FUNCTIONAL GROUP-GUIDED TOKENIZATION FOR MOLECULAR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As a foundation for molecular property prediction and drug discovery, molecular representation learning has attracted considerable attention. Molecules are naturally represented as graphs composed of atoms and bonds. So, most existing methods employ graph neural networks to learn molecular representations. However, these methods ignore functional groups, which play a critical role in determining molecular properties. As a result, the learned representations may fail to capture the intrinsic information that determines molecular properties, thereby limiting their out-of-distribution generalization. To address this issue, we propose a functional group-guided tokenization method that transforms a molecular into a token graph to capture both functional groups and their interactions. This information is then integrated to obtain the final molecular representation. Extensive experiments on benchmark data sets demonstrate that our method effectively improves the out-of-distribution generalization, showing its ability to learn intrinsic molecular representations that remain invariant across distributions.

1 INTRODUCTION

Molecular representation learning (MRL) aims to encode molecules into vectors that can be processed by machine learning models. It serves as an important foundation for molecular property prediction and drug discovery Wang et al. (2022); Liu et al. (2022b); Li et al. (2023). In real-world applications, training molecules usually cover only a limited chemical space, while test molecules may have different scaffolds, sizes, or assays. As a result, the training and test data no longer follow the independent and identically distributed assumption. Therefore, improving out-of-distribution (OOD) generalization becomes a core problem in MRL Yang et al. (2022); Zhuang et al. (2023).

OOD generalization aims to maintain reliable prediction performance when the training and test distributions differ and has long been an important machine learning problem. The key to OOD generalization is to learn invariant representations across distributions. Many methods have been developed from different perspectives, including invariant risk minimization, risk variance regularization, robust optimization, feature distribution alignment, adversarial domain adaptation, data augmentation, etc. Representative methods include IRM Arjovsky et al. (2019), VREx Krueger et al. (2021), GroupDRO Sagawa et al. (2020), CORAL Sun & Saenko (2016), DANN Ganin et al. (2016), and Mixup Zhang et al. (2018). However, these methods are designed for data in Euclidean spaces, which limits their application to molecular OOD generalization.

To address this issue, recent studies have explored graph OOD generalization methods, specifically for molecular graphs. Their main idea is to learn invariant graph representations across distributions. And the existing methods can be roughly divided into the following three categories.

(1) Latent representation disentanglement. These methods first encode atoms in a molecular graph and then disentangle invariant and spurious representations in the latent space. For example, DisC reduces the correlations among latent factors by representation disentanglement Fan et al. (2022). MoleOOD encourages the molecular encoder to focus on substructures that are stable across environments Yang et al. (2022). iMoLD further introduces a "first-encoding-then-separation" strategy to separate invariant and spurious information in a discrete latent space Zhuang et al. (2023). Similarly, CFD employs dedicated variational encoders to disentangle distribution-agnostic and spurious features and further enhances invariant representations Wu & Deng (2025). These methods benefit from the strong representation ability of deep models. However, their encoding process

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

treats atoms as the basic units and separates invariant representations from the highly entangled latent representations. As a result, the learned representations are difficult to explicitly associate with the key chemical structures, which limits their OOD generalization.

(2) Key subgraph mining. These methods aim to mining key subgraphs that are stably related to the prediction target. Specifically, CIGA identifies subgraphs containing causal information about the labels and then enhances their intra-class invariance using an information-theoretic objective Chen et al. (2022). MILI further identifies subgraphs that are stably related to molecular properties for molecular OOD generalization and jointly optimizes environment partitioning and invariant molecular learning Wang et al. (2024). CIGA extracts subgraphs by maximizing label-related invariant information Chen et al. (2022), while MILI builds its invariant molecular learning framework around subgraphs identification Wang et al. (2024). These methods move invariant representation learning from the latent space to subgraphs, providing an important basis for learning invariant representations. However, key subgraphs are usually identified using data-driven strategies without explicit chemical domain knowledge constraints. As a result, the model has to search for key subgraphs in a very large space, which not only increases the difficulty of learning but may also produce substructures that do not correspond to the chemical substructures that truly determine molecular properties.

(3) Interpretability-driven methods. These methods identify key rationales from the perspective of interpretability. DIR identifies causal stable rationales by constructing interventional distributions Wu et al. (2022). GSAT uses stochastic attention and the information bottleneck to suppress task-irrelevant graph structures and select subgraphs related to predictions Miao et al. (2022). GREA improves the identification of rationales through rationale-environment separation and environment replacement Liu et al. (2022a). These methods can locate subgraphs that affect predictions and help reduce the influence of environment-related information. However, these rationales are mainly determined by the prediction target and data distribution. Although the identified subgraphs may be predictive in a statistical sense, they do not necessarily have meaningful chemical semantics.

To sum up, learning invariant molecular representations requires not only identifying stable subgraphs but also using domain knowledge with clear chemical semantics to guide this process. In fact, functional groups (FGs) are specific structures of atoms with particular bonding patterns. They are key units that determine molecular properties and chemical behaviors, and their functions are relatively stable across different molecules (see **Section** 2.3 for details). Based on this observation, we propose **Functional Group-guided Tokenization for Molecular Representation Learning**. It uses FGs as knowledge to guide molecular graph tokenization which transfers a molecular graph into a functional-group-aware token graph. In this way, the model can explicitly capture FGs and their interactions, thereby learning intrinsic molecular representations that remain invariant. In summary, our main contributions are as follows.

1. We proposed a knowledge and data-driven method for MRL. Unlike the existing purely data-driven methods, we introduce the domain knowledge of FG to guide the model in capturing key chemical structures that determine molecular properties.

2. We develop a FG-guided tokenization method that decomposes the molecule into tokens. And the tokens explicitly capture the FGs in the molecule, while the connections between tokens further capture the interactions among different FGs.

3. We propose a unified framework allowing FG-guided tokenization, inter-token relation modeling, and token graph representation to be jointly optimized in an end-to-end manner.

4. Extensive experiments on molecular OOD benchmark data sets show that the proposed method achieves strong performance, demonstrating the effectiveness of FG-guided tokenization for learning invariant molecular representations.

2 Notation and Preliminary

2.1 Problem Formalization

A molecule consists of a set of atoms and the bonds between them, which naturally forms a graph. Formally, given a molecule graph

$$G = (X, E, A) \in \mathcal{G}. \tag{1}$$

There are n nodes (atoms) and m edges (bonds). Each node is described by a d_1-dimensional vector to encode information of the corresponding atom. $X \in \mathbb{R}^{n \times d_1}$ is formed by concatenating the description vectors of all the nodes. Similarly, $E \in \mathbb{R}^{m \times d_2}$ is formed by concatenating the description vectors of all the edges. And $A \in \{0,1\}^{n \times n}$ is the adjacency matrix of the graph.

The MPL aims to learn a mapping $f : \mathcal{G} \to \mathbb{R}^d$ to obtain an embedding vector of the molecule for graph-level tasks, e.g., molecule property prediction Ji et al. (2023). Sometimes, we need to obtain an embedding vector for each atom for node-level or edge-level tasks, e.g., predicting the 3D coordinates of atom and bound Xu et al. (2021), etc. This paper focuses on the former.

2.2 Graph Neural Network

As mentioned above, molecules are naturally presented as graphs. So MPL is a special type of graph representation learning. So far, graph neural networks (GNNs) have become the state-of-the-art approaches for graph processing. Existing GNNs can be broadly divided into two categories.

The first category is message-passing-based GNNs, which learn graph representations by iteratively aggregating information from neighboring nodes and edges. Representative methods include GCN Kipf & Welling (2017), GatedGCN Bresson & Laurent (2017), GIN Xu et al. (2019), PNA Corso et al. (2020), and MPNN Gilmer et al. (2017), which was specifically developed for molecular learning. These methods can effectively exploit local structures and edge information but are relatively limited in modeling long-range dependencies.

The second category is attention-based graph transformers. By employing attention mechanisms, these methods enable direct interactions between arbitrary pairs of nodes, thereby enhancing the modeling of global structures. Representative methods include Graph-Transformer Dwivedi & Bresson (2020), SAN Kreuzer et al. (2021), Graphormer Ying et al. (2021), GraphiT Mialon et al. (2021), and GraphTrans Wu et al. (2021), etc. These methods are often further equipped with positional or structural encodings to enhance the awareness of graph structures.

GPS Rampášek et al. (2022) integrates message-passing and attention within a unified layer. Specifically, given the input graph $G^l = (X^l, E^l, A)$, the l-th layer extracts features as follows,

$$X^{l+1}, E^{l+1} = \mathrm{GPS}^l\left(X^l, E^l, A\right), \tag{2}$$

$$X_M^{l+1}, E^{l+1} = \mathrm{MPNN}_e^l\left(X^l, E^l, A\right), \tag{2.a}$$

$$X_T^{l+1} = \mathrm{GlobalAttn}^l\left(X^l\right), \tag{2.b}$$

$$X^{l+1} = \mathrm{MLP}^l\left(X_M^{l+1} + X_T^{l+1}\right), \tag{2.c}$$

where Eqs.(2.a) and (2.b) are used to capture the local and global structure, respectively. It provides a general and flexible GNN framework and is used as the basic layer in the proposed method.

2.3 Functional Group

The functional group (FG) is an atom or a group of atoms that exhibits specific chemical behavior across different molecules IUPAC. It strongly influences molecular properties by altering local electron distribution, polarity, hydrogen-bonding capability, acidity/basicity, and chemical reactivity. Cramer et al. (2019) showed that the local structural changes may therefore induce substantial differences in macroscopic properties such as solubility, lipophilicity, and binding affinity. For example, hydroxyl ($-\mathrm{OH}$) can increase molecular polarity and enable hydrogen-bond formation. Carboxyl ($-\mathrm{COOH}$) and amino ($-\mathrm{NH_2}$) can significantly affect acidity and basicity, respectively.

Some theoretical models also explain the correlation between FGs and molecular properties. For example, Wildman & Crippen (1999) showed that physicochemical properties such as logP and molar refractivity can be estimated from contributions of atom types and their local chemical environments. Similarly, Ertl et al. (2000) demonstrated that molecular polar surface area can be approximated by summing contributions from polar fragments, providing strong evidence that local chemical substructures play a fundamental role in determining molecular properties.

In this paper, we collects 53 common FGs coming from 7 categories. Table 1 shows exemplars for each category. And all the FGs are listed in Table 6 of **Appendix** A.1. To unify the expression, these

FGs are represented as a set of graphs, i.e.,

$$\mathcal{F} = \{F_i = (\hat{X}_i, \hat{E}_i, \hat{A}_i)\}_{i=1}^{53}, \tag{3}$$

where F_i is the graph of the i-th FG with $\hat{n}_i$ atoms and $\hat{m}_i$ bounds. The meanings of $\hat{X}_i$, $\hat{E}_i$, and $\hat{A}_i$ are similar to those in Eq.(1), and are omitted.

Table 1: The typical FGs for the 7 categories.

Category	Exemplar	Smarts	Graph Structure
(1) Carbon–carbon Unsaturated Bonds	Alkene	$[C]=[C]$	
(2) Ring structures	Benzene	C1CCCCC1	
(3) Oxygen-containing	Alcohol	[CX4][OX2H]	
(4) Nitrogen-containing	Amide	$[NX3][CX3](=O)$	
(5) Sulfur-containing	Disulfide	[SX2][SX2]	
(6) Halogen-containing	Alkyl_halide	$[\#6][F, Cl, Br, I]$	
(7) Phosphorus-containing	Phosphate	$[P](=O)(O)(O)$	

3 METHODOLOGY

We propose a FG-guided tokenization for MPL, explicitly modeling the strong correlation between FGs and molecular properties. Figure 1 shows the overall design which consists of three parts, i.e., (1) molecular graph preprocessing, (2) molecular graph tokenization, and (3) token graph representation & predicting. Next, we use a molecular graph $G = (X, E, A)$ as an example to illustrate the process.

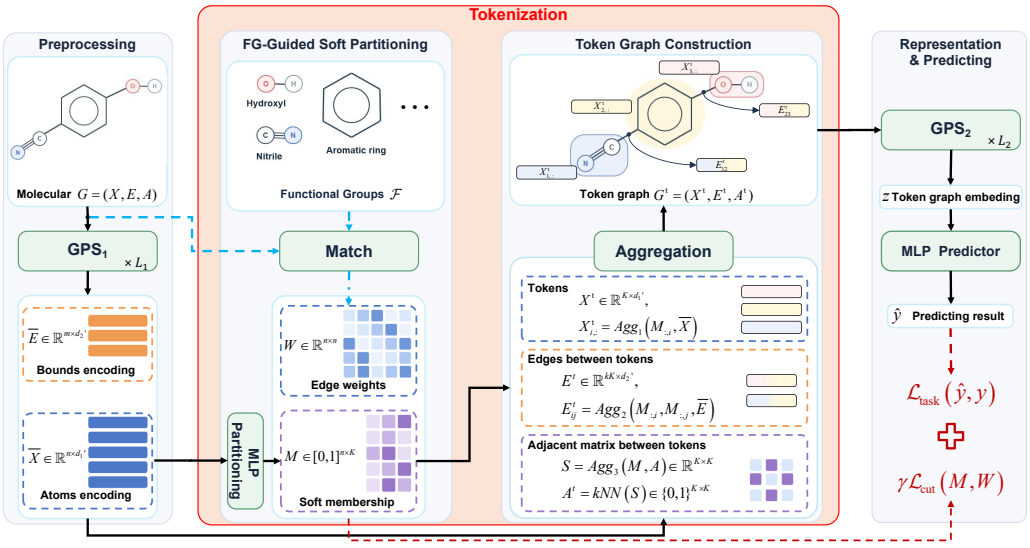


Figure 1: The overall design of the FG-guided MPL. −−: data streams solely for training.

3.1 Molecular Graph Preprocessing

For a molecular graph $G = (X, E, A)$, different data acquisition methods get different types of data. For example, the encoding of atoms can be discrete atomic numbers or manually designed continuous attributes. The same goes for bounds. To unify the subsequent procedures, the atoms and bonds are encoded as continuous vectors as follows.

$$\bar{G} = (\bar{X}, \bar{E}) = GPS_1(X, E, A), \tag{4}$$

where $\bar{X} \in \mathbb{R}^{n \times d_1'}$ and $\bar{E} \in \mathbb{R}^{m \times d_2'}$ are the encodings of atoms and bounds, respectively. GPS_1 is the encoding network with L_1 GPS layers.

3.2 Molecular Graph Tokenization

3.2.1 FG-Guided Soft Partitioning

Based on the graph $\bar{G} = (\bar{X}, \bar{E})$ in Eq.(4), we perform a soft partitioning on it to extract the information of FGs contained in the molecule. Specifically, it includes the following two steps.

Step 1. Construct FG-based weighted graph. Given the set of FGs $\mathcal{F}$ in Eq.(3), for a molecular $G = (X, E, A)$ in Eq.(1), the edges' weights matrix $W \in \mathbb{R}^{n \times n}$ is defined as

$$W = A + \sum_{F \in \mathcal{F}} Match(F, G), \tag{5}$$

where $Match(F, G)$ accumulates the adjacency matrix of F. If F does not appear in G, it returns a zero matrix. Otherwise, it returns a weighted matrix. To facilitate understanding, Figure 2 shows an example where F appears twice in G. The rigorous definition of $Match(F, G)$ is provided in **Algorithm** 1 of **Appendix** A.1.

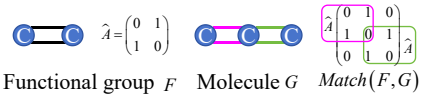


Figure 2: An example of $Match(F, G)$.

It should be noted that a molecule may contain various types of FGs and a FG may appear multiple times in a molecule. According to Eq.(5), for any two atoms p and q in G, there are three cases. (1) $W_{pq} = 0$, i.e., there is no bound between p and q. (2) $W_{pq} = 1$, i.e., there is a bound $<p, q>$ and it does not appear in any FG. (3) $W_{pq} > 1$, i.e., there is a bound $<p, q>$ and it appears in $W_{pq} - 1$ FGs. To sum up, the larger the W_{pq} is, the greater the contribution of the bound $<p, q>$ to FGs.

Step 2. Partition FG-based weighted graph. Based on W in Eq.(5), the atoms in G are cut into K clusters. And we use a multi-layer perceptron to learn the soft membership matrix.

$$M = MLP_{\text{par}}(\bar{X}) \in [0, 1]^{n \times K}, \tag{6}$$

where $\forall p = 1, 2, \cdots, n$, $\forall i = 1, 2, \cdots, K$, M_{pi} is the membership of atomic p to cluster i and $\sum_{j=1}^{K} M_{pj} = 1$. This can be achieved by the $softmax$ activation function. To ensure that the M can capture FGs in G, we introduce the following graph cut loss.

$$\mathcal{L}_{\text{cut}}(G) = \frac{1}{2} \sum_{p=1}^{n} \sum_{q=1}^{n} W_{pq} \left\| M_{p,:} - M_{q,:} \right\|_2^2 = tr(M^T L M), \tag{7}$$

where $L = D - W$ is the Laplacian matrix of W. And D is the degree matrix of W. Minimizing this loss can prevent FGs in G from being divided into different clusters.

3.2.2 Token Graph Construction

Based on the soft membership matrix M in Eq.(6), the graph $\bar{G} = (\bar{X}, \bar{E})$ in Eq.(4) is aggregated into a token graph

$$G^{\text{t}} = (X^{\text{t}}, E^{\text{t}}, A^{\text{t}}) = Agg(\bar{X}, \bar{E}, A; M). \tag{8}$$

In G^{t}, there are K nodes and kK edges. And each node corresponds to one token (cluster). The definitions of X^{t}, A^{t}, and E^{t} are as follows.

(1) $X^{\mathrm{t}} \in \mathbb{R}^{K \times d_1'}$ is the embeddings of all the tokens. And $\forall i = 1, 2, \cdots, K$, the embedding of the i-th token is computed as

$$X^{\mathrm{t}}_{i,:} = \left(\sum_{p=1}^{n} M_{pi} \bar{X}_{p,:} \right) / \left(\sum_{p=1}^{n} M_{pi} \right), \tag{9}$$

i.e., the weighted mean of the embeddings of the nodes in the i-th cluster.

(2) $A^{\mathrm{t}} \in \{0, 1\}^{K \times K}$ is the adjacency matrix between tokens and is computed as follows. $\forall i, j = 1, 2, \cdots, K$, $i \neq j$, the similarity between i-th token and j-th token is defined as the number of bonds between i-th cluster and j-th cluster, i.e.,

$$S_{ij} = \left(\sum_{p=1}^{n} \sum_{q=1}^{n} M_{pi} M_{qj} A_{pq} \right) / \left(\sum_{p=1}^{n} \sum_{q=1}^{n} M_{pi} M_{qj} \right). \tag{10}$$

Then, the kNN algorithm is applied on $S \in \mathbb{R}^{K \times K}$ to obtain the sparse 0-1 matrix A^{t}.

(3) $E^{\mathrm{t}} \in \mathbb{R}^{kK \times d_2'}$ are embeddings of kK edges in A^{t}. Specifically, $\forall i, j = 1, 2, \cdots, K$, $A^{\mathrm{t}}_{ij} = 1$, the embedding of edge $<i, j>$ is the weighted mean of the embeddings of the bounds between i-th cluster and j-th cluster, i.e.,

$$E^{\mathrm{t}}_{ij} = \left(\sum_{p=1}^{n} \sum_{q=1}^{n} M_{pi} M_{qj} \bar{E}_{pq} \right) / \left(\sum_{p=1}^{n} \sum_{q=1}^{n} M_{pi} M_{qj} \right)^{1}. \tag{11}$$

To sum up, in G^{t}, the tokens capture the information of FGs in molecule G, and the edges between tokens capture the interaction information between FGs in molecule G.

3.3 Token Graph Representation & Predicting

Given the token graph G^{t} in Eq.(8), the information of FGs is integrated by the subsequent GNNs to detect the molecule property. First, G^{t} is fed into a GPS to obtain the embedding of the token graph, i.e., $z = GPS_2\left(X^{\mathrm{t}}, E^{\mathrm{t}}, A^{\mathrm{t}}\right) \in \mathbb{R}^d$. Then, a multi-layer perceptron learns the predicting value(s) for the downstream task, i.e., $\hat{y} = MLP_{\mathrm{pre}}(z)$. Finally, the task-specific loss is used to train the model.

$$\mathcal{L}_{\mathrm{task}}(G, y) = \ell(\hat{y}, y), \tag{12}$$

where y is the true label of G and ℓ is the task-specific loss, typically the cross-entropy loss for the classification task and the mean squared error loss for the regression task.

3.4 Training and Test

Let $\{(G_i, y_i)\}_{i=1}^{N}$ be the training set. The training process can be written as

$$\underset{GPS_1,\, MLP_{\mathrm{par}},\, GPS_2,\, MLP_{\mathrm{pre}}}{min} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{task}}(G_i, y_i) + \gamma \mathcal{L}_{\mathrm{cut}}(G_i), \tag{13}$$

where $\gamma > 0$ is the tradeoff parameter. It should be noted that during the test phase, there is no need to calculate any loss. Therefore, there is no need to calculate the edges' weights matrix W in Eq. 5. See $--$ in Figure 1 for details.

4 Experiments

4.1 Experimental Setup

Data sets. For OOD molecular representation learning, we adopt the experimental protocol of iMoLD Zhuang et al. (2023) and evaluate our method on two real-world benchmarks. **GOOD** Gui et al. (2022) is a systematic benchmark developed for evaluating graph models under OOD distribution shifts. We consider two molecular data sets: GOOD-HIV Wu et al. (2018), a binary

[1]To simplify the expression, $\forall p, q = 1, 2, \cdots, n$, if $A_{pq} = 0$, the $\bar{E}_{pq}$ is regarded as the zero vector.

Table 2: Evaluation performance on GOOD benchmark. -: abnormal results caused by under-fitting. /: the method cannot be applied to the setting. **Boldface**: the best performance.

Method	GOOD-HIV ↑				GOOD-ZINC ↓			
	scaffold		size		scaffold		size	
	covariate	concept	covariate	concept	covariate	concept	covariate	concept
ERM	69.55	72.48	59.19	61.91	0.1802	0.1301	0.2319	0.1325
IRM	70.17	71.78	59.94	-	0.2164	0.1339	0.6984	0.1336
VREx	69.34	72.21	58.49	61.21	0.1815	0.1287	0.2270	0.1311
GroupDRO	68.15	71.48	57.75	59.77	0.1870	0.1323	0.2377	0.1333
Coral	70.69	72.96	59.39	60.29	0.1769	0.1303	0.2292	0.1261
DANN	69.43	71.70	62.38	65.15	0.1746	0.1269	0.2326	0.1348
Mixup	70.65	71.89	59.11	62.80	0.2066	0.1391	0.2531	0.1547
DisC	58.85	64.82	49.33	74.11	/	/	/	/
MoleOOD	69.39	69.08	58.63	55.90	0.2752	0.1996	0.3468	0.2275
CIGA	69.40	71.65	61.81	73.62	/	/	/	/
CAL	69.12	72.49	59.34	56.16	/	/	/	/
iMoLD	72.93	74.32	62.86	77.43	0.1410	0.1014	0.1863	0.1029
CFD	76.42	77.83	64.14	79.28	0.1187	0.0765	0.1421	0.0852
MoSIR	77.13	78.96	66.65	81.15	0.0913	0.0553	0.1317	0.0637
DIR	68.44	71.40	57.67	74.39	0.3682	0.2543	0.4578	0.3146
GSAT	70.07	72.51	60.73	56.96	0.1418	0.1066	0.2101	0.1038
GREA	71.98	70.76	60.11	60.96	0.1691	0.1157	0.2100	0.1273
Ours	**77.80**	**80.03**	**68.08**	**83.47**	**0.0738**	**0.0422**	**0.1011**	**0.0570**

classification dataset for predicting whether a molecule can inhibit HIV, and GOOD-ZINC Gómez-Bombarelli et al. (2018), a regression dataset for molecular solubility prediction. Each dataset is partitioned according to molecular scaffold and size, with both covariate and concept shifts considered for each partitioning strategy, yielding 8 evaluation settings in total. **DrugOOD** Ji et al. (2023) is an OOD benchmark designed for AI-assisted drug discovery. We consider three environment-partitioning strategies (assay, scaffold, and size) under two activity measurements (IC50 and EC50), resulting in 6 evaluation settings. Each setting defines a binary classification task for predicting drug–target binding affinity. More details are provided in Table 7 of the **Appendix** A.2.

Baselines. We compare the proposed method with ERM Vapnik (2013) and 4 groups of OOD methods. Specifically, (1) OOD algorithms used for Euclidean data, including IRM Arjovsky et al. (2019), VREx Krueger et al. (2021), GroupDRO Sagawa et al. (2020), Coral Sun & Saenko (2016), DANN Ganin et al. (2016), and Mixup Zhang et al. (2018). (2) Key substructure mining methods, including DisC Fan et al. (2022), MoleOOD Yang et al. (2022), and CIGA Chen et al. (2022). (3) Latent representation disentanglement methods, including CAL Sui et al. (2022), iMoLD Zhuang et al. (2023), MILI Wang et al. (2024), CFD Wu & Deng (2025), and MoSIR Li et al. (2026). (4) Interpretability-driven methods, including DIR Wu et al. (2022), GSAT Miao et al. (2022), and GREA Liu et al. (2022a).

Evaluation and implementation details. We use ROC-AUC to evaluate GOOD-HIV and DrugOOD because they involve binary classification. For the regression task on GOOD-ZINC, we adopt Mean Absolute Error (MAE). For $MLP_{\mathrm{par}}()$, we employ a two-layer fully connected network with 64 hidden units. For GPS, we adopt GINE for local message passing and a Transformer for global attention. Additional settings are detailed in Table 8 of the **Appendix** A.2.

4.2 Performance Analysis

Table 2 reports the results of the proposed method and comparison methods on GOOD benchmark. We have the following observations.

(1) Conventional OOD methods developed for Euclidean data, such as IRM, GroupDRO, DANN, and Mixup, exhibit relatively limited and unstable performance, suggesting that they cannot explicitly capture the structural semantics of molecular graphs.

(2) Key substructure mining methods, including DisC, MoleOOD, and CIGA, improve OOD generalization by identifying predictive molecular substructures. However, their purely data-driven selection processes may still capture unstable or spurious structural patterns.

(3) Latent representation disentanglement methods, such as CAL, iMoLD, MILI, CFD, and MoSIR, generally perform better by separating invariant information from variant information. However, these methods rely on atom-level encoding and disentangle invariant information from complex latent representations, limiting their ability to capture property-determining chemical structures and provide chemical interpretability.

(4) Interpretability-driven methods, including DIR, GSAT, and GREA, can identify prediction-relevant subgraphs but remain limited under certain distribution shifts.

(5) In our method, FG-guided tokenization incorporates stable chemical knowledge, while inter-token relation learning captures interactions among functional groups. Their end-to-end optimization enables robust molecular representations across scaffold, size, covariate, and concept shifts.

Table 3 reports the results of the proposed method and comparison methods on DrugOOD benchmark, where conclusions similar to those observed on GOOD can be drawn.

Table 3: Evaluation performance on DrugOOD benchmark. **Boldface**: the best performance.

Method	**DrugOOD-IC50** ↑			**DrugOOD-EC50** ↑		
	Assay	Scaffold	Size	Assay	Scaffold	Size
ERM	71.63	68.79	67.50	67.39	64.98	65.10
IRM	71.15	67.22	61.58	67.77	63.86	59.19
Coral	71.28	68.36	64.53	72.08	64.83	58.47
MixUp	71.49	68.59	67.79	67.81	65.77	65.77
DisC	61.40	62.70	61.43	63.71	60.57	57.38
MoleOOD	71.62	68.58	65.62	72.69	65.74	65.51
CIGA	71.86	69.14	66.92	69.15	67.32	65.65
CAL	70.09	65.90	66.42	74.54	65.19	61.21
iMoLD	72.11	68.84	67.92	77.48	67.79	67.09
MILI	72.67	69.58	68.40	77.11	68.07	65.97
CFD	73.86	70.02	69.73	78.32	69.13	67.62
MoSIR	74.65	71.50	72.03	79.52	70.25	68.56
DIR	69.84	66.33	62.92	65.81	63.76	61.56
GSAT	70.59	66.45	66.70	73.82	64.25	62.65
GREA	70.23	67.02	66.59	74.17	64.50	62.81
Ours	**75.74**	**72.61**	**73.44**	**80.57**	**71.28**	**69.63**

4.3 TOKENIZATION VISUALIZATION

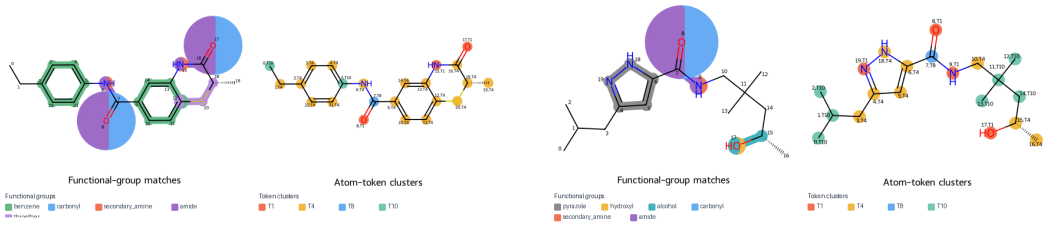
This section illustrates the mechanism of FG-guided tokenization.

First, we counted the number of FGs in each molecule and reported the average for each data set as presented in Table 9. We observe that a diverse range and substantial number of FGs can be successfully identified across molecules, demonstrating the feasibility of leveraging FGs to guide molecular representation learning.

Second, we visualize the FG matching results and the soft partitioning of atoms to molecular tokens, as shown in Figure 3. Additional visualization results are provided in Figure 4 of the **Appendix** A.2. As observed, atoms within the same FG tend to be assigned to the same token, thereby preserving the structural integrity of FGs and facilitating the learning of semantically invariant molecular representations for molecular property prediction.

4.4 PARAMETER ANALYSIS

In this section, we present sensitivity analyses on GOOD-ZINC for the number of tokens K and the number of neighbors k in Table 4, and for the graph cut loss weight γ in Table 5. From these tables, we can conclude that



(a) (b)

Figure 3: Visualization of functional group matching and atom clustering.

(1) The model achieves the best performance across all four distribution-shift settings when $K = 15$ and $k = 3$. A moderate number of tokens adequately represents molecular structures, whereas excessive tokens or neighbors may introduce redundancy and noise. The consistently better performance with a smaller k suggests that sparse token relations facilitate stable representation learning.

(2) The optimal value γ varies slightly across distribution-shift settings, indicating that the strength of the graph cut constraint should be adjusted accordingly. Nevertheless, all nonzero settings outperform $\gamma = 0.0$, demonstrating the effectiveness of the graph cut loss in identifying meaningful token structures and improving OOD generalization.

Table 4: Sensitivity analysis on the number of tokens K and neighbors k in kNN graph on GOOD-ZINC (MAE ↓).

K	k	Scaffold		Size	
		covariate	concept	covariate	concept
10	3	0.0816	0.0491	0.1098	0.0604
10	5	0.0905	0.0574	0.1103	0.0673
15	3	**0.0738**	**0.0422**	**0.1011**	**0.0570**
15	5	0.0796	0.0483	0.1072	0.0645
20	3	0.0824	0.0495	0.1134	0.0629
20	5	0.0878	0.0569	0.1165	0.0691

Table 5: Sensitivity analysis on the weight of the graph cut loss γ on GOOD-ZINC (MAE ↓).

γ	Scaffold		Size	
	covariate	concept	covariate	concept
1.0	**0.0738**	**0.0422**	0.1074	0.0648
0.1	0.0819	0.0433	**0.1011**	0.0588
0.01	0.0853	0.0459	0.1109	0.0612
0.001	0.0871	0.0472	0.1162	**0.0570**
0.0	0.1012	0.0603	0.1325	0.0762

5 Conclusion

Functional groups are important factors in determining molecular properties. In this paper, we incorporate this domain knowledge into molecular representation learning and propose a new knowledge and data-driven method. Through functional group-guided tokenization, a molecule is transformed into a token graph, where tokens capture the functional groups contained in the molecule and edges characterize the interactions between functional groups. This information is further integrated to obtain molecular representations that remain invariant across distributions, achieving state-of-the-art performance on molecular OOD generalization tasks. In future work, we plan to collect a broader set of functional groups to further improve the proposed method. Moreover, as molecular representation learning is a fundamental task, the success of our method suggests strong potential for extending it to a wide range of molecule-related applications.

AI USE STATEMENT

Generative AI tools were used solely for language editing. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

The source code is available at `https://anonymous.4open.science/r/FGT-MRL-DC51`.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*, 2017.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Conference on Neural Information Processing Systems*, 2022.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13260–13271, 2020.

Jonathan Cramer, Christoph P. Sager, and Beat Ernst. Hydroxyl groups in synthetic and natural-product-derived therapeutics: A perspective on a common functional group. *Journal of Medicinal Chemistry*, 62(20):8915–8930, 2019.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 2000.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. In *Conference on Neural Information Processing Systems*, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35, 2016.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272, 2017.

Rafael Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2): 268–276, 2018.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, pp. 2059–2073, 2022.

IUPAC. Functional group. Compendium of Chemical Terminology (Gold Book).

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. DrugOOD: Out-of-distribution dataset curator and benchmark for AI-aided drug discovery - a focus on affinity prediction problems with noise annotations. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14:7568, 2023. doi: 10.1038/s41467-023-43214-1.

Zhiqiang Li, Jianqing Liang, Zhiqiang Wang, Xizhao Luo, and Jiye Liang. Learning molecular semantic invariant representation with prototype constraint. In *International Conference on Machine Learning*, 2026.

Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph rationalization with environment-based augmentations. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1069–1078, 2022a.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022b.

Grégoire Mialon, Dexiong Chen, Margot Selosse, and Julien Mairal. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, volume 162, pp. 15524–15543, 2022.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1696–1705, 2022.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision - ECCV Workshops - Amsterdam, The Netherlands*, volume 9915, pp. 443–450, 2016.

Vladimir N Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Ruijia Wang, Haoran Dai, Cheng Yang, Le Song, and Chuan Shi. Advancing molecule invariant representation via privileged substructure identification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3188–3199, 2024.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, 2022. doi: 10.1038/s42256-022-00447-x.

Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

Aming Wu and Cheng Deng. CFD: learning generalized molecular representation via concept-enhanced feedback disentanglement. In *International Conference on Learning Representations*, 2025.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022.

Zhanghao Wu, Paras Jain, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13266–13279, 2021.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs. *arxiv.org/abs/2110.01717*, 2021.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. In *Conference on Neural Information Processing Systems*, 2022.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning invariant molecular representation in latent discrete space. In *Conference on Neural Information Processing Systems*, 2023.

A Appendix

A.1 Method Details

Table 6 shows the basic information of 53 functional groups that are from 7 categories.

Table 6: The basic information of 53 functional groups.

ID	Name	Smarts	Graph Structure
	(1) Carbon-carbon Unsaturated Bonds		
01	Alkene	[C]=[C]	
02	Alkyne	[C]#[C]	
	(2) Ring Structures		
03	Benzene	C1CCCCC1	

Continued on next page

Table 6 – continued from previous page

ID	Name	Smarts	Graph Structure
04	Pyrrole	[NH]1CCCC1	
05	Furan	C1CCOC1	
06	Thiophene	C1CCSC1	
07	Imidazole	C1NCC[NH]1	
08	Pyrazole	[NH]1NCCC1	
09	Oxazole	O1CNCC1	
10	Thiazole	S1CNCC1	
11	Pyridine	N1CCCCC1	
12	Pyridazine	N1NCCCC1	
13	Pyrimidine	N1CNCCC1	
14	Pyrazine	N1CCNCC1	
15	Naphthalene	C1CCC2CCCCC2C1	
16	Indole	C1CCC2[NH]CCC2C1	
17	Quinoline	C1CCC2NCCCC2C1	
18	Isoquinoline	C1CCC2CCNC2C1	
19	Benzofuran	C1CCC2OCCC2C1	
20	Benzothiophene	C1CCC2SCCC2C1	
		(3) Oxygen-containing	
21	Hydroxyl	[OX2H]	
22	Alcohol	[CX4][OX2H]	
23	Phenol	[C][OX2H]	

Continued on next page

Table 6 – continued from previous page

ID	Name	Smarts	Graph Structure
24	Ether	[OD2]([#6])[#6]	
25	Carbonyl	[CX3] = [OX1]	
26	Aldehyde	[CX3H1](= O)[#6]	
27	Ketone	[#6][CX3](= O)[#6]	
28	Carboxylic_acid	[CX3](= O)[OX2H1]	
29	Ester	[CX3](= O)[OX2][#6]	
30	Ccid_halide	[CX3](= O)[F, Cl, Br, I]	
31	Acid_anhydride	[CX3](= O)O[CX3](= O)	
32	Peroxide	[OX2][OX2]	
	(4) Nitrogen-containing		
33	Primary_amine	[NX3; H2]	
34	Secondary_amine	[NX3; H1]	
35	Tertiary_amine	[NX3; H0]	
36	Amide	[NX3][CX3](= O)	
37	Nitrile	[CX2]#[NX1]	
38	Nitro	[N+](= O)[O−]	
39	Nitroso	[N] = O	
40	Azo	[N] = [N]	
41	Imine	[CX3] = [NX2]	
42	Isocyanate	N = C = O	
	(5) Sulfur-containing		

Continued on next page

Table 6 – continued from previous page

ID	Name	Smarts	Graph Structure
43	Thiol	[SX2H]	
44	Thioether	[SX2]([#6])[#6]	
45	Disulfide	[SX2][SX2]	
46	Sulfoxide	[#6][S](= O)[#6]	
47	Sulfone	[#6][S](= O)(= O)[#6]	
48	Sulfonic_acid	[S](= O)(= O)[OX2H]	
49	Sulfonamide	[S](= O)(= O)N	
	(6) Halogen-containing		
50	Halogen	[F, Cl, Br, I]	
51	Alkyl_halide	[#6][F, Cl, Br, I]	
	(7) Phosphorus-containing		
52	Phosphate	[P](= O)(O)(O)	
53	Organophosphorus	[#6][P]	

Algorithm 1 provides the detailed process for calculating $Match(F, G)$. The $FindNext(F, G)$ in the 2nd and 9th lines find the next subgraph F in G. If the match is successful, it returns the sequence of node numbers corresponding to this subgraph in G. Otherwise, it returns **None**. For example, $I = (3, 4, 7, 9, 2) = FindNext(F, G)$ means that the subgraph consisting of atoms numbered 3, 4, 7, 9, 2 in G is exactly the functional group F. It should be noted that the order of elements in I is fixed.

A.2 Supplementary Experimental Details and Results

Data sets. In the experiments, we follow the official training, validation, and test splits. The environment-partitioning strategies used to construct distribution shifts, Specifically,

Scaffold: Molecules sharing the same scaffold belong to the same environment, while scaffold changes induce distribution shifts.

Size: Molecular size is measured by the number of atoms, and changes in size induce distribution shifts.

Assay: Molecules evaluated using the same assay belong to the same environment, while assay changes induce distribution shifts.

The detailed information of data sets are summarized in Table 7. Table 8 reports the detailed hyperparameters configurations in the experiments. These hyperparameters cover the tokenization module, the GPS architecture, and the optimizer.

Algorithm 1 $Match(F, G)$

Require: A functional group graph $F = (\hat{X}, \hat{E}, \hat{A})$ and a molecular graph $G = (X, E, A)$, where $\hat{A} \in \{0,1\}^{\hat{n}\times\hat{n}}$ and $A \in \{0,1\}^{n\times n}$ are the adjacency matrices of the functional group graph and the molecular graph, respectively.
Ensure: The cumulative adjacency matrix $C \in \mathbb{R}^{n\times n}$.

```
Initialize: ∀i, j = 1, 2, ··· n, C_ij ← 0.
I ← FindNext(F, G)
while I ≠ None do
    for i = 1 to n̂ do
        for j = 1 to n̂ do
            C_{I_i I_j} ← C_{I_i I_j} + Â_ij
        end for
    end for
    I ← FindNext(F, G)
end while
return C
```

Table 7: Statistics and evaluation metrics of the data sets.

Data set	Shift	Task	Metric	#Train	#Val	#Test	#Tasks
GOOD-HIV	scaffold/covariate	Binary Classification	ROC-AUC	24682	4133	4108	1
	scaffold/concept	Binary Classification	ROC-AUC	15209	9365	10037	1
	size/covariate	Binary Classification	ROC-AUC	26169	4112	3961	1
	size/concept	Binary Classification	ROC-AUC	14454	3096	10525	1
GOOD-ZINC	scaffold/covariate	Regression	MAE	149674	24945	24946	1
	scaffold/concept	Regression	MAE	101867	43539	60393	1
	size/covariate	Regression	MAE	161893	24945	17042	1
	size/concept	Regression	MAE	89418	19161	70306	1
DrugOOD-IC50	assay/covariate	Binary Classification	ROC-AUC	34953	19475	19463	1
	scaffold/covariate	Binary Classification	ROC-AUC	22025	19478	19480	1
	size/covariate	Binary Classification	ROC-AUC	37497	17987	16761	1
DrugOOD-EC50	assay/covariate	Binary Classification	ROC-AUC	4978	2761	2725	1
	scaffold/covariate	Binary Classification	ROC-AUC	2743	2723	2762	1
	size/covariate	Binary Classification	ROC-AUC	5189	2495	2505	1

Supplementary results. Table 9 reports the number of FGs in each molecule and reported the average for each data set. Figure 4 visualizes the additional FG matching results and the soft partitioning of atoms to molecular tokens.

Table 8: Configurations of the proposed method in the experiments.

Data set	Shift	Tokenization				GPS_1, GPS_2			Optimizer	
		γ	K	d'_1	k	n_l	n_h	dropout	bs	lr
GOOD-HIV	scaffold/covariate	0.1	15	128	3	5	4	0.5	256	0.001
	scaffold/concept	0.1	15	300	5	5	4	0.5	256	0.001
	size/covariate	0.001	15	128	3	5	4	0.5	256	0.001
	size/concept	0.001	15	300	5	5	4	0.5	256	0.001
GOOD-ZINC	scaffold/covariate	1.0	15	300	3	5	4	0.5	256	0.001
	scaffold/concept	1.0	15	300	3	5	4	0.5	256	0.001
	size/covariate	0.1	15	300	3	5	4	0.5	256	0.001
	size/concept	0.001	15	300	3	5	4	0.5	256	0.001
DrugOOD-IC50	assay/covariate	0.1	20	300	3	5	4	0.2	256	0.002
	scaffold/covariate	0.1	20	300	3	5	4	0.2	256	0.001
	size/covariate	0.1	20	300	3	5	4	0.2	256	0.001
DrugOOD-EC50	assay/covariate	1.0	20	300	3	5	4	0.2	256	0.001
	scaffold/covariate	0.1	20	300	3	5	4	0.2	128	0.002
	size/covariate	0.01	20	300	3	5	4	0.2	128	0.001

Table 9: The statistics of FG occurrences in molecules across different data sets.

Data set	Shift	#Molecule	#FG (Type)	#FG (Number)
GOOD-HIV	scaffold/covariate	32903	5.76	9.65
	scaffold/concept	34611	5.84	9.88
	size/covariate	34242	5.65	9.36
	size/concept	28075	5.77	9.71
GOOD-ZINC	scaffold/covariate	199565	6.50	9.01
	scaffold/concept	205799	6.60	9.21
	size/covariate	204240	6.51	9.01
	size/concept	178885	6.58	9.15
DrugOOD-IC50	assay/covariate	73891	7.26	11.90
	scaffold/covariate	60983	7.14	11.30
	size/covariate	72245	7.10	11.26
DrugOOD-EC50	assay/covariate	10464	7.47	17.04
	scaffold/covariate	8228	7.23	14.81
	size/covariate	10189	7.17	14.08

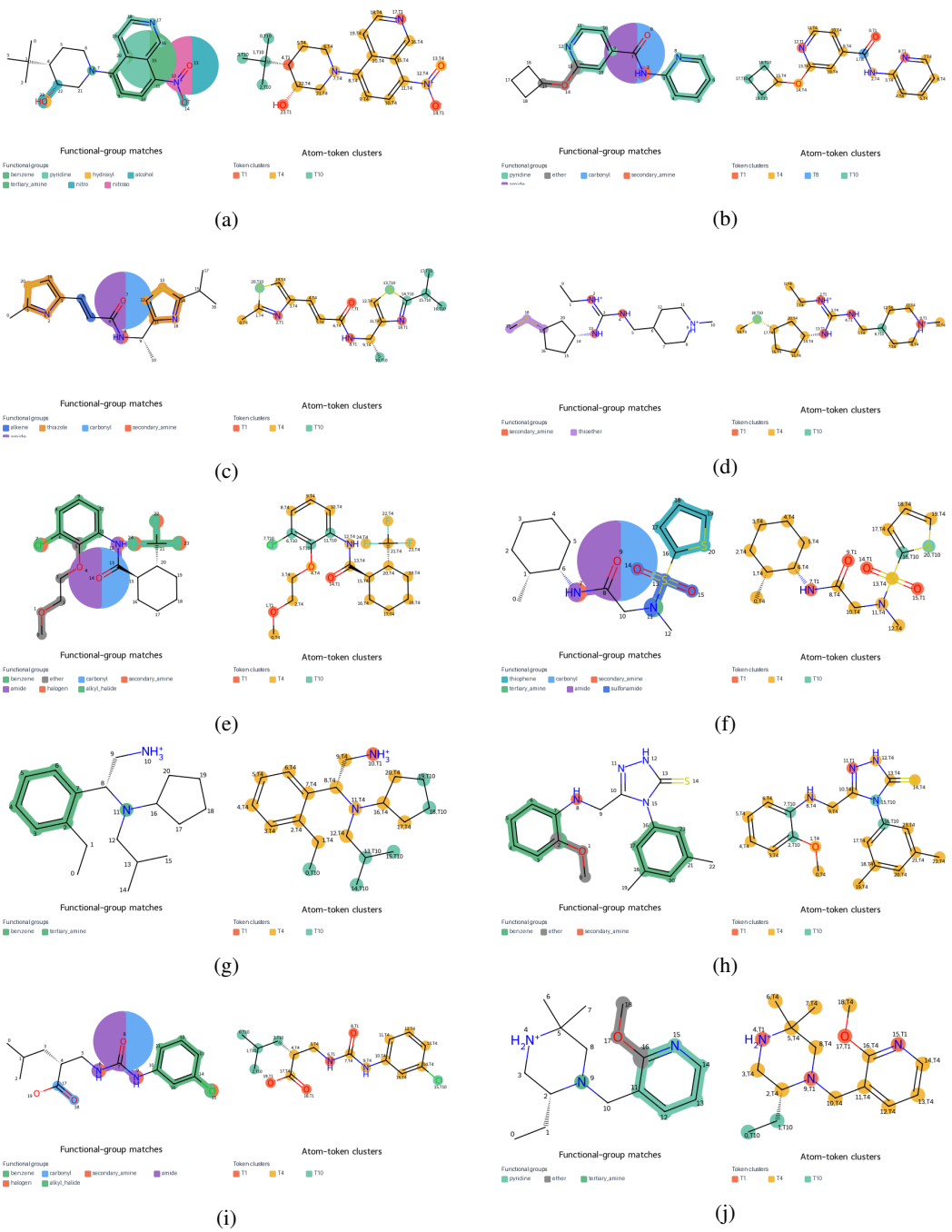


Figure 4: Visualization of functional group matching and atom clustering (continued from Figure 3).