A Multi-View Mixture-of-Experts based on Language and Graphs for Molecular Properties Prediction

Victor Shirasuna*[1] Eduardo Soares*[1] Emilio Vital Brazil*[1] Karen Fiorella Gutierrez[1] Renato Cerqueira[1]
Seiji Takeda[2] Akihiro Kishimoto[2]

Abstract

Recent progress in chemical-based machine learning utilizes a two-step process – pre-training on unlabeled corpora and fine-tuning on specific tasks – to enhance model capacity. Emphasizing the growing need for training efficiency, Mixture-of-Experts (MoE) efficiently scales model capacity, particularly vital for large-scale models. In the MoE architecture, sub-networks of multiple experts are selectively tailored through a gating network, optimizing overall model performance. Extending this, a Multi-View Mixture-of-Experts enhances model robustness and accuracy by fusing embeddings from different natures. Here, we introduce Mol-MVMoE, a novel approach for small molecules by fusing latent spaces from diverse chemical-based models. Utilizing a gating network to define and assign weights to different perspectives, Mol-MVMoE emerges as a robust framework for small molecule analysis. We assessed Mol-MVMoE using 11 benchmark datasets from MoleculeNet, where it outperformed competitors in 9 of them. We also provide a deep analysis of the results obtained with the QM9 dataset, where Mol-MVMoE consistently performed better than its state-of-the-art competitors. Our study highlights the potential of latent space fusion and different perspectives integration for advancing molecular property prediction. This not only signifies current advancement but also promises future refinements with the inclusion large-scale models.

1. Introduction

Chemical-based machine learning has gained widespread adoption for predicting molecular properties due to its efficiency in representing crucial structural aspects (Fang et al., 2022; Wieder et al., 2020; Shen & Nicolaou, 2019). Recent advancements leverage a two-step process, pre-training on unlabeled corpora and fine-tuning on specific tasks (Takeda et al., 2023; Soares et al., 2023; Horawalavithana et al., 2022), demonstrating success in scaling model capacity and enhancing performance.

While these strides in model development have significantly improved performance, there is a growing need to prioritize training efficiency (Pióro et al., 2024). Defined as the total computation required to surpass the quality of state-of-the-art systems (Shazeer et al., 2017), training efficiency gains importance in light of the increasing emphasis on green AI initiatives (Zhou et al., 2022). Mixture-of-Experts (MoE) emerges as a compelling solution for scaling model capacity within a fixed computational cost, playing a crucial role in enhancing the training efficiency of large-scale language models (Pióro et al., 2024; Jiang et al., 2024).

In the Mixture-of-Experts architecture, multiple experts operate as sub-networks, and their activation is selectively tailored, engaging only one or a few experts for each input (Zhou et al., 2022). The pivotal role of a gating network in this process is to efficiently route each input to the most suitable expert(s), optimizing the overall model performance (Pióro et al., 2024). Expanding on this concept, a Multi-View Mixture of Experts capitalizes on diverse perspectives from different sources or modalities, enhancing model robustness and accuracy. Through the selective activation of expert views based on input characteristics, Multi-View MoE models proficiently capture complex relationships in data, thereby fostering improved generalization across various tasks and domains.

In this paper, we introduce a Multi-View Mixture-of-Experts for small molecules (Mol-MVMoE) approach that leverages on the fusion of latent spaces from different natures generated by two state-of-the-art chemical-based models, a large language model based on the Transformer architec-

*Equal contribution [1]IBM Research, Rio de Janeiro, Brazil [2]IBM Research, Tokyo, Japan. Correspondence to: Eduardo Soares <eduardo.soares@ibm.com>, Emilio Vital Brazil <evital@br.ibm.com>.

Proceedings of the 41st International Conference on Machine Learning, AI for Science Workshop, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ture (Ross et al., 2022), and a graph-based approach for SMILES (Kishimoto et al., 2023). To achieve an optimized latent space tailored for specific tasks, a gating network is employed. This network serves to define and assign weights to the various views comprising the latent space. Through this intricate fusion of diverse perspectives, Mol-MVMoE emerges as a robust framework for enhancing the understanding and analysis of small molecules in chemical contexts.

Our findings demonstrate that our proposed Mol-MVMoE surpasses existing state-of-the-art algorithms, when it comes to tackling intricate tasks of small molecules. These challenging tasks are part of the MoleculeNet benchmark dataset (Wu et al., 2018). Furthermore, our approach exhibits superior performance in 9 out of 11 datasets studied during our experiments for both classification and regression tasks, including the QM9 dataset which is related to the quantum properties of the molecules. For this particular dataset we provide a deeper investigation over the 12 properties which are related to it. In this case, the best version of our proposed Mol-MVMoE was able to perform better in 7 out of the 12 properties within the QM9 dataset when compared with other recent state-of-the-art approaches.

It is crucial to underscore that the Mol-MVMoE approach involves the fusion of latent spaces from smaller models, consistently outperforming state-of-the-art larger models like MoLFormer-XL, trained on 1.1 billion molecules. This demonstration not only showcases a substantial improvement in performance with the potential to advance the field but also unveils opportunities for further refinement, particularly as our approach incorporates even large-scale models in the future.

2. Methodology

In this section, we explain the methodological framework outlined in this paper. In Figure 1, we illustrate the schema for latent space fusion using the proposed Mol-MVMoE approach. Our methodology relies on three key components: embeddings derived from molecular structures represented as graphs, embeddings rooted in chemical language, and a gating network that defines and assigns weights to the diverse perspectives constituting the latent space. Through this intricate fusion, Mol-MVMoE enhances the comprehension and analysis of small molecules within chemical contexts, capturing intricate relationships in data based on varying perspectives.

The molecular multi-view mixture-of-experts employs a network to weigh and fuse embeddings from different viewpoints. The graph-based architecture of MHG-GNN excels in accurately capturing molecular substructures compared to the language model-based MoLFormer. Conversely, the self-attention mechanism of MoLFormer offers an advantage in accounting for relationships between atoms, even when their distances exceed the radius covered by the graph approach. Details of the proposed method are given in the next subsection.

2.1. Multi-View Mixture-of-Expert Layer

As illustrated in Figure 1, the Multi-View Mixture-of-Experts layer comprises a set of n distinct ”expert networks” labeled as $E_1, E_2, \ldots, E_n$. Each expert is meticulously crafted to capture unique perspectives on the underlying data, spanning domains such as graphs, language, and more. Augmenting these experts is a gating network denoted as G, tasked with generating a sparse n-dimensional embedding space crucial for task evaluation within the proposed method.

Before applying the gating network, the feature extraction module maps the raw input SMILES into an embedding to be fed into the gating network. In this work, we map each SMILES into tokens and then convert the input tokens to fixed vectors of dimension 768. Finally, a mean pooling method is applied to all token embeddings in order to extract a single meaningful embedding of the molecule. We highlight that any feature extraction method can be applied as well to improve the representation of the molecules into the gating network.

Furthermore, the proposed architecture is enhanced with a router module responsible for determining the n experts that will be activated and receive inputs, which are based on SMILES, further refining the adaptability and specialization of the system.

Let $G(x)$ and $E_i(\hat{x})$ denote the output of the gating network and the output of the i-th expert network, respectively, for a given input $\hat{x}$ of SMILES and x, which is the embeddings derived from the feature extractor, following a similar notation as proposed in (Shazeer et al., 2017). The resulting output y of the Multi-View Mixture-of-Experts (MoE) approach is an embedding space of size 2048, defined as follows:

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

The resulting embedding space y is utilized to train a task-specific feed-forward network, where the loss function is chosen according to the specific task. The optimization process refines the parameters of $G(x)$ based on the incurred loss, enhancing the effectiveness of the gating network for the given task.

Still the output dimension size of the experts may diverge to be fed into the feed-forward network. To tackle this issue,

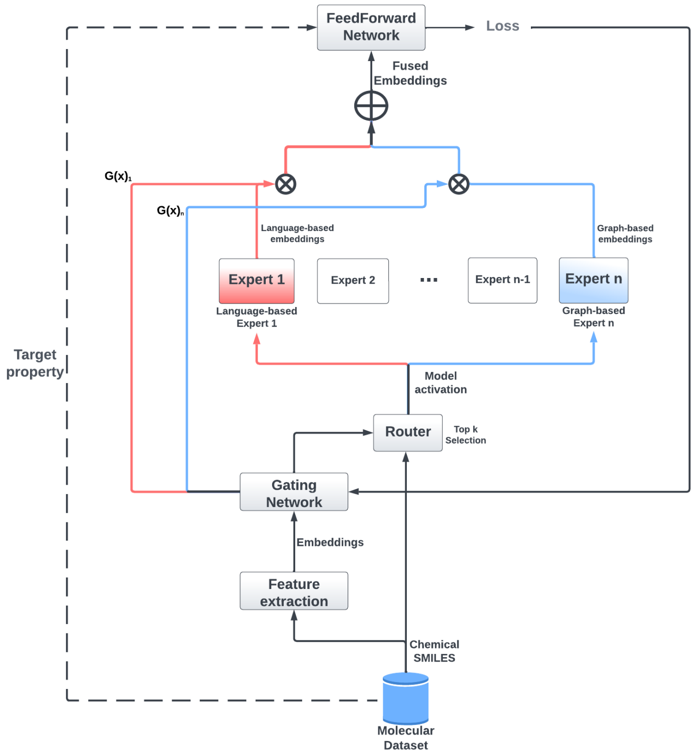

Figure 1. Architecture of the proposed Molecular Multi-view MoE (Mol-MVMoE) approach.

we first set the resulting size output y of the Multi-View Mixture-of-Experts (MoE) to be the largest experts output size, which in our case it was 2048. Thus, all the remaining spaces will be filled with a $.0$ value if the expert's output is less than the desired embedding space y size.

In our experiments, we selected experts to represent diverse perspectives of the data, including language and graphs. However, if needed, a larger number of experts from various sources could be incorporated. To manage computational complexity when dealing with a considerable number of experts, our strategy can employ a two-level hierarchical Mixture-of-Experts (MoE), similar to the approach presented in (Shazeer et al., 2017).

The sparse gating function utilized in the MoE is formulated by the multiplication of the input by a trainable weight matrix W_g, followed by the application of the *Softmax* function, as described by Equation (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-MVMoE to specific tasks, further refining the model's capabilities for diverse and task-specific objectives. The experts that composes Mol-MVMoE are detailed in the next subsections.

2.2. Graph-based model for small molecules

As a graph approach for small molecules we employ MHG-GNN (Kishimoto et al., 2023), which is an autoencoder that combines GNN with Molecular Hypergraph Grammar (MHG) introduced for MHG-VAE (Kajino, 2019).

Unlike existing autoencoders that receive their input and output in the same format, MHG-GNN receives them in a different format. MHG-GNN receives a molecular structure represented as a graph. The encoder constructed as Graph Isomorphism Network (GIN) (Xu et al., 2019) that additionally considers edges encodes that graph to its corresponding latent vector (Hu et al., 2020). 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.

MHG-GNN can inherit advantage of MHG-VAE that can always generate structurally valid molecular structures when decoding latent vectors. Additionally, MHG-GNN can always embed graph structures to their latent vectors, whereas the encoder of MHG-VAE cannot always; it cannot accept a molecule that cannot be represented by a set of production rules generated from the dataset for pre-training. Finally, thanks to GNN, MHG-GNN has more direct understanding to the structural information than language-based models, which may capture different characteristics than MoLFormer.

We used the model trained in the same steps described in (Kishimoto et al., 2023) 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.

2.3. Chemical language-based model

For chemical language-based model we employ MoLFormer (Ross et al., 2022), which is a large-scale masked chemical language model that processes inputs through a series of blocks that alternate between self-attention and feed-forward connections. MoLFormer was trained in a self-supervision manner with 1.1 billion molecules from PubChem and ZINC datasets and uses tokenization process, as detailed in (Schwaller et al., 2019).

MoLFormer is equipped with a self-attention mechanism that allows the network to construct complex representations that incorporate context from across the sequence of SMILES. By transforming the sequence features into queries (q), keys (k), and value (v) representations, attention mechanisms can weigh the importance of different elements within the sequence. MoLFormer optimizes relative encoding by using a modified version of the RoFormer (Su et al., 2021) attention mechanism. This involves position-dependent rotations (R_m) of the query and keys at position m. These rotations can be efficiently implemented as pointwise multiplications, ensuring that the computational complexity remains manageable (as shown in Eq (8)).

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
QM9	12 quantum mechanical calculations	133885	12	Average MAE	Regression
QM8	12 excited state properties of small molecules	21786	12	Average MAE	Regression
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

$$\mathrm{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle} \tag{8}$$

In Eq (8), $\mathrm{Attention}_m(Q,K,V)$ denotes the attention operation with queries (Q), keys (K), and values (V) at position m. The operation computes weighted sums of the value representations (v_n) based on the similarity of the transformed query ($\varphi(R_m q_m)$) and key ($\varphi(R_n k_n)$) representations. The relative position embeddings introduced through the rotations (R_m) allow the model to effectively capture positional information, leading to improved performance in molecular property predictions. In this work, we used the base version of the MoLFormer that was trained on a small portion of molecules compared to the MoLFormer-XL version. The MoLFormer-base version it is publicly available at `https://github.com/IBM/molformer`. Table 2 elucidates the hyper-parameters used to generate the specialized models for each regression task.

Table 2. MoLFormer Hyper-parameters for fine-tuning

Hyper-parameter	Values
Batch size	128
Learning Rate	$3e-5$
Number of embeddings	768
Dropout	0.1
Number of layers	12
Number of heads	12
Number of epochs (max)	500

3. Downstream Tasks Datasets

To evaluate the effectiveness of our proposed methodology, we conducted experiments using a comprehensive set of 11 distinct benchmark datasets sourced from MoleculeNet (Wu et al., 2018), as illustrated in Table 1. Specifically, we evaluated 6 datasets for the classification task and 5 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 (Wu et al., 2018).

3.1. Classification Tasks

For the classification task, we selected six distinctive classification tasks sourced from the MoleculeNet benchmark dataset. These specific tasks, namely BBBP, ClinTox, HIV, BACE, SIDER, and Tox21, were selected to represent a diverse array of chemical properties and biological activities, with their key characteristics thoughtfully summarized in Table 1. To ensure a consistent assessment, we employed the AUC-ROC metric to evaluate the performance of our models. Additionally, we leveraged scaffold splits as a reliable and established technique for the systematic evaluation of model performance.

3.2. Regression Tasks

For the regression task we choose five different regression tasks from the MoleculeNet. Specifically, the QM9 and QM8 subsets entail the prediction of various quantum chemical metrics, a challenging feat in the absence of exclusive 3D geometric information. Further details on the characteristics of these regression datasets can be found in Table 1. To evaluate the QM9 and QM8 datasets we report the average MAE, while RSME is reported for the remaining tasks.

4. Results

In this section, we present the analysis of the results obtained for the classification and regression tasks considered in this experiment, shedding light on the nuanced intricacies and outcomes derived from the experimentation process. Through this evaluation, we aim to provide a deeper understanding of the impact and potential of our proposed Multi-view methodology.

4.1. Ablation Studies

In this section, we compare our proposed methodology against the single models, MoLFormer and MHG-GNN, that we use to compose our proposed molecular Multi-view MoE, Mol-MVMoE, approach.

Table 3 highlights the enduring superiority of our fusion-based approach across all conducted experiments compared to the single MHG-GNN and MoLFormer (base and XL versions) methods. These results affirm our assertion that integrating diverse perspectives of the data yields a more comprehensive understanding of its intricacies, surpassing the performance of singular, large-scale models.

Furthermore, it is worth emphasizing that the Mol-MVMoE approach is built upon the foundation of MoLFomer-Base. While MoLFomer-Base initially achieved the worst results for the Tox21 dataset, the integration of multiple features views through our Mol-MVMoE approach led to a significant performance boost, elevating the model's performance from 43.2 to 85.6 (best performance).

4.2. Benchmark Tests with SOTA Methods

4.2.1. Results for classification tasks

Table 4 offers a comprehensive overview of the comparative performance between our proposed Multi-view MoE approach and state-of-the-art algorithms on various benchmark datasets for the classification task. A keen analysis of the table reveals that the Mol-MVMoE, which leverages the fusion of embeddings from different perspectives, outperforms its counterparts in all tested datasets, underscoring its potential to excel in diverse domains.

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.

Our proposed fusion-based Mol-MVMoE harnesses the power of latent spaces from transformers-based language approach and embeddings from graph-based, capitalizing on their complementary strengths to excel in a variety of challenging tasks. $k = 1$ means that we are selecting one or other expert, in counterparts, $k = 2$ means that we are mixing the experts in order to get the best of both perspectives. Results demonstrates that Mol-MVMoE performs better than state-of-the-art approaches as ChemBerta, Chemberta2, Galatica 30 and 120B, in all the experiments conducted. Mol-MVMoE also present better results in all the tested datasets when compared to the Multi-view approach proposed by (Yao et al., 2023), which is just based on graphs.

It is important to highlight that we use the fusion of latent spaces of two smaller models when compared to the state-of-the-art, MoLFormer-base and MHG-GNN. The fusion of these smalls models performed better than very large models as MoLFormer-XL, which was trained in 1.1 billion molecules, in all benchmarks datasets. This not only highlights our method's effectiveness but also paves the way for additional enhancements when our approach incorporates a larger-scale models from different natures.

4.2.2. Results for regression tasks

Next, we applied the proposed Mol-MVMoE to the prediction of chemical properties, tackling more intricate regression tasks sourced from the MoleculeNet database. The performance results across five challenging regression benchmarks, namely QM9, QM8, ESOL, FreeSolv, and Lipophilicity, are summarized in Table 5.

The regression tasks presented in the MoleculeNet benchmark datasets, especially the challenging QM9 and QM8 sets, pose a significant test for predictive models due to the intricate nature of quantum chemical measures. Table 5 elucidates the that Mol-MVMoE approach has not only surpassed the previous state-of-the-art performance achieved by MoLFormer-XL in both tasks (QM8 and QM9) but has also demonstrated reliability in handling the complexities embedded in these intricate quantum chemical datasets.

By harnessing the combined strengths of graph representations and the powerful linguistic insights embedded within a tailored language model for chemistry, our Mol-MVMoE approach has showcased significant advancements in performance, particularly in the QM9 dataset. This fusion of different perspectives over the same data has enabled our model to unravel the intricate relationships between molecular structures and the corresponding quantum chemical properties with greater precision and depth.

Furthermore, the Mol-MVMoE approach has displayed a clear competitive edge in predicting Lipophilicity when compared to other established methods, thereby highlighting its robustness and adaptability across diverse chemical property prediction tasks. While the performance on the ESOL and FreeSolv datasets aligns closely with that of the baseline approaches, the consistent and promising results obtained by our Mol-MVMoE strategy across various regression tasks underline its potential in the domain of chemical property prediction.

Table 3. Comparison between the proposed Mol-MVMoE approach and single models.

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
MoLFormer-XL (Ross et al., 2022)	93.7	94.8	82.2	88.2	69.0	84.7
MoLFormer-Base (Ross et al., 2022)	90.9	77.7	82.8	64.8	61.3	43.2
MHG-GNN	93.5	90.0	83.4	87.3	67.6	77.5
Mol-MVMoE (k=1)	92.9	91.9	76.5	87.4	66.6	83.5
Mol-MVMoE (k=2)	**93.8**	**95.9**	**84.2**	**89.1**	**69.1**	**85.6**

Table 4. Methods and Performance for the classification tasks of MoleculeNet benchmark datasets

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
RF (Ross et al., 2022)	71.4	71.3	78.1	86.7	68.4	76.9
SVM (Ross et al., 2022)	72.9	66.9	79.2	86.2	68.2	81.8
MGCN (Lu et al., 2019)	85.0	63.4	73.8	73.4	55.2	70.7
D-MPNN (Yang et al., 2019)	71.2	90.5	75.0	85.3	63.2	68.9
DimeNet (Gasteiger et al., 2020)	-	76.0	-	-	61.5	78.0
Hu, et al. (Hu et al., 2019)	70.8	78.9	80.2	85.9	65.2	78.7
N-Gram (Liu et al., 2019)	91.2	85.5	83.0	87.6	63.2	76.9
MolCLR (Wang et al., 2022)	73.6	93.2	80.6	89.0	68.0	79.8
GraphMVP (Liu et al., 2021)	72.4	77.5	77.0	81.2	63.9	74.4
GeomGCL (Liu et al., 2021)	-	91.9	-	-	64.8	85.0
GEM (Fang et al., 2022)	72.4	90.1	80.6	85.6	67.2	78.1
ChemBerta (Chithrananda et al., 2020)	64.3	90.6	62.2	-	-	-
ChemBerta2 (Ahmad et al., 2022)	71.94	90.7	-	85.1	-	-
Galatica 30B (Taylor et al., 2022)	59.6	82.2	75.9	72.7	61.3	68.5
Galatica 120B (Taylor et al., 2022)	66.1	82.6	74.5	61.7	63.2	68.9
Uni-Mol (Zhou et al., 2023)	72.9	91.9	80.8	85.7	65.9	79.6
Mixture of Collaborative Experts (MoCE) (Yao et al., 2023)	-	80.7	77.9	-	-	80.8
MoLFormer-XL (Ross et al., 2022)	93.7	94.8	82.2	88.2	69.0	84.7
Mol-MVMoE (k=1)	92.9	91.9	76.5	87.4	66.6	83.5
Mol-MVMoE (k=2)	**93.8**	**95.9**	**84.2**	**89.1**	**69.1**	**85.6**

4.2.3. A deeper analysis over the QM9 benchmark

In this subsection, we delve further into the exploration of results for individual tasks within the QM9 benchmark dataset, aiming to uncover nuanced insights and patterns inherent to each specific measure property. The twelve distinct properties of QM9, each accompanied by their respective units, are detailed in Table 6.

Within this paper, we compare the best version ($k = 2$) and standard version ($k = 1$) of our Mol-MVMoE approach against a selection of previously discussed baseline models, as well as four additional baselines. Our comparative analysis extends to benchmarking the MoE Multi-view approach against state-of-the-art models derived from three distinct categories: (i) Graph-based, (ii) Geometry-based, and (iii) SMILES-based methodologies for prediction of molecular properties. The included baselines models are: 123-gnn (Morris et al., 2019), a multitask neural net encoding the Coulomb Matrix (CM) (Rupp et al., 2012), and its GNN variant as in the deep tensor neural net (DTNN) (Schütt et al., 2017), we also considered the ChemBERTa (Chithrananda et al., 2020) approach in this study.

Table 7 presents a comprehensive comparison of the performance of various state-of-the-art models on the QM9 dataset, highlighting the effectiveness of different modeling strategies. Our proposed Multi-view MoE approach outperforms the current models in 7 out of the 12 properties in its best version, and presents the best and second best overall results for all the tasks in general.

The performance variation across different properties suggests that a one-size-fits-all approach might not be the most effective solution, as seen in the case of the property μ, where geometry-based models outperformed graph and SMILES-based approaches. This underscores the importance of considering a multiple perspective approach when dealing with such complex task.

These results highlights the potential benefits of leveraging a heterogeneous embedding spaces for accurate prediction of molecular properties. Furthermore, a notable observation from the results is that the 123-gnn model outperforms the MoLFormer-XL in a greater number of properties, but this difference has had a detrimental impact on the average mean absolute error (Avg MAE). Conversely, the fusion of views Mol-MVMoE has exhibited robust and consistent

Table 5. Methods and Performance for the regression tasks of MoleculeNet benchmark datasets.

Method	Dataset				
	QM9	QM8	ESOL	FreeSolv	Lipophilicity
GC (Altae-Tran et al., 2017)	4.35	0.0148	0.97	1.40	0.65
A-FP (Xiong et al., 2019)	2.63	0.0282	0.50	0.74	0.58
$GROVER_{Large}$ (Rong et al., 2020)	-	-	0.89	2.27	0.82
Padel-DNN (Zhang & Zhang, 2022)	-	-	0.62	0.91	-
ChemRL-GEM (Fang et al., 2022)	-	-	0.80	1.88	0.66
ChemBERTa-2 (Ahmad et al., 2022)	-	-	0.89	-	0.80
SPMM (Chang & Ye, 2023)	-	-	0.82	1.90	0.69
Uni-Mol (Zhou et al., 2023)	-	0.0156	0.79	1.48	0.60
MPNN (Gilmer et al., 2017)	3.18	0.0143	0.58	1.15	0.72
MoLFormer-XL (Ross et al., 2022)	1.59	0.0102	**0.28**	**0.23**	0.53
Multi-view GNN (Ma et al., 2020)	–	0.0127	0.80	1.84	0.60
Multi-view GNN (cross)(Ma et al., 2020)	–	0.0124	0.78	1.55	0.55
Mol-MVMoE (k=1)	1.51	0.0098	0.57	1.33	0.58
Mol-MVMoE (k=2)	**1.49**	**0.0097**	0.57	1.38	**0.52**

Table 6. Data description

Measure	Unit
α	$Bohr^3$
C_v	$cal/(mol * K)$
G	Hartree
gap	Hartree
H	Hartree
ϵ_{homo}	Hartree
ϵ_{lumo}	Hartree
μ	Debye
$\langle R^2 \rangle$	$Bohr^2$
U_0	Hartree
U	Hartree
ZPVE	Hartree

Table 7. Comparing state-of-the-art models performance on QM9 test set. **Blue** and **Orange** indicates best and second-best performing model, respectively.

	Graph-based			Geometry-based			SMILES-based		Mol-MVMoE	
Measure	A-FP	123-gnn	GC	CM	DTNN	MPNN	MoLFormer-XL	ChemBERTa	**k=1**	**k=2**
α	0.49	**0.27**	1.37	0.85	0.95	0.89	0.33	0.85	**0.26**	**0.26**
C_v	0.25	**0.09**	0.65	0.39	0.27	0.42	0.14	0.42	**0.11**	**0.11**
G	0.89	**0.05**	3.41	2.27	2.43	2.02	0.34	4.13	**0.084**	**0.084**
gap	0.0052	0.0048	0.01126	0.0086	0.0112	0.0066	**0.0038**	0.0052	**0.0037**	**0.0037**
H	0.89	**0.04**	3.41	2.27	2.43	2.02	0.25	4.08	**0.04**	**0.04**
ϵ_{homo}	0.0036	0.0034	0.0072	0.0051	0.0038	0.0054	**0.0029**	0.0045	**0.0028**	**0.0028**
ϵ_{lumo}	0.0041	0.0035	0.0092	0.0064	0.0051	0.0062	**0.0027**	0.0041	**0.0027**	**0.0027**
μ	0.451	0.476	0.583	0.519	**0.244**	**0.358**	0.3616	0.4659	0.369	0.369
$\langle R^2 \rangle$	26.84	22.90	35.97	46.00	**17.00**	28.5	17.06	86.15	17.1215	**16.88**
U_0	0.898	**0.0427**	3.41	2.27	2.43	2.05	0.3211	3.9811	**0.0435**	**0.0435**
U	0.89	**0.111**	3.41	2.27	2.43	2.00	0.25	4.38	**0.059**	**0.059**
ZPVE	0.00207	**0.00019**	0.00299	0.00207	0.0017	0.00216	**0.0003**	0.0023	**0.0003**	**0.0003**
Avg MAE	2.6355	1.9995	4.3536	4.7384	2.3504	3.1898	1.5894	8.7067	**1.5081**	**1.4879**

performance across all tested properties, as evidenced by the superior average performance metric.

This comprehensive evaluation not only emphasizes the effectiveness of the Mol-MVMoE approach in capturing the diverse aspects of molecular properties but also underscores the importance of a Multi-view Mixture-of-Experts as method to learn based on different perspectives in order to understand the intrinsically nuances of the data and demonstrate better performance across different challenging tasks.

In summary, the results presented for both classification and regression tasks underscore the exceptional capabilities of our proposed Mol-MVMoE approach, emphasizing its capacity to leverage different perspectives of the data for enhanced performance across a spectrum of complex tasks. Future research endeavors will be directed towards exploring different Mixture-of-Experts strategies and also the inclusion of more and larger models for enhanced predictions.

5. Conclusion

This paper presents Mol-MVMoE, a Multi-view Mixture-of-Experts framework that harnesses the synergies of distinct latent spaces derived from perspectives over the data to predict molecular properties. Evaluations on MoleculeNet benchmark datasets showcase the superiority of our proposed method. The proposed Mol-MVMoE outperforms state-of-the-art competitors on 9 out of 11 benchmark datasets emphasizing the robustness and adaptability of our model in addressing complex tasks within the molecular domain.

For the QM9 benchmark dataset, which encompasses quantum properties of molecules, the best version of our Mol-MVMoE approach surpass the current state-of-the-art models in 7 out of the 12 tested properties, suggesting that a one-size-fits-all approach might not be the most effective solution for such complex tasks.

By integrating embeddings from different natures, our model effectively captures nuanced structural features and intricate molecular interactions, leading to superior predictive performance. Future research directions will focus on exploring diverse fusion techniques for Mixture-of-Experts and incorporating more large scale models.

References

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

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

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

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

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

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

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

Horawalavithana, S., Ayton, E., Sharma, S., Howland, S., Subramanian, M., Vasquez, S., Cosbey, R., Glenski, M., and Volkova, S. 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*, pp. 160–172, 2022.

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

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

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

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

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

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

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

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

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

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4602–4609, 2019.

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

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Rupp, M., Tkatchenko, A., Müller, K.-R., and Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8 (1):13890, 2017.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

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

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

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

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Takeda, S., Kishimoto, A., Hamada, L., Nakano, D., and Smith, J. R. Foundation model for material science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15376–15383, 2023.

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

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.

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

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, 2018.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *ICLR*, 2019.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

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

Zhang, K. and Zhang, H. Predicting solute descriptors for organic chemicals by a deep neural network (dnn) using basic chemical structures and a surrogate metric. *Environmental Science & Technology*, 56(3):2054–2064, 2022.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: a universal 3d molecular representation learning framework. 2023.

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