



MOLECULE MAKER'S PLAYHOUSE

THRUST 1 AI-ENABLED SYNTHESIS PLANNING

Reaction Extraction from PDF: ChemScraper and
ReactionMiner Integration

Ayush Kumar Shah, as1211@rit.edu
Rochester Institute of Technology (RIT)

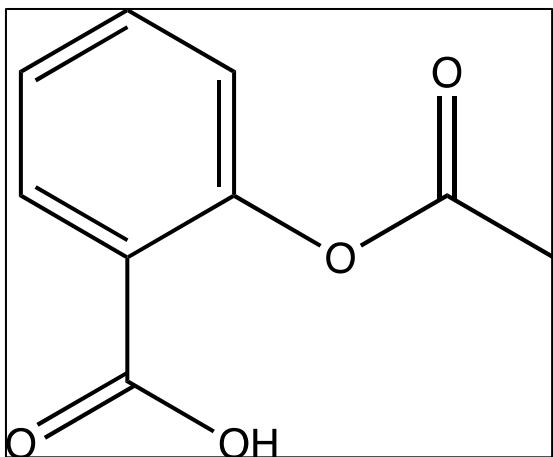
Jan 28, 2025



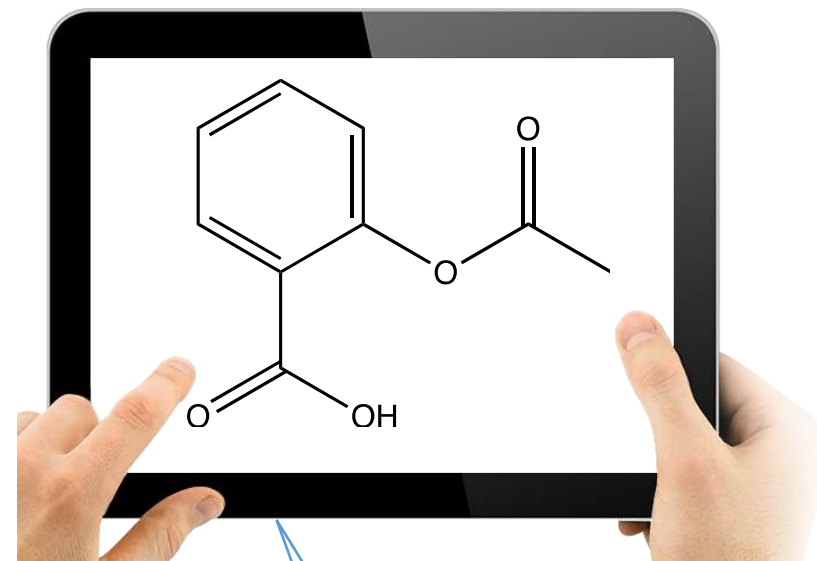
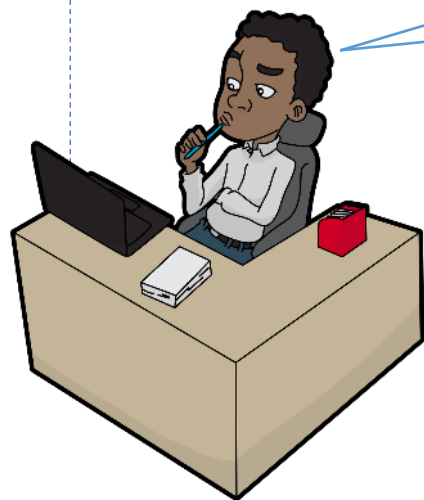
MOLECULE
MAKER LAB
INSTITUTE

Introduction

Motivation



Which catalysts can improve the yield of of this chemical (aspirin)?



Let's find reactions producing this molecule.

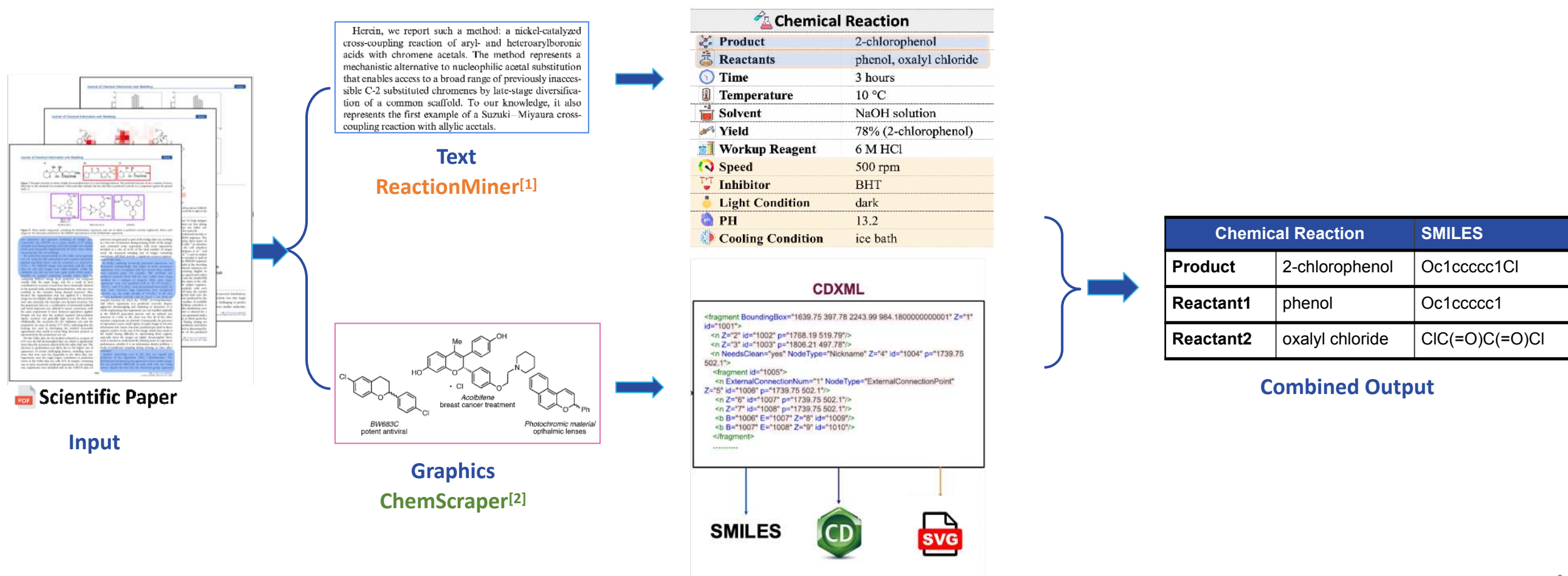
ChemScraper and Reaction Miner Integration

Task: Extract reactions from documents

1. **ReactionMiner:** Extracts **structured chemical reaction** information from literature (**document PDF text**)

2. **ChemScraper:** Parses **CDXML/SMILES** from literature (**document molecule diagrams**)

Integration: Match ReactionMiner information to parsed ChemScraper molecule diagrams (SMILES)



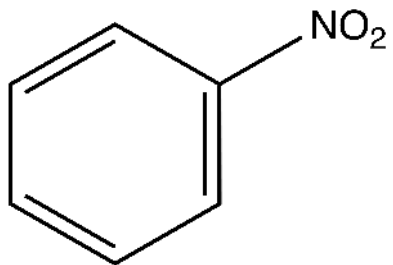
[1] Zhong et al., "Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data". In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 389-402).

[2] Shah et al., "ChemScraper: leveraging PDF graphics instructions for molecular diagram parsing." IJDAR, vol. 27, no. 3, pp. 395-414, Sep. 2024, doi: 10.1007/s10032-024-00486-7.

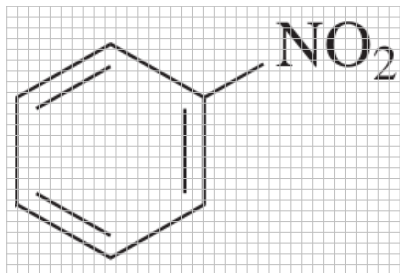
ChemScraper: Molecular Diagram Parsing

ChemScraper: Molecular Diagram Parsing

Vector (e.g. PDF)

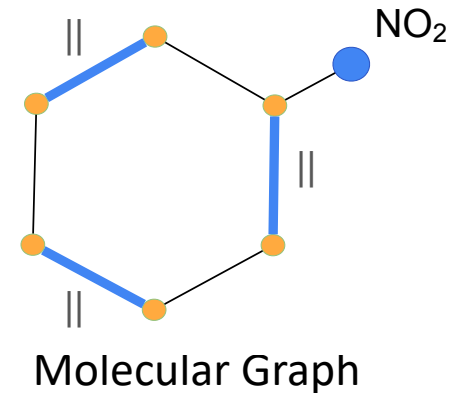


Raster (e.g. PNG)

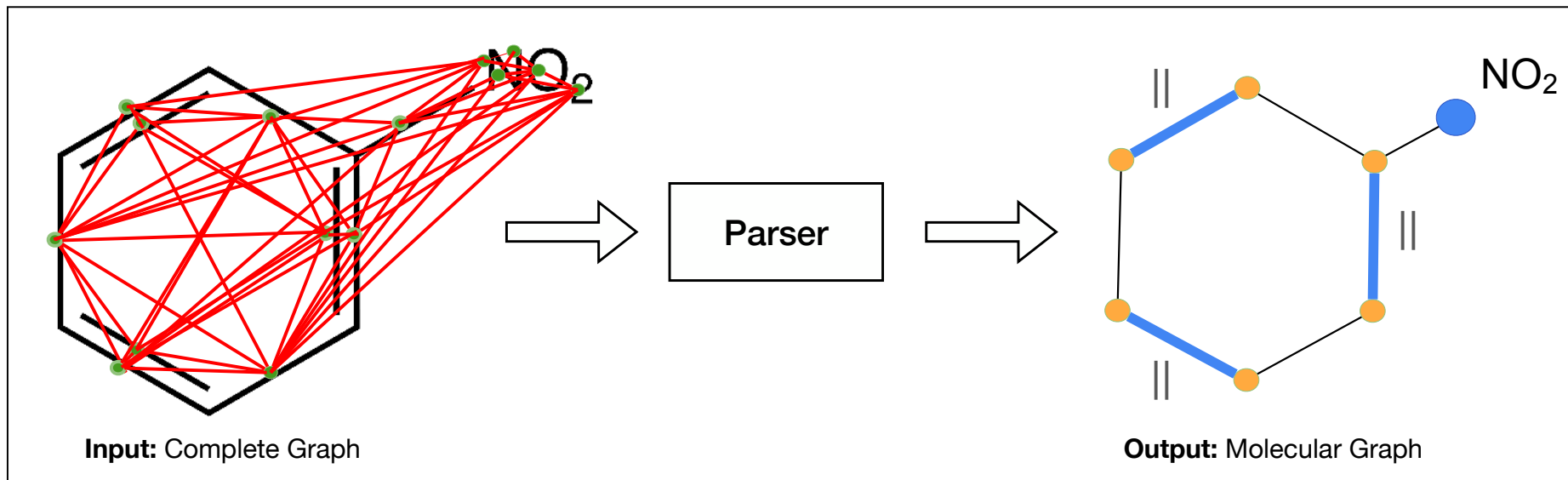


"C1=CC=C(C=C1)[N+](=O)[O-]"

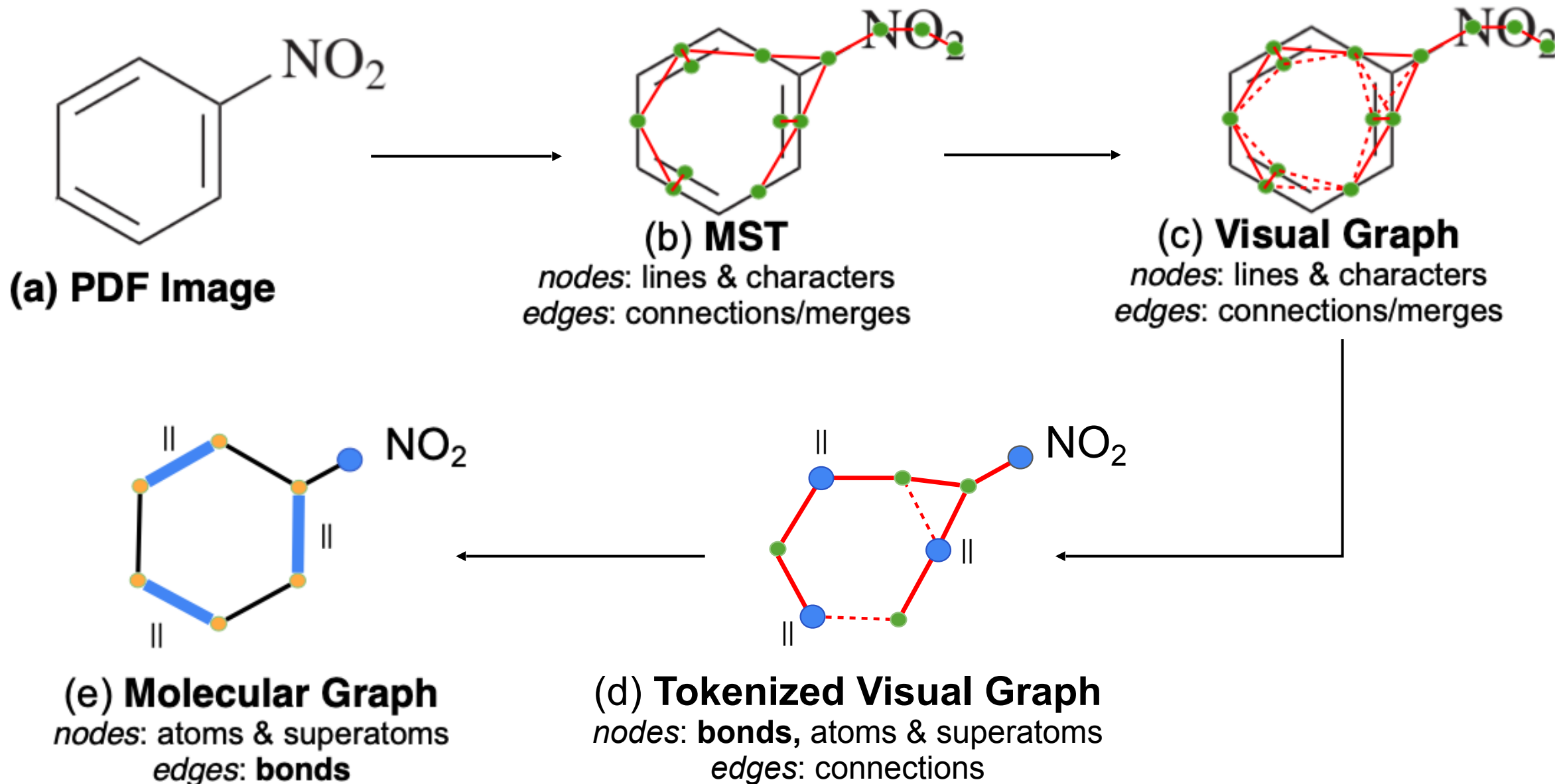
SMILES String



ChemScraper Parsing Model: graph decoder to convert input graph on primitives to output molecular graph



Born-digital Parser: PDF Molecule Image



ChemScraper and Reaction Miner Integration

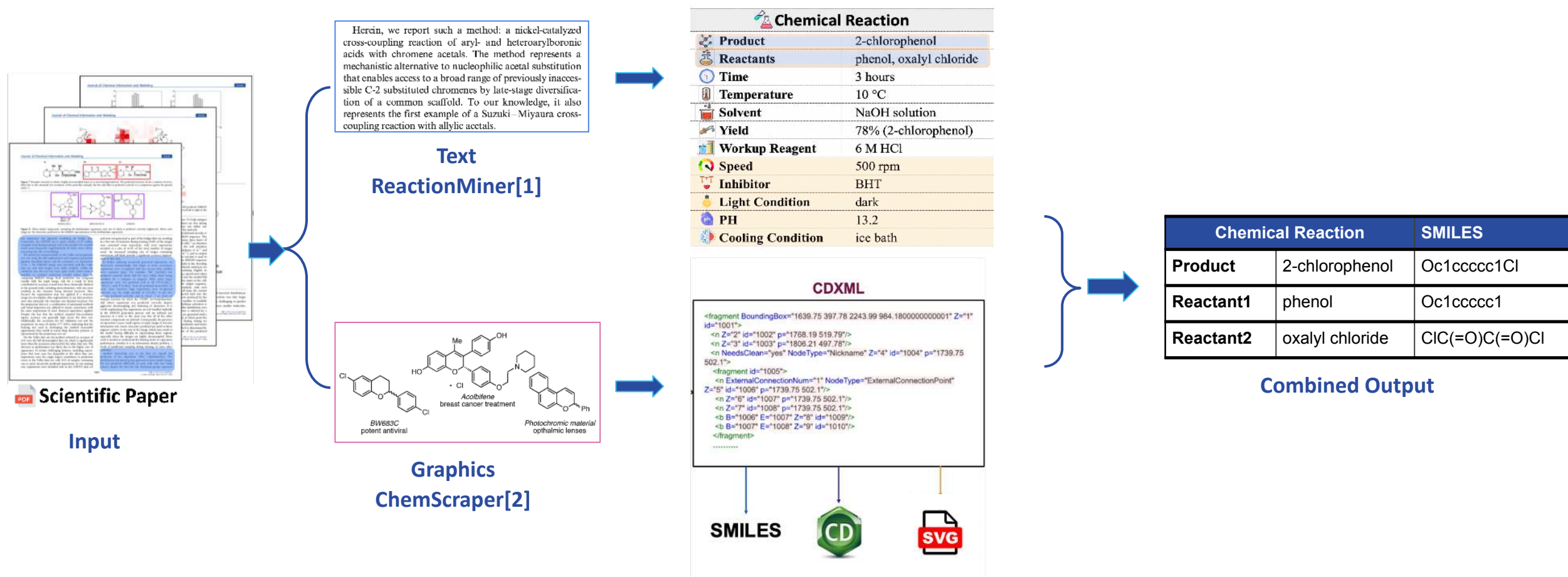
ChemScraper and Reaction Miner Integration

Task: Extract reactions from documents

1. **ReactionMiner:** Extracts **structured chemical reaction** information from literature (**document PDF text**)

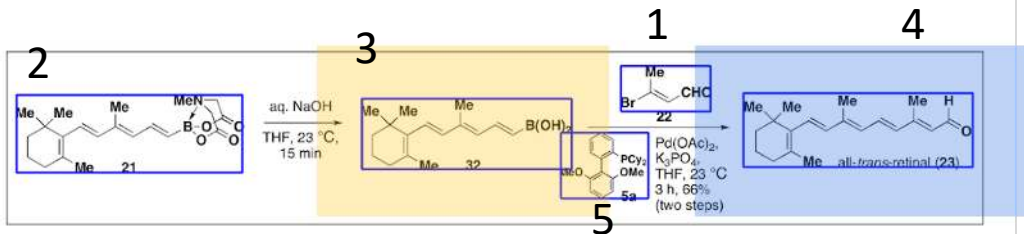
2. **ChemScraper:** Parses **CDXML/SMILES** from literature (**document molecule diagrams**)

Integration: Match ReactionMiner information to parsed ChemScraper molecule diagrams (SMILES)



ChemScraper and Reaction Miner Integration

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

Input Document (Paper9_Sla.pdf)

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)C4CCCC4)C=CC=C2</chem>

ChemScraper Output

Goal: Extract reactions from documents

- **Reaction Miner:** Structured reaction data from document PDF text
- **ChemScraper:** CDXML/SMILES from document molecule diagrams
- **Integration:** Match reaction miner data to chemscraper molecule diagrams and SMILES

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

Final combined output

Reaction	Reaction Miner Dictionary	
1	Text	A stirred solution of 32
	Product	all-trans-retinal (23)
	Reactant1	32
	Reactant2	22
	Yield	66% (all-trans-retinal (23))

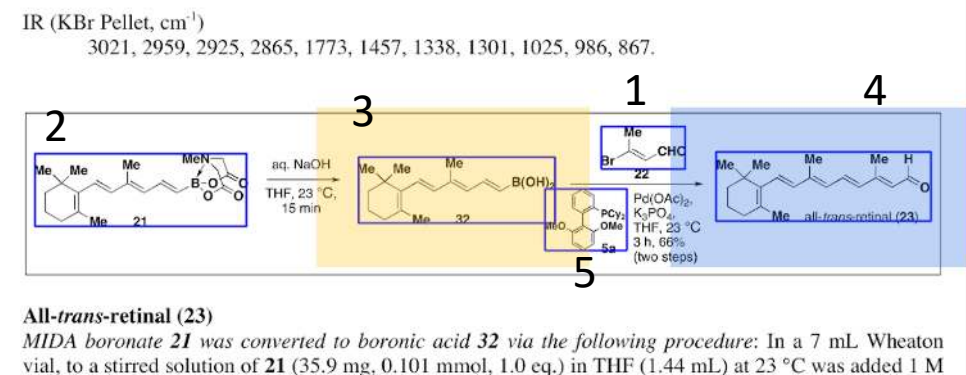
Reaction Miner Output

ChemScraper and Reaction Miner Integration

Matching Reaction Miner Data and Chemscraper Molecule Diagrams

1. String matching

- Molecule ID from reaction miner compared to words near molecules
- Levenshtein distance between nearest strings to select best match
- Example: 32, 22, compound 2a



Input Document (Paper9_Sla.pdf)

2. SMILES matching

- Compound name from Reaction Miner converted to SMILES using OPSIN API
- Tanimoto similarity on converted SMILES and chemscraper SMILES output to select best match
- Example: ethyl acetate, propanoate, vinyl iodide

Document	Page	Molecule	Predicted SMILES
Paper9_Sla	1	1	Br/C(C)=C/C=O
		2	CC1(C)C(/C=C/C(C)=C/C=C B(OC(C2)=O)OC(CN2C)=O)=C(C)CCC1
		3	CC1(C)C(/C=C/C(C)=C/C=C/B(O)O)=C(C)CCC1
		4	CC1(C)C(/C=C/C(C)=C/C=C/C(C)=C/C=O)=C(C)CCC1
		5	COC1=CC=CC(OC)=C1C2=C(P(C3CCCCC3)C4CCCC4)C=CC=C2

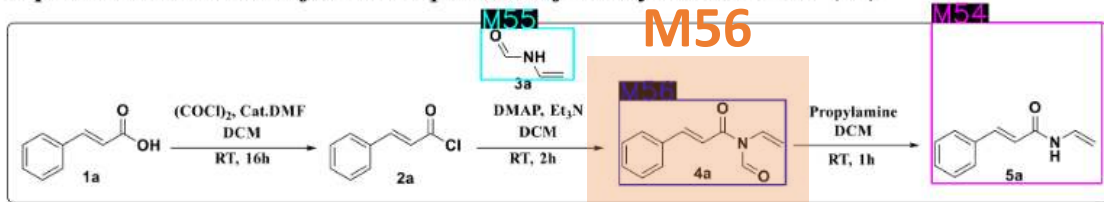
ChemScraper Output

Reaction	Reaction Miner Dictionary	
1	Text	A stirred solution of 32
	Product	all-trans-retinal (23)
	Reactant1	32
	Reactant2	22
	Yield	66% (all-trans-retinal (23))

Reaction Miner Output

ChemScraper and Reaction Miner Integration

Experimental Procedure for the Preparation of N-vinylcinnamamide (5a):



To a stirred solution of compound **1a** (2.0 g, 13.51 mmol) in DCM (20 mL) under N₂ atm was added oxalyl chloride (1.80 mL, 20.27 mmol) and cat. DMF at 0 °C and stirred the reaction mixture at RT for 16h. The excess oxalyl chloride was removed by vacuum to give the compound **2a**. In another RB flask compound **3a** (0.86 g, 12.51 mmol), DMAP (0.16 g, 1.35 mmol) and Et₃N (2.60 mL, 18.37 mmol) were taken in DCM (20 mL) and cooled the reaction mixture to 0 °C. To this reaction mixture, above prepared acid chloride (compound **2a**) was slowly added and stirred the reaction mixture at RT for 2h. The progress of the reaction was monitored by TLC (20% Ethyl acetate in petroleum ether) showed completion of the reaction. After completion of the reaction; reaction mixture was quenched with water (100 mL) and extracted with DCM (2×100 mL). The combined organic layers were washed with water, brine and dried over Na₂SO₄, filtered, filtrate was concentrated under reduced pressure to afford crude compound **4a** (4 g) as a brown liquid which was further proceeded to the next step without purification.

To a stirred solution of crude compound **4a** in DCM (60 mL) was added n-propyl amine (1.50 mL) dropwise at 0 °C and the resulting reaction mixture was stirred at RT for 1h. The progress of the reaction was monitored by TLC (20% Ethyl acetate in petroleum ether) showed

Input Document Page

```
},
"19": {
  "text": "To a stirred solution of crude compound 4a in DCM (60 mL) was added",
  "id_region": 19,
  "page_num": "9",
  "reactions": [
    {
      "Product": {
        "name": "crude compound 4b",
        "confidence_matching": 0.6666666666666667,
        "id_molecule": 10,
        "matching_type": "string",
        "id_chemscraper_tsv": 5,
        "SMILES": "[Bh]\\C=C/NC(=*)\\C=C\\*",
        "page_num": 2
      },
      "Reactant": {
        "name": "crude compound 4a, n-propyl amine",
        "confidence_matching": 1.0,
        "id_molecule": 56,
        "matching_type": "string",
        "id_chemscraper_tsv": 3,
        "SMILES": "C=CN(C=O)C(=O)\\C=C\\C1=CC=CC=C1",
        "page_num": 9
      },
      "text_region_id": 19,
      "Solvent": "DCM",
      "Temperature": "0 o C, RT",
      "Time": "1h",
      "id_reaction": 23
    }
  ]
},
"20": {
```

Integration Output JSON

Molecule with ID **56** (M56) matched with a reaction extracted from text region with id **19** (TR19)

ChemScraper and Reaction Miner Integration

Medicinal Chemistry & Drug Discovery

Synthesis of New Lansiumamide A/Alatamide (cis-Isomer) Based Unnatural α -Amino Acid Derivatives via the Suzuki-Miyaura Cross Coupling as Key Step

M. Sambaiah,^[a, b] Pradeep Kumar Thota,^[a] Shrinivas S. Kottawar,^[a] Satyanarayana Yennam,^[a] K. Shiva Kumar,^{*,[b]} and Manoranjan Behera^{*,[a]}

We have developed a new synthetic method for the preparation of Lansiumamide-amino acid (1) as well as Alatamide (cis-isomer)-amino acid (2) hybrids. The key step involves the Suzuki coupling reaction of N-((Z)-2-bromovinyl) cinnamamide with phenylboronic acid and (Z)-N-(2-iodovinyl)benzamide with 4-methoxy phenylboronic acid. The Suzuki coupling of N-((E)-2-bromo-2-phenylvinyl) cinnamamide as well as (E)-N-(2-iodo-2-

(4-methoxyphenyl)vinyl)benzamide with 4-borono-pinacol ester of phenylalanine to give the Lansiumamide-phenylalanine and Alatamide (cis-isomer)-phenylalanine amino acid hybrids in good yields. This is the first synthesis of C-linked Lansiumamide A and Alatamide (cis-isomer)-amino acid conjugates where both the natural products are linked to amino acid through C-C bond.

Introduction

Clausena lansium is a natural product which belongs to the genus of Clausena which a member of the Rutaceae family. The plant Rutaceae is generally spread in southern China and popular for its medicinal use. The leaves of Rutaceae family plants are utilized for the treatment of asthma, cough, gastrointestinal disorder and stomach related sickness while the seeds have been utilized for treatment of ulcers, chronic as well as inflammation in the gastrointestinal system. Based on the results of Jer-Huei Lin, Clausena plant leaves and seeds contain natural products like Lansiumamide A (Figure 1). Lansiumamide A was first isolated from the extracts of Clausena lansium Skeels by Lin in 1989, who also confirmed the structure of the molecule.^[1] Wan has explained in detail the biological activities of Lansiumamide.^[2] It has been shown to exhibit anticonvulsant

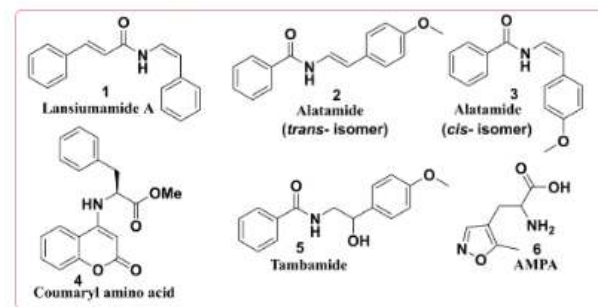
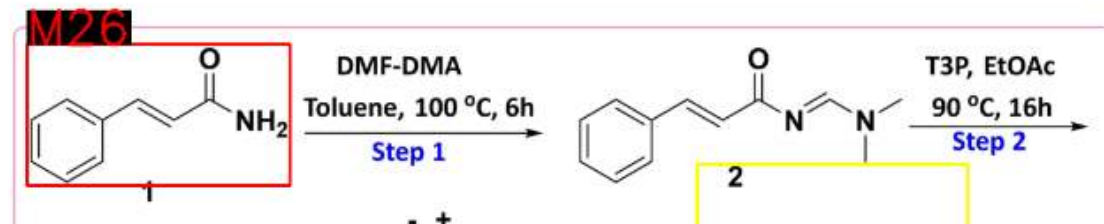


Figure 1. Chemical structures of Lansiumamide A and Alatamide.

Results and Discussion

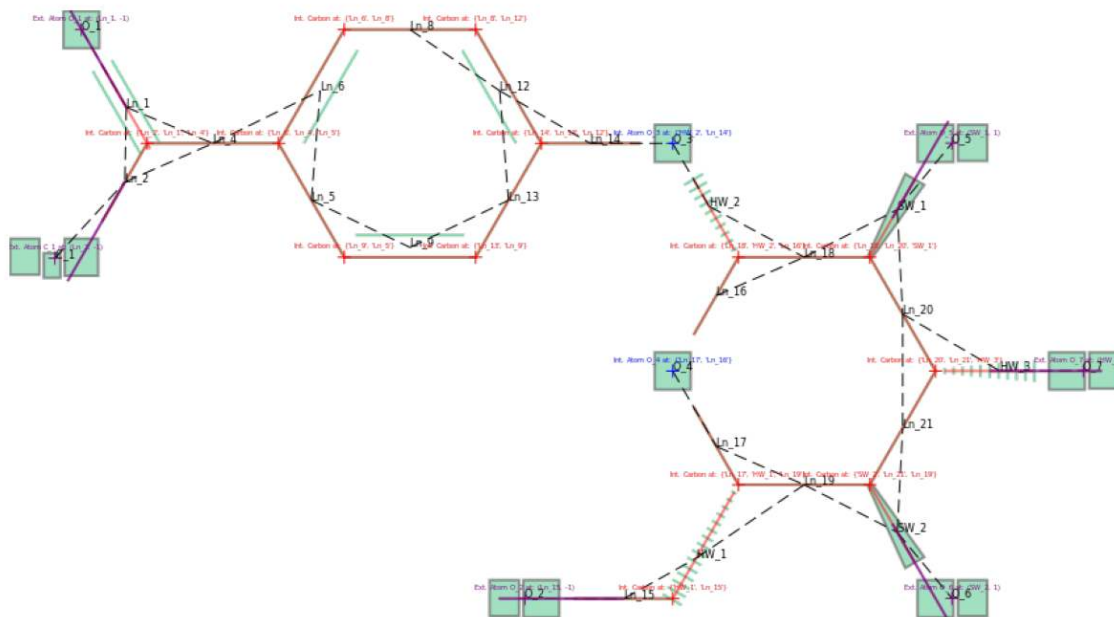
Initially we started our synthesis by taking amide 1 as starting material to get N-formyl amide by using reported method (Scheme 2).^[24] Thus compound 3 was reacted under Wittig reaction condition with (Bromomethyl)triphenylphosphonium bromide (Compound 4) in the presence of KO^tBu as base to afford compound 5. However, the reaction was not much successful, so, we started exploring various methods for the preparation of compound 6a and 4b (Scheme 3&4). To design and validate the synthetic scheme 3 & 4, we initiated the



More visualization outputs

Next Steps

- **Integration:** ReactionMiner + ChemScraper integrated online system
- **Born-digital Parser:** New parsing strategy focused on accurate line geometries
- **Visual Parser:** Handling imbalanced class distributions and robustness by adding image noise



Improved Born-digital parser strategy

Search results by SMILES or "Draw / Upload" another format: Draw / Upload

27 results

Molecule	1	Extracted Structure	Original Structure	SMILES	Flag
Unavailable			<chem>C/C(*)=C/C1=CC=CC=C1</chem>		

Results Summary
Found 1 time(s) in 1 document

Location in PDF
page: 1

Other Occurrences
page(s): 1

Description
Unavailable

Molecular Information
Formula: C₉H₈
Weight: 117.17 g/mol

Chemical Safety
Unavailable

External References
PubChem ID: 101209250

Structural Information
Extracted Structure: MeSEtPh
Original Structure:

Original Context
[Open in PDF viewer](#)

TABLE 2. CROSS-COUPINGS OF ALKYLBORON REAGENTS WITH ALKENYL ELECTROPHILES

Alkyl	Alkenyl Electrophile	Conditions	Products(s) and Yield(s) (%)	Ref.
Me		Pd(OAc) ₂ (5 mol %), PPh ₃ (10 mol %), Cs ₂ CO ₃ (1.5 eq), THF, 40°, 5 h	(66)	187
n-Bu		Pd(OAc) ₂ (5 mol %), SPhos (10 mol %), Cs ₂ CO ₃ (2 eq), toluene/water (4:1), 100°, 5 h	(71)	188

Navigation: << < 1 2 3 > >> 10

Online ChemScraper system

Thank You for your Attention

Acknowledgement:

DPRL: Bryan Amador, Abhisek Dey, Matt Langsenkamp, Ming Creekmore, Brandon Kirincich, and Patrick Philippy

NCSA: Matthew Berry, Yashdeep Thorat, Sara Lambert, and Kate Arneson

UIUC: Linyi Ding, Leo Luo, Siru Ouyang, Blake Ocampo, Dr. Jiawei Han, Dr. Scott Denmark

Any Questions?

