

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 4: 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 DOWNSTREAM TASKS DATA

In Table 5 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, 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, Sulfonic acid, Sulfonate ester, Thiocyanate, Phosphino, Phosphono, Phosphate, Phosphodiester, Phosphoryl, Borono, Boronate, Borino, Borinate, Silyl ether, Dichlorosilane, Trimethylsilyl, Fluoro, Chloro, Bromo, Iod.

Table 5: 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.

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 6a illustrates an example of naming a fused ring system based on the rules described above, and Figure 6b shows how FG-aware tokenization is applied.

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 cor-

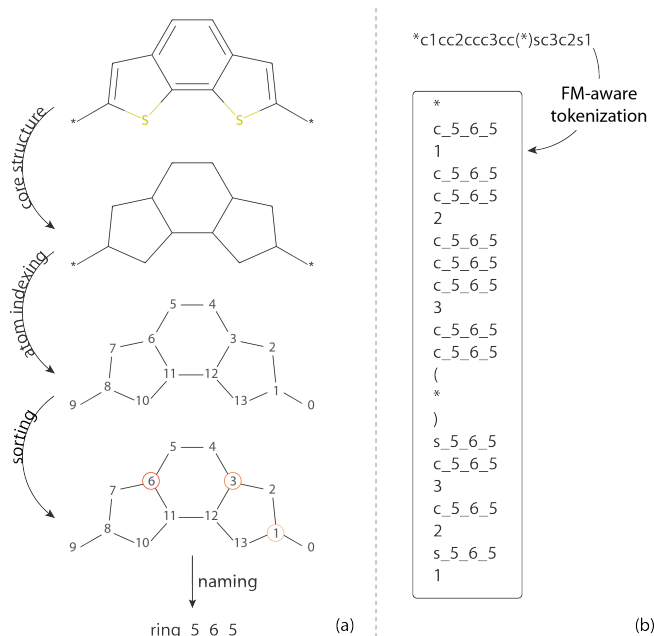


Figure 6: (a) Example of naming a fused ring system in 4 steps: generate the core structure of the functional group, index atoms using RDKit, select the smallest-index atom in each ring and sort, and name the fused ring system based on ring size. (b) Example of FG-aware tokenization.

responding to distinct FGs. The distribution of these elements is visualized in Figure 7, 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.

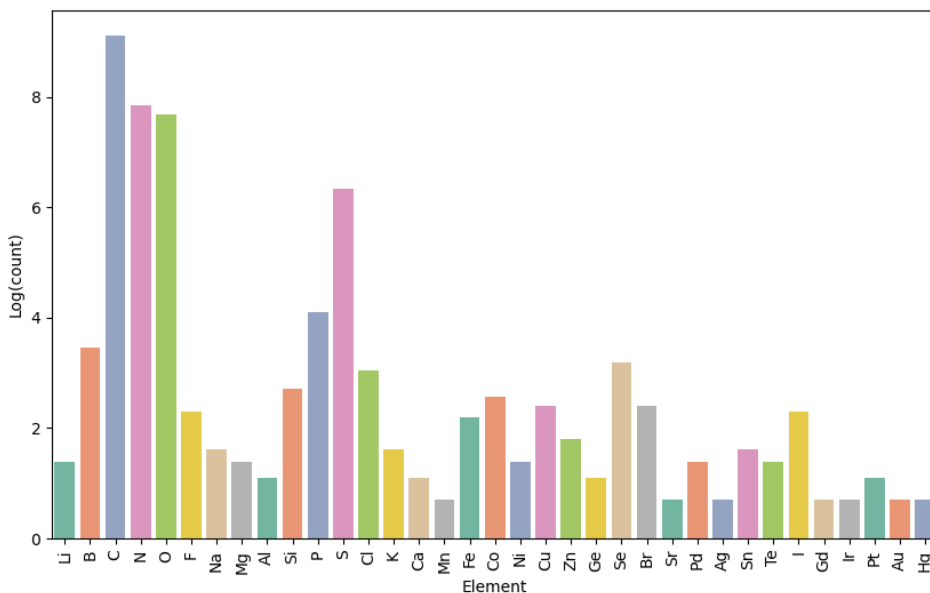


Figure 7: Number of functional groups associated with different chemical elements in the FG-enhanced SMILES dataset. The y-axis represents the natural logarithm (log, base e) of the count.

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:

Table 6: Key relations defined in the FG knowledge graph.

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.

- **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 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 7, indicate that a masking percentage of 0.35 yields the best performance across the considered downstream tasks.

Table 7: 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.

#tasks #samples Metric	BBBP 1 2039	BACE 1 1513	HIV 1 41127	Average	ESOL 1 1128	FreeSolv 1 642	Average	QM9 3 133885
		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 8 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.

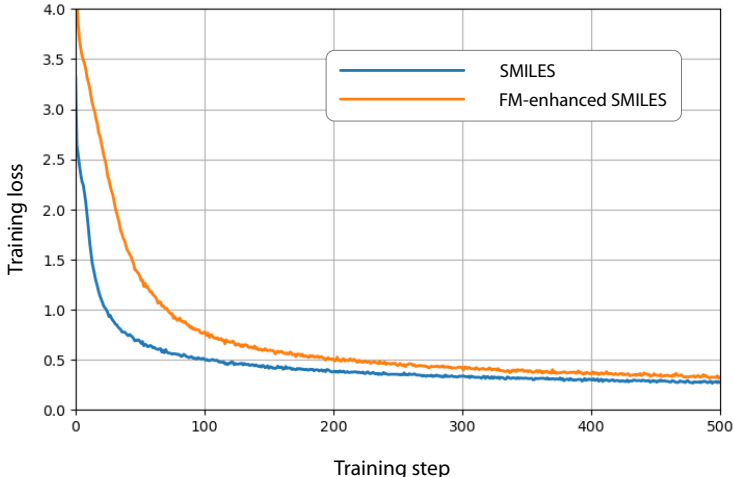


Figure 8: Loss curves for the masked language model (MLM) during training on two datasets: standard SMILES and functional group-enhanced SMILES.

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 (h, r, t) — where h is the head entity, r is the relation, and t is the tail entity — is represented as a complex vector:

$$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d \quad (1)$$

Scoring Function

The score for a given triple (h, r, t) is calculated as:

$$f(h, r, t) = \text{Re}(\mathbf{h}^T \mathbf{r} \cdot \mathbf{t}) \quad (2)$$

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:

$$\mathcal{L}_{Graph} = \sum_{(h,r,t) \in E^+} \sum_{(h',r,t') \in E^-} \max(0, \gamma + f(h', r, t') - f(h, r, t)) \quad (3)$$

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 5a.

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 module 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 4), followed by scoring potential edges with a multi-layer perceptron (MLP) (Equation 5). This score is used to calculate the probability between two nodes (Equation 6). 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 7. This approach effectively trains the model to distinguish between likely and unlikely connections between functional groups.

$$\mathbf{h}'_i = \text{ReLU} \left(\mathbf{W} \cdot \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j \right) \quad (4)$$

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.

$$s_{ij} = \text{MLP}(\mathbf{h}_i \oplus \mathbf{h}_j) \quad (5)$$

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 .

$$p_{ij} = \sigma(s_{ij}) \quad (6)$$

where σ is the sigmoid function.

$$\mathcal{L}_{\text{Link}} = -\frac{1}{|E^+|} \sum_{(i,j) \in E^+} \log p_{ij} - \frac{1}{|E^-|} \sum_{(i,j) \in E^-} \log(1 - p_{ij}) \quad (7)$$

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 9 shows the performance of the link prediction model. Similar to word embedding analogies in NLP, replacing one FG in a molecule

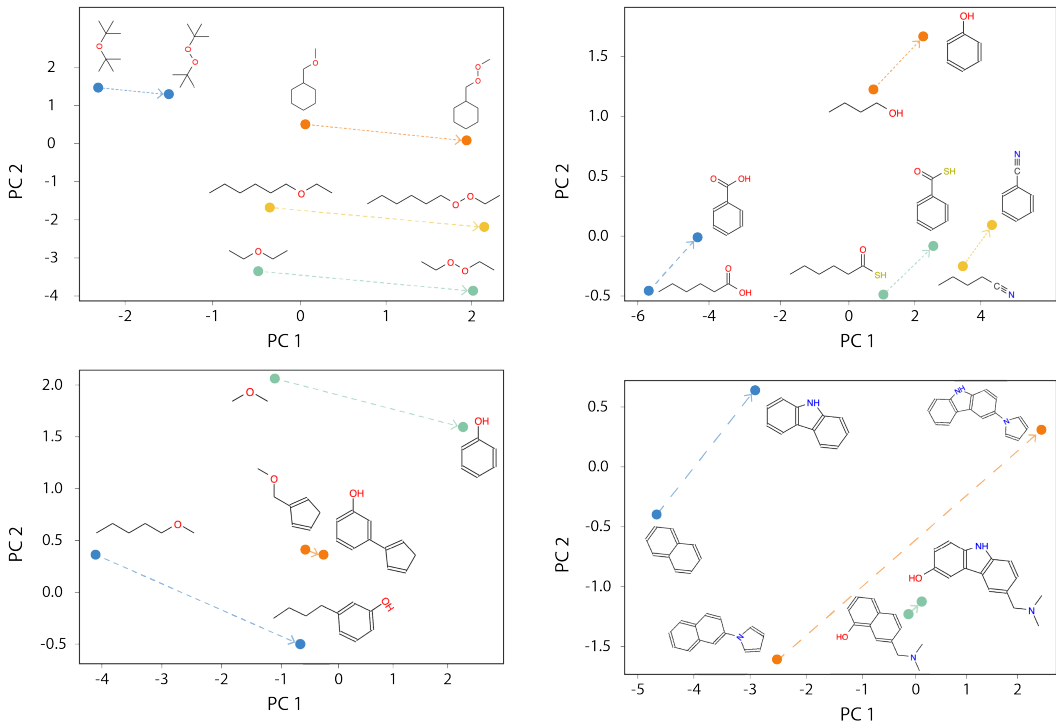


Figure 9: Link prediction model performance: Similar to word embedding analogies in NLP, replacing one functional group in a molecule with another produces parallel results across different molecules, demonstrating the model’s ability to capture chemical relationships effectively.

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 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](#).

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 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}_{MLM} is the SMILES representation derived from a pretrained BERT model and \mathbf{h}_{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}_{neg} is an augmented FG-graph, the contrastive loss can be written as:

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

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}_{MLM} represents the masked language modeling loss, which encourages the model to predict masked tokens in the input sequence effectively, and \mathcal{L}_{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.

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

To assess the effectiveness of each component in our architecture, we conducted a comprehensive ablation study across several MoleculeNet benchmark tasks. The first model, FM.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 8 presents the detailed results of the aforementioned models across various MoleculeNet tasks, illustrating the performance of each architecture.

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

<i>#tasks</i> <i>#samples</i> <i>Metric</i>	BBBP <i>1</i> <i>2039</i>	BACE <i>1</i> <i>1513</i> <i>ROC-AUC (↑)</i>	HIV <i>1</i> <i>41127</i>	Average	ESOL <i>1</i> <i>1128</i> <i>RMSE (↓)</i>	FreeSolv <i>1</i> <i>642</i>	Average	QM9 <i>3</i> <i>133885</i> <i>MAE (↓)</i>
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