

# ChemScraper: Pipeline for Parsing Raster and Vector Molecule Diagrams from PDFs

Ayush Kumar Shah<sup>1</sup>, Bryan Amador<sup>1</sup>, Abhisek Dey<sup>1</sup>, Patrick Philipppy<sup>1</sup>, Ming Creekmore<sup>1</sup>, Blake Ocampo<sup>2</sup>, Scott Denmark<sup>2</sup>, Richard Zanibbi<sup>1</sup>

<sup>1</sup>Document and Pattern Recognition Lab, Rochester Institute of Technology;

<sup>2</sup>University of Illinois Urbana-Champaign



Document and Pattern Recognition Lab

## Motivation

Create an end-to-end pipeline to locate, recognize, and convert molecular diagrams in PDF to page-level CDXML output, and SMILES for individual molecules. Can process PDF diagrams embedded as an (i) Image or a (ii) Vector Drawing.

- Born Digital leverages the drawing commands in a PDF allowing faster and accurate parsing
- Visual leverages the connected components in a molecule image to produce a unified output
- Born-digital parser generates training data for visual parser

## Enhancements:

- Faster visual parsing (~2.3 molecules/second on a machine with 1 GPU).
- Faster born digital parsing (~30.8 molecules/second on a machine with 1 GPU)
- Segmentation and relation hypothesis from complete graph instead of line-of-sight graph.
- Handling of atom numbering.

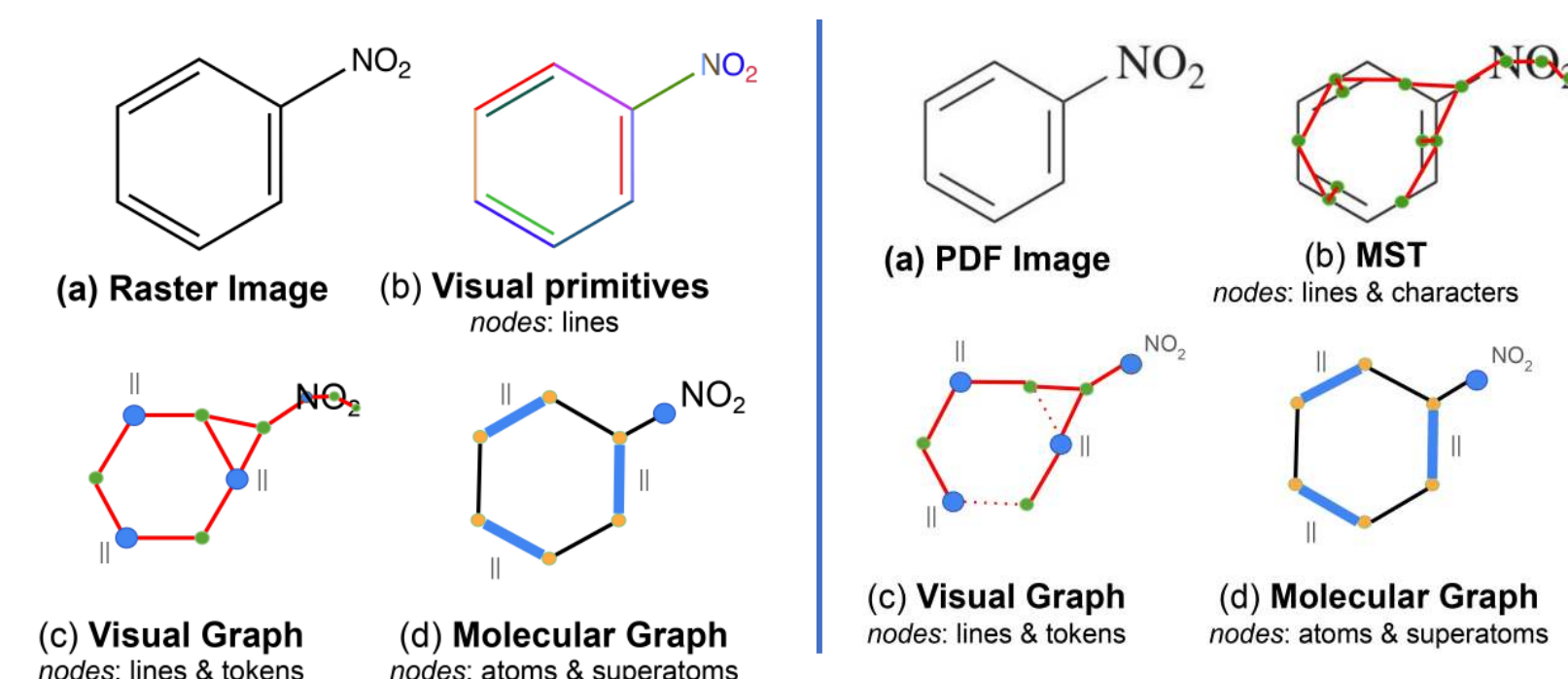


Figure 1. Visual and born-digital parsing stages

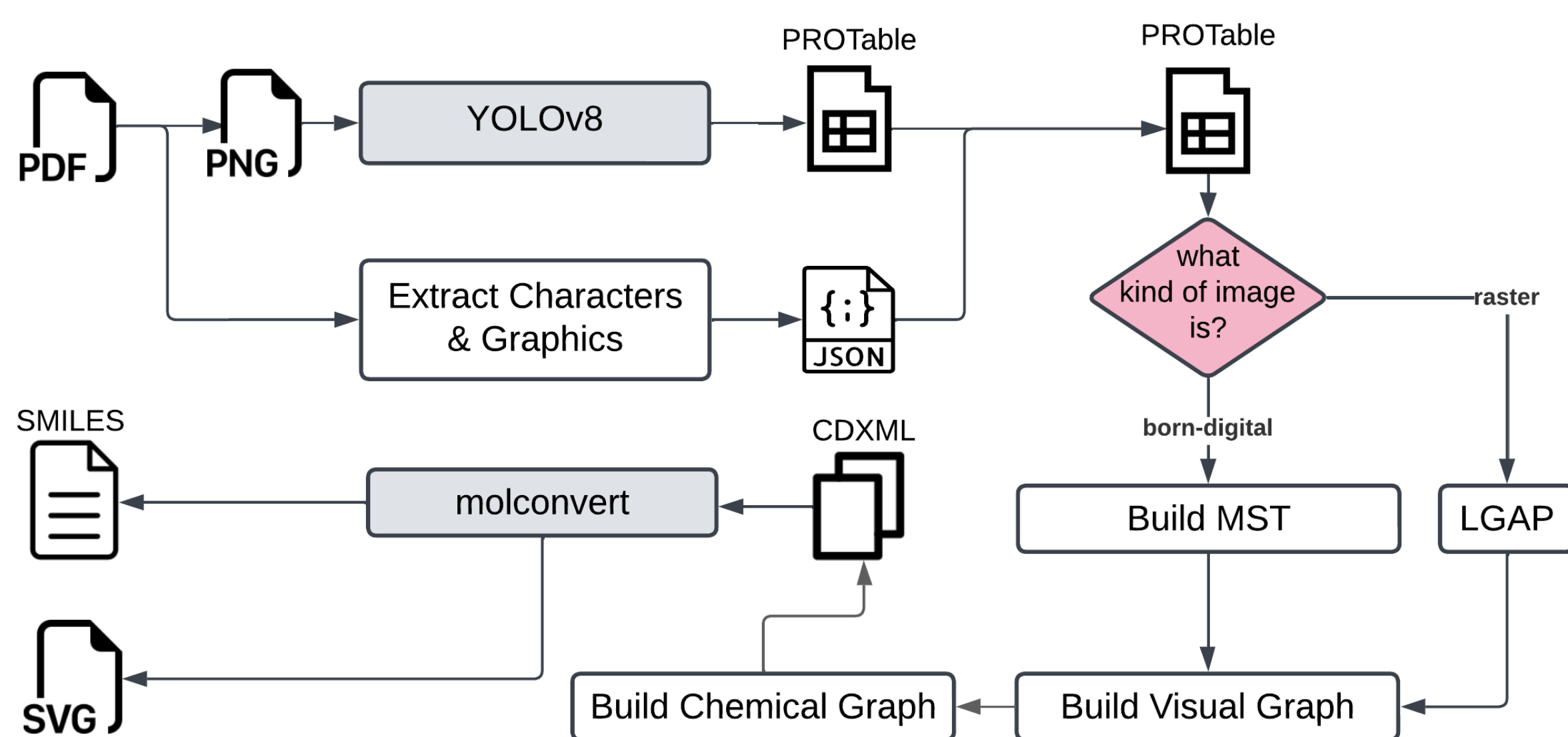


Figure 2. ChemScraper pipeline. Images are processed depending on its nature: vector image (using Born-Digital) or raster image (using LGAP).

Table 1. Born Digital & Visual Benchmarks. Values are percentage of exact SMILES matches.

| Model                               | Indigo-USPTO (5719) | CLEF-2012 (992) |
|-------------------------------------|---------------------|-----------------|
| MolVec 0.9.7                        | 95.40               | 83.80           |
| OSRA 2.1                            | 95.00               | 84.60           |
| Img2Mol                             | 58.90               | 62.70           |
| DECIMER                             | 69.60               | 62.70           |
| SwinOCSR                            | 74.00               | 30.00           |
| MolScribe                           | 97.50               | 88.90           |
| <b>ChemScraper (Born-Digital) *</b> | 98.41               | 90.42           |
| <b>ChemScraper (Visual) *</b>       | 91.00               | -               |

\* Evaluation Data for ChemScraper is a synthetically generated PDF version of the SMILES/Image based Indigo and CLEF

## References

Shah, A.K., Amador, B., Dey, A. et al. ChemScraper: leveraging PDF graphics instructions for molecular diagram parsing. IJDAR 27, 395–414 (2024)  
Shah, A.K., Dey, A., Amador, B., et al. Multimodal Search in Chemical Documents and Reactions. Feb. 24, 2025, arXiv: arXiv:2502.16865.

## Visual Parsing Methodology

Line-of-Sight with Graph Attention Parser (LGAP) uses SE-ResNeXt CNN backbone to generate a visual structure graph:

- Symbol Detection (Segmentation)
- Symbol Classification
- Relationship Classification

Model is run recurrently, updating/merging input primitives until symbol detection is unchanged.

Visual structure graph treats bonds as nodes & predicts relationships using binary classification (connected/non-connected)

## ReactionMiner Integration

Automated matching of recognized molecules with molecules present in reactions extracted by Reaction Miner.

- Molecule reference/id identification around molecule
- String matching with molecules from reactions

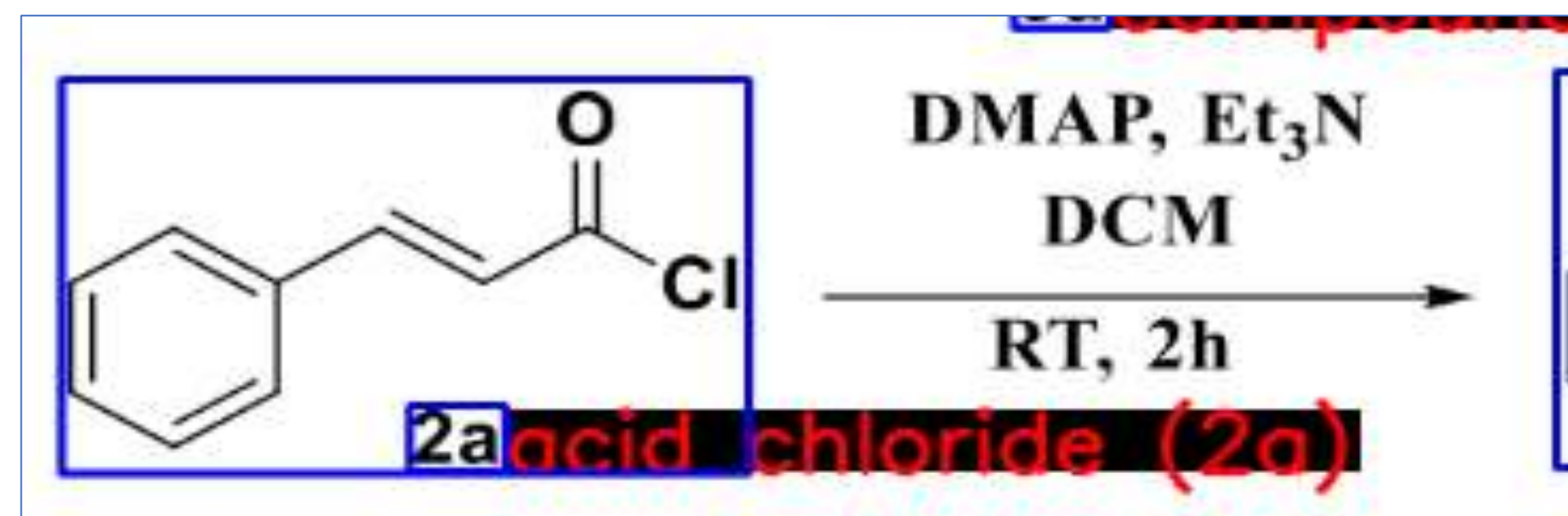


Figure 3. Matching of reactant/product "acid chloride" extracted by ReactionMiner with molecule "2a"

## Conclusions

- ChemScraper is an end-to-end pipeline for parsing vector or raster molecule images in PDFs.
- Tool available at AlphaSynthesis platform: <https://chemscraper.platform.moleculemaker.org/>

