Synergistic Fusion of Graph and Transformer Features for Enhanced Molecular Property Prediction

M V Sai Prakash  
mukkamala.prakash@quantiphi.com  
Applied Research, Quantiphi  
Mumbai, India

Siddartha Reddy N  
siddartha.reddy@quantiphi.com  
Applied Research, Quantiphi  
Bangalore, India

Ganesh Parab  
ganesh.parab@quantiphi.com  
Applied Research, Quantiphi  
Mumbai, India

Varun V  
varun.v@quantiphi.com  
Applied Research, Quantiphi  
Bangalore, India

Vishal Vaddina  
vishal.vaddina@quantiphi.com  
Applied Research, Quantiphi  
Toronto, Canada

Saisubramaniam  
gopalakrishnan.saisubramaniam@quantiphi.com  
Applied Research, Quantiphi  
Bangalore, India

ABSTRACT
Molecular property prediction is a critical task in computational drug discovery. While recent advances in Graph Neural Networks (GNNs) and Transformers have shown to be effective and promising, they face the following limitations: Transformer self-attention does not explicitly consider the underlying molecule structure while GNN feature representation alone is not sufficient to capture granular and hidden interactions and characteristics that distinguish similar molecules. To address these limitations, we propose SYN-FUSION, a novel approach that synergistically combines pre-trained features from GNNs and Transformers. This approach provides a comprehensive molecular representation, capturing both the global molecule structure and the individual atom characteristics. Experimental results on MoleculeNet benchmarks demonstrate superior performance, surpassing previous models in 5 out of 7 classification datasets and 4 out of 6 regression datasets. The performance of SYN-FUSION has been compared with other Graph-Transformer models that have been jointly trained using a combination of transformer and graph features, and it is found that our approach is on par with those models in terms of performance. Extensive analysis of the learned fusion model across aspects such as loss, latent space, and weight distribution further validates the effectiveness of SYN-FUSION. Finally, an ablation study unequivocally demonstrates that the synergy achieved by SYN-FUSION surpasses the performance of its individual model components and their ensemble, offering a substantial improvement in predicting molecular properties.

CCS CONCEPTS
- Applied computing → Molecular sequence analysis; Bioinformatics; Computing methodologies → Artificial intelligence; Transfer learning; Neural networks.

KEYWORDS
Graph Neural Networks, Transformer, Molecular Representation Learning, Molecular Property Prediction, Synergy

1 INTRODUCTION
Molecular property prediction [38] has rapidly evolved into an interdisciplinary field that leverages insights from chemistry, physics, and materials science. Predicting molecular properties is widely considered one of the most critical tasks in computational drug discovery. The ability to accurately predict molecular properties enables transformative applications in drug design, materials development, and reaction optimization [4, 11, 30]. Early approaches in molecular property prediction include quantum mechanics-based mathematical models describing atomic and molecular behavior [46], computational chemistry involving the study of chemical systems through computer simulations [32], and molecular mechanics and molecular dynamics [29] involving the simulation of larger and more complex molecules.

With the advent of computational methods, fingerprint techniques such as Morgan Fingerprint [28], Extended-Connectivity FingerPrints (ECFP) [34] have been developed for efficient molecular representation. Statistical methods and machine learning models like Quantitative Structure-Property Relationships (QSPR) [2] and Quantitative Structure-Activity Relationships (QSAR) [7] predict properties based on the molecular structure using these techniques. By utilizing large datasets to learn complex structure-property relationships [22, 39], deep learning approaches surpass traditional methods across various molecular tasks. Two such widely adopted approaches include modeling the molecule as (i) a sequence of atoms using either Simplified Molecular Input Line Entry System (SMILES) [45] or SELF-referencing Embedded Strings (SELFIES) [21] and (ii) a graph-based structure using Graph Neural Networks (GNN) [37].

Sequence-based molecular property prediction has witnessed significant growth in recent years [6, 42, 54]. This approach leverages the inherent sequential nature of molecular structures and protein sequences to accurately predict properties, opening new possibilities in the realm of drug discovery and materials science. Graph neural networks (GNNs) model the natural representation of molecules as graphs and use neighborhood aggregation strategies to predict molecular properties such as solubility, toxicity, binding affinity, etc. [13, 51] Related tasks, including molecular generation...
self-attention mechanisms that enable them to attend to different parts of the molecule and consider the relationships between atoms and bonds. However, while models such as Chemformer [17] and X-MOL [50], trained on SMILES data, have exhibited promising outcomes in classification and regression tasks, they have limitations in providing comprehensive insights into the underlying molecular structure, particularly in modeling the intricate connectivity patterns and spatial arrangements found in molecular graphs.

Graph Neural Networks (GNNs) have demonstrated significant potential in predicting molecular properties, as evidenced by previous studies [13, 19, 48]. In GNNs, molecules are represented as graphs, with atoms serving as nodes and bonds as edges. These networks leverage the graph structure to learn representations of the molecules. A widely adopted GNN-based approach is the Message Passing Neural Network (MPNN) [13]. MPNN utilizes a recursive message-passing mechanism to propagate information throughout the molecular graph structure. Another approach is the Graph Convolutional Network (GCN) [20], which uses graph convolution operations to learn node embeddings. GIN [48] handles graph isomorphism (the similarity between two graphs despite differences in node labels or orderings) by employing an aggregation function that is independent of the ordering of nodes or edges.

Molecules with similar overall structures can exhibit distinct functional groups or subtle variations that significantly impact their properties [31]. Self-supervised learning has gained significant attention in the field of molecular property prediction, enabling models to learn meaningful representations from unlabeled molecular data. Hu et al [16] introduces innovative strategies for molecule graphs, encompassing pre-training at both the node and graph levels. Contrastive learning [5, 49] is a machine learning technique that learns data representations by comparing pairs of instances. Instances that are similar are pulled closer together, while instances that are dissimilar are pushed further apart. In molecular property prediction, contrastive learning-based methods involve encoding molecular structures as feature vectors and comparing these vectors using a contrastive loss function. An illustration of this is MolCLR [44], which employs a contrastive loss function to learn representations of molecular structures that can subsequently be utilized for predicting molecular properties. MegaMolBART utilizes a transformer architecture based on BART [23] and is trained for small molecule drug discovery. It comprises a bidirectional encoder and an autoregressive decoder. The pretraining of MegaMolBART is built upon the foundation of Chemformer [17] and uses Masked Language Modelling and augmentation of SMILES input. Recent efforts have focused on integrating graph information into Transformer models. GROVER [36] uses a graph multi-head attention network with node vectors obtained through a specialized Graph Neural Network (GNN). Graphormer [52] is based on the standard Transformer, taking graph-structured encoded data directly as input and avoiding conversion to sequential formats. PharmHGT [18] utilizes various perspectives within the molecular graph for message passing, yielding distinct atom features, followed by attention aggregation for holistic molecule representation.

Transformers excel in learning complex relationships and hidden dependencies across the entire molecule, including interactions between atoms and bonds. However, they do not explicitly

2 RELATED WORKS

Molecular fingerprinting methods commonly used in cheminformatics and computational chemistry such as Extended-Connectivity FingerPrints (ECFP) [34] encode the structural features of molecules into fingerprint representations for similarity-based analyses and machine learning tasks. ECFP generates a binary fingerprint for each molecule based on the structure, taking into account the connectivity of atoms and the presence of chemical groups. ECFP is fast but limited in its ability to capture the diversity of molecular structures, as it only considers the presence (1) or absence (0) of specific sub-structures within a molecule.

The Simplified Molecular Input Line Entry System (SMILES) [45] is a linear representation designed to encode molecular structure in a machine-readable format. SMILES uses a distinctive set of characters to represent atoms, bonds, and functional groups within a molecule. Due to its machine-interpretability, SMILES has become a key molecular descriptor for training machine learning models. Advanced machine learning algorithms, including deep neural networks, can utilize SMILES strings to learn predictive models for various molecular properties. Transformer-based models [6, 42, 54] have been employed for molecular property prediction, primarily relying on SMILES representations. Transformers utilize

A pivotal challenge in property prediction is the limited availability of labeled data, an obstacle shared across different fields such as language and vision [9, 12]. The success of self-supervised learning in image and text domains [8, 15] has also been extended to molecular property prediction [53]. Contrastive learning [5] has been shown to be effective in learning better latent representations by pre-training a model to maximize the distance between positive and negative pairs from unlabeled data samples and learning downstream tasks with limited data [16, 35]. Masked Language Modelling (MLM) [8, 23] has been adopted as a pre-training strategy in sequence-to-sequence and discriminative cheminformatics tasks [17]. These approaches have proven beneficial for models to leverage the knowledge learned from larger datasets when fine-tuning smaller, focused tasks.

Both sequence-based transformers and graph-based models learn richer representations when they are initially pre-trained on large datasets of molecules in a self-supervised manner, followed by supervised fine-tuning on smaller datasets with specific properties of interest. This work investigates the benefits of fusing the pre-trained latent representations from both approaches and fine-tuning them towards the downstream task of property prediction.

The contributions of this work are as follows: (i) Learning the synergistic interaction between pre-trained features from graphs and transformers to create a more comprehensive molecular representation that captures both the global molecule structure and the characteristics of individual atoms, (ii) Conducting a detailed analysis of the learned synergistic fusion representation in various aspects, including loss, latent space, activation, and weights, through classification and regression case studies, and, (iii) an ablation study to showcase that the synergy effect of fusion is greater than the performances of the individual models and their ensemble.

[3], reaction prediction [27], and molecular docking [43] are other applications of GNNs.
3 METHODOLOGY

We propose a novel approach to learning the synergistic interaction between pre-trained features from graphs and transformers, termed Synergistic Fusion (SYN-FUSION). The overall framework is illustrated in Figure 1. The approach comprises two steps: In the first step, a single molecule is represented as a Graph and a SMILES string independently, and then encoded into two feature representations using a Graph Neural Network such as GIN and a Transformer based network such as MegaMolBART, respectively. In the second step, the two distinct features are concatenated and fused using a linear layer that learns the combined ‘synergistic’ information for enhanced downstream property prediction.

3.1 Graph Isomorphism Network

Graph Isomorphism Network (GIN) [48] learns a permutation-invariant representation of each input graph, achieved by applying a series of message-passing operations to the nodes and edges of the graph. Let \( G = (V, E) \) denote an undirected graph. Let \( V \) be the node feature matrix of the nodes in \( V \) and \( E \) be the edge feature matrix of the edges in \( E \). The message-passing operation computes a new feature vector \( h \) for each node in \( V \):

\[
h_i^{(k+1)} = \text{MLP}_{\text{atom}}^{(k+1)} \left( (1 + e_i^{(k)}) h_i^{(k)} \right) + \sum_{j \in \mathcal{N}(i)} \left( h_j^{(k)} + \text{MLP}_{\text{bond}}^{(k+1)} (e_{ij}) \right)
\]

where \( h_0 = V, e_{ij} \) is the edge feature vector of edge \( e \in E \) connecting atoms \( i \) and \( j \), \( k \) represents the \( k_{th} \) layer, MLP stands...
for multi layered perception, \(MLP_{atom}^{(k+1)}\) and \(MLP_{bond}^{(k+1)}\) are the \((k + 1)\)-th MLP layers on the atom- and bond-level respectively, \(N(i)\) is the set of neighboring nodes of \(i\), and \(\epsilon(k)\) is a learnable parameter that helps to avoid over-smoothing of the features.

The pooling operation aggregates the feature vectors of all nodes in the graph into a single vector:

\[
pooling(H) = \text{mean}(H) \tag{2}
\]

where \(H = [h_1^{(kn)}, h_2^{(kn)}, \ldots, h_N^{(kn)}]\) is the matrix of node features at the final layer \(kn\), \(N\) is the number of nodes in \(G\) and mean(\()\) is the element-wise mean operator.

The final output of the GIN network is obtained by passing the pooled feature vector through a linear layer:

\[
z_{\text{GIN}} = \text{linear}(pooling(H)) \tag{3}
\]

where \(z_{\text{GIN}}\) is the latent representation for the input graph \(G\), and linear is a multi-layer perceptron.

By stacking multiple layers of message passing and pooling operations, GIN is able to learn a hierarchical representation of the input graph that is invariant to node ordering and is capable of capturing complex structural patterns.

### 3.2 Transformer - MegaMolBART

The Chemformer [17] paper proposes a pre-training method for molecular property prediction based on the Bidirectional and Auto-Regressive Transformers (BART) [23] architecture. MegaMolBART employs an identical configuration for pre-training i.e. Masked Language Modeling (MLM). The goal of MLM is to predict the masked tokens based on the context provided by the other tokens in the sequence. The pre-training process commences by transforming each molecule in the batch into a non-canonical SMILES representation that aligns with the specific molecule. The SMILES strings are subsequently subjected to random masking, tokenization, and embedding into a vector sequence. This modified sequence is then fed into the bidirectional encoder, while the autoregressive decoder is tasked with predicting the initial SMILES sequence based on the same right-shifted sequence. A fully-connected layer is employed to process the decoder’s output, generating a distribution across the model’s vocabulary. To obtain the latent feature from MegaMolBART, only the encoder component is required.

\[
z_{\text{MMB}} = Encoder_{\text{MMB}}(x) \tag{4}
\]

where the bidirectional encoder of MegaMolBART, \(Encoder_{\text{MMB}}\) takes a SMILES representation of the molecule \(x\) as input and provides the corresponding latent representation \(z_{\text{MMB}}\).

### 3.3 Synergistic Fusion

Synergy refers to the phenomenon where the combined effect of two or more substances is greater than the sum of their individual effects. It occurs when the interaction between the substances enhances or amplifies their overall impact [33]. Let us consider two substances, denoted as \(A\) and \(B\), which each possess distinct effects represented by variables \(X\) and \(Y\), respectively. The combined effect resulting from the interaction of \(A\) and \(B\) can be represented as \(Z\). If there exists synergy between substances \(A\) and \(B\), it can be expressed through the equation \(Z > X + Y\). This equation denotes that the combined effect \((Z)\) surpasses the summation of the individual effects \((X + Y)\), thereby indicating the presence of a synergistic interaction.

Let us define the latent embeddings from MegaMolBART encoder as \(z_{\text{MMB}}\) and the latent embeddings from GIN as \(z_{\text{GIN}}\).

#### 3.3.1 Classification

For classification tasks using the cross-entropy loss, the objective function can be written as:

\[
L_c = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{5}
\]

where \(N\) is the number of samples, \(C\) is the number of classes, \(y_{ij}\) is the true label for sample \(i\) and class \(j\), and \(\hat{y}_{ij}\) is the predicted probability of sample \(i\) belonging to class \(j\).

We can now express the predicted probabilities as a function of the feature embeddings:

\[
\hat{y}_{ij} = \text{softmax} \left( W \left[ z_{\text{GIN}}^{(i)} z_{\text{MMB}}^{(i)} \right] + b \right)_{ij} \tag{6}
\]

where \(W\) and \(b\) are the weight matrix and bias vector of the linear layer, \([\cdot, \cdot]\) denotes concatenation, and \(\text{softmax}(\cdot)\) is the softmax function.

Finally, we can write the objective function for classification as:

\[
L_c = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log \left( \text{softmax} \left( W \left[ z_{\text{GIN}}^{(i)} z_{\text{MMB}}^{(i)} \right] + b \right)_{ij} \right) \tag{7}
\]

#### 3.3.2 Regression

For regression tasks using the Mean Squared Error loss, the objective function can be written as:

\[
L_r = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{8}
\]

where \(N\) is the number of samples, \(y_i\) is the true value for sample \(i\), and \(\hat{y}_i\) is the predicted value.

The predicted value \(\hat{y}_i\) can be defined as:

\[
\hat{y}_i = \left( W \left[ z_{\text{GIN}}^{(i)} z_{\text{MMB}}^{(i)} \right] + b \right)_i \tag{9}
\]

Finally, we can write the objective function for regression as:

\[
L_r = \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \left( W \left[ z_{\text{GIN}}^{(i)} z_{\text{MMB}}^{(i)} \right] + b \right)_i \right)^2 \tag{10}
\]

### 4 EXPERIMENTS

#### 4.1 Data

A series of experiments were conducted utilizing multiple molecular benchmarks obtained from MoleculeNet [47]. These benchmarks encompass classification and regression tasks derived from diverse studies. To divide the datasets into training and testing sets, the scaffold split method [1] was employed. A scaffold refers to a molecular substructure that exists within a group of molecules and serves to define a chemical series or class. To ensure the evaluation of the model’s generalization ability, this procedure maintains chemical distinctiveness between the training and test sets. The training set comprised molecules possessing a specific scaffold, while the test set comprised molecules lacking that specific scaffold. Using the scaffold split, the molecules in each dataset were divided into training,
Table 1: Classification Results on MoleculeNet datasets using scaffold split. SYN-FUSION approach outperforms the baselines in 5 out of 7 datasets. The best score for each dataset is indicated in bold and the second-best score is underlined.

<table>
<thead>
<tr>
<th>Classification (Higher is Better)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset</strong></td>
</tr>
<tr>
<td>Molecules</td>
</tr>
<tr>
<td>RF</td>
</tr>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>GIN [20]</td>
</tr>
<tr>
<td>D-MPNN [51]</td>
</tr>
<tr>
<td>Hu et al. [16]</td>
</tr>
<tr>
<td>MolCLR [44]</td>
</tr>
<tr>
<td>SYN-FUSION</td>
</tr>
<tr>
<td>SYN-FUSION_Hu et al.</td>
</tr>
<tr>
<td>SYN-FUSION_MolCLR</td>
</tr>
</tbody>
</table>

4.4 Baseline Methods

SYN-FUSION underwent a comprehensive evaluation for molecular property prediction, comparing its performance with other GNN and Transformer baseline methods. The evaluation process comprised of Random Forest (RF) and Support Vector Machine (SVM) which take molecular FPs as the input, GNN architectures that integrate edge features during the aggregation process such as GCN [20] and GIN [48], capturing quantum interactions within molecules such as D-MPNN [51], node-level (self-supervised) and graph-level (supervised) pre-training approaches as described in Hu et al. [16], contrastive pre-training approach such as MolCLR [44]. Transformer-Graph Combination networks that use both self-attention and graph features for training such as Graphformer [52], pharmHGT [18] and GROVER [36]. Due to inconsistencies between the reported results in the MolCLR paper and the reproduction using their provided code repository, the scores obtained from rerun using their code are presented and compared.

Transformer based models such as Chemformer [17], X-Mol [49], and MolBERT [25] were also included for comparison using random split as followed in their respective works. For a fair comparison, models that solely focus on 2D information were included, and models that incorporate 3D features like MGCN [26], GEM [10], etc. were excluded.

4.5 Results

4.5.1 Classification. The performance of SYN-FUSION was assessed on seven classification datasets. The comparison with GNN baselines using scaffold split is provided in Table 1. SYN-FUSION exhibited superior performance in 5 out of 7 such as ClinTox, HIV, SIDER, and MUV. SYN-FUSION has a relative improvement of (6.63, 0.4)% on BBBP, (-6.2, 4.3)% on Tox21, (19.89, 6.88)% on ClinTox, (4.36, 2.27)% on HIV, (-6.2, 2.4)% on BACE, (7.2, 0.15)% on SIDER, (9.21, 7.75)% on MUV datasets when comparing against its non-fusion counterparts Hu et. al [16] and MolCLR [44] approaches respectively. Also on ClinTox and SIDER datasets, SYN-FUSION outperforms the state-of-the-art model by 4.64% and 2.2% respectively. The results using random split are provided in Table 3. SYN-FUSION
4.5.2 Regression. The performance of SYN-FUSION was assessed on six regression benchmarks and the corresponding results are presented in Table 2. SYN-FUSION surpassed the performance of baseline methods on 4 out of the 6 datasets namely FreeSolv, ESOL, QM7, and QM9. These datasets encompass a diverse range of molecular properties, thereby providing a comprehensive evaluation of the fusion approach’s capabilities. SYN-FUSION has a relative improvement of (-10.6, 25.97)% on FreeSolv, (21.3, 31)% on ESOL, (5.4, 8.86)% on Lipo, (38.74, 29.79)% on QM7, (2.09, 3.2)% on QM8, (55.23, 35.49)% on QM9 datasets when comparing on Hu et al. [16] and the fusion approach’s capabilities. SYN-FUSION has a relative improvement of (-10.6, 25.97)% on FreeSolv, (21.3, 31)% on ESOL, (5.4, 8.86)% on Lipo, (38.74, 29.79)% on QM7, (2.09, 3.2)% on QM8, (55.23, 35.49)% on QM9 datasets when comparing on Hu et al. [16] and MolCLR [44] approaches respectively. Notably, the SYN-FUSION model demonstrated significant improvements over the previous best, achieving a remarkable 4.3% improvement on the ESOL dataset.

Table 2: Regression Results on MoleculeNet datasets using scaffold split. SYN-FUSION approach outperforms baselines in 4 out of 6 datasets. The best score for each dataset is indicated in bold and the second-best score is underlined.

<table>
<thead>
<tr>
<th>Metric</th>
<th>FreeSolv</th>
<th>ESOL</th>
<th>Lipo</th>
<th>QM7</th>
<th>QM8</th>
<th>QM9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>RMSE</td>
<td>MAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecules</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>2.10 ± 0.22</td>
<td>1.07 ± 0.19</td>
<td>0.88 ± 0.04</td>
<td>122.7 ± 4.2</td>
<td>0.0423 ± 0.0021</td>
<td>16.061 ± 0.019</td>
</tr>
<tr>
<td>SVM</td>
<td>3.14 ± 0.00</td>
<td>1.50 ± 0.00</td>
<td>0.82 ± 0.00</td>
<td>156.9 ± 0.00</td>
<td>0.0543 ± 0.0010</td>
<td>24.613 ± 0.144</td>
</tr>
<tr>
<td>GCN [20]</td>
<td>2.87 ± 0.14</td>
<td>1.43 ± 0.05</td>
<td>0.85 ± 0.08</td>
<td>122.9 ± 2.2</td>
<td>0.0366 ± 0.0011</td>
<td>7.596 ± 1.969</td>
</tr>
<tr>
<td>GIN [48]</td>
<td>2.76 ± 0.18</td>
<td>1.45 ± 0.02</td>
<td>0.85 ± 0.07</td>
<td>124.8 ± 0.7</td>
<td>0.0371 ± 0.0009</td>
<td>4.741 ± 0.912</td>
</tr>
<tr>
<td>D-MPNN [51]</td>
<td>2.18 ± 0.91</td>
<td>0.98 ± 0.26</td>
<td><strong>0.65 ± 0.05</strong></td>
<td>105.8 ± 13.2</td>
<td><strong>0.0143 ± 0.0022</strong></td>
<td>3.241 ± 0.119</td>
</tr>
<tr>
<td>Hu et al. [16]</td>
<td>2.83 ± 0.12</td>
<td>1.22 ± 0.02</td>
<td>0.74 ± 0.00</td>
<td>110.2 ± 6.4</td>
<td>0.0191 ± 0.0003</td>
<td>4.349 ± 0.061</td>
</tr>
<tr>
<td>MolCLR_GIN [44]</td>
<td>2.81 ± 0.03</td>
<td>1.29 ± 0.01</td>
<td>0.79 ± 0.00</td>
<td>92.3 ± 1.5</td>
<td>0.0187 ± 0.0012</td>
<td>2.933 ± 0.053</td>
</tr>
<tr>
<td>SYN-FUSION</td>
<td>3.13 ± 0.02</td>
<td>0.96 ± 0.007</td>
<td>0.70 ± 0.01</td>
<td>67.5 ± 1.3</td>
<td>0.0187 ± 0.0041</td>
<td><strong>1.947 ± 0.096</strong></td>
</tr>
<tr>
<td>SYN-FUSION</td>
<td><strong>2.08 ± 0.04</strong></td>
<td><strong>0.89 ± 0.02</strong></td>
<td><strong>0.72 ± 0.01</strong></td>
<td><strong>64.8 ± 1.4</strong></td>
<td><strong>0.0181 ± 0.0001</strong></td>
<td><strong>1.892 ± 0.042</strong></td>
</tr>
</tbody>
</table>

Table 3: Classification and Regression Results using random split. SYN-FUSION approach outperforms baselines on all datasets. The best score for each dataset is indicated in bold and the second-best score is underlined.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ROC-AUC (%)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBBP</td>
<td>2.039</td>
<td>1,513</td>
</tr>
<tr>
<td>BACE</td>
<td>1.513</td>
<td>1,476</td>
</tr>
<tr>
<td>ClinTox</td>
<td>1.476</td>
<td>41,127</td>
</tr>
<tr>
<td>HIV</td>
<td>41,127</td>
<td>1,128</td>
</tr>
<tr>
<td>Tasks</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chemformer [17]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X-Mol [49]</td>
<td>96.0</td>
<td>87.2</td>
</tr>
<tr>
<td>MolBERT [25]</td>
<td>87.5</td>
<td>92.3</td>
</tr>
<tr>
<td>SYN-FUSION</td>
<td><strong>96.5 ± 0.3</strong></td>
<td><strong>90.2 ± 0.4</strong></td>
</tr>
<tr>
<td>SYN-FUSION</td>
<td><strong>95.5 ± 0.2</strong></td>
<td><strong>88.4 ± 0.5</strong></td>
</tr>
</tbody>
</table>

Table 4: Comparison with Transformer-Graph combination networks on classification and regression datasets from MoleculeNet. The best score for each dataset is indicated in bold and the second-best score is underlined.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ROC-AUC (%)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClinTox</td>
<td>94.4 ± 2.1</td>
<td>68.2 ± 1.1</td>
</tr>
<tr>
<td>HIV</td>
<td>68.2 ± 1.1</td>
<td>65.8 ± 2.3</td>
</tr>
<tr>
<td>SIDER</td>
<td>65.8 ± 2.3</td>
<td>1.99 ± 0.07</td>
</tr>
<tr>
<td>FreeSolv</td>
<td>1.99 ± 0.07</td>
<td>1.10 ± 0.18</td>
</tr>
<tr>
<td>ESOL</td>
<td>1.10 ± 0.18</td>
<td>0.82 ± 0.01</td>
</tr>
<tr>
<td>Lipo</td>
<td>0.82 ± 0.01</td>
<td>1.70 ± 0.52</td>
</tr>
<tr>
<td>GROVER [36]</td>
<td><strong>94.4 ± 2.1</strong></td>
<td><strong>68.2 ± 1.1</strong></td>
</tr>
<tr>
<td>Graphormer [52]</td>
<td><strong>88.1 ± 3.8</strong></td>
<td><strong>78.9 ± 0.9</strong></td>
</tr>
<tr>
<td>PharmHGT [18]</td>
<td><strong>94.5 ± 0.4</strong></td>
<td><strong>80.6 ± 0.2</strong></td>
</tr>
<tr>
<td>SYN-FUSION</td>
<td><strong>94.6 ± 1.6</strong></td>
<td><strong>83.7 ± 1.9</strong></td>
</tr>
<tr>
<td>H紫外. al.</td>
<td><strong>94.7 ± 0.2</strong></td>
<td><strong>76.3 ± 1.3</strong></td>
</tr>
</tbody>
</table>
We conducted an extensive analysis of our synergistic fusion approach, SYN-FUSION, to evaluate its performance from both quantitative and qualitative perspectives. The analysis covered various aspects, such as examining its latent space, activation maps, loss interpolation, and weight distribution. All experiments conducted in this section compared SYN-FUSION as the synergistic model with MolCLR in GNN and MegaMolBART in Transformer as its individual models.

5 ANALYSIS ON SYNERGISTIC FUSION

5.1 Latent Space Visualization

Latent space visualization offers valuable insights into the encoded representations learned by a model, enabling a better understanding of the learned distribution and patterns. t-SNE plots were generated to compare the latent space representations of the proposed SYN-FUSION model and its individual components (MolCLR and MegaMolBART) on the ClinTox classification dataset, providing qualitative visualization for comparison as shown in Figure 2, with subfigures 2(a), 2(b), and 2(c) displaying the t-SNE plots of the embeddings derived from SYN-FUSION, MolCLR, and MegaMolBART, respectively. The red and blue points represent the projection of toxic and non-toxic molecule samples respectively. Figure 2(a) has a clear separation between the two classes, as the toxic samples cluster together at the top while the non-toxic molecules appear at the bottom. This observation indicates the successful learning and encoding of discriminative features pertaining to toxic and non-toxic molecules by SYN-FUSION, and the model’s ability to make accurate predictions regarding the toxicity of new molecules. In contrast, Figure 2(b) suggests that the latent representations of toxic and non-toxic molecules in MolCLR are intermingled instead of having a separation. This finding implies that MolCLR may face difficulties in accurately classifying molecules as toxic or non-toxic. MegaMolBART (Figure 2(c)) exhibits improved discrimination between toxic and non-toxic molecules, although there are still scattered instances of toxic molecules among the non-toxic ones, and the level of separation is not as pronounced as in SYN-FUSION. Confusion matrices obtained on evaluation of 1476 molecule samples from ClinTox presented in Figures 3(a), 3(b), and 3(c) indicate that better separation leads to fewer false predictions, and SYN-FUSION made fewer incorrect predictions in distinguishing between toxic and non-toxic molecules when compared to MolCLR and MegaMolBART.

5.2 1-D Activation Maps

Activation maps help to identify similar patterns across samples belonging to the same class and enable the model to distinguish between classes, leading to effective decision-making. To observe any learned patterns between toxic and non-toxic molecules, 100 samples from ClinTox were considered in equal proportions (50 toxic / 50 non-toxic), and 1-D activation maps were generated using the last layer of SYN-FUSION, MolCLR, and MegaMolBART models. Figure 4(a) showcases the stacked 1-D activation maps of SYN-FUSION, revealing distinct and clear activation patterns across samples in both classes, indicating that the model has learned to focus on relevant features for effective classification. In contrast, the activation maps of MolCLR in Figure 4(b) lack well-defined patterns, and it is difficult to differentiate the toxic from the non-toxic class. Activation maps generated by MegaMolBART, as depicted in Figure 4(c), demonstrate intermediate characteristics between the two models, revealing a few discernible patterns that are slightly more noticeable compared to MolCLR but not as prominent as SYN-FUSION.

5.3 Loss Interpolation

Loss interpolation plots offer a concise visualization of the transition between different loss values, providing valuable insights into the behavior and convergence of optimization over the course of training. Notably, the presence of Monotonic Linear Interpolation in a model’s loss trajectory signifies that the optimization of tasks is relatively easier [14]. Notable differences are observed in the loss trajectories of the models under comparison, as shown in Figure 5. Specifically, the loss curve of SYN-FUSION displays a remarkably high level of monotonicity, suggesting a smoother and more consistent optimization process, in contrast to the loss curves of MolCLR and MegaMolBART. Moreover, SYN-FUSION has lower initial and final loss values compared to the other two models, providing additional evidence of easier and better optimization. These findings substantiate the effectiveness of synergistic fusion surpassing individual models in terms of optimization and convergence.

5.4 Weight Histograms

A weight histogram shows the distribution of weights within a model, providing insights into the range and frequency of different weight values. Small weights (values close to 0.0) tend to yield sharper minimizers and exhibit greater sensitivity to perturbations [24]. Conversely, weight distribution with uniform variance (both positive and negative values) leads to flatter minima and contributes to better generalization. In light of this finding, the weight distributions of the final layer of SYN-FUSION, MolCLR, and MegaMolBART were investigated after completion of training, and the histogram of weights is presented in Figure 6. SYN-FUSION produces higher magnitude (both range and density) weights compared to MolCLR and MegaMolBART, indicating that fusion improves generalization and helps in easier and faster optimization. The impact of this phenomenon can be observed in the loss interpolation discussed in Section 5.3 Figure 5 where SYN-FUSION demonstrates a favorable initialization denoted by a lower initial loss value, undergoes a rapid minimization of the loss ending with a significantly lower final loss value.
Figure 2: Latent Space Visualization for SYN-FUSION, MolCLR, and MegaMolBART using t-SNE on ClinTox dataset. The red color represents toxic molecules, while the blue color represents non-toxic molecules. Compared to the other two methods, SYN-FUSION demonstrates a pronounced ability to achieve a clear separation between toxic and non-toxic molecules.

Figure 3: Confusion Matrix of SYN-FUSION, MolCLR, and MegaMolBART on ClinTox dataset. SYN-FUSION demonstrates enhanced efficiency in molecular classification and achieves lower rates of both false negatives and false positives.

Figure 4: Activation Maps of SYN-FUSION, MolCLR, and MegaMolBART on ClinTox dataset. The activation map is split into two parts, and each row of a part comprises a 1D vector of the pre-final layer. The resulting barcode per part is obtained by stacking 50 such samples belonging to the same class. The top part represents the activations of 50 samples from the toxic class while the bottom part contains 50 from the non-toxic class. The activation map of SYN-FUSION unveils notable patterns for both toxic and non-toxic molecules which are highlighted in the color red.

6 ABLATION STUDY

In order to experimentally verify the impact of synergy, we conducted a comparative analysis between the combined effect (represented by SYN-FUSION) and the individual models (MolCLR and MegaMolBART), as well as their (sum) ensemble, on both classification and regression tasks. In the ensemble approach, we handled the predictions generated by each individual model differently depending on the task at hand. For classification, if both models provided identical predictions, we retained the prediction as is. However, in cases where the models offered differing predictions, we considered the prediction with higher confidence. For regression, we computed the average of the two individual model predictions. The results are illustrated in Figure 7. In the absence of SYN-FUSION, the AUC% drops from 94.69% by 5.19%-6.24% on ClinTox, and the RMSE increases from 0.89 by 0.15-0.39 on ESOL. This demonstrates the synergy effect - the combined effect achieved through fusion is greater than the individual models and their ensemble.
Synergistic Fusion of Graph and Transformer Features for Enhanced Molecular Property Prediction

7 DISCUSSION AND CONCLUSION

We present SYN-FUSION, a novel approach that synergistically combines pre-trained features from Graph Neural Networks (GNNs) and Transformers to create a comprehensive molecular representation. Our method effectively captures both the global structure of molecules and the characteristics of individual atoms, addressing the limitations of existing approaches. Experimental results on various molecular benchmarks demonstrate the superior performance of SYN-FUSION compared to previous models in both classification and regression tasks. Furthermore, a detailed analysis of the learned fusion model provides insights into its effectiveness through aspects such as loss, activation, and weight distribution. The conducted ablation study demonstrates that the fusion approach outperforms individual models and their ensemble, offering a substantial improvement in predicting molecular properties. This work contributes to the advancement of molecular representation techniques, providing a promising solution for accurate molecule property prediction and generalization within the vast chemical space. We believe that the presented findings will make a substantial contribution to the field, and hold great potential for applications in drug discovery and chemical research.
Synergistic Fusion of Graph and Transformer Features for Enhanced Molecular Property Prediction

6-12, 2020, virtual. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/94af034441e3a38039beda1a1f9df5d-Abstract.html


