2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC) | 979-8-3503-7696-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/COMPSAC61105.2024.00317

Three-branch Molecular Representation Learning Framework for Predicting Molecular Properties in Drug Discovery

Yu Liu[†‡], Lihui Duo[⋆‡], Jonathan D. Hirst[*], Jianfeng Ren[†], Bencan Tang[⋆], Dave Towey[†]
[†]*School of Computer Science, University of Nottingham Ningbo China*
[⋆]*School of Chemistry Engineering, University of Nottingham Ningbo China*
[*]*School of Chemistry, University of Nottingham*
Contact: Jianfeng.Ren@nottingham.edu.cn, Bencan.Tang@nottingham.edu.cn, Dave.Towey@nottingham.edu.cn

***Abstract*—Graph Neural Networks (GNNs) have been widely used to model molecules with a graph representation. However, GNNs face inherent challenges in accurately modeling long-range atomic interactions and identifying complex molecular substructures. This research proposes a novel Three-branch Molecular Representation Learning Framework (TMRLF) for predicting molecular properties: it integrates one branch of a GNN that extracts local molecular structural information with two branches of fully connected networks that capture the chemical substructure based on two fingerprints. Specifically, to better capture the long-range interactions, the GNN is designed with an attention mechanism to enhance the atomic interactions. As the Morgan fingerprint effectively captures functional groups of molecules and another well-used molecular fingerprint in the field of drug discovery, the Extended Reduced Graph (ErG) Fingerprint specifically targets molecular features with pharmacological relevance. These two fingerprints are both utilized to complement the chemical information and long-range information processing at the level of key structural features that GNNs lack. The proposed TMRLF extracts a robust feature representation of molecules, crucial for accurately predicting molecular properties and identifying potential drug candidates. Our proposed TMRLF is compared against six state-of-the-art models on eight benchmark datasets. It demonstrates superior capability in predicting molecular properties. Its effectiveness is further highlighted through proof-of-concept validation in identifying potential inhibitors for the Son of Sevenless Homolog 1 (SOS1) protein in real-world drug discovery scenarios.**

***Index Terms*—Artificial Intelligence, Graph Neural Networks, Deep Learning, Drug discovery, Molecular Property Prediction, Molecular Representation Learning**

I. Introduction

Traditional drug discovery is often slow and costly, with a high risk of failure [20]. It may take more than a decade and can be delayed by the need for extensive experiments, and the complexity of biological interactions, leading to unpredictable outcomes [23]. Recently, various deep learning models such as Graph Neural Networks (GNNs) [25], Recurrent Neural Networks [4], [27] and transformers [40] have been increasingly utilized to accelerate this protracted drug discovery cycle [28]. These models can quickly analyze vast datasets, predict results, and simulate complex biological interactions, offering faster and more cost-effective drug development solutions [17].

The most impactful model in this field is the GNN, which has led to the discovery of potential antibiotics with novel structures through the utilization of its variant models [25], [35]. By mapping atoms to nodes and chemical bonds to edges in the graph [35], [38], various advanced GNN models have revolutionized molecular representations. Many GNNs have been developed to model molecular structures for subsequent drug discovery, including Message Passing Neural Networks (MPNNs) [8], Graph Convolutional Networks (GCNs) [13], Graph Attention Networks (GATs) [31], and Spectral-based GNNs [36].

Despite the advancements, two challenges become particularly evident in actual practical applications of GNNs for drug discovery. 1) GNNs face inherent challenges in their capacity to accurately model long-range interactions and identify complex molecular substructures [14], this primarily originates from the architecture's inherent emphasis on local connectivity, focusing on aggregating information from an atom's local graph neighborhood, limiting its capability to integrate and process signals from distant atoms. 2) Existing GNNs generally do not integrate vital chemical information such as functional groups and pharmacophores into their structural analysis [8], [13], while from a chemical perspective, particular molecular substructures are higher-order motifs representing unique atomic arrangements according to chemical rules, having a substantial impact on molecular properties [1]. In drug discovery, these motifs are fundamental for determining the biological activity of compounds, as they influence the binding affinity to target proteins, modulate pharmacokinetics, and contribute to the therapeutic and adverse effects of potential drugs [18], [39].

To tackle these challenges, we propose a novel Three-branch Molecular Representation Learning Framework (TMRLF), utilizing both molecular structure and fingerprints. The proposed TMRLF could effectively capture long-range atomic interactions by incorporating an attention mechanism, and integrate the graphical structure with molecular fingerprints to enrich the model with chemical information beyond pure

[‡] The authors contributed equally.

2836-3795/24/$31.00 ©2024 IEEE
DOI 10.1109/COMPSAC61105.2024.00317

topology. From another perspective, the GNN branch compensates for the potential shortcomings of fingerprints, which might not fully account for the complex topological information of molecules. The use of both molecular structure and fingerprints brings together the detailed chemical knowledge encoded in fingerprints with the deep topological insights provided by GNNs, creating a more robust and comprehensive model for drug discovery. Notably, our ablation experiments confirmed the superior performance of the proposed three-branch method compared to both singular branch methods and pairwise branch combination approaches in predicting the inhibitory activity of Son of Sevenless Homolog 1 (SOS1) [5].

In the proposed three-branch framework, in addition to the GNN branch, two fingerprinting techniques are utilized for the other two branches. Specifically, Morgan fingerprint [26] effectively captures functional group information essential for understanding molecular properties, while Extended Reduced Graph (ErG) Fingerprint [30] specifically targets key structures with pharmacological relevance. By combining these two fingerprinting techniques, we ensure a more comprehensive coverage of chemical information relevant to drug discovery. Our ablation studies show that this dual-fingerprint integration addresses the problem of incomplete chemical information that single fingerprints face.

The proposed TMRLF is compared with six baseline models on eight benchmarks. The results demonstrate its superior performance in molecular property prediction. Moreover, we highlight the practical applicability of TMRLF by employing it to identify potential inhibitors for anti-tumor targets, specifically focusing on the SOS1 protein [5]. The experimental results show that the proposed TMRLF performs much better on multiple drug discovery-related tasks, demonstrating its powerful capacity, effectiveness, and generalizability.

Our main contributions can be summarized as follows. 1) The proposed novel TMRLF captures the graphical structural information and chemical bonding between atoms both globally and locally. 2) The use of both the ErG fingerprint and the Morgan fingerprint can capture both the functional group information and the drug-related properties. 3) The proposed TMRLF achieves comparable or superior performance on eight benchmark datasets compared to six state-of-the-art methods and demonstrates its effectiveness in predicting the inhibitory activity of SOS1.

II. Related Work

A. Molecular Representation Learning

Artificial Intelligence (AI)-driven drug discovery relies on converting drug molecules into an informative representation for analysis [38]. Initially, fixed molecular descriptors and fingerprints are often used, including handcrafted molecular descriptors, covering substituent atoms, chemical bonds, structural fragments, and functional groups [3]. Additionally, molecules are represented by Simplified Molecular Input Entry System (SMILES) strings [32], which are fed into sequence learning models [21]. These methods rely heavily on complex feature engineering for satisfactory predictive performance. Recently, deep learning models have emerged as promising alternatives, offering more expressive molecular representations. For example, GNNs have garnered significant attention for their capacity to model graphical structural data, including molecular graphs [25]. They can be broadly categorized into four types, each tailored to specific aspects of molecular modeling. 1) MPNNs, which enhance the understanding of interactions by aggregating information from immediate neighbors [8]. 2) GCNs, which extend convolution operations to graph data, capturing essential local structural information efficiently [13]. 3) GATs, which introduce various attention mechanisms to prioritize information from significant neighbors [31]. 4) Spectral-based GNNs, which utilize graph spectral theory to analyze molecular structures at a more abstract level [36]. However, GNNs encounter many challenges, as illustrated earlier, which hinder the accurate modeling of molecules' structural information and may lead to sub-optimal performance in drug discovery.

B. Molecular Property Prediction

Molecular property prediction is a fundamental task in drug discovery [33] because many methods rely on predicted molecular properties to evaluate, select, and generate molecules [33]. The application of GNN models in this area is well documented. One of the initial instances of employing GNNs in quantitative structure-activity relationship modeling can be traced back to Merkwitrth [19], where a molecular graph was used to formulate a collection of descriptors, which, in turn, were employed in the construction of classification models aimed at predicting antiviral activity. A decade later, the field witnessed the advent of GCNs that are tailored for non-Euclidean spaces by effectively capturing the relational information among nodes in a simple and interpretable way [6]. As a variant of GCN, the MPNN model, initially introduced by Glimer *et al.* [8], extends upon the GCN framework by incorporating a node-update function. Instead of only using messages related to the vertices (atoms), directed message-passing neural networks (D-MPNN) employ messages associated with directed edges (bonds) to represent a molecule [38]. By gathering latent information from surrounding nodes and minimizing the demand for extensive feature engineering, these architectures effectively capture the spatial interrelationships among atoms within molecules.

III. Proposed Method

A. Overview of Proposed TMRLF

A block diagram of the proposed TMRLF is shown in Fig. 1. The proposed TMRLF extracts molecular features by merging atom-based embeddings from the GNN model and chemical substructure-based information from the molecular fingerprints — Morgan fingerprint [26] utilized to delineate the local topological features of molecules, particularly emphasizing the atom connectivity and cycle presence, and ErG fingerprint captures information about pharmacophore. In this manner, the framework offers a unified end-to-end solution that encompasses detailed structural and chemical information.

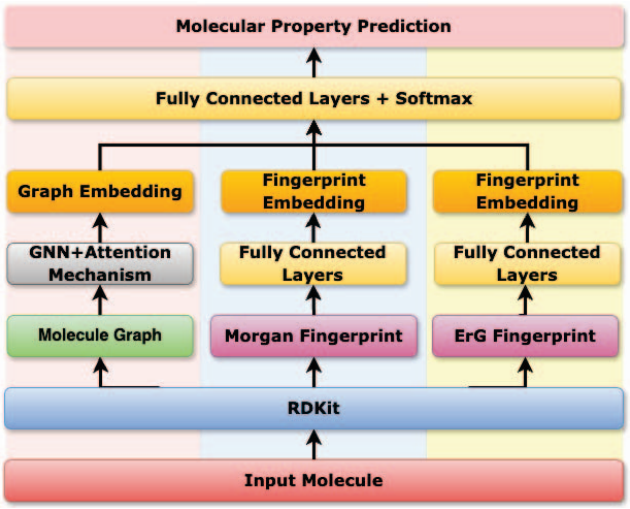

Fig. 1. Overview of proposed TMRLF. The proposed three-branch framework consists of a graph neural network with an attention mechanism to extract structural information from the molecule graph, and two sets of fully connected layers to extract long-range atomic interactions and chemical information from the Morgan and ErG fingerprints.

These three sets of features are then combined through a set of fully connected layers, refining the feature representation in preparation for property prediction.

B. Molecule Graph Embedding

Utilizing the open-source cheminformatics tool RDKit, each molecule's SMILES string [32] is converted into a molecular graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes, each representing an atom within the molecule, and $\mathcal{E}$ represents the set of edges, corresponding to the chemical bonds between atoms. The proposed TMRLF enhances the GNN with an attention mechanism to assess and dynamically prioritize the most informative features across multiple perspectives. We detail the molecule graph embedding process as follows.

Feature Transformation: The feature $\boldsymbol{h}_i$ of each atom is transformed by a weight matrix $\boldsymbol{W}$, mapping the initial features into a higher-dimensional space as:

$$\boldsymbol{h}'_i = \boldsymbol{W}\boldsymbol{h}_i. \tag{1}$$

Multi-Head Attention: The proposed attention mechanism derives attention coefficients multiple times independently across different "heads". For each head k, and each node pair (i, j), the attention mechanism assesses the significance of node j's features to node i through,

$$e_{ij}^{(k)} = \mathcal{F}_L\left(\boldsymbol{a}^{(k)T} \cdot [\boldsymbol{h}'_i \| \boldsymbol{h}'_j]\right), \tag{2}$$

where $\mathcal{F}_L$ denotes the 'LeakyReLU' operation.

Feature Normalization: Attention coefficients from each head are normalized using the softmax function. The model aggregates the transformed features $\boldsymbol{h}'_j$ of neighbors weighted by the normalized attention coefficients $\alpha_{ij}^{(k)}$, also incorporating the node's own transformed features $\boldsymbol{h}'_i$ to ensure self-inclusion,

$$\alpha_{ij}^{(k)} = \frac{\exp(e_{ij}^{(k)})}{\sum_{l\in\mathcal{N}(i)\cup\{i\}} \exp(e_{il}^{(k)})}, \tag{3}$$

where $\mathcal{N}(i)$ denotes the neighborhood of atom i.

Feature Aggregation: The final representation of each node, $\boldsymbol{h}''_i$, is obtained by aggregating contributions from all K heads,

$$\boldsymbol{h}''_i = \mathcal{F}_\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}(i)\cup\{i\}} \alpha_{ij}^{(k)}\boldsymbol{h}'_j\right), \tag{4}$$

where $\mathcal{F}_\sigma$ is an activation function.

Graph-level Output: The molecular graph embedding, $\boldsymbol{h}_G$, results from averaging the updated feature vectors $\boldsymbol{h}''_i$ of all nodes, offering a holistic view of the molecular structure:

$$\boldsymbol{h}_G = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \boldsymbol{h}''_i. \tag{5}$$

C. Molecular Fingerprints

The proposed TMRLF strategically integrates two distinct types of fingerprints: The Morgan fingerprint effectively captures functional group information essential for understanding molecular properties, while the ErG fingerprint specifically targets pharmacophore information crucial for the identification of potential drug candidates. By combining these two, we ensure a more comprehensive coverage of the chemical information relevant to molecular representation in drug discovery.

1) Morgan Fingerprint: A Morgan fingerprint encodes a wide array of molecular characteristics, serving as our primary tool for detailing functional groups within molecules. As an adaptation of the extended-connectivity fingerprint [26], it utilizes a circular substructure approach to encapsulate the molecule's local and global structural details. Specifically, it excels in identifying and encoding functional groups by iterating through each atom and examining the chemical environment up to a specified radius. The Morgan fingerprint $\boldsymbol{P}_i^M$ for the i-th atom is generated mathematically as follows:

$$\boldsymbol{P}_i^M = \bigoplus_{j\in\mathcal{N}(i,R)} \mathcal{F}_H(\boldsymbol{h}_j^{EC}, \boldsymbol{h}_{i,j}^{BC}, d_{i,j}), \tag{6}$$

where $\bigoplus$ denotes the bitwise XOR operation to combine features from different atoms into a unique fingerprint. $\mathcal{F}_H$ denotes the hash function that maps the set of local features into a unique identifier that minimizes collision and maximizes information preservation. $\mathcal{N}(i, R)$ is the neighborhood of atom i, defined by the radius R, which determines how far from atom i the algorithm searches for contributing features. $\boldsymbol{h}_j^{EC}$ (Elemental Composition) includes atomic number, valence, hybridization, and other relevant atomic properties of atom j that are indicative of functional groups. $\boldsymbol{h}_{i,j}^{BC}$ (Bond Characteristics) involves types of bonds (single, double, triple, aromatic) between atoms i and j, critical for determining the molecular structure and functional group attachment. $d_{i,j}$ represents the

topological distance between atoms i and j, important for understanding the spatial arrangement of functional groups.

The proposed TMRLF calculates the Morgan Fingerprint using RDKit with the most commonly employed radius of $R = 2$, efficiently transforming the SMILES string of each molecule into 1024 bits, where each bit signifies the presence or absence of specific substructural patterns including functional groups. To enhance the utility of these fingerprints in predictive modeling, the sparse and high-dimensional fingerprint is transformed into a compact and dense vector through fully connected neural network layers during an embedding process:

$$\boldsymbol{h}_i^M = \mathcal{F}_{MLP}(\boldsymbol{P}_i^M), \tag{7}$$

where $\mathcal{F}_{MLP}$ represents a Multi-Layer Perceptron (MLP).

2) ErG Fingerprint: The proposed TMRLF utilizes the ErG fingerprint [30] for its unique ability to capture pharmacophoric features essential for analyzing molecular structures. The ErG Fingerprint abstracts molecules into a graph representation where nodes and edges play crucial roles. We let $\mathcal{G}_{\text{ErG}} = (\mathcal{V}, \mathcal{E})$ denotes the extended reduce graph of the molecule, where $\mathcal{V}$ denotes a set of vertices representing pharmacophore types (such as hydrogen bond donors, acceptors, aromatic rings, and hydrophobic regions), and $\mathcal{E}$ comprises the edges encoding spatial relationships between these features. The ErG fingerprint $\boldsymbol{h}^E$ is then extracted as:

$$\boldsymbol{h}^E = \mathcal{F}_E(\mathcal{G}_{\text{ErG}}), \tag{8}$$

where $\mathcal{F}_E$ is a function to encode the graph $\mathcal{G}_{\text{ErG}}$ into a fixed-size descriptor, typically a binary vector. $\boldsymbol{h}^E$ is then transformed into a dense vector through an embedding process:

$$\boldsymbol{h}^N = \mathcal{F}_{MLP}(\boldsymbol{h}^E). \tag{9}$$

D. Training strategies

1) Loss Function: For classification tasks, the framework employs the Binary Cross-Entropy With Logits Loss $\mathcal{L}_C$ as,

$$\mathcal{L}_C(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \cdot \log(\sigma(\hat{y}_i)) + (1-y_i)\cdot\log(1-\sigma(\hat{y}_i))\big], \tag{10}$$

where $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]$ is the vector of predicted values, and $\hat{y}_i$ denotes the prediction for the i-th sample. $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ is the vector of ground-truth values, where y_i denotes the true value for the i-th sample. σ denotes the logistic sigmoid function and N represents the number of samples in the batch.

For regression tasks, Mean Squared Error Loss $\mathcal{L}_R$ is used,

$$\mathcal{L}_R(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2. \tag{11}$$

TABLE I
OVERVIEW OF THE EIGHT BENCHMARK DATASETS.

Dataset	Tasks	Molecules	Category	Task Type
BACE	1	1513	Biophysics	Classification
BBBP	1	2039	Physiology	Classification
SIDER	27	1427	Adverse Drug Reactions	Classification
Tox21	12	7831	Toxicity	Classification
ClinTox	2	1478	Toxicity	Classification
ESOL	1	1128	Physicochemical property	Regression
FreeSolv	1	642	Physicochemical property	Regression
Lipophilicity	1	4200	Physicochemical property	Regression

2) Hyper-parameter Optimization: The proposed TMRLF utilizes the Adam optimizer [12] for gradient descent optimization because of its efficient computation of the first and second moments of the gradients and its adaptive learning rate capabilities. Prior to model training, we conducted Bayesian optimization of hyper-parameters for each dataset using the Hyperopt Python package to determine the optimal configurations. These configurations include the dropout rate for the GNN branch, the number of multi-head attention mechanisms, the dimensionality of the attention layers, and the dimensions and dropout rates for the two fingerprint branches. Following this, the model was systematically trained, validated, and tested using the optimized parameters.

IV. EXPERIMENTAL RESULTS

A. Experimental Settings

1) Benchmark Datasets: To evaluate the proposed TMRLF for molecular property prediction, we selected benchmark datasets from MoleculeNet [37], covering over 40 essential molecular property prediction tasks. These datasets encompass various properties vital for drug discovery and toxicological studies, including physiology, toxicity, biophysics, adverse drug reactions, and physicochemical properties. Table I summarizes the eight benchmark datasets in this study, including the number of tasks, the number of molecules, the task type, and the category of each dataset.

On each dataset, we employed the scaffold split [15] to partition the dataset into training, validation, and test sets at a ratio of 8:1:1. Unlike random splitting, scaffold-split divides molecules based on molecular skeletons to ensure that the split subsets are more structurally different. This division strategy challenges the model's ability to generalize well beyond its training distribution, making prediction not only more difficult but also more reflective of real-world scenarios [37]. Finally, the average results obtained from ten independent runs under random seeds were reported.

2) Compared Methods: The proposed TMRLF was compared with six state-of-the-art models for molecular property prediction: DMPNN [38], TF_Robust [24], Weave [11], MPNN [8], MGCN [16], and TrimNet [15]. Among these models, TrimNet [15], as a graph-based approach, introduces

a novel triplet message-passing mechanism to learn molecular representation efficiently. MPNN [8] and its variants, DMPNN [38] and MGCN [16], are specialized models that incorporate edge features during message passing, thereby enhancing their ability to capture detailed molecular interactions. TF_Robust [24] is a multi-task deep neural network that integrates molecular fingerprints as inputs, combining traditional cheminformatics methods with modern machine learning techniques. Weave [11], in contrast, is a variant of GCNs, designed to process molecular graphs through the systematic integration of node features at each network layer.

3) Evaluation Metrics: We employed different evaluation metrics tailored to the nature of the tasks. For classification tasks, we utilized the area under the receiver operating characteristic curve (AUC) as the primary evaluation metric:

$$A^{macro} = \frac{1}{C}\sum_{i=1}^{C} A_i, \tag{12}$$

where C is the number of classes, and A_i is the AUC for the i-th class. This metric treats each class with equal importance, providing a balanced view of model performance across the dataset's diversity. For regression tasks, we employed the Root Mean Square Error (RMSE) as the key evaluation metric:

$$\epsilon_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{13}$$

where $\hat{y}_i$ is the predicted value, y_i is the actual value, and N is the number of observations.

For datasets involving multiple tasks, we reported the average performance metric across all tasks for each model.

B. Comparison with State-of-the-Art Methods

Tables II and III summarize the performance of various methods on the classification and regression tasks, respectively. Models that outperformed all others on the benchmark dataset are highlighted in bold. In the evaluation across different benchmark datasets for classification, the proposed TMRLF achieves the highest A^{macro} in three of the benchmark datasets, surpassing the other state-of-the-art models. When evaluating the TMRLF's performance on regression tasks, it demonstrates superior predictive accuracy on all three benchmarks. This superior performance of TMRLF across different benchmark datasets suggests its excellent ability to capture complex relationships and patterns within molecular data, leading to improved predictive accuracy.

Deep learning models often exhibit moderate performance on small datasets due to the limited information provided by the small number of samples. Empirical results on the BACE, SIDER, ESOL, and FreeSolv datasets, each comprising fewer than 2000 molecules, demonstrate the efficacy of our model in handling scenarios with limited sample sizes. Despite the small size of these datasets, our framework exhibits robust performance, highlighting its capacity to effectively generalize from limited data.

TABLE II
A^{macro} FOR FIVE CLASSIFICATION BENCHMARKS.

Model	BACE	SIDER	BBBP	Tox21	ClinTox
DMPNN [38]	0.852	0.632	0.919	**0.826**	0.897
TF_Robust [24]	0.824	0.607	0.860	0.698	0.765
Weave [11]	0.791	0.543	0.837	0.741	0.823
MPNN [8]	0.815	0.595	0.913	0.808	0.879
MGCN [16]	0.734	0.552	0.850	0.707	0.634
TrimNet [15]	0.843	0.606	0.892	0.812	**0.906**
TMRLF	**0.865**	**0.641**	**0.924**	0.814	0.815

TABLE III
ϵ_{RMSE} FOR THREE REGRESSION BENCHMARKS.

Model	ESOL	FressSolv	Lipo
DMPNN [38]	0.980	2.177	0.653
TF_Robust [24]	1.722	4.122	0.909
Weave [11]	1.158	2.398	0.813
MPNN [8]	1.167	2.185	0.672
MGCN [16]	1.266	3.349	1.113
TrimNet [15]	1.282	2.529	0.702
TMRLF	**0.860**	**2.009**	**0.640**

C. Ablation Studies

After establishing the benchmark performance of our TMRLF against state-of-the-art models, we next conducted ablation studies to delve deeper into the factors that influence the performance. We decided not to continue using the previously employed benchmark datasets for predicting molecular properties. Instead, our primary objective was to enhance the performance of downstream tasks and ultimately facilitate drug discovery. Aligned with this vision, we chose to utilize the SOS1 dataset [5] retrieved from the ChEMBL database [7] for ablation, which consists of 375 labeled molecules with known inhibitory activity against the SOS1 target. SOS1 is a gene that encodes a guanine nucleotide exchange factor vital for the activation of RAS proteins, thereby impacting cellular processes such as cell division and signal transduction pathways. Over-activation or mutations of SOS1 can lead to excessive RAS activation, which has been implicated in the development of various cancers due to uncontrolled cell proliferation. Therefore, inhibitors targeting SOS1 are considered promising for intervening in cancer progression by impeding the activation of subsequent pathways [29]. The selection of the SOS1 dataset supports our goal and effectively simulates the scenario of scarce labeled data encountered in real-world drug discovery. We evaluated various combinations of the three branches of features: A GNN with an attention mechanism, ErG fingerprint, and Morgan fingerprint. The effectiveness of each component individually was also examined. Model training was performed for each configuration, with hyperparameter optimization undertaken to ensure optimal model performance.

TABLE IV
ABLATION STUDIES OF PROPOSED TMRLF ON THE SOS1 DATASET.

GNN	ErG Fingerprint	Morgan Fingerprint	ϵ_{RMSE}
✓	✗	✗	0.72
✗	✓	✗	0.43
✗	✗	✓	0.38
✓	✓	✗	0.41
✓	✗	✓	0.36
✗	✓	✓	0.36
✓	✓	✓	0.34

Effect of Using Two Fingerprints: As shown in Table IV, the use of both fingerprints outperforms the employment of any single fingerprint alone, *e.g.*, using only the Morgan fingerprint yields an RMSE value of 0.38, and employing only the ErG fingerprint leads to an RMSE value of 0.43, while using both of them achieves an RMSE value of 0.36. This improvement demonstrates the synergistic effects of integrating Morgan and ErG fingerprints, which together provide a more comprehensive chemical and structural understanding of molecules.

Effect of Using Three Branches: As illustrated in Table IV, when using GNN alone, the performance is significantly lower compared to that of other individual components, as well as compared with the combination of all three. Notably, the incorporation of any single fingerprint, whether it relates to functional groups or pharmacophoric features, into the GNN model initially comprising solely the graphical network, leads to a notable performance enhancement. This improvement is further enhanced when both types of fingerprints are integrated into the framework, culminating in optimal performance with the RMSE reduced to a minimum value of 0.34.

These results indicate that while GNNs alone may not perform optimally on small datasets, the inclusion of fingerprints that carry functional group and pharmacophore information serves as a powerful complement. This finding also emphasizes the crucial role of integrating specific chemical information to effectively support AI-assisted drug discovery.

D. Proof-of-concept Validation in Real-world Drug Discovery

To evaluate TMRLF's capability of identifying SOS1 inhibitors accurately, we designed a virtual screening task that reflects the real-world common challenges of identifying potent inhibitors from a large database. Specifically, we created a dataset that includes 1,615 US Food and Drug Administration-approved (FDA) drugs collected from DrugBank [34]. Additionally, two known SOS1 inhibitors, BI-3406 [9] and BAY-293 [10], were strategically integrated into this dataset. These inhibitors, previously documented for their inhibitory activity, serve as reliable benchmarks for evaluation purposes. Importantly, none of these 1617 compounds were previously included in the training set of the model, ensuring a strict test of the model's ability to generalize to new molecules. The TMRLF that exhibited the best performance following extensive hyper-parameter optimization on the SOS1 dataset was used here.

The results were notably positive. The predictions generated by TMRLF were sorted by the pChEMBL values, which is a metric to quantify the potency of an inhibitor, from the highest to the lowest. BI-3406 [9] and BAY-293 [10] were ranked notably high within the dataset, placing the 4-th and the 26-th respectively out of 1,617 compounds, within the top 2% of all molecules. Additionally, the predictions matched well with their reported inhibitory activity, with BI-3406's predicted pChEMBL value of 8.27 compared to the reported value of 8.30 [9], and BAY-293's predicted value of 7.49 compared to the reported value of 7.68 [10]. The results not only confirm that the framework is capable of identifying potent inhibitors through learned molecular features, but also demonstrate its potential for broader adoption in the field of drug discovery.

E. Discussion

While TMRLF appears to be a promising approach, there are still opportunities to further enhance it. Our ablation studies on the SOS1 dataset reveal differences in the contributions of the two types of molecular fingerprints for predicting molecular properties. While these findings alone do not suffice to judge the superiority of one type of fingerprint over another, they inspire potential enhancements for TMRLF. Specifically, implementing a dynamic weighting mechanism that adjusts the influence of each fingerprint type based on the characteristics of datasets could optimize model performance across diverse datasets, which would allow TMRLF to adapt more flexibly and effectively to various datasets. Potentially, gated fusion could be a feasible solution to achieve this [2]. Furthermore, there is a critical requirement to improve the interpretability of TMRLF. Utilizing techniques such as attention mechanisms or saliency maps [22] can help reveal the key molecular features influencing predictions. Integrating domain expertise from chemistry and bioinformatics into interpretability frameworks could enhance the relevance and clarity of explanations.

V. CONCLUSION

In this paper, we have introduced TMRLF, an enhanced molecular representation learning framework for predictive modeling in drug discovery. It integrates one branch of a GNN with an attention mechanism and two branches of fully connected networks based on two fingerprints. This combined architecture is designed to efficiently capture both global atomic interactions and molecular substructures. Our experiments confirm that TMRLF effectively learns molecular features, enhancing its ability in downstream tasks of predicting molecular properties. Specifically for drug discovery, TMRLF demonstrates its capability to efficiently identify potential SOS1 inhibitors. We hope that TMRLF could serve as a robust tool in the field of drug discovery, helping researchers predict molecular properties accurately and streamline the development of new therapeutic agents.

VI. Acknowledgements

Professor Hirst acknowledges support from the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chairs in Emerging Technologies scheme. Other authors acknowledge the funding support (Grant No. 2022J171, 2020Z092 and 22171153).

References

[1] J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Generic strategies for chemical space exploration. *International Journal of Computational Biology and Drug Design*, 7(2-3):225–258, 2014.

[2] D. Budden, A. Marblestone, E. Sezener, T. Lattimore, G. Wayne, and J. Veness. Gaussian gated linear networks. *Advances in Neural Information Processing Systems*, 33:16508–16519, 2020.

[3] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[4] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018.

[5] L. Duo, Y. Chen, Q. Liu, Z. Ma, A. Farjudian, W. Y. Ho, S. S. Low, J. Ren, J. D. Hirst, H. Xie, et al. Discovery of novel SOS1 inhibitors using machine learning. *RSC Medicinal Chemistry*, 2024.

[6] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, 28, 2015.

[7] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.

[8] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

[9] R. C. Hillig, B. Sautier, J. Schroeder, D. Moosmayer, A. Hilpmann, C. M. Stegmann, N. D. Werbeck, H. Briem, U. Boemer, J. Weiske, et al. Discovery of potent SOS1 inhibitors that block RAS activation via disruption of the RAS–SOS1 interaction. *Proceedings of the National Academy of Sciences*, 116(7):2551–2560, 2019.

[10] M. H. Hofmann, M. Gmachl, J. Ramharter, F. Savarese, D. Gerlach, J. R. Marszalek, M. P. Sanderson, D. Kessler, F. Trapani, H. Arnhof, et al. BI-3406, a potent and selective SOS1–KRAS interaction inhibitor, is effective in KRAS-driven cancers through combined MEK inhibition. *Cancer Discovery*, 11(1):142–157, 2021.

[11] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30:595–608, 2016.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[14] A. Kosmala, J. Gasteiger, N. Gao, and S. Günnemann. Ewald-based long-range message passing for molecular graphs. In *International Conference on Machine Learning*, pages 17544–17563. PMLR, 2023.

[15] P. Li, Y. Li, C.-Y. Hsieh, S. Zhang, X. Liu, H. Liu, S. Song, and X. Yao. TrimNet: learning molecular representation from triplet messages for biomedicine. *Briefings in Bioinformatics*, 22(4):bbaa266, 2021.

[16] C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, and L. He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1052–1060, 2019.

[17] M. Lu, J. Yin, Q. Zhu, G. Lin, M. Mou, F. Liu, Z. Pan, N. You, X. Lian, F. Li, et al. Artificial intelligence in pharmaceutical sciences. *Engineering*, 27:37–69, 2023.

[18] K.-K. Mak, Y.-H. Wong, and M. R. Pichika. Artificial intelligence in drug discovery and development. *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*, pages 1–38, 2023.

[19] C. Merkwirth and T. Lengauer. Automatic generation of complementary descriptors with molecular graph networks. *Journal of Chemical Information and Modeling*, 45(5):1159–1168, 2005.

[20] J. G. Moffat, F. Vincent, J. A. Lee, J. Eder, and M. Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug discovery*, 16(8):531–543, 2017.

[21] N. R. Monteiro, B. Ribeiro, and J. P. Arrais. Drug-target interaction prediction: end-to-end deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6):2364–2374, 2020.

[22] T. N. Mundhenk, B. Y. Chen, and G. Friedland. Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*, 2019.

[23] K. Nicolaou. Advancing the drug discovery and development process. *Angew Chem Intl Ed*, 53(35):9128–9140, 2014.

[24] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[25] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, et al. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, 2022.

[26] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[27] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018.

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

[29] H. Sondermann, S. M. Soisson, S. Boykevisch, S.-S. Yang, D. Bar-Sagi, and J. Kuriyan. Structural analysis of autoinhibition in the Ras activator Son of sevenless. *Cell*, 119(3):393–405, 2004.

[30] N. Stiefl, I. A. Watson, K. Baumann, and A. Zaliani. ErG: 2D pharmacophore descriptions for scaffold hopping. *Journal of Chemical Information and Modeling*, 46(1):208–220, 2006.

[31] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. Graph attention networks. *Stat*, 1050(20):10–48550, 2017.

[32] D. Weininger, A. Weininger, and J. L. Weininger. SMILES. 2. algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

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

[34] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

[35] F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, 2024.

[36] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

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

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

[39] S.-Y. Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, 15(11-12):444–450, 2010.

[40] S. Zhang, R. Fan, Y. Liu, S. Chen, Q. Liu, and W. Zeng. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, 3(1):vbad001, 2023.