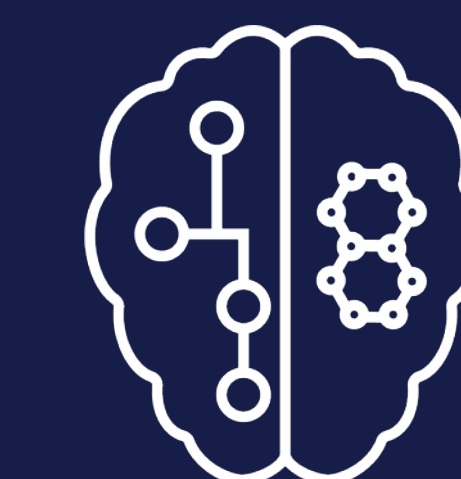




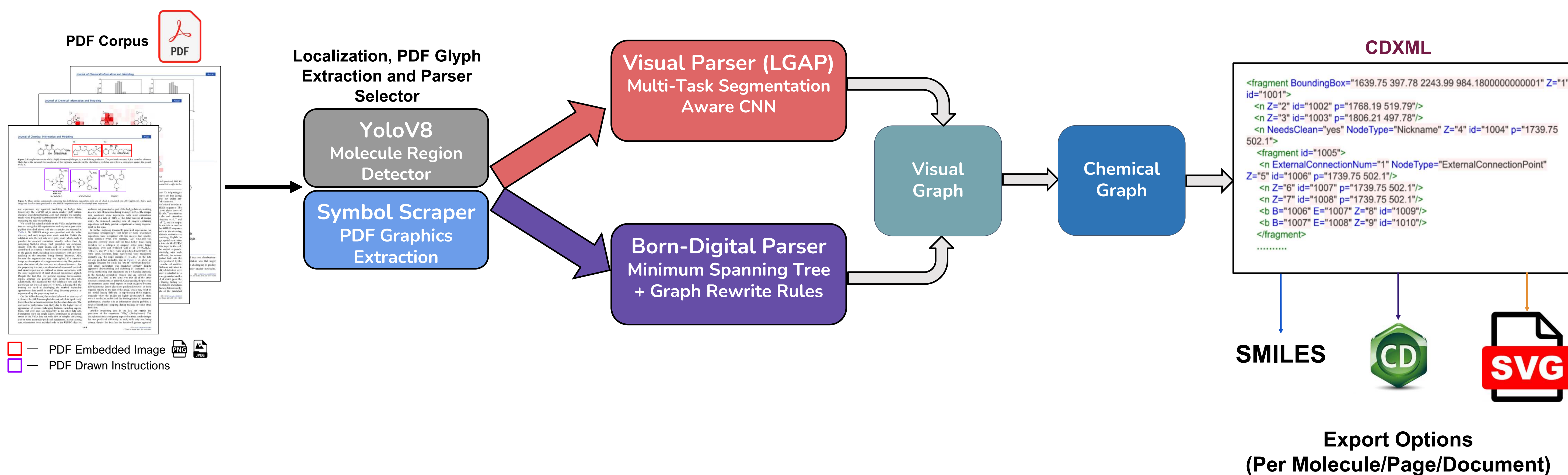
THRUST 1

ChemScraper: Extracting and Parsing Molecular Diagrams from PDFs and Images

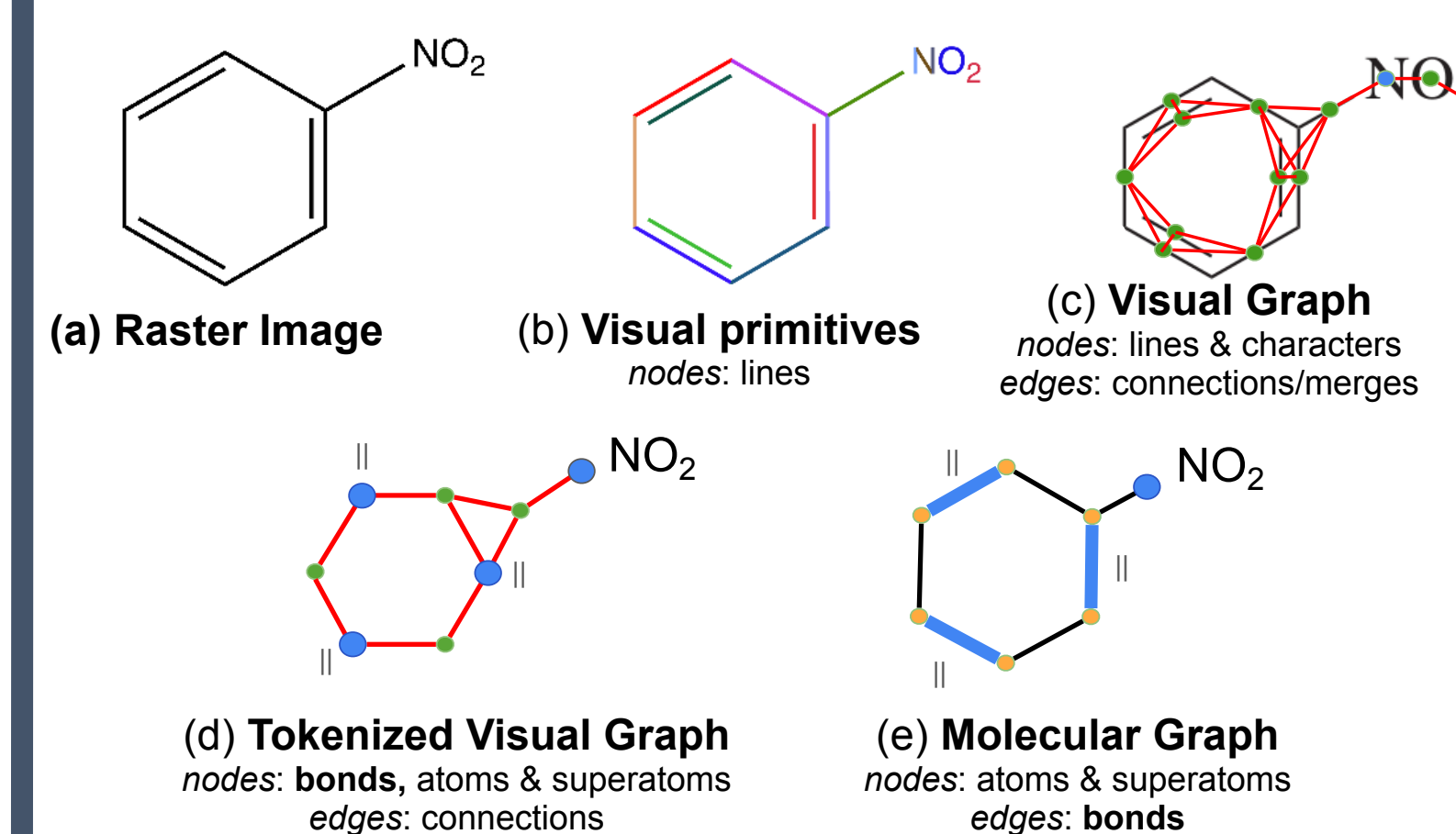
Ayush Kumar Shah, Bryan Amador, Abhisek Dey, Ming Creekmore and Richard Zanibbi
Document and Pattern Recognition Lab, Rochester Institute of Technology



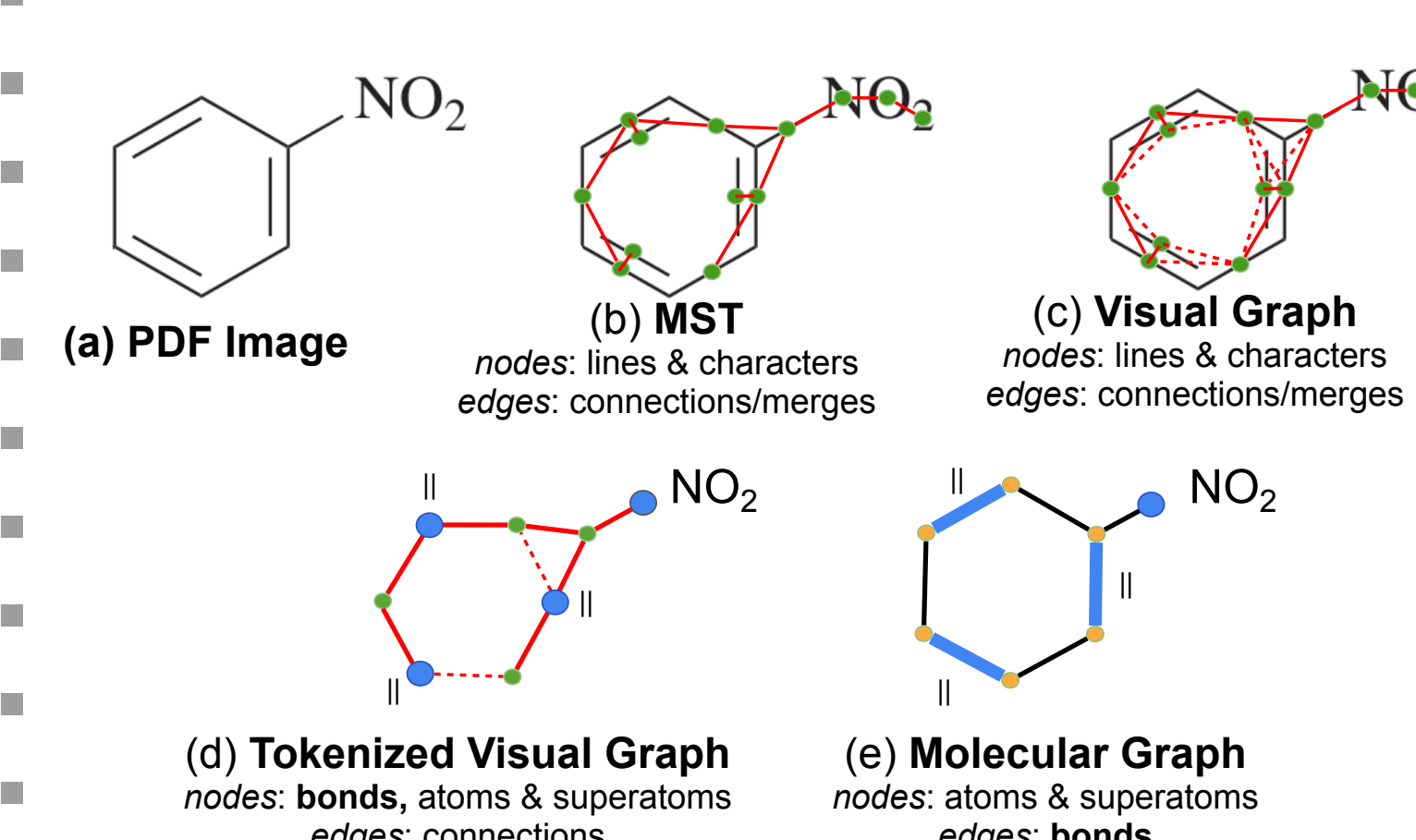
RIT



Visual Primitives + Parsing



Born-Digital Primitives + Parsing



Visual Parsing Methodology

Born-digital parser generates training data for visual parser

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

1. Symbol Detection (Primitive Merging/Segment.)
2. Symbol Classification
3. Relationship Classification

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

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

Conclusion and Future Work

1. ChemScraper is an end-to-end pipeline for extracting and parsing vector or raster molecule images in PDF pages.
2. Collaborating with Matt Berry et al. (**NCSA**) to deploy improved ChemScraper as a part of AlphaSynthesis
3. Collaborating with **Reaction Miner** group to include ChemScraper as a part of Reaction Miner

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

Enhancements:

1. Integrated with **Reaction Miner** to make the recognized SMILES available for extracted reactants and products
 2. Can now extract and parse chemical structures from molecule images using a visual parser
 3. Forwarding of molecule regions to either:
 - a. MST + Rewriting Parser if PDF (**Born Digital**)
 - b. Neural Image Parser if PNG/JPEG (**Visual**)
- 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

Results: Born Digital & Visual on Benchmarks

Exact SMILES Matches	Indigo (5719)	CLEF-2012 (992)	UoB (5740)
MolVec 0.9.7	95.40	83.80	80.60
OSRA 2.1	95.00	84.60	78.50
Img2Mol	58.90	48.84	78.18
DECIMER	69.60	62.70	88.20
SwinOCSR	74.00	30.00	44.90
MolScribe	97.50	88.90	87.90
ChemScraper (Born-Digital)	98.42	96.20	94.90
ChemScraper (Visual)	85.02	-	-

* Evaluation Data for ChemScraper is a synthetically generated PDF (born-digital) and PNG (visual) version of the SMILES/Image based Indigo, CLEF and UoB