Towards Theorem Landscapes Beyond Citation Graphs: Charting Theorems, Navigating Fields

Xiangnan Wei[1] and Pipi Hu*[1,2]

[1]MathonAI Team
[2]BIMSA

Abstract

Mathematical theorems form a vast web of dependencies, but we lack a map of it. Citation graphs are discrete, sparse, and fragmented across subfields; theorem provers based on large language models are powerful local search operators but have no global picture of where theorems sit relative to each other. Without such a picture, a prover can miss relevant tools from distant fields, and a mathematician cannot easily assess whether the infrastructure for attacking a conjecture already exists. We construct a continuous space over mathematics—the *Theorem Landscape*—by embedding a 10M-scale corpus containing 11.42M theorem-statement rows into a latent space where distance reflects mathematical relatedness. The key idea is that once theorems have coordinates, their relationships become measurable: nearby points can suggest related statements or extracted dependency links, clusters correspond to subfields, and sparse regions can motivate inspection rather than certify knowledge gaps. The space is built in two main stages: masked language model pre-training on theorem text, followed by theorem-level contrastive fine-tuning on proof-core dependency edges. Supporting analyses on the fixed 1.1M-theorem evaluation corpus suggest three navigational uses. **(1) Proof tool retrieval**: in the current strict Stage 2b rerun, nearest-neighbor search over one million candidates retrieves theorems from cited papers at 38.7% Recall@100, compared with 34.9% for Stage 1 and 31.7% for the random-edge control. **(2) Conjecture assessment**: for famous conjectures, the neighborhood shows plausible local toolkits—the Four Color Theorem sits among graph coloring tools, while the Collatz Conjecture is relatively sparse. (On 4,481 typical arXiv conjectures this density signal vanishes, $p = 0.41$; the pattern is not a general provability classifier.) **(3) Cross-domain bridge discovery**: current Stage 2b category-level bridge reruns surface low-citation category pairs such as complex variables–spectral theory and K-theory–symplectic geometry; these are navigation hints pending systematic expert validation. In the strict proof-core split, theorem-level fine-tuning achieves 0.9908 AUC on held-out theorem-to-theorem links against random negatives, raises paper-citation AUC from 0.8097 to 0.8921, and increases the conjecture-neighborhood density gap from 0.0617 to 0.1086. Stage 1 already reaches 0.9227 theorem-link AUC under this easy random-negative protocol, so we report hard-negative theorem-link tests separately. A shuffled-edge negative control collapses the density gap to 0.0097, showing that the effect is not a generic consequence of contrastive training. An all-edge ablation improves broad domain and citation metrics, but weakens the density signal, so we treat proof-core supervision as the main result and all-edge supervision as an ablation.

1 Introduction

The problem. Theorems do not exist in isolation. They build on earlier results, borrow techniques across disciplines, and form a web of dependencies far richer than any one paper can convey. Yet our main tool for navigating this web—the citation graph—is crude. Citations record only what authors choose to acknowledge, missing the many implicit connections [1]; they reflect convention as much as genuine intellectual debt [2]; and they rarely bridge across fields, even when deep analogies exist between, say,

*Email: hpp@bimsa.cn

algebraic geometry and combinatorics. In short, the citation graph is a lossy sketch of a much richer structure.

We take a first step toward filling this gap by embedding theorem *statements* into a continuous space, calibrated by citation and dependency signals. A complete representation would need proofs, definitions, and notational context as well, but the statement alone already turns out to carry substantial structural information. Graph neural networks on citation networks [3–6] inherit the limitations of the underlying graph. Large language models for theorem proving [7–9], despite impressive results, operate as local search: each proof attempt depends on the prompt, and without a broader picture of the theorem space, even a strong prover risks missing relevant tools from distant fields.

The gap. The study of scientific knowledge networks has revealed rich structural patterns in how ideas connect and evolve [10–12]. Pre-trained language models for scientific domains [13–15] and document-level embedding methods [16–18] provide useful representations, but are designed for general scientific text rather than mathematical theorems specifically, and focus on retrieval rather than constructing a global landscape. Large-scale mathematical corpora such as the theorem-search-dataset [19] and NaturalProofs [20] provide raw material, but no existing work embeds theorem statements into a continuous space calibrated by theorem-reference/dependency proxies and systematically evaluates the resulting structure.

Our contribution. We introduce **Theorem Landscape**. In the current 10M-scale setting, we train on 11.42M cleaned theorem-statement rows and evaluate a strict theorem-level contrastive model built from proof-core dependency edges. We compare this model against a Stage 1 language-model baseline, an all-edge ablation, and a shuffled-edge negative control. The comparison isolates what theorem-dependency supervision contributes beyond theorem text alone and beyond arbitrary theorem-pair contrastive training.

A useful analogy comes from computational physics. Molecular dynamics simulations explore atomic configurations one trajectory at a time, and each trajectory is local—determined by its starting point [21–23]. Energy landscape theory changed the game by mapping the global structure of the potential energy surface: basins, saddle points, transition pathways [24, 25]. The landscape does not replace simulation; it tells the simulator where to look. Theorem proving with LLMs faces the same issue. The field has advanced rapidly, from early neural premise selection [26, 27] through GPT-based provers [7, 28] to recent breakthroughs such as DeepSeek-Prover [8] and AlphaProof [9]; retrieval-augmented [29] and autoformalization [30] approaches have expanded what is possible (see [31] for a survey). But every one of these systems is a local search operator: each proof attempt is one trajectory through logical space, steered by the prompt. Without a broader picture of where theorems sit relative to each other, even a strong prover can miss tools from unfamiliar fields, or fail to recognize when a conjecture lacks the theoretical infrastructure needed for any proof to succeed.

At the sentence level, methods such as Sentence-BERT [32] and SimCSE [33] have advanced text representation learning, building on contrastive learning [34]. Knowledge graph embedding techniques [35, 36] offer complementary perspectives on relational structure, and NaturalProver [37] provides theorem-level reference graphs. Our two-stage approach draws on citation-informed representation learning [16, 17] to calibrate the semantic space toward extracted mathematical reference structure rather than relying on linguistic similarity alone.

The resulting landscape enables analyses that complement citation graphs and theorem provers. By embedding an open conjecture into the landscape, we can examine the density and nature of surrounding theorems. In curated analyses, famous conjectures show informative neighborhood content, while large-scale testing shows that density alone does not separate typical arXiv conjectures from proved-like statements ($p = 0.41$). The primary analytical value therefore lies in qualitative neighborhood content—identifying what specific mathematical tools are available near a conjecture. The continuous embedding also surfaces candidate cross-field neighbors: these are hypothesis-generating navigation hints, not validated discoveries of missing dependencies.

Measuring mathematical importance is notoriously hard. Mathematics has a low-citation culture that makes bibliometric proxies unreliable [38, 39], and earlier work on citation ranking [40], genealogy networks [41, 42], and mathematical knowledge graphs [43, 44] has given partial views. Philosophers of mathematics [45] emphasize depth, fertility, and interconnectedness—qualities that a continuous embedding might capture. In the landscape, we define hub theorems as those with high local density (many close neighbors). By combining this with publication time, we construct a *seminal score* that flags theorems appearing early in sparse regions that later attracted dense clusters of follow-up work. This is not a validated impor-

tance metric—it is a hypothesis-generating tool—but it picks out different theorems from citation counts ($r = 0.31$), and we examine its properties in Appendix B.9.

We employ dimensionality reduction techniques such as UMAP [46] and t-SNE [47] to visualize the resulting landscape and reveal its global structure. The landscape can also inform retrieval and tool selection: rather than relying solely on a mathematician's intuition to craft prompts, a theorem prover could consult the landscape to identify relevant tools and promising directions. In the current strict Stage 2b rerun, proof-core embeddings reach 38.7% Recall@100 over one million theorems, above Stage 1 and the random-edge control; the earlier Stage 2a retrieval suite is reported as auxiliary evidence in Appendix B.7. Integrating the landscape into an end-to-end theorem proving pipeline is a natural next step.

In experiments, proof-core Stage 2b improves over the Stage 1 baseline on domain classification (0.5428 vs. 0.4511), paper-citation prediction (AUC 0.8921 vs. 0.8097), held-out theorem-link prediction against random negatives (AUC 0.9908 vs. 0.9227), and conjecture-neighborhood density gap (0.1086 vs. 0.0617). A random-edge control performs much worse (theorem-link AUC 0.8062, density gap 0.0097), while an all-edge ablation improves broad citation/domain metrics but weakens the interpretability signal. Among famous conjectures, the Collatz Conjecture sits in a sparse neighborhood and the Four Color Theorem in a dense graph-coloring neighborhood, matching mathematical intuition; large-scale keyword-matched conjecture tests remain a caveat, so we treat this as a qualitative case study rather than a provability classifier.

The remainder of this paper is organized as follows. Section 2 describes our data collection and processing pipeline. Section 3 details the main two-stage training procedure. Section 4 presents the current 10M-scale results, a famous-conjecture case study, and short navigation examples. Section 5 discusses implications, failure analysis, and limitations, and Section 6 concludes. Detailed auxiliary analyses are collected in the appendix.

2 Data

2.1 Source Corpus

We use the theorem-search-dataset [19], a large-scale collection of mathematical theorem statements extracted from arXiv source files. The raw dataset contains 1,341,083 theorem statements from 209,777 papers in Parquet format. Each record includes the theorem body in LaTeX, a paper identifier, and paper-level metadata: arXiv primary category, publication date, and citation count.

We use only the theorem body (LaTeX source) as input, avoiding the dataset's LLM-generated natural language summaries to prevent introducing model-specific biases into the representation space. We restrict to theorem *statements* rather than proofs: the statement defines a theorem's position in mathematical space (its "coordinates"), while the proof describes a path to that position. Multiple distinct proofs may establish the same theorem, making proofs better suited as edge information in future work.

The current main training runs use a later cleaned 10M-scale theorem-statement corpus containing 11,424,541 rows. The 1,098,743-row theorem-search corpus described below is retained as the fixed downstream evaluation and auxiliary-analysis corpus because it provides theorem IDs, paper IDs, arXiv categories, publication dates, and citation metadata.

2.2 Preprocessing

For the fixed downstream evaluation corpus, we apply three filtering steps to obtain a clean, focused set:

1. **Domain filtering.** We retain only theorems from papers whose primary arXiv category belongs to the `math.*` hierarchy, removing theorems from computer science, physics, and other fields. This reduces the corpus from 1,341,083 to 1,114,926 theorems across 30 mathematical subcategories.

2. **Length filtering.** We remove theorems with body length below 20 or above 2,000 characters, excluding trivially short statements and anomalously long blocks that likely reflect parsing artifacts. This yields 1,098,743 theorems.

3. **Train/validation split.** For the earlier 1.1M auxiliary pipeline, we split the filtered corpus into 1,043,805 training and 54,938 validation theorems (95%/5%) with stratification by primary category.

A post-hoc language analysis reveals that 99.82% of the retained theorems are in English, with small minorities in French (0.14%), Russian (0.03%), Spanish (0.01%), and German (<0.01%). While we do not filter by language during training, we note that these non-English theorems form isolated clusters in the embedding space (Section B.4), driven by linguistic rather than mathematical dissimilarity.

The top five categories by size are combinatorics (`math.CO`, 123K), analysis of PDEs (`math.AP`, 110K), algebraic geometry (`math.AG`, 89K), number theory (`math.NT`, 86K), and probability (`math.PR`, 74K), providing broad coverage of major mathematical disciplines.

3 Method

3.1 Stage 1: Masked Language Model Pre-training

We initialize from RoBERTa-base [48] (125M parameters) and perform continued pre-training with the masked language modeling (MLM) objective on the 10M-scale cleaned theorem corpus. The training split contains 11,424,541 theorem-statement rows and the validation split contains 50,000 rows.

We choose RoBERTa over domain-specific alternatives such as SciBERT [13] or MathBERT [14] because (i) RoBERTa's dynamic masking and removal of the next-sentence prediction objective yield strong representations, (ii) domain-specific models were pre-trained on much smaller corpora without tokenizer improvements that would justify their use, and (iii) continued pre-training on 11.42M theorem statements provides substantial domain adaptation.

We use the standard RoBERTa tokenizer without modification. While mathematical LaTeX is split into subword tokens that do not correspond to meaningful mathematical units, the model learns compositional representations of these subtoken sequences through exposure to large amounts of domain data, analogous to how protein language models learn amino acid semantics from per-residue tokenization.

Training uses a maximum sequence length of 512, masking probability of 0.15, per-device batch size of 16 with gradient accumulation over 4 steps (effective batch size 64), learning rate of 2×10^{-5} with 10% linear warmup, and FP16 mixed precision. We train for one epoch on a single NVIDIA RTX 5090 GPU (32 GB), requiring approximately 13.8 hours. Table 1 reports the validation loss.

Table 1: 10M-scale MLM pre-training validation loss.

Epoch	1
Eval Loss	0.4412

3.2 Embedding Extraction and Whitening

After pre-training, we extract theorem embeddings by taking the `[CLS]` token representation from the final hidden layer. For downstream evaluation and auxiliary analyses, this produces a 768-dimensional vector for each theorem in the fixed 1,098,743-row evaluation corpus.

A well-known limitation of BERT-family models is *anisotropy*: the learned representations occupy a narrow cone in the embedding space, causing all pairwise cosine similarities to cluster near 1.0 regardless of semantic content. We confirm this phenomenon in our embeddings: random theorem pairs exhibit a mean cosine similarity of 0.992 with standard deviation 0.002, providing almost no discriminative signal.

We apply PCA whitening [49]: subtract the global mean, rotate by the principal components, and normalize variances. After whitening, random pair cosine similarities have mean ≈ 0.0 and standard deviation 0.036—enough to distinguish related from unrelated theorems. The effect on downstream metrics is large (Section 4).

3.3 Stage 2b: Theorem-Level Citation Fine-Tuning

Paper-level citations are an indirect and noisy proxy for theorem-level dependencies. When paper A cites paper B, typically only a few specific theorems in B are relevant. The earlier auxiliary Stage 2a pipeline

therefore contains unavoidable noise because it randomly pairs theorems across cited papers (Appendix B). We address this by extracting precise theorem-level citation edges directly from LaTeX source files.

Theorem-level citation extraction. We download the LaTeX source of 48,000+ arXiv math papers and parse two types of theorem-level references:

- *Internal references* (`\ref`): When a proof of Theorem A references Lemma B via `\ref{label_B}` within the same paper, we extract the directed edge $A \to B$. We identify the owning theorem of each proof environment and match `\ref` targets to theorem labels.
- *Cross-paper references* (`\cite`): Patterns such as "Theorem 3.1 of `\cite{key}`" are matched via regular expressions. We resolve cite keys to arXiv IDs through `\bibitem` parsing, then match the theorem number to entries in our dataset.

This extraction yields 562,309 internal references and 256,440 cross-paper references. After matching endpoints against the 10M-scale theorem corpus, we obtain 319,788 matched internal edges and 565,968 expanded cross-paper edges. The resulting all-edge set contains 885,749 theorem pairs. For the strict main setting, we use the *proof-core* subset: 278,856 internal proof-context edges with no duplicate normalized-statement pairs. These are high-confidence proxy edges of the form "the proof of theorem A references theorem B," subject to the extraction-noise caveats below.

An automatic extraction-consistency audit verifies the mechanical invariants needed for the proof-core split: all 278,856 proof-core rows are internal proof-context references between theorem-like environments, with zero missing endpoints, zero endpoint-paper mismatches, zero self-loops, zero directed duplicate rows, and zero endpoint-theorem overlap across train/validation/test. This audit does not estimate semantic precision; we therefore describe the edges as extracted proof-reference proxies rather than manually verified logical dependencies.

Training. The current Stage 2b proof-core model initializes from the 10M-scale Stage 1 checkpoint and uses the same symmetric InfoNCE objective with in-batch negatives. We split at the paper level into 223,384 training pairs, 28,168 validation pairs, and 27,304 test pairs; the split audit confirms zero overlap in anchor papers, endpoint theorem IDs, directed pairs, and undirected pairs. We train for 3 epochs with batch size 64, learning rate 2×10^{-5}, temperature 0.05, weight decay 0.01, and BF16 mixed precision. The best proof-core training loss is 0.0560. We also train all-edge and random-edge variants with the same architecture as ablations and controls.

4 Experiments

4.1 Current 10M-Scale Stage 2b Results

We first report the current 10M-scale Stage 2b results, which are the strict main results of this version of the paper. The training corpus contains 11,424,541 cleaned theorem-statement rows. The fixed downstream evaluation corpus contains 1,098,743 theorem statements with theorem IDs, paper IDs, arXiv categories, publication dates, and citation counts.

The main model is *Stage 2b proof-core split*: theorem-level contrastive fine-tuning initialized from the Stage 1 checkpoint and trained on high-confidence proof-core theorem-dependency edges. The split is at the paper level. We compare it against two controls. The *all-edge split* uses more reference edges, including many cross-paper edges, and is treated as an ablation rather than the strict main setting. The *random-edge split* preserves the proof-core split size and anchor distribution but shuffles positives while rejecting original proof-core directed and undirected edges.

Table 2 shows that proof-core Stage 2b improves over Stage 1 on domain classification, paper-citation prediction, random-negative theorem-link prediction, and conjecture-neighborhood density gap. The all-edge model has the best broad domain and citation metrics, but its density gap is close to Stage 1. The random-edge control collapses the density gap and performs far worse on held-out theorem-link prediction under the same negative sampling. This supports the narrower central claim that proof-core edge structure, not contrastive training alone, changes the local geometry; hard-negative theorem-link tests are the stricter check for whether this geometry goes beyond local topic/context matching.

Table 2: Main results on the 10M-scale theorem corpus. Stage 2b proof-core split is the strict main setting; all-edge split and random split are ablations. Theorem AUC uses held-out proof-core positives against random cross-paper negatives; hard-negative variants are reported separately when available.

Model	Domain Acc.	Macro F1	Citation AUC	Citation AP	Theorem AUC (rand. neg.)	Density Gap
Stage 1 MLM	0.4511	0.3296	0.8097	0.8255	0.9227	0.0617
Stage 2b proof-core split	0.5428	0.4316	0.8921	0.9046	0.9908	**0.1086**
Stage 2b all-edge split	**0.5639**	**0.4476**	**0.9147**	**0.9250**	**0.9943**	0.0648
Stage 2b random-edge split	0.3022	0.1864	0.6424	0.6683	0.8062	0.0097

Table 3: Hard-negative theorem-link evaluation on held-out proof-core positives. Same-section metadata is not available in the endpoint corpus, so the same-section setting falls back to same-paper non-edges and is omitted here. Lexical/length negatives are selected by length bucket and token-overlap similarity.

Negative set	Pairs	Proof-core AUC	Random-control AUC	Proof-core sim gap
Same-paper non-edge	9,583	0.6462	0.5957	0.1062
Same category/year non-edge	9,985	0.9704	0.7754	0.4922
Lexical/length-matched non-edge	10,000	0.9589	0.7134	0.4755

Table 3 addresses the main caveat in the theorem-link metric. When negatives are sampled from the same paper, proof-core AUC drops from 0.9908 to 0.6462, confirming that random cross-paper negatives are an easy setting and that local topic, notation, and writing style explain part of the signal. The proof-core model nevertheless remains above the random-control model on same-paper, same-category/year, and lexical/length-matched negatives. We therefore interpret theorem-link results as evidence for dependency-aware local organization, not as proof that distance alone recovers logical dependence.

Table 4 gives bootstrap intervals. The proof-core density gap remains above zero under resampling, while the random-edge interval is consistent with no meaningful gap. Table 5 records the leakage audit: proof-core and random are strict paper-level settings with no endpoint-theorem or pair overlap, while all-edge has endpoint reuse caused by cross-paper references. For this reason, we use proof-core as the main result and all-edge only as an ablation.

4.2 Result Scope

The strict current Stage 2b claims are those in Tables 2–6. We rerun famous-conjecture density, qualitative neighborhoods, proof-tool retrieval, and category-level bridge summaries for the current Stage 2b checkpoints and report the compact versions below. Detailed analyses from the earlier 1.1M-corpus pipeline—including Stage 2a, visualization, outliers, link prediction, retrieval, bridge discovery, centrality, ablations, and baseline comparisons—are reported in Appendix B. Those auxiliary results support the landscape framework but are not replacements for the current proof-core split.

4.3 Reproducibility and Traceability

The current 10M-scale experiments are tied to fixed data splits, checkpoints, and evaluation artifacts. Stage 1 uses the cleaned theorem-statement corpus in `data/stage1_full_clean_10m`; despite the historical "10M" name, this corpus contains 11,424,541 rows. Downstream evaluation uses the fixed 1,098,743-row theorem corpus with theorem IDs, paper IDs, arXiv categories, dates, and citation metadata.

The main artifacts are:

- Stage 1 checkpoint: `checkpoints/stage1_10m_roberta_base/best`.
- Stage 2b checkpoints: `checkpoints/stage2b_10m_proof_core_split/best`, `checkpoints/stage2b_10m_all_split/best`, and `checkpoints/stage2b_10m_random_split/best`.
- Stage 2b split data: `data/theorem_refs_10m_split`, `data/theorem_refs_10m_all_split`, and `data/theorem_refs_10m_random_split`.
- Evaluation outputs: Stage 1, proof-core, all-edge, and random-edge metrics JSON files.
- Stability and audit outputs: bootstrap confidence intervals and split-leakage audit reports.

Table 4: Same-protocol bootstrap confidence intervals. Values are point estimate with 95% percentile interval from 1,000 bootstrap samples. Theorem AUC uses random cross-paper negatives.

Model	Citation AUC	Theorem AUC (rand. neg.)	Density Gap
Stage 1 MLM	0.8097 [0.8022, 0.8168]	0.9227 [0.9190, 0.9268]	0.0623 [-0.0031, 0.1208]
Stage 2b proof-core split	0.8921 [0.8865, 0.8973]	0.9908 [0.9898, 0.9918]	**0.1069 [0.0418, 0.1767]**
Stage 2b all-edge split	**0.9147 [0.9095, 0.9193]**	**0.9943 [0.9935, 0.9950]**	0.0638 [-0.0529, 0.1530]
Stage 2b random-edge split	0.6424 [0.6322, 0.6523]	0.8062 [0.8001, 0.8121]	0.0105 [-0.0389, 0.0552]

Table 5: Split and leakage audit. The proof-core and random splits are strict paper-level splits with no endpoint-theorem or pair overlap. The all-edge split is retained as an ablation but is not the strict main setting because cross-paper edges introduce endpoint reuse across splits.

Split	Train pairs	Val pairs	Test pairs	Endpoint overlap	Pair overlap
Proof-core split	223,384	28,168	27,304	0	0
All-edge split	725,609	90,742	69,398	1,211	24 undirected
Random-edge split	223,384	28,168	27,304	0	0

- Reviewer-response audit scripts: hard-negative theorem-link evaluation, automatic theorem-reference extraction consistency audit, corpus-overlap audit, and Stage 2b retrieval/bridge reruns.

A normalized-statement overlap audit finds 69,997 overlaps between the 11.42M Stage 1 corpus and the fixed evaluation corpus, and 13,213 overlaps between the evaluation corpus and proof-core endpoint corpus. We treat these as memorization-risk diagnostics rather than split violations because the main supervised proof-core split itself has no endpoint-theorem or pair overlap across train, validation, and test. The paper tables are generated from these artifacts and summarized in the traceability file accompanying the paper sources.

4.4 Claim Scope

4.5 Famous-Conjecture Case Study

A central motivation for the Theorem Landscape is the hypothesis that a conjecture's position relative to known theorems can suggest the local mathematical tools around it. We test this by embedding ten famous conjectures—three proved and seven unsolved—into the landscape and analyzing their neighborhood density.

For each conjecture, we tokenize and encode its standard mathematical statement through the trained model, apply the same whitening transformation, and use UMAP's `transform` method to project it into the existing landscape coordinates. We quantify neighborhood density as the mean cosine similarity to the 50 nearest theorems in the full whitened embedding space (768 dimensions).

For this curated set (Table 8), the proved examples have higher neighborhood density than the unsolved examples, and the gap increases under current proof-core Stage 2b fine-tuning. With only 10 conjectures this is a case study, not a statistical claim; the real interest is in what each neighborhood contains.

- The **Four Color Theorem** (proof-core Stage 2b density 0.5683) is the densest proved theorem, deep inside the graph-coloring neighborhood.
- **Fermat's Last Theorem** and the **BSD Conjecture** land close together, both relying on elliptic curves and modular forms. Wiles's proof used the modularity theorem for semistable elliptic curves, which is directly related to BSD.
- The **Collatz Conjecture** (density 0.3295) remains the sparsest famous conjecture in proof-core Stage 2b. As Erdős put it, "Mathematics is not yet ready for such problems."
- The **Goldbach Conjecture** (proof-core Stage 2b density 0.4705) is the densest unsolved example, sitting near additive and prime-representation results. This matches the partial progress: Chen's theorem already gives every large even integer as a sum of a prime and a semiprime.

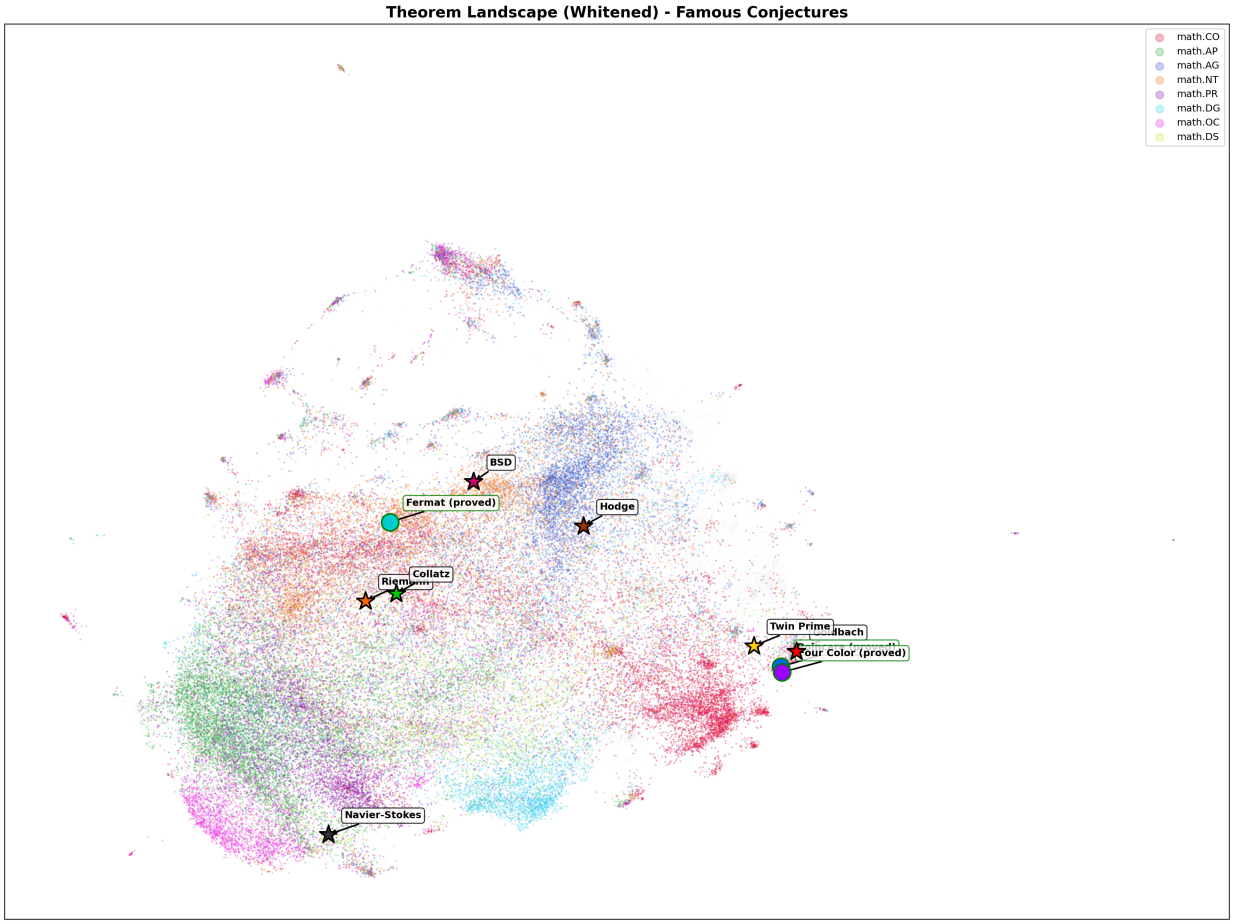

Figure 1: Auxiliary visualization of famous conjectures projected into the Theorem Landscape (whitened UMAP). Stars ($\star$) denote unsolved conjectures; circles ($\circ$) with green borders denote proved examples. The figure is illustrative; Table 8 reports the current 10M-scale density values.

Table 6: Negative-control summary. The random-edge control preserves the proof-core split size and anchor distribution, but shuffles positives within each split while rejecting original proof-core directed and undirected edges. Theorem AUC is measured with random cross-paper negatives.

Model	Training signal	Theorem AUC	Density Gap
Stage 2b proof-core split	True proof-core theorem dependencies	0.9908	0.1086
Stage 2b random-edge split	Shuffled theorem positives	0.8062	0.0097

Table 7: Scope of the main paper claims.

Claim	Status	Evidence
Proof-core Stage 2b improves dependency-aware geometry	Main claim	Strict proof-core split improves over Stage 1 on domain, citation, random-negative theorem-link, and density-gap metrics; random-edge control fails to reproduce the effect. Hard-negative theorem-link evaluation shows the same-paper setting is much harder, but proof-core remains above the random-control model.
All-edge supervision has a tradeoff	Main ablation	All-edge improves broad domain/citation metrics but weakens the density signal and has endpoint reuse across splits.
Famous-conjecture neighborhoods are interpretable	Qualitative case study	Ten famous examples show coherent local toolkits, but keyword-matched arXiv conjectures do not show a general density separation ($p = 0.41$).
Retrieval and bridge discovery are useful navigation tools	Supporting, not validated downstream claims	Current Stage 2b retrieval improves over Stage 1 and random control on cited-paper retrieval. Bridge discovery is category-level and unblinded; it should be read as navigation hints until expert validation is complete.
Centrality, temporal evolution, and generation	Exploratory	These are hypothesis-generating analyses and future directions, not validated main claims.

Some conjectures shift between stages. The **Hodge Conjecture** rises from 0.2846 to 0.3502 under proof-core Stage 2b, while all-edge supervision raises it further to 0.4207. **Navier–Stokes** rises from 0.3147 to 0.3953, reflecting nearby PDE and regularity statements. The **Riemann Hypothesis** changes only modestly from 0.3411 to 0.3588, and Collatz remains the sparsest example.

Neighborhood content analysis. Beyond aggregate density, we examine the specific theorems surrounding each conjecture to assess whether the neighborhood constitutes a "proof toolkit" (Table 9). This analysis reveals qualitative differences invisible to density alone.

The contrast between neighborhoods is visible in the actual retrieved statements. The **Four Color Theorem** is surrounded by direct graph-coloring results, and proof-core Stage 2b raises its top-neighbor similarity from 0.5675 in Stage 1 to 0.8397. The **Goldbach Conjecture** retrieves prime-representation and additive-number-theory statements. The **BSD Conjecture** concentrates on elliptic-curve arithmetic, including root numbers, rational torsion, and quadratic twists. The **Riemann Hypothesis** retrieves zeta-zero statements but also exposes a failure mode: some near neighbors come from general-mathematics or probability statements that share zeta/zero vocabulary without being proof tools.

Large-scale density analysis: the signal does not generalize. A critical question is whether the

Table 8: Current 10M-scale famous-conjecture density analysis (top-50 mean cosine similarity). Higher density is interpreted only as local-neighborhood richness in this curated case study, not as a general provability score.

Conjecture	Status	Stage 1	Stage 2b proof-core	Random control
Four Color Theorem	Proved (1976)	0.4169	0.5683	0.3395
Poincaré Conjecture	Proved (2003)	0.4224	0.4350	0.3627
Fermat's Last Theorem	Proved (1995)	0.3368	0.4715	0.2873
Proved average		*0.3920*	*0.4916*	*0.3298*
Goldbach Conjecture	Unsolved	0.4133	0.4705	0.3276
BSD Conjecture	Unsolved	0.2830	0.3717	0.2564
Twin Prime Conjecture	Unsolved	0.4213	0.4053	0.3681
Navier–Stokes	Unsolved	0.3147	0.3953	0.3340
Riemann Hypothesis	Unsolved	0.3411	0.3588	0.3724
Hodge Conjecture	Unsolved	0.2846	0.3502	0.3051
Collatz Conjecture	Unsolved	0.2541	0.3295	0.2769
Unsolved average		*0.3303*	*0.3830*	*0.3201*

Table 9: Representative nearest neighbors for selected conjectures in the current proof-core Stage 2b model. The neighborhood content is used as qualitative navigation evidence, not as calibrated proof-infrastructure measurement.

Conjecture	Representative Neighbors	Categories
Four Color Thm	Planar graph 4-colorability, planar triangle-free 3-colorability, cubic planar 3-edge-colorability	CO/GM
Goldbach Conj.	Sums of two squares for primes 1 mod 4, differences between square-primes, sums of three odd primes	NT/CO
BSD Conj.	Root numbers of semistable elliptic curves, elliptic curves with rational 2-torsion, quadratic twists	NT
Riemann Hyp.	Riemann zeta zeros on the critical line, zeros of related transforms, analytic properties of $\zeta(s)$	GM/NT/PR
Poincaré Conj.	Closed orientable 4-manifold embeddings, 3-manifolds homeomorphic to the 3-sphere, compact 3-orbifold quotients	GT/DG/DS

density pattern observed for famous conjectures holds more broadly. In the earlier Stage 2a auxiliary pipeline, we test this by identifying 4,481 conjecture-like statements (containing "conjecture," "open problem," etc.) and 10,000 proved-like statements (containing "theorem," "we prove," etc.) via keyword matching and comparing their neighborhood densities. The result is unambiguously negative: conjecture-like and proved-like theorems show no significant density difference (mean 0.326 vs. 0.328, $p = 0.41$).

Why the discrepancy? Typical arXiv conjectures—"We conjecture that Theorem 3.1 extends to the non-commutative case"—are incremental extensions of active research, sitting in the same dense neighborhoods as surrounding results. Famous open problems are different: they are famous *because* they resist existing frameworks, and the landscape reflects that isolation. Density can sometimes flag problems whose neighborhoods look thin for their complexity, but it cannot predict provability in general.

4.6 Navigation Examples

The main quantitative result is that proof-core dependency supervision changes the geometry in ways that text-only and random-edge training do not. To illustrate what such a geometry can be used for, we summarize two navigation reruns for the current Stage 2b checkpoints; longer historical analyses remain in Appendix B.7 and Appendix B.6.

First, in proof-tool retrieval over the fixed 1.1M-theorem evaluation corpus, nearest-neighbor search retrieves theorems from cited papers with 38.7% Recall@100 for proof-core Stage 2b, compared with 34.9% for Stage 1 and 31.7% for the random-edge control. The all-edge ablation reaches 40.1%, consistent with its stronger paper-citation metric. This is not yet an end-to-end theorem-proving result, but it shows that the current proof-core model improves cited-paper tool retrieval over text-only and shuffled-edge baselines.

Second, the current category-level bridge rerun finds low-citation category pairs that are close in embedding space, including complex variables–spectral theory, K-theory–symplectic geometry, and K-theory–operator algebras. These are not claims of new theorems; they are navigation hints that require blinded expert validation. We generate a bridge-validation sample for that purpose, while the current paper's main claim remains the controlled proof-core Stage 2b result.

5 Discussion

5.1 From Semantic to Citation-Calibrated Geometry

Stage 1 (MLM) produces a semantic landscape: theorems with similar vocabulary land nearby. The current proof-core Stage 2b model adds theorem-dependency supervision, moving the geometry toward proof-relevant structure. In the strict 10M-scale split, this raises domain accuracy from 0.4511 to 0.5428, paper-citation AUC from 0.8097 to 0.8921, random-negative theorem-link AUC from 0.9227 to 0.9908, and the conjecture-neighborhood density gap from 0.0617 to 0.1086. The random-edge control does not reproduce these gains, which is the key evidence that the edge structure matters.

5.2 Implications for Conjecture Assessment

For famous conjectures, neighborhood density and content are informative: Four Color sits in a dense graph-coloring cluster, while Collatz is relatively sparse. But this does not generalize—keyword-matched conjectures from arXiv show no density difference from proved theorems ($p = 0.41$). The useful signal is in *what* surrounds a conjecture, not how dense the neighborhood is.

In the current famous-conjecture analysis, the Hodge Conjecture moves from 0.2846 (Stage 1) to 0.3502 (proof-core Stage 2b), and to 0.4207 in the all-edge ablation. Its statement uses atypical vocabulary for algebraic geometry, so reference supervision appears to place it more appropriately. We treat this as qualitative evidence about neighborhood content, not as a calibrated provability score.

5.3 Landscape-Guided Proof Search

The current proof-core retrieval rerun gives 38.7% Recall@100 over one million theorems, above Stage 1 and the random-edge control but below the older Stage 2a auxiliary run. This suggests a practical workflow: query

the landscape for candidate lemmas and techniques before attempting a proof. Whether those candidates actually improve proving remains to be tested in an end-to-end theorem-proving pipeline.

5.4 Failure Analysis

Split-dependent metrics. The all-edge model improves broad domain and citation metrics but has endpoint reuse across splits and a weaker density signal. We therefore treat it as an ablation rather than the strict result. This distinction matters: without the split audit, the all-edge row would look like the strongest model by aggregate metrics alone.

Theorem-link negatives. The random-negative theorem-link AUC is useful for checking whether proof-core edges are closer than unrelated theorems, but it overstates dependency awareness because positives are internal proof references that share paper-level topic, notation, and style. The hard-negative audit confirms this: same-paper AUC is only 0.6462. Same-section negatives would be stronger still, but section metadata is not available in the current endpoint corpus.

Statement-only failures. Many theorem statements are opaque without definitions, notation, or surrounding assumptions. Short statements and label-dependent statements such as "the result follows from Theorem 3.2" provide little mathematical content to embed, regardless of the contrastive signal.

Auxiliary navigation failures. Retrieval and bridge discovery are useful for navigation, but they remain auxiliary analyses. A nearest neighbor can suggest a proof tool without being sufficient for a proof, and a high-similarity bridge can reflect shared notation or boilerplate rather than a mathematical relation.

Conjecture caveats. Famous conjecture neighborhoods are interpretable, but density does not generalize as a provability signal. High-density unsolved conjectures often sit inside active research programs, while low-density proved statements may simply omit context.

5.5 Limitations

Statement-only representation. We embed only theorem statements—no proofs, no definitions, no notational context. Many statements are opaque without context ("The theorem holds under the above assumptions"), and retrieval failures correlate with shorter bodies (337 vs. 445 characters). This is a *statement-language landscape*, not a complete theorem knowledge landscape.

Intra-paper dominance of theorem-level supervision. The strict proof-core Stage 2b edges are internal proof-context `\ref` dependencies. The model therefore learns high-confidence within-paper theorem structure, not a complete field-level dependency network. The all-edge ablation adds many cross-paper references, but its split has endpoint reuse and its density signal is weaker, so it is not the main setting.

Reference extraction noise. Proof-core edges come from LaTeX reference extraction and theorem endpoint matching. The automatic consistency audit verifies endpoint coverage, paper consistency, proof-context/core-environment shape, duplicate/self-loop rates, and split separation, but it cannot decide whether each reference is a genuine mathematical dependency. We therefore treat these edges as clean extracted proof-reference proxy supervision, not manually verified logical dependencies.

Conjecture sample size. The famous-conjecture density pattern rests on 10 examples and does not generalize ($p = 0.41$). The analysis is qualitative, not predictive.

Bridge validation. Current bridge results are category-level candidate bridges selected by embedding similarity and low citation counts. The automatic bridge audit supports only a navigation-candidate claim, not expert-validated mathematical relatedness. We have not tested whether knowing about these bridges actually helps downstream tasks.

Baseline coverage. We compare against five SOTA sentence embedding models (E5, GTE, BGE, SPECTER2, MiniLM) in addition to classical baselines, but do not evaluate reranking or late-interaction architectures. We also lack retrieval results for SOTA models on the full 1.1M corpus (encoding cost); a smaller hard-negative retrieval benchmark is the right controlled comparison.

Centrality and seminal score. These are exploratory. The seminal score ($r = 0.31$ with citations) captures something different from citation popularity, but we have no external gold standard for "importance." The score's design (top-50 neighbors, multiplicative formula) has not been optimized.

Corpus scope. Our corpus covers only arXiv-deposited mathematics (primarily 2008–2025), which

skews toward recent work and underrepresents classical results. The seminal theorem identification is consequently limited to "seeds" within this time window rather than historically foundational results.

6 Conclusion

We embedded a 10M-scale corpus of theorem statements into a continuous space and calibrated it with theorem-level dependency signals. In the strict proof-core split, Stage 2b improves over the Stage 1 language-model baseline on domain classification (0.5428 vs. 0.4511), paper-citation prediction (AUC 0.8921 vs. 0.8097), held-out theorem-link prediction against random negatives (AUC 0.9908 vs. 0.9227), and conjecture-neighborhood density gap (0.1086 vs. 0.0617). A random-edge control collapses the density gap to 0.0097 and lowers theorem-link AUC to 0.8062, showing that the proof-core edge structure matters under the same protocol. An all-edge ablation improves broad domain and citation metrics but weakens the density signal, so the proof-core split remains the main result. For famous conjectures, the landscape provides informative neighborhood characterizations, though the density signal does not hold for typical arXiv conjectures ($p = 0.41$).

Several directions follow naturally. First, expanding high-confidence cross-paper theorem-level supervision would address the main limitation of proof-core Stage 2b, which currently learns primarily within-paper structure. Second, integrating landscape-guided retrieval into a theorem proving pipeline would test whether the structural information we observe translates into measurable proving improvements. Third, generative theorem discovery is a speculative future direction; we discuss it separately in Appendix A.

We plan to release the embeddings, evaluation code, and data splits. The evaluation suite—domain classification, bridge discovery, conjecture analysis, link prediction, retrieval, and temporal evolution—can serve as a testbed for future work on mathematical representation learning.

Scaling theorem-level supervision. The current main training corpus is already 10M-scale, but the strict proof-core dependency graph remains dominated by internal proof references. Expanding high-confidence cross-paper theorem-level supervision would move the landscape from local proof structure toward a broader field-level dependency network.

A Generative Theorem Discovery

DiMA [50] showed that diffusion on protein language model embeddings can generate novel, functional protein sequences. An analogous approach for mathematics would train a diffusion model on theorem embeddings and decode the samples back to LaTeX. One could sample near an unsolved conjecture to propose auxiliary lemmas, or interpolate between distant fields to generate "bridge" statements.

The analogy has a catch. In protein design, AlphaFold provides cheap *in silico* validation of generated sequences. In mathematics, there is no equivalent: Lean [51] can check a proof, but it cannot tell you whether a statement is true without one. The pipeline would therefore be: generate a statement, send it to an LLM-based prover [8, 9], and only if the prover succeeds does Lean confirm correctness. Statements that resist proving are ambiguous—they could be new conjectures, false, or simply out of reach. This limitation is fundamental, not a matter of engineering.

Two things motivate this direction empirically. The conjecture neighborhood analysis suggests that embeddings can expose coherent local toolkits, while the current proof-core result shows that theorem-dependency supervision improves local structure under a strict split. The auxiliary derivation-chain analysis suggests that the landscape can surface missing intermediate lemmas—natural targets for generative sampling.

B Auxiliary Method and Analyses

B.1 Auxiliary Stage 2a: Paper-Level Citation Fine-Tuning

Stage 2a belongs to the earlier 1.1M auxiliary pipeline. It is useful for interpreting how paper-level citation signals shape the landscape, but it is not the strict current Stage 2b result.

While Stage 1 captures the statistical structure of mathematical language, it cannot distinguish between linguistic similarity and genuine mathematical dependency. Two theorems may use similar vocabulary (e.g., "Let G be a group…") without being logically related, or conversely, a result in algebraic geometry may depend critically on a theorem in number theory despite sharing no common terminology.

We calibrate the semantic space into a citation-calibrated structured space by fine-tuning with citation signals, following the paradigm of citation-informed representation learning [16, 17]. We construct the citation graph by querying the Semantic Scholar API for references among the 66K arXiv papers in our corpus, retaining only edges where both endpoints belong to our dataset.

Training pairs. We construct positive pairs from two sources: (i) *citation pairs*—a theorem from paper A paired with a theorem from paper B where A cites B or vice versa—and (ii) *same-paper pairs*—two theorems from the same paper, serving as a co-occurrence signal. Each training sample draws a citation pair with probability 0.7 and a same-paper pair otherwise. Note that citation pairs are *noisy*: when paper A cites paper B, we randomly pair a theorem from each paper, even though only specific theorems in B may be relevant to specific theorems in A.

Loss function. We use symmetric InfoNCE [34] with in-batch negatives. Given a batch of B anchor–positive pairs $\{(\mathbf{a}_i, \mathbf{p}_i)\}_{i=1}^{B}$, the loss is

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{e^{\text{sim}(\mathbf{a}_i, \mathbf{p}_i)/\tau}}{\sum_{j=1}^{B} e^{\text{sim}(\mathbf{a}_i, \mathbf{p}_j)/\tau}} + \log \frac{e^{\text{sim}(\mathbf{p}_i, \mathbf{a}_i)/\tau}}{\sum_{j=1}^{B} e^{\text{sim}(\mathbf{p}_i, \mathbf{a}_j)/\tau}} \right], \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and $\tau = 0.05$ is the temperature. The in-batch negatives naturally provide diverse contrastive signal: with batch size 32, each anchor is contrasted against 31 negatives spanning different papers, categories, and time periods.

Training details. We initialize from the Stage 1 checkpoint and fine-tune with learning rate 2×10^{-5}, batch size 32, 10% linear warmup, and FP16 mixed precision for 3 epochs on 500K sampled pairs. The lower learning rate (relative to Stage 1) prevents catastrophic forgetting of the semantic structure learned during MLM pre-training.

Stage 2a also resolves the anisotropy problem: the mean pairwise cosine similarity drops from 0.992 (Stage 1, raw) to 0.264 (Stage 2a, raw), so post-hoc whitening is no longer needed.

B.2 Auxiliary Domain Classification (Earlier 1.1M Pipeline)

To assess whether the embeddings capture mathematical domain structure, we train a k-nearest neighbors classifier ($k = 5$, cosine distance) to predict the arXiv primary category from the embedding vector. We sample 50,000 theorems (stratified by category), using an 80/20 train/test split.

Table 10: Domain classification results (30 arXiv math subcategories, KNN $k = 5$, cosine distance).

	Accuracy	Macro F1
Random baseline	0.033	—
Stage 1 (whitened)	0.431	0.303
Stage 2a (paper-level)	**0.567**	**0.467**
Stage 2b (theorem-level)	0.545	0.441

Stage 1 reaches 43.1% accuracy (13× the random baseline). Stage 2a raises this to 56.7%: citation signals sharpen domain boundaries beyond what vocabulary alone provides. Stage 2b sits at 54.5%, slightly below Stage 2a—paper-level citations encode domain membership more directly (papers cite within their field), while theorem-level dependencies can cross domain boundaries.

Per-field performance varies as expected. Combinatorics (F1=0.61) and PDEs (F1=0.56) separate well even in Stage 1, thanks to distinctive vocabulary. Dynamical systems (F1=0.22) and representation theory (F1=0.26) are hardest: they borrow techniques from multiple areas, and the embedding reflects this overlap.

B.3 Auxiliary Landscape Visualization (Earlier 1.1M Pipeline)

We apply UMAP [46] ($n_{\text{neighbors}} = 15$, min_dist $= 0.1$, cosine metric) to a random sample of 100,000 whitened embeddings. Figure 2 shows the resulting landscape colored by the eight largest mathematical subcategories, with each category displayed individually to reveal its spatial distribution.

Figure 3 compares the Stage 1 and Stage 2a landscapes. Stage 2a clusters are tighter and better-separated, matching the classification improvements in this auxiliary pipeline.

Fields with distinctive vocabulary—combinatorics, algebraic geometry, PDEs—form tight clusters. Cross-disciplinary fields like optimization (`math.OC`) and dynamical systems (`math.DS`) spread more broadly, as one would expect. Number theory (`math.NT`) sits in a localized region but overlaps with algebraic geometry and combinatorics.

B.4 Auxiliary Outlier Cluster Analysis

We apply DBSCAN ($\varepsilon = 0.8$, min_samples$= 5$) to the 10% of points farthest from the UMAP centroid to identify isolated clusters. This reveals 10 outlier groups, which fall into two distinct categories:

Linguistically-driven outliers. Four clusters consist of non-English theorems: French (157 theorems, primarily number theory), Russian (31 theorems, differential geometry and functional analysis), and a small group with encoding artifacts. Despite comprising only 0.18% of the corpus, these theorems form visually prominent isolated clusters because the tokenizer and MLM objective separate them by surface language rather than mathematical content.

Mathematically-driven outliers. Four clusters reflect genuine mathematical distinctiveness:

- A cluster of 13 pure symbolic logic theorems (`math.LO`) containing almost no natural language, only Boolean algebra expressions.
- A cluster of 200 abstract algebra theorems (`math.RA`) centered on highly specialized structures: Lie-Yamaguti algebras, Hom-Lie algebras, and Rota-Baxter systems.
- A cluster of 138 graph theory theorems (`math.CO`) with a distinctive structural vocabulary (k-connected, cubic, bridgeless).
- A cluster of 17 operator algebra theorems (`math.FA`) using the technical notation of g-fusion frames.

These mathematically-driven outliers are informative: they identify highly specialized subfields whose theorem statements are linguistically self-contained, forming "islands" in the landscape that may serve as markers of mathematical specialization. The linguistic outliers, by contrast, highlight a limitation of the current approach that could be addressed through multilingual pre-training or language filtering.

B.5 Auxiliary Link Prediction Analyses

This subsection reports earlier 1.1M-pipeline link-prediction runs. The current strict 10M-scale theorem-link result is Table 2; the older AUC values below are retained only as auxiliary evidence about cross-granularity behavior.

We evaluate whether embedding proximity predicts citation relationships between papers. From the 30,739 citation edges linking papers within our corpus, we hold out 20% (6,148 edges) for testing. For each test edge connecting paper A to paper B, we randomly sample one theorem from each paper and compute the cosine similarity of their embeddings. Negative pairs are sampled from theorems in unrelated papers.

DeepWalk [52], trained directly on the graph, achieves near-perfect AUC (0.999). Stage 2a reaches 0.948; Stage 2b, 0.902. The gap between 2a and 2b is unsurprising: this evaluation uses paper-level edges, which match Stage 2a's training signal. Stage 2b optimizes for theorem-level dependencies instead. DeepWalk wins on the graph it was trained on, but it cannot capture relationships absent from that graph (Section B.6).

Theorem-level link prediction. To provide a fair evaluation for Stage 2b, we also evaluate on theorem-level dependency edges. From the 247,149 internal `\ref` edges (where one theorem's proof cites another), we hold out 20% and measure whether embedding proximity predicts these theorem-to-theorem dependencies.

On theorem-level edges the pattern reverses (Table 12): Stage 2b reaches AUC 0.998, above Stage 2a's 0.991. Each model does best at the granularity it was trained on. More interesting is the cross-granularity

Figure 2: UMAP projection of 100K theorem embeddings (Stage 1, whitened). Top left: all eight categories overlaid. Remaining panels: each category shown individually against a grey background. Clear spatial clustering is visible for combinatorics (math.CO), algebraic geometry (math.AG), and analysis of PDEs (math.AP), while optimization (math.OC) and dynamical systems (math.DS) exhibit broader distributions reflecting their cross-disciplinary nature.

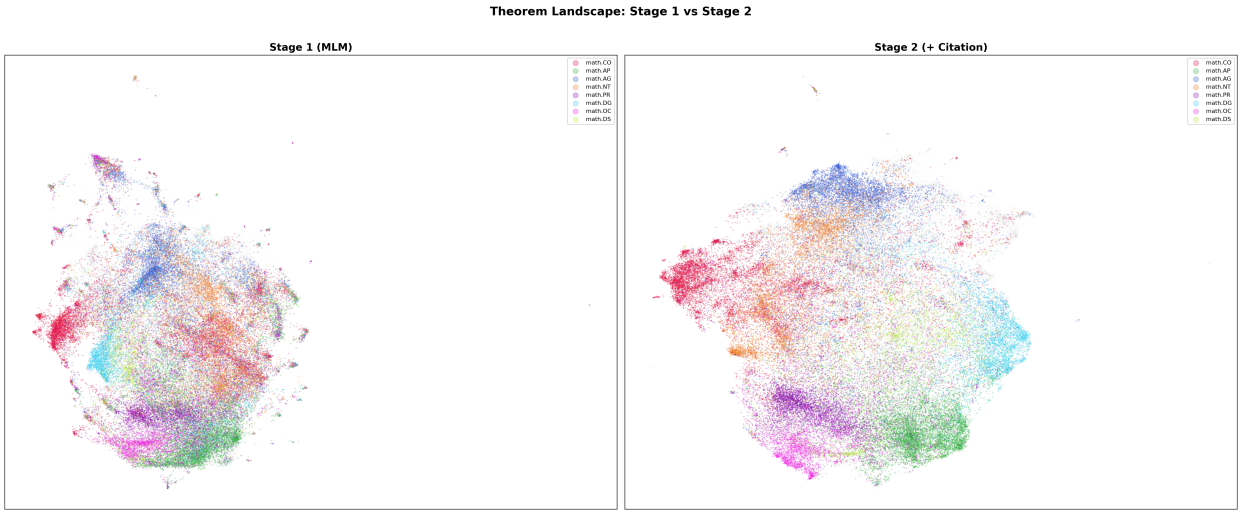

Figure 3: UMAP comparison of Stage 1 (whitened) vs. Stage 2a (raw) embeddings in the earlier auxiliary pipeline. Stage 2a produces tighter domain clusters and clearer inter-domain boundaries.

Table 11: Citation link prediction on held-out edges (5,000 positive / 5,000 negative pairs, full corpus). DeepWalk operates on the citation graph; Stage 1/2 use theorem embeddings. Note: Table 23 reports different values using a 50K-sample unified evaluation.

	AUC	AP
DeepWalk (citation graph)	**0.999**	**1.000**
Stage 1 (whitened)	0.774	0.818
Stage 2a (paper-level)	**0.948**	**0.950**
Stage 2b (theorem-level)	0.902	0.914

Table 12: Theorem-level link prediction on held-out internal `\ref` edges (10K positive / 10K negative pairs) in the earlier auxiliary pipeline. This random-negative evaluation tests whether extracted proof-reference endpoints are closer than unrelated theorems; it should not be read as direct recovery of logical dependence.

	AUC	AP
Stage 1 (whitened)	0.963	0.973
Stage 2a (paper-level)	0.991	0.991
Stage 2b (theorem-level)	**0.998**	**0.998**

transfer: Stage 2b still scores 0.902 on paper-level prediction (untrained), and Stage 2a scores 0.991 on theorem-level prediction. Both signals produce embeddings that generalize beyond their training edges, though paper-level training transfers better downward than theorem-level training transfers upward.

Domain classification comparison. On the subset of theorems in the citation graph, DeepWalk reaches 83.3% accuracy vs. 65.1% (Stage 2a) and 50.5% (Stage 1). Citation networks have strong category homophily—papers cite within their subfield—so a graph method naturally dominates here. The embedding trades classification accuracy for cross-domain sensitivity.

B.6 Auxiliary Cross-Domain Bridge Discovery

Can the embedding find connections that citation graphs miss? We look at two levels: category centroids and individual theorem pairs.

Category-level bridges. We compute the centroid embedding for each of the 30 arXiv subcategories and measure pairwise cosine similarities. Table 13 shows the closest category pairs.

Table 13: Top cross-domain bridges by category centroid similarity (Stage 2a, whitened).

Category 1	Category 2	Centroid Sim
Algebraic Topology (AT)	K-Theory (KT)	0.549
Quantum Algebra (QA)	Representation Theory (RT)	0.538
Geometric Topology (GT)	Symplectic Geometry (SG)	0.404
Functional Analysis (FA)	Operator Algebras (OA)	0.396
Probability (PR)	Statistics Theory (ST)	0.372
Quantum Algebra (QA)	Rings & Algebras (RA)	0.346
Differential Geometry (DG)	Symplectic Geometry (SG)	0.310

These are known to mathematicians: AT and KT share vector bundles and characteristic classes; QA and RT connect through quantum groups; GT and SG through Floer homology. But they are rarely explicit in citation graphs, where papers cite within their primary category.

Beyond-citation bridges. To quantify how the embedding space extends beyond the citation graph, we compare embedding similarity against citation edge density for all category pairs (Table 14). We identify pairs with high embedding similarity but zero or near-zero citation edges—connections invisible to graph-based methods.

Table 14: Top cross-domain bridges discovered by embedding but absent from the citation graph. Edge density is the fraction of possible inter-category citation edges that exist.

Category 1	Category 2	Emb. Sim	Edge Density
Complex Variables (CV)	Spectral Theory (SP)	0.902	0.000
General Topology (GN)	Operator Algebras (OA)	0.863	0.000
K-Theory (KT)	Symplectic Geometry (SG)	0.851	0.000
General Mathematics (GM)	Metric Geometry (MG)	0.836	0.000
Algebraic Topology (AT)	Representation Theory (RT)	0.917	0.006

Complex variables and spectral theory share operator-theoretic methods on function spaces, yet have zero citation edges in our corpus. K-theory and symplectic geometry interact through index theory, again with no citation edges. In total, 20 category pairs have embedding similarity above 0.80 with citation density below 0.01.

Theorem-level hidden connections. We search for *individual theorem pairs* that are close in the embedding space but share no citation link (neither paper-level nor theorem-level) and have no shared authors. Among all cross-category pairs sampled from 30 subcategories with cosine similarity above 0.5, we find 4,243 (Stage 2a) and 4,167 (Stage 2b) such candidate hidden connections. We inspect top examples from both stages as qualitative case studies rather than as a blinded validation set.

The discovered connections form a spectrum of mathematical relatedness, from shared linguistic patterns to candidate dependency-like chains:

1. **Derivation chains**: Theorem A establishes a condition; Theorem B uses that condition as a hypothesis. A researcher working on a specific instance could chain $A \to B$ to obtain new results.
2. **Shared core structures**: Both theorems study the same mathematical object (e.g., inverse semigroups, links in S^3) from different perspectives, using different methods. Neither implies the other, but awareness of both enriches understanding.
3. **Parallel constructions**: Both theorems develop analogous theories in different settings (e.g., polynomial identities in analysis vs. number theory).
4. **Linguistic similarity** (false positives): Boilerplate mathematical statements that match syntactically but carry no mathematical relationship.

Stage 2a and Stage 2b discover different types of connections. Stage 2a produces higher similarity scores overall (mean 0.733 vs. 0.626) but its top connections include more type-(4) false positives: the highest-scoring pair (GN↔QA, sim=0.907) matches "the relation $\cong$ is an equivalence relation" against "the map is an equivalence of categories." Stage 2b, trained on theorem-level dependencies, produces fewer high-similarity pairs but they are more content-driven, favoring types (1)–(3). This shift provides independent evidence that theorem-level fine-tuning reorients the embedding space from linguistic to logical similarity.

Table 15 presents verified examples spanning the spectrum, from derivation chains to shared structures.

Table 15: Verified hidden theorem connections across the relatedness spectrum. All connections have zero citation link and no shared authors. "Type" refers to the classification above.

Theorem 1	Theorem 2	Sim	Type	Connection
PI space characterization (math.MG)	Lipschitz regularity on PI spaces (math.SP)	0.80	(1)	*Derivation chain*: MG theorem establishes when a space is PI; SP theorem proves $\mathrm{lip}(f) = \lvert Df\rvert$ on PI spaces. Chaining: verify PI via MG $\to$ apply SP regularity
Link Floer homology (math.GT)	Fibred link Hopf plumbing (math.QA)	0.84	(2)	*Shared object*: both study links in S^3 via different machinery (Heegaard Floer vs. quantum $\mathfrak{sl}_2$ invariants)
Boolean inverse monoids (math.CT)	Abelianness in inverse semigroups (math.GR)	0.85	(2)	*Shared object*: both develop structural theory for inverse semigroups from categorical vs. group-theoretic perspectives
BiHom-Poisson algebras (math.RA)	Poisson cohomology (math.RT)	0.87	(3)	*Parallel development*: both extend Poisson algebra theory—one via BiHom generalizations, the other via cohomological deformation

Case study: derivation chain (MG↔SP). The most actionable type of hidden connection is a *derivation chain*, where one theorem's conclusion serves as another's hypothesis. Caputo [53] proves that a doubling metric measure space $(X, \mathsf{d}, \mathfrak{m})$ satisfies the Poincaré inequality if and only if it is (C, L)-1-set-connected—providing a *geometric criterion* for the PI property. De Ponti, Farinelli, and Violo [54], in a different arXiv category (math.SP), prove that on PI spaces, the local Lipschitz constant equals the minimal upper gradient: $\mathrm{lip}(f) = |Df|$ a.e.—a *regularity result* they use to establish the Pleijel nodal domain theorem.

Neither paper cites the other, yet together they form a logical chain: given a specific metric measure space, one can first apply Caputo's geometric criterion to verify the PI property, then invoke De Ponti et al.'s regularity result to study Laplacian eigenfunctions on that space. The landscape places these theorems as neighbors (sim=0.804), surfacing a bridge between geometric foundations (math.MG) and spectral analysis applications (math.SP) that a researcher in either field might not discover through standard literature search.

Case study: shared object across distant fields (GT↔QA). Binns and Dey [55] prove that rank bounds on link Floer homology $\widehat{HFL}$ constrain the topology of link complements in S^3, using Heegaard Floer theory (symplectic geometry). López Neumann and van der Veen [56], studying non-semisimple $\mathfrak{sl}_2$ quantum invariants (representation theory), invoke a classical characterization of fibred link surfaces via Hopf band plumbing. These papers study the *same* mathematical objects—links in the 3-sphere—through very different traditions that rarely interact at the citation level. Neither cites the other, yet the landscape recognizes their shared subject matter (sim=0.838). While these theorems cannot be derived from each other, a low-dimensional topologist aware of both perspectives may find new approaches: Floer-theoretic constraints could inform which links admit the fiber structures that quantum invariants exploit, and conversely, quantum invariant computations could suggest which links to investigate with Floer homology.

B.7 Auxiliary Proof Tool Retrieval

As an auxiliary test of the landscape's ability to guide theorem proving, we evaluate *proof tool retrieval*: given a theorem from paper A that cites paper B, can nearest-neighbor search in the embedding space retrieve theorems from B among the top-K results? We use FAISS [57] for efficient search over the full 1.1M theorem corpus.

Table 16: Proof tool retrieval: retrieving theorems from cited papers via nearest-neighbor search over 1.1M theorems (3,000 queries from held-out citation edges).

	MRR	R@10	R@50	R@100
TF-IDF	0.054	0.158	0.269	0.314
Stage 1 (whitened)	0.060	0.183	0.321	0.369
Stage 2a (whitened)	**0.063**	**0.195**	**0.381**	**0.448**

Stage 2a reaches Recall@100 of 44.8% over one million theorems, vs. 31.4% for TF-IDF (+13.4pp). Stage 1 already beats TF-IDF (+5.5pp) without any citation information, so the learned embeddings go beyond lexical overlap. The gap between Stage 1 and Stage 2a grows at higher K (R@100: +7.9pp): citation-aware training pulls related theorems closer without hurting precision at small K.

B.8 Auxiliary Landscape as Proof Guide: Case Studies

We embed five well-known theorems and examine their neighborhoods (Table 17).

Banach retrieves its generalizations (Geraghty contractions) and applications. The Cauchy Integral Formula retrieves Morera's theorem—its converse. Neighborhoods span multiple categories, so a category-restricted search would miss them.

Intra-paper coherence. Across 54,680 papers with 5–50 theorems, consecutive theorems have mean cosine similarity 0.650 vs. 0.261 for random cross-paper pairs (2.5×). Within each paper, the theorem with highest average similarity to all others typically corresponds to the paper's main result.

B.9 Auxiliary Topological Centrality

We define the *topological centrality* of a theorem as the mean cosine similarity to its 50 nearest neighbors in the whitened embedding space, measuring how densely connected a theorem is within the landscape. We compute this for a sample of 100,000 theorems.

Centrality by domain. Commutative algebra (`math.AC`, 0.250) and rings/algebras (`math.RA`, 0.249) have the highest centrality (Table 18)—algebraic structures produce tightly clustered theorems. Algebraic

Table 17: Nearest neighbors of foundational theorems in the landscape (Stage 2a). The landscape recovers mathematically meaningful related results.

Query Theorem	Top Neighbors	Category Mix
Banach Fixed Point	Geraghty contractions, iterated contractions on complete metric spaces	FA (40%), GN (20%), OC (14%)
Central Limit Theorem	Sub-Gaussian concentration, Berry-Esseen bounds, moment inequalities	ST (28%), PR (28%), CO (8%)
Hahn-Banach	Convex functionals on vector spaces, extension of linear operators	FA (44%), OC (10%), AP (8%)
Cauchy Integral	Morera's theorem, Schröder linearization, holomorphic function properties	PR (20%), CV (18%), AG (8%)
Fund. Thm. of Algebra	Reciprocal polynomials, root continuity, polynomial factorization	CV (24%), GR (10%), CA (6%)

topology (`math.AT`, 0.207) and symplectic geometry (`math.SG`, 0.208) have the lowest, their theorems spread across the landscape by connections to many other areas.

Table 18: Mean topological centrality by category (Stage 2a, whitened, top and bottom 5).

Highest Centrality			Lowest Centrality		
Category	Mean	n	Category	Mean	n
math.AC	0.250	1,940	math.AT	0.207	2,185
math.RA	0.249	2,727	math.SG	0.208	1,124
math.CV	0.237	1,809	math.SP	0.209	618
math.NT	0.234	7,818	math.DS	0.210	4,254
math.CA	0.233	2,351	math.KT	0.212	515

Temporal trends. Mean centrality declines from 0.235 (2010–2012) to 0.224 (2024–2025). As subfields emerge and specialize, the average theorem drifts farther from its neighbors.

Temporal evolution of fields. We analyze how mathematical fields evolve in the landscape by examining three dimensions across 3-year windows from 2005 to 2025 (Table 19). First, relative growth rates reveal the shifting focus of mathematical research: general topology (`math.GN`, 243×), statistics theory (`math.ST`, 187×), and mathematical logic (`math.LO`, 134×) have grown fastest relative to their 2005–2014 baselines, while K-theory (`math.KT`, 17×) and complex variables (`math.CV`, 25×) grew most slowly.

Table 19: Temporal evolution of selected fields: centrality by 3-year window (Stage 2a, whitened, 50K sample).

Category	08–10	11–13	14–16	17–19	20–22	23–25
math.NT	0.255	0.306	0.215	0.205	0.209	0.205
math.CO	0.191	0.205	0.205	0.196	0.204	0.201
math.DG	0.223	0.225	0.217	0.221	0.205	0.201
math.AP	0.199	0.201	0.213	0.209	0.201	0.200
math.FA	0.185	0.204	0.197	0.215	0.203	0.201

Second, field-level centrality trajectories reveal non-monotonic dynamics. Number theory (`math.NT`) exhibits a striking pattern: centrality rises sharply from 0.255 (2008–2010) to 0.306 (2011–2013)—driven by a burst of activity in q-analog special functions and Apostol-type polynomials—before declining to 0.205 (2023–2025), reflecting subsequent diversification within the field. In contrast, functional analysis (`math.FA`) shows a centrality peak in 2017–2019 (0.215), corresponding to activity in approximate amenability and Banach algebra characterizations. These non-monotonic trajectories demonstrate that the landscape captures field-specific dynamics beyond the global diversification trend.

Third, we measure centroid drift—the cosine distance between a field's mean embedding in 2005–2014 versus 2015–2025—to quantify how much each field has shifted in the embedding space. Number theory shows the largest drift (0.75), consistent with the transformation of the field through developments in arithmetic geometry and analytic methods. Differential geometry shows the smallest drift (0.39) among the top fields, reflecting greater thematic stability.

Seminal theorem identification. We combine topological centrality with temporal information to identify theorems that seeded subsequent lines of research. We define the *seminal score* of a theorem as

$$S(t) = C(t) \cdot F(t) \cdot T(t), \tag{2}$$

where $C(t)$ is the topological centrality (mean cosine similarity to 50 nearest neighbors), $F(t)$ is the follower ratio (fraction of the 50 nearest neighbors published *after* theorem t), and $T(t) = (y_{\max} - y_t + 1)/(y_{\max} - y_{\min} + 1)$ is the temporal priority (rewarding earlier publication). A high seminal score indicates a theorem that appeared early, achieved high centrality, and attracted a dense cluster of subsequent work—the signature of a foundational contribution.

Table 20: Top seminal theorems per category (Stage 2a, whitened, 100K sample). Year: publication year; C: centrality; F: follower ratio; S: seminal score.

Category	Year	C	F	S	Representative Content
math.NT	2010	0.481	0.98	0.419	q-analog identities for generalized polynomials
math.RA	2012	0.455	0.98	0.346	Novikov-Poisson algebra constructions
math.AC	2011	0.416	1.00	0.347	Depth of monomial ideal intersections
math.CO	2008	0.344	1.00	0.344	Total domination in trees
math.DG	2010	0.407	0.90	0.325	Conformal changes on 3-Sasakian manifolds
math.CA	2014	0.514	0.92	0.315	Maximal operators on Hardy spaces
math.LO	2012	0.390	0.92	0.291	Knot equivalence via ambient isotopy
math.FA	2010	0.333	0.94	0.279	Approximate contractibility of Banach algebras

The top scorers (Table 20) are q-analog special function theorems from 2010–2011 in number theory ($S = 0.419$), which seeded follow-up work on generalized Bernoulli, Euler, and Genocchi polynomials. In rings and algebras, Novikov-Poisson constructions from 2012 ($S = 0.346$) preceded the Hom-algebra wave. The combinatorics entry—total domination in trees, 2008—has $F = 1.00$: all 50 nearest neighbors were published later.

We note that the seminal theorems identified are specific to the arXiv corpus (primarily 2008–2025), and thus capture "seeds" within this time window rather than historically foundational results. The top 1% seminal score threshold is 0.160, with a mean of 0.022, confirming that high seminal scores are rare and selective.

Independence from citation counts. The seminal score has Spearman $r = 0.31$ with paper citation counts ($p \approx 0$). Centrality measures where a theorem sits; citations measure how often it is acknowledged. Among 100,000 theorems, 3,035 are "hidden gems" (top 5% seminal score, bottom 50% citations) and 307 are "popular but peripheral" (high citations, low seminal score).

B.10 Auxiliary Ablation Studies and Controls

Do the embeddings capture mathematical structure, or just vocabulary overlap? We run three ablations.

B.10.1 Vocabulary and Notation Ablation

To test whether the embeddings rely on mathematical keyword matching rather than deeper structural features, we perform three experiments.

Keyword masking. We mask common keywords ("theorem," "group," "space," "function") and re-embed 10,000 theorems. Accuracy drops only 3.5pp (47.8%→44.3%); cosine similarity between masked and unmasked embeddings is 0.94; nearest-neighbor overlap is 63.8%. The embeddings do not depend on a few domain-indicative keywords.

Paraphrase stability. On 5,000 theorem pairs: variable renaming ($x \to y$) gives cosine similarity 0.966; clause reordering, 0.975; LaTeX whitespace changes, 0.896. Content-preserving perturbations barely move the embedding. The lower whitespace stability is a tokenizer artifact.

Same-domain vs. cross-domain similarity. Same-domain pairs average 0.379 vs. 0.268 for cross-domain ($p < 10^{-10}$). The gap (0.11) is real but modest—the embedding captures within-domain structure, not just broad category separation.

B.10.2 Citation Artifact Ablation

Does the earlier Stage 2a pipeline just memorize citation patterns? We compare text-only, citation-only, and combined on a 50K sample (Table 21).

Table 21: Citation artifact ablation: text-only (Stage 1) vs. citation-only (DeepWalk) vs. combined (Stage 2a) on 50K sample.

Method	Accuracy	F1	R@100
Stage 1 (text only)	0.431	0.303	0.502
DeepWalk (citation only)	0.271	0.284	0.286
Stage 2a (combined)	**0.567**	**0.467**	**0.714**

Stage 2a beats both text-only and citation-only baselines in this auxiliary run; the signals are complementary. Stage 1 alone already outperforms DeepWalk on classification—theorem text carries structural information beyond what citations encode.

Bridge vocabulary overlap. All 498 cross-domain bridges have Jaccard vocabulary overlap below 0.15 (lowest: math.AG↔math.HO at 0.011). The bridges are not notational artifacts.

We note an important caveat throughout: our citation signals operate at the *paper* level rather than the *theorem* level. Two theorems from citing papers are treated as related even if they address different aspects. This paper-level granularity introduces noise into the training signal and may cause the embedding to reflect co-occurrence patterns within papers rather than direct logical dependencies between individual theorems.

B.10.3 Corpus Scale Sensitivity

We evaluate how corpus size affects embedding quality by training on random subsets of the 50K base corpus (Table 22).

Table 22: Effect of corpus scale on domain classification (50K base).

Fraction	N	Accuracy	F1
10%	5K	0.476	0.340
25%	12.5K	0.498	0.365
50%	25K	0.522	0.398
75%	37.5K	0.551	0.444
100%	50K	0.571	0.464

Accuracy improves monotonically with scale, but even 10% of the corpus gives 47.6% (14× random). Density statistics are stable across scales (mean ≈ 0.626, std ≈ 0.038).

Domain stability. Removing any one of the 10 largest subcategories drops accuracy by 11–14pp (worst: functional analysis, −13.8pp), but no removal kills discriminative power.

KNN k-sensitivity. Accuracy is stable across $k \in \{1, 3, 5, 10, 20, 50\}$ (range 55.6%–56.9%).

B.11 Auxiliary Baseline Comparison

In the earlier auxiliary baseline suite, we compare the three historical stages against both classical baselines (TF-IDF, RoBERTa-base, SciBERT) and state-of-the-art sentence embedding models (E5-large-v2 [58], GTE-large [59], BGE-large [60], SPECTER2 [18], all-MiniLM-L6-v2) on a unified 50K sample (Table 23). All embeddings are whitened; retrieval is over the full 1.1M corpus (available only for our models, which have pre-computed full-corpus embeddings).

Table 23: Baseline comparison on 50K sample. Domain classification: KNN k=5, cosine distance, 30 classes. Link prediction: AUC/AP on held-out citation edges. Retrieval: over full 1.1M corpus (N/A for models without full-corpus embeddings).

Method	Acc	F1	AUC	AP	R@10	R@100
Stage 2a (ours)	**0.578**	**0.499**	**0.808**	**0.856**	**0.195**	**0.448**
Stage 2b (ours)	0.552	0.466	0.793	0.838	0.167	0.355
Stage 1 (ours)	0.481	0.400	0.755	0.800	0.183	0.369
E5-large-v2	0.475	0.408	0.733	0.779	—	—
BGE-large	0.467	0.393	0.715	0.769	—	—
MiniLM-L6	0.464	0.385	0.728	0.781	—	—
GTE-large	0.460	0.385	0.726	0.780	—	—
SPECTER2	0.442	0.374	0.720	0.771	—	—
TF-IDF (50K, SVD)	0.328	0.241	0.746	0.783	0.085	0.185
RoBERTa-base	0.319	0.252	0.706	0.747	0.078	0.152
SciBERT	0.266	0.184	0.653	0.683	0.029	0.085

Stage 2a leads on every metric. The more interesting comparison is with SOTA general-purpose embeddings: E5-large-v2, the strongest among them, reaches 0.475 accuracy—below even our Stage 1 (0.481), which uses no citation information at all. Domain-specific continued pre-training on theorem text outweighs the advantage of larger, more diverse training corpora. SPECTER2, despite being designed for scientific documents, performs worst among the SOTA models (0.442), echoing SciBERT's poor showing: general scientific pre-training does not transfer well to mathematical LaTeX.

The gap between Stage 1 and Stage 2a (+9.7pp accuracy, +5.3pp AUC) shows what citation calibration adds on top of domain-adapted language modeling. On link prediction, TF-IDF (0.746 AUC) actually outperforms E5 (0.733) and BGE (0.715)—lexical overlap in mathematical notation is a surprisingly strong signal for citation relationships, and general-purpose embeddings lose this signal through their broader training.

References

[1] Chaomei Chen, Fidelia Ibekwe-SanJuan, and Jianhua Hou. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, 61(7):1386–1409, 2010.

[2] Michael H. MacRoberts and Barbara R. MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342–349, 1989.

[3] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[4] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

[5] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[6] Daniel Cummings and Marcel Nassar. Structured citation trend prediction using graph neural networks. *arXiv preprint arXiv:2104.02562*, 2021.

[7] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.

[8] Huajian Xin, Daya Guo, Zhihong Shao, et al. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

[9] Google DeepMind. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 2025.

[10] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[11] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359(6379):eaao0185, 2018.

[12] Aaron Clauset, Daniel B. Larremore, and Roberta Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017.

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of EMNLP*, 2019.

[14] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*, 2021.

[15] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben McGrew, and Dongwon Lee. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340*, 2021.

[16] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of ACL*, 2020.

[17] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of EMNLP*, 2022.

[18] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of EMNLP*, 2023.

[19] UW Math AI. Semantic search over 9 million mathematical theorems. *arXiv preprint arXiv:2602.05216*, 2026.

[20] Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural language. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

[21] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of Chemical Physics*, 151(7):070902, 2019.

[22] Rafael C. Bernardi, Marcelo C. R. Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta*, 1850(5):872–877, 2015.

[23] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2014.

[24] José N. Onuchic, Zaida Luthey-Schulten, and Peter G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997.

[25] Francesco Mallamace, Carmelo Corsaro, Domenico Mallamace, Sebastiano Vasi, Cirino Vasi, Piero Baglioni, Sergey V. Buldyrev, Sow-Hsin Chen, and H. Eugene Stanley. Topological and physical link between the energy landscape and the kinetics of protein folding and unfolding. *Proceedings of the National Academy of Sciences*, 113(12):3159–3163, 2016.

[26] Alex Alemi, François Chollet, Niklas Een, Geoffrey Irving, Christian Szegedy, and Josef Urban. Deepmath – deep sequence models for premise selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[27] Kshitij Bansal, Sarah Loos, Markus Rabe, Christian Szegedy, and Stewart Wilcox. Holist: An environment for machine learning of higher-order theorem proving. In *International Conference on Machine Learning (ICML)*, 2019.

[28] Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. Formal mathematics statement curriculum learning. In *International Conference on Machine Learning (ICML)*, 2022.

[29] Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[30] Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[31] Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Anima Anandkumar. A survey on deep learning for theorem proving. *COLM*, 2024.

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*, 2019.

[33] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, 2021.

[34] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[35] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[36] Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Computing Surveys*, 56(6), 2024.

[37] Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Naturalprover: Grounded mathematical proof generation with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[38] Robert Adler, John Ewing, and Peter Taylor. Citation statistics. *Statistical Science*, 24(1):1–14, 2009. Joint IMU/IMS/ICIAM report.

[39] Stephen J. Bensman, Lawrence J. Smolinsky, and Alexander I. Pudovkin. Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, 61(7), 2010.

[40] Eugene Garfield. The 200 'pure' mathematicians most cited in 1978 and 1979. *Essays of an Information Scientist*, 5:664–675, 1982.

[41] Sean A. Myers, Peter J. Mucha, and Mason A. Porter. Mathematical genealogy and department prestige. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(4):041104, 2011.

[42] Aleksandr Semenov, Alexander Veremyev, Alexey Nikolaev, Eduardo Pasiliao, and Vladimir Boginski. Network-based indices of individual and collective advising impacts in mathematics. *Computational Social Networks*, 2019.

[43] Flavio B. Gonzaga, Valmir Carneiro Barbosa, and Geraldo Xexeo. The network structure of mathematical knowledge according to the Wikipedia, MathWorld, and DLMF online libraries. *arXiv preprint arXiv:1212.3536*, 2012.

[44] Monika Cerinsek and Vladimir Batagelj. Network analysis of Zentralblatt MATH data. *arXiv preprint arXiv:1409.1726*, 2014.

[45] David Corfield. *Towards a Philosophy of Real Mathematics*. Cambridge University Press, 2003.

[46] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[47] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[49] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. In *arXiv preprint arXiv:2103.15316*, 2021.

[50] Viacheslav Dirks, Daniil Derkach, Nitya Agarwal, Valérie Frappier, Aditya Bhatt, Debsindhu Bhowmik, Kostas Karanasos, and Yann Vander Meersche. Diffusion on language model embeddings for protein sequence generation. *arXiv preprint arXiv:2403.03726*, 2024.

[51] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pages 625–635. Springer, 2021.

[52] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[53] Emanuele Caputo. Geometric characterizations of PI spaces: an overview of some modern techniques. *arXiv preprint arXiv:2501.19132*, 2025.

[54] Nicolò De Ponti, Sara Farinelli, and Ivan Yuri Violo. Pleijel nodal domain theorem in non-smooth setting. *arXiv preprint arXiv:2307.13983*, 2023.

[55] Fraser Binns and Subhankar Dey. Floer homology, clasp-braids and detection results. *arXiv preprint arXiv:2405.11224*, 2024.

[56] Daniel López Neumann and Roland van der Veen. Non-semisimple $\mathfrak{sl}_2$ quantum invariants of fibred links. *arXiv preprint arXiv:2407.15561*, 2024.

[57] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

[58] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[59] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[60] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.