MGSEKFORMER: MULTI-GRANULARITY GRAPH-SEQUENCE ENHANCED KNOWLEDGE TRANSFORMER FOR SINGLE-STEP RETROSYNTHESIS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-step retrosynthesis prediction serves as a fundamental building block for multi-step synthesis route planning in modern drug discovery and organic material synthesis. To improve both efficiency and accuracy, researchers have explored a variety of deep learning frameworks, most of which focus on molecular feature extraction at a single granularity—either the atomic or functional group level. However, these approaches have not sufficiently addressed the differences and interactions between these two distinct granularities in terms of their impact on chemical reactions, which limits further improvements in predictive performance. To this end, we propose the Multi-Granularity Graph-Sequence Enhanced Knowledge Transformer framework (MGSEKformer) for single-step retrosynthesis. Specifically, we employ two PAGformers to extract local structural priors and long-range dependencies at the atomic and functional group levels, respectively, and use cross-attention and gating units to model their interactions and aggregate complementary information adaptively. Furthermore, building upon the standard feedforward transformation, we design a knowledge-attention branch to replace the fixed-weight fusion based on simple concatenation in a conventional Kformer, enabling adaptive injection of variable-length chemical prior knowledge and overcoming the reliance on predefined reaction templates. Comparative experiments against 13 baseline models show that the MGSEKformer achieves optimal multi-candidate inference performance. Ablation studies further demonstrate that functional-group feature encoding and cross-granularity interactive fusion substantially improve the ranking accuracy of high-confidence candidates, effectively supporting practical retrosynthetic route design.

1 INTRODUCTION

Molecular retrosynthesis refers to the process of inferring the optimal reactants and synthesis pathway from the target molecule in reverse, which is one of the core tasks of the modern drug DMTA cycle and organic material synthesis (Mao et al., 2021; Fang et al., 2023a). As the fundamental unit of retrosynthetic analysis, single-step retrosynthesis critically determines the feasibility and economic viability of multi-step synthetic pathway planning, and is therefore of substantial importance for shortening research and development cycles and reducing production costs. In the early stages, chemists primarily relied on domain expertise and practical experience to propose potential reactants and design synthetic routes. However, with the continuously growing global demand for innovative drugs and novel materials, traditional retrosynthesis paradigms that depend on chemists' experience and manual search have become inadequate for addressing the large-scale and highly complex molecular synthesis requirements, thereby constraining innovation efficiency in related fields.

In recent years, inspired by the successful applications of artificial intelligence across diverse domains, researchers have introduced deep learning methods into single-step retrosynthesis prediction, gradually giving rise to three major technical paradigms: Template-Based (TB), Semi-Template-Based (ST) and Template-Free (TF) methods. TB methods are based on predefined reaction templates and offer relatively high interpretability. Nevertheless, constrained by the coverage of the template set, they exhibit limited generalization to unseen reaction types. ST methods decompose the retrosynthesis task into two subtasks—reaction center identification and synthon completion—

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

but remain insufficient in modeling the interactions between them. In contrast, TF methods typically model single-step retrosynthesis task as a molecular sequence translation task from product to reactant and implement it using models such as Transformers, which have demonstrated strong performance in sequence statistical inference. Owing to their flexible end-to-end modeling capability and superior generalization ability, these methods have attracted widespread attention. In this setting, SMILES serves as a linear textual representation of molecules and reflects multi-granularity semantic information: individual atoms correspond to elementary tokens—analogous to characters in natural language—and thus constitute the fine-grained level; functional groups correspond to contiguous subsequences—analogous to words or fixed collocations—and thus constitute the coarse-grained level. The essence of a chemical reaction is the breaking of existing chemical bonds and the formation of new ones. This process is influenced not only by molecular features at both the atomic and functional-group levels, but also by cross-granularity interactions that are closely associated with bond reactivity at specific positions and with the distinctive reactivity of functional groups. However, existing methods have yet to explicitly extract, interactively model, and fuse multi-granularity features from molecular SMILES sequences, which limits further improvements in the accuracy of molecular retrosynthesis prediction.

To this end, we propose a Multi-Granularity Graph–Sequence jointly Enhanced Knowledge Transformer framework (MGSEKformer) for single-step retrosynthesis. Specifically, based on the SMILES of the target molecule, we construct atom-level token sequences and molecular graphs, and employ MESPool (EGIN) to construct the corresponding functionalgroup-level sequences and graphs, which serve as fine-grained and coarse-grained molecular representations, respectively. These representations are fed into parallel atom-level and functionalgroup-level Position-Aligned Graph-injected Transformers (Atom/FuncGroup-level PAGformers) to thoroughly extract molecular structural features and global sequential information. Building upon this, we introduce cross-attention mechanism together with a gated fusion unit to enable cross-granularity feature interaction and adaptive weighted fusion between atom level and functional-group level. Finally, the SMILES of reactants is generated by Transformer decoder. The main contributions are as follows:

- Given the significant differences in the impact of various granularity features on chemical reactions, we feed the atom-level graphs and sequences, together with the functionalgroup-level representations constructed by MESPool (EGIN), into two parallel PAGformers, which can model the corresponding long-range sequential dependencies under the structural priors of each granularity, thereby achieving multi-granularity molecular representation and feature extraction.
- Considering that the effects of different granularity features on bond cleavage and formation are not mutually independent, we introduce cross-attention to capture interactions between atom-level and functionalgroup-level features, and design a gated fusion unit that adaptively reweights features according to their contributions to retrosynthesis prediction.
- Since the volume of retrieved knowledge varies dynamically with the number of atoms or functional groups in a molecule, we design a parallel knowledge attention branch based on the standard feedforward transformer to replace the simple concatenation of fixed knowledge weights used in traditional Kformers, which enables adaptive injection of variable-length atomic or functional-group prior knowledge and overcomes dependence on reaction templates.

2 Related Works

With the deep integration of machine learning and deep learning into chemistry, substantial progress has been achieved in single-step retrosynthesis prediction. According to whether the model relies on predefined reaction templates, existing methods can be broadly categorized into three classes: Template-Based (TB), Semi-Template-Based (ST), and Template-Free (TF) methods.

2.1 Template-Based (TB)

The accuracy of reaction template selection is a critical factor for improving the performance of this class of methods, which mainly includes three types: similarity matching, classification and

embedding. Similarity-matching methods retrieve templates via similarity metrics, such as Tanimoto similarity search based on Morgan molecular fingerprints (Coley et al., 2017) and subgraph pattern matching based on GLN (Dai et al., 2019). Classification methods treat reaction templates as category labels and train multi-class classifiers to match candidate templates for target molecules, exemplified by deep multi-class neural networks based on ECFP4 (Segler & Waller, 2017) and the multi-scale hierarchical DHN (Baylon et al., 2019). Embedding methods map both the target molecule and reaction templates into a comparable representation space for subsequent matching, including the Modern Hopfield Network (MHN) (Seidl et al., 2022) and GLN based on graph neural network embedding for template logical reasoning (Dai et al., 2019).

The limitations of TB models are pronounced: on the one hand, constructing a comprehensive template set requires extensive manual effort; on the other hand, these methods cannot predict novel reactions beyond the template set, resulting in limited generalization capability.

2.2 SEMI-TEMPLATE-BASED (ST)

This class of methods decomposes the single-step retrosynthesis task into two sequential subtasks: reaction center identification and leaving-group (LG) prediction. Wang et al. used two Transformers to solve the sequence inference tasks of product to synthon and synthon to reactant, respectively (Wang et al., 2021). Yan et al. retain a Transformer for LG matching while replacing the reaction-center predictor with EGAT (Yan et al., 2020). Shi et al. employ R-GCN to identify reaction centers and complete synthons into reactants via variational graph translation (Shi et al., 2020). Somnath et al. encode atoms and bonds with MPN, localize reaction centers through predefined graph-editing operations, and select leaving groups in an autoregressive manner (Somnath et al., 2021). Chen et al. propose G2Retro, which first predicts reaction centers on the molecular graph, and gradually attaches substructures to restore the synthons to reactants (Chen et al., 2023).

In ST models, the two subtasks are largely decoupled, making it difficult to adequately model their intrinsic interdependencies.

2.3 TEMPLATE-FREE (TF)

According to the representation type of the input, TF methods can be further divided into graph-based, sequence-based, and graph–sequence fusion approaches. Graph-based methods start from the graph representation of the target molecule to infer bond cleavage and formation, such as MEGAN (Sacha et al., 2021). Sequence-based methods formulate retrosynthesis as a molecular sequence translation task from product to reactant. Liu et al. use an attention-augmented LSTM to complete the transformation of SMILES (Liu et al., 2017). As one of the core architectures in natural language processing, Transformer has been widely adopted in retrosynthesis owing to its flexible end-to-end modeling capacity and parallelization advantages. Zheng et al. introduce a syntax-checking module on top of the Transformer to reduce the proportion of invalid SMILES (Zheng et al., 2020). Ucak et al. replace SMILES with the sequences composed of atom-environment (AE) sets as input to NMT (Ucak et al., 2022). Fang et al. extract substructures that are frequently retained in reactions and perform substructure-level sequence prediction with a Transformer (Fang et al., 2023a). Sequence-based methods excel at capturing global molecular features, but struggle to effectively model molecular structure. Graph-based methods are adept at extracting structural features and local interactions among substructures. Consequently, their fusion has become a key direction for further performance gains. Mao et al. fuse SMILES sequences with structural features learned by GAES to construct a molecular graph-enhanced Transformer (Mao et al., 2021). Lin et al. propose G2GT, which builds graph encoders and graph decoders upon the Transformer and incorporates a self-training strategy (Lin et al., 2023). Tu et al. encode molecular graphs with D-MPNN and then generate reactant SMILES via Transformer decoder (Tu & Coley, 2022).

Most TF models focus solely on capturing structural features and global semantic information at the fine-grained atomic level, while paying insufficient attention to coarse-grained features of macroscopic substructures such as functional groups. Their capacity for multi-granularity molecular feature extraction therefore remains to be improved.

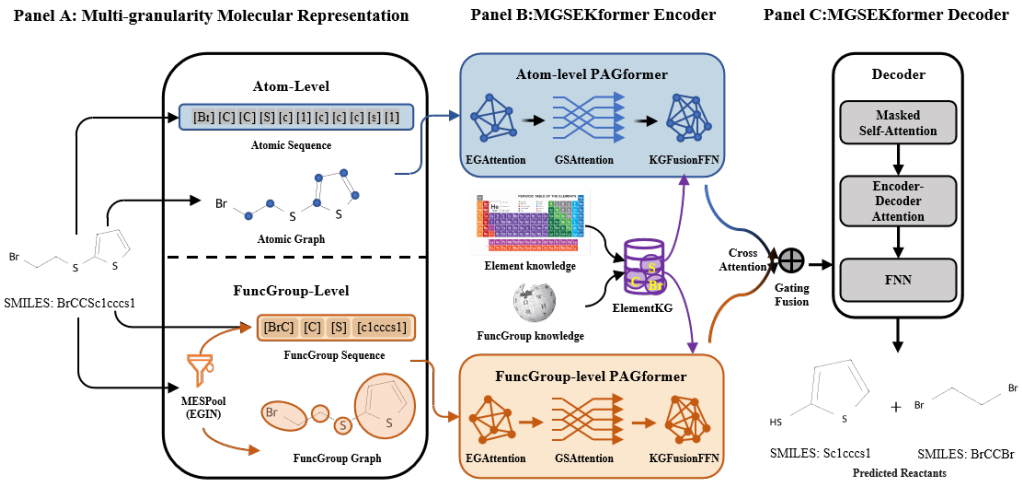


Figure 1: Overview of MGSEKformer. Panel A constructs the multi-granularity representation of product molecules by MESPool (EGIN). Panel B uses 2 parallel PAGformers to extract atom-level and functionalgroup-level features, and employs bidirectional cross attention and gating units to achieve their interaction and fusion. Meanwhile, prior knowledge with variable length is injected into them through KGFusionFFN. Panel C uses Transformer Decoder to generate reactant SMILES.

3 METHODOLOGY

3.1 OVERVIEW

Atom-level and functionalgroup-level molecular features exert substantially different influences on bond cleavage and formation. At the atomic level, the properties of the atoms connected by a bond (e.g., atomic size and electronegativity) directly affect the bond type, bond energy, and other properties. At the functional-group level, the properties of the functional groups attached to both ends of a bond (e.g., steric hindrance) influence the ease of bond cleavage or formation; moreover, structural characteristics among functional groups can also modulate reactivity at specific sites (for instance, the conjugated dicarbonyl structure in β-dicarbonyl compounds renders the intervening carbon an acidic site). In addition, the effects of features at different granularities on chemical reactions are not mutually independent, and cross-granularity interactions also play an important role: the cleavage and formation of chemical bonds within a functional group often depend on the corresponding special reactions; positional characteristics of functional groups relative to other structures (e.g., *ortho*/*meta*/*para* positions, or $\alpha/\beta/\gamma$ positions) can likewise activate or deactivate chemical bonds at particular sites. However, existing methods have not yet effectively distinguished and extracted molecular features at different granularities, nor have they achieved interaction and fusion between atom-level and functionalgroup-level features, thereby limiting the accuracy of molecular retrosynthesis prediction. To this end, we design MGSEKformer, a single-step molecular retrosynthesis framework, as illustrated in Figure 1.

Specifically, (1) to achieve multi-granularity molecular representation, we split the product SMILES at the character level to obtain the atom-level token sequence, and construct the corresponding atom-level molecular graph according to the one-to-one mapping between SMILES and the molecular graph, which serve as the fine-grained representation. We then apply MESPool (EGIN) to the atom-level graph to aggregate atom nodes and bonds within each functional group while preserving the original connectivity, thereby obtaining the functionalgroup-level molecular graph. We then re-segment the SMILES sequence based on the atomic-to-functional-group affiliation relationships, resulting in the coarse-grained representation. (2) The MGSEKformer encoder consists of a multi-granularity feature encoding layer and a cross-granularity feature interaction layer. In the multi-granularity feature encoding layer, we feed the atom-level and functionalgroup-level representations into parallel Atom/FuncGroup-level PAGformers to model their distinct effects on chemical reactions. Building upon the standard Transformer encoder (Vaswani et al., 2017), this module

introduces an Edge-aware Graph Attention layer (EGAttention) to process the molecular graph, aligning and integrating structural features into the sequence of the corresponding granularity, and employs a Graph-injected Self-Attention layer (GSAttention) to capture global dependencies while incorporating local structural priors. Meanwhile, following the knowledge injection paradigm of KFormer (Yao et al., 2022), we inject chemical knowledge into the FFN in the form of embeddings. Because the amount of retrieved knowledge varies dynamically with the numbers of atoms and functional groups in a molecule, the original method of fusing fixed knowledge weights through simple concatenation is not suitable. Therefore, we construct a knowledge attention branch based on the standard feedforward transformation to achieve adaptive injection of variable-length knowledge. In the cross-granularity feature interaction layer, we introduce a bidirectional cross-attention mechanism to capture cross-granularity feature interactions, and design a gating unit to perform adaptive weighted fusion of these features. (3) The final output of the encoder is processed by a Transformer decoder to autoregressively generate the SMILES sequence of the reactant molecule.

3.2 MESPool (EGIN)

This component consists of one EGIN layer and one MESPool pooling layer, which is used to identify the functional group structures of target molecules and construct coarse-grained representations. The theoretical derivation is detailed in the Appendix.

3.3 MGSEKformer Encoder

3.3.1 Multi-Granularity Feature Encoding Layer

The multi-granularity feature encoder comprises two parallel PAGformers, which process fine-grained atom-level representations and coarse-grained functionalgroup-level representations, respectively. Each PAGformer is composed of three components: EGAttention, GSAttention, and KGFusionFFN. We take the atom-level representations as an example to illustrate the processing pipeline.

EGAttention. Let the atom-level molecular graph for model input be $G = (V, \Sigma)$, where $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of atoms, and $\Sigma = \{\varepsilon_{ij} \mid v_i, v_j \in V\}$ denotes the set of chemical bonds. We linearly project the initial node feature h_i of atom v_i into a latent vector $h_i^0 = W_{\mathrm{EGA}} h_i + b_{\mathrm{EGA}}$, which is then layer-normalized to obtain h_i^{LN}. For the m-th attention head, the transformed node feature of atom v_i used for attention is $h_{\mathrm{EGA},i}^m = \Theta^m h_i^{\mathrm{LN}}$, and the transformed edge feature of bond ε_{ij} is $e_{\mathrm{EGA},ij}^m = \Phi^m e_{ij}$. Here, e_{ij} is the feature vector of bond ε_{ij}. The attention weight $a_{\mathrm{EGA},ij}^m$ is jointly determined by the node features $h_{\mathrm{EGA},i}^m$ and $h_{\mathrm{EGA},j}^m$ of atoms v_i and v_j, together with the edge feature $e_{\mathrm{EGA},ij}^m$ of bond ε_{ij}, thereby explicitly utilizing chemical bond characteristics, as formulated in equation 1:

$$a_{\mathrm{EGA},ij}^m = \mathrm{softmax}\big(\mathrm{LeakyReLU}\big(a_m^\top [h_{\mathrm{EGA},i}^m \| h_{\mathrm{EGA},j}^m \| e_{\mathrm{EGA},ij}^m]\big)\big). \tag{1}$$

Here, a_m is a learnable attention vector. Accordingly, the aggregated representation of atom v_i at the m-th attention head is $\tilde{h}_{\mathrm{EGA},i}^m = \sum_{j \in \mathcal{N}_i \cup \{i\}} a_{\mathrm{EGA},ij}^m h_{\mathrm{EGA},j}^m$, , where $\mathcal{N}_i$ denotes the first-order neighborhood of atom v_i, and the multi-head fused representation is $\tilde{h}_{\mathrm{EGA},i} = \mathrm{Concat}(\tilde{h}_{\mathrm{EGA},i}^1, \ldots, \tilde{h}_{\mathrm{EGA},i}^M)$. To incorporate neighborhood structural information while preserving the intrinsic features of atom v_i, we apply a linear transformation to $\tilde{h}_{\mathrm{EGA},i}$ and write it back to the node representation via a residual connection, yielding the structure-enhanced output $\tilde{h}_{\mathrm{EGA},i}^{\mathrm{out}} = h_i^0 + \mathrm{Dropout}(W_O \tilde{h}_{\mathrm{EGA},i} + b_O)$.

GSAttention. To enable Transformer to continue modeling long-range dependencies while considering local structural priors, we position-align and fuse the outputs of EGAttention with the token sequence of the corresponding granularity. At the fine-grained level, let the atom-level sequence of the target molecule be $S_{\mathrm{atom}} = [\langle \mathrm{SOS} \rangle, v_1, \ldots, v_N, \langle \mathrm{EOS} \rangle, \ldots]$, and let the corresponding token embeddings after positional encoding be denoted by $H_{\mathrm{atom}}^{\mathrm{tok}} = [h_1^{\mathrm{tok}}, \ldots, h_T^{\mathrm{tok}}]$. Since the tokens at both ends of the sequence do not correspond to graph nodes, we align $\tilde{h}_{\mathrm{EGA},i}^{\mathrm{out}}$ to the intervals occupied by atom tokens and perform element-wise addition, as formulated in equation 2. At the coarse-grained level, we first pad the functional-group supernodes obtained by MESPool into

fixed-length sequences for each sample, perform a two-stage injection on the supernode representations updated by EGAttention: (1) mean-pooling the valid supernodes within the same sample and broadcasting the result over the entire sequence as functional-group graph context; (2) adding each supernode representation onto the valid functional-group tokens at the corresponding positions.

$$h^0_{\mathrm{GSA},t} = \begin{cases} h^{\mathrm{tok}}_t + h^{\mathrm{out}}_{\mathrm{EGA},t-1}, & t = 2, 3, \ldots, N+1, \\ h^{\mathrm{tok}}_t, & \text{otherwise.} \end{cases} \tag{2}$$

Taking the atom level as an example, we project the ℓ-th layer input $H^{\ell}_{\mathrm{GSA}} = [h^{\ell}_{\mathrm{GSA},1}, \ldots, h^{\ell}_{\mathrm{GSA},T}]$ into the query matrix Q^{ℓ}_{GSA}, the key matrix K^{ℓ}_{GSA}, and the value matrix V^{ℓ}_{GSA} via three independent linear transformations in the same dimensions. Accordingly, the scaled dot-product attention weight $a^{\ell,m}_{\mathrm{GSA},ij}$ of the query $q^{\ell,m}_{\mathrm{GSA},i}$ with respect to the key $k^{\ell,m}_{\mathrm{GSA},j}$ in the m-th head of the ℓ-th layer is given by equation 3:

$$a^{\ell,m}_{\mathrm{GSA},ij} = \begin{cases} \mathrm{softmax}\left(\dfrac{(q^{\ell,m}_{\mathrm{GSA},i})^{\top} k^{\ell,m}_{\mathrm{GSA},j}}{\sqrt{d_1}} \right), & M_j = 0, \\ 0, & M_j = 1. \end{cases} \tag{3}$$

Here, M_j is the mask indicator. $M_j = 0$ indicates that the j-th token is visible and the softmax is performed normally; $M_j = 1$ indicates that the j-th token is masked, and its attention weight is set to zero so that invalid tokens do not participate in attention allocation. Consequently, the updated representation of the i-th token after fusing sequential contextual information under the mask constraint in the m-th attention head of the ℓ-th layer is $\tilde{h}^{\ell,m}_{\mathrm{GSA},i} = \sum_{j=1}^{L} a^{\ell,m}_{\mathrm{GSA},ij} v^{\ell,m}_{\mathrm{GSA},j}$, and the updated representation of the sequence is $\tilde{H}^{\ell,m}_{\mathrm{GSA}} = [\tilde{h}^{\ell,m}_{\mathrm{GSA},1}, \ldots, \tilde{h}^{\ell,m}_{\mathrm{GSA},L}]$. After concatenation and linear projection, we obtain $\tilde{H}^{\ell}_{\mathrm{GSA}} = \mathrm{Concat}(\tilde{H}^{\ell,1}_{\mathrm{GSA}}, \ldots, \tilde{H}^{\ell,M}_{\mathrm{GSA}}) W^{\ell}_{O} + b^{\ell}_{O}$. We further apply residual connection and layer normalization to produce the ℓ-th layer output $H^{\ell,\mathrm{out}}_{\mathrm{GSA}} = \mathrm{LN}(H^{\ell}_{\mathrm{GSA}} + \mathrm{Dropout}(\tilde{H}^{\ell}_{\mathrm{GSA}}))$, which introduces contextual updates while preserving the original sequential information and stabilizing deep training.

KGFusionFFN. To accommodate the fact that the amount of retrieved prior chemical knowledge varies with the number of atoms or functional groups contained in the product molecule, KGFusionFFN adopts a Pre-LayerNorm dual-branch architecture, in which a standard feed-forward transformation branch and a knowledge attention branch operate in parallel.

In the standard feed-forward transformation branch, we take the ℓ-th layer output $H^{\ell,\mathrm{out}}_{\mathrm{GSA}}$ of GSAttention as input, and apply two layers of linear transformations with layer normalization and GELU activation, as shown in equation 4:

$$H^{\ell}_{\mathrm{FFN}} = W^{\ell}_{2} \cdot \mathrm{GELU}\Big(W^{\ell}_{1} \cdot \mathrm{LN}(H^{\ell,\mathrm{out}}_{\mathrm{GSA}}) + b^{\ell}_{1} \Big) + b^{\ell}_{2}. \tag{4}$$

For the knowledge attention branch, we employ a knowledge injection strategy that is position-aligned with the atomic or functional-group structures. At the atomic level, we iterate through all tokens of the product molecule and sequentially retrieve the knowledge embedding vector z_t corresponding to the atom of the t-th token from the prior knowledge graph of elements, ElementKG. At the functional-group level, the knowledge embedding vector z_g corresponding to the g-th functional-group supernode token is given by equation 5:

$$z_g = \frac{1}{|A_g|} \sum_{i \in A_g} z_i. \tag{5}$$

Here, A_g is the set of all atom indices belonging to the g-th functional group. Taking the atom level as an example, we project each atomic knowledge embedding z_t into a key $k^{\ell}_{\mathrm{FFN},t}$ and a value $v^{\ell}_{\mathrm{FFN},t}$ via the linear transformations. The knowledge key and value vectors aligned with the corresponding atom-token positions are $K^{\ell}_{\mathrm{FFN}} = [k^{\ell}_{\mathrm{FFN},1}, \ldots, k^{\ell}_{\mathrm{FFN},L}]$ and $V^{\ell}_{\mathrm{FFN}} = [v^{\ell}_{\mathrm{FFN},1}, \ldots, v^{\ell}_{\mathrm{FFN},L}]$. Using H^{ℓ}_{FFN} as the query, we compute the attention weight matrix A^{ℓ}_{FFN} of the atom-level sequence over the corresponding knowledge, as formulated in equation 6:

$$A^{\ell}_{\mathrm{FFN}} = \mathrm{softmax}\left(\frac{H^{\ell}_{\mathrm{FFN}} (K^{\ell}_{\mathrm{FFN}})^{\top}}{\sqrt{d_2}} \right). \tag{6}$$

Accordingly, the knowledge-augmented sequence representation at the ℓ-th layer is $\tilde{H}^{\ell}_{\mathrm{FFN}} = A^{\ell}_{\mathrm{FFN}} \cdot V^{\ell}_{\mathrm{FFN}}$, and the final output is $H^{\ell,\mathrm{out}}_{\mathrm{FFN}} = H^{\ell,\mathrm{out}}_{\mathrm{GSA}} + \mathrm{Dropout}(H^{\ell}_{\mathrm{FFN}} + \tilde{H}^{\ell}_{\mathrm{FFN}})$, which realizes soft fusion of variable-length knowledge while preserving the residual learning paradigm of Transformer.

3.3.2 Cross-Granularity Feature Interaction Layer

The component comprises a bidirectional cross-attention layer (CrossAttention) and a gated fusion unit (GatingFusion). In CrossAttention, we take the atom-level and functionalgroup-level sequences finally produced by the multi-granularity feature encoding layer as inputs, and compute multi-head cross-attention in both directions—functional group to atom and atom to functional group—in parallel, so as to model how their interactions influence chemical reactions. In GatingFusion, we adaptively reweight features according to the contribution of each granularity to retrosynthesis prediction, thereby effectively integrating cross-granularity information while mitigating redundancy. The theoretical derivation is detailed in the Appendix.

3.4 MGSEKformer Decoder

The decoder follows the original Transformer architecture and consists of three components: a masked self-attention layer, an encoder–decoder attention layer and a position-wise FFN. Conditioned on the encoder output H_{out} and the outputs of previous decoding steps, the decoder iteratively generates the next SMILES token x_t until $x_t = \langle \mathrm{EOS} \rangle$.

4 Data Processing

The USPTO-FULL and USPTO-50K are the most widely used benchmark datasets for chemical reaction prediction. After removing duplicate samples, we randomly split the data into training, validation, and test sets at a ratio of 8:1:1. The resulting training, validation, and test sets contain 810,497, 101,312, and 101,312 samples, respectively. Based on this, we perform $5\times$ R-SMILES augmentation on the training and validation sets, while retaining a single normalized representation for the test set to ensure comparability in model performance evaluation. We pretrain MESPool (EGIN) and convert each chemical reaction sample into a quadruplet of the form (atom-level product sequence/graph, functionalgroup-level product sequence/graph, atom-to-functional-group membership mapping, reactant SMILES) for storage, thereby improving the training efficiency of MGSEKformer. The specific data processing process is detailed in the appendix.

We construct ElementKG, a chemical knowledge graph comprising both instance-level and class-level hierarchies (Fang et al., 2023b), with the required knowledge primarily derived from the Periodic Table of Elements (https://ptable.com). We employ OWL2Vec to learn continuous vector representations of entities and relations (Chen et al., 2021).

5 Experiments

5.1 Experimental Setup

We use the first three terms of the pre-stored reaction quadruple as inputs and the reactant SMILES as outputs to train MGSEKformer. The major hyperparameter settings of MGSEKformer are provided in the Appendix. Using the trained model, for each product molecule in the test set, we generate the top-k reactant candidate sequences with the highest cumulative log-probabilities, and compare them with the actual reactants after RDKit canonicalization. The inference hit rate $\mathrm{Top}\text{-}k$ is computed according to equation 7 as the evaluation metric, thereby eliminating bias caused by chemically equivalent molecules with different string representations:

$$\mathrm{Top}\text{-}k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big(\exists\, j \in \{1, \ldots, k\},\ y_i^{(j)} = y_i \Big). \tag{7}$$

Here, N is the number of valid samples in the test set, y_i is the canonicalized SMILES of the reactants in the i-th reaction sample, $\{y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}\}$ is the set of k best reactant candidates generated by the model.

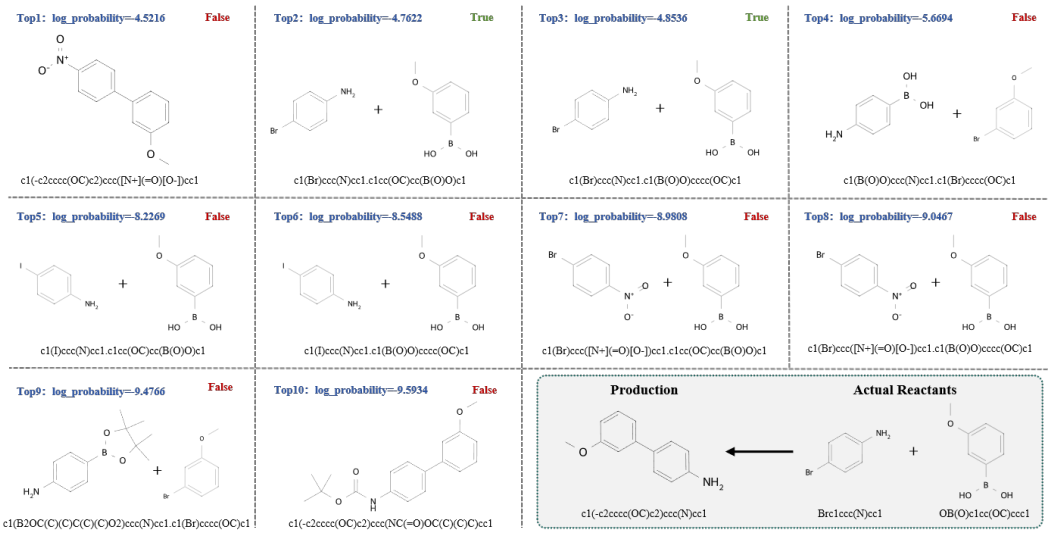


Figure 2: Top-10 reactant candidates predicted by MGSEKformer for No.US04859700. Top-2 and Top-3 are consistent with the actual reactants, and the other 8 candidates are also reasonable results with certain chemical basis.

5.2 Experimental Results

Analysis of Inference Results. Taking No.US04859700 reaction as an example, the top-10 reactant candidate sequences inferred by MGSEKformer with the highest cumulative log_probability are shown in Figure 2. (1) 8 predictions from Top-2 to Top-9 correctly identify the biaryl C–C bond in the biphenyl scaffold as the bond-cleavage site and output reactant pairs composed of two monocyclic aromatics. This indicates that MGSEKformer possesses a strong capability for reaction-site recognition. (2) Top-2 and Top-3 candidates match the ground-truth reactants; the only difference between them lies in the linearization starting point of the boronic acid fragment in SMILES, reflecting the strong ability of MGSEKformer to match the correct leaving group. Moreover, the log_probability gap between these two correct candidates is very small ($\Delta = 0.09$), indicating that the model assigns stable and relatively high confidence to the correct answers. (3) The remaining 8 candidates are likewise chemically plausible. For example, the nitro group (–NO_2) is an important precursor of the amino group (–NH_2). Top-7 and Top-8 replace the amino group of the actual reactants with the nitro group, which would require two steps: a coupling reaction (i.e., generating Top-1) followed by a reduction. It can be interpreted as a component of a multi-step synthetic route, suggesting that MGSEKformer exhibits a certain degree of multi-step retrosynthetic awareness.

Robustness. To further analyze model robustness, we train MGSEKformer with 10 different random seeds and compute the corresponding inference hit rates Top-1, Top-3, Top-5, and Top-10. The box plots are provided in the Appendix. It confirms that the inference hit rates of MGSEKformer remain stable across different random seeds, demonstrating strong robustness on the single-step retrosynthesis task.

Comparative Experiments. To assess the effectiveness of MGSEKformer, we compare it with 13 baseline methods on USPTO-FULL. The results are reported in Table 1. MGSEKformer achieves the best performance on Top-3, Top-5, and Top-10, reaching 63.4%, 68.6%, and 73.7%, respectively. Its Top-1 is 47.3%, ranking second and only 0.1 percentage points lower than RXN-Graphormer. These results indicate that MGSEKformer attains the strongest multi-candidate inference capability while maintaining a high top-1 hit rate. This finding is important for assisting practical retrosynthetic design, as chemists typically conduct a comprehensive evaluation of several leading candidate reactants rather than relying solely on the Top 1 sequence.

Table 1: The results of the comparative experiment on USPTO-FULL and USPTO-50K (Top-n, %).

	USPTO-FULL				USPTO-50K			
Model	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Template-Based (TB)								
RetroSim (Coley et al., 2017)	32.8	–	–	56.1	37.3	54.7	63.3	74.1
NeuralSym (Segler & Waller, 2017)	35.8	–	–	60.8	44.4	65.3	72.4	78.9
LocalRetro (Chen & Jung, 2021)	39.1	53.3	58.4	63.7	53.4	**77.5**	**85.9**	**92.4**
GLN (Dai et al., 2019)	39.3	–	–	63.7	52.5	74.6	80.5	86.9
Semi-Template-Based (ST)								
RetroPrime (Wang et al., 2021)	44.1	59.1	62.8	68.5	51.4	70.8	74.0	76.1
Graph2Edits (Zhong et al., 2023)	44.0	60.9	66.8	72.5	**55.1**	77.3	83.4	89.4
Template-Free (TF)								
MEGAN (Sacha et al., 2021)	33.6	–	–	63.9	48.1	70.7	78.4	86.1
Transformer (Vaswani et al., 2017)	42.9	–	–	66.8	42.4	58.6	63.8	67.7
Aug.Transformer (Tetko et al., 2020)	44.4	–	–	70.4	–	–	–	–
Graph2SMILES (Tu & Coley, 2022)	45.7	–	–	63.4	–	–	–	–
Substructure (Fang et al., 2023a)	46.0	–	–	68.5	–	–	–	–
GTA (Seo et al., 2021)	46.6	–	–	70.4	51.1	67.6	74.8	81.6
RXNGraphormer (Xu et al., 2025)	**47.4**	63.0	67.4	71.6	–	–	–	–
MGSEKformer	47.3	**63.4**	**68.6**	**73.7**	47.0	66.3	72.3	77.8

Table 2: The results of the ablation experiment (Top-n, %).

Variants	Top-1	Top-3	Top-5	Top-10
Baseline	46.0	62.7	68.3	**74.1**
Baseline+A	46.6	63.1	68.4	73.8
Baseline+A+B	47.1	63.1	68.3	73.8
Full Model	**47.3**	**63.4**	**68.6**	73.7

Ablation Experiments. To evaluate the contributions of functional group feature encoding (i.e., FuncGroup-level PAGformer, Component A), cross-granularity feature interaction (including CrossAttention and GatingFusion, Component B) and variable-length prior knowledge injection (i.e., KGFusionFFN, Component C), we use the Atom-level PAGformer with a standard FFN as the baseline and create 3 variants for ablation experiments. The results are shown in Table 2. The performance of Baseline+A and Baseline+A+B demonstrates that the extraction of functionalgroup-level features and the interaction with atom-level features primarily enhance its discriminability of the best candidates, thereby effectively improving prediction accuracy in the high-confidence regime, while contributing only limitedly to expanding the coverage of valid candidates. In contrast, the introduction of Component C comprehensively improves Top-1, Top-3 and Top-5, suggesting that the foundational knowledge provided by ElementKG enhances the chemical validity of the inferred SMILES. It not only further improves prediction accuracy in the high-confidence interval, but also moderately expands the coverage of feasible reaction pathways, thereby partially remedying the aforementioned limitation.

6 Conclusion

Considering the differing influences of atom-level and functionalgroup-level molecular features on bond cleavage and formation in chemical reactions, we propose MGSEKformer for single-step retrosynthesis. We train and evaluate it on USPTO-FULL, and find that MGSEKformer exhibits strong reaction-site identification and leaving-group matching capabilities on the large-scale dataset covering diverse reaction types. Comparative experiments against 13 baseline models confirm that MGSEKformer achieves the best multi-candidate inference performance. Ablation studies further verify that functionalgroup-level feature encoding and cross-granularity feature interaction effectively improve the ranking accuracy of high-confidence candidates, while the injection of variable-length prior chemical knowledge further expands the coverage of valid candidates.

AI USE STATEMENT

In this work, we have not used generative AI tools for research ideation, experimental design, data analysis, or figure generation.

REPRODUCIBILITY STATEMENT

We provide implementation details, dataset processing steps, and hyperparameter settings in the main text and Appendix. Upon acceptance, we will release source code and pretrained model checkpoints to facilitate reproduction of the reported results.

REFERENCES

Javier L. Baylon, Nicholas A. Cilfone, Jeffrey R. Gulcher, et al. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *Journal of Chemical Information and Modeling*, 59(2):673–688, 2019.

Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, et al. OWL2Vec*: Embedding of OWL ontologies. *Machine Learning*, 110(11):2857–2894, 2021.

Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

Ziqi Chen, Oluwaseyi R. Ayinde, James R. Fuchs, et al. G(2)Retro as a two-step graph generative model for retrosynthesis prediction. *Communications Chemistry*, 6(1):102, 2023.

Connor W. Coley, Luke Rogers, William H. Green, et al. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science*, 3(10):1237–1245, 2017.

Hanjun Dai, Chengtao Li, Connor Coley, et al. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, pp. 8870–8880, 2019.

Lin Fang, Junren Li, Ming Zhao, et al. Single-step retrosynthesis prediction by leveraging commonly preserved substructures. *Nature Communications*, 14(1):2446, 2023a.

Yin Fang et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553, 2023b.

Zaiyun Lin, Shiqiu Yin, Lei Shi, et al. G2GT: Retrosynthesis prediction with graph-to-graph attention neural network and self-training. *Journal of Chemical Information and Modeling*, 63(6): 1894–1905, 2023.

Bowen Liu et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3(9):1103–1113, 2017.

Kelong Mao et al. Molecular graph enhanced transformer for retrosynthesis prediction. *Neurocomputing*, 457:193–202, 2021.

Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, et al. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.

Marwin H. S. Segler and Mark P. Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry*, 23(10):5966–5971, 2017.

Philipp Seidl et al. Improving few- and zero-shot reaction template prediction using modern hopfield networks. *Journal of Chemical Information and Modeling*, 62(5):2111–2120, 2022.

Seung-Woo Seo et al. GTA: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 531–539. AAAI Press, 2021.

Chence Shi, Minkai Xu, Hongyu Guo, et al. A graph-to-graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning*. PMLR, 2020.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, et al. Learning graph models for retrosynthesis prediction. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Igor V. Tetko, Pavel Karpov, Ruud Van Deursen, et al. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11(1):1–11, 2020.

Zhengkai Tu and Connor W. Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling*, 62 (7):3503–3513, 2022.

Umit V. Ucak, Islambek Ashyrmamatov, Junsu Ko, et al. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature Communications*, 13(1):1186, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Xiaorui Wang et al. RetroPrime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845, 2021.

Li-Cheng Xu, Man-Jie Tang, Jin An, et al. A unified pre-trained deep learning framework for cross-task reaction performance prediction and synthesis planning. *Nature Machine Intelligence*, 2025.

Chaochao Yan et al. RetroXpert: Decompose retrosynthesis prediction like a chemist. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yunzhi Yao, Shaohan Huang, Li Dong, et al. Kformer: Knowledge injection in transformer feed-forward layers. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022*, pp. 131–143, Berlin, Heidelberg, 2022. Springer-Verlag.

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, et al. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1): 47–55, 2020.

Weihe Zhong, Zidong Yang, and Calvin Yu-Chian Chen. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1): 3009, 2023.

A Appendix

A.1 MESPool (EGIN)

This component consists of one EGIN layer and one MESPool layer, and is used to identify the functional-group structures of the target molecule.

EGIN. Let the atom-level molecular graph for model input be $G = (V, \Sigma)$, where $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of atoms, and $\Sigma = \{\varepsilon_{ij} \mid v_i, v_j \in V\}$ denotes the set of chemical bonds. EGIN updates the l-th layer edge feature $e_{i\to j}^{(l)}$ of bond ε_{ij} by using the $(l-1)$-th layer node features $h_i^{(l-1)}$ of the atom v_i, as shown in Eq. (8):

$$e_{i\to j}^{(l)} = \mathrm{ReLU}\Big(W_{\mathrm{EGIN}}\big(e_{i\to j}^{(l-1)} \,\|\, h_i^{(l-1)}\big) + b_{\mathrm{EGIN}}\Big). \tag{8}$$

Here, W_{EGIN} is a learnable transformation matrix. Through the above operator, the undirected edge ε_{ij} is converted into a directed edge with direction $i \to j$. The feature vector $e_{i\to j}$ of edge $\varepsilon_{i\to j}$ incorporates the feature h_i of its starting point. Therefore, $e_{i\to j}$ can be regarded as the weighted information sent from the neighboring atom node v_i to the central atom node v_j. By aggregating the l-th layer weighted messages from all atoms in set of nodes in the first-order neighborhood $N(j)$, the l-th layer feature $h_j^{(l)}$ of the central atom node v_j is updated as in Eq. (9):

$$h_j^{(l)} = \mathrm{MLP}\left(h_j^{(l-1)} \,\|\, \sum_{k\in N(j)} e_{k\to j}^{(l)}\right). \tag{9}$$

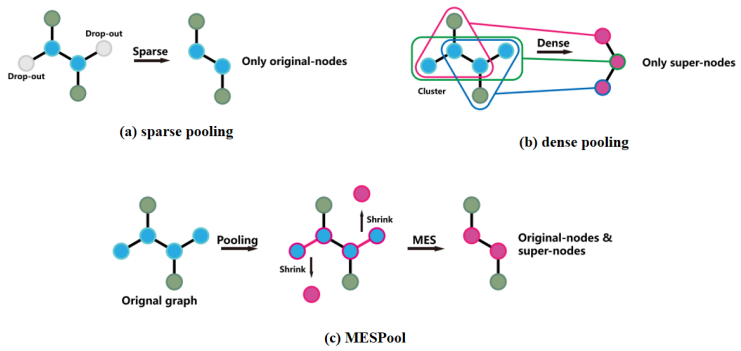


Figure 3: Comparison of MESPool with other pooling methods. (a) Sparse pooling results in the loss of some nodes; (b) Dense pooling allows a single atom to belong to multiple supernodes simultaneously; (c) MESPool forms functional group supernodes by aggregating redundant edges while preserving the original connectivity of the molecular graph.

MESPool. Figure 3 compares the MESPool process with other common molecular pooling methods. Sparse pooling retains only the important atom nodes from the original molecule and discards the others, as shown in Figure 3(a). Dense pooling aggregates multiple connected atoms into supernodes, where a single atom may simultaneously belong to multiple supernodes, as shown in Figure 3(b). In contrast, MESPool forms functional-group supernodes by aggregating redundant edges in the original molecule, and each atom node belongs to exactly one supernode, thereby preserving the original connectivity of the molecular graph, as depicted in Figure 3(c). Consequently, the pooling results of MESPool are suitable for the effective interaction between atom-level and functionalgroup-level molecular features, avoiding the adverse effects of missing or duplicated nodes. Therefore, we select the MESPool method to identify the functional group structures in the target molecules.

MESPool computes the directed score $s_{e(i,j)}$ of edge ε_{ij} based on the directed feature $e_{i\to j}$, as formulated in Eq. (10):

$$s_{e(i,j)} = \sigma\left(\frac{s^\top \mathrm{MLP}\left(e_{i\to j} + \sum_{k\in N(j)} e_{k\to i}\right)}{\|s\|_2}\right). \tag{10}$$

Here, s denotes a learnable scoring operator. Although chemical bonds are inherently undirected, their directed scores are typically asymmetric, as they reflect the differing roles of the two endpoint atoms within the local structure. For example, the directed score from a central atom to a peripheral atom within a functional group is often relatively high, whereas the reverse score tends to be lower. We set the pooling threshold as $\mathrm{threshold} = \min(\lambda, s_{\mathrm{mean}})$, and shrink edges for which either directed score falls below this threshold, so that nodes can be merged once affinity within a cluster is indicated from either direction. Here, λ is a manually specified hyperparameter, and s_{mean} denotes the average of all edge scores. The set of edges to be pooled is then defined as $\Sigma_{\mathrm{pool}} = \{\varepsilon_{ij} \mid s_{e(i,j)} < \mathrm{threshold} \vee s_{e(j,i)} < \mathrm{threshold}\}$, and the set of nodes involved in pooling is $V_{\mathrm{pool}} = \bigcup_{\varepsilon_{ij}\in\Sigma_{\mathrm{pool}}}\{v_i, v_j\}$. We apply a SCC algorithm to partition V_{pool} into mutually non-overlapping subsets $V_{\mathrm{super}(k)}$, each corresponding to all atoms belonging to functional group k, and contract each subset into a supernode that represents the functional group as a whole. The feature of the resulting functional-group supernode is given by $h_{\mathrm{super}(k)} = \sum_{v_i\in V_{\mathrm{super}(k)}} h_i$. Finally, a functionalgroup-level molecular graph $G_{\mathrm{new}} = (V_{\mathrm{new}}, \Sigma_{\mathrm{new}})$ is obtained, where $V_{\mathrm{new}} = (V \setminus V_{\mathrm{pool}}) \cup V_{\mathrm{super}}$, $\Sigma_{\mathrm{new}} = \big\{\{\phi(v_i), \phi(v_j)\} \mid \varepsilon_{ij} \in \Sigma,\ \phi(v_i) \neq \phi(v_j)\big\}$, and $\phi : V \to V_{\mathrm{new}}$.

A.2 Training Details of MESPool (EGIN)

Before training MGSEKformer, we pretrain MESPool (EGIN) on the product SMILES sequences from the training set to identify functional-group structures. The overall optimization objective $\mathcal{L}_{\mathrm{Total}}$ defined during the training stage, is given by Eq. (11):

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{Node}} + 0.2\,\mathcal{L}_{\mathrm{Pool}} + 0.1\,\mathcal{L}_{\mathrm{Score}} + 0.01\,\mathcal{L}_{\mathrm{Entropy}}. \tag{11}$$

Table 3: Hyperparameter of MESPool (EGIN).

Component	Meaning	Value
EGIN	Node in/out dim.	121
	Edge in/out dim.	13
	MLP layers	2
MESPool	Node feature dim.	121
	Edge feature dim.	13
	Edge selection threshold	0.5
MESPool (EGIN)	Batch size	64
	Epochs	10
	Learning rate	1×10^{-3}
	Weight decay	1×10^{-5}

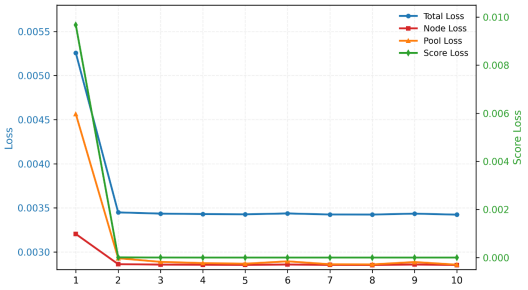


Figure 4: Training process of MESPool (EGIN).

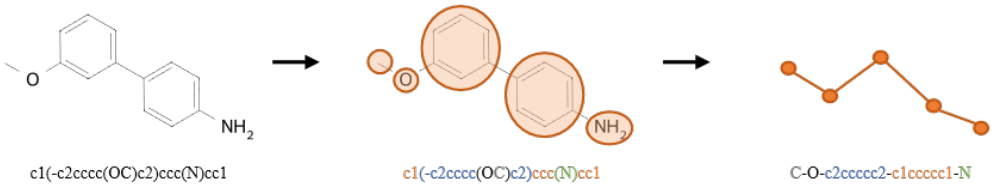


Figure 5: The pooling result for the product molecule of No.US04859700 by MESPool (EGIN).

Here, $\mathcal{L}_{\text{Node}}$ is the node-feature reconstruction MSE. It decodes and reconstructs the original atomic features after sampling functional group-level features according to the atomic-to-functional group affiliation relationships, thereby measuring the fidelity of the pooling process in preserving the original node information. $\mathcal{L}_{\text{Pool}}$ is the pooling-consistency MSE that constrains the upsampled functionalgroup-level representations to align with the atom-level latent representations, reflecting the structural stability of the aggregation–deaggregation mapping. $\mathcal{L}_{\text{Score}}$ is the edge-score supervision BCE, which uses bond-type priors to supervise edge scoring and guide the model in distinguishing contractible edges from those that should be retained. $\mathcal{L}_{\text{Entropy}}$ is an entropy regularization term for edge scoring probabilities. By minimizing the uncertainty in the scoring distribution, it suppresses ambiguous intermediate scores and promotes more definitive edge scoring decisions. The main model hyperparameters are provided in Table 3.

The training process of MESPool (EGIN) is shown in Figure 4. MESPool (EGIN) achieved rapid convergence in 1∼2 epochs, and the loss curves remained stable in 3∼10 epochs. The overall training process was smooth without significant oscillations. This indicates that MESPool (EGIN) has learned the main functional group pooling and molecular reconstruction modes in the early stage, and subsequently focused on minor optimizations, with good convergence speed and training stability. More specifically, (1) the $\mathcal{L}_{\text{Node}}$ curve and $\mathcal{L}_{\text{Pool}}$ curve stabilized at similar levels (approximately $2.85 \sim 2.93 \times 10^{-3}$) after the second round, and their trends were highly consistent, indicating that the model achieved a good balance between structural compression ability and information recovery ability; (2) the $\mathcal{L}_{\text{Score}}$ curve rapidly decays from its initial high value and approaches zero, indicating that the edge scoring branch established clear discriminative boundaries in the early stages of training.

Using the trained MESPool (EGIN), we perform one pooling pass over the product molecules of all reaction samples in USPTO-FULL and USPTO-50K to construct their corresponding coarse-grained molecular representations. Taking No.US04859700 reaction as an example, the pooling result of its product molecule is shown in Figure 5. This demonstrates that the trained MESPool (EGIN) component possesses a strong capability for identifying molecular functional-group structures and can generate functionalgroup-level representations from atom-level molecular representations with relatively high accuracy.

A.3 Cross-Granularity Feature Interaction Layer

This layer consists of two components: a bidirectional cross attention mechanism (CrossAttention) and a gated fusion unit (GatingFusion).

CrossAttention. For the query matrix Q, the key matrix K and the value matrix V, the basic multi-head cross-attention operator $\mathrm{MHA}(Q, K, V)$ is defined as in Eq. (12):

$$\begin{aligned} \mathrm{head}_m &= \mathrm{softmax}\left(\frac{(QW_Q^m)(KW_K^m)^\top}{\sqrt{d_k}} + M\right)(VW_V^m), \\ \mathrm{MHA}(Q,K,V) &= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_M)W_O'. \end{aligned} \tag{12}$$

We take the atom-level and functionalgroup-level sequences finally produced by the multi-granularity feature encoding layer as the inputs to the first layer, denoted by H_{atom}^0 and H_{fg}^0, respectively, and perform R layers of cross-attention transformations. At the r-th layer, we compute multi-head cross-attention in both directions—functional-group to atom and atom to functional-group—in parallel, as formulated in Eqs. (13) and (14):

$$H_{\mathrm{atom}}^r = \mathrm{LN}\Big(H_{\mathrm{atom}}^{r-1} + \mathrm{Dropout}\big(\mathrm{MHA}(H_{\mathrm{atom}}^{r-1}, H_{\mathrm{fg}}^{r-1}, H_{\mathrm{fg}}^{r-1})\big)\Big). \tag{13}$$

$$H_{\mathrm{fg}}^r = \mathrm{LN}\Big(H_{\mathrm{fg}}^{r-1} + \mathrm{Dropout}\big(\mathrm{MHA}(H_{\mathrm{fg}}^{r-1}, H_{\mathrm{atom}}^{r-1}, H_{\mathrm{atom}}^{r-1})\big)\Big). \tag{14}$$

GatingFusion. Because the lengths of the functionalgroup-level and atom-level sequences differ, the functionalgroup-level features must be upsampled to the atomic resolution before fusion. Based on the MESPool results, the atom-to-functional-group membership mapping under the sequence representation is defined as $\pi : \{1, 2, \ldots, N_{\mathrm{atom}}\} \to \{0, 1, \ldots, N_{\mathrm{fg}} - 1\} \cup \{-1\}$, where -1 denotes invalid positions (e.g., start/end tokens), and N_{atom} and N_{fg} are the numbers of atoms and functional groups, respectively. The aligned functional-group features are then given by Eq. (15):

$$H_{\mathrm{fg}}^{\mathrm{align}}[i] = \begin{cases} H_{\mathrm{fg}}^R[\pi(i)], & \pi(i) \geq 0, \\ 0, & \pi(i) < 0. \end{cases} \tag{15}$$

At the atomic resolution, we adaptively fuse the atom-level features H_{atom}^R with the aligned functional-group features $H_{\mathrm{fg}}^{\mathrm{align}}$, as formulated in Eq. (16):

$$\begin{aligned} G &= \sigma\big([H_{\mathrm{atom}}^R \,\|\, H_{\mathrm{fg}}^{\mathrm{align}}]W_g + b_g\big), \\ H_{\mathrm{GF}} &= G \odot H_{\mathrm{atom}}^R + (1 - G) \odot H_{\mathrm{fg}}^{\mathrm{align}}. \end{aligned} \tag{16}$$

Here, G is the gating coefficient matrix. Consequently, the final encoder output is $H_{\mathrm{out}} = \mathrm{LN}\big(\mathrm{Dropout}(H_{\mathrm{GF}})W_O'' + b_O''\big)$.

A.4 Training Details of MGSEKformer

During training, we adopt the AdamW optimizer with an 8,000-step linear learning-rate warm-up followed by a cosine annealing schedule, so as to stabilize early parameter updates and promote smooth mid-to-late-stage convergence, thereby improving generalization. The loss function is cross-entropy with label smoothing, which alleviates overfitting and overconfidence to enhance model robustness. Other major hyperparameter settings of MGSEKformer are provided in Table 4.

Table 4: Hyperparameter settings of MGSEKformer.

Meaning	Value	Meaning	Value	Meaning	Value
Hidden dimension	256	Attention heads	8	Feed-forward dimension	1024
Encoder layers	6	Decoder layers	6	Bidirectional cross-attn. layers	2
KG embedding dimension	133	Node feature dimension	121	Edge feature dimension	13
AdamW learning rate	2×10^{-4}	AdamW weight decay	0.01	AdamW momentum	$(0.9, 0.98)$
AdamW ε	1×10^{-9}	Samples per step	64	Maximum epochs	100

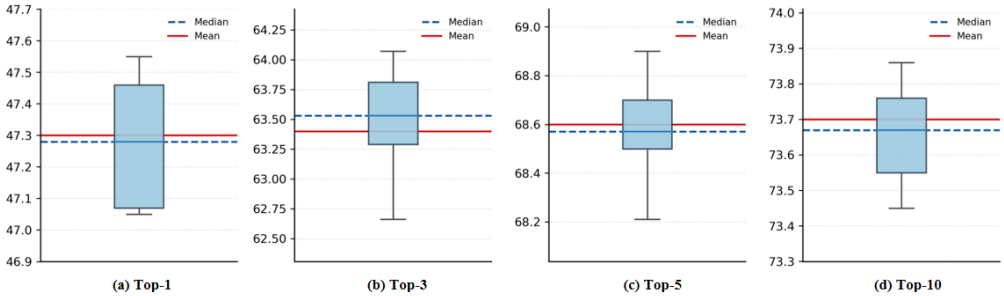


Figure 6: Box plots of Top-1, Top-3, Top-5 and Top-10 over 10 random seeds.

A.5 Robustness of MGSEKformer

Based on the fact that the output of neural network models is influenced by initialization parameters, we train MGSEKformer with 10 different random seeds and compute the corresponding inference hit rates Top-1, Top-3, Top-5, and Top-10. The box plots are provided in the Appendix. It confirm that the inference hit rates of MGSEKformer remain stable across different random seeds, demonstrating strong robustness on the single-step retrosynthesis task.