Project Tavnit

Data Science & Computational Humanities Briefing

March 2026 · tavnit.ashestoaltar.com

1. What This Project Is

Project Tavnit applies network analysis, spectral methods, and computational topology to a large-scale biblical intertextual graph. The dataset contains **1,411,192 edges** across **17 independently constructed layers**, connecting 3,151 paragraph-level nodes. Findings are validated through permutation null models, density-matched controls, cross-corpus comparisons, and a 3-LLM adversarial protocol. The project maintains 22 logged corrections and 51 methodology decisions.

2. The Dataset

Node Resolution

Nodes correspond to structural paragraph divisions: *petuchah/setumah* (Hebrew Masoretic open/closed sections) for OT and *kephalaia* (Greek manuscript chapter divisions) for NT. Yields 3,151 nodes. These boundaries predate modern chapter-and-verse numbering and reflect ancient editorial intent. Choice of resolution matters: the Bible's spectral profile sharpens dramatically from book-level (PC1 = 44%) to paragraph-level (PC1 = 2%), while Greek comparison literature remains flat (PC1 = 33% → 30%).

Edge Layers

Layer	Edges	Type	Sources
tsk	606,795	Curated cross-reference	Treasury of Scripture Knowledge
evangelical	254,598	LLM-extracted	9 teachers (Henry, MacArthur, Begg, Guzik, Baucham, Sproul, Smith, Wiersbe, McGee)
midrashic	170,832	LLM-extracted	26 Torah-positive teachers
strongs	129,338	Shared lexeme	Strong's concordance numbers
sefaria-commentary	72,185	Structured cross-reference	1,369 classical Jewish commentators (Sefaria API)
leitwort	69,235	Shared vocabulary	Hebrew/Greek shared-root co-occurrence
sefaria	42,499	Curated links	Sefaria structured cross-references
patristic	26,939	Co-citation	639 Church Fathers (CCEL ThML, HCF)
civic-political	17,664	LLM-extracted	9 political philosophy sources (Calvin, Edwards, Winthrop, OLL)
beale-carson	11,475	LLM-extracted + curated	25 NT scholars (Beale & Carson commentary)
chiastic	4,255	Structural	Chiastic structure pivot pairs
review	3,582	Mixed	Pending attribution review
precept-austin	551	Curated cross-reference	Precept Austin ministry
liturgical	434	Liturgical mapping	4 liturgical tradition sources
catholic-resources	369	Curated cross-reference	Catholic cross-reference database
bible-ca	324	Curated cross-reference	Bible.ca cross-reference lists
haftarah	117	Liturgical mapping	6 haftarah tradition sources

Total: 1,411,192 edges. All layers independently constructed. Connection registry enforces attribution validation (prevents cross-contamination between layers). Embeddings: BGE-M3 (1024-dim), 0 NULL embeddings across 2,694,361 chunks. The civic-political layer (17,664 edges) is included in the combined matrix but has not been individually analyzed in the current findings.

3. Methods

Core Algorithms

Five algorithms applied across nine matrix configurations (3 layers × 3 normalizations):

- **PCA** — 50-component extraction on paragraph-level adjacency matrices (raw, binary, length-normalized)

- **Leiden community detection** — genre purity as quality metric (76–92%, p = 0.000 permutation null)
- **Spectral analysis** — eigenvalue distribution, participation ratio, spectral entropy
- **Sinusoid fitting** — testing cyclicity hypothesis (rejected: R^2 max = 0.31 vs. 0.7 threshold)
- **Gravitational modeling** — center-of-mass computation per book, bootstrap CI

Null Models

- **Permutation null** — standard protocol for axis engagement: randomize edges within tradition-specific subgraph, recompute PCA cosine, 100 trials per tradition. z-score against null distribution. $|z| > 2.0$ threshold.
- **Erdős-Rényi baseline** — random graph at matched density for spectral comparison
- **Density-matched subsampling** — 50 rounds × 100 null trials, controlling for edge-count effects between layers/traditions
- **Layer ablation** — systematic removal of each layer to test structural contribution
- **Cross-corpus control** — 31 Greek works, 6 genres, 18,809 nodes, matched paragraph-level resolution

Validation Protocol

- **3-LLM adversarial protocol (MD-50):** Grok, Claude, GPT-4o. Two modes: (a) *review* — evaluate claims with full context; (b) *interpret* — blind data interpretation with zero project context. Validation LLMs operate with zero project context and no shared conversation history with the project's analytical systems. 30+ reports from 18+ briefs.
- **8 adversarial checkpoints** — internal Tamor system challenges each major finding before it propagates.
- **Jackknife bootstrap** — used for teacher-resampled metrics (standard bootstrap causes duplication artifact when unit of resampling = teacher).
- **Per-author null protocol** — individual commentator isolation: filter by layer + attribution, test all 13 axes, 100 null trials. Produces per-commentator "structural fingerprints."

4. Confirmed Findings

4.1 Non-Random Structure

Leiden community detection recovers genre-appropriate book groupings at 76–92% purity ($p = 0.000$ permutation null). Structure is not an artifact of edge density or layer selection.

4.2 Genre-Respecting Communities

Detected communities correspond to recognized biblical genre categories (Torah, prophets, wisdom, narrative) without genre labels being supplied to the algorithm.

4.3 Characteristic Dimensionality

The paragraph-level combined matrix (C_length_norm) distributes variance with no dominant axis: PC1 ≈ 2%, participation ratio ≈ 93. **13 validated interpretable axes** emerge above the noise floor, each with distinct loading profiles confirmed through cross-validation.

> **COR-21 correction:** Earlier versions reported "11 components for 80% of variance." This was a `np.searchsorted` truncation artifact in `run_phase1.py` (10-component extraction, searchsorted on 11.6% cumulative returns index 10, +1 = 11). True 80% threshold requires 200+ components. Broken-stick model (Jackson 1993) identifies 1–2 dominant axes. Parallel analysis (Horn 1965) finds all 50 components significant, but Erdős-Rényi null is too weak for structured graphs. Proper metrics: participation ratio (5–93 across configurations) and spectral entropy. The 13 interpretive axes are validated by loading profiles, ablation, four-bridge test, and blind review — not by variance counting.

4.4 No Cyclicity

Torah reading cycle tested via sinusoidal fitting: R^2 max = 0.31 (threshold 0.7). The annual reading cycle is a liturgical framework, not a property of the text's connection topology.

4.5 Layer Robustness / TSK Ablation

Removing TSK (47.8% of all edges): genre purity 89.4% → 89.1%, flat-spectrum preserved. TSK perturbs all 11 axes when tested in isolation (structural dominance), but removing it does not collapse the structure — other layers carry redundant signal. Strong's is most edge-efficient (5/11 axes recovered solo). Leitwort is fully redundant (0/11 axes unique). Sefaria punches above its weight (2/11 from 1.6% of edges).

4.6 Hub Concentration in Genesis

With TSK: Genesis + Deuteronomy are hubs. Without TSK: all top-10 hubs are Genesis paragraphs. Deuteronomy's hub status is TSK-dependent (COR-07). Genesis provides the vocabulary foundation independent of cross-reference scholarship.

4.7 Structural Keystone: Deuteronomy 28:15–68

The *Tokhekhah* (covenant curses) scores extreme on 9/11 originally tested axes. Sequential-discontinuity boundaries #1 and #2 in full OT. Rank 491/3,151 per-verse (84th percentile); raw extremity partly reflects passage length. Pending retest on 13-axis framework.

4.8 Four-Bridge Substrate Test

Four independent cross-testament data sources (TSK, Beale-Carson, Greek leitwort A, Greek leitwort B) tested for axis recovery. Result: four-category partition of the 13 axes:

Category	Count	Axes	Criterion
Substrate	5	PC1, 2, 5, 6, 9	Recovered by 2+ independent bridges
Tradition-specific	4	PC3, 4, 10, 11	Recovery depends on which bridge
Tradition-engaged	3	PC7, 8, 19*	Commentary traditions actively engage (reinforce or perturb)
Genre bedrock	1	PC14*	All traditions neutral; text-internal genre boundary

PC9 is triple-convergent (recovered by 3 of 4 bridges). The four-category partition supersedes the earlier three-way model (substrate / tradition-specific / unengaged). *PC14 and PC19 were identified through an extended-axis investigation (3-LLM blind interpret protocol) after the original 11-axis framework was established; see COR-21 addendum for dimensionality context.

4.9 Small-World Topology

The combined network exhibits small-world properties: high clustering coefficient relative to random graphs at matched density, short average path lengths.

4.10 Spectral Uniqueness

Cross-corpus comparison against 31 Greek literary works (Homer, Plato, Aristotle, Euripides, etc.; 6 genres; 18,809 nodes at paragraph resolution):

Metric	Bible	Greek Literature	Ratio
Participation ratio (length-norm)	18.1	6.0	3.0×
Persistent H1 loops	129	6	21×
Mean loop persistence	1.667	0.783	2.1×
Topology scaling	$N^{2.7}$ power law	0 loops at scale	—

The gap survives all tested controls: density matching (3× gap persists at equal density), lemmatization (CLTK stem-50% narrows ratio from 3.2× to 1.8×), resolution scaling, per-genre comparison (all 6 Greek genres PR < 9 vs. Bible 18.1). LXX (Greek Septuagint) shows Bible-like spectral profile (PR = 12.6–16.6), confirming the effect transcends language.

4.11 Kabbalistic Fingerprint (Per-Author Nulls)

Five medieval Jewish commentators tested individually (all 13 axes, 100 null trials each):

Author	Century	Method	Edges	Reinforced	Perturbed
Rashi	11th	Peshat/halakhic	4,095	9	0
Ibn Ezra	12th	Peshat/grammatical	1,634	2	0
Ramban	13th	Halakhic/kabbalistic	1,795	1	1
Rabbeinu Bahya	14th	Kabbalistic/ethical	3,334	2	1
Abarbanel	15th	Political/philosophical	1,806	4	0

Key result: Bahya PC10 z = −11.39 (extreme perturbation). This single commentator drives the aggregate Sefaria layer's z = −58 on PC10. Peshat commentators (Rashi z = +3.34, Abarbanel z = +3.16) reinforce PC10. Profiles cluster by hermeneutical method, not century. PC8 = universal axis (5/5 reinforce). Rashi = exceptional breadth (9/13).

4.12 Patristic Phase Transition

Diachronic decomposition of patristic layer (density-matched, 50 rounds × 100 null trials):

Era	Edges	Reinforced	Perturbed
Pre-Nicene (3,602 edges)	3,602	0	3 (PC3, PC6, PC10)
Post-Nicene (matched 3,602)	3,602	4.2 (mean)	0 (all 50 rounds)
Post-Nicene (full 7,014)	7,014	6	0

The qualitative asymmetry (zero perturbation for post-Nicene across all 50 density-matched rounds) is robust. Pre-Nicene disrupts; post-Nicene reinforces. PC3 flip is density-independent.

4.13 Bootstrap Confidence Intervals

Metric	Estimate	95% CI	Method
Midrashic convergence rate	9.5%	7.4–11.5%	Jackknife (26 teachers)
Midrashic-leitwort Jaccard	0.50	0.48–0.52	Bootstrap (1,000)
Sinai gravity (all layers)	22.5	21.8–23.1	Bootstrap (1,000)
Sinai gravity (leitwort)	25.8	25.6–26.0	Bootstrap (1,000)

4.14 Method Artifact Resolution

Discovery (Session 15): Matthew Henry co-citation perturbs PC9/PC10 (z = −7.69, −4.06) while Henry LLM extraction reinforces them (z = +2.92, +3.16) — same author, same text, opposite results by extraction method. Initially suspected as inherent co-citation vs. LLM confound.

Resolution (Sessions 16–19): three targeted tests showed the perturbation is **density-dependent, not method-inherent**:

1. **Density-matched Henry:** At 4,778 pairs (matched to LLM count), co-citation PC9 z = −1.59, PC10 z = −0.94 (unstable). The z = −7.69 was density-amplified from 73K pairs.

2. **Rashi co-citation:** 3,358 Sefaria-sourced co-citation pairs REINFORCE PC9 z = +2.92, PC10 z = +4.30. Breaks "co-citation inherently perturbs."
3. **Per-author null (Bahya):** The PC10 perturbation is author-specific (Bahya z = −11.39), not method-specific. Co-citation and LLM extraction can produce concordant results for the same tradition.

5. What's Novel

- **Scale and layer independence:** 1.4M edges across 17 layers is among the largest biblical intertextual datasets assembled, with each layer independently constructed.
- **Paragraph-level resolution:** Using ancient section markers (petuchah/setumah, kephalaia) rather than modern chapter-and-verse divisions as the unit of analysis.
- **Four-category axis partition:** substrate, tradition-specific, tradition-engaged, genre bedrock — derived empirically from four-bridge convergence test.
- **Per-author structural fingerprints:** Individual commentators tested across all 13 axes via permutation null, producing datable structural profiles.
- **Spectral comparison against classical literature:** Matched-methodology cross-corpus control demonstrates the Bible's spectral profile is distinct from 31 Greek literary works.
- **3-LLM blind interpret protocol:** Zero-context data interpretation by three independent LLMs as a guard against project-aware confirmation bias.
- **Diachronic tradition decomposition:** Density-matched era-by-era testing of interpretive traditions produces temporally anchored structural fingerprints.
- **22 corrections, 51 methodology decisions:** Full adversarial audit trail as part of the research output.

6. Open Questions

- **Cross-corpus null:** Would a comparably large commentary tradition on a secular corpus (Homer + scholia) produce similar "universal axis" patterns? Untested and identified as the most dangerous threat to finding specificity.
- **Tokhekhah on 13 axes:** The structural keystone was tested on 11 axes. Retest pending on the full 13-axis framework.
- **Configuration model null:** Erdős-Rényi is too weak for parallel analysis on structured graphs. A configuration-model null (preserving degree sequence) would provide a more stringent spectral baseline.
- **Density-matched multi-tradition comparison:** Unblocked after method-artifact resolution. Will normalize all five traditions to equal edge count before axis engagement comparison.
- **Kabbalistic fingerprint beyond medieval period:** Zohar-influenced and Hasidic commentary data would test whether the PC10 perturbation generalizes beyond Bahya/Ramban.
- **Broader corpus comparison:** How does the Bible's spectral profile compare to other long-edited, multi-author corpora (Talmud, Quran with tafsir, Hindu epics)?
- **Causal inference for patristic phase transition:** The pre-Nicene/post-Nicene structural shift is temporally localized but causally underdetermined. Canon stabilization, genre/preservation shifts, citation practice changes, and theological development are all confounded. Formal causal-inference methods (e.g., interrupted time-series or difference-in-differences with appropriate controls) would strengthen the finding beyond descriptive correlation.

7. Project Details

Item	Value
Total edges	1,411,192
Independent layers	17
Paragraph nodes	3,151
Library files	87,532 (2,694,361 chunks)
Teachers / commentators	2,000+ (9 evangelical, 26 midrashic, 1,369 Sefaria, 639 patristic, 25 Beale-Carson)
Corrections logged	22
Methodology decisions	51
Validation reports	30+ (3-LLM adversarial protocol)
Database	SQLite (local, single-user)
Embedding model	BGE-M3 (1024-dim, NVIDIA RTX 3060)
Key libraries	NumPy, SciPy, NetworkX, scikit-learn, ripser (persistent homology)
LLM extraction	Distributed (3 machines × 8 threads, ~122 files/min)