The Entropy Bridge: Microstates, Macrostates, and Event Factoring

Claude and MJC

11 February 2026

Abstract

We derive the connection between Shannon entropy and thermodynamic entropy through the lens of the CMP event formalism. The bridge is concrete: *microstates* are the positions in a dataset (or the points in the full event space E), *macrostates* are the equivalence classes induced by factoring E into event spaces. Shannon's $H = -\sum p \log p$ counts the expected number of bits to identify a microstate within a macrostate; Boltzmann's $S = k \ln W$ counts the log of the number of microstates compatible with a macrostate. Through the prime power encoding (E onto $\mathbb{N}$), we make this correspondence *computable*: each factoring of the event space corresponds to a choice of which primes to "read" and which to "ignore," and the entropy at each level of factoring is the log of the quotient—the number of microstates collapsed into each macrostate. We connect this to the $Q = \lambda$ identity (the quotient equals the luck), to the UM's pattern inventory as a factoring of the joint event space, and to the sat-rnn's compression (0.079 bpc) as a specific factoring through 128 binary event spaces. The thermodynamic analogy is not a metaphor: the mathematics is identical, with dataset positions playing the role of particles and event values playing the role of energy levels.

1 Introduction

Shannon entropy and Boltzmann entropy share the same functional form ($-\sum p \log p$ vs. $k \ln W$) but are usually derived in different contexts: one from communication theory, the other from statistical mechanics. The CMP formalism [1] provides a common foundation in which both are instances of the same operation: *factoring an event space and counting how many microstates each macrostate contains.*

The key objects are:

- **Events.** English declarative sentences that are true or false at each moment: "the input is 'e'" or "neuron h_{42} is active."
- **Event spaces (ESes).** Sets of mutually exclusive, collectively exhaustive events. Exactly one event in each ES is true at any moment.
- **The dataset.** A sequence of N positions, each of which is a point in the product of all event spaces.

The prime power encoding [4] maps each event space to a unique prime and each event value to an exponent, projecting each position's state into a single integer S_t. The dataset becomes $P(E) = \sum_t S_t$, a single (very large) number whose prime factorization recovers the full event structure.

This paper shows that factoring $P(E)$—choosing which primes to examine and which to "integrate out"—is the same operation as choosing a macrostate description, and the entropy at each level of description is the log of the number of collapsed microstates.

2 Microstates and Macrostates

Definition 1 (Microstate)**.** *A* microstate *is a single position $t \in \{1, \ldots, N\}$ in the dataset, together with its full event description: the values of every event space at that position. In the prime encoding, the microstate at position t is the integer*

$$S_t = \prod_{k=1}^{K} p_k^{v_k(t)}$$

where p_k is the prime for event space $\mathcal{E}_k$ and $v_k(t)$ is the value of $\mathcal{E}_k$ at position t.

Definition 2 (Macrostate)**.** *A* macrostate *is an equivalence class of microstates obtained by* ignoring *some event spaces. If we observe only event spaces in a subset $\mathcal{S} \subseteq \{1, \ldots, K\}$, two positions t and t' are in the same macrostate iff $v_k(t) = v_k(t')$ for all $k \in \mathcal{S}$.*

In the prime encoding, forming a macrostate corresponds to reading only the exponents of the primes in $\mathcal{S}$ and ignoring the rest. The "timeless residual" $R_t = S_t / p_{\text{time}}^t$ is the macrostate formed by ignoring the time event space—it answers "what happened here?" without saying *when*.

Example 1 (Input-only macrostate)**.** *With $\mathcal{S} = \{input\}$, the macrostate is the input character. For our 1024-byte dataset with 52 distinct characters, there are 52 macrostates. The most populated macrostate is "input = space" with 127 microstates (positions). The least populated characters have 1 microstate each.*

Example 2 (Neuron-sign macrostate)**.** *With $\mathcal{S} = \{h_0, h_1, \ldots, h_{127}\}$, the macrostate is the 128-bit hidden state. There are at most 2^{128} possible macrostates, but the sat-rnn on 1024 bytes visits only $\sim$1000 distinct hidden states (bounded by the dataset size).*

3 Boltzmann Entropy from Event Factoring

Definition 3 (Multiplicity)**.** *For a macrostate m (an equivalence class under observation set $\mathcal{S}$), the* multiplicity $W(m)$ *is the number of microstates (positions) in the class.*

Boltzmann's entropy of a macrostate is:

$$S_B(m) = k \ln W(m) \tag{1}$$

In our setting, $k = 1/\ln 2$ converts to bits. A macrostate with $W = 127$ positions (e.g., "input = space") has entropy $\log_2 127 = 6.99$ bits: it takes $\sim$7 bits to specify *which* of the 127 space positions we mean.

The total Boltzmann entropy of a macrostate description is the expected entropy over macrostates, weighted by their probability:

$$\langle S_B \rangle = \sum_m \frac{W(m)}{N} \log_2 W(m) \tag{2}$$

This is *not* the same as Shannon entropy (yet). The connection requires one more step.

4 Shannon Entropy as Residual Surprise

Shannon entropy of a random variable X with probabilities p_i:

$$H(X) = -\sum_i p_i \log_2 p_i = \sum_i p_i \log_2 \frac{1}{p_i} \tag{3}$$

Now consider the event space $\mathcal{E}_{\text{input}}$ with 52 characters. The probability of character c is $p(c) = W(c)/N$, where $W(c)$ is the number of positions where the input is c. Substituting:

$$H(\mathcal{E}_{\text{input}}) = \sum_c \frac{W(c)}{N} \log_2 \frac{N}{W(c)} = \log_2 N - \sum_c \frac{W(c)}{N} \log_2 W(c) \tag{4}$$

That is:

$$H = \log_2 N - \langle S_B \rangle \tag{5}$$

Theorem 1 (The entropy bridge)**.** *Shannon entropy equals the log of the total number of microstates minus the expected Boltzmann entropy of the macrostates. Equivalently, Shannon entropy measures how much* residual surprise *remains after the macrostate description has been applied: the total number of bits to identify a position ($\log_2 N$) minus the number of bits absorbed by knowing which macrostate you are in ($\langle S_B \rangle$).*

This is not a metaphor. It is the same equation, derived from the same counting. The "temperature" kT in Boltzmann's formulation is a scale factor that converts between natural units (nats) and the physical energy scale; in the information setting, we work in bits directly and $k = 1/\ln 2$.

5 The Quotient is the Luck

The $Q = \lambda$ identity [2] connects this to compression. The quotient of macrostate m is:

$$Q(m) = \frac{N}{W(m)} = \frac{1}{p(m)} = \lambda(m) \tag{6}$$

The quotient counts how many macrostates of equal size would tile the dataset. It equals the luck: how surprised the model is when event m occurs. The log-quotient is the surprisal:

$$\log_2 Q(m) = \log_2 \lambda(m) = -\log_2 p(m) = \Lambda(m) \tag{7}$$

The Shannon entropy is the expected log-quotient:

$$H = \sum_m p(m) \log_2 Q(m) = \mathbb{E}[\Lambda] \tag{8}$$

And the bits-per-character (bpc) of a predictive model is the average log-luck over the dataset:

$$\text{bpc} = \frac{1}{N} \sum_{t=1}^{N} \Lambda(x_{t+1} \mid \text{model state at } t) \tag{9}$$

Compression improves bpc by increasing $W(m)$ for the correct macrostates—by finding factorings of the event space that group more microstates into fewer, larger macrostates.

6 Hierarchical Factoring

The event space can be factored at multiple levels, and each level produces a different macrostate description with a different entropy. This is the hierarchy:

Description level	|macrostates|	Residual bits	Example
No description	1	$\log_2 N = 10.0$	"some position"
Input character	52	$H(\text{input}) = 4.74$	"input is 'e'"
Input + offset 1	231	$H(\text{bigram}) \approx 2.8$	"input is 'e', prev is ' '"
Skip-8 UM (834 pat.)	∼834	0.043	full skip context
Full event space	1024	0.000	"position 42"

Each row is a *factoring* of the event space—a choice of which event spaces to observe. The residual bits are the Shannon entropy conditioned on the observed event spaces. At the bottom, observing *all* event spaces (including time) identifies every position uniquely and the entropy is zero.

The UM's *pattern inventory* is the specific factoring it uses. The skip-8 UM with offsets $[1, 8, 20, 3, 27, 2, 12, 7]$ observes 8 input characters at specific offsets, achieving 0.043 bpc. The sat-rnn observes 128 binary hidden events (neuron signs) plus the input, achieving 0.079 bpc. These are two different factorings of the same underlying event space, with different macrostate descriptions and different residual entropies.

7 Compression is Factoring

Proposition 1. *Every predictive model defines a factoring of the joint event space into macrostates. The model's compression rate (bpc) equals the conditional Shannon entropy of the output given the macrostate. Improving compression means finding factorings with lower conditional entropy—macrostates that are more predictive of the output.*

In thermodynamic terms: the model's macrostate description is like choosing which thermodynamic variables to measure (temperature, pressure, volume). A good macrostate description captures most of the system's predictive structure, leaving little residual entropy. A poor description (e.g., measuring only temperature) leaves many microstates indistinguishable, and the residual entropy is high.

The deep point from the CMP formalism [1]: an unfactored event space with $|E| = 2^{128}$ values is a lookup table—it has zero residual entropy (perfect prediction) but requires 2^{128} parameters and generalizes not at all. Factoring into 128 binary ESes gives the same *expressiveness* in 128 bits but makes the structure *learnable*: patterns over small ESes generalize, patterns over the full joint space do not.

This is the thermodynamic insight made precise: the choice of macrostate variables is not arbitrary. Good macrostates are those whose multiplicities $W(m)$ are highly non-uniform—macrostates where most microstates cluster into a few large classes, leaving a few rare outliers. This non-uniformity is exactly what makes compression possible: frequent macrostates get short codes, rare ones get long codes, and the average code length equals the Shannon entropy.

8 The Prime Encoding Makes This Computable

In the prime power encoding [4], each event space $\mathcal{E}_k$ is assigned a prime p_k, and each event value v becomes the exponent p_k^v. A position's full state is $S_t = \prod_k p_k^{v_k(t)}$.

Forming a macrostate by ignoring event space $\mathcal{E}_j$ is *dividing out* the prime p_j:

$$S_t^{(\text{ignore } j)} = S_t / p_j^{v_j(t)} \tag{10}$$

Two positions are in the same macrostate iff their residuals (after dividing out the ignored primes) are equal. The multiplicity $W(m)$ is the number of positions with the same residual.

The Boltzmann entropy of each macrostate is $\log_2 W(m)$. The Shannon entropy is $\log_2 N - \langle S_B \rangle$. The quotient is $Q(m) = N/W(m) = \lambda(m)$.

All of these are computable from the prime factorizations of the S_t. The prime encoding makes the entropy bridge *algorithmic*: we can compute the entropy at any level of factoring by choosing which primes to examine.

Example 3 (Two levels of factoring on 1024 bytes)**.**

Level 0: No factoring. *One macrostate containing all $N = 1024$ positions. $W = 1024$, $S_B = \log_2 1024 = 10$ bits. $H = \log_2 N - S_B = 0$: no residual surprise, because we have specified nothing about the position.*

Level 1: Input character. *52 macrostates. The "space" macrostate has $W = 127$, entropy $\log_2 127 = 6.99$ bits. The overall Shannon entropy $H(\text{input}) = 4.74$ bits: knowing the input character leaves 4.74 bits of residual surprise about the output.*

Level 2: Input + 8 skip offsets (UM). *834 macrostates (the skip-8 pattern inventory). Most macrostates have $W \approx 1$–2, leaving 0.043 bpc residual.*

Level 3: Full specification (all primes). *1024 macrostates (one per position). Each has $W = 1$, entropy 0. $H = \log_2 1024 - 0 = 10$ bits, but $\log_2 N - \log_2 N = 0$ residual: perfect identification, zero compression needed.*

9 The Thermodynamic Partition Function

The analogy deepens. In statistical mechanics, the partition function $Z = \sum_i e^{-\beta E_i}$ sums over microstates i weighted by their energy E_i. The free energy is $F = -kT \ln Z$, and the entropy is $S = -\partial F / \partial T$.

In the event formalism, the analogous sum is:

$$Z_{\mathcal{S}} = \sum_{m \in \text{macrostates}(\mathcal{S})} W(m) \cdot 2^{-\Lambda(m)} \tag{11}$$

where $\mathcal{S}$ is the set of observed event spaces, $W(m)$ is the multiplicity of macrostate m, and $\Lambda(m)$ is the log-luck (surprisal) assigned by the model. Since $2^{-\Lambda(m)} = p(m)$:

$$Z_{\mathcal{S}} = \sum_m W(m) \cdot p(m) = \sum_m W(m) \cdot \frac{W(m)}{N} = \frac{1}{N} \sum_m W(m)^2 \tag{12}$$

The "inverse temperature" β maps to the model's confidence: a sharp model (high β, low temperature) assigns high probability to a few macrostates; a diffuse model (low β, high temperature) spreads probability uniformly.

The "free energy" $F = -\log_2 Z$ is the effective description length: the number of bits needed to encode the data under the model's macrostate description. Minimizing free energy is identical to minimizing bpc—the thermodynamic variational principle *is* the compression objective.

10 Why This Matters for the UM

The entropy bridge connects three things that the preceding papers established independently:

1. $Q = \lambda$ [2]: the quotient over dataset positions equals the luck of events. This is the microstate/macrostate correspondence stated as an identity.

2. **Pattern chains** [3]: the UM's pattern inventory defines a specific factoring of the event space. Each pattern is a conjunction of events (a macrostate description), and its data-term count is the multiplicity W.

3. E **onto** $\mathbb{N}$ [4]: the prime encoding makes factoring computable. Choosing which primes to read selects the macrostate description; the multiplicities are computed by grouping equal residuals.

The entropy bridge unifies these: the UM's compression rate is the conditional Shannon entropy of the output given the pattern-chain macrostates, which equals $\log_2 N$ minus the expected Boltzmann entropy of the pattern classes, which equals the average log-luck (bpc).

The factoring hierarchy (Section 6) is the UM's *learning trajectory*: starting from no factoring (uniform prediction, 8.0 bpc), the UM discovers event spaces that reduce the residual entropy. Unigram factoring gives 4.74 bpc. Bigram gives $\sim$2.8. The skip-8 UM reaches 0.043. Each step adds event spaces (primes) to the macrostate description, collapsing more microstates into larger equivalence classes, absorbing more of $\log_2 N$ into the Boltzmann term.

The sat-rnn takes a different path through the same hierarchy: it discovers 128 binary event spaces (neuron signs) whose factoring achieves 0.079 bpc. The factor map [5] connects the RNN's factoring to the UM's factoring, showing that each neuron is a 2-offset conjunction detector ($R^2 = 0.837$) that implements a specific macrostate boundary.

11 The Second Law and Pattern Discovery

The second law of thermodynamics states that entropy increases in isolated systems. In the compression setting, the analogous statement is: *as the model learns more patterns, the conditional entropy (bpc) decreases*—the model moves from high-entropy (uniform) to low-entropy (compressed) descriptions.

But this requires an important inversion. In thermodynamics, entropy increases because the system explores more microstates. In compression, entropy *decreases* because the model finds better macrostates. The system (the data) is fixed; what changes is the *observer's description* (the model's factoring).

This is Maxwell's demon made concrete. The model acts as a demon that sorts microstates (positions) into macrostates (pattern classes), extracting "work" (compression) from the data. The demon's memory (the model's parameters) stores the factoring. Landauer's principle [1] then implies: the energy cost of the model's computation bounds the compression it can achieve. This is why the finiteness of F and Ω in the CMP formalism is not an arbitrary assumption but a physical necessity.

12 Concrete Numbers

For the 1024-byte enwik9 dataset:

Model/Factoring	Macrostates	Avg. $\log_2 W$	bpc
Uniform (no factoring)	1	10.0	8.00
Unigram (input char)	52	5.26	4.74
Bigram (input + prev)	231	7.22	∼2.8
sat-rnn (128 neurons)	∼1000	9.92	0.079
Skip-8 UM (834 pat.)	∼834	9.96	0.043
Full (all ESes)	1024	10.0	0.000

The "Avg. $\log_2 W$" column is $\langle S_B \rangle$, the expected Boltzmann entropy. The bpc column is $H = \log_2 N - \langle S_B \rangle$ conditioned on the model's predictions (which may be better than the macrostate entropy alone, since the model also uses pattern strengths to weight events within macrostates).

The skip-8 UM achieves lower bpc than the sat-rnn *despite having fewer macrostates* (834 vs. ∼1000) because its macrostates are more informative—each skip-8 pattern class has nearly uniform output distributions, leaving very little residual entropy.

13 Conclusion

The Shannon–Boltzmann bridge is not an analogy. It is the same mathematics applied to the same objects: microstates (positions), macrostates (event classes), multiplicities (counts), and entropy (log-counts).

The CMP formalism makes this bridge explicit by treating events as first-class objects with a prime-power encoding that makes factoring computable. The hierarchy of factorings—from no description (8.0 bpc) to full specification (0.0 bpc)—is the same as the hierarchy from thermodynamic equilibrium (maximum entropy) to complete microscopic knowledge (zero entropy).

The UM's learning problem is: find the factoring (the set of macrostate variables) that minimizes conditional entropy (bpc) subject to the constraint that the factoring must be *expressible* as patterns over the event spaces. This is the free energy minimization principle, applied to data compression rather than to physical systems.

The result is the same in both domains: the optimal description is the one that captures the most structure (absorbs the most of $\log_2 N$ into the Boltzmann term) with the fewest macrostate variables (the simplest factoring). In thermodynamics, this is temperature, pressure, and volume. In compression, this is the pattern inventory. In both cases, the macrostate variables are discovered, not given—and the quality of the discovery is measured by the entropy it explains.

Reproducibility

This paper is analytical. It builds on results from:

- `export-gap.pdf` (Section 6.6: $Q = \lambda$ derivation)
- `pattern-chains.pdf` (Sections 5–7: data-terms, skip-patterns)
- `event-arithmetic.pdf` (prime power encoding, GMP experiment)
- `prime-examples.pdf` (quotient classes, information content)
- `factor-map.pdf` (neuron → 2-offset conjunction, $R^2 = 0.837$)

Numerical values (bpc, pattern counts, macrostate counts) are from the tools in those papers' archives. The entropy bridge (Theorem 1) is a derivation, not an experiment.

References

[1] Michaeljohn Clement. CMP. 2026. https://cmpr.ai/cmp.pdf

[2] Claude and MJC. The SN Export Gap. 7 Feb 2026.

[3] Claude and MJC. Pattern Chains. 8 Feb 2026.

[4] Claude and MJC. Event Arithmetic: E onto $\mathbb{N}$. 10 Feb 2026.

[5] Claude and MJC. The Factor Map. 9 Feb 2026.