

RFL: Simplifying Chemical Structure Recognition with Ring-Free Language

Qikai Chang¹, Mingjun Chen¹, Changpeng Pi², Pengfei Hu¹, Zhenrong Zhang¹, Jiefeng Ma¹, Jun Du^{1*}, Baocai Yin², Jinshui Hu²

¹NERC-SLIP, University of Science and Technology of China

²iFLYTEK Research

{qkchang, mingjunchen, pengfeihu, z zr666, jfma}@mail.ustc.edu.cn, jundu@ustc.edu.cn, {cp, bc, yin, jshu}@iflytek.com

Abstract

The primary objective of Optical Chemical Structure Recognition is to identify chemical structure images into corresponding markup sequences. However, the complex two-dimensional structures of molecules, particularly those with rings and multiple branches, present significant challenges for current end-to-end methods to learn one-dimensional markup directly. To overcome this limitation, we propose a novel Ring-Free Language (RFL), which utilizes a divide-and-conquer strategy to describe chemical structures in a hierarchical form. RFL allows complex molecular structures to be decomposed into multiple parts, ensuring both uniqueness and conciseness while enhancing readability. This approach significantly reduces the learning difficulty for recognition models. Leveraging RFL, we propose a universal Molecular Skeleton Decoder (MSD), which comprises a skeleton generation module that progressively predicts the molecular skeleton and individual rings, along with a branch classification module for predicting branch information. Experimental results demonstrate that the proposed RFL and MSD can be applied to various mainstream methods, achieving superior performance compared to state-of-the-art approaches in both printed and handwritten scenarios.

Code — <https://github.com/JingMog/RFL-MSD>

1 Introduction

Chemical structures, including molecules and formulas, are fundamental representations in chemistry. Optical chemical structure recognition (OCSR) (Rajan et al. 2020) accurately converts handwritten and printed chemical structures into machine-readable formats across various media. This technology offers significant advantages in drug development, scientific research, the chemical industry, and education (Hu et al. 2023; Oldenhof et al. 2024; Chen et al. 2024). Unlike one-dimensional common text and tree-structured mathematical expressions (Zhang et al. 2020), chemical structures are two-dimensional and graph-structured. Consequently, they contain more complex structures, such as rings and multiple branches. Therefore, models require not only robust capabilities to interpret the relationships between symbols and contexts but also advanced spatial analysis abilities.

*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

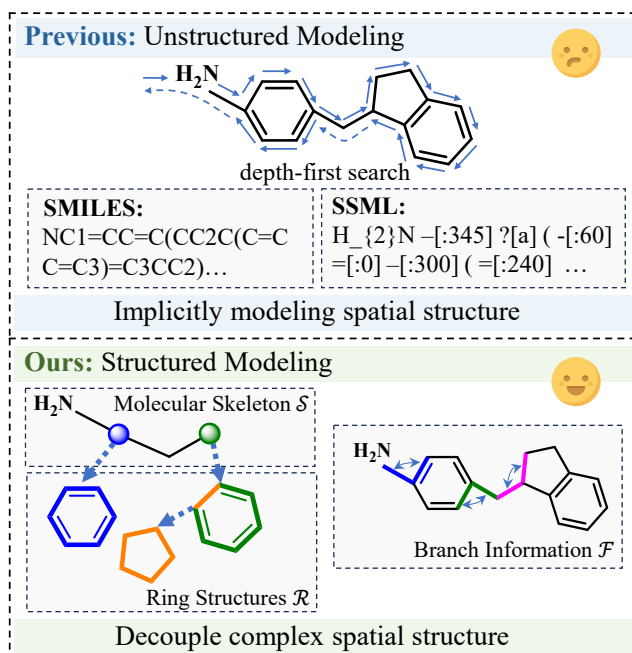


Figure 1: Comparison of Ring-Free Language with previous modeling method. Previous methods use depth-first search to establish one-dimension markup, where spatial structures are implicitly expressed. Our RFL significantly simplifies complex spatial structures by decoupling them into molecular skeleton \mathcal{S} , ring structures \mathcal{R} and branch information \mathcal{F} .

Recent advances in computer vision have driven the development of bottom-up based OCSR techniques (Xu et al. 2022a; Zhang et al. 2022; Morin et al. 2023; Oldenhof et al. 2024; Yao et al. 2024). These approaches use multiple models and OCR engines to separately detect atoms and bonds, then reconstruct molecular structures. However, these methods are time-consuming, prone to error accumulation. Consequently, many recent studies on OCSR employ end-to-end algorithms (Weir et al. 2021; Xu et al. 2022b,a; Hu et al. 2023; Qian et al. 2023), which are jointly optimized without additional modules. These algorithms mainly rely on image-to-markup frameworks, with markup typically in SMILES (Weininger 1988) or SSML (Hu et al. 2023). In addition,

many methods adopt motif-based modeling approaches to reduce sequence length (Yu and Gao 2022; Zang, Zhao, and Tang 2023; Jiang et al. 2023). These formats encapsulate complex two-dimensional molecular structures into one-dimensional markups. For instance, SSML uses depth-first traversal to generate strings, resulting in one-dimensional strings that represent molecular structures, as shown in Figure 1. Although this string-based representation is straightforward, there is no explicit design to guide the model in learning the spatial structure of the graph, leading to poor recognition performance in complex molecular structures with current end-to-end methods.

To address these issues, we propose a novel chemical formula modeling language named **Ring-Free Language (RFL)**. By equivalent conversion, RFL can decouple complex molecule structures and reducing the difficulty of the model’s learning process. For a molecular structure G , it will be equivalently converted into a molecular skeleton \mathcal{S} , individual ring structures \mathcal{R} and branch information \mathcal{F} , as shown in Figure 1. This approach effectively decouples and explicitly models the spatial structure of chemical molecules. Specifically, during graph traversal, individual rings are merged into a special atom, termed SuperAtom. Multi-ring structures are merged into a special bond, referred to as SuperBond, based on their adjacency until only individual rings remain, which are then merged into SuperAtom. The branch information \mathcal{F} represents the mapping relationships between skeleton bonds and ring bonds, ensuring accurate reconstruction of the original molecular structure.

Based on Ring-Free Language, we propose a universal Molecular Skeleton Decoder (MSD) including a skeleton generation module and a branch classification module. The skeleton generation module is tasked with predicting \mathcal{S} and \mathcal{R} , while branch classification module handles \mathcal{F} . Utilizing a divide-and-conquer strategy, the model first predicts the molecular skeleton, followed by ring structures, and finally reconstructs the molecular structure. Instead of parsing all complex molecular structures simultaneously, the model first parses the molecular skeleton and then individually predicts each ring structure, thereby reducing the errors.

We validate our method on the handwritten dataset EDU-CHEMC (Hu et al. 2023) and printed dataset Mini-CASIA-CSDB (Ding et al. 2022). We also select two representative mainstream methods as baselines to demonstrate the universality of our approach. Experiments show that our method significantly enhancing recognition ability for complex structures in both handwritten and printed scenarios.

Our primary contributions are as follows:

- We propose **Ring-Free Language (RFL)**, which decouples complex molecular structures into molecular skeleton, ring structures, and branch information, thereby significantly simplifying the prediction process.
- Based on RFL, we propose a universal Molecular Skeleton Decoder (MSD) that can be applied to various mainstream approaches.
- Comprehensive experiments show that our method surpasses the state-of-the-art methods with different baselines on both printed and handwritten scenarios.

2 Related Work

2.1 Bottom-Up Based OCSR

During the early stages of OCSR development from 1990 to 2017, bottom-up based methods dominated the field. These approaches relied on hand-crafted rules to process images and detect all elements of chemical molecules, including chemical bonds, atomic groups, charges, and so on. Subsequently, they performed image vectorization to construct connection tables or graphs, thereby reconstructing the molecular structures (Filippov and Nicklaus 2009; Fujiyoshi, Nakagawa, and Suzuki 2011; Smolov, Zentsev, and Rybalkin 2011; Ouyang and Davis 2011; Sadawi, Sexton, and Sorge 2012; Bukhari, Iftikhar, and Dengel 2019). However, rule-based systems typically struggle with noisy images and are challenging and time-consuming to develop.

With advancements in machine learning, numerous innovative techniques have emerged in OCSR. ChemGrapher (Oldenhof et al. 2020) and ABC-Net (Zhang et al. 2022) utilize Convolutional Neural Networks (CNNs) to predict molecular structures directly. MolGrapher (Morin et al. 2023) employs a ResNet (He et al. 2016) backbone to locate atom nodes within molecules and construct a Supergraph, which is then classified using a Graph Neural Network (GNN). Although graph-based methods perform well in complex printed scenarios, they often require additional OCR engines for support and are difficult to apply in complex handwritten scenarios. Consequently, many end-to-end methods have been developed to enhance recognition capabilities in various scenarios.

2.2 Image Caption Based OCSR

Most recent image caption based OCSR methods utilize an encoder-decoder end-to-end architecture. The encoder extracts features from the input image and passes them as input to the decoder. The decoder generates the final image caption, typically in SMILES (Weininger 1988), InChI (Heller et al. 2015) or SSML (Hu et al. 2023).

MSE-DUDL (Staker et al. 2019) employs a U-Net segmentation network (Ronneberger, Fischer, and Brox 2015) to extract and segment chemical structure diagrams from images. ChemPix (Weir et al. 2021), DECIMER (Rajan, Zielesny, and Steinbeck 2020), and Img2Mol (Clevert et al. 2021) use CNN encoders and RNN decoders, such as LSTM, GRU, or standard RNN, to detect chemical structures. Recent works have also incorporated Transformer decoders, such as Transformer (Khokhlov et al. 2022), Swin Transformer (Xu et al. 2022b) and Vision Transformer (Sundaramoorthy et al. 2021). Additionally, some methods used sampler to improve performance (Yao, Li, and Xiao 2024).

However, these methods treat chemical molecular captions as simple strings without explicitly modeling them to enhance recognition performance, resulting in weak recognition performance in complex scenarios with current end-to-end solutions. Our new modeling language addresses this limitation by decoupling molecular structures through a divide-and-conquer strategy, thereby enhancing recognition capability for complex structures in end-to-end solutions.

3 Ring-Free Language

To decouple molecular structures, we propose a novel chemical structure modeling language called **Ring-Free Language (RFL)**. For a given molecular structure G , RFL converts it into an equivalent representation consisting of the molecular skeleton \mathcal{S} , individual ring structures \mathcal{R} and branch information \mathcal{F} . The objective of the model is to separately predict \mathcal{S} , \mathcal{R} and \mathcal{F} , and ultimately reconstruct the original molecular structure. Each part of the RFL is separated by [ea] token.

The decoupling process adheres to specific principles: (1) Independent rings are merged into a SuperAtom, removed from G , and then appended into \mathcal{R} . (2) In cases involving adjacent multiple rings, the adjacency relationship between the rings, denoted as γ , is first resolved. The rings are then merged into a SuperBond in ascending order of γ , and subsequently appended into \mathcal{R} . The following sections will elaborate on the generation process of RFL, Splitting, and the graph restoration process, Restoring.

3.1 Splitting

The molecular structure is represented as a simple graph $G = (V, E)$, where V denotes the set of vertices and E denotes the set of edges. To decouple the ring structures, we first employ a modified depth-first search algorithm to detect all non-nested rings \mathcal{R} in G . The set \mathcal{R} is defined as follows:

$$\mathcal{R} = \{C \in \mathcal{C} \mid \forall C' \in \mathcal{C}, (C' \subsetneq C \Rightarrow C' \not\subseteq C)\} \quad (1)$$

where \mathcal{C} represents all rings in G . According to the principles, we must analyze the adjacency relationships γ between rings to determine the appropriate splitting strategy.

For a ring r_i , if $\gamma(r_i) = 0$, according to principle (1), r_i will be merged into a SuperAtom and separated from graph G . The connection relationships \mathcal{F} between r_i and G will be updated accordingly. \mathcal{F} is defined as follows:

$$\mathcal{F} = \{(b_{i,j}, b_{u,v}) \mid b_{i,j} \in E_g \wedge b_{u,v} \in E_r \wedge j = u\} \quad (2)$$

where $b_{i,j}$ represents the bond from v_i to v_j , E_g is the set of edges in G , and E_r is the set of edges in \mathcal{R} . The direction of chemical bonds is determined during graph traversal, with a clockwise traversal adopted for rings. If $\gamma(r_i) > 0$, according to principle (2), r_i will be merged into a SuperBond and separated from graph G . The SuperBond is one of the common bonds of r_i and r_j . \mathcal{F} will be updated accordingly. Ultimately, all r_i are separated from G , resulting in a ring-free structure and yielding the molecular skeleton \mathcal{S} .

As an example, we will demonstrate how to apply the Splitting process to the molecular shown in Figure 2, starting from step 0. Step 1: A modified depth-first search algorithm is applied to obtain $\mathcal{R} = \{r_1, r_2, r_3\}$. Step 2: The adjacency relationships are analyzed, $\gamma = \{r_1: 0, r_2: 1, r_3: 1\}$. In ascending order of γ , we merge the ring structures in sequence (r_1, r_2, r_3). Step 3: For r_1 , since $\gamma(r_1) = 0$, it is merged into SuperAtom v_{18} and appended into \mathcal{R} . The branch information \mathcal{F} is updated to $\mathcal{F} = \{(b_{7,2}, b_{2,1}), (b_{8,5}, b_{5,4})\}$. Step 4: For r_2 , since $\gamma(r_2) = 1$, it is merged into SuperBond $b_{10,11}$ and appended into \mathcal{R} . The branch information is updated to $\mathcal{F} \cup \{(b_{8,9}, b_{9,13})\}$. Step 5: For r_3 , since $\gamma(r_3) = 0$, it is merged into SuperAtom v_{19} . Finally, the resulting sets of SuperAtom $A = \{v_{18}, v_{19}\}$, SuperBond $B = \{b_{10,11}\}$.

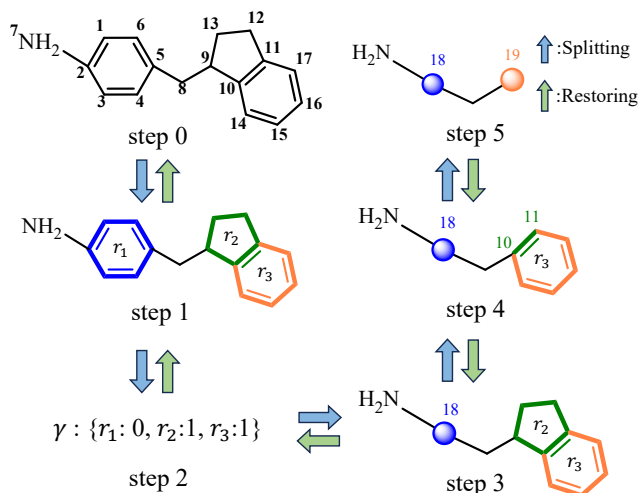


Figure 2: The step-by-step decoupling process of Ring-Free Language, including the stages of Splitting and Restoring.

3.2 Restoring

The process of Restoring is the reverse of Splitting. In this procedure, individual ring structures \mathcal{R} are sequentially restored into the molecular skeleton \mathcal{S} , replacing the SuperAtom and SuperBond. The branch information \mathcal{F} is consumed until all rings are restored, at which point $\mathcal{F} = \emptyset$.

Take the molecular skeleton in Figure 2 step 5 as an example. The set of SuperAtoms $A = \{v_{18}, v_{19}\}$ and SuperBonds $B = \{b_{10,11}\}$. In the reverse order of Splitting, r_3 is first restored into \mathcal{S} to replace SuperAtom v_{19} . At this stage $\mathcal{R} = \{r_1, r_2\}$, $A = \{v_{18}\}$, $B = \{b_{10,11}\}$ and $\mathcal{F} = \{(b_{7,2}, b_{2,1}), (b_{8,5}, b_{5,4}), (b_{8,9}, b_{9,13})\}$, as shown in step 4. Next, r_2 is restored into \mathcal{S} , replacing SuperBond $b_{10,11}$. At this stage, $\mathcal{R} = \{r_1\}$, $A = \{v_{18}\}$, $B = \emptyset$ and $\mathcal{F} = \{(b_{7,2}, b_{2,1}), (b_{8,5}, b_{5,4})\}$, as shown in step 3. Finally, r_1 is restored into \mathcal{S} , replacing SuperAtom v_{18} . At this final stage, $\mathcal{R} = \emptyset$, $A = \emptyset$, $B = \emptyset$ and $\mathcal{F} = \emptyset$, indicating the end of the restoration process, as shown in step 1.

4 Molecular Skeleton Decoder

Based on the RFL, we propose a universal Molecular Skeleton Decoder (MSD) designed to leverage the advantages of RFL and enhance the model’s recognition ability for complex molecules. The MSD consists of a skeleton generation module and a branch classification module. The overall architecture is shown in Figure 3. The skeleton generation module employs a hierarchical decoding approach to progressively predict molecular skeleton and rings. Subsequently, the node features of skeleton bonds and ring bonds are passed to the branch classification module to predict branch information, which represents the connection relationships between the molecular skeleton and individual rings. Finally, the complete molecular structure is restored.

4.1 Skeleton Generation Module

The skeleton generation module is responsible for predicting \mathcal{S} and \mathcal{R} . It employs an encoder-decoder architecture,

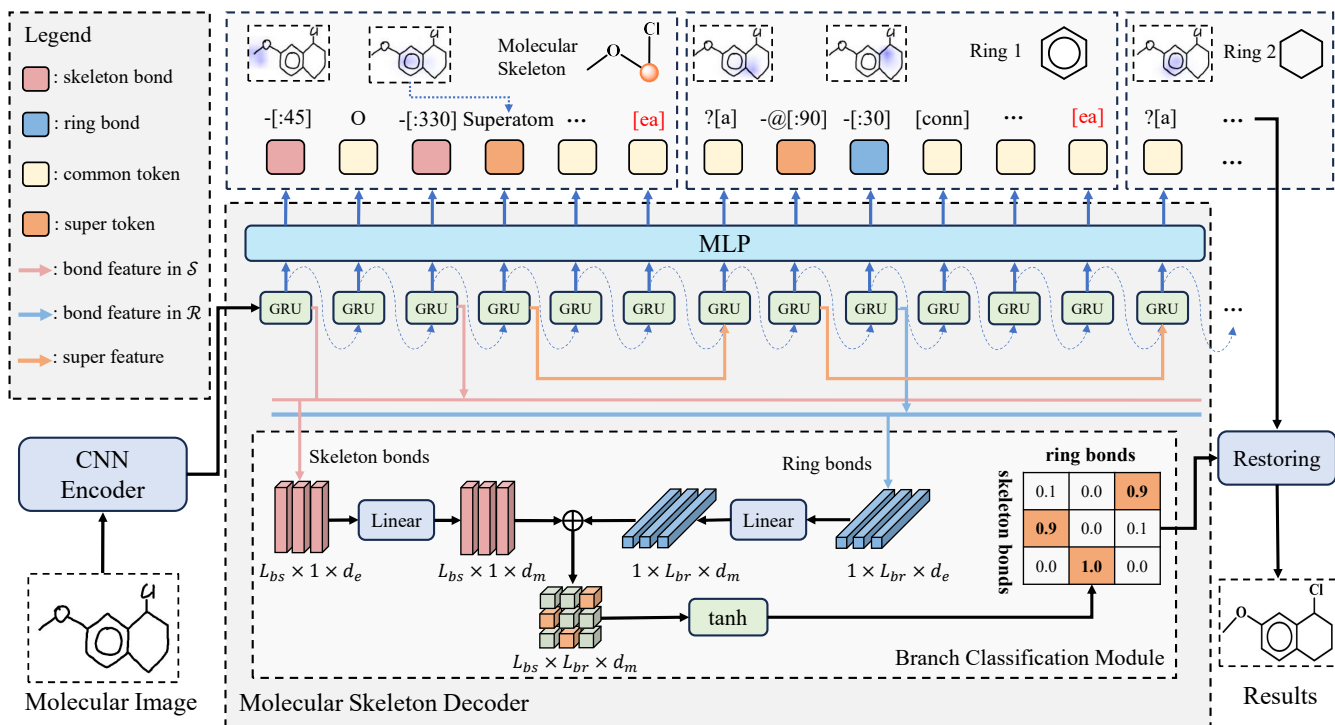


Figure 3: The architecture of our method. First, the molecular image is input into the CNN Encoder to extract deep features. Subsequently, the molecular skeleton decoder autoregressively decodes the skeleton \mathcal{R} and rings \mathcal{S} . During this process, the node features of skeleton and ring bonds are sent to the Branch Classification Module to derive branch information \mathcal{F} , which is used to restore the predicted molecular structure. The common token refers to tokens in RFL that are not specially processed.

with the encoder based on DenseNet (Huang et al. 2017) and the decoder is a GRU (Chung et al. 2014) model equipped with attention mechanisms (Zhang, Du, and Dai 2018). The DenseNet encoder extracts the feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ from the input images, where C represents the number of channels, W and H represent the width and height of the feature map. The decoder takes the feature map \mathbf{F} as input and decodes each element in an autoregression manner. It adopts a hierarchical decoding approach: first detecting the molecular skeleton \mathcal{S} , then using the node feature vectors of SuperAtom and SuperBond as conditional information to continue predicting individual ring structures \mathcal{R} .

During the decoding process, when the model encounters ring structures, it does not fully parse them in one step. Instead, it first identifies and temporarily stores the node feature \mathbf{f}^s of SuperAtom and SuperBond, then gradually decodes the ring structure only after the entire molecular skeleton has been fully decoded. With the node feature \mathbf{f}^s as conditional information, the decoder continues decoding the ring structures once the molecular skeleton decoding is complete. Each node feature of the SuperAtom and SuperBond corresponds to a ring structure that needs to be decoded.

$$\mathbf{f}_t^s = \text{Concat}(\mathbf{E}\mathbf{y}_t, \mathbf{s}_t, \mathbf{c}_t) \quad (3)$$

where $\mathbf{E} \in \mathbb{R}^{d_e \times V}$ represents word embedding matrix, $\mathbf{y}_t \in \mathbb{R}^V$ denotes the one-hot output vector at time t , $\mathbf{s}_t \in \mathbb{R}^{d_s}$ is the output state at time t , $\mathbf{c}_t \in \mathbb{R}^{d_c}$ is the context vector, and

V is the vocabulary size. \mathbf{c}_t is calculated as follows:

$$\mathbf{c}_t = \sum_{i=1}^{H \times W} \alpha_{ti} \mathbf{x}_i \quad (4)$$

$$\alpha_{ti} = \frac{e_{ti}}{\sum_{k=1}^{H \times W} e_{tk}} \quad (5)$$

$$\mathbf{e}_{ti} = \mathbf{w}^T \tanh(\mathbf{W}_x \mathbf{x}_i + \mathbf{W}_y \mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_s \mathbf{s}_t + \mathbf{W}_a \mathbf{a}_i + \mathbf{W}_f \mathbf{f}_u^s) \quad (6)$$

where \mathbf{w} , \mathbf{W}_x , \mathbf{W}_y , \mathbf{W}_s , \mathbf{W}_a , \mathbf{W}_f are trainable weights. The term e_{ti} represents the energy of \mathbf{x}_i at time step t . The current attention weight α_{ti} is obtained by applying the softmax function to e_{ti} . The coverage vector \mathbf{a}_i , used to address the lack of coverage with historical attention weight information (Tu et al. 2016), is calculated as:

$$\mathbf{a}_i = \sum_{k=1}^{t-1} \alpha_{tk} \quad (7)$$

To achieve hierarchical decoding, the node feature \mathbf{f}_u^s is input as an additional condition into the decoder to calculate the energy e_{ti} at the current time step in Eq. (6), ensuring that the model can continue decoding individual ring structures from the time step u corresponding to \mathbf{f}_u^s . When the model predicts [ea], it switches the prediction phase, at which point \mathbf{f}^s is a non-zero vector. The entire decoding

phase ends when there are no SuperAtoms or SuperBonds left to decode and the end token [END] is decoded.

At every time step, the decoded token is:

$$P(\mathbf{y}_t | \mathbf{y}_{\leq t-1}, \mathbf{x}) = g(\mathbf{W}_o h(\mathbf{W}_c \mathbf{c}_t + \mathbf{W}_y \mathbf{E} \mathbf{y}_{t-1} + \mathbf{W}_s \mathbf{s}_t)) \quad (8)$$

where g denotes the softmax activation function, h denotes the maxout activation function. A cross-entropy loss \mathcal{L}_{ce} is used as training objective for skeleton generation module:

$$\mathcal{L}_{ce} = - \sum_{t=1}^T \log P(\mathbf{y}_t | \mathbf{y}_{\leq t-1}, \mathbf{x}) \quad (9)$$

4.2 Branch Classification Module

The branch classification module is responsible for predicting \mathcal{F} , which represents the connection relationships between skeleton bonds and ring bonds, as described in Eq. (2). We construct a simple binary classifier to determine whether a connection exists between a skeleton bond b_s and a ring bond b_r .

$$\mathbf{q}_{sr} = \tanh(\mathbf{W}_b \mathbf{f}_{b_s}^s + \mathbf{W}_r \mathbf{f}_{b_r}^s) \quad (10)$$

where $\mathbf{W}_b, \mathbf{W}_r$ are trainable weights, \mathbf{f}_{b_s} is the node feature of skeleton bonds, \mathbf{f}_{b_r} is the node feature of ring bonds. The output $\mathbf{q}_{sr} \in \mathbb{R}^{L_{bs} \times L_{br}}$ is the probability distribution of \mathcal{F} . The loss function of the branch classification module is formulated as follows:

$$\mathcal{L}_{cls} = - \sum_{s=1}^{L_{bs}} \sum_{r=1}^{L_{br}} \mathbf{y}_{sr} \log(\mathbf{q}_{sr}) \quad (11)$$

where \mathbf{y}_{sr} is the one-hot vector of classification label.

The Branch Classification Module identifies the connected pairs from all candidate skeleton bonds and ring bonds. As illustrated in Figure 1 below, many bonds on the rings are not connected to the skeleton. If all candidate bonds are fed into the branch classification module, the resulting classification matrix would be excessively sparse. To address this, a [conn] token is added after the connected ring bonds, ensuring that only the filtered ring bonds need to be classified, significantly reducing the number of candidate bonds. Essentially, the [conn] token divides the branch prediction task into two parts: predicting ring bonds and predicting skeleton bonds. The skeleton generation module is responsible for predicting ring bonds, while the branch classification module handles the prediction of skeleton bonds. This approach significantly reduces classification errors.

5 Experiments

5.1 Implementation Details

To ensure a fair comparison with the baseline methods DenseWAP (Zhang, Du, and Dai 2018) and RCGD (Hu et al. 2023), we utilize the same DenseNet encoder (Huang et al. 2017), including three dense blocks. The growth rate and depth in each dense block are set to 24 and 32. The Molecular Skeleton Decoder (MSD) employs a GRU (Cho et al. 2014) with a hidden state dimension of 256. The embedding dimension is 256, and a dropout rate of 0.15 is applied.

The overall model is trained end-to-end. The training objective is to minimize the cross-entropy loss \mathcal{L}_{ce} and the branch classification loss \mathcal{L}_{cls} . The objective function for optimization is defined as follows:

$$\mathcal{O} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{cls} \quad (12)$$

In our experiments, we set $\lambda_1 = \lambda_2 = 1$. The Adam optimizer (Kingma and Ba 2014) is used with an initial learning rate of 2×10^{-4} , and the parameters are set as $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$. The learning rate adjustment strategy employs MultiStepLR with a decay factor $\gamma = 0.5$. All experiments are conducted on 4 NVIDIA Tesla V100 GPUs with 32GB RAM, using a batch size of 8 for the EDU-CHEMC dataset and 32 for the Mini-CASIA-CSDB dataset. The training epoch is set to 50, and the whole framework is implemented using PyTorch.

5.2 Datasets

We evaluate the performance of the proposed method on both handwritten dataset and printed dataset, as well as on our synthetic dataset.

EDU-CHEMC (Hu et al. 2023) contains 48,998 training samples and 2,992 testing samples of handwritten molecular structure images collected from various educational scenarios in the real world. The images are captured from various devices, including cameras, scanners, and screens. This dataset presents significant challenges due to its diverse writing styles and complex molecular structures.

Mini-CASIA-CSDB (Ding et al. 2022) contains 89,023 training samples and 8,287 testing samples of printed molecular structure images collected from the chemical database ChEMBL (Gaulton et al. 2017). The images are rendered using RDKit (Landrum 2013). We preprocess the dataset using the same method as in (Hu et al. 2023).

5.3 Metrics

For evaluation metrics, we use EM and Struct-EM to evaluate the performance of our model provided by CROCS-2024¹. EM measures the percentage of exact-match molecular images. Struct-EM measures the percentage of correctly identified structures in the predicted results, ignoring the non-chemical parts of the formulas. Struct-EM serves as an auxiliary evaluation metric.

5.4 Comparison With State-of-the-Art

To demonstrate the superiority of our method, we compare it with previous state-of-the-art (SOTA) methods. Table 1 presents the experimental results on the EDU-CHEMC dataset and Mini-CASIA-CSDB dataset. It can be observed that using DenseWAP (Zhang, Du, and Dai 2018) as the baseline, MSD-DenseWAP achieves significant improvements on both EDU-CHEMC and Mini-CASIA-CSDB. Additionally, it surpasses the latest SOTA method on EDU-CHEMC by a notable margin of 2.06%. This demonstrates the effectiveness of our method in enhancing the model’s recognition ability. The relatively smaller improvement on the printed dataset Mini-CASIA-CSDB is mainly due to the lower proportion of complex molecules in this dataset.

¹<https://crocs-ifyly-ustc.github.io/crocs/index.html>

Methods	Markup	Params (M)	Flops (G)	EDU-CHEMC		Mini-CASIA-CSDB	
				EM(%)	Struct-EM(%)	EM(%)	Struct-EM(%)
Imago (Smolov, Zentsev, and Rybalkin 2011) †	SMILES	-	-	0.00	0.00	38.80	38.88
WYGIWYS (Deng, Kanervisto, and Rush 2016) ‡	SMILES	-	-	-	-	78.55	-*
BTTR (Zhao et al. 2021) ‡	SSML	4.77	9.65	58.21	66.83	78.22	-*
ABM (Bian et al. 2022) ‡	SSML	22.45	19.39	58.78	67.24	-	-
CoMER (Zhao and Gao 2022) †	SSML	4.99	11.18	59.47	68.71	90.67	91.06
DenseWAP (Baseline) (Zhang, Du, and Dai 2018)	SSML	15.39	18.98	61.35	69.68	92.09	92.47
MSD-DenseWAP (Ours)	RFL	16.01	21.00	64.92	73.15	<u>94.10</u>	<u>94.44</u>
RCGD (Baseline) (Hu et al. 2023)	SSML	16.11	21.02	62.86	71.88	95.01	95.38
MSD-RCGD (Ours)	RFL	16.74	23.04	65.39	73.26	95.23	95.58

Table 1: Comparison with state-of-the-art methods on handwritten dataset (EDU-CHEMC) and printed dataset (Mini-CASIA-CSDB) in %. * indicates using SMILE as the learning target on Mini-CASIA-CSDB, resulting in the absence of Struct-EM. †denotes our reimplementations results, ‡refers to results from (Hu et al. 2023). MSD-DenseWAP and MSD-RCGD use DenseWAP and RCGD as baselines, respectively. All Params(M) and Flops(G) are measured with an input size of (1, 3, 200, 500).

To further validate the effectiveness and universality of our method, we select the latest SOTA method, RCGD (Hu et al. 2023), as a baseline to construct MSD-RCGD. As shown in Table 1, MSD-RCGD outperforms this baseline and achieves new SOTA results on both EDU-CHEMC and Mini-CASIA-CSDB. The qualitative comparison between MSD-RCGD and RCGD is illustrated in Figure 4. These results demonstrate that our method can be applied to various mainstream end-to-end approaches, significantly enhancing recognition performance.

To explore the efficiency of our proposed method, we evaluate the parameters and FLOPs of MSD-DenseWAP and MSD-RCGD. The additional parameters and computational cost primarily arise from the branch classification module. As shown in Table 1, the extra cost of our method is marginal. This demonstrates that our approach maintains efficiency while providing enhanced performance.

5.5 Ablation Study

To verify the effectiveness of our Ring-Free Language and Molecular Skeleton Decoder, we conduct ablation experiments through several designed systems in Table 2.

The effectiveness of MSD. The Molecular Skeleton Decoder (MSD) is designed to fully leverage the advantages of RFL. To validate its effectiveness, we devised the systems T1 and T3, as shown in Table 2. T3 utilizes MSD, whereas T1 does not, instead employing a standard string decoder. The experimental results show a significant improvement in the exact match rate when adopting MSD, demonstrating its effectiveness and necessity.

The effectiveness of [conn]. The branch prediction task includes two parts: predicting connected ring bonds and skeleton bonds. The [conn] token assigns these tasks to the Branch Classification Module and the Skeleton Generation Module, respectively. This approach decreases the classification errors. Therefore, we design systems T2 and T4, which incorporate the [conn] and achieve significant improvements compared to systems that do not use it. This is

because incorrect branch information can prevent the correct reconstruction of the molecule, even if the skeleton and rings are predicted accurately.

System	MSD	[conn]	EM	Struct-EM
T1	✗	✗	38.70	49.45
T2	✗	✓	44.02	55.77
T3	✓	✗	52.76	58.58
T4	✓	✓	64.96	73.15

Table 2: Ablation study on the EDU-CHEMC dataset, with all systems based on MSD-DenseWAP.

5.6 Generalization Analysis

To prove that RFL and MSD can simplify molecular structure recognition and enhance generalization ability, we design experiments on molecule complexity. The previous approach (Hu et al. 2023) defined complexity as the sum of atoms and bonds. However, for complex multi-ring structures, recognition difficulty is much higher compared to ring-free structures with the same number of atoms and bonds. Therefore, we redefine structural complexity as:

$$Complexity = n_{atom} + n_{bond} + 12 \times n_{ring} \quad (13)$$

where n_{atom} , n_{bond} , n_{ring} represent the number of atoms, bonds and rings, respectively. The coefficient 12 is chosen because the sum of atoms and bonds in the most common benzene rings is 12.

Due to the uneven distribution of structural complexities and the small proportion of complex samples in the Mini-CASIA-CSDB dataset, we created a new dataset from ChEMBL (Gaulton et al. 2017) to assess model generalization. The dataset is divided into five levels based on structural complexity, with each level containing a similar number of samples, as shown in Figure 5.

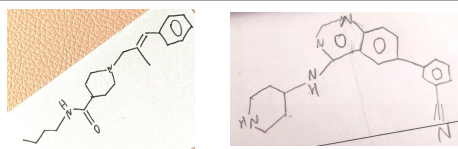
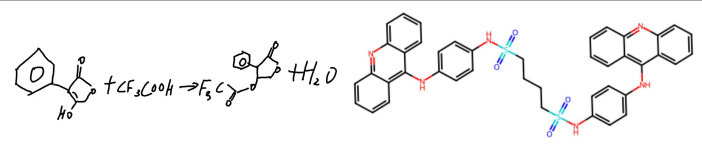
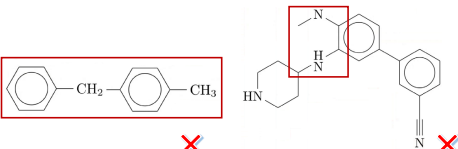
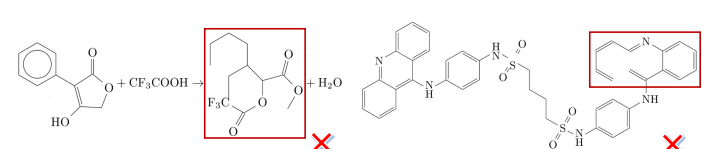
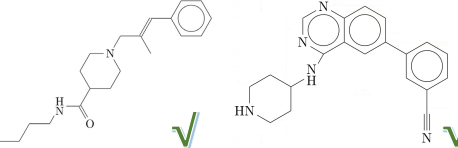
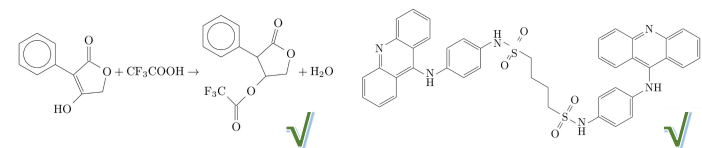
Input		
RCGD		
MSD-RCGD (Ours)		

Figure 4: A qualitative comparison of our proposed method with the SOTA method in the handwritten dataset EDU-CHEMC (a,b,c) and the printed dataset Mini-CASIA-CSDB (d). Compared to RCGD (Hu et al. 2023), our approach robustly recognizes molecular structures in challenging complex ring structures.

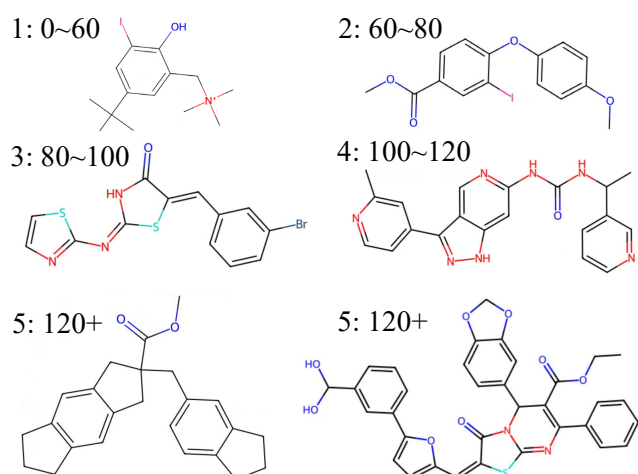


Figure 5: Illustration of chemical molecules with increased structural complexity. The number in the top left indicates the structural complexity level of each molecule, with the corresponding complexity range.

The experimental results are shown in Figure 6. DenseWAP-2 and MSD-DenseWAP-2 are trained with complexity levels {1,2}, while DenseWAP-3 and MSD-DenseWAP-3 use levels {1,2,3}. The test set includes all levels of structural complexity. It can be observed that when the structural complexity of chemical molecules in the test set exceeds the training set, DenseWAP-2 can hardly recognize them. However, MSD-DenseWAP-2 can still effectively recognize complex samples with previously unseen complexity. A similar trend is observed with MSD-DenseWAP-3, further confirming our conclusion.

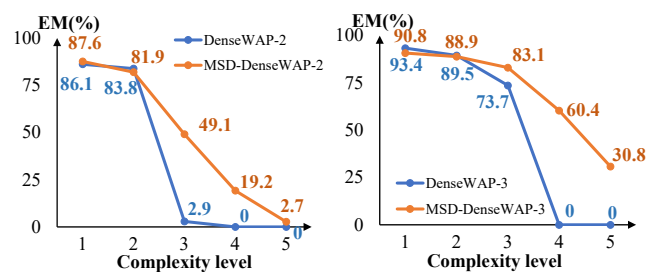


Figure 6: Exact match rate (in %) of DenseWAP and MSD-DenseWAP along test sets with different structural complexity. The left subplot is trained on complexity {1,2}, and the right subplot is trained on complexity {1,2,3}.

6 Conclusion and Future Work

In this study, we propose a novel chemical structure modeling language **Ring-Free Language (RFL)** to decouple complex ring structures in molecules. RFL decouples these structures into a molecular skeleton \mathcal{S} , individual rings \mathcal{R} , and branch information \mathcal{F} , thereby simplifying chemical structure prediction. Based on RFL, we propose a universal hierarchical Molecular Skeleton Decoder (MSD), which employs hierarchical decoding to progressively predict molecular skeleton \mathcal{S} and individual rings \mathcal{R} , followed by restoring the structure using branch information \mathcal{F} . To the best of our knowledge, this is the first end-to-end solution that decouples and models chemical structures in a structured form. In the future, we plan to explore the extension of structured-based modeling to tasks like tables, flowcharts, and diagrams. Our goal is to extend the RFL-based divide-and-conquer philosophy to a broader range of applications and develop a unified modeling framework.

References

- Bian, X.; Qin, B.; Xin, X.; Li, J.; Su, X.; and Wang, Y. 2022. Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 113–121.
- Bukhari, S. S.; Iftikhar, Z.; and Dengel, A. 2019. Chemical structure recognition (CSR) system: automatic analysis of 2D chemical structures in document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1262–1267. IEEE.
- Chen, M.; Wu, H.; Chang, Q.; Cheng, H.; Ma, J.; Hu, P.; Zhang, Z.; Liu, C.; Pi, C.; Hu, J.; et al. 2024. ICDAR 2024 Competition on Recognition of Chemical Structures. In *International Conference on Document Analysis and Recognition*, 397–409. Springer.
- Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Clevert, D.-A.; Le, T.; Winter, R.; and Montanari, F. 2021. Img2Mol—accurate SMILES recognition from molecular graphical depictions. *Chemical science*, 12(42): 14174–14181.
- Deng, Y.; Kanervisto, A.; and Rush, A. M. 2016. What you get is what you see: A visual markup decompiler. *arXiv preprint arXiv:1609.04938*, 10(32-37): 3.
- Ding, L.; Zhao, M.; Yin, F.; Zeng, S.; and Liu, C.-L. 2022. A Large-Scale Database for Chemical Structure Recognition and Preliminary Evaluation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1464–1470. IEEE.
- Filippov, I. V.; and Nicklaus, M. C. 2009. Optical structure recognition software to recover chemical information: OSRA, an open source solution.
- Fujiyoshi, A.; Nakagawa, K.; and Suzuki, M. 2011. Robust method of segmentation and recognition of chemical structure images in cheminfy. In *Pre-proceedings of the 9th IAPR international workshop on graphics recognition, GREC*, volume 1.
- Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. 2017. The ChEMBL database in 2017. *Nucleic acids research*, 45(D1): D945–D954.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; and Tchekhovskoi, D. 2015. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics*, 7: 1–34.
- Hu, J.; Wu, H.; Chen, M.; Liu, C.; Wu, J.; Yin, S.; Yin, B.; Yin, B.; Liu, C.; Du, J.; et al. 2023. Handwritten Chemical Structure Image to Structure-Specific Markup Using Random Conditional Guided Decoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8114–8124.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Jiang, Y.; Jin, S.; Jin, X.; Xiao, X.; Wu, W.; Liu, X.; Zhang, Q.; Zeng, X.; Yang, G.; and Niu, Z. 2023. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Communications Chemistry*, 6(1): 60.
- Khokhlov, I.; Krasnov, L.; Fedorov, M. V.; and Sosnin, S. 2022. Image2SMILES: Transformer-based molecular optical recognition engine. *Chemistry-Methods*, 2(1): e202100069.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Landrum, G. 2013. Rdkit documentation. *Release*, 1(1-79): 4.
- Morin, L.; Danelljan, M.; Agea, M. I.; Nassar, A.; Weber, V.; Meijer, I.; Staar, P.; and Yu, F. 2023. MolGrapher: Graph-based Visual Recognition of Chemical Structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19552–19561.
- Oldenhof, M.; Arany, A.; Moreau, Y.; and Simm, J. 2020. ChemGrapher: optical graph recognition of chemical compounds by deep learning. *Journal of chemical information and modeling*, 60(10): 4506–4517.
- Oldenhof, M.; De Brouwer, E.; Arany, A.; and Moreau, Y. 2024. Atom-Level Optical Chemical Structure Recognition with Limited Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17669–17678.
- Ouyang, T. Y.; and Davis, R. 2011. Chemink: a natural real-time recognition system for chemical drawings. In *Proceedings of the 16th international conference on Intelligent user interfaces*, 267–276.
- Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; and Barzilay, R. 2023. MolScribe: robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7): 1925–1934.
- Rajan, K.; Brinkhaus, H. O.; Zielesny, A.; and Steinbeck, C. 2020. A review of optical chemical structure recognition tools. *Journal of Cheminformatics*, 12: 1–13.
- Rajan, K.; Zielesny, A.; and Steinbeck, C. 2020. DECIMER: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1): 65.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III*, 234–241. Springer.

- Sadawi, N. M.; Sexton, A. P.; and Sorge, V. 2012. Chemical structure recognition: a rule-based approach. In *Document Recognition and Retrieval XIX*, volume 8297, 101–109. SPIE.
- Smolov, V.; Zentsev, F.; and Rybalkin, M. 2011. Imago: Open-Source Toolkit for 2D Chemical Structure Image Recognition. In *TREC*.
- Staker, J.; Marshall, K.; Abel, R.; and McQuaw, C. M. 2019. Molecular structure extraction from documents using deep learning. *Journal of chemical information and modeling*, 59(3): 1017–1029.
- Sundaramoorthy, C.; Kelvin, L. Z.; Sarin, M.; and Gupta, S. 2021. End-to-end attention-based image captioning. *arXiv preprint arXiv:2104.14721*.
- Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; and Li, H. 2016. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.
- Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1): 31–36.
- Weir, H.; Thompson, K.; Woodward, A.; Choi, B.; Braun, A.; and Martínez, T. J. 2021. ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning. *Chemical science*, 12(31): 10622–10633.
- Xu, Y.; Xiao, J.; Chou, C.-H.; Zhang, J.; Zhu, J.; Hu, Q.; Li, H.; Han, N.; Liu, B.; Zhang, S.; et al. 2022a. MolMiner: you only look once for chemical structure recognition. *Journal of Chemical Information and Modeling*, 62(22): 5321–5328.
- Xu, Z.; Li, J.; Yang, Z.; Li, S.; and Li, H. 2022b. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *Journal of Cheminformatics*, 14(1): 41.
- Yao, J.; Lai, Y.; Kou, H.; Wu, T.; and Liu, R. 2024. QE-BEV: Query evolution for bird’s eye view object detection in varied contexts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2927–2935.
- Yao, J.; Li, C.; and Xiao, C. 2024. Swift Sampler: Efficient Learning of Sampler by 10 Parameters. *arXiv preprint arXiv:2410.05578*.
- Yu, Z.; and Gao, H. 2022. Molecular representation learning via heterogeneous motif graph neural networks. In *International Conference on Machine Learning*, 25581–25594. PMLR.
- Zang, X.; Zhao, X.; and Tang, B. 2023. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1): 34.
- Zhang, J.; Du, J.; and Dai, L. 2018. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In *2018 24th international conference on pattern recognition (ICPR)*, 2245–2250. IEEE.
- Zhang, J.; Du, J.; Yang, Y.; Song, Y.-Z.; Wei, S.; and Dai, L. 2020. A tree-structured decoder for image-to-markup generation. In *International Conference on Machine Learning*, 11076–11085. PMLR.
- Zhang, X.-C.; Yi, J.-C.; Yang, G.-P.; Wu, C.-K.; Hou, T.-J.; and Cao, D.-S. 2022. ABC-Net: a divide-and-conquer based deep learning architecture for SMILES recognition from molecular images. *Briefings in bioinformatics*, 23(2): bbac033.
- Zhao, W.; and Gao, L. 2022. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *European conference on computer vision*, 392–408. Springer.
- Zhao, W.; Gao, L.; Yan, Z.; Peng, S.; Du, L.; and Zhang, Z. 2021. Handwritten mathematical expression recognition with bidirectionally trained transformer. In *Document analysis and recognition-ICDAR 2021: 16th international conference, Lausanne, Switzerland, September 5–10, 2021, proceedings, part II 16*, 570–584. Springer.