

Evaluation

- Test document of 24 pages, where 15 pages with molecules (**254 molecules in total**)
 - **90%** accuracy compared to **75%** of the old parser
- *Matching of smiles is not one-to-one:
 - Current evaluation script ignore brackets because they appear outside CDXML Fragments

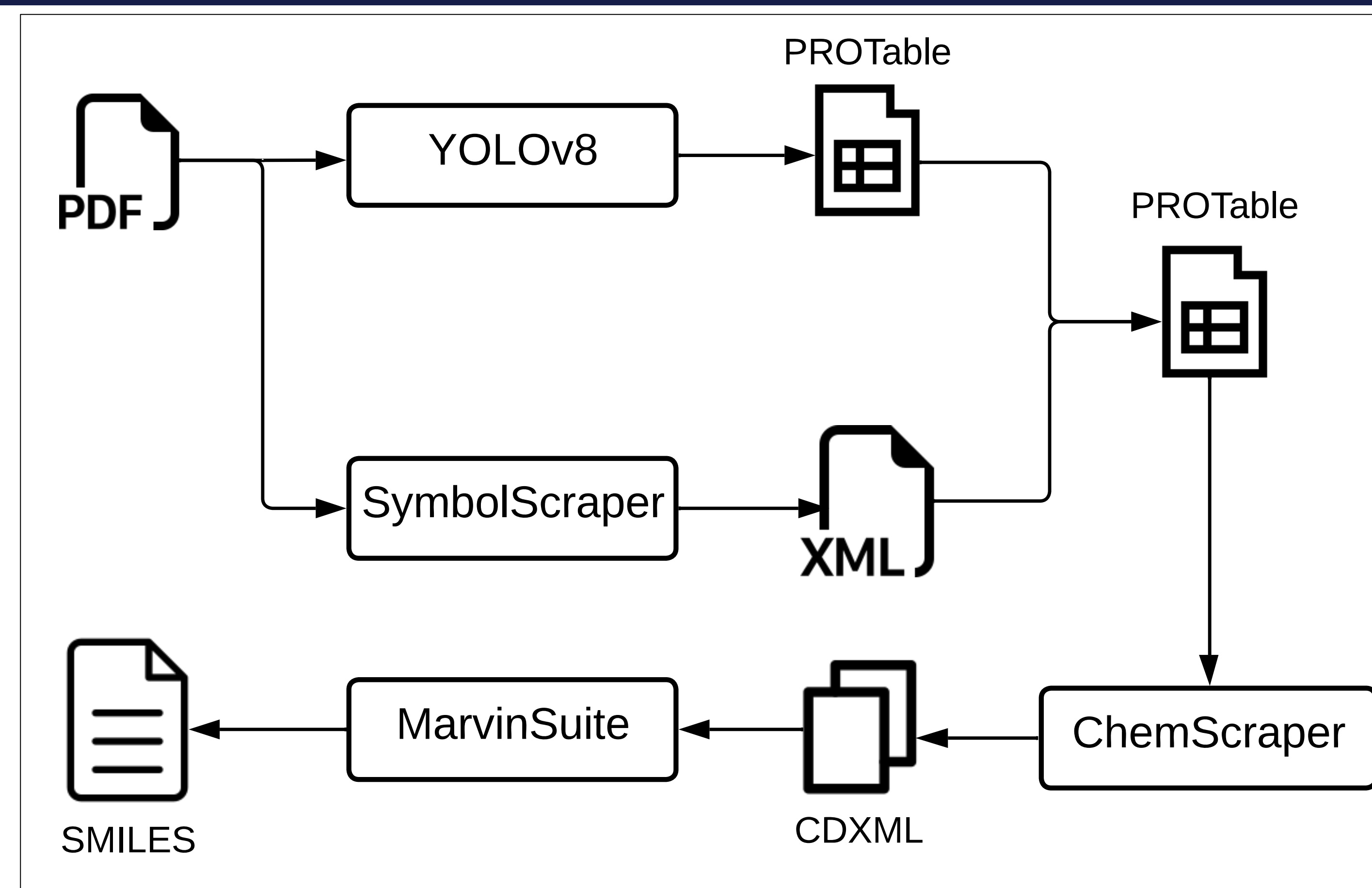
Evaluation Metric	Value
Accuracy of parser	90%
Correctly parsed molecules	229 of 254
Missed molecules (not detected)	13 of 254
Incorrectly parsed molecules	12 of 254
SMILES exact matches*(see above)	56.68%

- **Speed up**
 - Faster PDF to PNG images conversion (**x8 faster**)
 - Molconvert smiles generation (**x30 faster**)
 - Parsing in parallel across pages (**x2 faster**)

Stage	Time (s)
PDF command extraction	6.14
PDF to Image	0.74
YOLOv8 Detection	4.51
Merging YOLOv8 with PDF commands	2.42
Structure Parser	2.08
CDXML Translation	1.19
Extraction Time	17.08
SMILES Generation	0.07
Index Creation	0.004
Total Time	17.15

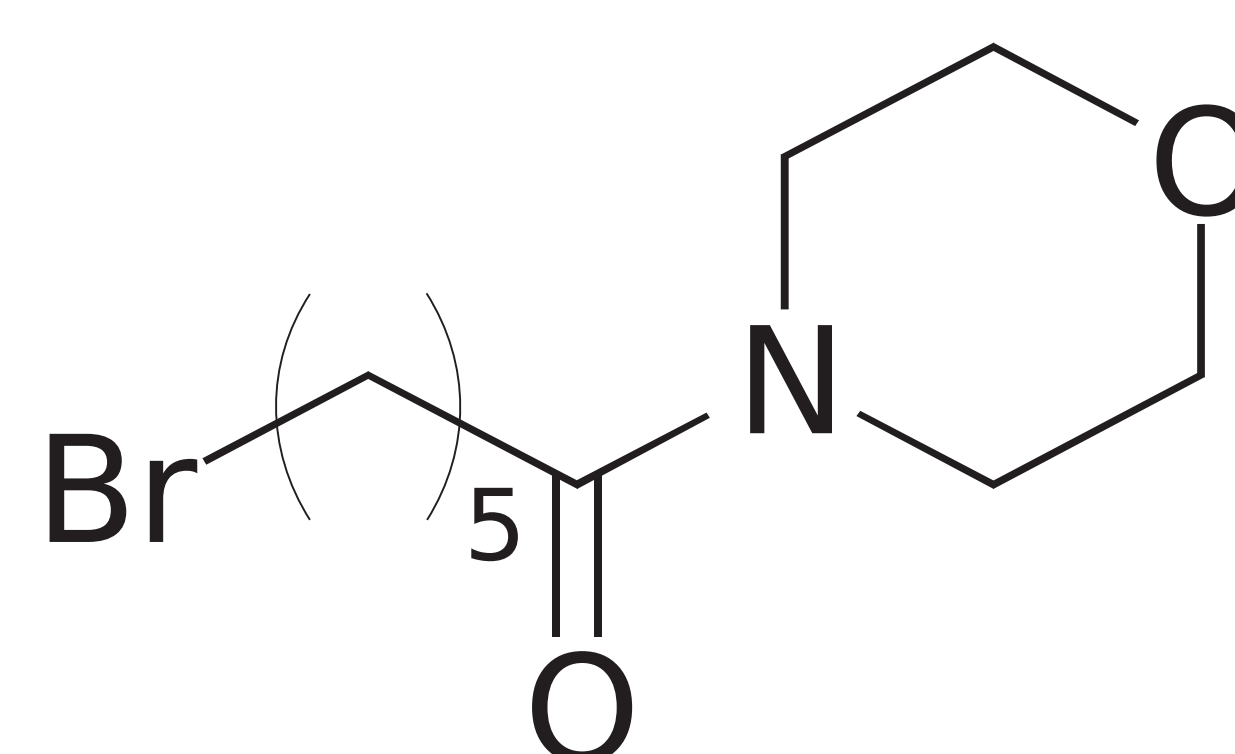
New structures handled

- Repetitive atoms inside brackets (Markush Structures)
- Dot between lines indicating carbons
- Charged atoms, wavy lines

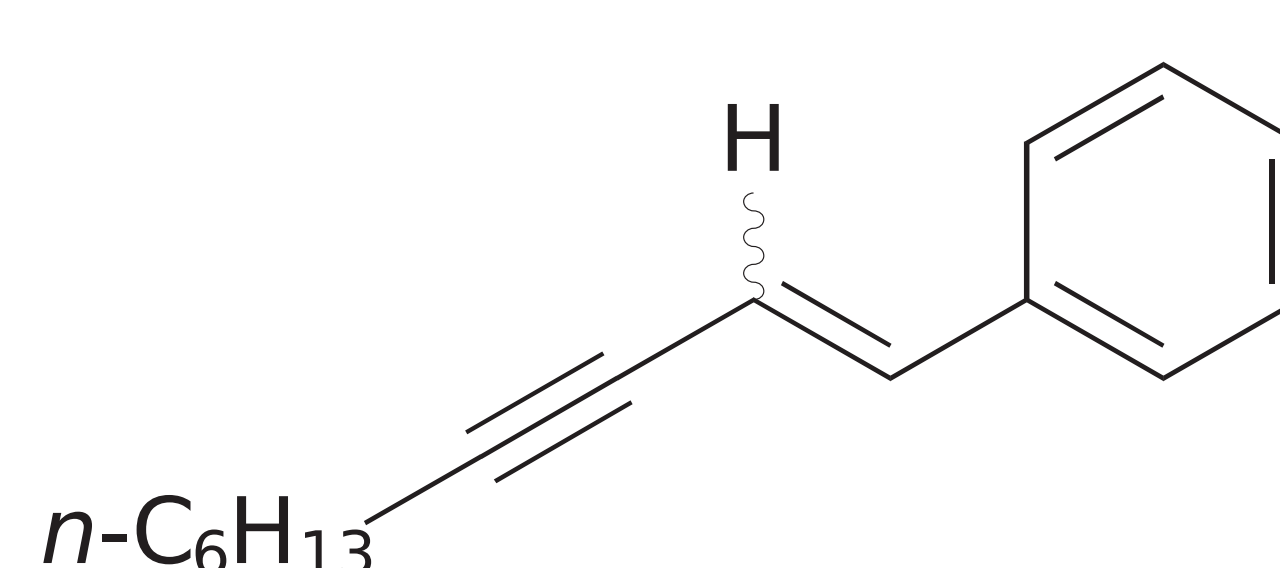


ChemScraper Extraction Pipeline

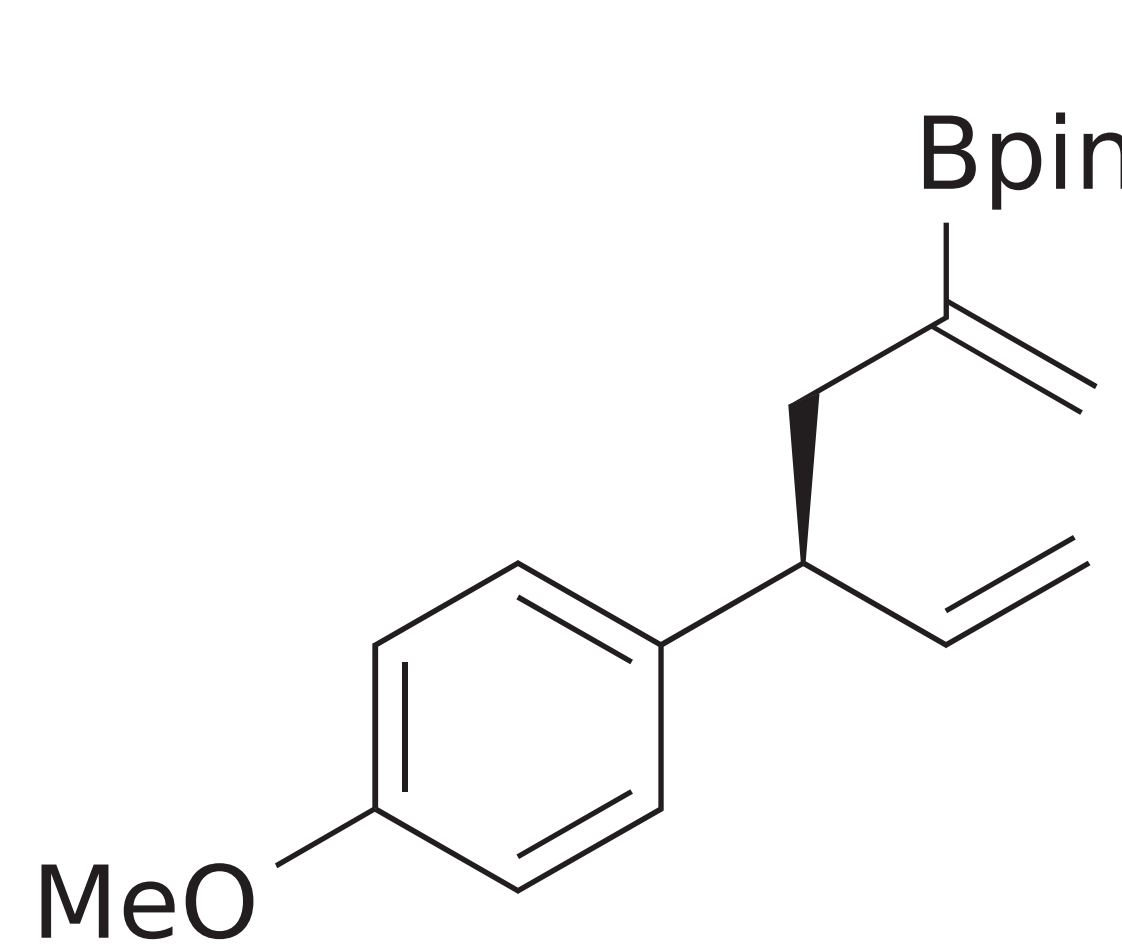
Ground Truth structures



Markush structures

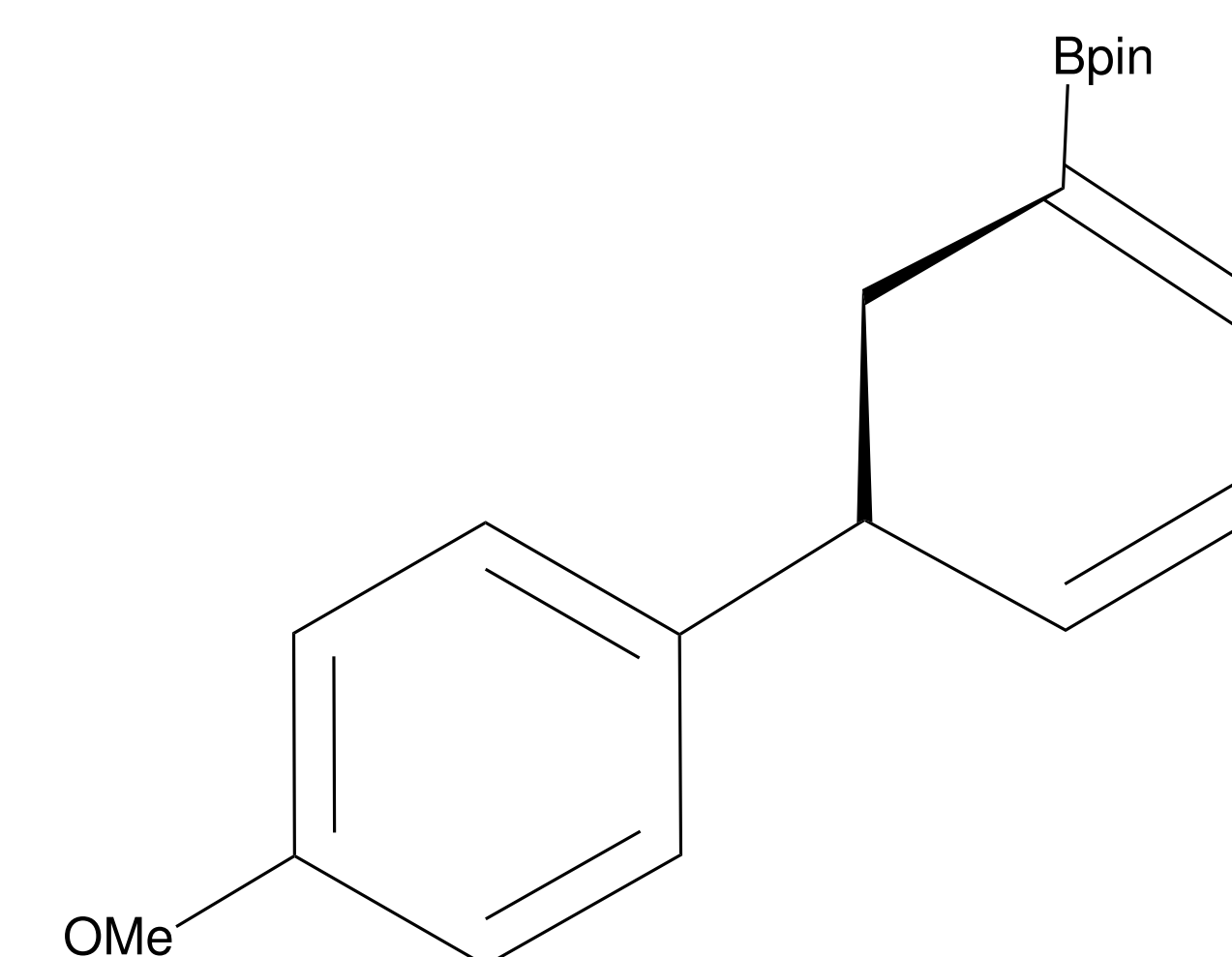
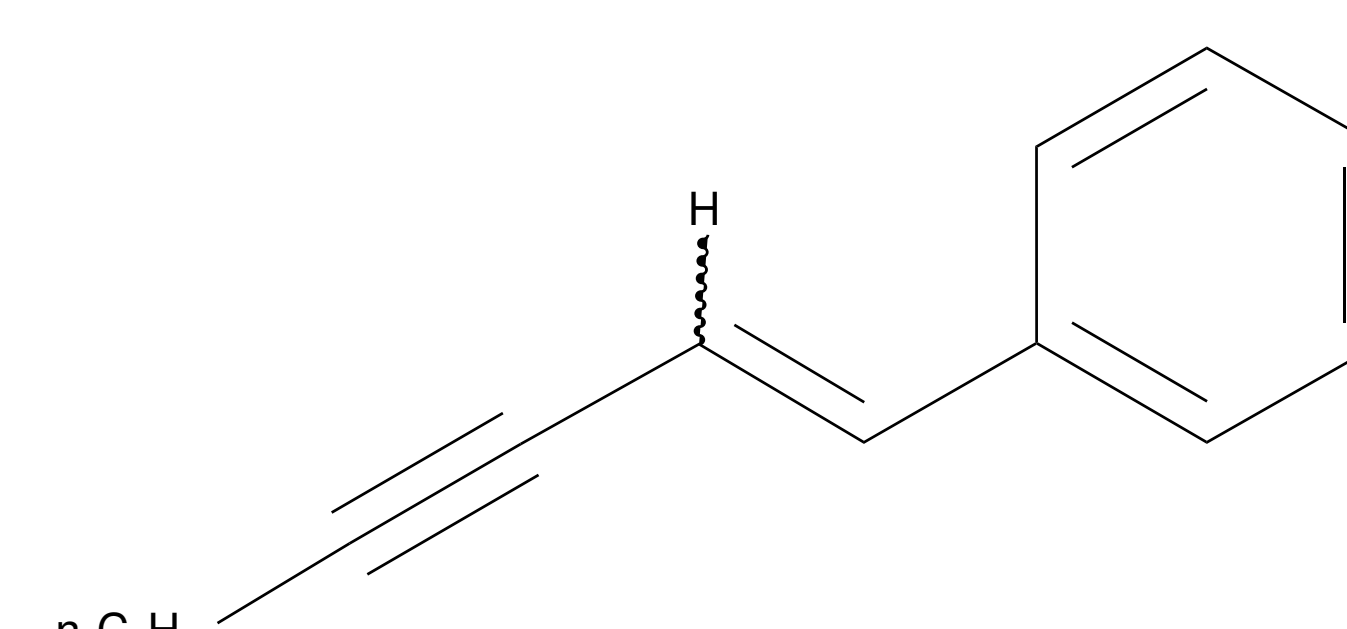
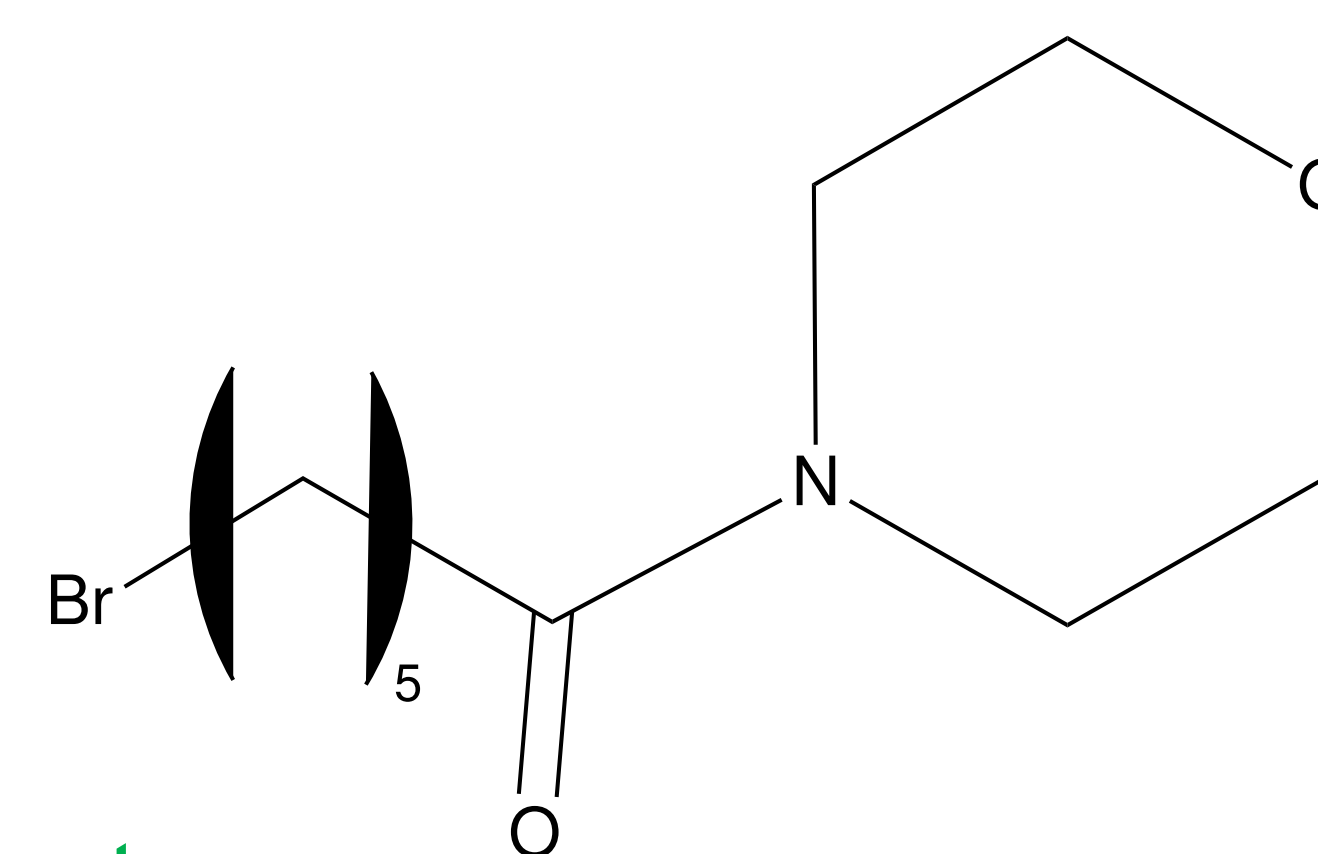


Wavy lines



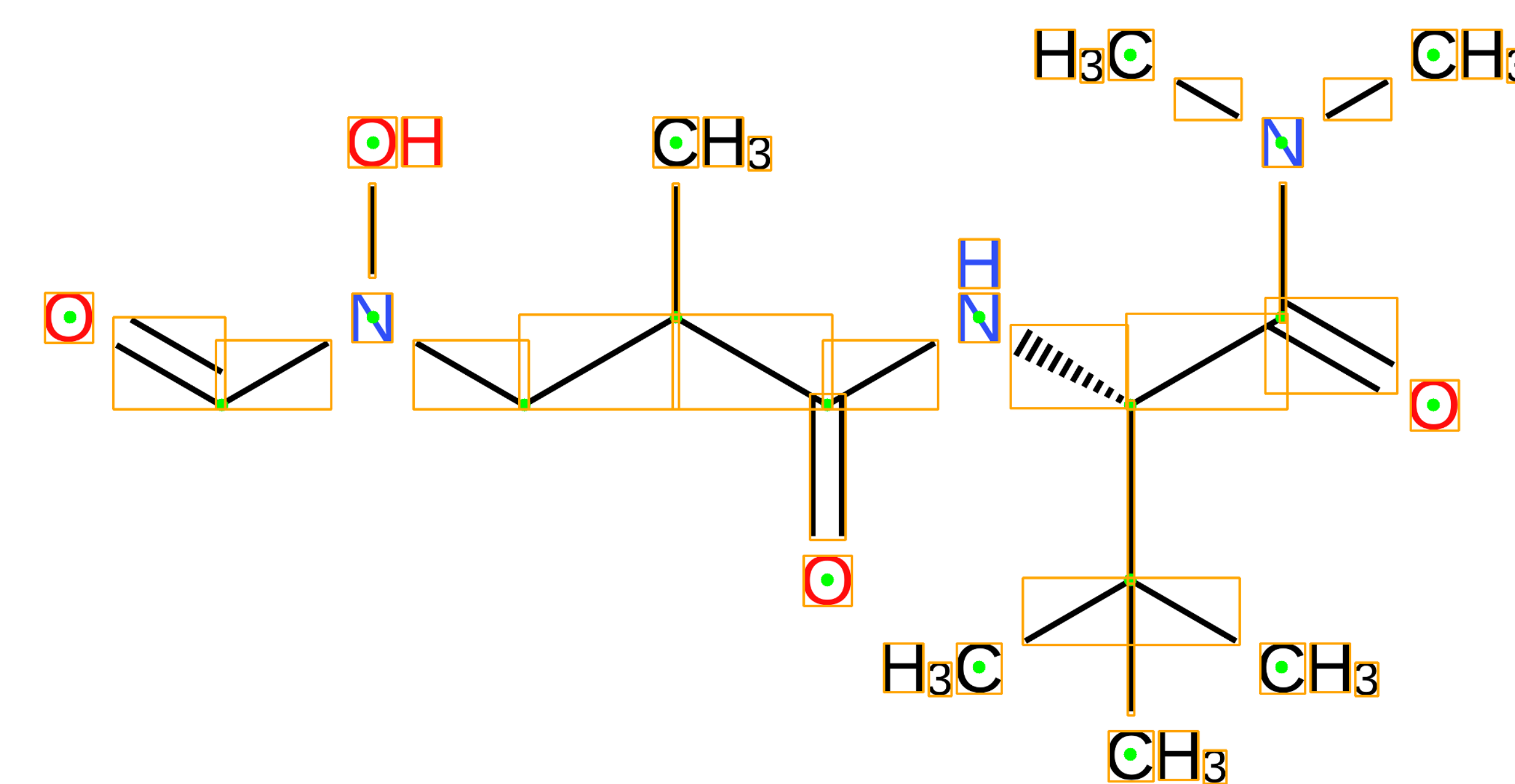
Solid Wedge Error

Predicted structures



Parsing Raster images

- **Dataset:** SMILES from PubChem and Indigo
- Training Dataset creation:
 - Generate PDF and PNG molecule from SMILES
 - **SymbolScraper:** atom labels, atom coordinates and line coordinates from PDF
 - **Indigo Toolkit:** atom labels, bond types, and atom coordinates from PNG
 - Combine both to get labeled graphs with symbols (atoms and bonds) and relationships (connections) for training LGAP^[1]
- LGAP scores symbols, relationships and segmentations using CNN and GNN on LOS graph over primitives
- Training and inference using LGAP to produce output SLT, including bond type information
- Post-processing to produce final SMILES or CDXML



Molecule image with information from SymbolScraper and Indigo Toolkit

In Progress

- Training and postprocessing for raster images with missing/damaged character information
- Address parser errors including solid wedges, merged characters, unrecognized bold lines

[1] Ayush Kumar Shah and Richard Zanibbi. **Line-of-Sight with Graph Attention Parser (LGAP) for Math Formulas**. ICDAR 2023

[2] Ayush Kumar Shah, Abhisek Dey, and Richard Zanibbi. 2021. **A Math Formula Extraction and Evaluation Framework for PDF Documents**, ICDAR 2021

[3] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. **MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation**. JCI 2023