Connexis South Tower, Singapore 138632, Republic of Singapore* yoon_ji_wei@i2r.a-star.edu.sg

1. Introduction

Generative models are advancing molecular discovery by enabling the design of novel molecules with desirable properties. This study proposes a Conditional Variational Autoencoder (CVAE) enhanced with foundation model features, which provide a comprehensive representation of molecular properties to better learn intricate patterns and dependencies in molecular data. This leads to improved reconstruction accuracy of generated molecules. Additionally, the framework employs SELFIES encoding to ensure chemically valid outputs. Furthermore, we leverage foundation models to enhance the optimization performance of traditional chemical encodings, showcasing their versatility and broad applicability in molecular design tasks.

2. Generation of molecules with specific properties

SMILES-based CVAEs have long supported molecular generation tasks by learning structure-property relationships [1]. However, SMILES often requires post-processing to ensure chemical validity. SELFIES improve this by guaranteeing chemically valid outputs, though their primary role is unable to address deeper learning challenges. Meanwhile, the integration of foundation model features represents a major step forward. Models like MACE-OFF [2] leverage extensive datasets to provide comprehensive molecular representations. Despite their promise, there is a lack of efforts to incorporate such features into generative frameworks [3]. In this work, foundation features are embedded into a SELFIES-based CVAE to significantly enhance molecular accuracy.

2.1 Methodology

The overview of the proposed CVAE framework is demonstrated in Fig. 1. Firstly, molecular structures, initially represented as SMILES, are converted into SELFIES and subsequently transformed into one-hot encoded vectors, forming the structural representation of molecules. Simultaneously, molecular condition and features derived from the MACE foundation model are extracted. These foundation features, enriched with information about atomic interactions and molecular environments, are concatenated with the one-hot encoded SELFIES and condition vectors to create a comprehensive input representation. The encoder processes this input and maps it into a latent space. The decoder then reconstructs from the latent representation to produce chemically valid molecular structures conditioned on input features. The training objective minimizes a composite loss function that combines reconstruction loss to measure fidelity to the input and KL divergence to regularize the latent space.

2.2 Experiments

The method was evaluated on two datasets. QM9 contains small organic molecules, while ESM includes fewer but more complex molecules. For the QM9 dataset, the conditions used for generation include mu, alpha, and cv, while for ESM, the generation conditions include temperature, viscosity, and density. For both datasets, 80% was used for training and 20% for validation. Models with and without foundation features were compared based on reconstruction accuracy, validity, and uniqueness. Results demonstrate that incorporating foundation features significantly improves all metrics, with greater gains observed for the more complex ESM dataset.

Table 1: Experimental results on two datasets across 5 random seeds.

	QM9 (w/o FF)	QM9 (w/ FF)	ESM (w/o FF)	ESM (w/ FF)
Accuracy (%)	98.5	99.0	70.2	78.2
Validity (%)	88.4	94.0	94.4	93.6
Uniqueness (%)	76.0	82.4	62.2	66.4

3. BO for optimization

3.1 Methodology

We leverage foundation model (FM) to enhance the optimization performance of existing traditional chemical encodings. Specifically, we concatenate the features generated by the foundation model and the traditional chemical encodings for the same molecule. To reduce dimensionality and remove potential redundancy, we apply Principal Component Analysis (PCA). The reduced features are then fed into the Bayesian Optimization (BO) model to opti-

[1*] First author.

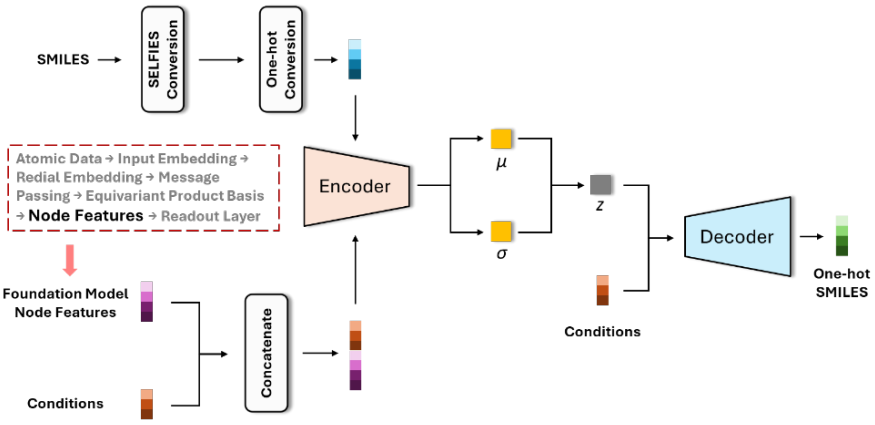

Fig. 1: Overview of the proposed CVAE framework.

mize the target objective.

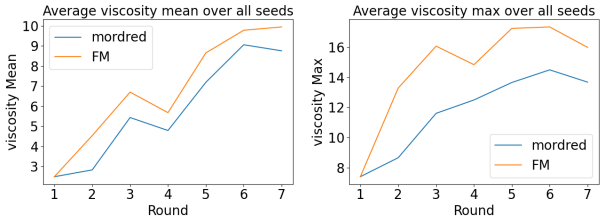

Fig. 2: Comparison of the proposed method and the benchmark approach based on the average viscosity results across 100 random seeds. Left: the average result of the 5 recommended conditions. Right: the maximum result among the 5 recommended conditions.

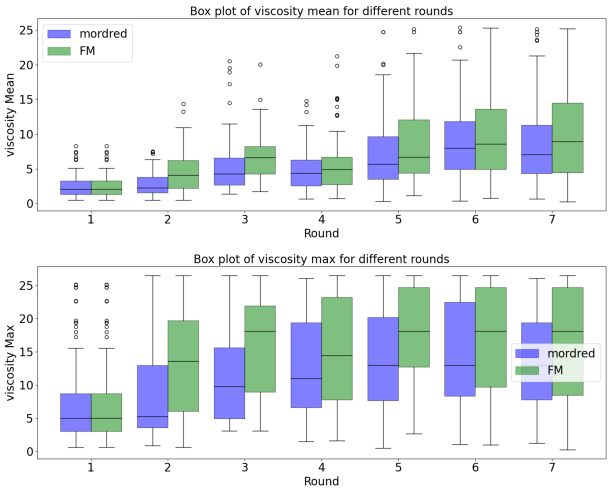

Fig. 3: Box plot comparison of the proposed method and benchmark approach. Up: the average result of the 5 recommended conditions. Down: the maximum result among the 5 recommended conditions.

3.2 *Experiment Setting*

We evaluate the proposed method using the ESM dataset [4], which provides viscosity measurements for different molecules at various temperatures. In this experiment, the optimization conditions are the molecule and temperature, while the optimization objective is to maximize viscosity. Since the dataset does not guarantee the presence of the optimal temperature corresponding to the model's output, we use the viscosity at the temperature closest to the optimal temperature for the given molecule in the dataset as the result corresponding to the model's optimal temperature output. We use viscosity as the evaluation metric. Since the objective is to maximize viscosity, higher values indicate better results. Each experiment consists of 7 rounds of condition optimization (including the initial random conditions), with 5 reaction conditions recommended in each round. We evaluated the proposed method and the baseline method across 100 different random seeds.

3.3 *Result*

We compare the proposed method, which enhances the SOTA chemical encoding Mordred, against using Mordred alone. For fair comparison, we apply PCA for dimensionality reduction to the features generated by Mordred encoding, ensuring that both methods have the same feature dimensionality when input into the BO model. As shown in Fig. 2, the performance of both methods improves progressively over the rounds, indicating that both are effectively identifying better reaction conditions. With the assistance of foundation model encoding, the performance surpasses that of using Mordred encoding alone across all 7 rounds. To provide a clearer visualization of the results and their variability, we also use a box plot to present the outcomes. As illustrated in Fig. 3, the proposed method demonstrates a significant improvement over Mordred encoding.

References

[1] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, 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, 2018.

[2] Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Venkat Kapil, William C Witt, Ioan-Bogdan Magdău, Daniel J Cole, and Gábor Csányi. Mace-off23: Transferable machine learning force fields for organic molecules. *arXiv preprint arXiv:2312.15211*, 2023.

[3] Chao Pang, Jianbo Qiao, Xiangxiang Zeng, Quan Zou, and Leyi Wei. Deep generative models in de novo drug molecule generation. *Journal of Chemical Information and Modeling*, 64(7):2174–2194, 2023.

[4] Alex K Chew, Matthew Sender, Zachary Kaplan, Anand Chandrasekaran, Jackson Chief Elk, Andrea R Browning, H Shaun Kwak, Mathew D Halls, and Mohammad Atif Faiz Afzal. Advancing material property prediction: using physics-informed machine learning models for viscosity. *Journal of Cheminformatics*, 16(1):31, 2024.