A Multi-View approach based on Graphs and Chemical Language Foundation Model for Molecular Properties Prediction

Anonymous submission

Abstract

Pre-trained Language Models have emerged as promising tools for predicting molecular properties, yet their development is in its early stages, necessitating further research to enhance their efficacy and address challenges such as generalization and sample efficiency. In this paper, we present a Multi-View approach that combines latent spaces derived from state-of-the-art chemical models. Our approach relies on two pivotal elements: the embeddings derived from MHG-GNN, which represent molecular structures as graphs, and MoLFormer embeddings rooted in chemical language. The attention mechanism of MoLFormer is able to identify relations between two atoms even when their distance is far apart, while the GNN of MHG-GNN can more precisely capture relations among multiple atoms closely located. In this work, we demonstrate the superior performance of our proposed Multi-view approach compared to existing state-of-the-art methods, including MoLFormer-XL, which was trained on 1.1 billion molecules, particularly in intricate tasks such as predicting the quantum mechanical properties of small molecules. We assessed our approach using 11 benchmark datasets from MoleculeNet, where it outperformed competitors in 8 of them. We also provide a deep analysis of the results obtained with the QM9 dataset, where our proposed approach surpass its state-of-the-art competitors in 9 out of the 12 tasks presented in this dataset. Our study highlights the potential of latent space fusion and feature integration for advancing molecular property prediction. In this work, we use small versions of MHG-GNN and MoLFormer, which opens up an opportunity for further improvement when our approach uses a larger-scale dataset.

Introduction

Chemical-based machine learning has gained widespread adoption as an efficient and accurate approach for predicting molecular properties, owing to its capacity to effectively represent crucial structural aspects of molecules (Fang et al. 2022; Wieder et al. 2020; Shen and Nicolaou 2019). Recent advancements in foundational models have shown promising results by leveraging chemical language representations through a two-step process of pre-training on extensive unlabeled corpora and subsequent fine-tuning on specific downstream tasks of interest (Takeda et al. 2023; Soares et al. 2023a; Horawalavithana et al. 2022).

Despite the emergence of pre-trained Language Models as viable options for molecular property prediction (White 2023; Pan 2023; White et al. 2022; Janakarajan et al. 2023), they are still in their nascent stages of development. There is a pressing need for further research to enhance their performance and generate better embedding space representations (Frey et al. 2023).

Furthermore, recent discussions have emphasized the pivotal role of enhancing embedding representations through the addition of features from different natures/views to improve the overall quality of generated models (Li et al. 2023; Soares et al. 2023b). The incorporation of high-quality data and/or representations has the potential to advance the state-of-the-art (SOTA) (Liu, Ren, and Ren 2023). This strategic approach not only broadens the scope of information considered but also enriches the contextual understanding within models (Desikan and Evans 2022). By incorporating features from different natures and viewpoints, the aim is to create a more comprehensive and nuanced representation of the underlying data.

In this study, we introduce a Multi-view approach that leverages on the fusion of latent spaces from different natures generated by two state-of-the-art chemical-based models, namely MoLFormer-base (Ross et al. 2022) which is based on Transformers, and MHG-GNN a graph-based approach. Our approach is geared towards enhancing the prediction of molecular properties. Our findings demonstrate that our proposed method surpasses existing state-of-the-art algorithms, including the chemical language-based MoLFormer-XL, when it comes to tackling intricate tasks like predicting quantum mechanical calculations of small molecules. These challenging tasks are part of the MoleculeNet benchmark dataset (Wu et al. 2018). Furthermore, our approach exhibits superior performance in 8 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 Multi-view approach was able to perform better in 9 out of the 12 properties within the QM9 dataset when compared with other recent state-of-the-art approaches.

It is also important to highlight that the proposed approach refers to a fusion of latent spaces of models smaller models, and consistently performed better than other SOTA larger

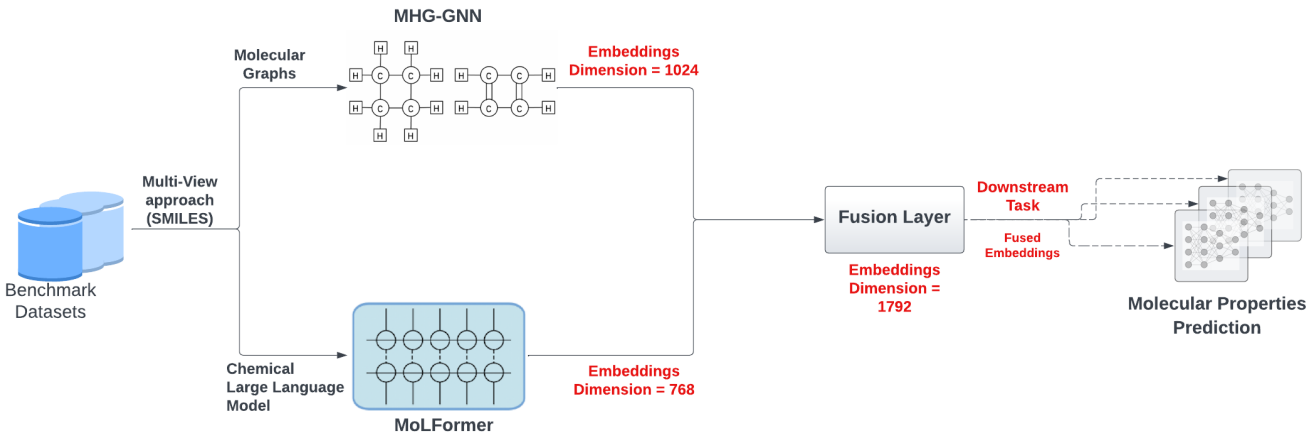

Figure 1: General architecture of the proposed Multi-view approach.

models, as the MoLFormer-XL which was trained on 1.1 billion molecules. By leveraging the fusion of latent spaces and feature sets, we have demonstrated a significant enhancement in performance that holds potential for advancing the field, it also opens up an opportunity for further improvement when our approach uses a larger-scale dataset.

Methodology

In this section, we explain the methodological framework delineated within this study. As depicted in Figure 1, we present an intricately devised schema for latent space fusion. Our approach relies on two pivotal elements: the embeddings derived from MHG-GNN, which represent molecular structures as graphs, and MoLFormer embeddings rooted in chemical language. This fusion of latent spaces is used for downstream tasks aiming at the prediction of molecular properties.

Our approach combines two orthogonal embeddings. A GNN architecture of MHG-GNN can more accurately capture molecular substructures than MoLFormer. On the other hand, a self-attention mechanism of MoLFormer has advantage of accounting for a relation between one atom to the other atoms even if their distances are larger than the radius of GNN.

We extract the embeddings for each SMILES contained in the dataset that we are exploring based on the pre-trained networks. For the MHG-GNN the embeddings space has the size of 1024, for MoLFormer-base the embeddings size is 768. Then, the resulting fused latent space has the size of 1792. Details of the employed models are described in the next subsections. XGBoost (Chen et al. 2015) with *optuna* (Akiba et al. 2019) optimizer was employed as predictor.

MHG-GNN

MHG-GNN (Kishimoto et al. 2023) 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:

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

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. 2:

$$h_G = \mathrm{CONCAT}\left(\left\{\sum_{i \in V_G} h_i^k \,\middle|\, k = 0, 1, \ldots, r\right\}\right) \tag{2}$$

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 3 (i.e., the iteration size for message passing step in GNN). With these configurations, MHG-GNN generates 1024 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.

MoLFormer

MoLFormer (Ross et al. 2022), 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). The MoLFormer vocabulary includes 2362 unique chemical tokens. These tokens are used to fine-tune or retrain the MolFormer model. To reduce computation time, the sequence length has been limited to a range of 202 tokens as 99.4% percent of all 1.1 billion molecules contain less than 202 tokens.

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 (3)).

$$\mathrm{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle \, v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle} \tag{3}$$

In Eq (3), $\mathrm{Attention}_m(Q,K,V)$ 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.

By leveraging the capabilities of MoLFormer and enhancing it with relative position embeddings, our approach offers an advanced and efficient solution for predicting complex molecular properties, providing valuable insights for various chemical applications. This enables the model to learn highly informative representations of the input data, making it a powerful tool for predicting molecular properties.

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.

It is important to highlight that for regression tasks, we also used a fine-tuned version of MoLFormer specialized for each of the tasks. Fine-tuning MoLFormer for each of the tasks has proved to improve the performance of the proposed Multi-view approach. 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

Fusion Layer

The fusion layer plays a pivotal role in amalgamating the embeddings obtained from both MHG-GNN, which adeptly represents molecular structures as graphs, and MoLFormer embeddings grounded in chemical language. Leveraging the unique strengths of each component, the attention mechanism of MoLFormer excels in discerning relationships between atoms, even when they are distantly positioned, while the GNN of MHG-GNN specializes in capturing intricate relations among closely situated multiple atoms. Specifically, the embeddings space within MHG-GNN boasts a size of 1024, whereas that of MoLFormer-base stands at 768. Consequently, the resultant fused latent space culminates in a size of 1792, synthesizing the nuanced insights derived from the complementary features of both models.

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

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 curated 6 datasets for the classification task and 5 datasets for regression tasks, encompassing a diverse array of objectives, ranging from the prediction of physical and biophysical properties to the characterization of physiological attributes of small-molecule chemicals. 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). This approach ensures the integrity of our evaluations and also enables a comprehensive and equitable comparison with existing methodologies.

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.

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.

Results

In this section, we present the analysis of the results obtained for the classification and regression tasks considered in this study, 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.

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 Multi-view approach.

Table 3 elucidates the consistent superiority of our fusion-based approach over the MHG-GNN and MoLFormer-Base methods in all conducted experiments. When compared with the MoLFomer-XL, our proposed multi-view approach obtained better results in 5 out of the 6 benchmark datasets tested. This pattern of across multiple datasets strongly suggests that the fusion of embeddings from different natures plays a pivotal role in enhancing the algorithm's performance.

Furthermore, it is worth emphasizing that the Multi-view 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 Multi-view approach led to a significant performance boost, elevating the model's performance from 43.2 to 80.5. Fig. 2 illustrates the ROC-AUC curve for the HIV, BACE, and BBBP datasets which are single-task. On these 3 tasks the Multi-view approach has demonstrated superior performance than the single model approaches.

Benchmark Tests with SOTA Methods

Results for classification tasks Table 4 offers a comprehensive overview of the comparative performance between our proposed Multi-view approach and state-of-the-art algorithms on various benchmark datasets for the classification task. A keen analysis of the table reveals that the Multi-view approach, which leverages the fusion of embeddings, outperforms its counterparts in 5 out of 6 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.

Table 3: Comparison between the multi-view approach and single models.

Method	Dataset					
	BBBP	ClinTox	HIV	BACE	SIDER	Tox21
MoLFormer-XL(Ross et al. 2022)	93.7	94.8	82.2	88.21	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
Multi-view approach	**94.2**	**98.8**	**86.1**	**90.4**	**69.9**	80.5

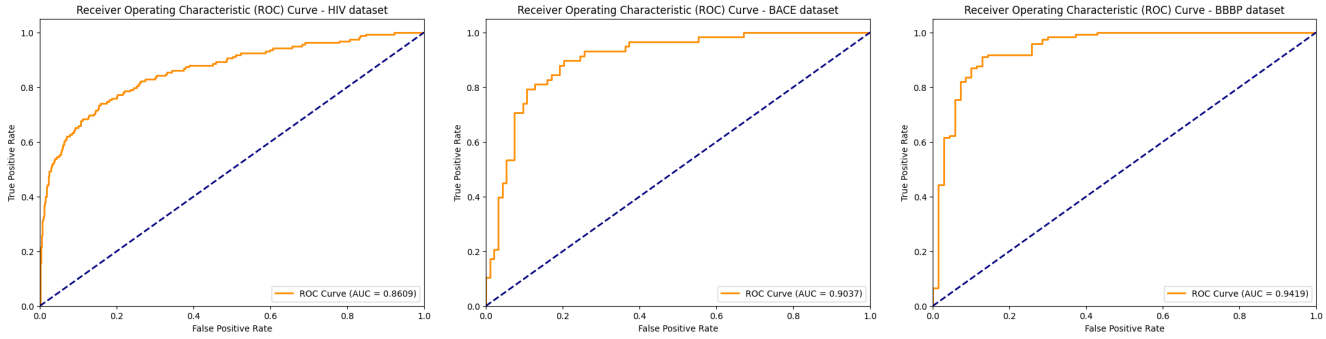

Figure 2: ROC-AUC curve for the single tasks HIV, BACE, and BBBP datasets.

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 approach harnesses the power of 768 embeddings from transformers-based MoLFormer-Base and 1024 embeddings from graph-based MHG-GNN, capitalizing on their complementary strengths to excel in a variety of challenging tasks. Results demonstrates that our proposed approach performs better than SOTA approaches as ChemBerta, Chemberta2, Galatica 30 and 120B, in all the experiments conducted. GraphMVP, presents the best result for the Tox21 tasks. However, our proposed methodology consistently presented better results in 5 out of the 6 benchmark datasets tested.

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 was pre-trained in a small portion of selected molecules from PubChem. The fusion of these smalls performed better than MoLFormer-XL which was trained in 1.1 billion molecules in 5 out of 6 benchmarks datasets. This not only highlights our method's effectiveness but also paves the way for additional enhancements when our approach leverages a larger-scale dataset.

Results for regression tasks Next, we applied the proposed Multi-view approach 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 importance of fine-tuning MoLFormer for these challenging tasks. With the fine-tunned MoLFormer, our Multi-view 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 Multi-view approach with fine-tuned MoLFormer has showcased significant advancements in performance, particularly in the demanding QM9 dataset. This amalgamation of diverse data modalities has enabled our model to unravel the intricate relationships between molecular structures and the corresponding quantum chemical properties with greater precision and depth.

To fine-tune the MoLFormer approach was crucial for the improved performance of our Multi-view approach. The best version of Multiview approach has demonstrated better performance than state-of-the-art models as Uni-Mol, ChemRL-GEM, SPMM, and larger models as ChemBERTa-2 and $GROVER_{Large}$.

Furthermore, the Multi-view approach (fine-tuned MoLFormer) 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

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, Groß, and Günnemann 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, Demirel, and Liang 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, Grand, and Ramsundar 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
MoLFormer-XL(Ross et al. 2022)	93.7	94.8	82.2	88.2	69.0	84.7
Multi-view approach	**94.2**	**98.8**	**86.1**	**90.4**	**69.9**	80.5

Table 5: Methods and Performance for the classification 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 and 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 and 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 approach - (Frozen Weights)	4.48	0.0128	0.69	1.47	0.61
Multi-view approach - (Fine-tuned MoLFormer)	**1.45**	**0.00961**	0.60	1.43	**0.51**

promising results obtained by our Multi-view strategy across various regression tasks underline its potential in the domain of chemical property prediction.

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 7.

Within this study, we compare the best version and standard version of our Multi-view approach against a selection of previously discussed baseline models, as well as four additional baselines. Our comparative analysis extends to benchmarking the 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, Grand, and Ramsundar 2020) approach in this study.

Table 6 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 approach outperforms the current models in 9 out of the 12 properties in tis best version. Notably, it also achieves the second-best performance in two properties, specifically C_v and ZPVE.

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

Table 6: 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		Multi-view based	
Measure	A-FP	123-gnn	GC	CM	DTNN	MPNN	MoLFormer-XL	ChemBERTa	**(Frozen Weights)**	**(Fine-tuned)**
α	0.49	**0.27**	1.37	0.85	0.95	0.89	0.33	0.85	0.96	**0.26**
C_v	0.25	**0.09**	0.65	0.39	0.27	0.42	0.14	0.42	0.44	**0.11**
G	0.89	**0.05**	3.41	2.27	2.43	2.02	0.34	4.13	2.63	**0.03**
gap	0.0052	0.0048	0.01126	0.0086	0.0112	0.0066	**0.0038**	0.0052	0.0061	**0.0036**
H	0.89	**0.04**	3.41	2.27	2.43	2.02	0.25	4.08	2.68	**0.02**
ϵ_{homo}	0.0036	0.0034	0.0072	0.0051	0.0038	0.0054	**0.0029**	0.0044	0.0045	**0.0028**
ϵ_{lumo}	0.0041	0.0035	0.0092	0.0064	0.0051	0.0062	**0.0027**	0.0041	0.0046	**0.0026**
μ	0.451	0.476	0.583	0.519	**0.244**	**0.358**	0.3616	0.4659	0.518	0.369
$\langle R^2 \rangle$	26.84	22.90	35.97	46.00	**17.00**	28.5	17.06	86.15	41.21	**16.52**
U_0	0.898	**0.0427**	3.41	2.27	2.43	2.05	0.3211	3.9811	2.6389	**0.0192**
U	0.89	**0.111**	3.41	2.27	2.43	2.00	0.25	4.38	2.69	**0.031**
ZPVE	0.00207	**0.00019**	0.00299	0.00207	0.0017	0.00216	0.0003	0.0023	0.00112	**0.0002**
Avg MAE	2.6355	1.9995	4.3536	4.7384	2.3504	3.1898	**1.5894**	8.7067	4.4837	**1.4485**

Table 7: 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

and SMILES-based approaches. This underscores the importance of choosing the appropriate methodology based on the specific property under consideration.

Our proposed Multi-view approach, which combines graph and transformer-based features, consistently demonstrates superior performance compared to other methods. This highlights the potential benefits of leveraging a diverse set of features 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 Multi-view embeddings has exhibited robust and consistent performance across all tested properties, as evidenced by the superior average performance metric.

This comprehensive evaluation not only emphasizes the effectiveness of the Multi-view approach in capturing the diverse aspects of molecular properties but also underscores the importance of choosing a modeling strategy tailored to the specific property under consideration.

In summary, the results presented for both classification and regression tasks underscore the exceptional capabilities of our proposed Multi-view approach, emphasizing its capacity to leverage diverse features for enhanced performance across a spectrum of complex tasks. Future research endeavors will be directed towards exploring various fusion strategies and integrating higher-quality features and embeddings, aiming to further refine and optimize our approach, thereby advancing the boundaries of predictive modeling in the field of chemical research.

Conclusion

This paper introduces a novel Multi-view approach that leverages the complementary latent spaces of two state-of-the-art algorithms, MoLFormer-base and MHG-GNN, for predicting molecular properties. Through extensive evaluations on the MoleculeNet dataset, our proposed method has demonstrated superior performance across various classification and regression tasks, outperforming the state-of-the-art competitors on 8 out of 11 benchmark datasets. This remarkable consistency highlights the robustness and versatility of our proposed approach.

It is also important to highlight that for the improved performance for the regression tasks we need to fine-tune the MoLFormer model for each of the tasks. With the fine-tuning our proposed Multi-view approach has demonstrated better results in 3 out 5 these challenging tasks when compared to recent state-of-the-art approaches, including very large models.

For the challenging QM9 benchmark dataset, which encompasses intricate quantum properties of molecules, the best version of our Multi-view approach excels, surpassing the current state-of-the-art models in 9 out of the 12 properties. Notably, it achieves the second-best performance in two properties, namely C_v and ZPVE. The performance of our Multi-view approach underscores the importance of combining features from different sources to achieve enhanced molecular properties prediction.

By integrating graph-based embeddings and language model representations, our model effectively captures nuanced structural features and intricate molecular interactions, leading to superior predictive performance. Future re-

search directions will focus on exploring diverse fusion techniques and incorporating high-quality features and embeddings to further refine our approach. These findings pave the way for promising advancements in the accuracy and efficacy of molecular properties prediction based on *in silico* modelling.

References

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

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

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

Chang, J.; and Ye, J. C. 2023. Bidirectional Generation of Structure and Properties Through a Single Molecular Foundation Model.

Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4): 1–4.

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

Desikan, B. S.; and Evans, J. 2022. Aggregate, integrate and align to embed everything: a multi-modal framework for measuring cultural dynamics. In *Cultures in AI/AI in Culture, NeurIPS 2022 Workshop; 2022 Dec 9; online*.

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

Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C. W.; and Gadepally, V. 2023. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 1–9.

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

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

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

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

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020. Strategies for Pre-training Graph Neural Networks. In *ICRL*.

Janakarajan, N.; Erdmann, T.; Swaminathan, S.; Laino, T.; and Born, J. 2023. Language models in molecular discovery. *arXiv preprint arXiv:2309.16235*.

Kajino, H. 2019. Molecular Hypergraph Grammar with Its Application to Molecular Optimization. In *ICML*, 3183–3191. 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. 2023. MHG-GNN: Combination of Molecular Hypergraph Grammar with Graph Neural Network. arXiv:2309.16374.

Li, C.; Mao, K.; Wang, S.; Yuan, Y.; and Wang, G. 2023. Self-Supervised Graph Information Bottleneck for Multi-View Molecular Embedding Learning. *IEEE Transactions on Artificial Intelligence*.

Liu, P.; Ren, Y.; and Ren, Z. 2023. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *arXiv preprint arXiv:2308.06911*.

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

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

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

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

Pan, J. 2023. Large language model for molecular chemistry. *Nature Computational Science*, 3(1): 5–5.

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

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

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

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

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

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

Soares, E.; Brazil, E. V.; Gutierrez, K. F. A.; Cerqueira, R.; Sanders, D.; Schmidt, K.; and Zubarev, D. 2023a. Beyond Chemical Language: A Multimodal Approach to Enhance Molecular Property Prediction. *arXiv preprint arXiv:2306.14919*.

Soares, E.; Sharma, V.; Brazil, E. V.; Cerqueira, R.; and Na, Y.-H. 2023b. Capturing Formulation Design of Battery Electrolytes with Chemical Large Language Model. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*.

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

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

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

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

White, A. D. 2023. The future of chemistry is language. *Nature Reviews Chemistry*, 1–2.

White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; et al. 2022. Do large language models know chemistry?

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

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.

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

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful Are Graph Neural Networks? In *ICLR*.

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

Zhang, K.; and Zhang, H. 2022. 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.

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