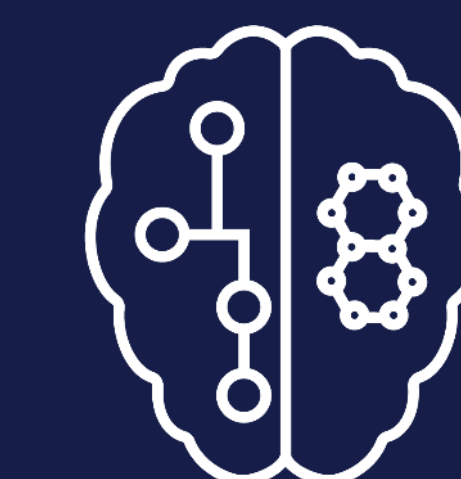




THRUST 1

ChemScraper: Extracting Molecule Diagrams from PDF Vector Images with Page-Level CDXML (ChemDraw) and SMILES Output

Abhisek Dey, Blake Ocampo, Matt Langsenkamp, Ming Creekmore, Bryan Amador, Ayush Kumar Shah, Scott Denmark, and Richard Zanibbi



RIT

Motivation

Create end-to-end pipeline to locate, recognize, and convert molecular diagrams in PDF to page-level CDXML output, and SMILES for individual molecules

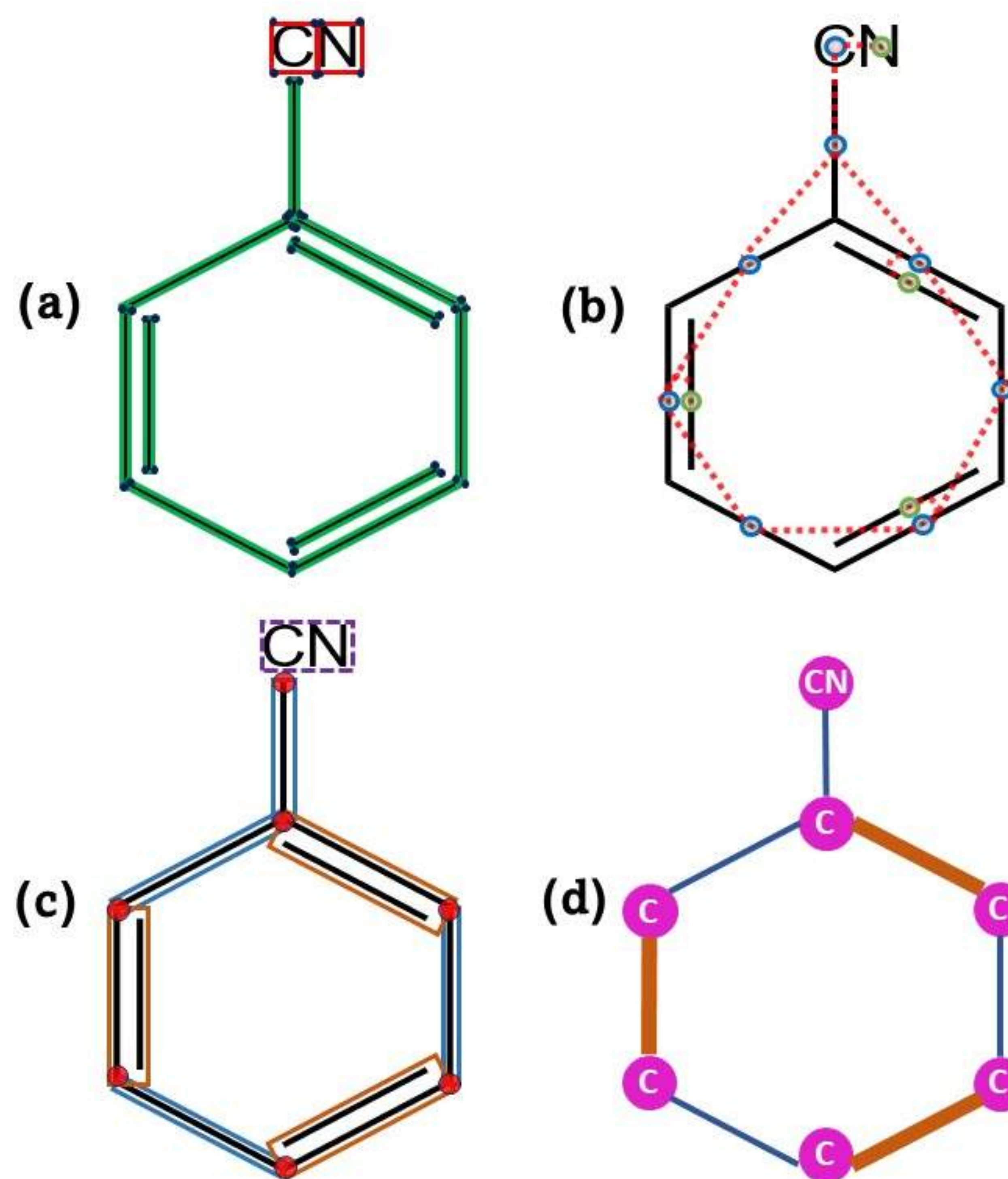
1. Construct/rewrite graphs to parse molecule structure and avoid detailed chemical heuristics
2. Use characters/lines and geometric information directly from PDF instructions where possible

Methodology

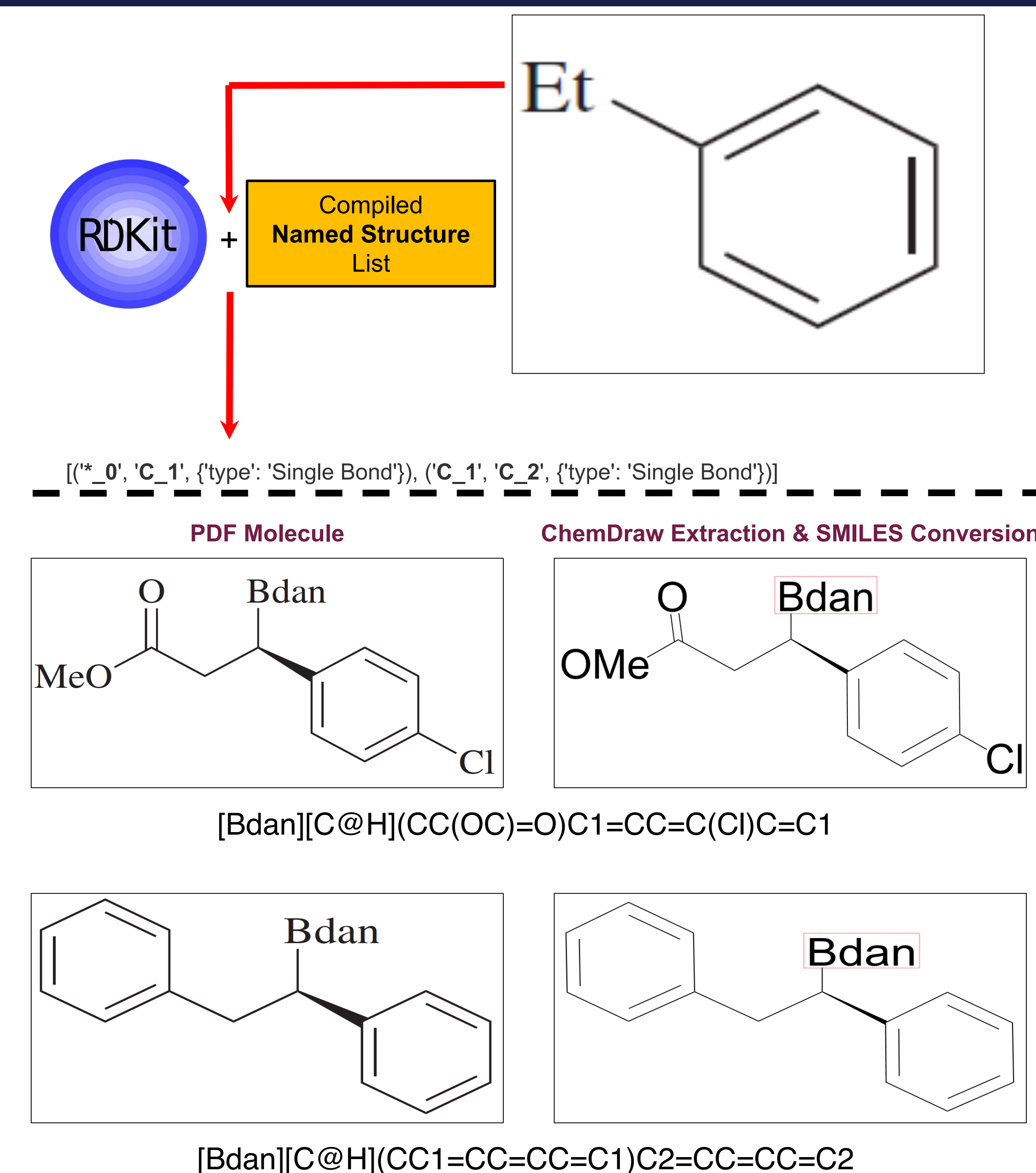
1. Characters and lines from PDF extracted by *SymbolScraper* to create initial graph nodes
2. Minimum Spanning Tree (MST) from line and character distances define initial graph edges
3. Hidden carbons added at line intersections
4. Merge nodes for parallel bond lines and names
5. Additional merges using geometric heuristics
6. CDXML generation: embeds structures for names using a lookup table, page locations

Extraction Highlights: Abbreviations, Chirality & SMILES

- 611 unique expanded structures of common superatom groups identified and compiled
- Solid and Hashed Wedge Bonds identified by checking parallelism and side lengths from the adjacency matrix generated by the MST
- CDXML (ChemDraw) chemdraw files for pages or individual molecules may be viewed/use, and translated to SMILES and other formats (e.g., using `molconvert`)



Stage	Total Time (s)	secs/molecule	secs/page
PDF command ext.	6.5	0.03	0.40
PDF to Image	10.85	0.05	0.68
YOLOv4 Detection	13.15	0.07	0.82
Structure Parser	13.07	0.07	0.82
CDXML Translation	1.46	0.01	0.09
Extraction Time	45.03	0.22	2.81
SMILES Generation	40.19	0.20	2.51
Index Creation	0.03	N/A	N/A
Total Time	85.25	0.42	5.32



Evaluation Metric	Value
Mean Normalized Levenshtein Distance	0.10
Mean Levenshtein Distance	1.94
Exact Matches	62.62%
SMILES Length (Min, Max, Mean)	(4, 30, 13)

In Progress

- Refactoring implementation for speed & robustness
- Adding Markush structures, charged molecules, detecting single-line ('linear') molecules
- Extensions for raster images, missing/damaged character information
- Collaboration w. NCSA on AlphaSynthesis (Kate Arneson, Yashdeep Thorat, Sara Lambert, Leigh Fu, and Matt Berry)

