FACT: Functional Group Alignment and Consistency in Token Space for Structure-aware Molecular Representation Learning

Hyeonyeong Nam[1], Woojae Choi[2], Deok-Joong Lee[1], Young-Han Son[1],

Sangwoon Lee[1], Bogyeong Kang[1], Eunjung Jo[1], Tae-Eui Kam[1]*

[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]School of Electrical Engineering, Korea University, Seoul, Republic of Korea
***Correspondence:** kamte@korea.ac.kr

Abstract

Molecular representation learning aims to capture chemically meaningful structures for various downstream tasks such as accurate molecular property prediction. However, incorporating functional group (FG) information into SMILES-based models remains challenging. The absence of explicit alignment between graph-defined FG atom sets and tokens in sequence prevents complete substructure masking, while multiple valid SMILES forms of the same molecule lead to inconsistent FG representations in token space. To address these challenges, we propose **FACT** (Functional Group Alignment and Consistency in Token Space), an end-to-end framework for structure-aware SMILES-based representation learning. FACT introduces an atom–token alignment module for complete FG span masking during pretraining and enforces FG consistency across different SMILES forms during fine-tuning. Experiments on MoleculeNet benchmarks show that FACT achieves state-of-the-art or competitive performance on eight tasks, demonstrating the effectiveness of alignment and consistency learning for molecular representation.

1 Introduction

Molecular representation learning aims to capture chemically meaningful structures. Among downstream tasks, molecular property prediction (MPP) is a fundamental task in drug discovery and material science, to efficiently estimate physicochemical and biological properties of molecules (Wieder et al., 2020; Dara et al., 2022). In recent days, deep learning approaches have achieved significant progress in this area through representation learning over large-scale molecular data (Son et al., 2024; Ji et al., 2024; Singh et al., 2026). Molecules can be represented in different forms, including SMILES strings (Weininger, 1988; Wang et al., 2019), graph-based (2D) representations (Li et al., 2022), and 3D conformations (Zhou et al., 2023). While graph-based and 3D representations provide rich structural information, they often require computationally expensive preprocessing steps such as graph construction or conformer generation that can limit scalability. In contrast, SMILES representations enable scalable learning with direct application of sequence-based models such as Transformers due to their simplicity and compatibility (Wang et al., 2019; Chithrananda et al., 2020; Ahmad et al., 2022).

Despite their achievement, existing approaches adopt pretraining objectives directly from natural language processing, particularly masked language modeling (MLM) (Liu et al., 2019), without accounting for the structural nature of molecules. While effective, such objectives do not fully reflect the structural nature of molecules. Random masking can produce chemically invalid patterns and, more importantly, disrupt meaningful chemical substructures such as functional groups (Li et al., 2023). Furthermore, unlike natural language, SMILES vocabulary is highly constrained, consisting of a limited set of atom symbols, bond types, and structural tokens (Leon et al., 2024). As a result, randomly masked tokens alone can often be predicted from simple local patterns without capturing the underlying chemical structure, limiting the effectiveness of the pretraining objective.

To better incorporate chemical structures into pretraining, recent works have introduced functional group (FG)-aware masking strategies, such as FG-BERT (Li et al., 2023) and MLM-FG (Peng et al., 2025). These approaches leverage structural information derived from molecular graphs to guide masking. FG-BERT operates masking directly on molecular graph with connectivity-aware attention (Li et al., 2023), and MLM-FG incorporates functional group information into SMILES-based pretraining through predefined SMARTS patterns (Peng et al., 2025). However, when such structural information is transferred to SMILES

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 695–703

token space, fundamental limitations arise.

First, existing SMILES-based FG masking approach lacks explicit FG alignment in token space. While functional groups are defined over atom indices in a molecular graph, SMILES sequences interleave atom token with structural non-atom tokens (e.g., bond symbols, branch parentheses, or ring closures). As a result, there is no deterministic alignment that identifies the complete token span of a functional group in the sequence, and FG-based masking is applied only over atom tokens within the group, leaving surrounding structural tokens unmasked. Although Transformer attention can in principle aggregate information across these tokens, prior FG-aware SMILES methods do not explicitly enforce complete substructure-level masking, and the extent to which models learn chemically meaningful substructures rather than relying on partially exposed local patterns has not been directly examined. We hypothesize that closing this alignment gap may benefit substructure-level representation learning, and investigate this in our framework.

Second, the one-to-many mapping between a molecule and its valid SMILES representations leads to inconsistent functional group localization in token space. Because a single molecule can be represented by multiple valid SMILES strings (Bjerrum, 2017), identical functional group may appear at different token positions across representations. This inconsistency is particularly pervasive in real-world settings, where datasets are aggregated from heterogeneous sources that may not share a unified standard for molecular structure representations. While prior work has used SMILES enumeration as a form of data augmentation (Bjerrum, 2017), existing FG-aware pretraining approaches do not explicitly enforce representational consistency of functional groups across SMILES forms during fine-tuning. Without such consistency, the model may rely on incidental token positions rather than learning invariant chemical structures.

To address the limitations, we propose **FACT** (Functional Group Alignment and Consistency in Token Space), an end-to-end framework for structure-aware molecular representation learning with SMILES data. We introduce an atom-token alignment module to deterministically identify precise FG spans in token space for complete substructure masking during pre-training. During fine-tuning, we introduce FG-preserving consistency loss to enforce representational invariance across different SMILES forms of the same molecule, leveraging the same alignment module to locate and align identical functional groups across different molecular views. We evaluate FACT framework on MoleculeNet benchmarks for downstream tasks (Wu et al., 2018) and achieve state-of-the-art or competitive performance across multiple tasks, demonstrating the effectiveness of explicit atom-token alignment and end-to-end structure-aware consistency learning for molecular representations.

Our contributions are as follows:

- We formalize the atom-token misalignment problem in SMILES-based pre-training and propose a deterministic alignment module that enables precise functional group spans identification.
- We propose FACT framework, an end-to-end framework that integrates complete functional group span masking and an FG-preserving consistency objective for invariant molecular representations.
- We achieve state-of-the-art or competitive performance on multiple MoleculeNet tasks, and provide representation-level and attribution-level visual analyses, demonstrating that our FACT framework learns chemically meaningful substructure representations.

2 Related Works

Molecular representation learning has increasingly adopted sequence-based approaches by leveraging the SMILES strings, which enables molecules to be processed using natural language models. Transformer-based architectures pretrained with masked language modeling (MLM) (Liu et al., 2019) have demonstrated solid performance across diverse molecular property prediction tasks. SMILES-BERT (Wang et al., 2019) applied BERT-style pretraining to molecular sequence, followed by ChemBERTa and its variants (Chithrananda et al., 2020; Ahmad et al., 2022) and MolFormer (Ross et al., 2022), which scaled pretraining to large-scale molecular data. While these models have shown the viability of SMILES-based pretraining, they adopt random token masking without adaptation to chemical structure. Various studies have also explored alternative training strategies and improved molecular representations to move beyond random masking. Broberg et al. (2022) pretrained Molecular

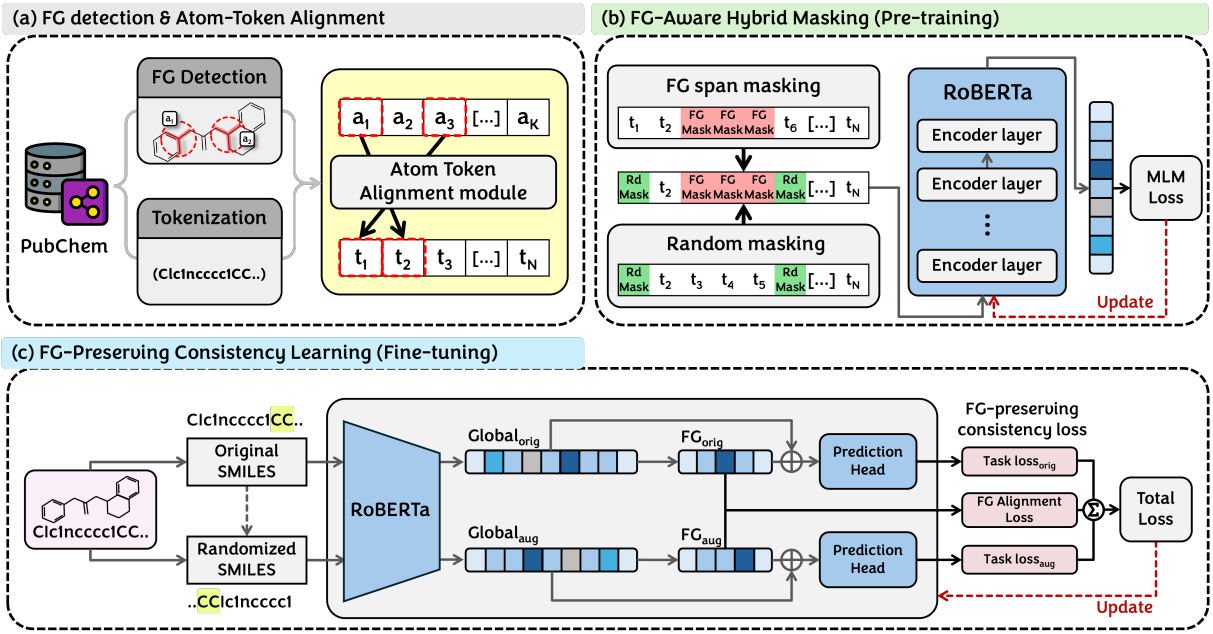

Figure 1: Overview of the proposed FACT framework. (a) FGs are first identified and mapped to their corresponding SMILES token spans via the atom-token alignment module. (b) The identified FG token spans are then used for hybrid masking, combining complete FG span masking with random token masking during pre-training. (c) During fine-tuning, two SMILES representations of the same molecule are encoded via the pretrained encoder and alignment module to obtain global and FG-level representations, which are concatenated for prediction. The FG-preserving consistency loss enforces representational invariance within the same molecule during model updates.

Transformer (Schwaller et al., 2019) using reaction template as a pre-training objective, demonstrating that chemically motivated training improves molecular representations. In parallel, fragment-based approaches have aimed to improve structural representations. t-SMILES (Wu et al., 2024) represents molecules as hierarchical fragment sequences derived from graph decomposition, while fragSMILES (Mastrolorito et al., 2025) encodes molecules at the fragment level. While these approaches improve structural representations, they do not explicitly incorporate chemically functional groups that can directly determine molecular properties.

FG-BERT (Li et al., 2023) performs pretraining on molecular graphs by masking predefined functional groups and reconstructing them to enforce chemically meaningful representations. MLM-FG (Peng et al., 2025) extends this idea to SMILES sequences by applying functional group-based masking over tokens associated with predefined functional groups, both relying on SMARTS-based patterns to specify functional groups. These approaches improve over random masking by introducing chemically meaningful substructures, yet remain constrained by the use of predefined SMARTS-based patterns for FG identification, thereby limiting adaptability across structurally novel chemical spaces (Ertl, 2017; Colmenarejo, 2025).

Moreover, when functional group information is expressed in SMILES token space, the absence of a deterministic mapping between atom indices and token positions prevents functional groups from being represented as complete substructures, leading to only partial structural encoding. In addition, structural supervision is limited to the pretraining stage, leaving fine-tuning without preserving functional group consistency. This becomes further problematic because a single molecule can be represented by multiple valid SMILES strings (Bjerrum, 2017), where identical functional groups may appear at different token positions, causing the model to rely on incidental token patterns rather than learning invariant chemical structures. Our FACT framework addresses these limitations through explicit atom–token alignment with pattern-free functional group detection, and functional group-preserving consistency learning.

3 Method

Our FACT framework is illustrated in Figure 1. The framework consists of three components: FG

detection and atom-token alignment, FG-aware pre-training, and FG-preserving consistency learning.

3.1 FG Detection and Atom-token Alignment

We first identify functional groups and establish atom–token alignment to enable FG-aware training with SMILES strings. Unlike prior approaches that rely on predefined SMARTS patterns (Li et al., 2023; Peng et al., 2025), we adopt the Ertl algorithm (Ertl, 2017) via the EFGs framework (Colmenarejo, 2025) for pattern-free functional group detection, which identifies FGs from local atomic environments without relying on predefined patterns.

We then implement a deterministic atom token alignment module that establishes a one-to-one correspondence between graph atoms and atom tokens in sequence while explicitly distinguishing structural tokens. Specifically, given a molecule, we obtain a SMILES representation and extract the corresponding atom output order provided by RDKit (RDKit, 2024), which defines the traversal of atoms in the sequence with respect to the molecular graph. We then tokenize the SMILES and check each token sequentially, assigning each token as either an atom token or a structural token, and mapping atom tokens to their corresponding graph atom indices according to the extracted order.

Based on this alignment, a functional group defined as a set of atom indices can be localized in the SMILES sequence by identifying all token positions mapped to those atoms. These positions are further expanded into a contiguous span by iteratively including adjacent structural tokens. This results in a complete token-level representation of each functional group, allowing both atom and structural tokens to be jointly masked during pre-training.

3.2 FG-aware Pre-training

Leveraging the aligned functional group spans, the model can be trained with complete FG span masking. However, applying FG masking alone is insufficient in practice. In our pre-training dataset, functional groups cover only a small fraction of tokens per sequence (median of 4 FGs covering roughly 4 atoms in a SMILES of length 39 tokens at median, i.e., about 10% of tokens; see Table 1). FG-only masking therefore provides limited training signal and leaves most context tokens unused as a reconstruction objective. To increase coverage while preserving substructure-level supervision, we propose a hybrid masking strategy that combines complete FG span masking with additional random token masking. This design ensures that functional groups are masked as complete substructures, while random masking covers the remaining context tokens that FG masking alone would leave untouched.

Table 1: Summary statistics of the 10M molecules used for pre-training.

Metric	Mean	Median	Std	Min	Max
FGs per SMILES	4.74	4.00	2.60	0.00	215
Atoms per FG	1.71	1.00	1.39	1.00	881
SMILES Token Length	43.64	39.00	20.32	4.00	2213

This process contains of two stages: (i) FGs are iteratively sampled and span-masked until a minimum FG coverage ratio is reached, and (ii) additional tokens are randomly masked until a predefined total masking ratio is satisfied. For long functional groups, sub-span sampling is further applied to increase diversity and encourage the model to learn beyond fixed substructures. We also exclude random token substitution, as it may produce chemically invalid sequences; all selected tokens are replaced exclusively with the `[MASK]` token. The specific choice of masking ratio is also detailed in Section 4.1.

3.3 FG-preserving Consistency Learning

The pretrained encoder is fine-tuned on downstream MPP tasks using a FG-preserving consistency objective. For each molecule, let A denote its original SMILES and B a randomized SMILES variant generated via RDKit's SMILES enumeration (Bjerrum, 2017). Both A and B are encoded by the shared encoder, although they describe the same molecular structure, differences in atom ordering yield different token-level representations, which we leverage to enforce representation consistency.

From these representations, we obtain two types of embeddings, a global representation via masked mean pooling over valid tokens, and a FG-level representation computed by aggregating tokens associated with the same functional groups. These representations are then concatenated to construct the final molecular embedding, which is fed into a prediction head for the downstream task. To enforce FG representational invariance, we introduce a FG-preserving consistency objective that encourages consistent representations of matched functional groups across different SMILES views of the

Table 2: Performance comparison on MoleculeNet classification benchmarks (ROC-AUC ↑). Best results are in **bold**, and second-best results are underlined.

Methods	Datasets						
	BACE	BBBP	ClinTox	Tox21	SIDER	HIV	MUV
RoBERTa	0.825(-)	0.857(-)	0.928(-)	0.753(-)	0.611(-)	0.700(-)	0.623(-)
MoLFormer	0.828(-)	0.904(-)	0.945(-)	0.773(-)	0.583(-)	0.763(-)	0.760(-)
BROBERG	0.817(0.106)	**0.921(0.101)**	**0.959(0.038)**	0.792(0.013)	0.578(0.039)	0.757(0.052)	-
FG-BERT	0.845(0.015)	0.702(0.009)	0.832(0.016)	0.784(0.080)	0.640(0.070)	0.774(0.010)	0.753(0.024)
MLM-FG (RoBERTa)	0.850(0.023)	0.897(0.043)	0.904(0.043)	0.838(0.020)	0.658(0.035)	0.800(0.027)	0.584(0.090)
MLM-FG (MoLFormer)	0.853(0.025)	0.893(0.034)	0.835(0.086)	0.850(0.015)	0.650(0.024)	0.801(0.038)	0.604(0.095)
Ours	**0.931(0.006)**	0.916(0.008)	0.950 (0.016)	**0.866(0.010)**	**0.685(0.006)**	**0.835(0.023)**	**0.782(0.028)**

same molecule. The overall training objective for the fine-tuning process is defined as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{fg}\mathcal{L}_{fg-align} \tag{1}$$

where $\mathcal{L}_{task}$ denotes the regular supervision loss for molecular property prediction task, $\mathcal{L}_{fg-align}$ enforces representational consistency between functional group representations, and λ_{fg} is a scalar hyperparameter, set to 0.001 in our experiments, that balances the two objectives. Since two SMILES representations are encoded per molecule, $\mathcal{L}_{\text{task}}$ is computed for both views and summed, matching the dual-branch task loss shown in Figure 1(c).

$$\mathcal{L}_{fg-align} = \frac{1}{|\mathcal{G}^*|} \sum_{g \in \mathcal{G}^*} \left\| \hat{\mathbf{h}}_g^{(A)} - \hat{\mathbf{h}}_g^{(B)} \right\|_2^2 \tag{2}$$

$$\mathcal{G}^* = \{g \mid g \in G_A \cap G_B\} \tag{3}$$

where $\hat{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|}$, and G_A and G_B denote the sets of functional groups detected in the original and randomized SMILES representations of the same molecule, respectively. $\mathcal{G}^*$ is their intersection, i.e., the set of functional groups that appear in both representations. $\hat{\mathbf{h}}_g^{(A)}$, $\hat{\mathbf{h}}_g^{(B)}$ denote the normalized representations of the functional group g in each representation.

4 Experiments

4.1 Experimental Setups

Datasets. For pre-training, we use 10 million SMILES sequences randomly sampled from Pubchem database (Kim et al., 2019), following the setup of MLM-FG (Peng et al., 2025). The dataset is split into training, validation, and test sets at a ratio of 98:1:1. For downstream molecular property prediction task, we conducted extensive experiments on the MoleculeNet benchmark (Wu et al., 2018), which contains binary classification, multi-label classification, and regression tasks.

Table 3: Performance comparison on MoleculeNet regression benchmarks (RMSE ↓). Best results are **bold**, and second-best results are underlined.

Methods	Datasets		
	Esol	Freesolv	Lipophilicity
RoBERTa	0.491(-)	4.444(-)	0.452(-)
MoLFormer	0.661(-)	4.449(-)	**0.446(-)**
BROBERG	**0.428(0.077)**	1.484(0.413)	0.700(0.035)
FG-BERT	0.944(0.025)	1.756(0.175)	0.655(0.009)
MLM-FG (RoBERTa)	0.605(0.068)	1.109(0.208)	0.662(0.032)
MLM-FG (MoLFormer)	0.609(0.053)	1.275(0.374)	0.612(0.014)
Ours	0.589(0.040)	**0.964(0.124)**	0.549(0.023)

Baselines. We evaluate our method within a SMILES-based Transformer framework, using RoBERTa (Liu et al., 2019) as the backbone architecture due to its strong performance in masked language modeling (MLM) and its widespread adoption in molecular representation learning. The baselines are categorized into two groups: random masking models (Broberg et al., 2022; Ross et al., 2022) and structure-aware models that incorporate domain knowledge into the pretraining objective, including a reaction prediction-based approach (Broberg et al., 2022) and FG-aware approaches (Li et al., 2023; Peng et al., 2025).

Masking Ratio. Building on the statistics indicated in Section 3.2 (FG-based masking naturally covering only about 10% of tokens at the median), we set the minimum FG coverage to 15% to ensure sufficient structural perturbation beyond this natural baseline, with the remainder contributed by random masking up to a total of 25%, balancing reconstruction difficulty with structural integrity.

Parameters and Metrics. To ensure comparability, we closely follow the pre-training configuration of MLM-FG framework (Peng et al., 2025), while certain configurations are adjusted

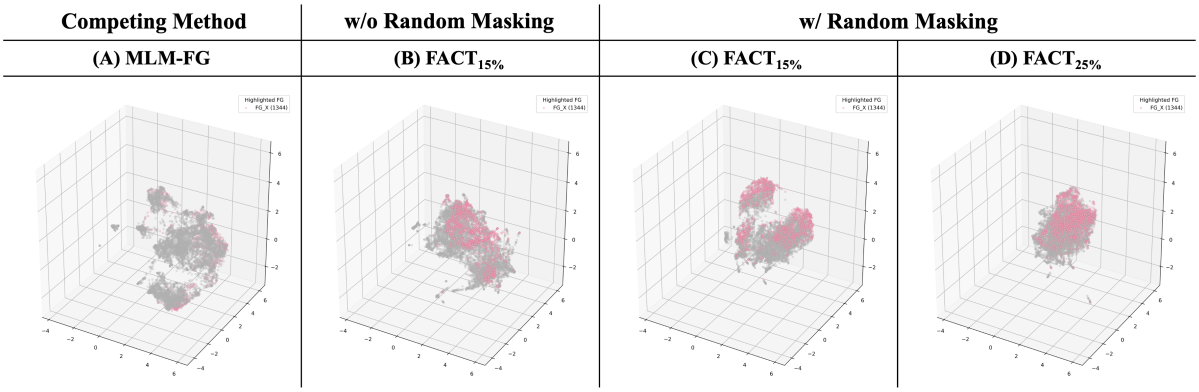

Figure 2: UMAP visualization of SMILES embeddings from pre-trained models under different masking strategies: (A) MLM-FG, the competing baseline, (B)-(D) FACT variants with different masking configurations, (B) FACT with FG-only masking at 15% (without random masking), (C) FACT with 10% FG + 5% random masking (reduced total masking ratio), and (D) FACT, our full configuration. SMILES samples containing a selected functional group (FG_X) are highlighted in pink.

due to practical constraints. Pre-training is conducted for 50 epochs using a batch size of 512 and a learning rate of 3e-5 with cosine decay, distributed across two NVIDIA H200 GPUs with the AdamW optimizer. SMILES sequences are tokenized using the *Pytoda* library, which is designed to handle large-scale chemical diversity. The vocabulary size is 576, including special tokens such as [PAD], [START], [END], and [MASK]. For fine-tuning, the pre-trained encoder is optimized on downstream tasks with a lightweight linear prediction head, using a batch size of 32 across six NVIDIA RTX 3090 GPUs. For evaluation, ROC-AUC is used for classification tasks and root mean squared error (RMSE) for regression tasks. The pre-trained encoder is available on Hugging Face[1], and the full training code will be released on Github.

4.2 Results

Classification Tasks. FACT achieves state-of-the-art performance on five out of seven classification tasks and competitive performance on the remaining two, with particularly strong gains on BACE (Table 2). On BBBP and ClinTox, FACT ranks second behind BROBERG, which may better capture global reactivity-related properties relevant to these tasks. Overall, these results suggest that explicit atom-token alignment and FG-preserving consistency learning provide consistent improvements across tasks where substructure-level representations are critical.

Regression Tasks. On regression tasks, FACT achieves state-of-the-art performance on FreeSolv, with a substantial improvement over all baselines (Table 3). However, performance on ESOL and Lipophilicity remains below the best baselines. We attribute this to the nature of these tasks: ESOL and Lipophilicity are strongly influenced by global molecular properties such as polarity and hydrophobicity, which may be better captured by models that encode broader molecular context rather than substructure-level supervision. This trade-off between substructure-level and global supervision is a direction for future work.

5 Analysis

5.1 Representation Analysis

To evaluate whether FACT learns chemically meaningful functional group representations, we visualize SMILES embeddings using UMAP (McInnes et al., 2018). We randomly sample 10,000 molecules from the PubChem database and extract representations from pre-trained models corresponding to four variants: (A) MLM-FG, one of our competing methods, (B) FACT variant, where only 15% of functional groups detected by the pattern-free Ertl algorithm are masked, (C) another FACT variant, where the total masking ratio is reduced to 15%, with 10% FG and 5% random masking, and (D) FACT. Molecules containing a specific functional group (FG_X) are highlighted in pink.

Because individual molecules usually contain multiple FGs (Table 1), we do not expect clear FG-specific clusters. Instead, we examine whether molecules sharing the highlighted functional group

[1]https://huggingface.co/Zaeus/FACT

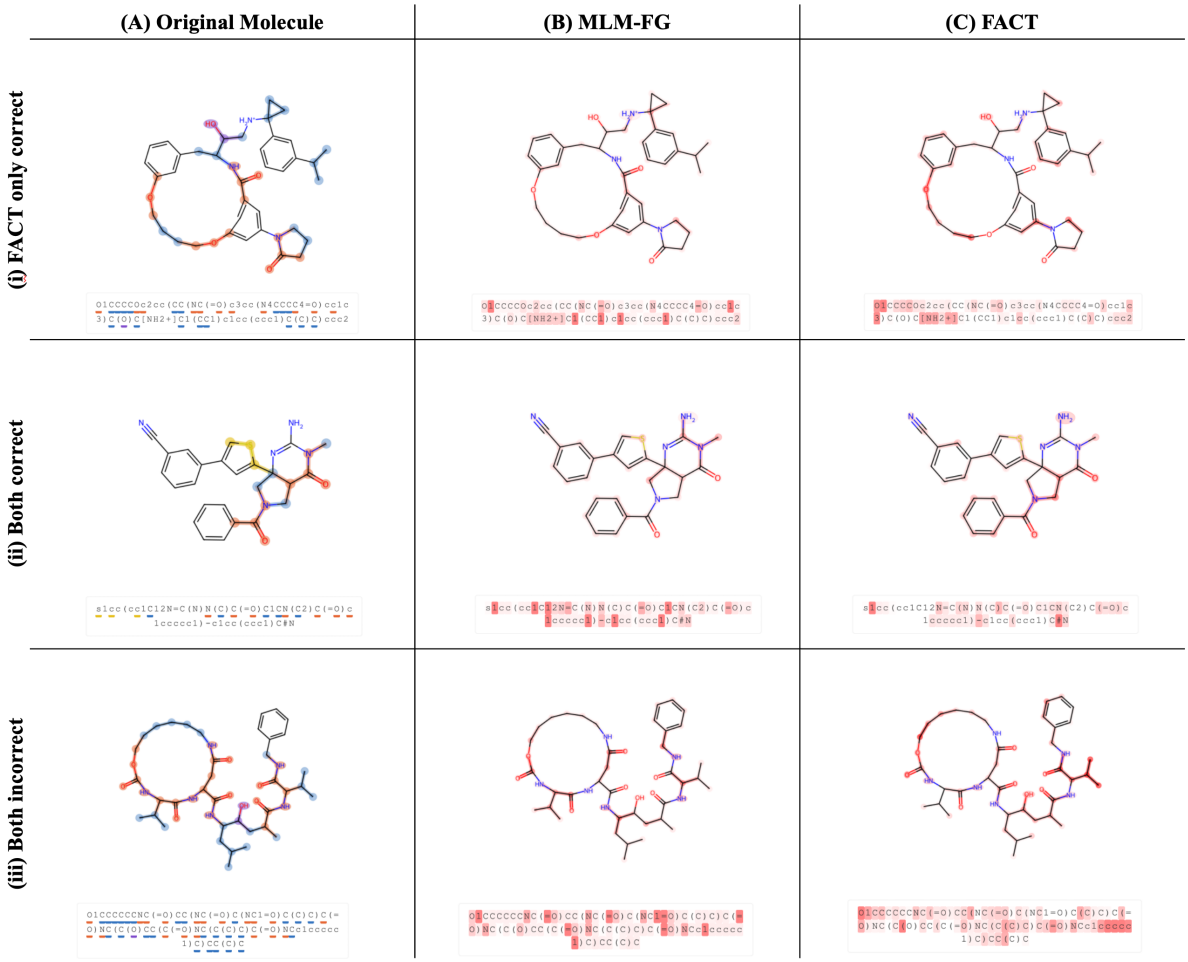

Figure 3: Attribution analysis on MoleculeNet BACE test samples under three prediction scenarios, where the ground-truth label indicates whether the molecule inhibits BACE-1: (i) FACT only correct (GT $= 0$), (ii) both models correct (GT $= 1$), and (iii) both models incorrect (GT $= 1$). Attribution scores are computed with respect to the positive class.

tend to appear in more localized regions of the embedding space. As shown in Figure 2, FACT and FACT's variants show more organized patterns compared to MLM-FG. Among these, (C) appears more organized than (B), with both gray and pink molecules forming tighter sub-regions, and (D) FACT further produces the most concentrated distribution, suggesting that combining FG-aware masking with random masking helps the model better capture FG-related structure.

5.2 Attribution Analysis

To examine whether FACT attends to chemically relevant substructures during prediction, we apply Layer Integrated Gradients (Sundararajan et al., 2017) to models fine-tuned on the BACE dataset, a binary classification task predicting whether a molecule inhibits BACE-1, computing model attribution scores with respect to the positive class score. Attribution scores are normalized and visualized on the 2D molecular structure and SMILES sequence in red, while ground-truth functional groups are color-coded by type.

To ensure a fair comparison, we analyze attribution patterns across multiple prediction scenarios on the MoleculeNet BACE test set. Specifically, we consider cases where only FACT predicts correctly (62 samples), where both models predict correctly (53 samples), and where both models fail (18 samples). From each group, we randomly sample representative molecules for visualization, labeled by their ground-truth class (GT). The selected examples correspond to (i) FACT only correct with GT $= 0$, (ii) both correct with GT $= 1$, and (iii) both incorrect with GT $= 1$.

In case (i), where GT $= 0$, FACT exhibits mini-

mal attribution with respect to the positive class, indicating that it does not rely on spurious signals for incorrect positive predictions. In case (ii), where $\mathrm{GT} = 1$, both models assign attribution to the molecule, but FACT shows stronger alignment with ground-truth functional groups, suggesting more chemically meaningful reasoning. In case (iii), where $\mathrm{GT} = 1$, although both models fail, FACT still assigns attribution to regions corresponding to ground-truth functional groups, indicating that it captures relevant substructures even when the final prediction is incorrect.

6 Conclusion

In this work, we identified two key limitations in existing SMILES-based FG-aware pretraining: the lack of explicit atom–token alignment and the absence of FG consistency under SMILES non-uniqueness. To address these, we proposed FACT, an end-to-end framework that introduces an alignment module for complete FG span identification and an FG-preserving consistency loss for molceular representational invariance. Experiments on MoleculeNet benchmarks demonstrate state-of-the-art or competitive performance, and analyses confirm that FACT learns chemically meaningful substructure representations.

Limitations

While FACT demonstrates strong performance across classification tasks, there remains room for further improvement and extension. The atom–token alignment module currently relies on RDKit-based SMILES parsing, which may present challenges for molecules with complex stereochemistry or non-standard representations, suggesting opportunities for more robust alignment strategies. In addition, the FG-preserving consistency objective introduces additional computation during fine-tuning, motivating more efficient learning mechanisms. Furthermore, while our analyses examine masking variants at the representation level (Figure 2), a complete downstream-task ablation isolating the contribution of the FG-preserving consistency loss from the alignment-based masking is left as future work. Finally, extending FACT beyond MoleculeNet to broader downstream tasks, including ADMET property prediction, represents a promising direction for future work.

CRediT authorship contribution statement

Hyeonyeong Nam: Conceptualization, Methodology, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Woojae Choi: Methodology, Investigation, Software, Validation, Writing – review & editing. Deok-Joong Lee: Visualization, Writing – review & editing. Young-Han Son, Sangwoon Lee, Bogyeong Kang, and Eunjung Jo: Writing – review & editing. Tae-Eui Kam: Supervision, Project administration, Writing – review & editing.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Graduate School Program at Korea University (No. RS-2019-II190079); by the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare and the Ministry of Science and ICT, Republic of Korea (No. RS-2025-16066488); and by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2023-00212498 and No. RS-2025-25302986).

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Esben Jannik Bjerrum. 2017. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*.

Johan Broberg, Maria Margareta Bånkestad, and Erik Ylipää Hellqvist. 2022. Pre-training transformers for molecular property prediction using reaction prediction. In *ICML 2022 2nd AI for Science Workshop*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Gonzalo Colmenarejo. 2025. Efgs: a complete and accurate implementation of ertl's functional group detection algorithm in rdkit. *Journal of Chemical Information and Modeling*, 65(3):1061–1066.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. 2022. Machine learning in drug discovery: a review. *Artificial intelligence review*, 55(3):1947–1999.

Peter Ertl. 2017. An algorithm to identify functional groups in organic molecules. *Journal of cheminformatics*, 9(1):36.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and 1 others. 2024. Unimol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Miguelangel Leon, Yuriy Perezhohin, Fernando Peres, Aleš Popovič, and Mauro Castelli. 2024. Comparing smiles and selfies tokenization for enhanced chemical language modeling. *Scientific Reports*, 14(1):25016.

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

Yuquan Li, Chang-Yu Hsieh, Ruiqiang Lu, Xiaoqing Gong, Xiaorui Wang, Pengyong Li, Shuo Liu, Yanan Tian, Dejun Jiang, Jiaxian Yan, and 1 others. 2022. An adaptive graph learning method for automated molecular interactions and properties predictions. *nature machine intelligence*, 4(7):645–651.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fabrizio Mastrolorito, Fulvio Ciriaco, Maria Vittoria Togo, Nicola Gambacorta, Daniela Trisciuzzi, Cosimo Damiano Altomare, Nicola Amoroso, Francesca Grisoni, and Orazio Nicolotti. 2025. fragsmiles as a chemical string notation for advanced fragment and chirality representation. *Communications Chemistry*, 8(1):26.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tianhao Peng, Yuchen Li, Xuhong Li, Jiang Bian, Zeke Xie, Ning Sui, Shahid Mumtaz, Yanwu Xu, Linghe Kong, and Haoyi Xiong. 2025. Pre-trained molecular language models with random functional group masking. *npj Artificial Intelligence*, 1(1):28.

RDKit RDKit. 2024. Open-source cheminformatics. *DOI: https://doi. org/10.5281/zenodo*, 591637.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, and 1 others. 2026. Chemberta-3: an open source training framework for chemical foundation models. *Digital Discovery*.

Young-Han Son, Dong-Hee Shin, and Tae-Eui Kam. 2024. Ftmmr: Fusion transformer for integrating multiple molecular representations. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4361–4372.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. 2020. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12.

Juan-Ni Wu, Tong Wang, Yue Chen, Li-Juan Tang, Hai-Long Wu, and Ru-Qin Yu. 2024. t-smiles: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1):4993.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework.