SARANGA: Structured ArXiv Representation, Automatic Network Generation and Analysis

Yaron Otmazgin • Or Zuk

Department of Statistics and Data Science, Hebrew University of Jerusalem

1 Abstract & Mid-Level Schema

Mathematical knowledge lives in unstructured LaTeX. SARANGA turns 100,000 arXiv papers into a typed knowledge graph - with a rule-based extractor that runs in seconds per paper and zero API cost.

We propose a *mid-level* representation that bridges raw LaTeX and formal logic: a typed directed acyclic graph whose nodes are mathematical objects (definitions, theorems, proofs, algorithms, ...) and whose edges are typed dependencies (USES, GENERALIZES, PROVES, ...). A rule-based extractor, distilled from LLM teachers, scales to 100K papers at **zero API cost** and ∼**3 s/paper**.

Nodes (16 types): definition, theorem, lemma, corollary, proposition, conjecture, claim, fact, observation, remark, problem, example, algorithm, notation, proof, sub-block.

Edges (7 types): USES, GENERALIZES, SPECIALIZES, PROVES, INSTANTIATES, CONTAINS, REFERENCES.

Theorem-like nodes carry separated *assumption/conclusion* fields. An *origin* field distinguishes paper-original content from cited material.

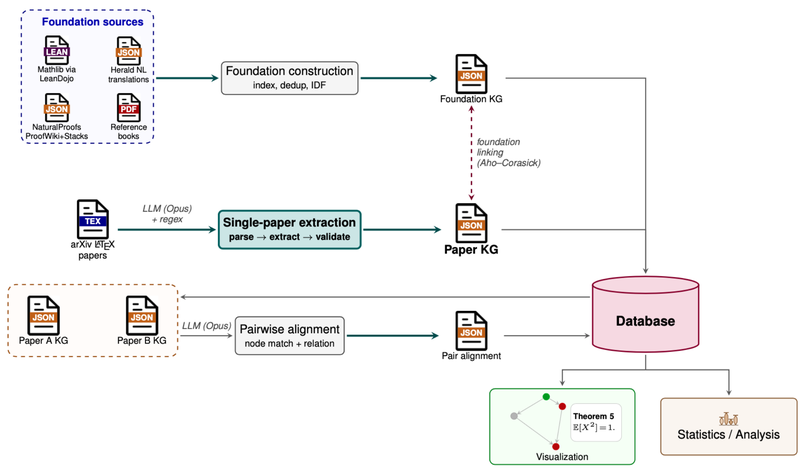
Foundation graph: ∼100 textbooks parsed (Lean Mathlib, NaturalProofs, encyclopedic references) → ∼298K entries / ∼572K relations; provides the shared vocabulary against which per-paper graphs are anchored and aligned.

2 Three Production Tiers

T1 - gold (10 papers, manual): ground truth, ∼330 nodes.
T2 - Opus (250 papers, Opus 4.6): high-quality extraction, ∼8.1K nodes.
T3 - rules (100K papers, regex): corpus scale, ∼10M nodes, zero API cost.

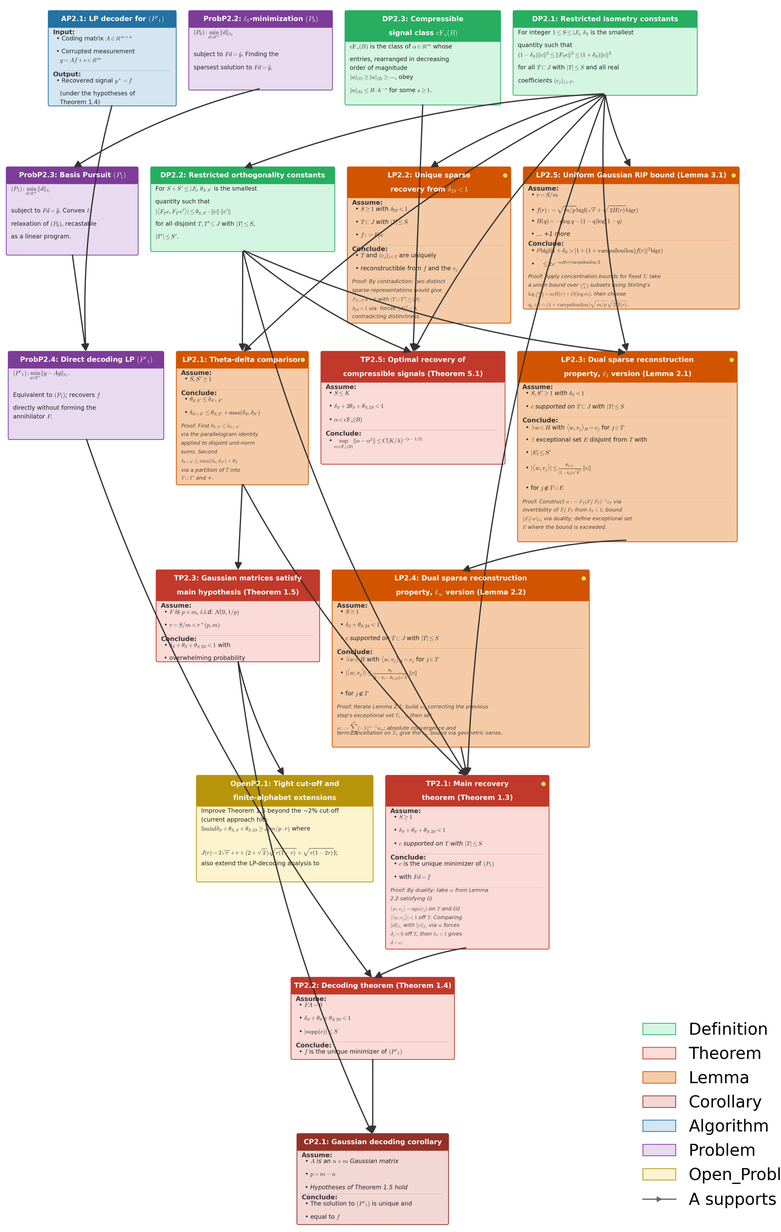
Tier	Method	Papers	Role
T1	manual annotation	10	ground truth
T2	Opus 4.6	250	high-quality silver
T3	regex (NtRED)	97,997	corpus scale

3 End-to-End Pipeline



SARANGA system overview. Foundation construction (Lean Mathlib, NaturalProofs, encyclopedic refs) merges with per-paper MONO extraction; outputs feed STEREO alignment and the interactive portal.

4 Worked Example: Candès & Tao (2005)



Candès & Tao, *Decoding by Linear Programming*. 18 nodes / 24 edges; recovered DAG with assumption/conclusion split and proof-step decomposition.

Across seven Opus prompt versions on this paper, soft F_1 ranges 0.15–0.40; edge $F_1 \leq 0.18$, locating dependency-edge recall as the next target.

5 Per-Node-Type Extraction

Per-paper macro F_1 on the 10-paper T1 gold set. Rules match LLMs on structured environments (lemma, corollary, conjecture); definitions, which often lack `\begin{definition}` markers, are hardest for all providers.

Greedy match on token-set Jaccard with threshold $\theta = 0.7$:

$$J(s, s') = \frac{|t(s) \cap t(s')|}{|t(s) \cup t(s')|} \geq \theta$$

Macro F_1 averages per-type F_1 over the 8 non-empty target types: $\overline{F_1} = \frac{1}{|T|} \Sigma_{t \in T} F_1^{(t)}$.

Node type	Opus 4.6	GPT-4o	Rules (T3)
Definition	0.61	0.39	0.52
Theorem	0.86	0.83	0.86
Lemma	1.00	0.55	1.00
Proposition	1.00	0.50	0.75
Corollary	1.00	0.75	1.00
Claim	1.00	0.67	1.00
Conjecture	1.00	1.00	1.00
Proof	0.99	0.59	0.70
Mean (8 types)	**0.93**	0.66	**0.85**

6 NtRED Distillation

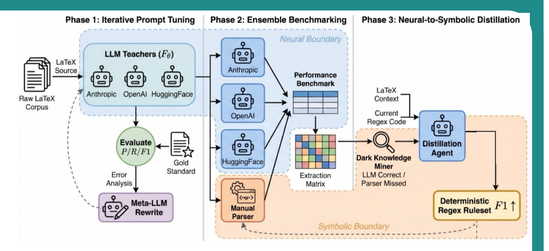
NtRED (Neural-to-Rule Ensemble Distillation). An LLM teacher ensemble proposes regex amendments to a rule student via an iterated dark-knowledge loop, each precision-reviewed before merging. Conclusion-separation rose from 24% to 93%; DAG acyclicity reached 99.6% - all at zero API cost at inference time.

Dark knowledge at iteration t - gold instances the student misses but ≥ 1 teacher recovers:

$$\mathcal{D}_{\text{dark}}^{(t)} = \{ s \in \mathcal{G} \mid \underbrace{\hat{y}_{\mathcal{R}^{(t)}}(s) = 0}_{\text{student FN}} \wedge \underbrace{\exists p \in \mathcal{P} : \hat{y}_p(s) = 1}_{\text{teacher TP}} \}$$

Discovery mode (no gold) via multi-teacher consensus: $\text{consensus}(s) = \mathbf{1}[\Sigma_p \hat{y}_p(s) \geq \tau]$, τ=2.

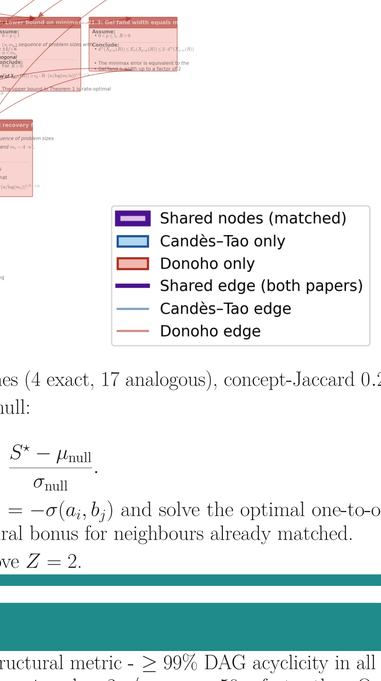
Rule update. Inspector LLM mines patterns; loop until F_1 plateaus: $\mathcal{R}^{(t+1)} = \mathcal{R}^{(t)} \cup \text{PatternMine}(\mathcal{D}_{\text{dark}}^{(t)})$.



Phase I prompt tuning → Phase II ensemble benchmark → Phase III neural-to-symbolic distillation. Conclusion separation 24% → 93%; DAG acyclicity → 99.6%; ***zero API at inference****.*

7 Per-Category Statistics & Graph Complexity

Category	#papers	nodes/p	edges/p	depth	% acyc	% concl	% anom	found/p
Pure mathematics (13 categories)								
math.LO	4,946	102.8	314.4	3.8	≥99	92	44	18.7
math.AT	4,985	165.8	448.0	4.3	≥99	92	53	19.7
math.CO	4,963	78.8	170.5	3.5	≥99	92	47	12.3
math.NT	4,954	128.8	171.9	3.5	≥99	90	69	16.1
math.AG	4,992	140.5	298.7	4.1	≥99	92	56	17.2
math.GR	4,982	103.3	353.2	4.1	≥99	93	38	21.5
math.RT	4,995	155.1	263.3	3.9	≥99	92	53	16.8
math.DG	4,966	130.4	240.4	3.9	≥99	89	69	14.6
math.AP	4,983	171.5	184.4	3.8	≥99	81	90	10.2
math.CA	4,929	123.3	139.5	3.3	≥99	85	84	14.0
math.FA	4,974	120.1	214.5	3.8	≥99	87	66	17.9
math.PR	4,964	145.3	184.8	3.8	≥99	88	75	14.7
math-ph	4,906	90.5	77.8	2.1	≥99	85	91	8.6
Theoretical CS (4 categories)								
cs.CC	4,847	86.9	216.0	3.4	≥99	89	42	13.5
cs.DS	4,781	58.4	129.3	2.8	≥99	85	50	7.1
cs.IT	4,660	63.9	67.5	2.5	≥99	82	80	5.9
cs.LG	4,602	36.1	32.9	1.5	≥99	75	81	4.3
Statistics & numerics (3 categories)								
math.OC	4,780	86.3	100.3	3.2	≥99	80	72	10.0
math.NA	4,882	91.0	78.3	2.8	≥99	75	76	7.5
stat.TH	4,906	114.3	125.4	3.4	≥99	86	76	12.0
Total / mean	**97,997**	**110.3**	**192.1**	**3.4**	**≥99**	**86**	**66**	**13.2**



Node/edge counts and DAG depth scale together (Pearson r=0.78 across 20 categories), with pure-math categories like `math.AT`/`math.GR` forming the deepest proof chains and applied tracks (`cs.LG`, `math-ph`) the flattest. DAG acyclicity exceeds 99% in every category, and `found/p` (foundation edges with IDF ≤10%) tracks topical specificity.

8 Cross-Paper Alignment

Cross-paper alignment. Candès–Tao ↔ Donoho on compressed sensing: 21 matches (4 exact, 17 analogous), concept-Jaccard 0.26.
Type-compatibility-weighted node similarity; significance via permutation null:

$$\sigma(a, b) = J(\phi_A(a),\, \phi_B(b)) \cdot K_{\tau(a),\tau(b)}, \qquad Z = \frac{S^\star - \mu_{\text{null}}}{\sigma_{\text{null}}}.$$

Hungarian assignment [Kuhn, 1955]: form cost matrix $C \in \mathbb{R}^{|V_A| \times |V_B|}$ with $C_{ij} = -\sigma(a_i, b_j)$ and solve the optimal one-to-one node matching $M^\star = \arg\min_M \Sigma_{(i,j) \in M} C_{ij}$ in $O(n^3)$; an iterative pass adds a structural bonus for neighbours already matched. Combined Jaccard widens the positive–negative gap to 1.20σ with zero negatives above $Z = 2$.

9 Summary & Future Work

Headline result. Post-fix T3 (rule-based) matches T2 (Opus 4.6) on the headline structural metric - ≥ 99% DAG acyclicity in all 20 categories and 92% conclusion separation on the 97,997-paper corpus - at **zero API cost** and ∼ 3 s/paper, ∼ 50× faster than Opus 4.6 (∼ 154 s/paper) and ∼ 34× faster than Sonnet 4 (∼ 102 s/paper). Running the same corpus through a frontier LLM would cost six figures of USD, motivating NtRED's neural-to-symbolic distillation. NtRED revisions raised conclusion separation from 24% to 93% on the gold set and a recent FP audit projects the rule false-positive rate from ∼ 48% to ∼ 6% on the audit sample.

Acyclicity is structural, not semantic. Edge recall is the next target: even the strongest extractor reaches only soft $F_1 \sim 0.18$ for dependency edges on the Sinkhorn case. T2 currently covers a single LLM family (Opus 4.6), and T1 has 10 papers from one annotator without inter-annotator agreement measured.

Released. 100K typed-DAG paper graphs (T1+T2+T3), ∼10M nodes, foundation KG (∼298K entries, ∼572K relations), 125-pair alignment benchmark, two node-level benchmarks with reference baselines, and an interactive portal - all without authentication.

Future work.

(i) Scale T2 toward the full T3 corpus and run further distillation passes against the residual ∼ 6% FP rate.

(ii) Expand the foundation KG beyond the current ∼100 textbooks.

(iii) Replace Jaccard σ with semantic/structural similarity to recover the four disjoint-vocabulary positive pairs (BRT/vdG, CT/BRT, CT/Donoho, SS/BSS) where $Z(\text{G}) < -1$.

(iv) **Disentangle structural complexity from writing style:** test whether graph-density metrics predict paper difficulty independently of prose verbosity - e.g., regress density against reader-time or citation-impact controls, and compare same-result papers (surveys vs. original proofs) to detect whether denser graphs reflect harder mathematics or just denser prose.

(v) Leverage SARANGA as the underlying infrastructure for downstream tasks - notably **open-problem discovery** (mining unresolved conjectures and dangling assumptions across the corpus) and **plagiarism / near-duplicate detection** (cross-paper structural matching at scale).

10 References

1. G. Hinton, O. Vinyals & J. Dean. **Distilling the knowledge in a neural network.** NeurIPS Deep Learning Workshop, 2015. (Classical neural-to-neural distillation; contrast for NtRED.)
2. E. Candès & T. Tao. **Decoding by linear programming.** IEEE Trans. Inf. Theory, 2005. (Worked example, Ch. 4.)
3. D. Donoho. **Compressed sensing.** IEEE Trans. Inf. Theory, 2006. (Cross-paper alignment partner, Ch. 8.)
4. H. W. Kuhn. **The Hungarian method for the assignment problem.** Naval Research Logistics Quarterly, 1955. (Node-alignment solver, Ch. 8.)
5. S. Welleck *et al.* **NaturalProofs: Mathematical theorem proving in natural language.** NeurIPS, 2021. (Closest NL theorem corpus.)
6. The Mathlib Community. **Mathlib: The Lean Mathematical Library**, 2024. (Foundation KG source.)

11 Portal

Browse SARANGA online. Every extracted typed-DAG graph for the 97,997-paper T3 corpus, the 250 T2 Opus graphs, and the 10 T1 hand-curated graphs is interactive: pan, zoom, inspect node text, follow typed dependency edges, and toggle assumption/conclusion overlays.

huggingface.co/spaces/RHCPP/saranga