Second Workshop on Data, Text, Web, and Social Network Mining
April 22, 2011

Text-based and Image-based Recognition and Extraction of Molecular Information from Figures and Figure Captions

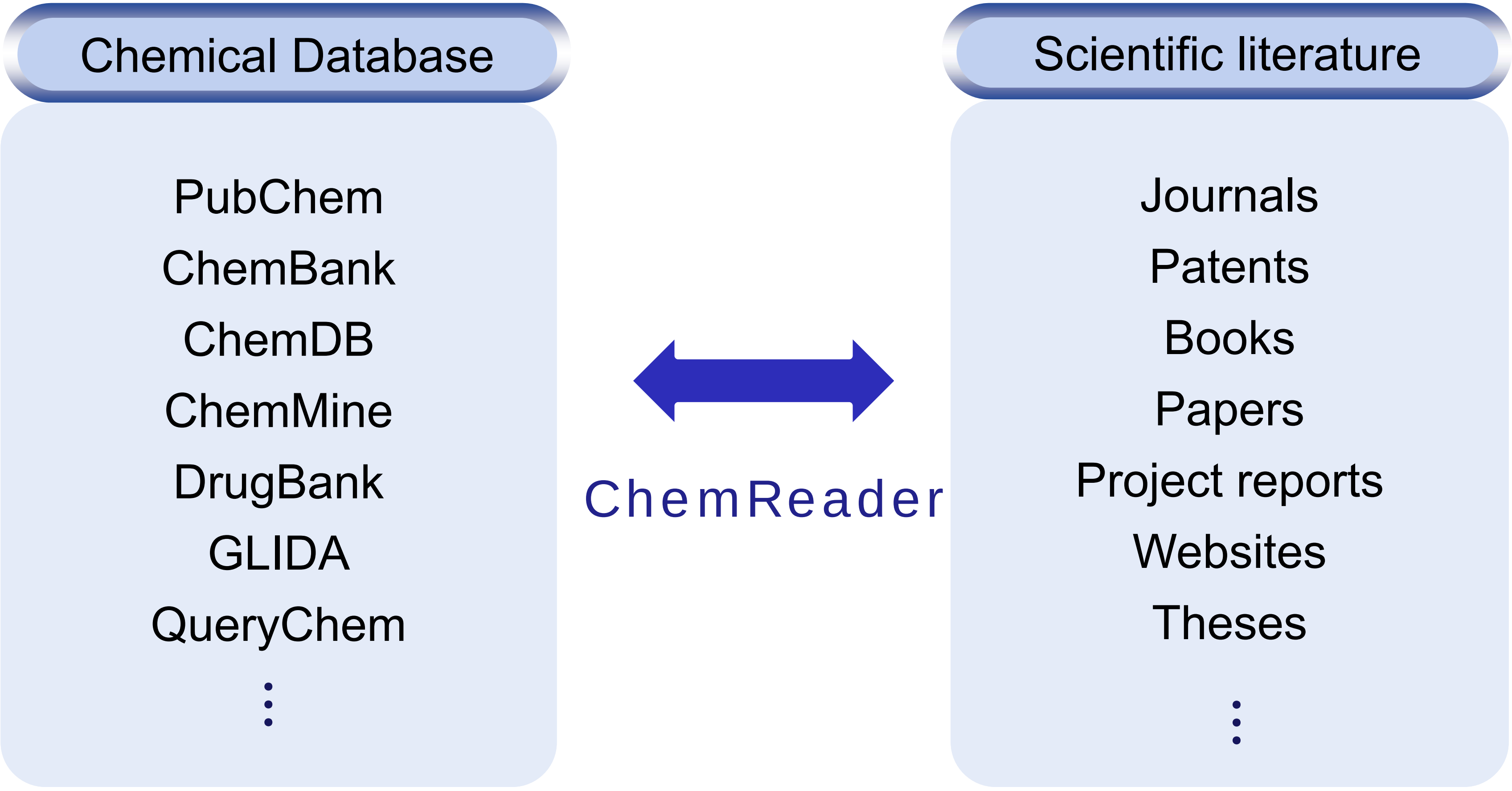
Jungkap Park, Gus R. Rosania & Kazuhiro Saitou

University of Michigan, Ann Arbor

Outline

- Overview of Image-based Annotation
- ChemReader
- Annotation Strategy and Test Result
- Chemical Literature Database
- Preliminary Statistics
- Future Works

Why ChemReader?



Chemical Database

PubChem
ChemBank
ChemDB
ChemMine
DrugBank
GLIDA
QueryChem
⋮

ChemReader

Scientific literature

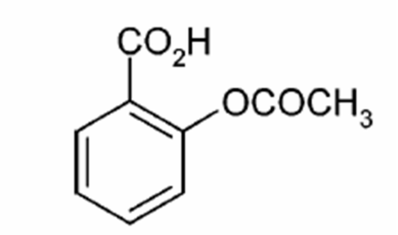
Journals
Patents
Books
Papers
Project reports
Websites
Theses
⋮

Searching for chemical information

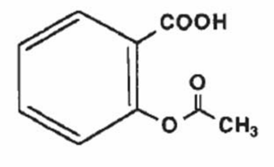
- **The problems:**
  - Too many synonyms
  - Often referenced by chemical structure diagrams

Ex) Aspirin

- Acetylsalicylic acid (ASA)
- 2-acetyloxybenzoic acid
- acetylsalicylate
- Acylpyrin
- Colfarit
- Ecotrin
- Enterosarein
- Acenterine
- Polopiryna
- ....



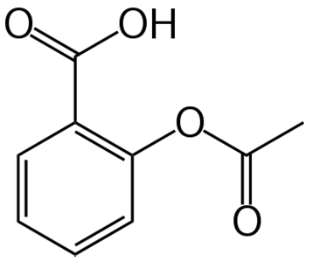
P Vishweshwar et al, J. Am. Chem. 2005



PJ Loll et al, Nat. Struct. Mol. Biol. 1995

Searching for chemical information

- **The problems**
  - Need to identify related compounds

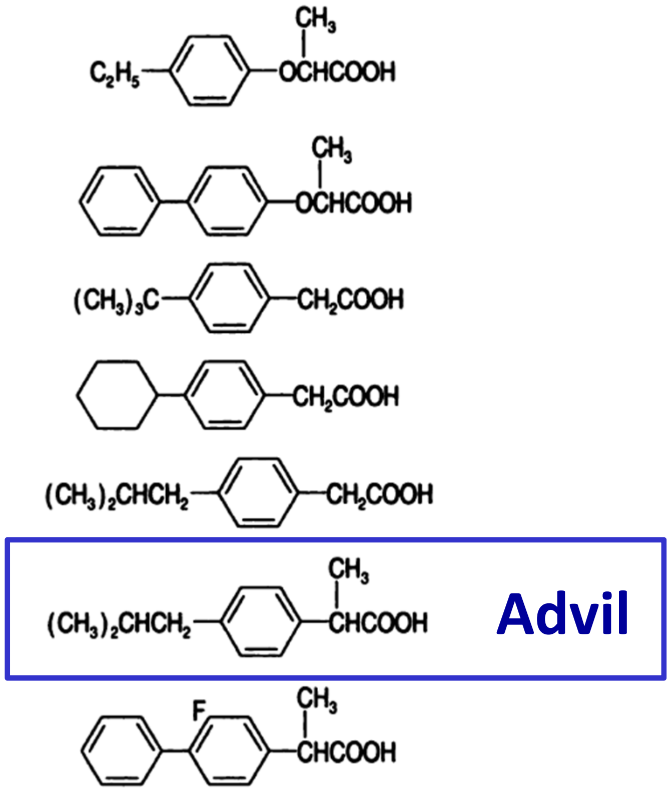


Aspirin

Similar structure

≈

Similar drug effect

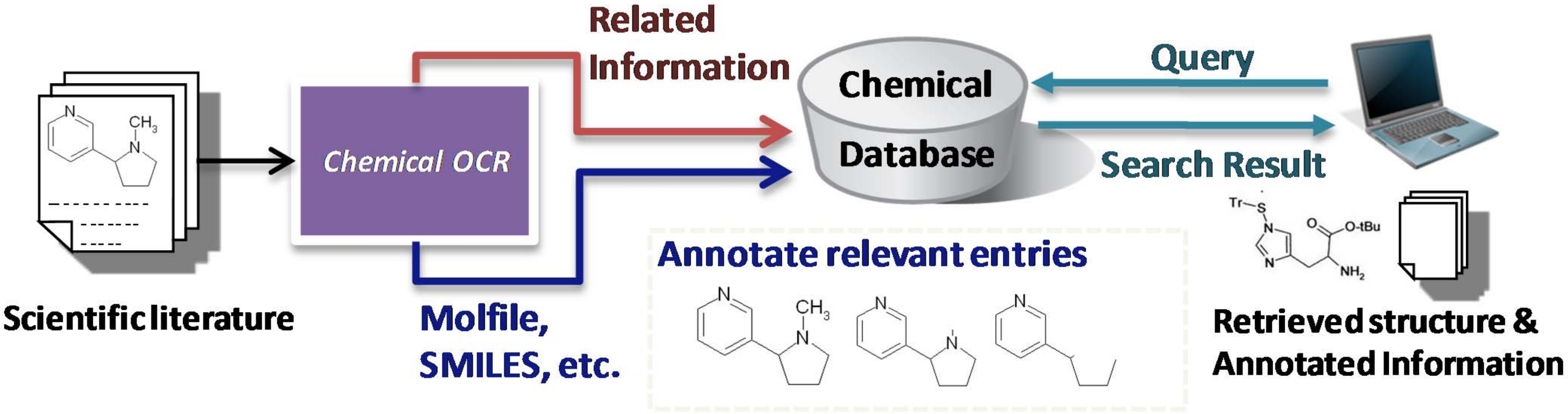


Advil

SS Adams, J. Clin. Pharmacol. 1992

Image Based Annotation

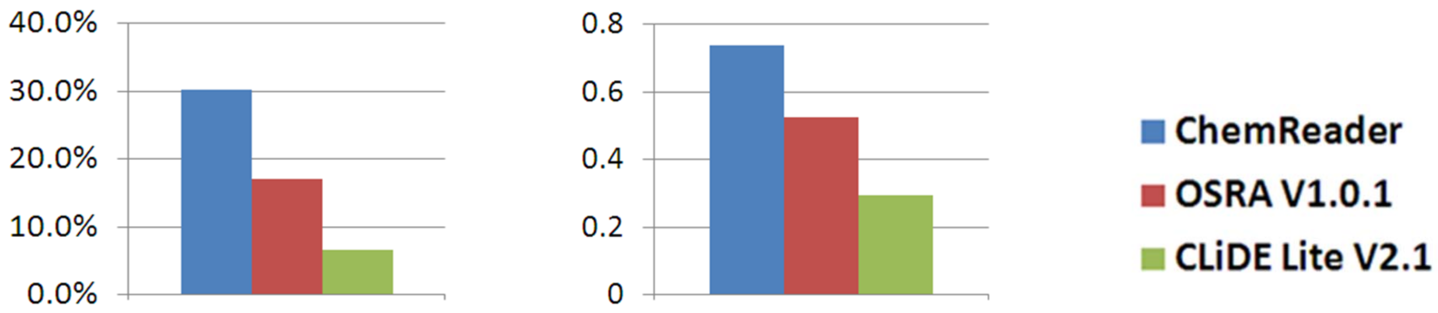
- **Chemical database annotation using Chemical OCR**



- **Chemical OCR system**
  - Extract 2D chemical structure diagram from literature
  - Convert tem to a standard chemical file format
  - CLiDE, ChemOCR, OSRA and **ChemReader**

Test Result

- **Recognition Test**



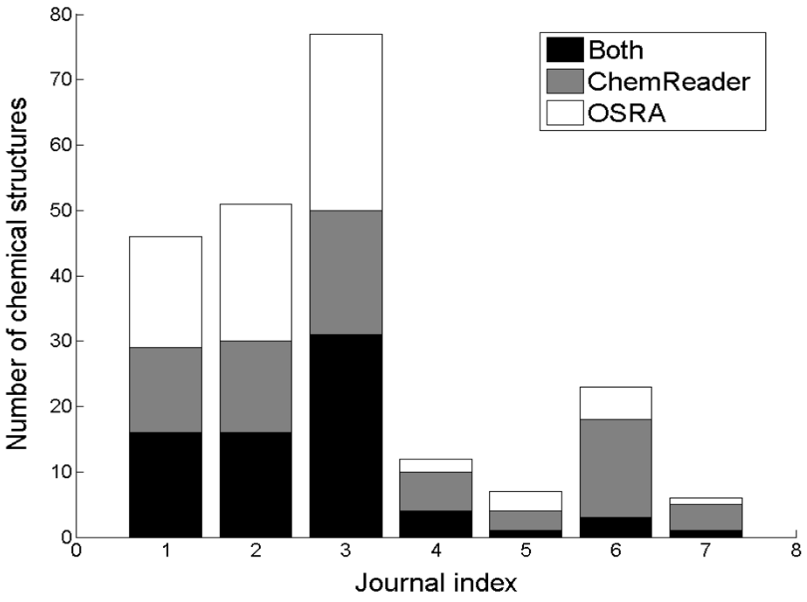
% of correct outputs **Avg. Tanimoto Similiarty**

- **Annotation Test**
  - Tunable annotation strategy: Two different conditions for screening output structures

	Avg. Recall	Avg. Precision
Test I	0.69	0.8
Test II	0.8	0.88

Ensemble Approach

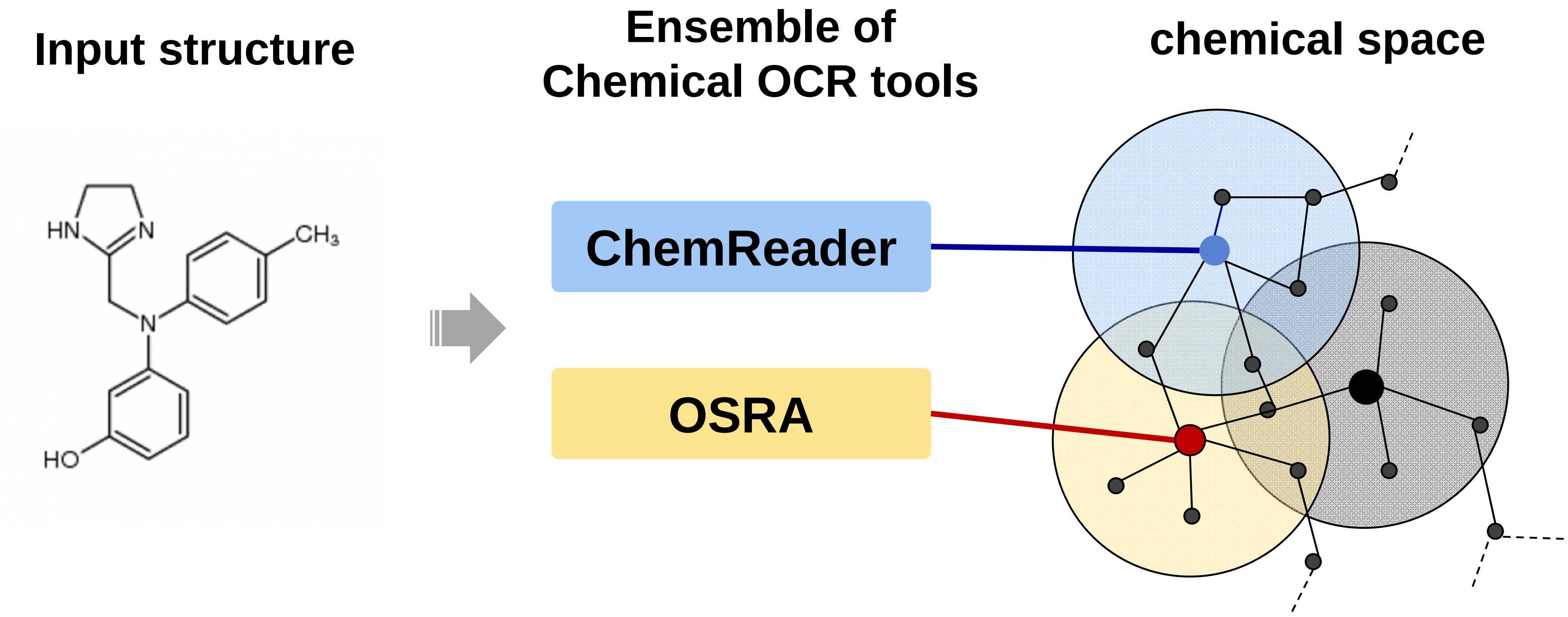
- **Motivation**
  - Maximize the chance of including correct structure information by combining strengths of multiple chemical OCR systems
- **Rationale**
  - Different machine-vision algorithms could have different strengths in particular types of structures



Number of successful outputs produced by ChemReader or OSRA grouped by journal index.

Ensemble Approach

- **Use of multiple chemical OCR tools**



Input structure

Ensemble of Chemical OCR tools

chemical space

ChemReader

OSRA

- Two output structures for the same input structure become members of the ensemble
- The ensemble approach enables to maximize chance of linking relevant entries in the annotation task

Annotation Test by Ensemble Approach

- **Result**
  - Total number of TP, FP and FN links

	TP	FP	FN
ChemReader	24592	30844	47631
OSRA	33105	21067	54995
Ensemble	45707	51535	55984

  - Averaged recall and precision rates

	Avg. Precision	Avg. Recall
ChemReader	0.563	0.569
OSRA	0.491	0.568
Ensemble	0.544	0.619

The need of image-based annotation

- **Motivation of Image-based annotation**
  - Many molecules are referenced by 2D structure diagrams in chemical literature due to the lack of standard names
  - Image-based mining can uncover knowledge on such molecules that is otherwise inaccessible in chemical databases
- **How to validate?**
  - How chemical entities are referred in research articles?
  - Comparison of text-based annotation and Image-based annotation

Ground truth for chemical literature mining

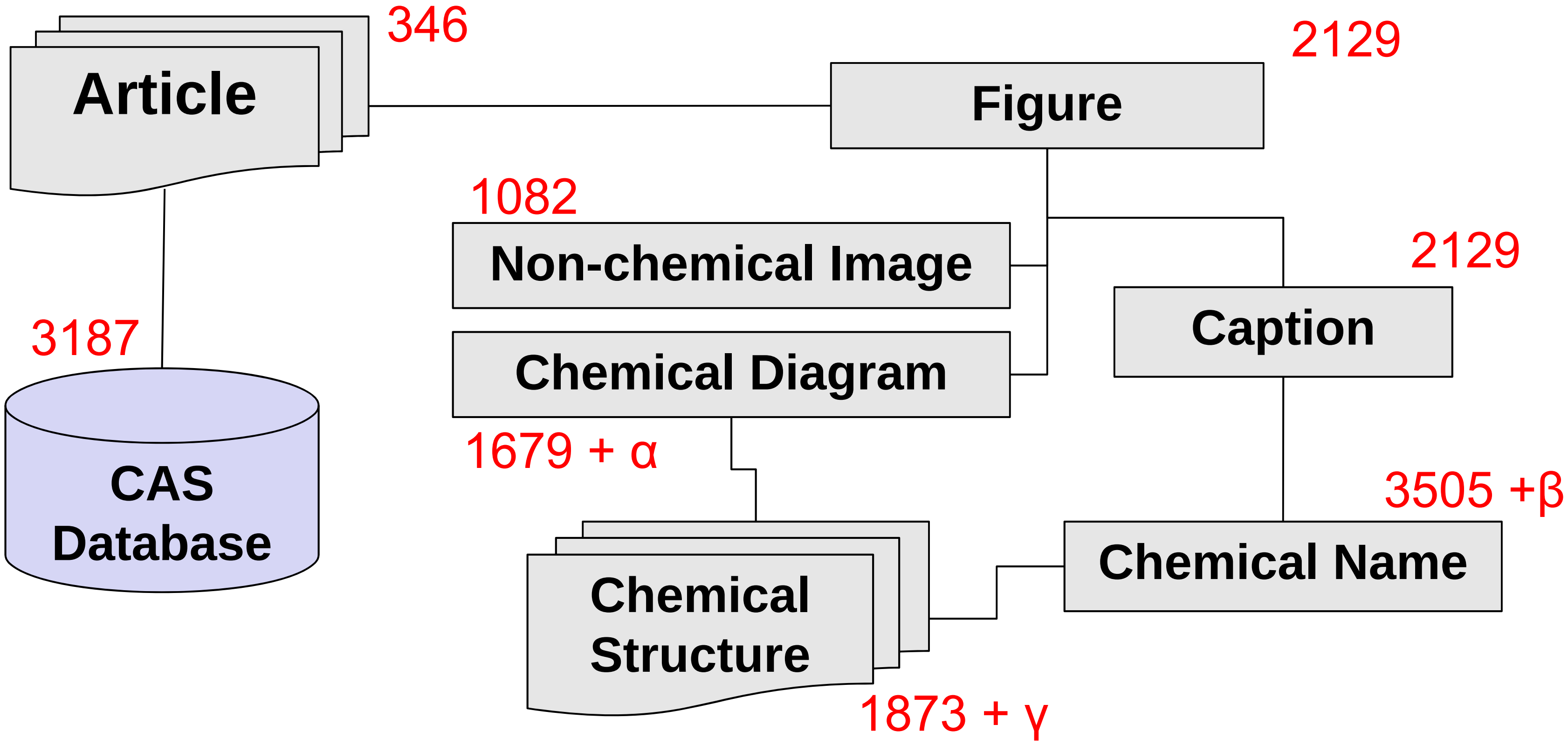
- **CAS Database**
  - The largest and commercially accessible chemical database
  - Links to cited references (journals or patents) dating back to the beginning of the late 19th century
- **Sample set**
  - Keywords search: “Diabetes” and “small molecule”
    - 822 Journal articles
  - Select 399 articles containing molecules being cited only once
  - Download PDF files from publishers’ website
    - Total **346** full-text articles in PDF format

Extraction of chemical info from figures

- **All figures and captions are extracted from articles**
- **Image extraction**
  - Export images without modification of color depth, size or resolution
  - Snapshot tool only for vector graphics
  - Separation of chemical structure images
- **Chemical structure extraction**
  - 2D Chemical structure diagram from image files
  - Chemical names from caption text
  - Extracted chemicals are indexed by CAS Registry numbers (or InChI strings)

Construction of chemical literature database

- **Extracted data is stored in a relational database as traceable assertions**



* Red numbers denote the number of records in the database

Preliminary statistics on current data

- **Identifying chemical diagrams or chemical names on progress**

Total number of linked molecules		
cited in captions	cited in diagram	cited in both
657 + α	1326 + β	110 + γ

- **Over 278 molecules cited in chemical diagrams are missed by CAS**

Text-based annotation using OSCAR3

- **OSCAR3**
  - Chemical documents processing tool (Corbett and Murray-Rust, 2008)
  - Identify chemical names, ontology terms and chemical data
- **Chemical names in caption text**
  - Number of captions tested : 334
  - Number of chemical names = 1087
  - Number of chemical names extracted by OSCAR= 1814
  - Number of correctly identified = 806
  - **Precision = 0.444**
  - **Recall = 0.741**

What we can do with the database

- **Statistical Analysis**
  - How molecules are cited first? By diagrams or names?
  - How many molecules are cited only by diagrams?
  - How many molecules are not indexed by CAS?

2D Chemical diagrams in articles are important data objects for mining chemical literature

Validation of Image-Based Annotation

- **ChemReader is effective?**
  - Chemical structures cited only by diagrams and missed by CAS
  - Chemical structures incorrectly annotated by text-based approach

Image-based approach can uncover knowledge that are inaccessible otherwise

Integration of Image-based and Text-based

- **Multi modal extraction from chemical literature**
  - Text-based mining enables to extract textual descriptors as well as chemical names
  - Graphical Mining
  - Uncover the contextual scientific knowledge

- **Ensemble approach**
  - Strengths of image-based and text-based techniques
  - Increase annotation accuracy

Conclusion

- **Significant fraction of molecules is referenced by chemical diagrams only, and a chemical OCR system can be effective in annotating articles with these molecules**

- **Constructed database will facilitate research in chemical literature mining for the design, training and testing of algorithms for chemical structure extraction and chemical database annotation**

Acknowledgement

- Polyergic Informatics, LLC
- Small Company Innovation Program, College of Engineering

- Michael Conlin
- Ye Li
- Christof Smith
- Caroline Yee
- Bethany Harris

Thank you!