A Deep Multimodal Representation Learning Framework for Accurate Molecular Properties Prediction

Yuxin Yang
Cleveland Clinic
Cleveland, Ohio, USA
yangy9@ccf.org

Zixu Wang
Kent State University
Kent, Ohio, USA
zwang55@kent.edu

Pegah Ahadian
Kent State University
Kent, Ohio, USA
pahadian@kent.edu

Abby Jerger
Pacific Northwest National Laboratory
Seattle, Washington, USA
abby.jerger@pnnl.org

Jeremy Zucker
Pacific Northwest National Laboratory
Richland, Washington, USA
jeremy.zucker@pnnl.org

Song Feng
Pacific Northwest National Laboratory
Richland, Washington, USA
song.feng@pnnl.org

Feixiong Cheng
Cleveland Clinic
Cleveland, Ohio, USA
chengf@ccf.org

Qiang Guan
Kent State University
Kent, Ohio, USA
qguan@kent.edu

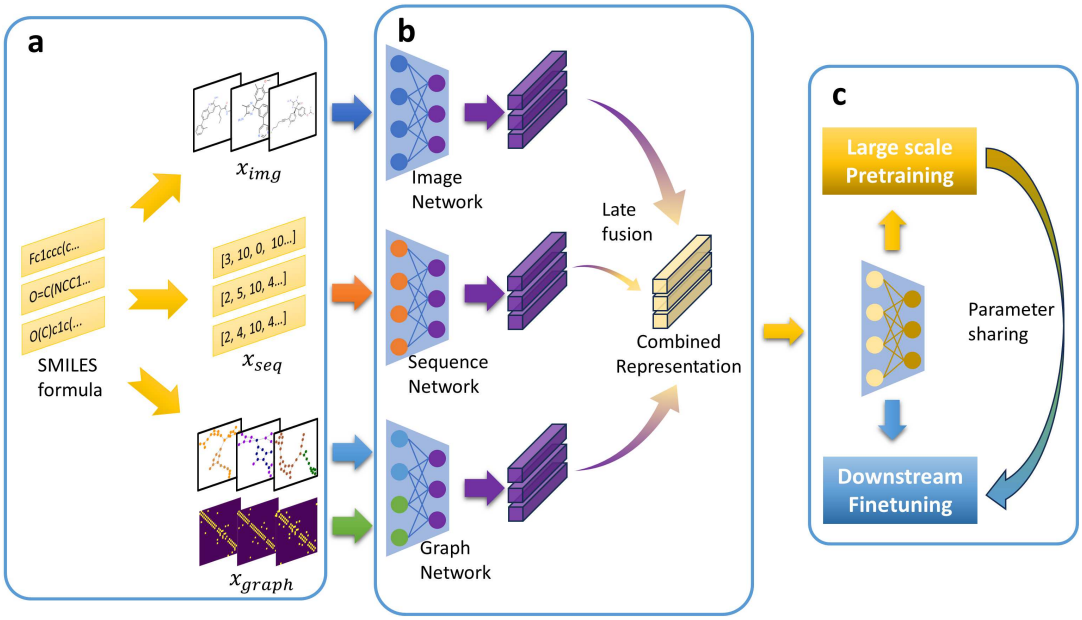

Figure 1: Schematic illustration of MRL-Mol framework. a. Data Processing. b. Modality Integration. c. Learning and Prediction.

ABSTRACT

Drug discovery is a challenging process, requiring the optimization of compounds to become safe and effective. Predicting molecular properties is an indispensable step in the drug discovery pipeline. Traditionally, this process is costly, involving multiple rounds of experiments, rendering it impractical for every candidate compound. Deep learning techniques have emerged as a promising approach to drug discovery to reduce the cost during the process. However, prevalent research in deep learning models focused on predicting molecular properties has primarily fixated on single-modal models, neglecting the potential benefits of combining different data modalities. To overcome this limitation, we introduce MRL-Mol: a deep Multimodal Representation Learning framework for accurate Molecular properties prediction. MRL-Mol harnesses three data modalities: sequence, graph, and image, augmenting the depth of

GLSVLSI '24, June 12–14, 2024, Clearwater, FL, USA

ACM ISBN 979-8-4007-0605-9/24/06
https://doi.org/10.1145/3649476.3660377

comprehension. Leveraging a large-scale unlabeled dataset (1M unique molecules), we pretrain MRL-Mol to extract inter- and intra-modal information. Our study demonstrates the superior performance of MRL-Mol in predicting molecular properties across six benchmark datasets. Notably, MRL-Mol outperforms other state-of-the-art molecular properties prediction models. These findings suggest that by combining information from multiple data modalities, MRL-Mol can comprehend molecules better than single-modal deep learning models and identify molecular properties with better accuracy.

CCS CONCEPTS

• **Applied computing → Computational biology**.

KEYWORDS

deep learning, multimodal learning, drug discovery

ACM Reference Format:

Yuxin Yang, Zixu Wang, Pegah Ahadian, Abby Jerger, Jeremy Zucker, Song Feng, Feixiong Cheng, and Qiang Guan. 2024. A Deep Multimodal Representation Learning Framework for Accurate Molecular Properties Prediction. In *Great Lakes Symposium on VLSI 2024 (GLSVLSI '24), June 12–14, 2024, Clearwater, FL, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649476.3660377

1 INTRODUCTION

Drug discovery persists as a formidable challenge in the realms of biology and medicine, marked by the costly, slow, and often unsuccessful conventional process [1]. One of the fundamental challenges is to accurately predict molecular properties. However, the landscape of drug discovery is undergoing a transformation, driven by substantial advancements in data analysis algorithms and computational capabilities. These strides have led to a new era of drug discovery, where the convergence of data science and molecular biology holds the promise of more efficient and effective approaches. Over the past decade, a significant turning point has been the accessibility of extensive compound data repositories, exemplified by the ChemBL database [19] and the PubChem database [13]. These vast datasets of chemical information have set the stage for the development of several deep learning techniques tailored to accurately predict molecular properties. These techniques, driven by the fusion of computational power and data, have begun to yield promising results, offering a glimpse into the potential of data-driven drug discovery.

Presently, most deep learning models for molecular properties prediction are engineered to harness a single data modality, such as sequence data (derived from compound's simplified molecular-input line-entry system (SMILES) formulas [25]), graph data (sourced from compound structures), or image data (derived from 2D snapshots of molecular structures). For instance, in the study outlined in [24], a graph neural network undergoes self-supervised learning to acquire molecular representations exclusively from molecular graphs. Similarly, other investigations, exemplified by [10, 12], advocate the use of compound SMILES formulas as input, employing BERT [5] which is based on transformer encoders [23] to process the sequence-based information. In [28], compound structures are encoded into images, and self-supervised learning techniques are employed to extract crucial molecular features. Additionally, [27] innovatively merges information from both compounds and protein targets, offering valuable insights into leveraging information from diverse data sources. While these individual works have undoubtedly enriched our understanding of molecular properties prediction and drug discovery, there is a growing recognition of the necessity to break free from the constraints of a single data modality. The convergence of sequence data, graph data, and image data, each offering unique insights into molecular properties, has the potential to revolutionize the field. This synergy among diverse data sources holds the promise of unraveling previously hidden patterns and relationships between molecular structures, offering a more accurate prediction of molecular properties.

In this work, we present MRL-Mol, a deep multimodal representation learning framework for molecular properties prediction. MRL-Mol integrates three modality networks specifically designed for graph data, image data, and sequence data, enabling the extraction of information-rich representations from these compound data modalities. Through these efforts, our goal is to elevate the precision and efficiency of molecular properties prediction, thereby contributing to the ongoing evolution of this critical field at the intersection of biology and technology.

We summarize the advancements of MRL-Mol over current state-of-the-art methods as follows:

- MRL-Mol integrates three modality networks, each designed for a specific data modality.
- Unlike current state-of-the-art methods using unsupervised pretraining or self-supervised pretraining, MRL-Mol employs a supervised pretraining method with pseudolabels generated from K-Means clustering [8] of MACCS fingerprints [2, 6] of molecules. This supervised pretraining procedure facilitates MRL-Mol in acquiring valuable insights into the features and structures of molecules.
- MRL-Mol achieves superior performance compared to state-of-the-art methods despite being pretrained on a smaller dataset.

In the following sections, we first delve into relevant prior research and existing literature in Section 2. Following that, in Section 3, we pivot towards the core contribution, meticulously detailing MRL-Mol's development and datasets employed for training and validation. In Section 4, we further provide a thorough presentation of all conducted numerical tests and corresponding results. Finally, Section 5 offers further discussion and draws conclusions.

2 RELATED WORKS

In this section, we offer a concise overview of multimodal deep learning, providing essential context for the subsequent discussions.

2.1 Multimodal deep learning

Multimodal deep learning is a powerful approach aimed at processing information by learning from diverse data modalities. This technique encompasses two primary strategies for fusing different modalities of data, known as early fusion and late fusion, distinguished by when the fusion of modalities takes place [21].

Early fusion in multimodal models involves the integration of various data types at the input level. Typically, early fusion techniques commence by projecting distinct modalities into a common lower-dimensional space, often using methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA). For instance, in the context of textual-visual sentiment analysis [3], an early fusion approach is employed to combine textual and visual features.

Conversely, late fusion multimodal models leverage separate sub-models, each tailored to handle a specific modality. These sub-models generate feature representations for their respective modalities, and the combination of these feature representations occurs just before the final prediction. In an example [11], visual and audio features are independently generated by separate models and then combined prior to the ultimate recognition of emotions.

Each of these studies offers valuable insights into the design of a multimodal deep learning model that combines three distinct modalities for drug discovery.

3 METHODOLOGY

Our proposed MRL-Mol framework encompasses three key phases (See Fig. 1): 1. Data processing, which converts input SMILES formulas to molecular images, sequences, and graphs. 2. Modality integration, which uses three modality networks to extract information-rich representations from each data modality and integrate them using a late-fusion manner. And 3. Learning and prediction, which leverages a large-scale pretraining dataset to enable MRL-Mol to learn useful patterns from diverse data modalities, and performs downstream finetuning to predict molecular properties.

3.1 Data Processing

We begin by delving into the technical details of the data processing phase, wherein SMILES formulas of input molecules are transformed into each modality (See Fig. 1a). Given a SMILES formula s, we use RDKit [16] to obtain the 2D molecular structure. From this, the molecular image x_{img} is generated by drawing atoms and chemical bonds based on the provided 2D coordinates. Simultaneously, the molecular graph x_{graph} is constructed by encoding atoms into nodes and connecting atoms via edges within the graph. The molecular sequence x_{seq} is derived by tokenizing the SMILES formula and encoding it into integers through embedding.

3.2 Modality Integration

We next turn our attention to the technical details of the modality integration phase, where the three modality networks learn representations from the three data modalities and merge these representations using late-fusion techniques (See Fig. 1b).

3.2.1 Image Network. The Image Network, built on Residual Networks (ResNet) [9], acts as a feature extractor $\mathcal{R}_\Phi$, mapping molecular images x_{img} to image representations $r_{img} \in \mathbb{R}^d$:

$$r_{img} = \mathcal{R}_\Phi(x_{img}), \tag{1}$$

where Φ is the trainable parameters of $\mathcal{R}$ and d is the dimensionality of the r_{img}.

3.2.2 Sequence Network. The Sequence Network, based on BERT [5] comprising transformer encoders [23], serves as a feature extractor $\mathcal{S}_\Psi$ transforming molecular sequence data x_{seq} into sequence representations $r_{seq} \in \mathbb{R}^d$:

$$r_{seq} = \mathcal{S}_\Psi(x_{seq}), \tag{2}$$

where Ψ is the trainable parameters of $\mathcal{S}$.

3.2.3 Graph Network. The Graph Network, consisting of a series of graph convolutional layers [14], maps molecular graphs x_{graph} to graph representations $r_{graph} \in \mathbb{R}^d$:

$$r_{graph} = \mathcal{G}_\Theta(x_{graph}), \tag{3}$$

where $\mathcal{G}_\Theta$ is the Graph Network with trainable parameters Θ.

3.2.4 Late Fusion. The representations from the three modality networks are integrated in a late-fusion manner:

$$r = r_{img} + r_{seq} + r_{graph}, \tag{4}$$

where $r \in \mathbb{R}^d$ is the combined representation with a dimension of d.

3.3 Learning and Prediction

To enable MRL-Mol to grasp inter- and intra-modality information, extract information-rich representations, and make accurate predictions, we adopt a pretraining-finetuning strategy (See Fig. 1c).

3.3.1 Large-scale Dataset Pretraining. The initial step involves pretraining MRL-Mol in a supervised fashion. Considering the absence of label information within the pretraining dataset, each molecule is assigned a pseudolabel based on the methodology outlined in [28]. Initially, MACCS fingerprints [2, 6] are extracted from all molecules in the pretraining dataset. Subsequently, employing K-Means clustering [8], the molecules are grouped into k clusters based on their MACCS fingerprints. Each molecule is then assigned the cluster index as its pseudolabel. MRL-Mol is then pretrained using these pseudolabels. To facilitate the prediction of pseudolabels, a fully connected projection head $\mathcal{F}_\Omega(\cdot)$ is appended after the combined representation. We then utilize Cross Entropy (CE) loss function with backpropagation to optimize MRL-Mol:

$$\begin{aligned} l_{CE}(u_i, \hat{u}_i) &= -\sum_{c=1}^{C} \log \frac{e^{\hat{u}_{i,c}}}{\sum_{j=1}^{C} e^{\hat{u}_{i,j}}} u_{i,c} \\ &= -\sum_{c=1}^{C} \log \frac{e^{\mathcal{F}_\Omega(r_{i,c})}}{\sum_{j=1}^{C} e^{\mathcal{F}_\Omega(r_{i,j})}} u_{i,c}, \end{aligned} \tag{5}$$

where $\hat{u}_{i,c}$ is the output of $\mathcal{F}_\Omega(\cdot)$ of i^{th} sample of cluster c and $u_{i,c}$ is the ground truth pseudolabel of i^{th} sample of cluster c.

3.3.2 Downstream Finetuning. Utilizing the pretrained MRL-Mol, we commence by substituting the original projection head with a new fully-connected prediction head $\mathcal{H}_\Gamma(\cdot)$ encompassing trainable parameters Γ. This prediction head is responsible for predicting the molecular property based on the combined representations:

$$\hat{y} = \mathcal{H}_\Gamma(r), \tag{6}$$

where $\hat{y}$ is the predicted molecular property.

Dataset	BACE	BBBP	ClinTox
# Molecules	1,513	2,039	1,478
CHEM-BERT [12]	0.8102	0.8240	0.9292
ImageMol [28]	0.7489	0.8743	0.9133
MolCLR [24]	0.7762	0.8099	0.9304
MRL-Mol (Ours)	**0.8353**	**0.9067**	**0.9527**

Table 1: Predictive performance of different models conducted on classification benchmark datasets measured by Aera Under Receiver Operating Characteristic (AUROC).

Dataset	ESOL	FreeSolv	Lipo
# Molecules	1,128	642	4,200
CHEM-BERT [12]	0.3486	0.3217	0.4019
ImageMol [28]	0.3618	0.3596	0.4068
MolCLR [24]	0.4829	0.5198	0.3972
MRL-Mol (Ours)	**0.3220**	**0.3040**	**0.3212**

Table 2: Predictive performance of different models conducted on regression benchmark datasets measured by Root Mean Squared Error (RMSE).

For optimizing the trainable parameters, in regression tasks, we employ the Mean Squared Error (MSE) loss function with backpropagation:

$$l_{MSE}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \tag{7}$$

where y_i is the ground truth molecular property of the i^{th} sample within the training dataset. In classification tasks, we continue using the Cross Entropy loss function with backpropagation to optimize the trainable parameters.

To maximize the benefits derived from the pretrained MRL-Mol, we follow the methodology outlined in [15] to finetune MRL-Mol on downstream datasets. This involves linear probing of the prediction head, followed by the subsequent finetuning of the entire model.

3.4 Dataset

We now delve into the details of the pretraining and downstream benchmark datasets, serving as the foundation for training and validating our MRL-Mol.

3.4.1 Pretraining dataset. Our pretraining dataset comprises selected ~1 million unique unlabeled drug-like and bioactive molecules' SMILES data sourced from PubChem database [13]. To process this dataset, we utilize RDKit [16] for converting SMILES data to molecule images and constructing molecule graphs.

3.4.2 Downstream datasets. For our downstream experiments, we select six datasets from MoleculeNet [26]. These datasets encompass various predictive tasks: (1) Molecular Target Prediction, focusing on the human β-secreatse 1 (BACE-1) target [22], (2) Blood-Brain Barrier Penetration (BBBP) prediction, predicting the permeability of molecules through the blood-brain barrier [18], (3) Molecular Toxicity Prediction, particularly targeting clinical trial toxicity (ClinTox) [7], (4) Molecular Solubility Prediction, including estimated solubility (ESOL) [4], free solvation (FreeSolv) [20], and lipophilicity (Lipo).

4 EXPERIMENTS

4.1 Experiment Setup

In the pretraining step, 95% of the molecules (9,499,921 molecules) from the pretraining dataset are utilized to train MRL-Mol, while the remaining 5% (499,996 molecules) serve as the validation dataset for hyperparameter selection purposes. We set the number of clusters k as 100. Each downstream benchmark dataset is randomly divided into three subsets, training, validation, and test, using an 80%/10%/10% ratio. Labels of regression benchmark datasets are normalized to the range of $[-1, 1]$:

$$y_{norm} = \left(\frac{y - \min(y)}{\max(y) - \min(y)} - 0.5 \right) \times 2, \tag{8}$$

where y_{norm} is the normalized regression label and y is the original regression label.

Both pretraining and finetuning procedures utilize the AdamW optimizer [17] with an initial learning rate of 1×10^{-5} and a weight decay coefficient of 1×10^{-3}. The pretraining step involves training the model for 50 epochs until convergence, while the downstream finetuning consists of a linear probing phase for 30 epochs followed by an additional 30 epochs of finetuning to reach convergence.

In our benchmark analysis, we compare the performance of MRL-Mol against three state-of-the-art single-modal molecular properties prediction models: a sequence-based model (CHEM-BERT [12]), a graph-based model (MolCLR [24]), and an image-based model (ImageMol [28]). Each baseline model was pretrained using a larger pretraining dataset comprising ~10M unique molecules. We conduct finetuning on the downstream benchmark datasets and adhere to their original training setups.

Evaluation metrics for classification datasets encompass the Aera Under Receiver Operating Characteristic (AUROC). For regression datasets, we measure Pearson's correlation coefficient R (Pearson R) and Root Mean Squared Error (RMSE) to evaluate performance.

4.2 Benchmark evaluation

We commence our assessment by benchmarking the performance of MRL-Mol against three baseline models across six downstream datasets. In classification tasks, MRL-Mol achieves remarkable results in terms of the AUROC metric on BACE (AUROC: 0.8353), BBBP (AUROC: 0.9067), and ClinTox (AUROC: 0.9527) datasets (See Table 1). Compared to the three baselines, MRL-Mol consistently outperforms them, showcasing an average improvement of 6.354% on AUROC. Notably, MRL-Mol achieves an enhancement of 11.952% at best when compared to MolCLR on the BBBP dataset (See Table 1).

In regression tasks, MRL-Mol exhibits robust predictive performance and low predictive error across ESOL (Pearson R: 0.6334 and RMSE: 0.3220), FreeSolv (Pearson R: 0.7287 and RMSE: 0.3040), and Lipo (Pearson R: 0.6551 and RMSE: 0.3212) datasets (See Fig. 2 and Table 2). Compared to the three baseline models, MRL-Mol consistently showcases superior performance, with an average improvement of ~20% improvement in RMSE. Particularly noteworthy is the 41.52% improvement over MolCLR on the FreeSolv dataset (See Fig. **??**b). These performance enhancements indicate the superiority of MRL-Mol, especially considering its utilization of a pretraining

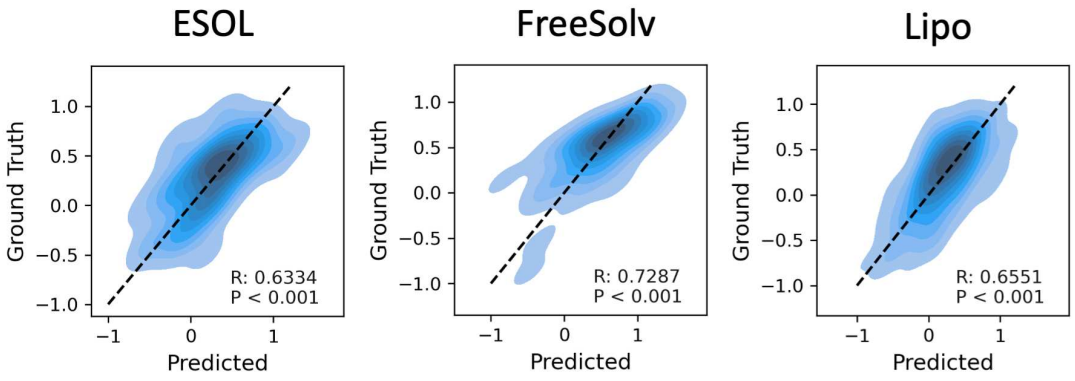

Figure 2: Predictive performance of our proposed MRL-Mol on the three regression benchmark datasets. Predicted label and ground truth label of each compound for each dataset are contour plotted with point density. Pearson's correlation coefficient (R) and P values are labeled.

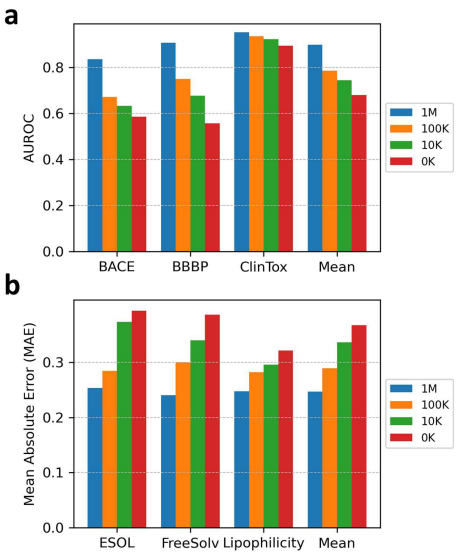

Figure 3: Ablation studies on the impact of pretraining dataset scale on the performance of MRL-Mol. The Means denote the average performance of MRL-Mol on benchmark datasets concerning different pretraining dataset sizes. a. Classification results measured by AUROC, and b. regression results measured by MAE.

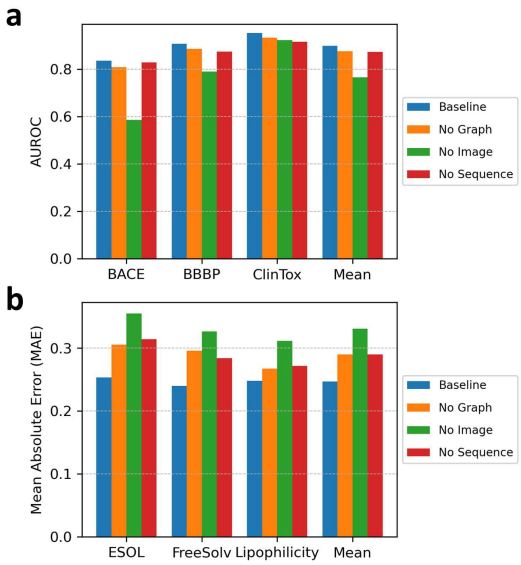

Figure 4: Ablation studies on the impact of each modality network in MRL-Mol. 'Baseline' refers to MRL-Mol without any modality network removed. 'No Graph' refers to removing the Graph Network. 'No Image' refers to removing the Image Network. 'No Sequence' represents MRL-Mol removing the Sequence Network. 'Mean' denotes the average performance of MRL-Mol across benchmark datasets with specific networks removed. a. Classification results measured by AUROC, and b. regression results measured by MAE.

dataset roughly one-tenth the size of the pretraining datasets used in the three baseline models (~1M vs. ~10M).

4.3 Interpreting the multimodal deep learning model

To delve deeper into understanding the impact of the pretraining process and individual modality networks on the predictive performance of MRL-Mol, we trained separate models with various sizes of pretraining datasets and removed one of the three modality networks. For the ablation study involving pretraining with different dataset sizes, we randomly sampled 10K and 100K unique molecules from our 1M pretraining dataset. Following the identical pretraining procedure for the MRL-Mol on the 1M dataset, we trained MRL-Mol using these smaller pretraining datasets. Subsequently, we benchmarked these pretrained models on six downstream benchmark datasets.

As shown in Fig. 3, the predictive performance of MRL-Mol elevates with the increase in the size of the pretraining dataset. The MRL-Mol trained on the largest pretraining dataset (~1M molecules)

outperforms the MRL-Mol trained on a medium-sized pretraining dataset (~100K molecules) by an average of 14.362% in terms of AUROC (0.8982 vs. 0.7854) and 14.523% in terms of MAE (0.2472 vs. 0.2892). Furthermore, it surpasses the performance of the MRL-Mol trained on the smallest pretraining dataset (~10K molecules) by an average of 20.710% in terms of AUROC (0.8982 vs. 0.7441) and an average of 26.538% in terms of MAE (0.2472 vs. 0.3365). Additionally, it outperforms the MRL-Mol without any pretraining by an average of 32.263% in terms of AUROC (0.8982 vs. 0.6791) and an average of 32.716% in terms of MAE (0.2472 vs. 0.3674). These results validate that the pretraining process and a larger pretraining dataset significantly enhance the performance and generalization ability of MRL-Mol.

Furthermore, we pretrained MRL-Mol by removing one of the three modality networks while using the ~1M pretraining dataset. Subsequently, we benchmarked these six downstream datasets. As shown in Fig. 4, we observed that MRL-Mol achieves optimal performance when all three modality networks are included during pretraining. This ablation study confirms that each of the three data modalities and their respective modality networks significantly contribute to the final prediction of molecular properties.

5 CONCLUSION

In this study, we introduced MRL-Mol, a robust multimodal deep learning framework tailored for predicting molecular properties. MRL-Mol integrates three diverse data modalities, image-based, sequence-based, and graph-based, each corresponding to specialized models. Leveraging a large-scale molecule dataset along with MACCS fingerprints, we pretrained MRL-Mol and conducted fine-tuning for downstream molecular properties prediction tasks. Our thorough evaluation assessed the performance of MRL-Mol across six diverse molecular properties prediction datasets encompassing both classification and regression tasks, showcasing its superiority over state-of-the-art methods in image-based, sequence-based, and graph-based molecular property prediction. Through comprehensive comparisons and analysis, our findings definitively demonstrate the substantial enhancement achieved by integrating these three data substantially. This innovative approach presents a significant stride in advancing the domain of drug discovery by leveraging the synergy of multiple data modalities.

ACKNOWLEDGMENTS

This material was supported by the National Science Foundation (NSF) under Grant #2212465, #2217021, #2217104, #2230111, #2238734, and #2311950.

REFERENCES

[1] John Arrowsmith and Philip Miller. 2013. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature reviews. Drug discovery* 12, 8 (2013), 569.
[2] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* 71 (2015), 58–63.
[3] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 163–171.
[4] John S Delaney. 2004. ESOL: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences* 44, 3 (2004), 1000–1005.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[6] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* 42, 6 (2002), 1273–1280.
[7] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology* 23, 10 (2016), 1294–1301.
[8] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[10] Shion Honda, Shoi Shi, and Hiroki R Ueda. 2019. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* (2019).
[11] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10 (2016), 99–111.
[12] Hyunseob Kim, Jeongcheol Lee, Sunil Ahn, and Jongsuk Ruth Lee. 2021. A merged molecular representation learning for molecular properties prediction with a web-based service. *Scientific Reports* 11, 1 (2021), 11028.
[13] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* 47, D1 (2019), D1102–D1109.
[14] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[15] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054* (2022).
[16] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8 (2013), 31.
[17] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
[18] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. 2012. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling* 52, 6 (2012), 1686–1697.
[19] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* 47, D1 (2019), D930–D940.
[20] David L Mobley and J Peter Guthrie. 2014. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* 28 (2014), 711–720.
[21] Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34, 6 (2017), 96–108.
[22] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling* 56, 10 (2016), 1936–1949.
[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[24] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* 4, 3 (2022), 279–287.
[25] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28, 1 (1988), 31–36.
[26] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.
[27] Yuxin Yang, Yunguang Qiu, Jianying Hu, Michal Rosen-Zvi, Qiang Guan, and Feixiong Cheng. [n. d.]. A Deep Learning Framework Streamlines Computational Drug Discovery Via Combining Molecular Image and Protein Structural Representation. *Available at SSRN 4710836* ([n. d.]).
[28] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. 2022. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* 4, 11 (2022), 1004–1016.