Optical Molecular Recognition From Chemical Reaction Mechanism Images

Ching Ting Leung, Yufan Chen & Hanyu Gao*

Department of Chemical and Biological Engineering,
Hong Kong University of Science and Technology

1 Background

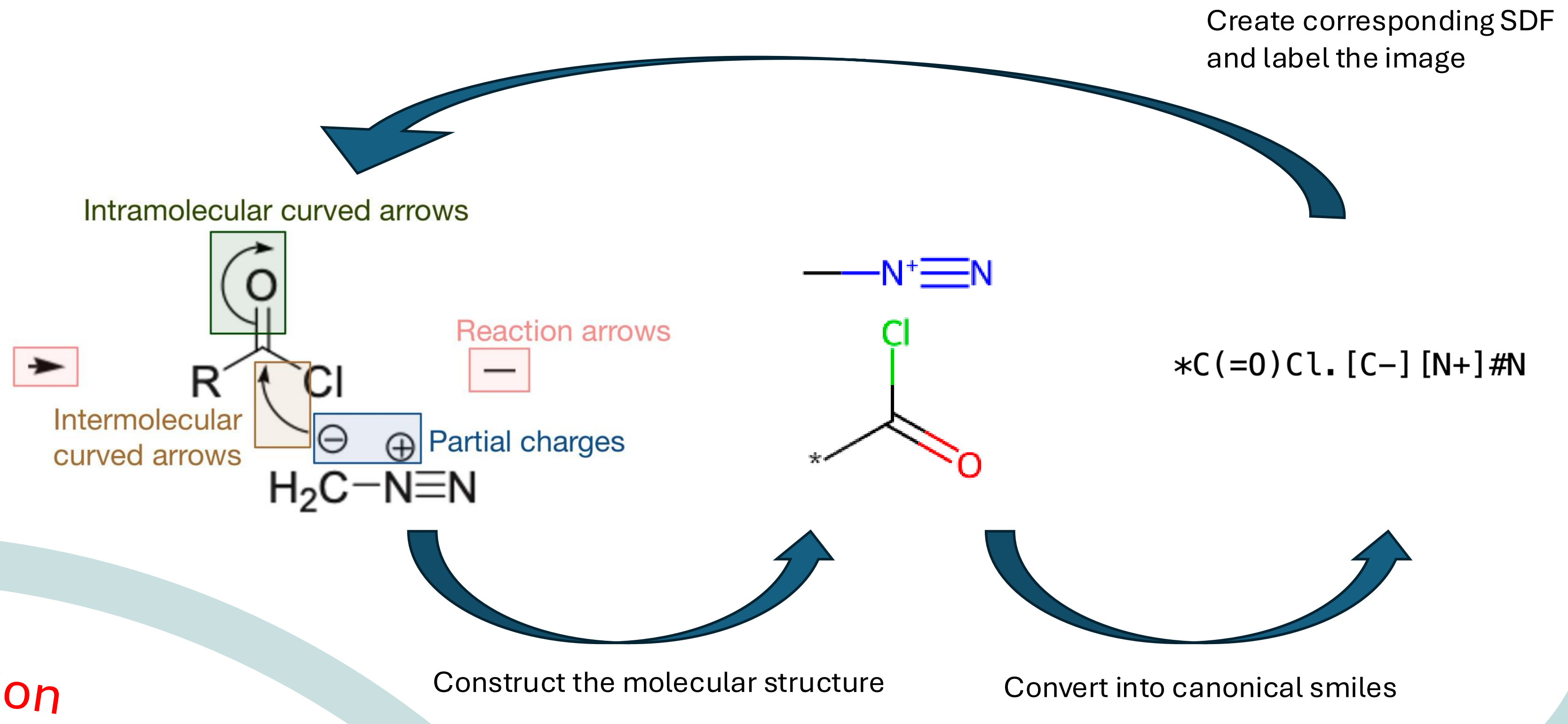
- Existing OCSR models mainly target properly drawn molecules
- Noises in chemical reaction mechanisms are essential for chemists, but they are structurally similar to the molecule, causing identification errors
- Relying purely on CNN or GNN models for filtering noises is computationally expensive

2 Highlights

- Proposed an image preprocessing technique for optical molecular recognition
- Propose an automated pipeline for processing molecular and reaction information from noisy data
- Created molecular and reaction datasets targeting specifically chemical reaction mechanisms

Molecular Dataset 3

- Created a dataset of 453 molecules from chemical reaction mechanisms [2]
- They include mechanistic features such as curved arrows and partial charges
- Manually annotated with ASKCOS [1] for their SMILES and RDKit for Structural Mol Files


Create corresponding SDF and label the image

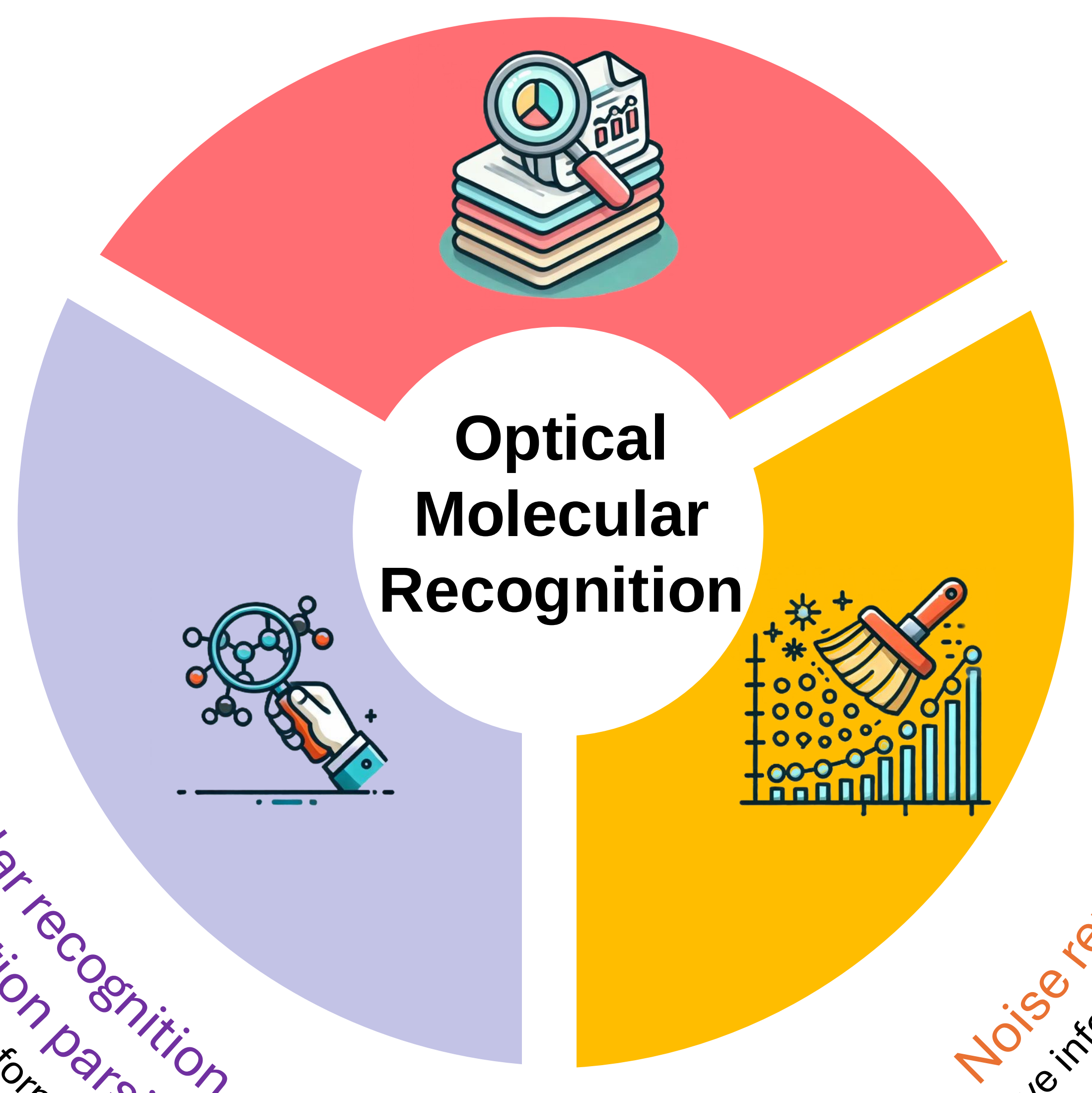
`*C(=O)Cl.[C-][N+]#N`

Construct the molecular structure

Convert into canonical smiles

Data collection

Collect target chemical data from literatures



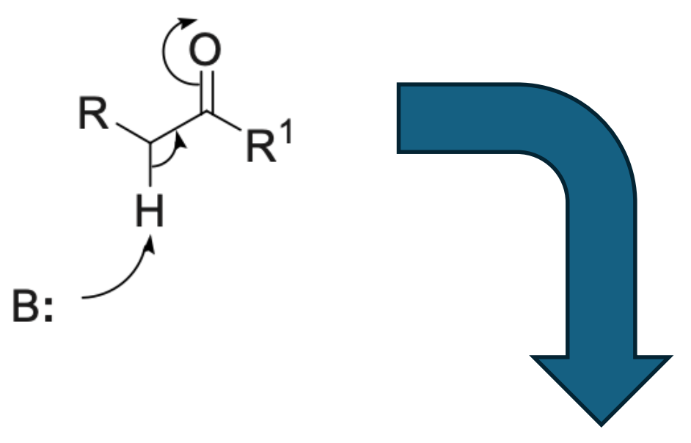
Molecular recognition and Reaction parsing

Extracting essential information such as molecular identities and positions

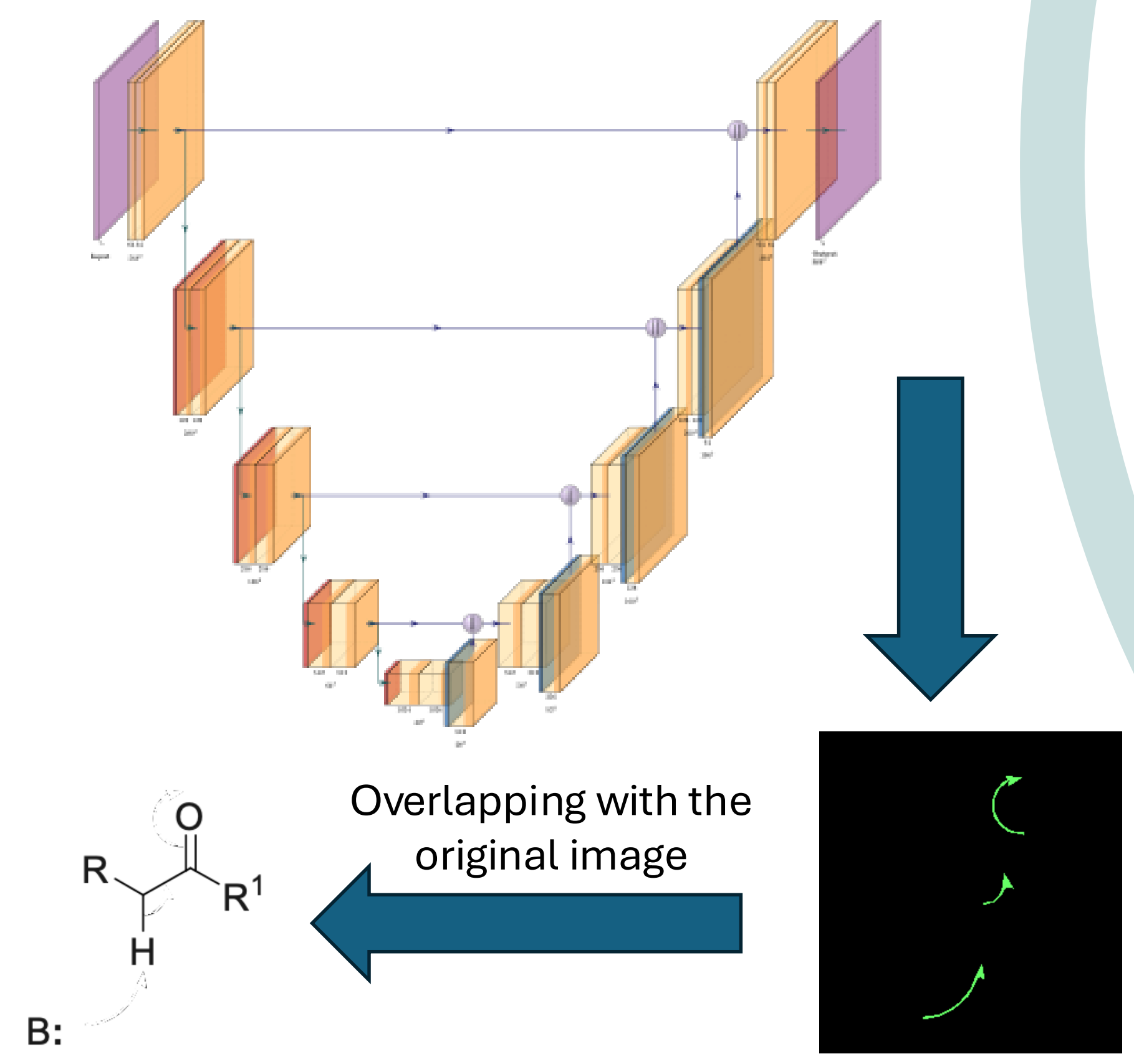
Noise removal

Remove information that is not necessary for identity recognition

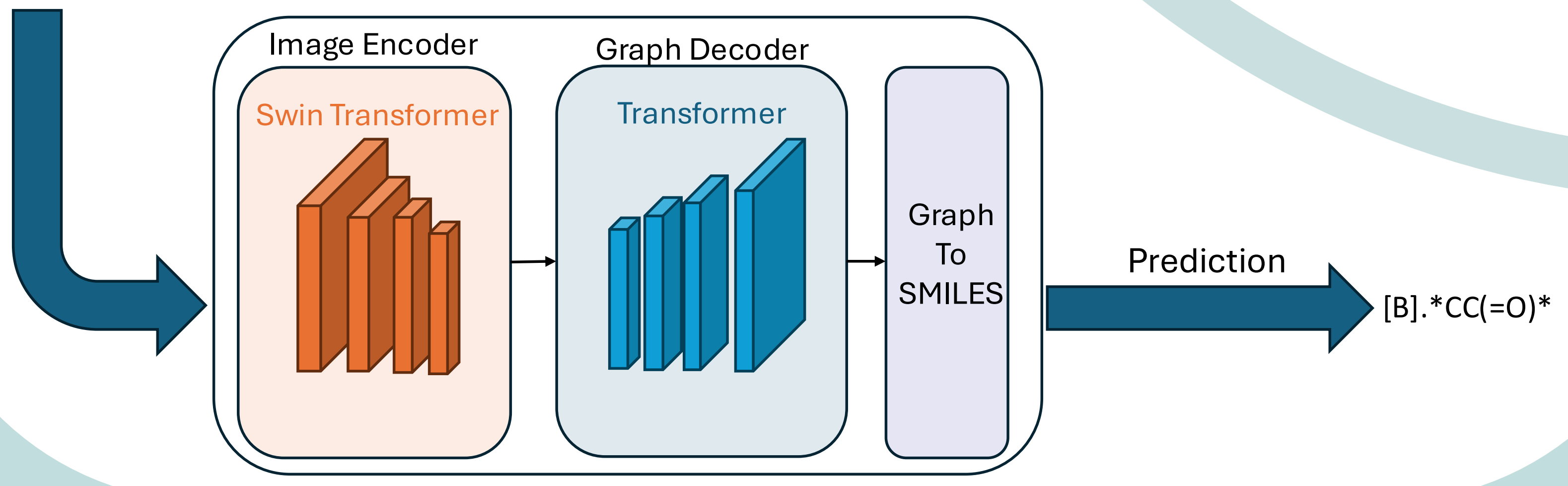
4 Recognition Pipeline



- Arrow segmentation based on U-net
- Manual annotated 150 images with CVAT for training

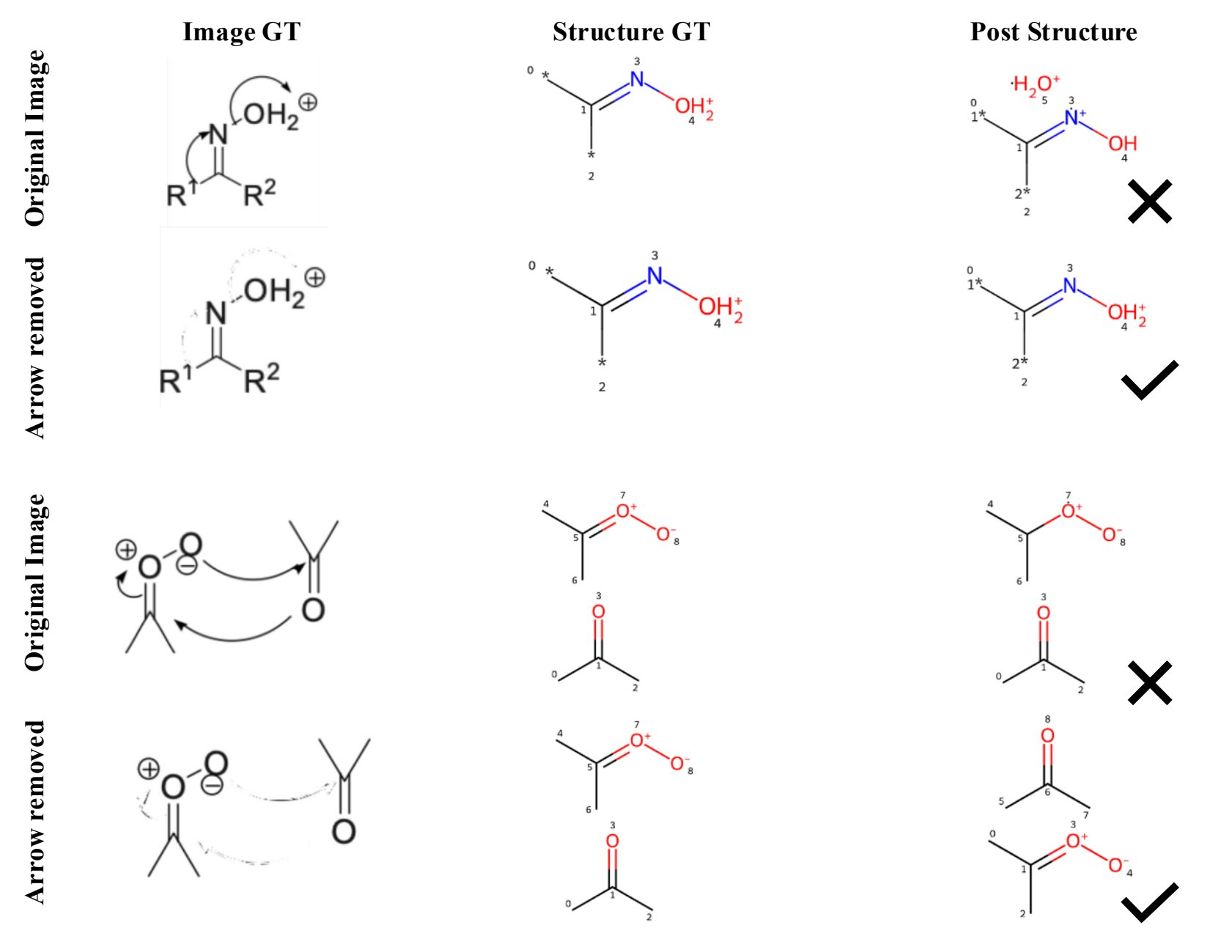

Overlapping with the original image

MolScribe [4] for molecular recognition

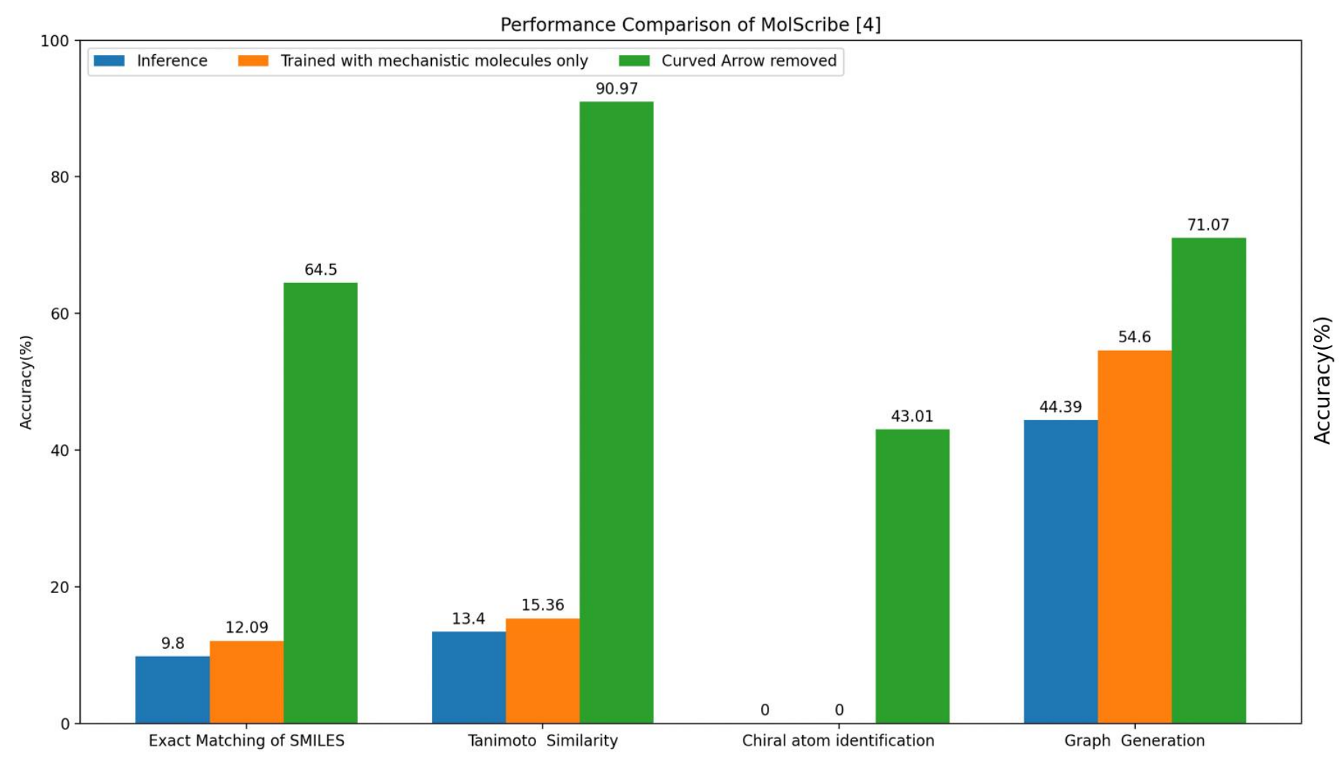


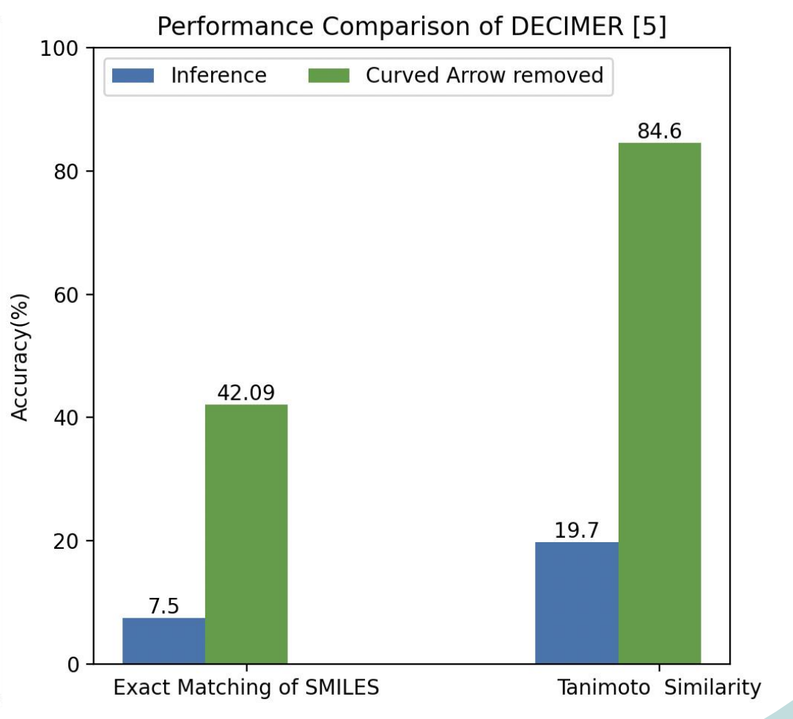
Results 5

- The presence of curved arrows significantly affects the identification of bonds
- Removal of curved arrows can help retain the essential information of molecule identification while remove noises.

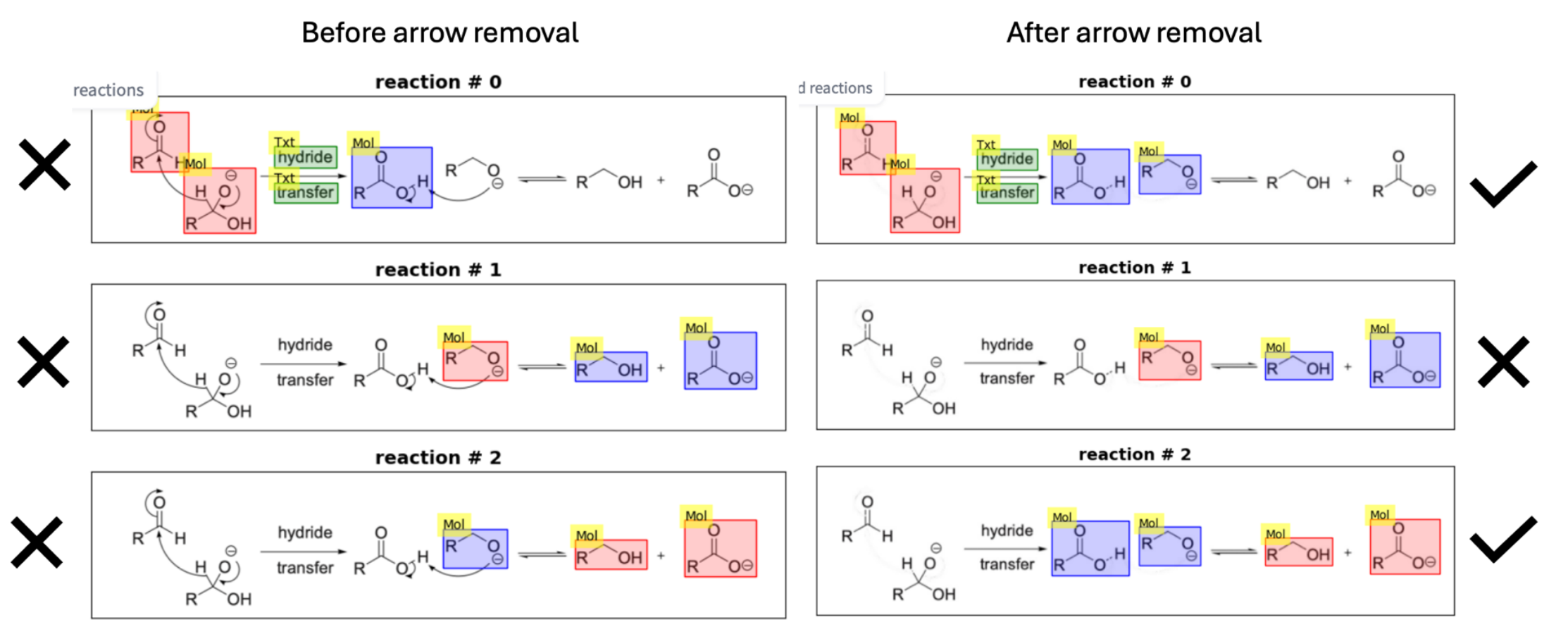


- Shows significant improvement in the used OCSR model evaluation metrics
- Performs as well as their collected dataset
- Perform well in both a sequence-to-sequence model or a CNN encoder scenario


Performance Comparison of MolScribe [4]

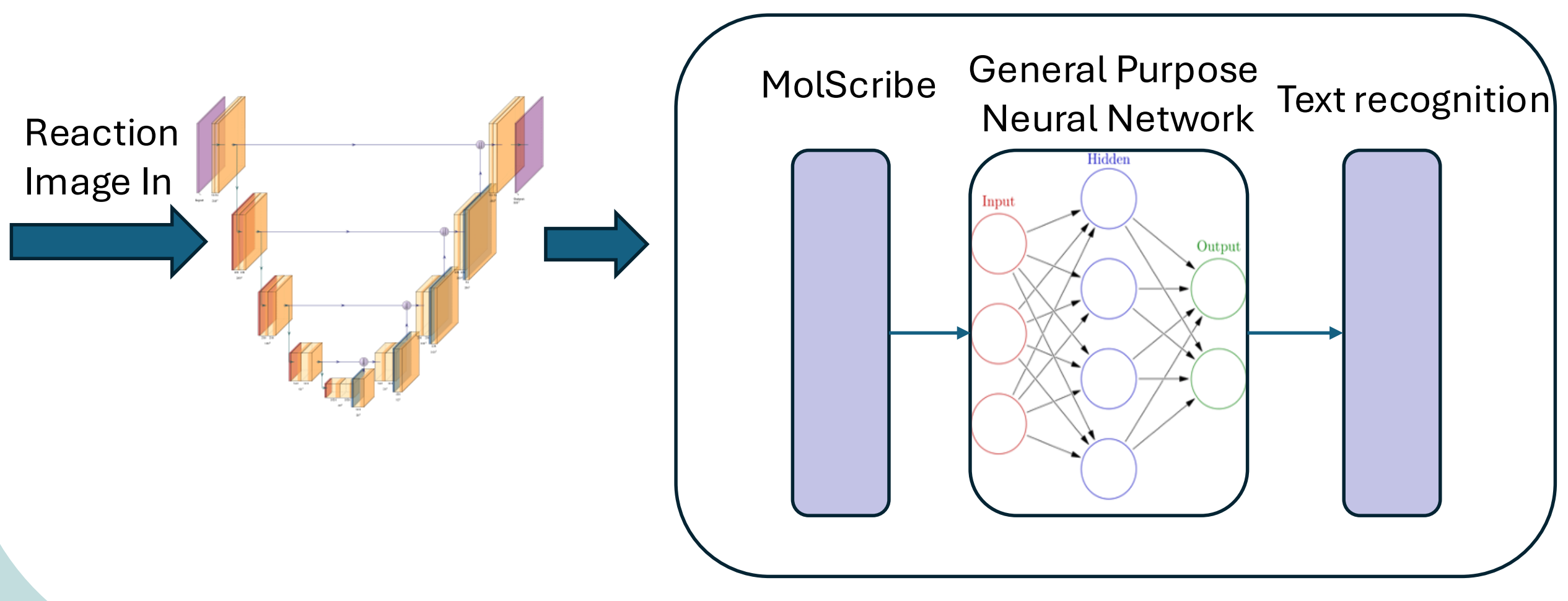

Performance Comparison of DECIMER [5]

6 Parsed Reaction Dataset



- Further application with increased accuracy on molecular identity and position
- RxnScribe [3] demo on chemical reaction mechanism parsing
- Develop pipeline for a centralized and unified collection of 296 named chemical reaction mechanisms [2]

RxnScribe



```
{“Rxn_name.png": [
    {"reactants": [
        {"category": "[Mol]",
         "bbox": [location],
         "smiles": ",
         "molfile": }],
    "conditions": [
         {"category": "[Txt]",
          "bbox": [location],
           "text": [] }],
    "products": [
         {"category": "[Mol]",
          "bbox": [location],
          "smiles": “”
          "molfile: }] },
```

Conclusion 7

- Justify the importance of data preprocessing
- Create datasets targeting chemical reaction mechanisms to further benefit both computer scientists and chemists

References 8

[1] C. Coley *et al*, "A robotic platform for flow synthesis of organic compounds informed by AI planning," *Science*, vol. 365, *(6453)*, pp. eaax1566, 2019. DOI: 10.1126/science.aax1566

[2] J. J. Li, *Name Reactions A Collection of Detailed Mechanisms and Synthetic Applications.* (4th ed.) 2009. DOI: 10.1007/978-3-642-01053-8

[3] Y. Qian *et al*, "RxnScribe: A Sequence Generation Model for Reaction Diagram Parsing," *Journal of Chemical Information and Modeling*, vol. 63, *(13)*, pp. 4030–4041, 2023. . DOI: 10.1021/acs.jcim.3c00439

[4] Y. Qian *et al*, "MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation," *Journal of Chemical Information and Modeling,* vol. 63, *(7)*, pp. 1925–1934, 2023. . DOI: 10.1021/acs.jcim.2c01480

[5] K. Rajan, A. Zielesny and C. Steinbeck, "DECIMER 1.0: deep learning for chemical image recognition using transformers," *J Cheminform*, vol. 13, *(1)*, pp. 61, 2021. . DOI: 10.1186/s13321-021-00538-8

Contact: ctleungaf@connect.ust.hk AIChE ASC 2024 28th October 2024 Github: github.com/ting2025