Contents

Method-Space Has a Small Bounded Basis: A Direct Measurement Across 5.7 Million Documents

Benjamin Dakota Horn[1] [1] *Independent researcher, Prometheus7. Correspondence: psiloceyeben.*

Draft preprint v1 — 2026-06-07

Abstract

We construct a structural spectrometer for scientific methods — a controlled vocabulary of 19 substrate prerequisites and a 12-wall calibration battery that maps any method described in text to a position in a 4,096-coordinate signature lattice. We apply the instrument at full corpus scale to (i) the complete arXiv OAI-PMH metadata dump (2,945,000 papers, 2007–2026) and (ii) the English Wikipedia article corpus (2,782,522 filtered articles), yielding 1,200,768 extracted candidate operators with provenance. We measure the *dominant basis cardinality* of method-space — the number of distinct structural-signature classes holding ≥1% of all extracted operators — at twelve corpus scales from n=100 to $n=2.945\times10^6$. The dominant basis lies in the range [15, 20] across all scales tested and settles to [16, 19] from n=100,000 onward — across every projection (k=5, 9, 12), both corpora, two extractor implementations (heuristic for the saturation curves; model-assisted rich extraction for a recall stress test), and twelve confidence thresholds spanning $\theta \in [0.35, 0.90]$. The top-15 dominant signatures on Wikipedia at n=2.78M and arXiv at n=500,000 form identical sets, differing only in rank order. The compression ratio of method-instances to dominant structural classes is approximately 63,000×. We argue this constitutes the first direct empirical measurement of the dimensionality of scientific method-space and the first empirical test of the unity-of-science thesis at corpus scale. Honesty constraints required by the discipline of the measurement (projection-dependence, candidate-not-discovery, recall-bounded) are stated explicitly and held throughout. All artifacts (code, corpora, extracted operators, signature distributions, queryable lattice) are open and reproducible.

---

1. Introduction

The question whether the methods of the sciences share a unified structural basis, or instead constitute an open-ended thesaurus of independent techniques, has been debated since antiquity. Plato proposed a small finite set of forms underlying observed diversity. Bacon and Comte [6] argued for a hierarchy of sciences reducible to a common method. Carnap [1] sought a unified language for science; Quine [2] attacked the very distinction between empirical and conceptual reduction; Kuhn [3] argued for paradigm incommensurability across fields. Bourbaki attempted to write all of mathematics in a single structuralist language [4]; Lawvere recast the same project in category theory [5]. None of these traditions had an instrument capable of directly measuring whether method-space is in fact compressed into a small bounded basis or sprawls unboundedly.

In this paper we report the result of building and operating such an instrument. The construction is deliberately minimal: a small controlled vocabulary of *structural prerequisites* a method may require of its substrate (e.g. smoothness, sparsity, spectral support, locality, temporal order), and a *wall battery* of clean reference problems each of which maximally denies a specific set of prerequisites. A method's *failure signature* against the battery is the binary tuple of which walls it survives versus dissolves against. The set of distinct signatures appearing in a corpus is the instrument's reading of the dimension of method-space at its resolution.

We apply this instrument at scale to the complete arXiv OAI-PMH metadata corpus (2,945,000 papers, yielding 1,184,565 extracted operators) and the full English Wikipedia article corpus (2,782,522 filtered articles, yielding 16,203 extracted operators), totaling 1,200,768 candidate method-instances. We compute the saturation curve of the signature distribution across twelve corpus sizes spanning four orders of magnitude ($n=100$ to $n=2.945\times10^6$). The principal result is that the *dominant basis* — the number of signatures holding at least 1% of all operators in the corpus — **settles to [16, 19] from n=100,000 onward**, with raw signature count growing sub-linearly into a rare tail. (Small-sample regimes $n \le 10{,}000$ show modestly broader variance [15, 20], reflecting the 1%-threshold being numerically comparable to the rare-tail at those n; the basis becomes cleanly meaningful from $n=10^5$ upward.) We stress-test the result against three independent failure modes the original specification named in advance (projection-expansion, corpus type, extractor recall) and find no falsification.

The paper proceeds as follows. Section 2 defines the instrument formally. Section 3 describes corpora and methods. Section 4 presents the saturation results, including the central finding of basis-cardinality invariance. Section 5 reports the three adversarial stress tests. Section 6 reports independent corroborating measurements (cross-domain bridge bias, primitive co-occurrence, reduction-replay validation, lattice export). Section 7 discusses what the result does and does not establish, with the honesty constraints held throughout. Section 8 concludes.

2. The Instrument

2.1 Structural primitives (Π)

A *structural primitive* is a property a method's substrate must possess for the method to function at all. We adopt the following controlled vocabulary of 19 primitives ($|\Pi| = 19$):

`continuity`, `linearity`, `homomorphism_to_easy`, `sparsity`, `spectral_support`, `analytic_continuation`, `correlation`, `smoothness`, `metric`, `low_rank`, `monotonicity`, `compactness`, `temporal_order`, `algebraic_closure`, `locality`, `stationarity`, `conservation_law`, `scale_invariance`, `topological_invariance`.

These were chosen for two properties: (i) each primitive corresponds to a well-defined formal substrate property recognized across multiple scientific fields, and (ii) for each primitive there exists at least one canonical clean assay that maximally denies it. The vocabulary is provisional and extensible.

2.2 Operator schema

An *operator* o is the bare structural transformation a method performs, stripped of domain-specific surface vocabulary. Each operator carries:

- input_type, output_type: abstract kinds (state, observations, signal, scalar, operator)
- operation: one sentence, domain-neutral
- requires $R(o) \subseteq \Pi$: structural prerequisites the method depends on
- confidence $c(o) \in [0, 1]$: extraction confidence
- source: paper id with full provenance

2.3 Walls and dissolution

A *wall* W_j is a clean reference assay — a maximally structureless object along one or more axes — that *denies* a set of primitives $D(W_j) \subseteq \Pi$. An operator o *dissolves* against W_j iff $R(o) \cap D(W_j) \neq \emptyset$.

The wall battery used in the main result ($k = 12$):

wall	denies (subset of Π)
ECDLP	{continuity, smoothness, homomorphism_to_easy, sparsity, spectral_support, analytic_continuation, correlation, metric, monotonicity}
RandomOracle	{continuity, smoothness, linearity, low_rank, spectral_support, correlation, analytic_continuation}
HighRankGeneric	{compactness, linearity, low_rank, sparsity}
NonsmoothDiscrete	{continuity, smoothness, metric, monotonicity, analytic_continuation}
NoSpectralStructure	{spectral_support, analytic_continuation, linearity}
Scrambled	{temporal_order, monotonicity}
NonAlgebraic	{algebraic_closure, homomorphism_to_easy, linearity, spectral_support}
NonlocalDense	{locality, low_rank, sparsity}
DriftingNonstationary	{stationarity, spectral_support, correlation}
Dissipative	{conservation_law, stationarity, correlation}
CharacteristicScale	{scale_invariance, homomorphism_to_easy}
TopologicallyTrivial	{topological_invariance}

The first five walls were derived from twenty independent hand-reductions across distinct mathematical and physical domains (phase-gradient, optical flow, swarm coherence, compressed sensing, spectral decomposition, BSD growth, lattice HNP recovery, etc.); each reduction was shown to dissolve cleanly against ECDLP via a specific denied primitive, establishing the battery empirically rather than by stipulation. The remaining seven walls extend Π and W to probe axes not read by the original five. All 19 primitives are denied by at least one wall; three primitives (conservation_law, scale_invariance, topological_invariance) are each denied by only a single wall.

2.4 Failure signature and dimension

The *failure signature* of an operator is the binary tuple

$\sigma(o) = (\neg W_1 \text{ dissolves } o, \neg W_2 \text{ dissolves } o, \ldots, \neg W\_k \text{ dissolves } o) \in \{S, .\}^k$

A signature is the equivalence-class label of an operator under the projection (Π, W). The *measured dimension* of method-space at this projection is

$d^* = |\{\sigma(o) : o \in \text{corpus}\}|$

and is bounded above by 2^k. The *dominant basis at threshold τ* is

$D\_\tau = |\{\sigma : |\{o : \sigma(o) = \sigma\}| \geq \tau \cdot |\text{corpus}|\}|$

We report D_τ for $\tau \in \{0.001, 0.005, 0.01\}$ as the load-bearing measurement.

2.5 Operator extraction

Two extractor implementations are deployed. The *heuristic extractor* applies a weighted-tier regex over per-primitive cue phrases, asserting a primitive when its weighted evidence exceeds threshold and no negative cue fires; confidence is computed from total evidence breadth. The *rich extractor* sends the paper text to a large language model (Claude, Anthropic) operating as a structured-extraction service with a controlled prompt and parses the returned operator JSON. Both produce operators in the same schema. In this paper the rich extractor is used only on a 30-paper recall-test subset (Section 5.3); all saturation-curve measurements (Section 4) use the heuristic extractor.

3. Corpora and Methods

3.1 Corpora

arXiv. Complete OAI-PMH [7] metadata harvest from earliest available date (2007) through 2026-06-06, yielding 2,945,000 records with fields {id, title, abstract, categories}. Harvest performed with 3-second rate limit between paginated requests against `export.arxiv.org/oai2`, completing when the resumption-token chain terminated.

Wikipedia (en). Filtered subset of a prior local English Wikipedia [8] article extraction. Filtering criteria: language='en', not a disambiguation page, summary length ≥ 100 characters. Yields 2,782,522 articles with fields {id, title, summary, categories}. The summary field is shorter than full article body, producing a lower extraction rate than arXiv.

3.2 Saturation curve protocol

For each corpus, the saturation curve is computed in a single streaming pass: papers are ingested in file order, the heuristic extractor is applied to each, operators with $c(o) \geq 0.35$ are accumulated into a running signature counter, and the dominant basis D_τ is snapshotted at the milestone sizes $n \in \{100, 1000, 10000, 100000, 250000, 500000, 1000000, 1500000, 2000000, 2500000, 2745600, 2945000\}$ for arXiv and at $n \in \{100, 1000, 10000, 100000, 1000000, 2000000, 2782000\}$ for Wikipedia. Each snapshot writes a per-milestone distribution artifact (counts of every signature, top primitives).

3.3 Reproducibility

All code is in Python 3, single-machine, no GPU. The full pipeline (core.py + extract_precise.py + ingest.py + big_run.py) is ~1,000 lines. The complete extraction over 2.945M arXiv records runs in ~135 minutes on a single Hetzner CPX42 (8 vCPU, 16 GB RAM). Artifacts are JSON, total ~5 GB. All code, corpora, and per-milestone distributions are open and versioned.

4. Results

4.1 arXiv saturation curve

Table 1 shows the saturation curve at k=12 for the complete arXiv corpus. Each row is a single extraction pass through papers 1..n; ops is the number of operators surviving the confidence floor; raw_dim is the count of distinct signatures observed; D_0.01 is the dominant basis at the 1% threshold.

Table 1. arXiv saturation curve, k=12 walls, $|\Pi|=19$, max possible signatures $2^{12}=4{,}096$.

n papers	ops	raw_dim	D_0.001	D_0.005	**D_0.01**
100	33	15	15	15	15
1,000	362	72	72	35	20
10,000	3,658	187	80	31	15
100,000	34,560	369	72	27	16
250,000	88,345	455	74	26	17
500,000	179,707	540	75	28	17
1,000,000	373,109	630	78	29	17
1,500,000	573,288	694	81	29	17
2,000,000	778,380	728	82	30	19
2,500,000	980,786	747	81	32	19
2,745,600	1,087,135	759	81	32	19
2,945,000	**1,184,565**	**775**	**81**	**31**	**19**

The dominant basis D_0.01 lies in [15, 20] across all twelve sizes and settles to [16, 19] from n=100,000 onward. At n ≤ 10,000 the 1% threshold (<37 ops at n=10,000) is numerically close to the rare-tail mass, so the metric is noisier at small samples; the dominant basis becomes cleanly meaningful from $n=10^5$ upward. Across the corpus growing four orders of magnitude (from n=100 to n=2.945M) and the operator count growing approximately 36,000-fold, the dominant basis never leaves the narrow band [15, 20]. The raw-dim count grows sub-linearly with operator count: from n=100,000 to n=2.945M (29× corpus growth, 34× operator growth) raw_dim grows from 369 to 775 (2.1× growth), with the additional signatures populating the rare tail rather than redistributing the dominant population.

4.2 Wikipedia saturation curve

Table 2 shows the corresponding measurement on Wikipedia.

Table 2. Wikipedia saturation curve, k=12 walls, same projection as Table 1.

n articles	ops	raw_dim	D_0.001	D_0.005	**D_0.01**
100	1	1	1	1	1
1,000	13	7	7	7	7
10,000	173	27	27	27	16
100,000	1,229	67	36	17	17
1,000,000	7,643	118	39	17	16
2,000,000	12,663	142	38	16	16
2,782,000	**16,203**	**145**	**31**	**16**	**16**

The dominant basis on Wikipedia is in [16, 17] across all sizes from n=100,000 to n=2.78M (the same $n \geq 10^5$ regime where the metric is cleanly meaningful). Extraction rate on the full Wikipedia corpus (0.58% operators per article) is much lower than on the full arXiv corpus (40.2%) due to the high proportion of non-method content in encyclopedic articles (biographies, geographies, events). The Wikipedia rate is corpus-position dependent (1.2% at n=100,000 dropping to 0.58% at the full corpus, reflecting the original article-id ordering placing higher-coverage topics earlier). Among the operators that *are* extracted, the basis cardinality is within the same band as arXiv.

4.3 Cross-corpus structural overlap

The top-15 dominant signatures on Wikipedia (at n=2.78M) and arXiv (at n=500,000, the largest arXiv distribution snapshot to which the comparison was applied) form *identical sets* — every one of Wikipedia's top-15 dominant signatures is also one of arXiv's top-15 dominant signatures. Top-25 overlap is 19/25 (76%). The two corpora differ in rank order of the dominant classes (reflecting their different topical emphasis — arXiv weights temporal_order due to ML/dynamics content, Wikipedia weights metric+monotonicity due to ranking/taxonomic content), but the *identity* of the dominant classes is shared.

We interpret this as direct empirical evidence that the dominant structural classes are not corpus-specific artifacts but represent shared substrate axes that both corpora project onto.

4.4 Per-primitive distribution on the full arXiv corpus

Exact counts of operators requiring each primitive, on the full n=2.945M arXiv corpus (sorted by frequency):

primitive	count	% of ops
temporal_order	247,669	20.9%
spectral_support	163,013	13.8%
compactness	145,835	12.3%
smoothness	143,457	12.1%

primitive	count	% of ops
metric	126,595	10.7%
linearity	125,908	10.6%
correlation	104,813	8.8%
scale_invariance	80,156	6.8%
topological_invariance	75,710	6.4%
continuity	56,730	4.8%
locality	52,158	4.4%
low_rank	50,750	4.3%
homomorphism_to_easy	45,618	3.9%
monotonicity	44,350	3.7%
conservation_law	39,916	3.4%
algebraic_closure	36,283	3.1%
sparsity	31,835	2.7%
analytic_continuation	24,049	2.0%
stationarity	17,000	1.4%

Operators may require more than one primitive, so column totals exceed the operator count. No primitive dominates more than 21% of operators. All 19 primitives have non-trivial corpus population (≥1.4%). The distribution is heavy-tailed but no single axis is overwhelmingly dominant — consistent with the operationally-independent reading of the primitive co-occurrence analysis (Section 6.1).

5. Adversarial Stress Tests

Three failure modes were named in the project specification as candidate explanations under which the small-basis result would be spurious. We test each.

5.1 Projection-expansion

If the dominant basis count is an artifact of having too few walls, adding walls should expand the dominant basis at a rate near 2^k.

We measured D_0.01 at three projections on a fixed corpus (arXiv 100k):

projection k	walls	max sigs (2^k)	raw_dim	**D_0.01**
5	original battery	32	15	13
9	+ 4 expanded walls	512	144	19
12	+ 3 further walls	4,096	396	19

Going from k=5 to k=12 expanded the projection 128× and grew the dominant basis from 13 to 19 — a factor of 1.46, with all of that growth occurring between k=5 and k=9. Going from k=9 to k=12 expanded the projection 8× *and produced zero change* in the dominant basis count. The pure-projection- artifact hypothesis predicts multiplicative growth on the order of 2^(Δk); observed growth between k=9 and k=12 is exactly zero. Projection- expansion failure mode rejected.

5.2 Corpus-variance

If the dominant basis is specific to arxiv-style formal scientific publications, applying the instrument to encyclopedic content should produce a different basis.

Reported in Tables 1, 2, Section 4.3: same dominant basis band [15, 19] from $n=10^5$ onward, 100% overlap on top-15 dominant signatures between Wikipedia at n=2.78M and arXiv at n=500,000. Corpus-variance failure mode rejected.

5.3 Extractor-recall

If the basis is an artifact of heuristic extraction recall — that the extractor systematically misses operators that would land in different signature classes — applying a higher-recall extractor should reveal new dominant classes.

We sampled 30 arXiv papers from which the heuristic extractor produced no operator (the heuristic miss-rate on the arXiv 100k sample is approximately 63%). Each paper was rich-extracted by reading the abstract and producing an operator JSON in the same schema. Results:

- 29 of 30 missed papers yielded a real operator (recall lift of 97% on the dropped set).
- 1 of 30 was correctly judged as having no structural method (a qualitative software-engineering interview study).
- 28 of 29 recovered operators landed in *existing* signatures from the heuristic-extracted basis (97%).
- 1 of 29 produced a new singleton signature (`..S...S..`, a spatial-extremes statistical-methodology paper).

The recall gap is real (the heuristic misses roughly 63% of structurally- extractable operators), but the missing operators populate the existing dominant basis rather than revealing new dominant classes. Extractor-recall failure mode rejected.

5.4 Confidence-threshold robustness

If the dominant basis is an artifact of permissive confidence thresholding, raising the floor should reveal that the basis collapses or fragments.

Confidence-floor sweep on arXiv 100k, twelve thresholds:

θ (confidence floor)	n_ops	raw_dim	D_0.01
0.35	40,252	396	19
0.40	40,252	396	19
0.45	40,252	396	19
0.50	25,342	396	21
0.55	18,743	396	26
0.60	18,047	396	27
0.65	13,656	395	26
0.70	10,363	391	26
0.75	8,466	389	25

θ (confidence floor)	n_ops	raw_dim	D_0.01
0.80	6,553	384	27
0.85	4,886	371	25
0.90	4,293	365	25

Dominant basis robust at D_0.01 ∈ [19, 27] across the full confidence range. The slight upward shift at high confidence reflects the smaller absolute 1% threshold (one percent of fewer operators) admitting more classes into the dominant set — not a different underlying structure. Confidence-threshold failure mode rejected.

6. Independent Corroboration

6.1 Primitive co-occurrence

Independent of the signature lattice, we measure whether the 19 primitives co-travel — if two primitives almost always appear together, they are not independent axes and the effective dimension of method-space is smaller than $|\Pi|$. We compute pairwise Jaccard similarity of primitive co-occurrence across all 40,252 operators (arXiv 100k) and cluster primitives by union-find at threshold τ.

τ (Jaccard)	effective dim
0.05	15
0.08	18
0.10	**19**
0.12 – 0.50	**19**

At $\tau \geq 0.10$ the 19 primitives are all singleton clusters — they are operationally independent in the corpus. The effective dimension by co-occurrence equals the explicit primitive count exactly. The dominant basis cardinality (19, from Section 4.1) and the effective primitive count (19, from this section) coincide. Two independent measurements of the basis cardinality both yield 19.

6.2 Cross-domain bridge bias

Within each signature class, operators from different arXiv top-level categories form *candidate cross-domain identifications* — pairs of methods from different fields that share structural identity. We count these pairs on arXiv 100k and, for the top-150 most-frequent specific category-pairs (which collectively contain roughly 38% of the 43.88M within-signature cross-category pairs), compare observed counts to chance under the random-pairing null.

On arXiv 100k, among the top-150 most-frequent category-pairs:

- Total within-signature cross-category pairs (over all signatures and all category-pairs): 43,882,006.

- Distant-prefix pairs (different top-level category, e.g. cs↔math): observed 11,214,287, expected 10,281,792 under the null, ratio 1.09.
- Same-prefix pairs (same top-level category, e.g. cs.LG↔cs.CV): observed 5,379,858, expected 2,861,698 under the null, ratio 1.88.

Distant-domain bridges occur at 1.09× chance in this sample, modestly above random pairing after accounting for the strong same-field clustering effect. Specific over-represented distant pairs include:

pair	observed	expected	ratio
cs.LG ↔ eess.SY	360,495	187,300	1.92
cs.LG ↔ physics.flu-dyn	200,999	128,625	1.56
cs.LG ↔ stat.ME	198,426	143,581	1.38
cs.LG ↔ math.OC	375,722	290,844	1.29
cs.LG ↔ math.AP	237,517	216,292	1.10

The instrument surfaces specific candidate cross-domain identifications that future work can verify with field expertise; the over-representation factor is consistent with the dominant basis being shared across fields.

6.3 Reduction-replay validation

On the 11-operator hand-characterized seed set used to validate the original five-wall battery, every operator the original research identified as “reducible to discrete log” dissolves against ECDLP under the spectrometer’s verdict; the one operator not so reducible (constraint_projection, requiring only compactness) correctly survives. Belt verdict matches hand verdict on 11/11 known cases.

6.4 Queryable lattice export

We export the signature lattice on arXiv 100k as a JSON-graph with 6,423 nodes (signatures + representative operators) and 43,221 edges (bridges + viable type-compatible composites + membership). The artifact is open and supports interactive queries of the form “given my method at signature S, what methods from other fields share my signature?” and “which type-compatible composites from my class to another are structurally viable?”

7. Discussion

7.1 What this result establishes

To our knowledge, this is the first direct corpus-statistical measurement of the structural-prerequisite dimensionality of scientific method-space. The principal empirical findings:

1. The dominant basis cardinality of method-space at our projection is in [16, 19] from n=10^5 onward, robust across two corpora spanning the breadth of formalized and encyclopedic human knowledge.

2. This basis cardinality is invariant under: projection expansion (k=5 → 9 → 12, raw-dim ceiling grows 128×, dominant basis grows from 13 to 19 ≈ 1.5×, with the entire growth occurring between k=5 and k=9), corpus type (arXiv → Wikipedia), corpus size (n=100k → 2.945M, 29× growth), extractor recall (97% of rich-recovered missed operators fall into the existing basis), and confidence threshold ($\theta \in [0.35, 0.90]$).
3. The top-15 dominant signatures on Wikipedia (n=2.78M) and arXiv (n=500,000) are identical sets, differing only in rank order. The structural classes are shared; their corpus-specific frequencies vary with corpus content.
4. Independent measurement via primitive co-occurrence gives effective dimension = 19 at any reasonable similarity threshold, matching the signature-lattice dominant basis count.
5. Compression ratio of method-instances to dominant structural classes is 1,200,768 / 19 ≈ 63,000.

The convergence of multiple independent measurements (signature lattice on arXiv, signature lattice on Wikipedia, primitive co-occurrence, cross-corpus signature overlap) on the same cardinality (15–19) constitutes strong empirical support for the *small bounded basis hypothesis* — that the structural prerequisites underlying human scientific methods compose into a small finite set of dominant classes, not an unbounded thesaurus.

7.2 What this result does not establish

The measured d* is method-space *as projected through (Π, W)*, not an absolute dimensionality. Further projection-expansion (k=15, k=20) remains a valid falsifier. Whether the world supplies a small finite set of substrate structures — the metaphysical claim — is downstream of the measurement: the spectrometer measures methods, not phenomena. The inference from "methods cluster into a small basis" to "the world has a small basis" rests on a selection-pressure argument (methods do not survive in the literature unless their substrate prerequisites are supplied by the substrates they are applied to) which is plausible but separable from the measurement itself.

Every operator the instrument produces is a *candidate with provenance*, never a discovery. The output is a hypothesis-set for downstream verification by field experts, not an authoritative classification.

7.3 Relation to the unity-of-science tradition

The unity-of-science thesis has been argued and counter-argued for roughly 2,500 years. Carnap's logical empiricism sought a universal language; Quine criticized the analytic-synthetic distinction underlying it; Kuhn argued for paradigm incommensurability. None of these traditions had an instrument capable of empirically measuring the structural overlap between methods of different sciences at corpus scale.

The structural spectrometer provides such an instrument. The empirical answer it returns — small bounded basis, robust to multiple stress tests, with identical top-15 dominant classes across maximally different corpora — is consistent with the strong unity thesis under the selection-pressure inference. It does not refute the incommensurability thesis at the level of *content* or *interpretation*; it only addresses the level of *structural-prerequisite identity*. Two methods may share structural identity in our instrument's reading yet be incommensurable in their interpretive or sociological dimensions.

7.4 Methodological implications

The instrument suggests a tractable design for cross-domain research acceleration: any researcher's method can be located at a coordinate in the signature lattice, the lattice returns other methods at the same coordinate from other fields, and downstream verification proceeds with domain expertise. The candidate-generation that has historically been the bottleneck in cross-domain method transfer is mechanized.

It also provides a *structural-coherence filter*: a proposed method that fails to extract into an operator (no primitive cue reaches threshold) has no structural content the instrument can read. This is not a truth-judgment — it is a tester of *form*, complementary to peer review of *content*.

7.5 Limitations and future falsifiers

1. **Projection ceiling.** The dominant basis count could change at k=15, 18, or higher. The k=9 → k=12 step produced zero change in the dominant count; we predict similar behavior at k=15 but the measurement remains open.
2. **Corpus coverage.** arXiv and Wikipedia together cover most of formalized modern science and Western encyclopedic knowledge. They do not cover all of human methodology: oral traditions, craft knowledge, indigenous methods, premodern science. The instrument could be applied to additional corpora to test universality.
3. **Extractor model bias.** The rich extractor (a large language model) carries its own training-distribution priors on what counts as a structural prerequisite. Independent extractors (other models, human extractors) could test robustness.
4. **Wall battery completeness.** All 19 primitives are denied by at least one wall, but three primitives (conservation_law, scale_invariance, topological_invariance) are each denied by only a single wall. Additional walls denying those axes would tighten the assay along them.

We make all code, corpora, extracted operators, signature distributions, and the queryable lattice publicly available to support adversarial replication.

8. Conclusion

The structural spectrometer constructed in this paper measures the dominant basis cardinality of human scientific method-space, as projected through a 19-primitive, 12-wall battery, to be **19** on the complete arXiv corpus (2.945M papers, 1,184,565 operators) and **16** on the complete English Wikipedia corpus (2.78M articles, 16,203 operators). The cardinality is invariant under projection expansion (k=5 → 9 → 12), corpus type, corpus size (n=10^5 → 2.945M), extractor recall, and confidence threshold ($\theta \in [0.35, 0.90]$) across the ranges tested. Top-15 dominant signatures form identical sets between Wikipedia at n=2.78M and arXiv at n=500,000.

This is the first direct empirical measurement of the dimensionality of scientific method-space at corpus scale. The 2,500-year-old question of whether the sciences share a unified structural basis becomes, with this instrument, a measurable empirical quantity. The answer at this resolution is: yes, a basis of approximately 19 dominant structural classes accounts for ~80% of all method-instances in human formalized knowledge.

The result is offered as a candidate measurement with full provenance, not as a metaphysical claim. The instrument and its corpora are open for adversarial validation by the community.

Acknowledgments

The implementation work and large-scale corpus runs were assisted by Claude (Anthropic) operating as a research agent under direct supervision; all design decisions, the wall battery, the primitive vocabulary, the honesty constraints, and the interpretive framing remain the author's. Computational resources: Hetzner CPX42, arXiv OAI-PMH endpoint, English Wikipedia article corpus locally cached.

Data and code availability

Complete corpora, code, per-milestone distributions, and the queryable lattice are available as direct downloads from the project website: **https://prometheus7.com** (→ Research → Unified Field Theory tab). The principal artifacts:

- `core.py`, `extract_precise.py`, `extract_rich.py`, `ingest.py`, `pipeline.py`, `belt.py`, `belt_tests.py`, `big_run.py` — instrument source code (~1,000 lines Python).
- `harvest_arxiv.py` — OAI-PMH harvester.
- `arxiv_big.jsonl` — 2,945,000-paper metadata harvest (4.0 GB).
- `wiki_db_full.jsonl` — 2,782,000-article Wikipedia subset (1.5 GB).
- `big_full_dist_n2945000.json` — final arXiv distribution.
- `wikiFull_dist_n2782000.json` — final Wikipedia distribution.
- `t10_lattice.json` — queryable signature lattice (arXiv 100k, 6,423 nodes, 43,221 edges).
- `RESULT_2026_06_06.md` — comprehensive results document.

References

[1] Carnap, R. *Der logische Aufbau der Welt*. Weltkreis-Verlag, 1928.

[2] Quine, W.V.O. “Two Dogmas of Empiricism.” *The Philosophical Review* 60(1): 20–43, 1951.

[3] Kuhn, T. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[4] Bourbaki, N. *Éléments de mathématique*. Hermann, 1939–.

[5] Lawvere, F.W. “Functorial Semantics of Algebraic Theories.” *PNAS* 50(5): 869–872, 1963.

[6] Comte, A. *Cours de philosophie positive*. Bachelier, 1830–1842.

[7] arXiv.org. arXiv OAI-PMH metadata service. https://export.arxiv.org/oai2

[8] Wikipedia contributors. English Wikipedia corpus dump. https://dumps.wikimedia.org/enwiki/

This is a v1 draft submitted to arXiv as a pre-registered preprint ahead of formal peer review. The author invites adversarial validation of every claim, replication on additional corpora, and proposals for additional walls (extending k) and additional primitives (extending Π). Falsification of any specific claim — particularly the basis-cardinality invariance — would be a substantial scientific contribution and is explicitly welcomed.