



THRUST 1

Reconstructing the Structure of Molecule Figures from PDF Documents using a CNN-Attention-Based Parsing Model

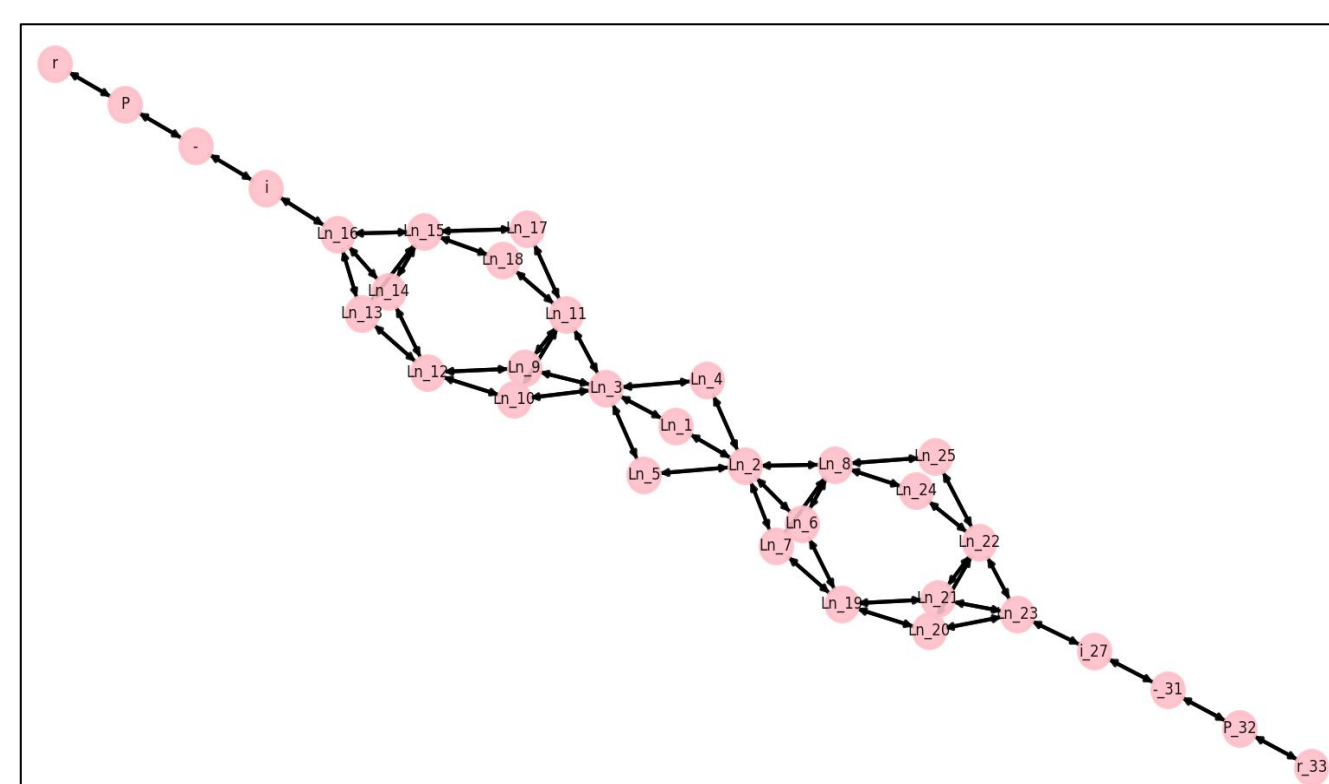
Abhisek Dey, Ayush K. Shah and Richard Zanibbi

Document and Pattern Recognition Lab, Rochester Institute of Technology

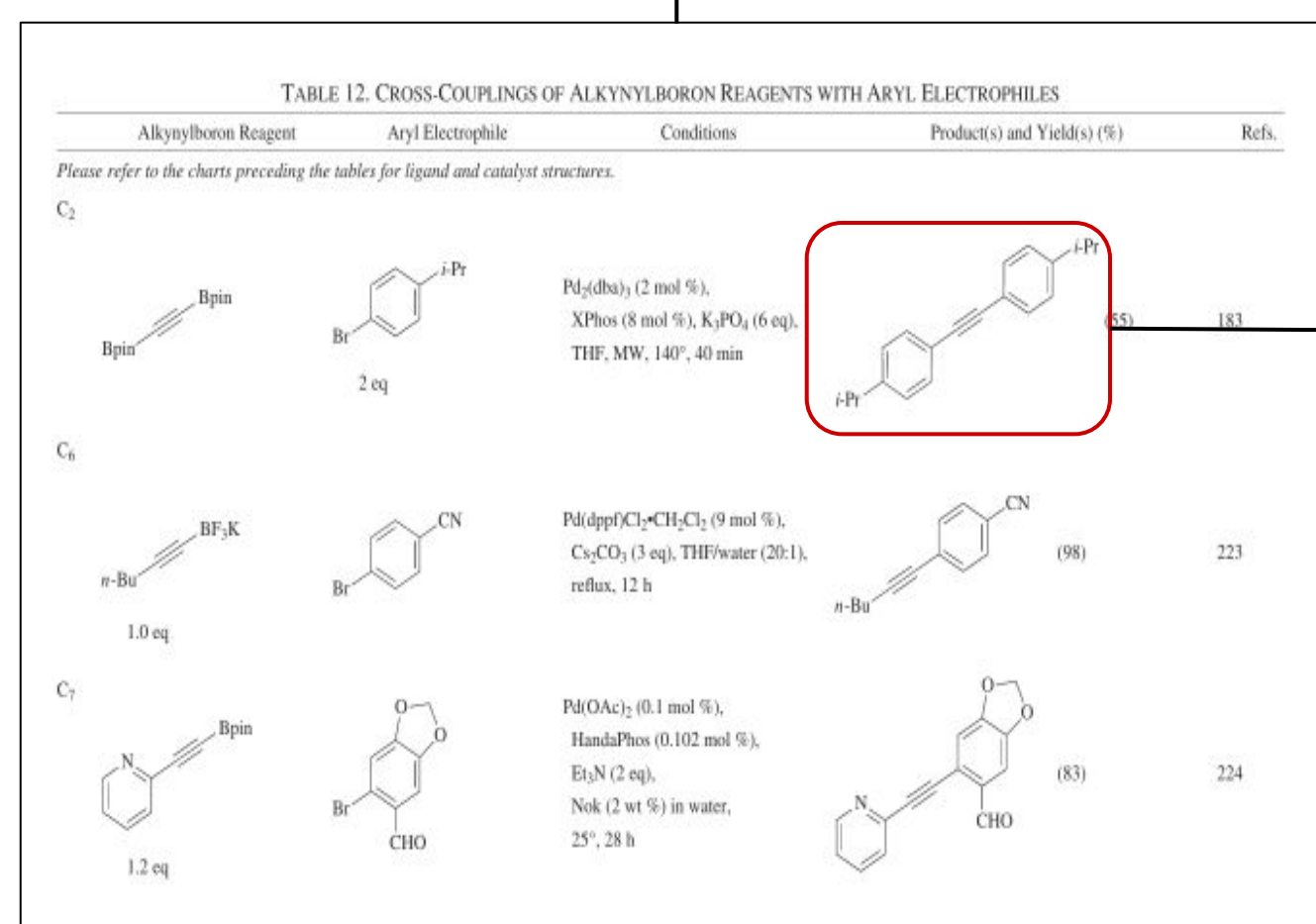


RIT

CDXML Visual Graph (For training)



Data Generation Process (with CDXML)



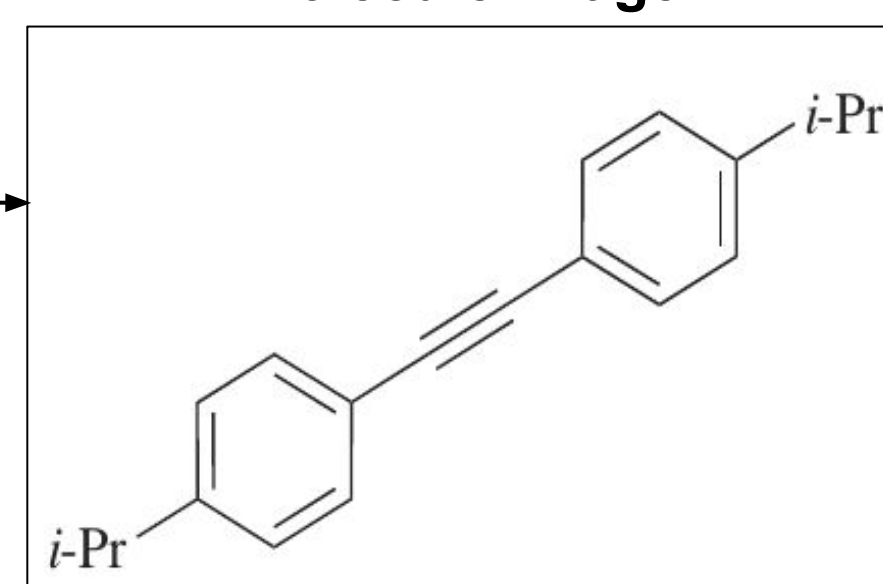
PDF Page

Input Label Graph File

```
# OBJECTS
O, 0, _, 1.0, 0
O, 1, _, 1.0, 1
O, 2, _, 1.0, 2

# SYMBOL LOCATIONS
#cc, 0, 66, 0, 85, 16
#cc, 1, 66, 18, 84, 24
#cc, 2, 52, 28, 72, 59
```

Molecule Image



Global Feature Map (CxHxW)

Convolution Block SE-ResNext Block with Downsample SE-ResNext Block without Downsample

Parsing Model

- Two input attention heads for edge and node masks
- Backbone CNN processing global molecule image features

Preliminary Results

	Nodes			Edges		
	F1	P	R	F1	P	R
Node Level	84.0	84.0	84.0	30.6	23.9	41.5

Output Label Graph File
Predicted Visual Graph (without CDXML)

```
# OBJECTS
O, i_3001_1, i, 1.0, 0
O, -, 3001_2, -, 1.0, 1
O, P_5, Ln, P, 1.0, 2

# RELATIONSHIPS
R, i_3001_1, -, MERGE, 1.0
R, -, 3001_1, i_3001_2, MERGE, 1.0
R, P_3001_2, -, 3001_1, MERGE, 1.0

# SYMBOL LOCATIONS
#cc, 0, 66, 0, 85, 16
#cc, 1, 66, 18, 84, 24
#cc, 2, 52, 28, 72, 59
```

Motivation

Reconstruct the structure of the molecule image from a PDF by training by a structure parsing model by using the data generated from PDF-CDXML set of files.

- Use molecule images and Visual Graph generated to train
- Reconstruct the figure visual graph when CDXML unavailable

Dataset

249 molecules from 26 PDF pages were divided into **191** for train and **58** for test.

In Progress

- Generate more training data to alleviate the issue of data sparsity
- Improve relation scores by selecting/developing a better graph traversal algorithm for extracting final connectivity graph.
- Convert the graph back to a CDXML representation to generate SMILES