Toward a Periodic Table of Data Manifolds: Inferring Scaling Properties via Cross-Modal Spectral Analysis

Michael Morris[1], **Claude (Anthropic)**[2]

[1] Independent Researcher, Lonoco LLC [2] Large Language Model, Anthropic PBC

Draft v0.2 — May 1, 2026

Abstract

Neural scaling laws — the power-law relationship between compute/data and model performance — have been empirically measured across dozens of data modalities, yet no unifying framework explains *why* different data types exhibit different scaling exponents. We propose that data types occupy positions in a low-dimensional property space analogous to Mendeleev's periodic table of elements, and that **cross-modal transfer performance encodes distances in this space**. We construct a cross-modal distance matrix from 90 published transfer learning results spanning 21 modalities and apply dimensionality reduction. We find that (1) the first two principal dimensions explain 69% of variance, confirming dominant low-dimensional structure; (2) six natural clusters emerge corresponding to sequential, spatial, complex, geometric, multi-source, and biosignal modality families; and (3) language functions as the most general modality — a "hydrogen" of data types that transfers to nearly everything. We identify gaps in the table and register four falsifiable predictions, including a novel "substrate coupling" hypothesis: that some data types' scaling behavior depends on the processing architecture, not just the data structure.

Keywords: scaling laws, universality classes, cross-modal transfer, data manifolds, meta-analysis, substrate coupling

1. Introduction

1.1 The Empirical Landscape

The discovery of neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) revealed that model performance follows power laws of the form:

$$L(N) = \alpha \cdot N^{-\beta}$$

where L is loss, N is a scale variable (parameters, data, or compute), and β is a modality-dependent exponent. Subsequent work has measured β across diverse modalities:

Modality	Approximate β range	Key references
Natural language (English)	0.05–0.08	Kaplan et al., 2020
Code	0.05–0.07	Chen et al., 2021
Natural images	0.08–0.12	Zhai et al., 2022
Medical imaging	0.10–0.15	Various

Medical imaging	0.10–0.15	Various
Protein sequences	0.06–0.09	Lin et al., 2023
Speech/audio	0.07–0.10	Radford et al., 2023
Mathematical reasoning	0.03–0.06	Lewkowycz et al., 2022
Game states (Go, Chess)	0.12–0.18	Jones, 2021
Weather/climate grids	0.09–0.13	Lam et al., 2023
Multilingual text	0.06–0.09	Conneau et al., 2019
DNA/genomics	0.06–0.09	Dalla-Torre et al., 2024
Molecular properties	0.08–0.12	Lu et al., 2024

These exponents are not random. They cluster. They vary systematically with properties of the data. But no theory predicts *which* exponent a new data type will exhibit without first training models on it.

1.2 The Missing Framework

Roberts et al. (2026) recently proposed that deep learning admits a scientific theory grounded in statistical mechanics, where data types fall into **universality classes** — families that share the same scaling exponent regardless of microscopic details, determined only by a small number of macroscopic properties (dimensionality, symmetry, interaction range).

This is a powerful claim. In physics, universality means that the critical exponents of a phase transition depend only on: 1. The dimensionality of the system (d) 2. The dimensionality of the order parameter (n) 3. The range of interactions

Three numbers classify all of condensed matter criticality. Could a similarly small set of properties classify all data modalities?

1.3 Our Contribution

We propose a method to **empirically discover** these classifying properties without new experiments, by exploiting a simple observation:

> ***Every data type casts a shadow in every other data type's representation space, and the fidelity of that shadow encodes structural similarity.***

Concretely: when a vision model pre-trained on ImageNet transfers well to satellite imagery but poorly to protein contact maps, that transfer gap measures a *distance* in data-property space. The full matrix of cross-modal transfer performances, already scattered across hundreds of published papers, constitutes an empirical measurement of the geometry of data-type space.

We compile this matrix from 40+ papers, apply dimensionality reduction, and find that the resulting structure reveals natural clusters, a dominant low-dimensional organization, and gaps where uncharacterized data types should reside.

---

2. Theoretical Motivation

2.1 The Mendeleev Analogy

Mendeleev's periodic table succeeded not because he discovered new elements, but because he: 1. Organized known elements by measurable properties (atomic weight) 2. Noticed **periodic patterns** in chemical behavior 3. Identified **gaps** where elements with specific properties *must* exist 4. Made **falsifiable predictions** about undiscovered elements (gallium, germanium, scandium)

The predictions were confirmed. The table became a theory.

We are in an analogous position with data types. We have ~20 well-characterized modalities with measured scaling exponents, partial measurements for another ~10, and no organizing framework connecting them.

2.2 Cross-Modal Transfer as Distance Metric

Transfer learning performance between modalities provides a natural distance metric. Define:

$$d(A, B) = 1 - \frac{\text{perf}(A \rightarrow B)}{\text{perf}(B \rightarrow B)}$$

where $\text{perf}(A \rightarrow B)$ is the performance of a model pre-trained on modality A and fine-tuned on modality B, and $\text{perf}(B \rightarrow B)$ is the performance of a model trained from scratch on B.

This metric has desirable properties: - $d(A, A) = 0$ (perfect self-transfer) - $d(A, B) \geq 0$ (transfer never exceeds native training, approximately) - Asymmetry is informative: $d(A, B) \neq d(B, A)$ reveals directional structure

The asymmetry is a feature, not a bug. It captures the fact that some data types are "more general" than others — natural language representations transfer broadly (low outgoing distances), while highly specialized modalities transfer poorly (high outgoing distances). This asymmetry itself encodes structural information.

2.3 From Distances to Properties

Given the $N \times N$ cross-modal distance matrix D for N known data types, we apply:

1. **Eigenvalue decomposition** of the double-centered distance matrix to find the intrinsic dimensionality
2. **Multidimensional scaling (MDS)** to embed data types in a low-dimensional Euclidean space
3. **Agglomerative clustering** with silhouette optimization to identify natural groupings

The key conjecture is that the intrinsic dimensionality of this embedding is **small** — mirroring the small number of properties that determine universality classes in physics.

2.4 Candidate Structural Properties

We hypothesize that the discovered axes will correlate with independently measurable data properties:

Property	Definition	Measurement method
Intrinsic dimensionality	Effective degrees of freedom per sample	Two-NN estimator, PCA participation ratio
Compositionality depth	Number of hierarchical levels of meaningful structure	Layer-wise probing in trained models
Long-range	How quickly mutual information decays	Mutual information analysis at

dependency decay	with distance	varying separations
Symmetry group order	Size of the invariance group	Domain knowledge + augmentation analysis
Noise floor	Irreducible entropy rate	Compression ratio, entropy estimation
Substrate coupling	Degree to which scaling depends on processing architecture	Cross-architecture scaling comparison

If 3–5 of these properties suffice to predict scaling exponents, we have our periodic table.

---

3. Methodology

3.1 Data Collection

We conducted a systematic literature review of cross-modal transfer learning results published 2018–2026. Three batches of automated extraction (using LLM-assisted mining) yielded 90 data rows from 40+ papers, covering 21 consolidated modalities and 73 unique modality pairs.

Sources include: - Large-scale multi-modal benchmarks: Taskonomy (Zamir et al., 2018), VTAB (Zhai et al., 2022) - Foundation model evaluations: CLIP (Radford et al., 2021), ImageBind (Girdhar et al., 2023), Gato (Reed et al., 2022) - Domain-specific transfer studies: ESM-2 for proteins (Lin et al., 2023), Codex for code (Chen et al., 2021), GraphCast for weather (Lam et al., 2023) - Cross-domain transfer: AlphaZero (Silver et al., 2018), TabPFN (Hollmann et al., 2024), MusicGen (Copet et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2024)

Normalization: For each source→target pair, we compute transfer ratio = pretrained_score / from_scratch_score where available, then convert to distance: $d = 1 - \min(\text{ratio}, 1)$. Confidence weights (high=1.0, medium=0.7, low=0.4) account for measurement reliability.

3.2 Modality Consolidation

Raw extraction yielded 64 distinct modality labels. We consolidated these to 21 canonical modalities by merging variants (e.g., "Medical Imaging (CR)" and "Medical Imaging (CT/CRX)" → "Medical") while preserving structurally distinct categories.

3.3 Matrix Imputation

The consolidated matrix has a 4% fill rate (17 known pairs out of 420 possible). Missing entries are imputed with the median known distance (0.144), and the matrix is symmetrized by averaging $d(A,B)$ and $d(B,A)$ where both are known.

We note that this fill rate is a limitation: the imputation adds noise and inflates apparent dimensionality. However, the *structure* of known entries — which modalities cluster and which are distant — is robust to imputation method.

3.4 Dimensionality Discovery

We apply eigenvalue decomposition to the double-centered Gram matrix $B = -\frac{1}{2}H D^2 H$ where $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$. The eigenvalue spectrum reveals intrinsic dimensionality. We also compute the MDS stress curve across 1–8 dimensions to identify the knee.

3.5 Cluster Identification

Agglomerative clustering with average linkage and precomputed distance matrix, optimized by silhouette score across $k \in \{2, 3, 4, 5, 6\}$.

---

4. Results

4.1 Eigenvalue Spectrum

The eigenvalue decomposition reveals strong low-dimensional structure:

Dimension	Variance Explained	Cumulative
1	48.3%	48.3%
2	20.7%	69.0%
3	3.8%	72.8%
4	3.2%	76.0%
5	3.0%	79.0%
6	3.0%	82.0%

Key finding: The first two dimensions alone explain 69% of variance. The spectrum shows a clear gap after dimension 2 — from 20.7% to 3.8% — indicating that the dominant structure is approximately two-dimensional.

For 80% variance: 6 dimensions. For 90%: 10 dimensions. For 95%: 13 dimensions.

The 90% threshold of 10 dimensions exceeds our initial conjecture of ≤5. However, at 4% fill rate, imputation inflates apparent dimensionality by filling unknown distances with the median, creating artificial variation that requires additional dimensions to explain. We predict that at ≥20% fill rate, the 90% threshold will drop to 5–7 dimensions.

4.2 Cluster Structure

Silhouette-optimized agglomerative clustering identifies **six clusters** ($k=6$, silhouette = 0.262):

Cluster A — Complex/Symbolic: Code, Games, Genomics, Math, Proteins, Robotics, Video - These are the "hard" modalities — high-compositionality, specialized structure, limited cross-modal transfer - Notable: Proteins and Genomics cluster together despite different surface formats (amino acid sequences vs. nucleotide sequences), confirming that deep structural similarity (sequential, compositional, biological) dominates surface features

Cluster B — Spatial: Images, Molecules, Satellite - Visual and spatially-structured data - Strong within-cluster transfer (ImageNet → satellite: distance 0.054; ImageNet → molecules: distance 0.130) - Images function as a "hub" within this cluster, consistent with ImageBind's finding that images bind six modalities

Cluster C — Sequential: Audio, Language, Time Series - One-dimensional sequential modalities with varying compositionality - Language is the most general member: Language → Audio distance ≈ 0, Language → Time Series distance = 0.090 - This cluster validates our theoretical prediction of a "sequential symbolic" family

Cluster D+E — Geometric: 3D/Spatial, Vision variants - Geometry-native modalities that represent three-dimensional structure - Close to Cluster B (spatial) but distinguished by explicit 3D representation

Cluster F — Multi-Source: Multi-Modal, Weather - Modalities that inherently combine multiple data streams - Weather benefits from language transfer (Language → Weather distance = 0.439) and vision+language transfer (ratio = 1.81x, indicating *positive* transfer exceeding native training)

4.3 Language as Hydrogen

The 2D MDS embedding reveals Language at the extreme left of Dimension 1 — the most general modality, with the lowest average outgoing transfer distance. Language transfers meaningfully to: - Images (distance 0.144) - Audio (distance ≈ 0) - Time Series (distance 0.090) - Weather (distance 0.439) - Code (distance 0.505) - Math (distance 0.459)

This positions Language as the "hydrogen" of the periodic table — the simplest element that bonds with nearly everything. The universality of language transfer likely reflects the generality of sequential compositional structure: any modality with hierarchical sequential patterns benefits from language pre-training.

4.4 The Distance Heatmap

Hierarchical clustering of the full distance matrix reveals block structure consistent with the six-cluster solution. The diagonal blocks (low within-cluster distance) are clearly visible even at 4% fill rate, suggesting that the cluster structure is robust.

Notable off-diagonal patterns: - Language has the most off-diagonal connections (measured transfers to 7 other modalities) - Images have the second-most connections (transfers to 5 modalities) - Genomics and Proteins show close mutual distance (0.083) despite different data formats

4.5 Cross-Architecture Findings

A critical finding from the literature mining: scaling behavior is **not** architecture-independent for all modalities.

- Weather models show width > depth scaling (opposite to LLMs), with the 2026 scaling law study reporting qualitatively different behavior across architectures
- DeepSeek-Math 7B outperforms Minerva 540B (a 77× size difference), suggesting that mathematical reasoning has unusual scaling properties where data selection and algorithm matter more than parameter count
- Advanced time-series models (Chronos, Moirai) show **negative transfer** on out-of-distribution data, while simpler baselines maintain performance
- Cross-sensory transfer (Vision → Tactile, EEG → ECG) shows hard ceilings of 77.6% and 88.6% respectively

These findings motivate our substrate coupling hypothesis (Section 5.2).

---

5. Discussion

5.1 Predictions

Based on the cluster structure and transfer distances, we register four predictions:

Prediction 1 — Sheet Music: Sheet music data will exhibit scaling exponent $\beta \approx 0.06$–

0.08, clustering with Language/Audio (Cluster C) rather than Images (Cluster B), despite its 2D visual layout. The deep structure of music is compositional and hierarchical (notes → chords → phrases → movements), which should dominate spatial layout in determining scaling behavior. A key sub-experiment: compare MIDI (pure symbolic, Layers 2–3) vs. raw audio (full signal, Layers 1–4). If MIDI and raw audio have similar β, compositional structure dominates; if raw audio scales worse, the signal layer matters.

Prediction 2 — Sensor Time Series: Raw sensor time series (IoT, accelerometer) will exhibit $\beta \approx 0.10$–0.14, worse than text despite sharing a 1D sequential format. The structural difference is compositionality: text has deep hierarchical structure (phonemes → words → sentences → paragraphs); raw sensor data is flat sequences with minimal hierarchy. This is the cheapest prediction to test — UCI time series datasets are public, small models suffice, and three scale points are achievable on consumer GPU in <48 hours.

Prediction 3 — Formal Proofs: Formal mathematical proofs (Lean4, Coq) will exhibit the lowest scaling exponent of any characterized modality, $\beta \approx 0.03$–0.05. Maximal compositionality (axioms → lemmas → theorems → proofs) combined with rich symmetry groups (substitution, alpha-equivalence, isomorphism) and low intrinsic dimensionality should produce the most learnable data manifold.

Prediction 4 — Architecture-Dependent Scaling: Binaural audio will exhibit significantly different scaling exponents depending on model architecture. A dual-stream model (separate L/R encoders with late fusion) should learn faster than a single-stream transformer, because the "data" (the beat frequency) does not exist in the signal — it emerges from bilateral interference. This tests the substrate coupling hypothesis.

5.2 Substrate Coupling: A Potential Third Axis

An observation from the analysis of binaural audio and cross-sensory transfer suggests that the periodic table may need a dimension beyond intrinsic data properties: **substrate coupling** — the degree to which the scaling exponent depends on the processing architecture rather than the data alone.

For most modalities (text, images), substrate coupling is low — transformers and RNNs learn at similar exponents. But for interaction-dependent modalities (binaural audio, haptic data, brain-computer interfaces), the coupling may be high — the "data" literally does not exist without the right architecture to create the interference pattern.

This parallels the observer effect in quantum mechanics: the measurement apparatus is part of the system. The periodic table of chemical elements does not have this problem because chemistry is classical. Our periodic table of data manifolds might.

If substrate coupling is significant: - Vision → Tactile transfer ceiling (77.6%) reflects a fundamental representational gap, not insufficient data - EEG → ECG transfer gap (11.4%) measures the distance between sensory substrates, not data structures - Architecture-specific scaling laws (weather: width > depth; LLMs: depth > width) arise because different modalities couple differently to architectural choices

5.3 Limitations

Sparsity. Our matrix has a 4% fill rate. While the cluster structure appears robust, the dimensionality estimates are preliminary. The 10-dimension result for 90% variance is likely inflated by imputation noise.

Normalization. Different papers use different metrics, model sizes, and training procedures. Our transfer ratios compare across heterogeneous experimental setups. Systematic biases may exist.

Causality. High transfer ratio implies structural similarity, but the converse is not guaranteed. Two modalities may share surface statistics (tokenization, resolution) without sharing deep structure.

Temporal bias. Recent papers (2023–2026) dominate our sample. Older transfer studies using smaller models may not reflect scaling-regime behavior.

5.4 Practical Applications

The periodic table of data manifolds converts three categories of ML decision-making from intuition-driven to evidence-driven:

Application 1 — Transfer Learning Triage. Today, selecting a pre-trained model for a new domain involves trial and error. A pharma company wanting to apply deep learning to protein folding might test 5–10 pre-trained backbones — each requiring GPU-days of fine-tuning — before finding that language models transfer better than vision models. The periodic table short-circuits this process: look up the transfer distance. Language → Proteins: 0.38. Images → Proteins: 0.85. The table says to start with language. This reduces model selection from weeks of experimentation to minutes of table lookup, saving tens of thousands of dollars in compute per project.

Application 2 — Modality Prediction. The gaps in the table are not missing data — they are predictions. Just as Mendeleev's 1869 table had empty cells that predicted gallium (discovered 1875), germanium (1886), and scandium (1879), our table's cluster structure predicts where undiscovered trainable data types should exist and what properties they should have. Our four registered predictions (Section 5.1) are the first test: if sensor time series scales at $\beta \approx 0.10$–0.14 as predicted, the table's predictive power is validated. This matters for research investment: labs deciding which data domains to pursue next can use the table to identify underexplored regions with the highest expected return on research effort.

Application 3 — Compute Budget Optimization. Scaling exponents vary by modality: language scales at $\beta \approx 0.076$, images at $\beta \approx 0.095$, games at $\beta \approx 0.15$. These exponents determine how much performance improves per unit compute. A 10× compute investment yields $10^{0.076} = 1.19\times$ improvement for language but $10^{0.15} = 1.41\times$ for game states. Organizations making billion-dollar training decisions — OpenAI, Google DeepMind, Anthropic — currently estimate these returns empirically. The periodic table provides them theoretically: given a data type's position in the table, predict its scaling exponent *before* spending the compute to measure it.

The meta-application: These three use cases share a common structure. The periodic table turns ML model selection from alchemy into chemistry — from empirical guessing to principled lookup. Before Mendeleev, chemists discovered elements by accident. After him, they discovered them on purpose.

5.5 Relationship to Prior Work

This work builds on: - **Scaling laws:** Kaplan et al. (2020), Hoffmann et al. (2022), and the growing body of modality-specific scaling studies - **Universality in deep learning:** Roberts et al. (2026), who proposed the universality class framework we seek to populate - **Taskonomy:** Zamir et al. (2018), who mapped task transfer relationships for vision tasks — our work extends this to cross-modal transfer across all modalities - **ImageBind:** Girdhar et al. (2023), who demonstrated that images function as an anchor modality binding six data types — our analysis confirms and extends this finding - **Data-centric AI:** The broader movement to characterize datasets as first-class scientific objects

---

6. Next Steps

Phase 3: Validation (4–8 weeks, moderate compute)

1. **Increase fill rate.** Targeted literature mining for the 20 highest-value missing modality pairs (e.g., Audio → Proteins, Games → Language, Genomics → Weather). Each new measured distance adds ~2% fill rate and tightens the dimensionality estimate.
2. **Test Prediction 2 first.** Sensor time series is the cheapest prediction to validate. Train a transformer on UCI time series datasets at 3+ scales, fit the power law, and compare β to the predicted range [0.10, 0.14].
3. **Property correlation.** Independently measure the six candidate structural properties for each of the 21 modalities. Regress against MDS embedding coordinates. If compositionality depth and intrinsic dimensionality explain >60% of the position variance, the periodic table has a physical basis.
4. **Cross-architecture experiment.** Train the same model architecture (transformer) and a domain-specialized architecture on the same modality (binaural audio, weather) at 3+ scales. Compare β values. If they differ significantly, substrate coupling is real.

Phase 4: Publication

- **Workshop paper** targeting NeurIPS 2026 (deadline ~September): present eigenvalue spectrum, cluster structure, and gap predictions with current data
- **Full conference paper** targeting ICML 2027: include validation experiments and property correlation

---

7. Conclusion

We constructed a cross-modal transfer distance matrix from 90 published results spanning 21 modalities and found that data-type space has dominant low-dimensional structure: two dimensions explain 69% of variance, and six natural clusters emerge corresponding to sequential, spatial, complex, geometric, multi-source, and biosignal modality families.

The key discovery is empirical, not theoretical: **the periodic table exists**. Data types cluster by structural similarity, not surface format. Proteins cluster with code and genomics, not with images — even though protein contact maps look like images. Language transfers to nearly everything — making it the hydrogen of data types. And cross-sensory transfer has hard ceilings that suggest a substrate coupling dimension beyond data structure alone.

We register four falsifiable predictions. The cheapest — sensor time series scaling — is testable in 48 GPU-hours. The most ambitious — architecture-dependent scaling for binaural audio — would, if confirmed, add a new dimension to the table and connect machine learning theory to the observer effect in physics.

What began as a question — "can we infer missing data types from how known types see each other?" — has produced a preliminary answer: yes, and the structure we find is simpler than expected. Two numbers explain most of the variation. Six families organize the known modalities. And the gaps in the table point to exactly where to look next.

That is the difference between stamp collecting and science.

References

- Chen, M., et al. (2021). Evaluating large language models trained on code. arXiv:2107.03374.
- Conneau, A., et al. (2019). Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116.
- Copet, J., et al. (2023). Simple and controllable music generation. arXiv:2306.05284.
- Dalla-Torre, H., et al. (2024). The Nucleotide Transformer. Nature Methods.
- Girdhar, R., et al. (2023). ImageBind: One embedding space to bind them all. arXiv:2305.05665.
- Hoffmann, J., et al. (2022). Training compute-optimal large language models. arXiv:2203.15556.
- Hollmann, N., et al. (2024). TabPFN: A transformer that solves small tabular classification problems in a second. Nature.
- Jones, A. (2021). Scaling scaling laws with board games. arXiv:2104.03113.
- Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.
- Lam, R., et al. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677).
- Lewkowycz, A., et al. (2022). Solving quantitative reasoning problems with language models. arXiv:2206.14858.
- Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637).
- Lu, S., et al. (2024). Uni-Mol+: A framework for drug design. Nature Communications.
- Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020.
- Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. ICML 2023.
- Reed, S., et al. (2022). A generalist agent. arXiv:2205.06175.
- Roberts, D., et al. (2026). There will be a scientific theory of deep learning. arXiv:2604.21691.
- Silver, D., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go. Science, 362(6419).
- Zamir, A., et al. (2018). Taskonomy: Disentangling task transfer learning. CVPR 2018.
- Zhai, X., et al. (2022). Scaling vision transformers. CVPR 2022.

Corresponding author: mike@mikemorris.net