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:

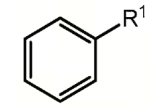
`SMILES<sep>EXTENSION`

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.

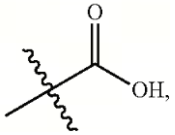
   An additional special token `<dum>` indicates a connection point.

   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.

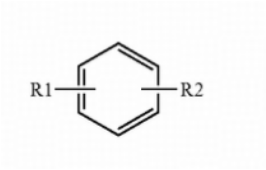
Figure 4 shows the usage of extended SMILES:



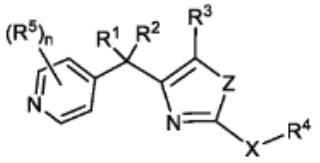
`*c1ccccc1<sep><a>0:R[1]</a>`



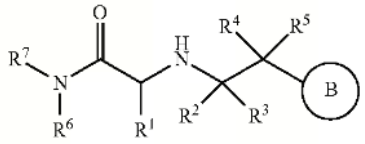
`*C(O)=O<sep><a>0:<dum></a>`



`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.

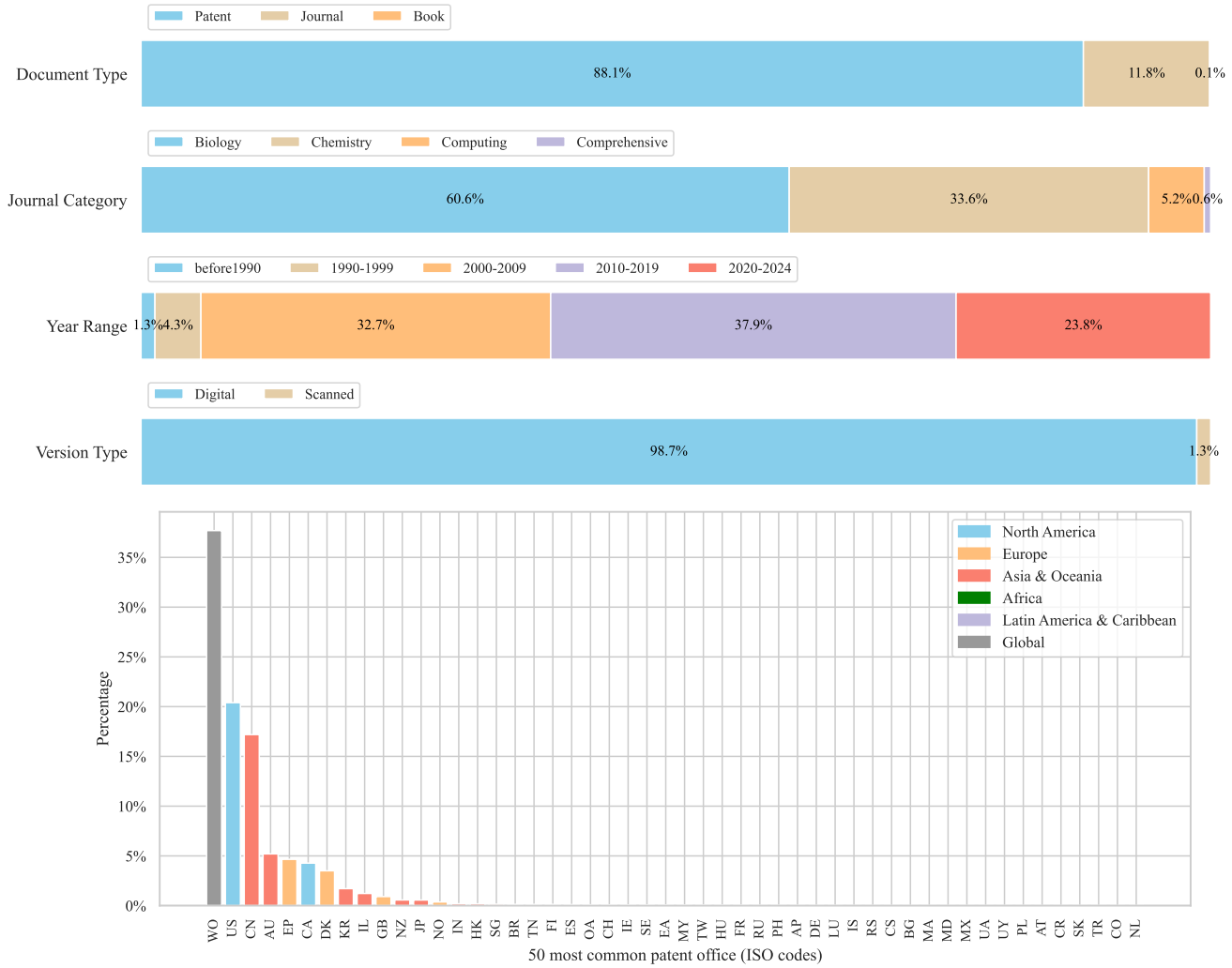
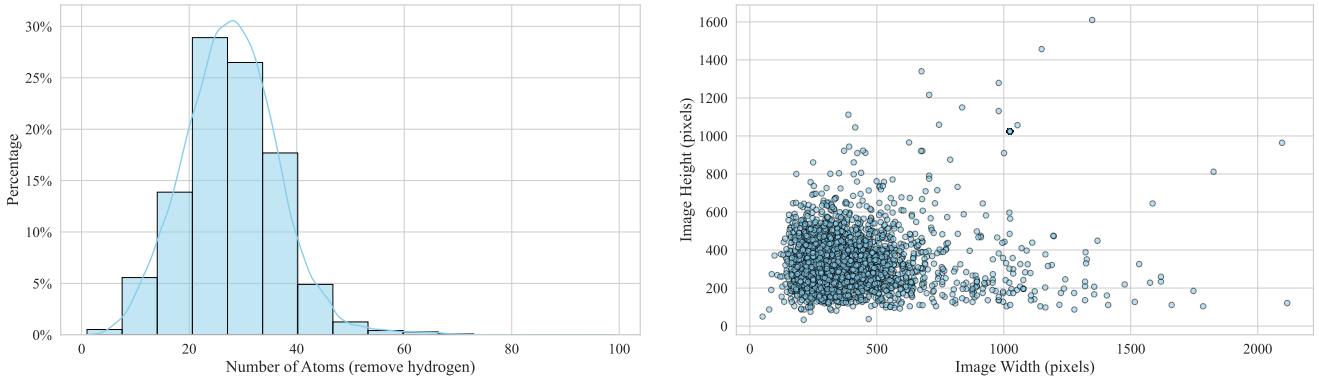


Figure 5. **Statistical analysis of MolParser-SFT original PDF database.** We compiled source information for the collected PDF files, including article type, publication date, PDF format, journal subject distribution, and patent office sources.



(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.

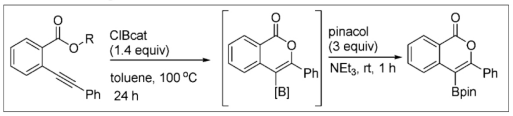
12. Case Study

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.

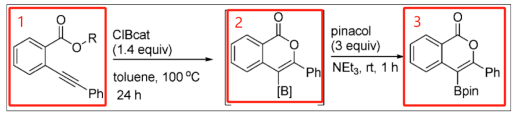
Reaction Diagram



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.

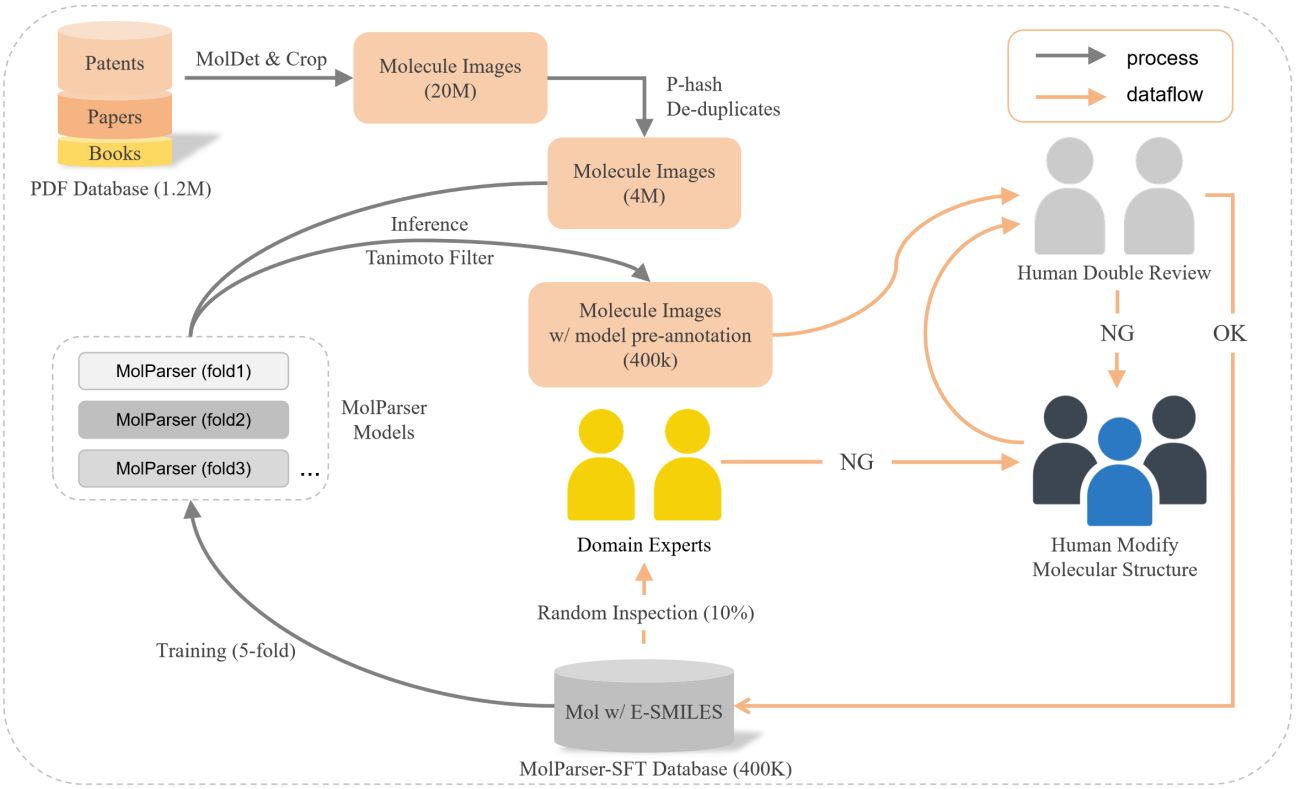


Figure 7. **MolParser data engine.** We design a human-in-the-loop active learning framework, using Tanimoto similarity scores of multiple model predictions to select molecules for training. Each molecule image is pre-labeled by the model, reviewed by two annotators, and subject to expert inspection.

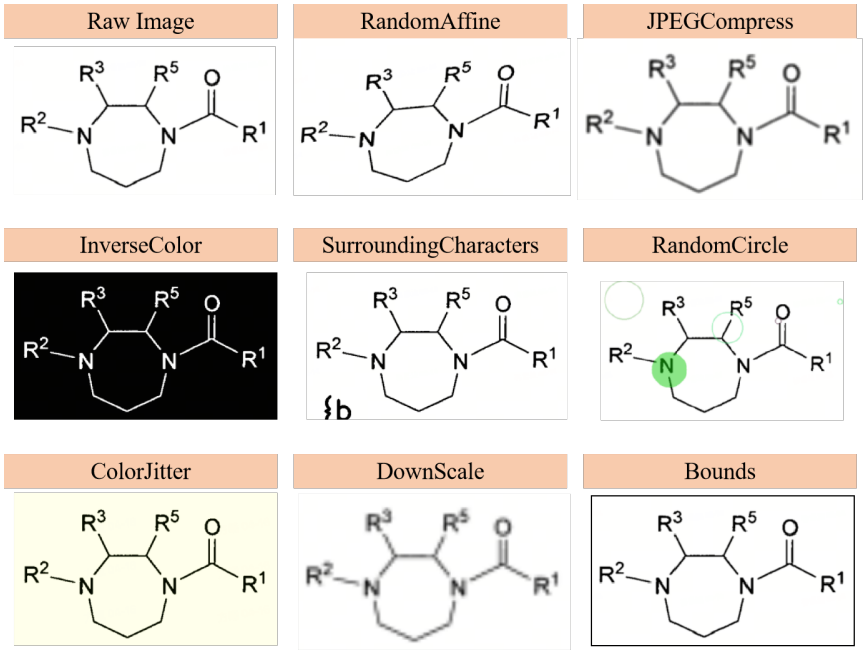


Figure 8. **Data augmentation in training.** We design the augm entation of the image according to the noise that may occur in real data, which cropped from scanned PDF files.

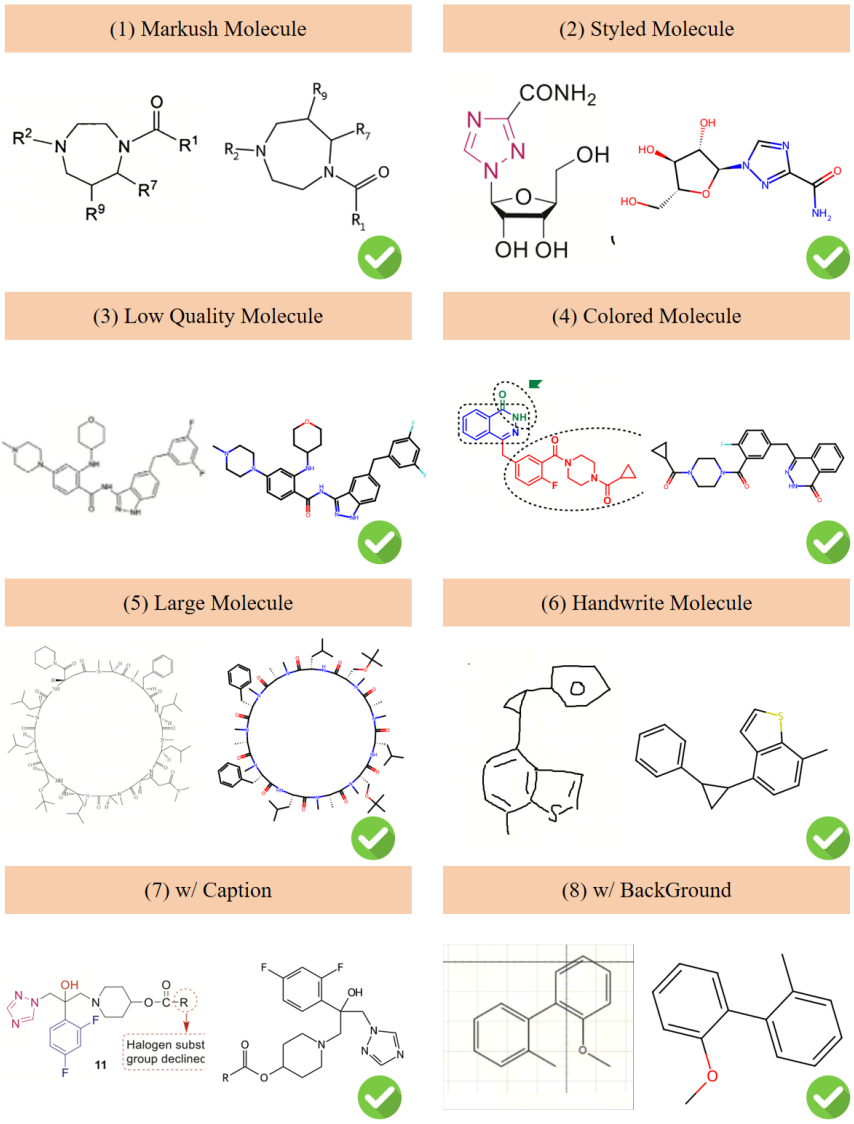


Figure 9. **MolParser qualitative evaluation.** The figure shows the broad diversity of predictions made by MolParser for input molecular images. The input image (left) is displayed alongside the predicted molecule rendered by E-SMILES prediction (right).

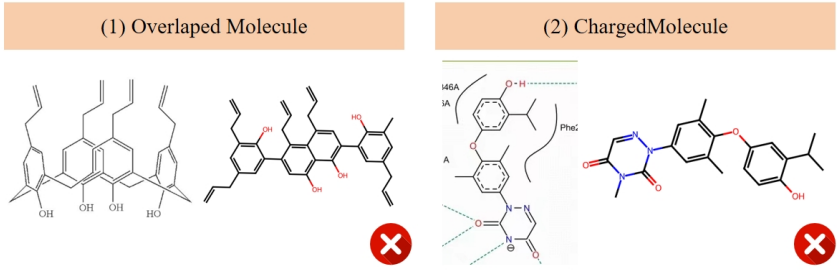


Figure 10. **MolParser failure case.** The figure shows the broad diversity of predictions made by MolParser for input molecular images. The input image (left) is displayed alongside the predicted molecule rendered by E-SMILES prediction (right).

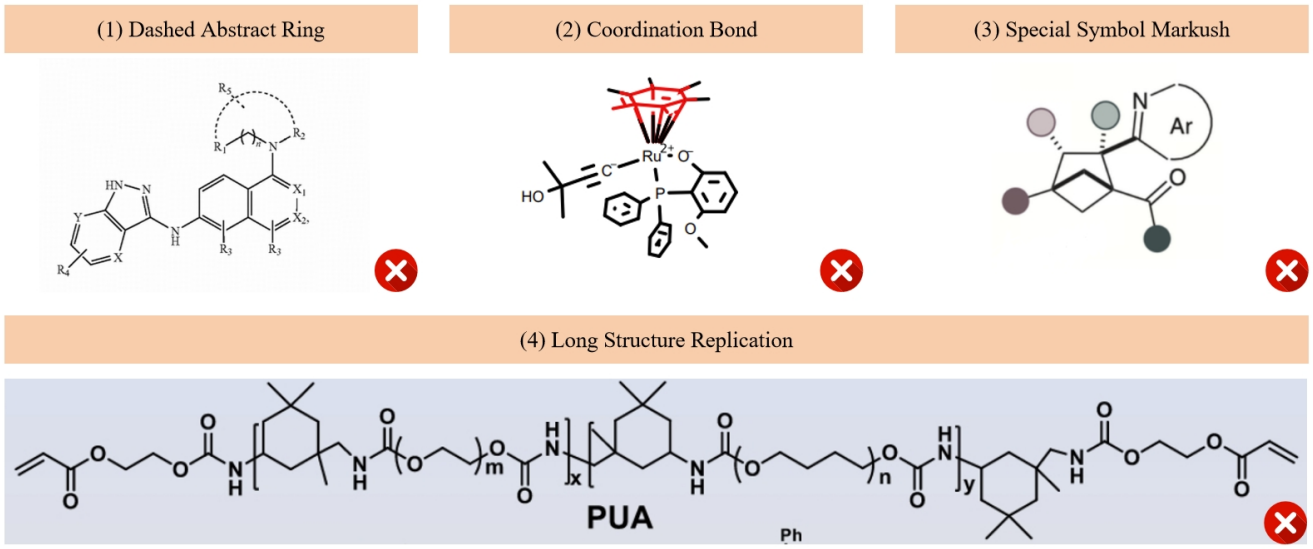


Figure 11. **E-SMILES failure case.** Molecular structures with dashed lines representing abstract rings, structures with coordination bonds, and Markush structures depicted using special patterns are not currently supported in E-SMILES notation. Additionally, the replication of long structural segments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

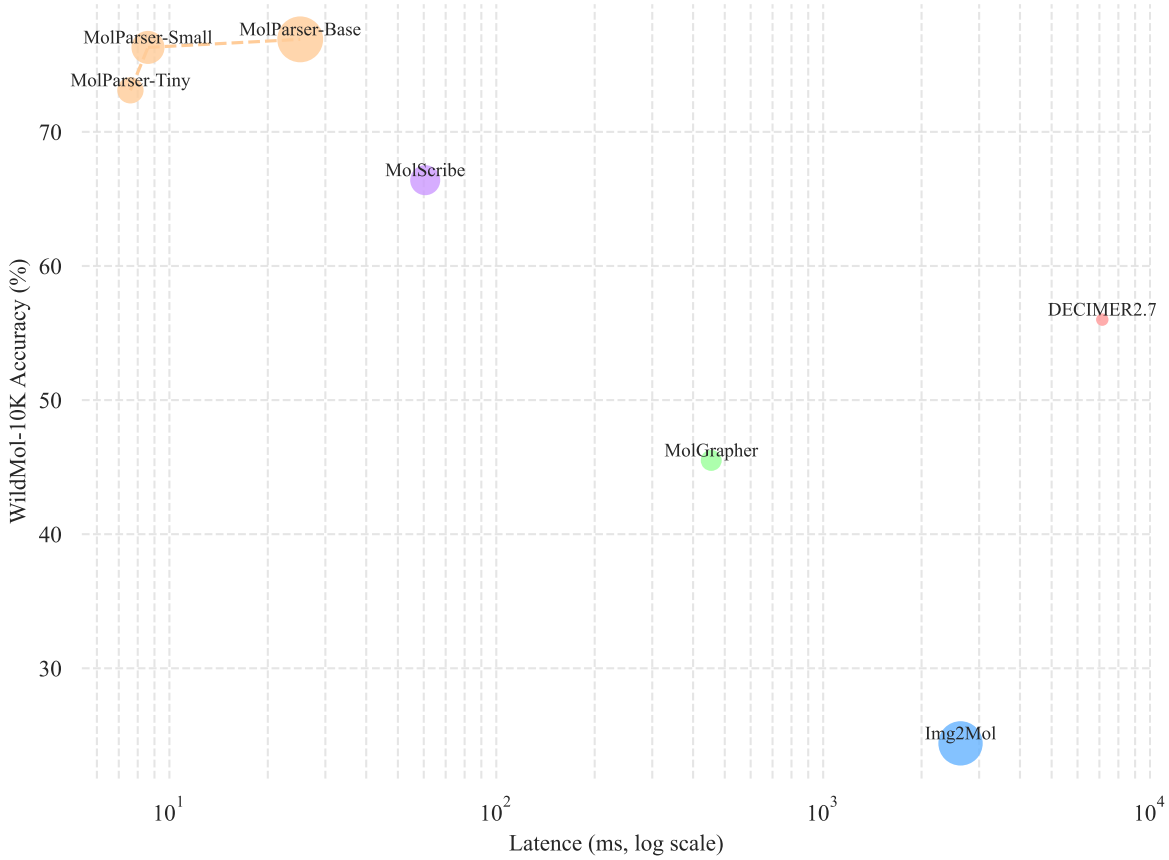


Figure 12. **The speed-accuracy Pareto curve of the OCSR system.** Models toward the top-left corner are better. The size of the circles represents the model's parameter count, and the time is tested on a single RTX-4090D GPU for the entire pipeline.