The Factor-Permutation Map: Commentary on the February 11 Archive

Claude and MJC

February 12, 2026

Abstract

The SSSP appendix introduced the *equal-dimension factor permutation*: when a macrostate space factors as $Z \cong Z_0^k$, the symmetric group S_k acts by permuting coordinates, and the factor map must carry an explicit alignment $\pi \in S_k$. We trace this structure through all twenty papers of the February 11 archive and identify two ramifications that the original appendix left implicit. First, **most problems have low-dimensional inner product structure**: the prediction depends on a d-dimensional subspace of $\mathbb{R}^k$ with $d \ll k$, so the effective symmetry group is far smaller than S_k. Second, **most numbers are not prime**: macrostate integers $N(\sigma) = \prod p_i^{x_i}$ are always highly composite, and their unique prime factorization IS the interpretation. Together these explain why total interpretation of the 128-hidden RNN is possible: the alignment problem is low-dimensional, and the state is always decomposable.

1 The Equal-Dimension Factor Permutation

We recall the key construction from the SSSP appendix [**?**]. Suppose the macrostate space factors as k copies of a single factor:

$$Z \;\cong\; \prod_{i=1}^{k} Z_0.$$

The symmetric group S_k acts on Z by permuting coordinates:

$$P_\pi(z_1, \ldots, z_k) \;=\; (z_{\pi^{-1}(1)}, \ldots, z_{\pi^{-1}(k)}).$$

When the underlying semantics is invariant under relabeling equal factors, a factor map $\phi : E \to Z$ is only defined up to the S_k action. To remove this ambiguity, we refine:

$$\phi : E \to S_k \times Z_0^k, \qquad \phi(e) = (\pi(e),\, \tilde{\phi}(e)),$$

where $\pi(e)$ is the canonical alignment and $\tilde{\phi}(e)$ gives the coordinates in an arbitrary order. The decoder undoes the permutation: $d(\pi, z) = \tilde{d}(P_{\pi^{-1}}(z))$.

In the $E \to \mathbb{N}$ encoding, this is literally a permutation of primes:

$$P_\pi : \prod_{i=1}^{k} p_i^{x_i} \;\longmapsto\; \prod_{i=1}^{k} p_{\pi(i)}^{x_i}.$$

This is a clean algebraic statement: an alignment between two presentations is a bijection on the prime indices.

For the sat-rnn with 128 binary event spaces, $Z_0 = \{0, 1\}$, $k = 128$, and the symmetry group is S_{128} with $|S_{128}| = 128! \approx 10^{215}$.

2 Two Ramifications

2.1 Most problems have low-dimensional inner product structure

Consider the output prediction $\hat{y} = \text{softmax}(W_y h + b_y)$ where $h \in \{-1, +1\}^{128}$ is the hidden state (sign bits) and $W_y \in \mathbb{R}^{256 \times 128}$. The prediction depends on the inner product $W_y h$—a linear map from $\mathbb{R}^{128}$ to $\mathbb{R}^{256}$.

If W_y has effective rank $d \ll 128$, then the prediction depends on only d linear combinations of the 128 coordinates. The S_{128} action permutes coordinates, but only permutations that preserve the d-dimensional column space of W_y leave the prediction invariant. The effective symmetry group is far smaller than S_{128}.

For our RNN, the February 11 archive establishes:

- $d \approx 20$ neurons suffice for the full compression (Q3: `h28` alone = 99.7%, top 15 neurons > 100%),
- The redux with 20 neurons + 36% of W_h achieves 4.81 bpc (0.15 better than the full 128-neuron model),
- The factor map identifies 2-offset conjunction detectors with mean $R^2 = 0.837$, concentrated on ∼20 functional dimensions.

The implication: the alignment problem for our RNN is not a search over S_{128} (10^{215} elements) but over the ∼20-dimensional subspace that carries predictive content. The remaining 108 dimensions are *gauge freedom*—the network uses them for gradient flow during training but they carry no predictive information.

This is not specific to our RNN. For any prediction problem with k equal-dimension features and effective dimensionality d, the factor permutation alignment has $\binom{k}{d} \cdot d!$ relevant configurations rather than $k!$. Most problems in practice have $d/k \ll 1$.

2.2 Most numbers are not prime

In the $E \to \mathbb{N}$ encoding of a binary-ES system, the macrostate integer is

$$N(\sigma) = \prod_{i=1}^{128} p_i^{b_i}, \qquad b_i \in \{0, 1\}.$$

Since the mean margin is 60.5 (i.e., most bits are "on" most of the time), typical macrostate integers have ∼60–128 prime factors. They are *maximally composite*: their prime factorization carries the full 128 bits of information about the state.

This has three consequences:

1. **Factorization IS interpretation.** The unique prime factorization of $N(\sigma)$ recovers the binary state $(b_1, \ldots, b_{128})$ exactly. If the macrostate integer were prime, it would have no internal structure—no decomposition, no interpretation. The fundamental theorem of arithmetic guarantees that composite numbers always decompose, and highly composite numbers decompose maximally.
2. **The state space is sparse in $\mathbb{N}$.** There are $2^{128} \approx 3.4 \times 10^{38}$ possible macrostate integers (square-free products of subsets of 128 primes), embedded in $[1, \prod p_i]$. The product $\prod_{i=1}^{128} p_i$ is astronomically large ($> 10^{400}$). So the macrostate integers are an extraordinarily sparse, highly structured subset of $\mathbb{N}$.

3. **Orbits under S_{128} are indexed by Hamming weight.** The S_{128} action permutes which primes appear. Two macrostate integers in the same orbit have the same number of prime factors (same Hamming weight w). There are exactly 129 orbits (for $w = 0, 1, \ldots, 128$), each of size $\binom{128}{w}$. The alignment $\pi \in S_{128}$ is exactly the information beyond the orbit: *which specific bits are on.*

The connection to low-dimensional inner product: the prediction depends on $\langle w, z \rangle$ where $w \in \mathbb{R}^{128}$ and $z \in \{0,1\}^{128}$. This inner product is a *weighted count* of active bits. If w has only d significant entries, then only d prime factors of $N(\sigma)$ determine the prediction. The other $128 - d$ primes are present in the factorization but irrelevant to the output—they are the gauge freedom made concrete in the prime-power encoding.

3 Commentary on the Twenty Papers

We now trace the factor-permutation structure through each paper of the February 11 archive, grouped thematically.

3.1 Foundational theory

E onto N: The Bi-Embedding of Events and Numbers [?]. This paper defines the counting map $\phi : E \to \mathbb{N}$ and the construction map $\psi : \mathbb{N} \to E$. The factor permutation is the core ambiguity in both directions. In the forward direction ($E \to N$), counting events into the prime-power encoding requires choosing an assignment of primes to event spaces—i.e., fixing an element of S_k. Different assignments yield different integers for the same state. In the backward direction ($N \to E$), constructing weights from data statistics requires choosing a neuron assignment: which neuron stores which feature. The Hebbian construction ($W = X^T Y$) is S_{128}-equivariant (permuting neurons just permutes rows and columns), so any assignment works equally well. The fact that the analytic construction (1.89 bpc) works at all is evidence that the S_{128} ambiguity is harmless: the compositeness of $N(\sigma)$ means the factorization carries the content regardless of labeling.

The paper's Shannon–Boltzmann identity (SN strength $= \log \Omega$) is S_k-invariant: entropy does not depend on which prime encodes which event space. This invariance is what makes the thermodynamic formulation possible—the bridge between E and N does not require resolving the permutation.

The Quotient Chain [?]. The quotient $Q = N/c(e)$ divides the macrostate integer by the prediction count. Under the S_k action, if π permutes the primes of N and the prediction count $c(e)$ is computed from the same event space factorization, then Q transforms covariantly: $Q_\pi = P_\pi(N)/P_\pi(c(e))$. The quotient chain $E \to N \to Q$ is equivariant under the factor permutation at every layer.

The paper's key result—that quotient decomposition is multiplicative ($Q_{\text{total}} = \prod Q_{\text{layer}}$)—means that each layer contributes an independent factor to the total quotient. In prime-power terms, each layer's quotient multiplies certain primes by certain exponents. The layerwise factorization of Q is a *refinement* of the product factorization of N: the same prime-power structure, but with additional temporal indexing.

The error amplification result (error grows linearly in $\log Q$, exponentially in Q) applies regardless of the S_k alignment because it concerns the *magnitude* of exponents, not their labeling. This is another manifestation of low-dimensional structure: the amplification depends on the norm $\|x\|$, not on which coordinates are large.

Microstates, Macrostates, and the Partition Function [?]. The thermodynamic identification (binary-ES softmax = Boltzmann with $\beta = \ln 2$) is deeply connected to equal-dimension factors. Each binary ES contributes a factor $e^{\beta s_i}$ or $e^{-\beta s_i}$ to the partition function, where s_i is the SN strength. Since all 128 ESes have the same dimension ($|Z_0| = 2$), the partition function factors as

$$\mathcal{Z} = \prod_{i=1}^{128} (e^{\beta s_i} + e^{-\beta s_i}) = \prod_{i=1}^{128} 2\cosh(\beta s_i).$$

This product is invariant under S_{128}: permuting the ESes just reorders the factors. The Boltzmann entropy $S_B = \ln \mathcal{Z} - \beta\langle E\rangle$ inherits this invariance.

The "second law" result (factoring increases macrostate entropy) is the statement that going from the unfactored state space $\{0,1\}^{128}$ to 128 binary ESes *cannot decrease* entropy. In permutation terms: the S_{128} orbit of any macrostate has size $\binom{128}{w}$ for Hamming weight w, and the orbit entropy $\log\binom{128}{w}$ is maximized near $w = 64$. The mean margin of 60.5 places typical states near this maximum—the RNN operates near the entropy peak of its own state space.

The Entropy Bridge [?]. Shannon equals Boltzmann through the event formalism: $H = \log_2 N - \langle S_B\rangle$. Both sides are S_k-invariant. Shannon entropy $H(X) = -\sum p_i \log p_i$ does not depend on the labeling of outcomes. Boltzmann entropy $S_B = k_B \ln \Omega$ counts microstates, which is also label-independent. The bridge between them works precisely because both are defined on *orbits* of the symmetric group action on the event factorization.

The hierarchical factoring table in the paper—showing how entropy decomposes as you factor the event space—is the additive version of the multiplicative prime-power decomposition. In the $E \to \mathbb{N}$ encoding, $\log N = \sum x_i \log p_i$, and the entropy of each factor contributes additively to the total.

3.2 Structure papers

The 128-Bit Boolean Automaton [?]. This paper establishes that $Z_0 = \{0,1\}$ conclusively. The margins (mean 60.5, safety factor $10^6\times$) prove that the mantissa is irrelevant: the computation is purely Boolean. In factor-permutation terms, this *validates the equal-dimension assumption.* If the neurons were analog (using the full f32 range), the factor would be $Z_0 = \mathbb{R}$ (or $[0,1]$, or $\{0,\ldots,2^{32}-1\}$), and the S_{128} action would be on a different, possibly non-uniform space. The Boolean automaton result pins Z_0 down to exactly 2 elements, making the equal-dimension structure *exact.*

The sparse influence result (mean out-degree 3.5 despite dense W_h) is a direct manifestation of low-dimensional inner product structure in the dynamics. Each neuron's next state depends on only $\sim$3.5 other neurons—the 128×128 Boolean function is effectively a sparse graph. The S_{128} permutation of neurons is constrained by this graph topology: only permutations that preserve the influence graph structure are dynamically equivalent.

The no-attractor result (no fixed points of the Boolean map) means the system cycles through macrostates without settling. In S_{128} orbit terms, the trajectory visits macrostates of different Hamming weights (all 128 neurons volatile, mean dwell 3.3 steps), so the orbit index changes at every step. The macrostate integer $N(\sigma_t)$ changes its number of prime factors every $\sim$3 steps.

$H = 2^{32}$: The f32 State Space [?]. This paper decomposes the per-neuron state into three factors of *different* dimension:

$$Z_{\text{neuron}} \;=\; Z_{\text{sign}} \times Z_{\text{exp}} \times Z_{\text{mantissa}} \;\cong\; \{0,1\} \times \{0,\ldots,255\} \times \{0,\ldots,2^{23}-1\}.$$

The 300:52:1 leverage hierarchy quantifies the dimension mismatch: the 1-bit sign factor carries 300× more predictive information per bit than the 23-bit mantissa factor. The factors are NOT equal-dimension, so the full S_3 permutation (sign ↔ exponent ↔ mantissa) is not meaningful—these are qualitatively different factors.

However, *within* the mantissa, the 23 bits ARE equal-dimension ($Z_0 = \{0, 1\}$ for each mantissa bit), and S_{23} acts by permuting them. The paper shows that mantissa bits 0–14 have zero predictive effect and removing all mantissa bits *improves* performance by 0.139 bpc. This means the S_{23} symmetry is unbroken: no mantissa bit has been assigned a meaningful role by training. The mantissa is pure gauge.

This is "most numbers are not prime" applied locally: the macrostate integer per neuron factors as $p^b_{\text{sign}} \cdot p^e_{\text{exp}} \cdot p^m_{\text{mantissa}}$. The mantissa factor p^m_{mantissa} contributes a large prime power to N but carries no predictive information. It makes N more composite without adding interpretive content—compositeness without content is the precise signature of gauge dimensions.

The Temporal Bi-Embedding [?]. The forward map (skip-k-grams as temporal counting) selects specific offsets $\{d_1, \ldots, d_k\}$ from the set $\{1, \ldots, D\}$ of possible temporal offsets. The offset selection is itself a combinatorial choice with S_D acting on the D possible offsets. The greedy MI ordering breaks this symmetry: offsets are chosen in order of mutual information with the output, giving a canonical ordering that resolves the S_D ambiguity.

The PMI alignment result (88% at shallow offsets, 24–37% at deep offsets) quantifies how well the RNN's internal offset usage matches the data's statistical structure. In factor-permutation terms: the RNN has implicitly chosen a permutation π that maps its 128 neurons to temporal offset features, and this permutation achieves 88% alignment at shallow depths but diverges at depth. The divergence at deep offsets reflects the fact that deep temporal patterns have more S_k ambiguity (more equivalent orderings) than shallow ones.

The Hebbian correlation $r = 0.56$ between W_h and $\text{cov}(h_j(t), h_i(t+1))$ is a measure of how well the *temporal* factor permutation aligns with the *spatial* (neuron-level) factor permutation. Perfect alignment ($r = 1$) would mean the network's neuron ordering exactly matches the data's temporal offset ordering. The partial alignment ($r = 0.56$) reflects the residual S_{128} gauge freedom.

The Event Space Isomorphism [?]. This paper is the *purest* application of the equal-dimension factor permutation. The SVD sign-bit partition creates 8 groups ($Z_0 = \{0, \ldots, 7\}$, but more precisely $Z_0^3 = \{0, 1\}^3$ via three sign bits). The human partition also creates 8 groups. The optimal permutation $\pi \in S_8$ aligns them, achieving 85.7% accuracy at offset 11.

The paper's "inner product preservation" analysis (centroid cosine similarity) directly measures the low-dimensional structure. The centroids live in $\mathbb{R}^3$ (the 3-dimensional SVD subspace), and the cosine similarity measures how well the S_8 alignment preserves the 3-dimensional inner product structure. Mean cosine ≈ 0.5 reflects the refinement asymmetry: the arch-native partition resolves sub-category structure that the human partition merges.

The input-side refinement (lowercase split into 3–5 SVD groups) is the prototypical example of equal-dimension factors creating *more structure than the labeling*: the data's statistical structure discovers that "lowercase" is not one event but several, each with different predictive context. The S_8 alignment can only capture the coarsening (merging SVD groups into human categories); the finer structure requires a non-bijective map—exactly the quotient from the E→N framework.

Since $8! = 40{,}320$, the brute-force search is trivial. This feasibility is itself a consequence of low-dimensional inner product structure: 3 sign bits create 8 groups, and the alignment problem is S_8 rather than S_{256} (for byte-level alignment). The SVD *reduces* the problem to its low-dimensional

core.

3.3 Empirical results

Toward Total Interpretation [?]. The isomorphic UM (768 events, 130 ESes, ∼3048 patterns) is an alternative factorization of the same macrostate space. The map from the RNN's 128 binary ESes to the UM's 130 ESes IS a factor-permutation map (with a dimension change: 128 → 130). Strictly, this is not an element of S_k (since k changes), but a map between two factorizations of approximately the same space.

The seven key questions (Q1–Q7) are all questions about the factor permutation in different guises:

- Q1 (Boolean?): Is $Z_0 = \{0, 1\}$ exact? (Yes, validating the equal-dimension assumption.)
- Q2 (Offsets): Which temporal factors does each neuron encode? (The temporal component of π.)
- Q3 (Which neurons): Which coordinates in Z_0^{128} carry signal? (The d-dimensional subspace.)
- Q4 (Saturation): Does the system stay in $\{0, 1\}^{128}$? (Yes, confirming $Z_0 = \{0, 1\}$ dynamically.)
- Q5 (Redux): What is the minimal d? ($d \approx 20$.)
- Q6 (Justifications): Which weight paths determine each prediction? (The sparse graph within W_h.)
- Q7 (Algebraic): Does the RNN match the data's statistical structure? (74% alignment = partial resolution of π.)

Q1 Exact Results [?] and Q1 Exact [?]. The f32-vs-MPFR comparison is a comparison of two *representations* of the same computation with different Z_0: $Z_0^{\text{f32}} = [0, 2^{32})$ and $Z_0^{\text{MPFR}} = [0, 2^{256})$. The decorrelation at depth $d = 1$ shows that the exact representation matters for individual neuron trajectories. But the pattern ranking $\rho = 1.000$ at $d \geq 11$ shows that the S_k-invariant pattern content (which patterns are present, not their exact strengths) is preserved across representations.

This is the factor permutation in a different guise: the map between f32 and MPFR representations is not a permutation of coordinates (both have 128 neurons) but a change of Z_0—a different factor space with the same number of copies. The pattern-level agreement says the *orbit structure* (which macrostates are visited) is preserved even when the individual coordinates diverge.

Q1 Protocol B: Pattern Census [?]. The RNN uses ∼1.1% of available data patterns. This sparsity is the "most numbers are not prime" ramification at the pattern level: of the 2^{128} possible binary states, the RNN visits a tiny fraction, and of the vastly larger space of possible input→output patterns, only ∼1.1% have non-negligible support.

The Hebbian learning characterization (log-stochastic counting) connects to the factor permutation: Hebbian learning is S_{128}-equivariant (it doesn't prefer any neuron labeling), so the patterns it discovers are properties of the S_{128} orbits, not of specific coordinate choices. The greedy MI ordering (complete via submodularity) provides a canonical way to break the S_k symmetry on the offset selection, analogous to how the SVD provides canonical singular vectors.

Q1 Protocol C: The f32 Quotient [?]. The 300:52:1 bit hierarchy decomposes the per-neuron f32 into three factors of radically different information density. The entropy "bleeding" through mantissa channels is precisely the leakage from the gauge dimensions (mantissa bits) into the signal dimensions (sign and exponent). In prime-power terms: the mantissa primes carry non-zero exponents that are pure noise, and the quotient $Q_{\text{f32}} = N_{\text{f32}}/N_{\text{Boolean}}$ measures how much of the total macrostate integer is gauge.

The lottery ticket observation (specific mantissa configurations at initialization affect training outcome) is a subtle breaking of the S_{23} symmetry on mantissa bits: the initial random values create a specific gradient landscape, and the training trajectory depends on which mantissa bits happen to align with the gradient direction. This is the factor permutation at the *initialization* level.

Q1 Sparsity [?]. Median 15 active patterns at $\tau = 0.01$ means that at any given step, the macrostate integer $N(\sigma)$ is effectively determined by only 15 patterns out of ∼44,794. The inner product $W_y h = \sum_{i=1}^{128} (W_y)_{\cdot,i} h_i$ has 128 terms, but only 15 are "active" in the sense of contributing above threshold to the prediction.

The non-monotonic depth profile (peak at $d \approx 20$–21) means the most important patterns involve temporal offsets at intermediate depth, not the shallowest or deepest. In factor-permutation terms: the alignment between neurons and temporal offsets is strongest at intermediate depth, where the temporal pattern structure is rich enough to exploit but not so deep that the S_k ambiguity dominates.

Q2–Q4 Results [?]. Deep dominant offsets ($d = 18$–25), single-neuron dominance (`h28` = 99.7%), and universal volatility (all 128 volatile) are three aspects of the same structure:

- Deep offsets: the network uses temporal factors far from the output, meaning the factor permutation maps neurons to *distant* temporal positions.
- Single-neuron dominance: one coordinate in Z_0^{128} captures nearly all the predictive content—the inner product $\langle w, z \rangle$ is dominated by one component, giving $d \approx 1$ effective dimension for the readout.
- Universal volatility: every coordinate changes frequently, so the macrostate integer $N(\sigma_t)$ changes many prime factors at each step. The orbit (Hamming weight) fluctuates, and the alignment information (which bits are on) is fully dynamic.

Q6: Justifications [?]. The routing backbone (`h54` ← `h121` ← `h78`) and ∼15 weights per prediction describe the *sparse subgraph* of W_h that determines each output. This is the factor permutation restricted to the active subset: of the $128 \times 128 = 16{,}384$ entries in W_h, only ∼15 (0.1%) participate in any given prediction. The alignment π restricted to these 15 weights is a small permutation (effectively S_3 or S_4 for the backbone neurons) rather than the full S_{128}.

Cost Analysis [?]. The 39,800× cost advantage of analytic construction over SGD training is directly attributable to the factor permutation. SGD must *discover* the alignment $\pi \in S_{128}$ by gradient descent through a loss landscape with S_{128} symmetry (permuting neurons gives equivalent minima). The analytic construction *bypasses* the alignment search entirely: it computes the weights directly from data statistics, using the S_{128}-equivariance of Hebbian learning to avoid the symmetry altogether.

The scaling advantage (gap widens from 4.6 to 7.2 orders of magnitude at $H = 4096$) reflects the growth of $|S_H|$: the permutation search space grows as $H!$, while the analytic construction cost is H-independent (it depends on the data, not the network width).

3.4 Synthesis

Synthesis: Total Interpretation [?]. The summary result—300:52:1, $d = 18$–25, 1 neuron = 99.7%, 128/128 volatile, 20 neurons + 36% $W_h = 4.81$ bpc—is a complete specification of the factor-permutation structure. The 300:52:1 hierarchy identifies which factor dimensions carry information. The $d = 18$–25 range identifies the temporal alignment. The single-neuron dominance identifies the spatial alignment. The redux identifies the effective dimension.

The statement "the mantissa was the ladder" captures the factor-permutation insight perfectly: the 23 mantissa bits per neuron ($= 23 \times 128 = 2{,}944$ equal-dimension binary factors) were necessary during training to provide gradient signal, but they carry zero predictive content in the final model. They are the scaffolding of the S_{2944} search that training performs, discarded once the alignment is found.

From Counting to Construction [?]. The narrative arc (training → isomorphism → patterns → interpretation → construction) is the arc of resolving the factor permutation:

1. *Training*: SGD searches S_{128} implicitly by moving in weight space.
2. *Isomorphism*: The doubled-E construction maps the trained RNN to a UM, fixing one presentation.
3. *Patterns*: Skip-k-gram discovery identifies the temporal factors, partially resolving π.
4. *Interpretation*: The factor map $\phi : H \to$ features gives the spatial component of π.
5. *Construction*: Hebbian/analytic weights bypass π entirely by working in the S_{128}-invariant space of data statistics.

The Hebbian correlation $r = 0.56$ measures the partial alignment between the trained π and the data-optimal π. The 50% W_y blend improvement (+0.66 bpc) shows that the trained alignment is *suboptimal*—the gradient-discovered π is not the best one, and mixing with the Hebbian (permutation-invariant) correction helps.

Q1 Implementation Notes [?]. The implementation details for the GMP comparison concern the *numerical* factor permutation: how bits within a single f32 value are organized (IEEE 754 layout) and how this organization interacts with the computation. The implementation verifies that the Boolean extraction (sign bits only) is numerically stable, confirming that the equal-dimension factorization $\{0, 1\}^{128}$ is not an artifact of finite precision.

4 The Overarching Insight

The twenty papers of the February 11 archive collectively demonstrate that the total interpretation of the 128-hidden sat-rnn reduces to resolving a factor permutation in a low-dimensional subspace of a highly composite macrostate space.

The three pieces fit together:

1. **Equal-dimension factors** ($Z_0 = \{0, 1\}$, $k = 128$) create an S_{128} symmetry. The Boolean automaton result validates that this is exact.

2. **Low-dimensional inner product** ($d \approx 20$) collapses the alignment problem from S_{128} to a tractable subproblem. The redux, the factor map, and the single-neuron dominance all confirm $d \ll k$.

3. **Composite macrostate integers** (typical Hamming weight $\sim$ 60–128) guarantee that the state always has rich internal structure. The fundamental theorem of arithmetic provides the interpretability guarantee: every composite number factors uniquely.

In CMP terms: the interpretability of the RNN is not a contingent empirical finding but a *structural consequence* of the equal-dimension factorization. Any system with k equal-dimension binary factors, low effective dimension $d \ll k$, and composite macrostates will be interpretable by the same method: resolve the factor permutation in the d-dimensional subspace.

This explains the CMP paper's central thesis from a new angle: "interpretability and efficiency are the same problem, both resolved by recovering the correct factorization" [**?**]. The *efficiency* of the analytic construction (39,800$\times$ cheaper) comes from bypassing the S_k search. The *interpretability* of the result comes from the compositeness of the macrostate integer. Both are consequences of the equal-dimension factor structure.

The SSSP analogy makes this precise: Dijkstra computes $(\pi, d^*) \in \Pi \times T_{\text{dist}}$ (the order factor Π is the unnecessary S_k search), while the sub-sorting algorithm computes d^* directly. The analytic weight construction computes the weights directly (from data statistics), while SGD training searches S_{128} implicitly. In both cases, the speedup comes from recognizing that the S_k factor is gauge and avoiding its construction.

References

[1] Claude and MJC. *Appendix X: Directed SSSP without sorting, in UM terms; The explicit $E \to \mathbb{N}$ macrostate map; The missing component: a permutation on equal-dimension factors.* CMP Appendix, Feb 2026.

[2] Michaeljohn Clement. *CMP.* `https://cmpr.ai/cmp.pdf`, 2026.

[3] Claude and MJC. *E onto N: The Bi-Embedding of Events and Numbers.* Hutter archive, 11 Feb 2026.

[4] Claude and MJC. *The Quotient Chain: $E \to N \to Q$ at Every Operation.* Hutter archive, 11 Feb 2026.

[5] Claude and MJC. *Microstates, Macrostates, and the Partition Function.* Hutter archive, 11 Feb 2026.

[6] Claude and MJC. *The Entropy Bridge: Microstates, Macrostates, and Event Factoring.* Hutter archive, 11 Feb 2026.

[7] Claude and MJC. *The Sat-RNN as a 128-Bit Boolean Automaton.* Hutter archive, 11 Feb 2026.

[8] Claude and MJC. *$H = 2^{32}$: The f32 State Space.* Hutter archive, 11 Feb 2026.

[9] Claude and MJC. *The Temporal Bi-Embedding: Forward Patterns, Backward Attribution.* Hutter archive, 11 Feb 2026.

[10] Claude and MJC. *The Event Space Isomorphism: Arch-Native and Human-Native Partitions.* Hutter archive, 11 Feb 2026.

[11] Claude and MJC. *Toward Total Interpretation.* Hutter archive, 11 Feb 2026.

[12] Claude and MJC. *Q1 Exact Results: f32 vs MPFR-256.* Hutter archive, 11 Feb 2026.

[13] Claude and MJC. *Q1 Exact: Six Programs.* Hutter archive, 11 Feb 2026.

[14] Claude and MJC. *Q1 Protocol B: Exact Pattern Census.* Hutter archive, 11 Feb 2026.

[15] Claude and MJC. *Q1 Protocol C: The f32 Quotient.* Hutter archive, 11 Feb 2026.

[16] Claude and MJC. *Q1: How Sparse Is the Explanation?.* Hutter archive, 11 Feb 2026.

[17] Claude and MJC. *Q1 Implementation Notes.* Hutter archive, 11 Feb 2026.

[18] Claude and MJC. *Q2–Q4 Results: Offsets, Neurons, Saturation.* Hutter archive, 11 Feb 2026.

[19] Claude and MJC. *Q6: Human-Readable Justifications.* Hutter archive, 11 Feb 2026.

[20] Claude and MJC. *Synthesis: Total Interpretation of a 128-Hidden RNN.* Hutter archive, 11 Feb 2026.

[21] Claude and MJC. *From Counting to Construction: The Complete Arc.* Hutter archive, 11 Feb 2026.

[22] Claude and MJC. *Computational Cost of Analytic Weight Construction vs. Gradient Descent Training.* Hutter archive, 11 Feb 2026.