arXiv:2411.11098v4 [cs.CV] 19 Sep 2025

# MolParser: End-to-end Visual Recognition of Molecule Structures in the Wild

Xi Fang Jiankun Wang Xiaochen Cai Shangqian Chen Shuwen Yang
Haoyi Tao Nan Wang Lin Yao Linfeng Zhang Guolin Ke
DP Technology
{fangxi, wangjiankun, caixiaochen, chenshangqian ,yangsw,
taohaoyi, wangnan, yaol, zhanglf, kegl}@dp.tech

## Abstract

*In recent decades, chemistry publications and patents have increased rapidly. A significant portion of key information is embedded in molecular structure figures, complicating large-scale literature searches and limiting the application of large language models in fields such as biology, chemistry, and pharmaceuticals. The automatic extraction of precise chemical structures is of critical importance. However, the presence of numerous Markush structures in real-world documents, along with variations in molecular image quality, drawing styles, and noise, significantly limits the performance of existing optical chemical structure recognition (OCSR) methods. We present MolParser, a novel end-to-end OCSR method that efficiently and accurately recognizes chemical structures from real-world documents, including difficult Markush structure. We use an extended SMILES encoding rule to annotate our training dataset. Under this rule, we build MolParser-7M, a large-scale OCSR dataset based on our E-SMILES representation. While utilizing a large amount of synthetic data, we employed active learning methods to incorporate substantial in-the-wild data, specifically samples cropped from real patents and scientific literature, into the training process. We trained an end-to-end molecular image captioning model, MolParser, using a curriculum learning approach. MolParser significantly outperforms classical and learning-based methods across most scenarios, with potential for broader downstream applications. The dataset is publicly available in huggingface.*

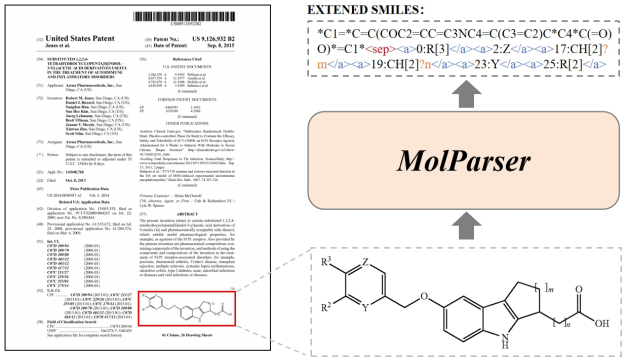



Figure 1. **MolParser** uses an end-to-end transformer to extract the chemical structures to string expression from the real patent or literature. We extend the SMILES format to enable the representation of more complex molecular structures including Markush.

## 1. Introduction

A significant portion of chemical information remains locked in unstructured formats within printed or digital documents, such as patents and scientific papers. In many of these documents, particularly in the field of chemistry, molecular structures are presented as images. These graphical depictions are essential for drug discovery, patent analysis, and chemical information retrieval. However, extracting them into machine-readable text remains a significant challenge. The automated interpretation of molecular structures from document images, a task known as Optical Chemical Structure Recognition (OCSR), is therefore of growing importance. With the rise of large language models (LLMs), increasing efforts are being directed toward applying them to the understanding of scientific literature. Converting molecular structure images into structured, interpretable text not only advances OCSR but also enables LLMs to more effectively process patents and scientific documents in chemistry-related domains.

OCSR aims to automatically convert chemical structure diagrams from scientific literature, patents, and other scanned documents into machine-readable string such as SMILES [57] representation. SMILES, though widely used for molecular representation, has notable limitations in handling complex chemical entities. It struggles with representing Markush structures, which are used in patents to describe a broad class of molecules by allowing variability at certain positions, enabling the protection of entire

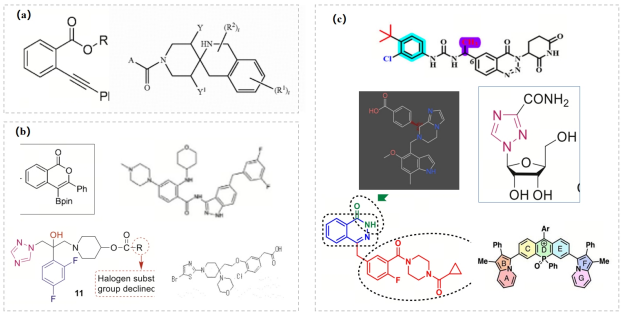



Figure 2. **In-the-wild problem in OCSR.** In real world patent and literature, we utilize object detection to locate and extract molecular images. However, there are several challenging cases that need to be addressed. These include (a) abbreviations and Markush structures, (b) image noise, blur and interference from surrounding elements, and (c) various drawing styles.

families of compounds. Additionally, SMILES cannot effectively handle connection points, abstract rings, ring attachments with uncertainty position, duplicated structures or polymers, all of which require a level of flexibility that its linear format does not support. Furthermore, SMILES is not well-suited for tasks involving large models, such as Markush-molecule matching, as its structure lacks the clarity and hierarchical organization needed for efficient interpretation by machine learning models. These limitations hinder its utility in advanced cheminformatics applications.

The complexity of OCSR arises from the intricate nature of chemical diagrams, which not only include atoms and bonds but also various annotations, nested structures, and ring connections. These elements make traditional Optical Character Recognition (OCR) techniques inadequate for this task. Another challenge of the OCSR task lies in the varying styles and visual representations of chemical diagrams, which can differ in terms of drawing styles, colors, and formatting. These variations complicate the extraction of molecular structures. Additionally, during the recognition process, non-chemical elements such as surrounding text or other images from the paper may be mistakenly captured, further hindering the accurate identification of the chemical structures. These issues, combined with the noise, distortions, and font variations often found in document images, make the task even more complex.

Early OCSR approaches follow a graph reconstruction paradigm, where molecular structures were rebuilt by identifying key components such as atoms, bonds, and charges. These were typically extracted using hand-crafted image processing techniques, with rules applied to connect the elements and form a graph representation. While some recent methods have introduced deep learning for atom and bond detection, they still rely on hand-crafted rules [12, 40, 51, 62, 69] for linking the recognized components. There are also some new methods that utilize Graph Neural Networks (GNNs) or transformers as replacements for traditional rule-based approaches [6, 38]. Constructing molecular structures uses graph representations and ultimately derives the string representation of these molecules (e.g. SMILES). This multi-step process makes training and model fine-tuning relatively complex and limits the robustness of these methods when dealing with noise and distortions commonly found in real-world documents and patents. As a result, their performance in handling noisy data remains suboptimal.

With the recent advancements in deep learning, end-to-end neural network models have become the dominant approach for OCSR. These methods combine image recognition with sequence generation tasks, directly converting input images into string molecular representations [7, 48, 63]. This method can be seen as a special case of image caption. However, despite significant progress, current models face real-world challenges, particularly when processing complex, noisy, or previously unseen chemical structures in patent documents. An important reason for this is that the training data utilized by these methods is often small-scale synthetic data, which differs significantly from real-world data scenarios. This highlights the need for developing more robust and versatile OCSR systems capable of handling greater diversity and complexity in real-world applications.

In order to better address the challenges of the OCSR task in real-world literature. We introduce a new end-to-end framework named MolParser for Optical Chemical Structure Recognition in the wild, illustrated in Figure 1 The main contributions of this paper are:

**Extended SMILES** We extend the SMILES representation to accommodate a broader range of specialized molecules commonly found in patents and literature, including Markush structures, connection points, abstract rings, ring attachments with uncertainty position, duplicated structures, and polymer structures. Additionally, this extended SMILES format is compatible with RDKit [25] and is also LLM-friendly, making it convenient for using LLMs to perform various analyses and processing on molecules.

**MolParser-7M Datasets** Based on our extended SMILES representation, we construct MolParser-7M , the largest annotated molecular recognition training dataset to our knowledge, with over 7 million paired image-SMILES data. MolParser-7M contains a large amount of diverse synthetic data, as well as in-the-wild data (cropped from real-world PDF scans). Additionally, we design a human-in-the-loop data engine to extract the most training-relevant molecular images from millions of patents and scientific papers, followed by meticulous manual annotation and cross review. We also provide a new OCSR benchmark, Wild-

Mol, including 10k ordinary molecules (WildMol-10k) and 10k Markush structures (WildMol-10k-M). All the samples are annotated by our extended SMILES (E-SMILES) format.

**MolParser Model** We regard OCSR tasks as a special type of image captioning task, where the content of the caption is an extended SMILES string. We develop MolParser model using an end-to-end image caption architecture, which includes a vision encoder, a feature compressor, and a BART [27] decoder to generate extended SMILES strings. We employed curriculum learning to train the MolParser model, first pretraining it on the diverse synthetic data of the MolParser7M dataset, gradually increasing the intensity of data augmentation during training. Afterward, we fine-tuned the model on a subset containing 400k in-the-wild real data. As a result, on the WildMol-10k benchmark, MolParser achieved a state-of-the-art accuracy of 76.9%, significantly outperforming existing methods, with MolScribe [45] at 66.4% and MolGrapher [38] at 45.5%. Additionally, with an inference speed of up to 40 FPS (131 FPS for the tiny version), MolParser is better suited for industrial applications compared to existing methods.

## 2. Related Works

Related work includes various molecular representation methods, such as SMILES, as well as algorithms for Optical Chemical Structure Recognition (OCSR).

**SMILES variants.** The Simplified Molecular Input Line Entry System (SMILES) provides a highly compact, linear string representation of molecular structures by encoding atoms and bonds efficiently. Its conciseness and simplicity have made SMILES a widely adopted standard in cheminformatics for molecular storage, retrieval, and similarity assessments. SMILES notation represents molecules, but cannot depict molecular templates like Markush structures. FG-SMILES suggested in Image2SMILES [23] attempts to solve this problem. This is an extension of standard SMILES, where a substituent or R-group can be written as a single pseudo-atom. However, this approach has limited scalability, as it struggles to support abstract rings, ring attachments with uncertainty position, duplicated structures, and polymer structures. At the same time, it is difficult to ensure compatibility with the current leading molecular processing tool, RDKit [25], which complicates subsequent processing and analysis.

**Image captioning based OCSR (End-to-End).** Most recent end-to-end deep learning approaches leverage image captioning techniques, which involve generating descriptive textual representations of images. These models employ an encoder to extract visual features from images and a decoder to convert them into SMILES [57] or InChI [17] sequences. Specifically, models like MSE-DUDL [52], DECIMER [46], Img2Mol [7], ChemPix [58], and MICER [65] utilize a convolutional encoder paired with various recurrent decoders (RNN, GRU or LSTM). Subsequent works have introduced transformer-based encoder-decoder architectures, such as DECIMER 1.0 [47], DECIMER 2.0 [48], SwinOCSR [63], IMG2SMI [5], Image2SMILES [23] and Image2InChI[28]. The advantage of these algorithms lies in their fast end-to-end speed and strong generalization performance. But a significant drawback of these image captioning methods is their requirement for large training datasets. Most of these methods rely on generated data and do not achieve satisfactory performance in in-the-wild scenarios such as patents or literature.

**Graph-based OCSR (Atom-Bond).** Traditional OCSR methods rely on hand-crafted image processing rule to detect molecular components and reconstruct the molecular graph [4, 12, 37, 41, 42, 50, 51]. Recent approaches utilize deep learning for component detection or segmentation [40, 62, 69]. While more recent utilize deep learning to build the graphs instead of using hand-crafted rule [44, 66]. However, this approach is complex and computationally slow, and its complexity makes extensive manual labeling nearly impossible, resulting in a heavy reliance on generated data. This reliance, in turn, makes it vulnerable to noise interference in real-world applications and contributes to lower generalization performance. Even though the latest MolGrapher [38] achieves state-of-the-art results on several benchmarks, it still encounters challenges in real-world literature scanning scenarios. Similar methods include MMSSC-Net [68] and MolScribe [45].

**Markush Recognition.** Markush structures are chemical representations that use variable groups to describe a family of related compounds. They are commonly found in patents to claim broad molecular classes. These structures often contain R-groups, repeating units, and variable attachments. Markush recognition [1, 16, 39, 55] remains a major challenge. Moreover, existing tools such as RDKit [25] cannot parse Markush structures. They do not support inputs with undefined groups or variable atoms. These tools also fail to render valid images of Markush structures. As a result, generating large-scale training data becomes more difficult. The lack of standardized representation and visualization limits data augmentation. This creates a bottleneck for training robust models on Markush recognition.

## 3. MolParser

### 3.1. Extended SMILES

We extend the SMILES representation method, abbreviated as E-SMILES, to more effectively represent the Markush structures commonly found in patents, as well as complex compositions such as connection points, abstract rings, ring attachments with uncertainty position, duplicated structures, and polymer structures. Additionally, we ensure

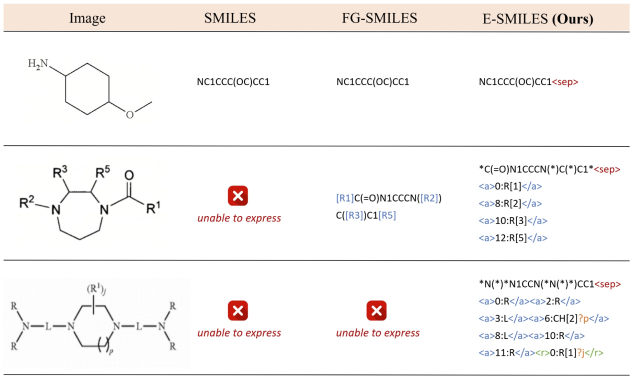



Figure 3. **Comparison of molecular expressions.** Our extended SMILES is able to express more complex Markush structures.

that this approach is compatible with RDKit and LLM-friendly, facilitating subsequent analysis and processing tasks. The extended SMILES format will be denoted as: `SMILES<sep>EXTENSION`. Where `SMILES` refers to the RDKit-compatible SMILES representation. Special token `<sep>` serves as the separator, and optional `EXTENSION` represents supplementary descriptive used to handle complex cases such as Markush structures.

In `EXTENSION`. We use some XML-like special tokens to represent certain special functional groups. For Markush R-groups and abbreviation groups, we add special tokens `<a>` and `</a>` to encapsulate descriptions of these special substituents. Similarly, we use special tokens `<r>` and `</r>` for ring attachments with uncertainty position; `<c>` and `</c>` for abstract rings. Additionally, there is a special token `<dum>` representing a connection point. For the specific description of each functional group, we use the following format: `[INDEX]:[GROUP_NAME]`. Figure 3 shows a example of our extended SMILES. Although there are numerous complex Markush structures in actual patents, our extended SMILES (E-SMILES) rules can still adequately address these cases. For more details, please refer to the supplementary materials 9.

### 3.2. Architecture

Motivated by image captioning approach [3, 26, 29, 63], we employ a single transformer only architecture to translate molecular structure images into our extended SMILES format. Our MolParser model has three components, an image encoder, a feature compresser, and a SMILES decoder.

We use an ImageNet [8] pretrained Swin-Transformer [33] as the image encoder for our MolParser. Similar to LLaVA [30], we employ a two-layer MLP as a vision-language connector to compress the feature dimension. Finally, we adopt a decoder-only Transformer architecture, BART-Decoder [27], to decode the compressed image features and autoregressively generate the extended SMILES (E-SMILES) sequence via next-token prediction.

### 3.3. Training

We employ autoregressive algorithm to generate the E-SMILES sequence. For paired training data, we first use a tokenizer to convert the E-SMILES sequence into a token sequence. Each token generally represents an atom, an abbreviated group name, a number, a special symbol, or a special token defined in E-SMILES. The training process is conducted in two stages: pretraining stage and supervised finetuning stage.

In the first stage, we conduct pretraining using a synthetic dataset comprising over 7.7 million paired training samples. Concurrently, we employ a curriculum learning [2] approach during pretraining to achieve better convergence. In the early phase of pretraining, we refrain from utilizing data augmentation and restrict our focus to simple molecules with a SMILES token count of fewer than 60. Subsequently, we progressively increase the intensity of data augmentation and incorporate molecules with longer token sequences into the training process. The detail of training dataset is described in Section 5.

In the second stage, we fine-tune the model with a set of human-annotated in-the-wild data. The purpose of this stage is to further enhance the generalization ability of MolParser in the real scene. Previous methods [7, 38, 48] only use synthetic data, which exhibited a relatively restricted distribution in images and consisted of relatively simple molecular structures, resulting in suboptimal performance in real-world applications. In contrast, we employ an active learning approach to construct a data engine that extracted approximately 400,000 molecular images deemed most valuable for model learning from over 1.22 million real patents and academic papers, supplemented by manual annotation and secondary review. After incorporating these data for fine-tuning, we further improve MolParser's ability in real application. The detail of our active leanring data engine is described in Section 4.

## 4. MolParser Data Engine

As paired OCSR training data is not abundant on the Internet, we build a data engine to enable the collection of our 7.7 million OCSR paired dataset, MolParser-7M, with low cost. The data engine has two stages: (1) start up phase with fully synthetic training data, (2) active learning with human in the loop. Here is the detail:

**Start up with synthetic data.** Due to the significant difficulty in annotating molecular data in SMILES format, the cost of manual annotation in large volumes is quite high. Our analysis shows that, on average, it takes over three minutes for an expert to annotate a single molecule from scratch. Therefore, we chose to initiate the data engine with synthetic data and gradually expand it to real data. We uti-

lized a diverse range of molecular structure sources and various image rendering methods, generating over 7 million pairs of images and extended-SMILES. Details of the synthetic data can be referenced in Section 5. Then we can use synthetic data to train the initial version of our MolParser model.

**Active Learning with Human in the Loop.** To further enhance the generalization performance of our model, we extract molecular training data directly from patents and scientific literature in PDF format and manually annotate the molecular structures. Initially, we train a YOLO11 [22] object detection model, named as MolDet, to locate molecules within these documents. We use 1.22 million real PDF files, including patent documents from various international patent offices and open-access papers from Internet such as bioRxiv, medRxiv, and ChemRxiv. From this dataset, we extract over 20 million molecular images. After employing image p-hash similarity analysis to remove duplicate or highly similar molecular images, this number reduces to 4 million. Due to the sheer volume of this dataset, manual annotation of all images is not feasible. Therefore, we introduce active learning algorithms to identify and select samples with higher importance.

We perform 5-fold training on synthetic data to obtain five distinct models. Each model independently generates predictions, resulting in five E-SMILES strings per molecule image. We extract the standard SMILES sequences before the <sep> token and compute the pairwise Tanimoto similarity, taking the average similarity score as the confidence score for OCSR prediction. We observe that data with low confidence scores often correspond to images with poor quality, whereas data with high confidence scores typically indicate a high probability of correct predictions. We then randomly select samples with confidence scores between 0.6 and 0.9, as we believe these samples present significant recognition challenges and are highly beneficial for training. We manually annotate the molecular structures for these selected samples.

During the annotation process, we use model predictions as pre-annotations, which annotators modify as needed, without editing molecular structures from scratch. Two different annotators independently review the model predictions or manual modifications to verify their correctness. Analysis shows that leveraging pre-annotations reduces annotation time per molecule to 30 seconds, achieving approximately 90% savings in manual labor compared to annotation from scratch.

After every 80,000 completed annotations, we incorporate the data into the training set, update the 5-fold models, and repeat the active learning cycle. This iterative process enhances both the model's generalization ability and the quality of the pre-annotated data. Through this loop, we construct a dataset of 400,000 manually annotated images.

| subset | ratio | source |
|---|---|---|
| Markush-3M | 40% | random groups replacement from PubChem [24] |
| ChEMBL-2M | 27% | molecules selected from ChEMBL [15] |
| Polymer-1M | 14% | random generated polymer molecules |
| PAH-600k | 8% | random generated fused-ring molecules |
| BMS-360k | 5% | molecules with long carbon chains from BMS [19] |
| MolGrapher-300K | 4% | training data from paper MolGrapher [38] |
| Pauling-100k | 2% | Pauling-style images drawn using epam.indigo [9] |

(a) Datasets used in pretraining stage.

| subset | ratio | source |
|---|---|---|
| MolParser-SFT-400k | 66% | manually annotated data obtained using data engine |
| MolParser-Gen-200k | 32% | synthetic data selected from pretraining stage |
| Handwrite-5k | 1% | handwritten modelcules selected from Img2Mol [7] |

(b) Datasets used in fine-tuning stage.

Table 1. **Summary of datasets used in MolParser training.** To construct MolParser-7M dataset, we use a very wide range of data sources.

In the final dataset, 56.04% of annotations directly use model pre-annotations, 20.97% pass review after a single manual correction, 13.87% are accepted after a second round of annotation, and 9.13% require three or more rounds of annotation.

## 5. MolParser-7M Dataset

Our dataset, MolParser-7M, consists of 7.7M diverse molecule structure images. As far as we know, MolParser contains the largest number of paired samples in open-sourced OCSR datasets and it is the only open-source training set that includes a significant amount of real molecule images croped from real patent and literature. The largest open-access OCSR paired dataset available before this work was MolGrapher-300K [38], which included only 300k paired samples, all of which are synthetic data.

**Synthetic Training Data Generation.** To launch our data engine, we first generated approximately 7M of paired OCSR training data. To obtain a more diverse distribution of data, we collected a substantial number of molecular structures from various sources and additionally generated a significant number of Markush structures at random. The sources from which we obtained molecular structure data include: ChEMBL [15] database, PubChem [24] database, Kaggle BMS [19] dataset. Training images are then generated from SMILES using the molecule drawing library RDKit [25] and epam.indigo [9]. Similar to previous work MolGrapher [38], in order to increase image diversity, rendering parameters are also randomly set. The specific data sources are listed in the Table 1a.

**Fine-tuning dataset construction.** We obtained approximately 400,000 manually annotated training data from the active learning data engine. In addition, we found that it is necessary to keep a part of synthetic data in the fine-tuning stage. To support handwritten molecule recogni-

tion, we also add some manually annotated handwritten molecules. The specific composition of fine-tuning data can be referred to Table 1b.

To assess the performance of the OCSR model in in-the-wild scenarios, we have also released an open-source OCSR test set, WildMol, comprising 20,000 human annotated molecule samples cropped from real PDF files. It presents greater difficulty and features an in-the-wild distribution compared to other open-source evaluation benchmarks.

## 6. Experiments

In this section, we conduct extensive experiments on various OCSR benchmarks. Moreover, we demonstrate the application of our MolParser method in downstream tasks and reveal an intriguing finding: the image encoder of our MolParser model shows promising utility in the field of molecular property prediction.

### 6.1. Evaluation datasets and metrics

To compare our method with previous state-of-the-art approaches, we evaluate the model on several classic publicly available benchmarks, including USPTO [12], Maybridge UoB [50], CLEF-2012 [43], and JPO [13]. However, these classic OCSR evaluation datasets are limited in size and contain systematic biases and annotation noise. To further assess the performance of our model, we conduct tests on a small but challenging in-the-wild OCSR dataset, ColoredBG [61], as well as a larger-scale dataset, USPTO-10k [38], containing 10,000 molecular images. Additionally, we evaluate model performance on our proposed WildMol dataset to comprehensively test the OCSR algorithm's performance in in-the-wild literature scenarios. WildMol-10K contains 10,000 regular molecules, and WildMol-10K-M contains 10,000 Markush structures. For evaluation metric, we use the classic accuracy metric, which is commonly applied in such tasks.

### 6.2. State-of-the-art comparison

Table 2 compares OCSR methods across different benchmarks, where our MolParser method consistently outperforms existing approaches, including the previous state-of-the-art, MolGrapher [38] and MolScribe [45]. On classical benchmarks such as USPTO, Maybridge UoB, JPO, ColoredBG and USPTO-10K, MolParser achieves satisfactory results. On our newly proposed, significantly more challenging test set, WildMol-10K, which consists of molecule images cropped from real patent literature, MolParser also demonstrates substantial improvements, confirming its ability to handle diverse molecular image data from various document sources.

We built a series of MolParser models with various sizes by using visual backbones of different scales, input resolutions, and BART decoders with varying parameter counts. As shown in Table 3, our model demonstrates clear advantages in both speed and accuracy compared to previous state-of-the-art model MolGrapher, achieving a significantly better Pareto frontier. The throughout is tested in RTX-4090D. The Tiny variant of MolParser achieves a parsing speed of over 130 molecular images per second with minimal accuracy loss, enabling rapid extraction and parsing of molecular structures in ultra-large-scale unstructured documents. The Base variant of MolParser achieves the highest accuracy with a recognition speed of 40 molecular images per second. Due to its end-to-end design that avoids complex preprocessing, postprocessing, and multiple inference stages, its speed significantly outperforms non-end-to-end algorithms. In comparison with existing methods, our MolParser achieves a significantly better speed-accuracy Pareto curve.

In our study, we also conducted a qualitative evaluation of MolParser and found it to be highly robust against noise present in various real-world data. It demonstrated strong parsing capabilities for Markush structures and performed well on many complex molecules—cases that have been challenging for previous methods to address. For more details, please refer to the supplementary materials.

### 6.3. Ablation study

**The importance of large scale training data.** Before our MolParser-7M, the largest paired molecule recognition open-source dataset was MolGrapher-300k [38], which included 300k artificially generated molecular images. We used the same model architecture and training methods of MolParser. As shown in Table 4, When we switched our pre-trained dataset to the significantly smaller MolGrapher, there was a noticeable drop in performance. It demonstrates the essential importance of scaling training data for end-to-end OCSR models.

**The importance of fine-tuning in real data.** As shown in Table 4. We demonstrate the effectiveness of our data engine in this study. We compared the performance of MolParser before and after fine-tuning with data obtained from the data engine. We find that training our end-to-end MolParser models solely on synthetic dataset do not yield satisfactory results across various benchmarks. The reason is that the data distribution of the benchmark differs significantly in style from the molecular images generated by RDKit. However, after incorporating real data from our data engine, a significant performance improvement is achieved. demonstrating that the real-world and in-the-wild data extracted through our active learning algorithm is essential.

**The impact of model scale.** In our study, we experiment with varying input resolutions of image sizes, the quantity of parameters in visual backbones, and the parameter count in transformer decoders. As shown in Table 3, it demon-

| Method | USPTO (5719) | UoB (5740) | CLEF (992) | JPO (450) | ColoredBG[61] (200) | USPTO-10K[38] (10000) | **WildMol-10K** (10000) |
|---|---|---|---|---|---|---|---|
| *Rule-based methods* | | | | | | | |
| OSRA 2.1 [12] * | 89.3 | 86.3 | **93.4** | 56.3 | 5.5 | 89.7 | 26.3 |
| MolVec 0.9.7 [42] * | 91.6 | 79.7 | 81.2 | 66.8 | 8.0 | 92.4 | 26.4 |
| Imago 2.0 [51] * | 89.4 | 63.9 | 68.2 | 41.0 | 2.0 | 89.9 | 6.9 |
| *Only synthetic training* | | | | | | | |
| Img2Mol [7] * | 30.0 | 68.1 | 17.9 | 16.1 | 3.5 | 33.7 | 24.4 |
| MolGrapher [38] †* | 91.5 | **94.9** | 90.5 | 67.5 | 7.5 | 93.3 | 45.5 |
| *Real data finetuning* | | | | | | | |
| DECIMER 2.7 [48] * | 59.9 | 88.3 | 72.0 | 64.0 | 14.5 | 82.4 | 56.0 |
| MolScribe [45] * | **93.1** | 87.4 | 88.9 | <u>76.2</u> | 21.0 | **96.0** | 66.4 |
| **MolParser-Tiny** (Ours) | <u>93.0</u> | 91.6 | <u>91.0</u> | 75.6 | **58.5** | 89.5 | 73.1 |
| **MolParser-Small** (Ours) | **93.1** | 91.1 | 90.8 | <u>76.2</u> | <u>57.0</u> | <u>94.8</u> | <u>76.3</u> |
| **MolParser-Base** (Ours) | <u>93.0</u> | <u>91.8</u> | 90.7 | **78.9** | <u>57.0</u> | 94.5 | **76.9** |

Table 2. **Comparison of our method with existing OCSR models.** We report the accuracy. We use **bold** to indicate the best performance and <u>underline</u> to denote the second-best performance. ∗: re-implemented results. †: results from original publications.

| Method | Vision Backbone | Resolution | Param Count | Throughout ↑ | WildMol-10K ↑ (10000) | WildMol-10K-M ↑ (10000) |
|---|---|---|---|---|---|---|
| *Open-sourced implements* | | | | | | |
| Img2Mol [7] | 8-Layer-CNN | 224*224 | 201M | 0.38 | 24.4 | - |
| MolGrapher [38] | Res18 + Res50 | 1024*1024 | 40M | 2.2 | 45.5 | - |
| DECIMER2.7 [48] | EfcientNet-B3 | 299*299 | 12M | 0.14 | 56.0 | - |
| MolScribe [45] | Swin-Base | 384*384 | 88M | 16.5 | 66.4 | - |
| *Our end-to-end Molparser* | | | | | | |
| MolParser-Tiny | Swin-Tiny | 224*224 | 66M | **131.6** | 73.1 | 15.3 |
| MolParser-Small | Swin-Small | 224*224 | 108M | 116.3 | 76.3 | 34.8 |
| **MolParser-Base** | Swin-Base | 384*384 | 216M | 39.8 | **76.9** | **38.1** |
| MolParser-InternVL | InternViT-300M [14] | 448*448 | 2200M | 1.5 | 72.9 | 33.7 |

Table 3. **Speed and accuracy evaluation in WildMol.** We report the throughout and accuracy in our WildMol benchmark. Throughput is measured on a single RTX 4090D GPU, and the time for preprocessing and postprocessing is also included in the calculation. Except for our MolParser model, existing models do not support the evaluation of extreme complex Markush data in WildMol-M Benchmark.

| Training Dataset | Fine-tuning | WildMol-10K ↑ |
|---|---|---|
| MolGrapher-300k | - | 22.4 |
| MolParser-7M (pt) | - | 51.9 |
| MolParser-7M (pt+ft) | - | 75.9 |
| MolParser-7M (pt) | MolParser-7M (ft) | **76.9** |

Table 4. **Ablation study in training and finetuning dataset.** We report the accuracy score in WildMol-10K benchmarks. 'pt' means the synthetic pretraining subset and 'ft' stand for fine-tuning subset, suggested in Section 5 .

strates that scaling model dimensions has a certain effect, yet it is less effective compared to scaling the dataset and fine-tuning in real data from our data engine. Additionally, we observe that employing excessively large end-to-end image caption models, such as Mini-InternVL [14], may render the training process more challenging.

**The impact of data augmentation.** We run controlled experiments. Table 5 reports the results. Data augmentation improves generalization on real scanned benchmarks. Curriculum learning helps further. It starts with weak augmentation and gradually increases strength during training.

| Data Augmentation | Curriculum Strategy | WildMol-10K ↑ |
|---|---|---|
| × | × | 40.1 |
| ✓ | × | 69.5 |
| ✓ | ✓ | **76.9** |

Table 5. **Ablation study in data augmentation and training strategy.** We report the accuracy score in WildMol-10K benchmarks.

### 6.4. Expanding applications: molecular property prediction

We make the unexpected observation that the image feature extractor of our MolParser, a Swin Transformer, can serve as an effective molecular fingerprint (or molecular embedding) for downstream molecular property prediction tasks

| **Method** | BBBP ↑ | Tox21 ↑ | ToxCast ↑ | BACE ↑ | SIDER ↑ | **Avg.** ↑ |
|---|---|---|---|---|---|---|
| *3D Conformation* | | | | | | |
| GEM[10] | 72.4 | 78.1 | - | 85.6 | 67.2 | - |
| 3D InfoMax[53] | 68.3 | 76.1 | 64.8 | 79.7 | 60.6 | 69.9 |
| GraphMVP[31] | 69.4 | 76.2 | 64.5 | 79.8 | 60.5 | 70.1 |
| MoleculeSDE[32] | 71.8 | 76.8 | 65.0 | 79.5 | **75.1** | 73.6 |
| Uni-Mol[70] | 71.5 | 78.9 | 69.1 | 83.2 | 57.7 | 72.1 |
| MoleBlend[67] | 73.0 | 77.8 | 66.1 | 83.7 | 64.9 | 73.1 |
| Mol-AE[64] | 72.0 | **80.0** | 69.6 | 84.1 | 67.0 | **74.5** |
| UniCorn[11] | **74.2** | 79.3 | 69.4 | **85.8** | 64.0 | **74.5** |
| *2D Graph* | | | | | | |
| AttrMask[20] | 65.0 | 74.8 | 62.9 | 79.7 | 61.2 | 68.7 |
| GROVER[49] | 70.0 | 74.3 | 65.4 | 82.6 | 64.8 | 71.4 |
| BGRL[54] | 72.7 | 75.8 | 65.1 | 74.7 | 60.4 | 69.7 |
| MolCLR[56] | 66.6 | 73.0 | 62.9 | 71.5 | 57.5 | 66.3 |
| GraphMAE[18] | 72.0 | 75.5 | 64.1 | 83.1 | 60.3 | 71.0 |
| Mole-BERT[60] | 71.9 | 76.8 | 64.3 | 80.8 | 62.8 | 71.3 |
| SimSGT[35] | 72.2 | 76.8 | 65.9 | 84.3 | 61.7 | 72.2 |
| MolCA + 2D[34] | 70.0 | 77.2 | 64.5 | 79.8 | 63.0 | - |
| *2D Image* | | | | | | |
| Swin-T (w/ ImageNet pretrained) | 62.5 | 77.9 | 67.4 | 76.0 | 60.5 | 68.9 |
| **Swin-T (w/ MolParser pretrained)** | 70.4 | 79.0 | **74.6** | 84.1 | 60.2 | 73.7 |

Table 6. **Comparison of molecular property prediction methods.** We report the average ROC-AUC scores after five runs.

after being trained on the MolParser-7M dataset. Specifically, for each molecule, we first use the RDKit toolkit [25] to render a 2D molecule structural image, then extract visual features using the MolParser vision backbone, and apply global average pooling to obtain a 2048-dimensional feature vector as the molecular representation. A simple two-layer MLP is then used to perform molecular property prediction.

We evaluate our approach alongside several baselines on five molecular property classification tasks from the MoleculeNet benchmark [59]. As shown in Table 6, our method achieves competitive performance in molecular property prediction. Notably, features extracted by MolParser, when paired with a lightweight MLP, perform on par with more complex models that rely on 2D or 3D graph-based representations. Moreover, our approach substantially outperforms other image-based feature extractors that are not pre-trained on the MolParser-7M dataset. These results indicate that 2D molecular structure images contain rich, chemically meaningful information. They also highlight the effectiveness of our end-to-end, large-scale OCSR training in learning high-quality visual representations for downstream chemical tasks.

### 6.5. Expanding applications: chemical reaction parsing

Following OmniParser [36], we input the molecular locations detected by our MolDet model and the E-SMILES sequences recognized by our MolParser model into GPT-4o [21] to enhance chemical reaction parsing. We draw bounding boxes around target molecules in chemical reaction images and label each with a corresponding molecule index, which is provided as input to GPT-4o. Additionally, we prepend the prompt with the E-SMILES recognized for each indexed molecule. This significantly enhances the ability of MLLMs like GPT-4o to process chemical reaction images. Refer to the appendix for details.

## 7. Conclusion

We propose a novel end-to-end OCSR algorithm that extends the SMILES representation and introduces a large-scale training dataset, MolParser-7M. The model is pre-trained on large scale synthesized data and fine-tuned on manually annotated in-the-wild samples. It outperforms existing methods on classical OCSR benchmarks and our newly introduced WildMol benchmark. The system achieves high processing speed and strong performance in extracting structured molecular information from real, unstructured scientific literature.

Despite its effectiveness, MolParser still has room for improvement. For instance, molecular chirality, which is closely related to chemical properties, is not yet fully exploited. In addition, scaling up the amount of real annotated training data may further boost performance. As future work, we aim to address the challenge of chirality in OCSR and scale up the volume of real-world training data. We also plan to use MolParser to extract molecules and Markush structures with their visual fingerprints from large-scale scientific literature and patents, enabling the creation of a comprehensive database for chemical information mining.

## References


[1] Edward J Beard and Jacqueline M Cole. Chemschematicresolver: a toolkit to decode 2d chemical diagrams with labels and r-groups into annotated chemical named entities. *Journal of chemical information and modeling*, 60(4):2059–2072, 2020. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 4

[3] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. 4

[4] Syed Saqib Bukhari, Zaryab Iftikhar, and Andreas Dengel. Chemical structure recognition (csr) system: automatic analysis of 2d chemical structures in document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1262–1267. IEEE, 2019. 3

[5] Daniel Campos and Heng Ji. Img2smi: translating molecular structure images to simplified molecular-input line-entry system. *arXiv preprint arXiv:2109.04202*, 2021. 3

[6] Yufan Chen, Ching Ting Leung, Yong Huang, Jianwei Sun, Hao Chen, and Hanyu Gao. Molnextr: A generalized deep learning model for molecular image recognition. *arXiv preprint arXiv:2403.03691*, 2024. 2

[7] Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2mol–accurate smiles recognition from molecular graphical depictions. *Chemical science*, 12(42): 14174–14181, 2021. 2, 3, 4, 5, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[9] epam. Indigo library. 5

[10] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022. 8

[11] Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. Unicorn: A unified contrastive learning approach for multi-view molecular representation learning. *arXiv preprint arXiv:2405.10343*, 2024. 8

[12] Igor V Filippov and Marc C Nicklaus. Optical structure recognition software to recover chemical information: Osra, an open source solution, 2009. 2, 3, 6, 7

[13] Akio Fujiyoshi, Koji Nakagawa, and Masakazu Suzuki. Robust method of segmentation and recognition of chemical structure images in cheminfty. In *Pre-proceedings of the 9th IAPR international workshop on graphics recognition, GREC*, 2011. 6

[14] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 7

[15] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012. 5

[16] Carina S Haupt. Markush structure reconstruction. In *Gesellschaft für Informatik (GI) publishes this series in order to make available to a broad public recent findings in informatics (ie computer science and informa-tion systems), to document conferences that are organized in co-operation with GI and to publish the annual GI Award dissertation.*, page 207, 2009. 3

[17] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5:1–9, 2013. 3

[18] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022. 8

[19] Addison Howard, Jacob Albrecht, and Yvette. Bristol-myers squibb – molecular translation., 2021. 5

[20] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019. 8

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8, 3

[22] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO11, 2024. 5, 3

[23] Ivan Khokhlov, Lev Krasnov, Maxim V Fedorov, and Sergey Sosnin. Image2smiles: Transformer-based molecular optical recognition engine. *Chemistry-Methods*, 2(1):e202100069, 2022. 3

[24] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1): D1102–D1109, 2019. 5

[25] Greg Landrum. Rdkit: Open-source cheminformatics software., 2023. 2, 3, 5, 8

[26] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 4

[27] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 3, 4

[28] Da-zhou Li, Xin Xu, Jia-heng Pan, Wei Gao, and Shi-rui Zhang. Image2inchi: Automated molecular optical image recognition. *Journal of Chemical Information and Modeling*, 64(9):3640–3649, 2024. 3

[29] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13094–13102, 2023. 4

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[31] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021. 8

[32] Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pages 21497–21526. PMLR, 2023. 8

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[34] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023. 8

[35] Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Rethinking tokenizer and decoder in masked graph modeling for molecules. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[36] Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. Omniparser for pure vision based gui agent. *arXiv preprint arXiv:2408.00203*, 2024. 8, 3

[37] Joe R McDaniel and Jason R Balmuth. Kekule: Ocr-optical chemical (structure) recognition. *Journal of chemical information and computer sciences*, 32(4):373–378, 1992. 3

[38] Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, Ingmar Meijer, Peter Staar, and Fisher Yu. Molgrapher: Graph-based visual recognition of chemical structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19552–19561, 2023. 2, 3, 4, 5, 6, 7

[39] Lucas Morin, Valéry Weber, Ahmed Nassar, Gerhard Ingmar Meijer, Luc Van Gool, Yawei Li, and Peter Staar. Markushgrapher: Joint visual and textual recognition of markush structures. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14505–14515, 2025. 3

[40] Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. Chemgrapher: optical graph recognition of chemical compounds by deep learning. *Journal of chemical information and modeling*, 60(10):4506–4517, 2020. 2, 3

[41] Tom Y Ouyang and Randall Davis. Chemink: a natural real-time recognition system for chemical drawings. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 267–276, 2011. 3

[42] Tyler Peryea, Daniel Katzel, Tongan Zhao, Noel Southall, and Dac-Trung Nguyen. Molvec v0.9.8, 2022. 3, 7

[43] Florina Piroi, Mihai Lupu, Allan Hanbury, Alan P Sexton, Walid Magdy, and Igor V Filippov. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *CLEF (notebook papers/labs/workshops)*, 2010. 6

[44] Yujie Qian, Zhengkai Tu, Jiang Guo, Connor W Coley, and Regina Barzilay. Robust molecular image recognition: A graph generation approach. Technical report, Technical Report, 2022. 3

[45] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. Molscribe: robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7):1925–1934, 2023. 3, 6, 7

[46] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):65, 2020. 3

[47] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13:1–16, 2021. 3

[48] Kohulan Rajan, Henning Otto Brinkhaus, M Isabel Agea, Achim Zielesny, and Christoph Steinbeck. Decimer. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature communications*, 14(1):5045, 2023. 2, 3, 4, 7

[49] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020. 8

[50] Noureddin M Sadawi, Alan P Sexton, and Volker Sorge. Chemical structure recognition: a rule-based approach. In *Document Recognition and Retrieval XIX*, pages 101–109. SPIE, 2012. 3, 6

[51] Viktor Smolov, Fedor Zentsev, and Mikhail Rybalkin. Imago: Open-source toolkit for 2d chemical structure image recognition. In *TREC*, 2011. 2, 3, 7

[52] Joshua Staker, Kyle Marshall, Robert Abel, and Carolyn M McQuaw. Molecular structure extraction from documents using deep learning. *Journal of chemical information and modeling*, 59(3):1017–1029, 2019. 3

[53] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pages 20479–20502. PMLR, 2022. 8

[54] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. *arXiv preprint arXiv:2102.06514*, 2021. 8

[55] Jie Wang, Zihao Shen, Yichen Liao, Zhen Yuan, Shiliang Li, Gaoqi He, Man Lan, Xuhong Qian, Kai Zhang, and Honglin Li. Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space. *Briefings in Bioinformatics*, 23(6), 2022. 3

[56] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022. 8

[57] David Weininger. 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, 1988. 1, 3

[58] Hayley Weir, Keiran Thompson, Amelia Woodward, Benjamin Choi, Augustin Braun, and Todd J Martínez. Chempix: automated recognition of hand-drawn hydrocarbon structures using deep learning. *Chemical science*, 12(31):10622–10633, 2021. 3

[59] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018. 8

[60] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Molebert: Rethinking pre-training graph neural networks for molecules. 2023. 8

[61] Jiacheng Xiong, Xiaohong Liu, Zhaojun Li, Hongzhong Xiao, Guangchao Wang, Zhenjiang Niu, Chaoyuan Fei, Feisheng Zhong, Gang Wang, Wei Zhang, et al. $\alpha$extractor: a system for automatic extraction of chemical information from biomedical literature. *Science China Life Sciences*, 67 (3):618–621, 2024. 6, 7

[62] Youjun Xu, Jinchuan Xiao, Chia-Han Chou, Jianhang Zhang, Jintao Zhu, Qiwan Hu, Hemin Li, Ningsheng Han, Bingyu Liu, Shuaipeng Zhang, et al. Molminer: you only look once for chemical structure recognition. *Journal of Chemical Information and Modeling*, 62(22):5321–5328, 2022. 2, 3

[63] Zhanpeng Xu, Jianhua Li, Zhaopeng Yang, Shiliang Li, and Honglin Li. Swinocsr: end-to-end optical chemical structure recognition using a swin transformer. *Journal of Cheminformatics*, 14(1):41, 2022. 2, 3, 4

[64] Junwei Yang, Kangjie Zheng, Siyu Long, Zaiqing Nie, Ming Zhang, Xinyu Dai, Wei-Ying Ma, and Hao Zhou. Mol-ae: Auto-encoder based molecular representation learning with 3d cloze test objective. *bioRxiv*, pages 2024–04, 2024. 8

[65] Jiacai Yi, Chengkun Wu, Xiaochen Zhang, Xinyi Xiao, Yanlong Qiu, Wentao Zhao, Tingjun Hou, and Dongsheng Cao. Micer: a pre-trained encoder–decoder architecture for molecular image captioning. *Bioinformatics*, 38(19):4562–4572, 2022. 3

[66] Sanghyun Yoo, Ohyun Kwon, and Hoshik Lee. Image-to-graph transformers for chemical structure recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3393–3397. IEEE, 2022. 3

[67] Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[68] Dehai Zhang, Di Zhao, Zhengwu Wang, Junhui Li, and Jin Li. Mmssc-net: multi-stage sequence cognitive networks for drug molecule recognition. *RSC advances*, 14(26):18182–18191, 2024. 3

[69] Xiao-Chen Zhang, Jia-Cai Yi, Guo-Ping Yang, Cheng-Kun Wu, Ting-Jun Hou, and Dong-Sheng Cao. Abc-net: a divide-and-conquer based deep learning architecture for smiles recognition from molecular images. *Briefings in bioinformatics*, 23(2):bbac033, 2022. 2, 3

[70] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2023. 8

# MolParser: End-to-end Visual Recognition of Molecule Structures in the Wild

## Supplementary Material

## 8. Open Source Materials

MolParser-7M dataset is open sourced in HuggingFace Dataset. The yolo11 model used for detection molecule structure is also avaliable in HuggingFace Model. We also provide a OCSR demo using our MolParser-Base model.

## 9. Extended SMILES Explanation

The extended SMILES format is defined as:

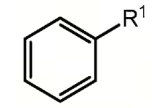
`SMILES<sep>EXTENSION`

1. `SMILES` represents an RDKit-compatible SMILES expression. Each molecule has a unique representation that can be generated (for non-Markush molecules) using the following method, where `rootedAtAtom=0` indicates that the SMILES generation starts from the atom indexed at 0.
2. `<sep>` is the delimiter separating the RDKit-compatible SMILES string from its extended description. The part before the delimiter is the RDKit-compatible SMILES, while the part after provides supplemental information (e.g., Markush groups, connection points, repeating groups).
3. `EXTENSION` is an optional component that supplements the preceding SMILES with descriptions written in XML format, including groups surrounded by special tokens of three types:
   (a) `<a>[ATOM_INDEX]:[GROUP_NAME]</a>` indicates a substituent.
   (b) `<r>[RING_INDEX]:[GROUP_NAME]</r>` represents a group connected at any position of a ring.
   (c) `<c>[CIRCLE_INDEX]:[CIRCLE_NAME]</c>` denotes abstract ring.

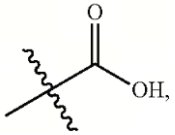
   An additional special token `<dum>` indicates a connection point.

   **Definitions:**
   - `ATOM_INDEX` refers to the atom index at which the substituent is located (starting from 0).
   - `RING_INDEX` denotes the ring index (starting from 0).
   - `GROUP_NAME` specifies the name of the substituent, which can be an abbreviated group, general substituent, or Markush group, such as `R, X, Y, Z, Ph, Me, OMe, CF3`, etc. It may also be `<dum>` to indicate a connection point. For Markush substituents with superscripts or subscripts, these can be represented within square brackets, e.g., `R[1], R[3]`.
   - `CIRCLE_INDEX` refers to the index of the named ring (starting from 0).
   - `CIRCLE_NAME` indicates the name of the ring.

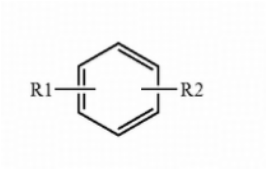
Figure 4 shows the usage of extended SMILES:



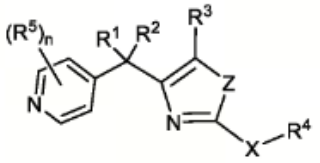

`*c1ccccc1<sep><a>0:R[1]</a>`



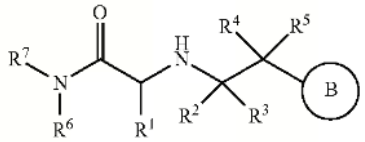

`*C(O)=O<sep><a>0:<dum></a>`




`c1ccccc1<sep><r>0:R[1]</r><r>0:R[2]</r>`




`**C1*C(*)=C(C(*)(*)C2=CC=NC=C2)N=1<sep>`
`<a>0:R[4]</a><a>1:X</a><r>1:R[5]?n</r>`
`<a>3:Z</a><a>5:R[3]</a><a>8:R[2]</a><a>9:R[1]</a>`




`*C(NC(*)(*)C(*)(*)*)C(=O)N(*)*<sep>`
`<a>0:R[1]</a><a>4:R[3]</a><a>5:R[2]</a><a>7:R[5]</a>`
`<a>8:R[4]</a><c>0:B</c><a>13:R[7]</a><a>14:R[6]</a>`

Figure 4. Molecule images examples with extended SMILES. The red parts are as follows: Markush group, attachment point, ring attachment with uncertainty position, duplicated structure and abstract ring.

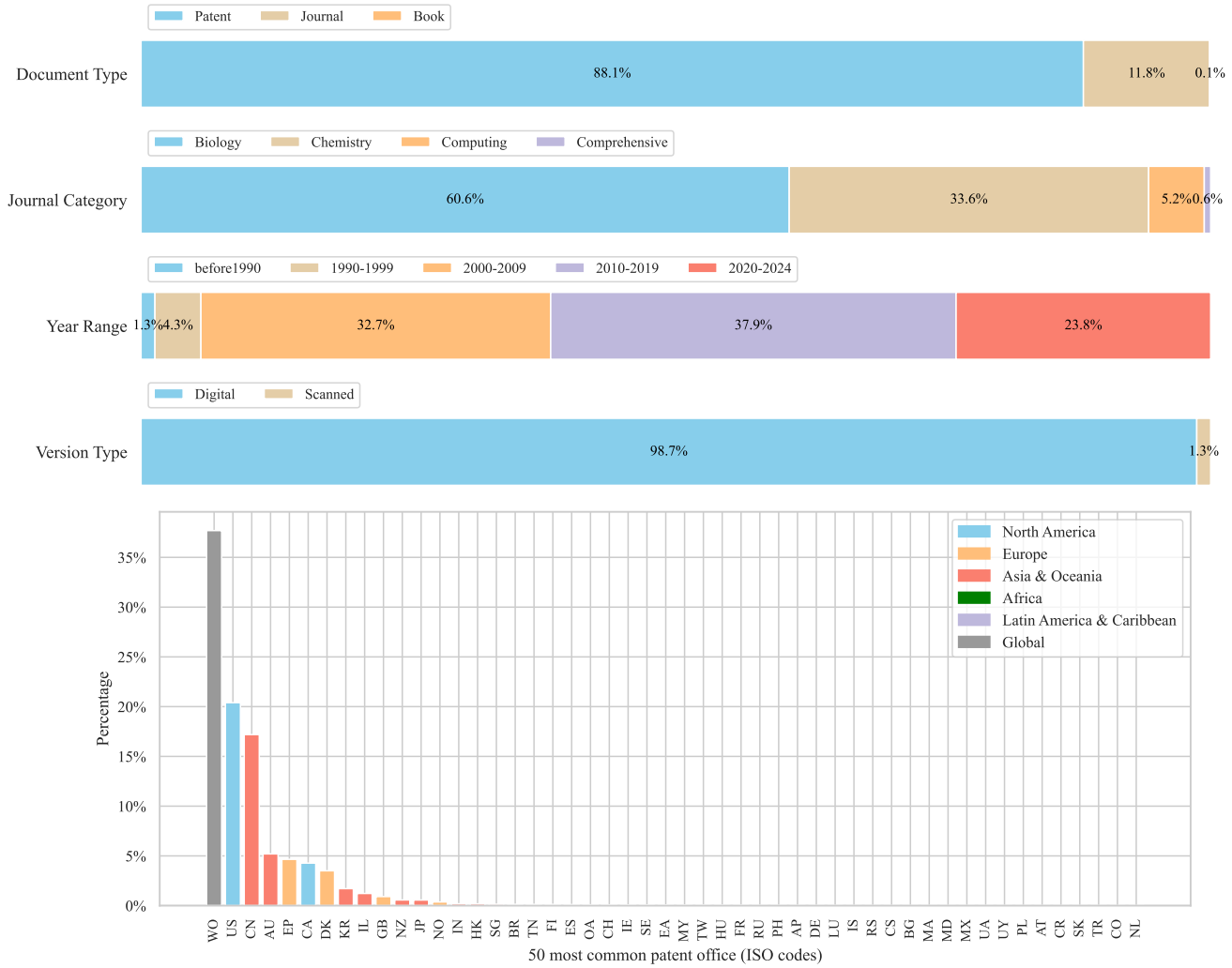
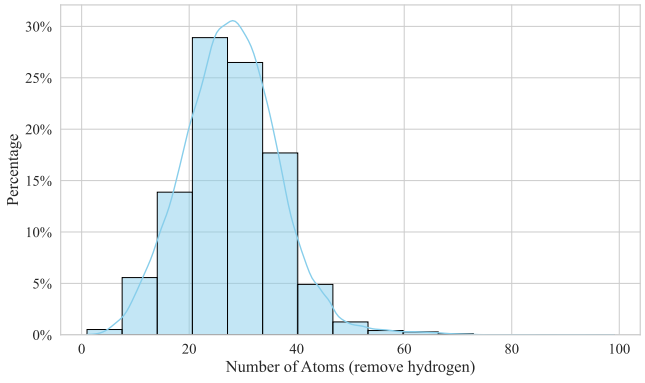
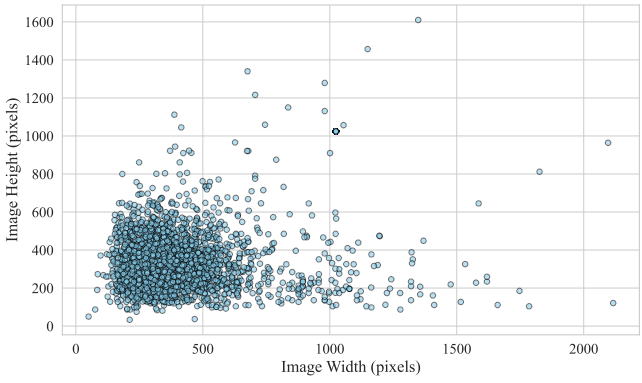



Figure 5. **Statistical analysis of MolParser-SFT original PDF database.** We compiled source information for the collected PDF files, including article type, publication date, PDF format, journal subject distribution, and patent office sources.




(a) Distribution of atom count




(b) Distribution of image size

Figure 6. **Statistical analysis of MolParser-7M dataset.** The distribution of molecule atom counts and molecule image sizes. Compared to the fixed-size synthetic datasets used in other studies, our dataset exhibits a wider range of image sizes and aspect ratios.

## 10. Dataset

### 10.1. Statistical information of MolParser-7M

Molparser-7M contains a total of 7,740,871 paired OCSR training data, making it the largest open-source paired OCSR dataset currently available. It is important to note that the open-source datasets MolGrapher-300k and Img2Mol are both subsets of our Molparser-7M. Additionally, as shown in Figure 5 and 6, the data distribution of our MolParser is more comprehensive.

### 10.2. Data Augmentation

**Render Augmentation** During synthetic data generation in our data engine, we incorporat several augmentations for rendering molecular structure diagrams, similar to those used in MolGrapher [38]. Augmentations such as bond width, font type, font size, rotation, and aromatic cycle representation are randomly applied during rendering.

**Image Augmentation** Whether for synthetic data or real data, we also apply image augmentations during training. We use several types of data augmentation, including RandomAffine, JPEGCompress, InverseColor, SurroundingCharacters, RandomCircle, ColorJitter, Downscale and Bounds. These type of augmentation are visualized in Figure 8.

**SMILES Augmentation** We apply SMILES augmentation only during pre-training. Since a molecule's SMILES representation varies with the choice of root atom, we randomly change the root atom to help the transformer learn SMILES syntax more robustly. During fine-tuning, augmentation is disabled and the root atom is fixed to index zero, reducing ambiguity during generation.

## 11. Experiment Setting

All variants of MolParser adopt a BART decoder with 12 transformer decoder layers and 16 attention heads. An MLP connector reduces the channel dimension of the visual encoder output by half. The Swin Transformer produces a feature map of size bs x n × n × d, which is flattened into a sequence and used as prefix tokens for the decoder.

During the pre-training stage, we train the model for 20 epochs using the AdamW optimizer with a learning rate of 1e-4, a weight decay of 1e-2, and a cosine learning rate schedule with warmup. We set label smoothing to 0.01. In the fine-tuning phase, we reduce the learning rate to 5e-5, decrease the number of training epochs to 4, and lower the label smoothing to 0.005. All experiments are conducted on 8 NVIDIA RTX 4090D GPUs.

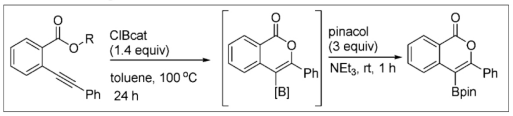
## 12. Case Study

We test numerous examples and conduct qualitative analysis. Our MolParser performs well on stylized and low-quality molecular images that challenge previous algorithms, though it struggles with overlapped large molecules, charged molecules, and some cases where E-SMILES cannot effectively represent the structures, which shown in figure 10 and figure 11.

## 13. Downstream Usage

In unstructured documents, extracting molecular structures and leveraging LLMs for structured information extraction has become a key application of Optical Chemical Structure Recognition (OCSR). We first convert each PDF page into an image and use a YOLO11 [22] model to detect molecular structures. The detected molecules are then parsed by our MolParser and converted into an extended, XML-like SMILES format that is more LLM-friendly. This representation allows LLMs to easily identify which groups undergo transformations in chemical reactions. Following OmniParser [36], we integrate molecular location and SMILES information into GPT-4o [21] to enhance MolParser's ability to parse full chemical reaction formulas.

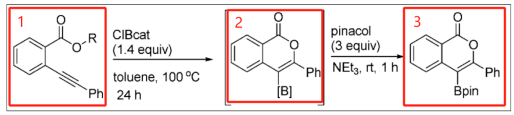
***Reaction Diagram***




**User:** Explain this reaction in detail.

**MolParser:** Applying molecule detection and OCSR.

***Reaction Diagram***




**MolParser:**
Extended SMILES of Molecules in Image:
Molecule Box ID 1: `*OC(=O)C1=CC=CC=C1C#C*<sep>`
`<a>0:R</a><a>12:Ph</a>`
Molecule Box ID 2: `O=c1c2c(cccc2)c(*)c(*)o1<sep>`
`<a>9:B</a><a>11:Ph</a>`
Molecule Box ID 3: `O=c1oc(*)c(*)c2ccccc12<sep>`
`<a>4:Ph</a><a>6:Bpin</a>`

**GPT-4o:**
This reaction diagram illustrates a two-step synthetic pathway:
First Step: Starting from molecule
`*OC(=O)C1=CC=CC=C1C#C*<sep>`
`<a>0:R</a><a>12:Ph</a>`,
which has an alkyne and ester group, the reaction proceeds with ClBcat (1.4 equivalents) in toluene at 100 °C for 24 hours. This yields the intermediate molecule
`"O=c1c2c(cccc2)c(*)c(*)o1<sep>`
`<a>9:B</a><a>11:Ph</a>"`,
which has a fused aromatic ring with a carbonyl group.
Second Step: The intermediate
`O=c1c2c(cccc2)c(*)c(*)o1<sep>`
`<a>9:B</a><a>11:Ph</a>`
is then reacted with pinacol (3 equivalents) and $NEt_3$ at room temperature for 1 hour, producing the final product
`O=c1oc(*)c(*)c2ccccc12<sep>`
`<a>4:Ph</a><a>6:Bpin</a>`,
where a Bpin group is introduced onto the fused ring structure.

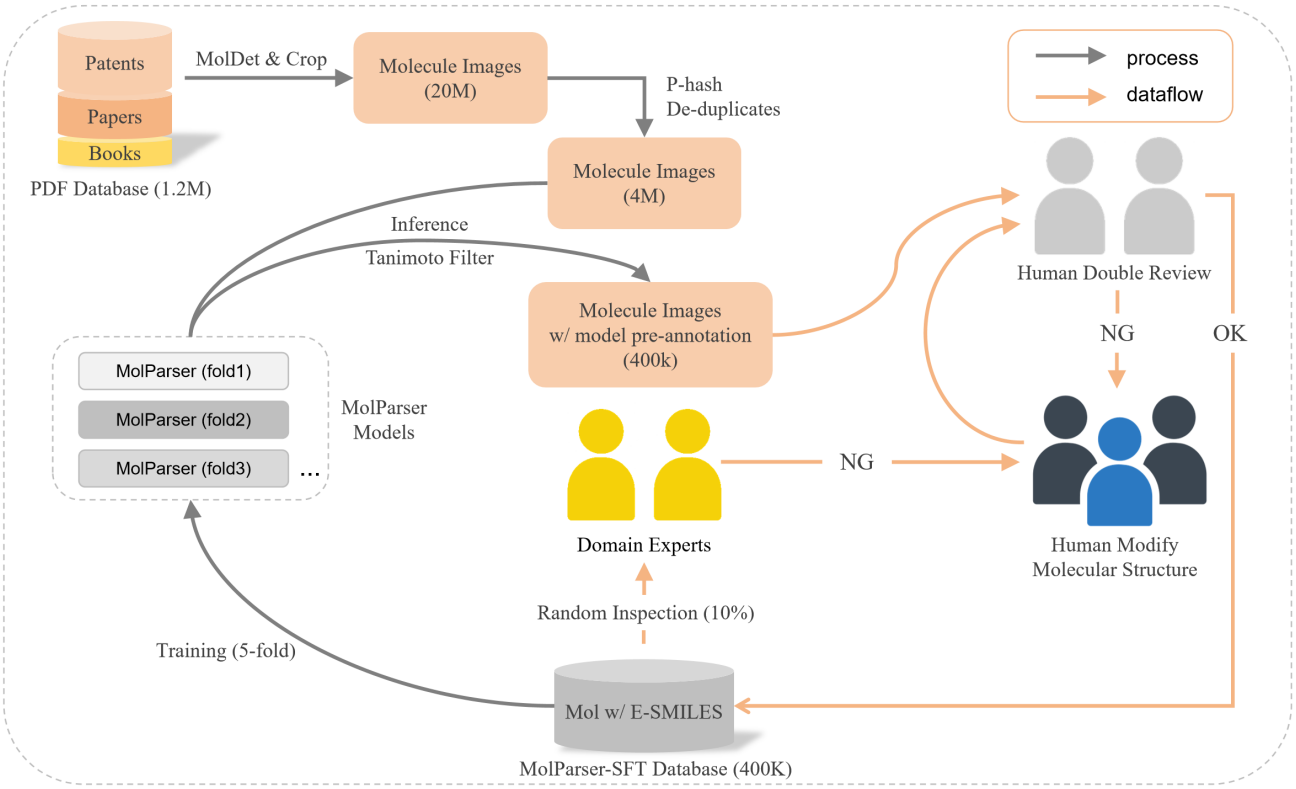



Figure 7. **MolParser data engine.** We design a human-in-the-loop active learning framework, using Tanimoto similarity scores of multiple model predictions to select molecules for training. Each molecule image is pre-labeled by the model, reviewed by two annotators, and subject to expert inspection.

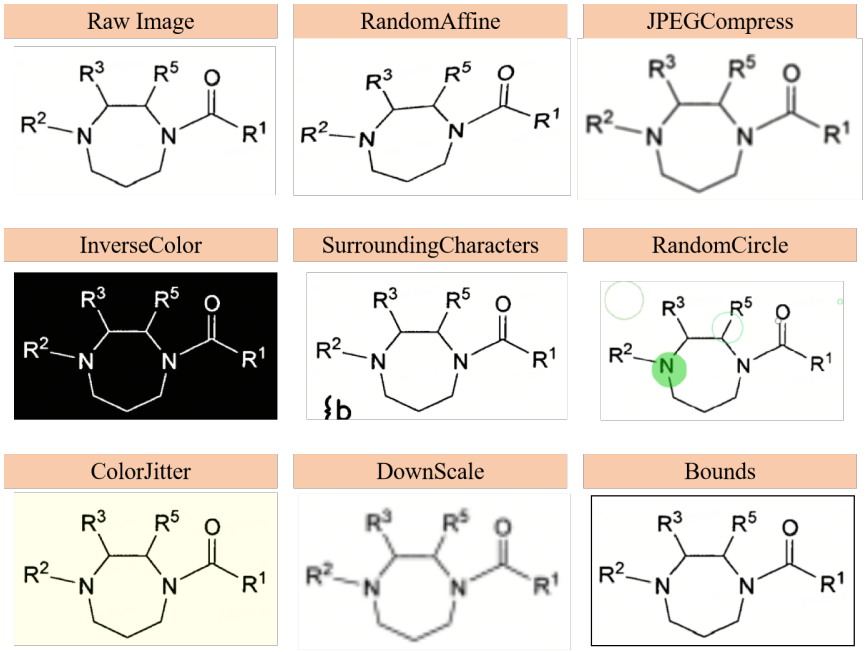



Figure 8. **Data augmentation in training.** We design the augm entation of the image according to the noise that may occur in real data, which cropped from scanned PDF files.

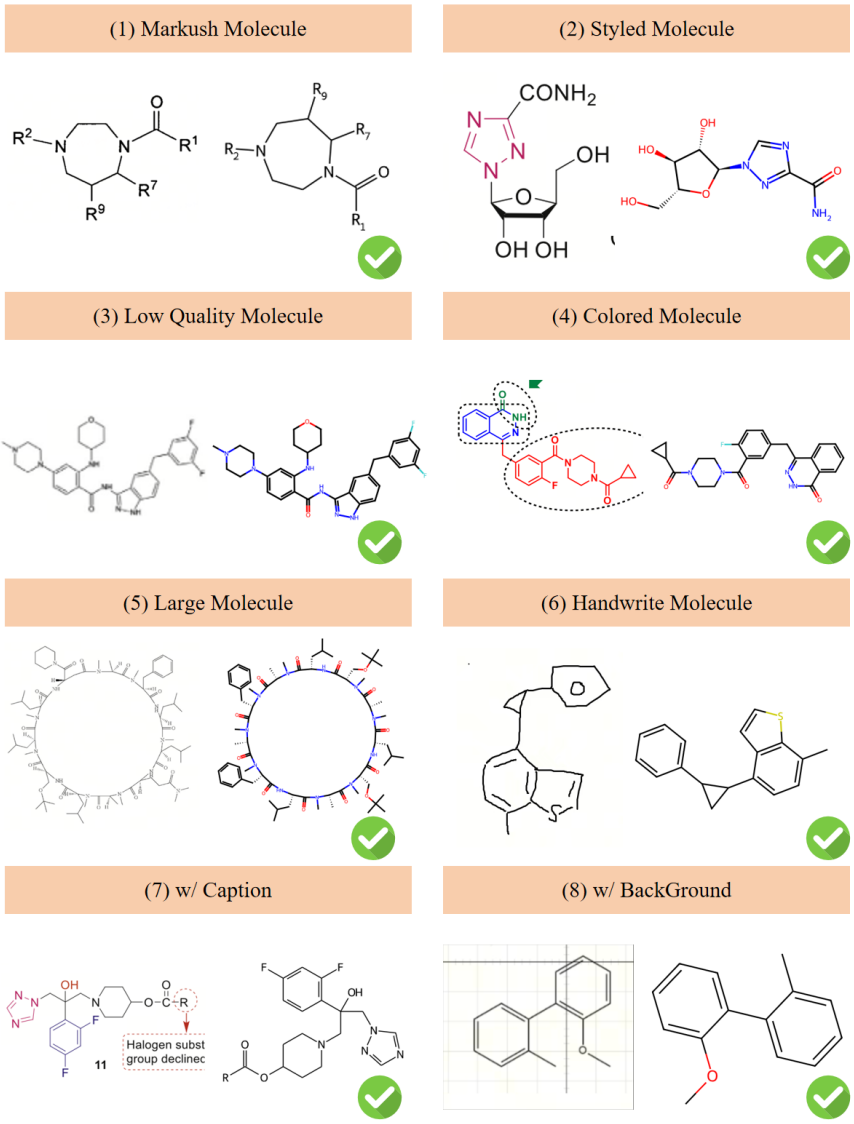



Figure 9. **MolParser qualitative evaluation.** The figure shows the broad diversity of predictions made by MolParser for input molecular images. The input image (left) is displayed alongside the predicted molecule rendered by E-SMILES prediction (right).

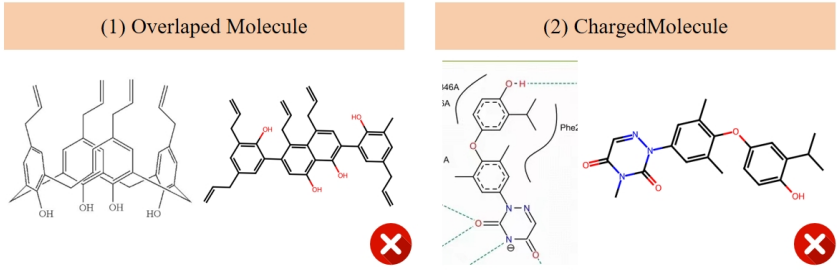



Figure 10. **MolParser failure case.** The figure shows the broad diversity of predictions made by MolParser for input molecular images. The input image (left) is displayed alongside the predicted molecule rendered by E-SMILES prediction (right).

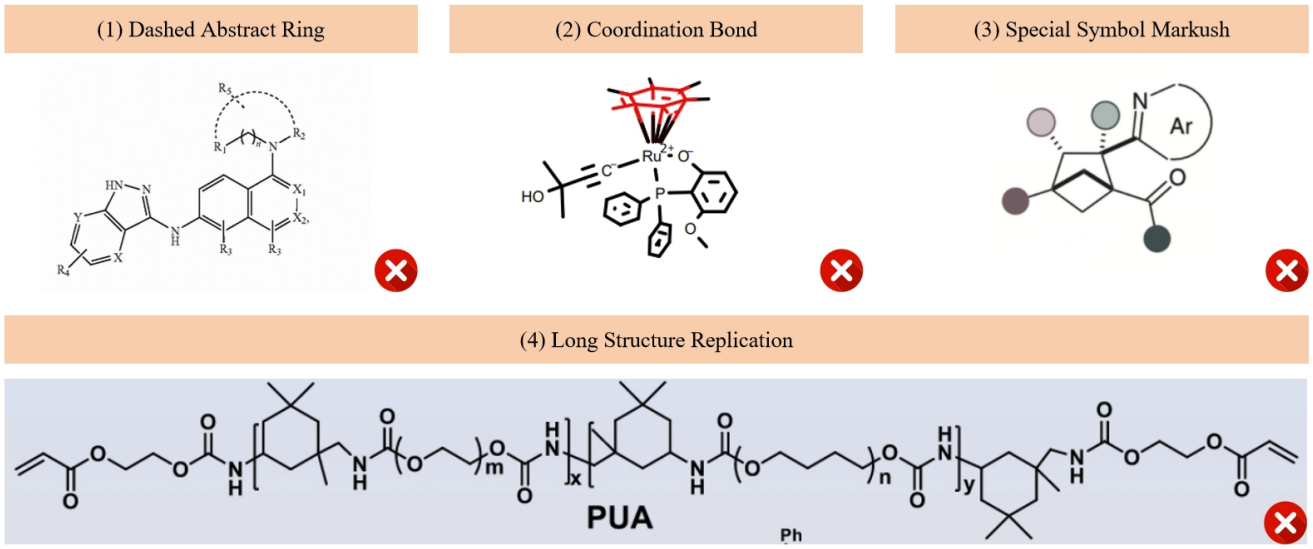



Figure 11. **E-SMILES failure case.** Molecular structures with dashed lines representing abstract rings, structures with coordination bonds, and Markush structures depicted using special patterns are not currently supported in E-SMILES notation. Additionally, the replication of long structural segments on the skeleton, rather than individual atoms, is also not supported by our E-SMILES format.

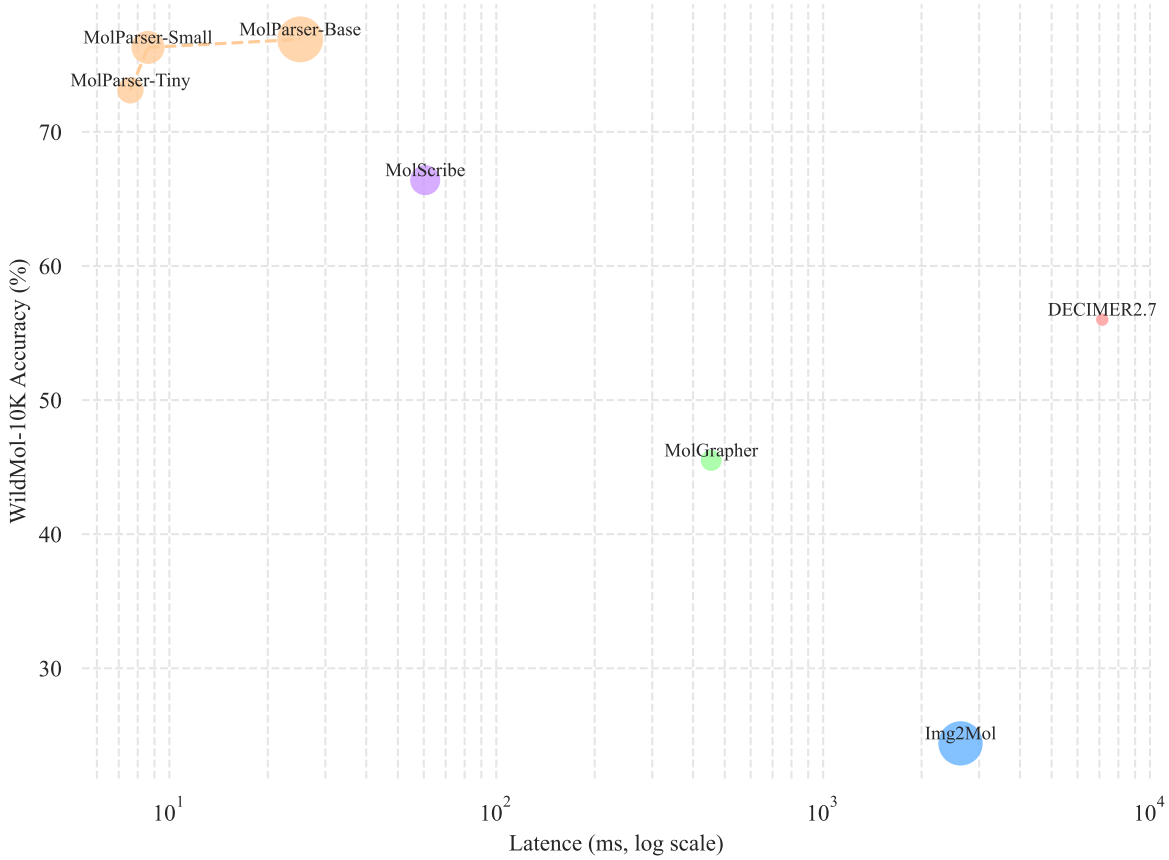



Figure 12. **The speed-accuracy Pareto curve of the OCSR system.** Models toward the top-left corner are better. The size of the circles represents the model's parameter count, and the time is tested on a single RTX-4090D GPU for the entire pipeline.