RiJEPA: Rule-Informed Neuro-Symbolic Joint-Embedding Predictive Architecture

Anonymous authors
Paper under double-blind review

Abstract

Modern Joint-Embedding Predictive Architectures (JEPAs) learn powerful representations from data but lack explicit mechanisms for incorporating symbolic knowledge, while traditional rule systems are interpretable but rely on rigid discrete representations and combinatorial search. We introduce **RiJEPA**, a Rule-Informed Neuro-Symbolic JEPA that connects symbolic reasoning and predictive representation learning in both directions. In the *rules-for-JEPA* direction, RiJEPA combines Energy-Based Constraints (EBC) with a multi-modal dual-encoder architecture to align continuous observations and symbolic rules in a shared latent space, shaping valid rules into low-energy basins while repelling contradictory or invalid rules. In the reverse *JEPA-for-rules* direction, the resulting differentiable rule-energy landscape replaces discrete rule search with gradient-guided exploration, supporting joint rule generation, forward inference, abductive reasoning, and marginal-predictive translation. On a controlled simulation, RiJEPA preserves near-zero energy for valid rules while increasing the energy of invalid OOD rules by approximately $85\times$ relative to a classic JEPA. On the UCI Heart Disease case study, RiJEPA organizes patient representations around symbolic diagnostic poles, enabling prediction by latent-space proximity without training a downstream classifier and supporting generative rule reasoning. Together, these results establish a bidirectional neuro-symbolic principle: symbolic rules can shape predictive representations, while the resulting representations provide a continuous, differentiable space for reasoning over and discovering rules.

1 Introduction

Self-supervised predictive architectures excel at capturing statistical structure in high-dimensional, noisy data, yet they act largely as black boxes with no inherent mechanism to internalize structured human knowledge or adhere to verifiable logic. Conversely, rule-based inference systems such as association rule mining (ARM) (Zhang & Zhang, 2002) and fuzzy logic (Zadeh, 1965) offer interpretable symbolic logic but suffer from rigid discrete boundaries, high sensitivity to thresholds, and combinatorial explosion at scale (Zheng et al., 2001). A promising paradigm lies in their synthesis: the continuous representational power of neural networks combined with the compositional rigour of symbolic logic.

Joint-Embedding Predictive Architectures (JEPAs) (LeCun, 2022) have emerged as a powerful foundation for latent-space representation learning, filtering out low-level stochasticity to focus on high-level semantics. However, current JEPAs are purely data-driven: they organize representations by statistical proximity alone, which leaves them vulnerable to spurious correlations, shortcut learning, and out-of-distribution (OOD) failures. We study the *bidirectional* synthesis of continuous predictive manifolds and discrete symbolic logic, and ask two questions:

1. **Rules for JEPA:** can rule-informed training inject explicit structural inductive biases into a JEPA, geometrically shaping its latent space so that representations are anchored to logical truths and robust to OOD inputs?
2. **JEPA for Rules:** can converting discrete, combinatorial rule spaces into continuous, differentiable embedding manifolds enable gradient-based generative rule discovery that bypasses traditional combinatorial search?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We provide evidence for both with a unified framework in which rule validity is expressed as a learned energy in the JEPA latent space, so that *logical validity equates to geometric proximity*. Our contributions are:

- **RbJEPA and RiJEPA.** We introduce Rule-based JEPA (RbJEPA), which distils pure symbolic knowledge into a continuous predictive space, and Rule-informed JEPA (RiJEPA), a hybrid objective that augments self-supervised prediction on raw data with *Energy-Based Constraints* (EBC) that carve low-energy basins for valid rules and repel corrupted rules into high-energy regions; at zero loss EBC provably separates valid and invalid rules by the margin, unlike pairwise supervision (section 3, Proposition 1).
- **Multi-modal dual-encoder architecture.** We propose an architecture with separate data and rule encoders and a single shared, modality-agnostic predictor, bridging the dimensional and semantic gap between continuous observations and discrete logic, and enabling classification by distance to symbolic poles without training a downstream classifier (section 3.4).
- **Continuous rule discovery.** We formalize rule manifolds as sublevel sets of the learned energy and introduce two generative discovery paradigms, joint energy-based Langevin diffusion (with unconditional, forward, and abductive modes) and marginal-predictive guided discovery, replacing combinatorial search with gradient-guided exploration via $\nabla_z E(z)$ (section 4).
- **Empirical evidence.** On a controlled simulation, EBC sculpts well-separated energy basins and raises the energy of OOD invalid rules by nearly two orders of magnitude over a classic JEPA. On a clinical case study, RiJEPA reaches 100% nearest-pole diagnostic accuracy without a downstream classifier and produces clinically plausible rules via forward and abductive inference (section 5).

2 Related Work

Joint-embedding predictive architectures. JEPAs predict the latent representations of target regions from encoded context rather than reconstructing pixels or tokens (LeCun, 2022). I-JEPA (Assran et al., 2023) learns semantic visual features without hand-crafted augmentations, using an exponential moving average (EMA) target encoder (Grill et al., 2020) to prevent collapse; V-JEPA (Bardes et al., 2024) extends this to video; JEPA world models (Terver et al., 2026; Destrade et al., 2025) plan directly in embedding space. Probabilistic (Huang, 2026b) and bidirectional (Huang, 2026a) variants have also been proposed. All remain fundamentally data-driven and have no mechanism for internalizing discrete, human-elicited structural priors.

Neuro-symbolic AI. Neuro-symbolic methods combine connectionist pattern recognition (Hinton, 1990) with symbolic reasoning (Besold et al., 2017; d'Avila Garcez et al., 2015; Dash et al., 2022). Differentiable ILP (Evans & Grefenstette, 2018) relaxes logic into continuous weights; Logic Tensor Networks (Serafini & d'Avila Garcez, 2016; Serafini et al., 2017; Badreddine et al., 2022) ground predicates in real-valued spaces; other lines use LLMs as symbolic executives (Grosvenor et al., 2025), logical compositions for question answering (Cao et al., 2021), or rules as regularizers for knowledge-graph embeddings (Cheng et al., 2024). Rather than using rules to regularize a final classifier, RiJEPA uses them as geometric constraints on a *self-supervised* joint-embedding manifold.

Metric learning. Deep metric learning shapes embedding spaces so that similar items are close and dissimilar ones far apart (Jain et al., 2008; Hoffer & Ailon, 2015; Wang et al., 2020; Mohan et al., 2023), via proxy anchors (Kim et al., 2020) or triplet margins (Weinberger & Saul, 2009; Schroff et al., 2015). Our EBC applies a related push–pull margin principle, but to *rule* pairs through a JEPA predictor, and our anchor loss ties data embeddings to symbolic poles.

Association rule mining. Classical ARM algorithms, e.g., Apriori (Agrawal & Srikant, 1994), FP-Growth (Han et al., 2000), and Eclat (Zaki et al., 1997), search discrete itemset spaces using support and confidence; they are threshold-sensitive and can produce super-exponential numbers of rules (Zheng et al., 2001; Pei, 2009). Bayesian (Tian et al., 2013; González et al., 2020) and reinforcement-learning or bandit-based methods (Fujioka & Shirahama, 2022; Huang, 2025) make the search adaptive but still operate in discrete combinatorial spaces. We use ARM only as a source of rules; devising new ARM methods is out of scope. To our knowledge, no prior framework unifies

self-supervised JEPA learning with discrete rule logic to support both rule-shaped representations and gradient-based generative and abductive rule reasoning. An extended discussion is given in Appendix A.

3 RULES FOR JEPA: RBJEPA AND RIJEPA

We assume rules of the form `IF` antecedent A `THEN` consequent C, extracted from data by any rule-producing mechanism (decision trees, ARM, fuzzy inference) or supplied by domain experts. Each rule is encoded as a structured object of feature–condition pairs together with statistical strengths (support, confidence).

3.1 RULE-BASED JEPA (RBJEPA): PURE SYMBOLIC DISTILLATION

RbJEPA trains a JEPA entirely on rules: the antecedent is the context and the consequent is the target, $z_c = f_c(A)$, $z_t = f_t(C)$, and a predictor g is trained with a strength-weighted loss

$$\mathcal{L}_{\text{RbJEPA}} = \textstyle\sum_{i=1}^{N} w_i \, \|g(z_c^{(i)}) - z_t^{(i)}\|_2^2, \tag{1}$$

where w_i is the confidence or support of rule i. Unlike a classic JEPA, which absorbs implicit correlations from raw observations, RbJEPA performs *structured abstraction distillation*: discrete logical transitions become continuous vector transformations in which similar antecedents cluster and valid rules are geometrically predictable. Because it predicts embeddings rather than rule strings, it is insensitive to low-level syntactic variation. The three-step pipeline and its relation to knowledge distillation (Hinton et al., 2015) and relational contrastive learning (Bordes et al., 2013) are detailed in Appendix B.

3.2 RULE-INFORMED JEPA (RIJEPA): ENERGY-BASED CONSTRAINTS

Rules summarize only the salient structure of data, so a purely rule-trained model can miss nuanced relations, while a purely data-driven JEPA entangles noise and spurious correlations. RiJEPA combines both sources: it presents the JEPA with raw data *and* rules, treating rules as an auxiliary modality that constrains the latent space. With $\mathcal{L}_{\text{JEPA}}$ the standard latent prediction loss on raw data,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{JEPA}} + \beta \, \mathcal{L}_{\text{EBC}}. \tag{2}$$

We define the energy of a rule as its predictive incompatibility, $E(A, C) = \|g(f_c(A)) - f_t(C)\|_2^2$, and impose a contrastive margin objective

$$\mathcal{L}_{\text{EBC}} = \textstyle\sum_{(A,C) \in \mathcal{R}_{\text{valid}}} E(A, C) \;+\; \lambda \sum_{(A,C_{\text{neg}}) \in \mathcal{R}_{\text{invalid}}} \max\big(0,\, m - E(A, C_{\text{neg}})\big), \tag{3}$$

where β balances rule and data terms, λ weights negatives, and m is the margin. Invalid pairs are generated on the fly by replacing a valid antecedent's consequent with a randomly sampled or structure-preserving corrupted alternative. This push–pull mechanism embeds valid logical transitions as low-energy basins and repels contradictory ones until they are at least m away. Crucially, although the energy has the same form as the RbJEPA loss, the two objectives differ in kind: *pairwise supervision learns a mapping, whereas energy-based training learns a landscape*, i.e., a bounded rule-validity surface that also constrains the space *between* valid rules (made precise in Section 3.3). RiJEPA therefore supervises *relationships* between entities rather than a teacher's outputs, which distinguishes it from classical knowledge distillation.

3.3 WHY ENERGY-BASED CONSTRAINTS? A MAPPING VERSUS A LANDSCAPE

The pairwise predictive supervision (PPS) of RbJEPA (Eq. (1)) and the EBC of RiJEPA (Eq. (3)) use the same energy $E(A, C)$ but constrain the optimization differently. PPS optimizes only valid pairs: it pulls them together but applies no force to invalid rules, so it learns a regression-like mapping $g(z_c) \approx z_t$ whose behaviour away from the training rules is unconstrained. EBC adds a margin on corrupted pairs, and the difference can be stated precisely.

Proposition 1 (Zero-loss energy separation)**.** *Let $\lambda > 0$, $m > 0$, and $w_i > 0$. (i) If $\mathcal{L}_{\text{EBC}} = 0$, then $E(A, C) = 0$ for every $(A, C) \in \mathcal{R}_{\text{valid}}$ and $E(A, C_{\text{neg}}) \geq m$ for every $(A, C_{\text{neg}}) \in$*

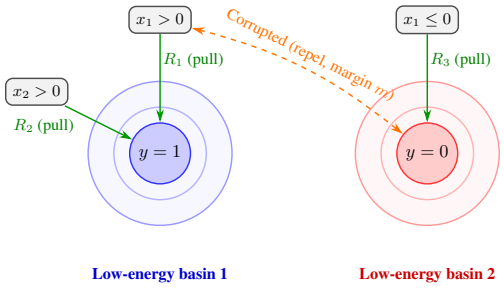


Figure 1: Geometry of EBC on three toy rules. Valid antecedents (grey) are pulled into low-energy basins around their consequents; corrupted pairings are repelled beyond the margin m.

Table 1: Pairwise predictive supervision (PPS, RbJEPA) versus energy-based constraints (EBC, RiJEPA).

PPS (RbJEPA)	**EBC (RiJEPA)**
Positive pairs only	Positive and negative pairs
Learns a conditional mapping	Learns a rule-validity landscape
Regression objective	Energy / margin objective
Invalid logic unconstrained	Invalid logic pushed to $E \geq m$
Admits collapse ($E \equiv 0$)	Collapse violates the margin
Stable and simple	More expressive, harder to optimize

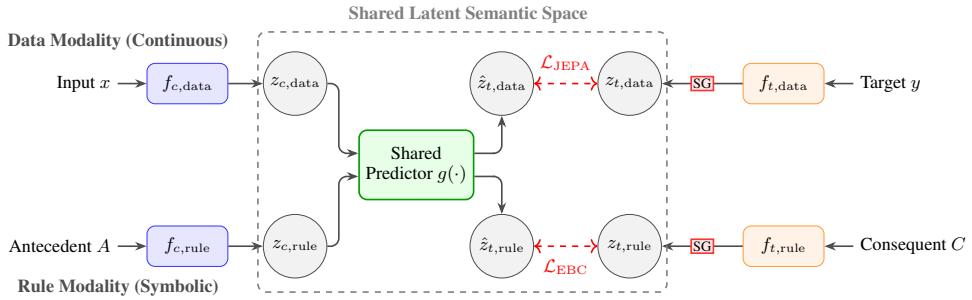


Figure 2: The multi-modal dual-encoder RiJEPA. Data contexts x and rule antecedents A are encoded by modality-specific encoders and processed by a single shared predictor g, whose outputs are compared with target encodings of data y and rule consequents C, jointly minimizing the data loss $\mathcal{L}_{\text{JEPA}}$ and the rule energy constraint $\mathcal{L}_{\text{EBC}}$. SG denotes stop-gradient. Masking the rule pathway recovers a standard data-only JEPA.

$\mathcal{R}_{\text{invalid}}$. *Consequently, under the Boltzmann density of Eq. (4), the latent pair of any valid rule is at least $e^{m/T}$ times as probable as that of any invalid rule. (ii) In contrast, $\mathcal{L}_{\text{RbJEPA}} = 0$ places no constraint on invalid pairs: in particular, the collapsed solution in which $g \circ f_c$ and f_t output the same constant attains zero PPS loss while assigning zero energy to* every *pair, valid or not.*

Proof. (i) Every term of $\mathcal{L}_{\text{EBC}}$ is non-negative, so a zero sum forces each valid energy to vanish and each hinge $\max(0, m - E(A, C_{\text{neg}}))$ to be zero, i.e., $E(A, C_{\text{neg}}) \geq m$. The density ratio is then $\exp\big((E_{\text{inv}} - E_{\text{val}})/T\big) \geq e^{m/T}$. (ii) If $g(f_c(A)) = f_t(C) = c$ for all A, C, then $E \equiv 0$, so the PPS loss is zero regardless of validity. □

Thus PPS *learns a mapping*, whereas EBC *learns a landscape*: absolute constraints on valid rules versus relative constraints that separate valid from invalid logic, which is what the generative paradigms of Section 4 rely on. Figure 1 illustrates the resulting geometry on three toy rules (R_1: $x_1 > 0 \to y = 1$, R_2: $x_2 > 0 \to y = 1$, R_3: $x_1 \leq 0 \to y = 0$): antecedents sharing a consequent collapse into one basin, R_3 forms a separate basin, and corrupted pairings such as $x_1 > 0 \to y = 0$ are repelled beyond the margin. The zero-loss statement is idealized; Section 5.1 verifies the finite-training behaviour empirically.

3.4 A MULTI-MODAL DUAL-ENCODER ARCHITECTURE

In simple settings (e.g., section 5.1) data and rules can share one encoder pair (f_c, f_t). In real systems the modalities differ fundamentally: raw data are high-dimensional, continuous, and noisy (images, sensor states), while rules are low-dimensional, discrete, and abstract (e.g., `IF red`

light THEN stop). We therefore use four encoders projecting into a *shared latent semantic space* (Figure 2): *data encoders* $(f_{c,\text{data}}, f_{t,\text{data}})$, e.g., CNNs or ViTs, map observations (x, y) to $(z_{c,\text{data}}, z_{t,\text{data}})$; *rule encoders* $(f_{c,\text{rule}}, f_{t,\text{rule}})$, e.g., text transformers or GNNs, map (A, C) to $(z_{c,\text{rule}}, z_{t,\text{rule}})$; and a *shared universal predictor* g operates on either modality. We apply g to each modality in separate forward passes with shared weights, rather than to a concatenation, which forces it to be *modality-agnostic*: a patient embedding reflecting high blood pressure and the embedding of the symbol trestbps_level=High must be transformed by the same function. If rule inputs are absent, the architecture reduces to a standard JEPA.

Inference by distance to symbolic poles. Because both modalities share one latent space, a data context can be compared directly with symbolic consequents. We construct fixed *poles* by encoding explicit outcomes with the rule target encoder (e.g., $z_{rh} = f_{t,\text{rule}}(\texttt{target_risk=1})$) and classify an unseen input by the nearest pole to $g(f_{c,\text{data}}(x_{\text{test}}))$. This answers queries such as "does the observed state satisfy logical condition C?" without test-time rule mining and without training a downstream classifier; each prediction comes with a geometric justification, its distances to the poles.

4 JEPA FOR RULES: CONTINUOUS RULE MANIFOLD LEARNING

The relationship between rules and representations is bidirectional. Classical rule discovery relies on non-differentiable, combinatorial search, e.g., Apriori (Agrawal & Srikant, 1994) or tree induction such as CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993). Once trained with EBC, the discrete rule space is relaxed into a smooth energy $E(z_c, z_t) = \|g(z_c) - z_t\|_2^2$, and the *rule manifold* is its sublevel set $\mathcal{M}_{\text{rule}} = \{(z_c, z_t) : E(z_c, z_t) \leq \epsilon\}$. Rules are no longer isolated symbols but points on a differentiable manifold, which we exploit for generative discovery.

Paradigm 1: joint energy-based diffusion. We interpret the learned energy as a Boltzmann density over the joint rule state $z = [z_c; z_t]$,

$$p(z_c, z_t) = \tfrac{1}{Z} \exp\big(-E(z_c, z_t)/T\big), \tag{4}$$

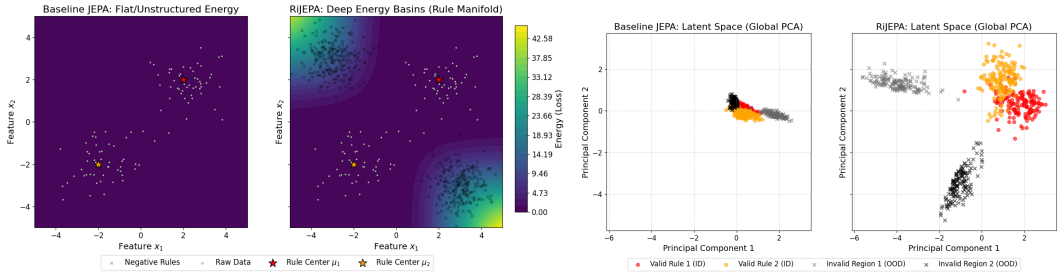
and sample low-energy (valid) configurations with Langevin dynamics,

$$z^{(k+1)} = z^{(k)} - \eta \nabla_z E(z^{(k)}) + \sqrt{2\eta T}\, \epsilon^{(k)}, \qquad \epsilon^{(k)} \sim \mathcal{N}(0, I). \tag{5}$$

The drift term moves latents towards the basins defined by g, and the diffusion term lets the chain explore and escape spurious minima; the temperature T controls conservativeness versus novelty. Restricting the gradient to subsets of z yields three modes: (i) *unconditional discovery* updates both z_c and z_t to generate new valid rules from scratch; (ii) *forward inference* fixes z_c and updates only z_t ("given this condition, what is the valid outcome?"); and (iii) *abductive reasoning* fixes z_t and updates only z_c ("what conditions would lead to this outcome?"). Since every update is driven by g, all generated rules are anchored in the structural logic learned during training. Because E is learned from training rules, samples are best viewed as interpolations consistent with the training rule distribution.

Paradigm 2: marginal-predictive guided discovery. Modeling the joint density can be costly and sensitive to landscape artifacts in high dimensions. Exploiting the asymmetric structure of JEPA, we instead (1) sample novel antecedents from a context marginal $p(z_c) \propto \exp(-E_{\text{ctx}}(z_c))$ with Langevin dynamics; (2) compute the implied consequent deterministically, $\hat{z}_t = g(z_c)$, relying on the EBC-regularized predictor; and (3) decode $(z_c, \hat{z}_t)$ into symbols $(\hat{A}, \hat{C})$ and accept it with a Metropolis–Hastings step $\alpha = \min\{1, \pi(\hat{A}, \hat{C})/\pi(A_{\text{old}}, C_{\text{old}})\}$, where the validity score π comes from a knowledge base, a simulator, or human feedback. This restricts diffusion to the context space and preserves the learned predictive mapping.

Interpretability requirements. Mapping generated geometry back to rules requires latents that represent distinct semantic entities, not arbitrary MLP features over flat text. We define a structured rule tuple $\mathcal{T}_{\text{rule}} = \langle X, \Phi, \Theta, Y, \mu, \Omega, \Sigma \rangle$ (input features, predicates, thresholds, outcomes, fuzzy memberships, logical operators, and rule-strength statistics) and a latent-to-symbolic decoder $\hat{\mathcal{T}}_{\text{rule}} = D(z)$. For example, IF age > 50 AND cholesterol > 200 → heart_risk = high (confidence 85%) gives $X = [\texttt{age}, \texttt{cholesterol}]$, $\Phi = [>, >]$, $\Theta = [50, 200]$, $\Omega = [\texttt{AND}]$, $Y = [\texttt{heart_risk = high}]$, and $\Sigma = [0.85]$. Decoding into this strict structure



(a) Energy landscapes (b) Shared PCA of context latents z_c

Figure 3: Simulation results. (a) The classic JEPA learns a flat, unstructured energy over the rule space, whereas RiJEPA carves deep basins at the valid rule centres and raises steep peaks over the negative (OOD) regions. (b) A single PCA fitted jointly to both models' embeddings: valid rules (circles) and invalid OOD regions (crosses) are entangled in the classic JEPA but clearly separated in RiJEPA.

Table 2: Mean energy assigned to the true rule mapping in the simulation. Lower is better for ID (valid) inputs; higher is better for OOD (invalid) inputs.

Model	**ID energy (valid)** $\downarrow$	**OOD energy (invalid)** $\uparrow$
Classic JEPA	**0.003**	0.150
RiJEPA (EBC)	0.006	**12.713**

guards against low-energy "rules" that are entangled, meaningless mixtures, and it enables *rule interpolation* (e.g., moving a threshold from `age > 50` to `age > 55`) and *compositionality* via vector arithmetic (Appendix C).

5 Experiments

We evaluate (i) whether EBC reshapes the latent geometry in a controlled simulation with known ground truth, and (ii) whether RiJEPA aligns real clinical data with symbolic knowledge and supports generative rule reasoning. In both settings the baseline is a *classic JEPA* obtained by removing the rule pathway from RiJEPA: it uses the same data encoders, predictor capacity, training batches, masking, data loss $\mathcal{L}_{\text{JEPA}}$, EMA momentum, and seed, and differs only in receiving no gradients from $\mathcal{L}_{\text{EBC}}$ (or $\mathcal{L}_{\text{anchor}}$). Any difference in latent structure or accuracy is therefore attributable to the rule constraints rather than to network capacity (Appendix E.3).

5.1 Simulation: a 3D energy landscape

Setup. We sample ground-truth rule antecedents in $\mathbb{R}^3$ from two Gaussians centred at $\mu_1 = [2, 2, 0]$ and $\mu_2 = [-2, -2, 0]$, apply the rule transformation $C = A + [1, 1, 1]$, and generate raw observations (x_i, y_i) coherent with these rules plus observation noise. Invalid (negative) rules have antecedents drawn from two OOD Gaussians centred at $\mu_{\text{neg1}} = [-3, 3, 0]$ and $\mu_{\text{neg2}} = [3, -3, 0]$, with randomly sampled consequents. Since data and rules share dimensionality here, a single shared encoder pair serves both. The classic JEPA minimizes $\mathcal{L}_{\text{JEPA}} = \mathbb{E}_{(x,y)} \|g(f_c(x)) - f_t(y)\|_2^2$; RiJEPA minimizes Eq. (2) with margin $m = 5$ and $\lambda = 1$. We evaluate the energy of the true rule mapping on an in-distribution (ID) test set near $\mu_{1,2}$ and an OOD test set near $\mu_{\text{neg1,2}}$. Full details are in Appendix D.

Results. Figure 3(a) shows that, driven solely by data correlations, the classic JEPA learns a nearly flat energy over the rule space: it implicitly assumes the mapping applies globally and establishes no boundary of logical validity. RiJEPA instead carves deep basins centred on μ_1 and μ_2 while pushing the negative clusters onto high-energy peaks. Quantitatively (Table 2), the classic JEPA assigns a low energy of 0.150 to OOD inputs, i.e., it over-extrapolates the rule into invalid regions, whereas RiJEPA assigns 12.713, an $85\times$ increase, while keeping ID energy near zero (0.006 vs. 0.003). The shared PCA (Figure 3(b)) explains the mechanism: in the classic JEPA, valid and invalid regions are

Table 3: Examples of the 4,801 clinical rules mined with FP-Growth, covering protective (low-risk) and severe (high-risk) profiles.

Risk profile	Antecedent → consequent	Support	Confidence
Low risk	`age_group=Middle` AND `thalach_level=High` → `target_risk=0.0`	21.5%	70.8%
Low risk	`cp=1.0` AND `restecg=2.0` → `target_risk=0.0`	4.2%	83.3%
High risk	`cp=4.0` AND `slope=3.0` → `target_risk=1.0`	4.2%	100.0%

Table 4: Quantitative evaluation on the UCI Heart Disease test set.

Metric	Classic JEPA	RiJEPA	Interpretation
Downstream linear probing	100%	100%	Both admit ex-post linear separation.
Zero-shot logical accuracy	N/A	100%	Only RiJEPA forms logical clusters that allow diagnosis by pole distance.
Fallback latent norm	N/A	5.87	Stable (finite, non-collapsed) output without symbolic input.

entangled in z_c, so the predictor cannot separate them, whereas EBC clusters valid rules and isolates invalid regions. Rule supervision thus does not merely adjust the output mapping; it reorganizes the internal representation to reflect logical structure.

5.2 Real-world case study: clinical rule alignment and discovery

Data and rules. We use the processed Cleveland subset of the UCI Heart Disease dataset (Detrano et al., 1989): 303 patients referred for coronary angiography, described by 13 clinical and non-invasive test variables (e.g., age, sex, chest-pain type, resting blood pressure, cholesterol, maximum heart rate, exercise-induced angina, ST-segment slope, number of calcified vessels) and a binary diagnosis `target_risk` (Appendix E.1). Continuous variables in the training split are discretized into clinically meaningful bins, and FP-Growth (Han et al., 2000) with minimum support 0.04 and minimum confidence 0.70 is used to mine rules. Keeping only rules whose consequent is `target_risk=1.0` (high risk) or `target_risk=0.0` (low risk) yields a logical prior of 4,801 rules spanning protective and severe diagnostic profiles (Table 3).

Model and objective. We instantiate the dual-encoder RiJEPA (Figure 2) with MLP encoders and a shared MLP predictor $g : \mathbb{R}^{32} \to \mathbb{R}^{32}$ (Appendix E.2). In each step g is applied independently to data contexts, valid rule antecedents, and corrupted rule antecedents. In addition to $\mathcal{L}_{\text{JEPA}}$ and $\mathcal{L}_{\text{EBC}}$, an *anchor loss* aligns each training patient's predicted state $\hat{z}_{t,\text{data}} = g(f_{c,\text{data}}(x_c))$ with the symbolic pole of their diagnosis $y \in \{0, 1\}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{JEPA}} + \beta\mathcal{L}_{\text{EBC}} + \gamma\mathcal{L}_{\text{anchor}}, \tag{6}$$

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{x,y}\big[y \,\|\hat{z}_{t,\text{data}} - z_{rh}\|_2^2 + (1-y)\,\|\hat{z}_{t,\text{data}} - z_{rl}\|_2^2\big], \tag{7}$$

with $\beta = 2$, $\gamma = 5$, and EBC margin $m = 2$; z_{rh}, z_{rl} are the high- and low-risk poles produced by $f_{t,\text{rule}}$. The rule encoders are trained by backpropagation, while the data target encoder is an EMA of the context encoder ($\tau = 0.99$) (Assran et al., 2023). At test time rule mining is switched off and only $f_{c,\text{data}}$ and g are used: $x_{\text{test}} \to f_{c,\text{data}} \to g \to$ nearest pole.

Evaluation. On the held-out test set we report: (i) *downstream linear probing*, a logistic regression on frozen embeddings $f_{c,\text{data}}(x)$; (ii) *zero-shot logical accuracy*, classification by the nearest symbolic pole with no task-specific classifier or fine-tuning ("zero-shot" refers to the absence of downstream training; the poles are aligned with training diagnoses through $\mathcal{L}_{\text{anchor}}$); and (iii) a *fallback latent norm* ablation, the L_2 norm of the prediction for an empty (all-zero) rule antecedent, which probes numerical stability when the rule modality is missing.

Manifold alignment. Table 4 and Figure 4 reveal a key distinction between *linear separability* and *semantic alignment*. Both models project patients into a 32-dimensional space that a linear probe separates perfectly, but the classic JEPA needs this ex-post supervised classifier to make diagnostic sense of its features: its manifold (Figure 4, left) shows a smooth, unclustered progression from low

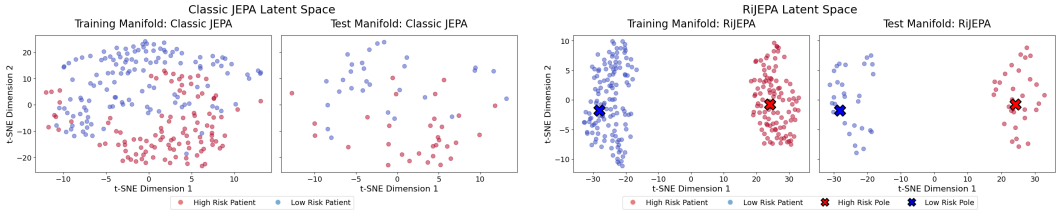


Figure 4: t-SNE of the latent spaces of the classic JEPA (left) and RiJEPA (right). The classic JEPA yields a continuous manifold on which risk profiles form a smooth, entangled gradient; RiJEPA's constraints pull patient embeddings (circles) into separable clusters anchored at the symbolic poles (crosses).

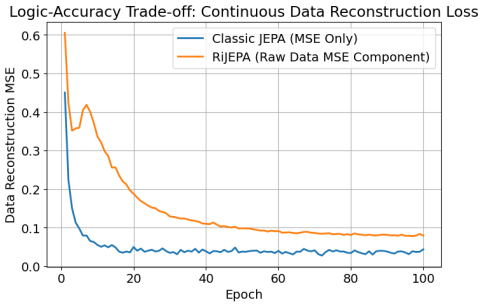


Figure 5: The logic–accuracy trade-off. The classic JEPA optimizes only data correlation and reaches a lower data-prediction MSE ($\mathcal{L}_{\text{JEPA}}$); RiJEPA absorbs a slightly higher data error in order to satisfy the injected rule constraints.

to high risk. RiJEPA's manifold is organized around the diagnostic logic during pretraining: patient embeddings form discrete clusters around their symbolic poles, which allows 100% accuracy by pole distance alone. The fallback norm of 5.87 shows that the shared predictor handles a missing rule modality gracefully, without collapsing, diverging, or producing `NaN`s.

Logic–accuracy trade-off. Tracking only the data-prediction component $\mathcal{L}_{\text{JEPA}}$ during training shows that the classic JEPA reaches a lower MSE than RiJEPA (Figure 5). RiJEPA accepts a slightly higher raw prediction error in exchange for satisfying the symbolic constraints. We interpret this as a *logical regularization* effect: statistical shortcuts that reconstruct the data efficiently but conflict with the injected rules are penalized.

Generative discovery and abductive reasoning. A classic JEPA cannot perform these tasks: without rule encoders its latent coordinates cannot be translated into discrete clinical concepts. RiJEPA defines a shared, symbolically grounded energy surface, and we navigate it with the Langevin paradigms of section 4 (Table 5). *Joint discovery* synthesizes a coherent high-risk profile (LV hypertrophy, downsloping ST); it also produces mixed age markers (young and senior), reflecting the interpolative nature of unconstrained diffusion near basin boundaries. *Forward inference* from "senior age" projects high risk with comorbid markers such as exercise angina and LV hypertrophy. *Abductive reasoning* from "high risk" recovers established clinical risk markers: exercise-induced angina (`exang=1`), low/medium maximum heart rate, downsloping ST (`slope=3`), and senior age. Finally, *marginal-predictive* translation, i.e., the feed-forward step of Paradigm 2 applied to a real test patient's context, maps a young patient with a high maximum heart rate to a protective, low-risk profile. Together these results show that RiJEPA is not merely a discriminative classifier but a bidirectional neuro-symbolic engine operating in a single semantic space.

6 Discussion and Limitations

Rules as multi-level regularization. The effect of rule injection depends on where the rules come from. When rules are mined from the *same* data, as in our clinical study, RiJEPA acts as a multi-

Table 5: Generative clinical discovery with Langevin dynamics on the RiJEPA energy ($T = 0.01$, $\eta = 0.1$, 100 steps). Latents are decoded to their nearest symbolic neighbours.

Paradigm	Query	Discovered clinical profile
Joint discovery	Unconditional	IF normal chol + elevated BP + LV hypertrophy + downsloping ST + typical angina ⇒ high risk + downsloping ST + high FBS + (mixed age markers)
Forward inference	IF senior THEN …	High risk + medium heart rate + exercise angina + asymptomatic pain + LV hypertrophy
Abductive reasoning	High risk BECAUSE …	Exercise angina + low/medium heart rate + downsloping ST + senior age
Marginal-predictive	Real patient vitals (young, high max HR)	Low risk + normal ECG + upsloping ST + high heart rate + no angina

level learner that integrates fine-grained continuous detail with macro-level logical abstractions and prevents the manifold from settling into entangled, spurious correlations. When rules come from an *external* source (clinical guidelines, expert priors), the same mechanism injects knowledge and forces representations to be coherent with domain expertise. More broadly, the joint energy view of Eq. (4) could equip any multi-modal JEPA with an explicit cross-modal energy surface, allowing Langevin-based generation across continuous modalities without autoregressive decoders.

Differentiable logic. Relaxing discrete rules into structured continuous embeddings removes the non-differentiable bottleneck that has confined rule discovery to combinatorial search: gradients $\nabla_z E(z)$ flow through logical constructs, so rule space can be navigated by gradient descent, thresholds interpolated, and abductive root causes inferred. A classic JEPA lacks both the symbolic vocabulary and the energy boundary of Proposition 1, and therefore cannot perform pole-based logical inference or generative rule discovery.

Limitations. Our experiments are deliberately small-scale: a low-dimensional simulation and a single tabular clinical dataset with 303 patients, compared against a classic JEPA. We assume rules are available and reasonably reliable; noisy or conflicting rules may distort the landscape. Our decoder $D(z)$ is currently a nearest-neighbour lookup over the rule vocabulary, and the cost and mixing of joint Langevin diffusion in high-dimensional latent spaces remain to be characterized. Also, the anchor loss uses fixed absolute poles; a natural extension is contrastive triplet training (Schroff et al., 2015; Weinberger & Saul, 2009) over (antecedent, valid consequent, corrupted consequent) triplets, which may better capture hierarchical rule structure (Appendix F).

7 Conclusion

We introduced a bidirectional framework that integrates discrete symbolic rules with continuous joint-embedding predictive architectures. In the *rules for JEPA* direction, RbJEPA distils pure symbolic knowledge and RiJEPA injects rules into raw-data training through energy-based constraints and a multi-modal dual-encoder with a shared predictor, turning rule validity into a geometric energy landscape. In the *JEPA for rules* direction, this landscape enables continuous rule discovery, including unconditional generation, forward inference, abductive reasoning, and marginal-predictive translation, via gradient-guided Langevin dynamics rather than combinatorial search. Simulations show that EBC carves well-separated basins and strongly penalizes OOD invalid rules, and a clinical case study shows that RiJEPA intrinsically aligns patient data with symbolic diagnostic poles, classifying by geometric proximity without a downstream classifier and offering a transparent, distance-based explanation for each decision. This bridge between neural pattern recognition and symbolic logic offers a foundation for interpretable neuro-symbolic representation learning.

AI use statement

In this work, we used AI tools to assist with reformatting the manuscript into the conference template, condensing the text to meet the page limit, and language editing. All AI-assisted work were carefully reviewed and the authors take responsibility for the final content of this work.

Ethics statement

This work uses the publicly available, de-identified UCI Heart Disease (Cleveland) dataset (Detrano et al., 1989); no new human-subject data were collected. The clinical experiment is a proof of concept for neuro-symbolic representation learning and is not a validated diagnostic tool; rules mined from a small, single-centre cohort may encode population-specific biases, and any clinical deployment would require prospective validation and expert oversight. By making predictions explainable through distances to explicit symbolic poles, the proposed approach aims to support rather than replace transparent, auditable clinical reasoning.

Reproducibility statement

The complete data-generation process, network architectures, hyperparameters, random seeds, and computing infrastructure for the simulation are given in Appendix D. For the clinical study, the dataset source and variable encoding (Appendix E.1), the rule-mining settings (section 5.2), all encoder and predictor specifications, training hyperparameters, and Langevin sampling settings (Appendix E.2), and the construction of the controlled baseline (Appendix E.3) are fully specified. All losses are defined in Equations 2, 3, and 6.

References

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pp. 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558601538.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pp. 207–216, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915925. doi: 10.1145/170035.170072. URL https://doi.org/10.1145/170035.170072.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artif. Intell.*, 303(C), February 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103649. URL https://doi.org/10.1016/j.artint.2021.103649.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation, 2017. URL https://arxiv.org/abs/1711.03902.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1 edition, 1984. doi: 10.1201/9781315139470.

Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. Association for Computing

Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Shulin Cao, Zijun Yao, Lei Hou, and Juanzi Li. A general neural-symbolic architecture for knowledge-intensive complex question answering. *Neurosymbolic Artificial Intelligence*, 1(0): 1–4, 2021.

Kewei Cheng, Nesreen K. Ahmed, Ryan A. Rossi, Theodore Willke, and Yizhou Sun. Neural-symbolic methods for knowledge graph reasoning: A survey. *ACM Trans. Knowl. Discov. Data*, 18(9), November 2024. ISSN 1556-4681. doi: 10.1145/3686806. URL https://doi.org/10.1145/3686806.

Tirthankar Dash, Shubham Chitlangia, Anshuman Ahuja, and et al. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12: 1040, 2022. doi: 10.1038/s41598-021-04590-0. URL https://doi.org/10.1038/s41598-021-04590-0.

Artur d'Avila Garcez, Tarek R. Besold, Luc de Raedt, Peter Földiák, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C. Lamb, Risto Miikkulainen, and Daniel L. Silver. Neural-symbolic learning and reasoning: Contributions and challenges. In *Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*. AAAI Press, 2015. Papers from the 2015 AAAI Spring Symposium.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models, 2025. URL https://arxiv.org/abs/2601.00844.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. ISSN 0002-9149. doi: https://doi.org/10.1016/0002-9149(89)90524-9. URL https://www.sciencedirect.com/science/article/pii/0002914989905249.

Katlyn Edwards, Corrie Scalisi, Julianne DeMars-Smith, and Key Lee. Google colab for teaching cs and ml. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024, pp. 1925, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704246. doi: 10.1145/3626253.3635432. URL https://doi.org/10.1145/3626253.3635432.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *J. Artif. Int. Res.*, 61(1):1–64, January 2018. ISSN 1076-9757.

Kazuma Fujioka and Kimiaki Shirahama. Generic itemset mining based on reinforcement learning. *IEEE Access*, 10:5824–5841, 2022. ISSN 2169-3536. doi: 10.1109/access.2022.3141806. URL http://dx.doi.org/10.1109/ACCESS.2022.3141806.

Luis Ignacio Lopera González, Adrian Derungs, and Oliver Amft. A bayesian approach to rule mining, 2020. URL https://arxiv.org/abs/1912.06432.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vision*, 129(6):1789–1819, June 2021. ISSN 0920-5691. doi: 10.1007/s11263-021-01453-z. URL https://doi.org/10.1007/s11263-021-01453-z.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

Allan Grosvenor, Abdul Wahab, Kyrylo Bohachov, Anton Zemlyansky, Ryland Adams, and Dwyer Deighan. Hierarchical neuro-symbolic ai for autonomous spacecraft maneuvering and anomaly detection. In *The Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, pp. 94, sep 2025.

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pp. 1–12, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335372. URL https://doi.org/10.1145/342009.335372.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval, 2018. URL https://arxiv.org/abs/1803.06189.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Geoffrey I. Hinton. *Connectionist learning procedures*, pp. 555–610. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 0934613095.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog (eds.), *Similarity-Based Pattern Recognition*, pp. 84–92, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24261-3.

Yongchao Huang. Probabilistic and reinforced mining of association rules, https://arxiv.org/abs/2506.18155, 2025. URL https://arxiv.org/abs/2506.18155.

Yongchao Huang. Bijepa: Bi-directional joint embedding predictive architecture for symmetric representation learning, https://doi.org/10.5281/zenodo.18659573, February 2026a. URL https://doi.org/10.5281/zenodo.18659573.

Yongchao Huang. Vjepa: Variational joint embedding predictive architectures as probabilistic world models, https://arxiv.org/abs/2601.14354, 2026b. URL https://arxiv.org/abs/2601.14354.

Prateek Jain, Brian Kulis, Inderjit Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/aa68c75c4a77c87f97fb686b2f068676-Paper.pdf.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yann LeCun. A path towards autonomous machine intelligence. Preprint, jun 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Jiashi Lin, Lifang Wang, Xinyu Lu, Zhongtian Hu, Wei Zhang, and Wenxuan Lu. Improving knowledge graph completion with structure-aware supervised contrastive learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13948–13959, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.772. URL https://aclanthology.org/2024.emnlp-main.772/.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJgMlhRctm.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. doi: 10.1038/nature14236.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016. URL https://arxiv.org/abs/1602.01783.

Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraj. Deep metric learning for computer vision: A brief overview, 2023. URL https://arxiv.org/abs/2312.10046.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Jian Pei. *Association Rules*, pp. 140–142. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_26. URL https://doi.org/10.1007/978-0-387-39940-9_26.

J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA, 1993. ISBN 978-1-55860-240-3.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge, 2016. URL https://arxiv.org/abs/1606.04422.

Luciano Serafini, Ivan Donadello, and Artur d'Avila Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, SAC '17, pp. 125–130, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450344869. doi: 10.1145/3019612.3019642. URL https://doi.org/10.1145/3019612.3019642.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models?, 2026. URL https://arxiv.org/abs/2512.24497.

David Tian, Ann Gledson, Athos Antoniades, Aristo Aristodimou, Ntalaperas Dimitrios, Ratnesh Sahay, Jianxin Pan, Stavros Stivaros, Goran Nenadic, Xiao-jun Zeng, and John Keane. A bayesian association rule mining algorithm. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3258–3264, 2013. doi: 10.1109/SMC.2013.555.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rklz9iAcKQ.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning, 2020. URL https://arxiv.org/abs/1904.06627.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009. URL http://jmlr.org/papers/v10/weinberger09a.html.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965. doi: 10.1016/S0019-9958(65)90241-X.

M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, KDD'97, pp. 283–286. AAAI Press, 1997.

Chengqi Zhang and Shichao Zhang (eds.). *Association Rule Mining: Models and Algorithms*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 1 edition, 2002. ISBN 978-3-540-43533-4. doi: 10.1007/3-540-46027-6.

Jinglei Zhang, Tiancheng Lin, Yi Xu, Kai Chen, and Rui Zhang. Relational contrastive learning for scene text recognition, 2023. URL https://arxiv.org/abs/2308.00508.

Zijian Zheng, Ron Kohavi, and Llew Mason. Real world performance of association rule algorithms. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 401–406, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 158113391X. doi: 10.1145/502512.502572. URL https://doi.org/10.1145/502512.502572.

A Extended Related Work

JEPA family. Since the conceptual introduction of hierarchical JEPA (H-JEPA) (LeCun, 2022), many variants have proven effective. I-JEPA (Assran et al., 2023) extracts semantic visual features without data augmentation, using an asymmetric EMA target encoder (Mnih et al., 2015; 2016; Lillicrap et al., 2019; Grill et al., 2020) to prevent representation collapse. V-JEPA (Bardes et al., 2024) extends this to the temporal domain, learning motion dynamics through causal prediction. JEPA world models (Terver et al., 2026) and value-guided JEPA (Destrade et al., 2025) formalize latent-space planning, optimizing action sequences directly in embedding space without decoding to observations.

Neuro-symbolic integration. Early neuro-symbolic systems focused on extracting knowledge from neural networks or injecting propositional logic via specialized losses (d'Avila Garcez et al., 2015; Dash et al., 2022). Within structured relational data, neural-symbolic integration has substantially advanced knowledge-graph reasoning (Cheng et al., 2024), with methods ranging from logical rules as structural regularizers of graph embeddings to fully differentiable rule learners that execute logical queries end to end. Our framework shares the goal of differentiable logic but moves it from supervised or discriminative models to self-supervised joint-embedding spaces.

Metric learning and latent-space shaping. Metric learning learns task-specific distance functions from (weakly) supervised data instead of fixed Euclidean or cosine metrics (Mohan et al., 2023). Proxy-anchor and centroid-based losses (Kim et al., 2020) pull embeddings towards fixed class coordinates, establishing absolute semantic locations; triplet losses, popularized by LMNN (Weinberger & Saul, 2009) and FaceNet (Schroff et al., 2015) and later extended (He et al., 2018; Wang et al., 2020), enforce $d(A, P) < d(A, N) - m$ for an anchor A, positive P, and negative N. Our anchor loss follows the absolute (centroid) view, while EBC follows the relative, margin-based view.

Association rule mining. Foundational ARM (Agrawal et al., 1993) uses frequency measures (support, confidence, lift). Apriori performs breadth-first search with candidate generation, Eclat performs depth-first search on vertical layouts, and FP-Growth avoids candidate generation with compressed prefix trees. Beyond threshold sensitivity, ARM struggles with rare but significant associations and produces many redundant rules (Pei, 2009); Zheng et al. (2001) showed that small changes to the minimum support can cause a super-exponential explosion in the number of rules, far beyond practical human use. Bayesian formulations (Tian et al., 2013; González et al., 2020) evaluate belief rather than raw frequency, and RL/bandit approaches (Fujioka & Shirahama, 2022;

Huang, 2025) adapt the search strategy, yet all still operate in discrete combinatorial spaces. Extracting optimal rules is outside our scope; we assume rules are either extractable from data or supplied as prior knowledge.

B RbJEPA Pipeline

RbJEPA is implemented as a three-step pipeline (Figure 6). It pursues three goals: (i) compressing discrete symbolic knowledge into a continuous latent predictive space; (ii) grounding a model in structured, causal-like abstractions rather than noisy input spaces; and (iii) using verified logical rules as a high-level form of self-supervision. The approach intersects with knowledge distillation (Hinton et al., 2015; Buciluundefined et al., 2006; Gou et al., 2021), neuro-symbolic learning (d'Avila Garcez et al., 2015; Besold et al., 2017; Mao et al., 2019; Evans & Grefenstette, 2018), relational contrastive learning (Zhang et al., 2023; Bordes et al., 2013; Veličković et al., 2019), and structure-aware self-supervision (LeCun, 2022; Assran et al., 2023; Lin et al., 2024). Compared with generative or autoregressive rule-learning models, predicting latent embeddings instead of rule strings ignores low-level syntactic stochasticity, and learning a representation space instead of exact symbolic outputs yields a more general foundation for downstream reasoning.

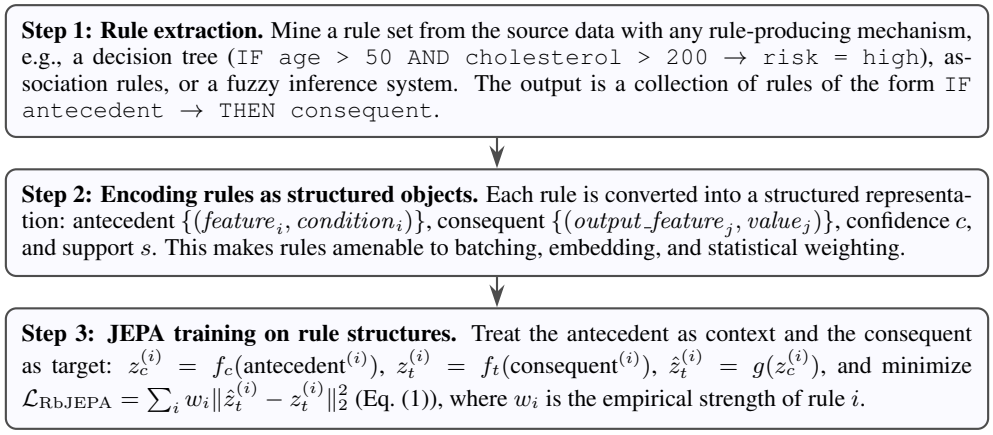


Figure 6: The three-step RbJEPA pipeline: rule extraction, structured encoding, and representation learning by weighted latent prediction.

C Structured Rule Embeddings

The rule tuple $\mathcal{T}_{\text{rule}} = \langle X, \Phi, \Theta, Y, \mu, \Omega, \Sigma \rangle$ has the following components: X, input features (e.g., `age`); Φ, relational predicates ($<, >, \leq$); Θ, numeric thresholds or categorical states; Y, outcomes (e.g., `heart_risk = high`); μ, side information such as fuzzy membership degrees; Ω, logical operators (AND, OR, NOT); and Σ, rule-strength statistics (support, confidence, lift). For instance, `IF age > 50 AND cholesterol > 200 → heart_risk = high` (confidence 85%) gives $X = [\texttt{age}, \texttt{cholesterol}]$, $\Phi = [>, >]$, $\Theta = [50, 200]$, $\Omega = [\texttt{AND}]$, $Y = [\texttt{heart_risk = high}]$, and $\Sigma = [0.85]$. Relaxing discrete logic into an energy landscape risks low-energy "rules" that are entangled, meaningless mixtures; a latent-to-symbolic decoder $D(z)$ that maps back to this strict structure guards against this. Structured embeddings further enable *rule interpolation*, i.e., discovering nearby valid rules never explicitly mined (e.g., adjusting Θ from `age > 50` to `age > 55`), and *symbolic compositionality* via vector arithmetic (e.g., $z_{\text{rule A}} + z_{\text{rule B}} \approx z_{\text{rule C}}$).

D Simulation Details

Data generation. The environment is a 3D feature space ($\mathbb{R}^3$) with the third coordinate fixed to 0 for clear visualization. The generative process is: (1) *latent rule sampling*: ground-truth rule

embeddings are drawn from two Gaussians $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$ with $\mu_1 = [2, 2, 0]$, $\mu_2 = [-2, -2, 0]$, and $\Sigma = 0.5\mathbf{I}$; (2) *rule decoding*: each embedding is mapped to a rule $A \to C$ with the additive transformation $C = A + [1, 1, 1]$; (3) *raw data*: observations $\mathcal{D}_{\text{raw}} = \{(x_i, y_i)\}$ are sampled around the rule centres with covariance 1.5Σ, labelled by applying the rules, with additive noise $\mathcal{N}(0, 0.2^2)$ on the targets; (4) *negative rules*: antecedents are drawn from OOD Gaussians centred at $\mu_{\text{neg1}} = [-3, 3, 0]$ and $\mu_{\text{neg2}} = [3, -3, 0]$, with randomly sampled consequents that contradict the ground truth. Sizes: $N_{\text{data}} = 1000$, $N_{\text{rules}} = 200$, $N_{\text{neg}} = 500$, and $N_{\text{test}} = 300$ for each of the ID and OOD test sets.

Networks. The context encoder f_c, target encoder f_t, and predictor g are MLPs: `Linear(3, 32)` $\to$ GELU $\to$ `Linear(32, 16)`, i.e., input dimension 3, hidden dimension 32, and latent dimension 16. The energy of a pair is $E(x, y) = \|g(f_c(x)) - f_t(y)\|_2^2$.

Training. Both models use full-batch AdamW with learning rate 10^{-3} for 500 epochs; RiJEPA additionally uses EBC with margin $m = 5.0$ and $\lambda = 1.0$. PyTorch (Paszke et al., 2019) and NumPy (Harris et al., 2020) are seeded with 111.

Compute. All simulations ran on Google Colab (Edwards et al., 2024) CPUs (Intel Xeon @ 2.20 GHz, 2 logical cores, 13.6 GB RAM); no GPU is required.

Additional results. Figure 7 visualizes the mean energy per latent region reported in Table 2.

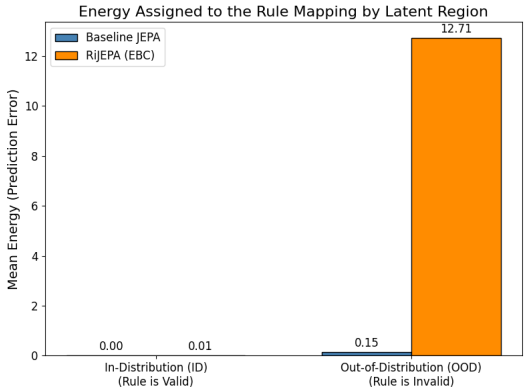


Figure 7: Mean energy assigned by latent region in the simulation. RiJEPA applies a large energy penalty to invalid rules outside the data distribution, while the classic JEPA does not.

E Clinical Case Study Details

E.1 Dataset

The UCI Heart Disease dataset[1] (Detrano et al., 1989) comprises records of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic. Of the 76 attributes in the full database, we use the standard processed subset of 14 variables (Table 6): four clinical variables (age, sex, chest-pain type, resting blood pressure), nine routine or non-invasive test results, and the diagnosis. Examples of the mined rules are given in Table 3 in the main text.

E.2 Architecture and Training

All encoders ($f_{c,\text{data}}, f_{t,\text{data}}, f_{c,\text{rule}}, f_{t,\text{rule}}$) and the shared predictor g use a common latent dimension $d = 32$.

Encoders. Data and rule encoders are MLPs: `Linear(`d_{in}`, 64)` $\to$ GELU $\to$ LayerNorm(64) $\to$ `Linear(64, 32)`, where d_{in} depends on the modality.

[1] `https://archive.ics.uci.edu/dataset/45/heart+disease`; we use the processed Cleveland data.

Table 6: The 14 clinical and non-invasive test variables.

Variable	Type	Description and encoding
`age`	Integer	Patient age in years
`sex`	Categorical	Patient sex (1 = male, 0 = female)
`cp`	Categorical	Chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal, 4: asymptomatic)
`trestbps`	Integer	Resting blood pressure on admission (mm Hg)
`chol`	Integer	Serum cholesterol (mg/dl)
`fbs`	Categorical	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
`restecg`	Categorical	Resting ECG (0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)
`thalach`	Integer	Maximum heart rate achieved during exercise testing
`exang`	Categorical	Exercise-induced angina (1 = yes, 0 = no)
`oldpeak`	Continuous	ST depression induced by exercise relative to rest
`slope`	Categorical	Slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
`ca`	Integer	Number of major vessels (0–3) coloured by fluoroscopy
`thal`	Categorical	Thalassemia (3: normal, 6: fixed defect, 7: reversible defect)
`target_risk`	Categorical	Angiographic disease status (0: $< 50\%$ diameter narrowing, 1: $\geq 50\%$ diameter narrowing)

Shared predictor. g maps $\mathbb{R}^{32} \to \mathbb{R}^{32}$ through an expansion–compression cycle (Figure 8). In each training step it is called three times, with shared weights, on (1) data contexts, giving $g(z_{c,\text{data}}) \approx z_{t,\text{data}}$ (data path); (2) valid rule antecedents, giving $g(z_{c,\text{rule}}) \approx z_{t,\text{rule}}$ (rule path); and (3) corrupted rule antecedents (negatives). The data-path predictions also enter the anchor loss (anchor path). The embeddings are never concatenated, because a predictor trained on concatenated inputs would learn to expect both modalities at all times. The gradients from all passes accumulate in g, so it satisfies the continuous data constraints and the symbolic constraints simultaneously. In the EBC term of Eq. (6), $\lambda = 1$.

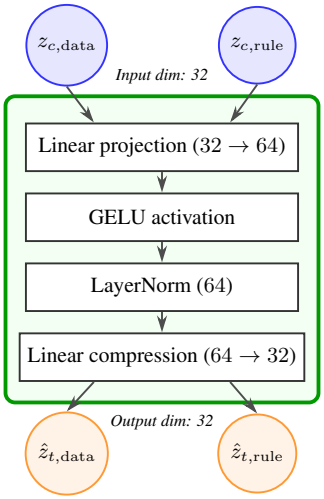


Figure 8: The shared universal predictor g. It accepts either a data context embedding $z_{c,\text{data}}$ or a rule antecedent embedding $z_{c,\text{rule}}$ (in separate passes) and applies the same non-linear transformation to produce the corresponding target prediction.

Target encoder. The rule encoders $f_{c,\text{rule}}$ and $f_{t,\text{rule}}$ are updated by backpropagation. The data target encoder $f_{t,\text{data}}$ receives no gradients; its parameters are an EMA of the context encoder's, $\theta_{t,\text{data}} \leftarrow \tau\theta_{t,\text{data}} + (1-\tau)\theta_{c,\text{data}}$, with $\tau = 0.99$.

Hyperparameters. AdamW with learning rate 10^{-3}; 100 epochs; batch size 32; loss weights $\beta = 2.0$ (EBC) and $\gamma = 5.0$ (anchor); EBC margin $m = 2.0$; gradient-norm clipping at 1.0. Langevin discovery uses temperature (noise scale) 0.01, step size $\eta = 0.1$, and 100 iterations; generated latents are decoded to their nearest symbolic neighbours in the rule vocabulary.

Fallback latent norm. To simulate a missing rule, we feed a zero tensor with the dimension of the rule vocabulary, i.e., an empty antecedent "IF [nothing]", through $f_{c,\text{rule}}$ and g and report the L_2 norm of the resulting prediction z_{null}. A finite, moderate value indicates that the shared predictor handles a masked rule modality without collapsing, diverging, or producing `NaN`s.

E.3 CONTROLLED CLASSIC JEPA BASELINE

The classic JEPA is a clone of the dual-encoder model with the rule-specific components removed. (1) *Identical data pathway:* it uses the same encoder class for $f_{c,\mathrm{data}}$ and $f_{t,\mathrm{data}}$ and the same predictor g (expansion to 64, GELU, LayerNorm, compression to 32), so data-processing capacity is identical and gains cannot be attributed to a larger network. (2) *Identical training data and masking:* both models process the same batches of targets x_t and masked contexts x_c, compute the same data loss $\mathcal{L}_{\mathrm{JEPA}} = \|g(f_{c,\mathrm{data}}(x_c)) - f_{t,\mathrm{data}}(x_t)\|_2^2$, and use the same EMA momentum ($\tau = 0.99$). (3) *Isolated variable:* the classic JEPA receives gradients only from $\mathcal{L}_{\mathrm{JEPA}}$, while RiJEPA receives the same data gradients plus those of $\mathcal{L}_{\mathrm{EBC}}$ and $\mathcal{L}_{\mathrm{anchor}}$. Differences in latent structure or accuracy can therefore be attributed to the rule constraints rather than to architectural scaling.

F FURTHER DISCUSSION

Why embeddings. Real-world knowledge comes in two modalities: raw observations, which are comprehensive and fine-grained but noisy, and abstracted rules (from decision trees, fuzzy inference, or ARM (Huang, 2025)), which are clean and interpretable but lose nuance. Projecting both into shared *embeddings*, rather than modeling logic at the level of pixels, observations, or tokens, filters out low-level stochasticity and lets the network focus on high-level semantic relations and stable structure.

Differentiable logic. Relaxing discrete rules into structured continuous embeddings removes the non-differentiable bottleneck that has confined rule discovery to combinatorial search. Gradients $\nabla_z E(z)$ flow through logical constructs, so the model can navigate rule space by gradient descent, interpolate decision thresholds, infer abductive root causes, and compose rules algebraically. A classic JEPA, lacking a symbolic vocabulary and an energy boundary, cannot perform zero-shot logical inference or generative rule discovery.

Contrastive triplet manifolds. Our anchor loss is an absolute centroid alignment towards fixed poles, and EBC treats positive and negative rule pairs separately. A triplet formulation (Schroff et al., 2015; Weinberger & Saul, 2009) would instead evaluate an antecedent A, a valid consequent P, and a corrupted consequent N jointly, enforcing $d(A, P) + m < d(A, N)$, and would enable *contrastive JEPA rule manifold learning*. It may better capture hierarchical rule relationships and fine-grained compositionality. Future work will also study the computational limits of joint Langevin diffusion in higher-dimensional semantic spaces and more expressive latent-to-symbolic decoders.