FARM: Enhancing Molecular Representations with Functional Group Awareness

Anonymous authors
Paper under double-blind review

Abstract

We introduce **F**unctional Group-**A**ware **R**epresentations for Small **M**olecules (**FARM**), a novel foundation model designed to bridge the gap between SMILES (a linear string representation of molecular structures), natural language, and molecular graphs. The key innovation of FARM lies in its functional group (FG) annotation at the atomic level, which enables both **FG-enhanced SMILES** and **FG graphs**: SMILES are enriched with FG information to specify which functional group each atom belongs to, while the FG graph captures the molecular backbone by showing how the functional groups are connected. This tokenization not only injects chemical knowledge into SMILES but also expands the chemical lexicon, effectively bridging the gap between SMILES and natural language in terms of vocabulary size, making the sequences more suitable for Transformer-based models. FARM then learns molecules from two complementary perspectives to fully encode both functional and structural information. Masked language modeling on FG-enhanced SMILES captures atom-level features enriched with FG context, while graph neural networks encode higher-level molecular topology by representing how FGs are connected. Contrastive learning then aligns these two views into a unified embedding, ensuring that atom-level details and FG-level structure are jointly represented. This dual-level modeling is central to FARM's ability to predict molecular properties accurately. We rigorously evaluate FARM on the MoleculeNet dataset, achieving state-of-the-art performance on 11 out of 13 tasks, and further validate its generalization on the photostability dataset for quantum mechanics properties. These results highlight FARM's potential to improve molecular representation learning, demonstrate its strong transfer learning capabilities across drug discovery and material design domains, and make way for broad applications in pharmaceutical research and functional materials development. The code is available at: `https://anonymous.4open.science/r/farm-E3CB`.

1 Introduction

Artificial intelligence (AI) has emerged as a transformative tool in accelerating scientific discovery, particularly in drug development and material design. In these domains, it is increasingly employed for tasks such as molecular property prediction, drug–target interaction prediction, quantitative structure–activity relationship (QSAR) modeling, and the discovery of novel materials with tailored electronic, mechanical, or optical properties. (Chen et al., 2016; Wen et al., 2017; Shen and Nicolaou, 2019; Walters and Barzilay, 2020; Achary, 2020; Wang et al., 2022a; Edwards et al., 2022; Zhang et al., 2023a; Nguyen et al., 2024a; Edwards et al., 2024b;a). However, one of the central challenges in this field is the scarcity of large labeled datasets required for traditional supervised learning methods. This has shifted the focus towards self-supervised pre-trained models that can extract meaningful patterns from vast amounts of unlabeled molecular data (Shen and Nicolaou, 2019). As a result, the development of robust foundation models for molecular representations is now more critical than ever.

Despite significant advancements in other domains, such as natural language processing (NLP) and computer vision, there is still no dominant foundation model tailored to molecular representation in drug discovery (Zhang et al., 2023b). This paper begins to address this pressing gap by introducing an innovative *functional group (FG)-aware* approach, in which classical FGs are identified according to established chemical principles, while

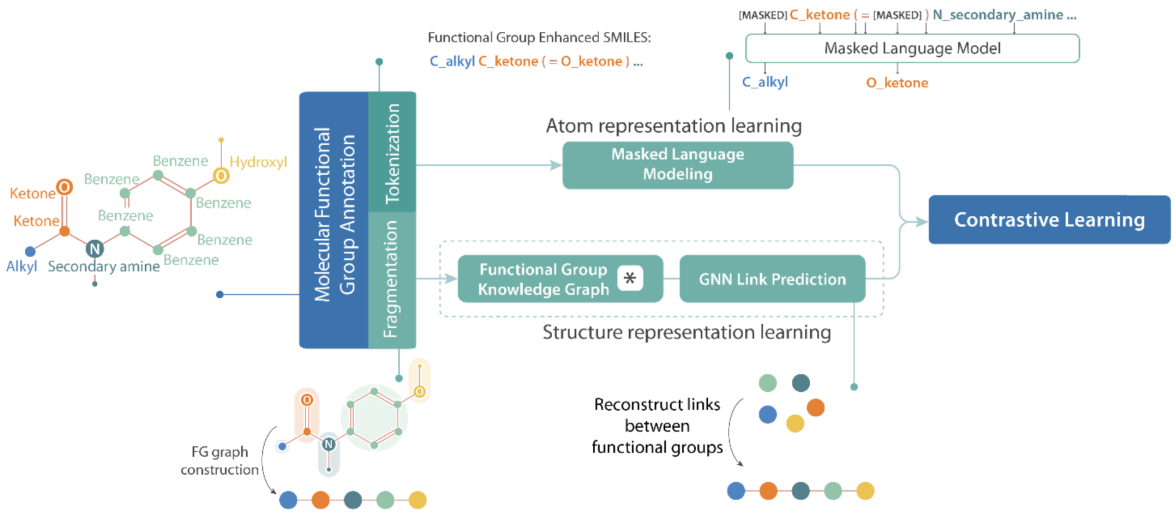

Figure 1: FARM's molecular representation learning framework. Given a molecule as input, we apply FG-aware tokenization to its SMILES representation and FG-based fragmentation to its molecular graph. Each resulting representation is then processed through a self-supervised learning objective. Finally, the two views are aligned using contrastive learning to make the final embedding robust by capturing both atom-level information from the SMILES and structural information from the graph.

also considering additional chemically meaningful substructures that strongly influence molecular properties. Functional groups play a central role in determining the chemical and biological behavior of molecules, and even small changes can lead to significant differences in activity. For instance, salicylic acid and aspirin share the same molecular backbone but differ by a single functional group: the hydroxyl group in salicylic acid is replaced by an acetyl group in aspirin. This subtle modification dramatically alters their properties and biological activities, as the acetyl group in aspirin enables irreversible inhibition of cyclooxygenase (COX) enzymes, producing blood-thinning effects that help prevent heart attacks and strokes, effects not observed with salicylic acid. By rigorously grounding tokenization in **chemical rules**, our approach aims to more effectively capture these critical molecular properties and functions.

In the strictest chemical sense, functional groups refer to well-defined moieties such as hydroxyl, carbonyl, or amino groups. In this work, we adopt a broader interpretation of FGs to include not only these classical groups but also ring systems and fused ring systems, which, although not always labeled as functional groups in textbooks, often play critical roles in molecular recognition, stability, and activity.

The key innovation of our approach lies in how molecules are tokenized. In SMILES representations, we enrich detail by annotating each atom with its FG, while in molecular graph representations, we abstract detail to emphasize the molecular backbone, namely which FGs are present and how they are connected. This dual strategy allows molecular embeddings to more faithfully capture functional information, enabling more accurate prediction of molecular properties and functions.

For example, instead of representing every oxygen atom with the same token ("O"), we assign different tokens based on their role in functional groups, such as "O_ketone" or "O_hydroxyl"(Figure 1). Traditional sequence-based models typically treat atoms only by type and ignore the functional groups they belong to, even though a single atom can play very different roles depending on its chemical context. This is analogous to natural language, where a word like "bank" can mean either a river bank or a financial institution. By specifying an atom's functional group, our approach enriches SMILES with chemically contextual information. We refer to this representation as **FG-enhanced SMILES**, which enriches molecular sequences with chemically meaningful context and expands the vocabulary beyond the limited set of atom symbols in standard SMILES.

This vocabulary expansion is particularly important because standard SMILES have a very limited and repetitive vocabulary, essentially atom and bond types, with the majority being C, O, and N. Training models

on such a restricted corpus often leads to rapid convergence, but the representations capture little about functional or property-related patterns. As a result, downstream performance can suffer and even exhibit negative transfer (Xia et al., 2023). By enriching SMILES with FG annotations, we **enlarge the vocabulary** and bring it closer in spirit to natural language, the domain for which Transformer models are designed. This richer representation helps bridge the gap between SMILES and natural language, thereby mitigating negative transfer and improving the model's ability to learn molecular properties and functions.

While FG-enhanced SMILES captures fine-grained chemical details, sequential representations like SMILES often struggle to represent the overall molecular backbone– the higher level arrangement of functional groups that defines molecular topology. To address this, we construct a **FG graph**, where each node corresponds to a functional group and edges encode their connectivity. For example, a ketone group (C=O) is represented as a single FG node rather than two separate atoms, producing a high-level molecular backbone. Each FG node is assigned a chemically informed embedding, learned from a **FG knowledge graph** that captures both the intrinsic properties of each FG and their relationships with other FGs across molecules. These embeddings are then used to initialize a graph neural network (GNN), which is trained via a **link prediction** task to reconstruct connections between FGs. By predicting which FGs are likely to be connected, the GNN learns the molecule's high-level structure, relationships that are often implicit in SMILES representations.

Finally, **contrastive learning** is employed to align the two views: the **FG-enhanced SMILES** and the **FG graph**, encouraging the model to learn unified representations that are sensitive to chemically meaningful structures. Figure 1 and Figure 2a present an overview of our molecular representation learning model.

In summary, our key contributions include:

- **FG-aware tokenization and fragmentation:** We introduce an FG annotation algorithm that labels each atom in a molecule with the FG to which it belongs. The FG-annotated molecular graph is then used for FG-aware tokenization and fragmentation, producing both FG-enhanced SMILES and FG graphs. This approach enriches each atom with chemical context, bridges the gap between SMILES and natural language through a FG–aware vocabulary, and still captures higher-level structural information in the graph.
- **Structural representation learning:** We construct a FG knowledge graph to obtain robust FG embeddings, and learn molecular structure by predicting the links between FGs, thereby capturing the high-level structural organization of molecules.
- **Atom-feature and structural representation integration:** We combine a masked language model to learn atom-level features with GNNs to capture structural information, and align these two views using contrastive learning.
- **Robustness in downstream tasks:** FARM demonstrates strong transfer learning capabilities, the core goal of pretrained models, outperforming other methods in 11 out of 13 tasks from MoleculeNet, which spans physiology, biophysics, physical chemistry, and quantum physics. Beyond MoleculeNet, FARM further shows robust generalization on quantum mechanics datasets, including the photostability dataset.

2 Related Work

Functional group-Aware Molecular Representations In recent years, there has been growing recognition that incorporating functional group (FG) information into molecular representations can significantly enhance model performance in downstream tasks. Approaches in this domain can be broadly classified into two categories: those leveraging language models (LMs) (Li et al., 2023; Xia et al., 2023) and those utilizing graph neural networks (GNNs) (Zhang et al., 2020; 2021; Yu and Gao, 2022; Yang et al., 2022; Wang et al., 2023; Wu et al., 2023; Han et al., 2023; Fang et al., 2023). Within these categories, methods either employ rule-based functional group detection, relying on predefined chemical rules to identify FGs, or adopt unsupervised strategies that infer substructures or motifs from the data. Regardless of the approach, these models consistently demonstrate that enriching molecular representations with functional group information leads to improved performance across a wide range of molecular property prediction tasks (Fang et al., 2023; Han et al., 2023; Wang et al., 2023). This underlines the importance of functional group awareness in achieving accurate and generalizable molecular representations.

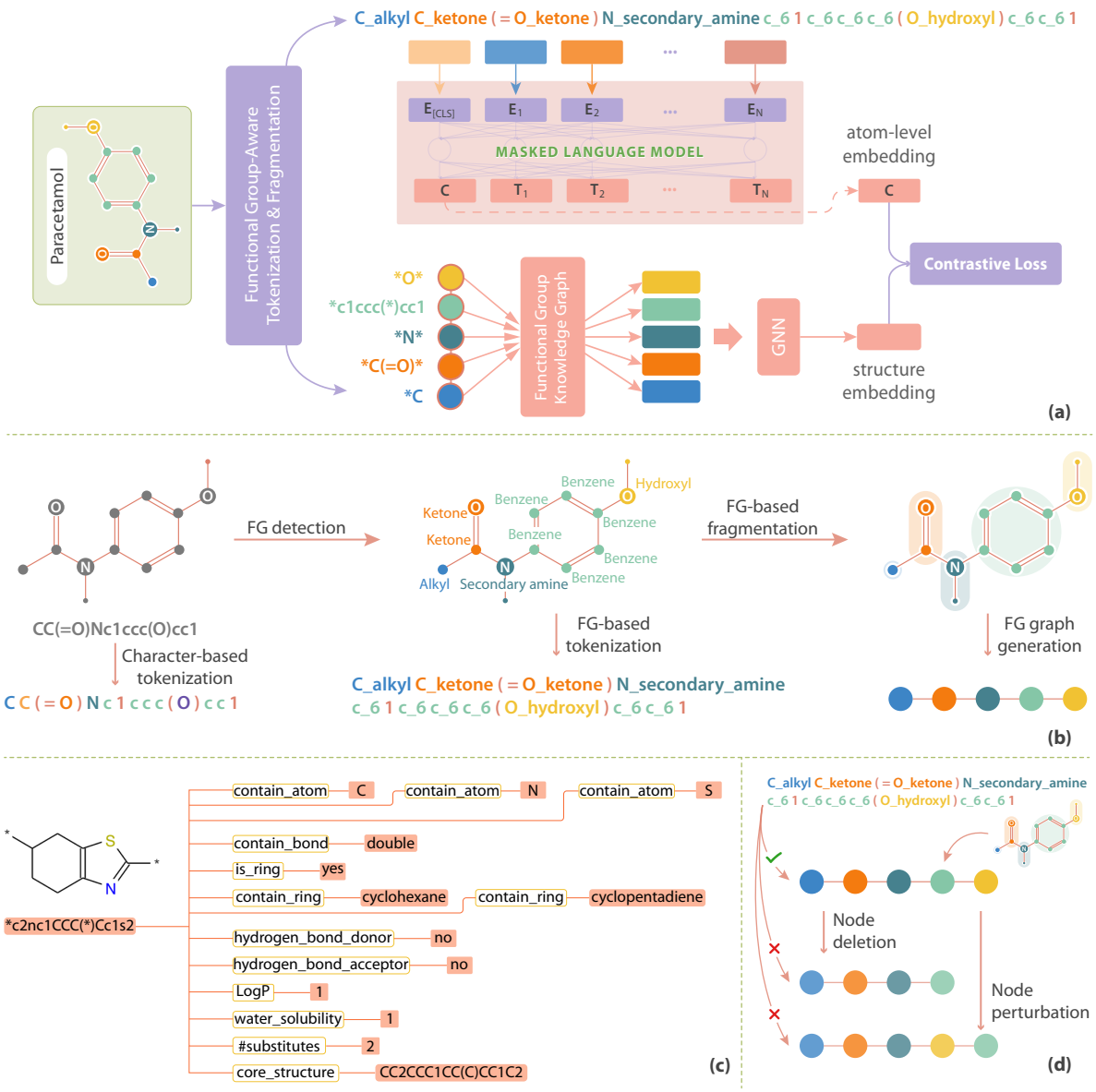

Figure 2: (a) FARM's molecular representation learning model architecture. (b) Functional group-aware tokenization and fragmentation algorithm. (c) Snapshot of the functional group knowledge graph. (d) Generation of negative samples for contrastive learning.

In studies such as (Zhang et al., 2021; Han et al., 2023; Chen et al., 2024; Yang et al., 2022), the BRICS algorithm (Degen et al., 2008) is employed to fragment molecules, with some extending the approach through additional rules to achieve finer-grained segmentation. However, this method focuses on general reaction-based bond breaking, rather than targeting functional group-specific characteristics. Additionally, while BRICS groups atoms into fragments, it does not explicitly identify or label functional groups, limiting its utility in tasks that require functional groups-specific information. Other works (Li et al., 2023; Chen et al., 2024) leverage RDKit (Landrum, 2010) for functional group detection. While RDKit is effective in identifying common functional groups, its capabilities are limited to well-known groups and do not extend to detecting more complex functional groups, such as ring systems that often dominate molecular datasets and play a critical role in molecular function. In this work, we present a rule-based functional group detection algorithm that accurately identifies 101 common groups and all ring-containing ones. Unlike frequent subgraph mining or motif-based methods, our approach is explicitly grounded in chemical principles, ensuring reliable and chemically meaningful detection aligned with expert definitions.

Contrastive Learning-Based Molecular Representations Contrastive learning, a self-supervised learning technique, aims to learn representations by maximizing agreement between positive pairs while distinguishing them from negative samples. In the context of molecular representation, (Wang et al., 2022b) applies this technique by augmenting each molecular graph to create slightly different versions, which are then treated as negative examples to enhance the learning process. In (Pinheiro et al., 2022), the authors use a molecular graph encoder and a SMILES encoder to encode both molecular graphs and SMILES, employing contrastive loss to maximize the agreement between these embeddings. This method enriches SMILES with topology information and the molecular graph with sequence context, making the final embeddings of molecules more robust. In (Zhang et al., 2020), the authors minimize the distance between the representation of a molecule and the representations of its constituent substructures. Collectively, these works highlight the versatility of contrastive learning in unifying diverse molecular representations, leading to improved downstream performance in molecular tasks. Our method extends this approach by using contrastive learning to maximize the agreement between FG-enhanced SMILES representations and FG graphs, thereby enhancing the alignment between sequence-based and graph-based representations, ensuring a more comprehensive and chemically informed embedding.

Algorithm 1 Functional group annotation for molecular graph

Require: Molecular graph G, functional group definitions FG_list (ordered largest to smallest)
for each atom v in G **do**
  if v does not have a functional group assigned **then**
    $FG \leftarrow$ largest functional group in FG_list that matches v and its neighborhood
    $matched_atoms \leftarrow$ atoms involved in the match
    for each atom a in $matched_atoms$ **do**
      a.SetProp("FunctionalGroup", FG.name)
    end for
  end if
end for

3 Methodology

We present our approach for enhancing molecular representation learning through FG-enhanced SMILES learning with Transformers, FG graph backbone modeling, and dual-perspective contrastive learning, jointly capturing atom-level chemical context and FG-level molecular structure. Section 3.1 introduces FG-aware tokenization and fragmentation, which injects detailed chemical context into both sequence and graph representations. Section 3.2 delves into the masked atom prediction task, a self-supervised technique that enables the model to learn atom-level representations. In Section 3.3, we describe our method for capturing the core molecular structure. Finally, Section 3.4 presents contrastive learning, which aligns FG-enhanced SMILES strings with FG graph embeddings to achieve a comprehensive, unified molecular representation.

3.1 Functional Group-Aware Tokenization and Fragmentation

We propose an FG-aware tokenization and fragmentation method that incorporates detailed functional group information into molecular representations, designed for both SMILES and graph-based models. Our approach defines a set of functional groups and uses an algorithm to detect them within the molecular graph, annotating each atom with its corresponding FG. The algorithm traverses the molecular graph, evaluating each atom based on criteria such as atom type, neighboring atom types and bonds, number of neighbors, atom charge, and bonded hydrogen atoms to identify its functional group. This approach ensures precise detection that strictly aligns with the chemical definitions of functional groups within molecular graphs. The pseudocode is provided in Algorithm 1. For instance, a carbon atom with a charge of 0 and three neighbors, including a double-bonded oxygen atom, is classified as part of a ketone group (RCOR'), with the oxygen also contributing to the group. Additionally, we address cases where a functional group may be a subset of another. The algorithm first checks for the presence of the larger functional group. If it's not identified, the

Table 1: Evaluation of the FG-enhanced SMILES lexicon size across various databases. *Min frequency* indicates the minimum number of occurrences required for a token to be included in the lexicon.

Database	**#molecules**	**#atom types**	**Length**	Min frequency = 1	Min frequency = 5
ZINC15	3M	10	[5; 63]	1,151	1,089
ChEMBL25	1.8M	35	[1; 867]	15,269	10,016
Collected dataset	20M	46	[1; 867]	22,364	14,741

algorithm then checks for smaller functional groups, ensuring correct identification even in complex structures. Appendix B.1 provides a full list of 101 non-ring-containing functional groups.

The results of the FG annotation algorithm are utilized for two distinct processes:

- **FG-aware tokenization:** The molecular graph, where each node (atom) is assigned to a specific functional group, is converted back into a SMILES string that incorporates functional group information. For example, a ketone-containing group originally depicted in SMILES as **C(=O)** is transformed into an FG-enhanced SMILES string like **C_ketone(=O_ketone)**. This FG-enhanced SMILES embeds additional chemical context directly into the molecular representation while remaining fully compliant with traditional SMILES rules (if the FG information is removed, the FG-enhanced SMILES reverts to its standard SMILES form).
- **FG-aware fragmentation:** Once FGs are identified, the molecule is segmented according to the bonds connecting these groups, as illustrated in Figure 2(b). These bonds serve as edges in a graph, with nodes corresponding to the functional groups. This representation, referred to as the FG graph, captures the molecule's structural backbone.

3.2 Atom-Level Representation Learning

We employ a masked language model architecture that takes FG-enhanced SMILES as input, using atom-level tokenization and leveraging masked atom prediction as a self-supervised task to train the model. Specifically, we adopt BERT model (Devlin, 2018) with the following loss:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P_\theta(x_i \mid x_{\setminus i}),$$

where $\mathcal{M}$ denotes the set of indices corresponding to the masked tokens, x_i is the original token at position i, and $x_{\setminus i}$ represents the input sequence with the token at position i masked. To evaluate the impact of masking, we conduct experiments with varying masking percentages and test the model's performance on six downstream tasks. The results indicate that a masking percentage of 35% yields the highest average performance across tasks. Detailed results and information about the training process are provided in Appendix D.1.

We examine the attention mechanism of the BERT model trained with FG-enhanced SMILES by visualizing the attention scores for a query atom (Figure 3). The attention map reveals that the model pays more attention to atoms that are strongly connected to the query atom than to those that are merely adjacent in the SMILES string. In detail, the query atom at position 23 shows higher attention to the atom at position 0, which is part of the same ring, rather than to the atom at position 26, which is closer in the SMILES string but not directly connected. This demonstrates that the model effectively learns the syntax and semantics of SMILES, capturing the underlying molecular structure rather than merely relying on the linear sequence of SMILES.

3.3 Molecular Structure Learning

To learn the structure of molecules, we employ a two-step process. First, we derive **FG embeddings** from the **FG knowledge graph**. Next, we use these embeddings as inputs for a GNN **link prediction model** to predict interactions between FGs. This approach effectively captures the relationships among FGs, both in terms of their structural characteristics and interactions, thereby facilitating the implicit learning of molecular structure.

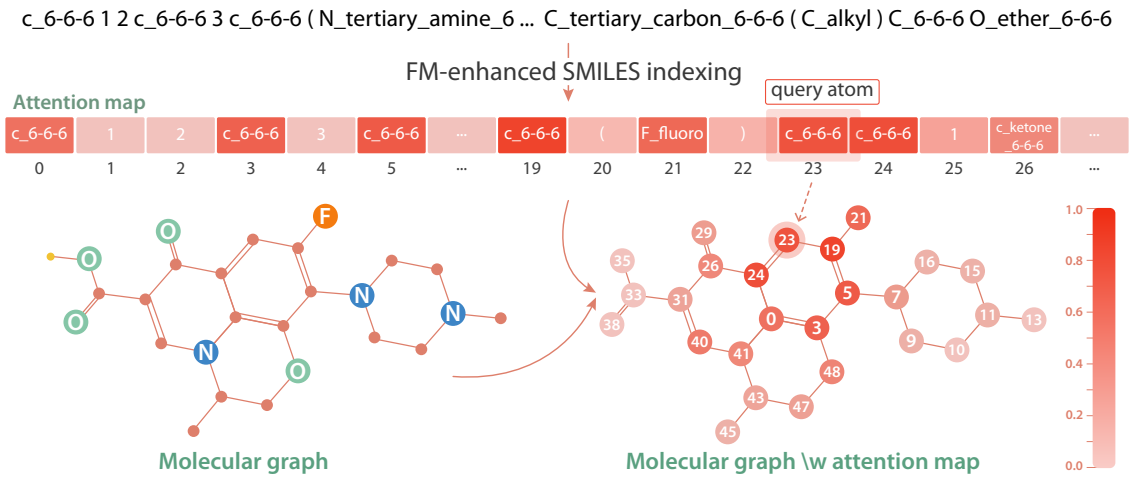

Figure 3: Visualization of the attention map of the BERT model trained with functional group-enhanced SMILES, demonstrating the model's ability to learn and recognize the syntax of SMILES.

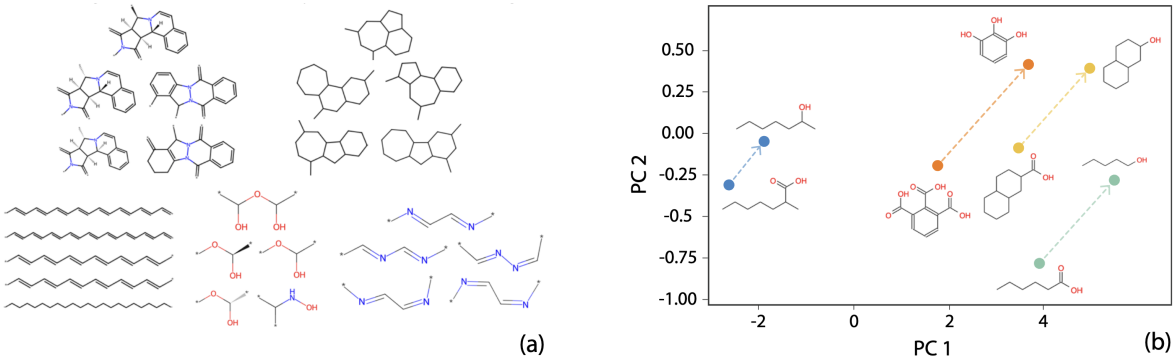

Figure 4: (a) Visualization of functional group knowledge graph embedding space: Clusters of five functional groups with closely related embeddings. (b) Link prediction performance: Substituting one functional group in a molecule with another generates parallel results across different molecules.

FG knowledge graph. The FG knowledge graph models each FG with various relations, such as the atoms, bonds, and ring structures it contains, as well as properties like water solubility, lipophilicity (logP), and more. A comprehensive list of FG knowledge graph relations is provided in the Appendix B.1. A snapshot of the FG knowledge graph is shown in Figure 2(c).

The FG knowledge graph embedding is learned by the ComplEx model (Trouillon et al., 2016) to obtain node embeddings. ComplEx is a matrix factorization model specifically designed to learn embeddings from multi-relational data. It is particularly effective at capturing complex, asymmetric relationships and handling one-to-many relations in knowledge graphs, making it well-suited for modeling intricate relational structures. Specifically, each element in a triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ — where $\mathbf{h}$ is the head entity, $\mathbf{r}$ is the relation, and $\mathbf{t}$ is the tail entity — is represented as a complex vector. The score for a given triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ is calculated as: $f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \mathrm{Re}\left(\mathbf{h}^T \mathbf{r} \cdot \mathbf{t}\right)$. ComplEx employs a margin-based ranking loss function defined as:

$$\mathcal{L}_{\text{Graph}} = \sum_{\substack{(\mathbf{h},\mathbf{r},\mathbf{t}) \\ \in E^+}} \sum_{\substack{(\mathbf{h}',\mathbf{r},\mathbf{t}') \\ \in E^-}} \max\left(0, \gamma + f(\mathbf{h}', \mathbf{r}, \mathbf{t}') - f(\mathbf{h}, \mathbf{r}, \mathbf{t})\right) \tag{1}$$

where E^+ denotes the set of positive triples, E^- denotes the set of negative triples, and γ represents the margin. Optimizing this loss function will minimize the score of positive triples $(\mathbf{h}, \mathbf{r}, \mathbf{t})$, while maximizing the score of negative triples $(\mathbf{h}', \mathbf{r}, \mathbf{t}')$, with a margin γ separating the two. This is achieved through margin-based ranking loss, where the function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ evaluates the plausibility of the triples. The goal is to ensure that the score for positive triples is higher than that for negative triples by at least the margin γ, thus pushing

Table 2: Performance comparison of FARM and baseline models on MoleculeNet classification tasks. The first 10 models incorporate FG (functional group) information to enhance representation learning.

	Physiology					Biophysics		
Dataset	**BBBP**	**Tox21**	**ToxCast**	**SIDER**	**ClinTox**	**BACE**	**MUV**	**HIV**
#tasks	*1*	*12*	*617*	*27*	*2*	*1*	*17*	*1*
#samples	*2,039*	*7,831*	*8,575*	*1,427*	*1,478*	*1,513*	*93,807*	*41,127*
Evaluation Metric	*ROC-AUC (%)* (↑)					*ROC-AUC (%)* (↑)		
MICRO (Zhang et al., 2020)	84.4 ± 1.1	77.0 ± 0.8	65.2 ± 0.8	56.7 ± 0.9	77.0 ± 2.0	77.2 ± 2.0	-	75.1 ± 1.1
MGSSL (Zhang et al., 2021)	69.7 ± 0.9	76.5 ± 0.3	64.1 ± 0.7	61.8 ± 0.8	80.7 ± 2.1	79.1 ± 0.9	78.7 ± 1.5	78.8 ± 1.2
MoleOOD (Yang et al., 2022)	71.0 ± 0.8	-	-	63.4 ± 0.7	-	84.3 ± 1.1	-	79.4 ± 0.5
MCM (Wang et al., 2023)	90.0 ± 3.1	80.2 ± 1.5	-	62.7 ± 2.8	65.5 ± 14	82.0 ± 5.5	-	-
HimGNN (Han et al., 2023)	92.8 ± 2.7	80.7 ± 1.7	-	64.2 ± 2.3	91.7 ± 3.0	85.6 ± 3.4	-	-
FG-BERT (Li et al., 2023)	70.2 ± 0.9	78.4 ± 0.8	63.3 ± 0.8	64.0 ± 0.7	83.2 ± 1.6	84.5 ± 1.5	75.3 ± 2.4	77.4 ± 1.0
HiMol (SMALL) (Zang et al., 2023)	71.3 ± 0.6	76.0 ± 0.2	-	62.5 ± 0.3	70.6 ± 2.1	84.6 ± 0.2	-	-
HiMol (LARGE) (Zang et al., 2023)	73.2 ± 0.8	76.2 ± 0.3	-	61.3 ± 0.5	80.8 ± 1.4	84.3 ± 0.3	-	-
Mole-BERT (Xia et al., 2023)	71.9 ± 1.6	76.8 ± 0.5	62.8 ± 1.1	62.8 ± 1.1	78.9 ± 3.0	80.8 ± 1.4	78.6 ± 1.8	78.2 ± 0.8
MolCLR (Wang et al., 2022b)	73.6 ± 0.5	79.8 ± 0.7	-	**68.0 ± 1.1**	**93.2 ± 1.7**	89.0 ± 0.3	78.9 ± 2.3	88.6 ± 2.2
GLAD (Nguyen et al., 2024b)	80.4 ± 1.5	-	-	64.7 ± 1.8	87.3 ± 1.2	85.7 ± 0.9	-	-
N-GRAM (Hu et al., 2019)	70.8 ± 1.5	78.7 ± 0.4	66.5 ± 0.3	62.7 ± 0.8	72.6 ± 1.5	84.5 ± 0.7	81.3 ± 2.1	79.9 ± 0.7
GROVER (Rong et al., 2020)	86.8 ± 2.2	80.3 ± 2.0	65.3 ± 0.5	61.2 ± 2.5	70.3 ± 13.7	82.4 ± 3.6	67.3 ± 1.8	68.2 ± 1.1
GraphMVP (Liu et al., 2021)	72.4 ± 1.6	75.9 ± 0.5	63.1 ± 0.4	63.1 ± 0.4	79.1 ± 2.8	81.2 ± 0.9	77.7 ± 0.6	77.0 ± 1.2
GEM (Fang et al., 2022)	88.8 ± 0.4	78.1 ± 0.4	69.2 ± 0.4	63.2 ± 1.5	90.3 ± 0.7	87.9 ± 1.1	75.3 ± 1.5	81.3 ± 0.3
UniMol (Zhou et al., 2023)	72.9 ± 0.6	79.6 ± 0.5	69.6 ± 0.1	65.9 ± 1.3	**91.9 ± 1.8**	85.7 ± 0.2	82.1 ± 1.3	82.8 ± 0.3
FARM (Ours)	**93.3 ± 0.2**	**80.8 ± 1.1**	**69.9 ± 0.5**	**65.9 ± 0.7**	82.2 ± 0.7	**89.6 ± 0.4**	**82.7 ± 2.1**	**83.5 ± 0.5**

positive triples closer and negative triples farther apart in the embedding space. The detailed implementation of ComplEx is described in Appendix D.2.

By embedding FGs through the knowledge graph, the model can capture both structural and property-based features of each FG, leading to richer FG representations. Figure 4(a) illustrates clusters of FGs in the FG knowledge graph embedding space, showing that similar FGs are closely positioned, suggesting effective structural and property-based grouping.

Link prediction. For link prediction using a graph convolutional network (GCN), the FG graph is used as input. We employ embeddings from the FG knowledge graph as node features for the GCN. During training, node embeddings are computed through graph convolution:

$$\mathbf{h}_i' = \mathrm{ReLU}\left(\mathbf{W} \cdot \frac{1}{|\tilde{\mathcal{N}}(i)|} \sum_{j \in \tilde{\mathcal{N}}(i)} \mathbf{h}_j\right)$$

where $\mathbf{h}_i'$ is the updated embedding for node i, computed by averaging the embeddings $\mathbf{h}_j$ of its neighbors $\tilde{\mathcal{N}}(i)$, including itself due to the added self-loop, applying the weight matrix $\mathbf{W}$, and passing through a ReLU activation function.

The score estimates the probability of connections between nodes is computed with a multi-layer perceptron (MLP) and a sigmoid function: $p_{ij} = \sigma(\mathrm{MLP}(\mathbf{h}_i \oplus \mathbf{h}_j))$.

We then sample positive edges E^+ and negative edges E^-, optimizing the model to maximize scores for positive edges and minimize those for negative ones using the loss function:

$$\mathcal{L}_{\text{Link}} = -\frac{1}{|E^+|} \sum_{(i,j) \in E^+} \log p_{ij} - \frac{1}{|E^-|} \sum_{(i,j) \in E^-} \log(1 - p_{ij})$$

Figure 4(b) and Figure 8 (Appendix D.3) demonstrates the capability of our molecular structure representation model, showing that, akin to word pair analogy tasks in NLP, replacing one functional group in a molecule with another (in this case, replacing -OH with -COOH) produces parallel results across different molecules. This demonstrates the model's ability to effectively capture and preserve chemical analogies, highlighting its robustness in learning and representing molecular structures.

Table 3: Performance comparison of FARM and baseline models on MoleculeNet regression tasks. Performance is evaluated by RMSE and MAE.

	Physical chemistry			Quantum mechanics	
Dataset	**ESOL**	**FreeSolv**	**Lipophilicity**	**QM8**	**QM9**
#tasks	*1*	*1*	*1*	*12*	*3*
#samples	*1,128*	*642*	*4,200*	*21,786*	*133,885*
Evaluation Metric	*RMSE* (↓)			*MAE* (↓)	
HimGNN (Han et al., 2023)	0.870 ± 0.154	1.921 ± 0.474	0.632 ± 0.016	-	-
FG-BERT (Li et al., 2023)	0.944 ± 0.025	-	0.655 ± 0.009	-	-
Mole-BERT (Xia et al., 2023)	1.015 ± 0.003	-	0.676 ± 0.002	-	-
N-GRAM (Hu et al., 2019)	1.100 ± 0.030	2.510 ± 0.191	0.880 ± 0.121	0.0320 ± 0.003	0.00964 ± 0.00031
GROVER (Rong et al., 2020)	1.423 ± 0.288	2.947 ± 0.615	0.823 ± 0.010	0.0182 ± 0.001	0.00719 ± 0.00208
GEM (Fang et al., 2022)	0.813 ± 0.028	1.748 ± 0.114	0.674 ± 0.022	0.0163 ± 0.001	0.00562 ± 0.00007
MolCLR (Wang et al., 2022b)	1.113 ± 0.023	2.301 ± 0.247	0.789 ± 0.009	0.0185 ± 0.013	0.00480 ± 0.00003
UniMol (Zhou et al., 2023)	0.788 ± 0.029	1.480 ± 0.048	**0.603 ± 0.010**	0.0156 ± 0.001	0.00467 ± 0.00004
FARM (Ours)	**0.761 ± 0.031**	**1.097 ± 0.033**	0.778 ± 0.005	**0.0146 ± 0.001**	**0.00456 ± 0.00001**

3.4 Molecular Structure Integration via Contrastive Learning

To integrate atom-level representations with molecular structure information, we employ contrastive learning to align FG-enhanced SMILES representations with FG graph embeddings. This method captures both the atom-level and topological aspects of molecular structures, allowing the model to develop a unified representation that integrates chemical context with overall molecular architecture.

In this framework, each molecule is treated as a pair of representations: the FG-enhanced SMILES and its corresponding FG graph. The contrastive learning task encourages the embeddings of these two representations to be as similar as possible for the same molecule, while pushing apart the representations of different molecules. This allows the model to capture both local chemical features (from the FG-enhanced SMILES) and global molecular topology (from the FG graph).

To enhance the learning process and make it more robust, we generate negative examples by augmenting the FG graph. We apply two types of augmentations: (1) node deletion, where one or more functional groups are removed from the graph, and (2) node swapping, where functional groups are randomly exchanged with one another. Figure 2(d) illustrates how these augmentations are applied to generate negative examples from a FG graph. These augmentations create harder negative examples that force the model to better understand the correct structure and connectivity between functional groups, making it more effective at learning meaningful molecular representations.

Specifically, given a positive pair $(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{pos}})$, where $\mathbf{h}_{\text{MLM}}$ is the atom-level representation derived from a pretrained BERT model and $\mathbf{h}_{\text{pos}}$ is the corresponding structure representation from a graph neural network (GNN), and a negative pair $(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{neg}})$, where $\mathbf{h}_{\text{neg}}$ is a augmented FG-graph, the contrastive loss can be written as:

$$\mathcal{L}_{\text{CL}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K}\max\left(0,\ \gamma - \cos\left(\mathbf{h}_{\text{MLM}}^{(i)}, \mathbf{h}_{\text{pos}}^{(i)}\right) + \cos\left(\mathbf{h}_{\text{MLM}}^{(i)}, \mathbf{h}_{\text{neg}}^{(j)}\right)\right)$$

where γ is the margin parameter and N is the number of training examples (or contrastive pairs). The final objective function for integration is: $\mathcal{L}_{\text{Integration}} = \lambda_{\text{MLM}} \cdot \mathcal{L}_{\text{MLM}} + \lambda_{\text{CL}} \cdot \mathcal{L}_{\text{CL}}$, where λ_{MLM} and λ_{CL} are hyperparameters that control the contribution of each loss to the overall objective. Details on the implementation of contrastive training can be found in the Appendix D.4.

The effect of the contrastive learning objective can be observed in Figure 4(b) and Figure 8. The learned embedding space exhibits meaningful chemical relationships between molecules. The learned embedding space exhibits meaningful chemical relationships between molecules. In particular, similar to the parallel relationships observed in word embeddings in natural language processing, such as *king : queen* being analogous to *man : woman* or *husband : wife*, substituting one functional group with another produces consistent directional shifts in the molecular embedding space across different molecules. This suggests that functional group substitutions correspond to approximately parallel transformations in the learned

representation space, indicating that the model captures systematic chemical relationships between functional groups and their structural contexts.

4 Pre-training Data Collection and Diversity Assessment

The performance and generalizability of a machine learning model is heavily dependent on the quality and diversity of its training data. Given the vastness of chemical space, it is vital that the training data represents a sufficiently representative subset of this space. Traditional approaches assess the diversity of datasets based on criteria such as the number of chemical elements and the number of atoms per molecule. However, to the best of our knowledge, no previous studies have systematically evaluated the diversity of large chemical datasets like ZINC (Irwin and Shoichet, 2005) or ChEMBL (Gaulton et al., 2011) in terms of functional groups. This gap is significant, as functional groups provide deep insights into the structural and functional complexity of molecules.

In this work, we introduce a novel criterion to assess dataset diversity by using the size of the FG-enhanced SMILES lexicon as a metric. By analyzing the size of this lexicon, we can assess whether the dataset captures a comprehensive range of chemical functionalities, which is crucial for building robust molecular foundation models. Table 1 presents the size of the FG-enhanced SMILES lexicon for the ZINC15 and ChEMBL25 datasets, as well as the dataset we collected for training our foundation model. The table shows that despite its widespread use for training foundation models in small molecule representation, ZINC15 is significantly less diverse compared to ChEMBL25. This is largely due to the fact that ZINC15 is primarily designed to include molecules that adhere to Lipinski's Rule of Five for drug-likeness, excluding more exotic or less common elements that are less relevant to pharmaceutical chemistry. This limited diversity may negatively impact the model's performance on out-of-distribution datasets, which contain a broader range of atom types and functional groups not present in ZINC15. In contrast, ChEMBL25 is far more diverse and thus better suited for training foundation models that can be fine-tuned for various downstream tasks. Based on this insight, we collected a dataset that includes the entire ChEMBL25 database, libraries from chemical drug suppliers, and a subset of ZINC15 to ensure a more comprehensive coverage of chemical space. Detailed information about the collected data can be found in the Appendix A.

Table 4: Performance of FARM on the photostability dataset for quantum mechanics property prediction. R^2 scores are reported with standard deviation over 10 runs.

Class	$\mathbf{R^2}$	**Class**	$\mathbf{R^2}$
HOMOm1 (eV)	0.927 ± 0.0148	T5	0.905 ± 0.0315
HOMO (eV)	0.895 ± 0.0298	S1	0.936 ± 0.0161
LUMO (eV)	0.928 ± 0.00948	S2	0.948 ± 0.0156
LUMOp1 (eV)	0.922 ± 0.0280	S3	0.911 ± 0.00849
DipoleMoment (Debye)	0.755 ± 0.0532	S4	0.919 ± 0.00792
LogP	0.995 ± 0.00566	S5	0.924 ± 0.0143
PrimeExcite (eV)	0.877 ± 0.0491	O1	0.839 ± 0.0340
PrimeExcite (osc)	0.837 ± 0.0243	TDOS2.0	0.868 ± 0.0990
T1	0.929 ± 0.00587	TDOS2.5	0.877 ± 0.00410
T2	0.925 ± 0.0182	TDOS3.0	0.815 ± 0.0447
T3	0.900 ± 0.0101	TDOS3.5	0.895 ± 0.0327
T4	0.921 ± 0.0196	TDOS4.0	0.919 ± 0.0313
		TDOS4.5	0.938 ± 0.0183

5 Experimental Results

5.1 Data, Splits and Evaluation Metrics

MoleculeNet We consider 13 benchmark tasks in the MoleculeNet dataset (Wu et al., 2018). Following previous works, the data is split using a scaffold split into training, validation, and test sets with an 8:1:1 ratio, ensuring fair and consistent evaluation across all models.

We use ROC-AUC as the evaluation metric for classification tasks due to the high imbalance in some datasets. For physical chemistry tasks (ESOL, FreeSolv, and Lipophilicity), we use RMSE, and for quantum mechanics tasks (QM8 and QM9), we use MAE, following previous works. For each downstream task, we split the data using three random seeds, train the models, and report the average and standard deviation of the results.

D-B-A Oligomer Photostability Dataset (Angello et al., 2024) This dataset comprises a chemical library of light-harvesting donor–bridge–acceptor (D–B–A) oligomers, constructed using a modular approach with 22 donor–bridge blocks and 100 acceptor blocks. This design results in a theoretical chemical space of 2,200 unique molecules, which were synthesized and experimentally characterized to study photostability in solution. We adopt a hard-case scaffold split: 3 donor–bridge blocks and 15 acceptor blocks are held out to create the test set, including all molecules containing any of the held-out donors or acceptors, ensuring that no molecules in the training or validation sets share the same donor or acceptor structures. A similar procedure is used to construct the validation set, and the process is repeated for ten random seeds to ensure robust evaluation across multiple train/validation/test splits.

The tasks include the prediction of frontier molecular orbital energies (HOMO, LUMO, HOMO-1, LUMO+1), dipole moment, logarithm of the partition coefficient (LogP), excitation energies (PrimeExcite), oscillator strengths (PrimeExcite (osc)), triplet and singlet excitation energies (T1–T5, S1–S5), and time-dependent density of states (TDOS 2.0–4.5). We use R^2 as the evaluation metric to quantify the accuracy of predicted properties against experimental or high-level computational values. Models are trained on the training set and validated on the held-out validation set, with performance reported on the test set across all ten random scaffold splits to ensure robustness and assess generalization to unseen donor or acceptor blocks.

5.2 Baselines

MoleculeNet We consider works that incorporate functional group information to enhance representation learning (Zhang et al., 2020; 2021; Yang et al., 2022; Wang et al., 2023; Han et al., 2023; Li et al., 2023; Zang et al., 2023; Wang et al., 2022b; Xia et al., 2023; Nguyen et al., 2024b), and other approaches that utilize masked atom prediction as a self-supervised training task (Rong et al., 2020; Hu et al., 2019; Liu et al., 2021; Fang et al., 2022; Zhou et al., 2023).

D-B-A Oligomer Photostability For this dataset, we do not find any prior benchmark studies, as most foundation models for small molecules have focused exclusively on MoleculeNet. We include this dataset to evaluate our model on a diverse set of tasks and to assess its generalizability beyond standard benchmarks.

5.3 Results

MoleculeNet Table 2 and 3 present the performance of FARM alongside other baseline models on seven MoleculeNet classification tasks. FARM consistently outperforms baseline models across the benchmarks, demonstrating its robustness and versatility. Notably, FARM achieves particularly strong performance on BBBP and BACE, which are often more learnable because these tasks involve well-defined functional group patterns associated with blood–brain barrier penetration and beta-secretase inhibition, respectively—patterns that FG-aware tokenization captures effectively. FARM also excels on solubility-related tasks such as Lipo, FreeSolv, and ESOL, likely due to its ability to model atom-level chemical context and functional groups that strongly influence solubility. Overall, its strong performance across both classification and regression tasks highlights FARM as a highly effective pretrained model, well-suited for a broad range of downstream applications in molecular property prediction.

D-B-A Oligomer Photostability We evaluate FARM on the D–B–A oligomer photostability dataset across 25 property prediction tasks (results shown in Table 4). The model achieves a lowest R^2 of 0.815 for TDOS 3.0 and a highest R^2 of 0.995 for LogP, with 16 out of 25 tasks exceeding R^2 values of 0.9. These tasks are particularly challenging not only due to the hard-case scaffold split, in which the training set contains no molecules sharing donor or acceptor blocks with the validation and test sets, but also due to the nature of the targets, including difficult-to-predict properties such as dipole moment, excited-state energies (S1–S5 and T1–T5), and total density of states (TDOS). Despite these challenges, FARM demonstrates consistently high performance across all properties. These results highlight FARM's ability to accurately model complex

Table 5: Performance comparison of the model using BRICS fragmentation versus our FG-based fragmentation across six MoleculeNet tasks.

	BBBP	**BACE**	**HIV**	**Average**	**ESOL**	**FreeSolv**	**Average**	**QM9**
#tasks	*1*	*1*	*1*		*1*	*1*		*3*
#samples	*2039*	*1513*	*41127*		*1128*	*642*		*133885*
Metric	*ROC-AUC* (↑)				*RMSE* (↓)			*MAE* (↓)
BRICS	88.87 ± 1.5	89.54 ± 1.09	70.09 ± 2.24	82.83	1.340 ± 0.235	0.823 ± 0.14	1.081	0.0085 ± 0.00034
OURS	96.23 ± 0.7	96.19 ± 0.65	82.13 ± 1.10	**91.43**	0.734 ± 0.039	0.308 ± 0.08	**0.521**	**0.0038 ± 0.00014**

quantum mechanical behaviors by leveraging functional group information and capturing nuanced molecular interactions, underscoring its robustness and generalizability for real-world chemical property prediction tasks.

5.4 Ablation Study on FARM Component Analysis

To assess the contribution of each component in our architecture—FG-enhanced SMILES modeling via BERT, FG embeddings from the FG knowledge graph, and molecule backbone modeling using FG graphs with GNN link prediction—we conducted a comprehensive ablation study across several MoleculeNet benchmark tasks (results shown in Table 10). For additional details on the baselines, please refer to Appendix E. The results demonstrate that each component of FARM improves performance over the baseline, and the combination of all components yields the strongest overall performance across all test tasks.

Ablation Study on Fragmentation Methods We compare our FG-based fragmentation method with BRICS fragmentation. When treating each BRICS fragment as a functional group, the FG-enhanced SMILES vocabulary expands to 575,612 tokens (min frequency ≤ 5), compared to just 14,741 tokens with our chemically grounded FG detection method. This difference arises because BRICS generates larger, less reusable fragments, leading to a bloated vocabulary with many rare tokens, which complicates training and increases the risk of out-of-vocabulary issues. As shown in Table 5, our FG-based method consistently outperforms BRICS across all six MoleculeNet tasks under the same setup.

6 Conclusion, Limitations, and Future Work

In summary, FARM demonstrates robust performance across various MoleculeNet classification tasks, outperforming or matching baseline models. The integration of functional group information and the alignment of FG-enhanced SMILES representations with FG graph embeddings through contrastive learning significantly enhance its effectiveness. While FARM shows strong performance, there is a key limitation that should be addressed in future work. The current model does not incorporate a full 3D molecular representation, which is critical for capturing stereochemistry and spatial configurations that influence molecular properties. Integrating 3D structural information (Yan et al., 2024) could further improve the model's ability to make accurate predictions.

Looking ahead, our ultimate goal is to develop a pre-trained atom embedding that parallels the capabilities of pre-trained word embeddings in NLP. This would enable a richer and more nuanced understanding of molecular properties and behaviors at the atomic level. Similarly, we aim to achieve molecule-level representations that are as expressive and versatile as sentence-level embeddings in NLP, capturing both local and global molecular features. By bridging the gap between atom-wise embeddings and holistic molecule representations, FARM paves the way for more accurate, generalizable molecular predictions across a variety of tasks.

References

Patnala GR Achary. Applications of quantitative structure-activity relationships (qsar) based virtual screening in drug design: A review. *Mini Reviews in Medicinal Chemistry*, 20:1375–1388, 2020.

Nicholas H Angello, David M Friday, Changhyun Hwang, Seungjoo Yi, Austin H Cheng, Tiara C Torres-Flores, Edward R Jira, Wesley Wang, Alán Aspuru-Guzik, Martin D Burke, et al. Closed-loop transfer enables artificial intelligence to yield chemical knowledge. *Nature*, 633(8029):351–358, 2024.

Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, and Yongdong Zhang. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*, 17: 696–712, 2016.

Yongqiang Chen, Quanming Yao, Juzheng Zhang, James Cheng, and Yatao Bian. Hight: Hierarchical graph tokenization for graph-language alignment. *arXiv preprint arXiv:2406.14021*, 2024.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3:1503, 2008.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proc. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022)*, 2022.

Carl Edwards, Aakanksha Naik, Tushar Khot, Martin Burke, Heng Ji, and Tom Hope. Synergpt: In-context learning for personalized drug synergy prediction and drug design. In *Proc. 1st Conference on Language Modeling (COLM2024)*, 2024a.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. L+m-24: Building a dataset for language + molecules acl 2024. In *Proc. ACL 2024 Workshop on Language+Molecules*, 2024b.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5 (5):542–553, 2023.

A Gaulton, L Bellis, J Chambers, M Davies, A Hersey, Y Light, S McGlinchey, R Akhtar, F Atkinson, AP Bento, et al. Chembl: A large-scale bioactivity database for chemical biology and drug discovery. *Nucleic Acids Research. Database*, page D1, 2011.

Shen Han, Haitao Fu, Yuyang Wu, Ganglan Zhao, Zhenyu Song, Feng Huang, Zhongfei Zhang, Shichao Liu, and Wen Zhang. Himgnn: a novel hierarchical molecular graph representation learning framework for property prediction. *Briefings in Bioinformatics*, 24(5):bbad305, 2023.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45:177–182, 2005.

G Landrum. Rdkit: Open-source cheminformatics. `https://www.rdkit.org`, 2010. Accessed: 2024-09-19.

Biaoshun Li, Mujie Lin, Tiegen Chen, and Ling Wang. Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics*, 24(6):bbad398, 2023.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Thao Nguyen, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, and Heng Ji. Glad: Synergizing molecular graphs and language descriptors for enhanced power conversion efficiency prediction in organic photovoltaic devices. In *Proc. 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*, 2024a.

Thao Nguyen, Tiara Torres-Flores, Changhyun Hwang, Carl Edwards, Ying Diao, and Heng Ji. Glad: Synergizing molecular graphs and language descriptors for enhanced power conversion efficiency prediction in organic photovoltaic devices. *33rd ACM International Conference on Information and Knowledge Management*, 2024b.

Gabriel A Pinheiro, Juarez LF Da Silva, and Marcos G Quiles. Smiclr: contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *Journal of Chemical Information and Modeling*, 62(17):3948–3960, 2022.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

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

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54:263–270, 2020.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin Burke. Chemical-reaction-aware molecule representation learning. In *Proc. The International Conference on Learning Representations (ICLR2022)*, 2022a.

Yifei Wang, Shiyang Chen, Guobin Chen, Ethan Shurberg, Hang Liu, and Pengyu Hong. Motif-based graph representation learning with application to chemical molecules. In *Informatics*, volume 10, page 8. MDPI, 2023.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b.

Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16:1401–1409, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5312–5320, 2023.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9:513–530, 2018.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63:8749–8760, 2019.

Keqiang Yan, Xiner Li, Hongyi Ling, Carl Ashen, Kenna; Edwards, Raymundo Arroyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Qian Xiaoning, and Shuiwang Ji. Invariant tokenization for language model enabled crystal materials generation. In *Proc. the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS2024)*, 2024.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978, 2022.

Zhaoning Yu and Hongyang Gao. Molecular representation learning via heterogeneous motif graph neural networks. In *International Conference on Machine Learning*, pages 25581–25594. PMLR, 2022.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.

Shichang Zhang, Ziniu Hu, Arjun Subramonian, and Yizhou Sun. Motif-driven contrastive learning of graph representations. *arXiv preprint arXiv:2012.12533*, 2020.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, YuQing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. In *arxiv*, 2023a.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023b.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

A Molecular Datasets

A.1 Training data

We collected a diverse dataset to train our **FARM** model from various sources, including ChEMBL25, ZINC15, and several chemical suppliers. The number of compounds in each dataset is reported as follows:

Table 6: List of compound suppliers and number of compounds

Supplier	Number of Compounds	Source
Targetmol	22,555	https://www.targetmol.com/
Chemdiv	1,741,620	https://www.chemdiv.com/
Enamine	862,698	https://enamine.net/
Life Chemical	347,657	https://lifechemicals.com/
Chembridge	1,405,499	https://chembridge.com/
Vitas-M	1,430,135	https://vitasmlab.biz/
InterBioScreen	560,564	https://www.ibscreen.com/
Maybridge	97,367	https://chembridge.com/
Asinex	601,936	https://www.asinex.com/
Eximed	61,281	https://eximedlab.com/
Princeton BioMolecular	1,647,078	https://princetonbio.com/
Otava	9,203,151	https://www.otava.com/
Alinda Chemical	733,152	https://www.alinda.ru/synthes_en.html
ChEMBL 25	1,785,415	https://www.ebi.ac.uk/chembl/
ZINC15	4,000,000	https://zinc15.docking.org/
Total	20,000,000	

A.2 Lexicon Growth for FG-enhanced SMILES Dataset

As part of our investigation into the relationship between corpus size and vocabulary size, we conducted experiments to observe how vocabulary growth evolves as the dataset increases. Our results show that while the vocabulary size increases with the corpus size, it eventually stabilizes. Specifically, when expanding the corpus from 20 million tokens to 200 million tokens, the vocabulary grew from 14,741 to 42,364 tokens, and with a 1 billion-token corpus, the vocabulary reached a plateau of 42,391 tokens. This indicates that the lexicon tends to converge around 42.3K tokens as the corpus size increases, suggesting that further growth in corpus size may not result in substantial increases in vocabulary.

Even with the smaller vocabulary of 14,741 tokens derived from the 20 million-token dataset, the model was able to achieve excellent coverage of common tokens. In fact, the 14,741 tokens from the 20M dataset accounted for 99.997% of all tokens in the 200M dataset, with 2,932,883,078 out of 2,932,984,142 tokens already represented. This demonstrates that a relatively smaller vocabulary can effectively capture the majority of the data, even with larger corpora. Moreover, the ratio of the UNKNOWN token in downstream datasets remains consistently below 0.1

A.3 Downstream Tasks Data

In Table 7, we provide an overview of the datasets used for evaluating the performance of our model on various downstream tasks. Each dataset is denoted by its name, followed by the number of tasks it encompasses, the total number of samples available in each dataset, and a brief description. These datasets cover a range of chemical and biological properties, enabling comprehensive evaluation of the model's performance across different tasks in molecular representation learning.

B FG-aware Tokenization and Fragmentation

B.1 The List of Functional Groups

The exhaustive list of 101 functional groups that can be detected by the functional group detection algorithm includes: Tertiary carbon, Quaternary carbon, Alkene carbon, Cyanate, Isocyanate, Hydroxyl, Ether, Hydroperoxy, Peroxy, Haloformyl, Aldehyde, Ketone, Carboxylate, Carboxyl, Ester, Hemiacetal, Acetal, Hemiketal, Ketal, Orthoester,

Table 7: Overview of downstream tasks, corresponding sample sizes, and dataset descriptions.

Dataset	# Tasks	# Samples	Description
BBBP	1	2,039	Benchmark for Blood-Brain Barrier permeability prediction, assessing whether compounds can cross the blood-brain barrier.
Tox21	12	7,831	Toxicology data containing multiple assays for evaluating the toxicity of compounds across various endpoints.
SIDER	27	1,427	Side Effect Resource dataset that includes drug side effects associated with FDA-approved drugs, focusing on adverse drug reactions.
ClinTox	2	1,478	Clinical Toxicology dataset designed to predict the toxicity of drug-like compounds based on clinical data.
BACE	1	1,513	Data for predicting activity against the beta-secretase enzyme, relevant for Alzheimer's disease drug discovery.
MUV	17	93,807	Multiple Unrelated Variables dataset aimed at assessing the ability to predict various molecular properties and activities.
HIV	1	41,127	Dataset focused on predicting the activity of compounds against the HIV virus, crucial for antiviral drug development.
ESOL	1	1,128	Dataset used for estimating the solubility of organic compounds in water, useful for understanding compound behavior in biological systems.
FreeSolv	1	642	Dataset containing free energy of solvation values for small organic molecules in water, aiding in solvation energy predictions.
Lipophilicity	1	4,200	Data focused on predicting the octanol-water partition coefficient, a key measure of a compound's lipophilicity.
QM8	12	21,786	Quantum Mechanics dataset that provides a range of molecular properties computed using quantum mechanical methods for small organic molecules.
QM9	3	133,885	Quantum Mechanics dataset providing molecular properties for a large set of small organic compounds.

Carbonate ester, Orthocarbonate ester, Amidine, Carbamate, Isothiocyanate, Thioketone, Thial, Carbothioic S-acid, Carbothioic O-acid, Thiolester, Thionoester, Carbodithioic acid, Carbodithio, Trifluoromethyl, Difluorochloromethyl, Bromodifluoromethyl, Trichloromethyl, Bromodichloromethyl, Tribromomethyl, Dibromofluoromethyl, Triiodomethyl, Difluoromethyl, Fluorochloromethyl, Dichloromethyl, Chlorobromomethyl, Chloroiodomethyl, Dibromomethyl, Bromoiodomethyl, Diiodomethyl, Alkyl, Alkene, Alkyne, Carboxylic anhydride, Primary amine, Secondary amine, Amide, Imide, Tertiary amine, 4-ammonium ion, Hydrazone, Primary ketimine, Primary aldimine, Secondary ketimine, Secondary aldimine, Nitrile, Azide, Azo, Nitrate, Isonitrile, Nitrosooxy, Nitro, Nitroso, Aldoxime, Ketoxime, Sulfhydryl, Sulfide, Disulfide, Sulfinyl, Sulfonyl, Sulfur dioxide, Sulfuric acid, Sulfino, Sulfonic acid, Sulfonate ester, Thiocyanate, Phosphino, Phosphono, Phosphate, Phosphodiester, Phosphoryl, Borono, Boronate, Borino, Borinate, Silyl ether, Dichlorosilane, Trimethylsilyl, Fluoro, Chloro, Bromo, Iod.

B.2 Naming functional groups with Rings in Fused Ring Systems

Fused ring systems are a diverse and prevalent class of functional groups, accounting for 99.37% of the total functional groups in our dataset (147,564 out of 148,507 FGs). Despite their importance, many of these systems lack standardized nomenclature. To address this, we propose a systematic approach to naming these ring systems based on their ring sizes and core structures.

Each ring in a fused ring system is named according to its size. For instance, a six-membered aromatic ring like benzene is named ring_6. This straightforward approach provides a clear identifier for individual rings within a system. For systems composed of multiple fused rings, we use the following naming convention:

- **Identification:** Determine the smallest atom index for each ring within the system.
- **Sorting:** Arrange the rings by increasing atom indices.
- **Construction:** Combine the ring sizes in ascending order. For example, a fused system with a six-membered ring and a five-membered ring would be named ring_5_6.

This systematic naming helps in identifying and categorizing complex fused ring systems by focusing on their core structure. The core structure is defined as the central framework of interconnected rings that forms the fundamental backbone of the molecule. The core structure of a ring system is important because it influences the molecule's reactivity, stability, and biological activity. In SMILES notation, which uses lowercase characters to indicate atoms within aromatic rings, we can enhance the representation by combining the atom symbol (uppercase or lowercase) with the core structure, thereby providing a comprehensive depiction of the ring system. Figure 5a illustrates an example

of naming a fused ring system based on the rules described above, and Figure 5b shows how FG-aware tokenization is applied.

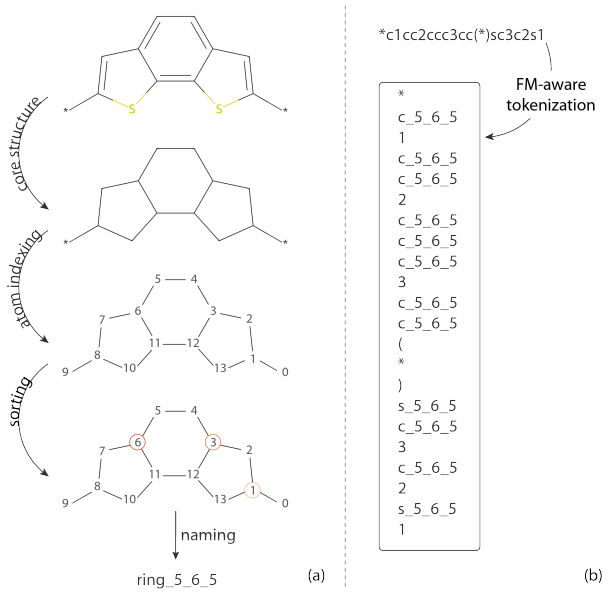

Figure 5: (a) Example of naming a fused ring system in 4 steps: generate the core structure of the functional group, index atoms using RDKit, select the smallest-index atom in each ring and sort, and name the fused ring system based on ring size. (b) Example of FG-aware tokenization.

After completing the naming process, we derive a new FG-enhanced SMILES representation for the molecules. We then analyze our collected dataset, which comprises 20 million samples of FG-enhanced SMILES, to evaluate the results. This dataset includes representations of 46 different elements. Notably, 11 elements are represented by only a single form, indicating their rare occurrence within the dataset (excluding hydrogen). These elements are: H, Ti, V, Cr, Rb, Mo, Rh, Sb, Ba, Pb, and Bi. In contrast, the remaining 35 elements feature at least two representations, each corresponding to distinct FGs. The distribution of these elements is visualized in Figure 6, highlighting the diversity of representations in our dataset. The most prevalent element in our dataset is Carbon, with 9,112 FGs containing it. Nitrogen follows as the second most prevalent element, represented in 2,549 FGs, while Oxygen and Sulfur appear in 2,156 and 571 FGs, respectively.

C FG Knowledge Graph

The FG knowledge graph is designed to capture both the structural and property-related information of FGs. The list of relations includes:

- **List of functional groups that act as hydrogen bond donors:** Hydroxyl, Hydroperoxy, Primary amine, Secondary amine, Hydrazone, Primary ketimine, Secondary ketimine, Primary aldimine, Amide, Sulfhydryl, Sulfonic acid, Thiolester, Hemiacetal, Hemiketal, Carboxyl, Aldoxime, Ketoxim.
- **List of functional groups that act as hydrogen bond acceptors:** Ether, Peroxy, Haloformyl, Ketone, Aldehyde, Carboxylate, Carboxyl, Ester, Ketal, Carbonate ester, Carboxylic anhydride, Primary amine, Secondary amine, Tertiary amine, 4-Ammonium ion, Hydrazone, Primary ketimine, Secondary ketimine, Primary aldimine, Amide, Sulfhydryl, Sulfonic acid, Thiolester, Aldoxime, Ketoxi.

D Implementation Details

D.1 Training Masked Language Model for SMILES Representation

We trained the BERT model using Hugging Face (Wolf et al., 2020) on the masked molecule prediction task with both conventional SMILES and FG-enhanced SMILES from our collected dataset. To assess the impact of different

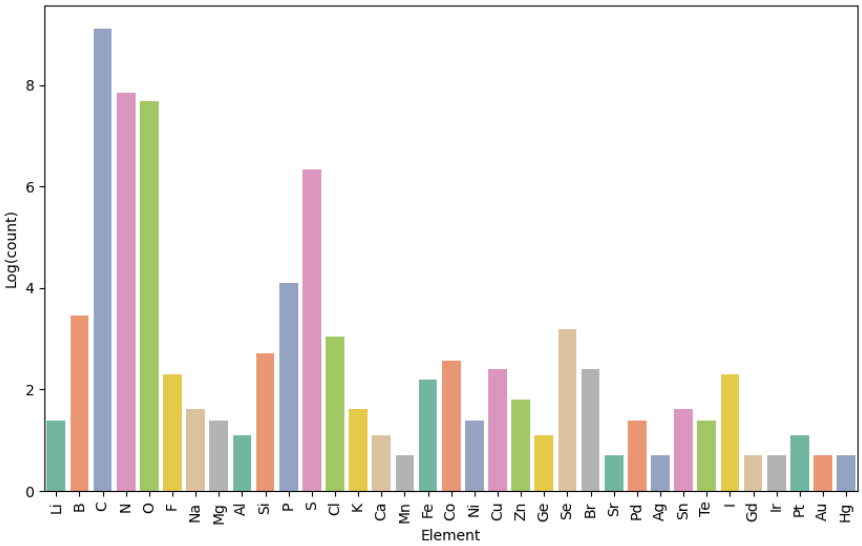

Figure 6: Number of functional groups associated with different chemical elements in the FG-enhanced SMILES dataset. The y-axis represents the natural logarithm (log, base e) of the count.

Table 8: Key relations defined in the FG knowledge graph. (Note: Continuous values, such as LogP and water solubility, are discretized by rounding to the nearest integer.)

Relation	**Description**
contain_atom	Identifies atoms present in the FG (e.g., C, H, O, N).
contain_bond	Specifies types of bonds in the FG (e.g., single, double, triple, aromatic).
functional_group	Recognizes functional groups in the FG (e.g., hydroxyl, carboxyl, amine).
contain_ring_[n]	Indicates the presence of a non-aromatic ring of size n in the FG.
contain_aromatic_ring_[n]	Indicates the presence of an aromatic ring of size n in the FG.
num_substitutes	Specifies the number of substituents (e.g., alkyl or aryl groups) in the FG.
is_hydrogen_bond_donor	Identifies whether the FG contains a functional group capable of donating hydrogen bonds.
is_hydrogen_bond_acceptor	Identifies whether the FG contains a functional group capable of accepting hydrogen bonds.
logp	Measures the lipophilicity of the FG using the logP value (calculated via RDKit). In the collected dataset, values range from -35 to 31.
water_solubility	Predicts the solubility of the FG in water, based on logP, molecular weight, and TPSA. In the collected dataset, values range from -5 to 8.
core_smiles	The SMILES representation of the core structure of the FG.

masking percentages, we trained BERT models with masking percentages of 0.15, 0.25, 0.35, 0.45, and 0.55. The models were then evaluated on seven MoleculeNet tasks, including three classification tasks and four regression tasks, to determine the optimal masking percentage. The results, presented in Table 9, indicate that a masking percentage of 0.35 yields the best performance across the considered downstream tasks.

Table 9: Performance of BERT models with varying masking percentages across six MoleculeNet tasks. The data is split using a random split into training, validation, and test sets with an 8:1:1 ratio.

	BBBP	**BACE**	**HIV**	**Average**	**ESOL**	**FreeSolv**	**Average**	**QM9**
#tasks	*1*	*1*	*1*		*1*	*1*		*3*
#samples	*2039*	*1513*	*41127*		*1128*	*642*		*133885*
Metric	*ROC-AUC (↑)*				*RMSE (↓)*			*MAE (↓)*
0.25	93.01 ± 0.9	94.31 ± 1.08	80.17 ± 1.5	89.16	0.688 ± 0.033	0.622 ± 0.007	0.655	0.0091 ± 0.00001
0.25	93.59 ± 1.7	93.94 ± 1.4	81.03 ± 1.9	89.52	0.543 ± 0.030	0.714 ± 0.010	0.629	**0.0032 ± 0.00001**
0.35	94.36 ± 0.5	94.54 ± 0.4	81.93 ± 1.7	**90.27**	0.608 ± 0.031	0.507 ± 0.030	**0.558**	0.0041 ± 0.00001
0.45	93.48 ± 1.3	94.36 ± 0.90	80.12 ± 1.7	89.32	0.795 ± 0.028	0.493 ± 0.008	0.644	0.0048 ± 0.00001
0.55	92.85 ± 1.1	88.68 ± 1.0	79.89 ± 0.90	87.14	0.734 ± 0.030	0.599 ± 0.005	0.667	0.0097 ± 0.00001

Additional details of the training setup include training the BERT model on 20 million SMILES for 15 epochs using two NVIDIA Tesla V100 GPUs. The learning rate was set to $1e-5$, with a batch size of 128, and model checkpoints were saved after every 10,000 batches. This setup was also applied to the baseline model, which used conventional SMILES for comparison.

Figure 7 illustrates the convergence behavior of the models trained on different representations of molecular data. The model utilizing FG-enhanced SMILES exhibits a slower convergence rate, attributed to the increased complexity of its vocabulary, reflecting its closer resemblance to natural language. The SMILES model converges by step 200 (after processing 25,600 SMILES), while the FG-enhanced SMILES model achieves convergence by step 300 (after processing 38,400 SMILES). Notably, despite the larger prediction vocabulary (14,714 vs. 93), the FG-enhanced model ultimately reaches a lower loss, suggesting its enhanced capacity to capture intricate molecular representations and improve generalization in complex tasks. This indicates the model's ability to leverage functional group information effectively, potentially leading to better performance in downstream applications.

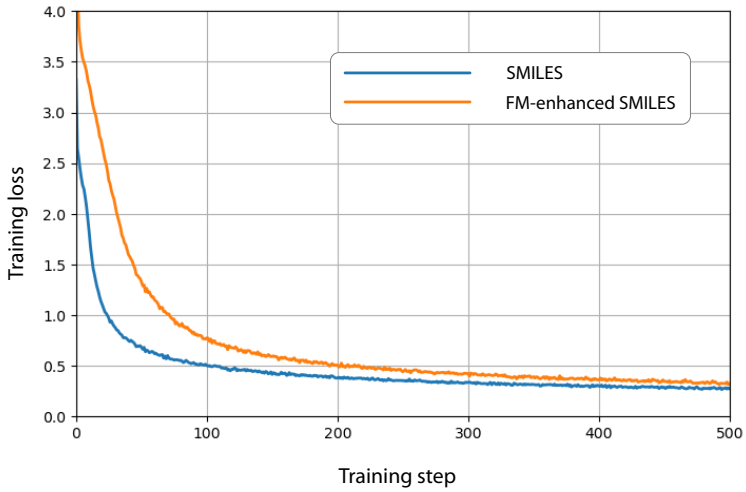

Figure 7: Loss curves for the masked language model (MLM) during training on two datasets: standard SMILES and functional group-enhanced SMILES.

D.2 Training FG Knowledge Graph Embedding Model for Molecular Structure Representation

Once the FG knowledge graph is constructed as detailed in Section C, we utilize the ComplEx model to learn embeddings for the functional groups. The knowledge graph comprises 148,507 unique nodes: 147,564 corresponding to ring systems and 943 representing non-ring functional groups. Training is conducted with a batch size of 64, a learning rate of 1×10^{-3}, over 50 epochs, with model checkpoints saved at the end of each epoch.

ComplEx Model Representation

In the ComplEx model (Trouillon et al., 2016), each element in a triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ — where $\mathbf{h}$ is the head entity, $\mathbf{r}$ is the relation, and $\mathbf{t}$ is the tail entity — is represented as a complex vector:

$$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d \tag{2}$$

Scoring Function

The score for a given triple $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ is calculated as:

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \text{Re}\left(\mathbf{h}^T \mathbf{r} \cdot \mathbf{t}\right) \tag{3}$$

where $\mathbf{r}$ is a complex-valued vector, and the dot product is performed in the complex space.

Loss Function

ComplEx employs a margin-based ranking loss function defined as:

$$\mathcal{L}_{\text{Graph}} = \sum_{(h,r,t)\in E^+} \sum_{(\mathbf{h}',\mathbf{r},\mathbf{t}')\in E^-} \max\left(0, \gamma + f(\mathbf{h}',\mathbf{r},\mathbf{t}') - f(\mathbf{h},\mathbf{r},\mathbf{t})\right) \tag{4}$$

where E^+ denotes the set of positive triples, E^- denotes the set of negative triples, and γ represents the margin.

To assess the quality of the learned embeddings, we randomly sample clusters of five closely related embedding vectors and analyze their arrangement in the embedding space. The results of this evaluation are presented in Figure 4a.

D.3 Link Prediction Model Using GNNs

For link prediction using the GCN model, we start by segmenting molecules into functional groups via FG-aware molecular segmentation, where each group is connected by single bonds. We then use embeddings from the FG knowledge graph embedding model as node features for the GCN. The training process involves computing node embeddings through graph convolution (Equation 5), followed by scoring potential edges with a multi-layer perceptron (MLP) (Equation 6). This score is used to calculate the probability between two nodes (Equation 7). Positive and negative edges are sampled, and the model is optimized to maximize scores for positive edges while minimizing scores for negative edges using the loss function in Equation 8. This approach effectively trains the model to distinguish between likely and unlikely connections between functional groups.

$$\mathbf{h}_i' = \text{ReLU}\left(\mathbf{W} \cdot \frac{1}{|\mathcal{N}(i)|} \sum_{j\in\mathcal{N}(i)} \mathbf{h}_j\right) \tag{5}$$

where $\mathbf{h}_i'$ is the updated embedding for node i. It is computed by averaging the embeddings $\mathbf{h}_j$ of neighboring nodes $\mathcal{N}(i)$, applying the weight matrix $\mathbf{W}$, and then passing through the ReLU activation function.

$$s_{ij} = \text{MLP}(\mathbf{h}_i \oplus \mathbf{h}_j) \tag{6}$$

where s_{ij} denotes the score assigned to the potential edge between nodes i and j. The score is computed using a multi-layer perceptron (MLP), which takes as input the concatenated node embeddings of i and j, denoted as $\mathbf{h}_i \oplus \mathbf{h}_j$. Here, $\mathbf{h}_i$ and $\mathbf{h}_j$ represent the node embeddings for nodes i and j, respectively. The operator $\oplus$ indicates the concatenation of these embeddings. The MLP processes this concatenated vector to produce a score that reflects the likelihood of an edge existing between i and j.

$$p_{ij} = \sigma(s_{ij}) \tag{7}$$

where σ is the sigmoid function.

$$\mathcal{L}_{\text{Link}} = -\frac{1}{|E^+|} \sum_{(i,j)\in E^+} \log p_{ij} - \frac{1}{|E^-|} \sum_{(i,j)\in E^-} \log(1 - p_{ij}) \tag{8}$$

where $\mathcal{L}$ is the loss function for link prediction. It computes the average log-likelihood of positive edges E^+ and negative edges E^-, where p_{ij} is the predicted probability of an edge between nodes i and j. The loss penalizes the model for incorrect predictions, encouraging high probabilities for true edges and low probabilities for false edges.

The GCN model for link prediction is trained as follows: For each molecule, represented as a FG graph, we generate all possible combinations of nodes, encompassing both positive pairs (nodes that are linked) and negative pairs (nodes that are not linked). In cases where the graph contains more than three nodes (FGs), we select 60% of all possible combinations along with all positive pairs to form the training data for each graph. The model is subsequently trained for three epochs on a comprehensive dataset consisting of 20 million data points. Figure 8 shows the performance of the link prediction model. Similar to word embedding analogies in NLP, replacing one FG in a molecule with another produces parallel results across different molecules, demonstrating the model's ability to capture chemical relationships effectively.

D.4 Contrastive Learning: Align SMILES and Structure Representation

In this work, we propose a contrastive learning strategy to align SMILES-based representations of molecules with their corresponding graph-based molecular structures. The goal of this approach is to capture both the sequential

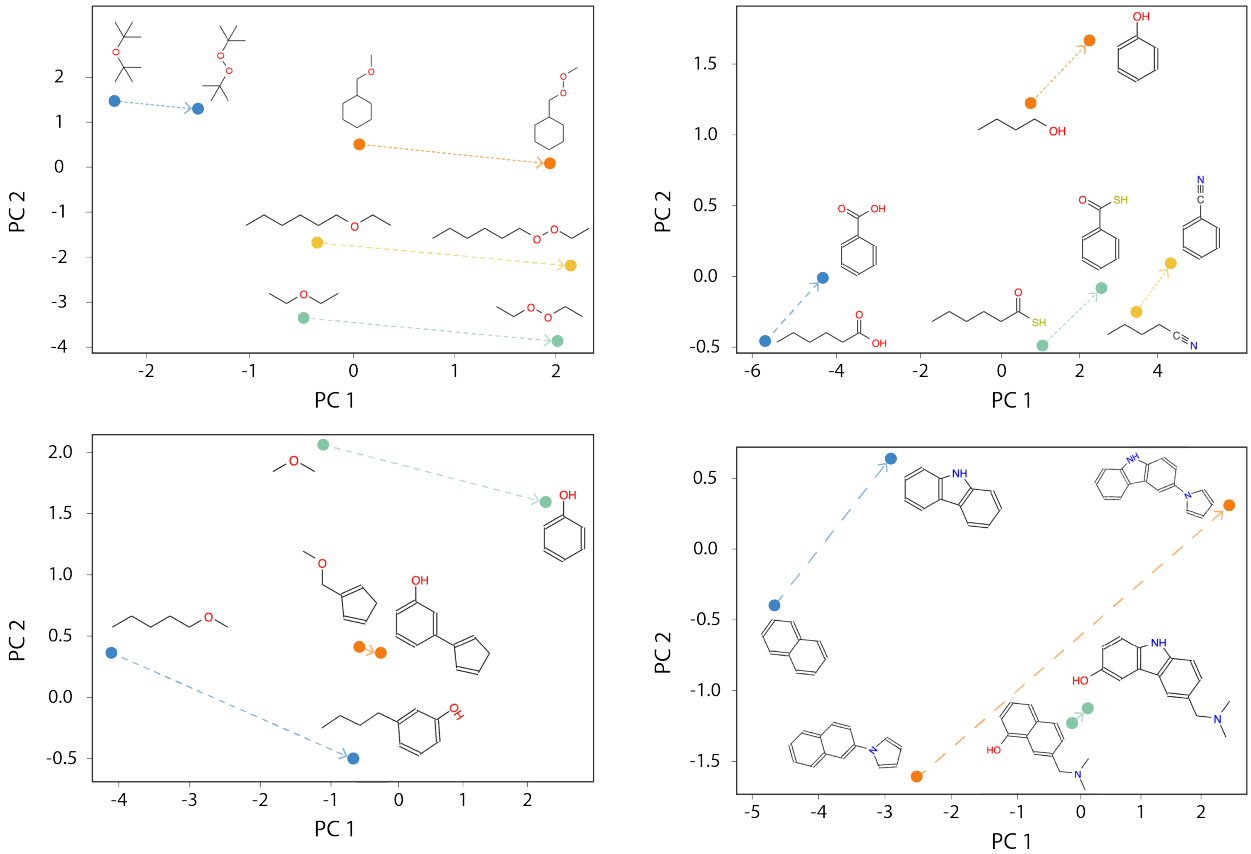

Figure 8: Link prediction model performance: Similar to word embedding analogies in NLP, replacing one functional group in a molecule with another produces parallel results across different molecules, demonstrating the model's ability to capture chemical relationships effectively.

information from SMILES and the structural relationships encoded in graph representations, thus allowing the model to learn a more comprehensive molecular representation that bridges these two modalities.

To measure the similarity between representations derived from the FG-enhanced SMILES and FG graph, we utilize cosine similarity, which is defined as: The cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as:

$$\text{cosine_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Here, $\mathbf{u}$ and $\mathbf{v}$ represent the embeddings from two different modalities, such as the SMILES-based BERT output and the GNN output for the molecular graph. This similarity score helps ensure that embeddings of positive (i.e., matched) SMILES and graph representations are closer in the latent space.

To align these two types of representations, we use contrastive loss, a popular technique in self-supervised learning that enforces representations from the same sample (positive pair) to be more similar than those from different samples (negative pair). Given a positive pair $(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{pos}})$, where $\mathbf{h}_{\text{MLM}}$ is the SMILES representation derived from a pretrained BERT model and $\mathbf{h}_{\text{pos}}$ is the corresponding representation from a graph neural network (GNN), and a negative pair $(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{neg}})$, where $\mathbf{h}_{\text{neg}}$ is a augmented FG-graph, the contrastive loss can be written as:

$$\mathcal{L}_{\text{CL}} = \frac{1}{N} \sum_{i=1}^{N} \max\left(0, \gamma - \cos(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{pos}}) + \cos(\mathbf{h}_{\text{MLM}}, \mathbf{h}_{\text{neg}})\right)$$

Where:

- γ is the margin parameter, ensuring that the positive similarity is significantly larger than the negative similarity.
- N is the number of training examples (or contrastive pairs)

The objective function is

$$\mathcal{L} = \lambda_{\text{MLM}} \cdot \mathcal{L}_{\text{MLM}} + \lambda_{\text{CL}} \cdot \mathcal{L}_{\text{CL}}$$

where $\mathcal{L}_{\text{MLM}}$ represents the masked language modeling loss, which encourages the model to predict masked tokens in the input sequence effectively, and $\mathcal{L}_{\text{CL}}$ denotes the contrastive loss, which aligns the SMILES and structural representations. The coefficients λ_{MLM} and λ_{CL} are hyperparameters that control the contribution of each loss to the overall objective. By tuning these coefficients, we can balance the learning process between the two tasks, allowing the model to learn rich and meaningful representations from both the sequential and structural aspects of the molecular data.

This combined loss function enables the model to leverage the strengths of both masked language modeling and contrastive learning, fostering a more comprehensive understanding of molecular representations that can enhance performance in downstream tasks such as property prediction, molecular generation, and structure-based drug discovery.

In our contrastive learning model, we set the margin $\gamma = 0.5$ and the weights $\lambda_{MLM} = 1.0$ and $\lambda_{CL} = 0.5$. We train the contrastive BERT model using a batch size of 126 for a total of 5 epochs. This training configuration mirrors the setup used for learning atom representations with the BERT model, as described in Section D.1.

The overall model architecture consists of three main modules that require training: a BERT model, a FG knowledge graph, and GNN. The FG knowledge graph is represented using a ComplEx model and trained first to generate FG embeddings. These embeddings are then used to train the GNN on a link prediction task. Once the GNN training converges, molecular structure embeddings are extracted and a small projection layer is trained to align them with BERT's representations via contrastive learning. This alignment is optimized jointly with the masked language modeling loss and the contrastive loss, updating both the BERT model and the projection layer.

D.5 Downstream task finetuning

MoleculeNet tasks are treated as downstream tasks for our **FARM** model. We freeze all layers of FARM and pair it with a GRU head for both classification and regression tasks. For classification, we use cross-entropy as the loss function, while for regression, we employ mean squared error. The Adam optimizer is applied with a learning rate of

$1e-4$ and a cosine annealing learning rate schedule with a period of 20 epochs. The training process spans 100 epochs with a batch size of 16, using an 80-10-10 train-validation-test split with scaffold splitting. To address imbalanced datasets, we implement a weighted loss function, assigning a weight of 5 to classes with fewer samples. For each task, we conduct three runs with different train-validation-test splits and report the average and standard deviation of the results.

E Ablation Study

E.1 FARM Component Analysis

To assess the effectiveness of each component in our architecture, we conducted a comprehensive ablation study across several MoleculeNet benchmark tasks. The first model, FG_KGE + GAT, utilizes FG knowledge graph embeddings as input for a Graph Attention Network (Veličković et al., 2017) (GAT) to predict molecular properties. Although its performance on these tasks is not the strongest, the model still demonstrates its capacity to learn underlying chemical rules (syntax and semantics) from the data to a certain degree.

The second model, AttentiveFP (Xiong et al., 2019), performs a masked atom prediction task on the molecular graph, predicting atom types such as carbon, hydrogen, oxygen, and nitrogen. Its variation, FG AttentiveFP, shares the same architecture as AttentiveFP, but it predicts both the atom type and the associated functional group. Experimental results indicate that incorporating functional group information significantly improves the model's performance on downstream tasks.

We also evaluate the BERT model trained on canonical SMILES strings, and its counterpart, FG BERT, which is trained on FG-enhanced SMILES. Results show that providing additional chemical context about functional groups boosts model performance in downstream tasks.

Finally, **FARM** (FG BERT with contrastive learning) integrates molecular structure representations from link prediction embeddings. **FARM** consistently achieves the highest performance across 6 out of 7 downstream tasks, demonstrating the power of combining FG-enhanced SMILES and contrastive learning.

Table 10 presents the detailed results of the aforementioned models across various MoleculeNet tasks, illustrating the performance of each architecture. For these experiments, we used random splitting to divide the downstream datasets into training, validation, and test sets in an 8:1:1 ratio. While random splitting is used consistently across models in this ablation study, scaffold splitting is applied for benchmarking to ensure a fair comparison with other methods.

Table 10: Performance of various models across 6 MoleculeNet tasks. The data is split using a random split into training, validation, and test sets with an 8:1:1 ratio.

	BBBP	**BACE**	**HIV**	**Average**	**ESOL**	**FreeSolv**	**Average**	**QM9**
#tasks	*1*	*1*	*1*		*1*	*1*		*3*
#samples	*2039*	*1513*	*41127*		*1128*	*642*		*133885*
Metric	*ROC-AUC (↑)*				*RMSE (↓)*			*MAE (↓)*
FG_KGE + GAT	73.23 ± 1.93	76.44 ± 1.27	71.65 ± 0.98	73.77	2.35 ± 0.210	4.32 ± 0.29	3.335	0.0139 ± 0.00014
AttentiveFP	77.71 ± 1.30	77.15 ± 0.78	78.81 ± 0.99	77.89	1.63 ± 0.042	2.11 ± 0.94	1.87	0.0056 ± 0.00012
FG AttentiveFP	85.57 ± 1.32	87.30 ± 0.90	81.21 ± 0.92	84.5	1.02 ± 0.034	1.08 ± 0.14	1.05	0.0053 ± 0.00034
BERT	82.12 ± 1.45	85.12 ± 0.76	83.03 ± 1.12	83.42	1.45 ± 0.056	1.89 ± 0.09	1.67	0.0059 ± 0.00012
FG BERT	94.36 ± 0.50	94.54 ± 0.40	81.93 ± 1.70	90.27	0.608 ± 0.031	0.507 ± 0.03	0.558	0.0041 ± 0.00017
FARM	96.23 ± 0.7	96.19 ± 0.65	82.13 ± 1.10	**91.43**	0.734 ± 0.039	0.308 ± 0.08	**0.521**	**0.0038 ± 0.00014**

E.2 Fragmentation Methods

In this section, we compare the performance of our FG-based fragmentation method with BRICS fragmentation. When using BRICS as the fragmentation method and enriching SMILES with functional group information derived from BRICS (treating each BRICS fragment as a functional group), the resulting vocabulary size is 575,612 (with a minimum frequency of 5), compare to 14,741 with our FG-based fragmentation. This significant difference arises because BRICS tends to generate larger fragments, causing the vocabulary to grow rapidly. This larger vocabulary can make the model harder to train and converge, as it includes a high number of rare tokens. Consequently, the likelihood of encountering out-of-vocabulary tokens during validation and testing increases, which negatively impacts model performance. As shown in Table 5, under the same experimental setup, our FG-based fragmentation method outperforms BRICS fragmentation across all six MoleculeNet tasks.

F Benchmarking on ADMET Tasks

Table 11 presents the performance of FARM on the ADMET datasets. The ADMET leaderboard[1] evaluates model performance across ADMET tasks (Absorption, Distribution, Metabolism, Excretion, Toxicity) using a standardized train/validation/test split, which we adopt for consistency. FARM achieves state-of-the-art results on 3 out of 7 tasks and demonstrates competitive performance with other top models on the remaining tasks, highlighting its robustness in ADMET prediction. Notably, none of the foundation models for small molecules included in Table 2 and Table 3 have reported results on these ADMET tasks, primarily due to the limitations of the datasets themselves.

Many ADMET tasks in TDC are limited by small and highly imbalanced datasets, often containing fewer than 100 positive examples. These constraints make it difficult for models to learn meaningful, generalizable patterns, particularly for complex endpoints like toxicity or clearance, which involve diverse and mechanistically unrelated biological pathways. For instance, a dataset labeling 1,000 compounds as "toxic" may include compounds that exert toxicity through entirely different mechanisms, preventing models from reliably capturing features that generalize to unseen cases. Consequently, performance on many TDC ADMET tasks is often near random, limiting their utility for assessing generalization. Specific tasks such as HIA_Hou, Clearance_Hepatocyte_AZ, and CYP3A4_Substrate_CarbonMangels exemplify this limitation, where even top-performing models fail to achieve meaningful results. To mitigate this, we evaluated FARM on a subset of more reliable TDC ADMET tasks, comparing its performance against all leading models, including some whose code is not publicly runnable (e.g., MapLight models). Many leaderboard approaches rely heavily on hand-crafted features and complex pipelines, which are orthogonal to our focus on end-to-end, transferable representation learning; combining FARM with such strategies presents a promising direction for future work.

Table 11: FARM's performance on ADMET tasks

Dataset	ADMET task	Unit	Metric	Task	SOTA	FARM	FARM's ranking
Bioav	Absorption	%	AUROC	Binary	0.753	0.709	5
AqSol		log mol/L	MAE	Regression	0.761	**0.739**	**1**
PPBR	Distribution	%	MAE	Regression	7.526	**7.376**	**1**
VDss		L/kg	Spearman	Regression	0.724	0.652	4
CYP2C9 Inhibition	Metabolism	%	AUPRC	Binary	0.859	0.798	4
CYP3A4 Inhibition		%	AUPRC	Binary	0.916	0.877	5
Ames	Excretion	%	AUROC	Binary	0.871	**0.875**	**1**

G Computational Resources

For the BERT model, even with an expanded vocabulary of approximately 14,000 tokens, the model size remains smaller than the English BERT vocabulary and does not require additional computational resources compared to training a standard BERT model. Each BERT model, with around 100 million parameters, was trained for 20 epochs on a machine with an NVIDIA A100 GPU (40GB memory), 256GB of RAM, and a 32-core AMD EPYC 7502 CPU. Training each model took approximately 6 hours. The experiments were conducted on an internal computing cluster, and no additional compute was required beyond the primary experiments included in the paper.

For other models, such as those used for ComplEx and link prediction GNN, the computational requirements are relatively lightweight, and the same machine used for BERT training was sufficient without the need for additional resources.

[1] https://tdcommons.ai/benchmark/admet_group/overview/