Causal-Coding Energy Models with Differentiable Hamiltonian Relational Constraints

Ben Goertzel

December 7, 2025

Abstract

We describe a self-contained way to combine (i) continuous predictive-coding style inference and learning with (ii) discrete, logic-like relational constraints expressed as Hamiltonian energies on graphs. The central idea is to add a differentiable relational energy term to the model's free energy, treat the relational machinery as a context-gated module (causal coding), and ensure smooth gradients by (a) relaxing binary states to probabilities in $[0, 1]$ and (b) keeping graph topology fixed while predicting soft edge weights. The same relational energies can also bias transformer attention using sparse candidate pairs to avoid dense $O(T^2)$ relational computation.

1 Problem and design goals

Many tasks mix two kinds of structure:

- **Inductive structure:** smooth statistical regularities well captured by standard neural representations (continuous vectors, linear maps, attention).
- **Abductive relational structure:** role and constraint patterns such as "these two variables satisfy relation r" or "these nodes form a known part/motif." These are naturally expressed as graph constraints and logic-like patterns.

We want a single training/inference objective in which both structures contribute to prediction and to causal estimation, without breaking differentiability or computational budgets.

2 Energy-based inference with context-gated modules

2.1 Predictive-coding free energy

Let x be data and $z = \{z^{(\ell)}\}_{\ell=1}^{L}$ be latent states in a multilayer model. A standard energy (free energy) form is

$$F_{\mathrm{PC}}(x, z; \theta) = \sum_{\ell=1}^{L} \frac{1}{2\sigma_\ell^2} \left\| z^{(\ell-1)} - f_\ell(z^{(\ell)}; \theta_\ell) \right\|^2 + \mathrm{priors}(z, \theta), \tag{1}$$

with $z^{(0)} = x$. Inference updates latents by gradient descent:

$$z \leftarrow z - \eta_z \nabla_z F_{\mathrm{PC}}(x, z; \theta). \tag{2}$$

Learning updates parameters by gradient descent (possibly after inference converges):

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta F_{\mathrm{PC}}(x, z; \theta). \tag{3}$$

2.2 Causal coding as context-gated modular updates

Partition parameters into modules $\theta = (\theta_1, \ldots, \theta_K)$. Let c denote a context (e.g., task, domain, regime). Introduce gates $g_{k,c} \in \{0, 1\}$ so that only a context-specific support set S_c updates:

$$\theta_k \leftarrow \theta_k - \eta_\theta \, g_{k,c} \, \nabla_{\theta_k} F(x, z; \theta), \qquad k = 1, \ldots, K. \tag{4}$$

This reduces interference across contexts by limiting which modules change in each context.

3 Hamiltonian relational constraints on graphs

3.1 Binary relations as "zero when satisfied" energies

Consider binary node states $\psi_i \in \{0, 1\}$. For a pair $(x, y) \in \{0, 1\}^2$ and a relation type r from a small set $\mathcal{R}$, define a quadratic energy

$$E_r(x, y) = \begin{bmatrix} x & y \end{bmatrix} H_r \begin{bmatrix} x \\ y \end{bmatrix} + k_r, \tag{5}$$

chosen so that

$$E_r(x, y) = 0 \text{ if } (x, y) \text{ satisfies relation } r, \qquad E_r(x, y) > 0 \text{ otherwise.}$$

Thus E_r measures constraint violation.

3.2 Graph energy and parts

For a labeled graph G on n nodes with state vector $\psi \in \{0, 1\}^n$, compose pairwise energies into a global quadratic energy

$$E_G(\psi) = \psi^\top H_G \psi + k_G, \tag{6}$$

where H_G is sparse and equals a sum of embedded 2×2 blocks derived from the edges and their relation types.

A **part** (or motif) is a subgraph p with its own energy $E_p(\psi)$. A state matches a part if $E_p(\psi) = 0$. Collections of parts can represent reusable relational patterns and role structure.

4 Differentiable interface: soft states and binary-consistent relaxation

Predictive-coding latents z are continuous. Directly thresholding z into $\psi \in \{0, 1\}^n$ is non-differentiable. We therefore use a soft relaxation.

4.1 Soft node states

Let $u \in \mathbb{R}^n$ be logits derived from latents (for example $u = W z^{(\ell^\star)} + b$). Define relaxed node states

$$p = \sigma(u/T) \in (0, 1)^n, \tag{7}$$

where σ is sigmoid and $T > 0$ is a temperature. Interpret p_i as the probability that node i is "on."

4.2 Binary-consistent (pseudo-Boolean) energy extension

The quadratic form in Eq. (5) is designed for $x, y \in \{0, 1\}$, where $x^2 = x$ and $y^2 = y$. To avoid unintended behavior on $(0, 1)$, rewrite the pair energy in a polynomial that is equivalent on binary inputs and then extend that polynomial to $[0, 1]$.

Write

$$E_r(x, y) = h_{11}x^2 + (h_{12} + h_{21})xy + h_{22}y^2 + k_r.$$

On $\{0, 1\}$ this equals

$$E_r(x, y) = h_{11}x + (h_{12} + h_{21})xy + h_{22}y + k_r. \tag{8}$$

Define the relaxed pair energy by the same polynomial for $(p, q) \in [0, 1]^2$:

$$\widetilde{E}_r(p, q) = h_{11}p + (h_{12} + h_{21})pq + h_{22}q + k_r. \tag{9}$$

This equals $\mathbb{E}[E_r(X, Y)]$ under independent Bernoulli variables $X \sim \text{Bern}(p)$ and $Y \sim \text{Bern}(q)$, which is a natural relaxation for logic-like energies.

Similarly, rewrite any graph energy into linear and bilinear terms (dropping squares using $x^2 = x$) and extend it by substituting p_i for ψ_i.

4.3 Optional binarization regularizer

Relational constraints often encourage near-binary roles. To make this explicit and tunable, add

$$F_{\text{bin}}(p) = \beta \sum_{i=1}^{n} p_i(1 - p_i), \tag{10}$$

which is minimized at $p_i \in \{0, 1\}$. Temperature T controls saturation: larger T keeps gradients alive earlier; annealing T later can sharpen roles if desired.

5 Smooth graph builder on a fixed edge template

Building a graph by hard thresholding edges can create discontinuous energies. Instead, fix a candidate edge set and learn soft edge weights and edge-type mixtures.

5.1 Fixed candidate edges and soft edge types

Let E_0 be a fixed set of candidate edges, such as:

- local neighbors (vision),
- known relations (knowledge graph),
- syntax neighborhood (language),
- or a learned sparse proposal set.

For each $(i, j) \in E_0$, predict:

- a soft mask $m_{ij}(u) \in [0, 1]$ (whether the edge matters),
- relation-type mixture weights $w_{ij,r}(u) \geq 0$ with $\sum_{r \in \mathcal{R}} w_{ij,r}(u) = 1$.

All of these can be produced by small neural heads applied to $z^{(\ell^\star)}$.

5.2 Relational energy on the fixed template

Define a relational energy as a weighted sum of relaxed pair energies:

$$F_{\mathrm{rel}}(u) = \sum_{(i,j)\in E_0} m_{ij}(u) \sum_{r\in\mathcal{R}} w_{ij,r}(u)\, \widetilde{E}_r(p_i(u), p_j(u)), \tag{11}$$

optionally plus additional terms that compose edges into parts/motifs (still using continuous masks and weights on the fixed template). Because E_0 is fixed and all operations are continuous, the energy landscape remains smooth enough for gradient-based inference.

6 Combined model: PC plus relational constraints with causal-coding gates

Choose a layer $\ell^\star$ whose latent state drives the relational module. Let $u = u(z^{(\ell^\star)})$, $p = \sigma(u/T)$, and F_{rel} as above. Define the total free energy:

$$F_{\mathrm{total}}(x, z; \theta, \theta^{(H)}) = F_{\mathrm{PC}}(x, z; \theta) + \lambda F_{\mathrm{rel}}(u(z^{(\ell^\star)}); \theta^{(H)}) + F_{\mathrm{bin}}(p(u)). \tag{12}$$

Here $\theta^{(H)}$ are the parameters of the relational module (edge masks, edge-type predictors, part templates).

Inference updates latents by Eq. (2) using F_{total}. Learning updates parameters using gated module updates (Eq. (4)) applied to both base parameters and relational-module parameters. In contexts where relational constraints are useful, gates keep relational modules active; otherwise they can be disabled to save compute and reduce interference.

Interpretation. F_{PC} explains data using continuous inductive structure, while F_{rel} rewards configurations that satisfy learned relational motifs. Their sum yields a unified objective in which abductive relational regularities can directly shape latent inference and downstream predictions.

7 Transformer integration: Hamiltonian attention bias on sparse candidate pairs

Let a sequence have length T. Standard attention logits for head h are

$$\ell_{ij}^{(h)} = \frac{\langle q_i^{(h)}, k_j^{(h)} \rangle}{\sqrt{d}}.$$

Define relaxed node states p_i for tokens (from token embeddings) and optional edge features e_{ij} (distance, dependency label, etc.). A relational (Hamiltonian) bias can be added as

$$\ell_{ij}^{(h)} \leftarrow \ell_{ij}^{(h)} - \gamma\, \widetilde{E}_{\mathrm{pair}}^{(h)}(p_i, p_j, e_{ij}), \tag{13}$$

where $\widetilde{E}_{\mathrm{pair}}^{(h)}$ is a relaxed pair energy of the form in Eq. (9) (possibly conditioned on e_{ij}).

To avoid dense $O(T^2)$ relational work, compute Eq. (13) only for a sparse candidate set. For each token i, let $C(i)$ be a set of $k \ll T$ indices (top-k under base logits, a local window, or a syntax neighborhood). Apply the bias only for $j \in C(i)$, giving $O(Tk)$ relational computation. As in the PC case, treat Hamiltonian heads as modules and gate them by context.

8 Optional: structure selection by evidence/weakness

A separate controller can compare alternative modular factorizations f (different module partitions, gating schemes, and relational templates) using a local evidence score:

$$Z_{n,c}(W; f) = \int_{W(f)} \exp(-n\mathcal{L}_c(\theta))\, \pi(\theta)\, d\theta, \tag{14}$$

where $\mathcal{L}_c$ is a context-conditional loss, π a prior, and $W(f)$ a neighborhood consistent with factorization f. A relational module contributes additional evidence by achieving low $\widetilde{E}$ on context c. The controller can allocate relational capacity only when it improves evidence enough to justify its complexity.

9 Summary

The approach is defined by three concrete choices:

1. **Differentiable states:** use soft node states $p \in (0, 1)^n$ derived from continuous latents.
2. **Binary-consistent energies:** extend logic-designed energies to $[0, 1]$ using pseudo-Boolean (relaxed expected) forms, optionally with an explicit binarization regularizer.
3. **Smooth, sparse graphs:** keep a fixed candidate edge set and predict soft masks and edge-type mixtures; in transformers, apply relational biases only to sparse candidate pairs.

These choices let Hamiltonian relational constraints act as ordinary energy factors and as ordinary context-gated modules, so abductive relational motifs and inductive neural features jointly guide inference, prediction, and continual learning.