# Appendix D: Free Energy Correspondence

**For:** *Memory Without Storage, Learning Without a Learner* (Neale, 2026)

**Status:** Sketch. A full derivation connecting the observer-inference framework to variational and thermodynamic free energy will appear in a forthcoming paper (Neale, in preparation). This appendix establishes the structural correspondences that motivate that development.

---

## D.1 Overview

The three observer measures — M, L, D — map onto components of the variational free energy as defined by Friston (2010) and, through a separate correspondence, onto aspects of the Helmholtz free energy in classical thermodynamics. These are not analogies. The measures are computed from the same mathematical objects (prediction errors under incomplete observation) that appear in both frameworks, reinterpreted as properties of the observation relation rather than of the system.

---

## D.2 Variational Free Energy Decomposition

Friston's variational free energy for an observer with generative model q observing data o from a system with hidden states s is:

$$\mathbf{F = D\_KL[q(s) \parallel p(s|o)] - \ln p(o)}$$

This decomposes into:

**F = Complexity − Accuracy**

where Complexity = $D\_KL[q(s) \parallel p(s)]$ measures how far the observer's model deviates from the prior, and Accuracy = $E\_q[\ln p(o|s)]$ measures how well the model predicts the observations.

In our framework, the observer maintains a rolling mean model q (the prediction from recent emissions) and evaluates it against each new observation $o = E(t)$. The three measures correspond to distinct modes of failure in this predictive cycle.

---

## D.3 Mapping M, L, D to Free Energy Components

### L (Learning) ↔ Free Energy Minimisation

Our L measure is the fractional decrease in mean surprise between early and late phases:

$$\mathbf{L = (\bar{s}\_early - \bar{s}\_late) / \bar{s}\_early}$$

Surprise $s(t)$ is the observer's prediction error — the empirical proxy for $-\ln p(o|q)$, the negative log-evidence under the model. Mean surprise over a phase is therefore a proxy for the expected free energy over that phase.

$L > 0$ means $F_{late} < F_{early}$: the observer's free energy decreased. This is precisely Friston's free energy minimisation, with two reframings:

1. The "system" that minimises free energy is the observer's model, not the observed system. The system follows its own dynamics; the observer's model improves because the emission stream becomes more predictable. Whether the system "is learning" or the observer "is learning about the system" produces identical emission profiles — and therefore identical L values.
2. The activity gate (late emissions must still be variable) ensures that F decreased through improved accuracy, not through the trivial route of the system ceasing to produce data. A dead system has $F \to 0$, but this is not learning — it is the absence of observation.

**Formal correspondence:**

Let $F(t) = -\ln p(E(t) \mid q_t)$ be the instantaneous free energy under the observer's model at time t. Then:

$$\mathbf{L \approx (\langle F\rangle_{early} - \langle F\rangle_{late}) / \langle F\rangle_{early}}$$

where $\langle\cdot\rangle$ denotes phase averages. L is the fractional free energy reduction, gated by the system remaining an active source of observations.

**M (Memory) ↔ Model Divergence (Complexity)**

Our $M_{pred}$ measure compares two models: one trained on early data ($q_{early}$) and one trained on late data ($q_{late}$). Memory is detected when $q_{early}$ fails on late data:

$$\mathbf{M = \langle F(E_{late} \mid q_{early})\rangle - \langle F(E_{late} \mid q_{late})\rangle}$$

This is the excess free energy incurred by using the wrong model — which is structurally identical to a KL divergence between the early and late generative distributions:

$$\mathbf{M \approx D_{KL}[p_{late}(o) \parallel p_{early}(o)]}$$

In Friston's framework, this corresponds to the Complexity term: the divergence between the observer's current model and its prior. A system that "remembers" ($M > 0$) is one where the emission distribution shifted persistently, forcing the observer to update its model. The free energy cost of failing to update — of clinging to the prior — is exactly M.

**Formal correspondence:**

The observer's free energy under the early model applied to late data is:

**F(q_early, late) = D_KL[q_early ‖ p_late] − Accuracy_late**

Under the late model:

**F(q_late, late) = D_KL[q_late ‖ p_late] − Accuracy_late**

The difference eliminates the Accuracy term:

**M = F(q_early, late) − F(q_late, late) = D_KL[q_early ‖ p_late] − D_KL[q_late ‖ p_late]**

This is the excess divergence of the stale model. When the emission regime has shifted (p_late ≠ p_early), q_early is far from p_late and q_late is close — the difference is large and M is positive.

### D (Decision) ↔ Free Energy Spikes

Our D measure detects timesteps where surprise exceeds the local baseline by more than 2 MAD:

**D(t) = 1 if s(t) > median(s_local) + 2 × MAD(s_local)**

In free energy terms, a D event is a timestep where the instantaneous free energy F(t) spikes above the local running average. The observer's model, which was tracking the emission stream adequately, suddenly fails. The system produced an emission that the model assigned very low probability.

In Friston's framework, these correspond to transitions between regimes — moments where the generative model encounters data outside its current basin of attraction. The system may have crossed a boundary in its hidden state space, entered a new attractor, or undergone a phase transition. The observer experiences this as a sudden increase in prediction error — a free energy spike.

**Formal correspondence:**

**D(t) = 1 if F(t) ≫ ⟨F⟩_local**

where ⟨F⟩_local is the running average free energy and ≫ is operationalised as exceeding 2 MAD. The D_rate is the fraction of timesteps where the observer's free energy exceeds its local equilibrium — the fraction of time the system is in transition, as seen by the observer.

---

## D.4 Summary Mapping

| Observer measure | Free energy component | Friston term | Interpretation |
|---|---|---|---|
| L (learning) | ΔF < 0 | Free energy minimisation | Observer's model improving |

| Observer measure | Free energy component | Friston term | Interpretation |
|---|---|---|---|
| M (memory) | D_KL[q_old ‖ p_new] | Model complexity / divergence | Observer's model went stale |
| D (decision) | F(t) spike | Regime transition | Observer's model suddenly failed |
| M∩L∩D (overlap) | Structured F minimisation with transitions and regime shifts | Active inference | Full predictive cycle |

---

## D.5 Thermodynamic Correspondence

The Helmholtz free energy F = U − TS contains a term TS that represents the entropy of the system's microscopic state — the information content of configurations the observer cannot see. This term is the thermodynamic counterpart of the observation gap.

| Observer measure | Thermodynamic aspect | Connection |
|---|---|---|
| M | The TS landscape shifted | The entropy of accessible microstates changed persistently. The system's macroscopic observables moved because the hidden microstate distribution changed and stayed changed. |
| L | The observer's estimate of TS improved | The observer can better predict which microstates are being visited, reducing the effective TS uncertainty. The macroscopic description became more informative. |
| D | Sudden rearrangement of microstates | A phase transition, barrier crossing, or regime change in the hidden state space produced a macroscopic observable that violates the observer's running estimate of TS. |

The key insight: in both the variational and thermodynamic framings, M/L/D are aspects of the observer's relationship to the hidden degrees of freedom. They measure how the observer's incomplete model performs, not what the system intrinsically does. The variational framing locates this in the KL divergence between model and truth. The thermodynamic framing locates it in the TS term that the observer cannot directly measure. Both point to the same structure: cognition-as-inference arises from the gap between what the observer sees and what the system does.

---

## D.6 The Hopfield Correspondence

Hopfield networks (Hopfield 1982) provide a concrete instantiation of the correspondence. The energy function $E = -\frac{1}{2} \Sigma w_{ij} s_i s_j$ defines a landscape in configuration space. The network's dynamics are gradient descent on this landscape. An observer watching the network's state (or its output emissions) would measure:

**M:** The network has settled into a specific attractor basin. The early model (when the network was in a high-energy state exploring the landscape) fails on late data (when the network has reached the basin floor). The emission regime shifted persistently.

**L:** The observer's surprise decreases as the network descends toward an attractor. The trajectory becomes more predictable as the network commits to a basin.

**D:** Transitions between basins — when the network's state jumps from one attractor to another (due to noise, perturbation, or ambiguous input) — produce surprise spikes. The observer registers a "decision."

The Hopfield capacity limit (0.14N memories for N neurons) is the point where attractor basins overlap sufficiently that the observer's model cannot reliably predict which basin the network will reach — the landscape becomes too complex for the observer to track. Beyond this limit, M remains high (the network still settles into basins) but L decreases (the observer can't predict which basin) and D becomes frequent (the network crosses basin boundaries unpredictably). The $M \cap L \cap D$ overlap collapses — the observer no longer infers "intelligent" behaviour but "confused" behaviour.

This is a prediction our framework makes about Hopfield networks that the standard analysis does not: the observer's attribution of memory and learning to the network should degrade at capacity in a specific, measurable way ($L \to 0$, $D \to$ high), not just in the network's retrieval accuracy.

---

## D.7 Scope and Limitations

This appendix establishes structural correspondences, not proofs of equivalence. Several steps require elaboration:

1. **The rolling mean predictor is not a full variational model.** A proper variational treatment would use a parameterised family $q_\theta$ updated by gradient descent on F. Our rolling mean is a special case (the posterior is a Dirac delta on the running mean). The correspondence holds structurally but the quantitative relationship between our L and Friston's $\Delta F$ depends on the model family. Note, however, that this limitation is conservative: a more capable model would produce lower surprise and therefore lower M, L, and D values. The Cognitive

Hierarchy (Section 2.2) predicts that observer sophistication and cognitive attribution are inversely related. The rolling mean produces an upper bound on the correspondence values, not a lower bound.

2. **The M mapping to D_KL is approximate.** The early model is a static mean, not a full generative distribution. The KL divergence interpretation is exact only if both models are Gaussian with equal variance — which is approximately true for slowly varying emission streams but not guaranteed.
3. **The thermodynamic mapping requires an equilibrium assumption.** The Helmholtz free energy is defined at equilibrium; our systems are generally far from equilibrium. The correspondence is structural (both involve hidden degrees of freedom inaccessible to the observer) but the quantitative relationship between our measures and physical entropy production requires the non-equilibrium framework developed in Neale (2025).
4. **The Hopfield predictions are untested.** The predicted degradation pattern at capacity (L → 0, D → high, M stable) is a specific, falsifiable prediction that could be tested computationally. We note this as proposed future work.

These limitations do not affect the main results of the paper, which rest on the computational demonstrations (Sections 3–4) rather than on the free energy correspondence. The correspondence motivates and contextualises the results; it does not underwrite them.

---

## References for Appendix D

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11(2), 127–138.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris* 100(1-3), 70–87.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79(8), 2554–2558.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical Review* 106(4), 620–630.

Neale, D. (2025). Being from Nothingness: Deriving the Structure of Existence from a Single Axiom. Preprint, goleudy.ai.