



THRUST 1

AI-ENABLED SYNTHESIS
PLANNING



THRUST 4

AI-ENABLED MOLECULE
DISCOVERY

Integration of Reaction Miner and ChemScraper: Mining Chemical Reactions from Literature



MMLI All-Institute Meeting
June 4, 2024

Presented By:

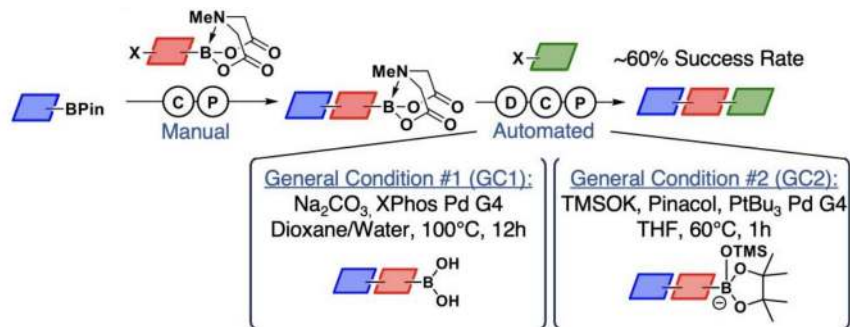
Linyi Ding (Thrust 1)

Ayush Kumar Shah (Thrust 1)

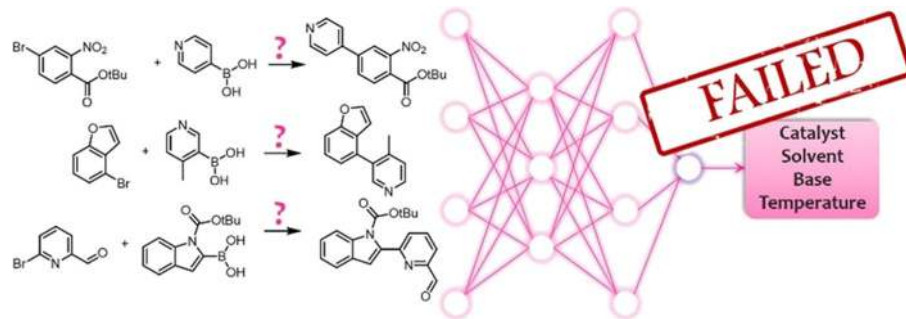
Jason Wu (Thrust 4)

Background

Thrust 4 (OPV)



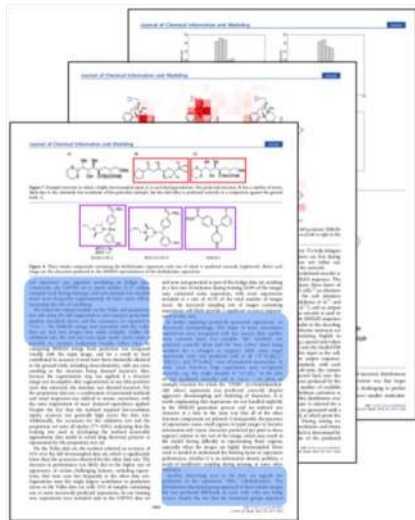
Literature data is insufficient for ML



Large peer-reviewed Suzuki-coupling database will

- Facilitate reaction outcome prediction
- Enable general reaction condition discovery
- Accelerate closed-looped synthesis campaigns

Overview



Dihydroxylation.pdf

 **Scientific Paper**

 Chemical Reaction	
 Product	2-chlorophenol
 Reactants	phenol, oxalyl chloride
 Time	3 hours
 Temperature	10 °C
 Solvent	NaOH solution
 Yield	78% (2-chlorophenol)
 Workup Reagent	6 M HCl
 Speed	500 rpm
 Inhibitor	BHT
 Light Condition	dark
 PH	13.2
 Cooling Condition	ice bath

Reaction Miner Output



CDXML

```
<fragment BoundingBox="1639.75 397.78 2243.99 984.1800000000001" Z="1" id="1001">
  <n Z="2" id="1002" p="1768.19 519.79">
    <n Z="3" id="1003" p="1806.21 497.78">
      <n NeedsClean="yes" NodeType="Nickname" Z="4" id="1004" p="1739.75 502.1">
        <fragment id="1005">
          <n ExternalConnectionNum="1" NodeType="ExternalConnectionPoint" Z="5" id="1006" p="1739.75 502.1">
            <n Z="6" id="1007" p="1739.75 502.1">
              <n Z="7" id="1008" p="1739.75 502.1">
                <b B="1006" E="1007" Z="8" id="1009">
                  <b B="1007" E="1008" Z="9" id="1010">
                    </fragment>
                  </fragment>
                </fragment>
              </fragment>
            </fragment>
          </fragment>
        </fragment>
      </fragment>
    </fragment>
  </fragment>
```

↓

SMILES  

ChemScraper Output

Chemical Reaction		Reference molecule	SMILES
Product	2-chlorophenol	Dihydroxylation_Page2_No005	Oc1ccccc1Cl
Reactant1	phenol	Dihydroxylation_Page2_No003	Oc1ccccc1
Reactant2	oxalyl chloride	Dihydroxylation_Page2_No004	ClC(=O)C(=O)Cl

Combined Output

Reaction Miner

- Task Definition: Chemical Reaction Extraction

- Input: **PDF file** of a scientific paper
- Output: **All structured** chemical reactions



Scientific Paper

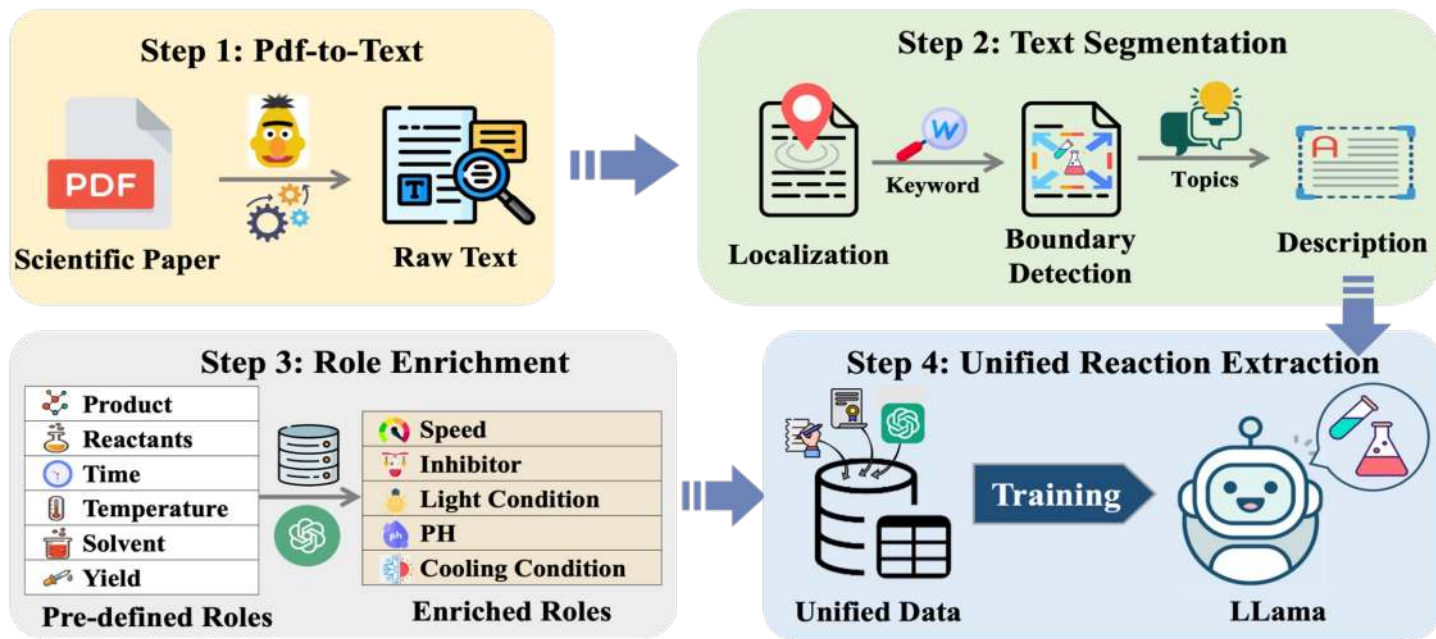


 Chemical Reaction		
	Product	2-chlorophenol
	Reactants	phenol, oxalyl chloride
	Time	3 hours
	Temperature	10 °C
	Solvent	NaOH solution
	Yield	78% (2-chlorophenol)
	Workup Reagent	6 M HCl
	Speed	500 rpm
	Inhibitor	BHT
	Light Condition	dark
	PH	13.2
	Cooling Condition	ice bath

Reaction Miner

- Workflow

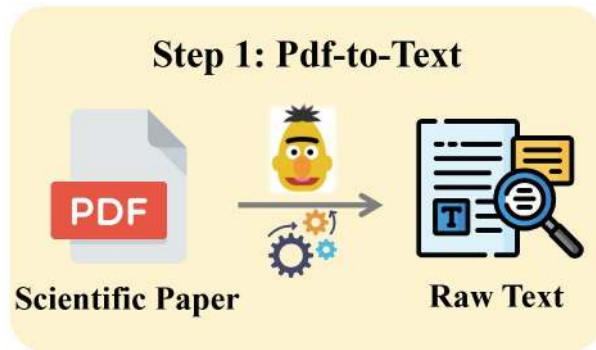
- PDF-to-text → Text Segmentation → Role Enrichment → Reaction Extraction



Reaction Miner

- Step 1: PDF-to-Text

- No powerful PDF-to-Text tool for chemistry papers
 - Templates for CS conference papers are limited
 - Templates for chemistry journal papers are much more diverse
- 1: Convert PDF into XML and excludes the figures/tables.
- 2: Maintain the anchor embedding obtained by Sentence-BERT to solve the **incoherence** issue (filter out irrelevant information like header, footer)



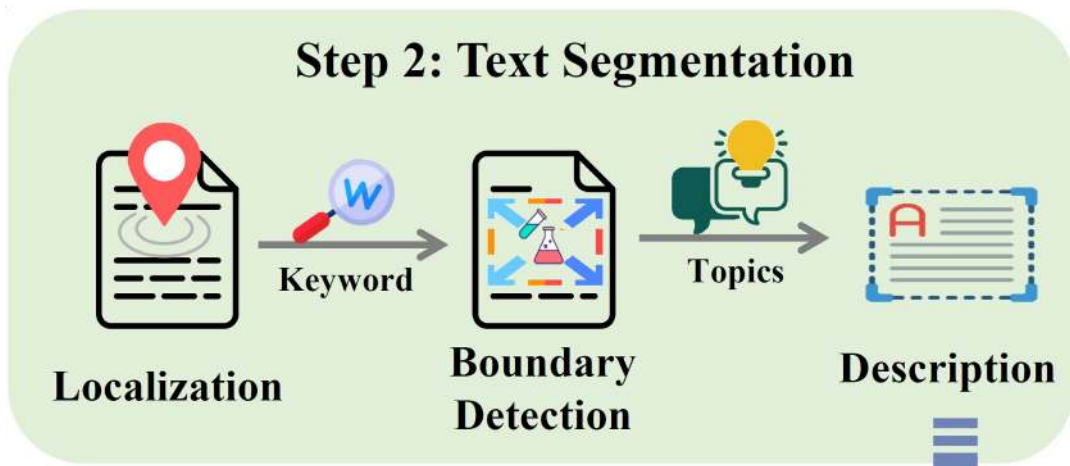
Reaction Miner

- Step 2: Text Segmentation

- Keyword-based central sentence localization (e.g., yields)

*“Removal of the TBS moiety of 17 was carried out with TBAF/AcOH in MeCN at 60 °C to give diol 20 in 86% **yield**”*

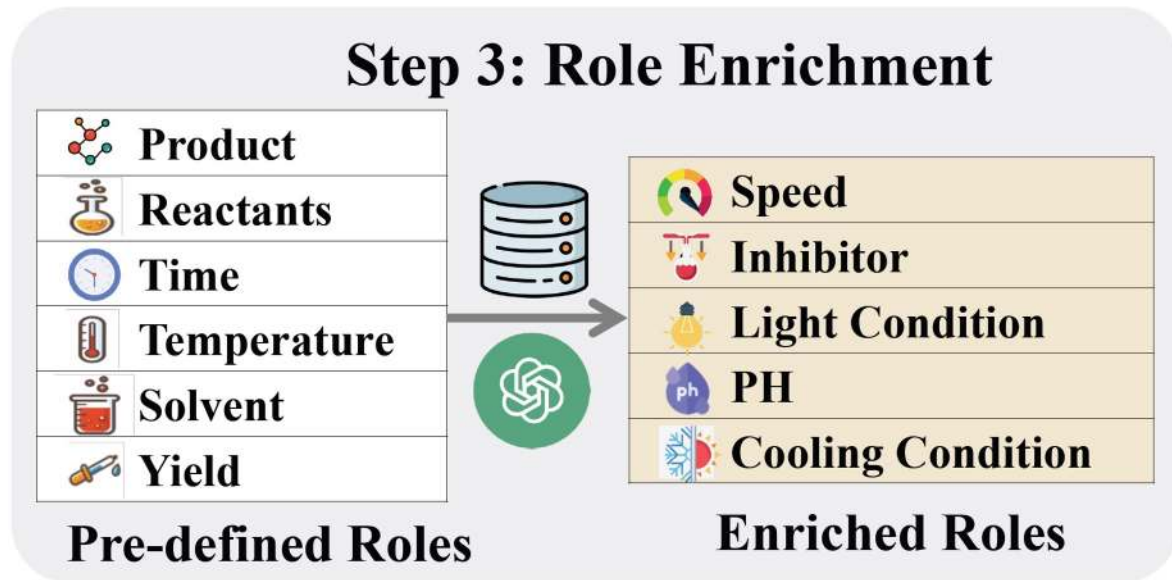
- Topic-aided boundary detection



Reaction Miner

- Step 3: Role Enrichment

- Previous work only extracted a small number of predefined chemical reaction roles
- We use event mining^[1] to automatically enrich more reaction roles to be extracted



Reaction Miner

- Step 4: Unified Reaction Extraction

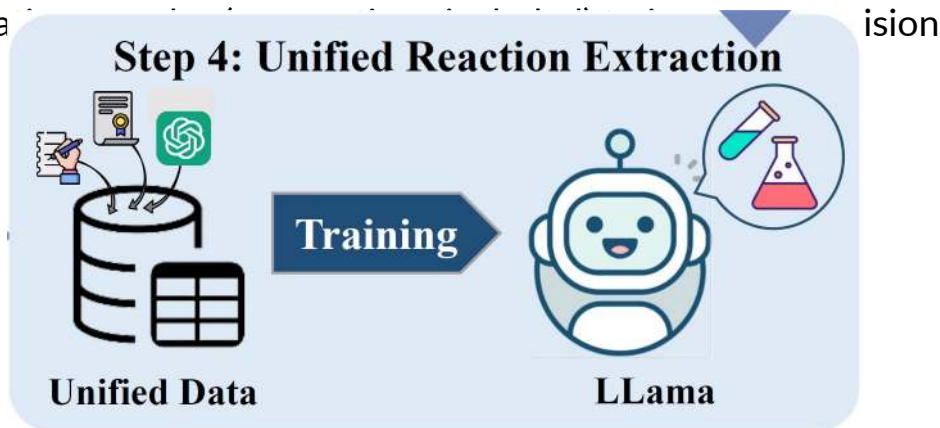
- Various data source

- Synthetic data obtained using pattern mining

Frequent patterns: [Chem] was obtained ; produced [Chem] ; be transformed to [Chem]

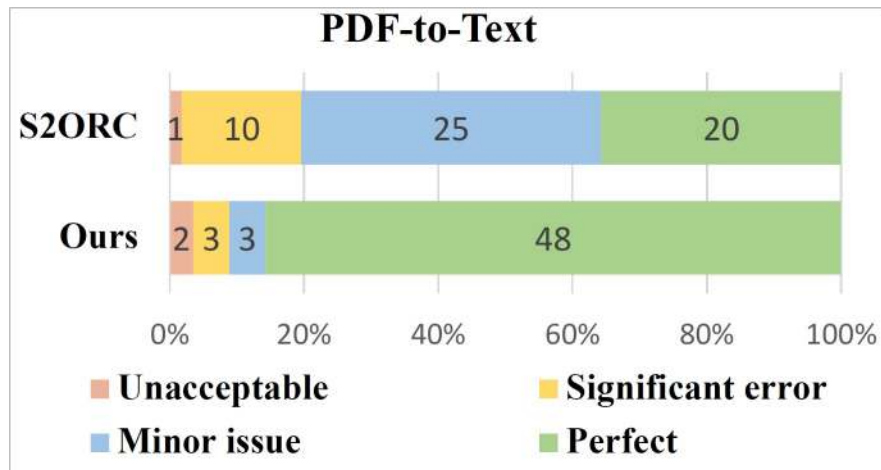
- Patent literature (well-formatted)

- Negative



Reaction Miner

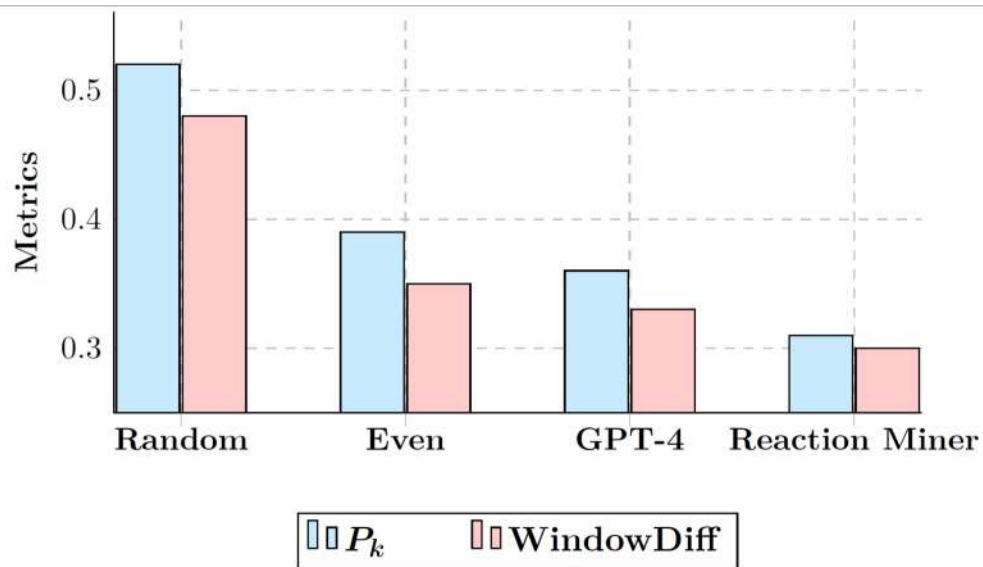
- Experimental Result for PDF-to-Text



- ❑ S2ORC is the most popular pdf-to-text tool developed by Allen AI
 - ❑ Only 35% of pdf conversion is perfect in the chemistry domain
- ❑ Our tool: convert the text perfectly in 86% samples

Reaction Miner

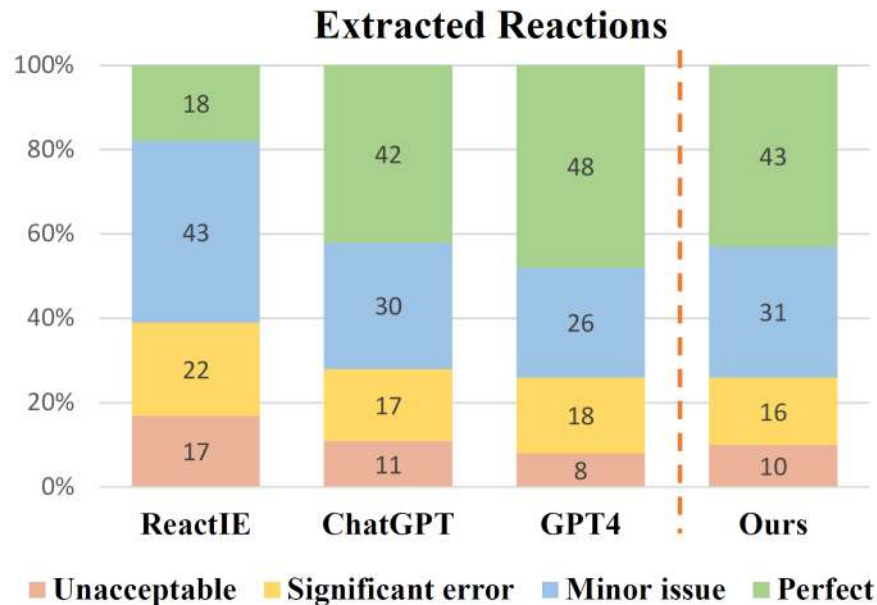
- Experimental Result for Text Segmentation



- Lower value indicates better performance
- Reaction Miner > GPT-4

Reaction Miner

- Experimental Result for Reaction Extraction



- Reaction Miner is comparable to ChatGPT and GPT-4
- Notably, Reaction Miner has much less parameters than GPTs

Case Study

Input text: Further, compound 5a was subjected to bromination reaction by employing N-bromosuccinimide in DCM solvent at room temperature, however an expected yield was not obtained and many side products were observed. Therefore, we modified the reaction condition by taking 1 equ. of N-bromosuccinimide as well as 1.1 equ. of Et3N at 0 0C which lead to a higher yield of compound 6a.

Reaction Miner (perfect)

Product : compound 6a,

Reactant : compound 5a, N-bromosuccinimide, Et3N

Solvent : DCM,

Temperature : 0 0C

Yield : higher (compound 6a)

GPT 4 (error)

Product : compound 6a,

Reactant : compound 5a (missing reactant)

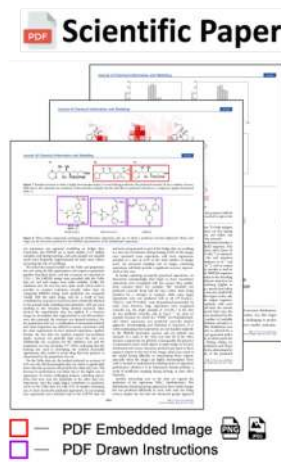
Solvent: dcm solvent,

Temperature: room temperature initially, later modified to 0°C (confused with wrong settings)

Yield: not obtained for initial reaction, higher yield obtained after modifying reaction condition

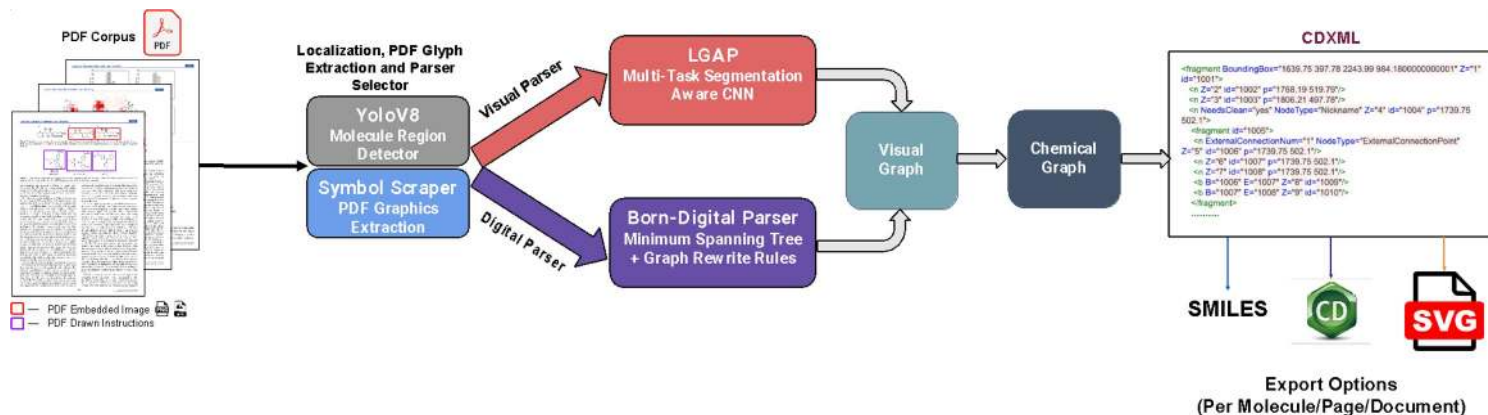
ChemScraper: Extraction of Molecules

- Task Definition: Extraction of Chemical Diagrams from PDFs
 - Input: **PDF file** of a scientific paper
 - Output: **All molecule CDXMLs/SMILES**



ChemScraper: Extraction of Molecules

1. **YOLOv8 Detector:** detect molecule regions
2. **Symbol Scraper:** extract character and graphics from PDF encodings
3. **Graph Construction and Transformations:**
 - **Born-digital parser:** extracts visual graph from PDF symbols
 - **Visual parser:** multi-task CNN (LGAP) extracts visual graph from Image primitives
4. **Chemical graph generation**



[1] Shah, A. K., Amador, B. M., Dey, A., Creekmore, M., Ocampo, B., Denmark, S., & Zanibbi, R. (2024). *ChemScraper: Leveraging PDF Graphics Instructions for Molecular Diagram Parsing* (arXiv:2311.12161). arXiv. <https://doi.org/10.48550/arXiv.2311.12161>, accepted at ICDAR 2024

[2] Shah, A. K., & Zanibbi, R. (2023). Line-of-Sight with Graph Attention Parser (LGAP) for Math Formulas. In G. A. Fink, R. Jain, K. Kise, & R. Zanibbi (Eds.), *Document Analysis and Recognition—ICDAR 2023* (pp. 401–419). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41734-4_25

ChemScraper: Benchmarks

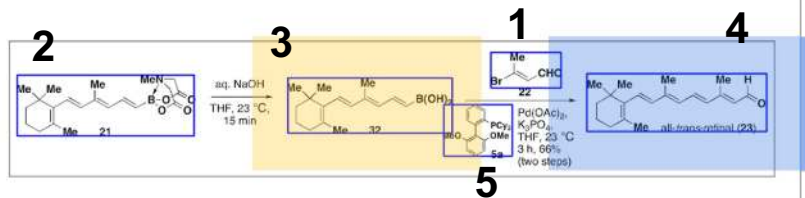
Exact SMILES Matches	Indigo (5719)	CLEF-2012 (992)	UoB (5740)
MolVec 0.9.7	95.40	83.80	80.60
OSRA 2.1	95.00	84.60	78.50
Img2Mol	58.90	48.84	78.18
DECIMER	69.60	62.70	88.20
SwinOCSR	74.00	30.00	44.90
MolScribe	97.50	88.90	87.90
ChemScraper (Born-Digital - PDF) *	98.42	96.20	94.90
ChemScraper (Visual)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

ChemScraper: Integration with Reaction Miner

IR (KBr Pellet, cm^{-1})

3021, 2959, 2925, 2865, 1773, 1457, 1338, 1301, 1025, 986, 867.



All-trans-retinal (23)

MIDA boronate **21** was converted to boronic acid **32** via the following procedure: In a 7 mL Wheaton vial, to a stirred solution of **21** (35.9 mg, 0.101 mmol, 1.0 eq.) in THF (1.44 mL) at 23 °C was added 1 M aq. NaOH (0.30 mL, 0.30 mmol, 3.0 eq.) and the resulting mixture was stirred for 15 min. The reaction was then quenched with the addition of 0.5 M pH 7 phosphate buffer (1.5 mL) and diluted with Et₂O (1.5 mL). The

Paper9_Sla.pdf

Reaction	Reaction Miner Dictionary	
1	Product	all-trans-retinal (23)
	Reactant1	32
	Reactant2	22
	Yield	66% (all-trans-retinal (23))

Reaction Miner Output

Document	Page	Molecule	Predicted SMILES
Paper9_Sla	1	1	<chem>Br/C(C)=C/C=O</chem>
		2	<chem>CC1(C)C(/C=C/C(C)=C/C=C/B(OC(C2)=O)OC(CN2C)=O)=C(C)CCC1</chem>
		3	<chem>CC1(C)C(/C=C/C(C)=C/C=C/B(O)O)=C(C)CCC1</chem>
		4	<chem>CC1(C)C(/C=C/C(C)=C/C=C/C(C)=C/C=O)=C(C)CCC1</chem>
		5	<chem>COC1=CC=CC(OC)=C1C2=C(P(C3CCCCC3)C4CCCCC4)C=CC=C2</chem>

ChemScraper Output

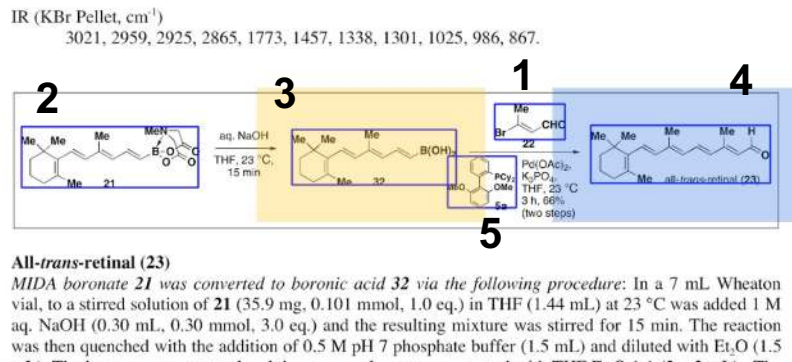
ChemScraper: Integration with Reaction Miner

Algorithm:

- For each detected molecule, find word candidates within a window (5% of document height) with distance from the molecule

Page	Molecule	Word	Distance
1	4	Me	0.00
		Me	0.00
		H	0.00
		...	
		all-trans-retinal (23)	0.00
1	3	Me	0.00
		...	
		B(OH) ₂	0.00

Word Candidates for molecules



Paper9_S1a.pdf

Reaction	Reaction Miner Dictionary	
1	Product	all-trans-retinal (23)
	Reactant1	32
	Reactant2	22
	Yield	66% (all-trans-retinal (23))

Reaction Miner Output

ChemScraper: Integration with Reaction Miner

Algorithm:

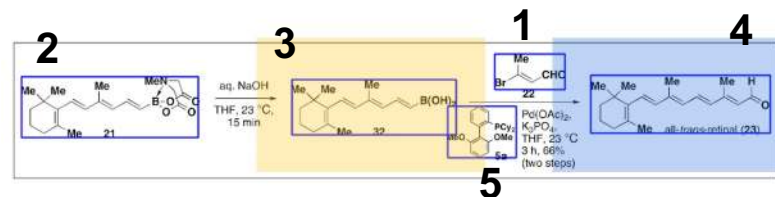
2. For each **word query (Reactant, Product, Catalyst)** in the reaction miner output dictionary,
 - a. Compute word similarity with each word candidates obtained in Step 1
 - b. Select the molecule which has a candidate word with maximum similarity and minimum distance

Page	Molecule	Word	Distance	Similarity
1	4	Me	0.00	8.33
		Me	0.00	8.33
		H	0.00	0.00
		...		
		all-trans-retinal (23)	0.00	100.00
1	3	Me	0.00	8.33
		...		
		B(OH)2	0.00	14.28

Query: all-trans-retinal (23)

IR (KBr Pellet, cm⁻¹)

3021, 2959, 2925, 2865, 1773, 1457, 1338, 1301, 1025, 986, 867.



All-trans-retinal (23)

MIDA boronate **21** was converted to boronic acid **32** via the following procedure: In a 7 mL Wheaton vial, to a stirred solution of **21** (35.9 mg, 0.101 mmol, 1.0 eq.) in THF (1.44 mL) at 23 °C was added 1 M aq. NaOH (0.30 mL, 0.30 mmol, 3.0 eq.) and the resulting mixture was stirred for 15 min. The reaction was then quenched with the addition of 0.5 M pH 7 phosphate buffer (1.5 mL) and diluted with Et₂O (1.5 mL). The mixture was then extracted with Et₂O (1.5 mL) and the organic layer was dried with Na₂SO₄ and concentrated under reduced pressure to give **32** as a white solid.

Paper9_S1a.pdf

Reaction	Reaction Miner Dictionary	
1	Product	all-trans-retinal (23)
	Reactant1	32
	Reactant2	22
	Yield	66% (all-trans-retinal (23))

Reaction Miner Output

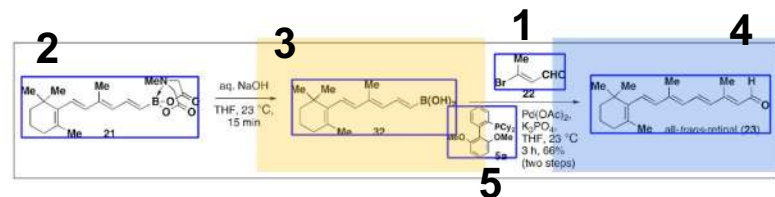
ChemScraper: Integration with Reaction Miner

Page	Molecule	Word	Distance	Similarity
1	4	Me	0.00	8.33
		Me	0.00	8.33
		H	0.00	0.00
		...		
		all-trans-retinal (23)	0.00	100.00
1	3	Me	0.00	8.33
		...		
		B(OH)2	0.00	14.28

Query: all-trans-retinal (23)

IR (KBr Pellet, cm⁻¹)

3021, 2959, 2925, 2865, 1773, 1457, 1338, 1301, 1025, 986, 867.



All-trans-retinal (23)

MIDA boronate **21** was converted to boronic acid **32** via the following procedure: In a 7 mL Wheaton vial, to a stirred solution of **21** (35.9 mg, 0.101 mmol, 1.0 eq.) in THF (1.44 mL) at 23 °C was added 1 M aq. NaOH (0.30 mL, 0.30 mmol, 3.0 eq.) and the resulting mixture was stirred for 15 min. The reaction was then quenched with the addition of 0.5 M pH 7 phosphate buffer (1.5 mL) and diluted with Et₂O (1.5 mL). The mixture was then extracted with Et₂O (3 x 1 mL) and the combined organic layers were dried over anhydrous Na₂SO₄ and concentrated under reduced pressure to give **32** as a white solid (10.5 mg, 0.033 mmol, 33% yield).

Paper9_Sla.pdf

Chemical Reaction 1		Reference molecule	SMILES
Product	all-trans-retinal (23)	Paper9_Sla_Page1_No004	CC1(C)C(/C=C/C(C)=C/C=C/C(C)=C/C=O)=C(C)CCC1
Reactant1	32	Paper9_Sla_Page1_No003	CC1(C)C(/C=C/C(C)=C/C=C/B(O)O)=C(C)CCC1
Reactant2	22	Paper9_Sla_Page1_No001	Br/C(C)=C/C=O
Yield	66% (all-trans-retinal (23))	-	-

Final combined output

Conclusion

- Reaction Miner extracts structured chemical reactions from scientific pdfs
- Reaction Miner achieves satisfactory outcome in 75% cases
- ChemScraper produces molecular structure graphs in CDXML, which can be used with ChemDraw, Marvin, and converted to other representations (SMILES, MOL, InChI) easily
- ChemScraper obtains 98.4% and 85% recognition on the synthetic USPTO benchmark
- ChemScraper annotates predicted SMILES/CDXML for reactants, products and other molecules extracted by Reaction Miner
- The integrated tool automates mining of chemical reactions from literature

Future Work

- Improve robustness of ChemScraper born-digital and visual parser (data augmentation)
- Improve formula alignment with candidate words for more complex cases

Thank you!

We thank Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao, Jiawei Han who contributed to Reaction Miner

We thank Matt Langsenkamp, Matt Berry, Kate Arneson, other members in the NCSA team, Bryan Amador, Abhisek Dey, Ming Creekmore, Blake Ocampo, Scott Denmark, Richard Zanibbi who contributed to the ChemScraper system