# **ChemScraper:** Leveraging PDF Graphics Instructions for Molecular Diagram Parsing

Ayush Kumar Shah[1], Bryan Amador[1], Abhisek Dey[1], Ming Creekmore[1], Blake Ocampo[2], Scott Denmark[2], Richard Zanibbi[1]

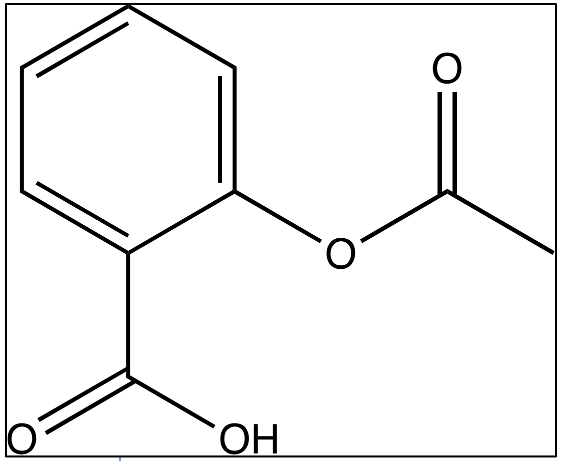
[1]Rochester Institute of Technology, NY, USA
[2]University of Illinois at Urbana-Champaign, IL, USA

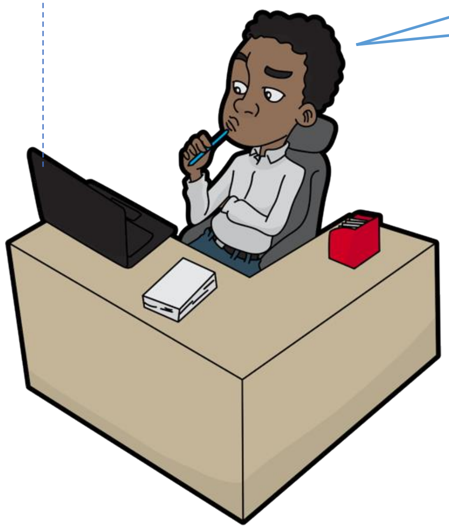
Contact:
as1211@rit.edu

Document and Pattern Recognition Lab

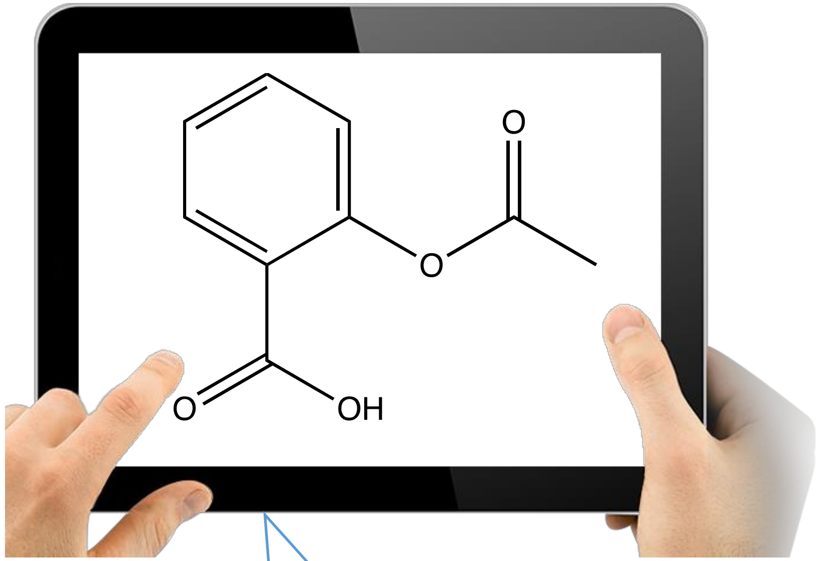
ICDAR Athens Greece 2024
International Conference on Document Analysis and Recognition

# Motivation





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



Let’s find reactions producing this molecule.

# Contributions

- SymbolScraper: Improved PDF character and graphics information extractor
- Born-digital parser: Parsing molecules from vector graphics information (simple, fast and accurate)
- Data generation: Annotated raster images for molecular diagram recognition and other tasks
- Visual Parser trained using generated annotated data (low data requirement and fewer model parameters)
- Graph-based evaluation of chemical structure

# Overview

**Task:** Parsing molecules from documents

**Input:** A scientific paper (PDF)

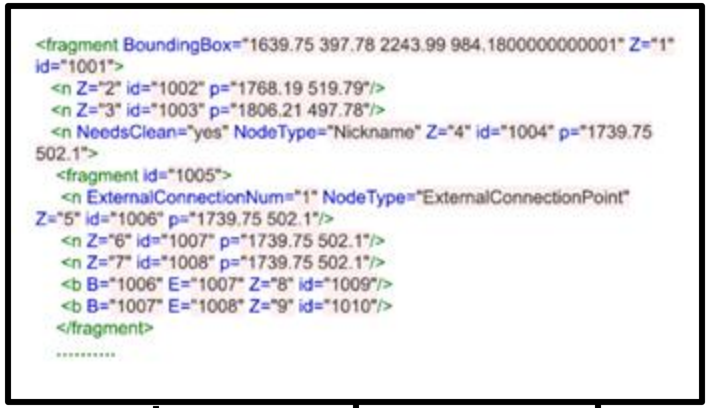
- Embedded raw images
- Drawn vector instructions

🟥 Embedded raw Image

🟪 Drawn vector Instructions

**Output:** All molecule CDXMLs/SMILES

CDXML file



```
<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>
   ..........
```

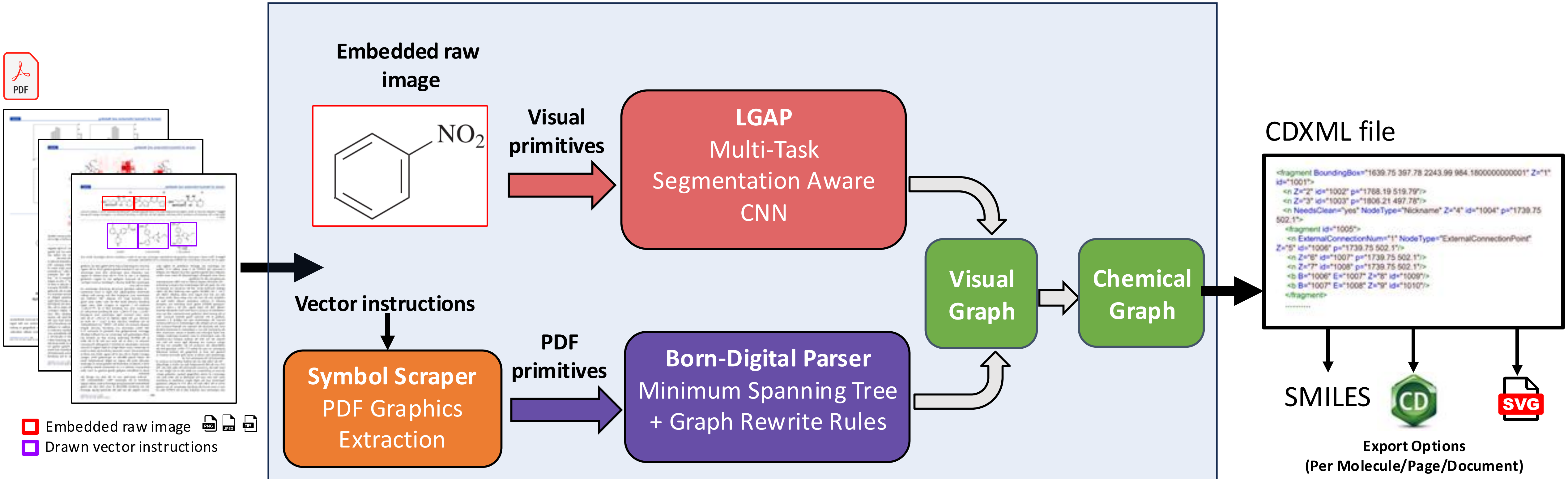
SMILES

CD

SVG

**Export Options**
**(Per Molecule/Page/Document)**

# Overview



Embedded raw image

Visual primitives

**LGAP**
Multi-Task
Segmentation Aware
CNN

Vector instructions

**Symbol Scraper**
PDF Graphics
Extraction

PDF primitives

**Born-Digital Parser**
Minimum Spanning Tree
+ Graph Rewrite Rules

Visual Graph

Chemical Graph

CDXML file

```
<fragment BoundingBox="1639.75 397.78 2243.99 984.180000000001" 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>
   ..........
```

SMILES

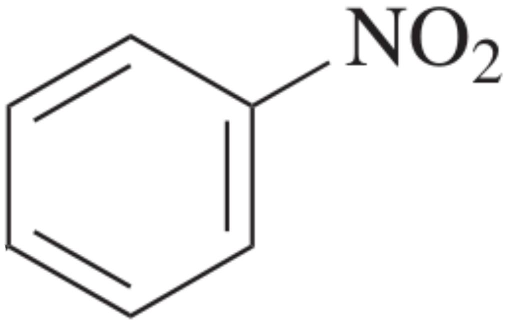
**Export Options**
**(Per Molecule/Page/Document)**

🟥 Embedded raw image

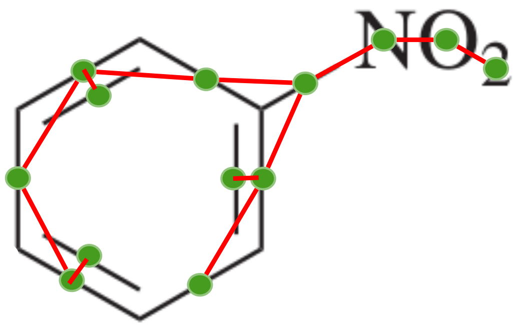
🟪 Drawn vector instructions

[1] Shah, A. K., & Zanibbi, R. (2023). Line-of-Sight with Graph Attention Parser (LGAP) for Math Formulas., Document Analysis and Recognition—ICDAR 2023 (pp. 401–419).

# Born-digital Parser: PDF Molecule Image

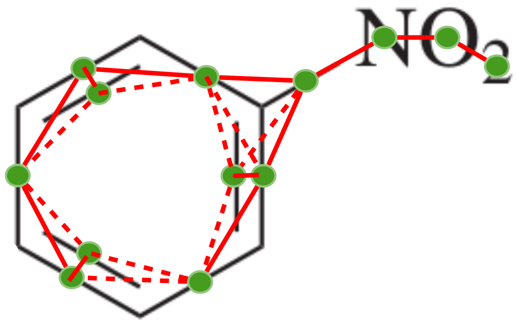


**(a) PDF Image**



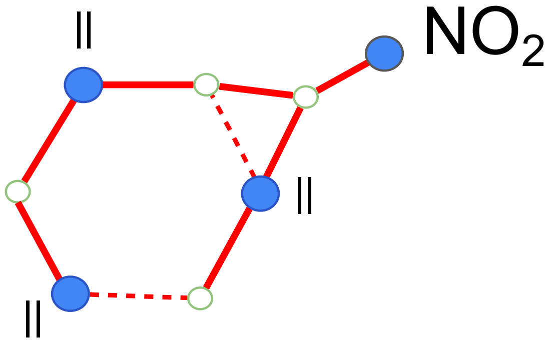
(b) **MST**

*nodes*: lines & characters
*edges*: connections/merges



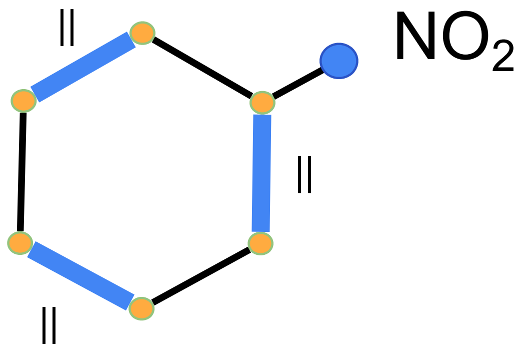
(c) **Visual Graph**

*nodes*: lines & characters
*edges*: connections/merges



(d) **Tokenized Visual Graph**

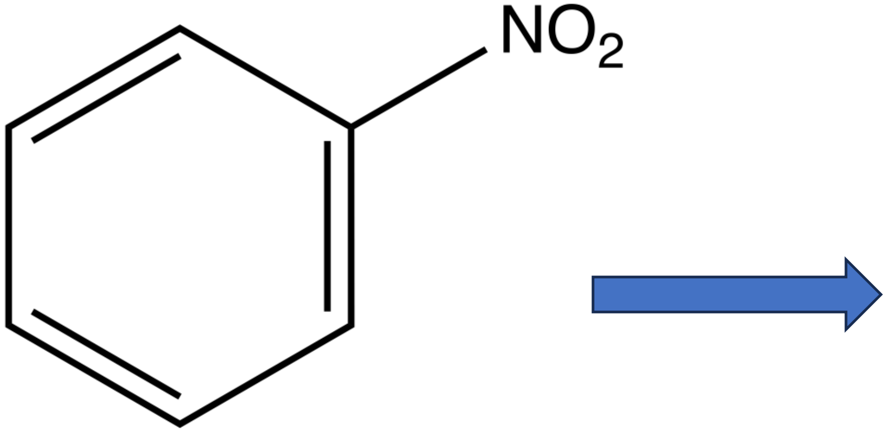
*nodes*: **bonds,** atoms & superatoms
*edges*: connections



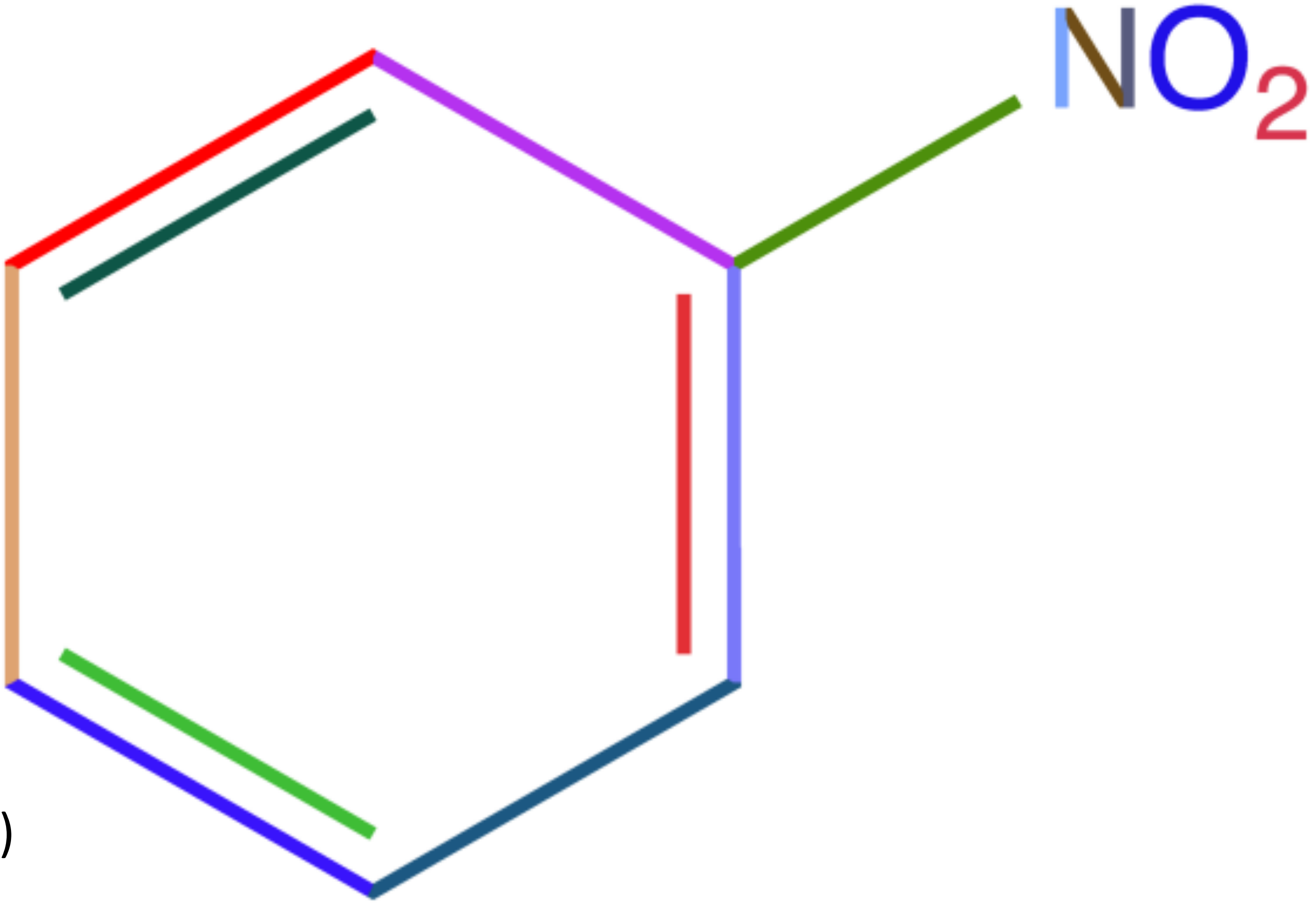
(e) **Molecular Graph**

*nodes*: atoms & superatoms
*edges*: **bonds**

# Visual Primitives for Raster Images (PNG)



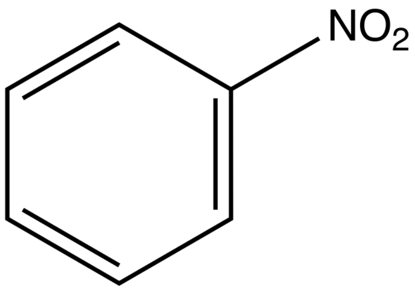
Rendered PDF image



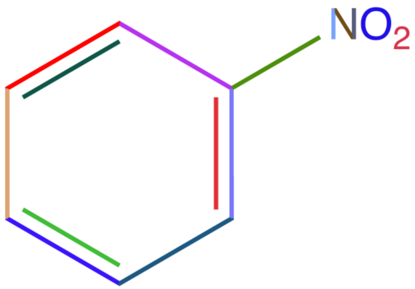
- **Build polygons** from Connected Components (CCs)
- **Extract skeletons** from medial axis of parallel lines
- **Segment CCs:** flood fill assigning each pixel to its closest skeleton

https://pypi.org/project/shapely/

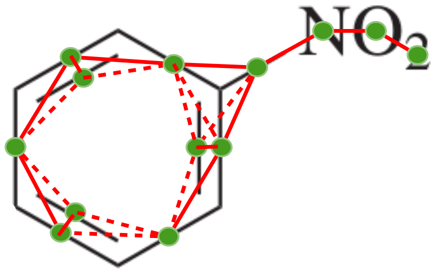
# Annotated Data Generation



Rendered PDF image
(from SMILES)



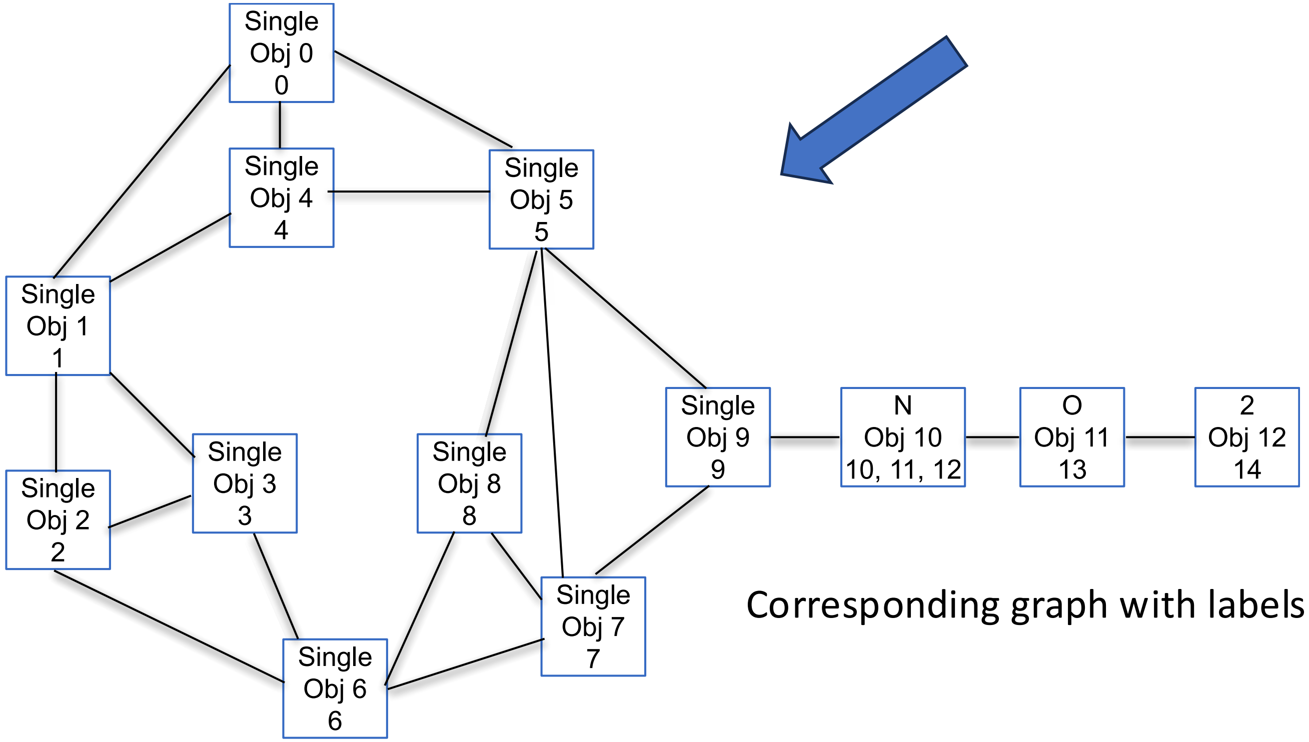
Visual primitives



Visual graph generated
by born-digital parser

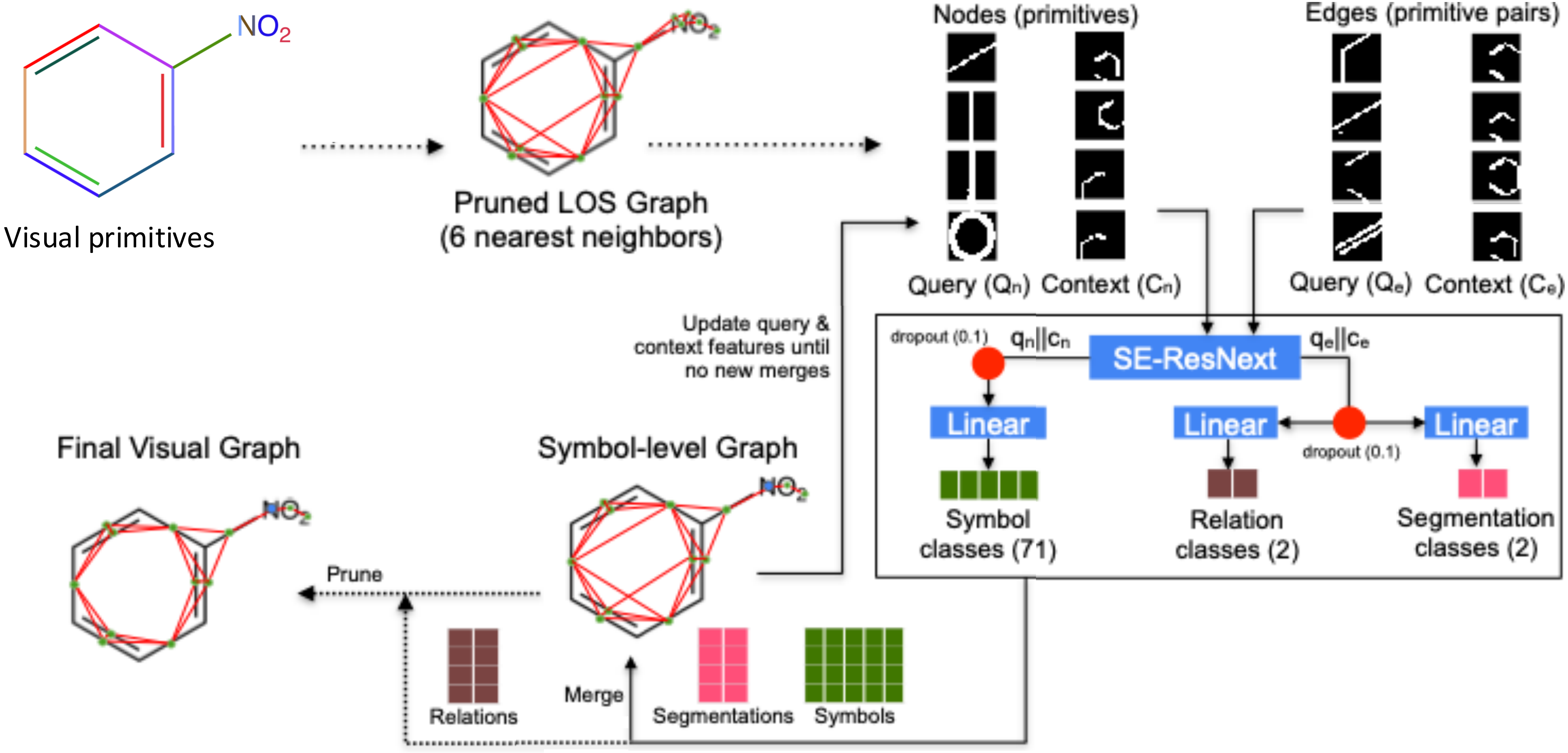
```
# [ OBJECTS ]
# Objects (O): 10
# Format: O, objId, class, 1.0, [primitiveId list]
O, Obj0, Single, 1.0, 0
O, Obj1, Single, 1.0, 1
O, Obj10, N, 1.0, 10, 11, 12
...

# [ RELATIONSHIPS ]
# Relationships (R): 11
# Format: R, parentId, childId, class, 1.0 (weight)
R, Obj0, Obj4, CONNECTED, 1.0
R, Obj0, Obj1, CONNECTED, 1.0
R, Obj1, Obj3, CONNECTED, 1.0
...

# [PRIMITIVE FEATURES]
#contours, 0, 58, 139, 56, 141, 55, 141, ...
#contours, 0, 78, 98, 77, 99, 76, 99, ...
#contours, 1, 80, 395, 80, 397, 81, 398, ...
...
```

Label graph file



Corresponding graph with labels

# Visual Parser



Visual primitives

Pruned LOS Graph
(6 nearest neighbors)

Nodes (primitives)

Edges (primitive pairs)

Query ($Q_n$) Context ($C_n$)

Query ($Q_e$) Context ($C_e$)

Update query &
context features until
no new merges

dropout (0.1)

$q_n||c_n$

SE-ResNext

$q_e||c_e$

Linear

Linear

Linear

dropout (0.1)

Symbol
classes (71)

Relation
classes (2)

Segmentation
classes (2)

Final Visual Graph

Symbol-level Graph

Prune

Relations

Merge

Segmentations

Symbols

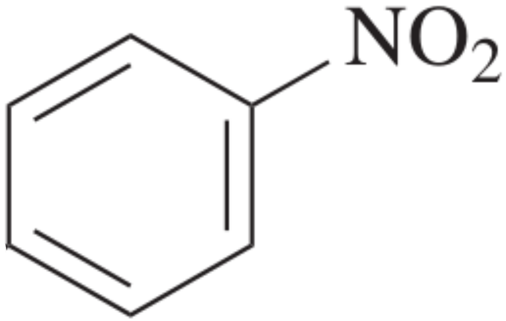
# Results

**Training data source:** Pubchem 1 million

- **Born-digital:** 5,000 molecules
- **Visual:** 3,416 molecules (validated from 5000)

**Metrics:**

- **Exact SMILES match:** String based metrics



c1ccc(cc1)[N+](=O)[O-]

| Systems | Exact SMILES Matches | | |
|---|---|---|---|
| | Indigo<br>(5719) | CLEF-2012<br>(992) | UoB<br>(5740) |
| MolVec 0.9.7 | 95.40 | 83.80 | 80.60 |
| OSRA 2.1 | 95.00 | 84.60 | 78.50 |
| MolScribe | 97.50 | 88.90 | 87.90 |
| MolGrapher | - | **90.50** | **94.90** |
| ChemScraper (Born-Digital – PDF input) | ***98.16*** | *89.32* | *94.41* |

| Systems | Exact SMILES Matches | | |
|---|---|---|---|
| | Indigo<br>(5719) | CLEF-2012<br>(992) | UoB<br>(5740) |
| MolVec 0.9.7 | 95.40 | 83.80 | 80.60 |
| OSRA 2.1 | 95.00 | 84.60 | 78.50 |
| MolScribe | 97.50 | 88.90 | 87.90 |
| MolGrapher | - | 90.50 | **94.90** |
| ChemScraper (Born-Digital – PDF input) | *98.16* | *89.32* | *94.41* |
| ChemScraper (Born-Digital – PDF input)<br>* Skipping rendering errors | ***98.42*** | ***96.20*** | *94.41* |

| Systems | Exact SMILES Matches | | |
|---|---|---|---|
| | Indigo<br>(5719) | CLEF-2012<br>(992) | UoB<br>(5740) |
| MolVec 0.9.7 | 95.40 | 83.80 | 80.60 |
| OSRA 2.1 | 95.00 | 84.60 | 78.50 |
| MolScribe | 97.50 | 88.90 | 87.90 |
| MolGrapher | **-** | 90.50 | **94.90** |
| ChemScraper (Born-Digital – PDF input) | *98.16* | *89.32* | *94.41* |
| ChemScraper (Born-Digital – PDF input)<br>* Skipping rendering errors | ***98.42*** | ***96.20*** | *94.41* |
| ChemScraper (Visual – PNG input) | 85.02 | - | - |

# Conclusion

**Born-digital parser**

1. **Simple:** no OCR, vectorization or GPU, simple geometrical and chemical constraints
2. **Interpretability:** visual correspondence of output symbols with the input PDF
3. **Accessible:** output CDXML directly editable in ChemDraw, easily converted to other formats (SMILES, MOL, InChI)

# Conclusion

**Annotated data generation**

1. **Efficiency:** reduces time and effort for generating large datasets
2. **Consistency:** uniform and accurate annotations
3. **Generalizability:** generalizable to other visual parsing tasks

# Conclusion

**Visual Parser**

1. **Pruned LOS Graph:** efficiently captures spatial relationships, reducing complexity and improving accuracy.
2. **Visual primitives:** computational geometry-based, deterministic
3. **Discrete Attention:** updates query and context images based on predicted segmentation
4. **Training:** on annotated data generated by born-digital parser

# Thank You

This material is based on upon work supported by the National Science Foundation (USA) under Grant No. 2019897 (Molecule Maker Lab Institute project)

We thank Matt Langsenkamp, Matt Berry, Kate Arneson, and other members in the NCSA team, who contributed to the online ChemScraper online system

Code

gitlab.com/dprl/graphics-extraction

System

chemscraper.frontend.staging.mmli1.ncsa.illinois.edu

NSF

MOLECULE MAKER LAB INSTITUTE

dprl

Document and Pattern Recognition Lab