Multimodal Search in Chemical Documents and Reactions

Ayush Kumar Shah*[1], Abhisek Dey*[1], Leo Luo*[2], Bryan Amador*[1], Patrick Philippy[1],
Ming Zhong[2], Siru Ouyang[2], David Mark Friday[2], David Bianchi[2], Nick Jackson[2], Richard Zanibbi[1], Jiawei Han[2]
[1] Document and Pattern Recognition Lab, Rochester Institute of Technology, NY, USA *[2] Department of Chemistry/NCSA, University of Illinois at Urbana-Champaign, IL, USA*

Overview

- **Multimodal search** for molecular diagrams, reactions, and associated text in PDF chemical documents
- **Text-diagram links** for diagram locations, SMILES strings, and passage descriptions
- **Reaction extraction** via LLM-guided parsing (LLaMA3.1)

cs.rit.edu/~dprl/reactionminer-demo-landing/

Linking Names & Diagrams

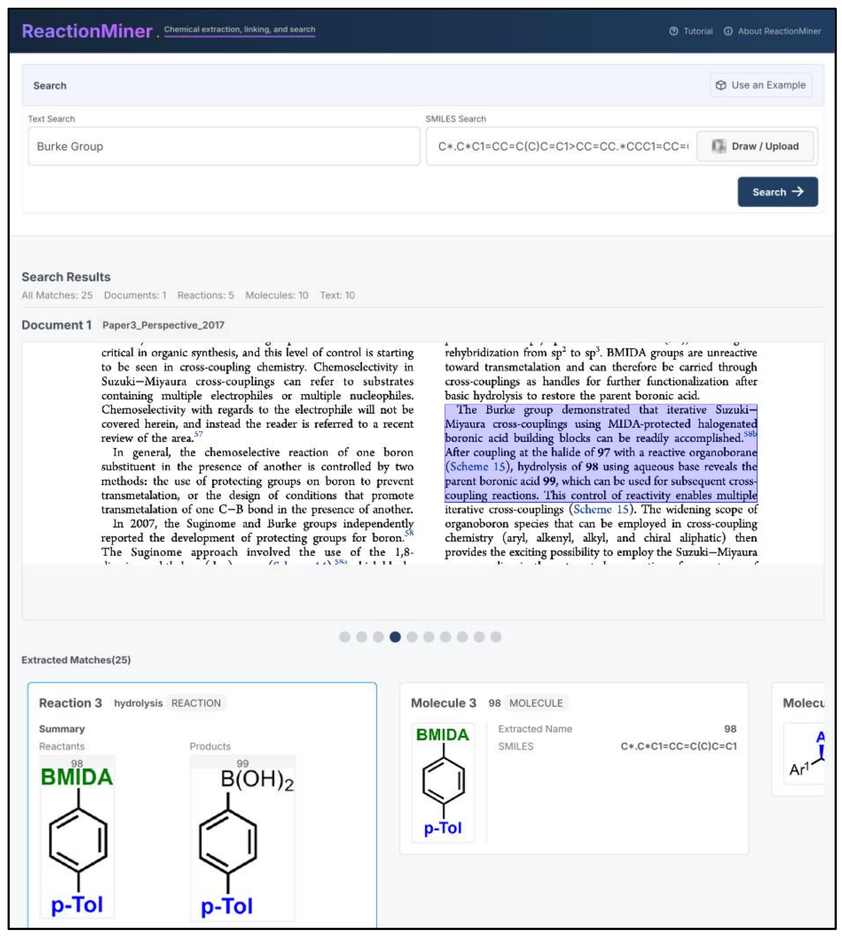
- **Text label matching:** Match text references e.g., "Compound 3" with diagrams by highest normalized Levenshtein similarity
- **Structure matching:** Convert names & diagrams to SMILES using ChemDataExtractor and OPSIN, matches using Tanimoto similarity
- **Highest match score** used when text label and SMILES links available

Indexing

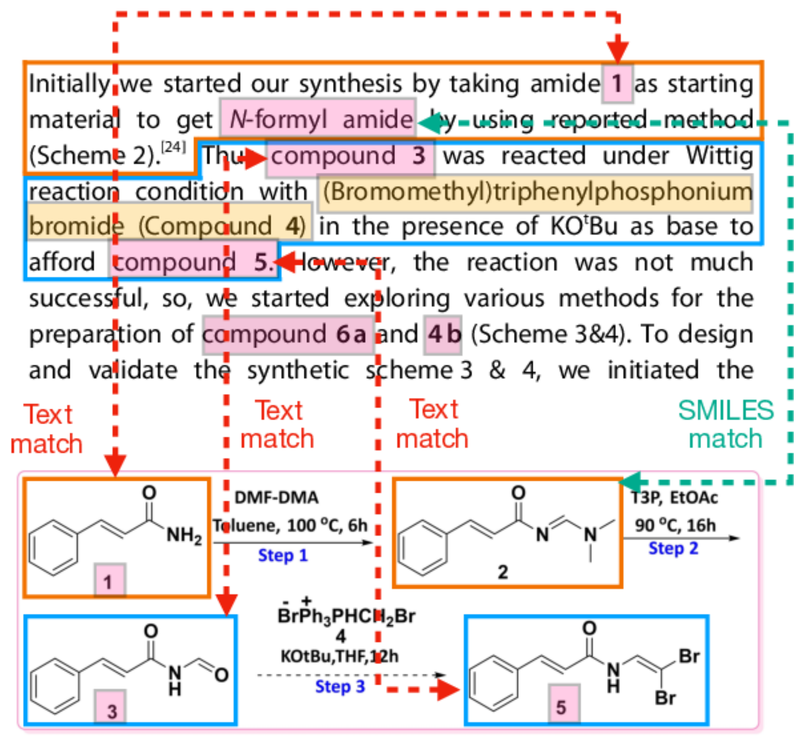
- **Text** sent to **ReactionMiner**[3]
- Inverted index for text tokens
- **Diagrams and molecule names converted to SMILES,** numeric ids
- **Diagrams linked to text passages**
- Index metadata: document, page no., bounding boxes

Search

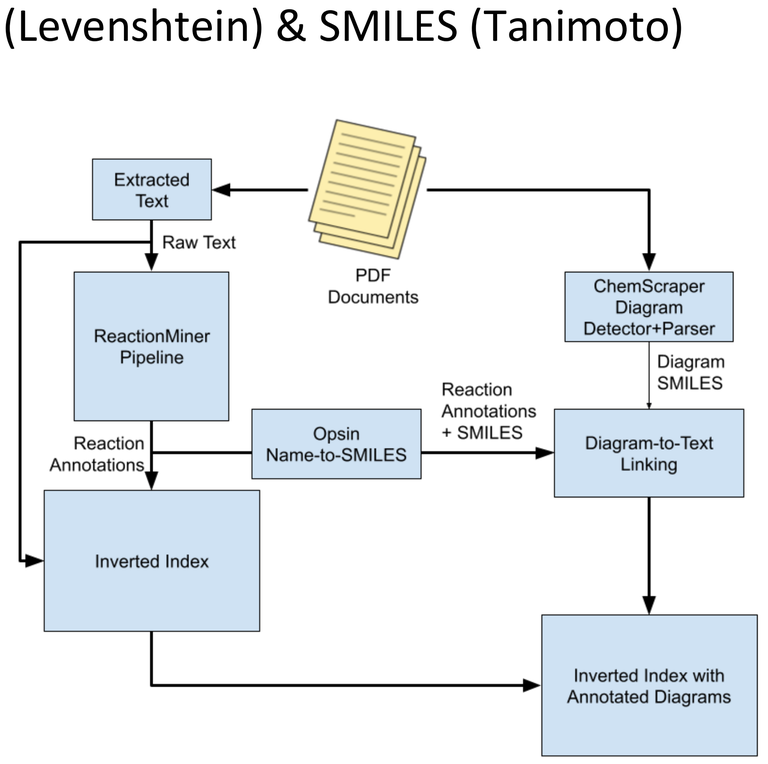
- **Text queries:** BM25 search
- **SMILES/SMARTS queries:**
  (1) RDKit substructure search
  (2) Tanimoto similarity
- **Multimodal queries:** fuse text/structure scores by re-ranking passages based on BM25 score and number of matched SMILES
- For 10 page document:
  Extraction time: 53.78 sec
  Linking & Indexing time: 23.51 sec
- **Retrieval time:** <1 sec



Text + SMARTS reaction query result.
Matched passages, molecules, reactions have clickable navigation within PDF docs



Linking compound names to diagrams.
Links obtained using text references (Levenshtein) & SMILES (Tanimoto)



Indexing Pipeline

Search Results View

- **Structure search:** Exact match for text + SMILES/SMARTS, similarity search for SMILES-only queries
- Highlights matched text and diagrams in PDF documents
- **Navigation:** Scrollable molecule and reaction cards, clicking jumps to PDF diagrams and passages
- **Reaction information** in reaction cards: reactant, product, catalyst, temperature, yield

PDF Document Collection

- **Suzuki coupling reactions**
- **13 PDF documents:**
  7 regular papers + 6 supplementary
- 538 passages, 383 unique SMILES
- 219 extracted reaction records

Expert Evaluation

Chemists at University of Illinois tested all three retrieval modes

Pros

- Liked clickable molecule cards and **navigation** to reaction details
- Facilitates **faster**, more **complete** literature search for chemists

Cons

- Ranking can make influence of text vs. SMILES unclear
- Some molecule derivatives not named; missing text/diagram links
- **No filtering tool** for molecule, reaction, and text matches
  (***added later**)

References

[1] Lowe et al., "**Chemical Name to Structure: OPSIN, an Open Source Solution**," J. Chem. Inf. Model., 2011
[2] Shah et al., "**ChemScraper: leveraging PDF graphics instructions for molecular diagram parsing**," IJDAR, 2024
[3] Zhong et al., "**Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data**," EMNLP 2023
[4] Mavračić et al., "**ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science**," J. Chem. Inf. Model., 2021
[5] Robertson et al., "**Okapi at TREC-3**," TREC, 1994
[6] Grattafiori et al., "**The Llama 3 Herd of Models**," 2024
[7] Wang et al., "**Scaled-YOLOv4: Scaling Cross Stage Partial Network**," CVPR, 2021

NSF
MOLECULE MAKER LAB INSTITUTE
dprl Document and Pattern Recognition Lab
RIT Rochester Institute of Technology