Textures of Modern Formal Mathematics

J. R. Landers

Remarks on LeanDojo Benchmark 4

Formal libraries can now be studied not only theorem by theorem, but as empirical mathematical objects. Machine-readable traces and AI-assisted analysis make it possible to pose, test, and revise a question about thousands of proofs in an afternoon. That speed is useful only when suggestive pictures are kept separate from measured evidence.

Formal proofs have at least two vocabularies. A tactic sequence records *how* a proof proceeds by rewriting, simplifying, splitting cases, introducing hypotheses, or building intermediate facts. The premises named by those tactics record more of *what* it is about: lists, filters, measure theory, category theory, or polynomial algebra. Predictive systems often combine these signals; separating them instead distinguishes procedure from subject.

Two soft topic models fitted to 10,000 human-written Lean 4 tactic proofs find stable structure in each view, but little alignment *between* them. Common proof idioms appear to travel across subjects, while one subject admits several modes of formal argument.

Two representations. From all 52,187 tactic-traced proofs in the benchmark's random split, a seed-0 uniform sample selects $n = 10{,}000$ proofs, indexed by i. Let L_i be the number of tactics in proof i, $(h_{i1}, \dots, h_{iL_i})$ its tactic-head sequence, and P_i the set of premises explicitly named in its traced tactics. The *style* document contains tactic unigrams h_{ij} and adjacent bigrams $(h_{ij}, h_{i,j+1})$; the *domain* document contains each $p \in P_i$ and its top-level namespace. Each collection is TF–IDF weighted, giving nonnegative proof-by-feature matrices $X^{(s)}$ and $X^{(d)}$, where s and d denote the style and domain views. If $c_{iw}^{(v)}$ counts feature w in proof i for view $v \in \{s, d\}$ and $d_w^{(v)} = |\{i : c_{iw}^{(v)} > 0\}|$, the unnormalized entry is

$$\widetilde{X}_{iw}^{(v)} = \left(1 + \log c_{iw}^{(v)}\right)\left(\log \frac{1+n}{1+d_w^{(v)}} + 1\right) \quad (c_{iw}^{(v)} > 0), \qquad X_{i\cdot}^{(v)} = \frac{\widetilde{X}_{i\cdot}^{(v)}}{\|\widetilde{X}_{i\cdot}^{(v)}\|_2}.$$

Here $|\cdot|$ denotes set size and $\|\cdot\|_2$ Euclidean length; entries with $c_{iw}^{(v)} = 0$ are zero, as is any row with zero length. Sublinear term frequency prevents repeated calls from growing linearly in importance; inverse document frequency emphasizes features that distinguish proofs. The matrices are factorized independently:

$$\min_{W^{(v)}, H^{(v)} \geq 0} \left\| X^{(v)} - W^{(v)} H^{(v)} \right\|_F^2, \qquad v \in \{s, d\},$$

where the Frobenius norm $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ sums squared matrix entries. Here $W_{ik}^{(v)}$ is proof i's loading on topic k, while $H_{kw}^{(v)}$ is topic k's loading on feature w. With K_v topics in view v, row normalization gives $\theta_{ik}^{(v)} = W_{ik}^{(v)} / \sum_{\ell=1}^{K_v} W_{i\ell}^{(v)}$; the dominant topic $z_i^{(v)} = \arg\max_k \theta_{ik}^{(v)}$ is used only for comparison and coloring. Subsampling stability and the reconstruction-curve elbow select $K_s = 8$ style topics and $K_d = 10$ domain topics, an exploratory resolution rather than a true number of proof kinds. The normalized entropy

$$E_i^{(v)} = -\frac{1}{\log K_v} \sum_{k=1}^{K_v} \theta_{ik}^{(v)} \log \theta_{ik}^{(v)} \in [0, 1]$$

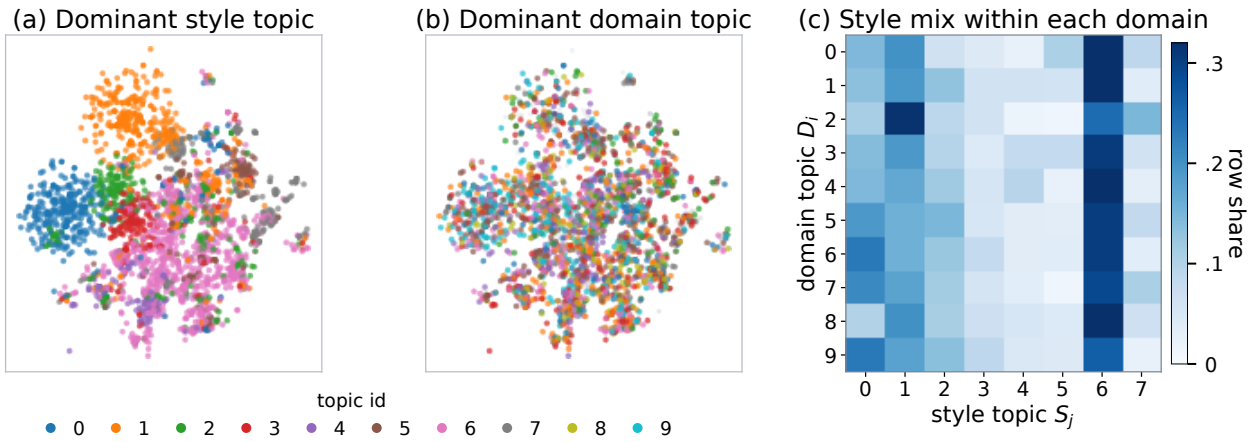

Figure 1: A fixed 3,000-proof sample of one style-derived t-SNE slice, colored by style (a) and domain (b), plus the full row-normalized domain–style table (c). Gray means no retained topic. Examples: S_0 `rw`, S_1 `simp`, S_6 `have/apply/refine'`, S_7 `rfl/cases`; D_0 category theory, D_2 lists, D_4 measure theory, D_8 filters.

records whether a proof concentrates on one component or mixes several; retaining this soft description avoids turning the dominant label into an ontological claim. As usual, $0 \log 0$ is interpreted as zero.

Measuring alignment. For the $n_c = 8{,}882$ proofs assigned in both views, let $N_{ab} = |\{i : z_i^{(d)} = a, z_i^{(s)} = b\}|$ count proofs in domain topic a and style topic b. Write $p_{ab} = N_{ab}/n_c$, $p_{a+} = \sum_b p_{ab}$, and $p_{+b} = \sum_a p_{ab}$. Mutual information compares the joint table with the independent table:

$$I(D;S) = \sum_{a,b:p_{ab}>0} p_{ab} \log \frac{p_{ab}}{p_{a+}p_{+b}}.$$

With $H(D) = -\sum_a p_{a+} \log p_{a+}$ and $H(S) = -\sum_b p_{+b} \log p_{+b}$,

$$\mathrm{NMI}(D,S) = \frac{I(D;S)}{\sqrt{H(D)H(S)}}.$$

Here D and S denote the random domain- and style-topic labels, and H is Shannon entropy (natural logarithms are used throughout). Because a finite contingency table has positive mutual information even under random labels, adjusted mutual information (AMI) subtracts its expected value under fixed marginals and scales by the corresponding entropy ceiling. AMI is therefore zero in expectation under the permutation null and one for identical partitions. Cramér's V supplies a complementary effect size based on squared cell deviations. Neither statistic uses the plotted coordinates.

Different structures, weak correspondence. Panel (a)'s coordinates and colors both derive from style, so its coherence is expected. The evidence comes from domain colors on the unchanged coordinates and, independently of any projection, the contingency table.

The style topics separate short normalization and rewriting proofs from longer structured arguments: `rw` and `simp` proofs average 1.4 and 1.8 steps, while the `have/apply/refine'` topic averages 8.3. Domain topics instead recover category theory, sets, lists, measure theory, polynomial algebra, filters, and number systems.

Among the 8,882 proofs with nonzero assignments in both views, adjusted mutual information is only 0.015 (normalized mutual information 0.017).

Let $N_{a+} = \sum_b N_{ab}$ and $N_{+b} = \sum_a N_{ab}$ be the row and column totals. Pearson's chi-squared statistic χ^2, its independence-model expected counts E_{ab}, and Cramér's effect size V are given below; the resulting $V = 0.102$ is a small association:

$$\chi^2 = \sum_{a,b} \frac{(N_{ab} - E_{ab})^2}{E_{ab}}, \qquad E_{ab} = \frac{N_{a+} N_{+b}}{n_c}, \qquad V = \sqrt{\frac{\chi^2}{n_c \min(K_s - 1, K_d - 1)}} = 0.102,$$

None of 20,000 permutations that preserve those totals produced so large a statistic (Monte Carlo $p < 5 \times 10^{-5}$), but the effect size says *detectable preference, not close alignment*. The largest within-domain style shares range from 26.2% to 36.6%. Domain topics align much more with mathlib modules than style topics do (AMI 0.241 versus 0.025). An initial 1,940-proof analysis gave nearly identical effect sizes (cross-view AMI 0.0157, $V = 0.113$).

Why this is interesting. Branches are usually related through shared objects and theorems. This comparison points to a second, procedural relation: subjects with disjoint vocabularies may reuse normalization, extensionality, cases, construction, or rewriting, while neighboring subjects support different styles. A map based on proof-style distributions need not reproduce a premise graph; it could identify methodological kinship, a relation based on how arguments are organized.

Subject is not irrelevant: Graph2Tac finds nearby concepts with similar proof structures [4]. Our weak global association leaves room for such local conventions alongside portable idioms. This complements ML4PG's proof-pattern mining [1, 2] and LeanDojo's separation of premise retrieval from tactic generation [3]. SEPIA models tactic sequences as inferred automata [5], while TacticToe learns which tactic or mathematical technique suits a proof state from recorded human proofs [6]; both show why procedure is a meaningful object of study in its own right.

Limitations and next test. The 10,000 proofs are a sample rather than the complete corpus and remain short (median two tactics); annotations omit lemmas used internally by automation; and labels reflect Lean/mathlib conventions. A stronger test would embed state transitions, compare full mixtures, and build branch networks from premises versus procedures. Their disagreement would test for methodological geometry distinct from dependency geometry.

This is a modest example of AI-assisted computational mathematics. Proof assistants guarantee local correctness; analysis asks global questions about practice. AI makes those questions cheaper to iterate, not to validate. The prize is not an automated verdict, but a testable map of formalized mathematical practice.

References

[1] E. Komendantskaya, J. Heras, and G. Grov. "Machine Learning in Proof General: Interfacing Interfaces." EPTCS 118, 2013.

[2] J. Heras and E. Komendantskaya. "Proof Pattern Search in Coq/SSReflect." arXiv:1402.0081, 2014.

[3] K. Yang et al. "LeanDojo: Theorem Proving with Retrieval-Augmented Language Models." NeurIPS Datasets and Benchmarks, 2023.

[4] L. Blaauwbroek et al. "Graph2Tac: Online Representation Learning of Formal Math Concepts." arXiv:2401.02949, 2024.

[5] T. Gransden, N. Walkinshaw, and R. Raman. "SEPIA: Search for Proofs Using Inferred Automata." CADE-25, pp. 246–255, 2015.

[6] T. Gauthier, C. Kaliszyk, J. Urban, R. Kumar, and M. Norrish. "TacticToe: Learning to Prove with Tactics." Journal of Automated Reasoning 65(2), pp. 257–286, 2021.