

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

Defense Chair: Dr. Dan Phillips

PhD Dissertation Defense

Golisano College of Computing and Information Sciences

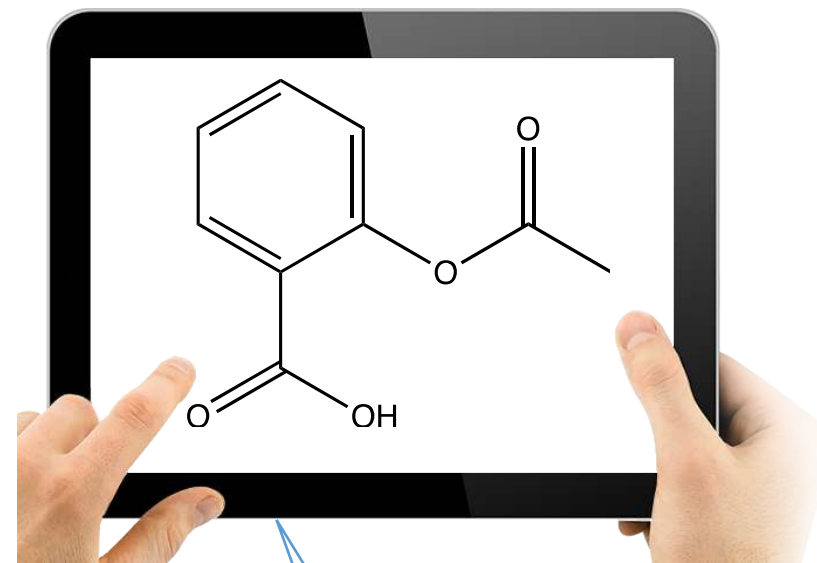
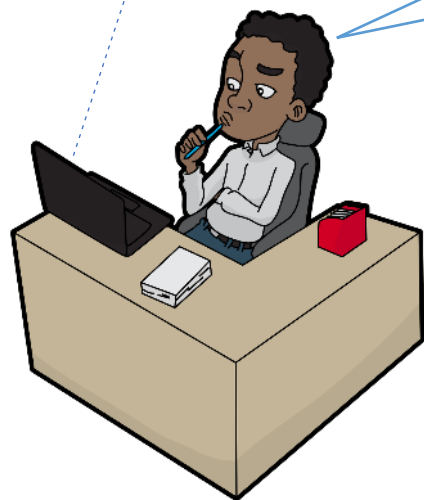
June 27, 2025

Introduction

Motivation

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

What is the name
of this formula?

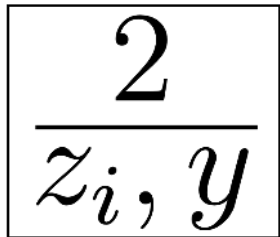


Let's find reactions
producing this molecule.

Related Work

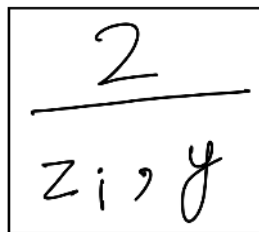
Parsing Math Formulas

Vector (e.g. PDF)



A vector representation of the fraction $\frac{2}{z_i, y}$ in a serif font, enclosed in a black rectangular box.

Raster (e.g. PNG)



A raster representation of the fraction $\frac{2}{z_i, y}$ in a handwritten-style font, enclosed in a black rectangular box.

Handwritten



`"\frac{2}{z_i, y}"`

LATEX String

Symbol Layout Tree (SLT)

Parsing Chemical Diagrams

Vector (e.g. PDF)

Raster (e.g. PNG)

Handwritten



`"C1=CC=C(C=C1)[N+](=O)[O-]"`

SMILES String

Molecular Graph

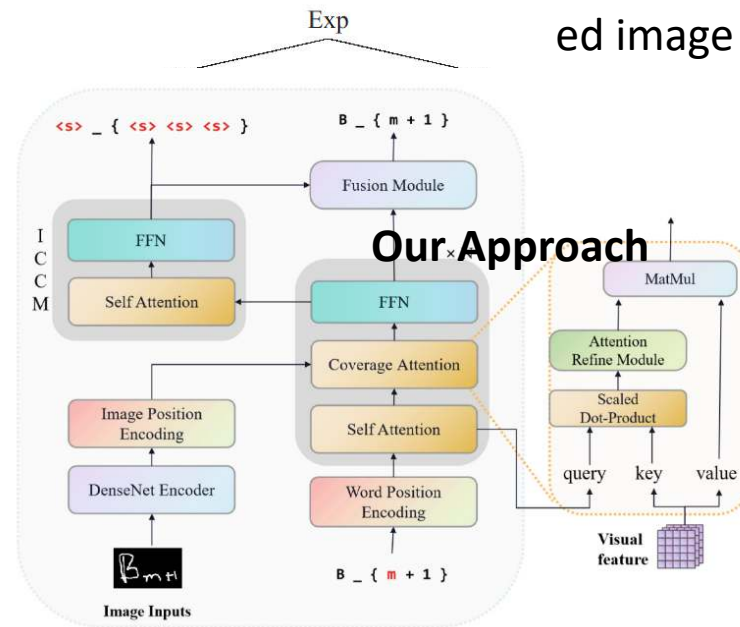
Related Work

1. Syntactic Models:

1. Math: generate **parse tree**, linearize it to LATEX **string** using **context-free grammars** (CFGs, SCFGs)
2. Chem: generate SMILES **string** using image processing and chemical/geometrical **constraints (rules)**

2. Encoder-Decoder Models: 2D image pixels or strokes **encoder** and 1D LATEX/SMILES **string decoder** (with attention)

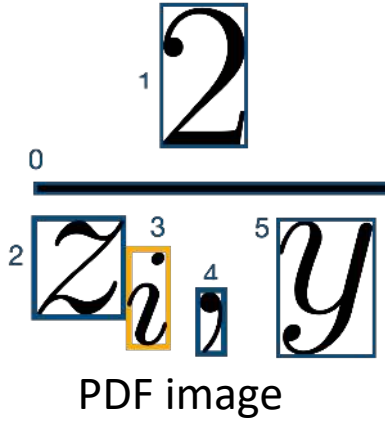
3. Graph-based Models: graph input to **graph**: SLT or molecular graph



- [1] Truong et al., "A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and GNN models," *Pattern Recognition*, vol. 153, p. 110531, Sep. 2024
- [2] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in Symposium on Interactive Systems for Experimental Applied Mathematics, Aug. 1967
- [3] Álvaro et al., "An integrated grammar-based approach for mathematical expression recognition," *Pattern Recognition*, Mar. 2016
- [4] Zhang et al., "Track, Attend, and Parse (TAP): An End-to-End Framework for Online Handwritten Mathematical Expression Recognition," *IEEE Transactions on Multimedia*, Jan. 2019
- [5] Zhu et al., "ICAL: Implicit Character-Aided Learning for Enhanced Handwritten Mathematical Expression Recognition," in *Document Analysis and Recognition - ICDAR 2024*
- [6] Xie et al., "Stroke-Level Graph Labeling with Edge-Weighted Graph Attention Network for Handwritten Mathematical Expression Recognition," in *Document Analysis and Recognition - ICDAR 2024*

Inputs/Outputs and Evaluation

Our Inputs and Outputs



Vector source

Raster or handwritten source

Inputs

Outputs

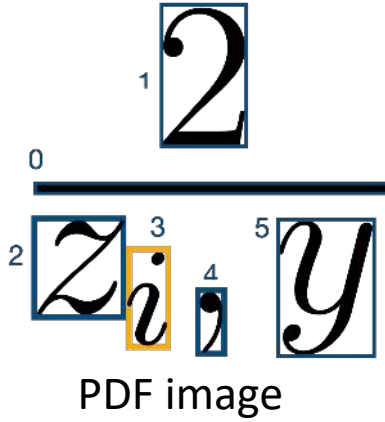
Primitive level graph (PDF)

PNG image with CCs

Handwritten strokes

Primitive level graph (CCs/strokes)

Our Inputs and Outputs



Vector source

Raster or handwritten source

Inputs

Outputs

Symbol layout tree (PDF symbols)

PNG image with CCs

Handwritten strokes

Symbol layout tree (Raster symbols)

Graph-based Evaluation Metrics

$$\text{Detection Precision} = \frac{|\text{Correct detections}|}{|\text{Predicted detections}|}$$

$$\text{Detection Recall} = \frac{|\text{Correct detections}|}{|\text{Target detections}|}$$

$$1. \text{ Symbol Detection F1} = HM(P, R) = HM\left(\frac{5}{7}, \frac{5}{6}\right)$$

$$2. \text{ Relation Detection F1} = HM(P, R) = HM\left(\frac{4}{6}, \frac{4}{5}\right)$$

$$3. \text{ Structure rate} = \frac{|\text{Formulas with correct detections}|}{|\text{Formulas}|}$$

$$\text{Detection + Class Precision} = \frac{|\text{Correct det + classes}|}{|\text{Predicted classes}|}$$

$$\text{Detection + Class Recall} = \frac{|\text{Correct det + classes}|}{|\text{Target classes}|}$$

$$4. \text{ Symbol Detection + Class F1} = HM(P, R) = HM\left(\frac{4}{7}, \frac{4}{6}\right)$$

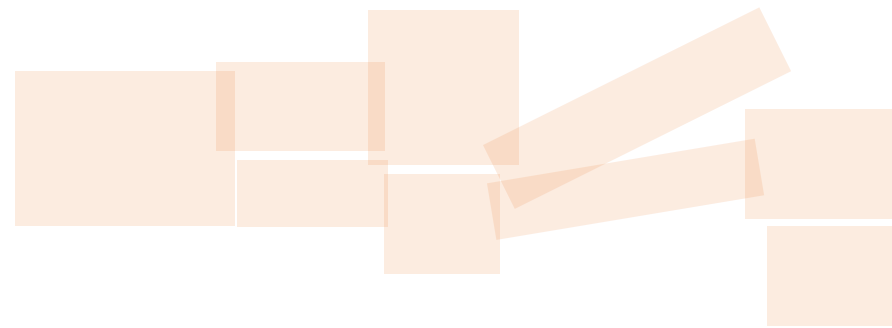
$$5. \text{ Relation Detection + Class F1} = HM(P, R) = HM\left(\frac{3}{6}, \frac{3}{5}\right)$$

$$6. \text{ Expression rate} = \frac{|\text{Correct formulas}|}{|\text{Formulas}|}$$



Ground Truth

Predicted



Research Questions

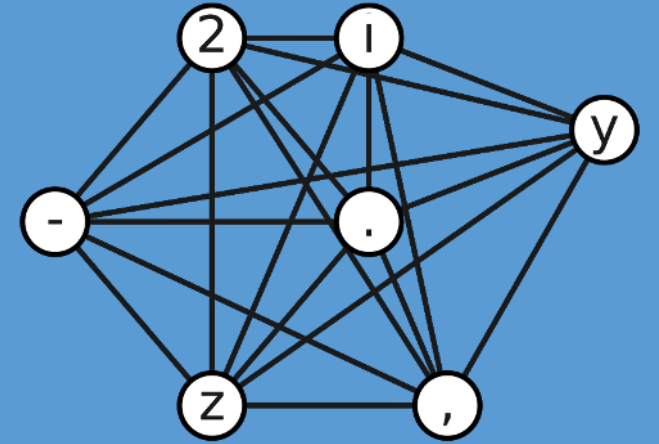
Research Questions

RQ1 How can **primitive features** be constructed and updated efficiently to support **iterative refinement of segmentation** and improve structural parsing?

RQ2 What types of **input graph representations** are most effective for parsing mathematical formulas and chemical diagrams, and do these representations differ between the two domains?

RQ3 How can **graph context and class distributions** be leveraged to improve contextual representation and **task interaction** in multi-task parsing?

RQ4 Can **visual and structural noise** at the primitive level, together with **class balancing strategies**, improve the robustness of the visual parser?

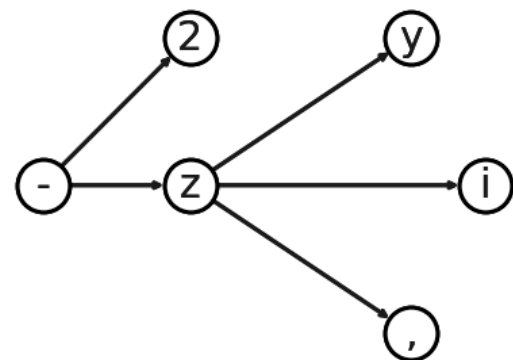


RQ2: Input Graph Representation

RQ2: Input Graph Representation

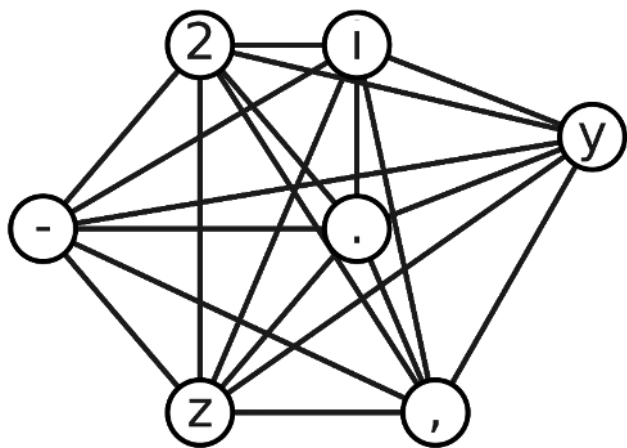
1. Complete Graphs
2. Line-of-Sight (LOS) Graphs
3. K-nearest neighbor (KNN) Graphs:

1. 2NN Graph
2. 6NN Graph

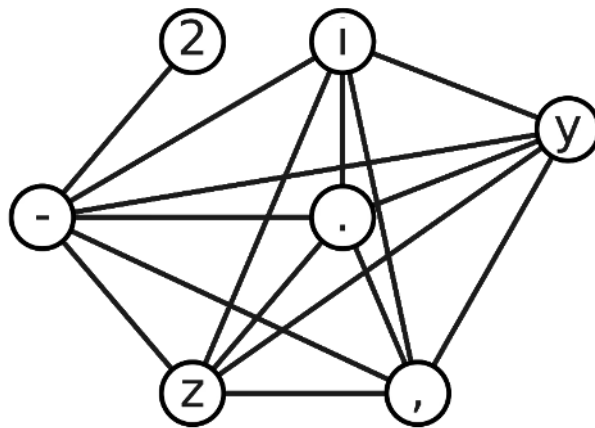


Formula image

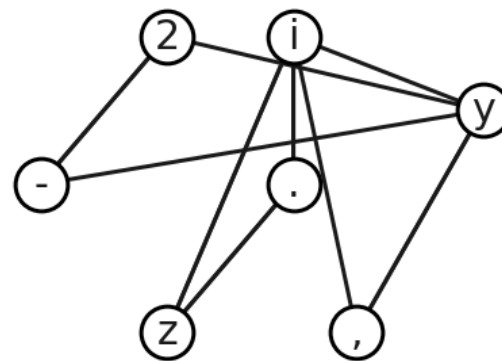
Ground truth Graph
(Symbol level)



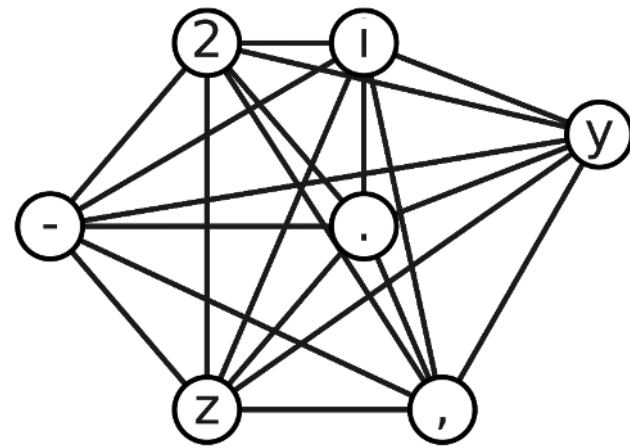
Complete Graph



LOS Graph



2NN Graph



6NN Graph

RQ2: Input Graph Representation

Maths: Infty-MCDB2 Train dataset (12551 formulas)

Expression covered = % of formulas with 100% edge recall

LOS graph is the best choice

RQ2: Input Graph Representation

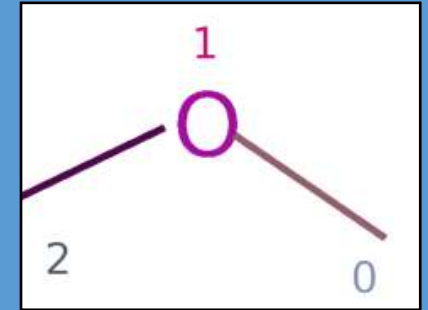
Chemistry: Pubchem-5k Train dataset (4590 formulas)



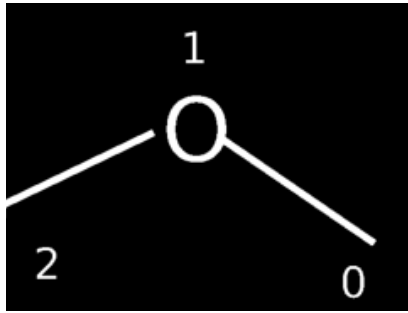
Expression covered = % of formulas with 100% edge recall

6NN graph is the best choice

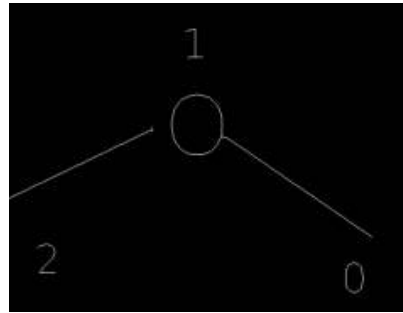
RQ1: Visual primitives and iterative refinement



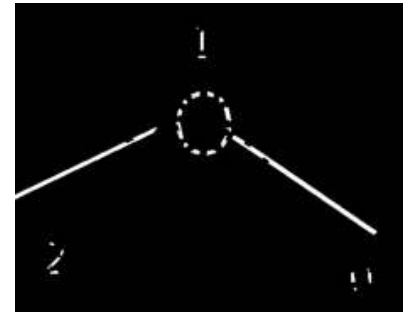
RQ1: Visual features



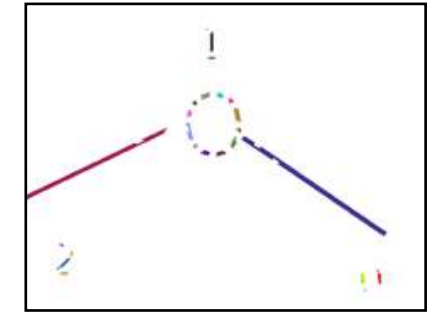
(a) Input image
5 primitives



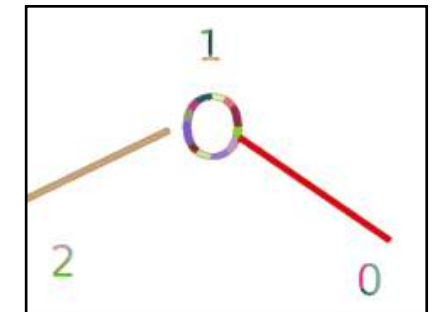
(b) Skeletonized
5 primitives



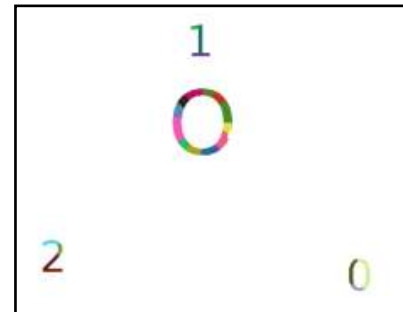
(c) LSD applied
30 primitives



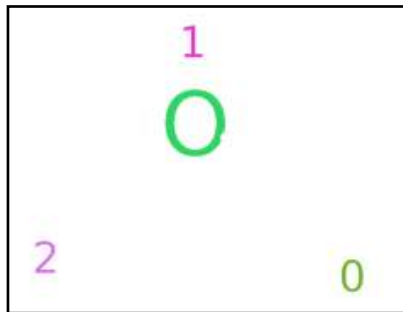
(d) CC Markers
30 primitives



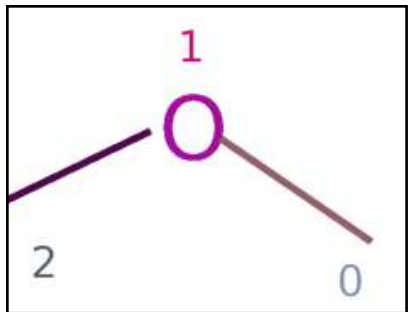
(e) Watershed applied
30 primitives



(f) Long lines removed
28 primitives



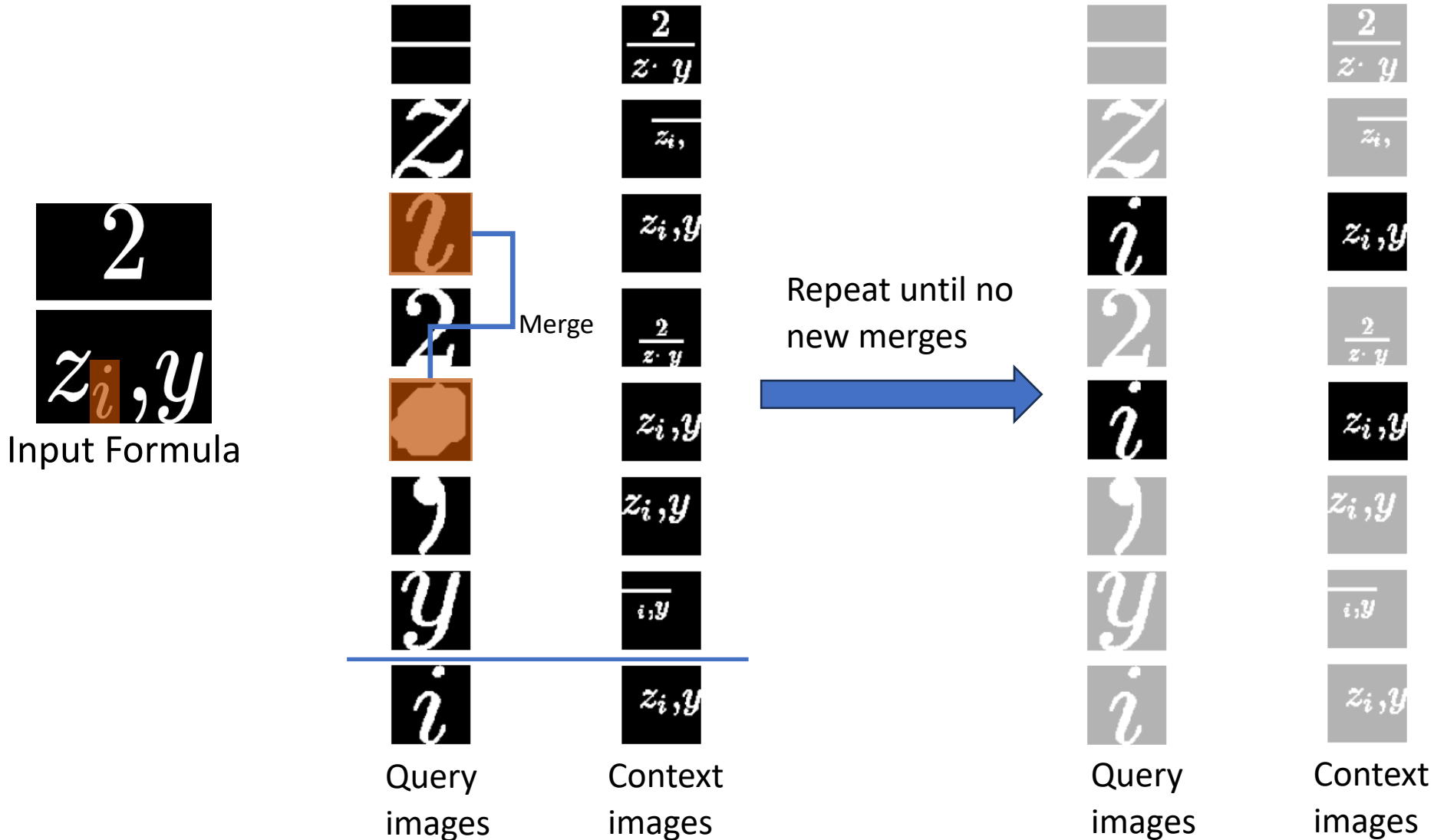
(g) CCs merged
4 primitives



(h) Output primitives
6 primitives

Line Segment Detection (LSD) based visual primitive extraction pipeline

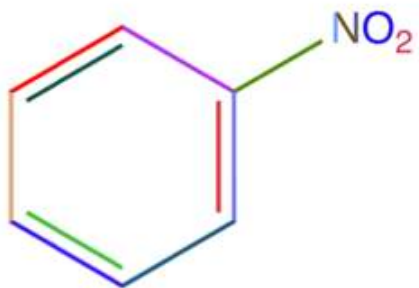
RQ1: Visual features



Iterative refinement of primitives

RQ3: Graph context and task interaction

RQ3: Graph context and task interaction



Visual primitives



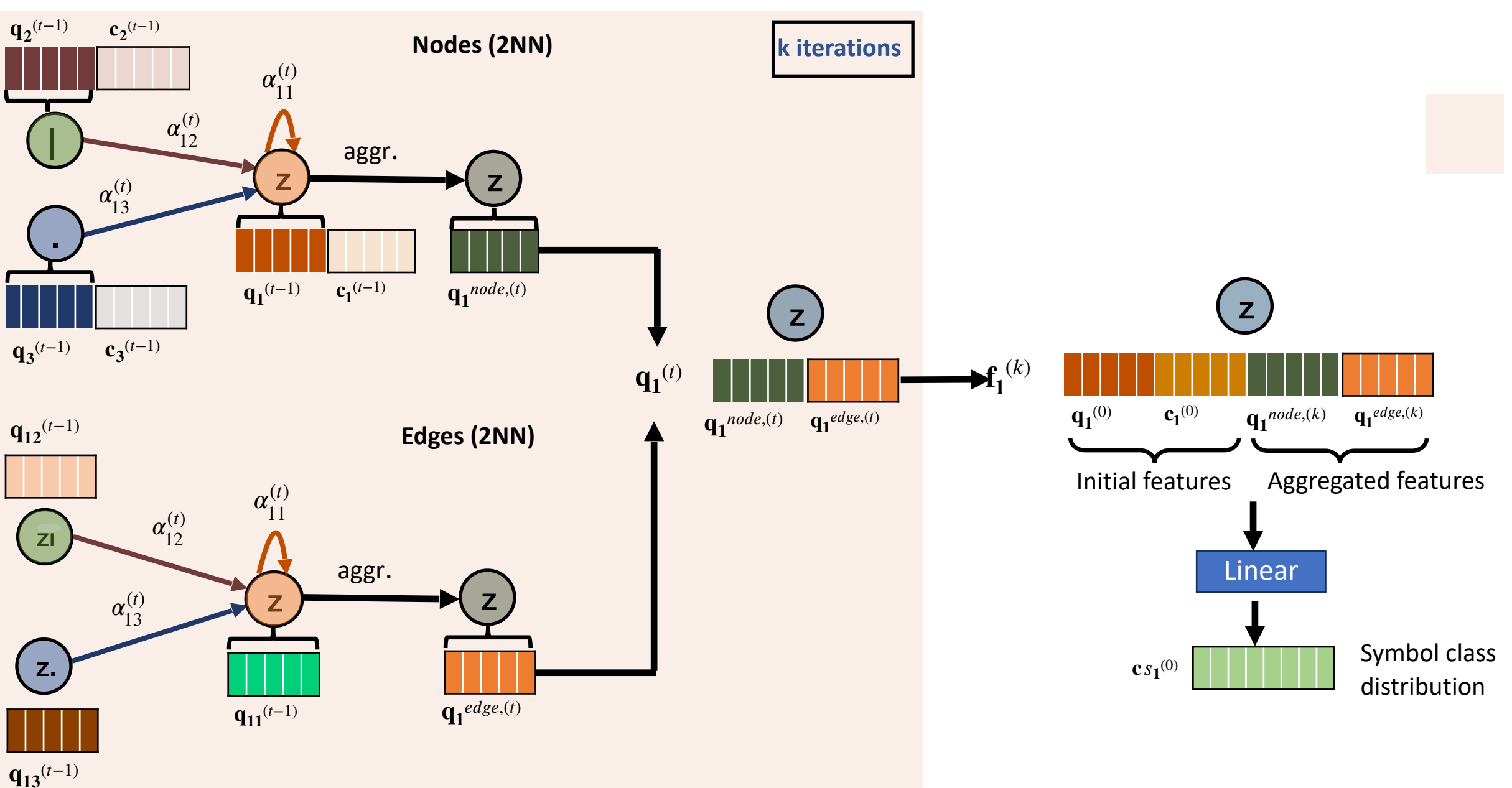
6NN Graph



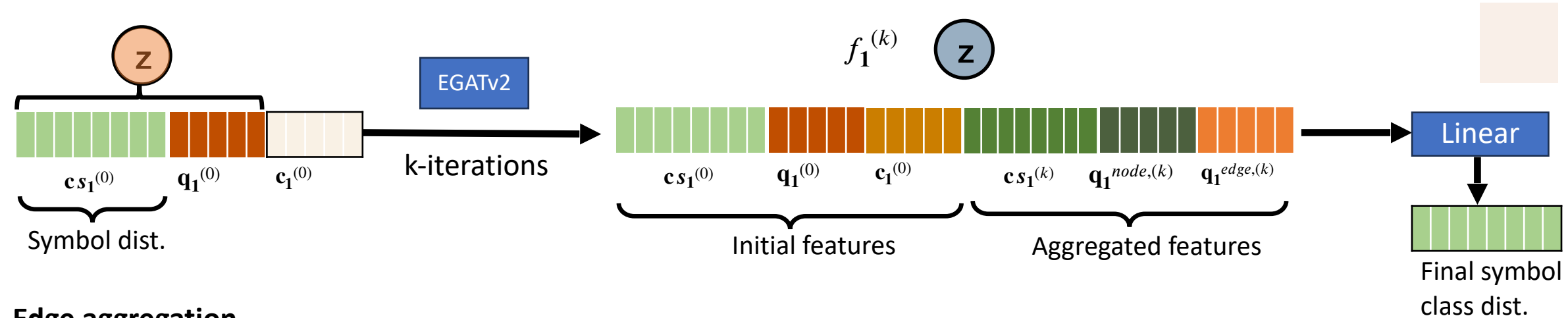
RQ3: Graph context and task interaction



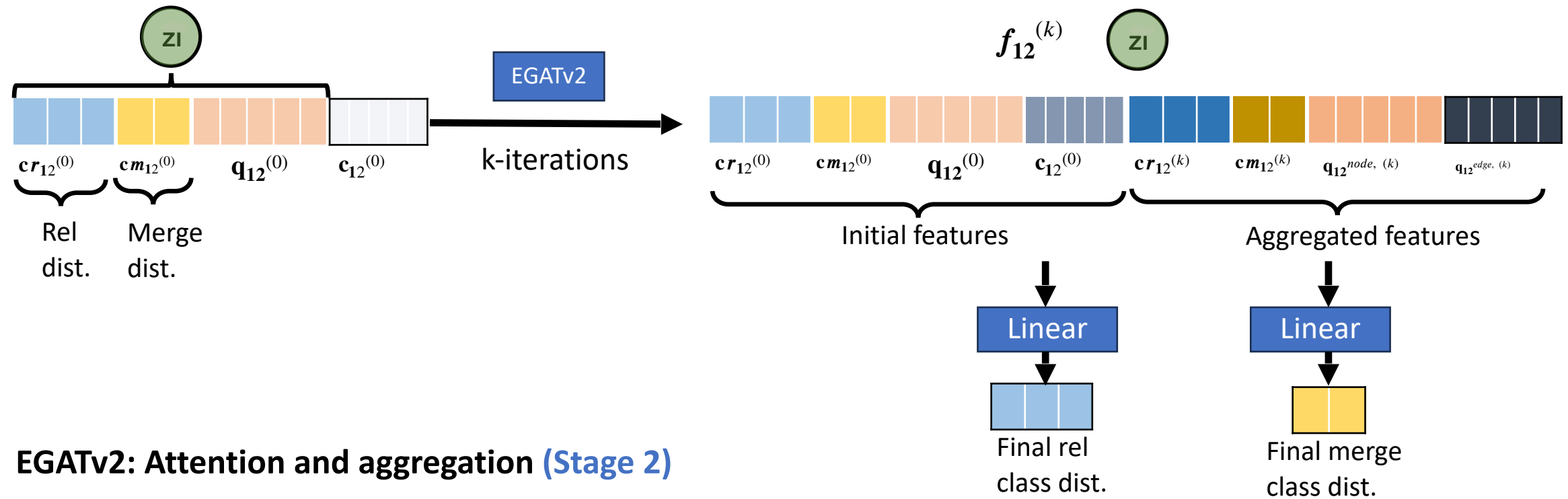
EGATv2: Edge attention weights computation



Node aggregation



Edge aggregation



EGATv2: Attention and aggregation (Stage 2)

RQ3: Graph context and task interaction

Results and error analysis

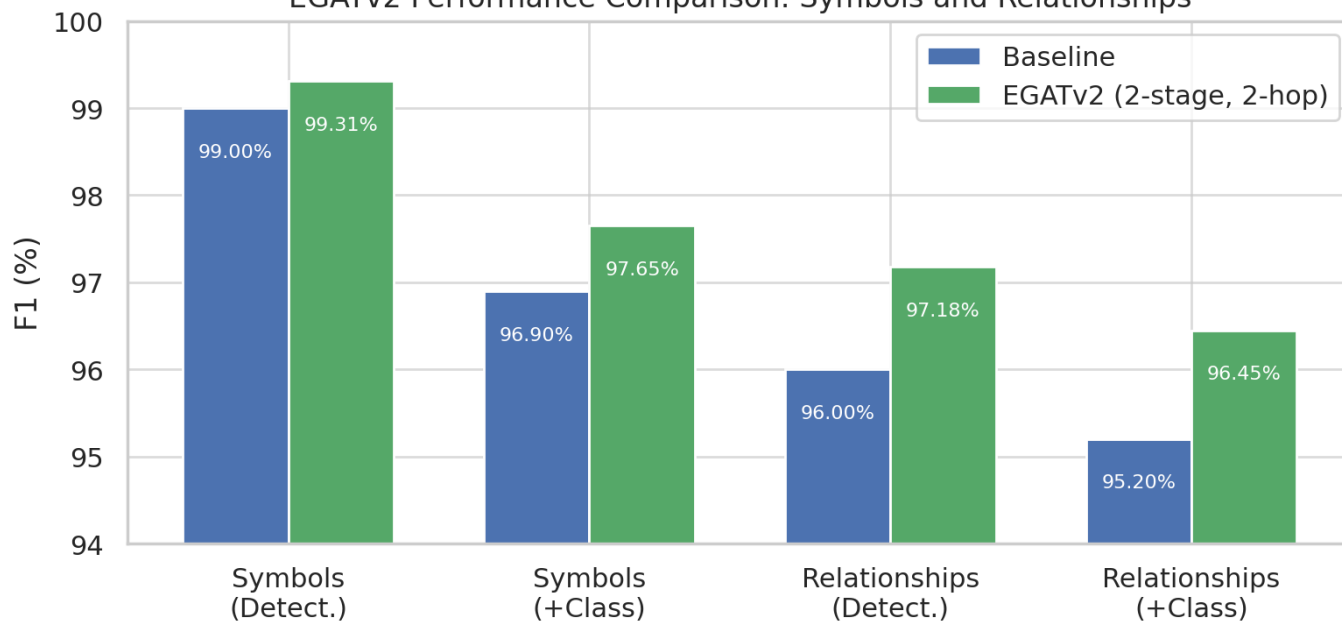
RQ3: Graph context and task interaction

Maths: Infty-MCDB2

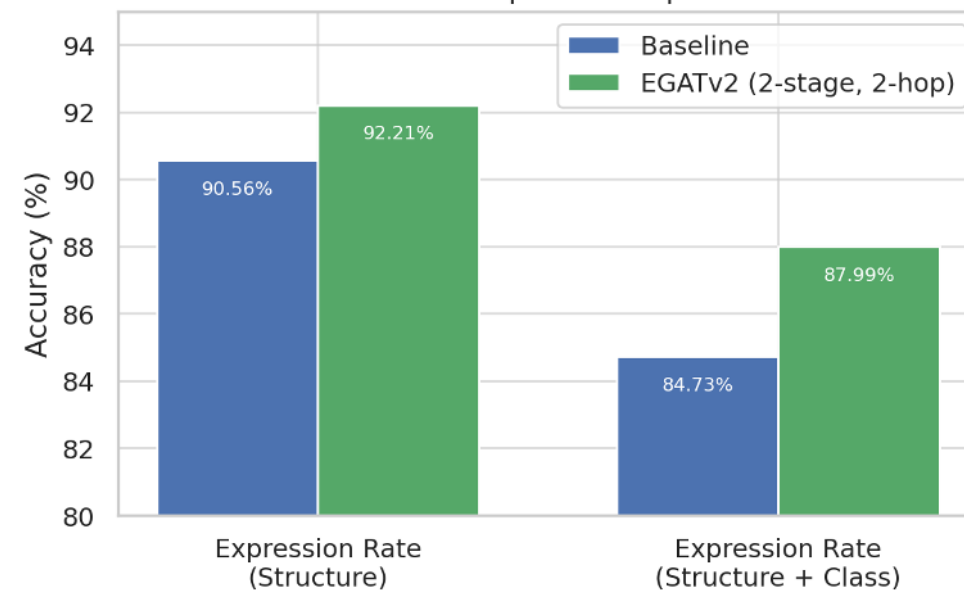
Training: 12,551 formulas

Test: 6,830 formulas

EGATv2 Performance Comparison: Symbols and Relationships



EGATv2 Performance Comparison: Expression-Level Rates



No EGATv2

$$KC \equiv -3d$$

$$k \equiv m_0 > m_1$$

$$\xi^i \quad f_{i_k} \rightarrow f \quad S_{ij}$$

EGATv2, 2-hop 2-stage

$$D + C_1 + C_2$$

$$\text{to } \check{\beta} \cap R_{\text{--}} = \emptyset$$

$$a \equiv 3$$

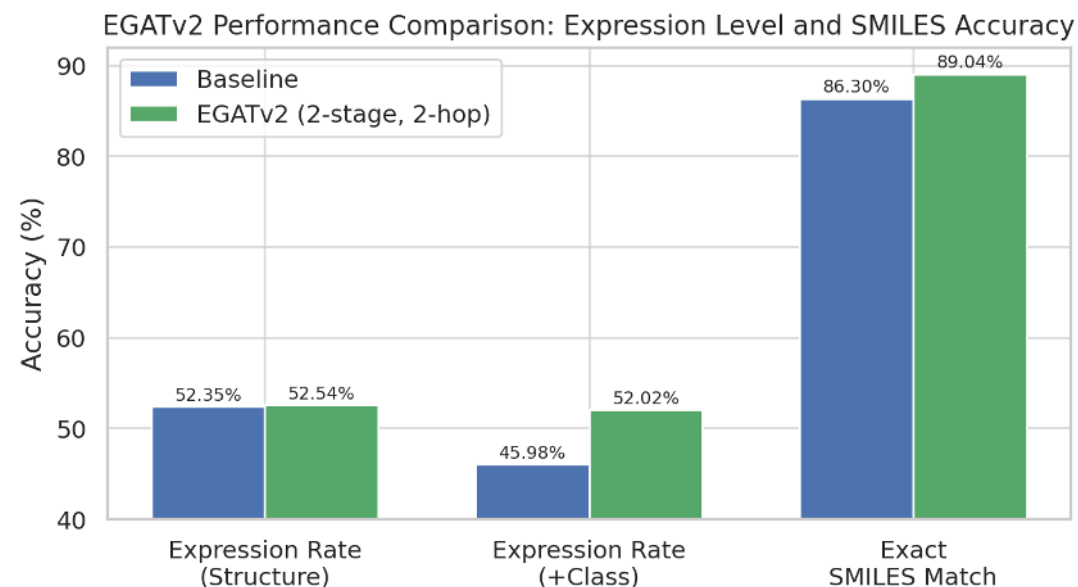
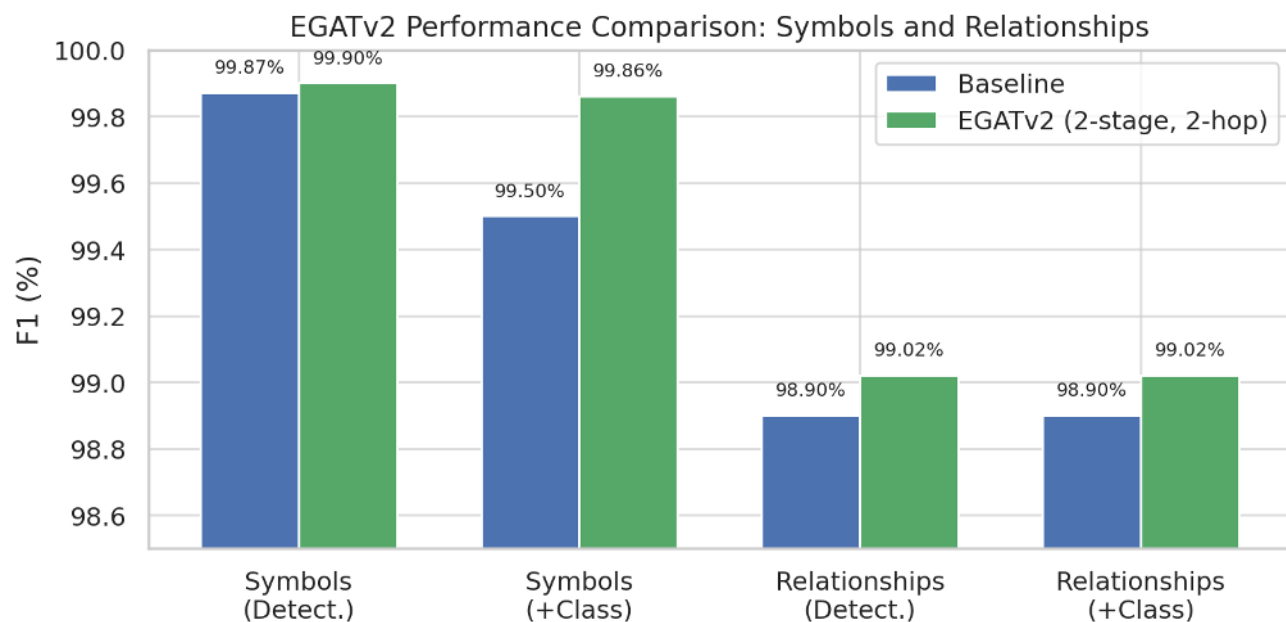
$$x^\beta \equiv v \, x^a \, v^{-1}$$

RQ3: Graph context and task interaction

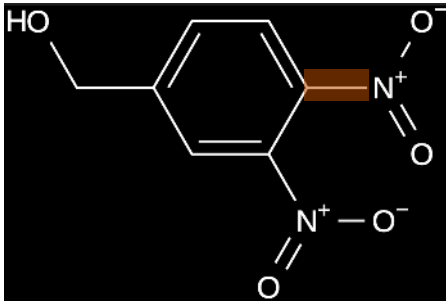
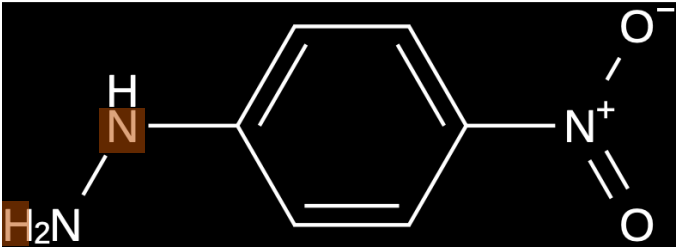
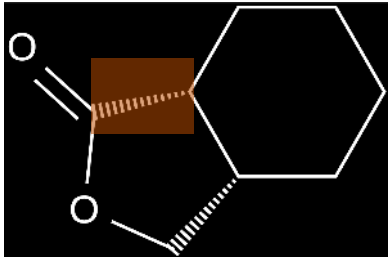
Chemistry: Pubchem-5k

Training: 4,590 molecules

Test: 5,719 molecules (USPTO)

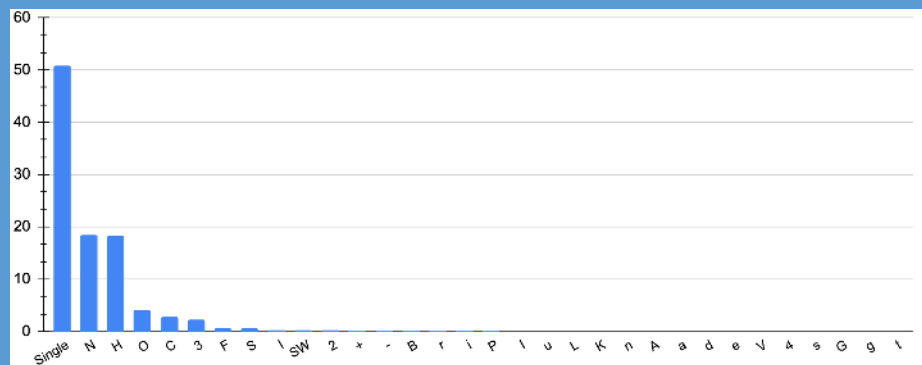
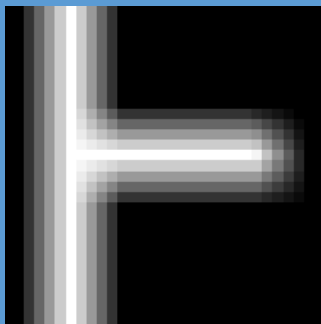


No EGATv2



EGATv2, 2-hop 2-stage

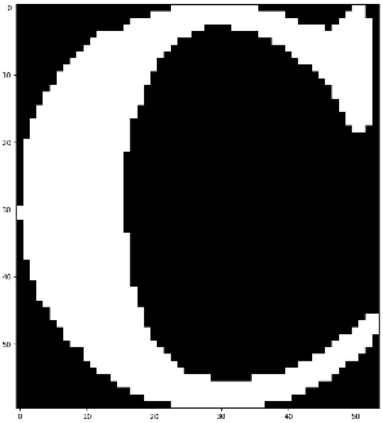




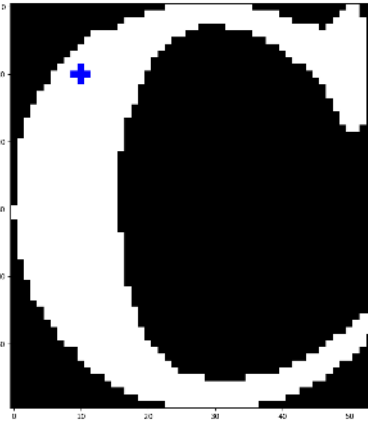
RQ4: Noise and Class Imbalance

RQ4: Noise and Class Imbalance

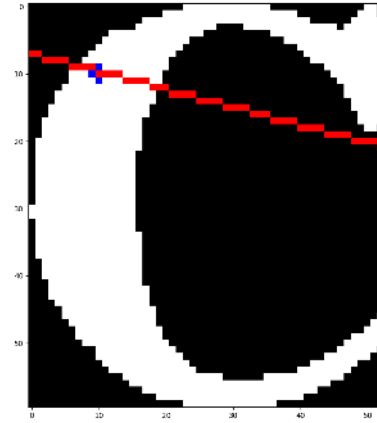
Structure Noise



Original



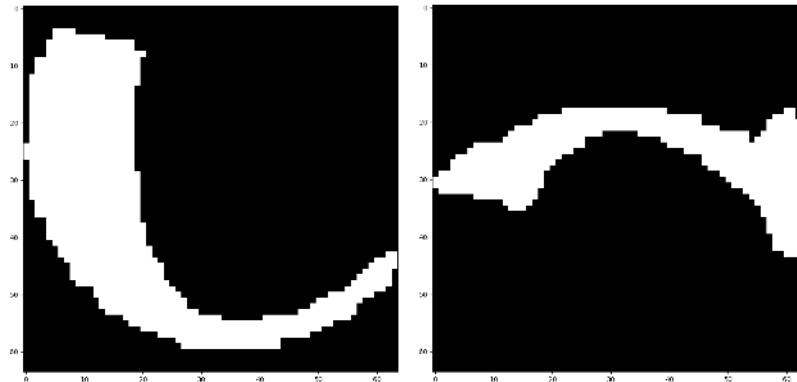
Random point



Random direction



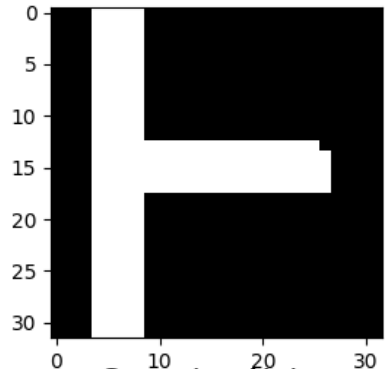
Split into two



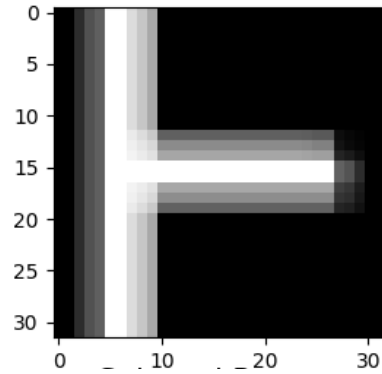
New split primitives

RQ4: Noise and Class Imbalance

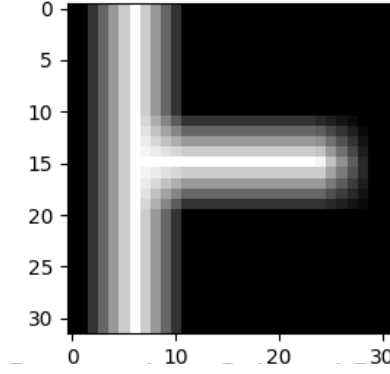
Visual Noise



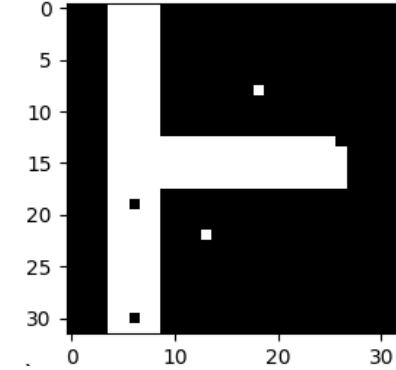
(a) Original



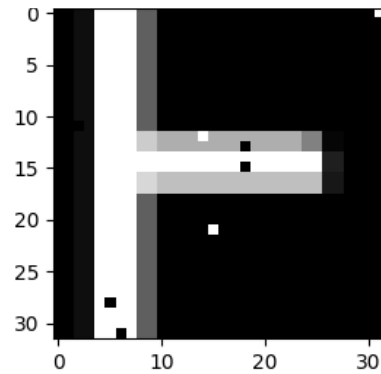
(b) Downscale



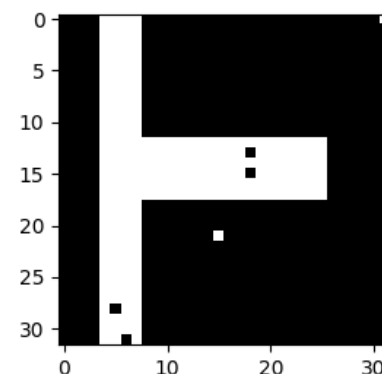
(c) Blur



(d) Salt and Pepper



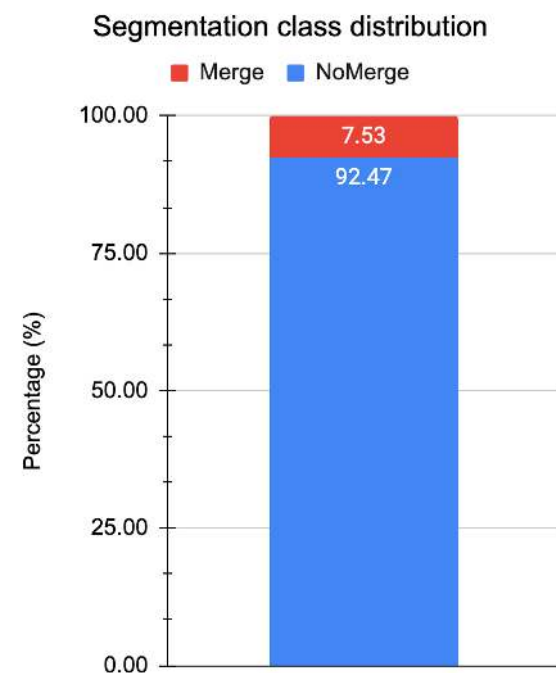
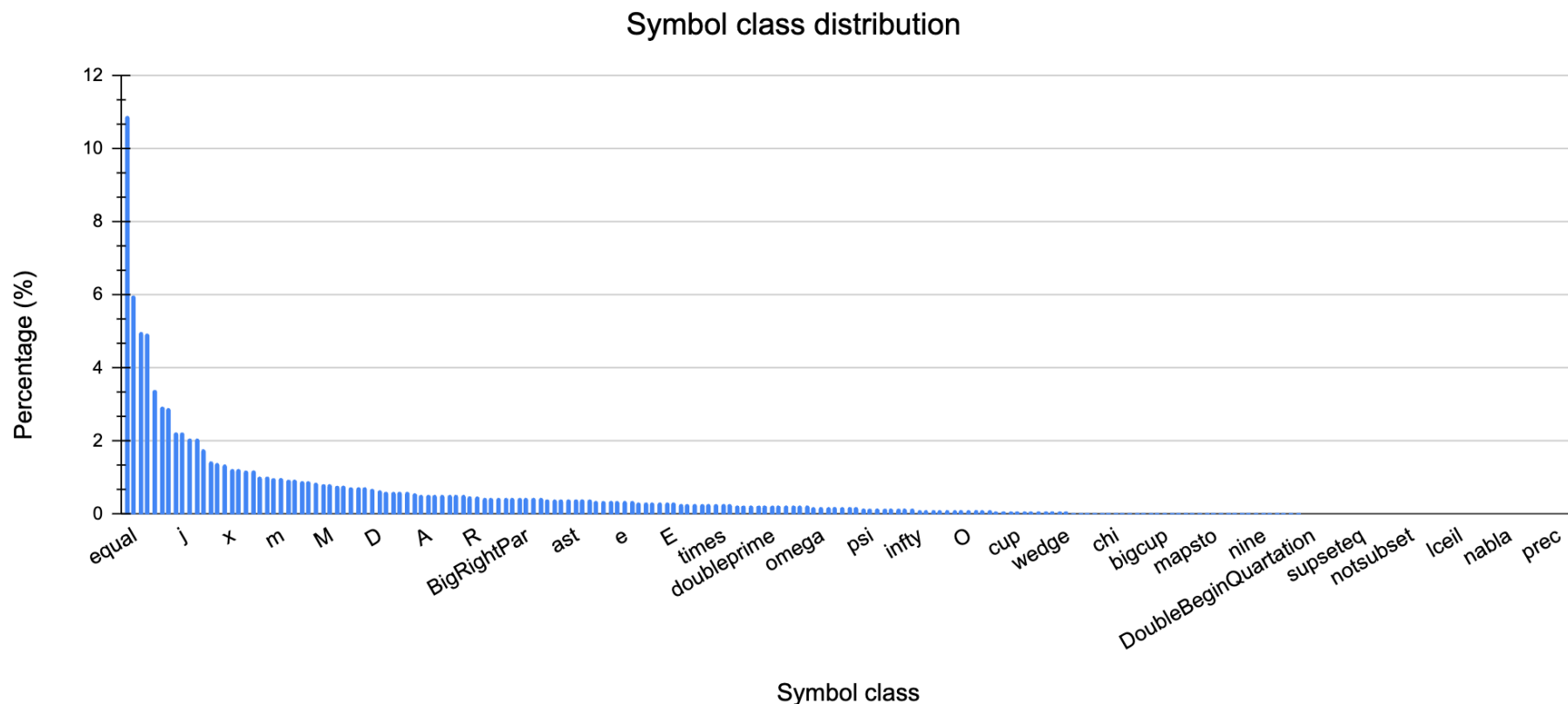
(e) Combination



(f) Binarized (OTSU)

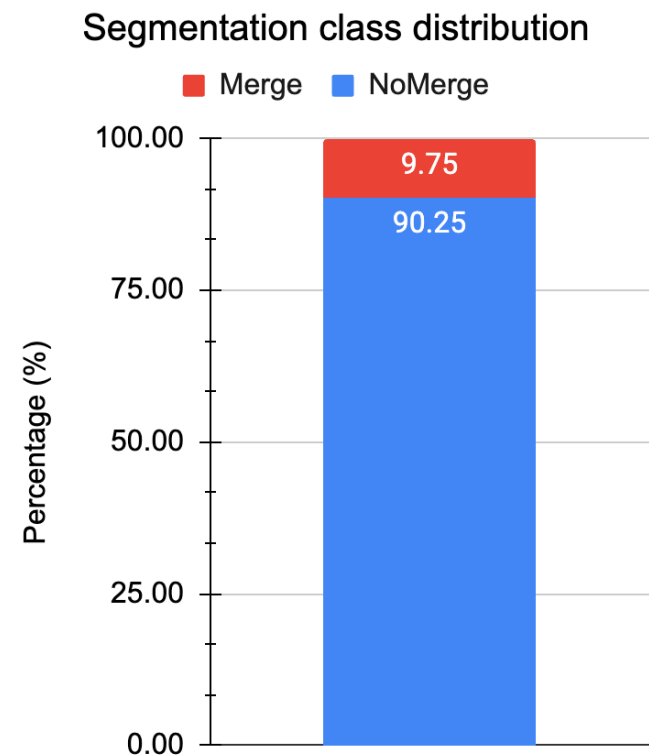
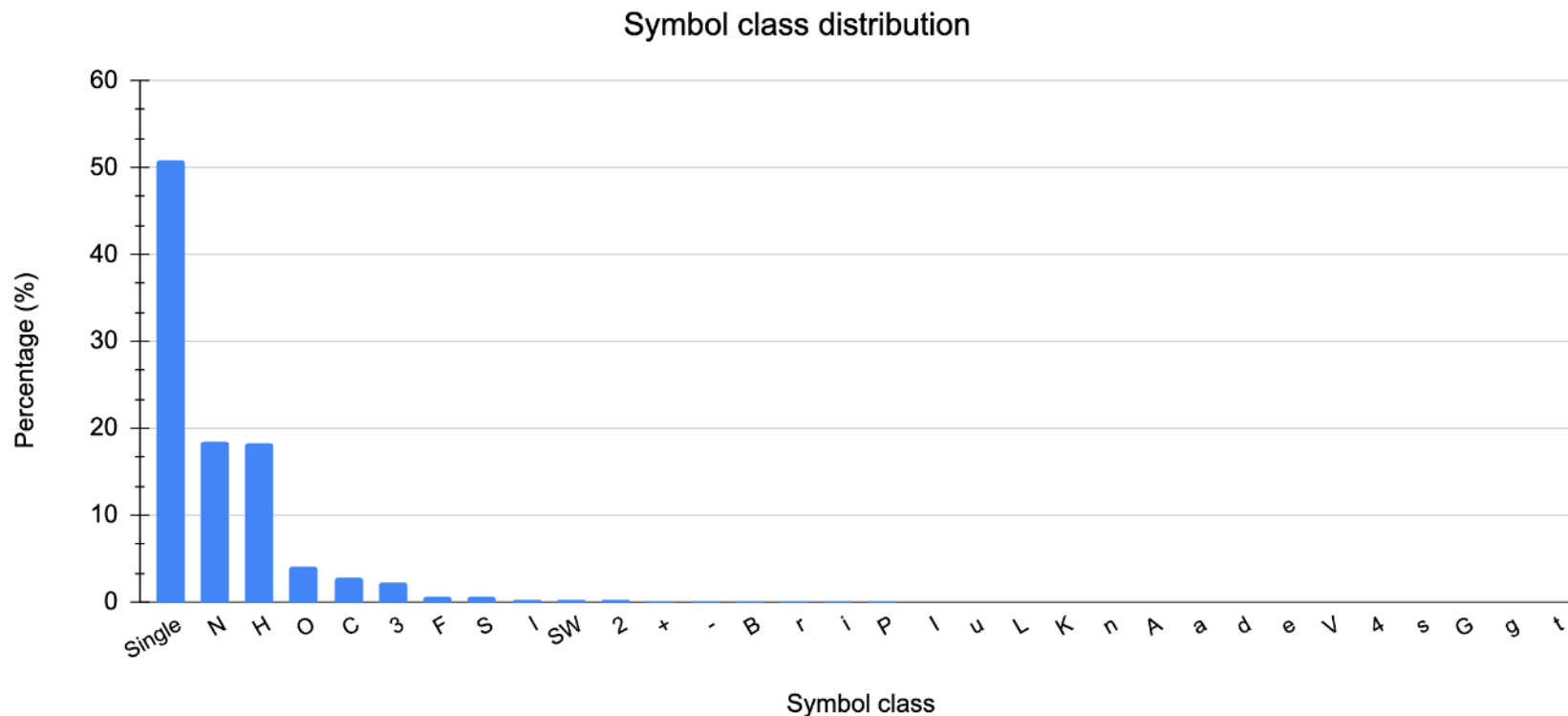
RQ4: Noise and Class Imbalance

Maths: Infty-MCDB2 Train dataset (12551 formulas)



RQ4: Noise and Class Imbalance

Chemistry: Pubchem-5k Train dataset (4590 formulas)



RQ4: Noise and Class Imbalance

Weighted Loss Functions

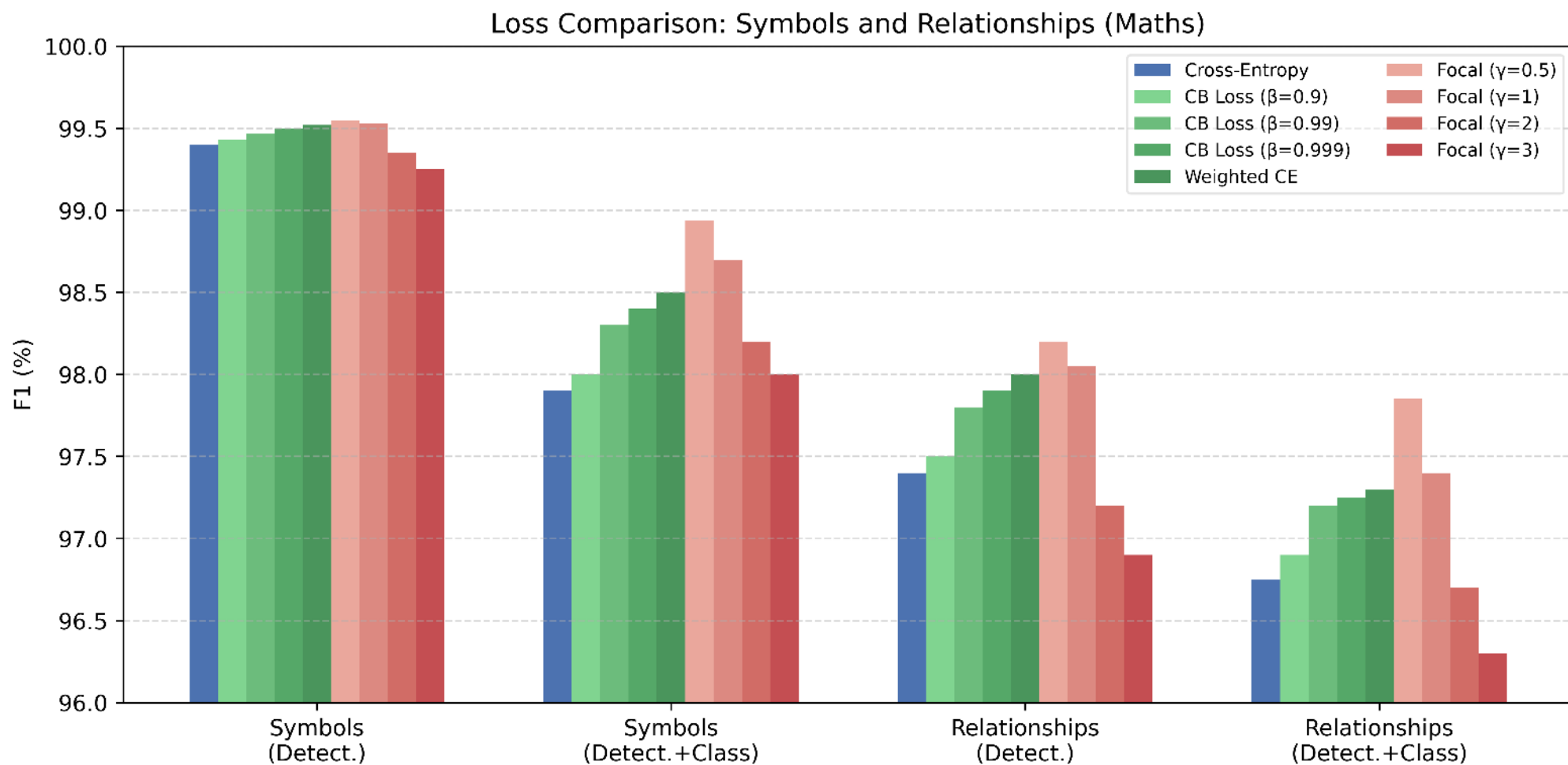
Loss Type	Final Formulation (per query q)	Parameter Range
Cross-Entropy (CE)	$\mathcal{L}^{CE}(q) = - \sum_{i=1}^C y_i \log p_i$	–
Weighted Cross-Entropy (WCE)	$\mathcal{L}^{WCE}(q) = - \sum_{i=1}^C w_i y_i \log p_i$	–
Class-Balanced (CB)	$\mathcal{L}^{CB}(q) = - \sum_{i=1}^C w_i^{CB} y_i \log p_i,$ where $w_i^{CB} = \frac{1-\beta}{1-\beta^{n_i}}$ normalized to mean 1	$\beta \in \{0.9, 0.99, 0.999\}$
Focal Loss (FL)	$\mathcal{L}^{FL}(q) = - \sum_{i=1}^C \frac{\alpha_i (1 - p_i)^\gamma y_i \log(p_i)}{\frac{1}{ \mathcal{B} } \sum_{j \in \mathcal{B}} \alpha_{t_j}},$ with α_i as normalized class weights w_i	$\gamma \in \{0.5, 1, 2, 3\}$

RQ4: Noise and Class Imbalance

Maths: Infty-MCDB2

Training: 12,551 formulas

Test: 6,830 formulas

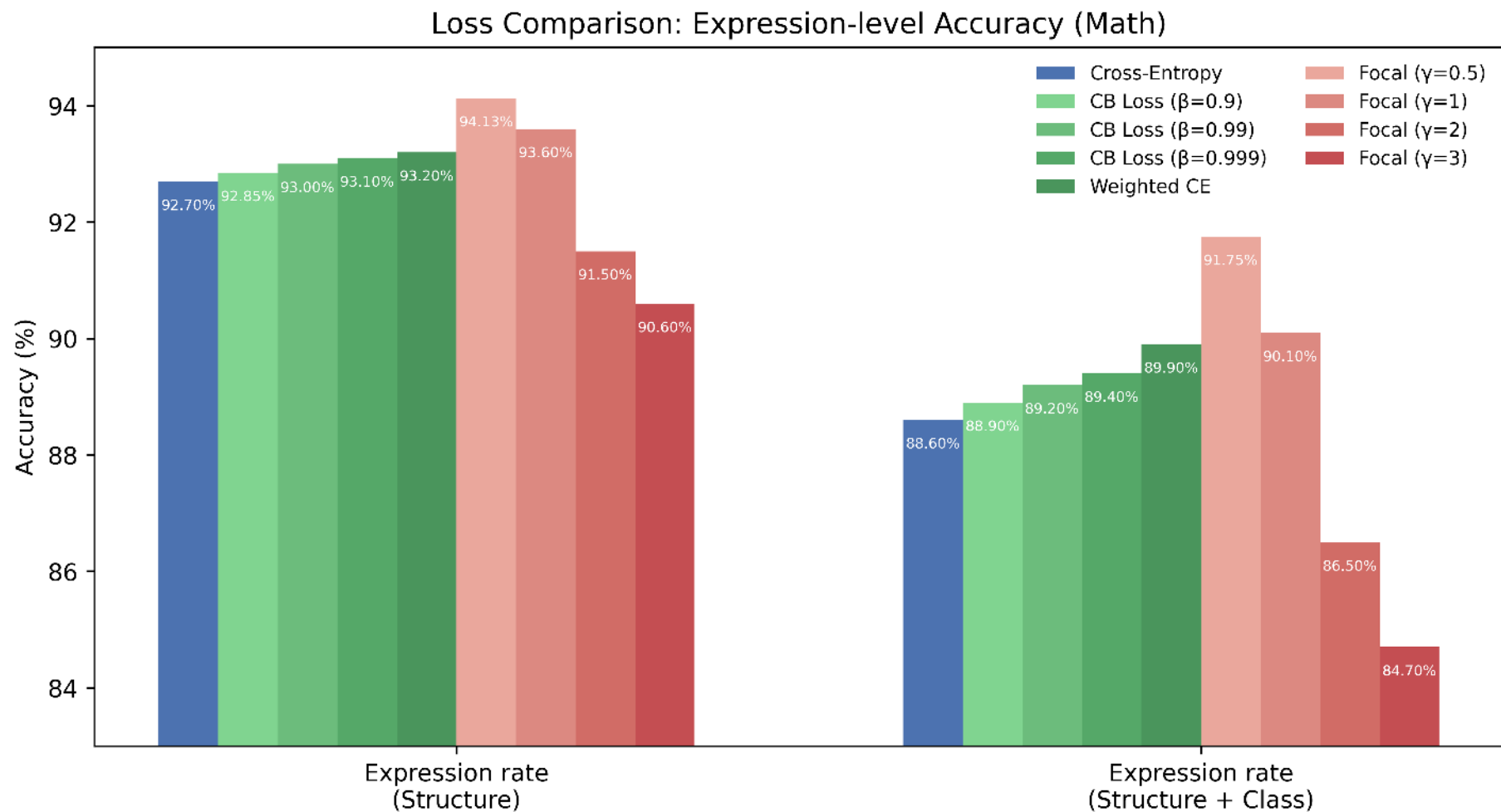


RQ4: Noise and Class Imbalance

Maths: Infty-MCDB2

Training: 12,551 formulas

Test: 6,830 formulas

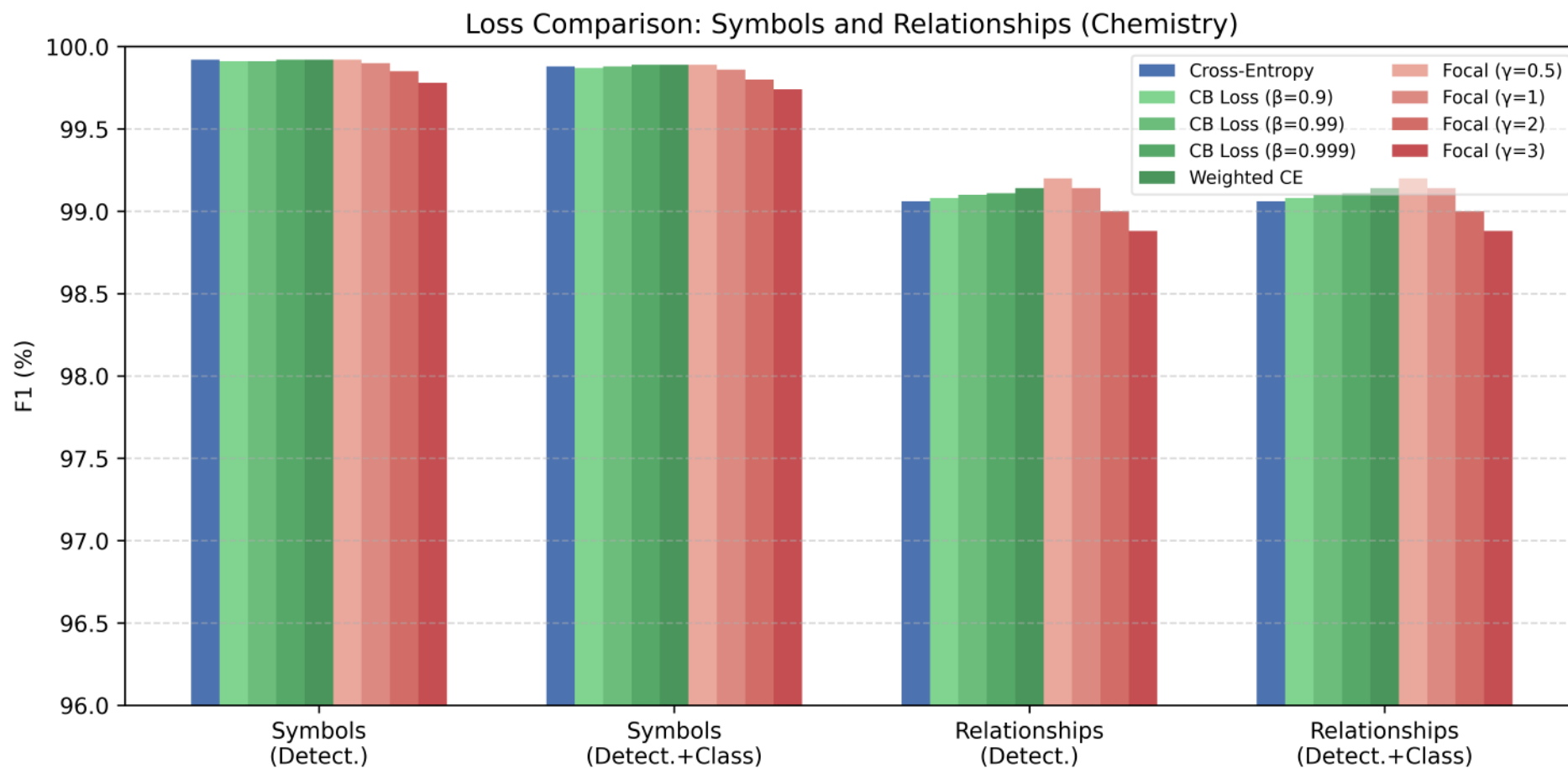


RQ4: Noise and Class Imbalance

Chemistry: Pubchem-5k

Training: 4,590 molecules

Test: 5,719 molecules (USPTO)

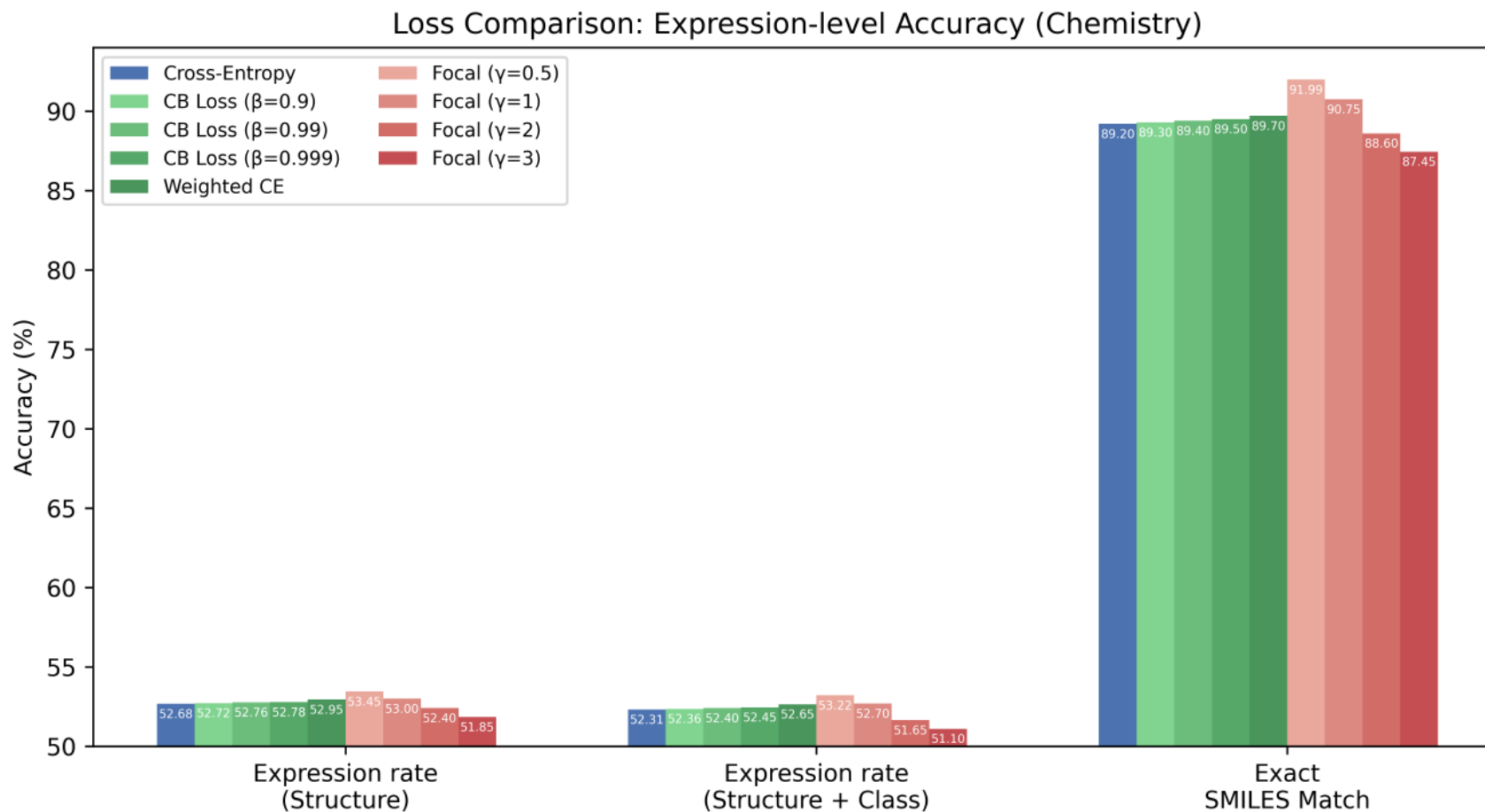


RQ4: Noise and Class Imbalance

Chemistry: Pubchem-5k

Training: 4,590 molecules

Test: 5,719 molecules (USPTO)



RQ4: Noise and Class Imbalance

Complemented Harmonic Mean for Loss Aggregation

1. Sum aggregation:

Mean

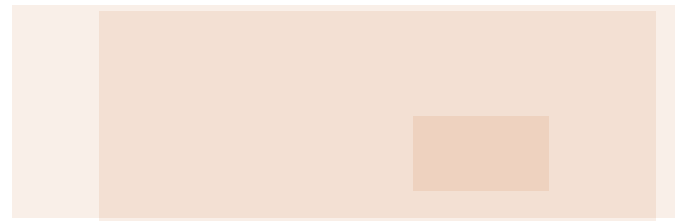
2. Complemented HM aggregation:

HM

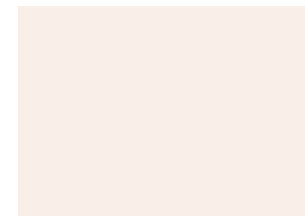
(Normalization)

CHM

(Complement)



(Scale)



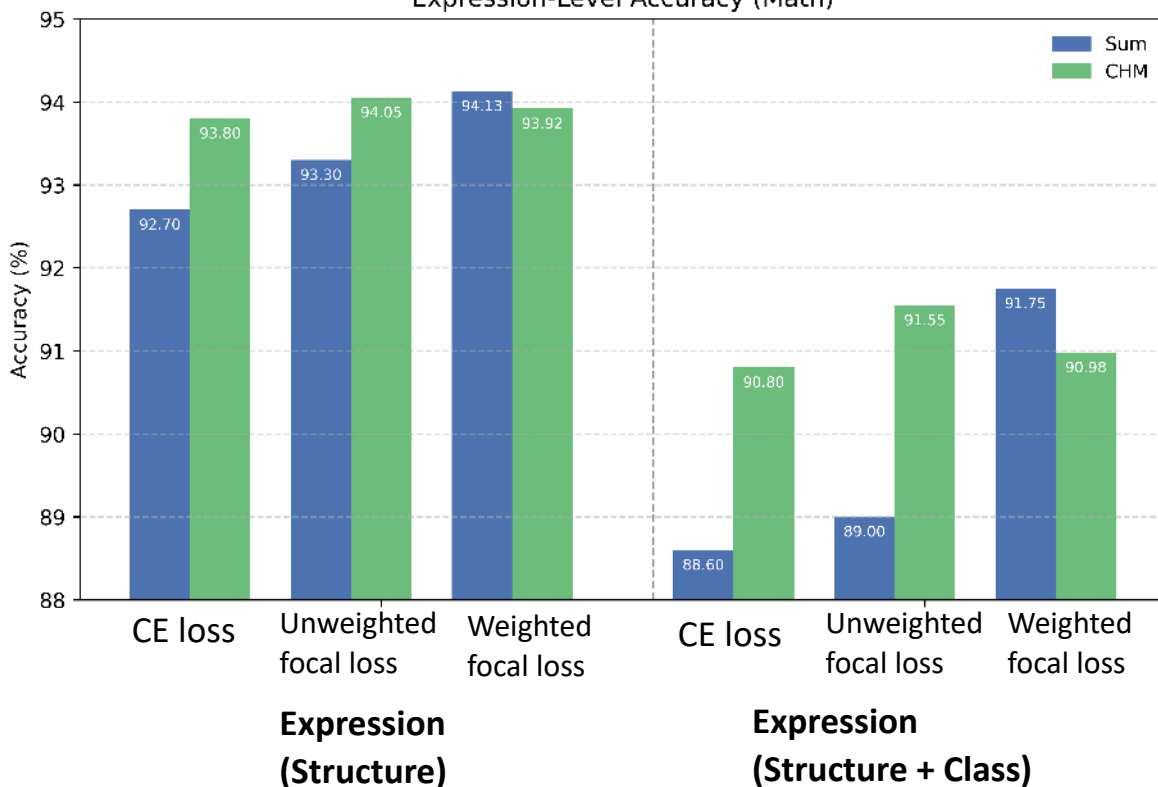
RQ4: Noise and Class Imbalance

Maths: Infty-MCDB2

Training: 12,551 formulas

Test: 6,830 formulas

Aggregation Strategy Comparison:
Expression-Level Accuracy (Math)

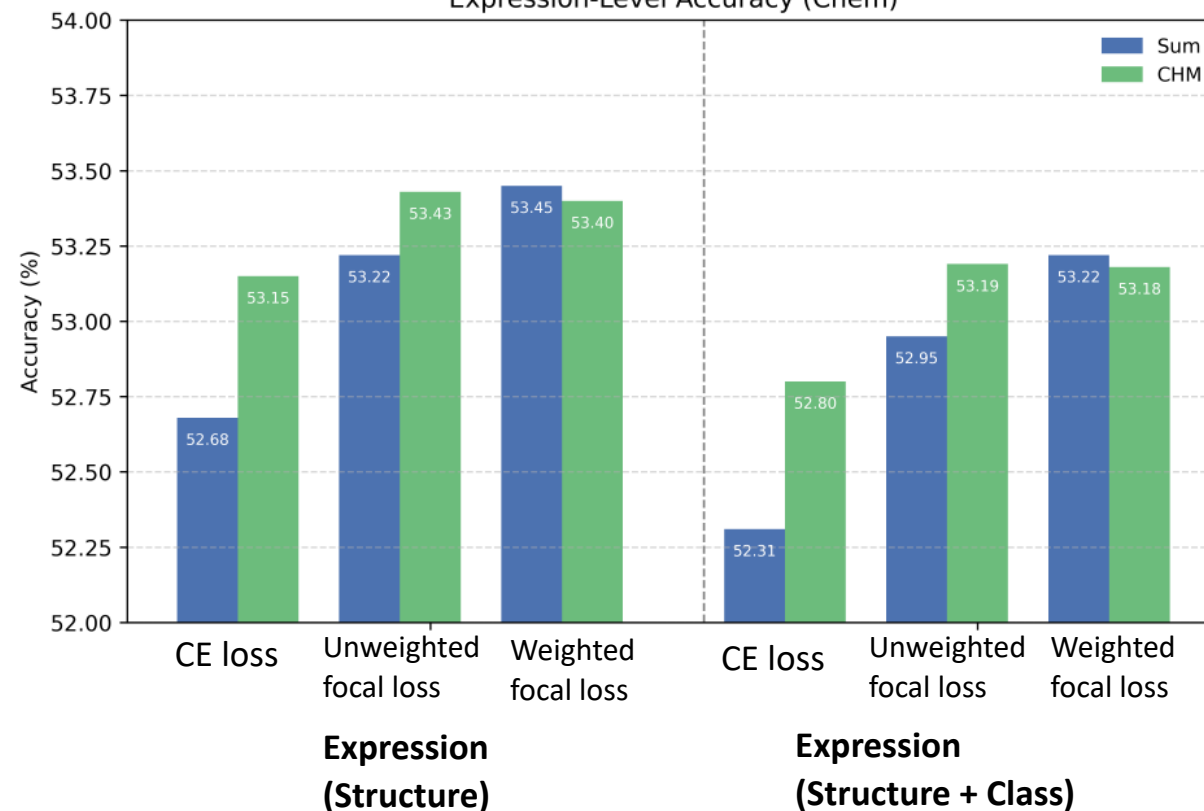


Chemistry: Pubchem-5k

Training: 4,590 molecules

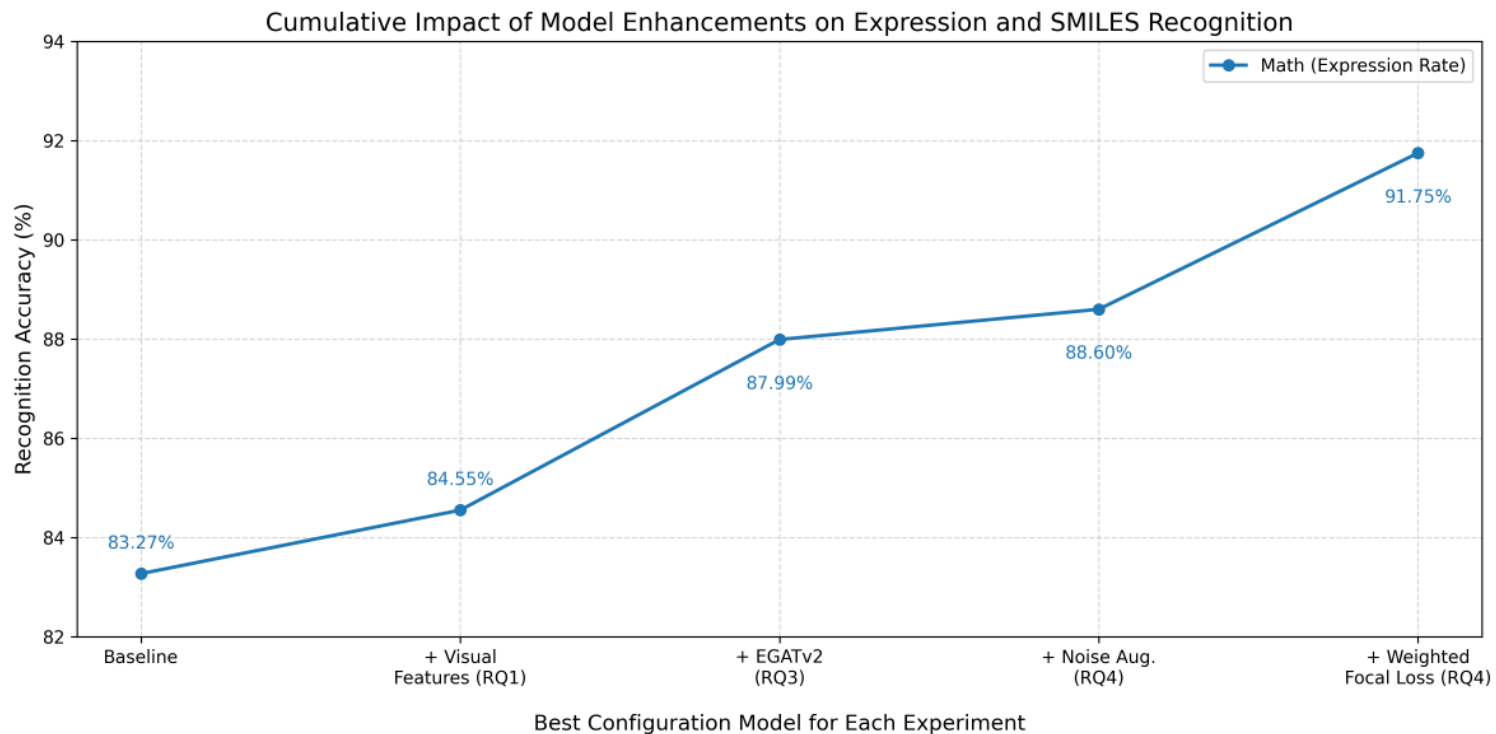
Test: 5,719 molecules (USPTO)

Aggregation Strategy Comparison:
Expression-Level Accuracy (Chem)



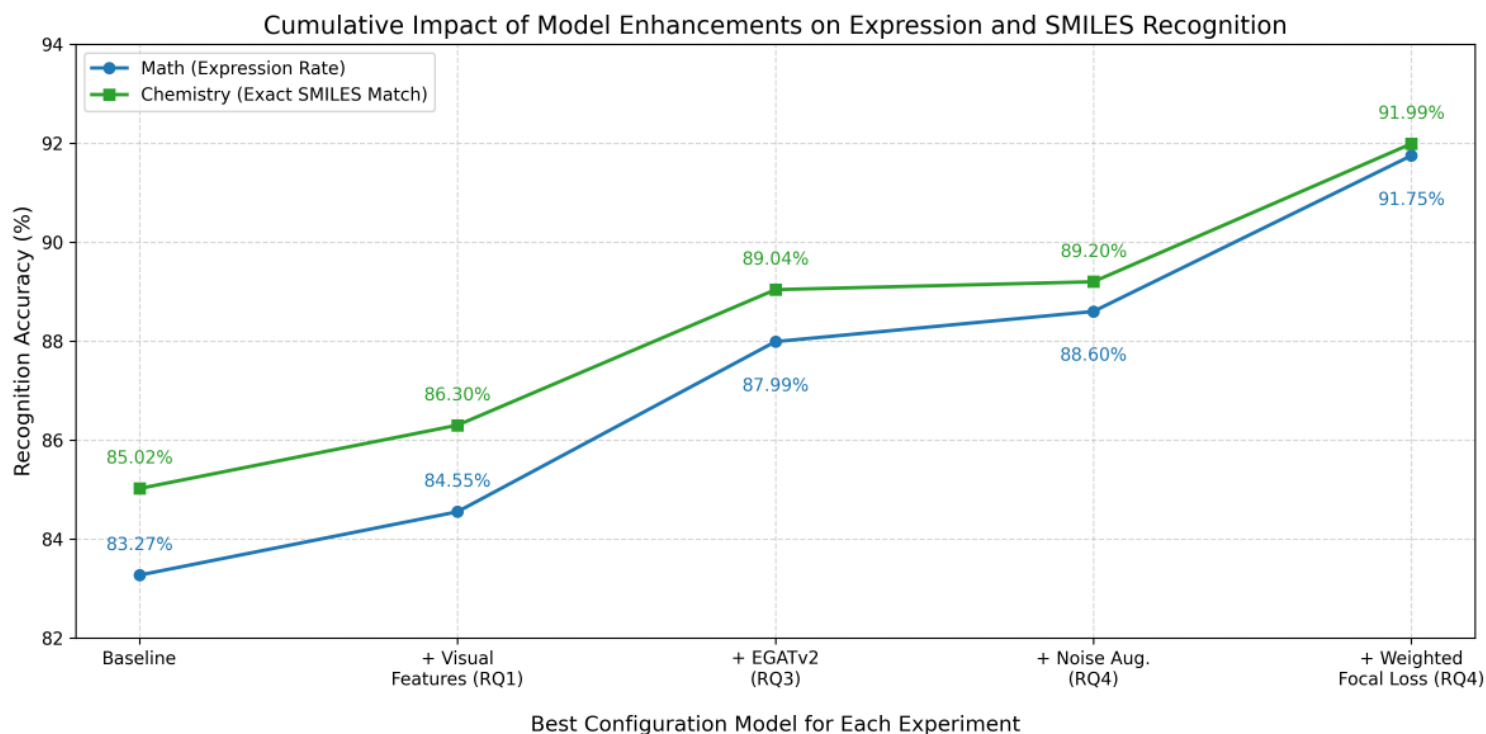
Summary: Maths

Best Configuration Model	RQ	Errors (Math)
Baseline	-	
+ Better Visual features	RQ1	Fewer symbol class confusions on isolated symbols and symbols with similar patterns (e.g., alpha/a, i/j, 1/l)
+ Edge attention (EGATv2)	RQ3	Fewer confusions between similar symbols (e.g., minus vs equal)
+ Structural and visual noise augmentation	RQ4	Reduction in errors on symbol segmentation and classification, especially for noisy test images
+ Weighted Focal loss	RQ4	Lower errors on high frequency classes (minus, equal, NoRelation); fewer missed relationships

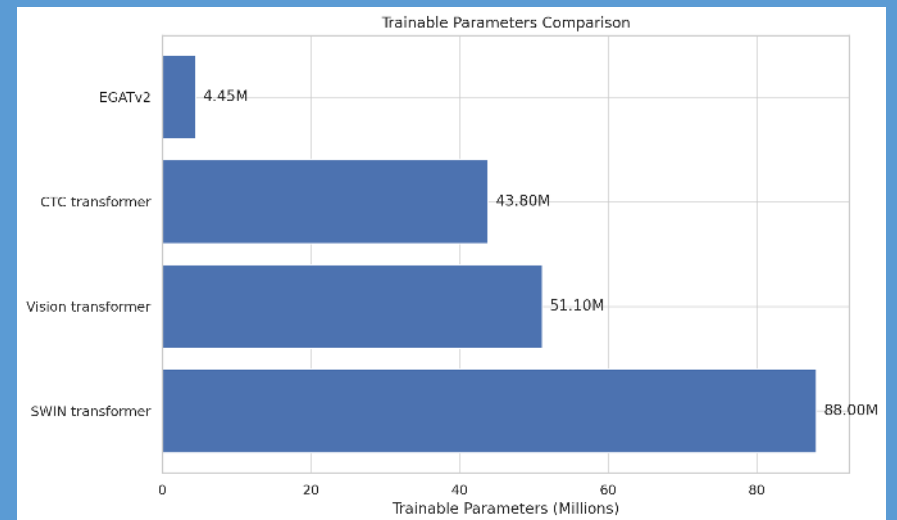


Summary: Chemistry

Best Configuration Model	RQ	Errors (Chemistry)
Baseline	-	
+ Better Visual features	RQ1	Fewer symbol confusions on multi-primitive symbols (e.g., H, N) and symbols with similar patterns (e.g., W/M, i/j, 1/l)
+ Edge attention (EGATv2)	RQ3	Fewer misclassifications of Single lines; improved connection relations
+ Structural and visual noise augmentation	RQ4	Reduction in errors on symbol segmentation and classification, especially for noisy test images
+ Weighted Focal loss	RQ4	Lower errors on high frequency classes (e.g., Single line); fewer missed connections (NoRelation)



Benchmarks



Results: Molecule Diagram Recognition

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

Systems	Synthetic: Indigo (5719)		No. of parameters
	SMILES Match	# Mol / 1% (K)	
MolVec 0.9.7	95.40		
OSRA 2.1	95.00		
Img2Mol	58.90	25.47	30 M
SwinOCSR	74.00	60.81	120 M
MolScribe	97.50	17.23	88 M
MolGrapher	-		45 M
EGATv2 (Ours)	91.99	0.04	4.45 M

Benchmarks: Math Formula Recognition

Handwritten Math Formula Recognition

Dataset	No. of expressions
CROHME 2019 Training	9993
CROHME 2019 Test	1199

Formula recognition results on CROHME test sets

Model Category	Model	Expressions (Structure + Class)			Expressions (Structure)	No. of parameters
		ExpRate	≤1	≤2		
Rule-based	MyScript	79.15	86.82	89.82	90.66	
	Samsung	79.82	87.82	89.15	89.32	
Encoder-decoder	MathType	60.13	74.4	78.57	79.15	
	USTC-iFLYTEK	80.73	88.99	90.74	91.49	~30 M
	ICAL	60.51	78	84.63		7.37 M
	PAL-v2	62.55	74.98	78.4	79.15	
Graph Neural Network (GNN)	QD-GGA	40.65	60.01	64.96	60.22	8 M
	GETD	54.13				~30 M
	GGM-EGAT	60.72	71.14	76.73	83.74	
	Ours (EGATv2)	49.54	66.17	73.92	73.79	4.45 M

Conclusion

Contributions

RQ1: Visual Primitive Features

- Contour-based visual features and iterative refinement to improve segmentation

RQ2: Input Graph Representations

- Compared complete, LOS, and KNN graphs across math and chemistry datasets
- **LOS graphs best for math** (long-range alignment)
6NN graphs best for chemistry (local structure preservation)

Contributions

RQ3: Graph Context and Task Interaction

- **EGATv2**, enables **bidirectional node-edge** multi-hop attention
- Graph attention on **class distributions** with visual features for better task interaction

RQ4: Visual Noise and Loss Functions

- Applied **localized structural and visual noise** robustness
- Applied **weighted focal loss** and **CHM aggregation** to handle class imbalance
- Achieved better generalization for noisy and imbalanced settings

Contributions

- **Less data and fewer model parameters**, improving **data efficiency**
- **Interpretability** through primitive-level **error analysis** of symbol/edge confusions
- Visual primitive extraction, EGATv2, CHM loss aggregation **adaptable** to other diagrams (charts, tables) and multi-task learning problems with class imbalance

Future Work

Extend primitive extraction to math: Apply robust visual primitive extraction to handwritten and typeset math

Enhanced graph attention: Improve EGATv2 with context window aggregation

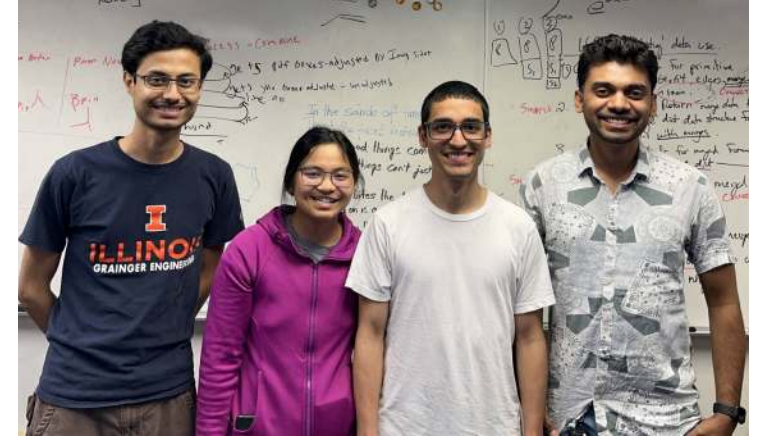
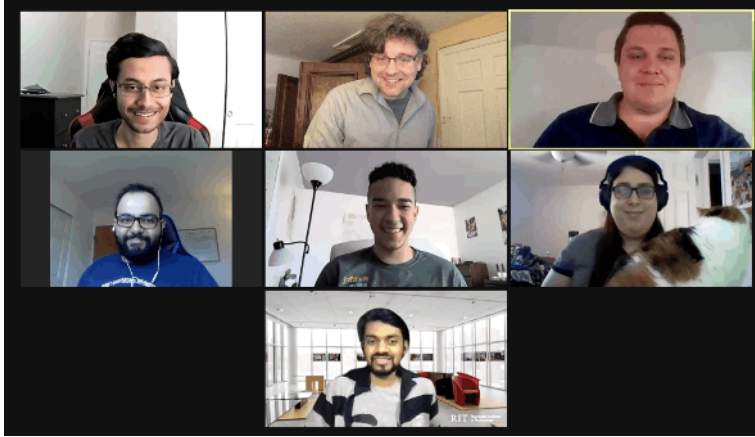
Advanced noise modeling: Simulate realistic document artifacts and structural distortions beyond simple splits

Refined loss functions: Applying CHM within each task loss to better handle class imbalance

Contrastive learning: Use pairwise objectives to improve feature separability between visually similar classes

Handwritten diagrams: Extend model to stroke-based chemical diagrams and establish benchmarks

Thank You



Acknowledgement:

DPRL: Bryan Amador, Abhisek Dey, Matt Langsenkamp, Ming Creekmore, Brandon Kirincich, and Patrick Philippy

NCSA: Matthew Berry, Yashdeep Thorat, Sara Lambert, and Kate Arneson

UIUC: Dr. Scott Denmark, Blake Ocampo

Funding: Alfred P. Sloan Foundation (G-2017-9827), National Science Foundation (IIS-1717997 and 2019897)



Alfred P. Sloan
FOUNDATION

Thank You



Questions?

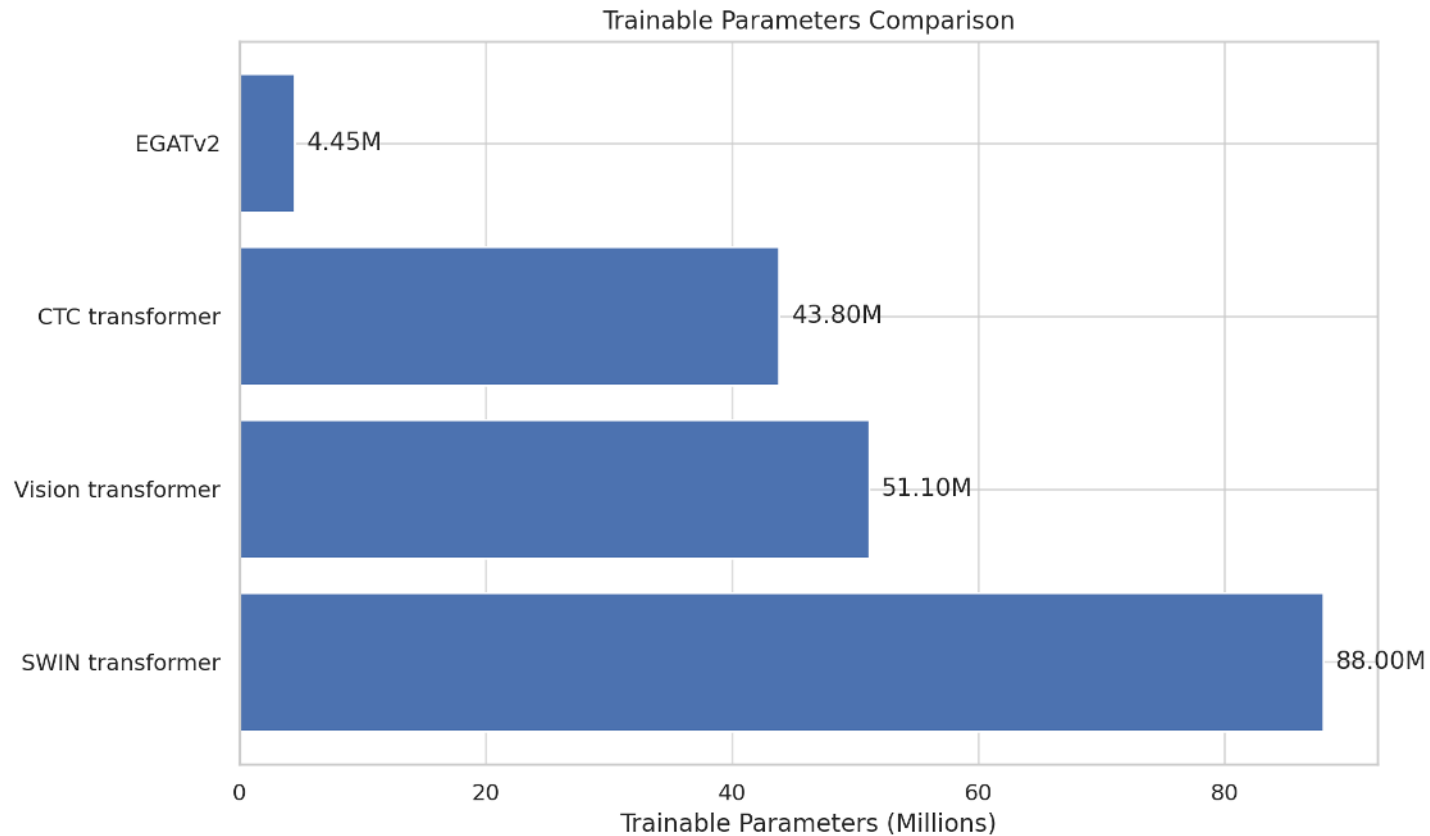
List of Publications

Title	Authors	Venue	Year
Multimodal Search in Chemical Documents and Reactions	A. K. Shah , A. Dey, L. Luo, B. M. Amador, P. Philippy, M. Zhong, et al.	International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)	2025
ChemScraper: Leveraging PDF Graphics Instructions for Molecular Diagram Parsing	A. K. Shah , B. M. Amador, A. Dey, M. Creekmore, B. Ocampo, S. Denmark, and R. Zanibbi	International Journal on Document Analysis and Recognition (IJDAR)	2024
Line-of-Sight with Graph Attention Parser (LGAP) for Math Formulas	A. K. Shah and R. Zanibbi	International Conference on Document Analysis and Recognition (ICDAR)	2023
Searching the ACL Anthology with Math Formulas and Text	B. M. Amador, M. Langsenkamp, A. Dey, A. K. Shah , and R. Zanibbi	International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)	2023
A Math Formula Extraction and Evaluation Framework for PDF Documents	A. K. Shah , A. Dey, and R. Zanibbi	International Conference on Document Analysis and Recognition (ICDAR)	2021

EXTRA SLIDES

Benchmarks

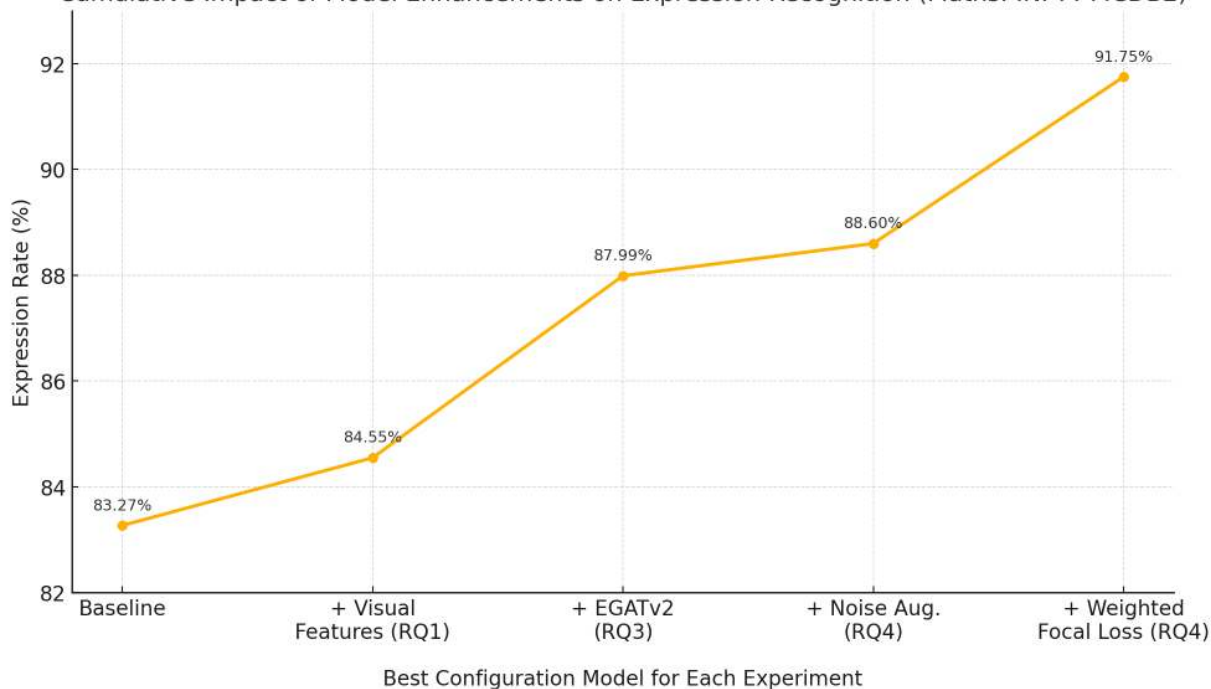
Comparison of Model Parameters



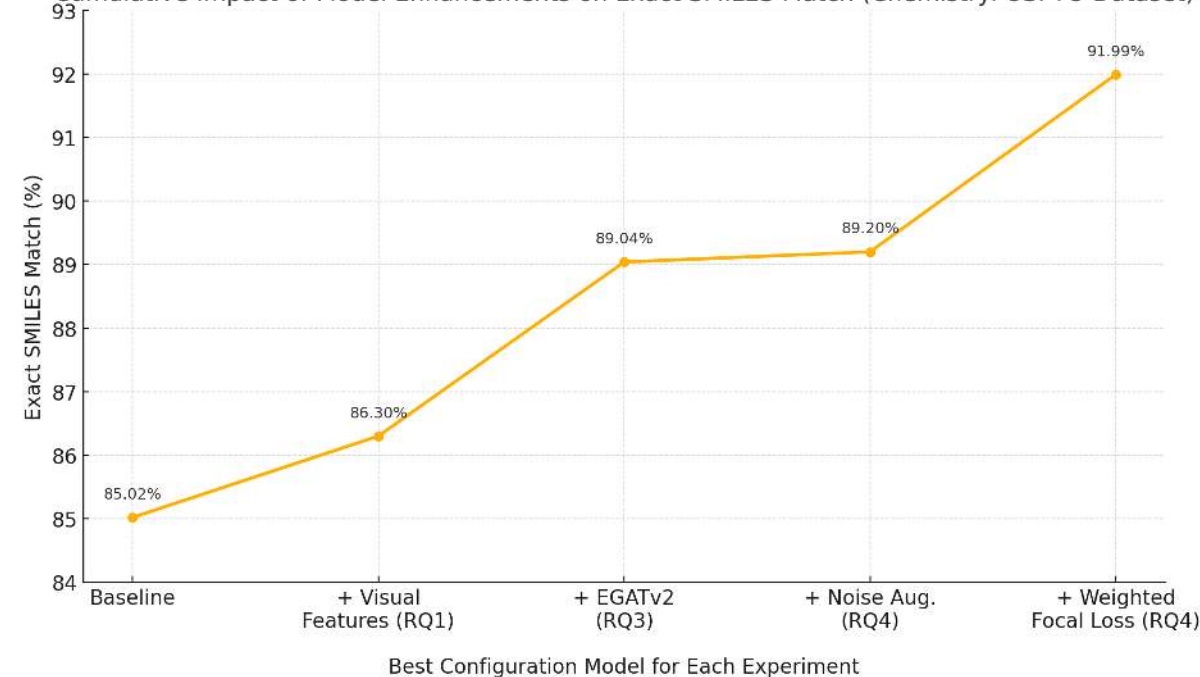
Summary

Best Configuration Model	RQ	Errors (Math)	Errors (Chemistry)
Baseline	-		
+ Better Visual features	RQ1	Fewer symbol class confusions on isolated symbols	Fewer symbol confusions on multi-primitive and symbols (e.g., H, N) and symbols with similar patterns (e.g., W/M, i/j, 1/l)
+ Edge attention (EGATv2)	RQ3	Fewer confusions between similar symbols (e.g., minus vs equal)	Fewer misclassifications of Single lines; improved bond connections
+ Structural and visual noise augmentation	RQ4	Improvements in symbol segmentation and classification, especially for noisy test images	Reduction in errors on symbol segmentation and classification, especially for noisy test images
+ Weighted Focal loss	RQ4	Lower errors on high frequency classes (minus, equal, NoRelation) by emphasizing difficult cases	Lower overprediction of Single class; fewer missed connections (NoRelation)

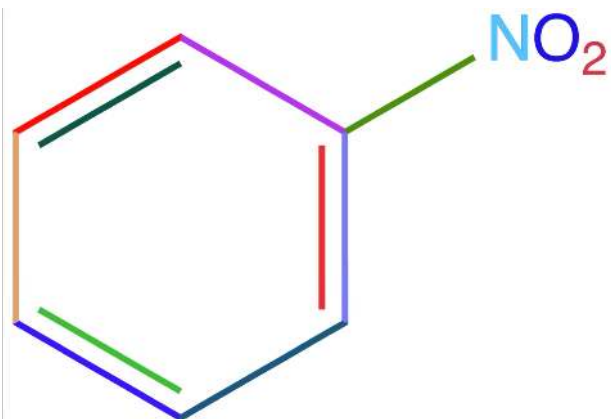
Cumulative Impact of Model Enhancements on Expression Recognition (Maths: INFTY-MCDB2)



Cumulative Impact of Model Enhancements on Exact SMILES Match (Chemistry: USPTO Dataset)



Our Inputs and Outputs



PDF image with PDF primitives

Vector source

Raster source

Inputs

Primitive Level Graph (PDF)

(Visual Graph)

nodes: lines & characters

edges: connections/merges

Symbol Level Graph

(Tokenized Visual Graph)

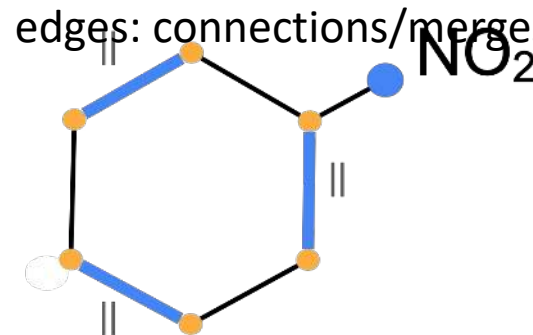
nodes: **bonds**, atoms & atom groups

Primitive Level Graph (Raster)

(Visual Graph)

nodes: lines & characters

edges: connections/merges



Molecular Graph

nodes: atoms & atom groups

edges: **bonds**

PNG image with visual primitives

Graph-based Evaluation Metrics

Symbol detection (segmentation)

True Positive (TP) = Correctly segmented symbols

False Positive (FP) = Incorrectly merged symbols (oversegmentation)

False Negative (FN) = Missed symbols (undersegmentation)

Relationship detection (presence of edge)

TP = Correctly detected edge between correctly segmented symbols

FP = Incorrectly detected non-existent edge in the ground truth

FN = Missed ground truth edge

$$Precision (P) = \frac{|Correct|}{|Detected|} = \frac{TP}{TP + FP}$$

$$Recall (R) = \frac{|Correct|}{|Target|} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

|



Ground Truth

Predicted

Primitive level graph (with errors)

Graph-based Evaluation Metrics

1. Symbol & Relationship Detection F1-scores
2. Symbol & Relationship Detection + classification F1-scores
3. Unlabeled structure recognition rate: Both symbols and relationships detected correctly

$$\text{Unlabeled structure rate} = \frac{|\text{Formulas with correctly detected symbols and relations}|}{|\text{Formulas in the dataset}|}$$

4. Labeled structure recognition rate: Both symbols and relationships detected and classified correctly

$$\text{Labeled structure rate} = \frac{|\text{Formulas with correctly detected and classified symbols and relations}|}{|\text{Formulas in the dataset}|}$$



Ground Truth

Predicted

Primitive level graph (with errors)

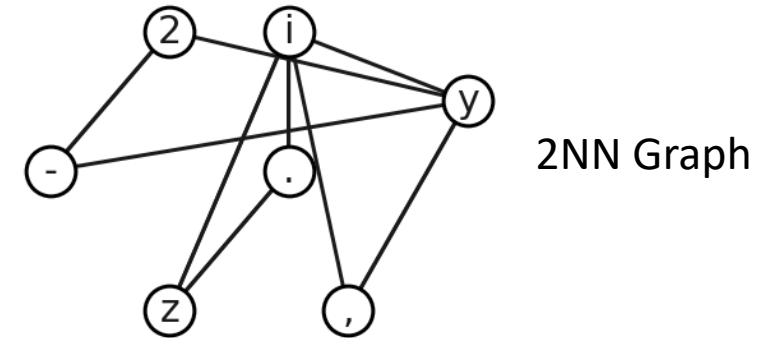
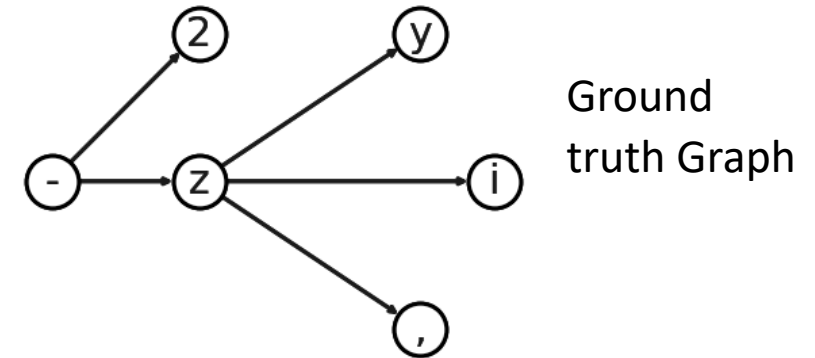
RQ2: Input Graph Representation

$$\text{Edge Recall} = \frac{\# \text{ ground truth edges in the input graph}}{\# \text{ edges in the ground truth graph}}$$

$$\text{Edge Precision} = \frac{\# \text{ ground truth edges in the input graph}}{\# \text{ edges in the input truth graph}}$$

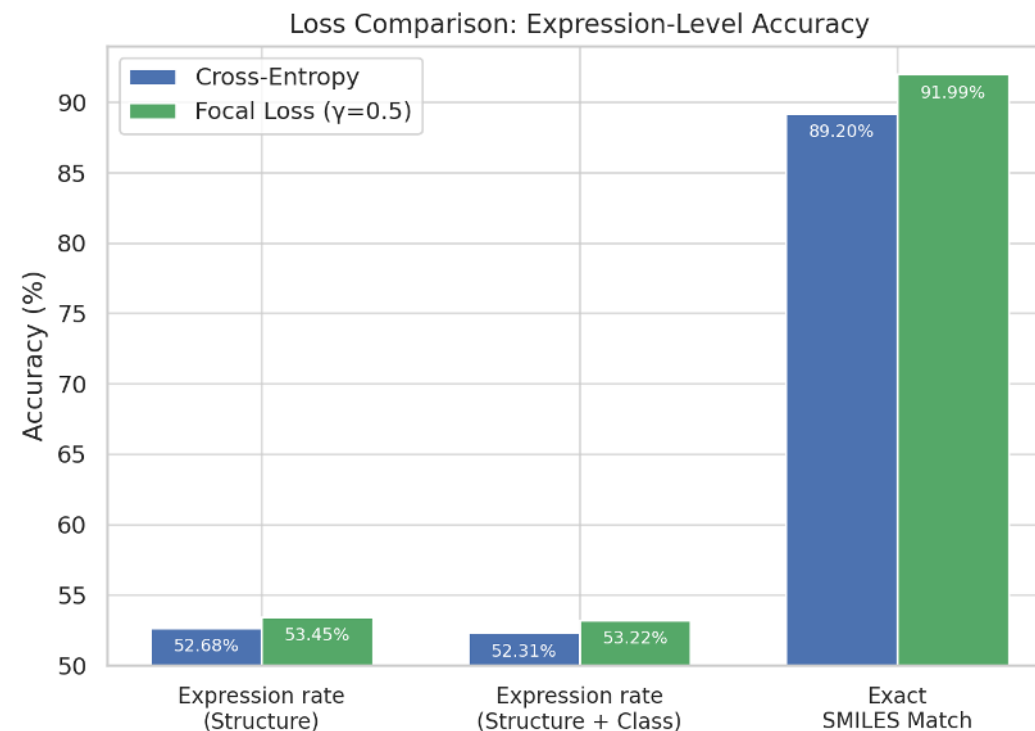
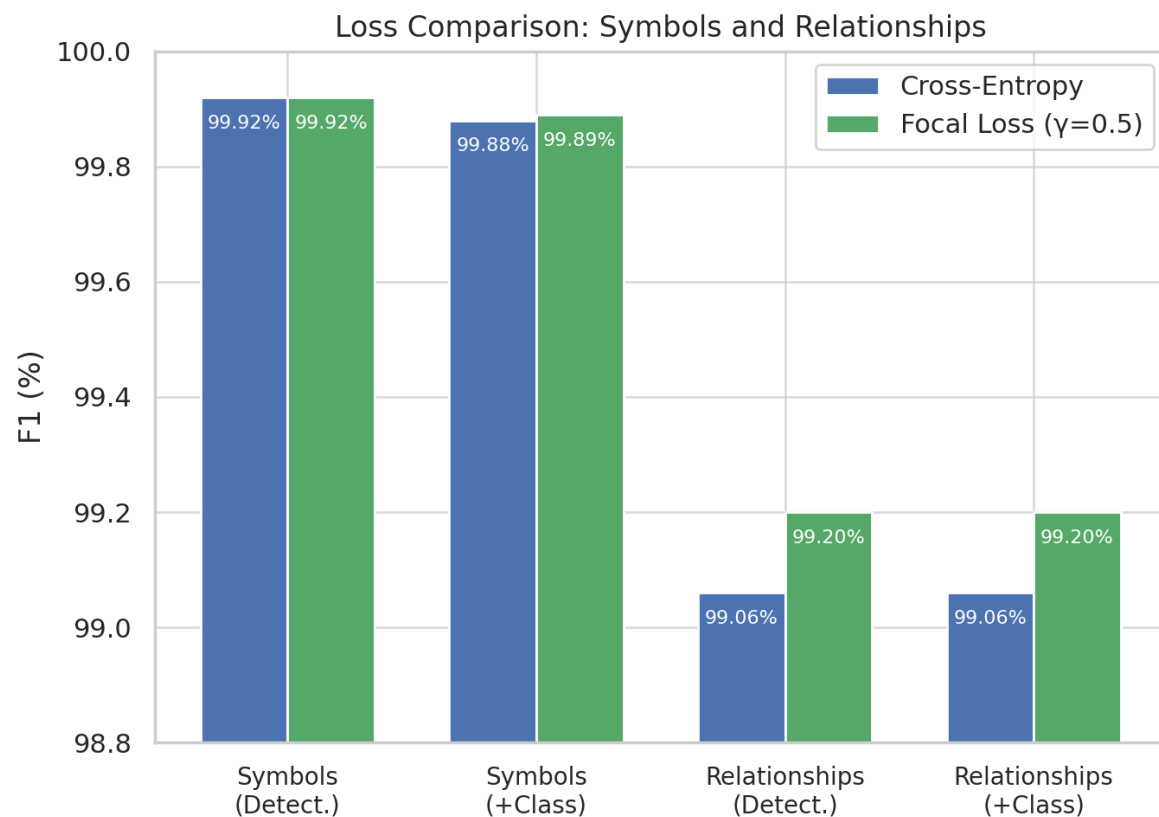
$$\text{Edge F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Expression coverage} = \frac{\# \text{ expressions where all ground truth edges are present}}{\# \text{ total expressions in the dataset}}$$



RQ4: Noise and Class Imbalance

Chemistry: Pubchem-5k



Typeset Math Formula Recognition

Dataset	No. of expressions
Infty-MCDB2 Training	12551
Im2latex-100k Test	10354

Formula recognition results on Im2latex-100k test sets

Model Category	Model	Edit %	BLEU-4 %	Image Match %	Image Match (-ws) %	Data per 1% BLEU-4
Encoder-decoder	IM2TEX		87.73	77.46	79.88	956.150
	WYGIWYS	88.6	90.3	78.6		928.937
	ConvMath	90.8	88.33	83.41		949.655
	EDPA	91.39	92.31	82.07		908.710
	DNN-RL	96.5	94.1	84.2		891.424
	MI2LS	92.3	90.3	82.3	84.8	928.937
	HybridViT		89.9		86.5	933.070
	MathNet	94.7	94.5	63.4		887.651
Graph-based	INFTY	53.82	66.65	15.6	26.66	1258.560
	GNN		90.2	81.8		929.967
	Im2Latex-GNN		90.19	81.82		930.070
	Ours (EGATv2)	55.23	46.05	19.7	17.49	217.155

Results: Molecule Diagram Recognition

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

Systems	Synthetic: Indigo (5719)		Scanned: CLEF-2012 (992)		No. of parameters
	SMILES Match	# Mol / 1% (K)	SMILES Match	# Mol / 1% (K)	
MolVec 0.9.7	95.40		83.80		
OSRA 2.1	95.00		84.60		
Img2Mol	58.90	25.47	48.84	30.71	30 M
SwinOCSR	74.00	60.81	30.00	150.00	120 M
MolScribe	97.50	17.23	88.90	18.90	88 M
MolGrapher	-		90.50	3.31	45 M
EGATv2 (Ours)	91.99	0.04	39.61	0.09	4.45 M

Motivation

The screenshot displays the Math Deck search interface. At the top, the search bar contains the query "language model $P(X|x)$ ". Below the search bar, the results are displayed for the query "language model $P(X|x)$ ". The first result is titled "(1996) An Overview of Corpus-Based Statistics-Oriented ({CBSO}) Techniques for Natural Language Processing" by Su, Keh-Yih and Chiang, Tung-Hui and Chang, Jing-Shin. The result includes a BibTeX entry and a link to the ACL Anthology. The second result is titled "(1999) Unsupervised Models for Named Entity Classification" by Collins, Michael and Singer, Yoram, also with a BibTeX entry and a link to the ACL Anthology. The interface includes a "Math Deck" logo, a "HIDE RESULTS" button, and a "By rank" sorting option. A search bar at the top shows the query "language model $P(X|x)$ ".

language model $P(X|x)$

language model $P(X|x)$

HIDE RESULTS $P(X|x)$ × add formula... × | +

Results for query: language model $P(X|x)$ By rank

(1996) An Overview of Corpus-Based Statistics-Oriented ({CBSO}) Techniques for Natural Language Processing

BibTeX Su, Keh-Yih and Chiang, Tung-Hui and Chang, Jing-Shin $P(X|x)$ Found 2 hits in 1 page

<https://aclanthology.org/O96-2004>

$$J^{\text{PE}} = 1 - P(c_i | x)$$

Minimizing the risk thus reduces to finding the minimum error probability $1 - P(c_i | x)$ or maximum *a posteriori* probability $P(c_i | x)$. The minimum error rate classifier is normally used when the penalty of all types of errors are equally weighted.

In a part-of-speech (POS) tagging task, the minimum error rate classifier can be derived

(1999) Unsupervised Models for Named Entity Classification

BibTeX Collins, Michael and Singer, Yoram $P(X|x)$ Found 2 hits in 2 pages

<https://aclanthology.org/W99-0613>

ChemScraper System

Sort results by SMILES or "Draw / Upload" another format:

Draw / Upload

Location In PDF

27 results

Molecule	Extracted Structure	Original Structure	SMILES	Flag
Unavailable			<chem>C\C(*)=C\C1=CC=CC=C1</chem>	

View Similar Structures

Results Summary

Found 1 time(s) in 1 document

Location in PDF

page: 1

Other Occurrences

page(s): 1

Description

Unavailable

Molecular Information

Formula: C₉H₉
Weight: 117.17 g/mol

Chemical Safety

Unavailable

External References

[PubChem ID: 101209250](#)

Structural Information

Extracted Structure

MeSEtPh

Original Structure

Original Context

TABLE 2. CROSSCOUPLINGS OF ALKYLBORON REAGENTS WITH ALKENYL ELECTROPHILES

Reactant	Alkenyl Electrophile	Conditions	Product(s) and Yield(s) (%)	Refs.
		Pd(OAc) ₂ (5 mol %), PPh ₃ (10 mol %), C ₆ CO ₂ (1.5 eq), THF, 40°, 5 h	(66)	187
		Pd(OAc) ₂ (5 mol %), SPhos (10 mol %), C ₆ CO ₂ (2 eq), toluene/water (4:1), 100°	(71)	188

<<

<

1

2

3

>

>>

10

chemscraper.frontend.staging.mml1.ncsa.illinois.edu

ChemScraper System

Results: Math Formula Recognition

Typeset Formula Recognition (Raster Images)

Typeset formula dataset

Dataset	No. of expressions
InftyMCCDB-2 Training set	12,551
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

Summary

Typeset math formulas: InftyMCDB-2

Best Configuration Model	RQ	Table	Exp Rate	Errors
Baseline		5.6	83.27	
+ Better Visual features	RQ1	5.6	84.55	Fewer symbol class confusions on isolated symbols
+ Edge attention (EGATv2)	RQ3	5.9	87.99	Fewer confusions between similar symbols (e.g., minus vs equal)
+ Structural and visual noise augmentation	RQ4	6.3	88.60	Minor improvements in symbol segmentation and classification, especially for noisy test images
+ Weighted Focal loss	RQ4	6.5	91.75	Lower errors on high frequency classes (minus, equal, NoRelation) by emphasizing difficult cases

Summary

Synthetic chemical diagrams: USPTO

Best Configuration Model	RQ	Table	SMILES Match	Errors
Baseline	-	5.7	85.02	
+ Better Visual features	RQ1	5.7	86.30	Fewer symbol confusions on multi-primitive and symbols (e.g., H, N) and symbols with similar patterns (e.g., W/M, i/j, 1/l)
+ Edge attention (EGATv2)	RQ3	5.10	89.04	Fewer misclassifications of Single lines; improved bond connections
+ Structural and visual noise augmentation	RQ4	6.4	89.20	Reduction in errors on symbol segmentation and classification, especially for noisy test images
+ Weighted Focal loss	RQ4	6.6	91.99	Lower overprediction of Single class; fewer missed connections (NoRelation)

Proposed Work



Multiple virtual global nodes for enriched global context in graph-based parsers



Edge-specific attention for improved context and structural relationships



Iterative self-attention to enhance multi-task learning (MTL) performance



Robust over-segmentation strategies and **synthetic noise augmentation** for parsing noisy images

Research Questions

- **Loss Aggregation:** Sum vs. Complemented Harmonic Mean vs. Contraharmonic Mean

$$\mu_{CHM} = 1 - 3 \cdot \left(\sum_{t \in \{S, M, R\}} \frac{1}{1 - \mathcal{L}_t} \right)^{-1}$$

$$\mu_{Contra} = \frac{\sum_{t \in \{S, M, R\}} \mathcal{L}_t^2}{\sum_{t \in \{S, M, R\}} \mathcal{L}_t}$$

Contributions

Parsing models and techniques



Parsing mathematical formulas and molecular diagrams using vector information



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



Recurrent segmentation-aware parser for raster molecule images

Completed Work

Data generation and evaluation



Annotated ground truth generation for visual parsing of raster images



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



Utilized fine-grained graph-based evaluation metrics for chemical diagrams parsing

Future Work

- Unified graph-based parsing framework for math and chemistry
- Insights into shared challenges and solutions between math and chemical parsing
- Understand the effect of input primitives on parsing
- Understand the utilization of graph structure for context and attention in parsing
- Extend the use of global nodes at multiple levels and sub-levels
- Extend MTL interaction using graph specific structure and attention models

Diao et al., "Relational Attention: Generalizing Transformers for Graph-Structured Tasks," presented at the The Eleventh International Conference on Learning Representations, Sep. 2022

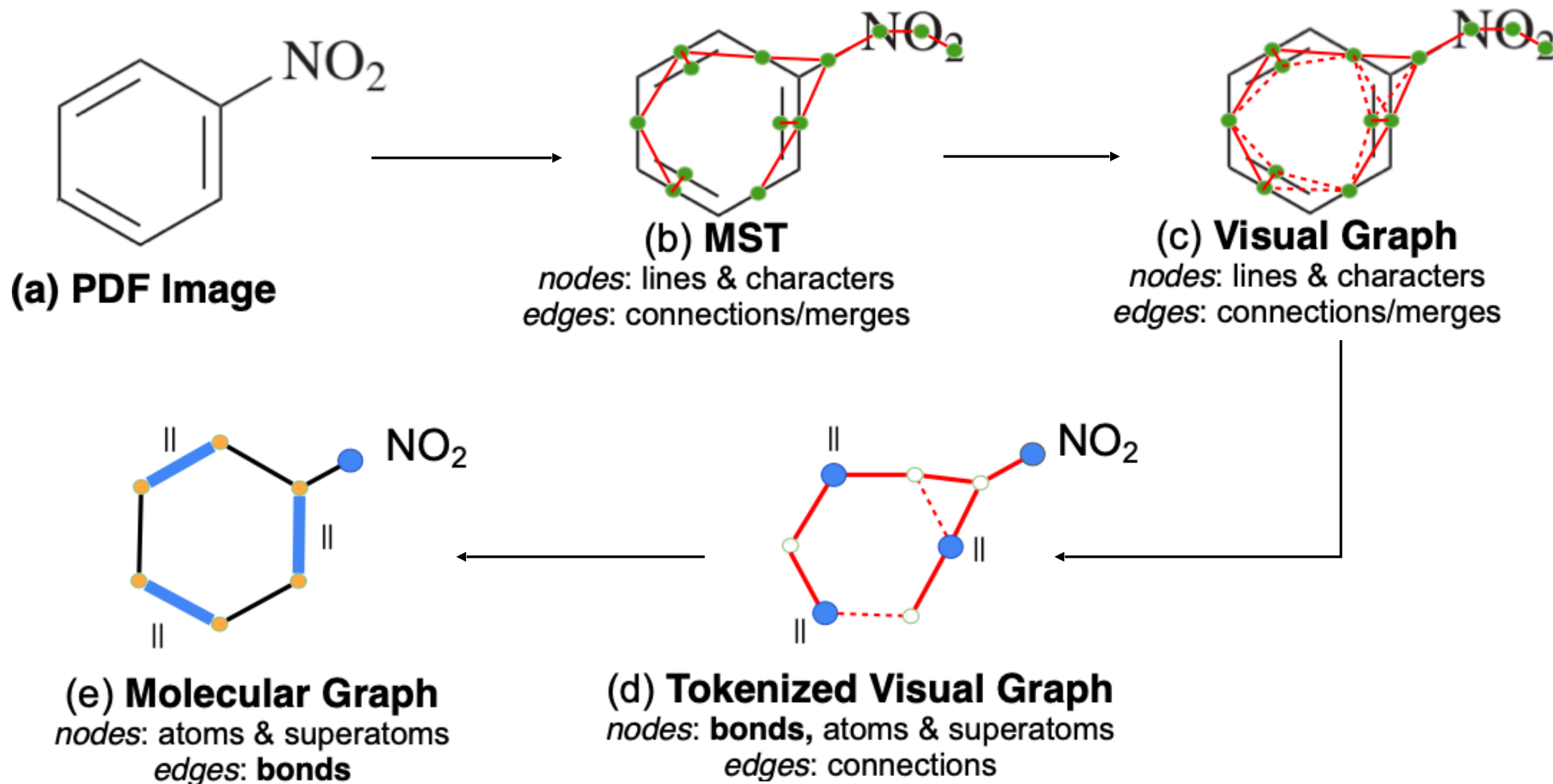
Xie et al., "Local and Global Graph Modeling with Edge-weighted Graph Attention Network for Handwritten Mathematical Expression Recognition," Oct. 24, 2024, *arXiv*: arXiv:2410.18555

Ying *et al.*, "Do Transformers Really Perform Badly for Graph Representation?," presented at the Advances in Neural Information Processing Systems, Nov. 2021.

Xu et al., "PAD-Net: Multi-tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018.

Zhang et al., "Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation," in Computer Vision – ECCV 2018.

Born-digital Parser for Molecules



Results: Molecule Diagram Recognition

Training data source: Pubchem 1 million (5,000 molecules)

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

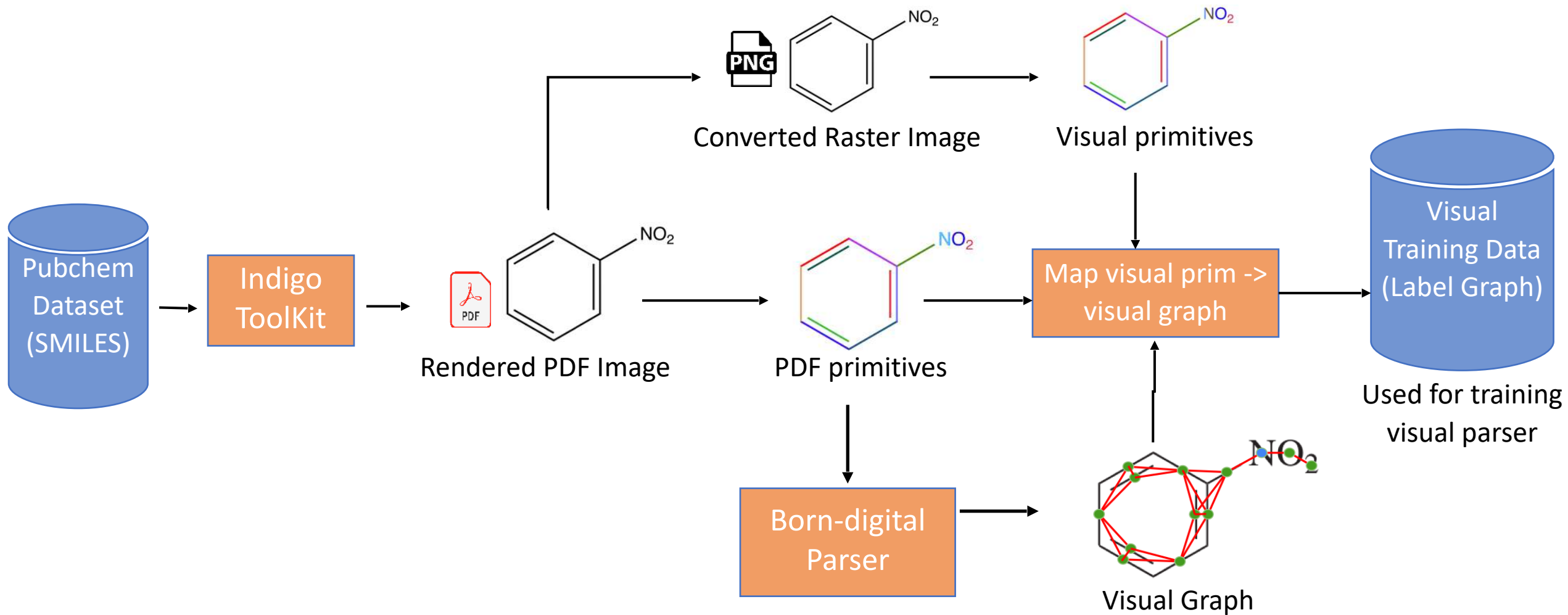
Results: Molecule Diagram Recognition

Training data source: Pubchem 1 million (5,000 molecules)

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)	<i>98.16</i>	<i>89.32</i>	<i>94.41</i>
ChemScraper (Born-Digital – PDF input) * Skipping rendering errors	98.42	96.20	<i>94.41</i>

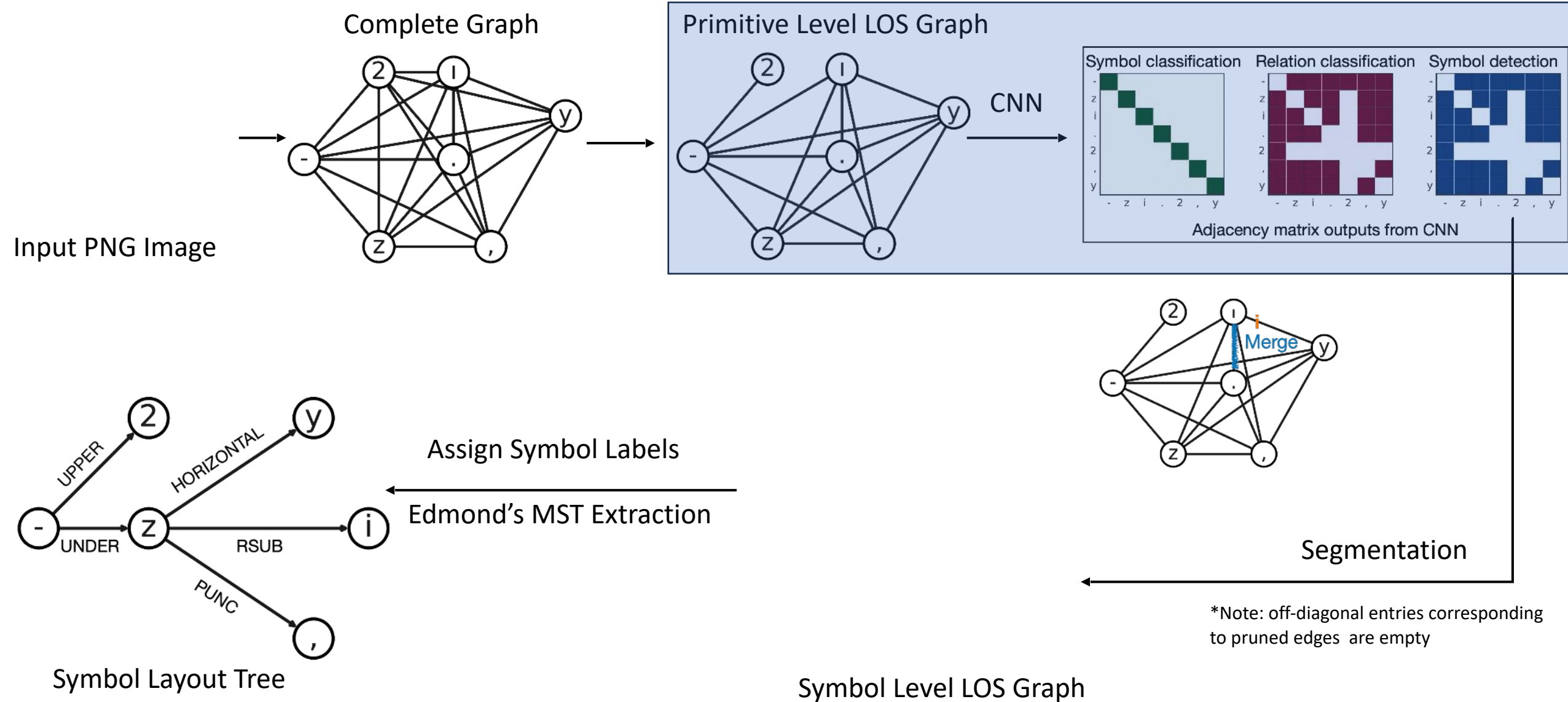
Annotated Data Generation for Visual Parser

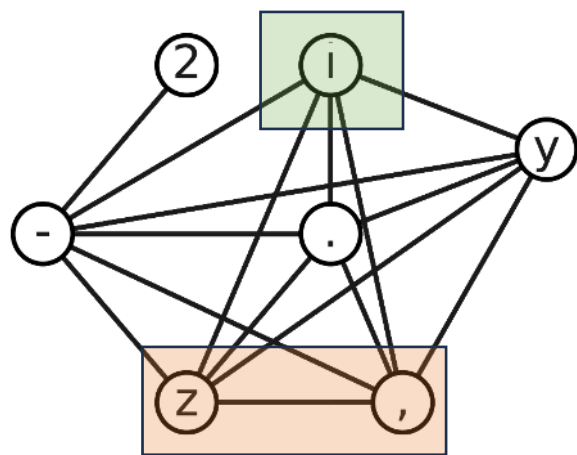


Visual Parser for Math Formulas

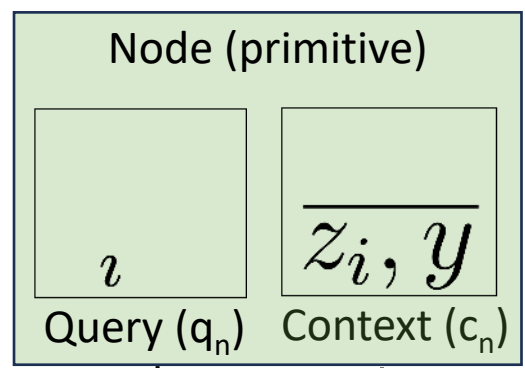
Line-of-sight Graph Attention Parser (LGAP) for Raster Math Formula Image

Visual Parser: Raster Math Formula Image

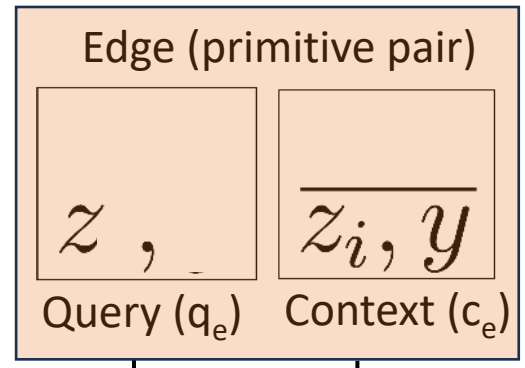
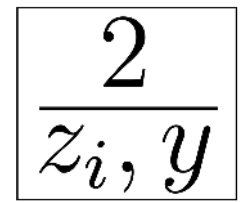




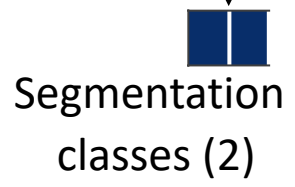
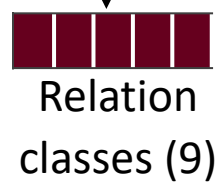
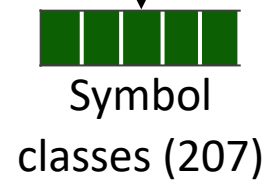
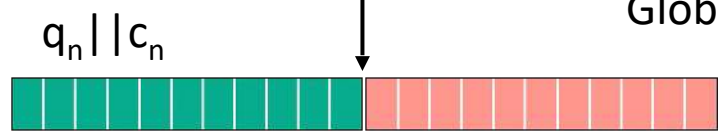
Primitive Level LOS Graph



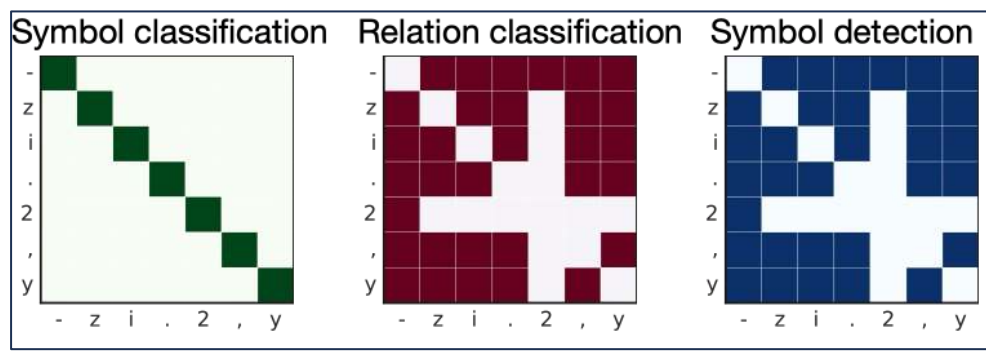
Formula image



Global Feature Map

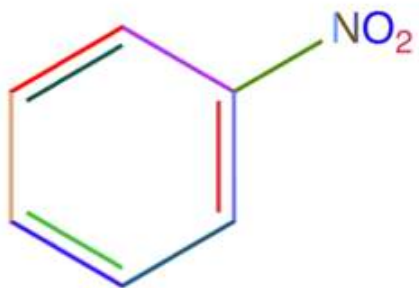


Probability outputs



Line-of-sight Graph Attention Parser (LGAP)

Visual Parser: Raster Molecule Diagram (LCGP)



Visual primitives

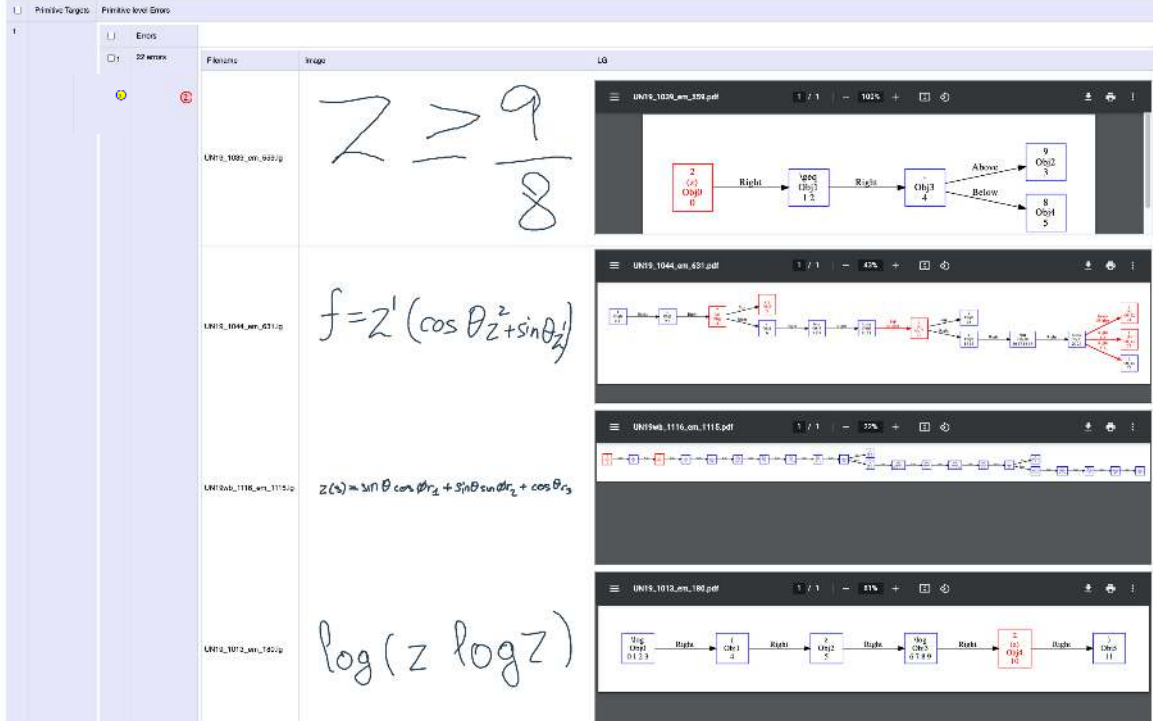
LgEval Extension: Error Visualization

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
						
		☐ 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](#)

Benchmarks

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)	<i>98.16</i>	<i>89.32</i>	<i>94.41</i>
ChemScraper (Born-Digital – PDF input) * Skipping rendering errors	98.42	96.20	<i>94.41</i>
ChemScraper (Visual – PNG input)	85.02	-	-

Results: Math Formula Recognition

Handwritten Formula Recognition

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

Formula recognition results on CROHME test sets

Systems	2014	2016	2019
	Str. + Class	Str. + Class	Str. + Class
SON	37.63	-	-
IM2TEX	38.74	-	-
WAP	46.55	44.55	-
TAP	48.47	44.81	-
QD-GGA	32.04	32.84	-
Graph-to-Graph	54.46	52.05	-
ICAL	60.63	58.79	60.51
GGM-EGAT	-	56.67	60.72
USTC-iFLYTEK	-	-	80.73
Sunia	-	-	88.24
LGAP (Ours)	42.23	41.67	40.70

Results: Math Formula Recognition

Typeset Formula Recognition (Raster Images)

Typeset formula dataset

Dataset	No. of expressions
InftyMCCDB-2 Training set	12,551
InftyMCCDB-2 Test set	6,830

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

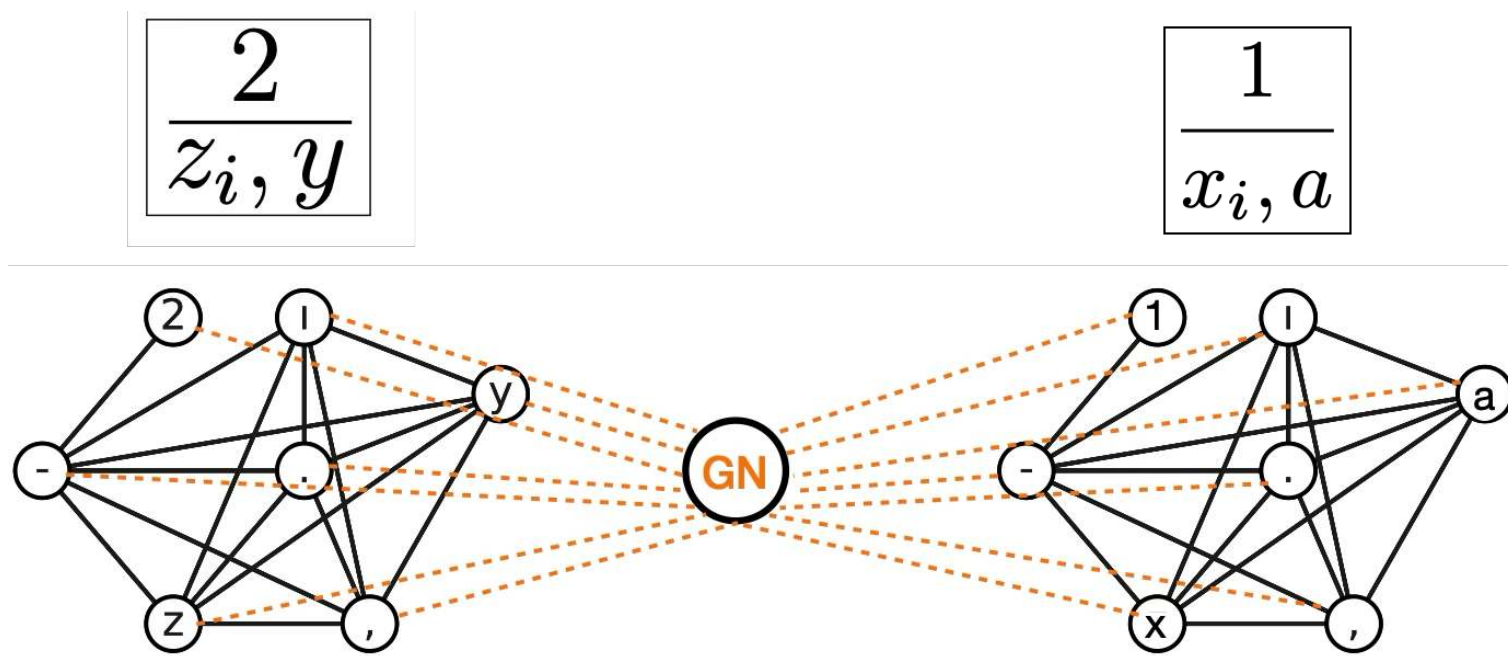
MST Model	Symbols		Relationships		Formulas	
	Detection	+Classification	Detection	+Classification	Structure	+Classification
QD-GGA	98.50	94.54	94.43	93.96	87.77	76.72
LPGA _{RF}	99.34	98.51	97.83	97.56	93.81	90.06
LPGA _{CNN}	99.35	98.95	97.97	97.74	93.37	90.89
LGAP (Ours)	98.32	95.66	94.85	94.35	89.27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

Proposed Work

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



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 global nodes per formula**

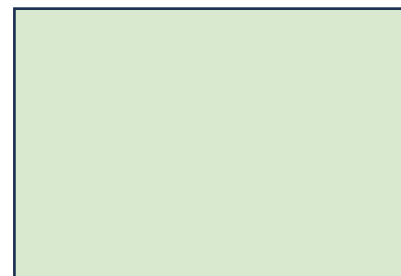
- **Auxiliary tasks:** e.g., binary classification: digit vs. non-digit

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 global nodes per formula**

- **Auxiliary tasks:** e.g., binary classification: digit vs. non-digit



X_S , X_M , and X_C are cross entropy losses given the correct target response vectors (1-hot) and softmax distributions for (S)ymbol classification, primitive (M)erge, and primitive (C)onnected outputs, respectively

X_D : Binary cross entropy loss for digit classification

Q_n : set of node queries (primitives)

Q_e : set of edge queries (primitive pairs)

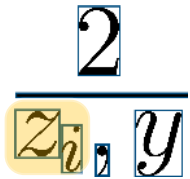
Proposed Work: Edge Attention

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?

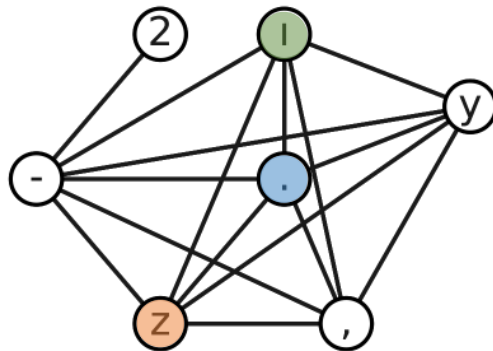
Attention mechanisms: None vs. GAT vs. GAT with edge attention

Approach: Use of edge features to guide attention

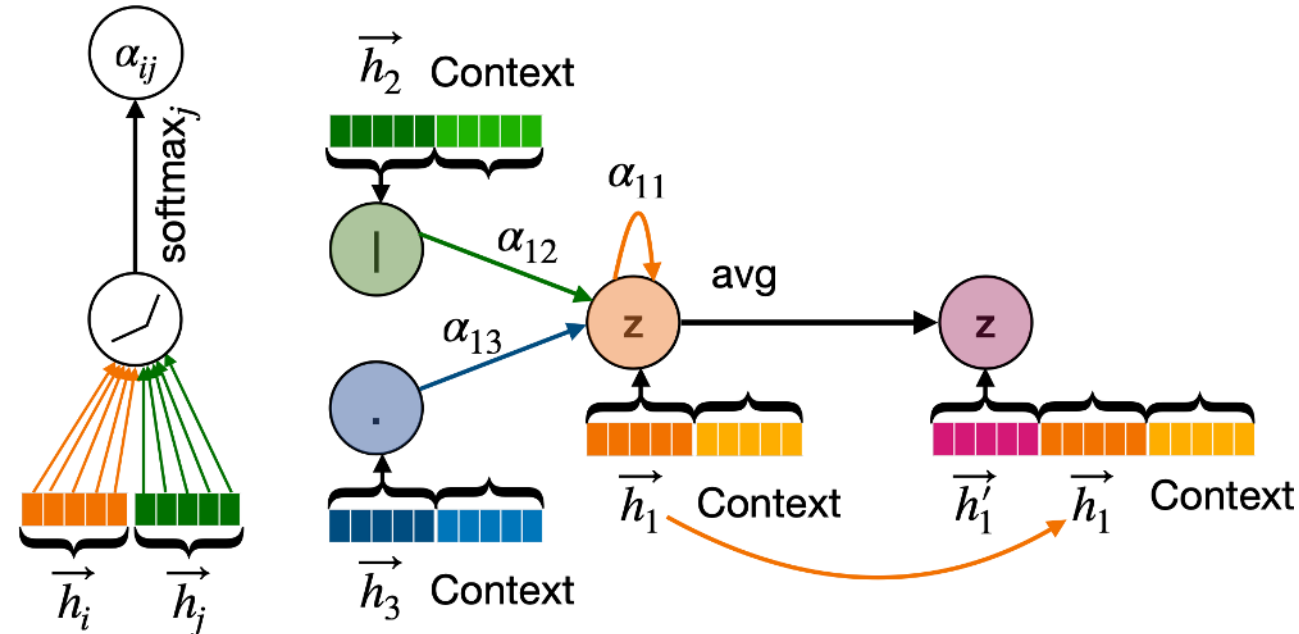
- Add node embeddings during edge attention
- Add edge embeddings during node attention



Formula image



Primitive LOS Graph



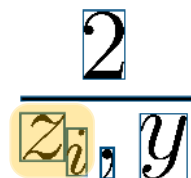
Graph Attention with k=2 LOS Nearest Neighbors

Proposed Work: Edge Attention

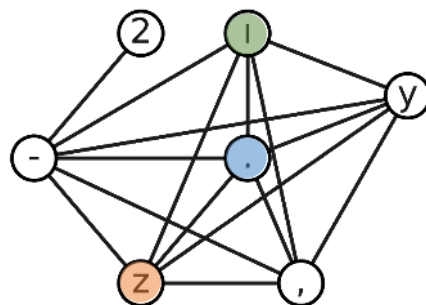
RQ2: Does the incorporation of **edge-specific attention mechanism**, relational transformer attention, improve the performance of graph-based parsers compared to not incorporating edge features, using GAT as baseline?

Approach: Use of edge features to guide attention

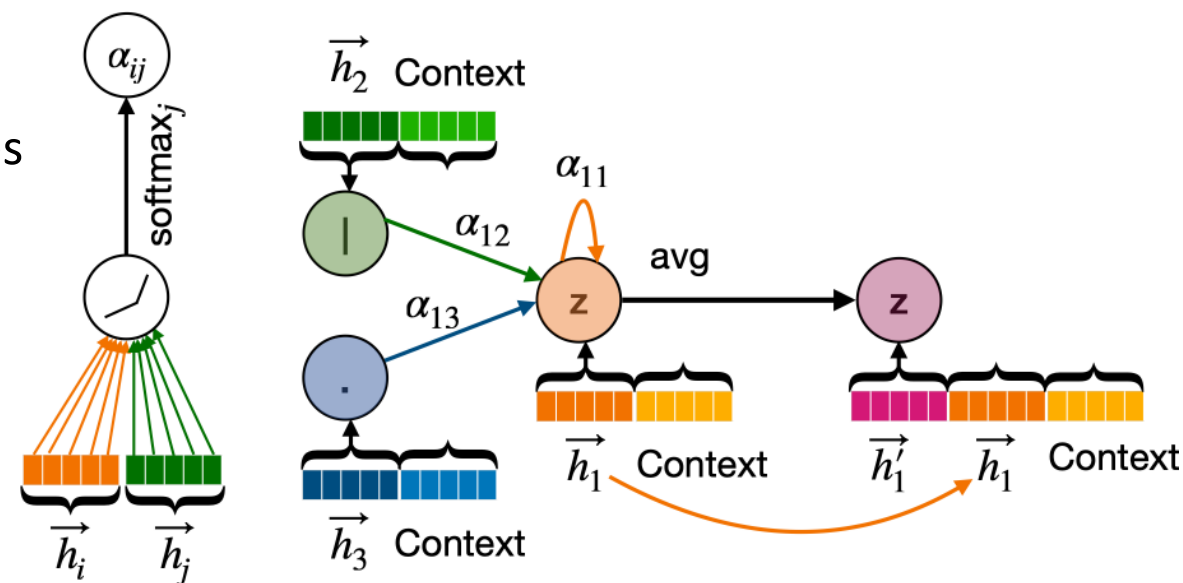
- Bias in attention: edge encodings, edge + spatial distances
- Aggregate edge encodings along the shortest path, average path.



Formula image



Primitive LOS Graph



Graph Attention with k=2 LOS Nearest Neighbors

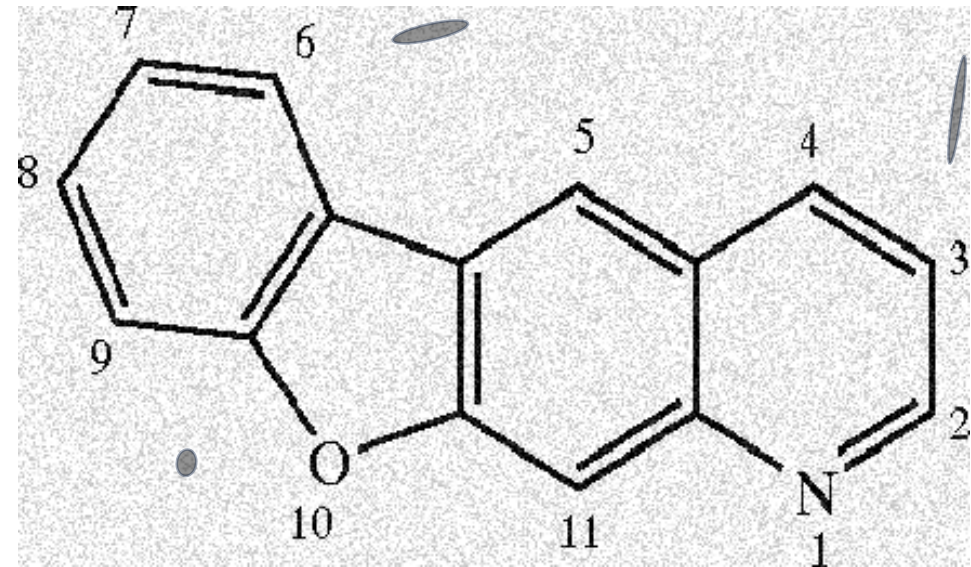
Proposed Work: Over-segmentation and Noise

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

- Control amount of over-segmentation
- Add “noise” symbol class

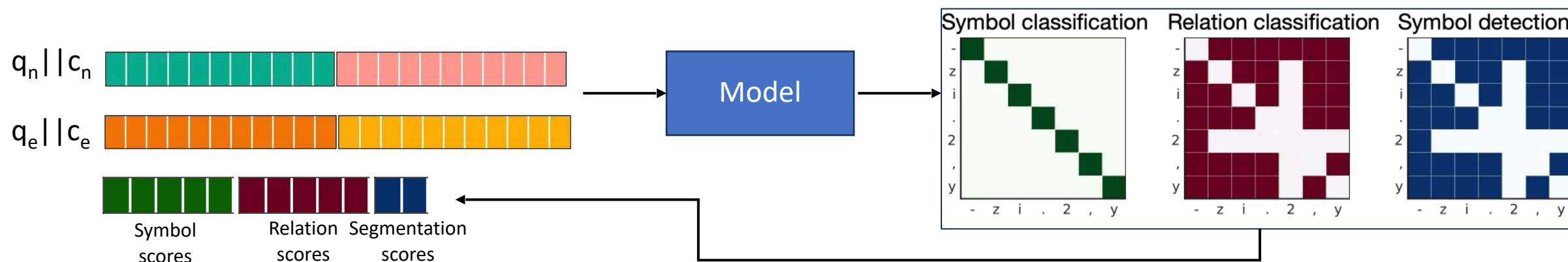


Proposed Work: Multi-Task Learning

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

1. Self-attention and cross-attention between tasks
2. Node-edge interaction through attention



Ying *et al.*, "Do Transformers Really Perform Badly for Graph Representation?," presented at the Advances in Neural Information Processing Systems, Nov. 2021.

L. Morin *et al.*, "MolGrapher: Graph-based Visual Recognition of Chemical Structures," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

Xu *et al.*, "PAD-Net: Multi-tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018.

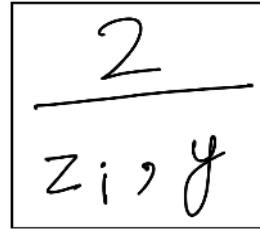
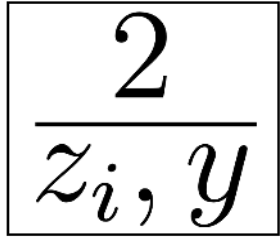
Zhang *et al.*, "Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation," in Computer Vision – ECCV 2018.

Related Work

Vector (e.g. PDF)

Raster (e.g. PNG)

Handwritten



`"\frac{2}{z_i, y}"`

LATEX String

Symbol Layout Tree (SLT)

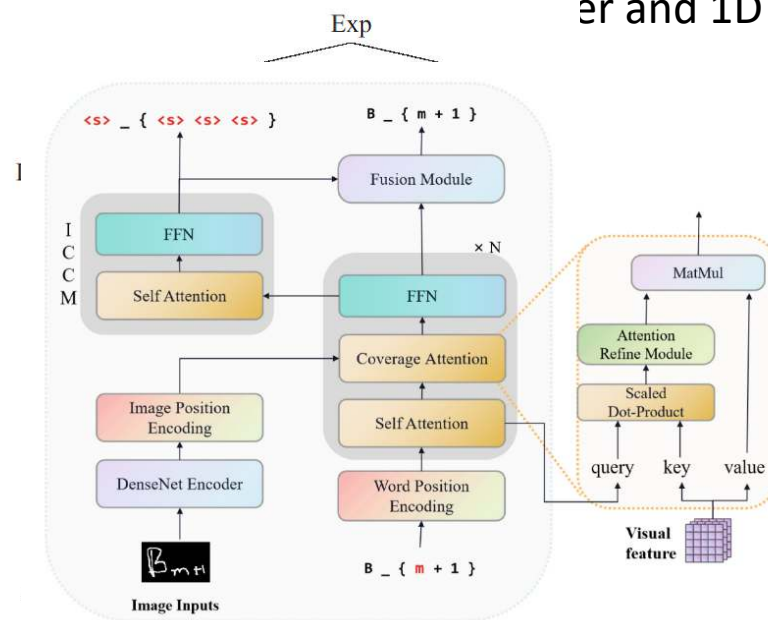
Syntactic Models: generate parse tree, and linearize it to LATEX string using context-free grammars (CFGs, SCFGs)

Encoder-Decoder Models: 2D

er and 1D LATEX string decoder (with attention)

Graph-based Models: graph d

h input to output graph (symbol layout tree)



R. H. Anderson, "Syntax-directed recognition of hand-printed mathematical expressions," *Journal of Experimental Applied Mathematics*, Aug. 1967

Álvaro et al., "An integrated grammar-based approach for mathematical expression recognition," *Pattern Recognition*, Mar. 2016

Zhang et al., "Track, Attend, and Parse (TAP): An End-to-End Framework for Online Handwritten Mathematical Expression Recognition," *IEEE Transactions on Multimedia*, Jan. 2019

Zhu et al., "ICAL: Implicit Character-Aided Learning for Enhanced Handwritten Mathematical Expression Recognition," in *Document Analysis and Recognition - ICDAR 2024*

Xie et al., "Stroke-Level Graph Labeling with Edge-Weighted Graph Attention Network for Handwritten Mathematical Expression Recognition," in *Document Analysis and Recognition - ICDAR 2024*

Nakagawa et al., "A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and GNN models," *Pattern Recognition*, vol. 153, p. 110531, Sep. 2024

Molecule Diagram Parsing

Vector (e.g. PDF) Raster (e.g. PNG) Handwritten



"C1=CC=C(C=C1)[N+](=O)[O-]"

SMILES String

Syntactic Models: generate SMILES string using image processing and chemical/geometrical constraints

Molecular Graph

Encoder-Decoder Models: 2D image pixels or strokes encoder and 1D SMILES string (or others) decoder

Graph-based Models: graph decoder to convert encoded image or graph input to output graph

Our Approach

Ray et al., Finding chemical records by digital computers. *Science*, 1957

McDaniel et al., "Kekule: OCR-optical chemical (structure) recognition," *J. Chem. Inf. Comput. Sci.*, Jul. 1992

Filippov et al., "Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution," *J. Chem. Inf. Model.*, Mar. 2009

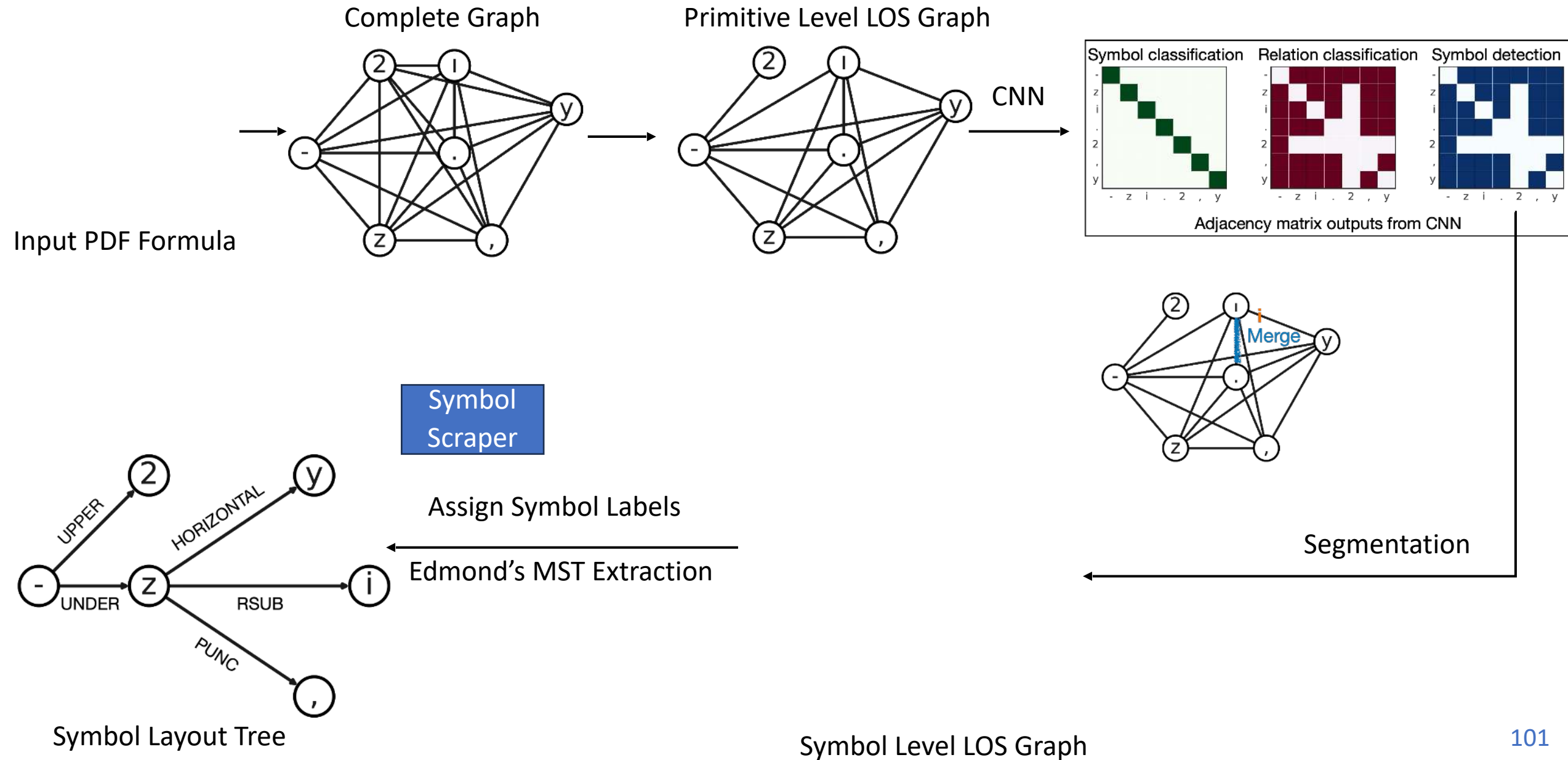
Rajan et al., "Advancements in hand-drawn chemical structure recognition through an enhanced DECIMER architecture," *Journal of Cheminformatics*, Jul. 2024

Qian et al., "MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation," *J. Chem. Inf. Model.*, Apr. 2023

Yoo et al., "Image-to-graph transformers for chemical structure recognition," in *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, May 2022

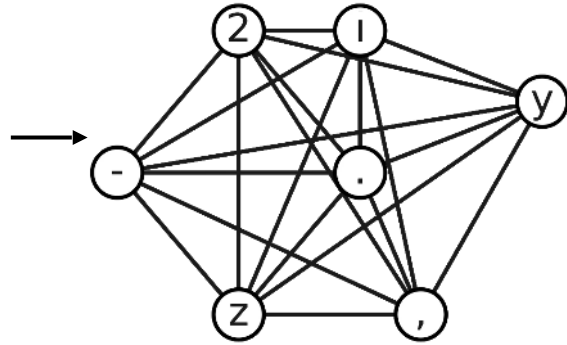
EXTRA SLIDES OLD

Born-digital Parser: Math Formula (PDF)

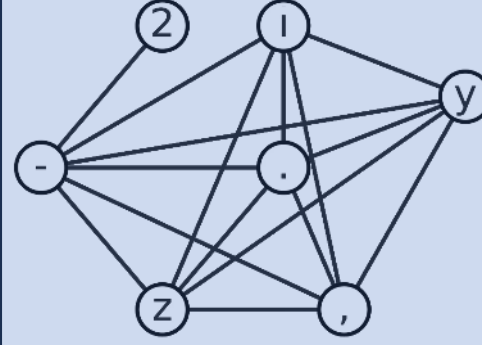


Visual Parser: Raster Math Formula Image

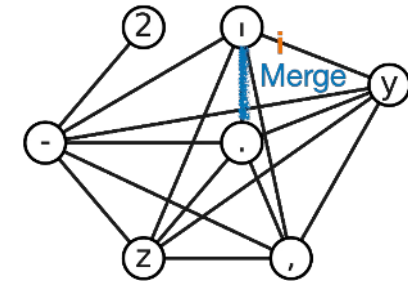
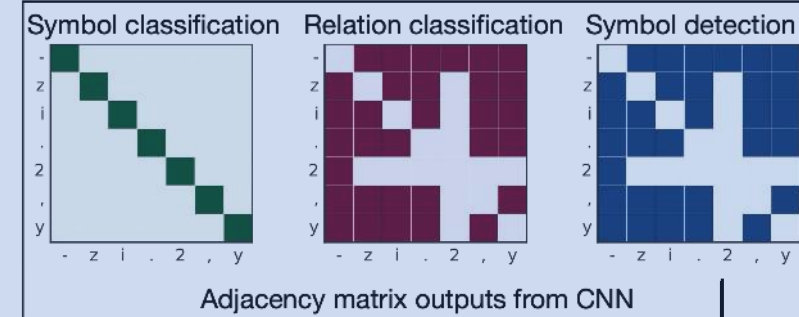
Complete Graph



Primitive Level LOS Graph



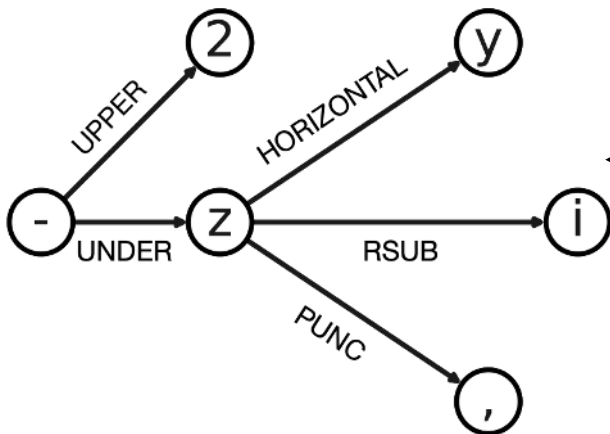
CNN



Segmentation

Assign Symbol Labels

Edmond's MST Extraction



Symbol Layout Tree

Symbol Level LOS Graph

Visual Primitives Extraction for Raster Images (PNG)

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

Input Raster Image



Simplified Polygons

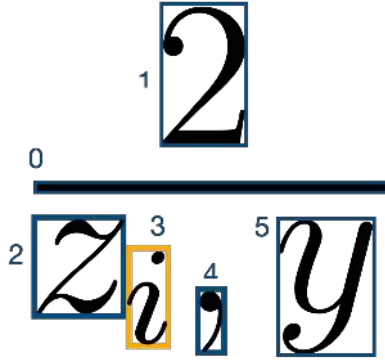


Skeletal Lines



Visual Primitives

Our Inputs and Outputs



Vector source

PDF image

Raster or handwritten source

Inputs **Outputs**

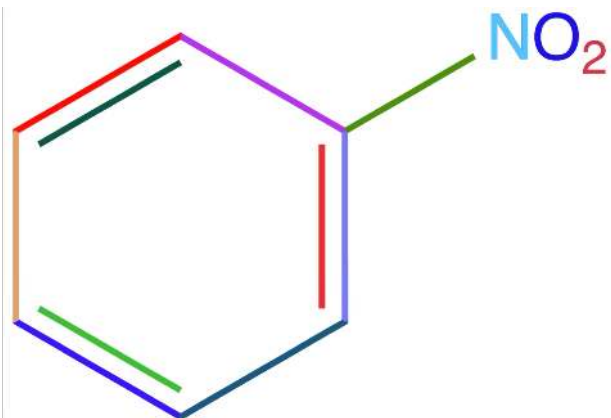
LATEX String

MathML String

PNG image with CCs Handwritten strokes

Symbol layout tree (symbols)

Our Inputs and Outputs



PDF image with PDF primitives

Vector source

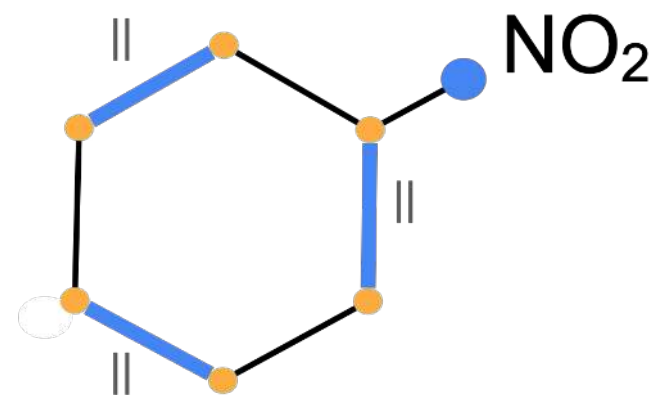
Raster source

Inputs

SMILES String

CDXML File

Outputs

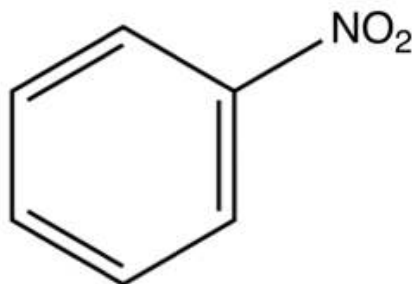


Molecular Graph

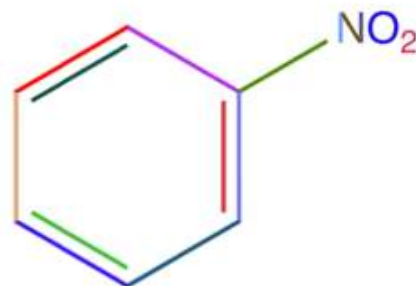
nodes: atoms & atom groups

PNG image with visual primitives

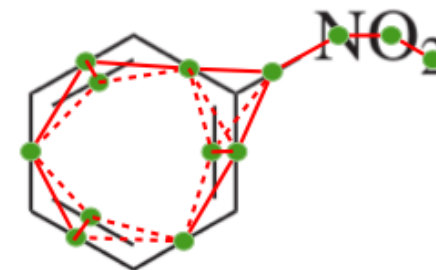
Annotated Data Generation for Visual Parser



Rendered PDF Image
(from SMILES)



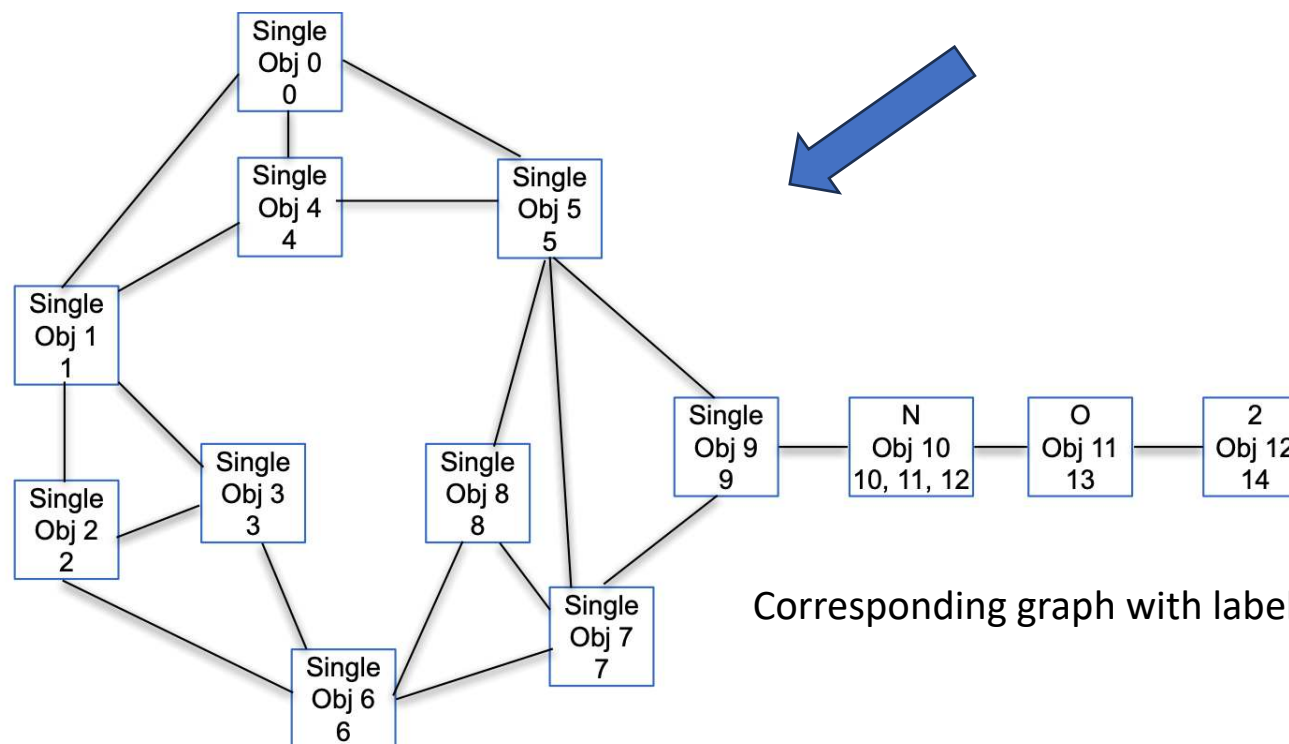
Visual primitives



Visual graph generated
by born-digital parser

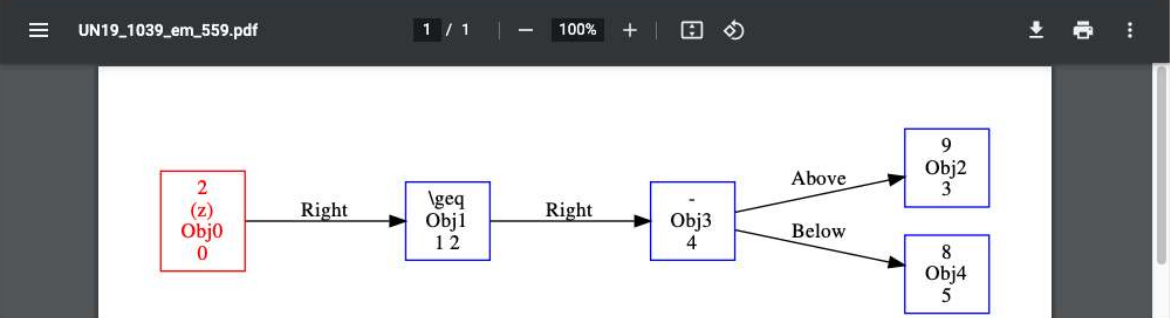
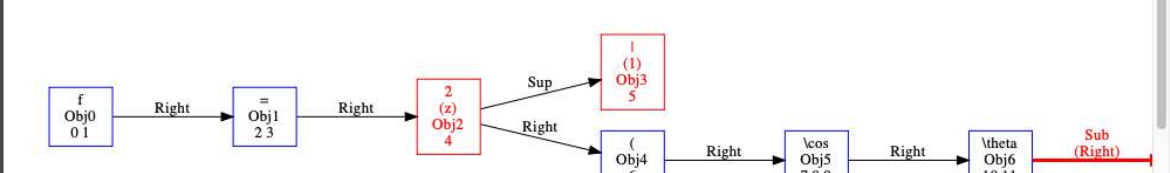
```
# [ OBJECTS ]  
# Objects (O): 10  
# Format: O, objId, class, 1.0, [primitivelId list]  
O, Obj0, Single, 1.0, 0  
O, Obj1, Single, 1.0, 1  
O, Obj10, N, 1.0, 10, 11, 12  
...  
  
# [ RELATIONSHIPS ]  
# Relationships (R): 11  
# Format: R, parentId, childId, class, 1.0 (weight)  
R, Obj0, Obj4, 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

LgEval Extension: Error Visualization

Primitive Targets		Primitive level Errors		Filename	Image	LG
1		☐ Errors ☐ 1 22 errors	2	UN19_1039_em_559.lg		
				UN19_1044_em_631.lg		

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

ChemScraper System

Sort results by SMILES or "Draw / Upload" another format: Draw / Upload Location In PDF Filter

27 results

Molecule ↑↓	Extracted Structure	Original Structure	SMILES	Flag
Unavailable			<chem>C\C(*)=C\C1=CC=CC=C1</chem>	

⊕ View Similar Structures

Results Summary

Found 1 time(s) in 1 document

Location in PDF

page: 1

Other Occurrences

page(s): 1

Description

Unavailable

Molecular Information

Formula: C₉H₉
Weight: 117.17 g/mol

Chemical Safety

Unavailable

External References

PubChem ID: 101209250

Structural Information

Extracted Structure
MeSEtPh

Original Structure

Original Context

Open in PDF viewer

TABLE 2. CROSSCOUPLINGS OF ALKYLBORON REAGENTS WITH ALKENYL ELECTROPHILES

Reactant	Alkenyl Electrophile	Conditions	Product(s) and Yield(s) (%)	Refs.
		Pd(OAc) ₂ (5 mol %), PPh ₃ (10 mol %), C ₆ CO ₂ (1.5 eq), THF, 40°, 5 h	(66)	187
		Pd(OAc) ₂ (5 mol %), SPhos (10 mol %), C ₆ CO ₂ (2 eq), toluene/water (4:1), 40°, 5 h	(71)	188

<< < 1 2 3 > >>

10

chemscraper.frontend.staging.mml1.ncsa.illinois.edu

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

Handwritten Formula Recognition

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	2019
	Str. + Class	Str. + Class	Str. + Class
SON	37.63	-	-
IM2TEX	38.74	-	-
WAP	46.55	44.55	-
TAP	48.47	44.81	-
MAN	54.05	50.56	-
QD-GGA	32.04	32.84	-
Graph-to-Graph	54.46	52.05	-
ICAL	60.63	58.79	60.51
GGM-EGAT	-	56.67	60.72
USTC-iFLYTEK	-	-	80.73
Sunia	-	-	88.24
LGAP (Ours)	42.23	41.67	40.70

Results: Molecule Diagram Recognition

truth

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

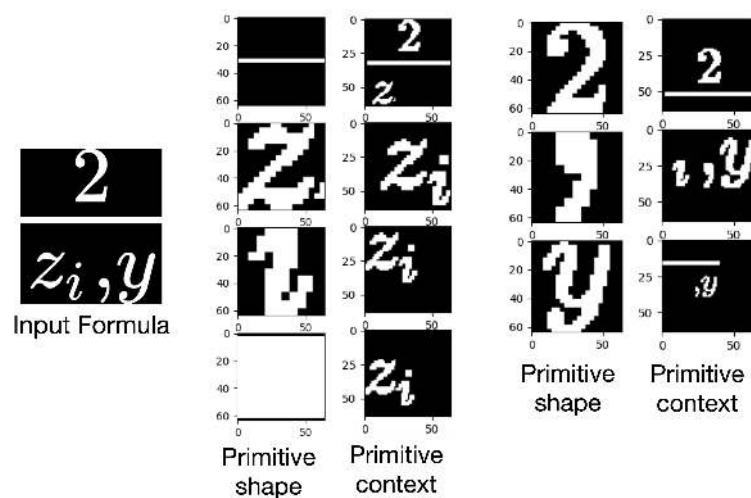
Default: terminal-hetero, implicit hydrogens, thickness=1

Hardest: all, implicit hydrogens, thickness=1.5

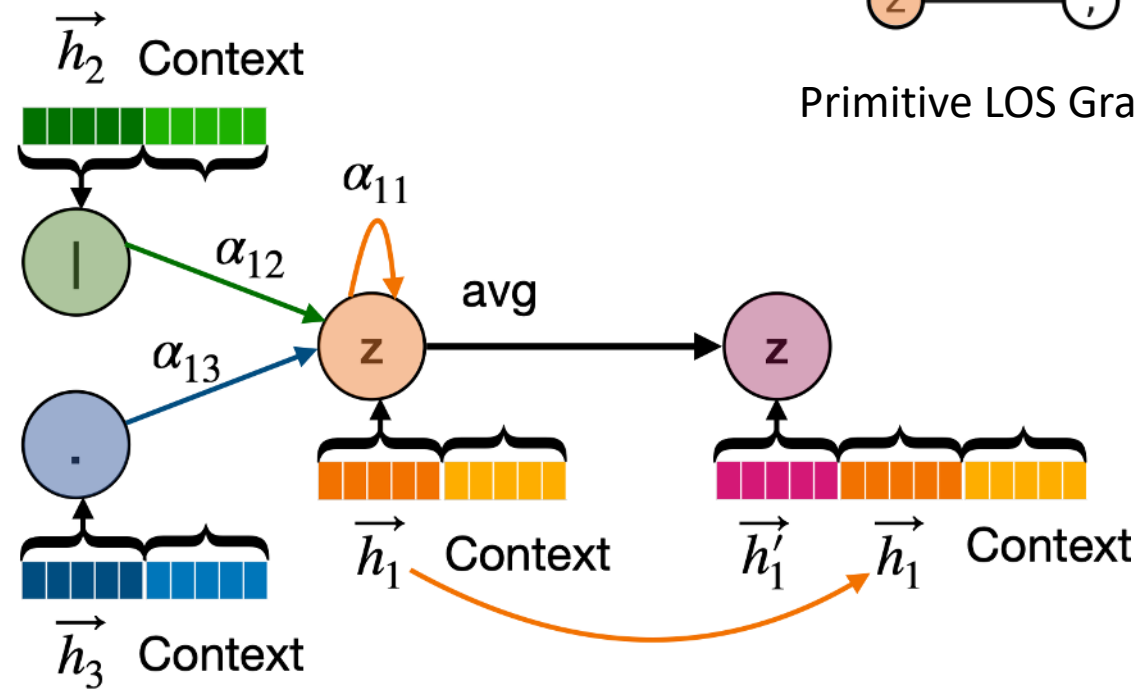
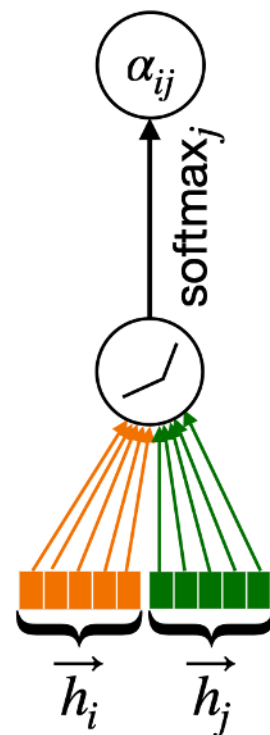
Rendering a molecule with different parameters (Indigo toolkit)

Graph Attention Network (GAT)

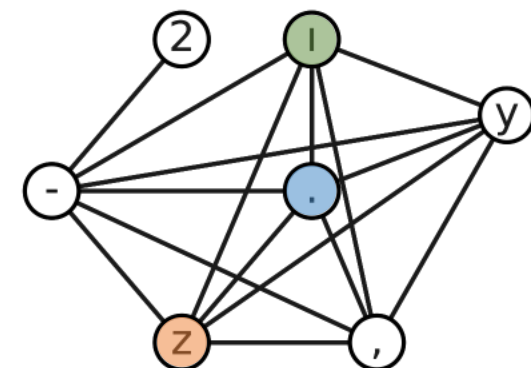
- Consistent separation of shape and context visual features for all tasks
- Modified GAT[3] with KNN for learning correlation between individual primitives
- GAT for interaction among three classification distributions
- 5 % improvement in recognition rate



Shape and Context Features



Graph Attention with k=2 LOS Nearest Neighbors



Primitive LOS Graph

Recognition Results Visualization (HTML)

MathSeer Pipeline Results Visualization

Pdf name: K15-1002

Page: 4

Previous page Home Next page

Page image

2.3 Joint Inference Framework

We extend expression (1) to facilitate joint inference on mention heads and coreference as follows:

$$\begin{aligned} \arg \max_{\mathbf{y}} \sum_{u,v \in M'} \log \phi(y_{u,v}) - \sum_{m \in M} \log \phi(y_m) \\ \text{s.t. } \sum_{u,v \in M'} y_{u,v} \leq 1, \forall u,v \in M', \\ \sum_{m \in M} y_m \leq 1, \forall m \in M, \\ y_{u,v} \in \{0, 1\}, \forall u,v \in M', \\ y_m \in \{0, 1\}, \forall m \in M. \end{aligned}$$

Here, M' is the set of all mention head candidates, $y_{u,v}$ is the decision variable for mention head candidate $y_{u,v}$ if and only if the mention head u is chosen. To consider coreference decisions and mention head decisions together, we add the constraint $\sum_{u,v \in M'} y_{u,v} \leq 1$, which ensures that if a candidate mention head n is not chosen, then it will not have coreference links with other mention heads.

2.4 Joint Learning Framework

To support joint learning of the parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ described above, we define a joint training objective function $\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2)$ for mention head detection and coreference, which uses a max-margin approach to learn both weight vectors. Suppose we have a collection of documents $\mathcal{D}$, and we generate $\mathcal{M}_d$ mention head candidates for each document d ($d \in \mathcal{D}$). We use an indicator function $\mathbb{I}(a, b)$ to represent whether mention heads u, v are in the same coreference cluster based on gold annotations ($\mathbb{I}(u, v) = 1$ iff they are in the same cluster). Similarly, $\mathbb{I}(m)$ is an indicator function representing whether mention head m is valid in the gold annotations.

For simplicity, we first define

$$\begin{aligned} u' &= \arg \max_{u \in \mathcal{M}_d} \theta(u, v) - \delta(u, v), \\ u'' &= \arg \max_{u \in \mathcal{M}_d, u \neq u'} \theta(u, v) - \delta(u, v). \end{aligned}$$

We then minimize the following joint training objective function $\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2)$:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2) = \frac{1}{D} \sum_{d \in \mathcal{D}} \sum_{u,v \in \mathcal{M}_d} \{ C_{\text{coref},m}(u,v) \\ + C_{\text{head},m}(u,v) + C_{\text{coref},m}(u,v) - R(\mathbf{w}_1, \mathbf{w}_2) \}. \end{aligned}$$

$\mathcal{L}(\mathbf{w}_1, \mathbf{w}_2)$ is composed of four parts. The first part is the loss function for coreference, where we have

$$\begin{aligned} C_{\text{coref},m}(u,v) &= -v_2 \phi(u', v') \Omega(u'), \\ &+ [1 - v_2 \phi(u', v') - \delta(u', v')] \Omega(u') \phi(u''). \end{aligned}$$

It is similar to the loss function for a latent left-linking coreference model⁵. As the second component, we have the quadratic loss for the mention head detection model,

$$C_{\text{head},m}(u,v) = \frac{1}{2} (v_1 \phi(u') - \Omega(u'))^2.$$

Using the third component, we further maximize the margin between valid and invalid mention head candidates when they are selected as the best-left-link mention heads for any valid mention head. It can be represented as

$$C_{\text{margin},m}(u,v) = \frac{1}{2} (v_1 \phi(u') - \Omega(u'))^2 - \Omega(u')^2.$$

The last part is the regularization term

$$R(\mathbf{w}_1, \mathbf{w}_2) = \frac{\lambda_1}{2} \|\mathbf{w}_1\|^2 + \frac{\lambda_2}{2} \|\mathbf{w}_2\|^2.$$

2.5 Stochastic Subgradient Descent for Joint Learning

For joint learning, we choose stochastic subgradient-descent (SGD) approach to facilitate performing SGD on a per mention head basis. Next, we describe the weight update algorithm by defining the subgradients.

The partial subgradient w.r.t. mention head m for the head weight vector $\mathbf{w}_1$ is given by

$$\begin{aligned} \nabla_{\mathbf{w}_1, m} \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2) = \\ \frac{1}{D} \sum_{d \in \mathcal{D}} \{ \nabla_{\mathbf{w}_1, m} C_{\text{coref},m}(u,v) + \nabla_{\mathbf{w}_1, m} C_{\text{head},m}(u,v) \} + \lambda_1 \mathbf{w}_1. \end{aligned} \quad (2)$$

where

$$\begin{aligned} \nabla_{\mathbf{w}_1, m} C_{\text{coref},m}(u,v) &= (v_1 \phi(u') - \Omega(u')) \phi(u''), \\ \nabla_{\mathbf{w}_1, m} C_{\text{head},m}(u,v) &= (v_1 \phi(u') - \Omega(u')) \phi(u') \Omega(u'). \end{aligned}$$

The partial subgradient w.r.t. mention head u for the coreference weight vector $\mathbf{w}_2$ is given by

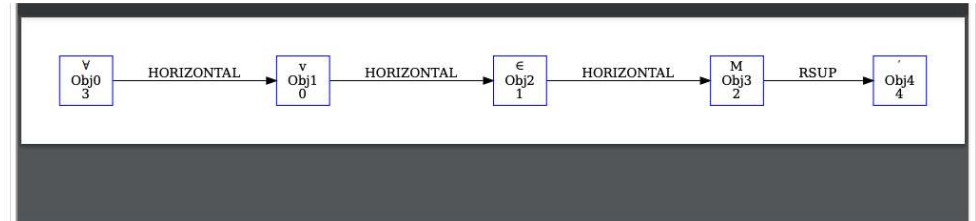
$$\begin{aligned} \nabla_{\mathbf{w}_2, u} \mathcal{L}(\mathbf{w}_1, \mathbf{w}_2) = \lambda_2 \mathbf{w}_2 + \\ \begin{cases} \phi(u', m) - \phi(u'', m) & \text{if } \Omega(u') = 1 \\ \phi(u', u) - \phi(u'', u) & \text{if } \Omega(u') = 0 \text{ and } \Omega(u'') = 1 \\ 0 & \text{if } \Omega(u') = 0 \text{ and } \Omega(u'') = 0 \end{cases} \quad (3) \end{aligned}$$

Here λ_1 and λ_2 are regularization coefficients which are tuned on the development set. To learn the mention head detection model, we consider two different parts of the gradient in expression (2). $\nabla_{\mathbf{w}_1, m} C_{\text{coref},m}(u,v)$ is exactly the local gradient of mention head m while we add $\nabla_{\mathbf{w}_1, m} C_{\text{head},m}(u,v)$ to represent

⁵More details can be found in (Peng et al., 2013). The difference here is that we also consider the validity of mention heads using $\Omega(u')$ and $\Omega(u'')$.

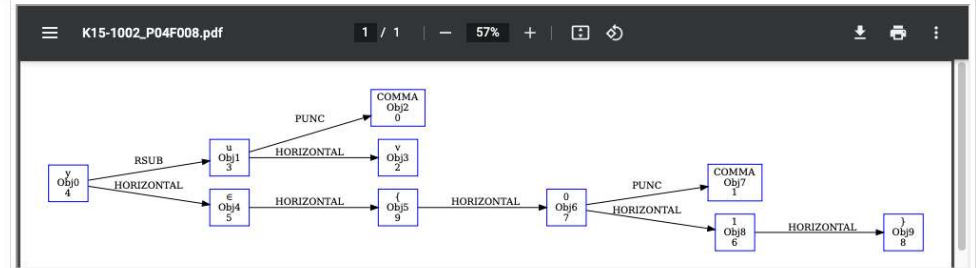
$$\forall v \in M'$$

$$\forall v \in M'$$



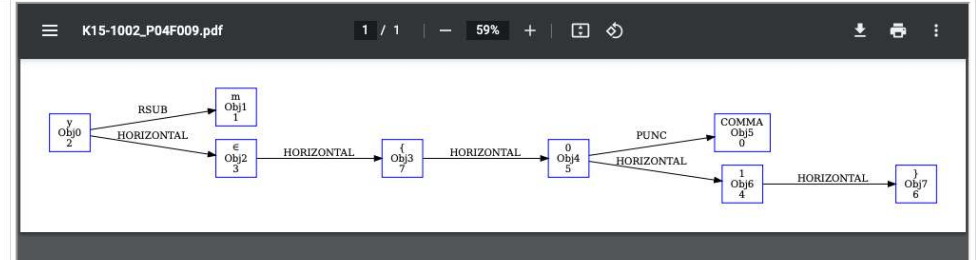
$$y_{u,v} \in \{0, 1\}$$

$$y_{u,v} \in \{0, 1\}$$



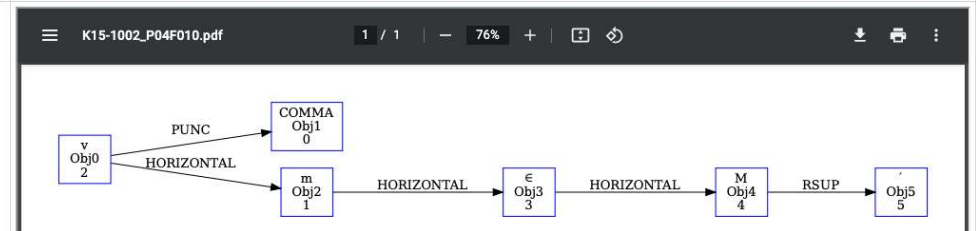
$$y_m \in \{0, 1\}$$

$$y_m \in \{0, 1\}$$



$$v, m \in M'$$

$$v, m \in M'$$



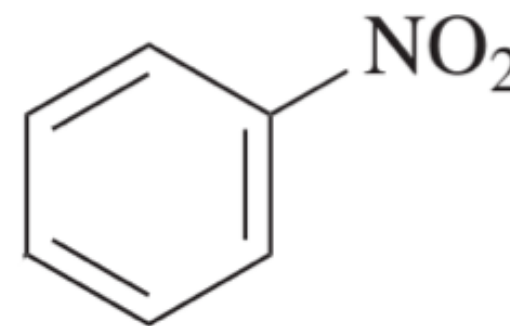
Results

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-]

Summary of Results:

- **Born-digital parser:** accurate recognition for PDF inputs
- **Visual parser:** competitive performance with far less training data and smaller model

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

Completed Work

Parsing models and techniques



Parsing mathematical formulas and molecular diagrams using vector information



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



Recurrent segmentation-aware parser for raster molecule images

Completed Work

Data generation and evaluation



Annotated ground truth generation for visual parsing of raster images



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



Utilized fine-grained graph-based evaluation metrics for chemical diagrams parsing

Conclusion

Completed Work

1. **Parsing from PDF:** no OCR, vectorization or GPU, geometrical and chemical constraints
2. **Interpretability:** visual correspondence of output symbols with the input primitives
3. **Ground truth generation** tool, generalizable to other visual parsing tasks
4. **Iterative segmentation aware** parsing for raster images
5. **Enhanced visualization** and utilized fine-grained **graph-based evaluation**

Conclusion

Proposed Work

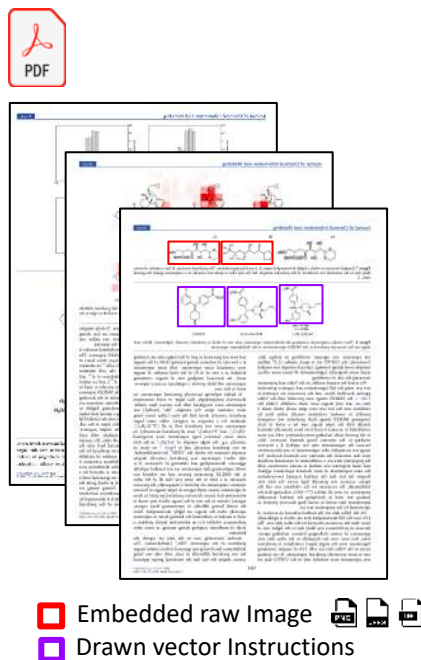
1. **Edge-specific attention** for improved context and structural relationships
2. **Multiple virtual global nodes** for enriched global context in graph-based parsers
3. **Iterative self-attention** to enhance multi-task learning (MTL) performance
4. **Robust over-segmentation** strategies and **synthetic noise augmentation** for parsing noisy raster images

Overview

Task: Parsing molecules from documents

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>
    .....
  </n>
</fragment>
```

SMILES

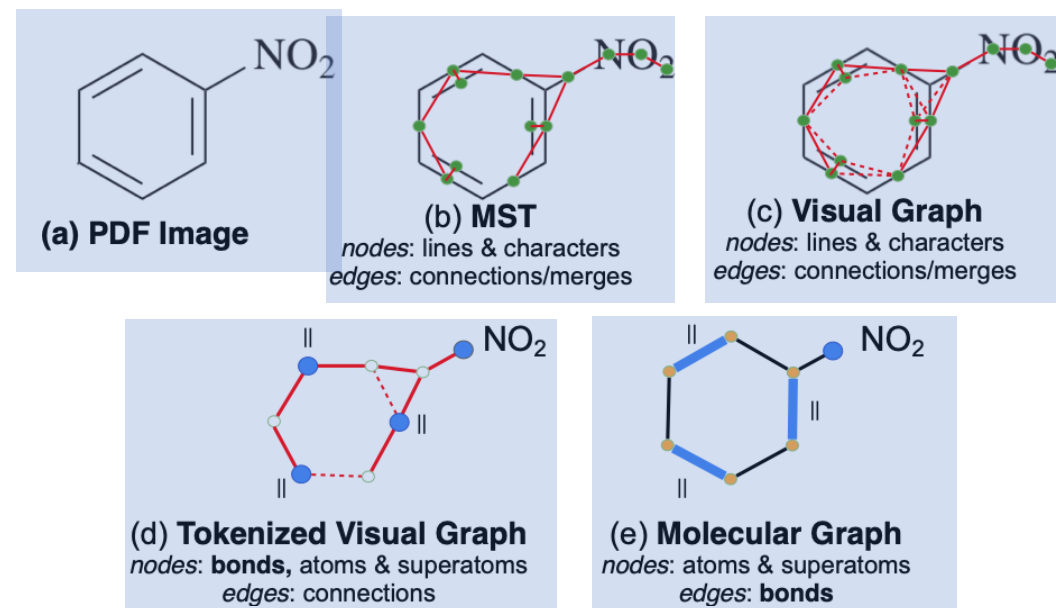


Export Options
(Per Molecule/Page/Document)

Born-digital Parser

Input: Born-Digital PDF Molecule Image

1. Extract Symbols from PDF
2. Build Minimum Spanning Tree (MST)
3. MST → Visual Graph
 - i. Detect negative charges (vs. other lines)
 - ii. Restructure MST: add and delete edges
 - iii. Tokenization: characters and parallel lines
4. Visual Graph → Molecular Graph
 - i. Convert line intersections into Carbons
 - ii. Replace bond nodes by edges
 - iii. Annotate names with subgraphs
 - iv. Generate CDXML



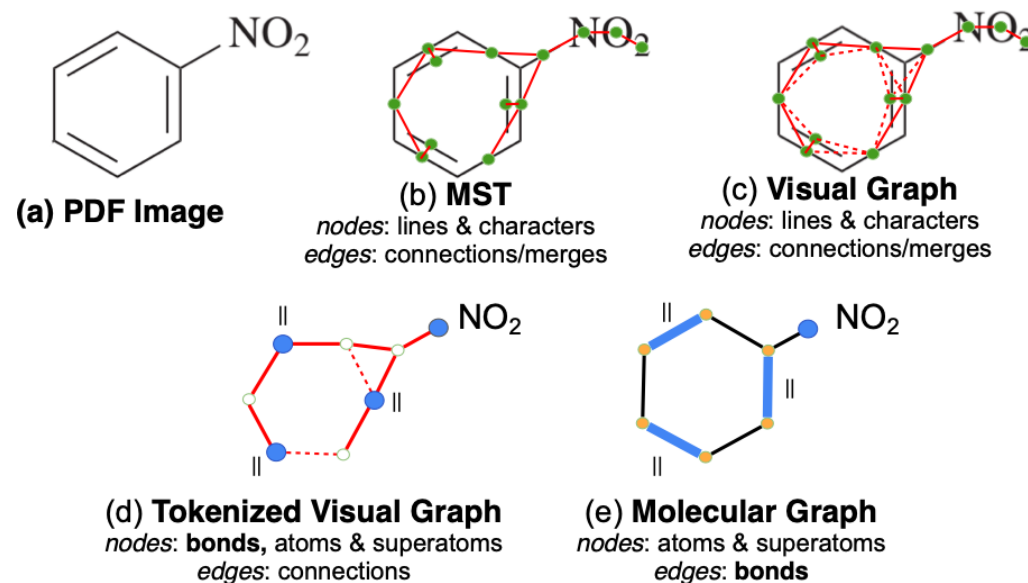
Parsing Nitrobenzene from a PDF image

Output: Editable molecular diagram (CDXML)

Born-digital Parser

Effectiveness of born-digital parser

- Does not require OCR, vectorization or GPU
- Uses both chemical and visual information
- Uses simple geometrical and chemical constraints
- Visual correspondence of output symbols with the input PDF
- Output CDXML directly editable in ChemDraw
- Easily converted to other formats (SMILES, MOL, InChI)

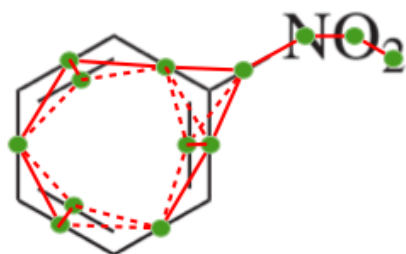


Parsing Nitrobenzene from a PDF image

Annotated Data Generation

Algorithm

1. **Convert** PDF molecule to raster image (PNG)
2. **Extract** visual line primitives
3. **Align** visual primitives with PDF primitives using maximum overlap
4. **Create** annotated ground truth data using visual graphs generated by born-digital parser
5. **Validate** output SMILES with ground truth

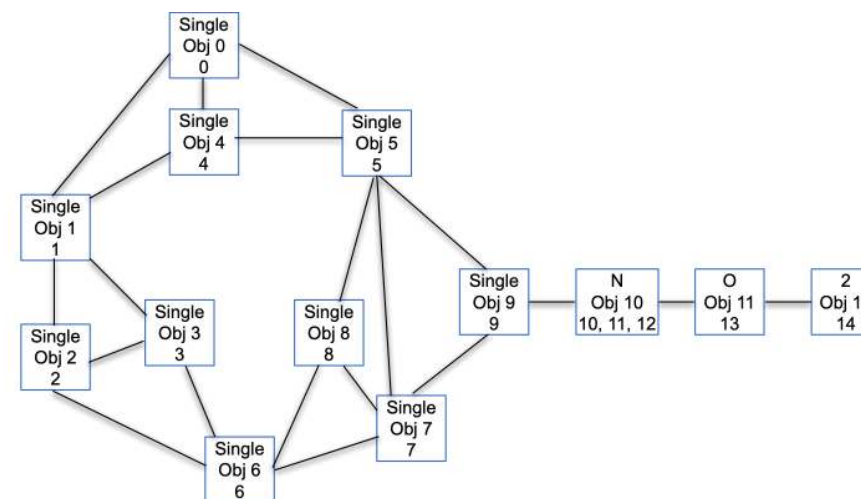


Visual graph generated
by born-digital parser

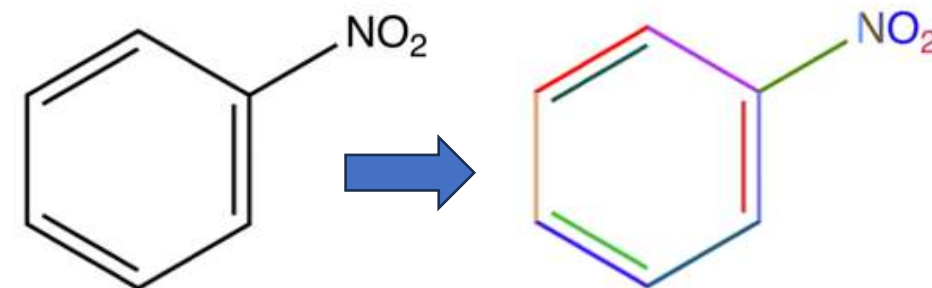


```
# [ OBJECTS ]  
# Objects (O): 10  
# Format: O, objId, class, 1.0, [primitiveId list]  
O, Obj0, Single, 1.0, 0  
O, Obj1, Single, 1.0, 1  
O, Obj10, N, 1.0, 10, 11, 12  
...  
  
# [ RELATIONSHIPS ]  
# Relationships (R): 11  
# Format: R, parentId, childId, class, 1.0 (weight)  
R, Obj0, Obj4, 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



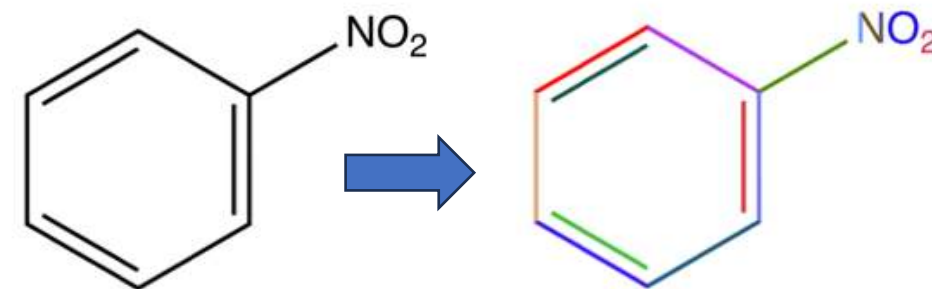
Extraction of visual primitives

Annotated ground truth data

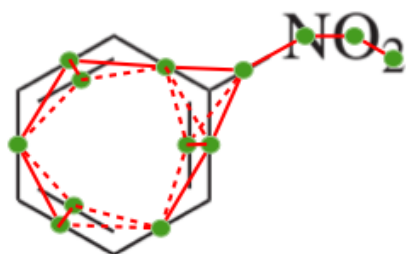
Annotated Data Generation

Advantages

1. **Efficiency:** Reduces time and effort for generating large datasets
2. **Consistency:** uniform and accurate annotations
3. **Scalability:** Easily scalable for extensive datasets
4. **Generalizability:** Easily generalizable to other visual parsing tasks



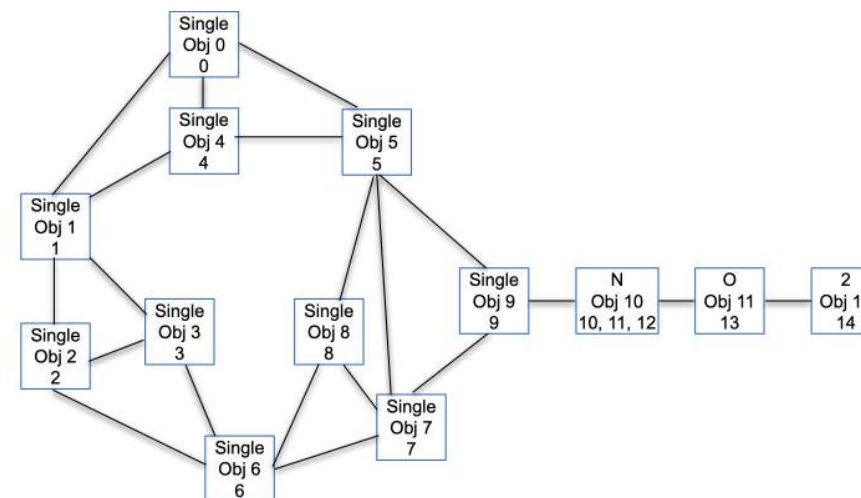
Extraction of visual primitives



Visual graph generated
by born-digital parser

```
# [ OBJECTS ]  
# Objects (O): 10  
# Format: O, objId, class, 1.0, [primitivelId 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

Annotated ground truth data

Results

Training data source: Pubchem 1 million

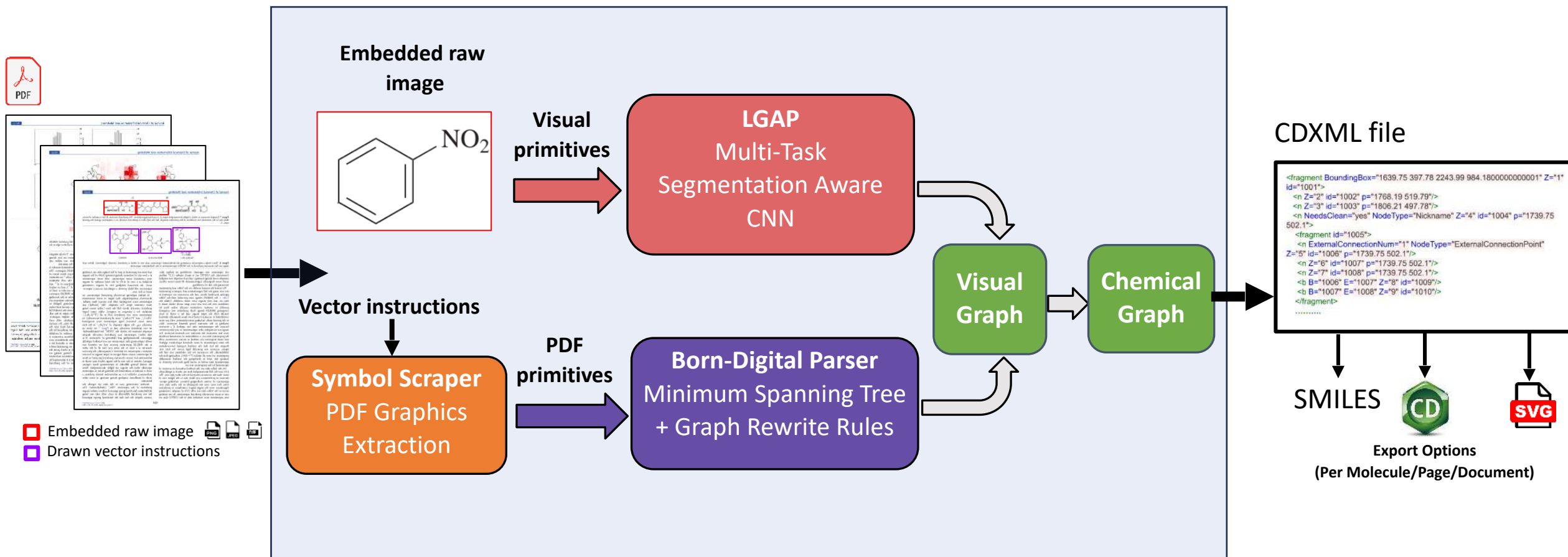
- **Born-digital:** 5,000 molecules
- **Visual:** 3416 molecules (validated from 5000)

Summary of Results:

- The born-digital parser has accurate recognition for PDF inputs
- The visual parser has competitive performance with far less training data

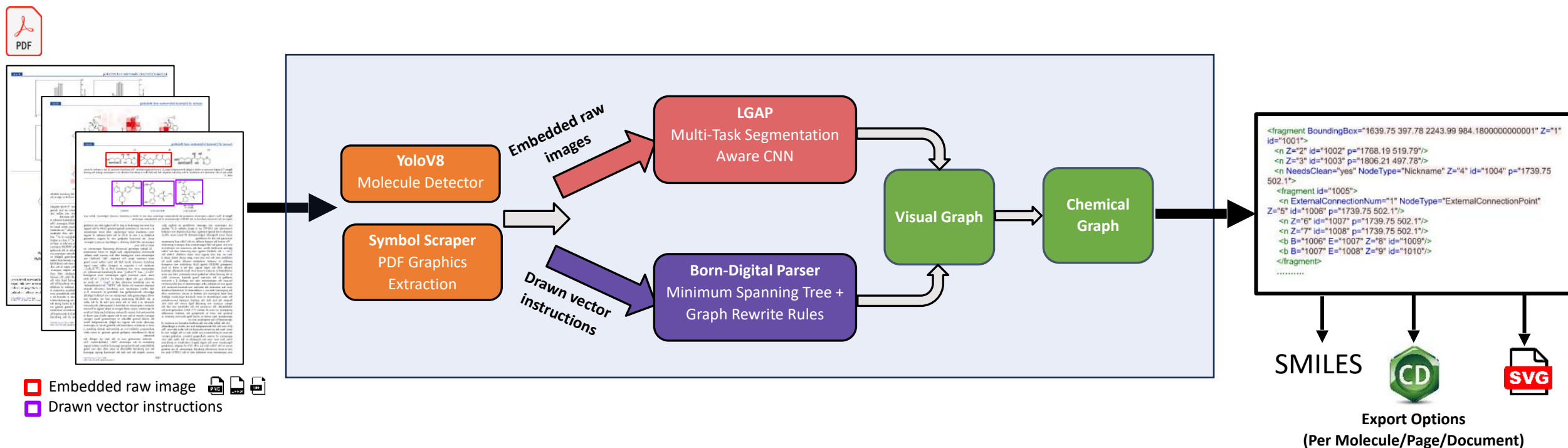
Exact SMILES Matches	Indigo (5719)	CLEF-2012 (992)	UoB (5740)
MolVec 0.9.7	95.40	83.80	80.60
OSRA 2.1	95.00	84.60	78.50
Img2Mol	58.90	48.84	78.18
DECIMER	69.60	62.70	88.20
SwinOCSR	74.00	30.00	44.90
MolScribe	97.50	88.90	87.90
MolGrapher	-	90.50	94.90
ChemScraper (Born-Digital – PDF input)	98.16	89.32	94.41
ChemScraper (Born-Digital – PDF input) * Skipping rendering errors	98.42	96.20	94.41
ChemScraper (Visual – PNG input)	85.02	-	-

Overview



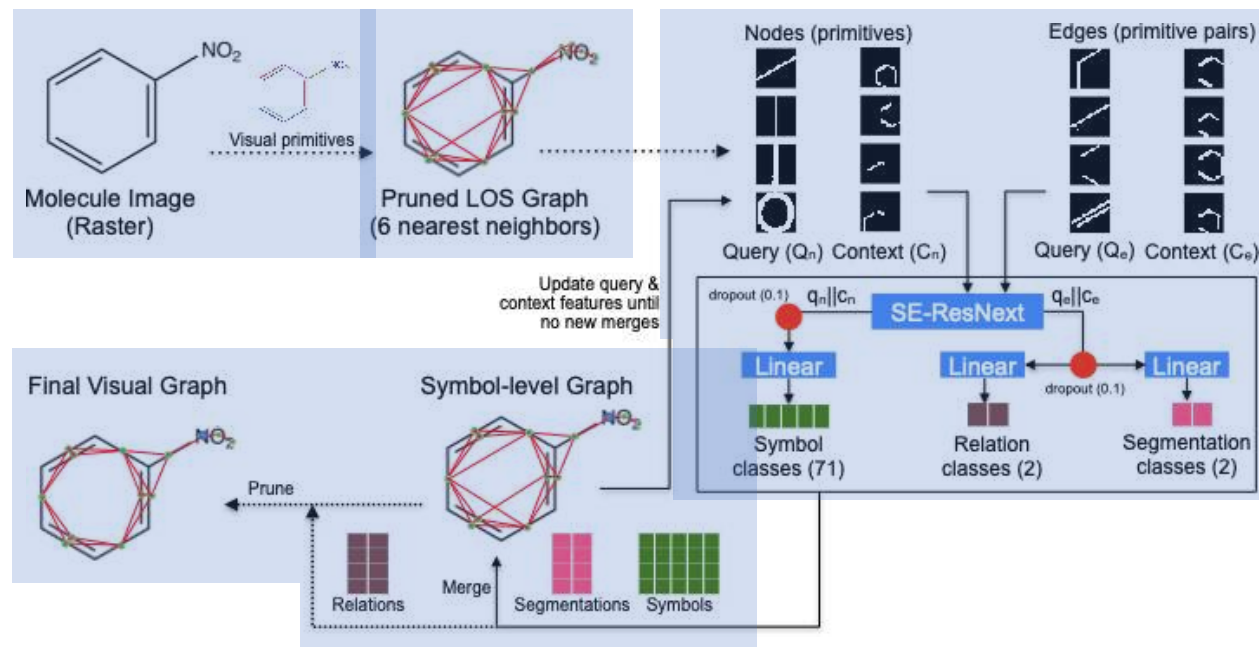
Overview

1. **YOLOv8 Detector:** detect molecule regions
2. **Symbol Scraper:** extract character and graphics from PDF encodings
3. **Graph Construction and Transformations:**
 - **Born-digital parser:** extracts visual graph from PDF symbols using minimum spanning tree and graph rules
 - **Visual parser (LGAP):** multi-task CNN extracts visual graph from image primitives
4. **Chemical graph generation**



Visual Parser

1. **Input:** Raster image with extracted line primitives
2. **Pruned LOS Graph:** LOS graph over visual primitives pruned to 6 nearest neighbors
3. **SE-ResNext Backbone:** Processes query and context images
4. **Classification:** MLPs for
 1. Symbols (71 classes),
 2. Segmentation (MERGE and NOT_MERGE),
 3. Relationships (CONNECTED and NOT_CONNECTED)
5. **Symbol-label Graph:** Merge primitives
6. **Output:** Final visual graph with labeled nodes (visual symbols) and edges (connections)



Visual Parser

Novel Techniques in Visual Parser

1. **Pruned LOS Graph:** Efficiently captures spatial relationships, reducing complexity and improving accuracy.
2. **Discrete Attention:** Updates query and context images based on predicted segmentation
3. **Training** on annotated data generated by born-digital parser

EXTRA EXTRA SLIDES

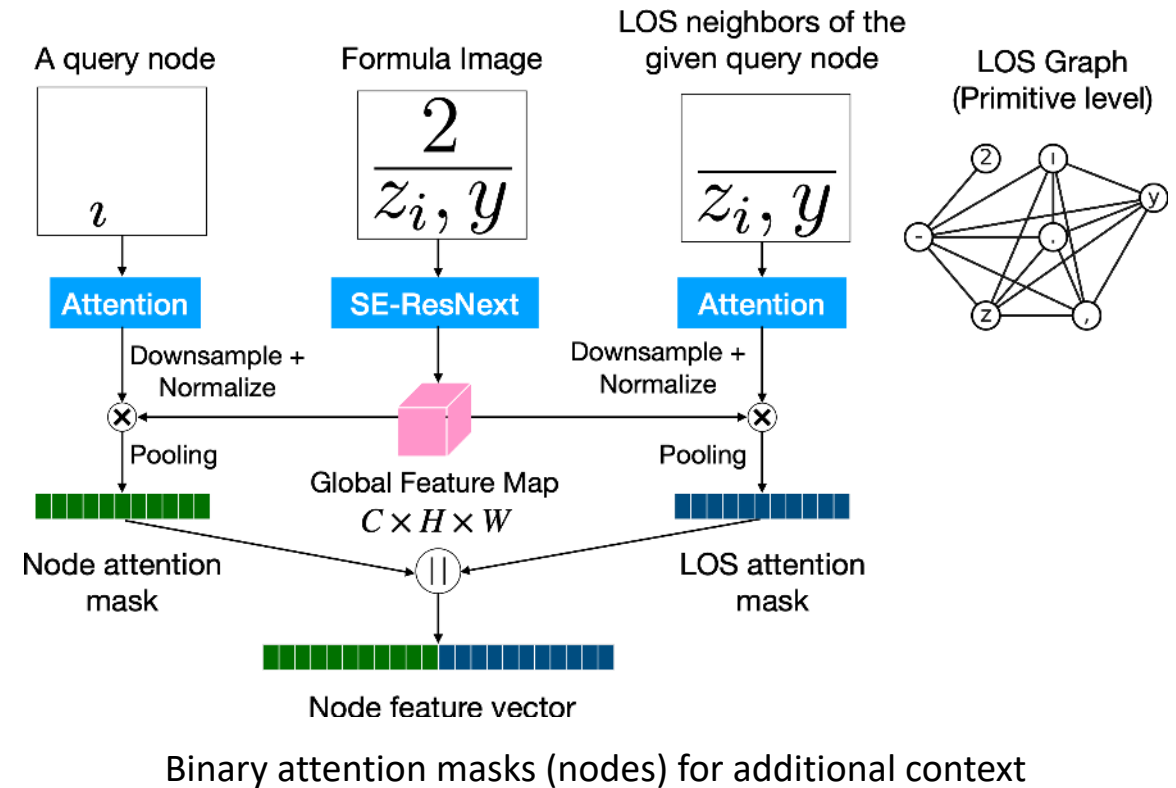
LOS Context Masks

LOS Context Masks

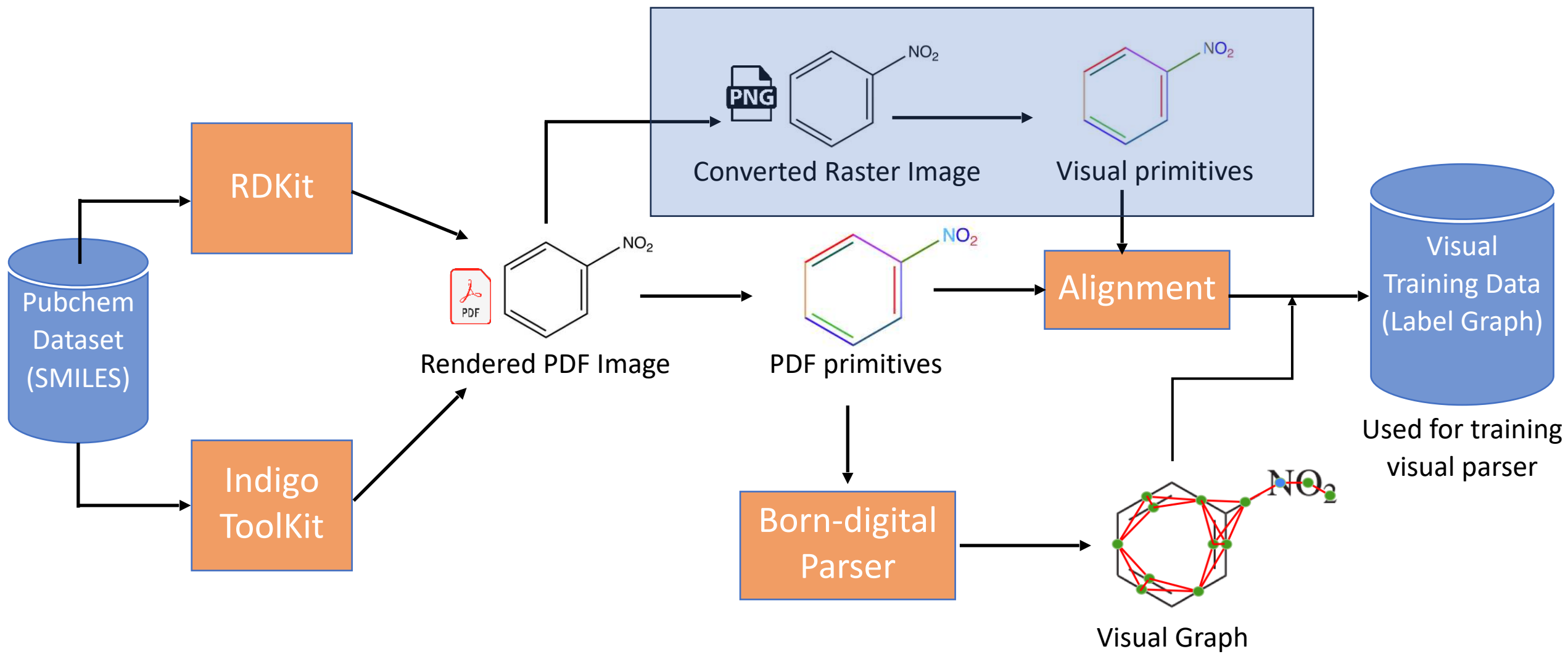
- **Goal:** using additional context from line-of-sight graph neighbors in visual features
- Concatenate LOS attention masks to node attention masks to form the node feature vector
- **2.49% improvement** in formula recognition rate

Punctuation Representation

- **Goal:** Representing punctuation relationships more consistently in ground truth
- **1.23% improvement** in formula recognition rate



Annotated Data Generation for Visual Parser



Born-Digital Parser

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

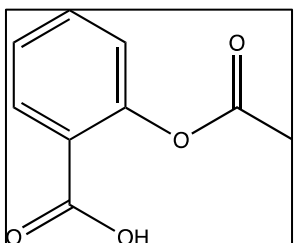
Conclusion

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

Motivation

- Enhancement of Chemical Informatics
- Advancement in Drug Discovery and Chemical Synthesis
- Development of Graphics-Aware Search Engines
- Support for Academic Research



Which catalysts can improve the yield of of this chemical?



Let's find reactions producing this molecule.



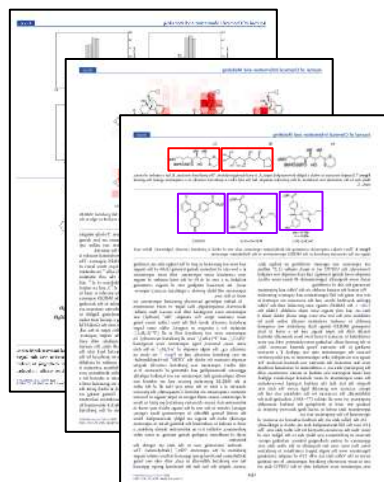
Overview

Task: Detecting and parsing molecules from documents

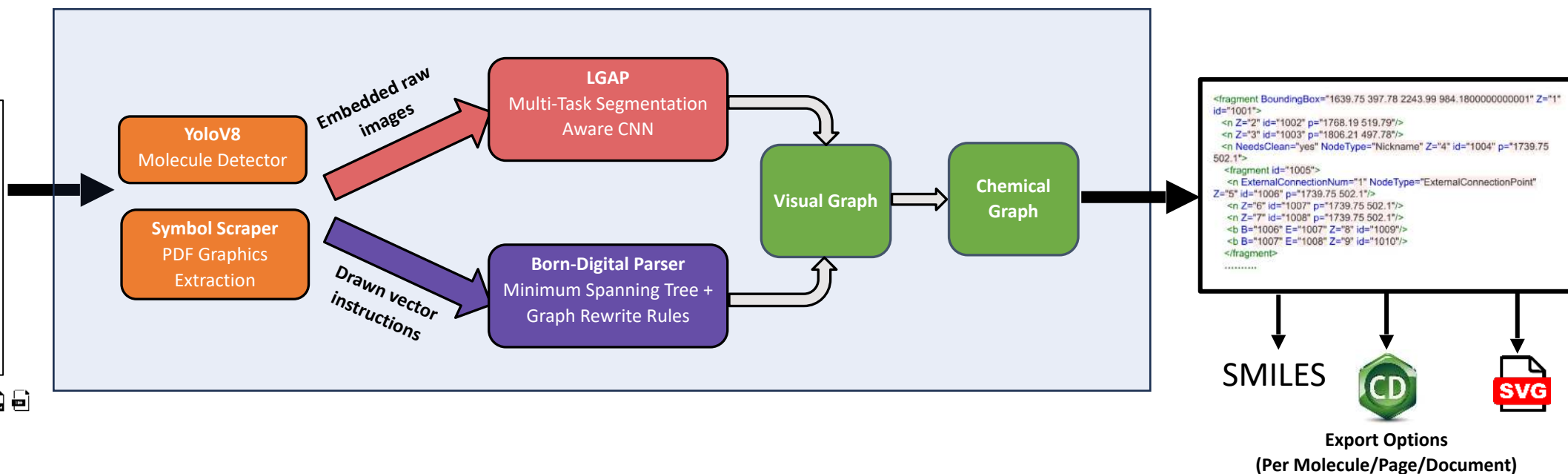
- **Input:** A scientific paper (PDF)

- Embedded raw images
- Drawn vector instructions

Output: All molecule CDXMLs/SMILES

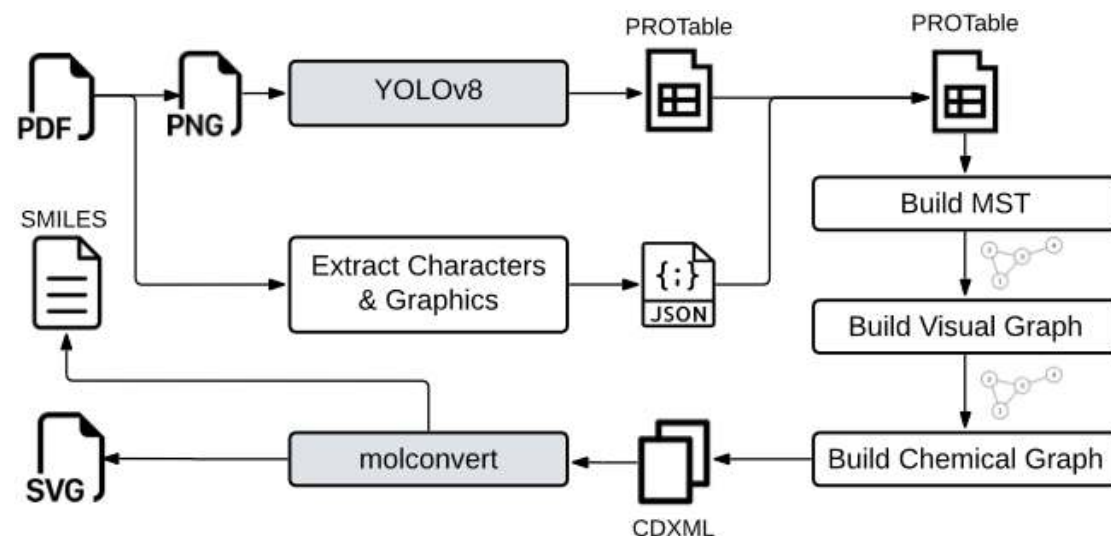


- Embedded raw image
- Drawn vector instructions

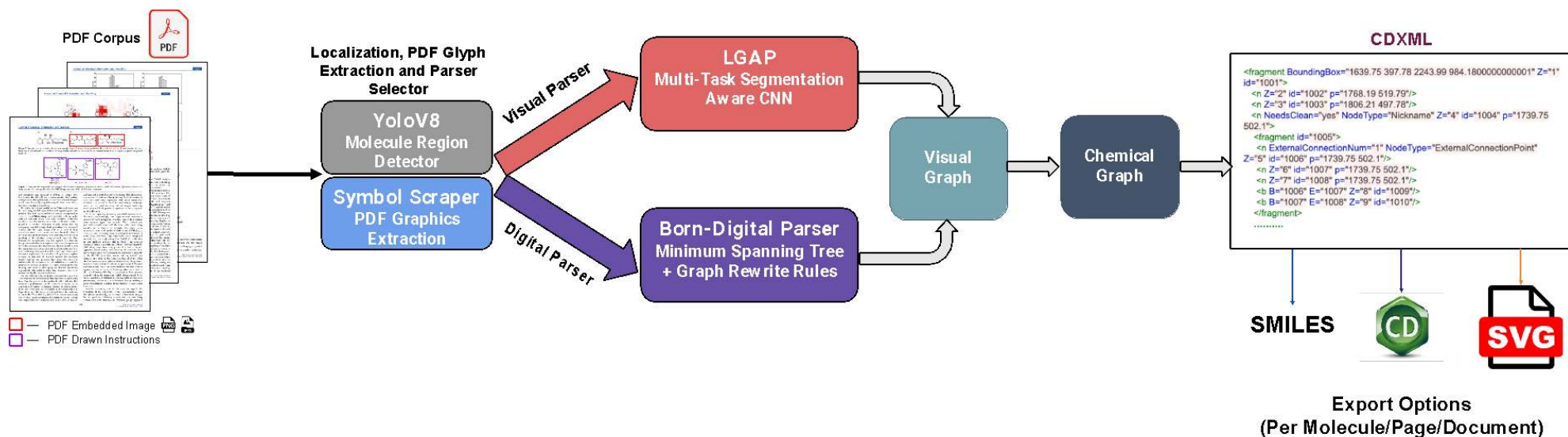


Born-digital Parser

1. **YOLOv8 Detector:** detect molecule regions
2. **Symbol Scraper:** extract character and graphics from PDF encodings
3. **Graph Construction and Transformations:** PDF primitives -> MST -> Visual graph
4. **Chemical graph generation**
5. **CDXML generation**

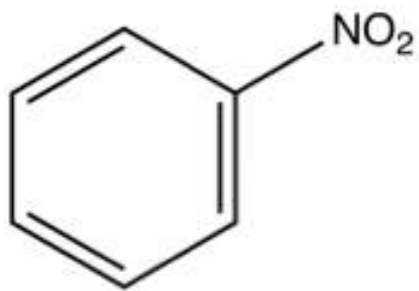


1. **YOLOv8 Detector:** detect molecule regions
2. **Symbol Scraper:** extract character and graphics from PDF encodings
3. **Graph Construction and Transformations:**
 - **Born-digital parser:** extracts visual graph from PDF symbols
 - **Visual parser:** multi-task CNN (LGAP) extracts visual graph from Image primitives
4. **Chemical graph generation**

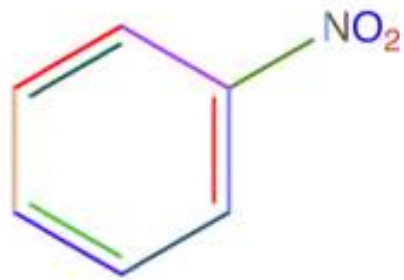


[1] Shah, A. K., Amador, B. M., Dey, A., Creekmore, M., Ocampo, B., Denmark, S., & Zanibbi, R. (2024). *ChemScaper: Leveraging PDF Graphics Instructions for Molecular Diagram Parsing* (arXiv:2311.12161). arXiv. <https://doi.org/10.48550/arXiv.2311.12161>, accepted at ICDAR 2024

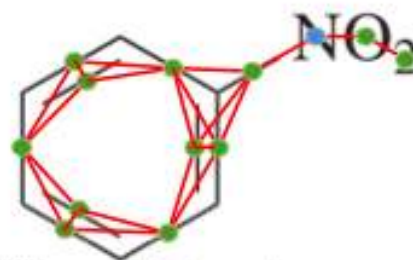
[2] Shah, A. K., & Zanibbi, R. (2023). Line-of-Sight with Graph Attention Parser (LGAP) for Math Formulas. In G. A. Fink, R. Jain, K. Kise, & R. Zanibbi (Eds.), *Document Analysis and Recognition—ICDAR 2023* (pp. 401–419). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41734-4_25



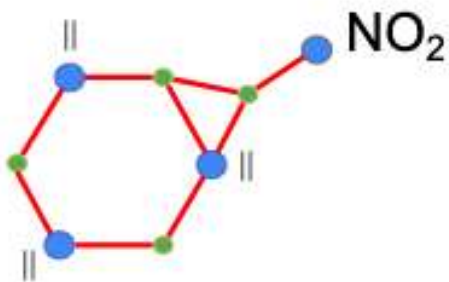
(a) Raster Image



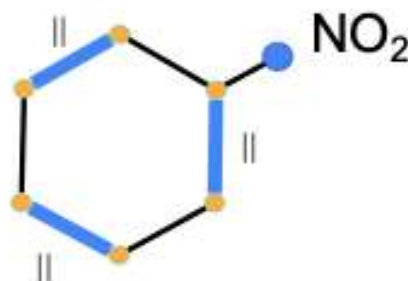
(b) Visual primitives
nodes: lines



(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**

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

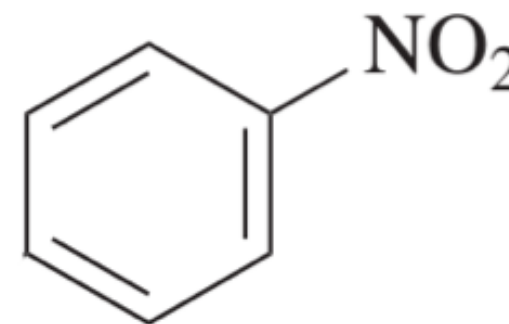


Ground Truth

Predicted

String-based Metrics

- Exact SMILES match (chemistry)



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

Conclusion and Future Work

- Introduced a simple, fast born-digital parser for parsing molecules from vector graphics
- Automated the ground truth data generation process
- Created a competitive visual parser for parsing molecules from raster images trained on data generated by born-digital parser
- Output CDXML directly editable in ChemDraw and easily converted to other formats

Future Work

- Exploring GNNs, graph rewriting systems, encoder-decoders for born-digital parsing
- Applying graph-based evaluation methods to other graphics including math and tables
- Applying the parser to index molecules and other graphics in PDF collections for graphics-aware search applications

Results: Molecule Diagram Recognition

Training data source: Pubchem 1 million

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

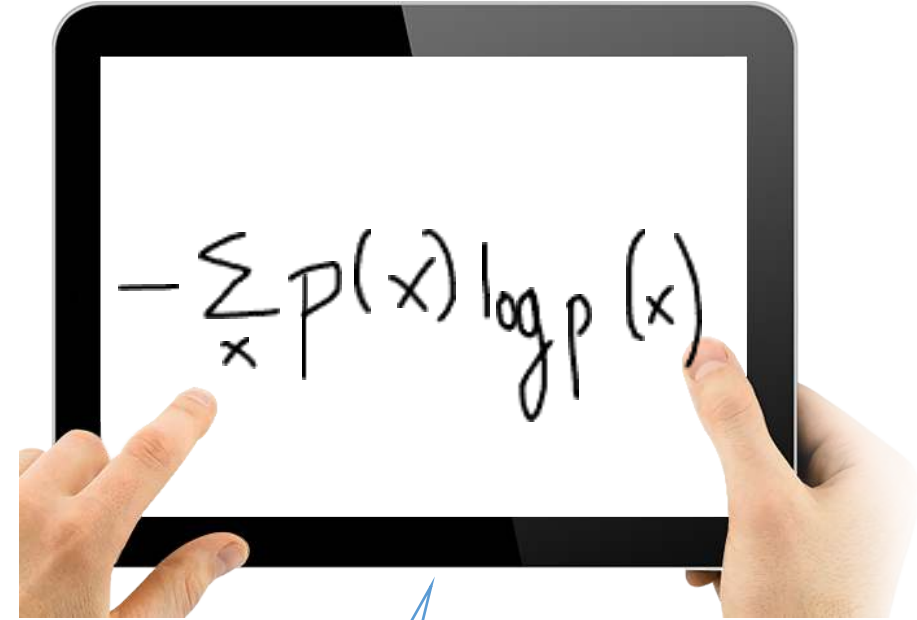
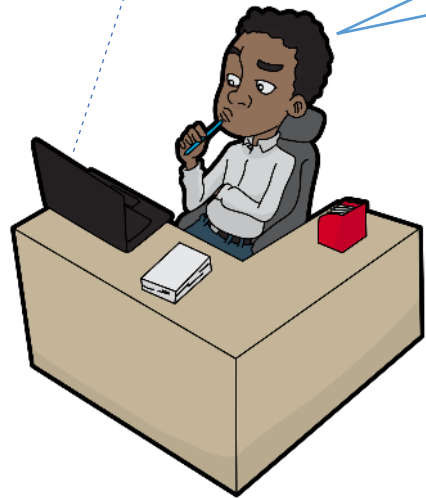
Additional Metrics:

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

Motivation

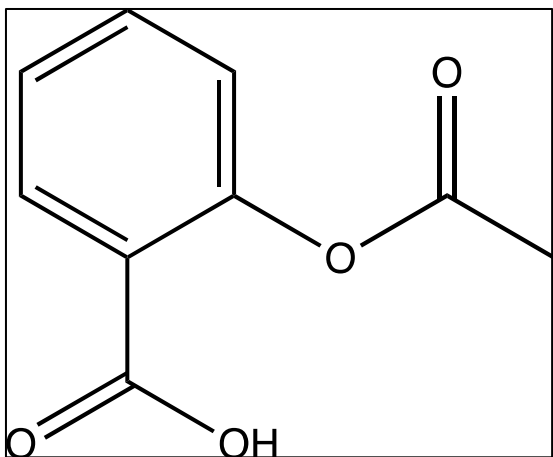
$$-\sum_x p(x) \log p(x)$$

What is the name
of this formula?

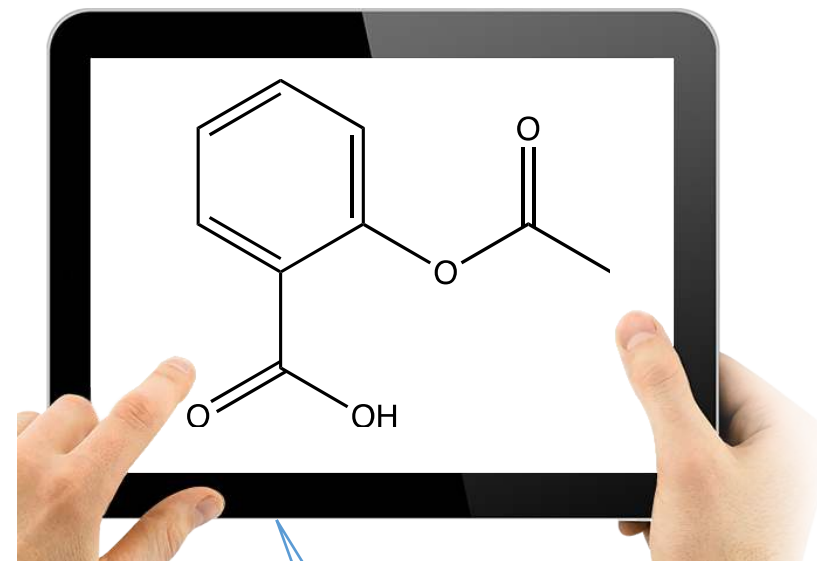
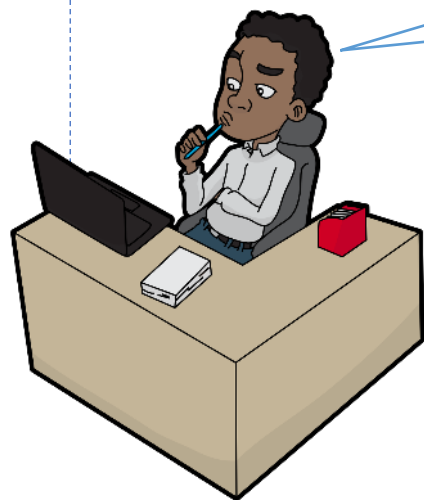


Let's find documents
describing this formula.

Motivation



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



Let's find reactions producing this molecule.