

Parsing of Math Formulas and Chemical Diagrams using Graph-Based Representation and Attention Models

Ayush Kumar Shah, as1211@rit.edu

Adviser: Dr. Richard Zanibbi, rxzvcs@rit.edu

Dissertation Committee: Dr. Qi Yu, Dr. Weijie Zhao, Dr. Masaki Nakagawa

PhD Proposal Defense

Golisano College of Computing and Information Sciences

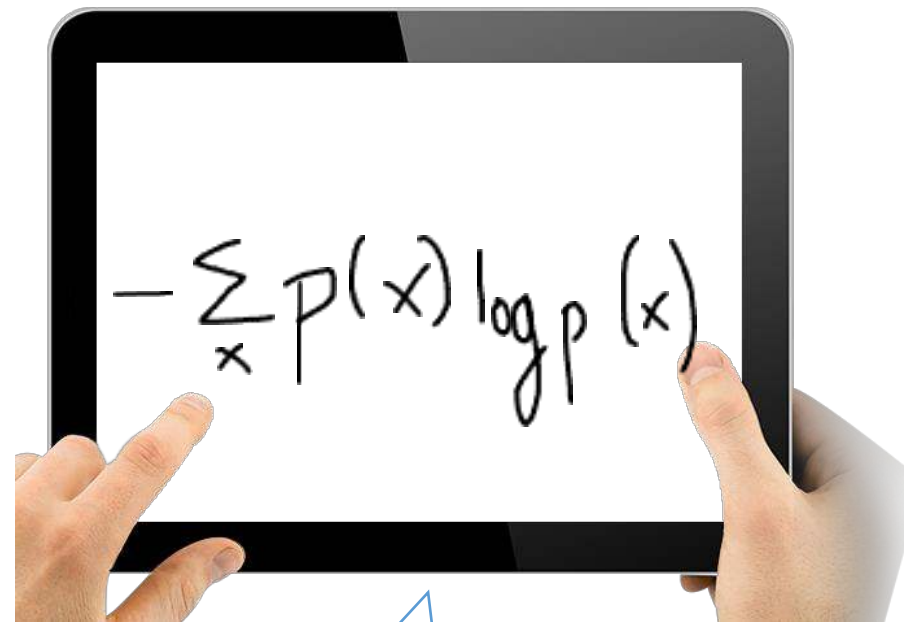
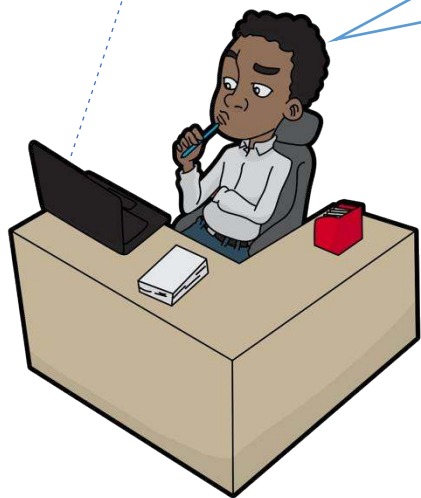
Dec 19, 2024

Introduction

Motivation

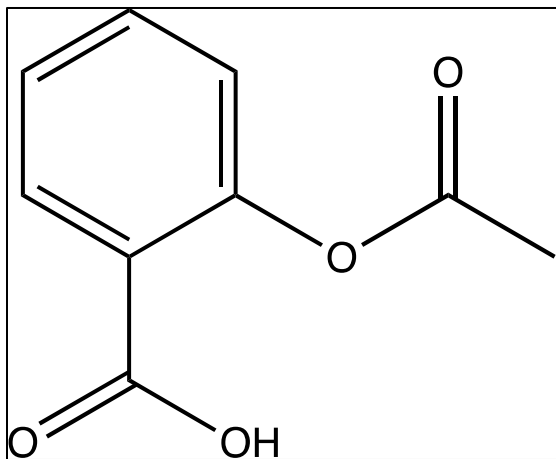
$$-\sum_x p(x) \log p(x)$$

What is the name
of this formula?

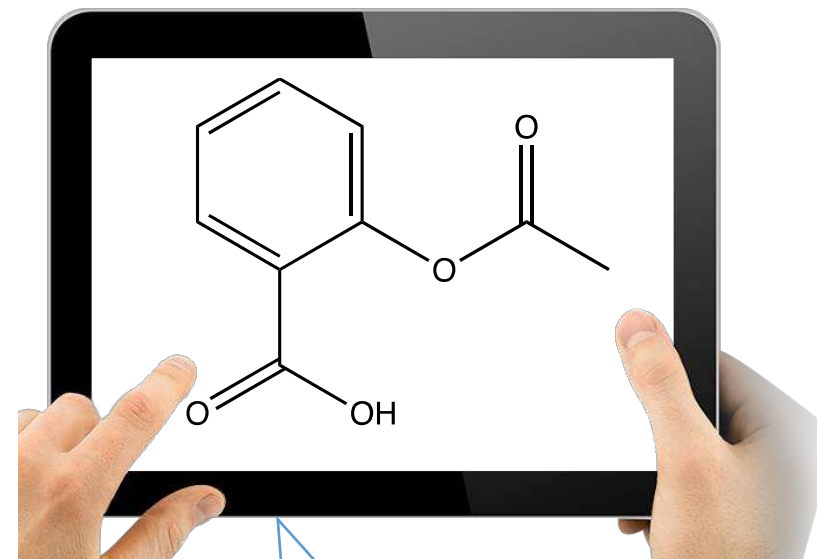
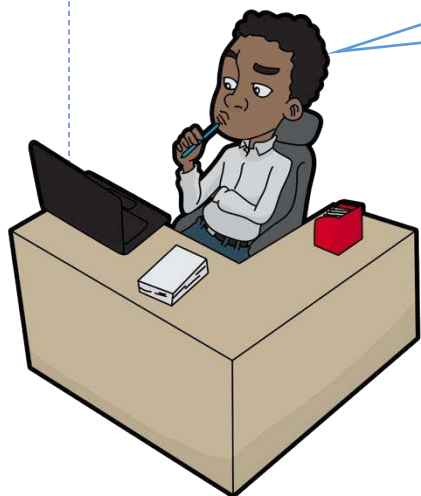


Let's find documents
describing this formula.

Motivation



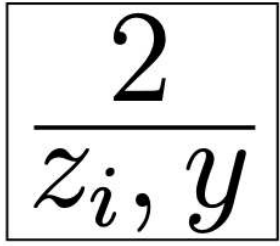
Which catalysts can improve the yield of of this chemical (aspirin)?



Let's find reactions producing this molecule.

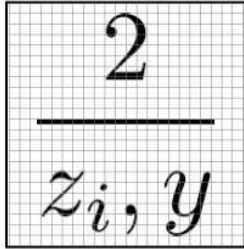
Introduction: Math Formula Parsing

PDF



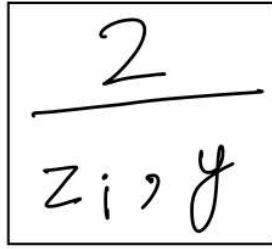
PDF representation of the formula $\frac{2}{z_i, y}$. The numerator '2' is above a horizontal line, and the denominator 'z_i, y' is below it.

Raster

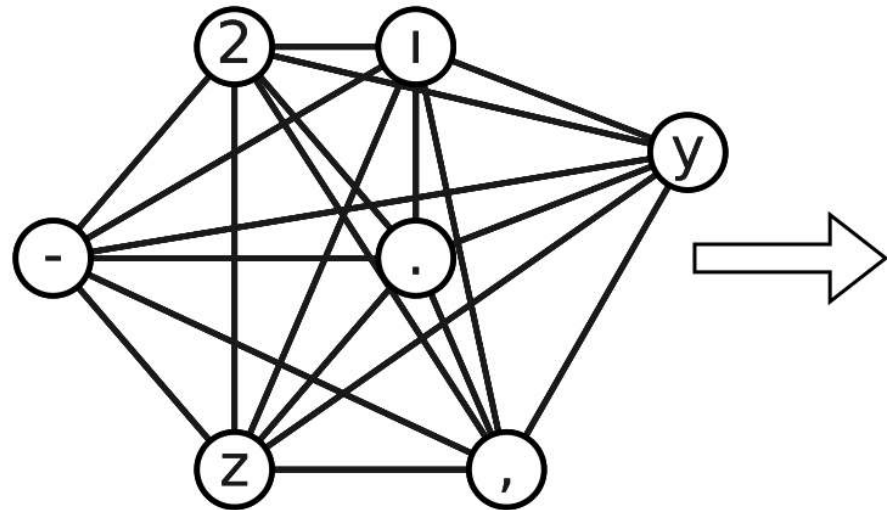


Raster representation of the formula $\frac{2}{z_i, y}$. The formula is overlaid on a grid.

Handwritten

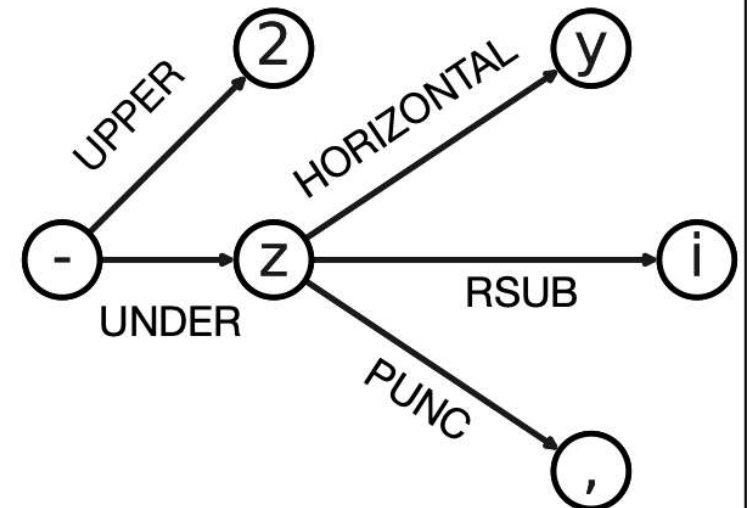


Handwritten representation of the formula $\frac{2}{z_i, y}$.



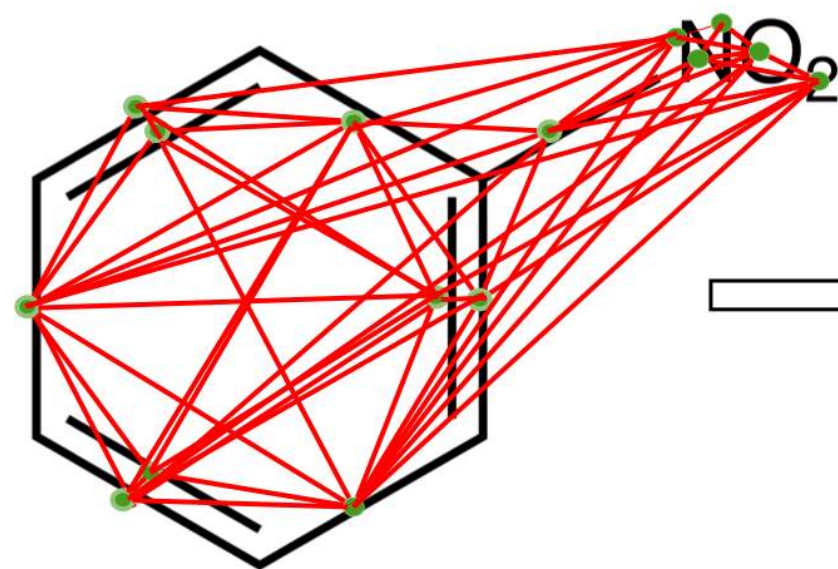
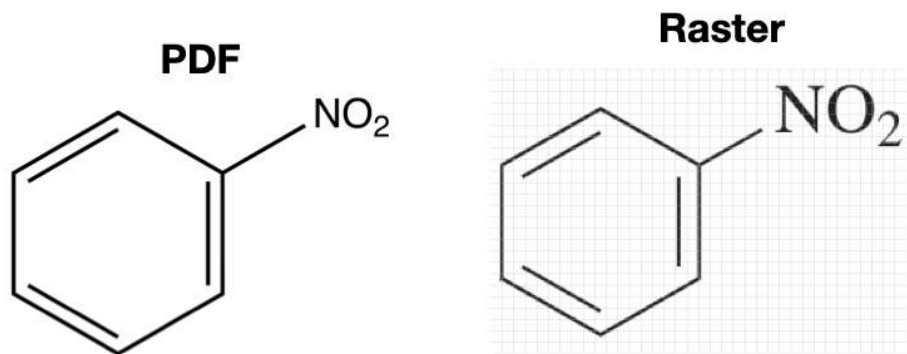
Input: Complete Graph

Parser

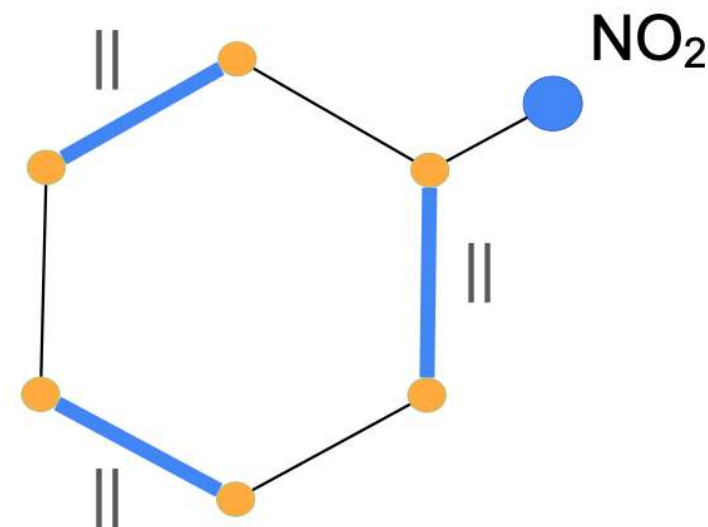
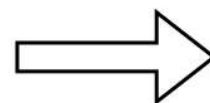
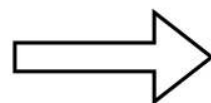


Output: Symbol Layout Tree (SLT)

Introduction: Molecule Diagram Parsing



Input: Complete Graph



Output: Molecular Graph

Contributions



Parsing mathematical formulas and molecular diagrams using vector information



Shape geometry based annotated ground truth generation for molecular diagrams



Graph attention parsing model using line-of-sight (LOS) neighbor context for images



Recurrent segmentation-aware parser for raster molecule images

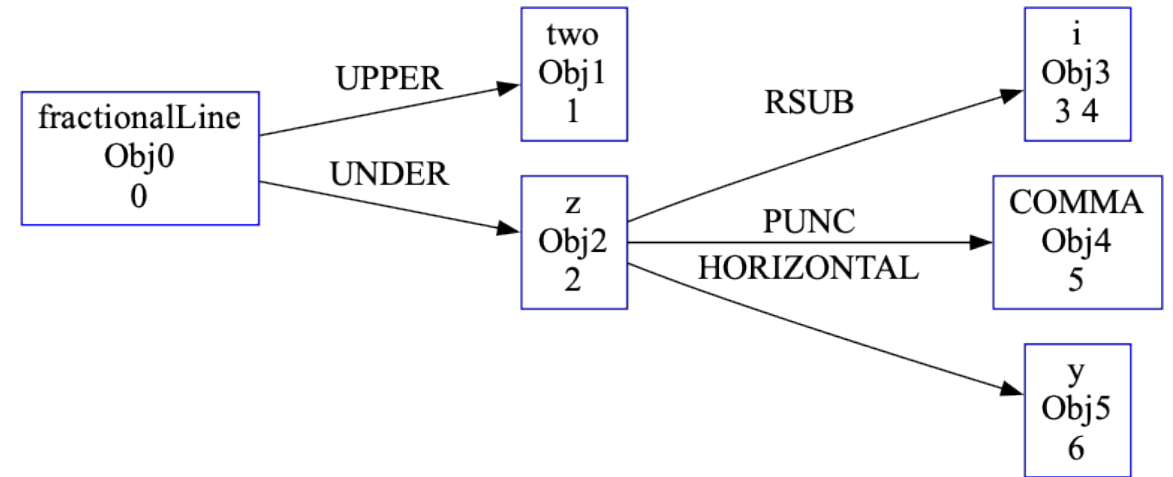
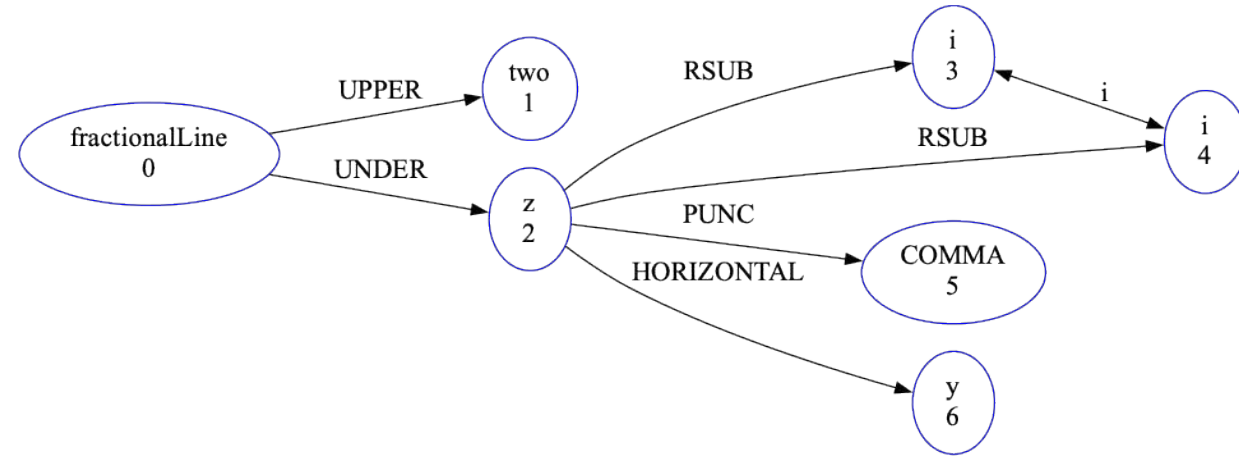
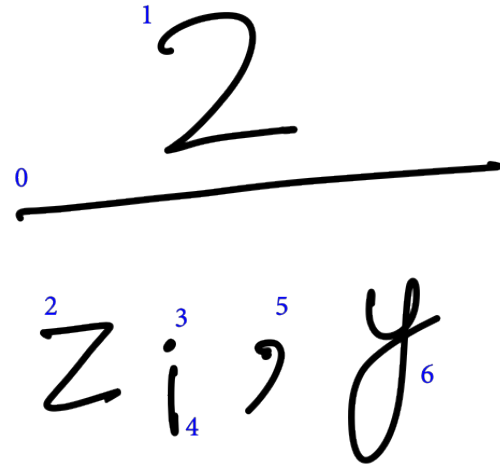
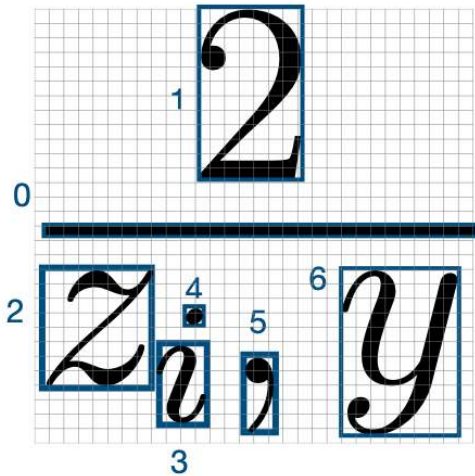
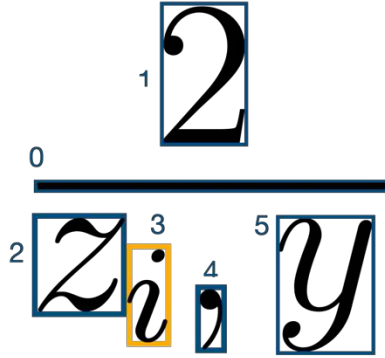


Enhanced visualization tool for analyzing graph-based parsing errors in context

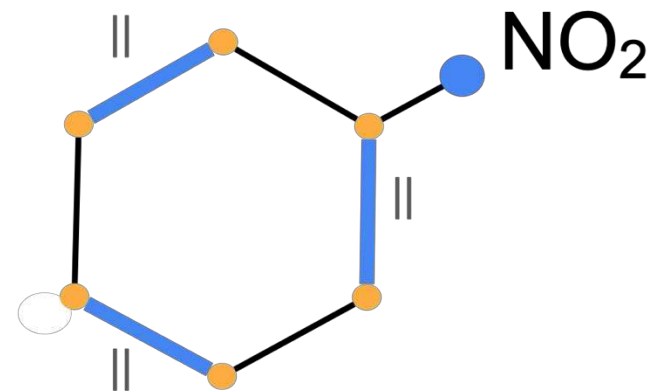
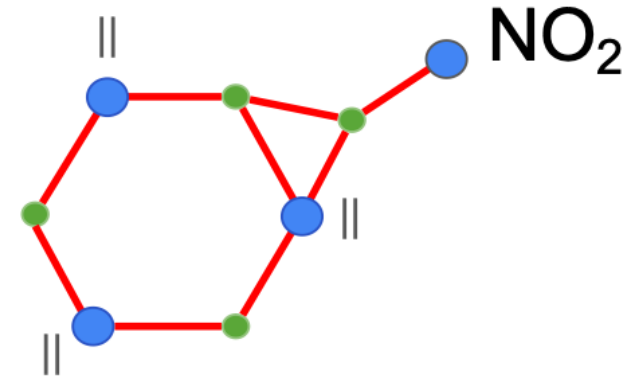
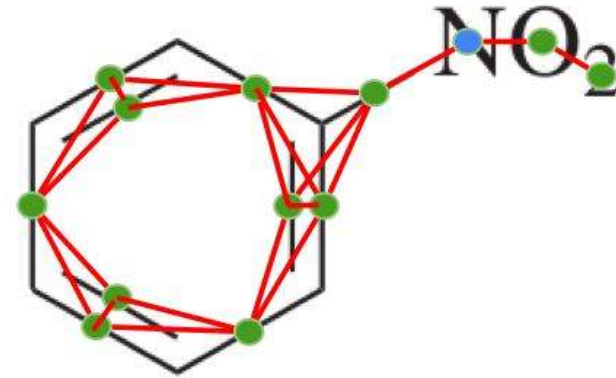
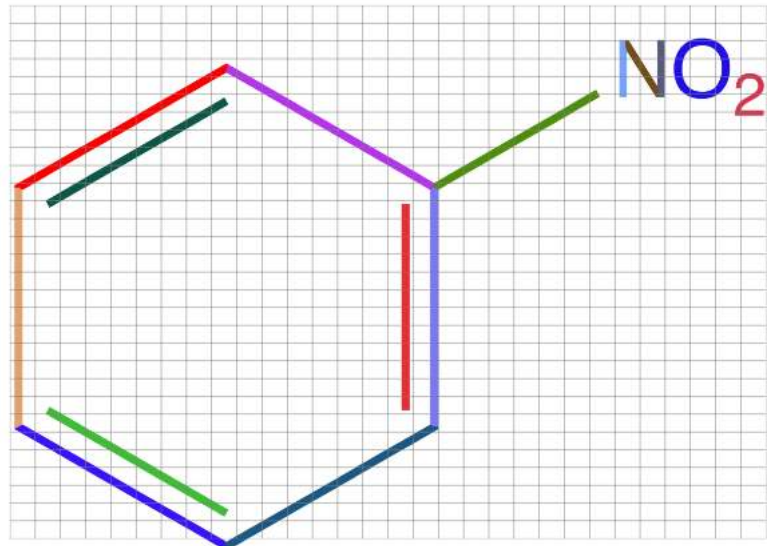
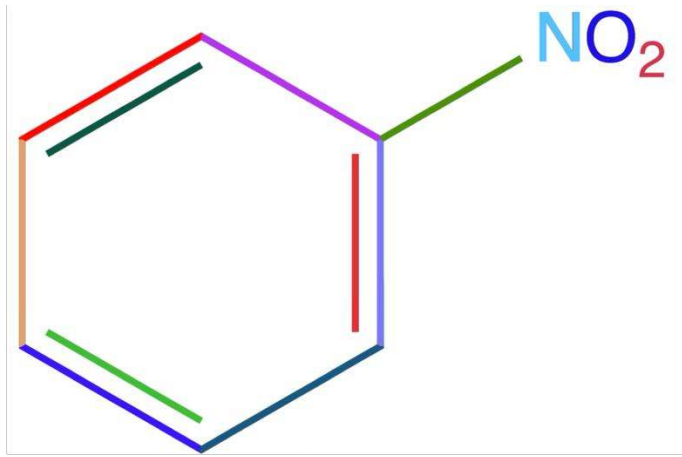


Fine-grained graph-based evaluation framework and metrics for chemical diagrams

Inputs and Outputs: Math Formula

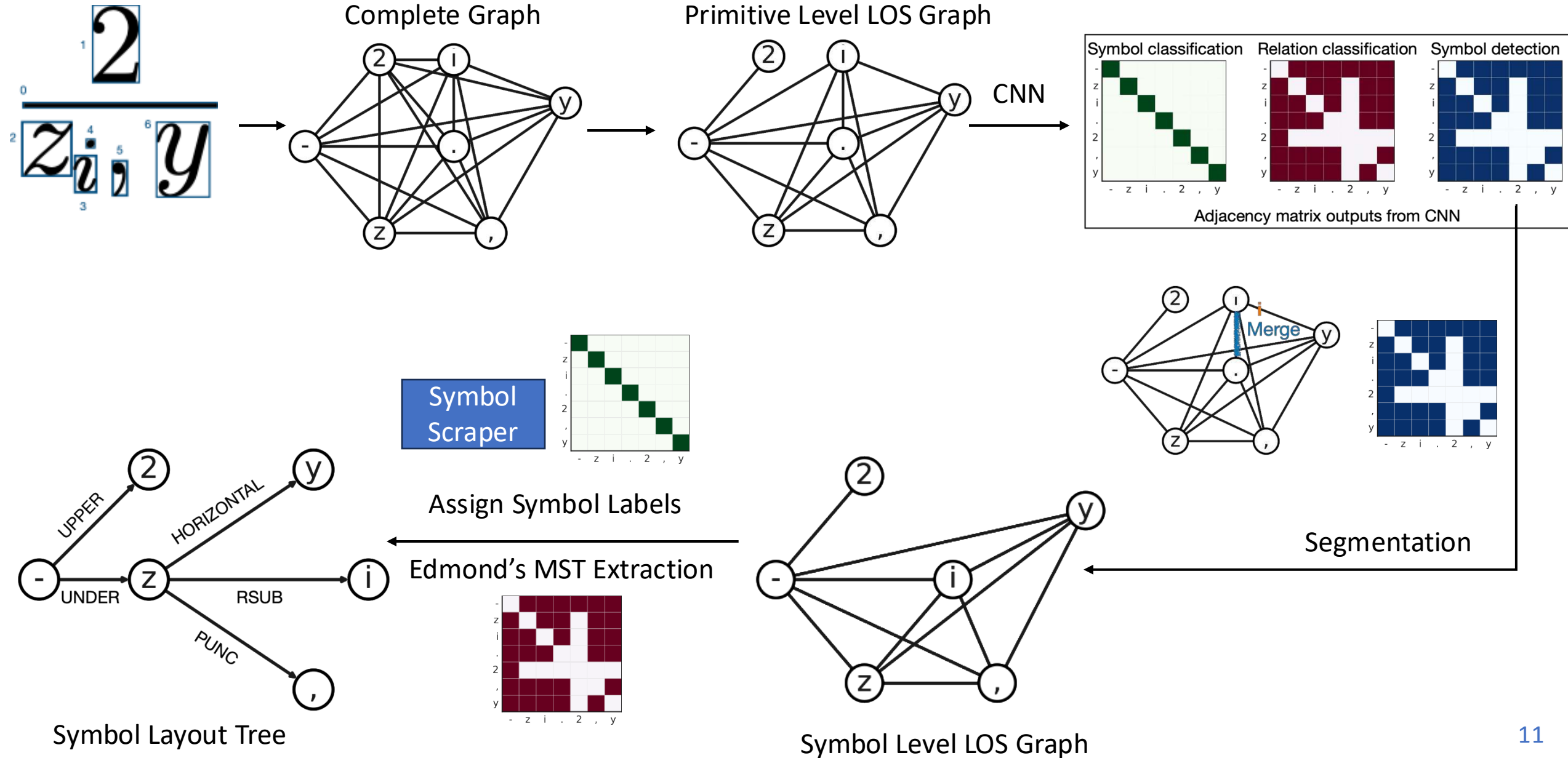


Inputs and Outputs: Molecule Diagram

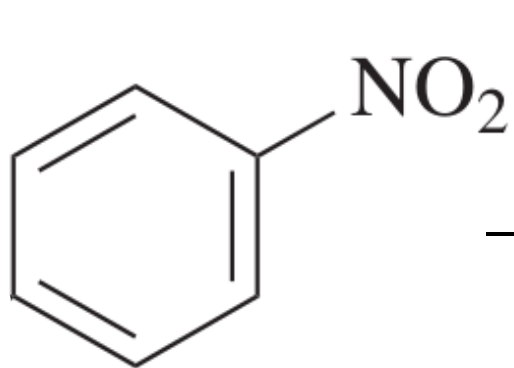


Born-digital Parser

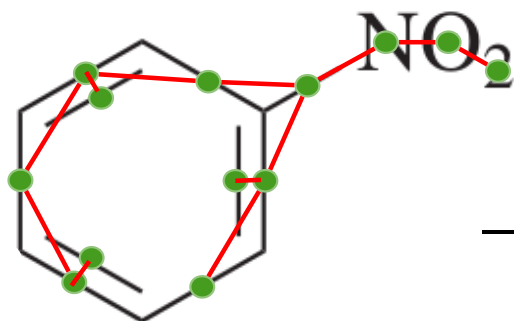
Born-digital Parser: Math Formula (PDF)



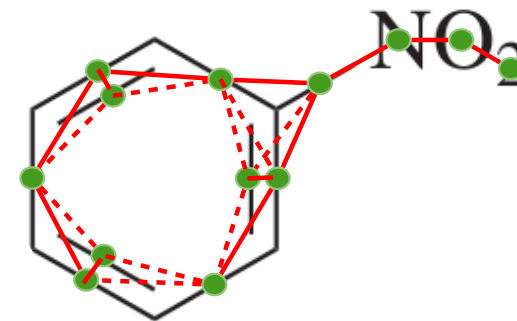
Born-digital Parser: PDF Molecule Image



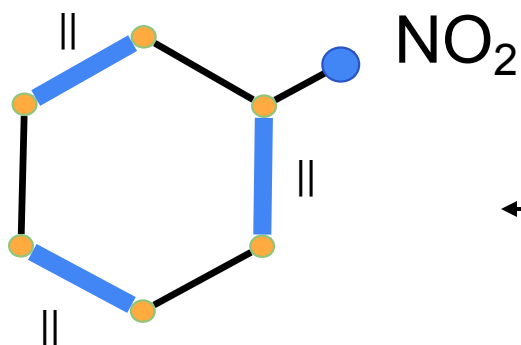
(a) **PDF Image**



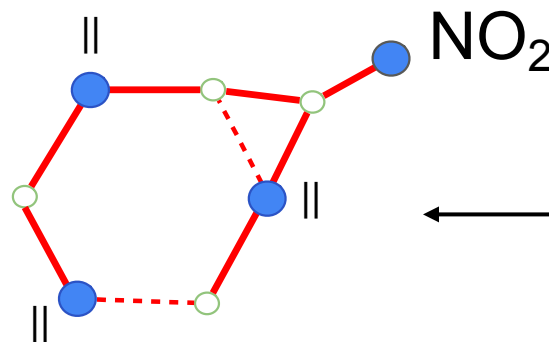
(b) **MST**
nodes: lines & characters
edges: connections/merges



(c) **Visual Graph**
nodes: lines & characters
edges: connections/merges

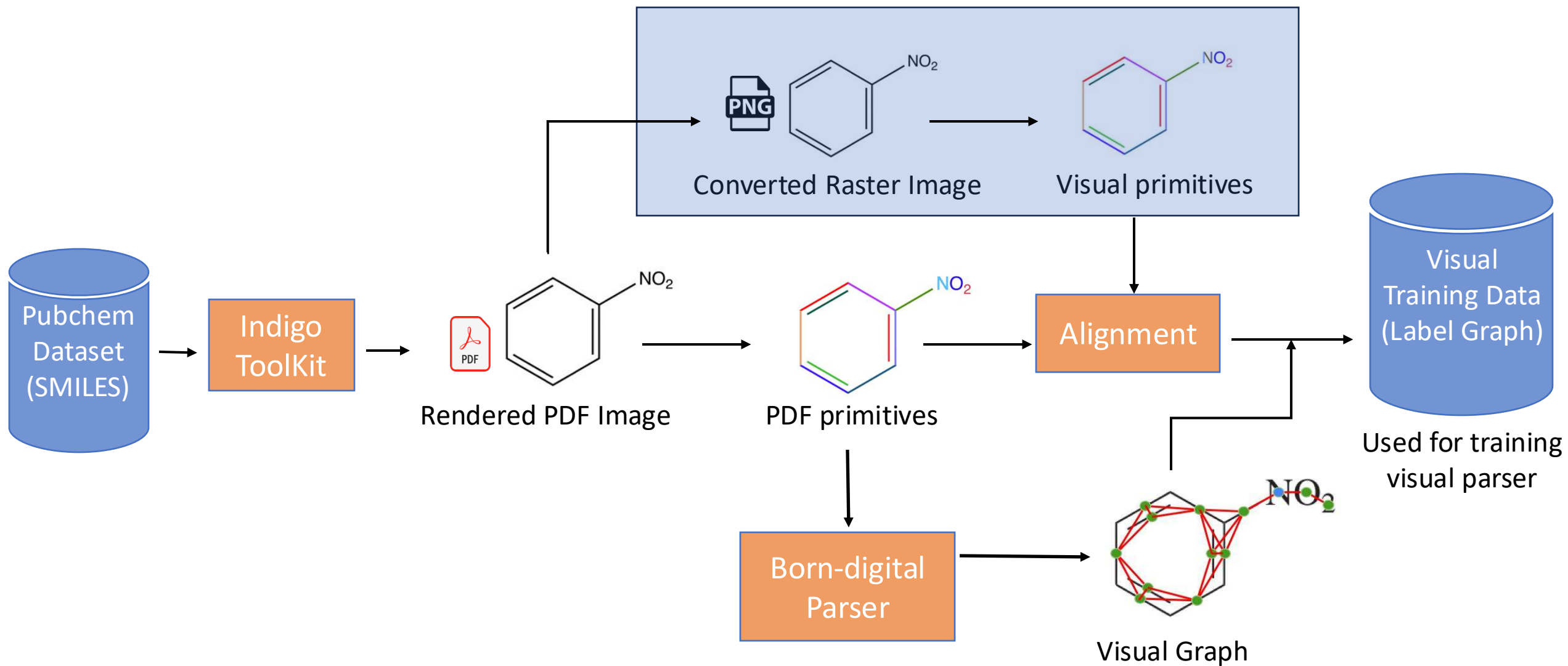


(e) **Molecular Graph**
nodes: atoms & superatoms
edges: bonds

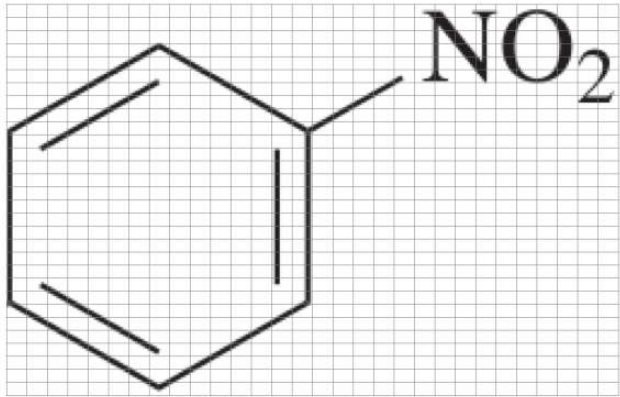


(d) **Tokenized Visual Graph**
nodes: bonds, atoms & superatoms
edges: connections

Annotated Data Generation for Visual Parser

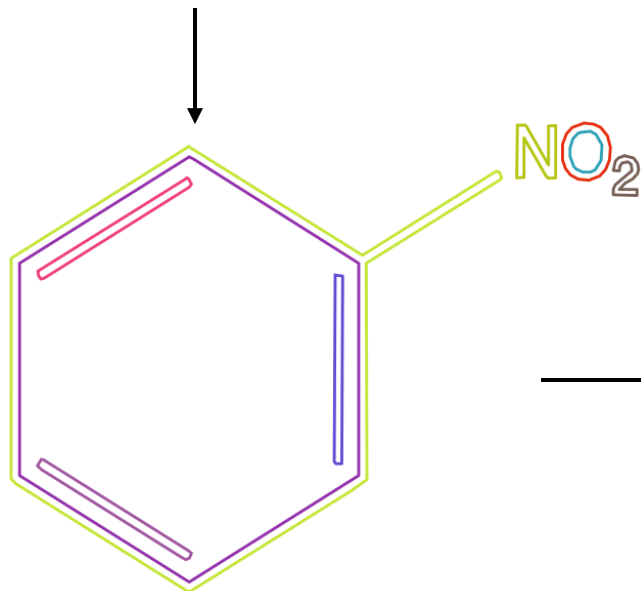


Visual Primitives Extraction for Raster Images (PNG)

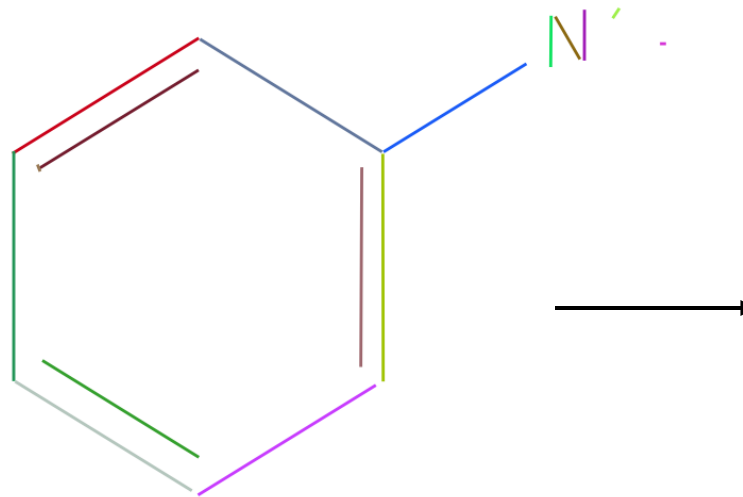


Input Raster Image

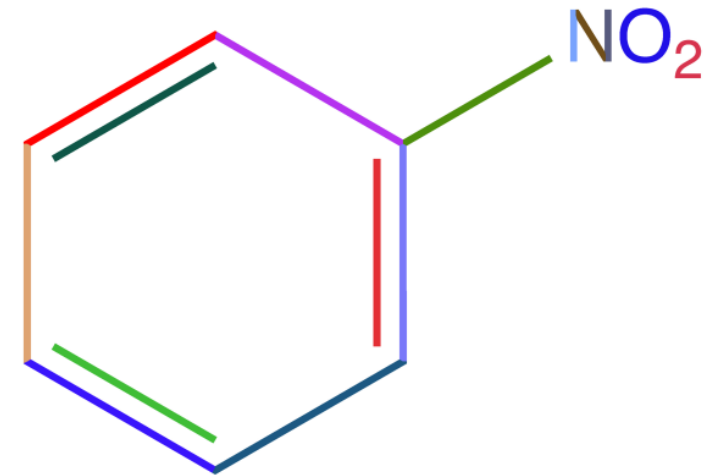
- **Build polygons** from Connected Components (CCs)
- **Extract skeletons** from medial axis of parallel lines
- **Segment CCs:** flood fill assigning each pixel to its closest skeleton



Simplified Polygons



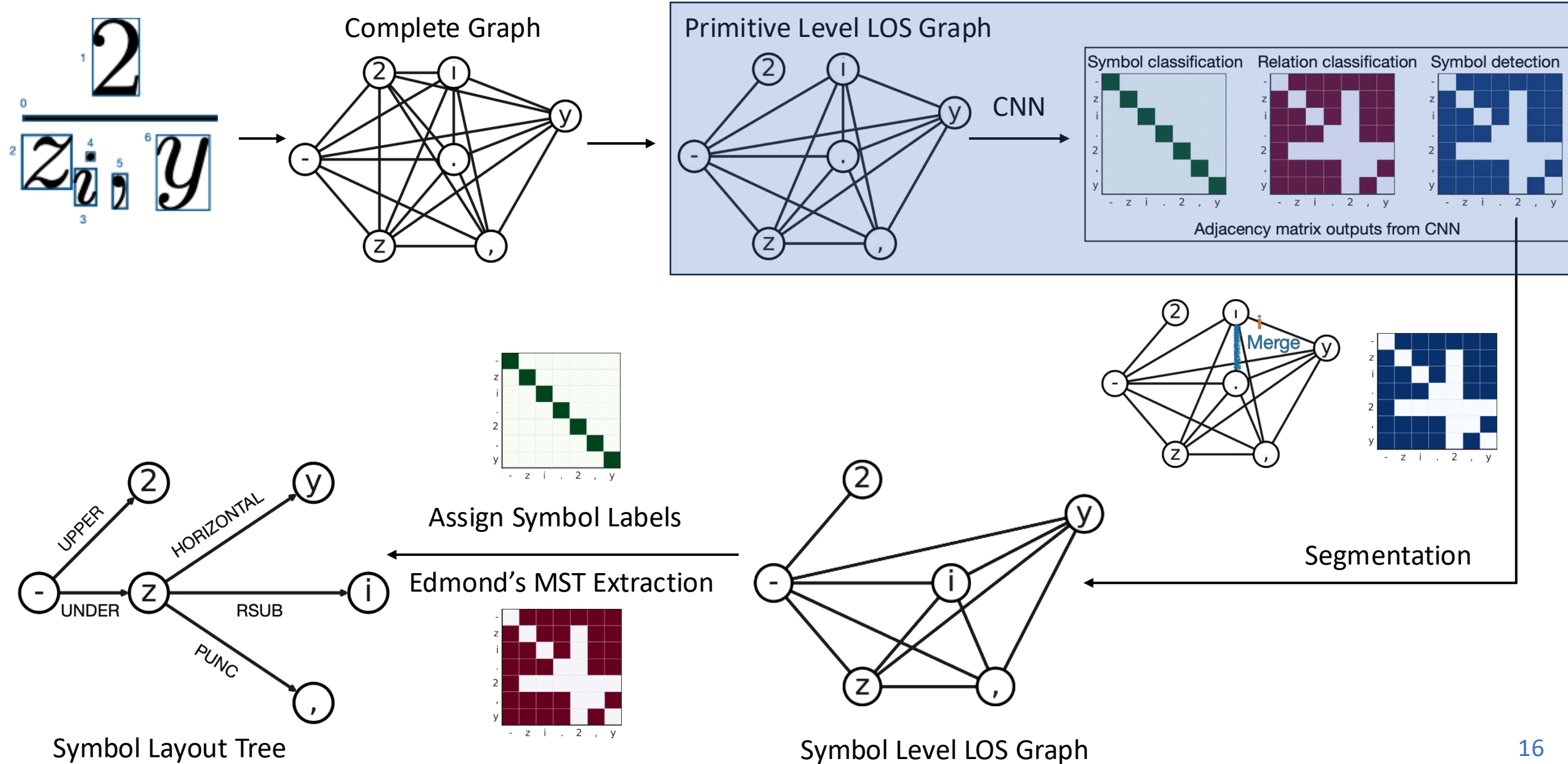
Skeletal Lines

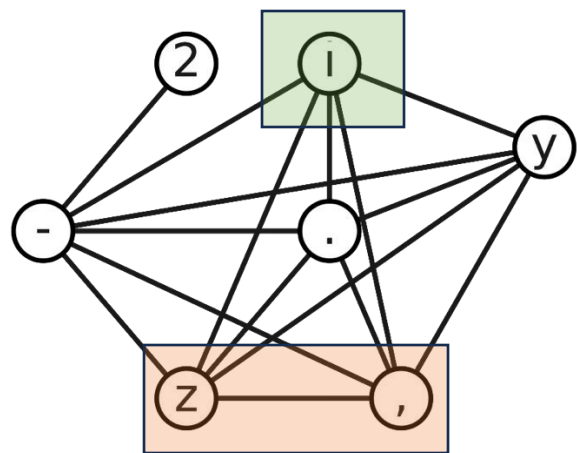


Visual Primitives

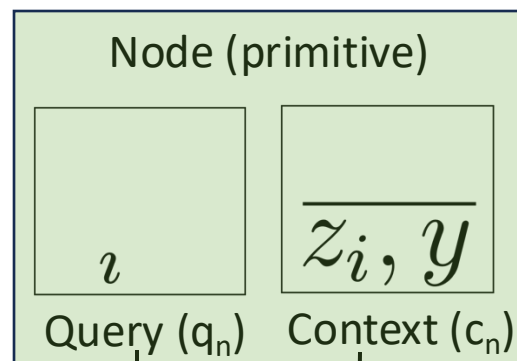
Visual Parser

Visual Parser: Raster Math Formula Image





Primitive Level LOS Graph

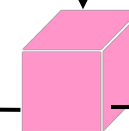


Attention Block

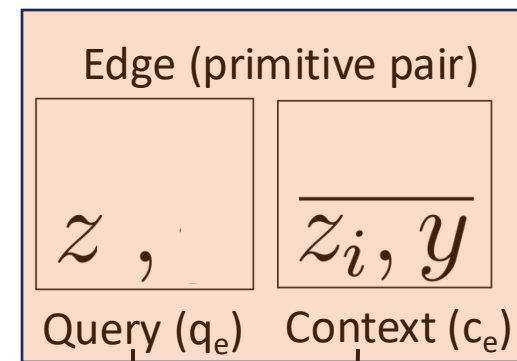
Formula image

$$\frac{2}{z_i, y}$$

SE-ResNext

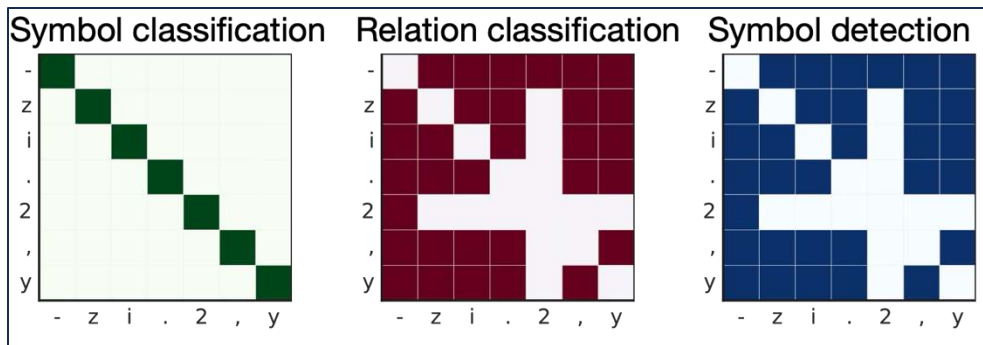


Global Feature Map



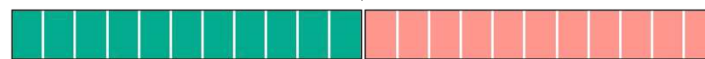
Attention Block

Probability outputs



$\otimes$

$q_n || c_n$



Linear



Symbol
classes (207)

$\otimes$

$q_e || c_e$



Linear



Relation
classes (9)

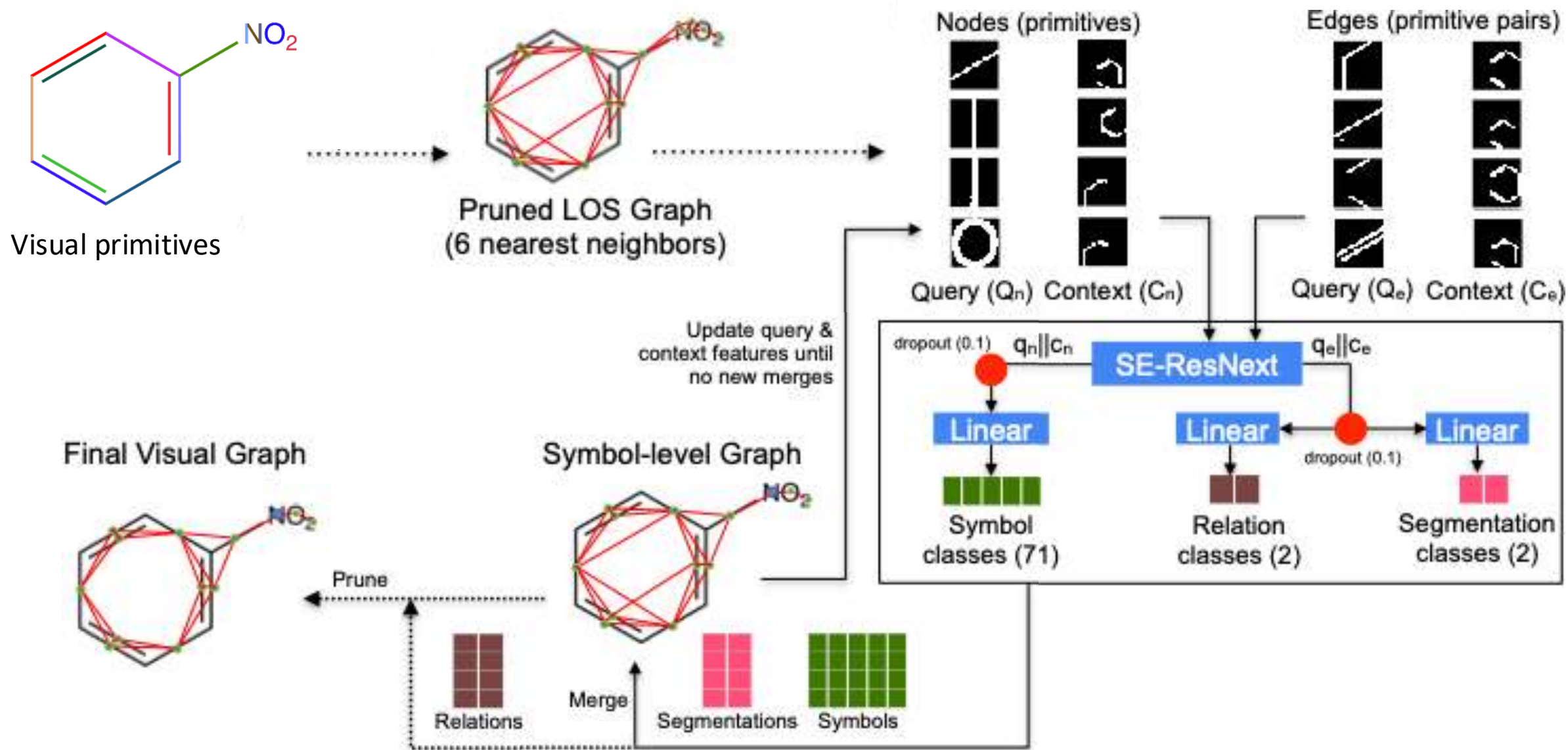
Linear



Segmentation
classes (2)

Line-of-sight Graph Attention Parser (LGAP)

Visual Parser: Raster Molecule Diagram (LCGP)

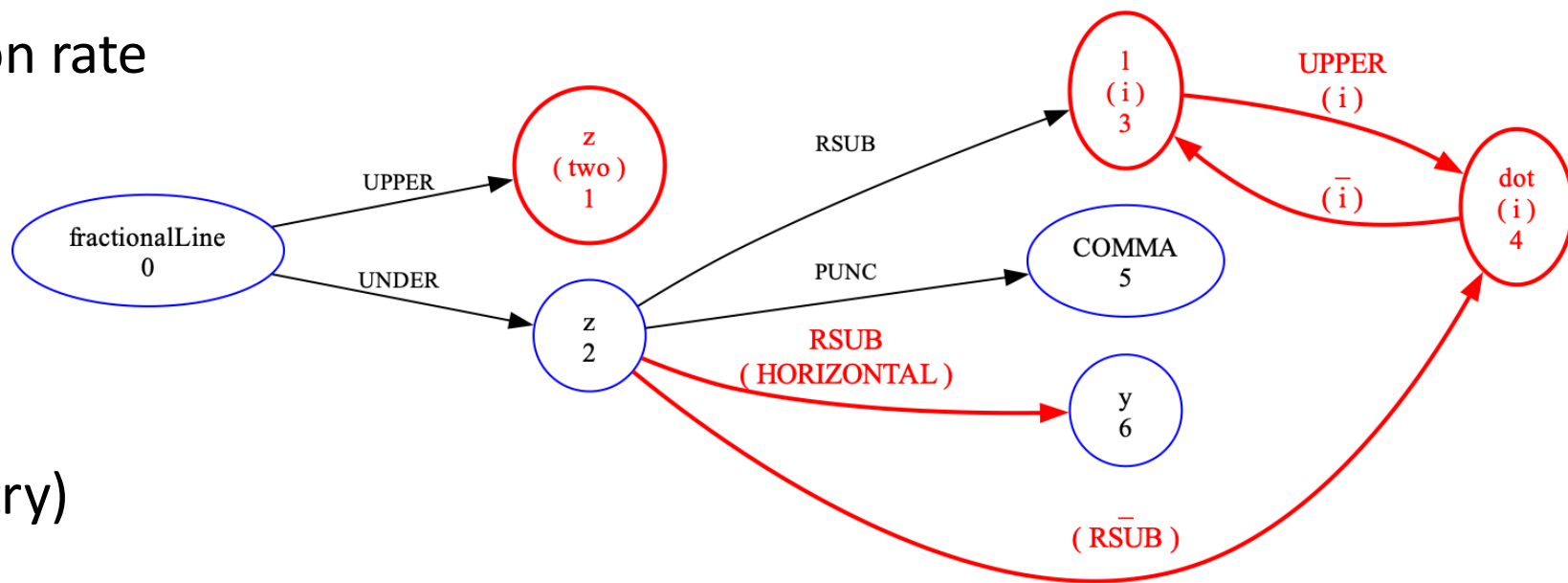
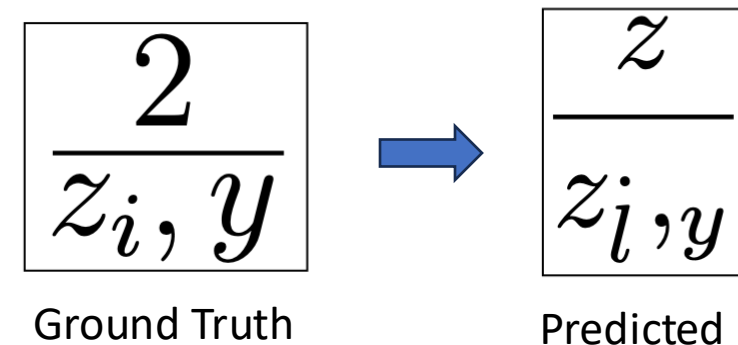


Results

Results: Evaluation Metrics

Graph-based Metrics (LgEval)

1. Detection F-scores (symbol & relationship)
2. Detection + classification F-scores (symbol & relationship)
3. Unlabeled structure recognition rate
4. Labeled structure recognition rate



String-based Metrics

- Exact SMILES match (chemistry)

Primitive level graph (with errors)

Results

Math Formula Recognition

Results: Math Formula Recognition

Handwritten Formula Recognition (Handwritten Strokes)

Datasets

Dataset	No. of expressions
CROHME 2019 Training set	9993
CROHME 2019 Test set	1199
CROHME 2016 Test set	1147
CROHME 2014 Test set	986

Benchmark results on CROHME 2019 test set

Systems	Symbols		Relationships		Formulas	
	Detection	Det. + Class	Detection	Det. + Class	Structure	Str. + Class
USTC-iFLYTEK	-	-	-	-	91.49	80.73
QD-GGA	97.30	89.09	92.19	90.76	65.55	38.81
LGAP (Ours)	99.10	93.59	94.39	93.51	70.23	40.70

Benchmark results on CROHME 2014 and 2016 test sets

Systems	2014	2016
	Str. + Class	Str. + Class
SON	37.63	-
IM2TEX	38.74	-
WAP	46.55	44.55
TAP	61.6	57.02
MAN	54.05	50.56
QD-GGA	32.04	32.84
Graph-to-Graph	54.46	52.05
LGAP (Ours)	42.23	41.67

Results: Math Formula Recognition

Typeset Formula Recognition (Raster Images)

Typeset formula dataset

Dataset	No. of expressions
InftyMCCDB-2 Training set	12,251
InftyMCCDB-2 Test set	6,830

Effect of adding LOS neighborhood masks to LGAP (InftyMCCDB-2 test set)

LOS context vector	Feature vector size	Symbols		Relationships		Formulas	
		Detection	Detection + Classification	Detection	Detection + Classification	Structure	Structure + Classification
False	1408	97.96	94.48	94.36	93.89	88.70	78.90
True	2816	98.32	95.66	94.85	94.35	89.27	83.27
False	512	98.39	95.21	94.63	94.28	89.21	81.77
True	1024	98.39	95.49	94.85	94.46	89.36	81.89

Results: Math Formula Recognition

Typeset Formula Recognition (Raster Images)

Typeset formula dataset

Dataset	No. of expressions
InftyMCCDB-2 Training set	12,251
InftyMCCDB-2 Test set	6,830

Benchmarking MST-based parsing models on InftyMCCDB-2 test set

MST Model	Symbols		Relationships		Formulas	
	Detection	+Classification	Detection	+Classification	Structure	+Classification
QD-GGA	98.50	94.54	94.43	93.96	87.77	76.72
LPGA _{RF}	99.34	98.51	97.83	97.56	93.81	90.06
LPGA _{CNN}	99.35	98.95	97.97	97.74	93.37	90.89
LGAP (Ours)	98.32	95.66	94.85	94.35	89.27	83.27
LGATI (Ours)	99.48	98.60	97.11	96.61	91.67	88.11

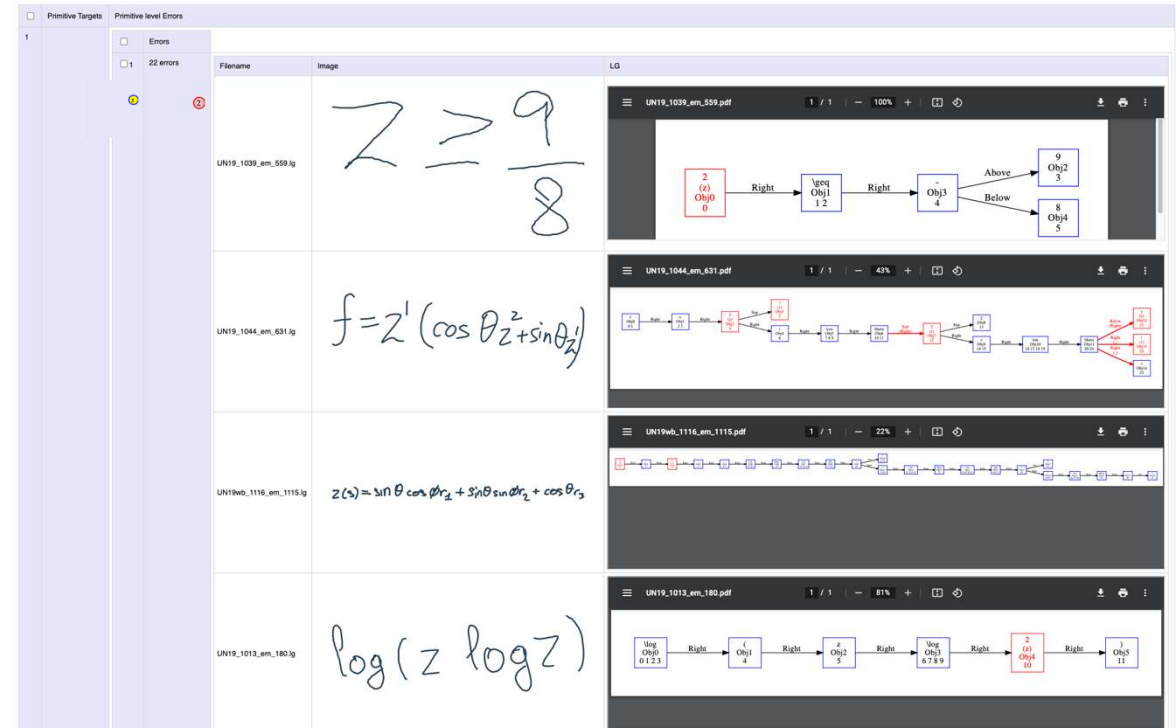
LgEval Extension: Error Visualization

Object Confusion Histograms

Object structures recognized incorrectly are shown at left, sorted by decreasing frequency. 95 incorrect targets, 1418 errors.

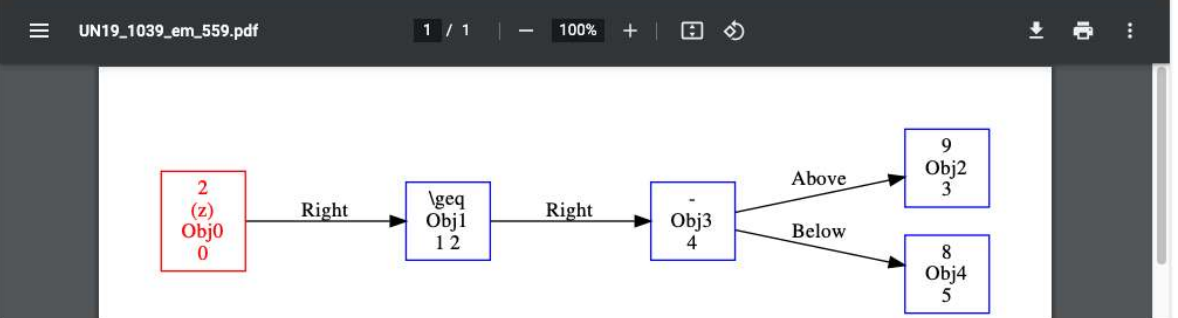
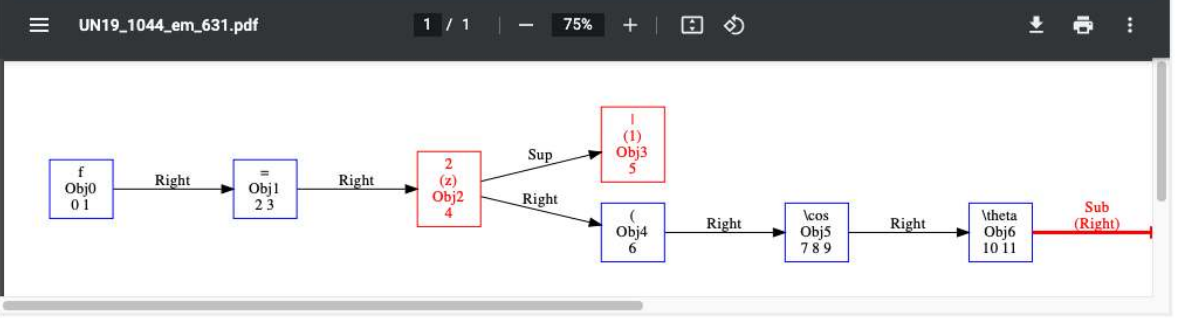
☐ Object Targets		Primitive Targets and Errors						
3	55 errors 	☐ Targets						
		☐ 1	51 errors 	☐ 22 errors 	☐ 7 errors 	☐ 5 errors 	☐ 4 errors 	☐ 3 errors 
		☐ 2	4 errors  ... 	☐ 2 errors  ... 	☐ 2 errors  ... 			

Errors organized by decreasing frequency



Specific instances where 'z' is misclassified as '2',
seen after [clicking on the '22 errors' link](#)

LgEval Extension: Error Visualization

Primitive Targets		Primitive level Errors		
1		☐ Errors		
		☐ 1 22 errors		
			Filename	Image
			UN19_1039_em_559.lg	
			UN19_1044_em_631.lg	
				LG
				
				

Zoomed in: Specific instances where 'z' is misclassified as '2,' seen after [clicking on the '22 errors' link](#)

Results

Molecular Diagram Recognition

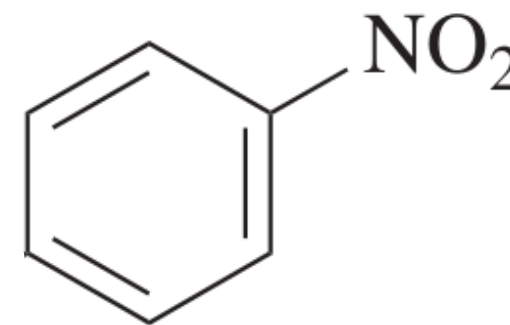
Results: Molecule Diagram Recognition

Training data source: Pubchem 1 million

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

Additional Metrics:

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



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

Results: Molecule Diagram Recognition

Molecular structure recognition benchmarks showing percentages of generated SMILES matching ground truth

Systems	Exact SMILES Matches		
	Indigo (5719)	CLEF-2012 (992)	UoB (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

Results: Molecule Diagram Recognition

Molecular structure recognition benchmarks showing percentages of generated SMILES matching ground truth

Systems	Exact SMILES Matches		
	Indigo (5719)	CLEF-2012 (992)	UoB (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) * Skipping rendering errors	98.42	96.20	94.41

Results: Molecule Diagram Recognition

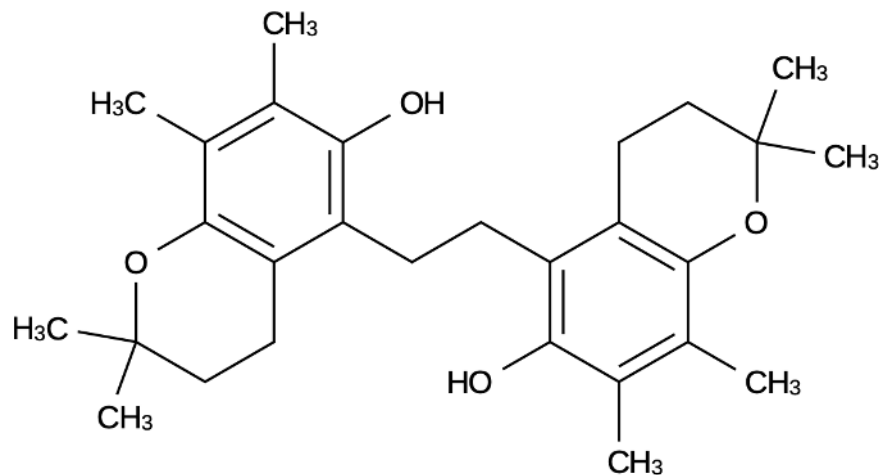
Molecular structure recognition benchmarks showing percentages of generated SMILES matching ground truth

Systems	Exact SMILES Matches		
	Indigo (5719)	CLEF-2012 (992)	UoB (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) * Skipping rendering errors	98.42	96.20	94.41
ChemScraper (Visual – PNG input)	85.02	-	-

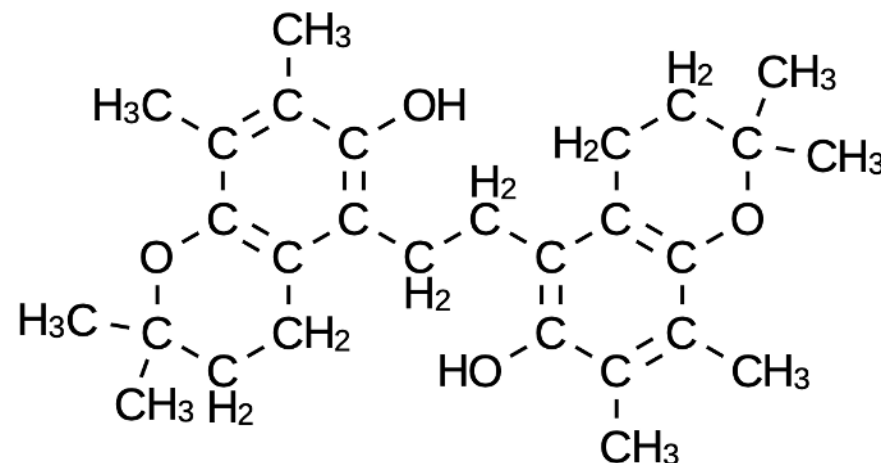
Results: Molecule Diagram Recognition

Born-digital parser label graph metrics for different rendering parameters (5719 molecules).
Shown are F1 measures for symbol labels, correct labels, and complete graphs

Render	Label Mode	Implicit Hydrogens Visible	Relative Thickness	Correct Node (Labels) F1	Correct Edge (Labels) F1	Molecules Struct.	Molecules Struct.+Class
Default	terminal-hetero	true	1	99.96	99.84	98.49	97.62
Hardest	all	true	1.5	99.65	99.01	81.89	81.12



Default: terminal-hetero, implicit hydrogens, thickness=1



Hardest: all, implicit hydrogens, thickness=1.5

Rendering a molecule with different parameters (Indigo toolkit)

Proposed Work

Proposed Work

Thesis statement:

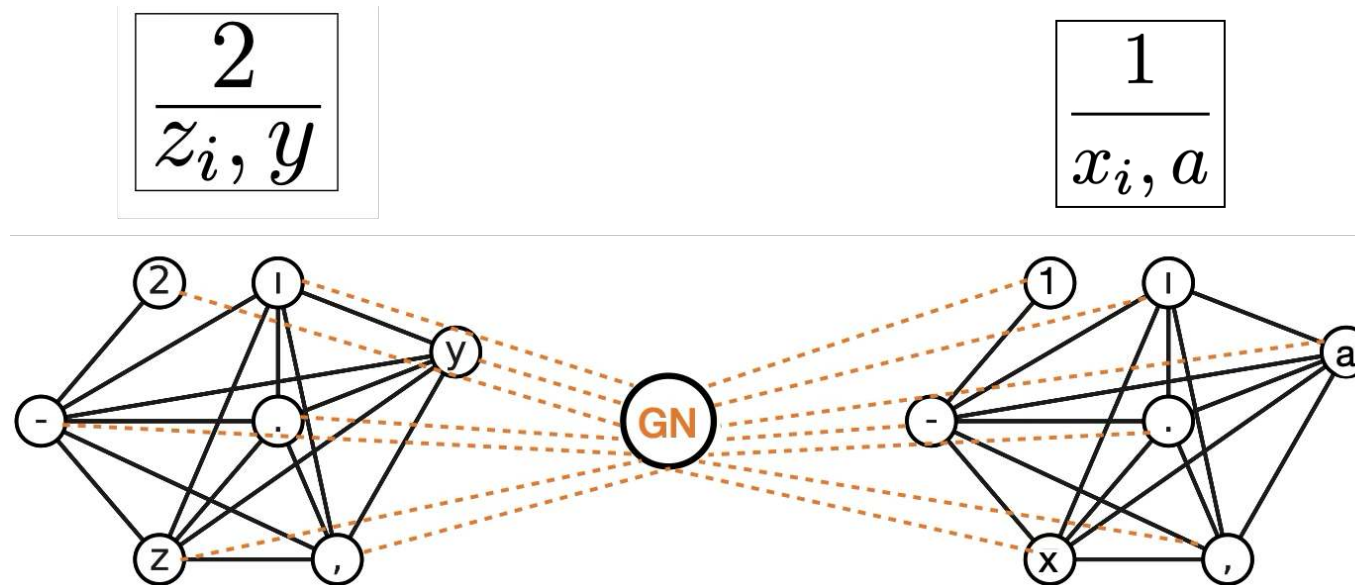
Recognition accuracy of mathematical formulas and chemical diagrams can be increased by

1. enhancing local and global context representation using edge-specific attention models and adding multiple virtual global nodes, and
2. improving interactions among multiple tasks by employing self attention mechanisms, adapted for graph transformers.

Proposed Work: Global Nodes

RQ1: Can the use of **multiple virtual global nodes** improve the global contextual information propagation in graph-based parsers compared to a single global node?

Approach: 1 global node shared across multiple formulas vs multiple nodes per formula



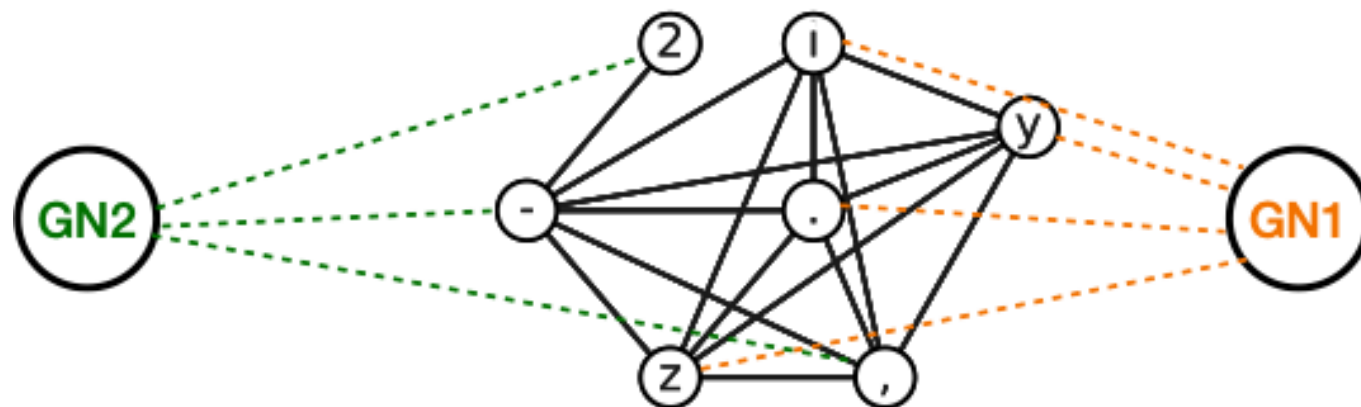
- [1] C. Ying *et al.*, “Do Transformers Really Perform Badly for Graph Representation?,” presented at the Advances in Neural Information Processing Systems, Nov. 2021. Available: <https://openreview.net/forum?id=OeWooOxFwDa>
- [2] C. Diao and R. Loynd, “Relational Attention: Generalizing Transformers for Graph-Structured Tasks,” presented at the The Eleventh International Conference on Learning Representations, Sep. 2022. Available: <https://openreview.net/forum?id=cFuMmbWiN6>
- [3] Y. Xie, R. Zanibbi, and H. Mouchère, “Local and Global Graph Modeling with Edge-weighted Graph Attention Network for Handwritten Mathematical Expression Recognition,” Oct. 24, 2024, *arXiv*: arXiv:2410.18555. doi: [10.48550/arXiv.2410.18555](https://doi.org/10.48550/arXiv.2410.18555).

Proposed Work: Global Nodes

RQ1: Can the use of **multiple virtual global nodes** improve the global contextual information propagation in graph-based parsers compared to a single global node?

Approach: 1 global node shared across multiple formulas vs **multiple nodes per formula**

$$\frac{2}{z_i, y}$$



- [1] C. Ying *et al.*, "Do Transformers Really Perform Badly for Graph Representation?," presented at the Advances in Neural Information Processing Systems, Nov. 2021. Available: <https://openreview.net/forum?id=OeWooOxFwDa>
- [2] C. Diao and R. Loynd, "Relational Attention: Generalizing Transformers for Graph-Structured Tasks," presented at the The Eleventh International Conference on Learning Representations, Sep. 2022. Available: <https://openreview.net/forum?id=cFuMmbWiN6>
- [3] Y. Xie, R. Zanibbi, and H. Mouchère, "Local and Global Graph Modeling with Edge-weighted Graph Attention Network for Handwritten Mathematical Expression Recognition," Oct. 24, 2024, *arXiv*: arXiv:2410.18555. doi: [10.48550/arXiv.2410.18555](https://doi.org/10.48550/arXiv.2410.18555).

Proposed Work: Edge Attention

RQ2: Does the incorporation of **edge-specific attention mechanism**, relational transformer attention, improve the performance of graph-based parsers compared to not incorporating edge features, using GAT as the baseline?

Approach: Use of edge features to guide attention

- Node embeddings added during edge attention and
- Edge embeddings added to nodes during node attention

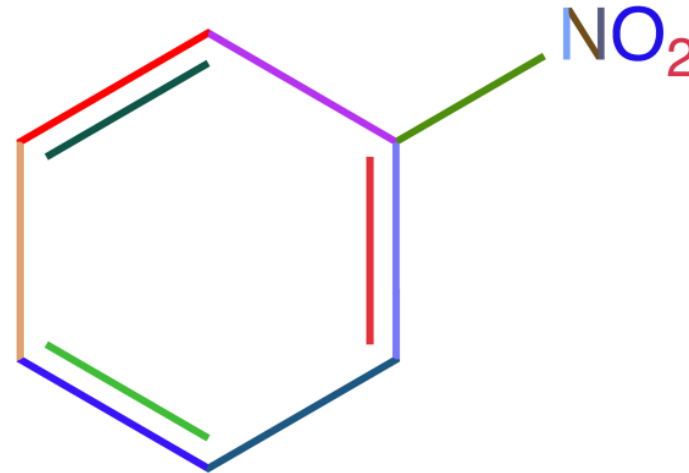
[2] C. Diao and R. Loynd, "Relational Attention: Generalizing Transformers for Graph-Structured Tasks," presented at the The Eleventh International Conference on Learning Representations, Sep. 2022. Available: <https://openreview.net/forum?id=cFuMmbWiN6>

[4] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations*, Feb. 2018. Available: <https://openreview.net/forum?id=rJXMpikCZ>

Proposed Work: Over-segmentation and Noise

RQ3: What level of over-segmentation of primitives is helpful in improving the recognition rates for noisy raster images, and can synthetic document noise improve the robustness of the parser for such images?

Approach: Analyze the effect of over-segmentation and addition of noise on recognition of real-life raster images



Proposed Work: Global Nodes

RQ4: Can incorporating self-attention with iterative classification refinements improve multi-task learning (MTL) performance compared to the baseline models (LGAP, LCGP)

Approach: Utilize initial outputs from segmentation, symbol classification, and relationship detection tasks as inputs to subsequent network iterations

Conclusion