1 Supplementary Materials

To evaluate the effectiveness of our proposed method, we conducted two sets of ablation experiments. The first set aimed to verify the effectiveness of the network model by comparing the performance of single-stream, dual-stream, and multi-stream networks based on different inputs, as shown in Table 3. Firstly, the results indicate that the performance of multi-stream networks with ECFP, SMILES and combination representation as input (MSNR) is better than the compared single-stream network and dual-stream network using other representations in different top-k accuracies, which illustrates the effectiveness of the MSNR for the research problem. Secondly, the performance of the dual-stream network using ECFP and combination representation as input is superior to the single-stream network using ECFP as input, and the performance of the dual-stream network using SMILES and combination representation as input surpasses the single-stream network using SMILES as input. Besides, the performance of the multi-stream network using ECFP, SMILES and combination representation as input (MSNR) outperforms the dual-stream network using ECFP and SMILES as input. These phenomena prove that the combination representation is a value representation, and it is useful for the retrosynthesis prediction.

The second set of ablation experiments aimed to validate the effectiveness of the selected molecular representations. We collected common molecular representations used in existing retrosynthesis literature, including direct molecular structure descriptions (e.g., SMILES and graph) and representations of molecular sub-structures or fragments (e.g., ECFP and MACCS Keys). As shown in Table 4, Each composition of the representations in this ablation experiment consists of a molecular structural representation and a representation describing the molecules by molecular fragments or sub-structural composition. The MSNR using ECFPs and SMILES as representations outperform the method utilizing other representations, which indicates the sub-structure composition information in ECFPs and the atomic composition information in SMILES are beneficial for the retrosynthesis task compared to other compositions of representations. Furthermore, the results demonstrate that features extracted by the Text-CNN outperform those extracted by the transformer encoder for retrosynthesis prediction. This suggests that Text-CNN can more effectively capture local molecular structure information, which is critical for retrosynthesis prediction, without losing important original features in comparison to the transformer encoder.

Table 3 Top-k exact match accuracy of single-stream networks, dual-stream networks, and multi-stream networks on USPTO-50k dataset.

Method	Top-1%	Top-3%	Top-5%	Top-10%
Single-Stream(ECFP)	40.2	56.6	61.1	65.3
Single-Stream (SMILES)	44.9	62.3	67.0	71.4
Single-Stream (Combination)	45.6	61.1	65.3	69.9
Dual-Stream(ECFP+SMILES)	41.3	57.6	62.3	66.8
Dual-Stream(SMILES+ Combination)	52.0	67.4	71.3	75.0
Dual-Stream(ECFP+ Combination)	52.5	67.6	71.5	75.1
Multi-Stream (ECFP+SMILES + Combination)	53.0	68.4	72.2	75.6

Table 4 Top-k exact match accuracy of the molecular representations on the USPTO-50k dataset.

Method	Top-1%	Top-3%	Top-5%	Top-10%
MSNR(MACCS Keys+Graph)	29.5	46.3	52.6	59.4
MSNR(MACCS Keys+SMILES(Transformer))	38.2	55.5	61.1	66.6
MSNR(MACCS Keys+SMILES(Text-CNN))	49.0	65.6	69.7	73.8
MSNR(ECFP+Graph)	40.5	57.6	62.8	67.9
MSNR(ECFP+SMILES(Transformer))	46.2	61.9	66.3	70.4
MSNR(ECFP+SMILES(Text-CNN))	53.0	68.4	72.2	75.6