Information Processing and Management 63 (2026) 104771

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

From graphs to tokens: Substructure-aware molecular representation for large language models

Runze Wang [a], Zijie Xing [b], Xingyue Liu [b], Mingqi Yang [c], Che He [a], Yanming Shen [a,b,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China
[b] School of Future Technology, Dalian University of Technology, Dalian, 116024, China
[c] Department of Electronic Business, South China University of Technology, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
Substructure-aware tokenization
Molecular graph
Large language models
Substructure token embedding
Substructure token alignment

ABSTRACT

Enabling large language models (LLMs) to process graph-structured data holds significant potential for advancing molecular research, especially in tasks related to molecular structure analysis and understanding. However, a key challenge remains in effectively tokenizing molecular graphs to align with the capabilities of LLMs. To bridge the gap, this paper proposes **S**ub**S**tructure-aware **Token**ization (S^2Token) that fragments molecules into chemically meaningful, reusable substructures. The key idea is to generate semantically enriched substructure token embeddings and represent the complex relationships between tokens in the LLM embedding space, allowing LLMs to learn and generalize chemical patterns effectively. To achieve this, S^2Token begins by decomposing molecular graphs based on frequency and functionality to build a vocabulary. Then, we establish a dual-view token embedding method that jointly encodes fine-grained structural and functional attributes of molecular substructures. Furthermore, we propose an efficient token alignment strategy to associate the dependencies between non-sequential substructures with the LLM token embedding space. In this way, our method activates the LLMs' generalization ability to unseen molecular graphs. To evaluate this, we curate four task-specific molecular datasets designed to investigate generalization. Extensive experiments on three standard benchmarks and four curated datasets demonstrate that S^2Token consistently yields substantial performance improvements. On the molecular caption benchmark, S^2Token achieves a 12.6% average performance gain across six metrics, outperforming the best LLM-based method. For forward reaction and retrosynthesis tasks, it enhances fingerprint similarity for synthesized molecules by 6.6% and 8.6%, respectively. Moreover, S^2Token shows advantages on four generalist evaluation tasks, including 9.7% and 3.0% reductions in MAE on the molecular property prediction task, compared to graph- and node-centric tokenization methods. Code is available at https://github.com/GraphMoLab/S2Token.

1. Introduction

Large Language Models (LLMs) have shown strong potential in processing molecular graph data to capture the complexity of molecular systems. By leveraging their vast prior knowledge, LLMs offer new opportunities through their ability to generalize across

* Corresponding author.
E-mail addresses: runze_wang@mail.dlut.edu.cn (R. Wang), kokojjj@mail.dlut.edu.cn (Z. Xing), aroma@mail.dlut.edu.cn (X. Liu), yangmq@scut.edu.cn (M. Yang), heche@mail.dlut.edu.cn (C. He), shen@dlut.edu.cn (Y. Shen).

https://doi.org/10.1016/j.ipm.2026.104771
Received 14 October 2025; Received in revised form 21 February 2026; Accepted 21 March 2026
Available online 28 March 2026

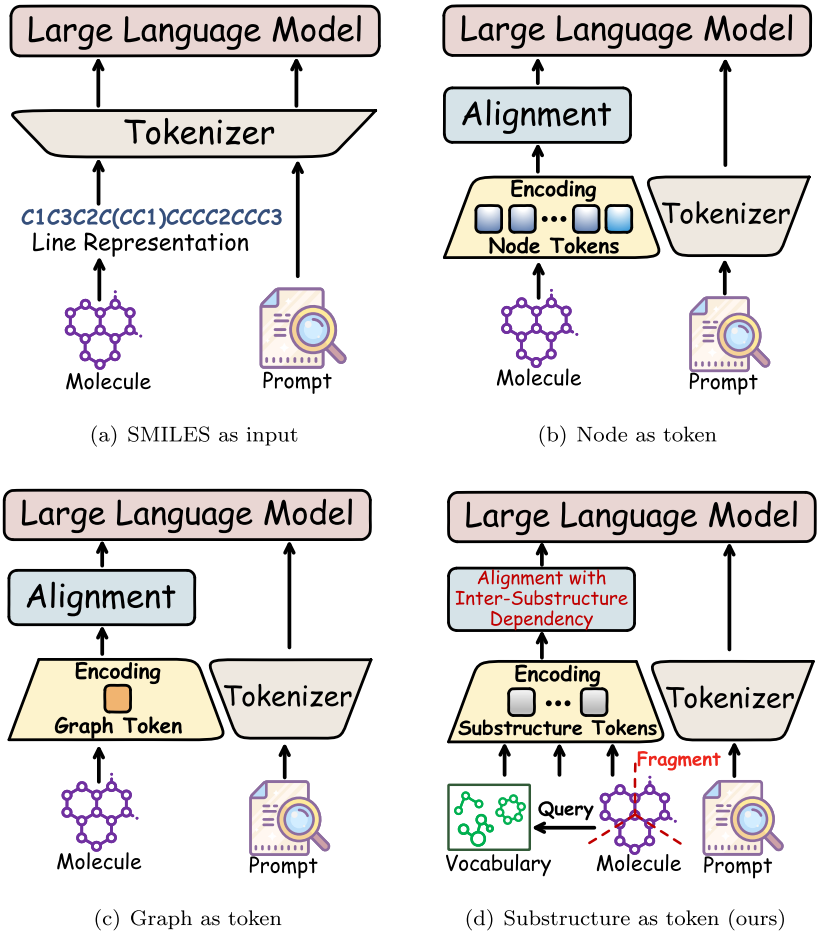


Fig. 1. Different molecular graph tokenization approaches for LLMs.

a wide range of tasks, such as predicting forward reaction and molecular properties (Chen et al., 2025; Park et al., 2024; Wang et al., 2025b). However, optimal graph tokenization strategies remain the critical bottleneck for fully bridging LLMs and molecular graphs. The inherently non-Euclidean nature of molecular structures necessitates specialized tokenization approaches that preserve graph-specific semantics. Designing an effective molecular graph tokenizer is key to unlocking the full potential of LLMs for molecular research.

Tokenization in LLMs, which treats subwords rather than individual characters as semantic units (Gastaldi et al., 2025), strikes a balance between *representativeness* and *generalization*. While this paradigm facilitates the processing of sequence molecular representations such as SMILES (Fang et al., 2024; Yu et al., 2024) (Fig. 1(a)), it fails to capture the non-Euclidean structural constraints intrinsic to molecular graphs. To address the limitation, recent approaches typically integrate graph neural networks (GNNs) to pre-encode molecular graphs prior to aligning their embeddings with the LLM token space. One straightforward method involves node-level tokenization (Fig. 1(b)), where atomic nodes are treated as independent tokens. The resulting node embeddings are either projected through linear alignment layers (Cao et al., 2025; Chen et al., 2025), or compressed into fixed-length vectors (Li et al., 2024b) before being input into LLMs. An alternative strategy encodes the entire molecular graph into a unified vector representation (Fig. 1(c)), abstracting the structural information into a holistic embedding (Le et al., 2024; Liang et al., 2024; Wang et al., 2025b).

However, from a tokenization perspective, both strategies fall short in preserving the critical trade-off between *representativeness* and *generalization*. Node-level token leads to excessive fragmentation, which obscures higher-order structural semantics (e.g., carbon atoms in benzene rings are treated the same as those in methane, losing important chemical distinctions). And graph-level token embeddings sacrifice atomic granularity and relational topology, both of which are essential for chemical interpretation.

Molecule graphs exhibit chemically meaningful substructures (e.g., aromatic rings) that align with high-order structures and correspond to functional behaviors. Recurring substructures across molecules enable the reuse of learned features, promote parameter sharing, and support generalization to unseen molecules containing the same substructures. This mirrors the advantages of subword-level tokenization in LLMs, which captures meaningful units and generalizes better than character-level approaches.

Therefore, an intuitive solution is to treat substructures as 'chemical subwords', analogous to token units in LLMs. Yet, molecular properties emerge from the interplay between the intrinsic features of constituent substructures (**intra-substructure**) and the interactions between these substructures (**inter-substructure**). The non-sequential nature of both characteristics poses challenges: (1) How to obtain semantical-rich substructure embeddings? and (2) Can the aligned tokens preserve inter-substructure dependencies in non-Euclidean space?

In this paper, we propose **S^{2}Token**, a **S**ub**S**tructure-aware molecular graph tokenizer tailored to enhance both *representativeness* and *generalization* in molecular LLMs (Fig. 1(d)). Inspired by subword tokenization, S^{2}Token treats molecular substructures as discrete tokens, leveraging their chemical relevance and frequent recurrence to ensure both *representativeness* and *generalization*. S^{2}Token first decomposes molecular graphs using a ring-driven strategy to build a balanced vocabulary of reasonable size. To obtain semantically rich substructure embeddings, we propose a dual-view embedding approach that captures intra-substructure relationships through a hierarchical representation and encodes chemical identifiers as functional features. To align these substructure token embeddings with LLMs, we introduce a learning mechanism that captures inter-substructure dependencies using a substructure-level Transformer, guided by structural encodings. For evaluation, we construct cross-dataset molecular benchmarks spanning four tasks to assess S^{2}Token's generalization capability. Experimental results demonstrate that our method consistently achieves strong performance in both *generalization* and *representativeness*, fulfilling the core design objectives. Our contributions are as follows:

- We propose to take functional molecular substructure as tokens and introduce a substructure-aware tokenizer, aligning intra-substructure relationships and inter-substructure dependencies to LLM token embeddings.
- We construct generalist evaluation datasets across four molecular tasks to assess the generalization performance of various molecular tokenization approaches.
- Extensive experiments on both standard benchmarks and newly generalist datasets validate the effectiveness of our approach.

2. Related work

2.1. Graph representation learning for molecules

Accurate molecular representations are fundamental for extracting feature embeddings that capture both atomic properties and structural relationships (Atz et al., 2021), which are key to predicting various physicochemical properties. Early methods, such as handcrafted descriptors and fingerprint-based approaches, represent molecules as vectors from predefined features (Li et al., 2022). While useful in some cases, these methods fail to capture the complex atomic interactions within molecular structures.

Given that molecules naturally form graph-structured data, where atoms are modeled as nodes and bonds as edges, graph-based models have emerged as a more suitable framework for molecular representation (Li et al., 2025a; Wang et al., 2025a). In recent years, Graph Neural Networks (GNNs) have attracted significant attention for their capacity to learn powerful representations directly from molecular graphs (Corso et al., 2024; Sypetkowski et al., 2024). To further enhance the expressive power, numerous studies have focused on improving the architectural capacity of GNNs (Zhang et al., 2024a; Zhao et al., 2026) and also explored the Graph Transformers (GTs) (Hong et al., 2026; Yuan et al., 2025) to better capture long-range dependencies and alleviate over-smoothing. Both GNNs and GTs aim to improve the discriminative capacity of learned molecular representations, allowing the extraction of more informative structural features to improve downstream task performance. In addition to architectural innovations, augmentative techniques, such as the positional encodings (Dwivedi et al., 2022) and certain structural features (Chen & Schwaller, 2024) within molecular graphs, have been proposed to further enhance the expressiveness of graph models.

Inspired by the success of Vision-Language Models (VLMs) (Zhang et al., 2024b; Zijing et al., 2025) that align images with textual descriptions in a shared embedding space, molecule-language models (Liu et al., 2024, 2023b) align the molecular structures (represented as graphs) with textual descriptions that convey the molecular properties, functions, or names. The central idea is to train a graph encoder and a text encoder, to produce aligned representations in a shared latent space. This alignment is typically achieved through a contrastive pre-training objective (Liu et al., 2023b; Yangding et al., 2025). Some studies also use cross-modal contrastive learning (Seidl et al., 2023) or encoder-decoder (Zhao et al., 2023) architectures to align the molecular graph structure with textual task-specific instructions, ultimately improving the accuracy of molecular property prediction.

2.2. Large language models for molecules

Recent studies have demonstrated the potential of LLMs on various molecular tasks (Liu et al., 2025; Zhang et al., 2025; Zheng et al., 2025). Some works (Guo et al., 2023; Sadeghi et al., 2024; Zhong et al., 2024) investigate the application of LLMs to molecular representations such as SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020), and InChI (Heller et al., 2013), treating molecules as text sequences processed directly by an LLM tokenizer. To further enhance the capability of molecular understanding, efforts (Fang et al., 2024; Pei et al., 2024; Yu et al., 2024) have focused on instruction tuning on curated high-quality datasets.

Inspired by advances in multi-modal LLMs (Xuyang et al., 2025), recent work has extended LLMs to molecular graph representations. These approaches often adopt Qformer architectures to align structured molecular inputs with LLM embeddings (Lee et al., 2025; Li et al., 2024b; Liu et al., 2023c; Park et al., 2024). Several studies employ specifically designed GNNs as molecular tokenizers, encoding atom-level embeddings (Cao et al., 2025) or quantized molecular features (Chen et al., 2025) as input to LLMs. Others explore whole-graph embeddings aligned to LLMs via adapter modules or projection layers (Le et al., 2024; Liang et al., 2024; Wang et al., 2025b). Some methods (Chen et al., 2025; Hu et al., 2024) treat hierarchical structures, such as node and function motifs,

as uniform tokens, arranging them into a structural sequence for LLMs. Overall, the absence of representative and generalizable molecular graph tokens limits the generalization performance of LLMs across molecular tasks.

2.3. Large language models for graph reasoning

Recent efforts have sought to adapt LLMs for graph tasks that involve implicit or explicit structural reasoning (Wang et al., 2025c), with particular attention to text-attributed graphs where nodes carry rich textual annotations. One line of work serialize graphs into natural language, enabling end-to-end inference (Fatemi et al., 2024; Wang et al., 2023; Zhang et al., 2024c). However, NLGIFT (Zhang et al., 2024c) reveals that the observed improvements often stem from memorization rather than true generalizable reasoning. An alternative direction focuses on alignment-based methods that embed graph topological structure into a shared representation space with text to enhance reasoning (Chen et al., 2024; Tang et al., 2024; Wang et al., 2024). For instance, GraphGPT (Tang et al., 2024) aligns LLMs with graph structures through graph-text alignment and the instruction tuning, achieving strong zero-shot transferability. These insights motivate the design to explicitly encode molecular graph structures, rather than relying solely on textual descriptions. Unlike text-attributed graphs, molecular graphs lack rich semantic annotations, and their properties are fundamentally governed by the underlying topology. Consequently, effective topological reasoning becomes the central challenge for enabling LLMs to understand and reason over molecular systems.

2.4. Tokenization and molecular substructures

Among various tokenization solutions (Barrault et al., 2024), subword-level tokenization has emerged as the prevailing strategy, primarily due to its balance in capturing semantic *representativeness* and *generalization*. For molecular graphs, substructures, such as rings, chains, and functional groups, serve as recurring patterns. These substructures capture higher-order interactions (Du et al., 2025) beyond simple pairwise atom connections and are often associated with functional properties (Ma et al., 2024). Tokenizing molecular graphs into substructures presents several advantages. First, substructures provide chemically interpretable and functional units, which can enhance explainability and discriminability, aiding GNNs in distinguishing non-isomorphic graphs (Han et al., 2023; Zeng et al., 2023b). Second, taking substructures as tokens promotes better generalization to unseen molecules (Wollschläger et al., 2024), which are largely composed of various motifs that have been seen before Li et al. (2025b). FLAG (Zhang et al., 2023) utilizes substructures as the basic units for 3D molecular geometry generation. To achieve conformation-aware generation, it prioritizes the cleavage of rotatable bonds while preserving conformational validity. In parallel, GROVER (Rong et al., 2020) demonstrates the efficacy of pretraining GNN models via motif-based self-supervised learning, substantially improving molecular property prediction. PharmHGT (Jiang et al., 2023b) pretrains a heterogeneous graph transformer model that integrates pharmacophore information and reaction details, thereby learning multi-view molecular representations for enhanced predictive performance. Recent advances in subgraph-based interpretability (Wu et al., 2023), grounded in subgraph matching theory, underscore the importance of subgraph-level reasoning for understanding and explaining the behavior of GNNs.

2.5. Comparison with existing methods

Our proposed S^2Token framework distinguishes itself from existing approaches by fundamentally rethinking the graph tokenization units for LLMs. As illustrated in Fig. 1, the key difference between our approach (Fig. 1(d)) and other molecular tokenization strategies for LLMs is highlighted.

Comparison with Text-based LLMs. As depicted in Fig. 1(a), a line of work directly processes molecular SMILES strings using standard LLM tokenizers (Fang et al., 2024; Guo et al., 2023). While efficient, they inherently face challenges in capturing explicit structural information, which is crucial for defining atomic relationships in molecular graphs. In contrast, our S^2Token operates directly on the graph structure, preserving structure information that is lost in linearized string representations.

Comparison with Graph-based LLMs. The core work in graph-based LLM methods is to bridge GNNs and LLMs. One common approach is node-level tokenization (Cao et al., 2025), where each atom is treated as an independent token (see Fig. 1(b)). However, this leads to excessive fragmentation, obscuring higher-order semantics (e.g., chemically meaningful functional groups). Alternatively, some methods employ graph-level tokenization (Liang et al., 2024; Wang et al., 2025b), encoding the entire graph into a single, holistic embedding vector before feeding it into the LLM backbone. While this simplifies the representation (see Fig. 1(c)), it omits fine-grained structural and atomic details, limiting the LLM's ability to perform reasoning that requires topological information. In contrast, our S^2Token (illustrated in Fig. 1(d)) strikes a balance by adopting substructures as tokens. This maintains atomic-level granularity within a functional context, while enabling cross-molecular knowledge transfer through a reusable vocabulary.

3. Research objective

The primary goal of this research is to bridge the gap between molecular graph-structured data and large language models (LLMs). While LLMs are promising in molecular tasks, their effectiveness is limited by the inability of current tokenization strategies to represent non-Euclidean molecular graphs adequately. Existing approaches, such as node-level or graph-level tokenization, either excessively fragment molecular graphs or oversimplify structural relationships. As a result, they struggle to balance *representativeness* and *generalization*, which are crucial for LLMs to leverage molecular tokens.

To address these limitations, this study introduces S^2Token, a SubStructure-aware Tokenization framework, with the following specific research objectives:

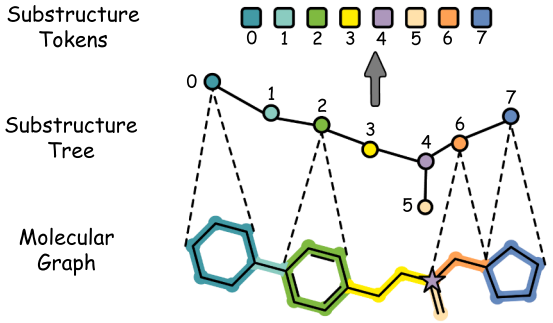


Fig. 2. Overview of molecular graph fragmentation. The molecular graph is fragmented and then converted into a substructure tree, where nodes (0, 2, 7) denote rigid scaffolds, nodes (1, 3, 5, 6) denote flexible linkers, and node (4) functions as a junction connecting three substructures (corresponding to the purple star node in the molecular graph). All nodes in the tree are treated as tokens to represent the molecular graph for large language model input. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- Design a substructure-aware graph tokenization strategy that decomposes molecular graphs into chemically meaningful and frequently recurring substructures, forming a reusable vocabulary analogous to subwords in natural language processing.
- Develop a token embedding method that jointly encodes intra-substructure relationships and functional attributes of molecular substructures, facilitating semantically meaningful token representations.
- Propose an efficient token alignment mechanism that preserves the inter-substructure dependencies between non-sequential token relationships within the LLM token embedding space.
- Evaluate the proposed method on both standard benchmarks and newly curated generalist datasets, demonstrating its superiority in terms of representativeness and generalization.

By achieving these objectives, this work seeks to enable LLMs to effectively handle molecular graph data, thereby broadening their applicability in molecular research.

4. Method

S^2Token is a substructure-aware tokenization framework tailored for aligning molecular graphs with large language models. In this way, it mitigates the challenges faced by graph tokenization approaches in molecular LLMs, which struggle to achieve both *representativeness* and *generalization*.

4.1. Substructure vocabulary construction

Similar to subword tokenization in language models, the molecular graph vocabulary consists of pure substructure tokens, derived by fragmenting molecules following three principles: high frequency, chemical validity, and non-redundancy. However, overly coarse fragmentation reduces cross-molecular token reuse and impairs generalization, whereas excessively fine splitting loses structural semantics, such as breaking pharmacophores. To trade off fragmentation granularity, we prioritize **rings as rigid scaffolds**, as rings constitute the structural backbone of over 85% of bioactive molecules. Treating rings as indivisible units (e.g., benzene, piperazine) preserves pharmacophore integrity, avoids arbitrary bond cleavage that could disrupt conjugated systems, and enhances the expressive power of GNNs. Non-ring regions connecting critical nodes are mapped to chemically meaningful functional connectors like carboxyl (C(=O)O) and alkyl chains, which consist of multiple bonds forming linear path-like topologies (as illustrated in yellow color in Fig. 2). Based on the above, we naturally model **paths as flexible linkers** beyond ring systems.

Given a molecular dataset, we begin by extracting a set of substructures to construct a token vocabulary. As shown in Fig. 2, each molecular graph is then converted into a substructure tree based on the extracted substructures. The nodes in the resulting tree are subsequently extracted and linearized into a graph token sequence. The overall fragmentation workflow is inspired by Wollschläger et al. (2024), which first extracts all minimal rings, and then connects the remaining edges at nodes of degree two to form paths. In cases where three or more fragments converge at a node, a junction node is introduced in the higher-level graph to organize connections, as illustrated by the purple node in the substructure tree in Fig. 2.

4.2. Substructure token embedding

In language models, each token is assigned a unique identifier and mapped to a dense vector representation via a learned embedding table. Molecular substructures, however, present a fundamentally different challenge, as they serve dual roles: acting as chemically functional identifiers and hierarchical graph structures. Chemical identifiers govern molecular properties and support effective model generalization across molecules. As hierarchical structures, substructures carry the intra-substructure relationships between atomic nodes and fine-grained graph topology. While it is possible to embed discrete substructure identifiers analogously to

language tokens in LLMs, such an approach fails to handle the structural characteristics. We introduce a dual-view substructure token embedding framework that integrates (1) intra-substructure relationships and (2) chemical identifiers, enabling token embeddings that are both structurally discriminative and generalizable to novel molecules (Algorithm 1).

Algorithm 1 Dual-view substructure token embedding framework.

Input: Molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$;
substructure tokens $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$, with associated node set $\{\mathcal{V}_{s_1}, \mathcal{V}_{s_2}, \ldots, \mathcal{V}_{s_{|\mathcal{S}|}}\}$;
node representations produced by GNN encoder $\mathbf{H}$;
RWPE positional encodings $\mathbf{P}$;
weight matrices $\mathbf{W}^0$, $\mathbf{W}^{PE}$
Output: Substructure token embedding set $\mathbf{S}$ of $\mathcal{G}$

Initialize: Substructure token embedding set $\mathbf{S} \leftarrow \emptyset$
for s_k in $\mathcal{S}$ **do**
  # Intra-substructure relationships embedding
  Aggregate atomic features
  $\mathbf{e}_k = \text{Pooling}(\mathbf{W}^0 \mathbf{H}_{\mathcal{V}_{s_k}})$
  Aggregate structural encodings
  $\mathbf{p}_k = \text{Pooling}(\mathbf{W}^{PE} \mathbf{P}_{\mathcal{V}_{s_k}})$
  $\widehat{\mathbf{e}}_k = \text{Sum}(\mathbf{e}_k, \mathbf{p}_k)$
  # Chemical identifiers encoding
  Determine structural class
  $c_k = \text{class}(s_k) \in \{\text{path}, \text{cycle}, \text{junction}\}$
  Compute bond-count scalar encoding
  $b_k = \text{EmbedBondCount}(|E_{s_k}|)$
  Form chemical identifier vector
  $d_k = \text{EmbedClass}(c_k) * b_k$
  Obtain final dual-view substructure token embeddings
  $\widetilde{\mathbf{e}}_k = \widehat{\mathbf{e}}_k \,\|\, d_k$
  $\mathbf{S} \leftarrow \mathbf{S} \cup \widetilde{\mathbf{e}}_k$
end for
Return: $\mathbf{S} = \{\widetilde{\mathbf{e}}_1, \ldots, \widetilde{\mathbf{e}}_{|\mathcal{S}|}\}$

4.2.1. *Intra-substructure relationships embedding*

To generate fine-grained and expressive token embedding for LLM alignment, the intra-substructure relationships embedding covers: (1) atomic attributes and arrangement, derived through atom-level message passing; and (2) detailed substructure topology, modeled via explicit structural encoding. For atom-level message passing, we adopt a lightweight design via a frozen GNN encoder pretrained on large-scale molecule-text paired data (Liu et al., 2023b).

Given a molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the sets of nodes and edges, respectively, the graph is decomposed into a sequence of substructure tokens $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$. Let the encoder be an L-layer GNN, where the node features at the lth layer are denoted as $\mathbf{H}^{(l)} \in \mathbb{R}^{|\mathcal{V}| \times d_n}$, with d_n being the hidden dimension. For a substructure token $s_k \in \mathcal{S}$, defined by its associated node set $\mathcal{V}_{s_k} \subseteq \mathcal{V}$, we project the node features $\{\mathbf{H}_v^{(l)} | v \in \mathcal{V}_{s_k}\}$ into substructure embedding space and aggregate them to form the embedding of s_k:

$$\mathbf{e}_k^{(l)} = \frac{1}{|\mathcal{V}_{s_k}|} \sum_{i \in \mathcal{V}_{s_k}} \mathbf{W}^0 \mathbf{h}_i^{(l)}, \quad \mathbf{h}_i^{(l)} \in \mathbf{H}^{(l)}, \tag{1}$$

where $\mathbf{W}^0 \in \mathbb{R}^{d_s \times d_n}$ is a learnable projection matrix and d_s is the dimension of substructure embedding.

To capture fine-grained topology, we incorporate explicit structural encodings derived from Random Walk Positional Encoding (RWPE) (Dwivedi et al., 2022), which has shown significant capabilities in distinguishing molecular graphs. Specifically, RWPE features are computed at the node level and subsequently projected into the substructure embedding space. For each substructure s_k, the projected positional encodings of its constituent nodes are aggregated to obtain a structural representation $\mathbf{p}_k$. This structural vector is then combined with the embedding $\mathbf{e}_k$ via summation to produce a new substructure representation $\widehat{\mathbf{e}}_k$.

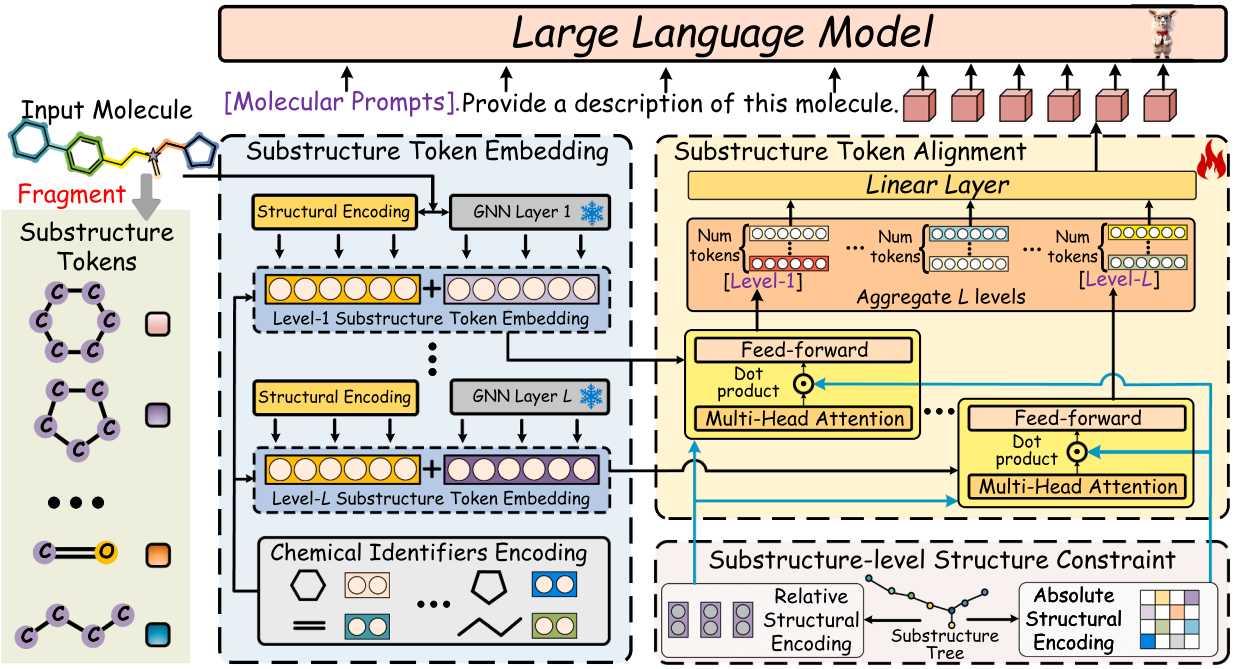


Fig. 3. Illustration of S^{2}Token on aligning substructure tokens with LLM token space. The framework consists of two key components: substructure token embedding and substructure token alignment. A ring-driven fragmentation rule decomposes the input molecule into substructure tokens, which are embedded using intra-substructure relationships (GNN-based node arrangement, structural topology, and chemical identifiers) and enriched with multi-level graph semantics. Substructure token alignment captures the inter-substructure dependencies via multi-head attention with structural constraints on the coarsened substructure tree. The aggregated attention outputs are projected into the LLM token space, and combined with molecular prompts (e.g., SMILES) and task-specific instructions to form the final inputs to the LLM backbone.

4.2.2. *Chemical identifiers encoding*

Substructure tokens serve as chemical identifiers and require an encoding scheme that facilitates parameter sharing among the same substructures, analogous to word embeddings in language models. To this end, we encode each substructure ID using two components: (1) a structural class embedding based on the fragment type, $class(s_k) \in \{path, cycle, junction\}$, and (2) a scaled scalar embedding that captures the number of chemical bonds (i.e., edges) within the substructure. These two elements are combined to form the chemical identifier encoding. This design enables the model to flexibly represent a theoretically unbounded vocabulary of substructures, promoting generalization across chemically similar functional units. The final substructure embedding $\widetilde{\mathbf{e}}_k$ is then obtained by concatenating the chemical identifier encoding with $\widehat{\mathbf{e}}_k$.

4.2.3. *Extensive multi-level graph representations*

Graph neural networks exhibit a layer-wise learning behavior, with different layers capturing structural and semantic information at varying levels of granularity. In molecular graphs, this aligns naturally with chemical intuition: shallow layers primarily encode local atomic environments (e.g., bond types, hybridization states), intermediate layers emphasize functional group patterns, and deeper layers capture global scaffold-level interactions. To fully exploit this layer-wise representation, we preserve the complete set of node embeddings from all L layers of the GNN encoder, denoted as $\{\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(L)}\}$. As illustrated in the substructure token embedding module in Fig. 3, we apply the **Substructure Token Embedding** operation to each layer, producing a corresponding sequence of substructure-level embeddings $\{\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(L)}\}$, where each $\mathbf{S}^{(l)} = \{\widetilde{\mathbf{e}}^{l}_{1}, \ldots, \widetilde{\mathbf{e}}^{l}_{|S|}\}$ encodes the substructures extracted at the lth layer.

4.3. *Substructure token alignment with inter-substructure dependency*

The goal of this section is to effectively align substructure embeddings with the token embedding space of LLMs. It is well established that both intra- and inter-substructure relationships jointly determine molecular properties and behaviors. However, unlike language tokens, inter-substructure dependencies in molecular graphs are non-Euclidean data, which results in the loss of crucial interaction information when directly aligning substructures with LLM token embeddings. To address this challenge, we propose an alignment method that fuses the intrinsic interactions between substructures into a shared token embedding space. This approach leverages a multi-head attention network to capture the interaction between substructures, as shown in the substructure token alignment module in Fig. 3, augmented by complementary absolute and relative structural encodings that serve as topological constraints in the non-Euclidean space.

Given the set $\{\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(L)}\}$, along with a coarsened substructure tree, we calculate the relative structural encoding using the RWPE derived from the substructure tree. The next step involves projecting each embedding $\mathbf{S}^{(l)}$ and the corresponding structural encoding $\mathbf{P}^{(l)}$ into a shared alignment space of dimension d_a. This projection is followed by layer normalization, as detailed below:

$$\tilde{\mathbf{S}}^{(l)} = \mathrm{LayerNorm}\left(\mathbf{W}^1\mathbf{S}^{(l)} + \mathbf{W}^2\mathbf{P}^{(l)}\right), \tag{2}$$

where $\mathbf{W}^1 \in \mathbb{R}^{d_{\tilde{s}} * d_a}$ and $\mathbf{W}^2 \in \mathbb{R}^{d_p * d_a}$ are learnable weight matrices. $d_{\tilde{s}}$, d_p and d_a define the dimensions of $\mathbf{S}^{(l)}$, relative structural encodings, and alignment space, respectively. Subsequently, the normalized embeddings are passed through a standard Transformer network. We further derive an absolute structural encoding on the coarsened substructure tree. Given the molecular graph adjacency matrix $\mathbf{A}^{\mathcal{G}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and the vertex sets $\{\mathcal{V}_{s_1}, \ldots, \mathcal{V}_{s_k}\}$ associated with the substructures, we construct a coarsened adjacency matrix $\mathbf{A}^{S}$ over the substructure tree as follows:

$$\mathbf{A}^{S}_{ij} = |\mathcal{V}_{s_i} \cap \mathcal{V}_{s_j}|, \tag{3}$$

which is then used to modulate the attention matrix via a Hadamard product $\odot$, guiding the attention mechanism:

$$\mathrm{Attention}^{(l)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}^{(l)})^{\top}}{\sqrt{d}} \odot \mathbf{A}^{S}\right)\mathbf{V}^{(l)}, \tag{4}$$

where $\mathbf{Q}^{(l)} = \tilde{\mathbf{S}}^{(l)}\mathbf{W}_Q$, $\mathbf{K}^{(l)} = \tilde{\mathbf{S}}^{(l)}\mathbf{W}_K$, $\mathbf{V}^{(l)} = \tilde{\mathbf{S}}^{(l)}\mathbf{W}_V$ are the query, key, and value matrices, respectively. Finally, we aggregate the attention outputs from all L layers:

$$\mathbf{Z} = [\mathrm{Attention}^{(0)} \oplus \cdots \oplus \mathrm{Attention}^{(L-1)}] \in \mathbb{R}^{d_a * L}, \tag{5}$$

where $\oplus$ denotes concatenation along the feature dimension. This procedure inherently integrates inter-substructure dependencies across shallow, intermediate, and deeper levels, which are subsequently projected into the LLM token embedding space with linear layers (Liu et al., 2023a).

4.4. Training S^2Token

Similar to recent multi-modal LLMs, S^2Token is trained via a two-stage pipeline: pre-training on molecule-text pairs for substructure-token alignment, and (2) instruction tuning for downstream task adaptation.

4.4.1. Stage 1

We freeze the LLM backbone and train only the substructure token alignment components, along with a few parameters from the substructure embedding module, such as $\mathbf{W}^0$ in Eq. (1), which projects the features to share a unified dimension. Both intra-substructure relationship encoding and chemical identifier encoding are performed offline, and therefore do not introduce extra trainable parameters. The model is trained to generate an answer $\mathcal{A}$ under an autoregressive next-token prediction paradigm:

$$p_\theta(\mathcal{A}|\mathcal{T}, S) = \Pi_{i=1}^{|\mathcal{A}|} p_\theta(a_i|\mathcal{T}, f(\mathrm{Emb}(S)), a_{<i}), \tag{6}$$

where $a_{<i}$ indicates the generated token sequences until ith token and $\mathcal{T}$ is the instruction-based prompt. $\mathrm{Emb}(S)$ denotes the substructure token embeddings, and $f(\cdot)$ is the token alignment function.

4.4.2. Stage 2

To improve the instruction-following capabilities of the LLM within molecular contexts, we employ a parameter-efficient fine-tuning strategy based on Low-Rank Adaptation (LoRA), applied to the [q_proj, v_proj] modules. During this stage, we jointly optimize the substructure-token alignment module to better adapt to more complex and specific downstream tasks.

5. Generalist evaluation datasets

Current evaluations of LLMs on molecular tasks primarily focus on in-distribution performance within a single dataset. To more rigorously assess cross-dataset transfer and out-of-distribution (OOD) generalization, we construct generalist datasets that cover the following tasks:

- Molecular property prediction: Estimate the physicochemical or biological properties of a molecule from its structure.
- Forward reaction prediction: Predict the products of a given set of reactants and conditions.
- Retrosynthesis: Decompose a target molecule into plausible reactant candidates through reverse reaction reasoning.
- Molecular caption generation: Generate natural language descriptions of molecular structures or functions.

The specific data format and detailed statistics are presented in Fig. 4.

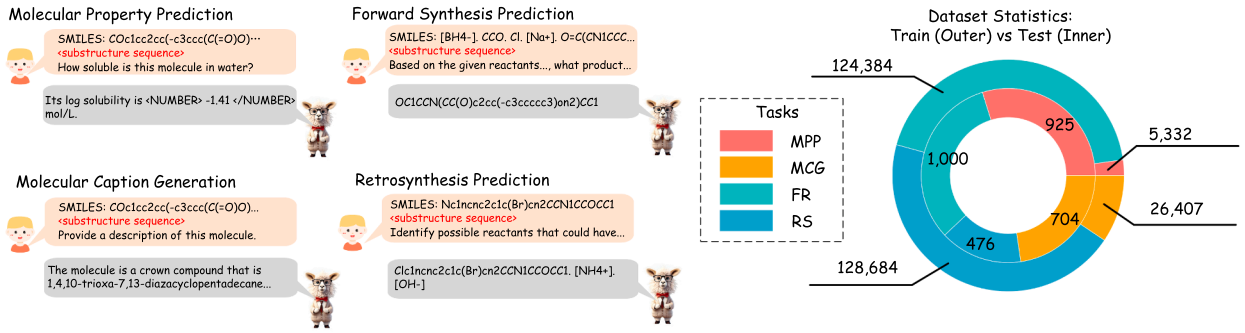


Fig. 4. An overview of molecular question-answer tasks and detailed statistics in the constructed generalist evaluation dataset.

5.1. Molecular property prediction (MPP): log solubility

For LogS prediction, we construct an OOD evaluation split using the ESOL (Wu et al., 2018) and AqSolDB (Sorkun et al., 2019) datasets. The ESOL dataset provides experimental water solubility values (log Solubility in mol/L) for small organic molecules, while AqSolDB is a curated collection of aqueous solubility measurements from multiple sources. All molecules are standardized using RDKit's canonicalization pipeline (Landrum, 2013), including cleanup, normalization, and reionization. To prevent data leakage, standardized SMILES from ESOL and AqSolDB are cross-checked. Molecules present in both are removed from AqSolDB. To ensure consistency during testing, ESOL entries with solubility differences greater than 0.1 log units are excluded. To define OOD samples, we apply a structural dissimilarity criterion. Morgan fingerprints (Zhou & Skolnick, 2024) are computed, and for each AqSolDB molecule, the maximum Tanimoto similarity to ESOL compounds is obtained. The bottom 40% (least similar) are selected as OOD samples, while the remaining 60% serve as in-distribution references.

The final ESOL dataset thus contains unique, standardized molecules without solubility conflicts, whereas the OOD AqSolDB subset comprises structurally novel compounds suitable for testing model generalization under distributional shifts. The resulting AqSolDB subset is used for instruction finetuning, while the ESOL dataset serves as the OOD test set.

5.2. Molecular caption generation (MCG)

For the MCG task, we use ChEBI-20 (Edwards et al., 2022) as the in-distribution training set and construct an extended set of molecule-description pairs from the ChEBI database (Hastings et al., 2016) to serve as the OOD test set. To ensure structural diversity, we adopt a scaffold-based split. A molecular scaffold refers to the core structural framework of a molecule, consisting of its ring systems and linkers with peripheral substituents removed. In a scaffold split, all molecules sharing the same scaffold are assigned to the same split, ensuring that the test set contains only molecules built on scaffolds unseen during training.

To construct this dataset, all SMILES strings from ChEBI-20 and the collected set are standardized to their canonical forms, and scaffold analysis is performed using RDKit to identify OOD molecules, labeling any molecule whose scaffold does not appear in ChEBI-20 as OOD. Molecule-description pairs are further curated to include fine-grained structural descriptions that emphasize functional groups and substituent positions. Text descriptions are tokenized using the LLaMA2 tokenizer, and those shorter than 30 tokens or longer than 512 tokens are discarded to ensure well-formed, meaningful content. The resulting OOD dataset consists of molecules that are structurally and semantically distinct from those in ChEBI-20, enabling a rigorous evaluation of a model's ability to generalize beyond familiar scaffolds.

5.3. Reaction prediction: forward reaction (FR) and retrosynthesis (RS)

For reaction prediction, we adopt a scaffold-based OOD split to ensure structural novelty. Training data are drawn from Mol-Instruction (Fang et al., 2024), which compiles synthetically relevant reactions from the USPTO database (Wei et al., 2010) of U.S. patents and patent applications. For OOD evaluation, we collect reactions from the Open Reaction Database (ORD) (Wigh et al., 2024), excluding those sourced from USPTO and any with scaffolds overlapping the training set.

Accepted reactions are reformatted into task-specific input-output pairs: reactants and reagents as input with the product as output for forward prediction, and the product as input with reactants and reagents as output for retrosynthesis. To enforce chemical validity, retrosynthesis data undergoes an atom-type consistency check, ensuring that no new element types appear in outputs. For forward prediction, reactions are subsampled to match the test set size of Mol-Instruction. The resulting OOD datasets comprise structurally novel yet chemically valid reactions, providing a rigorous evaluation of the robustness and generalization of reaction models under distributional shift.

6. Experiments

We conduct extensive experiments on standard molecular question-answering benchmarks as well as our curated generalist evaluation datasets. Specifically, we aim to address the following research questions:

- **RQ1**: How does the proposed model perform on standard molecular question-answering benchmarks compared to recent state-of-the-art methods?
- **RQ2**: Can S^2Token effectively generalize to out-of-distribution (OOD) molecular datasets?
- **RQ3**: Do the learned embeddings of substructure tokens for large language models exhibit interpretability?
- **RQ4**: What is the contribution of each key component of S^2Token to the model's overall performance?

6.1. Experimental settings

Recent molecular LLMs commonly use LLaMA2-based instruction-tuned models as the backbone. For a fair comparison, we adopt LLaMA2-7B-Chat as our LLM backbone. For substructure token embeddings, we employ a GNN architecture based on the Graph Isomorphism Network (GIN) (Xu, Hu, Leskovec, & Jegelka, 2026) consisting of 5 layers, which is initialized using a pretrained molecule-text alignment model from MoleculeSTM (Liu et al., 2023b). In Stage 1 training, we use a molecule-answer dataset from Fang et al. (2024), consisting of 298k molecule-text pairs collected from PubChem. In Stage 2, the model is instruction-tuned using a training split of forward reaction and retrosynthesis instruction datasets from Fang et al. (2024). For the molecular caption generation task, we evaluate on the ChEBI-20 dataset (Edwards et al., 2022).

Our implementation leverages the PyTorch framework for model development and training. In Stage 2 training, we utilize Low-Rank Adaptation (LoRA) to fine-tune the LLM. LoRA offers a parameter-efficient method to adapt large pre-trained models to domain-specific tasks with minimal overhead. For the optimizer, we use the AdamW optimizer with the following settings:

- In Stage 1, the initial learning rate is set to 1×10^{-5}, with a minimum learning rate of 1×10^{-6} and a warmup learning rate of 1×10^{-7}. We apply a cosine learning rate scheduler with 1000 warmup steps.
- In Stage 2, the learning rate is increased to 5×10^{-5}, with a minimum of 5×10^{-6} and a warmup learning rate of 5×10^{-7}.

Based on a trade-off between performance and computational efficiency, we set the number of epochs to 5 for Stage 1 and 10 for Stage 2. It is worth noting that while increasing the number of epochs leads to a gradual performance improvement, these gains exhibit diminishing returns. We save the trained model weights after each stage, which are later used for evaluation on general-purpose datasets. For the generalist evaluation, we fine-tune baseline models such as InstructMol and Graph-Token using the same hyperparameter settings. These models are adapted with a multi-level graph representation approach, where the alignment between the graph and language tokens is achieved via a two-layer linear projector in the alignment layers.

6.2. Benchmark datasets

- The ChEBI-20 dataset (Edwards et al., 2022) is a benchmark for evaluating molecular caption generation models. It contains 33,010 molecule-description pairs curated from the ChEBI database (Hastings et al., 2016), with each description spanning over 20 words and providing rich chemical details, such as descriptions of the molecule's acidic/basic variants. The dataset is split into 80% training, 10% validation, and 10% test sets.
- The forward reaction prediction task uses the Mol-Instruction dataset (Fang et al., 2024), which is derived from the USPTO dataset (Wei et al., 2010) containing organic reactions in SMILES format from U.S. patents and patent applications. For each entry, the input consists of the reactants and reagents involved in the reaction, separated by periods ('.'), while the output shows the reaction product. We follow the Mol-Instruction split with 124,384 samples for training and 1000 for testing.
- The benchmark dataset for the retrosynthesis prediction task is derived from Mol-Instruction, which focuses solely on single-step retrosynthesis. The data is sourced from the curated USPTO_500MT dataset (Lu & Zhang, 2022). Each entry consists of a product as the input, with the output being the reactants, where individual reactants are separated by a period ('.'). Consistent with the previous data split, the dataset includes 128,684 entries for training and 1000 entries for testing.

Here, we report the average number of substructure tokens per molecule for the benchmark datasets in Table 1. On average, S^2Token achieves about 2× reduction in sequence length compared to node-level tokenization. This compression arises from representing multi-node structural substructures—rings (≥ 3 nodes) and paths (≥ 2 nodes)—as single tokens. S^2Token improves sequence efficiency and enables better context utilization for LLMs.

6.3. Baselines

We compare our S^2Token with: (1) LLM-based generalist models, though not specialized for molecular tasks, leverage their strong instruction-following abilities to perform zero-shot or few-shot reasoning on textual molecular representations (e.g., SMILES):

- Alpaca (Dubois et al., 2023): A LLaMA-based model fine-tuned on a curated collection of instruction-following data, aimed at enhancing general-purpose NLP capabilities.

Table 1
Comparisons of average number of tokens.

Dataset	Avg. Num. Substructure Tokens	Avg. Num. Node Tokens
PubChem298K	~ 18.5	~ 33.7
ChEBI-20	~ 17.7	~ 32.3
Forward Reaction	~ 20.9	~ 41.3
Retrosynthesis	~ 12.5	~ 25.1

Table 2
Performance on both forward reaction prediction and retrosynthesis. ‡: few-shot ICL results from Fang et al. (2024).

Model	Exact ↑	BLEU ↑	Levenshtein ↓	RDK FTS ↑	MACCS FTS ↑	Morgan FTS ↑
Forward Reaction Prediction						
Alpaca‡	0.000	0.065	41.989	0.004	0.024	0.008
ChatGLM‡	0.000	0.183	40.008	0.050	0.100	0.044
Vicuna‡	0.000	0.057	41.690	0.007	0.016	0.006
Mol-Instruction	0.045	0.654	27.262	0.313	0.509	0.262
InstructMol-GS	0.536	**0.967**	10.851	0.776	0.878	0.741
HIGHT-GS	0.293	0.935	16.687	0.774	0.618	0.566
LLaMo	0.584	0.894	6.162	0.857	0.918	0.841
S^2Token (Ours)	**0.696**	0.938	**4.432**	**0.919**	**0.959**	**0.911**
Retrosynthesis						
Alpaca‡	0.000	0.063	46.915	0.005	0.023	0.007
ChatGLM‡	0.000	0.117	48.365	0.056	0.075	0.043
Vicuna‡	0.000	0.057	46.877	0.025	0.030	0.021
Mol-Instruction	0.009	0.705	31.227	0.283	0.487	0.230
InstructMol-GS	0.407	**0.941**	13.967	0.753	0.852	0.714
HIGHT-GS	0.202	0.914	20.194	0.772	0.623	0.577
LLaMo	0.341	0.830	12.263	0.793	0.868	0.750
S^2Token (Ours)	**0.496**	0.891	**9.505**	**0.866**	**0.915**	**0.838**

- ChatGLM (Zeng et al., 2023a): A bilingual LLM built on the GLM architecture, optimized for dialogue and general reasoning tasks.
- Vicuna (Chiang et al., 2023): A LLaMA derivative fine-tuned on user-shared multi-turn conversations to improve open-domain conversational performance.
- GPT-3.5-Turbo (Li et al., 2024a): A proprietary model from OpenAI, widely adopted for its robust instruction-following and reasoning abilities across a range of domains.

(2) Molecular Instruction Tuning LLMs, explicitly fine-tuned on molecular tasks using textual encodings like SMILES or SELFIES:

- Mol-Instruction (Fang et al., 2024): A domain-adapted LLM trained on a curated dataset of molecular instructions, designed for tasks such as property prediction, retrosynthesis, and molecular generation.
- BioMedGPT-10B (Luo et al., 2023): A large-scale biomedical model that integrates multimodal inputs (e.g., molecules, protein sequences, and natural language) to support a broad spectrum of biomedical reasoning tasks.
- LlaSMol (Yu et al., 2024): A LLaMA-based variant fine-tuned on domain-specific molecular instruction datasets to enhance performance on drug discovery and chemistry-related tasks.

(3) Graph modal alignment large language models, which go beyond text-only representations by incorporating molecular graphs and designing cross-modal alignment strategies with LLMs:

- InstructMol (G / GS) (Cao et al., 2025): InstructMol-G integrates a GNN encoder to represent molecular graphs, aligning node embeddings with LLMs for instruction following. InstructMol-GS further augments prompts with SELFIES-based representations to enhance textual grounding.
- HIGHT (G / GS) (Chen et al., 2025): HIGHT-G leverages a hierarchical GNN architecture, where functional motifs are treated as supernodes. The model concatenates the atomic and supernode embeddings into a token sequence for alignment with LLMs. HIGHT-GS introduces SELFIES-based prompt extensions for enhanced instruction comprehension.
- LLaMo (Park et al., 2024): LLaMo utilizes a multi-level graph projector with cross-modal attention layers to align molecular graph embeddings with LLM token embeddings. SMILES representations are also incorporated into prompts to provide additional semantic cues.

6.4. Experimental results

To answer RQ1, we conduct experiments on three molecular benchmarks. Table 2 summarizes the results of forward synthesis and retrosynthesis, and Table 3 presents the performance of molecular caption generation.

Table 3
Model performance comparison on molecular caption generation (%) task.

Model	BLEU-2 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	METEOR ↑
Text-based LLM Model						
GPT-3.5-turbo	10.3	5.0	26.1	8.8	20.4	16.1
BioMedGPT-10B	23.4	14.1	38.6	20.6	33.2	30.8
Mol-Instruction	24.9	17.1	33.1	20.3	28.9	27.1
Graph-based LLM Model						
InstructMol-G	46.6	36.5	54.7	36.5	47.9	49.1
InstructMol-GS	47.5	37.1	56.6	39.4	50.2	50.9
HIGHT-G	50.4	40.5	57.0	39.7	50.2	52.4
HIGHT-GS	49.8	39.7	58.2	41.4	51.8	52.5
S^2Token (Ours)	**56.0**	**47.6**	**62.8**	**46.8**	**56.7**	**59.3**

6.4.1. Forward reaction and retrosythesis prediction

As shown in Table 2, our method achieves the highest performance on five out of six metrics for both reaction tasks. S^2Token outperforms the strongest baselines by nearly 19% in Exact Match accuracy, suggesting that substructure tokens effectively capture the canonical structure and precise sequence of the reaction outputs. In terms of Levenshtein Distance, S^2Token achieves the lowest values (4.432 and 9.505), indicating the syntactically and chemically accurate output, likely due to its enhanced awareness of local chemical environments encoded in the learned tokens. For the Fingerprint Similarity Metrics (FTS), our model achieves consistent average improvements of 6.6% and 8.6% on the two tasks, highlighting that the substructure-centric representation fosters a deeper understanding of functional group interactions and reactivity patterns. Compared to InstructMol, the lower BLEU score may stem from differences in the LLM backbone and the molecular prompt (SMILES vs. SELFIES).

6.4.2. Molecular caption generation

Table 3 evaluates the capability of molecular LLMs to connect structures with language descriptions. S^2Token outperforms other tokenization strategies by encoding higher-order chemical patterns rather than individual atoms, providing a richer semantic and molecular context for LLM comprehension.

6.5. Generalization evaluation

To answer RQ2, we investigate the generalization capabilities of various molecular tokenization strategies using curated out-of-distribution (OOD) datasets in a strict zero-shot setting. Comparisons include node-token methods (e.g., InstructMol, LLaMo), graph-level tokenization approaches (Liang et al., 2024), and molecule-specialist LLMs (LlaSMol). All models assessed on the molecular captioning tasks are trained using the ChEBI-20 dataset. For the molecular property prediction, we fine-tune the generalist LLMs, LLaMA2-7B-Chat and Mistral-7B-Instruct, on the training set via LoRA-based fine-tuning, and subsequently evaluate them in an OOD setting. Fig. 5(a) and (b) show results for forward synthesis prediction and retrosynthesis, while Fig. 5(c) and (d) summarize performance on molecular captioning and property prediction.

While all models perform well in-distribution, they degrade on OOD tasks. This underscores the challenges associated with generalization in molecular LLMs. Notably, our proposed method, S^2Token, consistently outperforms all baseline models across all OOD tasks, demonstrating superior performance, especially in molecular captioning and property prediction tasks. We attribute this performance gain to the use of substructures as transferable token units. By treating frequently occurring substructures as tokens, S^2Token captures domain-relevant features that are more robust to distribution shifts. These substructures often encode key molecular motifs that determine molecular behavior, making them especially beneficial for tasks like caption generation. Similarly, for molecular solubility prediction, which is highly influenced by substructure properties, the use of substructure-level tokens enhances generalization.

6.6. Substructure analysis

For RQ3, we assess the ability of S^2Token to identify key substructure tokens in molecular graphs. Unlike recent tokenization studies of molecular graphs for LLMs, S^2Token derives substructures as graph tokens, enabling the model to reason about molecular properties through the semantics of substructure-level tokens. One critical task for evaluating such models is the ring count task, which measures performance on graph-based molecular properties. To this end, we benchmark S^2Token against other molecular tokenization strategies, such as SMILES as input and nodes as tokens. For this evaluation, we construct a specialized dataset based on the Lipo dataset (Wu et al., 2018), containing 4200 entries designed to address a ring-structure counting task. Each entry is associated with a question-answer instruction, allowing the model to adapt to this specific task and answer format. The task instructions are as follows: '*Identify all rings in the molecular structure and output in valid JSON format. "rings": "3-membered": <int>, "4-membered": <int>, "5-membered": <int>, "6-membered": <int>, "7-membered": <int>.*' We evaluate ring occurrences from 3- to 7-membered rings under two settings: **In-datasets**, where both training and testing use the Lipo dataset, and **Cross-datasets**, where models are trained on Lipo and tested on BBBP (Wu et al., 2018). While Lipo measures lipophilicity, BBBP focuses on blood-brain barrier penetration, providing distinct molecular distributions for assessing generalization.

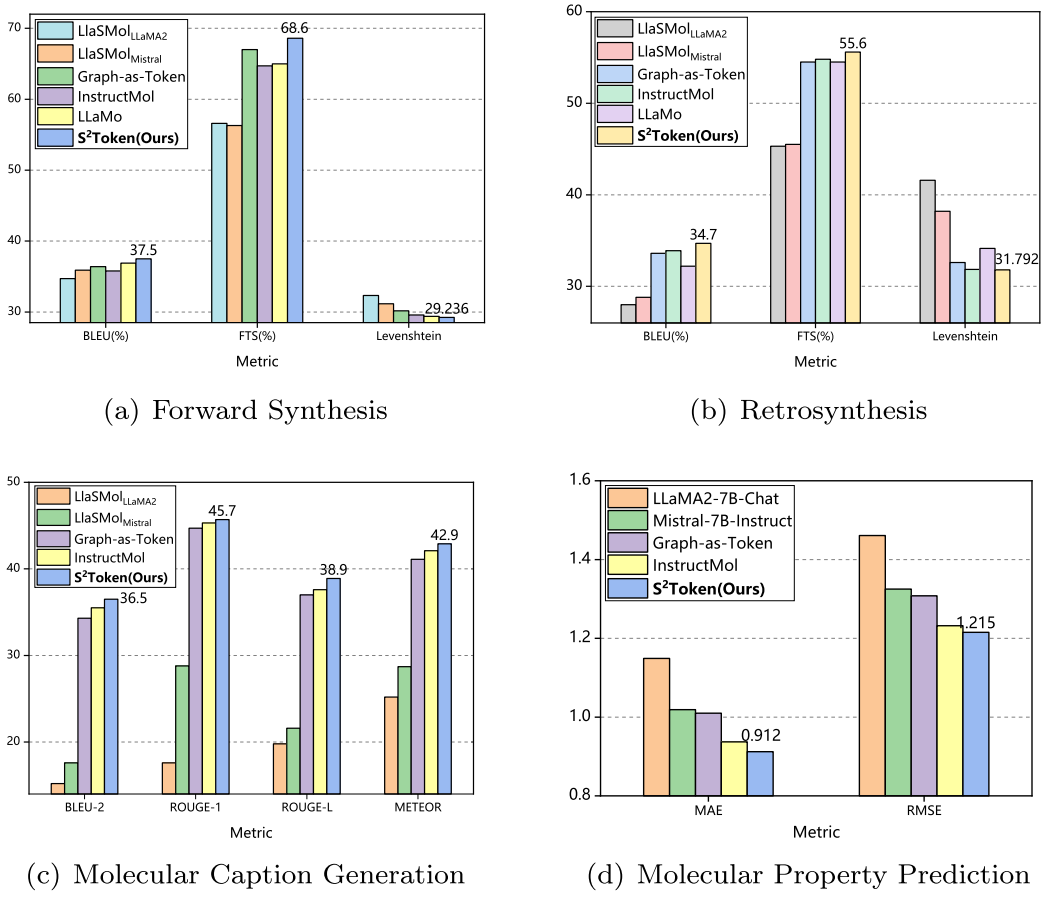


Fig. 5. Evaluation of OOD generalization between S^{2}Token and other tokenization methods for LLMs.

The results, presented in Table 4, show the accuracy of substructure identification for each ring type (3 to 7-membered rings). Our method, S^{2}Token, which tokenizes substructures directly, significantly outperforms both SMILES-based input (using LLaMA2-7B-Chat) and node-based tokenization (using InstructMol) in both in-dataset and cross-dataset evaluations. This demonstrates the superior interpretability of substructure tokenization in LLMs, where learned token embeddings encapsulate meaningful substructure semantics. Moreover, we observe that relying on textual representations, such as SMILES, makes it challenging for LLMs to effectively capture the graph-based structure of molecules, even though SMILES is derived from traversing molecular structures. In contrast, when graph features are incorporated, the performance of LLMs improves substantially, highlighting the benefits of a graph-centric tokenization approach. In summary, S^{2}Token offers a more interpretable and effective method for substructure identification, outperforming existing molecular tokenization techniques. This provides strong evidence that incorporating substructures as tokens enhances the ability of LLMs to reason about molecular properties, thereby achieving superior task performance.

6.7. Backbone comparison

Since S^{2}Token is backbone-agnostic, in this section, we further investigate the impact of adopting a domain-specific backbone on molecular task performance, motivated by the hypothesis that our framework can benefit from stronger backbone capabilities. We conduct additional experiments by replacing LLaMA2 with BioMistral (Labrak et al., 2024), which is built upon Mistral-7B Instruct (Jiang et al., 2023a). The experimental protocol, including data splits, model pipeline, and tuning procedure, remains the same. And the results, reported in Tables 5 and 6, show that pairing S^{2}Token with the stronger backbone yields consistent improvements, validating our hypothesis that S^{2}Token complements backbone pretraining and effectively leverages enhanced model capacity. We attribute the advantage of BioMistral over LLaMA2 primarily to domain-adaptive pretraining on biomedical literature, including the PubMed Central Open Access Subset, which exposes the model to reaction-oriented language patterns that are rare in general-purpose corpora. When combined with instruction-style fine-tuning on biomedical tasks, this domain adaptation further improves the model's ability to align task prompts with the desired output distribution. At the same time, we observe only marginal gains on molecular caption tasks with ChEBI-style descriptions after switching backbones. This is likely because the ChEBI-20 captions follow highly templated naming conventions and repetitive functional-group phrasing. Instruction tuning on ChEBI-20 already enables the backbone to capture most caption-specific linguistic patterns, leaving limited room for additional gains from backbone specialization.

Table 4
Comparison of accuracy (%) in substructure counts for common rings in the In-dataset and Cross-dataset, between S^{2}Token and other molecular tokenization methods for LLMs.

	Model	3-member	4-member	5-member	6-member	7-member
In-datasets	LLaMA2-7B-Chat	92.8	98.3	58.1	53.3	96.7
	InstructMol	98.8	98.5	95.5	90.2	98.0
	S^{2}Token(Ours)	**99.6**	**99.8**	**97.2**	**94.6**	**98.6**
Cross-datasets	LLaMA2-7B-Chat	96.0	89.3	52.0	48.0	85.3
	InstructMol	97.3	90.0	84.4	73.0	91.4
	S^{2}Token(Ours)	**97.9**	**96.8**	**89.2**	**79.1**	**95.2**

Table 5
Performance comparison of different backbones on both forward reaction prediction and retrosynthesis.

Model	Exact ↑	BLEU ↑	Levenshtein ↓	RDK FTS ↑	MACCS FTS ↑	Morgan FTS ↑
Forward Reaction Prediction						
S^{2}Token$_{\text{LLaMA2}}$	0.696	0.938	4.432	0.919	0.959	0.911
S^{2}Token$_{\text{BioMistral}}$	**0.757**	**0.956**	**2.450**	**0.937**	**0.967**	**0.926**
Retrosynthesis						
S^{2}Token$_{\text{LLaMA2}}$	0.496	0.891	9.505	0.866	0.915	0.838
S^{2}Token$_{\text{BioMistral}}$	**0.548**	**0.895**	**7.798**	**0.867**	**0.919**	**0.843**

Table 6
Performance comparison of different backbones on molecular caption generation(%) task.

Model	BLEU-2 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	METEOR ↑
S^{2}Token$_{\text{LLaMA2}}$	56.0	47.6	62.8	46.8	56.7	**59.3**
S^{2}Token$_{\text{BioMistral}}$	**56.3**	**47.9**	**63.3**	**47.3**	**57.3**	**59.3**

Table 7
Ablation studies (%) on caption generation task.

Model	BLEU-2 ↑	ROUGE-L ↑	METEOR ↑
w/o Chemical Identifier	36.2	70.5	62.9
w/o Inter-substructure Alignment	32.8	68.7	60.7
w/o Multi-level Representations	31.5	67.7	59.5
w/o Structural Constraints$_{\text{Absolute}}$	35.9	70.4	63.1
w/o Structural Constraints$_{\text{Relative}}$	35.6	70.7	62.7
w/o Structural Constraints	35.5	70.1	62.2
S^{2}Token (Ours)	**36.8**	**71.2**	**63.6**

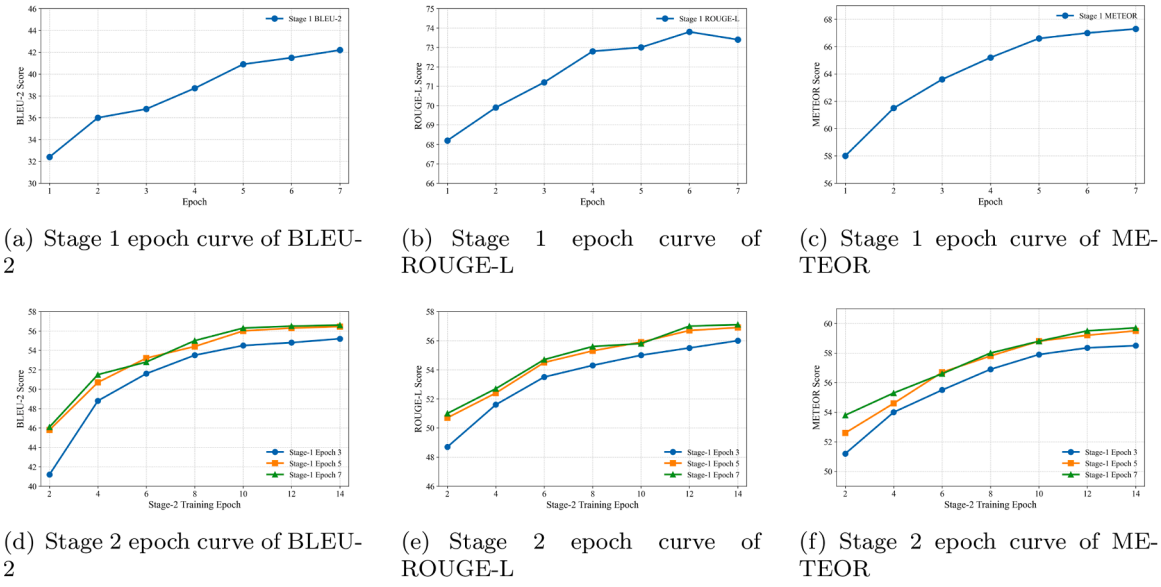
6.8. Ablation study

This section addresses RQ4 by analyzing the role of each key component in driving the performance gains of S^{2}Token. We conduct an ablation study, i.e., chemical identifier encoding (-**Chemical Identifier**), substructure token alignment with inter-substructure dependency (-**Inter-substructure Alignment**), and multi-level graph representations (-**Multi-level Representations**). Here, we further quantify the contribution of structural constraints in alignment (-**Structural Constraints**), and conduct a more fine-grained ablation to assess the individual effects of absolute (-**Structural Constraints$_{\text{Absolute}}$**) and relative structural encoding (-**Structural Constraints$_{\text{Relative}}$**). Each variant is evaluated on the caption generation task for three training epochs on the PubChem298K dataset (Fang et al., 2024) using the stage 1 protocol.

As shown in Table 7, each component plays a critical role in aligning molecular substructures with LLMs. Removing the multi-level graph representation (-Multi-level Representations) or inter-substructure alignment (-Inter-substructure Alignment) leads to substantial performance drops, underscoring the necessity of semantically rich token embeddings and effective alignment strategies. S^{2}Token integrates shallow, intermediate, and deep graph representations to capture multi-granularity molecular features, enhancing

Table 8
Comparison (%) of different fragmentation rules.

Method	BLEU-2 ↑	ROUGE-L ↑	METEOR ↑
InstructMol	29.2	65.0	55.3
BRICS-based Fragmentation	35.2	69.2	60.1
S^2Token (Ours)	**36.8**	**71.2**	**63.6**



(a) Stage 1 epoch curve of BLEU-2

(b) Stage 1 epoch curve of ROUGE-L

(c) Stage 1 epoch curve of METEOR

(d) Stage 2 epoch curve of BLEU-2

(e) Stage 2 epoch curve of ROUGE-L

(f) Stage 2 epoch curve of METEOR

Fig. 6. Performance with various numbers for training epochs.

token-level semantic expressiveness. Inter-substructure alignment further enables the model to capture interactions between substructures, which is essential for LLMs to express molecular behavior.

6.9. Evaluation of different fragmentation strategies

To evaluate the sensitivity of S^2Token to different fragmentation strategies, we employ the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS) (Degen et al., 2008) algorithm, which decomposes molecules into chemically meaningful building blocks based on retrosynthetic rules. Due to its ability to generate interpretable fragments, BRICS has been widely adopted in graph-based molecular representation learning (Chen et al., 2025; Zhang et al., 2021). In our default setup, ring systems are treated as indivisible units to preserve key substructures such as aromatic systems. The evaluation is performed on the PubChem298K dataset using the stage 1 protocol with three training epochs. As shown in Table 8, S^2Token consistently outperforms the node-based tokenization method and remains robust across reasonable fragmentation choices.

The results show that the BRICS-based fragmentation rule yields slightly lower performance. A possible reason is that the BRICS algorithm produces a much larger vocabulary space, leading to a higher proportion of rare fragments with low occurrence frequency. These infrequent fragments receive limited training, resulting in weaker representation learning and reduced overall performance compared to the more frequent ring-path substructure vocabulary.

6.10. Training epochs analysis

In this section, we conduct an additional analysis of varying the numbers of training epochs in both Stage 1 alignment training and Stage 2 instruction tuning, while keeping all other experimental settings identical to those used in the main experiments. Specifically, Stage 1 training is performed with epoch numbers [3, 5, 7], seen in Fig. 6(a)–(c). We further explore combinations of Stage 1 epochs [3, 5, 7] and Stage 2 epochs [2, 4, 6, 8, 10, 12, 14] on the ChEBI-20 dataset, summarized in Fig. 6(d)-(f). Model performance is evaluated using BLEU-2, ROUGE-L, and METEOR metrics. The results demonstrate the performance improvement as the number of training epochs increases. However, the magnitude of improvement gradually decreases, and the performance curves begin to plateau, indicating that the model approaches convergence under the current training regime. Although slight metric gains can still be observed with additional training, the marginal benefits become increasingly limited.

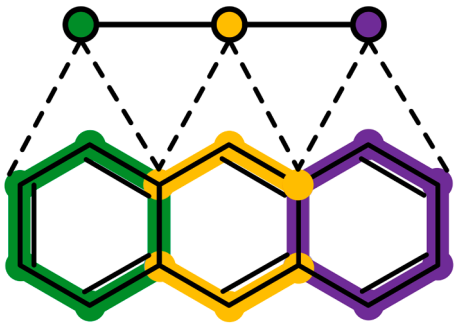


Fig. 7. Illustration of fragmentation of anthracene.

7. Discussion

In this work, we introduce S^{2}Token, a substructure-aware tokenization framework designed to bridge molecular graphs with LLMs. Our experimental results demonstrate that S^{2}Token consistently outperforms existing tokenization strategies across a range of molecular tasks, including forward reaction prediction, retrosynthesis, and molecular caption generation. The improvements are also pronounced in OOD settings, underscoring the superior generalization capabilities of our approach. The success of S^{2}Token carries significant implications for both the field of graph learning and molecular science.

7.1. Interpretation of results

S^{2}Token excels due to its ability to tokenize molecular graphs into chemically meaningful substructures, which allows LLMs to align with fundamental chemistry principles. These substructures, like a benzene ring, capture molecular properties and reactivity more effectively than purely syntactic or atom-based approaches. This granularity enables LLMs to better understand and interpret molecular data, as evidenced by strong performance on fingerprint similarity and exact match accuracy in reaction tasks. S^{2}Token's focus on general, reusable substructures also contributes to its strong generalization ability, particularly in OOD settings, where it handles unseen molecular data without overfitting to rare configurations.

7.2. Case study for fused ring system

For fused-ring systems such as anthracene, a sequence of atoms with alternating single and double bonds (e.g., C1 = CC = C2C…) enables continuous overlap of p-orbitals and delocalization of electrons. These structures challenge molecular representation in preserving conjugation information. In a case study on anthracene, our ring-driven fragmentation mechanism addresses this by decomposing the fused systems into three conjugated units (each represented by the SMILES 'c1ccccc1' like a benzene), as shown in Fig. 7. The shared junction atoms serve as bridges that preserve inter-ring electronic coupling. Each indivisible conjugated unit is embedded with atom-level and bond-level features such as hybridization, aromaticity, and conjugation. Through hierarchical inter-substructure learning, the model aligns substructure tokens to capture spatial and linkage characteristics between adjacent rings. The hierarchical treatment allows LLMs to handle the corresponding structural properties for these specific conjugate systems.

7.3. Research implications for LLMs on graph learning

Our work introduces a method that enables LLMs to process non-sequential graph data by translating its relational topology into a tokenized format. S^{2}Token advances traditional graph representations by tokenizing substructures (subgraphs) as semantic units, rather than treating individual nodes as isolated 'words' or the entire graph as a single 'sentence.' This approach parallels subword tokenization in natural language processing, allowing finer-grained yet context-aware encoding of structural information. Beyond molecular graphs, S^{2}Token provides a blueprint for tokenizing other complex graphs (e.g., social networks and biological pathways) into structural components that share common characteristics, opening a possibility to enhance an LLM's ability to generalize to unseen graph data. Moreover, by incorporating explicit structural encodings as a graph inductive bias directly into the LLM's input, S^{2}Token strengthens the model's ability to reason over graph connectivity and structural topology.

7.4. Research implications for LLMs on molecular science

For molecular research, S^{2}Token serves as a powerful translator, converting the language of chemistry into a format that LLMs can effectively process. It unlocks the potential of LLMs for tasks such as molecular synthesis and optimization by following researchers' instructions (e.g., '*Based on the given reactants and reagents, suggest a possible product.*'). Taking reaction prediction as an example, S^{2}Token can assist researchers in designing efficient synthesis pathways and discovering novel compounds, thereby acting as an intelligent assistant in molecular discovery. By reasoning over substructure interactions, the model can generate informed predictions.

Since S^{2}Token operates with chemically meaningful units, its predictions become more interpretable, allowing chemists to trace model outputs back to specific structural features such as rings or functional groups.

8. Conclusion

In this paper, we propose S^{2}Token, a substructure-aware molecular graph tokenization framework that aligns fragmented substructures as discrete tokens for LLMs. S^{2}Token exploits the chemical significance and recurrence of substructures across diverse molecules, providing a more semantically consistent and transferable token space. By integrating structural and functional information and introducing inter-substructure dependency alignment, S^{2}Token preserves chemical and topological properties. Evaluations on both in-distribution and OOD tasks show that our method significantly improves molecular representation and generalization for LLMs.

CRediT authorship contribution statement

Runze Wang: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization; **Zijie Xing:** Validation, Investigation, Data curation; **Xingyue Liu:** Validation, Investigation, Data curation; **Mingqi Yang:** Methodology, Formal analysis, Conceptualization; **Che He:** Visualization, Validation, Methodology, Formal analysis; **Yanming Shen:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Formal analysis, Conceptualization.

Data availability

I have made the code and data publicly available in the manuscript.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62276044.

References

Atz, K., Grisoni, F., & Schneider, G. (2021). Geometric deep learning on molecular representations. *Nature Machine Intelligence*, *3*(12), 1023–1032. https://doi.org/10.1038/s42256-021-00418-8.

Barrault, L., Duquenne, P.-A., Elbayad, M., Kozhevnikov, A., Alastruey, B., Andrews, P., Coria, M., Couairon, G., Costa-jussà, M. R., Dale, D. et al. (2024). Large concept models: Language modeling in a sentence representation space. https://doi.org/10.48550/arXiv.2412.08821.

Cao, H., Liu, Z., Lu, X., Yao, Y., & Li, Y. (2025). Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st international conference on computational linguistics* (pp. 354–379). https://aclanthology.org/2025.coling-main.25/.

Chen, J., & Schwaller, P. (2024). Molecular hypergraph neural networks. *The Journal of Chemical Physics*, *160*(14). https://doi.org/10.1063/5.0193557.

Chen, R., Zhao, T., Jaiswal, A. K., Shah, N., & Wang, Z. (2024). Llaga: Large language and graph assistant. In *International conference on machine learning* (pp. 7809–7823). PMLR. https://proceedings.mlr.press/v235/chen24bh.html.

Chen, Y., Yao, Q., Zhang, J., Cheng, J., & Bian, Y. (2025). Hierarchical graph tokenization for molecule-language alignment. In *Forty-second international conference on machine learning*. https://openreview.net/forum?id=wpbNczwAwV.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E. et al. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. See (accessed 14 April 2023), *2*, 6. https://vicuna.lmsys.org, https://lmsys.org/blog/2023-03-30-vicuna/.

Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. *Nature Reviews Methods Primers*, *4*(1), 17. https://doi.org/10.1038/s43586-024-00294-7.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., & Rarey, M. (2008). On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, *3*(10), 1503. https://pubmed.ncbi.nlm.nih.gov/18792903/.

Du, W., Zhang, S., Cai, Z., Li, X., Liu, Z., Fang, J., Wang, J., Wang, X., & Wang, Y. (2025). Molecular merged hypergraph neural network for explainable solvation gibbs free energy prediction. *Research*, *8*, 0740. https://doi.org/10.34133/research.0740.

Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P. S., & Hashimoto, T. B. (2023). Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, *36*, 30039–30069. https://openreview.net/forum?id=-Aw0rrrPUF.

Dwivedi, V. P., Luu, A. T., Laurent, T., Bengio, Y., & Bresson, X. (2022). Graph neural networks with learnable structural and positional representations. In *The tenth international conference on learning representations*. https://openreview.net/forum?id=wTTjnvGphYj.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., & Ji, H. (2022). Translation between molecules and natural language. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 375–413). https://doi.org/10.18653/v1/2022.emnlp-main.26.

Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., & Chen, H. (2024). Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The twelfth international conference on learning representations*. https://doi.org/10.48550/arXiv.2306.08018.

Fatemi, B., Halcrow, J., & Perozzi, B. (2024). Talk like a graph: Encoding graphs for large language models. In *The twelfth international conference on learning representations*. https://openreview.net/forum?id=IuXR1CCrSi.

Gastaldi, J. L., Terilla, J., Malagutti, L., DuSell, B., Vieira, T., & Cotterell, R. (2025). The foundations of tokenization: Statistical and computational concerns. In *The thirteenth international conference on learning representations*. https://doi.org/10.48550/arXiv.2407.11606.

Guo, T., Nan, B., Liang, Z., Guo, Z., Chawla, N., Wiest, O., Zhang, X. et al. (2023). What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, *36*, 59662–59688. https://doi.org/10.48550/arXiv.2305.18365.

Han, S., Fu, H., Wu, Y., Zhao, G., Song, Z., Huang, F., Zhang, Z., Liu, S., & Zhang, W. (2023). HimGNN: A novel hierarchical molecular graph representation learning framework for property prediction. *Briefings in Bioinformatics*, *24*(5), bbad305. https://doi.org/10.1093/bib/bbad305.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). Inchi-the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, *5*(1), 7. https://doi.org/10.1186/1758-2946-5-7.

Hong, Z., Li, J., Sun, L., & Liu, G. (2026). HT-geogt: A hierarchical twin-stream geometric graph transformer with graph representation learning architecture. *Information Processing & Management*, *63*(2, Part A), 104412. https://doi.org/10.1016/j.ipm.2025.104412.

Hu, C., Li, H., Yuan, Y., Li, J., & Tsang, I. (2024). Exploring hierarchical molecular graph representation in multimodal LLMs. https://doi.org/10.48550/arxiv.2411.04708.
Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023a). Mistral 7b. https://arxiv.org/abs/2310.06825.
Jiang, Y., Jin, S., Jin, X., Xiao, X., Wu, W., Liu, X., Zhang, Q., Zeng, X., Yang, G., & Niu, Z. (2023b). Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Communications Chemistry*, *6*(1), 60. https://doi.org/10.1038/s42004-023-00857-x.
Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, *1*(4), 045024. https://iopscience.iop.org/article/10.1088/2632-2153/aba947.
Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: Acl 2024* (pp. 5848–5864). https://aclanthology.org/2024.findings-acl.348/.
Landrum, G. (2013). Rdkit documentation. *Release*, *1*(1-79), 4. https://www.rdkit.org/docs/.
Le, K., Guo, Z., Dong, K., Huang, X., Nan, B., Iyer, R., Zhang, X., Wiest, O., Wang, W., & Chawla, N. V. (2024). Molx: Enhancing large language models for molecular learning with a multi-modal extension. https://doi.org/10.48550/arXiv.2406.06777.
Lee, C., Ko, H., Song, Y., Jeong, Y., Hormazabal, R., Han, S., Bae, K., Lim, S., & Kim, S. (2025). Mol-LLM: Multimodal generalist molecular LLM with improved graph utilization. https://doi.org/10.48550/arXiv.2502.02810.
Li, J., Liu, Y., Fan, W., Wei, X.-Y., Liu, H., Tang, J., & Li, Q. (2024a). Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 6071–6083. https://doi.org/{10.1109/TKDE.2024.3393356}.
Li, L., Zhang, Y., Wang, G., & Xia, K. (2025a). Kolmogorov–Arnold graph neural networks for molecular property prediction. *Nature Machine Intelligence*, (pp. 1–9). https://doi.org/10.1038/s42256-025-01087-7.
Li, S., Liu, Z., Luo, Y., Wang, X., He, X., Kawaguchi, K., Chua, T.-S., & Tian, Q. (2024b). Towards 3D molecule-text interpretation in language models. In *The twelfth international conference on learning representations*. https://openreview.net/forum?id=xI4yNlkaqh.
Li, Y., Fang, Y., Zhang, M., & Shi, C. (2025b). Advancing molecular graph-text pre-training via fine-grained alignment. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 2* (pp. 1589–1599). https://doi.org/{10.1145/3711896.3736834}.
Li, Z., Jiang, M., Wang, S., & Zhang, S. (2022). Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, *27*(12), 103373. https://doi.org/10.1016/j.drudis.2022.103373.
Liang, Y., Zhang, R., Li, Y., Huo, M., Ma, Z., Singh, D., Gao, C., Rahmani, H., Bandi, S., Zhang, L. et al. (2024). Multi-modal large language model enables all-purpose prediction of drug mechanisms and properties. *bioRxiv*, 2024–09. https://doi.org/10.1101/2024.09.29.615524.
Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023a). Visual instruction tuning. *Advances in Neural Information Processing Systems*, *36*, 34892–34916. https://doi.org/10.48550/arXiv.2304.08485.
Liu, P., Ren, Y., Tao, J., & Ren, Z. (2024). Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, *171*, 108073. https://doi.org/10.1016/j.compbiomed.2024.108073.
Liu, P., Tao, J., & Ren, Z. (2025). A quantitative analysis of knowledge-learning preferences in large language models in molecular science. *Nature Machine Intelligence*, *7*(2), 315–327. https://doi.org/10.1038/s42256-024-00977-6.
Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., & Anandkumar, A. (2023b). Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, *5*(12), 1447–1457. https://doi.org/10.1038/s42256-023-00759-6.
Liu, Z., Li, S., Luo, Y., Fei, H., Cao, Y., Kawaguchi, K., Wang, X., & Chua, T.-S. (2023c). MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15623–15638). https://doi.org/10.18653/v1/2023.emnlp-main.966.
Lu, J., & Zhang, Y. (2022). Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, *62*(6), 1376–1387. https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c01467.
Luo, Y., Zhang, J., Fan, S., Yang, K., Wu, Y., Qiao, M., & Nie, Z. (2023). BiomedGPT: Open multimodal generative pre-trained transformer for biomedicine. https://doi.org/10.48550/arxiv.2308.09442.
Ma, Y., Yu, S., & Shen, Y. (2024). Pretraining molecules with explicit substructure information. In *Proceedings of the 2024 SIAM international conference on data mining (SDM)* (pp. 517–525). SIAM. https://doi.org/10.1137/1.9781611978032.60.
Park, J., Bae, M., Ko, D., & Kim, H. J. (2024). Llamo: Large language model-based molecular graph assistant. *Advances in Neural Information Processing Systems*, *37*, 131972–132000. https://proceedings.neurips.cc/paper_files/paper/2024/file/ee46288ab2aaf5c6e53aebebe719712c-Paper-Conference.pdf.
Pei, Q., Wu, L., Gao, K., Liang, X., Fang, Y., Zhu, J., Xie, S., Qin, T., & Yan, R. (2024). Biot5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the association for computational linguistics ACL 2024* (pp. 1216–1240). https://doi.org/10.18653/v1/2024.findings-acl.71.
Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, *33*, 12559–12571. https://proceedings.neurips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html.
Sadeghi, S., Bui, A., Forooghi, A., Lu, J., & Ngom, A. (2024). Can large language models understand molecules? *BMC Bioinformatics*, *25*(1), 225. https://link.springer.com/article/10.1186/s12859-024-05847-x.
Seidl, P., Vall, A., Hochreiter, S., & Klambauer, G. (2023). Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International conference on machine learning* (pp. 30458–30490). PMLR. https://proceedings.mlr.press/v202/seidl23a.html.
Sorkun, M. C., Khetan, A., & Er, S. (2019). AqsolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, *6*(1), 143. https://doi.org/10.1038/s41597-019-0151-1.
Sypetkowski, M., Wenkel, F., Poursafaei, F., Dickson, N., Suri, K., Fradkin, P., & Beaini, D. (2024). On the scalability of GNNs for molecular graphs. *Advances in Neural Information Processing Systems*, *37*, 19870–19906. https://proceedings.neurips.cc/paper_files/paper/2024/hash/2345275663a15ee92a06bc957be54a2c-Abstract-Conference.html.
Tang, J., Yang, Y., Wei, W., Shi, L., Su, L., Cheng, S., Yin, D., & Huang, C. (2024). GraphGPT: Graph instruction tuning for large language models. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 491–500). https://doi.org/10.1145/3626772.3657775.
Wang, D., Zuo, Y., Li, F., & Wu, J. (2024). LLMs as zero-shot graph learners: Alignment of GNN representations with LLM token embeddings. *Advances in Neural Information Processing Systems*, *37*, 5950–5973. https://proceedings.neurips.cc/paper_files/paper/2024/file/0b77d3a82b59e9d9899370b378087faf-Paper-Conference.pdf.
Wang, H., Feng, S., He, T., Tan, Z., Han, X., & Tsvetkov, Y. (2023). Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, *36*, 30840–30861. https://proceedings.neurips.cc/paper_files/paper/2023/file/622afc4edf2824a1b6aaf5afe153fa93-Paper-Conference.pdf.
Wang, R., Ma, Y., Liu, X., Xing, Z., & Shen, Y. (2025a). Motif-driven molecular graph representation learning. *Expert Systems with Applications*, *269*, 126484. https://doi.org/10.1016/j.eswa.2025.126484.
Wang, R., Yang, M., & Shen, Y. (2025b). Bridging molecular graphs and large language models. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 21234–21242). (*vol. 39*). https://doi.org/10.1609/aaai.v39i20.35422.
Wang, S., Huang, J., Chen, Z., Song, Y., Tang, W., Mao, H., Fan, W., Liu, H., Liu, X., Yin, D. et al. (2025c). Graph machine learning in the era of large language models (LLMs). *ACM Transactions on Intelligent Systems and Technology*, *16*(5), 1–40. https://doi.org/{10.1145/3732786}.
Wei, J.-M., Yuan, X.-J., Hu, Q.-H., & Wang, S.-Q. (2010). A novel measure for evaluating classifiers. *Expert Systems with Applications*, *37*(5), 3799–3809. https://doi.org/10.1016/j.eswa.2009.11.040.
Weininger, D. (1988). A chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, (pp. 1549–9596). https://pubs.acs.org/doi/pdf/10.1021/ci00057a005.
Wigh, D. S., Arrowsmith, J., Pomberger, A., Felton, K. C., & Lapkin, A. A. (2024). Orderly: Data sets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling*, *64*(9), 3790–3798. https://doi.org/10.1021/acs.jcim.4c00292.

Wollschläger, T., Kemper, N., Hetzel, L., Sommer, J., & Günnemann, S. (2024). Expressivity and generalization: Fragment-biases for molecular GNNs. In *International conference on machine learning* (pp. 53113–53139). PMLR. https://proceedings.mlr.press/v235/wollschlager24a.html.

Wu, F., Li, S., Jin, X., Jiang, Y., Radev, D., Niu, Z., & Li, S. Z. (2023). Rethinking explaining graph neural networks via non-parametric subgraph matching. In *International conference on machine learning* (pp. 37511–37523). PMLR. https://proceedings.mlr.press/v202/wu23j.html.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). Moleculenet: A benchmark for molecular machine learning. *Chemical Science*, *9*(2), 513–530. https://doi.org/10.1039/C7SC02664A.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2026). How powerful are graph neural networks? In *The seventh international conference on learning representations*. https://openreview.net/forum?id=ryGs6iA5Km&noteId=rkl2Q1Qi6X&noteId=rkl2Q1Qi6X.

Xuyang, Z., Ye, W., Fei, T., Hong, Y., & Qun, L. (2025). Hierarchical chat-based strategies with MLLMs for spatio-temporal action detection. *Information Processing & Management*, *62*(4), 104094. https://doi.org/10.1016/j.ipm.2025.104094.

Yangding, L., Yangyang, Z., Xiangchao, Z., Jiawei, C., Hao, F., Shaobin, F., Cui, Y., & Shichao, Z. (2025). GNN-transformer contrastive learning explores homophily. *Information Processing & Management*, *62*(4), 104103. https://doi.org/10.1016/j.ipm.2025.104103.

Yu, B., Baker, F. N., Chen, Z., Ning, X., & Sun, H. (2024). LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First conference on language modeling*. https://openreview.net/forum?id=lY6XTF9tPv.

Yuan, C., Zhao, K., Kuruoglu, E. E., Wang, L., Xu, T., Huang, W., Zhao, D., Cheng, H., & Rong, Y. (2025). A survey of graph transformers: Architectures, theories and applications. https://arxiv.org/abs/2502.16533.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X. et al. (2023a). GLM-130B: An open bilingual pre-trained model. In *The eleventh international conference on learning representations*. https://openreview.net/forum?id=-Aw0rrrPUF.

Zeng, D., Liu, W., Chen, W., Zhou, L., Zhang, M., & Qu, H. (2023b). Substructure aware graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11129–11137). (*vol. 37*). https://doi.org/10.1609/aaai.v37i9.26318.

Zhang, B., Fan, C., Liu, S., Huang, K., Zhao, X., Huang, J., & Liu, Z. (2024a). The expressive power of graph neural networks: A survey. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2024.3523700.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024b). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625–5644. https://doi.org/10.1109/TPAMI.2024.3369699.

Zhang, Q., Ding, K., Lv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z. et al. (2025). Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, *57*(6), 1–38. https://dl.acm.org/doi/abs/10.1145/3715318.

Zhang, Y., Wang, H., Feng, S., Tan, Z., Han, X., He, T., & Tsvetkov, Y. (2024c). Can LLM graph reasoning generalize beyond pattern memorization? In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 2289–2305). https://aclanthology.org/2024.findings-emnlp.127/.

Zhang, Z., Liu, Q., Wang, H., Lu, C., & Lee, C.-K. (2021). Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, *34*, 15870–15882. https://proceedings.neurips.cc/paper_files/paper/2021/file/85267d349a5e647ff0a9edcb5ffd1e02-Paper.pdf.

Zhang, Z., Min, Y., Zheng, S., & Liu, Q. (2023). Molecule generation for target protein binding with structural motifs. In *The eleventh international conference on learning representations*. https://openreview.net/forum?id=Rq13idF0F73.

Zhao, H., Liu, S., Chang, M., Xu, H., Fu, J., Deng, Z., Kong, L., & Liu, Q. (2023). Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, *36*, 5850–5887. https://proceedings.neurips.cc/paper_files/paper/2023/file/129033c7c08be683059559e8d6bfd460-Paper-Conference.pdf.

Zhao, Y., Wang, W., Wang, S., Dong, J., & Duan, H. (2026). Over-smoothing problem of heterogeneous graph neural networks: A heterogeneous graph neural network with enhanced node differentiability. *Information Processing & Management*, *63*(2, Part A), 104395. https://doi.org/10.1016/j.ipm.2025.104395.

Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T. N., May, L. T., Webb, G. I., & Pan, S. (2025). Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, (pp. 1–11). https://www.nature.com/articles/s42256-025-00994-z.

Zhong, Z., Zhou, K., & Mottin, D. (2024). Benchmarking large language models for molecule prediction tasks. https://doi.org/10.48550/arxiv.2403.05075.

Zhou, H., & Skolnick, J. (2024). Utility of the Morgan fingerprint in structure-based virtual ligand screening. *The Journal of Physical Chemistry B*, *128*(22), 5363–5370. https://doi.org/10.1021/acs.jpcb.4c01875.

Zijing, T., Zhonghong, O., Yifan, Z., Shuai, L., Hanyu, Z., Jinghua, X., & Meina, S. (2025). Multi-SEA: Multi-stage semantic enhancement and aggregation for image-text retrieval. *Information Processing & Management*, *62*(5), 104165. https://doi.org/10.1016/j.ipm.2025.104165.