Multi-View Mixture-of-Experts for Predicting Molecular Properties Using SMILES, SELFIES, and Graph-Based Representations

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Indra Priyadarsini
IBM Research Tokyo
Tokyo, Japan
indra.ipd@ibm.com

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Victor Shirasuna
IBM Research Brazil
São Paulo, SP, Brazil
victor.shirasuna@ibm.com

Seiji Takeda
IBM Research Tokyo
Tokyo, Japan
seijitkd@jp.ibm.com

Abstract

Recent advancements in chemical machine learning have adopted a two-step approach—pre-training on unlabeled data followed by fine-tuning on specific tasks—to boost model capacity. With the increasing demand for training efficiency, Mixture-of-Experts (MoE) has become essential for scaling large models by selectively activating sub-networks of experts through a gating network, thereby optimizing performance.This paper presents MoL-MoE, a Multi-view Mixture-of-Experts framework designed to predict molecular properties by integrating latent spaces derived from SMILES, SELFIES, and molecular graphs. Our approach leverages the complementary strengths of these representations to enhance predictive accuracy. Here, we evaluate the performance of MoL-MoE with a total of 12 experts, organized into 4 experts for each modality (SMILES, SELFIES, and molecular graphs). We evaluate MoL-MoE on a range of benchmark datasets from MoleculeNet, demonstrating its superior performance compared to state-of-the-art methods across all nine datasets considering two different routing activation settings: $k = 4$ and $k = 6$. The results underscore the model's robustness and adaptability in handling various complex molecular prediction tasks. Our analysis of routing activation patterns reveals that MoL-MoE dynamically adjusts its use of different molecular representations based on task-specific requirements. This adaptability highlights the importance of representation choice in optimizing model performance.

1 Introduction

Chemical-based machine learning has become widely used for predicting molecular properties due to its effectiveness in capturing essential structural features [1, 2, 3, 4**?**]. Recent developments utilize a two-step process—pre-training on unlabeled data followed by fine-tuning on specific tasks [5, 6, 7, 8]—which has proven successful in scaling model capacity and improving performance.

Despite these advancements, there is an increasing focus on training efficiency [9], defined as the computational cost required to exceed the performance of current state-of-the-art systems [10]. This focus is driven by the growing emphasis on green AI initiatives [11]. The Mixture-of-Experts (MoE)

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

framework offers a promising solution for scaling model capacity while maintaining computational efficiency, significantly enhancing the training efficiency of large-scale models [9, 12, 13].

In MoE architectures, multiple experts act as sub-networks, with a gating network selectively activating only the most relevant experts for each input [11]. This gating mechanism ensures efficient routing and optimizes overall model performance [9]. Building on this, the Multi-View Mixture-of-Experts approach leverages diverse perspectives from various sources or modalities to improve model robustness and accuracy. By selectively activating different expert views based on input characteristics, Multi-View MoE models enhance their ability to capture complex relationships and generalize across different tasks and domains.

In this paper, we present MoL-MoE, a Multi-View Mixture-of-Experts approach tailored for small molecule analysis. MoL-MoE integrates latent spaces from three advanced chemical models: a large SMILES-based encoder-decoder [14], a BART-based SELFIES encoder-decoder [15], and a graph-based SMILES model [16]. A gating network is employed to optimize the latent space by defining and assigning weights to these diverse views. This fusion of multiple perspectives positions MoL-MoE as a robust framework for small molecule analysis.

Our experiments demonstrate that MoL-MoE outperforms existing state-of-the-art methods on complex small molecule tasks, as evaluated using the MoleculeNet benchmark dataset [17]. In this paper, we evaluate the performance of our model with a total of 12 experts, organized into 4 experts for each modality (SMILES, SELFIES, and molecular graphs). We investigate two different routing activation settings: $k = 4$ and $k = 6$. Specifically, our approach shows superior performance in all the 9 datasets for both classification and regression tasks when $k = 6$. Additionally, we provide an in-depth analysis of the MoE configurations explored in this study.

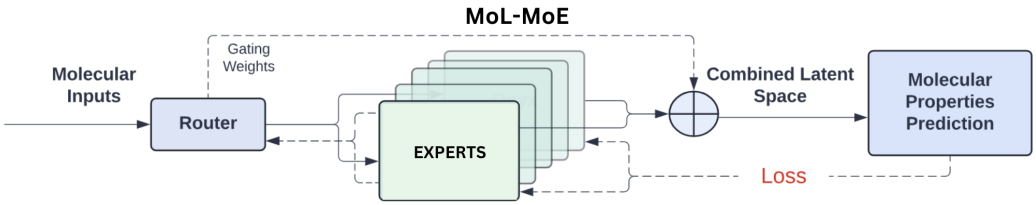

Figure 1: Architecture of the proposed Molecular Multi-view MoE (MoL-MoE) approach.

2 Methodology

In this section, we outline the methodological framework of our proposed approach. Fig. 1 illustrates the schema for latent space fusion using the MoL-MoE model. Our methodology integrates four key components: embeddings from SMILES representations, embeddings from SELFIES representations, molecular structures depicted as graphs, and a gating network that assigns and optimizes weights across these diverse perspectives.

The molecular Multi-View Mixture-of-Experts framework utilizes a network to effectively weigh and fuse embeddings from different perspectives viewpoints. The following subsection provides a detailed description of our proposed method.

2.1 Multi-View Mixture-of-Expert Layer

As shown in Fig. 1, the Multi-View Mixture-of-Experts (MoE) layer consists of n distinct "expert networks," labeled $E_1, E_2, \ldots, E_n$. Each expert is designed to focus on different aspects of the data, such as graphs and language. A gating network, denoted as G, helps to combine these experts by creating a sparse n-dimensional embedding space that is important for evaluating the task.

Before the gating network is used, the feature extraction module converts raw SMILES input into embeddings for the gating network. Each SMILES string is tokenized, and these tokens are turned into fixed 768-dimensional vectors. Mean pooling is then used to create a single embedding for the molecule. Other feature extraction methods can also be used to improve the molecule's representation.

The architecture includes a router module that selects which n expert networks will be activated based on the SMILES input. This improves the system's ability to adapt and specialize.

Let $G(x)$ be the output of the gating network and $E_i(\hat{x})$ be the output of the i-th expert network for a given SMILES input $\hat{x}$, where x is the embedding from the feature extractor. The output y of the Multi-View Mixture-of-Experts approach is a 768-dimensional embedding, given by:

$$y = \sum_{i=1}^{n} G(x)_i E_i(\hat{x}) \tag{1}$$

The resulting embedding space y is used to train a task-specific feed-forward network, with the loss function tailored to the specific task at hand. The optimization process adjusts the parameters of $G(x)$ based on this loss, which helps improve the gating network's performance for the task.

To address potential size mismatches between the experts' outputs and the feed-forward network's requirements, we standardize the output dimension of the Multi-View MoE to match the largest expert output size. In our case, this size was set to 768. If any expert's output is smaller than this size, the remaining space is filled with zeros.

In our experiments, we used experts to cover different perspectives of the data, such as language and graphs. However, more experts from various sources can be added if needed. To handle the computational complexity of using many experts, we can use a two-level hierarchical MoE, similar to the method described in [10].

The gating function in the MoE is implemented by multiplying the input by a trainable weight matrix W_g and then applying the *Softmax* function, as described in Eq. (2):

$$G_\sigma = \text{Softmax}(x \cdot W_g) \tag{2}$$

This formulation ensures that the gating mechanism appropriately distributes attention across the diverse set of experts, facilitating effective information integration from multiple sources.

Before applying the *Softmax* function, we introduce a router layer which is composed by tunable Gaussian noise and subsequently retain only the top k values, setting the remaining values to $-\infty$ (which effectively assigns corresponding gate values as 0). This sparsity-inducing step serves to optimize computational efficiency, as discussed previously. The magnitude of noise for each component is regulated by a second trainable weight matrix W_{noise}.

The formulation is expressed as follows:

$$G(x) = \text{Softmax}(\text{KeepTopK}(H(x), k)) \tag{3}$$

$$H(x)_i = (x \cdot W_g)_i + \text{StdNormal}() \cdot \text{Softplus}((x \cdot W_{noise})_i) \tag{4}$$

$$\text{KeepTopK}(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ of } v \\ -\infty & \text{otherwise} \end{cases} \tag{5}$$

This noise injection and sparsity-inducing mechanism contribute to the adaptability of the gating function, enabling it to effectively focus on relevant expert networks while controlling computational overhead. When opting for a value of k greater than 1, the gate values for the top k experts exhibit non-zero derivatives concerning the weights of the gating network. Furthermore, gradients propagate backward through the gating network to its inputs. The feed-forward network is employed for the adaptation of the MoL-MoE to specific tasks, further refining the model's capabilities for diverse and task-specific objectives. The experts that composes MoL-MoE are detailed in the next subsections.

In this paper, we evaluate the performance of our model with a total of 12 experts, organized into 4 experts for each modality (SMILES, SELFIES, and molecular graphs). We investigate two different routing activation settings: $k = 4$ and $k = 6$.

2.2 SMILES-based foundation model

For our proposed approach, we utilized the SMI-TED$_{289M}$ foundation model as the SMILES encoder [14]. SMI-TED$_{289M}$ is a large-scale, open-source encoder-decoder model pre-trained on a curated dataset of 91 million SMILES samples from PubChem, encompassing 4 billion molecular tokens. This model has demonstrated superior performance compared to state-of-the-art methods across various molecular tasks. The general architecture of SMI-TED$_{289M}$ is depicted in Fig. 2.

To construct the vocabulary of SMI-TED$_{289M}$, was employed the molecular tokenizer proposed by [18]. All 91 million molecules curated from PubChem were utilized in the tokenization process, resulting in a set of 4 billion molecular tokens. The unique tokens extracted from the resulting output provided a vocabulary of 2988 tokens plus 5 special tokens.

Pre-training of of SMI-TED$_{289M}$ was performed for 40 epochs through the entire curated PubChem dataset with a fixed learning rate of 1.6e-4 and a batch size of 288 molecules on a total of 24 NVIDIA V100 (16G) GPUs parallelized into 4 nodes using DDP and *torch run*. It involved two distinct phases: i) Learning of token embeddings through a masking process; ii) Subsequently, the token embeddings were mapped into a common latent space that encapsulates the entire SMILES string.

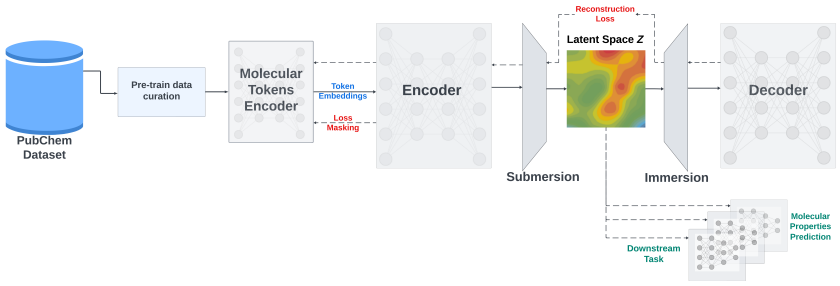

Figure 2: This figure illustrates the architecture of SMI-TED$_{289M}$.

2.3 SELFIES-based foundation model

SELFIES-BART, the SELFIES-based foundation model is an encoder-decoder architecture derived from the BART (Bidirectional Auto-Regressive Transformer) model [**?**]. The encoder processes the sequence of input token bidirectionally and the decoder generates the sequence autoregressively. The SELFIES-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 [19]. During pre-training the model is trained with a denoising objective function. The model is trained using the ZINC-22 [20] and PubChem [21] datasets. The dataset consists of molecules represented in SMILES notation. We convert these SMILES strings to SELFIES using the SELFIES API [19]. 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.

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

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.
- $P(Y_t|Y_{<t}, X_{\text{corrupt}}; \theta)$ is the probability predicted by the model for token Y_t, conditioned on the corrupted input and the previously generated tokens.

We hypothesize that the encoder-decoder structure of the SELFIES-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.

2.4 Graph-based model for small molecules

As a graph approach for small molecules we employ MHG-GNN [16], which is an autoencoder that combines GNN with Molecular Hypergraph Grammar (MHG) introduced for MHG-VAE [22]. MHG-GNN receives a molecular structure represented as a graph. The encoder constructed as Graph Isomorphism Network (GIN) [23] that additionally considers edges encodes that graph to its corresponding latent vector [24].

In the MHG-GNN framework, individual atoms forming a molecule are encoded using specific chemical characteristics, including attributes such as atomic number, formal charge, and aromaticity. Consequently, each atom feature is transformed into a vector of equal dimensions, aligning with the corresponding node in the GIN (Graph Isomorphism Network). The collective embedded representations of the atom features are then aggregated to create an initial vector, denoted as h_i^0, corresponding to the GIN node i. Similarly, the edges within the molecular structure, such as bond types, are also transformed into embedding vectors, designated as $e_{i,j}^0$, associated with the undirected edge in the GIN linking nodes j and i. Throughout the k-th iteration, the encoder executes what is termed as "message passing" for each node i, a process that can be defined as follows:

$$h_i^{k+1} = \mathrm{MLP}\left((1+\epsilon)h_i^k + \sum_{j \in N(i)} \mathrm{ReLU}(h_j^k + e_{j,i})\right) \tag{6}$$

where $N(i)$ is a set of direct neighbors of i, and ϵ is a trainable parameter, MLP is a neural network module, and ReLU is a Rectified Linear Unit. The entire representation h_G of graph G is defined by Eq. 7:

$$h_G = \mathrm{CONCAT}\left(\left\{\sum_{i \in V_G} h_i^k | k = 0, 1, \ldots, r\right\}\right) \tag{7}$$

CONCAT is used to concatenate vectors, V_G is a set of nodes in G, and r is the maximum iteration size. The entire representation h_G can be used as a latent vector for different downstream tasks.

The decoder is constructed as GRU and with several neural network models decodes that latent vector to the original molecular structure represented as a sequence of production rules on molecular hypergraphs. The production rules are generated from the dataset for pre-training. We used the model trained in the same steps described in [16] and with a radius, r, of 7 (i.e., the iteration size for message passing step in GNN). With these configurations, MHG-GNN generates 2048 dimensional embeddings. MHG-GNN was pre-trained on 1,381,747 molecules extracted from the PubChem database in its training part. This process generates 16,362 production rules that represent these molecules.

3 Experiments

To evaluate the effectiveness of our proposed methodology, we conducted experiments using a comprehensive set of 9 distinct benchmark datasets sourced from MoleculeNet [17], as illustrated in Table 1. Specifically, we evaluated 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 [17].

Table 1: MoleculeNet Benchmark datasets for classification task

Dataset	Description	# compounds	# tasks	Metric	Type
BBBP	Blood brain barrier penetration dataset	2039	1	ROC-AUC	Classification
Tox21	Toxicity measurements on 12 different targets	7831	12	ROC-AUC	Classification
Clintox	Clinical trial toxicity of drugs	1478	2	ROC-AUC	Classification
HIV	Ability of small molecules to inhibit HIV replication	41127	1	ROC-AUC	Classification
BACE	Binding results for a set of inhibitors for β – secretase 1	1513	1	ROC-AUC	Classification
SIDER	Drug side effect on different organ classes	1427	27	ROC-AUC	Classification
ESOL	Water solubility dataset	1128	1	RMSE	Regression
FreeSolv	Hydration free energy of small molecules in water	642	1	RMSE	Regression
Lipophilicity	Octanol/water distribution coefficient of molecules	4200	1	RMSE	Regression

4 Results and Discussion

In this section, we present the results of our classification and regression tasks. We will provide an overview of the experiment details and findings, evaluating the performance of our proposed Multi-View approach. Additionally, we will examine the specific MoE settings used in these experiments.

4.1 Results for classification tasks

Table 2 provides a detailed comparison of the performance of our Multi-View Mixture-of-Experts (MoL-MoE) approach against various state-of-the-art algorithms across different classification benchmarks. Our analysis reveals that MoL-MoE, which combines embeddings from SMILES, SELFIES, and molecular graphs, consistently achieves superior results across all datasets.

An important aspect to note is the complex nature of the classification tasks, as they encompass multi-task datasets such as Tox21, which comprises 12 tasks, Clintox with 2 tasks, and SIDER with a comprehensive 27-task dataset. This intricate and diverse task composition underscores the challenge posed by these classification tasks, making the consistent performance of our proposed approach across these datasets a testament to its reliability and robustness in handling complex and varied data.

Table 2: Methods and Performance for the classification tasks of MoleculeNet benchmark datasets

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
RF [25]	71.4	71.3	78.1	86.7	68.4	76.9
SVM [25]	72.9	66.9	79.2	86.2	68.2	81.8
MGCN [26]	85.0	63.4	73.8	73.4	55.2	70.7
D-MPNN [27]	71.2	90.5	75.0	85.3	63.2	68.9
DimeNet [28]	-	76.0	-	-	61.5	78.0
Hu, et al. [29]	70.8	78.9	80.2	85.9	65.2	78.7
N-Gram [30]	91.2	85.5	83.0	87.6	63.2	76.9
MolCLR [31]	73.6	93.2	80.6	89.0	68.0	79.8
GraphMVP [32]	72.4	77.5	77.0	81.2	63.9	74.4
GeomGCL [32]	-	91.9	-	-	64.8	85.0
GEM [1]	72.4	90.1	80.6	85.6	67.2	78.1
ChemBerta [33]	64.3	90.6	62.2	-	-	-
ChemBerta2 [34]	71.94	90.7	-	85.1	-	-
Galatica 30B [35]	59.6	82.2	75.9	72.7	61.3	68.5
Galatica 120B [35]	66.1	82.6	74.5	61.7	63.2	68.9
Uni-Mol [36]	72.9	91.9	80.8	85.7	65.9	79.6
Mixture of Collaborative Experts (MoCE) [37]	-	80.7	77.9	-	-	80.8
MoLFormer-XL [25]	93.7	94.8	82.2	88.2	69.0	84.7
MoL-MoE (k=4)	93.1	91.4	78.3	85.2	66.6	82.1
MoL-MoE (k=6)	**94.2**	**96.8**	**84.3**	**89.6**	**69.1**	**85.1**

The Table 2 highlights the complexity of the classification tasks addressed. For instance, Tox21 includes 12 separate tasks, Clintox consists of 2 tasks, and SIDER covers 27 tasks. The ability of MoL-MoE to maintain high performance across these varied and complex datasets underscores its robustness and effectiveness in handling challenging classification problems.

Our results indicate that MoL-MoE outperforms other leading methods such as ChemBerta, Chemberta2, and Galactica 30B and 120B. Specifically, MoL-MoE with $k = 6$ (which uses six out of

twelve available experts) demonstrates the best performance across all tested datasets. This version of MoL-MoE shows significant improvements compared to methods based on single representations, such as ChemBerta and Chemberta2, and those based solely on graphs, such as the approach by [37].

Additionally, the performance of MoL-MoE with $k = 4$ (which uses four experts) also exceeds that of several other state-of-the-art methods. This suggests that while using fewer experts still yields strong results, leveraging more experts (as in $k = 6$) further enhances the model's ability to integrate diverse perspectives and capture more nuanced patterns in the data.

Overall, our findings demonstrate that the fusion of multiple expert views in MoL-MoE effectively combines the strengths of different molecular representations, leading to improved classification performance. This method's ability to outperform existing approaches across a range of benchmarks highlights its potential as a robust tool for molecular property prediction and other related tasks

4.2 Results for regression tasks

In this section, we evaluate the performance of the proposed Multi-View Mixture-of-Experts (MoL-MoE) approach on complex regression tasks using the MoleculeNet dataset. The focus here is on predicting chemical properties through three challenging benchmarks: ESOL, FreeSolv, and Lipophilicity. Table 3 summarizes the performance results for these regression tasks.

Table 3: Methods and Performance for the regression tasks of MoleculeNet benchmark datasets.

Method	Dataset		
	ESOL	FreeSolv	Lipophilicity
GC [38]	0.97	1.40	0.65
$GROVER_{Large}$ [39]	0.89	2.27	0.82
Padel-DNN [40]	0.62	0.91	-
ChemRL-GEM [1]	0.80	1.88	0.66
ChemBERTa-2 [34]	0.89	-	0.80
SPMM [41]	0.82	1.90	0.69
Uni-Mol [36]	0.79	1.48	0.60
MPNN [42]	0.58	1.15	0.72
Multi-view GNN [43]	0.80	1.84	0.60
Multi-view GNN (cross)[43]	0.78	1.55	0.55
MoL-MoE (k=4)	0.57	1.23	0.52
MoL-MoE (k=6)	**0.53**	**1.12**	**0.51**

Table 3 shows that MoL-MoE achieves strong results across all three datasets. For ESOL, MoL-MoE with $k = 6$ achieves results compared to the state-of-the-art, outperforming other methods, including those based on graph convolutions and multi-view approaches. This indicates that MoL-MoE effectively captures the nuances of chemical properties in this dataset.

In the FreeSolv dataset, which is known for its complexity due to its large and diverse set of solubility, MoL-MoE with $k = 6$ achieves the best performance with an error of 1.12. This result is better than those from other approaches, such as GC and GROVER, demonstrating MoL-MoE's capability in handling intricate regression tasks involving solvation energy predictions.

For Lipophilicity, MoL-MoE with $k = 6$ also delivers the best result with an error of 0.51. This performance surpasses the results of other methods, highlighting the model's effectiveness in predicting this chemical property, which is crucial for understanding molecular interactions and drug design.

Overall, MoL-MoE shows robust performance across all tested regression benchmarks, consistently outperforming other state-of-the-art methods. The results demonstrate that MoL-MoE effectively integrates multiple molecular representations and expert views to achieve accurate predictions. The model's ability to handle different types of chemical properties with high precision further supports its potential for broader applications in chemical informatics and related fields.

4.3 A deeper analysis over the different MoE settings

In this section, we analyze the activation patterns of the experts within the MoL-MoE model for various tasks investigated in our experiments. Fig. 3 displays the activation distribution of the MoL-MoE model when $k = 6$ for the BACE dataset.

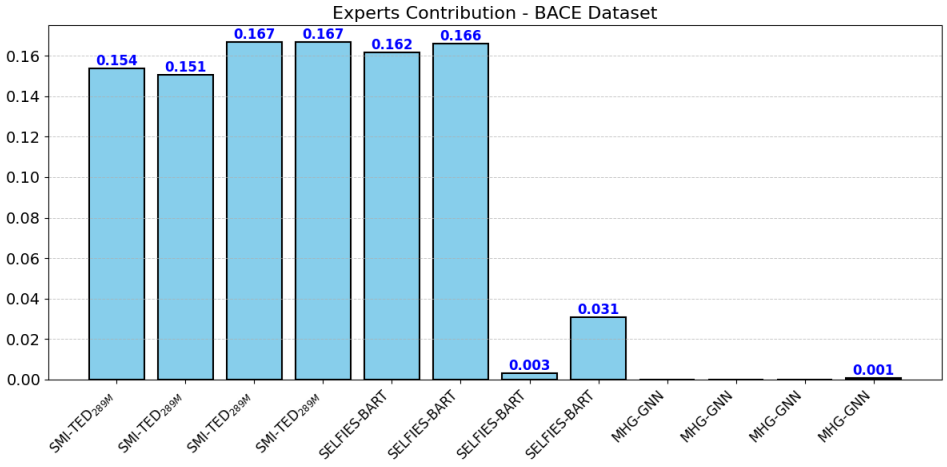

Figure 3: This figure illustrates the activation patterns of the MoL-MoE model with $k = 6$ for the BACE dataset.

From Fig. 3, it is evident that the routing algorithm predominantly activates the experts based on SMILES and SELFIES representations more frequently than those based on molecular graphs. This observation suggests that the SMILES and SELFIES representations may offer more effective or complementary features for the classification task associated with the BACE dataset.

The preference for SMILES and SELFIES over molecular graphs could indicate that these representations capture features that are particularly relevant for the classification of BACE inhibitors.

In contrast, Fig. 4 shows that the MoL-MoE model exhibits a more balanced activation pattern for the ESOL regression task. Specifically, the gating network distributes activation more evenly across the different representations—SMILES, SELFIES, and molecular graphs. This balanced contribution suggests that for the ESOL task, which may involve a complex interplay of features, each type of representation contributes valuable information. The equal activation across these representations indicates that the model leverages the diverse information from each representation effectively, highlighting the model's ability to integrate and utilize multiple data sources for accurate regression predictions. This balanced activation could enhance the model's robustness and overall performance by ensuring that no single representation dominates, allowing for a more comprehensive understanding of the chemical properties being predicted.

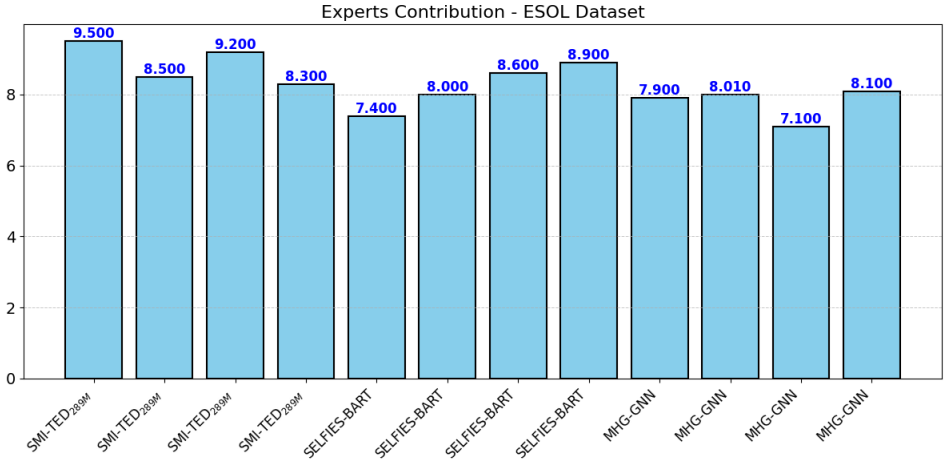

Figure 4: This figure illustrates the activation patterns of the MoL-MoE model with $k = 6$ for the ESOL dataset.

This pattern of expert activation highlights the adaptive nature of the MoL-MoE model in leveraging different molecular representations based on the task requirements. It also suggests that the choice of representation plays a crucial role in the performance of the model, particularly for tasks with specific characteristics or complexities.

5 Conclusion

This paper introduces MoL-MoE, a Multi-view Mixture-of-Experts framework that integrates multiple latent spaces from SMILES, SELFIES, and molecular graphs to predict molecular properties. Our evaluations using MoleculeNet benchmark datasets reveal that MoL-MoE consistently outperforms state-of-the-art methods across all nine datasets. This highlights the model's robustness and flexibility in addressing a wide range of complex molecular prediction tasks.

The routing activation patterns demonstrate that MoL-MoE dynamically adjusts its focus on different molecular representations based on the specific needs of each task. This indicates that the choice of representation is crucial for optimizing model performance, especially for tasks with distinct characteristics or complexities. For instance, some tasks may benefit more from SMILES or SELFIES, while others may require a focus on molecular graphs.

By combining embeddings from diverse sources, MoL-MoE effectively captures detailed structural features and complex molecular interactions. This comprehensive approach enhances the model's ability to address the subtleties of molecular data, leading to improved predictive performance.

References

[1] 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.

[2] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, and T. Langer, "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, 2020.

[3] J. Shen and C. A. Nicolaou, "Molecular property prediction: recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, vol. 32, pp. 29–36, 2019.

[4] E. Soares, F. Cipcigan, D. Zubarev, and E. V. Brazil, "A framework for toxic pfas replacement based on gflownet and chemical foundation model," in *NeurIPS 2023 AI for Science Workshop*, 2023.

[5] E. A. Soares, V. Shirasuna, E. A. V. Brazil, R. F. de Gusmao Cerqueira, D. Zubarev, T. Callahan, and S. Capponi, "Molmamba: A large state-space-based foundation model for chemistry," in *American Chemical Society (ACS) Fall Meeting*, 2024.

[6] S. Takeda, A. Kishimoto, L. Hamada, D. Nakano, and J. R. Smith, "Foundation model for material science," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, pp. 15 376–15 383.

[7] E. Soares, E. V. Brazil, K. F. A. Gutierrez, R. Cerqueira, D. Sanders, K. Schmidt, and D. Zubarev, "Beyond chemical language: A multimodal approach to enhance molecular property prediction," *arXiv preprint arXiv:2306.14919*, 2023.

[8] S. Horawalavithana, E. Ayton, S. Sharma, S. Howland, M. Subramanian, S. Vasquez, R. Cosbey, M. Glenski, and S. Volkova, "Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned," in *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022, pp. 160–172.

[9] E. Soares, A. Kishimoto, E. V. Brazil, S. Takeda, H. Kajino, and R. Cerqueira, "Improving molecular properties prediction through latent space fusion," *arXiv preprint arXiv:2310.13802*, 2023.

[10] M. Pióro, K. Ciebiera, K. Król, J. Ludziejewski, and S. Jaszczur, "Moe-mamba: Efficient selective state space models with mixture of experts," *arXiv preprint arXiv:2401.04081*, 2024.

[11] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[12] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.

[13] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[14] V. Y. Shirasuna, E. Soares, E. V. Brazil, K. F. A. Gutierrez, R. Cerqueira, S. Takeda, and A. Kishimoto, "A multi-view mixture-of-experts based on language and graphs for molecular properties prediction," in *ICML 2024 AI for Science Workshop*, 2024.

[15] E. Soares, V. Shirasuna, E. V. Brazil, R. Cerqueira, D. Zubarev, and K. Schmidt, "A large encoder-decoder family of foundation models for chemical language," *arXiv preprint arXiv:2407.20267*, 2024.

[16] 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.

[17] A. Kishimoto, H. Kajino, M. Hirose, J. Fuchiwaki, I. Priyadarsini, L. Hamada, H. Shinohara, D. Nakano, and S. Takeda, "Mhg-gnn: Combination of molecular hypergraph grammar with graph neural network," 2023.

[18] 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.

[19] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] H. Kajino, "Molecular hypergraph grammar with its application to molecular optimization," in *ICML*, 2019, pp. 3183–3191, also see the supplementary material available at http://proceedings.mlr.press/v97/kajino19a/kajino19a-supp.pdf.

[24] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *ICLR*, 2019.

[25] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," in *ICRL*, 2020.

[26] 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.

[27] 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, no. 01, 2019, pp. 1052–1060.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

[38] X. Yao, S. Liang, S. Han, and H. Huang, "Enhancing molecular property prediction via mixture of collaborative experts," *arXiv preprint arXiv:2312.03292*, 2023.

[39] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS central science*, vol. 3, no. 4, pp. 283–293, 2017.

[40] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, "Self-supervised graph transformer on large-scale molecular data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 559–12 571, 2020.

[41] K. Zhang and H. Zhang, "Predicting solute descriptors for organic chemicals by a deep neural network (dnn) using basic chemical structures and a surrogate metric," *Environmental Science & Technology*, vol. 56, no. 3, pp. 2054–2064, 2022.

[42] J. Chang and J. C. Ye, "Bidirectional generation of structure and properties through a single molecular foundation model," 2023.

[43] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[44] H. Ma, Y. Bian, Y. Rong, W. Huang, T. Xu, W. Xie, G. Ye, and J. Huang, "Multi-view graph neural networks for molecular property prediction," *arXiv preprint arXiv:2005.13607*, 2020.