MOSAIC: MODE-AWARE PROTOTYPICAL NEUROSYMBOLIC AI FOR CONCEPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

NeuroSymbolic (NeSy) predictors combine the expressive power of neural networks with the safety and interpretability of symbolic methods by reasoning over learned intermediate concepts. Yet state-of-the-art methods remain vulnerable to *reasoning shortcuts* (RSs): incorrect input-to-concept mappings that nevertheless achieve high task accuracy under distant supervision. Many mitigation strategies require dense concept annotations, while recent prototypical approaches avoid both shortcuts and dense supervision by grounding each concept in a single *prototype*. However, they implicitly assume that each concept induces a single-mode distribution in the embedding space. In this paper, we show that the unimodality assumption gives rise to previously overlooked RSs at the latent-mode level, which we term *implicit reasoning shortcuts* (IRSs), in contrast to *explicit reasoning shortcuts* (ERSs) operating at the concept level. We prove that mode-blind predictors are structurally susceptible to IRSs: their optima hurts concept accuracy by conflating distinct modes and this ambiguity persists even with full concept supervision and complete data coverage. To address this limitation, we introduce **M**ode-Aware Pr**O**totypical Neuro**S**ymbolic **AI** for **C**oncept Learning (**MOSAIC**), which discovers latent concept modes from unlabelled support data using infinite mixtures. We derive its embedding updates, characterise its deterministic optima, and establish, through an identifiability analysis, that optimal mode-aware predictors recover the intended concept semantics up to a permutation of latent modes. Empirically, MOSAIC substantially improves concept accuracy over existing reasoning-shortcut mitigation methods on the new `Omniglot-EvenOdd` benchmark and on the real-world, safety-critical `BDDOIA` task.

1 INTRODUCTION

NeuroSymbolic AI (d'Avila Garcez & Lamb, 2023; Raedt et al., 2020) (NeSy) recently emerged as a method to deliver safe, interpretable and annotation efficient models, by combining the adaptability of neural networks with the reliability of symbolic reasoners. Despite their progresses, NeSy models are still plagued by *reasoning shortcuts* (RSs), unintended associations among intermediate neural predicates (or *concepts*), which though satisfy the distant symbolic supervision.

Classical mitigation strategies (Marconato et al., 2023b) often resort to expensive measures that require to densely annotate the intermediate concepts or introduce an additional network reconstructing them. Conversely, prototypical NeSy models (Andolfi & Giunchiglia, 2025) mitigate RSs with minimal supervisions (called the *support set*) by grounding each concept into a learned *prototype*, i.e., an anchor computed as the mean of the embeddings associated to the example datapoints for that concept in the support set. However, these models impose a single-centre representation on each concept. Real-world concepts do not necessarily have such a simple geometry: instances sharing the same symbolic label may form multiple distinct regions of the embedding space, corresponding to different underlying realisations of the con-



Figure 1: *"Obstacle"* examples in `BDDOIA`.

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.

cept. For example, in the BDD-OIA dataset used in our experimental analysis, the concept obstacle may be instantiated by visually distinct entities such as a construction vehicle and a bus (Figure 1).

In this paper, we first study the consequences of mode-blindness in concept predictors by extending the data-generation process of (Marconato et al., 2023b) with latent modes underlying each concept. This reveals two distinct forms of shortcut reasoning: explicit reasoning shortcuts, arising from spurious concept assignments, and implicit reasoning shortcuts, arising from spurious assignments among their latent modes. This analysis motivates **M**ode-Aware Pr**O**totypical Neuro**S**ymbolic **AI** for **C**oncept Learning (**MOSAIC**), which automatically discovers latent modes from the data and allocates a separate prototype to each. We then derive MOSAIC's embedding updates, showing how each embedding is updated relative to a mode-aware belief centre, with individual prototypes attracting or repelling it according to both their current responsibility and the symbolic supervision. Finally, we study concept recovery through the lens of identifiability theory: we characterise the deterministic optima of mode-aware predictors and establish conditions under which the identifiable class of optimal predictors recovers the intended concept semantics up to a permutation of latent modes, while mode-blind predictors remain susceptible to implicit reasoning shortcuts. We extensively validate MOSAIC against existing reasoning-shortcut mitigation strategies, demonstrating substantial improvements in concept accuracy on both variants of our new Omniglot-EvenOdd benchmark and on the real-world, safety-critical BDDOIA task.

2 Problem Statement

Notation. We denote scalar constants in lower-case x, random variables (RVs) X in upper case, their support $\texttt{supp}(X)$, and ordered sets of constants $\mathbf{x}$ and RVs $\mathbf{X}$ in bold typeface. We also will denote the set $[1, \dots n]$ as $[n]$ and the *prior knowledge* as K.

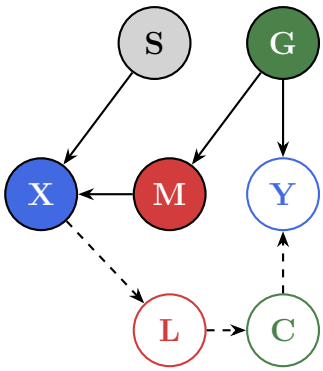


Figure 2: Mode-aware Data Generation Process. Filled nodes and solid arrows denote ground-truth variables and dependencies; empty nodes and dashed arrows denote their predicted counterparts. Colors indicate: explicit, implicit, observed.

Setting. We extend the underlying ground-truth data generation process $p^*(\mathbf{X}, \mathbf{Y}; \mathrm{K})$ from Marconato et al. (2023a;b) by introducing latent modes underlying each explicit concept, as illustrated in Figure 2: (i) $\mathbf{X}$ is the ordered set of RVs with domain $\mathcal{X}$, representing the *input* data *observed* in the learning problem, (ii) $\mathbf{Y}$ is the ordered set of r discrete RVs with domain $\mathcal{Y}$, representing the *observed* ground truth labels for $\mathbf{X}$ to be predicted, (iii) $\mathbf{S}$ is a finite ordered set of RVs influencing the style of the observed data $\mathbf{X}$ but not their ground truth labels, (iv) $\mathbf{G}$ is the ordered set of k *unobserved* discrete RVs ranging in $\mathcal{G} = [h_1] \times \ldots \times [h_k]$, called the **explicit** concepts (EC) and determining their corresponding ground truth label, (v) $\mathbf{M}$ is the ordered set of k *unobserved* discrete RVs ranging in $\mathcal{M} = [m_1] \times \ldots \times [m_k]$, called the **implicit** concepts (IC) which specify the modes for the explicit ones and, together with the style $\mathbf{S}$, generate the observed data $\mathbf{X}$. The ECs influence $\mathbf{X}$ solely through $\mathbf{M}$ (i.e., $\mathbf{X} \perp \mathbf{G} \mid (\mathbf{M}, \mathbf{S})$). We assume the data generation process defines a taxonomy, given by a known deterministic surjection $\tau : \mathcal{M} \to \mathcal{G}$, which maps each vector of implicit concepts $\mathbf{m} = (\mu_1, \dots, \mu_k)$ to its explicit counterpart $\tau(\mathbf{m}) = (\tau_1(\mu_1), \dots, \tau_k(\mu_k)) = \mathbf{g}$, where $\tau_i : [m_i] \to [h_i]$ maps the i-th implicit concept to its corresponding explicit concept. Given $c \in [h_i]$, we call the set $\tau^{-1}(c)$ the *mode-lift* of c.

Let $\beta_{\mathrm{K}} : \mathcal{G} \to \mathcal{Y}$ be a deterministic mapping inducing $p^\star(\mathbf{Y} \mid \mathbf{G}; \mathrm{K}) = \mathbb{1}(\mathbf{Y} = \beta_{\mathrm{K}}(\mathbf{G}))$, then the joint distribution for the mode-aware data generation process can be written as:

$$p^\star(\mathbf{X}, \mathbf{Y}, \mathbf{G}, \mathbf{M}, \mathbf{S} \mid \mathrm{K}) = p^\star(\mathbf{X} \mid \mathbf{M}, \mathbf{S}) \cdot p^\star(\mathbf{S}) \cdot p^\star(\mathbf{G}) \cdot p^\star(\mathbf{M} \mid \mathbf{G}) \cdot \mathbb{1}(\mathbf{Y} = \beta_{\mathrm{K}}(\mathbf{G}))$$

Notice that whenever $|\tau^{-1}(c)| = 1, \forall c \in [h_i]$ and $i \in [k]$, we recover the single mode generation process from Marconato et al. (2023a;b). Indeed, τ becomes a bijection identifying $\mathbf{G}$ from $\mathbf{M}$ and the predicted pathway $\mathbf{X} \to \mathbf{L} \to \mathbf{C}$ collapses to $\mathbf{X} \to \mathbf{C}$ as in the single mode setting.

In our settings, a *NeSy predictor* is a model that infers the labels $\mathbf{Y}$ by reasoning over the knowledge K and a set of k discrete EC $\mathbf{C}$, derived from a set of k discrete latent concepts $\mathbf{L}$ predicted from the input $\mathbf{X}$. We assume K is known, $(\mathbf{g}, \mathbf{y}) \not\models \mathrm{K}$ entails $p^*(\mathbf{y} \mid \mathbf{g}; \mathrm{K}) = 0$ and $\mathbf{C}$, $\mathbf{L}$ take values in $\mathcal{C} = \mathcal{G}$ and $\mathcal{L} = \mathcal{M}$ respectively.

Example 1 *(Mode-aware digit addition) Consider a pair of numbers* $\mathbf{x} = (\boxed{6}, \boxed{\text{III}})$ *featuring two modes determined by their alphabet: Arabic (a) and Roman (r). The corresponding IC vector is* $\mathbf{m} = (6_a, \text{III}_\text{r})$*, while while the EC vector is* $\mathbf{g} = (\tau(6_a), \tau(\text{III}_\text{r})) = (6, 3)$*. The latter determine the final label* $\mathbf{y} = 9$ *under* $\text{K} = (Y = G_1 + G_2)$*.*

Goal. Our goal is to learn the parameters θ of a mode-aware NeSy predictor recovering the intended concepts semantics, up to a permutation of latent modes within each explicit concept, i.e., such that there exists a family of permutations ϱ_i of $[m_i]$ with $\tau_i \circ \varrho_i = \tau_i$ and $p_\theta(\mathbf{L} \mid \mathbf{X}) = p^*(\varrho(\mathbf{M}) \mid \mathbf{X})$, with $\varrho = (\varrho_1, \ldots, \varrho_k)$. By the taxonomy, this also implies $p_\theta(\mathbf{C} \mid \mathbf{X}) = p^*(\mathbf{G} \mid \mathbf{X})$, since $\mathbf{C} = \tau(\mathbf{L})$ and $\mathbf{G} = \tau(\mathbf{M})$ deterministically.

Definition 2.1 (Explicit and Implicit Reasoning Shortcuts) *Consider a data generation process as in Figure 2. Let* p_θ *be a NeSy predictor attaining maximal log-likelihood* ℓ *on a training dataset* $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$*, that is* $\ell(p_\theta, \mathcal{D}, \text{K}) = \max_{\theta' \in \Theta} \ell(p_{\theta'}, \mathcal{D}, \text{K})$*. We say that* p_θ *is affected by an* explicit reasoning shortcut *(ERS) if* $p_\theta(\mathbf{C} \mid \mathbf{X}) \neq p^\star(\mathbf{G} \mid \mathbf{X})$*. We also say that* p_θ *is affected by an* implicit reasoning shortcut *(IRS) if there exists no family of permutations* ϱ_i *of* $[m_i]$ *such that* $\tau_i \circ \varrho_i = \tau_i$ *and* $p_\theta(\mathbf{L} \mid \mathbf{X}) = p^\star(\varrho(\mathbf{M}) \mid \mathbf{X})$*, where* $\varrho = (\varrho_1, \ldots, \varrho_k)$*.*

Observe that by the concept taxonomy every ERS implies an IRS, but the converse does not hold: a model might rely on spurious implicit-concept assignments that nonetheless map to the correct EC under τ. For this reason, modeling $\mathbf{L}$ directly is not an additional source of shortcut reasoning, but rather a useful diagnostic tool: mode-blind predictors expose only $p_\theta(\mathbf{C} \mid \mathbf{X})$, so an IRS they suffer leaves no signature in their outputs. We characterise this phenomenon in Section 4.

Example 2 (Ex. 1 Cont'd.) *Let* $\mathcal{D}$ *be a dataset of number pairs, whose latent concept vectors take values in* $\{(6_a, \text{III}_\text{r}), (6_\text{a}, 3_\text{a}), (2_\text{a}, \text{VIII}_\text{r}), (2_\text{a}, 8_\text{a})\}$*, then:* $\tau^{-1}(3) = \{3_a, \text{III}_\text{r}\}$, $\tau^{-1}(8) = \{8_a, \text{VIII}_\text{r}\}$ *while* $\mathbf{g} = (6, 3)$, $\mathbf{g}' = (2, 8)$ *are the only explicit-concept configurations realised in* $\mathcal{D}$*. A NeSy predictor* p_θ *takes an ERS if for example* $p_\theta((7, 2) \mid \boxed{6}, \boxed{\text{III}}) \approx 1.0$ *and* $p_\theta((5, 5) \mid \boxed{2}, \boxed{\text{VIII}}) \approx 1.0$*. Here, the mapping preserves* $\mathbf{Y}$ *under* $\text{K} = (Y = G_1 + G_2)$*, but the concept are mapped erroneously. A NeSy predictor* $p_{\theta'}$ *takes an IRS if* $p_{\theta'}((6_a, 3_a) \mid \boxed{6}, \boxed{\text{III}}) \approx p_{\theta'}((6_a, 3_a) \mid \boxed{6}, \boxed{3}) \approx 1.0$*: the distribution defines a non-injective modes mapping with no permutation between them: the resulting IRS yields correct ECs and correct sum labels.*

3 MODE-AWARE PROTOTYPICAL NEUROSYMBOLIC AI

Mode-blind Setting. Let $\mathcal{D}$ be a finite dataset. For each $i \in [k]$, (single-prototype) prototypical NeSy predictors (Andolfi & Giunchiglia, 2025) use embedding functions $f_\theta^i : \mathbb{R}^d \to \mathbb{R}^{\eta_i}$ (η_i being the embedding size for the ith concept) with learnable parameters θ, called *prototype extractors*, to form prototypes (or *centroids*) as the mean of labelled *support* datapoints belonging to each EC c:

$$\mathbf{c}_c = \frac{1}{|\hat{\mathcal{S}}_c|} \sum_{\mathbf{x} \in \hat{\mathcal{S}}_c} f_\theta^i(\mathbf{x}), \quad \text{for } i \in [k], \quad \text{given } \hat{\mathcal{S}}_c \subset \mathcal{S}_c = \{\mathbf{x} \mid (\mathbf{x}, \mathbf{v}, \mathbf{y}) \in \mathcal{D}, c \in \mathbf{v}\}, \tag{1}$$

where $\mathbf{v}$ is (a partially specified) vector of EC-level labels for $\mathbf{x}$. Training under K jointly shapes the embeddings around prototypes while satisfying the constraints; inference maps each input (or *query*) to the closest prototype class and predictions are aggregated with the inference layer of choice.

We call a prototypical NeSy predictor p_θ *mode-aware* if, for every $i \in [k]$ and $c \in [h_i]$, its prototypes are indexed by the latent modes $\mu \in \tau_i^{-1}(c)$, associating a distinct prototype μ with each mode. Conversely, p_θ is *mode-blind* if its prototypes are indexed only by the ECs, associating a single prototype c_c with each $c \in [h_i]$, irrespective of how many modes belong to $\tau_i^{-1}(c)$. The prototypical NeSy predictors introduced above are therefore mode-blind.

MOSAIC. To lift mode-blind prototypical predictors to the mode-aware data generation process, MOSAIC employs Infinite Mixtures, following the approach in Allen et al. (2019). Supports are first clustered by the mode-blind initialisation in Eq.(1) and new centroids representing the ICs are dynamically discovered from the data: for each support point $\mathbf{x}$, the minimum distance of $\mathbf{z}_i = f_\theta^i(\mathbf{x})$ to all clusters for the same class c is computed and, whenever it exceeds a threshold λ, a new cluster is generated whose mean is $\mathbf{z}_i$. Thus, MOSAIC requires no explicit annotations for the ICs. Under squared Euclidean distance, the prototype for mode μ can be seen as the mean

$\mathbf{c}_\mu$ of a Gaussian mixture with isotropic covariance $\sigma_i^2\mathbf{I}$; hence, the mode-conditional likelihood is $p_\theta(\mathbf{z}_i \mid L_i = \mu) = \mathcal{N}(\mathbf{z}_i \mid \mathbf{c}_\mu, \sigma_i^2\mathbf{I})$. As in Allen et al. (2019), we compute the threshold as:

$$\lambda_i = -2\sigma_i^2 \log \upsilon + \eta_i \sigma_i^2 \log\Big(1 + \frac{\rho}{\sigma_i^2}\Big) \;=\; 2\sigma_i^2 \log\Big(\frac{(1+\rho/\sigma_i^2)^{\eta_i/2}}{\upsilon}\Big), \tag{2}$$

where υ is the concentration parameter for the Chinese restaurant process (CRP) (Aldous, 1985) and ρ the variance of the base distribution from which clusters are drawn in the CRP. Given the prior π_μ for each centroid, MOSAIC produces a distribution over the ICs based on the softmax over distances between $\mathbf{z}_i$ and mode centroids (sums always range over the modes MOSAIC has discovered):

$$p_\theta(L_i = \mu \in [m_i] \mid \mathbf{x}) = \frac{\pi_\mu \exp(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2)}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \exp(-||\mathbf{z}_i - \mathbf{c}_{\mu'}||_2^2/2\sigma_i^2)} \tag{3}$$

From the per-mode soft assignments in Eq.(3) the ECs distribution is obtained through τ_i as:

$$p_\theta(C_i = c \in [h_i] \mid \mathbf{x}) = \frac{\sum_{\mu \in \tau_i^{-1}(c)} \pi_\mu \exp(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2)}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \exp(-||\mathbf{z}_i - \mathbf{c}_{\mu'}||_2^2/2\sigma_i^2)} \tag{4}$$

Differently from Allen et al. (2019), the concept distribution MOSAIC hands to the inference layer is not the softmax over distances to the *closest* centroid cluster in each class, but rather a soft marginalization over the possible mode-lifts of c wrt τ_i. Given $i \in [k]$ and $c \in [h_i]$, by setting $\pi_{\mu \in \tau^{-1}(c)} = \frac{1}{h_i|\tau^{-1}(c)|}$ and taking the logarithms we can equivalently express Eq.(4) as:

$$\log p_\theta(C_i = c \in [h_i] \mid \mathbf{x}) = \text{logsumexp}_{\mu \in \tau_i^{-1}(c)}(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2) - \log Z_c^i - b \tag{5}$$

where $Z_c^i = |\tau^{-1}(c)|$ and b denotes the constant across c expressing the shared log-denominator. The complete derivation of Eq.(3, 4, 5) is reported in the Appendix A.1. The discriminant analysis in Figure 3 makes tangible the need of multiple prototypes per class in `Omniglot-EvenOdd`: in MOSAIC the datapoints fall close to one (of many) prototypes matching their intended EC label. Conversely, by committing to a single prototype, mode-blind predictors push in the wrong decision region characters dissimilar to the digit's mean. Following (Andolfi & Giunchiglia, 2025), prototype extractors in MOSAIC are trained to (i) correctly classify annotated query points in so-called *episodes* and (ii) make predictions compliant with the constraints for all $\mathbf{x} \in \mathcal{D}$. MOSAIC training is shown in Algorithm 1.

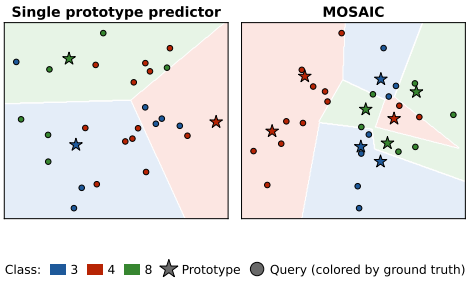


Figure 3: Decision regions in single prototype extractors (left) and MOSAIC (right) over three `Omniglot-EvenOdd` alphabets (ECs).

Theorem 3.1 (MOSAIC Embedding Update) *Let $\mathcal{G} = [h_1]\times\ldots\times[h_k]$ and $\mathcal{M} = [m_1]\times\ldots\times[m_k]$ be the sets of ECs and ICs related through the surjection $\tau : \mathcal{M} \to \mathcal{G}$. Consider a mode-aware prototypical NeSy predictor with output $\hat{\mathbf{y}}$ trained using $\mathcal{L}^{\text{NeSy}}(\hat{\mathbf{y}}, \mathrm{K})$ encoding the satisfaction of knowledge K by $\hat{\mathbf{y}}$. For every $i \in [k]$ the gradient of the loss calculated wrt the embedding $\mathbf{z}_i$ is:*

$$\nabla_{\mathbf{z}_i}\mathcal{L}^{\text{NeSy}}(\hat{\mathbf{y}}, \mathrm{K}) = \frac{1}{\sigma_i^2} \sum_{c \in [h_i]} \underbrace{\frac{\partial \mathcal{L}^{\text{NeSy}}}{\partial \hat{y}_c^{\text{EC}}}}_{\text{symbolic signal}} \sum_{\mu \in \tau_i^{-1}(c)} \underbrace{\hat{y}_\mu}_{\substack{\text{mode}\\ \text{weight}}} \underbrace{(\mathbf{c}_\mu - \bar{\mathbf{c}}_i)}_{\substack{\text{mode-specific}\\ \text{direction}}}, \qquad \bar{\mathbf{c}}_i = \underbrace{\sum_{\mu' \in [m_i]} \hat{y}_{\mu'} \mathbf{c}_{\mu'}}_{\text{mode-aware belief centre}}$$

where $\hat{y}_\mu = p_\theta(L_i = \mu \mid \mathbf{z}_i)$ is the softmax over the distances and $\hat{y}_c^{\text{EC}} = \sum_{\mu \in \tau_i^{-1}(c)} \hat{y}_\mu$.

Proof in Appendix A.2. The theorem shows that, even though the supervision is provided at the ECs level, the resulting gradient induces mode-specific updates in the embedding space. Indeed, the mode-aware belief centre $\bar{\mathbf{c}}_i = \mathbb{E}_{\mu' \sim p_\theta(L_i \mid \mathbf{z}_i)}[\mathbf{c}_{\mu'}]$ is the expected prototype under the predictor's posterior over modes. The EC-level symbolic signal is distributed across its modes according to the model's confidence in each mode, $\hat{y}_\mu$, with each contribution directed by the displacement of the corresponding prototype from the belief centre.

4 IDENTIFIABILITY OF EXPLICIT AND IMPLICIT CONCEPTS

We now study the conditions to learn concepts with the intended semantics under the lens of identifiability theory. Following (Marconato et al., 2023b; 2025), we assume the conditions of *deterministic knowledge* [**A1**] and *extrapolability* [**A2**], specified at the latent modes level. Under [**A1**] the distribution $p^*(\mathbf{Y} \mid \mathbf{G}; \mathbf{K})$ is induced by a deterministic mapping $\beta_{\mathbf{K}} \colon \mathbf{g} \to \mathbf{y}$, while [**A2**] induces the distribution $p^*(\mathbf{X} \mid \mathbf{G}, \mathbf{M}, \mathbf{S}) = p^*(\mathbf{X} \mid \mathbf{M}, \mathbf{S}) = \delta\{\mathbf{X} = \phi_\tau(\mathbf{m}, \mathbf{s})\}$ through a map $\phi_\tau \colon (\mathbf{m}, \mathbf{s}) \mapsto \mathbf{x}$, where ϕ_τ is invertible, and smooth over $\mathbf{s}$. Lastly, we work under *modes clusterability* [**A3**]: for every $i \in [k]$ all datapoints generated from a given mode μ_i share the same nearest prototype, and every prototype is nearest to datapoints from at least one mode. Distinct modes may nevertheless share the same prototype. Equivalently, the map

$$\mathbf{x} \mapsto \arg\min_j \left\| f_\theta^i(\mathbf{x}) - \mathbf{c}_j \right\|_2^2 \tag{6}$$

depends on $\mathbf{x} = \phi_\tau(\mathbf{m}, \mathbf{s})$ only through μ_i. In the deterministic limit $\sigma_i \to 0$, $p_\theta(L_i \mid \mathbf{x})$ concentrates on the index of the nearest prototype. We write $\gamma_i : [m_i] \to [m_i]$ for the resulting mapping, where $\gamma_i(\mu_i) = j$ whenever $\mathbf{c}_j$ is nearest to datapoints generated from mode μ_i. Unlike the EC-level clusterability assumption in Andolfi & Giunchiglia (2025), γ_i is *not* required to be injective.

Proposition 4.1 (Number of Deterministic Optima) *Consider a mode-aware NeSy predictor inducing $p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{K})$. Let $\mathcal{G} = [h_1] \times \ldots \times [h_k]$, $\mathcal{M} = [m_1] \times \ldots \times [m_k]$ and $\mathcal{D}$ be a dataset realising every $\mathbf{m} \in \mathit{supp}(\mathbf{M})$. Under **[A1,2,3]** the number of deterministic optima for $p_\theta(\mathbf{L} \mid \mathbf{X})$ is:*

$$\sum_{\gamma \in \Gamma} \mathbb{1}\Big(\bigwedge_{\mathbf{m} \in \mathit{supp}(\mathbf{M})} \Big[\Big(\bigwedge_{i \in [k]} \left(\mu_i \in \mathcal{S}_i^{\text{sup}} \;\Rightarrow\; \tau_i(\gamma_i(\mu_i)) = \tau_i(\mu_i) \right) \Big) \;\wedge\; \left((\beta_{\mathbf{K}} \circ \tau \circ \gamma)(\mathbf{m}) = (\beta_{\mathbf{K}} \circ \tau)(\mathbf{m}) \right) \Big] \Big),$$

where $\mathcal{S}_i^{\text{sup}}$ denotes the set of modes occurring in at least one EC-annotated datapoint in $\mathcal{D}$ and $\Gamma \subseteq \prod_{i \in [k]} \{\gamma_i : [m_i] \to [m_i]\}$ is the set of coordinate-wise assignments induced by the deterministic distributions $p_\theta(\mathbf{L} \mid \mathbf{X})$ the predictor can express, acting as $\gamma(\mathbf{m}) = (\gamma_1(\mu_1), \ldots, \gamma_k(\mu_k))$.

Proof in Appendix A.3. If $|\tau_i^{-1}(c)| = 1$, for all $i \in [k]$ and $c \in [h_i]$, τ_i is a bijection and hence mode and EC-level clusterability coincide. Setting $\alpha_i = \tau_i \circ \gamma_i \circ \tau_i^{-1}$ turns Proposition 4.1 into the shortcut count of Andolfi & Giunchiglia (2025), with $\mathcal{S}_i^{\text{sup}}$ and Γ resp. in the role of their $\mathcal{H}_i$ and $\mathcal{A}$. A mode-blind predictor exposes only $p_\theta(\mathbf{C} \mid \mathbf{X})$, making IRSs that preserve the EC assignments invisible. Mode-aware predictors can instead distinguish such latent assignments, provided the corresponding modes are represented in the support set. We formalise this requirement through mode coverage.

Definition 4.2 (Mode Covering Set) *Let $\mathcal{S}$ be the subset of datapoints in $\mathcal{D}$ receiving EC-level supervision and used to construct the prototypes. For each $\mathbf{x} \in \mathcal{S}$, let $\mathbf{m}(\mathbf{x}) = \big(\mu_1(\mathbf{x}), \ldots, \mu_k(\mathbf{x})\big)$ be its mode vector, uniquely determined under **[A2]**. The* mode covering set *of $\mathcal{S}$ at $i \in [k]$ is*

$$\textsc{Cov}_i(\mathcal{S}) \;=\; \left\{ \mu_i(\mathbf{x}) \;\middle|\; \mathbf{x} \in \mathcal{S} \right\} \;\subseteq\; [m_i].$$

We call $|\textsc{Cov}(\mathcal{S})| = \big|\bigsqcup_{i \in [k]} \textsc{Cov}_i(\mathcal{S})\big| = \sum_{i \in [k]} |\textsc{Cov}_i(\mathcal{S})|$ the *mode coverage* of $\mathcal{S}$. A support set $\mathcal{S}$ achieves *complete mode coverage* whenever $\textsc{Cov}_i(\mathcal{S}) = [m_i]$ for every $i \in [k]$. Complete mode coverage is necessary for mode-awareness: prototypes are created only from support datapoints (c.f. steps (a), (b) in Algorithm 1), and hence $|\text{im}(\gamma_i)| \leq |\textsc{Cov}_i(\mathcal{S})|$.

Definition 4.3 (Equivalence up to Euclidean Transformations) *Given two prototype extractors f_θ^i and $f_{\theta'}^i$ with embedding dimension η, let $\mathcal{O}(\eta)$ and $\mathcal{E}(\eta)$ be the orthogonal and Euclidean groups on $\mathbb{R}^\eta$, respectively. We say that f_θ^i and $f_{\theta'}^i$ are* equivalent up to a Euclidean transformation, *written as $f_\theta^i \sim_{\mathcal{E}(\eta)} f_{\theta'}^i$, if there exist $U \in \mathcal{O}(\eta)$ and $t \in \mathbb{R}^\eta$ such that $f_{\theta'}^i(\mathbf{x}) = U f_\theta^i(\mathbf{x}) + t, \forall \mathbf{x} \in \mathcal{X}$.*

Given a NeSy prototypical predictor p_θ, let $f_\theta = (f_\theta^1, \ldots, f_\theta^k)$ be its *joint prototype extractor*, whose ith component f_θ^i induces the embedding space associated with the i-th concept coordinate. We say that two joint extractors f_θ and $f_{\theta'}$ are *equivalent up to coordinate-wise Euclidean transformations*, written $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$, if $f_\theta^i \sim_{\mathcal{E}(\eta)} f_{\theta'}^i, \forall i \in [k]$. We characterise each extractor through its *query-to-centroid distance profile* $\Phi(f_\theta^i)$:

$$\Phi(f_\theta^i)(\mathbf{x}, B) = \left\| f_\theta^i(\mathbf{x}) - \mathbf{c}_B(f_\theta^i) \right\|_2^2, \quad \mathbf{c}_B(f) = \tfrac{1}{|B|} \textstyle\sum_{\mathbf{b} \in B} f(\mathbf{b}), \quad \text{for } B \subseteq \mathcal{X} \text{ non-empty and finite.}$$

Table 1: Comparison of MOSAIC and mode-blind prototypical predictors (PNet) on the `Omniglot-EvenOdd` 10-alphabet and 40-alphabet splits — 2 labelled images per character – and on BDD-OIA — 20 positively annotated support datapoints per EC (1.63% of the dataset).

	`Omniglot-EvenOdd` (10 Alphabets)			`Omniglot-EvenOdd` (40 Alphabets)			`BDD-OIA`		
Model	**Acc(C)%↑**	**F1(C)%↑**	**Cls(C)%↓**	**Acc(C)%↑**	**F1(C)%↑**	**Cls(C)%↓**	**Acc(C)%↑**	**F1(C)%↑**	**Cls(C)%↓**
DPL	8.85±3.17	3.09±0.91	0.71±0.03	14.89±2.44	3.62±1.47	0.76±0.07	59.21±0.62	38.90±0.43	0.19±0.03
+PNet	72.00±0.01	68.00±0.01	**0.00±0.00**	35.24±0.75	32.80±0.72	**0.00±0.00**	72.52±1.61	54.81±0.93	**0.00±0.00**
+ MOSAIC	**90.00±0.00**	**87.00±0.01**	**0.00±0.00**	**71.83±0.42**	**70.71±0.44**	**0.00±0.00**	**89.13±0.12**	**57.59±1.29**	**0.00±0.00**
CCN+	11.29±3.41	3.03±1.00	0.76±0.05	9.76±1.92	2.74±1.38	0.77±0.06	52.96±11.68	31.98±6.15	0.34±0.08
+PNet	69.12±0.92	65.12±1.10	**0.00±0.00**	42.89±4.33	40.82±3.70	**0.00±0.00**	73.05±1.01	55.17±0.51	**0.00±0.00**
+MOSAIC	**89.58±0.47**	**87.12±0.45**	**0.00±0.00**	**71.94±0.28**	**70.76±0.37**	**0.00±0.00**	**89.17±0.12**	**57.17±0.89**	**0.00±0.01**
SL	8.10±3.44	3.06±1.65	0.70±0.00	16.04±3.27	6.24±1.55	0.70±0.00			
+PNet	71.00±0.01	67.00±0.01	**0.00±0.00**	45.59±0.51	43.42±0.48	**0.00±0.00**			
+MOSAIC	**90.00±0.00**	**87.00±0.00**	**0.00±0.00**	**71.81±0.37**	**70.60±0.36**	**0.00±0.00**			
Scallop	9.56±4.55	3.63±1.77	0.71±0.03	7.87±0.47	2.94±1.00	0.71±0.05			
+PNet	69.79±1.20	65.89±1.29	**0.00±0.00**	45.90±0.76	43.71±0.92	**0.00±0.00**			
+MOSAIC	**89.44±0.43**	**86.98±0.48**	**0.00±0.00**	**71.72±0.50**	**70.51±0.55**	**0.00±0.00**			
CTSketch	9.74±3.51	3.36±1.19	0.71±0.03	14.84±2.31	3.68±1.38	0.75±0.05			
+PNet	70.17±0.97	66.36±1.08	**0.00±0.00**	45.82±0.49	43.70±0.54	**0.00±0.00**			
+MOSAIC	**89.66±0.54**	**87.22±0.59**	**0.00±0.00**	**71.71±0.30**	**70.51±0.28**	**0.00±0.00**			

Red: exact methods (K always satisfied) for `Omniglot-EvenOdd` and `BDDOIA`

Green: approximate methods (K likely satisfied) for `Omniglot-EvenOdd`

Theorem 4.4 (Identifiability of Prototype Extractors) *For any pair of joint prototype extractors f_θ and $f_{\theta'}$ it holds that:*

$$f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'} \implies \Phi(f_\theta^i)(\mathbf{x}, B) = \Phi(f_{\theta'}^i)(\mathbf{x}, B) \quad \text{for all } i \in [k],\ \mathbf{x} \in \mathcal{X} \text{ and all finite } B \subseteq \mathcal{X}.$$

Conversely, if for every $i \in [k]$ there is a finite $\mathcal{B}_i \subset \mathcal{X}$ and an integer $s_i \in [|\mathcal{B}_i| - 1]$ so that:

- ***(C1)*** *$\Phi(f_\theta^i)(\mathbf{x}, B) = \Phi(f_{\theta'}^i)(\mathbf{x}, B)$ for every $\mathbf{x} \in \mathcal{X}$ and every $B \subseteq \mathcal{B}_i$ with $|B| = s_i$ and*
- ***(C2)*** *$f_\theta^i(\mathcal{X}) \subseteq \mathtt{aff}\big(f_\theta^i(\mathcal{B}_i)\big)$, where $\mathtt{aff}(\cdot)$ is the affine hull.*

Then $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$.

Proof in Appendix A.4. In the converse statement, **(C1)** ranges over all s_i-subsets of $\mathcal{B}_i$, not only those an episode returns; **(C2)** asks that $f_\theta^i(\mathcal{B}_i)$ affinely span $f_\theta^i(\mathcal{X})$: such $\mathcal{B}_i$ always exists with at most $\eta + 1$ points, and without it the two extractors may differ off $\mathtt{aff}(f_\theta^i(\mathcal{B}_i))$ while leaving every distance in **(C1)** unchanged. To lift the geometric equivalence of Theorem 4.4 to conclusions about the distributions induced by the predictors, we compare predictors sharing the same υ, σ_i, and episode stream. Otherwise, two predictors with identical joint extractors could induce different IC distributions. We denote by $[p_\theta]_{\mathcal{E}(\eta)} = \{p_{\theta'} \mid f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}\}$ the class of predictors whose joint extractors are equivalent up to coordinate-wise Euclidean transformations to f_θ.

Lemma 4.5 (Distribution Equivalence over Equivalence Class) *Let p_θ and $p_{\theta'}$ be prototypical NeSy predictors. If $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$, then there is a family of permutations ϱ_i of $[m_i]$ with $\tau_i \circ \varrho_i = \tau_i$ such that $p_\theta(L_i = \mu \mid \mathbf{x}) = p_{\theta'}\big(L_i = \varrho_i(\mu) \mid \mathbf{x}\big)$ for every $i \in [k]$, $\mu \in [m_i]$, $\mathbf{x} \in \mathcal{X}$, and hence $p_\theta(\mathbf{L} = \mathbf{m} \mid \mathbf{X}) = p_{\theta'}(\mathbf{L} = \varrho(\mathbf{m}) \mid \mathbf{X})$ with $\varrho = (\varrho_1, \ldots, \varrho_k)$.*

Proof in Appendix A.5. A consequence of Lemma 4.5 is that two prototypical NeSy predictors also share the same deterministic optima up to mode permutation when their embeddings agree up to an element of the Euclidean group $\mathcal{E}(\eta)$. The proofs for the next results are deferred to Appendix A.6.

Theorem 4.6 (Identifiable class with Intended Concept Semantics) *Given a dataset $\mathcal{D}$ and a support set $\mathcal{S}$ with complete mode coverage, let p_θ be a prototypical NeSy predictor attaining maximal log-likelihood on $\mathcal{D}$. Under **[A1,2,3]**, the predictor p_θ is mode-aware if and only if $[p_\theta]_{\mathcal{E}(\eta)}$ learns concepts with the intended semantics up to mode permutation.*

Corollary 4.7 (Implicit Reasoning Shortcuts in mode-blind models) *In the same settings of Theorem 4.6, assume the taxonomy to be* non-degenerate*, that is $|\tau_i^{-1}(c)| \geq 2$ for at least one $i \in [k]$ and one $c \in [h_i]$. Under **[A1,2,3]**, if p_θ is mode-blind then its set of deterministic optima has*

Table 2: F1(C) comparison across mitigation strategies on Omniglot-EvenOdd (10-alphabet split) and BDD-OIA. C denotes concept supervision, E entropy and PRE pretraining.

	Omniglot-EvenOdd					BDD-OIA	
Mitigation	**DPL**	**SL**	**Scallop**	**CTSketch**	**CCN+**	**DPL**	**CCN+**
+C	5.19±2.30	12.31±2.67	4.35±1.53	5.88±2.96	8.10±3.72	41.10±0.58	35.54±2.57
+E	3.01±0.50	4.54±2.83	4.60±1.49	4.03±2.13	4.59±2.57	47.73±0.23	43.93±0.42
+PRE	8.40±1.57	9.57±2.09	7.96±2.27	6.91±2.37	4.76±3.31	45.57±2.33	50.49±1.64
+C+E	12.35±4.16	31.07±7.54	12.77±4.76	18.44±14.68	8.96±4.69	49.96±0.36	46.54±0.64
+C+PRE	8.56±2.24	22.75±7.02	8.48±1.76	7.82±3.74	10.88±4.92	47.28±2.05	52.62±0.76
+E+PRE	9.70±0.82	7.53±1.61	9.68±0.12	9.53±0.29	3.81±2.76	51.30±0.13	48.59±0.46
+C+E+PRE	48.19±9.15	37.47±9.26	52.92±5.94	54.24±5.66	10.84±4.30	52.41±0.52	50.79±0.68
+MOSAIC	**87.00** ±0.01	**87.00** ±0.00	**86.98** ±0.48	**87.22** ±0.59	**87.12** ±0.45	**57.59** ±1.29	**57.17** ±0.89

cardinality $\prod_{i\in[k]}\prod_{c\in[h_i]}|\tau_i^{-1}(c)|$*, and every one of them is an IRS. In particular, no mode-blind predictor learns the intended concept semantics, whatever the mode coverage of* $\mathcal{S}$*.*

5 Experimental Analysis

Metrics. Coherently with Andolfi & Giunchiglia (2025), we measure accuracy and F1 macro scores wrt. ECs and the task final label — denoted resp. as Acc(C), F1(C), Acc(Y) and F1(Y) — together with the concept collapse Cls(C), i.e., the fraction of ECs that the predictor never assigns.

Tasks. (i) Omniglot-EvenOdd is the mode-aware variant of MNIST-EvenOdd (Bortolotti et al., 2024) we introduce. Each input is a pair of Omniglot (Lake et al., 2015) characters (ICs) labelled with the sum (task label) of the digits associated to their alphabets (ECs). Only pairs were both digits are even or odd are present. In the 10 (resp. 40) alphabets split each alphabet (resp. 4 alphabets) corresponds to the same digit from 0 to 9. (ii) BDD-OIA (Xu et al., 2020) is an autonomous driving task that proposes frames each featuring reasons (ECs with unknown ICs) causing the ego-vehicle to undertake an action (task label).

Models. We consider DeepProbLog (DPL) (Manhaeve et al., 2018), Semantic Loss (SL) (Xu et al., 2018), Scallop (Li et al., 2023), CTSketch (Choi et al., 2025) and coherent-by-construction CCN+ (Giunchiglia et al., 2024). We integrate each of these reasoning layers with the existing mitigation strategies (and combinations thereof), together with mode-blind prototypes and MOSAIC.

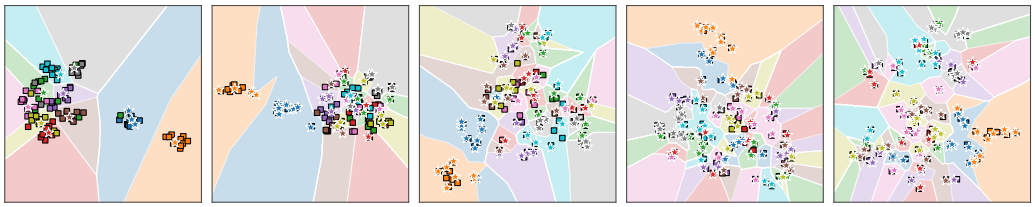
5.1 RQ.1: Can MOSAIC avoid Explicit and Implicit Reasoning Shortcuts?

For Omniglot-EvenOdd, we randomly label with the corresponding alphabet digit **just two** images per character and then use rotation for data augmentation (Allen et al., 2019). We then train both mode-blind and mode-aware NeSy prototypical models with support set size 500 to ensure complete coverage (Definition 4.2). For BDDOIA, we provide 20 *positive* annotations for each of the 21 concepts, then sample the *negative* ones for each EC automatically within the same aggregated pool. No data augmentation is used. We compare our method with the standard NeSy baselines and the single prototype models. The results are shown in Table 1. The *single* prototype constructed by PNet models for each EC is a noisy average of different modalities labeled that concept, compromising their ability to recover the ECs. Conversely, MOSAIC correctly learns the concepts for both tasks, with accuracy of $\approx 90\%$ consistently across the board.

Table 3: Mode purity and IRS-rate comparison for CCN+ MOSAIC and PNet on Omniglot-EvenOdd, 10-alphabets.

Model	**Purity%**(↑)	**IRS-rate**%(↓)
PNet	2.93±0.00	97.14±0.00
MOSAIC	**80.98** ±1.08	**2.94** ±0.51

Table 3 provides the complementary perspective in terms of mode purities and IRS-rate. Here purity measures how internally consistent the nearest-prototype assignment $\gamma_i(x) = \arg\min_j ||f_\theta^i(x) - c_j||^2$ is with true character identity, while the IRS-rate is computed as $1 - |\mathrm{im}(\hat{\gamma}_i)|/m_i$. This tells the fraction of true modes whose empirical image collides onto the same prototype. Clearly, MOSAIC



(a) 10 (b) 50 (c) 100 (d) 200 (e) 500

Figure 4: MOSAIC embedding space at increasing support per EC ($\star$: prototype, $\square$: query).

keeps $\hat{\gamma}_i$ almost injective, whereas PNet's single prototype per concept forces near-total mode collision, which empirically confirms Corollary 4.7. Both the mode-blind and aware prototypical models succeed in spreading their EC-level predictions, with zero concept collapse, differently from the vanilla baselines leveraging few ECs to optimise task accuracy. Full results in Appendix D.1.

5.2 RQ.2: How does MOSAIC compare with other mitigation strategies?

We consider the mitigation strategies, and combinations thereof, proposed in Marconato et al. (2023b); Andolfi & Giunchiglia (2025). (i) *Shannon-entropy regularisation* (E) encourages the models to spread their predictions across the ECs; (ii) *concept supervision* (C) provides auxiliary cross-entropy loss over a subset of densely annotated ECs. As in Marconato et al. (2023b) we label *all* the ECs "4" and "9"; (iii) *pretraining* (PRE) trains the encoder with the labelled datapoints in the support set $\mathcal{S}$ of prototypical models before learning the final task. We compare them with MOSAIC, under the same settings of RQ1, in Table 2. The results show that the high diversity of ICs — combined with restricted supports — causes even the combination of multiple mitigations to struggle to learning the ECs. Appendix D.2 also tests the *reconstruction loss* mitigation.

5.3 RQ.3: Can greater encoder capacity avoid implicit Reasoning Shortcuts?

MOSAIC's and PNet backbone is realised by the four-convolutional stack from Snell et al. (2017). To test whether greater backbone capacity can capture all the EC modes within one prototype, we compare MOSAIC against the PNet using a ResNet12 encoder, under the same setup of the previous RQs. The results for CCN+ are show in Table 4 (other baselines in Appendix D.3). For `Omniglot-EvenOdd`, while still below MOSAIC of 12%, ResNet12 improves F1(C) by 10%. However, in `BDDOIA` it overfits the limited EC annotations and F1(C) drops from 55.17%, achieved with the standard backbone, to 37.29%.

Table 4: F1(C) comparison of CCN+ MOSAIC vs. ResNet-PNet on `Omniglot-EvenOdd` (10-alphabet split) and `BDD-OIA`.

Model	Omniglot	BDD-OIA
ResNet-PNet	75.14±1.00	37.29±10.04
MOSAIC	**87.12** ±0.45	**57.17** ±0.89

5.4 RQ.4: How effectively does MOSAIC discover modes from the data?

We study how MOSAIC's ability to *automatically* discover modes from the data changes as we progressively drop *complete mode coverage* from Definition 4.2, and the impact on concepts learning. We ablate on the support set size in `Omniglot-EvenOdd`: when using only 10 support points per EC, the UMAP plot in Figure 4(a) shows that MOSAIC discovers single, collapsed embedding regions for each: its concept accuracy reduces to the one of PNet, as these embeddings can be captured by *one* prototype each. Conversely, in Figure 4(e) for 500 supports — corresponding to complete coverage — MOSAIC separates the embeddings in diverse, non-adjancent regions, each realising an EC mode. Figure 4(b),(c),(d) show a smooth transition between the mode-blind, collapsed representation to the faithful modeling of all the modes. Finally, we show the impact of reduced EC support sets on ACC(C). Figure 5(left) shows that in `BDDOIA` with a support set of size 5 DPL+MOSAIC recovers the same F1(C) of PNet in Table 1 ($\approx 55\%$). Increasing it up to 20 brings the mean score close to 58% as MOSAIC better differentiates the ICs. Similarly, as the number of discovered prototypes in `Omniglot-EvenOdd` increases (Figure 5(right)), ACC(C) rises (Figure 5(center)).

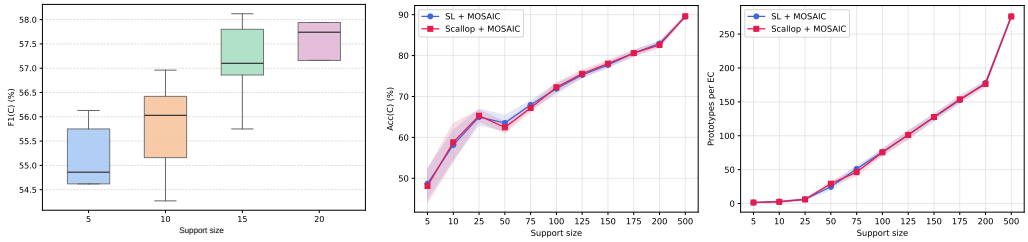


Figure 5: `BDDOIA`: DPL+MOSAIC F1(C) (left), `Omniglot-EvenOdd`: SL/Scallop + MOSAIC Acc(C) (center) and prototypes number per EC (right) when increasing support size.

5.5 RQ.5: Can MOSAIC recover mode-blind prototypical predictors under complete mode coverage?

Here we carry out a single-mode ablation study to test whether, under complete mode coverage, MOSAIC behaves as the mode-blind PNet predictor in Andolfi & Giunchiglia (2025). This is expected to happen whenever (i) either the EC in the data come with single modes or (ii) λ is arbitrary large. To test (i) we run MOSAIC and PNet over the RS benchmark `MNIST-EvenOdd` (Bortolotti et al., 2024), where each $\{0, \cdots, 9\}$ concept is represented by a MNIST digit, and pairs thereof are labeled with their sum. Conversely, to test (ii) we train MOSAIC on `Omniglot-EvenOdd` setting $\upsilon = 0$ in Eq. (2) and then compare with PNet. The results are shown in Table 5; MOSAIC recovers mode-blind prototypes in both settings. Full results in Appendix D.5.

Table 5: MOSAIC and PNet comparison on: (above) `Mnist-EvenOdd` and (below) `Omniglot-EvenOdd`, 10 alphabets.

Model	**F1(C)↑**
Scallop	
+PNet	96.22±1.81
+MOSAIC	**97.56** ±0.68
CCN+	
+PNet	65.12±1.10
+MOSAIC (υ=0)	**65.51** ±1.03

6 Related Works

NeSy AI develops various approaches integrating machine learning and formal logics often smoothing non-differentiable symbolic reasoning within probabilistic (Hoernle et al., 2022; Pryor et al., 2023; van Krieken et al., 2023; Marra & Kuzelka, 2021; Ahmed et al., 2022; Yang et al., 2020; Maene et al., 2024) or fuzzy logics frameworks (van Krieken et al., 2022; Badreddine et al., 2022; Scassola et al., 2024; Diligenti et al., 2012; Donadello et al., 2017). Despite their progresses (van Krieken et al., 2025; Derkinderen et al., 2026), Marconato et al. (2023a; 2024) shows that most of them are affected by RSs and even costly mitigation strategies fail in challenging settings (Marconato et al., 2025). Conversely, prototypical NeSy models (Andolfi & Giunchiglia, 2025) improve concept alignment (Colamonaco et al., 2026) with minimal data requirements, but they assume the concepts induce a single-mode distribution. Following Bortolotti et al. (2025), we connect identifiability to RSs. Our formalisation follows Khemakhem et al. (2020) for equivalence classes and Roeder et al. (2021) for identifiability from matched outputs, while, as in Zimmermann et al. (2021), the ambiguity class is derived from a geometric observable. Our results establish the representation-similarity characterisation in the lines of Theorem 2.2 in Nielsen et al. (2026) for prototypical models.

7 Conclusion

In this paper we shed light on the previously overlooked shortcut reasoning at the level of implicit modalities in the data. We showed that mode-blind predictors are structurally unable to prevent them and introduced **M**ode-Aware Pr**O**totypical Neuro**S**ymbolic **AI** for **C**oncept Learning (**MOSAIC**): our approach automatically construct a prototype for each mode discovered from the support set and during training it jointly optimises query-prototypes proximity and constraints satisfaction. We characterised MOSAIC's deterministic optima and the conditions needed to learn the intended implicit and explicit concept semantics, bridging the gap with identifiability theory. Finally, we validated the RSs mitigation capabilities of MOSAIC with extensive experiments across the `Omniglot-EvenOdd` benchmark we introduced and the autonomous driving task `BDDOIA`.

8 AI DISCLOSURE STATEMENT

In this work, we used generative AI tools for improving the readability of the paper and for mechanical software-engineering tasks, namely migrating the public Infinite Mixture Prototypes codebase Allen et al. (2019) from PyTorch 0.3.1 to PyTorch 2.6, writing test cases, and debugging. We have not used generative AI tools to generate data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims or write proofs, propose hypotheses, design experiments, implement the proposed method, or interpret results. Translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code, as described above, and for editing the manuscript to improve readability. We have carefully reviewed all AI-assisted work: the migrated codebase was checked against the original implementation and the results reported by Allen et al. (2019), and every AI-assisted code change was reviewed and tested by the authors before being used in any experiment. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

9 ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work involves no human subjects, user studies, or collection of new personal data. `BDD-OIA` consists of real driving footage that may depict pedestrians and vehicles; we use it only in the pre-extracted embedding form of prior work, without attempting to identify individuals.

We consider the main societal effect of this work as positive, since it makes NeSy predictors more faithful to their intended concept semantics and hence easier to audit.

10 REPRODUCIBILITY STATEMENT

The code for reproducing the results in the paper has been submitted with the manuscript. All the dataset used are publicly available (`BDDOIA`) or they can be generated using the code provided from open-source datasets (`Omniglot-EvenOdd`). All the details about reproducibility can be found in Appendix C.

REFERENCES

Kareem Ahmed, Stefano Teso, Kai-Wei Chang, Guy Van den Broeck, and Antonio Vergari. Semantic probabilistic layers for neuro-symbolic learning. In *Proceedings of NeurIPS*, 2022.

David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pp. 1–198. Springer, Berlin, 1985. URL http://www.springerlink.com/content/c31v17440871210x/fulltext.pdf.

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 232–241. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen19b.html.

Luca Andolfi and Eleonora Giunchiglia. Right for the right reasons: Avoiding reasoning shortcuts via prototypical neurosymbolic ai. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 157309–157343. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e6bb63d00f597c380c012bd99c68b0c1-Paper-Conference.pdf.

Samy Badreddine, Artur S. d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303, 2022.

Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. In *Proceedings of NeurIPS*, 2024.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 173204–173243. Curran Associates, Inc., 2025. doi: 10.52202/085713-5763. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fcdccd419c4dc471fa3b73ec97b53789-Paper-Conference.pdf.

Seewon Choi, Alaia Solko-Breslin, Rajeev Alur, and Eric Wong. CTSketch: compositional tensor sketching for scalable neurosymbolic learning. In *Proceedings of NeurIPS*, 2025.

Stefano Colamonaco, David Debot, Pietro Barbiero, and Giuseppe Marra. Prototype-grounded concept models for verifiable concept alignment. In *Proceedings of the 43rd International Conference on Machine Learning*, ICML '26, 2026.

Vincent Derkinderen, Robin Manhaeve, Rik Adriaensen, Lucas Van Praet, Lennert De Smet, Giuseppe Marra, and Luc De Raedt. The deeplog neurosymbolic machine, 2026. URL https://arxiv.org/abs/2508.13697.

Michelangelo Diligenti, Marco Gori, Marco Maggini, and Leonardo Rigutini. Bridging logic and kernel machines. *Machine Learning*, 2012.

Ivan Donadello, Luciano Serafini, and Artur S. d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proceedings of IJCAI*, pp. 1596–1602, 2017.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 2023.

Eleonora Giunchiglia, Alex Tatomir, Mihaela Catalina Stoian, and Thomas Lukasiewicz. CCN+: A neuro-symbolic framework for deep learning with requirements. *International Journal of Approximate Reasoning*, 171, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, 2016.

Nicholas Hoernle, Rafael-Michael Karampatsis, Vaishak Belle, and Kobi Gal. MultiplexNet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of Association for the Advancement of Artificial Intelligence*, 2022.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Ziyang Li, Jiani Huang, and Mayur Naik. Scallop: A language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages*, 7(PLDI), 2023.

Jaron Maene, Vincent Derkinderen, and Luc De Raedt. On the hardness of probabilistic neurosymbolic learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Proceedings of NeurIPS*, 2018.

Emanuele Marconato, Gianpaolo Bontempo, Elisa Ficarra, Simone Calderara, Andrea Passerini, and Stefano Teso. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. In *Proceedings of ICML*, 2023a.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. In *Proceedings of NeurIPS*, 2023b.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Antonio Vergari, Andrea Passerini, and Stefano Teso. BEARS make neuro-symbolic models aware of their reasoning shortcuts. In *Proceedings of UAI*, 2024.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Paolo Morettin, Elena Umili, Antonio Vergari, Efthymia Tsamoura, Andrea Passerini, and Stefano Teso. Symbol grounding in neuro-symbolic ai: A gentle introduction to reasoning shortcuts, 2025. URL https://arxiv.org/abs/2510.14538.

Giuseppe Marra and Ondrej Kuzelka. Neural markov logic networks. In *Proceedings of UAI*, 2021.

Beatrix Miranda Ginn Nielsen, Emanuele Marconato, Andrea Dittadi, and Luigi Gresele. When does closeness in distribution imply representational similarity? an identifiability perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=mQOaC3uH2r.

Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Yang Wang, and Lise Getoor. Neupsl: Neural probabilistic soft logic. In *Proceedings of IJCAI*, 2023.

Luc De Raedt, Sebastijan Dumancic, Robin Manhaeve, and Giuseppe Marra. From statistical relational to neuro-symbolic artificial intelligence. In *Proceedings of IJCAI*, 2020.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/roeder21a.html.

Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:1–1, 01 2022. doi: 10.1109/ACCESS.2022.3167702.

Davide Scassola, Sebastiano Saccani, Ginevra Carbone, and Luca Bortolussi. Zero-shot conditioning of score-based diffusion models by neuro-symbolic constraints, 2024. URL https://arxiv.org/abs/2308.16534.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of NeurIPS*, 2017.

Emile van Krieken, Erman Acar, and Frank van Harmelen. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 2022.

Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank Van Harmelen, and Annette Ten Teije. A-neSI: A scalable approximate method for probabilistic neurosymbolic inference. In *Proceedings of NeurIPS*, 2023.

Emile van Krieken, Pasquale Minervini, Edoardo Ponti, and Antonio Vergari. Neurosymbolic diffusion models, 2025. URL https://arxiv.org/abs/2505.13138.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of ICML*, 2018.

Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhun Yang, Adam Ishay, and Joohyung Lee. Neurasp: Embracing neural networks into answer set programming. In *Proceedings of IJCAI*, 2020.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642, 2018.

Roland Simon Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231942400.

A THEORETICAL APPENDIX

A.1 IMPLICIT AND EXPLICIT CONCEPTS DISTRIBUTION IN MOSAIC.

Given the likelihood $p_\theta(\mathbf{z}_i \mid L_i = \mu) = \mathcal{N}(\mathbf{z}_i \mid \mathbf{c}_\mu, \sigma_i^2\mathbf{I})$, we apply Bayes rule:

$$p_\theta(L_i = \mu \mid \mathbf{z}_i) = \frac{p_\theta(\mathbf{z}_i \mid L_i = \mu)p(L_i = \mu)}{p_\theta(\mathbf{z}_i)} = \frac{\pi_\mu \mathcal{N}(\mathbf{z}_i \mid \mathbf{c}_\mu, \sigma_i^2\mathbf{I})}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \mathcal{N}(\mathbf{z}_i \mid \mathbf{c}_{\mu'}, \sigma_i^2\mathbf{I})} \tag{7}$$

where $p_\theta(\mathbf{z}_i) = \sum_{\mu' \in [m_i]} \pi_{\mu'} \mathcal{N}(\mathbf{z}_i \mid \mathbf{c}_{\mu'}, \sigma_i^2\mathbf{I})$ by the law of total probability. Since the multivariate Gaussian density in η_i dimensions with mean μ and covariance Σ is equal to:

$$\mathcal{N}(\mathbf{z} \mid \mathbf{c}_\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^{\eta_i} det\Sigma}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{c}_\mu)^\top \Sigma^{-1}(\mathbf{z} - \mathbf{c}_\mu)\right)$$

the isotropic $\Sigma = \sigma_i^2\mathbf{I}$ case gives:

$$\mathcal{N}(\mathbf{z} \mid \mathbf{c}_\mu, \sigma^2\mathbf{I}) = \frac{1}{(2\pi)^{\eta_i/2}\sigma_i^{\eta_i}} \exp\left(\frac{-||\mathbf{z} - \mathbf{c}_\mu||_2^2}{2\sigma_i^2}\right) = (2\pi\sigma_i^2)^{-\eta_i/2} \exp\left(\frac{-||\mathbf{z} - \mathbf{c}_\mu||_2^2}{2\sigma_i^2}\right) \tag{8}$$

By replacing Eq.(8) into Eq.(7) we obtain:

$$p_\theta(L_i = \mu \mid \mathbf{z}_i) = \frac{\pi_\mu \exp(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2)}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \exp(-||\mathbf{z}_i - \mathbf{c}_{\mu'}||_2^2/2\sigma_i^2)}$$

which is the implicit concepts distribution of Eq.(3). The distribution for each explicit concept c is then obtained by aggregating the distances between $\mathbf{z}_i$ and the centroids of all the modes belonging to the mode-lifts of c wrt. τ:

$$p_\theta(C_i = c \in [h_i] \mid \mathbf{z}_i = f_\theta^i(\mathbf{x})) = \frac{\sum_{\mu \in \tau_i^{-1}(c)} \pi_\mu \exp(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2)}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \exp(-||\mathbf{z}_i - \mathbf{c}_{\mu'}||_2^2/2\sigma_i^2)}.$$

In plain words, the concept posterior is the sum of the mode-lifts posterior and this matches Eq.(4). Then, assuming all the modes μ have the same per-class weight $\frac{1}{h_i Z_c^i}$, where $Z_c^i = |\tau_i^{-1}(c)|$, we obtain:

$$p_\theta(C_i = c \in [h_i] \mid \mathbf{x}) = \frac{\frac{1}{Z_c^i} \sum_{\mu \in \tau_i^{-1}(c)} \exp(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2)}{\sum_{c' \in [h_i]} \frac{1}{Z_{c'}^i} \sum_{\mu' \in \tau_i^{-1}(c')} \exp(-||\mathbf{z}_i - \mathbf{c}_{\mu'}||_2^2/2\sigma_i^2)}$$

Now, taking the logarithms yields:

$$\log p_\theta(C_i = c \in [h_i] \mid \mathbf{z}_i = f_\theta^i(\mathbf{x})) = -\log(Z_c^i) + \text{logsumexp}_{\mu \in \tau_i^{-1}(c)}(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2) - b$$

where b is constant with respect to c and expresses the shared log-denominator. Rearranging the terms give Eq.(5).

A.2 MOSAIC EMBEDDING UPDATE - THEOREM 3.1.

Statement: Let $\mathcal{G} = [h_1] \times \ldots \times [h_k]$ and $\mathcal{M} = [m_1] \times \ldots \times [m_k]$ be the sets of ECs and ICs related through the surjection $\tau : \mathcal{M} \to \mathcal{G}$. Consider a mode-aware prototypical NeSy predictor with output $\hat{\mathbf{y}}$ trained using $\mathcal{L}^{\text{NeSy}}(\hat{\mathbf{y}}, \text{K})$ encoding the satisfaction of knowledge K by $\hat{\mathbf{y}}$. For every $i \in [k]$ the gradient of the loss calculated wrt the embedding $\mathbf{z}_i$ is:

$$\nabla_{\mathbf{z}_i} \mathcal{L}^{\text{NeSy}}(\hat{\mathbf{y}}, \text{K}) = \frac{1}{\sigma_i^2} \sum_{c \in [h_i]} \sum_{\mu \in \tau_i^{-1}(c)} \frac{\partial \mathcal{L}^{\text{NeSy}}}{\partial \hat{y}_c^{EC}} \hat{y}_\mu (\mathbf{c}_\mu - \sum_{\mu' \in [m_i]} \hat{y}_{\mu'} \mathbf{c}_{\mu'}),$$

where $\hat{y}_\mu = p_\theta(L_i = \mu \mid \mathbf{z}_i)$ is the softmax over the distances and $\hat{y}_c^{\text{EC}} = \sum_{\mu \in \tau_i^{-1}(c)} \hat{y}_\mu$.

Proof.
First observe that $\mathcal{L}^{\text{NeSy}}$ depends on $\hat{y}^{EC}_{\tau_i(\mu)}$, which in turn depends on the mode μ determined by the distance between $\mathbf{c}_\mu$ and $\mathbf{z}_i$. Hence we can write:

$$\nabla_{\mathbf{z}_i}\mathcal{L}^{\text{NeSy}}(\hat{\mathbf{y}}, \mathrm{K}) = \sum_{c\in[h_i]} \frac{\partial \mathcal{L}^{\text{NeSy}}}{\partial \hat{y}^{\text{EC}}_c} \sum_{\mu\in[m_i]} \frac{\partial \hat{y}^{\text{EC}}_c}{\partial \hat{y}_\mu} \nabla_{\mathbf{z}_i}\hat{y}_\mu, \tag{9}$$

where,

$$\hat{y}^{\text{EC}}_c := \sum_{\mu\in\tau_i^{-1}(c)} \hat{y}_\mu, \qquad \frac{\partial \hat{y}^{\text{EC}}_c}{\partial \hat{y}_\mu} = \mathbb{1}(\tau_i(\mu) = c).$$

and

$$\nabla_{\mathbf{z}_i}\hat{y}_\mu = \nabla_{\mathbf{z}_i} \frac{\pi_\mu \exp\big(-\|\mathbf{z}_i - \mathbf{c}_\mu\|_2^2/2\sigma_i^2\big)}{\sum_{\mu'\in[m_i]} \pi_{\mu'} \exp\big(-\|\mathbf{z}_i - \mathbf{c}_{\mu'}\|_2^2/2\sigma_i^2\big)}. \tag{10}$$

The priors π_μ do not depend on $\mathbf{z}_i$, so they are not included in the derivation below. We consider one element j of $\mathbf{z}_i$ and set $d_\mu = -||\mathbf{z}_i - \mathbf{c}_\mu||_2^2/2\sigma_i^2$. Then Eq.(10) becomes:

$$\frac{\partial \hat{y}_\mu}{\partial \mathbf{z}_i^j} = \sum_{\mu'\in[m_i]} \frac{\partial \hat{y}_\mu}{\partial d_{\mu'}} \frac{\partial d_{\mu'}}{\partial \mathbf{z}_i^j}. \tag{11}$$

The derivative of the softmax $\frac{\partial \hat{y}_\mu}{\partial d_{\mu'}}$ has Kronecker delta equal to $\hat{y}_\mu(\delta_{\mu\mu'} - \hat{y}_{\mu'})$. Substituting the delta in (11) and accounting for all the mode-lifts for each concept yields:

$$\frac{\partial \hat{y}_\mu}{\partial \mathbf{z}_i^j} = \hat{y}_\mu[(1-\hat{y}_\mu)\frac{\partial d_\mu}{\partial \mathbf{z}_i^j} - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')|\mu'\neq\mu} \hat{y}_{\mu'} \frac{\partial d_{\mu'}}{\partial \mathbf{z}_i^j}]. \tag{12}$$

Given that

$$\frac{\partial d_\mu}{\partial \mathbf{z}_i^j} = \frac{\partial(-||\mathbf{z}_i - \mathbf{c}_\mu||_2^2)/2\sigma_i^2}{\partial \mathbf{z}_i^j} = -\frac{\mathbf{z}_i^j - e_\mu^j}{\sigma_i^2},$$

where e_μ^j is the j-th component of $\mathbf{c}_\mu$, we further rewrite Eq.(12) as:

$$\begin{aligned}\frac{\partial \hat{y}_\mu}{\partial \mathbf{z}_i^j} &= \hat{y}_\mu[\frac{(1-\hat{y}_\mu)}{\sigma_i^2}(-(\mathbf{z}_i^j - e_\mu^j)) - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')|\mu'\neq\mu} \hat{y}_{\mu'}(-\frac{\mathbf{z}_i^j - e_{\mu'}^j}{\sigma_i^2})] = \\ &= \frac{1}{\sigma_i^2}\hat{y}_\mu[-(\mathbf{z}_i^j - e_\mu^j) + \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'}(\mathbf{z}_i^j - e_{\mu'}^j)]\end{aligned} \tag{13}$$

We observe that the last term can be decomposed as:

$$\begin{aligned}&\sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'}(\mathbf{z}_i^j - e_{\mu'}^j) = \\ &= \mathbf{z}_i^j \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'} - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'} e_{\mu'}^j = \\ &= \mathbf{z}_i^j - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'} e_{\mu'}^j\end{aligned}$$

since $\sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'}$ equals 1. Substituting the previous expression into Eq.(13) yields:

$$\begin{aligned}&\frac{1}{\sigma_i^2}\hat{y}_\mu[-\mathbf{z}_i^j + e_\mu^j + \mathbf{z}_i^j - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'} e_{\mu'}^j] = \\ &= \frac{1}{\sigma_i^2}\hat{y}_\mu[e_\mu^j - \sum_{c'\in[h_i]} \sum_{\mu'\in\tau_i^{-1}(c')} \hat{y}_{\mu'} e_{\mu'}^j]\end{aligned} \tag{14}$$

To conclude the proof notice that the possible mode-lifts partition the mode set $[m_i] = \bigsqcup_{c' \in [h_i]} \tau_i^{-1}(c')$. By stacking all the j-th components of e and using the partition Eq.(14), we obtain the final expression for $\nabla_{\mathbf{z}_i} \mathbf{y}_\mu$ as:

$$\nabla_{\mathbf{z}_i} \hat{y}_\mu = \frac{1}{\sigma_i^2} \hat{y}_\mu \Big(\mathbf{c}_\mu - \sum_{\mu' \in [m_i]} \hat{y}_{\mu'} \mathbf{c}_{\mu'} \Big) \tag{15}$$

And substituting in Eq.(9) finally gives:

$$\nabla_{\mathbf{z}_i} \mathcal{L}^{\text{NeSy}}(\hat{y}, \mathrm{K}) = \frac{1}{\sigma_i^2} \sum_{c \in [h_i]} \sum_{\mu \in \tau_i^{-1}(c)} \frac{\partial \mathcal{L}^{\text{NeSy}}}{\partial \hat{y}_c^{\text{EC}}} \hat{y}_\mu (\mathbf{c}_\mu - \sum_{\mu' \in [m_i]} \hat{y}_{\mu'} \mathbf{c}_{\mu'})$$

which is the claim, where $\frac{\partial \hat{y}_c^{\text{EC}}}{\partial \hat{y}_\mu} = \mathbb{1}(\tau_i(\mu) = c)$ is expressed in the second sum. □

A.3 Number of Deterministic Optima of Mode-Aware Predictors - Proposition 4.1.

Statement: Consider a mode-aware NeSy predictor inducing $p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathrm{K})$. Let $\mathcal{G} = [h_1] \times \ldots \times [h_k]$, $\mathcal{M} = [m_1] \times \ldots \times [m_k]$ and $\mathcal{D}$ be a dataset realising every $\mathbf{m} \in \texttt{supp}(\mathbf{M})$. Under **[A1,2,3]** the number of deterministic optima for $p_\theta(\mathbf{L} \mid \mathbf{X})$ is:

$$\sum_{\gamma \in \Gamma} \mathbb{1}\Big(\bigwedge_{\mathbf{m} \in \texttt{supp}(\mathbf{M})} \Big[\Big(\bigwedge_{i \in [k]} \big(\mu_i \in \mathcal{S}_i^{\text{sup}} \;\Rightarrow\; \tau_i(\gamma_i(\mu_i)) = \tau_i(\mu_i) \big) \Big) \;\wedge\; \big((\beta_{\mathrm{K}} \circ \tau \circ \gamma)(\mathbf{m}) = (\beta_{\mathrm{K}} \circ \tau)(\mathbf{m}) \big) \Big] \Big),$$

where $\mathcal{S}_i^{\text{sup}}$ denotes the set of modes occurring in at least one EC-annotated datapoint in $\mathcal{D}$ and $\Gamma \subseteq \prod_{i \in [k]} \{\gamma_i : [m_i] \to [m_i]\}$ is the set of coordinate-wise assignments induced by the deterministic distributions $p_\theta(\mathbf{L} \mid \mathbf{X})$ the predictor can express, acting as $\gamma(\mathbf{m}) = (\gamma_1(\mu_1), \ldots, \gamma_k(\mu_k))$.

Proof. First, recall the working assumptions:

- *Deterministic knowledge* ([**A1**]): the distribution $p^*(\mathbf{Y} \mid \mathbf{G}; \mathrm{K})$ is induced by a deterministic mapping $\beta_{\mathrm{K}} \colon \mathbf{g} \to \mathbf{y}$
- *Modes extrapolability* ([**A2**]): the distribution $p^*(\mathbf{X} \mid \mathbf{G}, \mathbf{M}, \mathbf{S}) = p^*(\mathbf{X} \mid \mathbf{M}, \mathbf{S}) = \delta\{\mathbf{X} = \phi_\tau(\mathbf{m}, \mathbf{s})\}$ is induced through a map $\phi_\tau \colon (\mathbf{m}, \mathbf{s}) \mapsto \mathbf{x}$, where ϕ_τ is invertible, and smooth over $\mathbf{s}$.
- *Modes clusterability* ([**A3**]): for every $i \in [k]$ all datapoints generated from a given mode μ_i share the same nearest prototype, and every prototype is nearest to datapoints from at least one mode. Distinct modes may nevertheless share the same prototype. Equivalently, the map

  $$\mathbf{x} \mapsto \arg\min_j \left\| f_\theta^i(\mathbf{x}) - \mathbf{c}_j \right\|_2^2 \tag{16}$$

  depends on $\mathbf{x} = \phi_\tau(\mathbf{m}, \mathbf{s})$ only through μ_i. In the deterministic limit $\sigma_i \to 0$, $p_\theta(L_i \mid \mathbf{x})$ concentrates on the index of the nearest prototype.

Observe that, under [**A2**], the data distribution is induced by a smooth, invertible map and every datapoint gets written as $\mathbf{x} = \phi_\tau(\mathbf{m}, \mathbf{s})$. The proof proceeds in two steps.

Step 1. We show that the correspondence of the mappings $\gamma \in \Gamma$ with the deterministic optima for $p_\theta(\mathbf{C} \mid \mathbf{X})$ is many-to-one; however the same mappings define a bijection with the deterministic optima of $p_\theta(\mathbf{L} \mid \mathbf{X})$, on which we base the count.

Step 2. We derive the counting expression using these latent-modes mappings.

Step 1. Under [**A2**] every datapoint is written as $\mathbf{x} = \phi_\tau(\mathbf{m}, \mathbf{s})$ and, by [**A3**], the nearest prototype at coordinate i depends on $\mathbf{x}$ only through its i-th mode μ_i, and equals $\mathbf{c}_{\gamma_i(\mu_i)}$. Hence every deterministic optimum of $p_\theta(\mathbf{L} \mid \mathbf{X})$ is constant over the support of $p^\star(\mathbf{X} \mid \mathbf{m})$ apart from a zero-measure subspace $\mathcal{X}_{\mathbf{m}}^0$, and is realised by a coordinatewise assignment $\gamma = (\gamma_1, \ldots, \gamma_k) \in \Gamma$ through $p_\theta(\mathbf{L} \mid \mathbf{x}) = \mathbb{1}(\mathbf{L} = \gamma(\mathbf{m}))$; conversely every $\gamma \in \Gamma$ realises exactly one such optimum, so the correspondence is a bijection and we base the count on it. Observe that $\gamma(\mathbf{m}) \in \mathcal{M}$ and need not lie in $\texttt{supp}(\mathbf{M})$, since the joint support is not in general a product of its marginals and a factorised predictor may assign a combination of modes never realised jointly in the data; as τ and β_{K} are defined on the whole of $\mathcal{M}$, the conditions below remain well posed.

The same is not true of $p_\theta(\mathbf{C} \mid \mathbf{X})$. Its deterministic optima are the mappings $\alpha = \tau \circ \gamma$, acting as $\alpha(\mathbf{m})_i = \tau_i(\gamma_i(\mu_i))$, and the correspondence $\gamma \mapsto \alpha$ is many-to-one: every γ that permutes modes inside a mode-lift is transparent to α, by definition of the taxonomy τ. There is thus no bijective correspondence between the deterministic optima of $p_\theta(\mathbf{C} \mid \mathbf{X})$ and the α mappings, which is why we count over Γ and push forward through τ.

Step 2. In Step 1 we showed that by characterising γ we capture all the optimal deterministic distributions $p_\theta(\mathbf{L} \mid \mathbf{X})$, and hence, through τ, all the optimal deterministic $p_\theta(\mathbf{C} \mid \mathbf{X})$. Fix $\mathbf{m} \in \mathtt{supp}(\mathbf{M})$ and let $\mathbf{g} = \tau(\mathbf{m})$ be *its own* EC vector. Along a coordinate i whose mode carries an EC annotation, that is $\mu_i \in \mathcal{S}_i^{\text{sup}}$, attaining maximal likelihood requires the predictor to return that annotation,

$$\tau_i(\gamma_i(\mu_i)) = g_i = \tau_i(\mu_i),$$

while a coordinate carrying no annotation is unconstrained by the explicit supervision. Note that the constraint is attached to $\mathbf{m}$ through $\tau(\mathbf{m})$, and not to every $\mathbf{g} \in \mathtt{supp}(\mathbf{G})$ independently. The considered mappings should also satisfy the background knowledge to attain maximum likelihood; under **[A1]** this happens if and only if $(\beta_{\mathrm{K}} \circ \tau \circ \gamma)(\mathbf{m}) = (\beta_{\mathrm{K}} \circ \tau)(\mathbf{m})$. By **[A3]** every $\gamma \in \Gamma$ acts coordinatewise, since the nearest prototype at coordinate i depends on $\mathbf{x}$ only through μ_i. Counting all the assignments in Γ satisfying both conditions over $\mathtt{supp}(\mathbf{M})$ yields our counting expression.

We remark that by **[A3]** the prototypical extractor realises γ on the modes via nearest-centroid assignment, and the explicit supervision constrains it only through $\tau_i \circ \gamma_i$. It is enough for $\mathcal{D}$ to comprise one explicit-concept level annotation inducing the creation of a prototype for each mode to match the same guarantees of dense annotation strategies (Marconato et al., 2023b). □

A.4 Identifiability of Prototypical Extractors - Theorem 4.4.

Statement: For any pair of joint prototype extractors f_θ and $f_{\theta'}$ it holds that:

$$f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'} \implies \Phi(f_\theta^i)(\mathbf{x}, B) = \Phi(f_{\theta'}^i)(\mathbf{x}, B) \quad \text{for all } i \in [k],\ \mathbf{x} \in \mathcal{X} \text{ and all finite } B \subseteq \mathcal{X}.$$

Conversely, if for every $i \in [k]$ there is a finite $\mathcal{B}_i \subset \mathcal{X}$ and an integer $s_i \in [|\mathcal{B}_i| - 1]$ so that:

(C1) $\Phi(f_\theta^i)(\mathbf{x}, B) = \Phi(f_{\theta'}^i)(\mathbf{x}, B)$ for every $\mathbf{x} \in \mathcal{X}$ and every $B \subseteq \mathcal{B}_i$ with $|B| = s_i$ and
(C2) $f_\theta^i(\mathcal{X}) \subseteq \mathtt{aff}\big(f_\theta^i(\mathcal{B}_i)\big)$, where $\mathtt{aff}(\cdot)$ is the affine hull.

Then $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$.

Proof. We prove the forward and the converse implication separately.

Proof of the forward implication ($\Longrightarrow$)

We fix $i \in [k]$ and drop it together with θ from the notation. By Definition 4.3 we can write $f' = Uf + t$ with $U \in \mathcal{O}(\eta)$, $t \in \mathbb{R}^\eta$. Given weights $w_{\mathbf{x}} \geq 0$, by definition of a centroid we have $\mathbf{c}_w(f) := \big(\sum_{\mathbf{x}} w_{\mathbf{x}} f(\mathbf{x})\big) / \big(\sum_{\mathbf{x}} w_{\mathbf{x}}\big)$ for the corresponding weighted mean of embeddings. Now affine maps commute with weighted means,

$$\mathbf{c}_w(Uf + t) = \frac{\sum_{\mathbf{x}} w_{\mathbf{x}} \big(Uf(\mathbf{x}) + t\big)}{\sum_{\mathbf{x}} w_{\mathbf{x}}} = U\, \mathbf{c}_w(f) + t, \tag{17}$$

and for a centroid over a finite set B, $w_{\mathbf{x}} = \mathbb{1}(\mathbf{x} \in B)$ and $\mathbf{c}_B(f') = U\mathbf{c}_B(f) + t$. Hence, for every $\mathbf{x} \in \mathcal{X}$ and every finite $B \subseteq \mathcal{X}$,

$$\Phi(f')(\mathbf{x}, B) = \big\|Uf(\mathbf{x}) + t - U\mathbf{c}_B(f) - t\big\|_2^2 = \big\|U\big(f(\mathbf{x}) - \mathbf{c}_B(f)\big)\big\|_2^2 = \Phi(f)(\mathbf{x}, B).$$

□

Proof of the converse ($\Longleftarrow$)

For any two joint prototype extractors sharing the same distance profile, the proof constructs an orthogonal matrix U_i so that $f_{\theta'}^i(\mathbf{x}) = U_i f_\theta^i(\mathbf{x}) + t_i$, for all $\mathbf{x} \in \mathcal{X}$ and $i \in [k]$ with $t_i \in \mathbb{R}^\eta$. In order to be orthogonal, U_i must preserve all the pairs of inner products. However, observe that the

hypothesis only provides us with the equality $\Phi(f^i_\theta)(\mathbf{x}, B) = \Phi(f^i_{\theta'})(\mathbf{x}, B)$, expressing datapoints-to-centroids distances and not the inner product between datapoints. This additional information must be first derived throughout the proof that builds the desired transformation matrix U_i. Below we lay the grounds for the proof by introducing some additional necessary notation and deriving through it the consequences of the hypothesis. Then we show how U_i is constructed in four steps.

We fix $i \in [k]$ and drop i and θ from the notation for readability. Let s be an integer as in the statement. Fix now the origin point for the embedding space as $\mathbf{x}_0 \in \mathcal{B}$ and set

$$a_{\mathbf{x}} := f(\mathbf{x}) - f(\mathbf{x}_0) \in \mathbb{R}^\eta, \qquad a'_{\mathbf{x}} := f'(\mathbf{x}) - f'(\mathbf{x}_0) \in \mathbb{R}^\eta, \qquad \mathbf{x} \in \mathcal{X},$$

so that $a_{\mathbf{x}_0} = a'_{\mathbf{x}_0} = 0$ by construction. For a block B put $\bar{a}_B := \frac{1}{|B|}\sum_{\mathbf{b}\in B} a_{\mathbf{b}} = \mathbf{c}_B(f) - f(\mathbf{x}_0)$, and $\bar{a}'_B$ analogously. Since $f(\mathbf{x}) - \mathbf{c}_B(f) = \big(f(\mathbf{x}) - f(\mathbf{x}_0)\big) - \big(\mathbf{c}_B(f) - f(\mathbf{x}_0)\big) = a_{\mathbf{x}} - \bar{a}_B$, the reference point cancels out and we get:

$$\Phi(f)(\mathbf{x}, B) = \big\|a_{\mathbf{x}} - \bar{a}_B\big\|_2^2 = \|a_{\mathbf{x}}\|_2^2 - 2\langle a_{\mathbf{x}}, \bar{a}_B\rangle + \|\bar{a}_B\|_2^2, \tag{18}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product. In particular, $\Phi(f)(\mathbf{x}_0, B) = \|\bar{a}_B\|_2^2$, so the *centred profile* has expression:

$$\Delta(\mathbf{x}, B) := \Phi(f)(\mathbf{x}, B) - \Phi(f)(\mathbf{x}_0, B) = \|a_{\mathbf{x}}\|_2^2 - 2\langle a_{\mathbf{x}}, \bar{a}_B\rangle, \tag{19}$$

and, crucially, it is free of the centroid-norm term, and likewise for Δ'. Evaluating (C1) at the queries $\mathbf{x}$ and $\mathbf{x}_0$ gives

$$\Delta(\mathbf{x}, B) = \Delta'(\mathbf{x}, B) \qquad \text{for all } \mathbf{x} \in \mathcal{X} \text{ and all } B \subseteq \mathcal{B},\ |B| = s. \tag{20}$$

The proof proceeds in four steps: **Step 1** recovers the inner products between arbitrary embeddings and support embeddings, **Step 2** the norms, **Step 3** builds an orthogonal U on the span of the support, **Step 4** lifts it to all of $\mathcal{X}$.

Step 1. This step proves that inner products are preserved. Let $\mathbf{y}, \mathbf{z} \in \mathcal{B}$ with $\mathbf{y} \neq \mathbf{z}$. Since $|\mathcal{B}| \geq s + 1$ we can pick $D \subseteq \mathcal{B} \setminus \{\mathbf{y}, \mathbf{z}\}$ with $|D| = s - 1$ and set:

$$B_1 := \{\mathbf{y}\} \cup D, \qquad B_2 := \{\mathbf{z}\} \cup D \qquad \Longrightarrow \qquad |B_1| = |B_2| = s,$$

as two admissible blocks differing in exactly one element, so that:

$$\bar{a}_{B_1} - \bar{a}_{B_2} = \frac{1}{s}\Big(a_{\mathbf{y}} + \sum_{\mathbf{d}\in D} a_{\mathbf{d}}\Big) - \frac{1}{s}\Big(a_{\mathbf{z}} + \sum_{\mathbf{d}\in D} a_{\mathbf{d}}\Big) = \frac{1}{s}\big(a_{\mathbf{y}} - a_{\mathbf{z}}\big). \tag{21}$$

Taking the difference of the two centred profiles cancels the query norm: by Eq. (19) and Eq. (21), for every $\mathbf{x} \in \mathcal{X}$,

$$\Delta(\mathbf{x}, B_1) - \Delta(\mathbf{x}, B_2) = -2\big\langle a_{\mathbf{x}},\, \bar{a}_{B_1} - \bar{a}_{B_2}\big\rangle = -\frac{2}{s}\big\langle a_{\mathbf{x}},\, a_{\mathbf{y}} - a_{\mathbf{z}}\big\rangle. \tag{22}$$

The same identity holds for f' with the same constant s, which does not depend on the extractor. Hence combining Eq. (22) with Eq. (20) we get:

$$\big\langle a_{\mathbf{x}},\, a_{\mathbf{y}} - a_{\mathbf{z}}\big\rangle = \big\langle a'_{\mathbf{x}},\, a'_{\mathbf{y}} - a'_{\mathbf{z}}\big\rangle \qquad \text{for all } \mathbf{x} \in \mathcal{X},\ \mathbf{y}, \mathbf{z} \in \mathcal{B}, \tag{23}$$

(the case $\mathbf{y} = \mathbf{z}$ being trivial). Choosing $\mathbf{z} = \mathbf{x}_0$, which is legitimate because $\mathbf{x}_0 \in \mathcal{B}$, and using $a_{\mathbf{x}_0} = a'_{\mathbf{x}_0} = 0$ yields:

$$\boxed{\big\langle a_{\mathbf{x}}, a_{\mathbf{y}}\big\rangle = \big\langle a'_{\mathbf{x}}, a'_{\mathbf{y}}\big\rangle \quad \text{for all } \mathbf{x} \in \mathcal{X},\ \mathbf{y} \in \mathcal{B}.} \tag{24}$$

Note that Eq. (24) holds in particular for $\mathbf{x} \in \mathcal{B}$, i.e. the two extractors induce the same Gram matrix on the support.

Step 2. This step proves that norms are preserved. Let $\mathbf{x} \in \mathcal{X}$ be arbitrary and let $B \subseteq \mathcal{B}$, $|B| = s$. By Eq. (24) applied to each $\mathbf{b} \in B \subseteq \mathcal{B}$,

$$\big\langle a_{\mathbf{x}}, \bar{a}_B\big\rangle = \frac{1}{s}\sum_{\mathbf{b}\in B}\big\langle a_{\mathbf{x}}, a_{\mathbf{b}}\big\rangle = \frac{1}{s}\sum_{\mathbf{b}\in B}\big\langle a'_{\mathbf{x}}, a'_{\mathbf{b}}\big\rangle = \big\langle a'_{\mathbf{x}}, \bar{a}'_B\big\rangle. \tag{25}$$

Rearranging Eq. (19) as $\|a_{\mathbf{x}}\|_2^2 = \Delta(\mathbf{x}, B) + 2\langle a_{\mathbf{x}}, \bar{a}_B\rangle$ and matching the two terms on the right – the first by Eq. (20), the second by Eq. (25) – gives

$$\|a_{\mathbf{x}}\|_2^2 = \|a'_{\mathbf{x}}\|_2^2 \qquad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{26}$$

This step is what extends the information of Eq. (24), which involves the support, to *every* point of $\mathcal{X}$ and it is used in Steps 3 and 4 that build the actual matrix U.

Step 3. We now proceed to construct an orthogonal map on the support span. Any matrix U realising the equivalence must map each $a_{\mathbf{s}}$ to $a'_{\mathbf{s}}$, so we build it first on $\mathrm{span}\{a_{\mathbf{s}}\}$. Here we start building U from the support points and in the following (final) step we generalise to arbitrary queries.

Let $V := \mathrm{span}\{a_{\mathbf{s}} \mid \mathbf{s} \in \mathcal{B}\}$ and $r := \dim V$. Choose $\mathbf{s}_1, \dots, \mathbf{s}_r \in \mathcal{B}$ such that $a_{\mathbf{s}_1}, \dots, a_{\mathbf{s}_r}$ is a *basis* of V and define the linear map $U_0 : V \to \mathbb{R}^\eta$ by

$$U_0\, a_{\mathbf{s}_j} := a'_{\mathbf{s}_j}, \qquad j = 1, \dots, r,$$

extended linearly. Constructing U_0 on a basis makes it well defined, since the family $\{a_{\mathbf{s}}\}_{\mathbf{s}\in\mathcal{B}}$ need not be linearly independent. Observe that:

(i) U_0 is a linear isometry. For $v = \sum_j \lambda_j a_{\mathbf{s}_j}$,

$$\big\|U_0 v\big\|_2^2 = \sum_{j,l} \lambda_j \lambda_l \big\langle a'_{\mathbf{s}_j}, a'_{\mathbf{s}_l}\big\rangle \overset{(24)}{=} \sum_{j,l} \lambda_j \lambda_l \big\langle a_{\mathbf{s}_j}, a_{\mathbf{s}_l}\big\rangle = \|v\|_2^2, \tag{27}$$

so U_0 preserves norms on V and inner products: $\langle U_0 u, U_0 v\rangle = \langle u, v\rangle$ for all $u, v \in V$.
In particular U_0 is injective.

(ii) $U_0 a_{\mathbf{s}} = a'_{\mathbf{s}}$ for every $\mathbf{s} \in \mathcal{B}$ and not only for the basis points. Indeed, we can write: $a_{\mathbf{s}} = \sum_j \lambda_j a_{\mathbf{s}_j}$. Then, by Eq. (24),

$$\big\langle U_0 a_{\mathbf{s}}, a'_{\mathbf{s}}\big\rangle = \sum_j \lambda_j \big\langle a'_{\mathbf{s}_j}, a'_{\mathbf{s}}\big\rangle \overset{(24)}{=} \sum_j \lambda_j \big\langle a_{\mathbf{s}_j}, a_{\mathbf{s}}\big\rangle = \big\langle a_{\mathbf{s}}, a_{\mathbf{s}}\big\rangle = \|a_{\mathbf{s}}\|_2^2 \overset{(26)}{=} \|a'_{\mathbf{s}}\|_2^2,$$

while $\|U_0 a_{\mathbf{s}}\|_2^2 = \|a_{\mathbf{s}}\|_2^2 = \|a'_{\mathbf{s}}\|_2^2$ by Eq. (27) and Eq. (26). Hence, we obtain:

$$\big\|U_0 a_{\mathbf{s}} - a'_{\mathbf{s}}\big\|_2^2 = \|U_0 a_{\mathbf{s}}\|_2^2 - 2\big\langle U_0 a_{\mathbf{s}}, a'_{\mathbf{s}}\big\rangle + \|a'_{\mathbf{s}}\|_2^2 = \|a'_{\mathbf{s}}\|_2^2 - 2\|a'_{\mathbf{s}}\|_2^2 + \|a'_{\mathbf{s}}\|_2^2 = 0.$$

Consequently, $V' := U_0(V) = \mathrm{span}\{a'_{\mathbf{s}} \mid \mathbf{s} \in \mathcal{B}\}$ and $\dim V' = \dim V = r$.

(iii) U_0 has an extension to $\mathbb{R}^\eta$. Let $V^\perp = \{\mathbf{w} \in \mathbb{R}^\eta \mid \langle \mathbf{w}, \mathbf{v}\rangle = 0\ \forall \mathbf{v} \in V\}$, so that $\mathbb{R}^\eta = V \oplus V^\perp$ and

$$\dim V^\perp = \eta - \dim V = \eta - \dim V' = \dim V'^\perp.$$

Pick any linear isometry $\iota : V^\perp \to V'^\perp$ sending an orthonormal basis of $V^\perp$ to one of $V'^\perp$ and define $U := U_0 \oplus \iota$, i.e., $U(v + \mathbf{w}) := U_0 v + \iota \mathbf{w}$ for $v \in V$, $\mathbf{w} \in V^\perp$. Since $U_0 v \in V'$ and $\iota \mathbf{w} \in V'^\perp$ are orthogonal, Pythagoras applies and it gives:

$$\|U(v + \mathbf{w})\|_2^2 = \|U_0 v\|_2^2 + \|\iota \mathbf{w}\|_2^2 = \|v\|_2^2 + \|\mathbf{w}\|_2^2 = \|v + \mathbf{w}\|_2^2,$$

so U is a norm-preserving linear bijection of $\mathbb{R}^\eta$, that is $U \in \mathcal{O}(\eta)$, and $U a_{\mathbf{s}} = a'_{\mathbf{s}}$ for every $\mathbf{s} \in \mathcal{B}$ by (ii).

Step 4. To conclude the proof, we show how to lift U to all of $\mathcal{X}$. Consider an arbitrary $\mathbf{x} \in \mathcal{X}$. By **(C2)**, $f(\mathbf{x}) \in \mathtt{aff}(f(\mathcal{B}))$, so there are $\mathbf{s}_j \in \mathcal{B}$ and scalars λ_j with $\sum_j \lambda_j = 1$ and $f(\mathbf{x}) = \sum_j \lambda_j f(\mathbf{s}_j)$. Subtracting $f(\mathbf{x}_0) = \sum_j \lambda_j f(\mathbf{x}_0)$, which is legitimate precisely because the coefficients sum to one, gives

$$a_{\mathbf{x}} = \sum_j \lambda_j a_{\mathbf{s}_j} \in V, \qquad \text{hence} \qquad U a_{\mathbf{x}} = U_0 a_{\mathbf{x}} \in V'. \tag{28}$$

Also notice this is the only point in which the proof needs **(C2)**.

Now set $\mathbf{w} := a'_{\mathbf{x}} - U a_{\mathbf{x}}$. For every $\mathbf{s} \in \mathcal{B}$, using $a'_{\mathbf{s}} = U a_{\mathbf{s}}$ from Step 3(ii), then Step 3(i), then Eq. (24),

$$\big\langle \mathbf{w}, a'_{\mathbf{s}}\big\rangle = \big\langle a'_{\mathbf{x}}, a'_{\mathbf{s}}\big\rangle - \big\langle U a_{\mathbf{x}}, U a_{\mathbf{s}}\big\rangle = \big\langle a'_{\mathbf{x}}, a'_{\mathbf{s}}\big\rangle - \big\langle a_{\mathbf{x}}, a_{\mathbf{s}}\big\rangle = 0,$$

so $\mathbf{w} \perp V'$, the $a'_{\mathbf{s}}$ spanning V'. Together with Eq. (28) the decomposition $a'_{\mathbf{x}} = U a_{\mathbf{x}} + \mathbf{w}$ is orthogonal ($U a_{\mathbf{x}} \in V'$, $\mathbf{w} \in V'^\perp$), so by Pythagoras and Step 3(iii)

$$\|a'_{\mathbf{x}}\|_2^2 = \|U a_{\mathbf{x}}\|_2^2 + \|\mathbf{w}\|_2^2 = \|a_{\mathbf{x}}\|_2^2 + \|\mathbf{w}\|_2^2.$$

By Eq. (26) the two outer norms coincide, whence $\mathbf{w} = 0$ and $a'_{\mathbf{x}} = U a_{\mathbf{x}}$ for every $\mathbf{x} \in \mathcal{X}$. Unfolding the definitions of a and a',

$$f'(\mathbf{x}) = U f(\mathbf{x}) + t, \qquad t := f'(\mathbf{x}_0) - U f(\mathbf{x}_0), \qquad \forall\, \mathbf{x} \in \mathcal{X}.$$

Repeating the argument for each $i \in [k]$ yields pairs (U_i, t_i) with $U_i \in \mathcal{O}(\eta)$, i.e. $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$. $\square$

A.5 Distribution Equivalence over Equivalence Class - Lemma 4.5

Statement: Let p_θ and $p_{\theta'}$ be prototypical NeSy predictors. If $f_\theta \sim_{\mathcal{E}(\eta)} f_{\theta'}$, then there is a family of permutations ϱ_i of $[m_i]$ with $\tau_i \circ \varrho_i = \tau_i$ such that $p_\theta(L_i = \mu \mid \mathbf{x}) = p_{\theta'}\big(L_i = \varrho_i(\mu) \mid \mathbf{x}\big)$ for every $i \in [k]$, $\mu \in [m_i]$, $\mathbf{x} \in \mathcal{X}$, and hence $p_\theta(\mathbf{L} = \mathbf{m} \mid \mathbf{X}) = p_{\theta'}(\mathbf{L} = \varrho(\mathbf{m}) \mid \mathbf{X})$ with $\varrho = (\varrho_1, \ldots, \varrho_k)$.

Proof. We consider two arbitrary joint prototype extractors, related by $\sim_{\mathcal{E}(\eta)}$. The taxonomy τ, the mode priors $\pi_\mu = 1/h_i(|\tau_i^{-1}(\tau_i(\mu))|)$, the cluster variances σ_i^2, and the hyperparameters are shared. Without this, even two predictors with identical extractors could already induce different IC distributions.

We fix $i \in [k]$ and let $U_i \in \mathcal{O}(\eta)$, $t_i \in \mathbb{R}^\eta$ be so that $f^i_{\theta'} = U_i f^i_\theta + t_i$. Both exist by hypothesis and Definition 4.3. Consider $\mathbf{z}_i = f^i_\theta(\mathbf{x})$ and $\mathbf{z}'_i = f^i_{\theta'}(\mathbf{x})$, and let $\mathbf{c}_\mu$, $\mathbf{c}'_\mu$ be the centroids created by the extractors f^i_θ and for $f^i_{\theta'}$ respectively.

By induction along the execution of Algorithm 1 the two runs take the same branches: every branch depends on the embeddings only through distances, which are preserved by the forward implication of Theorem 4.4; the threshold λ is a function of σ_i, υ, ρ and η_i alone, with ρ the spread of the prototypes and hence invariant under U_i; and centroids transform according to Eq. (17). The two runs therefore create their prototypes in one-to-one correspondence and terminate with the same number of them, so both label sets are $[m_i]$. Corresponding prototypes are created inside the same annotated class (lines 5 and 13 of Algorithm 1) and hence carry the same EC label, so the induced bijection ϱ_i of $[m_i]$ satisfies $\tau_i \circ \varrho_i = \tau_i$ and

$$\mathbf{c}'_{\varrho_i(\mu)} = U_i \mathbf{c}_\mu + t_i, \qquad \mu \in [m_i]. \tag{29}$$

Consequently, for every $\mathbf{x} \in \mathcal{X}$ and every $\mu \in [m_i]$,

$$\big\|\mathbf{z}'_i - \mathbf{c}'_{\varrho_i(\mu)}\big\|_2^2 = \big\|U_i(\mathbf{z}_i - \mathbf{c}_\mu)\big\|_2^2 = \big\|\mathbf{z}_i - \mathbf{c}_\mu\big\|_2^2. \tag{30}$$

By Eq. (3). the IC posterior, we also have that:

$$p_\theta(L_i = \mu \in [m_i] \mid \mathbf{x}) = \frac{\pi_\mu \exp\big(-\|\mathbf{z}_i - \mathbf{c}_\mu\|_2^2/2\sigma_i^2\big)}{\sum_{\mu' \in [m_i]} \pi_{\mu'} \exp\big(-\|\mathbf{z}_i - \mathbf{c}_{\mu'}\|_2^2/2\sigma_i^2\big)}.$$

The prior π_μ depends only on the taxonomy and σ_i^2 is shared: by Eq. (30), every exponent at numerator and denominator has the same value for the two predictors, over the same index set. The priors satisfy $\pi_{\varrho_i(\mu)} = \pi_\mu$, since $\tau_i \circ \varrho_i = \tau_i$ leaves the cardinality of the mode-lift unchanged, and σ_i^2 is shared. Hence by Eq. (30) the numerator of $p_{\theta'}(L_i = \varrho_i(\mu) \mid \mathbf{x})$ equals that of $p_\theta(L_i = \mu \mid \mathbf{x})$, while the two denominators coincide because ϱ_i is a bijection of $[m_i]$. Since the prototype layer predicts each IC from its own extractor, the joint posterior is obtained coordinate-wise and the equality lifts to $p_\theta(\mathbf{L} = \mathbf{m} \mid \mathbf{X}) = p_{\theta'}(\mathbf{L} = \varrho(\mathbf{m}) \mid \mathbf{X})$ with $\varrho = (\varrho_1, \ldots, \varrho_k)$. □

A.6 Identifiable class with Intended Concept Semantics - Theorem 4.6.

Remark. Recall we called the prototypical NeSy predictor p_θ *mode-aware* when, for every $i \in [k]$, it maintains a distinct prototype for each mode, so that its IC distribution separates the elements of each mode-lift $\tau_i^{-1}(c)$; in terms of the permutation of **[A3]**, this is the injectivity of every γ_i. On the other hand, it is called *mode-blind* when it maintains a single prototype per EC, that is when every γ_i factors through τ_i; this is the case of the unimodal predictors above. The two notions are not exhaustive: a predictor whose γ_i is neither injective nor constant on the mode-lifts is *partially* mode-aware, and Theorem 4.6 covers it together with the mode-blind case.

Statement. Given a dataset $\mathcal{D}$ and a support set $\mathcal{S}$ with complete mode coverage, let p_θ be a prototypical NeSy predictor attaining maximal log-likelihood on $\mathcal{D}$. Under **[A1,2,3]**, the predictor p_θ is mode-aware if and only if $[p_\theta]_{\mathcal{E}(\eta)}$ learns concepts with the intended semantics up to mode permutation.

Proof. Recall that p_θ learns concepts with the intended semantics whenever it holds that: $p_\theta(\mathbf{L} \mid \mathbf{X}) = p^\star(\varrho(\mathbf{M}) \mid \mathbf{X})$ (the Goal of Section 2), where ϱ is a family of modes permutation. Under **[A2]** every datapoint is written $\mathbf{x} = \phi_\tau(\mathbf{m}, \mathbf{s})$ with ϕ_τ invertible, so $\mathbf{m}$ is determined by $\mathbf{x}$ (Marconato et al., 2023b) and $p^\star(\mathbf{M} = \mathbf{m}' \mid \mathbf{x}) = \mathbb{1}(\mathbf{m}' = \mathbf{m})$ is a point mass. Since ϱ defines a bijection,

and $p^\star(\mathbf{M} = \mathbf{m}' \mid \mathbf{x}) = \mathbb{1}(\mathbf{m}' = \mathbf{m})$, it follows that $p^\star(\varrho(\mathbf{M}) = \mathbf{m}' \mid \mathbf{x}) = \mathbb{1}(\mathbf{m}' = \varrho(\mathbf{m}))$ also is a point mass. Finally, recall from Appendix A.3 that a deterministic optimum of $p_\theta(\mathbf{L} \mid \mathbf{X})$ is realised by the nearest-centroid assignment $\gamma\colon [m_i] \to [m_i]$, expressing the induced labelling of the modes under **[A3]**, and that Proposition 4.1 indexes the deterministic optima by such γ. Learning the intended semantics thus means assigning every datapoint to its own mode, corresponding to the ground-truth ones up to a permutation of their labels according to Definition 2.1. When τ restricted to $\texttt{supp}(\mathbf{M})$ is a bijection, every EC admits a single mode-lift and mode-blind and mode-aware predictors coincide: this is the unimodal regime of the control task in the experimental analysis, in which no IRS exists and the distinction we draw here is empty.

We prove separately the forward and converse implication of the claim.

($\Longrightarrow$) Mode-awareness implies intended semantics for the whole class.
Step 1: complete mode coverage makes γ a bijection. By Definition 4.2, $\mathcal{S}$ realises every mode, so for each $i \in [k]$ and each $\mu \in [m_i]$ there is a support datapoint of mode μ. Being mode-aware, p_θ maintains a distinct prototype for each mode, i.e. γ_i is injective by **[A3]**, and an injective self-map of the finite set $[m_i]$ is a bijection. Note that coverage is also necessary here: prototypes are created only from support datapoints, so without it $|\mathrm{im}(\gamma_i)| < m_i$ and no predictor is mode-aware.

Step 2: the EC supervision confines γ to the mode-lifts. Every support datapoint is EC-annotated, so complete mode coverage gives $\mathcal{S}_i^{\mathrm{sup}} = [m_i]$, and since p_θ attains maximal log-likelihood its assignment is one of the optima enumerated by Proposition 4.1: $\tau_i(\gamma_i(\mu)) = \tau_i(\mu)$ for every $\mu \in [m_i]$. Hence $\mathbf{C} = \tau(\mathbf{L})$ carries the ground-truth ECs and p_θ is free of ERS.

Step 3: the modes are recovered. By Steps 1 and 2 each γ_i is a permutation of $[m_i]$ with $\tau_i \circ \gamma_i = \tau_i$, that is a permutation of every mode-lift. Therefore $p_\theta(\mathbf{L} = \mathbf{m}' \mid \mathbf{x}) = \mathbb{1}(\mathbf{m}' = \gamma(\mathbf{m})) = p^\star(\gamma(\mathbf{M}) = \mathbf{m}' \mid \mathbf{x})$, and p_θ is free of IRS in the sense of Definition 2.1.

Step 4: the property holds for the entire Euclidean class. Let $p_{\theta'} \in [p_\theta]_{\mathcal{E}(\eta)}$. By Lemma 4.5, there exist a permutation ϱ for which $p_{\theta'}(\mathbf{L} \mid \mathbf{X}) = p_\theta(\varrho(\mathbf{L}) \mid \mathbf{X})$, and Step 3 applies for each member of the class.

($\Longleftarrow$) Intended semantics implies mode-awareness. We prove the contrapositive. If p_θ is not mode-aware then, by **[A3]**, γ_i is not injective for some $i \in [k]$: there are $\mu \neq \mu'$ in $[m_i]$ with $\gamma_i(\mu) = \gamma_i(\mu')$, so the datapoints of the two modes are assigned to one and the same prototype and the partition of $\mathcal{X}$ induced by $\mathbf{L}$ is strictly coarser than the one induced by $\mathbf{M}$. A coarser partition is not the image of a finer one under any permutation, so no permutation of the modes recovers the intended semantics and p_θ is affected by an IRS in the sense of Definition 2.1; by Lemma 4.5 the same holds for every predictor in $[p_\theta]_{\mathcal{E}(\eta)}$. Mode-blind predictors are the special case in which γ_i factors through τ_i, and they fail to be mode-aware exactly when the taxonomy is not a bijection; the claim for the case of partially mode-aware predictors is proved through the same argument. □

IMPLICIT REASONING SHORTCUTS IN MODE-BLIND MODELS - COROLLARY 4.7.

Statement. In the same settings of Theorem 4.6, assume the taxonomy to be *non-degenerate*, that is $|\tau_i^{-1}(c)| \geq 2$ for at least one $i \in [k]$ and one $c \in [h_i]$. Under **[A1,2,3]**, if p_θ is mode-blind then its set of deterministic optima has cardinality $\prod_{i\in[k]} \prod_{c\in[h_i]} |\tau_i^{-1}(c)|$, and every one of them is an IRS. In particular, no mode-blind predictor learns the intended concept semantics, whatever the mode coverage of $\mathcal{S}$.

Proof. Being mode-blind, p_θ maintains a single prototype per EC, built by Eq. (1) from the EC-c annotated support points, and the episode loss in Algorithm 1(c) is minimised exactly when every query point is nearest to the centroid of its own annotated EC. By **[A3]** the nearest prototype of a datapoint depends on it only through its mode, so at maximal log-likelihood the assignment factors through the taxonomy:

$$\gamma_i = \omega_i \circ \tau_i, \qquad \omega_i : [h_i] \to [m_i] \quad \text{with} \quad \tau_i(\omega_i(c)) = c, \tag{31}$$

one ω_i per coordinate i, selecting the mode that labels the prototype of each EC. Distinct families $(\omega_i)_{i\in[k]}$ give distinct assignments γ, and every family is optimal, since the ECs they induce are the

ground-truth ones. For each i there are $\prod_{c\in[h_i]} |\tau_i^{-1}(c)|$ choices of ω_i, which in turn gives

$$\#\{\text{deterministic optima of } p_\theta(\mathbf{L} \mid \mathbf{X})\} = \prod_{i\in[k]} \prod_{c\in[h_i]} |\tau_i^{-1}(c)|.$$

Each $\gamma_i = \omega_i \circ \tau_i$ is constant on every mode-lift, hence non-injective as soon as some lift contains two modes. Under the non-degeneracy assumption this happens for at least one i and one c, so no family $(\omega_i)_{i\in[k]}$ yields an injective γ, no permutation of the modes repairs the resulting partition, and all of the counted optima are IRSs in the sense of Definition 2.1. Since every counted γ is non-injective, by Theorem 4.6 none of them learns the intended semantics, which is therefore unattainable for a mode-blind predictor. □

B MOSAIC TRAINING DETAILS

MOSAIC training algorithm is shown in Algorithm 1. For every prototypical extractor f^i, Lines 2–6 carry out the mode-blind initialisation in Eq. (1): l_i ECs and a disjoint support/query split $(\hat{\mathcal{S}}_c, \mathcal{Q}_c)$ per EC are sampled. One initial mode μ is created per EC with $\mathbf{c}_\mu$ set to the mean embedding of $\hat{\mathcal{S}}_c$ and $\tau(\mu) = c$. The active set $\mathcal{A}_i$ is used to track every mode created , across both this initialisation and the following discovery step. Following Allen et al. (2019), Line 7 instantiates the threshold λ_i of Eq. (2): ρ is estimated as the empirical variance of the class means, and σ_i^2 is the (learned) cluster variance shared by all modes for the considered extractor i. Lines 9–16 then run mode discovery: for each support point $\mathbf{x}$ whose EC is c, we measure its distance to the *nearest* existing prototype carrying the label of the same EC only, and build a new mode centred at $\mathbf{z}_i = f_\theta^i(\mathbf{x})$ if and only if that distance exceeds λ_i. This is the Chinese-restaurant-process step (Aldous, 1985): λ_i plays the role of a data-driven "new table" threshold, controlled through the concentration hyperparameter υ, so the number of modes discovered per EC is inferred rather than fixed, and no mode-level annotation is ever used to decide where a new mode is placed. If at least one new mode was created (line 17), every support point's responsibility is recomputed as the normalised Gaussian density of $\mathbf{z}_i = f_\theta^i(\mathbf{x})$ under the modes of its own EC (line 21, c.f. Eq. (3)). Every mode centroid $\mathbf{c}_\mu$ for $\mu \in \mathcal{A}_i$ is then re-estimated as the responsibility-weighted mean of the support embeddings (lines 25–27). Lines 28–38 score the query set.

For a query point $\mathbf{x} \in \mathcal{Q}_c$ and every EC c' in the episode, $\mu^\star_{c'}(\mathbf{x})$ is the closest mode belonging to c': as in Allen et al. (2019), we adopt here *modes-masking* not to penalise having multiple modes per EC: the *episode* loss is formed, so that the loss only ever pulls $\mathbf{z}_i$ towards the mode that already explains $\mathbf{x}$ best, never towards every mode of its EC simultaneously. The resulting term is a softmax cross-entropy over ECs (c.f. Eq. (5)) with $-\|\mathbf{z}_i - \mathbf{c}_{\mu^\star_{c'}(\mathbf{x})}\|_2^2 / 2\sigma_i^2$ standing in for the logit of EC c'.

Finally, line 40 accounts the NeSy loss $\mathcal{L}^{\text{NeSy}}(\mathbf{y}, \mathrm{K})$ for every $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$, evaluated against the soft-marginal EC distribution of Eq. (4) rather than the arg-max modes used for the episode loss above: this is what lets gradients flow back into every mode of a wrongly-predicted EC in proportion to its current responsibility $\hat{y}_\mu$, exactly as characterised by Theorem 3.1.

C EXPERIMENTAL APPENDIX

C.1 TASK, MODELS AND METRICS DETAILS

The task `Omniglot-EvenOdd` is the generalisation of `MNIST-EvenOdd` from Bortolotti et al. (2024) to multi-modes, based on `Omniglot`(Lake et al., 2015). We will distribute the benchmark under the CC-BY-SA 4.0 license — as well as the code under a BSD 3 license — over GitHub in case of acceptance. `BDDOIA` is a task belonging to Bortolotti et al. (2024) distributed under the CC-BY-SA 4.0 license. We provide below the details about each task. The code for reproducing the results in this work has been submitted with the manuscript.

(i) `Omniglot-Add-EvenOdd` features input pair of handwritten Omniglot characters, labelled with their sum (task label): the EC is the digit 0–9 assigned to the alphabet a character belongs to, while the ICs are the characters realising it, so that every EC is spread over tens of visually disjoint modes. For example, given $\mathbf{x} = ($, $)$ the characters are from the alphabet `Angelic` and the one of the `Magi`, whose alphabet indexes and digit are 0 and 2. Thus the sum of the two characters is 2.

Algorithm 1 MOSAIC training episode loss. l_i is the number of classes sampled per set $[h_i]$ and episode; s_i (resp. q_i) the number of support (resp. query) examples per concept; λ the threshold in 2; σ_i^2 the shared cluster variance. Each mode μ carries a centroid $\mathbf{c}_\mu$ and a label $\tau(\mu) \in [h_i]$; $\mathcal{A}_i$ is the active set of modes for extractor i, tracking mode coverage.

Require: (i) Dataset $\mathcal{D}$, (ii) support set $\mathcal{S}_c$ for every concept c, (iii) knowledge K.
Ensure: Loss $\mathcal{L}$ for a randomly generated training episode.

$\mathcal{L} \leftarrow 0$
$\mathcal{E} \leftarrow \textsc{RandomSample}([h_i], l_i)$; $\mathcal{A}_i \leftarrow \emptyset$ ▷ Select labelled classes; init. mode set
for c in $\mathcal{E}$ **do** ▷ **(a) Create one initial mode per class**
  $\hat{\mathcal{S}}_c \leftarrow \textsc{RandomSample}(\mathcal{S}_c, s_i)$; $\mathcal{Q}_c \leftarrow \textsc{RandomSample}(\mathcal{S}_c \setminus \hat{\mathcal{S}}_c, q_i)$
  new mode μ: $\mathbf{c}_\mu \leftarrow \frac{1}{s_i}\sum_{\mathbf{x}\in\hat{\mathcal{S}}_c} f_\theta^i(\mathbf{x})$; $\tau(\mu) \leftarrow c$; $\mathcal{A}_i \leftarrow \mathcal{A}_i \cup \{\mu\}$
end for
Estimate λ as in 2 ▷ Infer the number of modes using infinite mixtures)
`created` $\leftarrow$ `False`
for c in $\mathcal{E}$ **do**
  for $\mathbf{x}$ in $\hat{\mathcal{S}}_c$ **do**
    $\mathbf{z}_i \leftarrow f_\theta^i(\mathbf{x})$
    if $\min_{\mu\in\tau^{-1}(c)} \|\mathbf{z}_i - \mathbf{c}_\mu\|_2^2 > \lambda$ **then** ▷ Threshold exceeded
      new mode $\mu^\star$: $\mathbf{c}_{\mu^\star} \leftarrow \mathbf{z}_i$; $\tau(\mu^\star) \leftarrow c$; $\mathcal{A}_i \leftarrow \mathcal{A}_i \cup \{\mu^\star\}$
      `created` $\leftarrow$ `True`
    end if
  end for
end for
if `created` **then**
  for c in $\mathcal{E}$ **do**
    for $\mathbf{x}$ in $\hat{\mathcal{S}}_c$ **do** ▷ **(b) Compute soft assignment and mode update**
      $\mathbf{z}_i \leftarrow f_\theta^i(\mathbf{x})$
      $z_{\mathbf{x},\mu} \leftarrow \dfrac{\mathcal{N}(f_\theta^i(\mathbf{x});\, \mathbf{c}_\mu, \sigma_i^2\mathbf{I})}{\sum_{\mu'\in\tau^{-1}(c)} \mathcal{N}(f_\theta^i(\mathbf{x});\, \mathbf{c}_{\mu'}, \sigma_i^2\mathbf{I})}$ for all $\mu \in \tau^{-1}(c)$
    end for
  end for
  for $\mu \in \mathcal{A}_i$ **do**
    $\mathbf{c}_\mu \leftarrow \big(\sum_{\mathbf{x}} z_{\mathbf{x},\mu}\, f_\theta^i(\mathbf{x})\big) \big/ \big(\sum_{\mathbf{x}} z_{\mathbf{x},\mu}\big)$
  end for
end if
for c in $\mathcal{E}$ **do** ▷ **(c) Compute masked query loss**
  for $\mathbf{x}$ in $\mathcal{Q}_c$ **do**
    for $c' \in \mathcal{E}$ **do**
      $\mu_{c'}^\star(\mathbf{x}) \leftarrow \arg\max_{\mu\in\tau^{-1}(c')} \left(-\frac{\|f_\theta^i(\mathbf{x}) - \mathbf{c}_\mu\|_2^2}{2\sigma_i^2}\right)$
    end for
    $\mathbf{z}_i \leftarrow f_\theta^i(\mathbf{x})$
    $\mathcal{L} \leftarrow \mathcal{L} + \dfrac{1}{l_i q_i}\left[\log \sum_{c'\in\mathcal{E}} \exp\left(-\frac{\|\mathbf{z}_i - \mathbf{c}_{\mu_{c'}^\star(\mathbf{x})}\|_2^2}{2\sigma_i^2}\right) + \frac{\|\mathbf{z}_i - \mathbf{c}_{\mu_c^\star(\mathbf{x})}\|_2^2}{2\sigma_i^2}\right]$ ▷ $\star$: closest class cluster
  end for
end for
for $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ **do**
  $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}^{\text{NeSy}}(\mathbf{y}, \text{K})$ ▷ **(d) NeSy training with soft marginalization over all clusters**
end for

As in MNIST−EvenOdd, only same-parity pairs are observed, which alone admits 49 ERSs (Marconato et al., 2023b), while the admissible IRSs follow from Proposition 4.1. We consider two configurations of increasing mode multiplicity: 10 alphabets, assigning one alphabet to each digit, and 40 alphabets. The latter split retains all Omniglot alphabets with $\geq$10 characters, mapping alphabets to digit labels via their alphabet digit modulo 10. In this way multiple alphabets collide onto each digit and the space of underlying character-level modes is far larger and more entangled than in the 10-alphabet split. Table 6 and Table 8 report the exact alphabet-to-digit assignment used in the 10- and 40-alphabet configurations, respectively. The list of all possible training combinations and their labels, under the 10-alphabet configuration, is given in Table 7. Note how the same EC (e.g. digit 6, resp. digit 9) is realised by a visually distinct character every time it recurs across pairs — unlike MNIST−EvenOdd, where a digit's appearance is fixed, this makes explicit that a mode-blind predictor cannot tell these instances apart from EC supervision alone. Overall, the training set contains 6720 data, the validation set 1920, and the test set 960.

Table 6: Alphabet-to-digit mapping used in the 10-alphabet Omniglot−EvenOdd split (Table 7).

Alphabet	Digit
Angelic	0
Grantha	1
N'Ko	2
Aurek-Besh	3
Japanese (hiragana)	4
Malay (Jawi – Arabic)	5
Asomtavruli (Georgian)	6
Sanskrit	7
Ojibwe Canadian Aboriginal Syllabics)	8
Korean	9

Table 7: Pairs of digits and their label available in Omniglot−EvenOdd (10-alphabet split). The alphabet realising each digit is fixed, but the specific character sampled to realise a recurring digit (here digits 6 and 9) changes across pairs, since each EC is spread over many IC-level modes.

Example Pair	Label	Example Pair	Label
(,)	6	(,)	6
(,)	10	(,)	10
(,)	10	(,)	10
(,)	12	(,)	12

Finally, we retain MNIST−EvenOdd(Bortolotti et al., 2024) itself as a unimodal control: there τ is a bijection and every EC admits a single mode-lift, so a mode-aware predictor is expected to recover its mode-blind counterpart.

(ii) BDDOIA (Xu et al., 2020) is an autonomous driving task built on ∼22.5k real-world frames from BDD-100k (Yu et al., 2018), where the goal is to correctly classify the reasons (concepts) behind the action taken by the ego-vehicle (task label, one or more of {*MoveForward*, *Stop*, *TurnLeft*, *TurnRight*}) among 21 possible concepts, related to the action and to each other through a background knowledge K of concept-to-concept and concept-to-action rules (e.g. *Person* $\vee$ *RedLight* $\rightarrow$ *Stop*). Frames are pre-processed into 2048-dimensional embeddings via a Faster-RCNN backbone (Ren et al., 2015), following (Sawada & Nakamura, 2022). Here multiple modes naturally occur: a single EC such as obstacle_on_the_right is realised by parked cars, cyclists, pedestrians or road barriers, and the corresponding ICs are latent and never annotated. The dataset is further challenging due to heavy concept class imbalance, label noise (e.g. frames occasionally labelled with the mutually-exclusive *MoveForward* and *Stop*), loss of temporal context between consecutive frames, and single frames sometimes lacking the surrounding context needed to tell whether a visible concept (e.g. a pedestrian) is actually the cause of the labelled action. We refer to the experimental Appendix C.2.5 of (Marconato et al., 2023b) for the complete specification of K and we provide

Table 8: Alphabet-to-digit mapping used in the 40-alphabet `Omniglot-EvenOdd` split. Digit = alphabet index modulo 10, where the alphabet index (shown in parentheses) follows order of first appearance in the Vinyals training split; only the 33 alphabets with $\geq$10 characters survive the split's filtering. Multiple alphabets collide onto the same digit, so every EC is realised by several visually disjoint modes at the alphabet level alone, on top of the many characters within each alphabet.

Digit	Colliding Alphabets (alphabet index)
0	*Angelic* (0), *Arcadian* (10), *Burmese* (20), *Tifinagh* (30)
1	*Grantha* (1), *Greek* (11), *Avesta* (21), *Latin* (31)
2	*N'Ko* (2), *Alphabet of the Magi* (12), *Gujarati* (22), *Inuktitut* (32)
3	*Aurek-Besh* (3), *Blackfoot* (13), *Ge'ez* (23)
4	*Japanese* (4), *Futurama* (14), *Syriac (24)*
5	*Malay* (5), *Gurmukhi* (15), *Atlantean (25)*
6	*Asomtavruli* (6), *Tagalog* (16), *Japanese (katakana)* (26)
7	*Sanskrit* (7), *Anglo-Saxon Futhorc* (17), *Balinese* (27)
8	*Ojibwe* (8), *Braille* (18), *Atemayar Qelisayer* (28)
9	*Korean* (9), *Cyrillic* (19), *Glagolitic* (29)

here an example:

$$\begin{aligned}&\mathit{RedLight} \rightarrow \neg \mathit{GreenLight},\\ &\neg \mathit{Car} \vee \mathit{Person} \vee \mathit{Rider} \vee \mathit{OtherObstacle} \leftrightarrow \mathit{RoadClear},\\ &\mathit{GreenLight} \vee \mathit{Follow} \vee \mathit{RoadClear} \rightarrow \mathit{MoveForward}.\end{aligned}$$

Models. For all tasks, we consider DeepProbLog (DPL) (Manhaeve et al., 2018), Semantic Loss (SL) (Xu et al., 2018), Scallop (Li et al., 2023) and CTSketch (Choi et al., 2025) and coherent-by-construction CCN+ (Giunchiglia et al., 2024). We integrate each of them with MOSAIC and compare against (i) the same inference layers equipped with mode-blind prototypes (PNet) (Andolfi & Giunchiglia, 2025), and (ii) against the mitigation strategies of Marconato et al. (2023b). These are: reconstruction (R), entropy regularisation (E) concept supervision (C), concept pretraining (PRE). To assess whether sheer encoder capacity can substitute for mode awareness, we further equip the mode-blind prototypes with a ResNet12 backbone (He et al., 2016), while MOSAIC retains the standard four-block convolutional extractor of Snell et al. (2017).

Metrics. Following Andolfi & Giunchiglia (2025), we measure accuracy and F1-score wrt. both final labels and ECs, denoted resp. as Acc(Y), F1(Y), Acc(C) and F1(C), together with the concept collapse Cls(C), i.e. the fraction of ECs that the predictor never assigns. In addition, for the `BDDOIA` task we also measure the *weighted recall* on the ECs. This metrics is crucial in an autonomous-driving setting: missing the concept that should have triggered a *Stop* is a costlier error than a false positive. On `Omniglot-EvenOdd` we additionally report two controls: a random classifier, predicting each digit uniformly at random and independent of the input, as a chance-level floor; and Infinite Mixture Prototypes Allen et al. (2019) trained on EC-labeled episodes only (i.e. Line 1 to 36 in Algorithm 1) to isolate the contribution of the symbolic loss expressing the background knowledge. Finally, to directly probe mode-level behaviour on `Omniglot-EvenOdd` we report *mode purity* and *IRS-rate* — the fraction of held-out query points whose nearest prototype agrees with its true character identity, and the fraction of character-level modes whose empirical image under $\hat{\gamma}_i$ collides with another mode's (Theorem 3.1), respectively — and on `BDDOIA`, where no ground-truth mode identity exists, the analogous *mode usage* statistics (fraction of maintained prototypes actually used, and the entropy of that usage) as a proxy.

C.2 Experimental Setup

Neural Architectures As backbone concept extractors for the baseline NeSy models we use the architecture from Bortolotti et al. (2024). As prototypical extractor for `Omniglot-EvenOdd` (and, as a unimodal control, `MNIST-EvenOdd`) we use the same four-block convolutional network as in (Snell et al., 2017): a first 1×64 block followed by three 64×64 blocks, each a 3×3 convolution, batch normalisation (Ioffe & Szegedy, 2015), ReLU, and 2×2 max-pooling, yielding

a 64-dimensional embedding. One such extractor is shared across both pairing coordinates $i = 1$ and $i = 2$. For the encoder-capacity control (Section 5) we substitute it with a ResNet12 (He et al., 2016), leaving MOSAIC's extractor unchanged.

For `BDD-OIA`, each frame is processed following (Sawada & Nakamura, 2022) into a 2048-dimensional embedding via a Faster-RCNN backbone pre-trained on MS-COCO and fine-tuned on BDD-100k (Yu et al., 2018). We employ 21 separate prototypical extractors, one per concept, each a single linear layer mapping the 2048-dimensional embedding to a 256-dimensional space, followed by batch normalisation and ReLU.

Hyperparameters All `Omniglot-EvenOdd` experiments are run across ten seeds (0, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768); `BDDOIA` across five seeds (0–4). Prototype extractors are trained with RMSprop (lr $= 10^{-3}$, $\alpha = 0.9$, eps $= 10^{-10}$, no momentum) and a step learning-rate schedule (halved every 10 epochs), shared by both tasks. `Omniglot-EvenOdd` trains for up to 50 epochs of 1000 episodes each; `BDD-OIA` for up to 40 epochs of 100 episodes each, with early stopping (patience 15). The CRP concentration υ (Eq. 2) is 0.1 for `Omniglot-EvenOdd` and 0.75 for `BDD-OIA`. Per-concept encoders on `BDD-OIA` use Adam with the same step schedule.

For `Omniglot-EvenOdd`, mitigations are trained with $\mathcal{L} = \mathcal{L}_{\text{cls}} + \gamma\,(w_{\text{rec}}\mathcal{L}_{\text{rec}} + w_{\text{e}}\mathcal{L}_{\text{e}} + w_{\text{c}}\mathcal{L}_{\text{c}})$ across all five NeSy baselines (DeepProbLog, Semantic Loss, CTSketch, Scallop, CCN+), each with a corresponding reconstruction variant R; PRE composes independently of the loss and adds encoder pretraining on the same support set pool constructed for the prototypical NeSy models. This is a 2-shot support pool and pretraining proceeds with Adam, lr $= 10^{-3}$, weight decay 10^{-4}, `ReduceLROnPlateau` with patience 3/factor 0.5 and early stopping patience 5. DeepProbLog, CTSketch, and Scallop share coefficients $\beta_{\text{KL}} = 2$, $w_{\text{rec}} = 1$, $w_{\text{e}} = 0.5$, $w_{\text{c}} = 1$ and the tuned schedule $(\text{lr}, \gamma) = (10^{-3}, 5)$ for the baseline, $(5\times10^{-3}, 1)$ for R, $(5\times10^{-4}, 1)$ for E, $(5\times10^{-3}, 0.1)$ for C, $(1 \times 10^{-3}, 2)$ for C+E, $(5 \times 10^{-3}, 1)$ for R+E, $(1 \times 10^{-3}, 0.1)$ for R+C, and $(5 \times 10^{-3}, 1)$ for R+C+E. Semantic Loss and CCN+ instead share $\beta_{\text{KL}} = 0.1$, $w_{\text{rec}} = 0.1$, $w_{\text{e}} = 2$, $w_{\text{c}} = 5$ and the schedule $(\text{lr}, \gamma) = (10^{-3}, 5)$ for the baseline, $(5 \times 10^{-3}, 1)$ for R, E, and C individually, and $(5 \times 10^{-3}, 2)$ for every two- and three-term combination (C+E, R+E, R+C, R+C+E). Concept supervision C uses $c_{\text{sup}} = 1.0$ on digits $\{4, 9\}$ (a 6.33% annotation budget) for all five backbones. For `BDD-OIA`, mitigations on both the `DPL` and `CCN+` share $n_{\text{epochs}} = 40$, early stopping patience 15 (selection metric: binary F1), Adam with lr $= 10^{-3}$, batch size 64, an exponential decay schedule (exp_decay $= 0.9$), and a uniform mitigation weight $\gamma = 1$ with KL weight $\beta = 0.99$; negative concept examples are sampled at the dataset's natural per-concept imbalance ratio with no capping. The same mitigation definitions apply to both backbones: C applies concept supervision with $c_{\text{sup}} = 0.5$ restricted to 4 concepts (green light, red light, car, pedestrian); C(S) applies that same supervised loss only on the fixed support-set pool, reapplied every epoch; PRE(S) pretrains the encoder on the support pool for up to 50 epochs (patience 5) once before training under the distant symbolic supervision; E sets the entropy weight $w_{\text{e}} = 1$ (vs. $w_{\text{e}} = 0$ when absent).

Computing Infrastructure All the experiments run on a machine equipped with the following:

- AMD EPYC 9334 32-Core Processor, 32 Cores, 64 Threads (per socket, 2 sockets, 128 threads total);
- 3 NVidia RTX 4500 Ada Generation GPUs, 24 GB GDDR6 each.

whose operating system is Red Hat Enterprise Linux 9.8 (Plow) and CUDA Version 12.4. For realizing the code of our experiments we used Python 3.11.15 and PyTorch 2.6.0.

D Additional Experiments

D.1 RQ.1: Can MOSAIC avoid Explicit and Implicit Reasoning Shortcuts?

Tables 9–18 report the complete `Omniglot-EvenOdd` results for all five NeSy reasoning layers on both the 10 and 40 alphabets splits; Tables 19–23 are the complete `BDDOIA` results, and Tables 24–26 the mode-level diagnostics. Below dashed line we report two references: a random classifier, and Infinite Mixture Prototypes (IMP) (Allen et al., 2019) trained on the EC-labelled episodes alone,

i.e. MOSAIC's prototypical extractor without the distant symbolic supervision provided thorugh the background knowledge.

Starting with `Omniglot-EvenOdd`, we can observe that on the 10-alphabet split MOSAIC attains Acc(C) in the range $[89.44\%, 90.00\%]$ and F1(C) in $[86.98\%, 87.22\%]$, with per-seed maximum standard deviation 0.59%; on the 40-alphabet split, Acc(C) ranges over $[71.71\%, 71.94\%]$ and F1(C) over $[70.51\%, 70.76\%]$. Replacing the inference layer — exact (DPL, CCN+) for approximate (SL, Scallop, CTSketch) — moves F1(C) by at most 0.24 and 0.25 points respectively, i.e. by less than the seed-level standard deviation of any baseline in the same table. This is the empirical counterpart of Theorem 4.6: once the mode-lift is identified, which ECs are recovered does not depend on how K is relaxed. However, modifying the inference layer enforcing K moves the unmitigated baselines' Acc(Y) between 55.13% and 76.34%: therefore the way reasoning is implemented is decisive for task performance, which is what the symbolic objective is designed to optmise against.

Now, with reference to the random classifier — predicting each EC uniformly and independently of the input — *every* unmitigated baseline is below it on F1(C) in both splits (2.74%–6.24%), and four of five are below it on Acc(C) in the 10-alphabet split. A NeSy model trained by distant supervision alone is thus strictly less informative than chance, because it concentrates its mass on one to three ECs (as reflected by the concept collapse Cls(C) $= 0.70\%$–0.90%). This reference baseline also shows that zero collapse is necessary but far from sufficient — the random classifier has Cls(C) $= 0.00\%$ as well — and that only the prototypical predictors pair it with above-chance concept accuracy.

On the other hand, looking at the final label scores, F1(Y) is averaged over $\mathcal{Y}_\cup$, the union of sum values occurring as either true or predicted label. Every $y \in \mathcal{Y}_\cup$ that is never a true label has no positives and hence contributes exactly 0, so, writing n for the number of sum values actually realised in the test set, $\mathrm{F1}(Y) \leq n/|\mathcal{Y}_\cup|$ *whatever* the accuracy. In `Omniglot-EvenOdd` the alphabet-to-digit map (Table 6) together with the same-parity constraint leaves only the sums $\{6, 10, 12\}$ realisable (Table 7), while all 19 remain predictable: each spurious value a model predicts even once enters the denominator and adds a zero to the average. The random control makes the size of the artefact concrete — moving from chance to MOSAIC raises Acc(Y) by 72.3% points ($7.96\% \to 80.22\%$) but F1(Y) by only 12.4% ($1.89\% \to 14.30\%$), a factor of $\approx 6\%$ between the two metrics' absolute gains on the very same predictions. On `MNIST-EvenOdd` (Table 30) MOSAIC recovers 97.68% of the concepts and 95.40% of the sums exactly, and still records F1(Y) 20.76%. On these benchmarks F1(Y) measures the support of the label distribution rather than the quality of the predictor; we report it only for comparability with Andolfi & Giunchiglia (2025).

Finally the IMP reference in `Omniglot-EvenOdd` isolates the contribution of K during training: the comparison shows that MOSAIC training under the *distant supervision alone* does not reintroduce shortcut reasoning: relative to IMP over the 10-alphabets split, MOSAIC gains $+1.37\%$ to $+1.93\%$ Acc(C), $+1.32\%$ to $+1.56\%$ F1(C), $+2.37\%$ to $+2.77\%$ Acc(Y) and $+0.15\%$ to $+0.53\%$ F1(Y), at unchanged Cls(C) $= 0.00\%$. Hence, first, concept fidelity is never traded for task performance: the distantly-supervised phase leaves the mode structure discovered during episodic training intact, and MOSAIC ends up *above* the purely few-shot learner on every concept metric. Second, the gain is largest on F1(Y): the symbolic loss spends its capacity re-positioning the prototypes for label prediction, which is what it is designed for, while the EC-level solution do not change. Similarly over the 40-alphabets split, MOSAIC gains wrt. IMP $+5.22$ to $+5.45\%$ Acc(C), $+5.45\%$ to $+5.70\%$ F1(C), $+7.11\%$ to $+7.45\%$ Acc(Y) and $+1.06\%$ to $+1.10\%$ F1(Y), again at unchanged Cls(C) $= 0.00\%$ for both. The gap is roughly 3–4$\times$ larger than on the 10-alphabet split across every metric. This is because the 40-alphabet configuration exhibits far more character-level modes under each EC (Table 8), so there is more room for the mode-aware embedding update of Theorem 3.1 to refine the representativeness of the prototypes in the embedding space.

Tables 24 and 25 probe the nearest-prototype map directly. MOSAIC reaches within-EC purity 80.98% (10 alphabets) and 69.32% (40 alphabets) against PNet's 2.93% and 1.00%, and IRS-rate 2.94% and 7.14% against 97.14% and 99.03%. PNet's measured IRS-rate coincides with the floor $1 - h_i/m$ predictable from Corollary 4.7, so its mode collision is structural rather than an optimisation failure: no training signal can lift it while a single prototype per EC is maintained. Note also that MOSAIC's global and within-EC purities differ by only 3.29 points (77.69% vs. 80.98%): the few mode-level mistakes it does make stay inside the correct EC, i.e. they are IRs and not ERSs. For PNet the two coincide. On `BDDOIA`, where no ground-truth mode identity exists, Table 26 reports

the analogous usage statistics: MOSAIC's held-out mass is spread across the maintained prototypes (top-1 share 32.29%, entropy ratio 73.68%), whereas PNet's $(100\%, 0\%)$ is pinned by construction and carries no mode-level information.

Table 9: **Omniglot-EvenOdd, DPL, 10-alphabet split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using DPL as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset .

Model/Mitigation	Acc(C)%↑	F1(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
Baseline	8.85 ± 3.17	3.09 ± 0.91	56.47 ± 6.49	**35.10 ± 5.12**	0.71 ± 0.03
+C	15.40 ± 3.77	5.19 ± 2.30	49.27 ± 6.03	29.28 ± 3.52	0.73 ± 0.07
+E	14.56 ± 1.75	3.01 ± 0.50	48.99 ± 5.31	32.32 ± 5.75	0.73 ± 0.05
+PRE	19.65 ± 3.11	8.40 ± 1.57	57.29 ± 5.24	29.27 ± 3.15	0.71 ± 0.03
+C+E	24.75 ± 5.26	12.35 ± 4.16	25.11 ± 4.09	7.78 ± 0.54	0.47 ± 0.13
+C+PRE	19.88 ± 4.49	8.56 ± 2.24	55.40 ± 11.03	25.70 ± 4.27	0.69 ± 0.07
+E+PRE	21.51 ± 1.24	9.70 ± 0.82	53.97 ± 2.85	25.24 ± 2.29	0.65 ± 0.05
+C+E+PRE	57.72 ± 9.98	48.19 ± 9.15	38.98 ± 9.85	7.97 ± 1.67	0.01 ± 0.03
+R	14.33 ± 1.90	2.50 ± 0.30	51.81 ± 11.76	22.45 ± 4.26	0.90 ± 0.00
+R+C	16.00 ± 1.65	4.42 ± 1.86	48.09 ± 5.68	31.09 ± 5.30	0.76 ± 0.10
+R+E	14.33 ± 1.90	2.50 ± 0.30	55.53 ± 0.00	23.80 ± 0.00	0.90 ± 0.00
+R+PRE	15.99 ± 2.23	6.52 ± 1.68	59.97 ± 3.47	33.03 ± 4.85	0.75 ± 0.05
+R+C+E	14.56 ± 0.00	2.54 ± 0.00	55.53 ± 0.00	23.80 ± 0.00	0.90 ± 0.00
+R+C+PRE	22.69 ± 1.10	10.11 ± 0.46	52.63 ± 1.63	24.56 ± 2.93	0.68 ± 0.04
+R+E+PRE	17.21 ± 3.12	7.25 ± 1.66	59.30 ± 5.48	32.42 ± 2.81	0.73 ± 0.05
+R+C+E+PRE	16.12 ± 4.33	3.67 ± 2.74	51.68 ± 4.57	18.33 ± 3.21	0.85 ± 0.07
+PNet	72.00 ± 0.01	68.00 ± 0.01	53.00 ± 0.02	10.00 ± 0.00	**0.00 ± 0.00**
+ResNet-PNet	64.39 ± 1.31	60.88 ± 1.31	42.14 ± 1.52	8.55 ± 0.28	**0.00 ± 0.00**
+MOSAIC (ours)	**90.00 ± 0.00**	**87.00 ± 0.01**	**80.00 ± 0.01**	14.00 ± 0.00	**0.00 ± 0.00**
IMP	88.07 ± 0.49	85.66 ± 0.58	77.61 ± 0.81	13.85 ± 0.35	**0.00 ± 0.00**
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	**0.00 ± 0.00**

D.2 RQ.2: How does MOSAIC compare with other mitigation strategies?

Tables 9–18 and 22–23 report every mitigation and combination thereof, including *reconstruction loss* (R), omitted from Table 2 for space.

On `Omniglot-EvenOdd`, R lowers F1(C) with respect to the corresponding R-free variant in almost every metric and drives Cls(C) to 0.90% — nine of ten ECs never predicted — at *zero* variance across ten seeds: DPL +R and +R+E both report Acc(Y) 51.81%/55.53% and F1(Y) 22.45%/23.80% with standard deviation 0.00%, and Scallop and CTSketch behave identically.

R is only useful as a third or fourth term on SL (+R+C+E+PRE, F1(C) 43.07%, +5.60% over C+E+PRE), where it acts as a regulariser on an already heavily supervised objective. The only configuration that pushes F1(C) past 45% on `Omniglot-EvenOdd` is the full C+E+PRE stack, and on DPL, Scallop and CTSketch it pays for it on the final task labels: DPL moves F1(C) 3.09% → 48.19% while Acc(Y) falls 56.47% → 38.98% and F1(Y) 35.10% → 7.97%; Scallop 3.63% → 52.92% with Acc(Y) 55.13% → 45.08%; CTSketch 3.36% → 54.24% with Acc(Y) 56.23% → 44.47%. The tension lies in the fact that the only way a mode-blind model can spread its EC predictions is to abandon the shortcut it was using to fit $\mathbf{Y}$. MOSAIC is the only entry in these tables that is simultaneously best on F1(C) (86.98%–87.22%) and best on Acc(Y) (79.98%–80.38%). SL is the single exception where mitigations improve both axes (+R+C+E+PRE: F1(C) 43.07%, Acc(Y) 82.15%, the highest Acc(Y) in the 10-alphabet tables), yet it reaches half of MOSAIC's F1(C) and 55.83% Acc(C) against 90.00%, while consuming the 6.33% dense EC annotation budget of C (all images of digits 4 and 9) plus reconstruction plus encoder pretraining, against MOSAIC's two labelled images per character.

Table 10: **Omniglot-EvenOdd, DPL, 40 alphabets split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using DPL as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 6.33% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%↑**	**F1(C)%↑**	**Acc(Y)%↑**	**F1(Y)%↑**	**Cls(C)%↓**
Baseline	14.89 ± 2.44	3.62 ± 1.47	38.72 ± 6.60	26.75 ± 3.58	0.76 ± 0.07
+C	12.05 ± 3.44	2.56 ± 1.13	37.75 ± 5.83	21.52 ± 4.57	0.80 ± 0.09
+E	14.78 ± 2.39	3.37 ± 1.15	41.72 ± 4.83	$\mathbf{28.29 \pm 2.92}$	0.79 ± 0.10
+PRE	17.09 ± 0.39	6.50 ± 0.72	44.04 ± 3.47	22.77 ± 1.07	0.70 ± 0.00
+C+E	29.81 ± 1.62	21.53 ± 1.79	19.14 ± 1.08	4.69 ± 0.15	0.16 ± 0.05
+C+PRE	13.03 ± 3.97	4.43 ± 1.70	42.19 ± 6.65	22.94 ± 2.50	0.70 ± 0.05
+E+PRE	17.04 ± 0.23	6.56 ± 0.44	44.08 ± 2.03	22.48 ± 0.27	0.67 ± 0.05
+C+E+PRE	31.14 ± 1.47	23.73 ± 1.67	18.67 ± 0.53	4.41 ± 0.12	0.06 ± 0.05
+R	15.39 ± 0.00	2.67 ± 0.00	18.12 ± 0.00	10.23 ± 0.00	0.90 ± 0.00
+R+C	13.10 ± 4.05	4.01 ± 1.95	41.33 ± 4.44	24.46 ± 4.01	0.77 ± 0.09
+R+E	15.39 ± 0.00	2.67 ± 0.00	18.12 ± 0.00	10.23 ± 0.00	0.90 ± 0.00
+R+PRE	14.37 ± 3.37	3.81 ± 1.47	39.49 ± 6.00	22.86 ± 3.05	0.71 ± 0.03
+R+C+E	10.76 ± 0.00	1.94 ± 0.00	14.49 ± 23.34	6.52 ± 10.49	0.90 ± 0.00
+R+C+PRE	17.91 ± 0.54	7.24 ± 0.79	44.04 ± 2.77	18.16 ± 3.57	0.64 ± 0.05
+R+E+PRE	11.96 ± 4.21	3.00 ± 1.75	45.58 ± 5.75	24.25 ± 3.82	0.72 ± 0.04
+R+C+E+PRE	22.17 ± 1.89	14.35 ± 1.80	16.88 ± 2.08	5.92 ± 0.45	0.50 ± 0.00
+PNet	35.24 ± 0.75	32.80 ± 0.72	15.07 ± 0.41	3.68 ± 0.09	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	44.21 ± 0.93	42.72 ± 0.96	21.69 ± 0.75	5.07 ± 0.14	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{71.83 \pm 0.42}$	$\mathbf{70.71 \pm 0.44}$	$\mathbf{52.22 \pm 0.58}$	10.37 ± 0.09	$\mathbf{0.00 \pm 0.00}$
IMP	66.49 ± 0.46	65.06 ± 0.46	44.91 ± 0.58	9.28 ± 0.09	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 11: **Omniglot-EvenOdd, SL, 10-alphabet split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using SL as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%**↑	**F1(C)%**↑	**Acc(Y)%**↑	**F1(Y)%**↑	**Cls(C)%**↓
Baseline	8.10 ± 3.44	3.06 ± 1.65	76.34 ± 2.07	72.46 ± 2.62	0.70 ± 0.00
+C	23.81 ± 1.82	12.31 ± 2.67	79.60 ± 1.50	74.94 ± 2.52	0.54 ± 0.05
+E	11.66 ± 6.05	4.54 ± 2.83	77.89 ± 1.84	73.84 ± 2.96	0.70 ± 0.00
+PRE	20.60 ± 1.34	9.57 ± 2.09	69.24 ± 2.67	63.68 ± 5.01	0.69 ± 0.03
+C+E	43.56 ± 7.14	31.07 ± 7.54	81.88 ± 1.19	78.43 ± 1.80	0.28 ± 0.11
+C+PRE	32.63 ± 5.35	22.75 ± 7.02	81.01 ± 2.31	77.13 ± 2.77	0.42 ± 0.12
+E+PRE	17.64 ± 2.74	7.53 ± 1.61	70.04 ± 3.73	64.66 ± 6.64	0.71 ± 0.03
+C+E+PRE	51.40 ± 9.42	37.47 ± 9.26	81.83 ± 1.17	78.41 ± 1.63	0.18 ± 0.09
+R	11.68 ± 5.95	4.71 ± 3.16	77.47 ± 1.87	73.68 ± 2.33	0.70 ± 0.00
+R+C	42.12 ± 10.46	28.53 ± 9.08	81.33 ± 1.02	77.88 ± 1.38	0.38 ± 0.11
+R+E	12.29 ± 6.35	5.39 ± 4.07	77.27 ± 3.55	73.81 ± 4.27	0.63 ± 0.07
+R+PRE	16.51 ± 3.41	6.95 ± 1.63	71.01 ± 2.64	67.82 ± 2.30	0.70 ± 0.00
+R+C+E	47.08 ± 6.59	33.29 ± 6.43	81.98 ± 0.83	78.66 ± 1.34	0.28 ± 0.06
+R+C+PRE	46.40 ± 3.83	35.93 ± 4.15	81.95 ± 1.08	78.85 ± 1.27	0.23 ± 0.07
+R+E+PRE	17.97 ± 3.45	7.72 ± 1.67	71.00 ± 2.99	66.73 ± 5.54	0.70 ± 0.00
+R+C+E+PRE	55.83 ± 7.25	43.07 ± 7.20	**82.15 ± 1.05**	**79.09 ± 1.08**	0.14 ± 0.05
+PNet	71.00 ± 0.01	67.00 ± 0.01	50.00 ± 0.01	10.00 ± 0.00	**0.00 ± 0.00**
+ResNet-PNet	78.42 ± 1.27	74.71 ± 1.27	61.82 ± 1.88	11.78 ± 0.40	**0.00 ± 0.00**
+MOSAIC (ours)	**90.00 ± 0.00**	**87.00 ± 0.00**	80.00 ± 0.01	14.00 ± 0.00	**0.00 ± 0.00**
IMP	88.07 ± 0.49	85.66 ± 0.58	77.61 ± 0.81	13.85 ± 0.35	**0.00 ± 0.00**
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	**0.00 ± 0.00**

Table 12: **Omniglot-EvenOdd, SL, 40 alphabets split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using SL as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 6.33% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%↑**	**F1(C)%↑**	**Acc(Y)%↑**	**F1(Y)%↑**	**Cls(C)%↓**
Baseline	16.04 ± 3.27	6.24 ± 1.55	57.57 ± 1.62	49.46 ± 4.19	0.70 ± 0.00
+C	16.02 ± 2.63	7.71 ± 2.13	56.36 ± 1.33	46.22 ± 2.69	0.57 ± 0.07
+E	11.39 ± 4.75	2.71 ± 1.44	50.28 ± 2.16	30.80 ± 8.19	0.79 ± 0.10
+PRE	17.16 ± 0.11	6.69 ± 0.19	56.85 ± 0.64	49.25 ± 1.17	0.70 ± 0.00
+C+E	23.55 ± 2.37	15.65 ± 2.15	56.73 ± 1.97	49.14 ± 3.24	0.33 ± 0.12
+C+PRE	16.07 ± 0.87	9.12 ± 1.51	58.95 ± 2.21	52.77 ± 3.09	0.51 ± 0.03
+E+PRE	15.52 ± 2.89	5.15 ± 1.27	54.08 ± 1.19	43.28 ± 2.31	0.70 ± 0.00
+C+E+PRE	27.74 ± 1.36	19.86 ± 1.41	$\mathbf{60.12 \pm 1.09}$	$\mathbf{55.25 \pm 1.76}$	0.22 ± 0.04
+R	11.19 ± 4.16	2.93 ± 1.40	50.00 ± 1.48	31.37 ± 6.80	0.76 ± 0.10
+R+C	21.45 ± 1.67	13.22 ± 2.06	57.12 ± 1.49	49.23 ± 2.40	0.51 ± 0.03
+R+E	11.80 ± 4.23	2.80 ± 1.43	49.86 ± 2.01	29.37 ± 8.24	0.81 ± 0.10
+R+PRE	15.68 ± 2.90	5.21 ± 1.35	54.72 ± 1.72	43.46 ± 2.58	0.70 ± 0.00
+R+C+E	24.49 ± 1.54	16.42 ± 1.61	57.68 ± 1.38	50.29 ± 2.36	0.34 ± 0.08
+R+C+PRE	24.15 ± 0.88	16.26 ± 0.89	59.82 ± 1.05	54.82 ± 1.45	0.44 ± 0.07
+R+E+PRE	15.52 ± 2.89	5.18 ± 1.29	54.25 ± 1.07	42.17 ± 1.75	0.70 ± 0.00
+R+C+E+PRE	27.74 ± 1.10	19.67 ± 0.95	59.64 ± 0.65	54.79 ± 1.03	0.22 ± 0.06
+PNet	45.59 ± 0.51	43.42 ± 0.48	22.49 ± 0.46	5.28 ± 0.10	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	64.87 ± 1.12	63.86 ± 1.17	42.94 ± 1.38	8.92 ± 0.22	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{71.81 \pm 0.37}$	$\mathbf{70.60 \pm 0.36}$	52.18 ± 0.48	10.36 ± 0.07	$\mathbf{0.00 \pm 0.00}$
IMP	66.49 ± 0.46	65.06 ± 0.46	44.91 ± 0.58	9.28 ± 0.09	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 13: **Omniglot-EvenOdd, CCN+, 10-alphabet split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using CCN+ as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%↑**	**F1(C)%↑**	**Acc(Y)↑**	**F1(Y)%↑**	**Cls(C)%↓**
Baseline	11.29 ± 3.41	3.03 ± 1.00	63.78 ± 8.42	35.65 ± 9.69	0.76 ± 0.05
+C	24.44 ± 6.74	8.10 ± 3.72	38.57 ± 6.29	18.29 ± 3.37	0.67 ± 0.10
+E	13.28 ± 3.80	4.59 ± 2.57	58.35 ± 12.51	38.20 ± 13.41	0.55 ± 0.29
+PRE	12.86 ± 6.04	4.76 ± 3.31	65.44 ± 6.71	$\mathbf{46.75 \pm 9.33}$	0.75 ± 0.05
+C+E	23.17 ± 7.10	8.96 ± 4.69	46.36 ± 8.15	19.90 ± 2.67	0.54 ± 0.22
+C+PRE	29.72 ± 8.45	10.88 ± 4.92	32.67 ± 6.78	16.79 ± 1.84	0.68 ± 0.10
+E+PRE	12.94 ± 3.16	3.81 ± 2.76	53.05 ± 10.01	27.80 ± 10.90	0.81 ± 0.15
+C+E+PRE	28.03 ± 6.15	10.84 ± 4.30	30.70 ± 7.19	12.84 ± 3.07	0.59 ± 0.24
+R	11.22 ± 4.00	2.92 ± 1.10	59.43 ± 9.43	34.01 ± 13.41	0.72 ± 0.06
+R+C	16.90 ± 2.73	4.91 ± 1.35	46.37 ± 10.01	14.22 ± 5.08	0.76 ± 0.12
+R+E	15.79 ± 4.85	6.07 ± 4.18	60.27 ± 4.84	29.19 ± 9.32	0.48 ± 0.35
+R+PRE	9.23 ± 0.53	2.60 ± 0.99	69.45 ± 2.55	45.35 ± 6.19	0.75 ± 0.05
+R+C+E	27.92 ± 6.39	12.41 ± 3.99	43.12 ± 6.64	18.83 ± 2.83	0.33 ± 0.26
+R+C+PRE	28.97 ± 8.51	10.72 ± 4.86	31.27 ± 7.40	15.83 ± 3.65	0.72 ± 0.06
+R+E+PRE	12.56 ± 4.52	3.83 ± 5.51	56.43 ± 2.71	25.75 ± 5.85	0.81 ± 0.27
+R+C+E+PRE	29.29 ± 7.11	11.14 ± 4.54	30.54 ± 6.54	13.82 ± 3.20	0.63 ± 0.16
+PNet	69.12 ± 0.92	65.12 ± 1.10	48.38 ± 1.21	9.65 ± 0.33	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	78.87 ± 0.89	75.14 ± 1.00	62.49 ± 1.35	12.10 ± 0.36	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{89.58 \pm 0.47}$	$\mathbf{87.12 \pm 0.45}$	$\mathbf{80.22 \pm 0.88}$	14.30 ± 0.25	$\mathbf{0.00 \pm 0.00}$
IMP	88.07 ± 0.49	85.66 ± 0.58	77.61 ± 0.81	13.85 ± 0.35	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 14: **Omniglot-EvenOdd, CCN+, 40 alphabets split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using CCN+ as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 6.33% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%**↑	**F1(C)%**↑	**Acc(Y)%**↑	**F1(Y)%**↑	**Cls(C)%**↓
Baseline	9.76 ± 1.92	2.74 ± 1.38	48.81 ± 3.91	$\mathbf{30.58 \pm 8.81}$	0.77 ± 0.06
+C	15.55 ± 4.98	5.28 ± 1.23	7.79 ± 7.30	2.54 ± 0.99	0.56 ± 0.20
+E	10.70 ± 2.86	2.78 ± 0.89	46.72 ± 4.40	20.69 ± 2.05	0.80 ± 0.12
+PRE	12.88 ± 2.87	4.21 ± 1.33	46.76 ± 3.96	27.11 ± 2.25	0.69 ± 0.08
+C+E	18.12 ± 1.52	6.37 ± 0.80	8.45 ± 5.18	4.33 ± 0.43	0.36 ± 0.25
+C+PRE	23.97 ± 4.52	16.32 ± 9.58	8.30 ± 2.47	3.34 ± 0.92	0.34 ± 0.35
+E+PRE	16.20 ± 5.06	7.19 ± 6.53	43.16 ± 11.58	19.33 ± 7.91	0.62 ± 0.27
+C+E+PRE	19.95 ± 1.39	8.37 ± 1.74	8.64 ± 3.36	3.32 ± 0.34	0.06 ± 0.08
+R	10.31 ± 3.15	2.77 ± 1.27	44.63 ± 5.29	22.12 ± 4.69	0.81 ± 0.07
+R+C	14.73 ± 4.60	4.82 ± 1.23	7.59 ± 7.43	2.45 ± 1.06	0.58 ± 0.23
+R+E	10.03 ± 2.92	3.39 ± 2.09	45.48 ± 5.68	21.09 ± 2.38	0.69 ± 0.23
+R+PRE	13.03 ± 2.65	3.96 ± 1.26	41.19 ± 6.50	24.09 ± 6.88	0.76 ± 0.05
+R+C+E	18.07 ± 1.58	6.46 ± 1.12	8.94 ± 4.86	4.58 ± 0.57	0.26 ± 0.15
+R+C+PRE	28.24 ± 0.72	25.59 ± 0.76	10.59 ± 0.39	2.51 ± 0.10	$\mathbf{0.00 \pm 0.00}$
+R+E+PRE	12.55 ± 3.58	5.40 ± 3.72	42.72 ± 7.04	20.81 ± 6.75	0.51 ± 0.31
+R+C+E+PRE	22.06 ± 3.84	12.28 ± 7.47	9.02 ± 2.32	3.19 ± 0.45	$\mathbf{0.00 \pm 0.00}$
+PNet	42.89 ± 4.33	40.82 ± 3.70	21.02 ± 2.10	4.93 ± 0.52	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	62.04 ± 5.95	60.90 ± 5.74	39.79 ± 6.10	8.43 ± 1.11	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{71.94 \pm 0.28}$	$\mathbf{70.76 \pm 0.37}$	$\mathbf{52.36 \pm 0.39}$	10.38 ± 0.06	$\mathbf{0.00 \pm 0.00}$
IMP	66.49 ± 0.46	65.06 ± 0.46	44.91 ± 0.58	9.28 ± 0.09	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 15: **Omniglot-EvenOdd, Scallop, 10-alphabet split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using Scallop as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%↑**	**F1(C)%↑**	**Acc(Y)%↑**	**F1(Y)%↑**	**Cls(C)%↓**
Baseline	9.56 ± 4.55	3.63 ± 1.77	55.13 ± 10.74	31.45 ± 6.24	0.71 ± 0.03
+C	12.05 ± 4.22	4.35 ± 1.53	55.77 ± 6.09	28.22 ± 2.88	0.73 ± 0.05
+E	12.44 ± 4.18	4.60 ± 1.49	54.24 ± 5.62	26.63 ± 3.44	0.68 ± 0.07
+PRE	19.06 ± 4.31	7.96 ± 2.27	55.04 ± 5.12	28.28 ± 3.11	0.72 ± 0.04
+C+E	25.74 ± 5.63	12.77 ± 4.76	32.51 ± 8.26	10.10 ± 1.44	0.50 ± 0.08
+C+PRE	19.13 ± 3.16	8.48 ± 1.76	53.23 ± 10.23	26.31 ± 5.17	0.72 ± 0.07
+E+PRE	22.32 ± 0.22	9.68 ± 0.12	53.14 ± 0.47	27.30 ± 0.31	0.69 ± 0.03
+C+E+PRE	63.46 ± 3.92	52.92 ± 5.94	45.08 ± 3.03	9.09 ± 0.78	0.03 ± 0.06
+R	10.56 ± 4.82	1.87 ± 0.81	51.81 ± 11.16	22.45 ± 4.04	0.90 ± 0.00
+R+C	12.91 ± 5.31	4.38 ± 2.18	51.55 ± 10.28	28.69 ± 5.54	0.74 ± 0.08
+R+E	10.56 ± 4.82	1.87 ± 0.81	51.81 ± 11.16	22.45 ± 4.04	0.90 ± 0.00
+R+PRE	17.57 ± 1.27	7.68 ± 0.65	61.42 ± 0.75	$\mathbf{32.32 \pm 1.92}$	0.77 ± 0.05
+R+C+E	13.94 ± 1.68	2.64 ± 0.68	48.70 ± 13.36	19.96 ± 5.98	0.86 ± 0.12
+R+C+PRE	22.84 ± 1.26	10.02 ± 0.75	52.68 ± 2.28	25.42 ± 3.13	0.68 ± 0.04
+R+E+PRE	16.55 ± 3.40	6.96 ± 1.87	61.60 ± 0.74	31.00 ± 2.29	0.76 ± 0.05
+R+C+E+PRE	14.75 ± 0.55	2.76 ± 0.58	50.24 ± 7.49	17.83 ± 2.27	0.87 ± 0.05
+PNet	69.79 ± 1.20	65.89 ± 1.29	49.12 ± 1.66	9.68 ± 0.30	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	78.98 ± 1.05	75.34 ± 1.14	62.67 ± 1.62	11.90 ± 0.23	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{89.44 \pm 0.43}$	$\mathbf{86.98 \pm 0.48}$	$\mathbf{79.98 \pm 0.75}$	14.20 ± 0.33	$\mathbf{0.00 \pm 0.00}$
IMP	88.07 ± 0.49	85.66 ± 0.58	77.61 ± 0.81	13.85 ± 0.35	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 16: **Omniglot-EvenOdd, Scallop, 40 alphabets split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using Scallop as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 6.33% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%**↑	**F1(C)%**↑	**Acc(Y)%**↑	**F1(Y)%**↑	**Cls(C)%**↓
Baseline	7.87 ± 0.47	2.94 ± 1.00	44.89 ± 5.91	23.67 ± 2.91	0.71 ± 0.05
+C	9.21 ± 3.01	2.80 ± 1.69	45.21 ± 4.39	23.10 ± 3.79	0.77 ± 0.08
+E	11.41 ± 3.39	3.37 ± 1.40	44.81 ± 5.31	$\mathbf{24.33 \pm 3.08}$	0.71 ± 0.03
+PRE	16.87 ± 0.30	6.22 ± 0.67	42.59 ± 3.65	21.73 ± 1.26	0.70 ± 0.00
+C+E	22.39 ± 6.10	13.08 ± 6.67	20.26 ± 10.16	6.07 ± 2.28	0.36 ± 0.12
+C+PRE	8.43 ± 1.17	2.20 ± 0.80	49.61 ± 7.26	20.19 ± 2.54	0.70 ± 0.04
+E+PRE	16.84 ± 0.31	6.02 ± 0.78	41.38 ± 3.47	21.65 ± 1.11	0.69 ± 0.03
+C+E+PRE	31.02 ± 0.64	23.20 ± 0.81	17.61 ± 0.62	4.26 ± 0.18	0.13 ± 0.08
+R	9.26 ± 4.22	1.67 ± 0.69	48.31 ± 0.00	21.72 ± 0.00	0.90 ± 0.00
+R+C	8.87 ± 0.84	3.14 ± 1.29	42.50 ± 5.32	22.75 ± 3.89	0.73 ± 0.08
+R+E	9.26 ± 4.22	1.67 ± 0.69	48.31 ± 0.00	21.72 ± 0.00	0.90 ± 0.00
+R+PRE	11.20 ± 3.98	3.00 ± 1.53	43.54 ± 7.93	20.75 ± 2.86	0.72 ± 0.04
+R+C+E	13.03 ± 1.96	2.30 ± 0.31	42.35 ± 9.95	19.58 ± 3.68	0.90 ± 0.00
+R+C+PRE	17.05 ± 3.03	6.83 ± 1.86	45.03 ± 2.28	18.77 ± 3.63	0.65 ± 0.07
+R+E+PRE	10.53 ± 3.84	2.89 ± 1.71	48.50 ± 3.95	21.88 ± 3.03	0.72 ± 0.04
+R+C+E+PRE	23.03 ± 0.71	15.44 ± 0.56	16.49 ± 1.21	6.24 ± 0.44	0.50 ± 0.00
+PNet	45.90 ± 0.76	43.71 ± 0.92	22.73 ± 0.66	5.33 ± 0.13	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	64.69 ± 0.77	63.66 ± 0.81	42.67 ± 0.97	8.89 ± 0.16	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{71.72 \pm 0.50}$	$\mathbf{70.51 \pm 0.55}$	$\mathbf{52.05 \pm 0.65}$	10.34 ± 0.09	$\mathbf{0.00 \pm 0.00}$
IMP	66.49 ± 0.46	65.06 ± 0.46	44.91 ± 0.58	9.28 ± 0.09	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 17: **Omniglot-EvenOdd, CTSketch, 10-alphabet split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using CTSketch as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset.

Model/Mitigation	**Acc(C)%**↑	**F1(C)%**↑	**Acc(Y)%**↑	**F1(Y)%**↑	**Cls(C)%**↓
Baseline	9.74 ± 3.51	3.36 ± 1.19	56.23 ± 6.76	33.33 ± 5.96	0.71 ± 0.03
+C	15.96 ± 4.34	5.88 ± 2.96	52.68 ± 4.50	30.33 ± 4.50	0.72 ± 0.06
+E	15.68 ± 2.31	4.03 ± 2.13	50.20 ± 5.59	31.12 ± 5.82	0.73 ± 0.06
+PRE	16.80 ± 4.55	6.91 ± 2.37	59.98 ± 2.73	30.65 ± 3.18	0.76 ± 0.05
+C+E	30.86 ± 15.55	18.44 ± 14.68	27.54 ± 8.83	8.16 ± 0.84	0.42 ± 0.21
+C+PRE	18.57 ± 6.47	7.82 ± 3.74	54.77 ± 13.38	25.05 ± 3.15	0.70 ± 0.09
+E+PRE	21.97 ± 0.82	9.53 ± 0.29	53.39 ± 1.97	27.62 ± 1.52	0.70 ± 0.00
+C+E+PRE	63.40 ± 6.07	54.24 ± 5.66	44.47 ± 6.28	8.96 ± 1.10	$\mathbf{0.00 \pm 0.00}$
+R	14.33 ± 1.80	2.50 ± 0.29	51.81 ± 11.16	22.45 ± 4.04	0.90 ± 0.00
+R+C	18.00 ± 3.58	5.70 ± 3.31	51.08 ± 4.64	31.22 ± 8.48	0.73 ± 0.08
+R+E	14.33 ± 1.80	2.50 ± 0.29	55.53 ± 0.00	23.80 ± 0.00	0.90 ± 0.00
+R+PRE	18.40 ± 2.01	8.03 ± 0.90	61.86 ± 0.59	$\mathbf{33.83 \pm 0.90}$	0.77 ± 0.05
+R+C+E	14.56 ± 0.00	2.54 ± 0.00	55.53 ± 0.00	23.80 ± 0.00	0.90 ± 0.00
+R+C+PRE	22.39 ± 1.45	9.77 ± 0.84	51.89 ± 2.98	24.05 ± 2.34	0.66 ± 0.05
+R+E+PRE	17.26 ± 2.85	7.40 ± 1.59	61.22 ± 0.94	32.71 ± 2.13	0.77 ± 0.05
+R+C+E+PRE	14.68 ± 0.36	2.71 ± 0.47	54.76 ± 1.06	17.84 ± 0.08	0.87 ± 0.05
+PNet	70.17 ± 0.97	66.36 ± 1.08	49.51 ± 1.48	9.73 ± 0.36	$\mathbf{0.00 \pm 0.00}$
+ResNet-PNet	78.81 ± 0.64	75.16 ± 0.78	62.36 ± 1.14	11.66 ± 0.30	$\mathbf{0.00 \pm 0.00}$
+MOSAIC (ours)	$\mathbf{89.66 \pm 0.54}$	$\mathbf{87.22 \pm 0.59}$	$\mathbf{80.38 \pm 0.96}$	14.38 ± 0.11	$\mathbf{0.00 \pm 0.00}$
IMP	88.07 ± 0.49	85.66 ± 0.58	77.61 ± 0.81	13.85 ± 0.35	$\mathbf{0.00 \pm 0.00}$
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	$\mathbf{0.00 \pm 0.00}$

Table 18: **Omniglot-EvenOdd, CTSketch, 40 alphabets split.** Comparison of MOSAIC against RS mitigation strategies (C, E, PRE, R, and their combinations) and prototypical-network baselines (PNet, ResNet-PNet) using CTSketch as the NeSy layer. All models share a fixed annotation budget of 2 labeled images per character, corresponding to 6.33% of the full Omniglot dataset .

Model/Mitigation	Acc(C)%↑	F1(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
Baseline	14.84 ± 2.31	3.68 ± 1.38	38.71 ± 6.76	25.61 ± 5.01	0.75 ± 0.05
+C	10.61 ± 2.98	2.31 ± 1.01	41.10 ± 4.78	22.11 ± 4.24	0.80 ± 0.09
+E	14.81 ± 2.28	3.26 ± 1.26	41.70 ± 5.24	27.34 ± 3.73	0.82 ± 0.10
+PRE	16.96 ± 0.48	6.24 ± 0.81	42.56 ± 4.09	22.42 ± 0.97	0.70 ± 0.00
+C+E	30.04 ± 1.30	21.63 ± 1.26	19.01 ± 1.09	4.70 ± 0.10	0.14 ± 0.05
+C+PRE	13.27 ± 3.84	4.04 ± 1.71	44.16 ± 7.14	22.08 ± 5.44	0.67 ± 0.05
+E+PRE	16.91 ± 0.28	6.30 ± 0.75	43.22 ± 3.04	22.55 ± 0.48	0.69 ± 0.03
+C+E+PRE	30.89 ± 1.12	23.07 ± 1.26	18.42 ± 0.72	4.32 ± 0.14	0.08 ± 0.04
+R	15.39 ± 0.00	2.67 ± 0.00	18.12 ± 0.00	10.23 ± 0.00	0.90 ± 0.00
+R+C	13.24 ± 3.86	4.33 ± 1.85	41.38 ± 4.05	23.75 ± 4.24	0.74 ± 0.09
+R+E	15.39 ± 0.00	2.67 ± 0.00	21.14 ± 9.06	11.38 ± 3.45	0.90 ± 0.00
+R+PRE	15.15 ± 2.41	3.96 ± 1.29	37.00 ± 7.01	23.94 ± 2.92	0.70 ± 0.00
+R+C+E	10.76 ± 0.00	1.94 ± 0.00	14.49 ± 22.14	6.52 ± 9.95	0.90 ± 0.00
+R+C+PRE	17.74 ± 0.54	7.03 ± 0.84	42.99 ± 3.49	17.45 ± 3.04	0.63 ± 0.05
+R+E+PRE	15.04 ± 2.38	3.88 ± 1.34	41.33 ± 3.61	**27.38 ± 4.57**	0.73 ± 0.05
+R+C+E+PRE	21.67 ± 2.37	13.87 ± 2.64	16.02 ± 0.55	6.07 ± 0.36	0.52 ± 0.06
+PNet	45.82 ± 0.49	43.70 ± 0.54	22.72 ± 0.42	5.32 ± 0.07	**0.00 ± 0.00**
+ResNet-PNet	64.91 ± 0.76	63.91 ± 0.78	42.94 ± 0.93	8.93 ± 0.15	**0.00 ± 0.00**
+MOSAIC (ours)	**71.71 ± 0.30**	**70.51 ± 0.28**	**52.02 ± 0.39**	10.34 ± 0.06	**0.00 ± 0.00**
IMP	66.49 ± 0.46	65.06 ± 0.46	44.91 ± 0.58	9.28 ± 0.09	**0.00 ± 0.00**
Random	10.02 ± 0.10	9.30 ± 0.13	7.96 ± 0.23	1.89 ± 0.05	**0.00 ± 0.00**

Table 19: **BDD-OIA, DPL baseline, support-set size $S = 5$ per concept.** Comparison of MOSAIC ($\upsilon = 0.75$) against reasoning-shortcut mitigation strategies (C, C(S), PRE(S), E, and their combinations) and prototypical-network baselines under the DeepProbLog NeSy layer. For each of the 21 concepts, $S = 5$ positive support examples are annotated, with negatives sampled at the natural per-concept imbalance ratio of the dataset (no capping). This yields 1998 annotated (image, concept) pairs, corresponding to 0.42% of the full BDD-OIA dataset (22,924 images × 21 concepts = 481,404 possible annotations).

Model/Mitigation	Acc(C)%↑	F1(C)%↑	WRec(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
DPL	59.43 ± 0.95	39.16 ± 0.35	59.43 ± 0.95	75.51 ± 0.11	61.44 ± 0.40	0.20 ± 0.03
+C	60.16 ± 1.17	41.25 ± 1.05	60.15 ± 1.17	75.40 ± 0.10	61.22 ± 0.48	0.13 ± 0.01
+C(S)	69.53 ± 0.38	48.10 ± 0.66	69.52 ± 0.38	75.59 ± 0.13	61.19 ± 0.11	0.07 ± 0.03
+PRE(S)	60.25 ± 2.80	39.31 ± 1.69	60.25 ± 2.80	75.62 ± 0.26	61.15 ± 0.19	0.22 ± 0.04
+E	63.87 ± 0.59	47.87 ± 0.22	63.88 ± 0.59	**75.83 ± 0.18**	62.74 ± 0.19	**0.00 ± 0.00**
+PRE(S)+C	60.71 ± 1.66	41.14 ± 0.95	60.70 ± 1.66	75.82 ± 0.29	61.13 ± 0.40	0.16 ± 0.05
+PRE(S)+C(S)	70.23 ± 1.12	48.11 ± 0.67	70.23 ± 1.13	75.68 ± 0.16	61.12 ± 0.09	0.07 ± 0.02
+PRE(S)+E	63.77 ± 0.73	48.22 ± 0.71	63.78 ± 0.73	75.57 ± 0.15	62.61 ± 0.18	**0.00 ± 0.00**
+C+E	67.40 ± 0.84	50.01 ± 0.25	67.40 ± 0.84	75.77 ± 0.26	**62.76 ± 0.36**	**0.00 ± 0.01**
+C(S)+E	64.98 ± 0.32	49.18 ± 0.21	64.98 ± 0.32	75.45 ± 0.19	62.42 ± 0.59	**0.00 ± 0.00**
+PRE(S)+C+E	67.49 ± 1.34	50.21 ± 0.94	67.49 ± 1.34	75.55 ± 0.16	62.43 ± 0.66	0.01 ± 0.01
+PRE(S)+C(S)+E	65.03 ± 0.33	49.22 ± 0.21	65.04 ± 0.33	75.48 ± 0.21	62.57 ± 0.30	**0.00 ± 0.00**
+PNet	72.74 ± 4.87	51.86 ± 2.40	72.75 ± 4.87	58.51 ± 3.61	45.43 ± 3.21	**0.00 ± 0.00**
+ResNet-PNet	47.54 ± 7.26	36.78 ± 5.65	47.54 ± 7.27	47.34 ± 0.00	25.76 ± 0.00	0.10 ± 0.08
+MOSAIC	**87.42 ± 0.37**	**54.71 ± 1.54**	**87.42 ± 0.37**	60.92 ± 3.01	42.89 ± 1.83	**0.00 ± 0.01**

Table 20: **BDD-OIA, DPL baseline, support-set size $S = 10$ per concept.** Comparison of MOSAIC ($\upsilon = 0.75$) against reasoning-shortcut mitigation strategies (C, C(S), PRE(S), E, and their combinations) and prototypical-network baselines under the DeepProbLog NeSy layer. For each of the 21 concepts, $S = 10$ positive support examples are annotated, with negatives sampled at the natural per-concept imbalance ratio of the dataset (no capping). This yields 3941 annotated (image, concept) pairs, corresponding to 0.82% of the full BDD-OIA dataset (22,924 images $\times$ 21 concepts = 481,404 possible annotations).

Model/Mitigation	Acc(C)%↑	F1(C)%↑	WRec(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
DPL	59.54 ± 0.57	39.25 ± 0.35	59.53 ± 0.56	75.50 ± 0.16	61.18 ± 0.25	0.20 ± 0.03
+C	60.09 ± 0.71	41.32 ± 0.63	60.08 ± 0.71	75.43 ± 0.12	61.02 ± 0.19	0.14 ± 0.03
+C(S)	75.37 ± 0.65	50.98 ± 0.17	75.37 ± 0.65	75.54 ± 0.09	61.25 ± 0.27	0.05 ± 0.02
+PRE(S)	69.48 ± 2.59	44.32 ± 2.08	69.48 ± 2.60	75.70 ± 0.22	61.00 ± 0.30	0.21 ± 0.06
+E	63.61 ± 0.61	47.76 ± 0.45	63.62 ± 0.61	75.75 ± 0.28	62.65 ± 0.41	**0.00 ± 0.00**
+PRE(S)+C	68.96 ± 2.43	45.99 ± 1.67	68.95 ± 2.43	75.64 ± 0.24	61.01 ± 0.33	0.14 ± 0.05
+PRE(S)+C(S)	75.66 ± 0.70	51.05 ± 0.41	75.67 ± 0.70	75.57 ± 0.10	61.27 ± 0.24	0.05 ± 0.02
+PRE(S)+E	67.85 ± 1.16	50.46 ± 0.91	67.85 ± 1.16	75.15 ± 0.30	61.72 ± 0.90	**0.00 ± 0.00**
+C+E	67.21 ± 0.77	49.76 ± 0.45	67.20 ± 0.77	**75.79 ± 0.18**	**62.70 ± 0.63**	0.01 ± 0.01
+C(S)+E	65.69 ± 0.36	49.76 ± 0.19	65.70 ± 0.36	75.32 ± 0.18	62.34 ± 0.27	**0.00 ± 0.00**
+PRE(S)+C+E	69.14 ± 0.66	51.59 ± 0.64	69.13 ± 0.65	75.43 ± 0.27	62.07 ± 0.89	0.01 ± 0.01
+PRE(S)+C(S)+E	67.68 ± 1.40	50.69 ± 0.64	67.68 ± 1.40	75.35 ± 0.15	62.32 ± 0.60	**0.00 ± 0.00**
+PNet	76.19 ± 0.91	55.11 ± 0.45	76.18 ± 0.91	57.16 ± 2.59	43.35 ± 1.99	**0.00 ± 0.00**
+ResNet-PNet	49.66 ± 4.90	38.80 ± 3.81	49.66 ± 4.90	47.39 ± 0.07	25.83 ± 0.13	0.08 ± 0.06
+MOSAIC	**88.60 ± 0.11**	**55.77 ± 1.06**	**88.60 ± 0.11**	62.65 ± 1.65	42.29 ± 4.48	**0.00 ± 0.01**

Table 21: **BDD-OIA, DPL baseline, support-set size $S = 15$ per concept.** Comparison of MOSAIC ($\upsilon = 0.75$) against reasoning-shortcut mitigation strategies (C, C(S), PRE(S), E, and their combinations) and prototypical-network baselines under the DeepProbLog NeSy layer. For each of the 21 concepts, $S = 15$ positive support examples are annotated, with negatives sampled at the natural per-concept imbalance ratio of the dataset (no capping). This yields 5810 annotated (image, concept) pairs, corresponding to 1.21% of the full BDD-OIA dataset (22,924 images $\times$ 21 concepts = 481,404 possible annotations).

Model/Mitigation	Acc(C)%↑	F1(C)%↑	WRec(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
DPL	59.88 ± 1.22	39.12 ± 0.48	59.88 ± 1.22	75.71 ± 0.22	61.44 ± 0.38	0.22 ± 0.04
+C	60.23 ± 0.51	41.42 ± 0.30	60.22 ± 0.51	75.49 ± 0.24	61.23 ± 0.28	0.14 ± 0.02
+C(S)	76.29 ± 0.88	51.57 ± 0.39	76.29 ± 0.88	75.66 ± 0.12	60.97 ± 0.46	0.03 ± 0.02
+PRE(S)	66.41 ± 2.04	44.24 ± 1.92	66.41 ± 2.04	75.60 ± 0.13	60.86 ± 0.37	0.17 ± 0.06
+E	63.60 ± 0.41	47.82 ± 0.35	63.62 ± 0.41	**75.81 ± 0.24**	**62.63 ± 0.34**	**0.00 ± 0.00**
+PRE(S)+C	69.72 ± 6.12	46.76 ± 3.72	69.71 ± 6.12	75.51 ± 0.26	58.70 ± 4.24	0.12 ± 0.05
+PRE(S)+C(S)	76.86 ± 0.75	51.75 ± 0.52	76.86 ± 0.74	75.76 ± 0.12	60.97 ± 0.53	0.04 ± 0.01
+PRE(S)+E	67.27 ± 1.38	50.40 ± 0.91	67.27 ± 1.38	75.53 ± 0.20	61.98 ± 0.60	**0.00 ± 0.00**
+C+E	67.31 ± 0.27	49.93 ± 0.27	67.31 ± 0.27	75.66 ± 0.25	62.49 ± 0.24	**0.00 ± 0.01**
+C(S)+E	65.85 ± 0.54	49.93 ± 0.23	65.85 ± 0.53	75.48 ± 0.13	62.22 ± 0.23	**0.00 ± 0.00**
+PRE(S)+C+E	69.18 ± 1.12	51.78 ± 0.72	69.18 ± 1.12	75.73 ± 0.12	62.31 ± 0.46	0.01 ± 0.01
+PRE(S)+C(S)+E	67.01 ± 0.87	50.56 ± 0.57	67.00 ± 0.87	75.48 ± 0.22	62.18 ± 0.54	**0.00 ± 0.00**
+PNet	73.93 ± 2.18	54.95 ± 1.06	73.92 ± 2.18	58.09 ± 1.66	46.46 ± 1.87	**0.00 ± 0.00**
+ResNet-PNet	50.71 ± 5.30	39.52 ± 3.90	50.71 ± 5.29	47.40 ± 0.12	25.90 ± 0.30	0.07 ± 0.06
+MOSAIC	**89.00 ± 0.20**	**57.13 ± 0.92**	**89.00 ± 0.20**	65.58 ± 2.40	46.50 ± 3.56	**0.00 ± 0.00**

Table 22: **BDD-OIA, DPL baseline, support-set size $S = 20$ per concept.** Comparison of MOSAIC ($\upsilon = 0.75$) against reasoning-shortcut mitigation strategies (C, C(S), PRE(S), E, and their combinations) and prototypical-network baselines under the DeepProbLog NeSy layer. For each of the 21 concepts, $S = 20$ positive support examples are annotated, with negatives sampled at the natural per-concept imbalance ratio of the dataset (no capping). This yields 7832 annotated (image, concept) pairs, corresponding to 1.63% of the full BDD-OIA dataset (22,924 images × 21 concepts = 481,404 possible annotations).

Model/Mitigation	Acc(C)%↑	F1(C)%↑	WRec(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
DPL	59.21 ± 0.62	38.90 ± 0.43	59.20 ± 0.61	75.55 ± 0.13	61.12 ± 0.27	0.19 ± 0.03
+C	59.95 ± 0.75	41.10 ± 0.58	59.94 ± 0.75	75.29 ± 0.09	61.08 ± 0.26	0.13 ± 0.01
+C(S)	77.15 ± 0.81	52.33 ± 0.10	77.15 ± 0.81	75.51 ± 0.13	61.27 ± 0.18	**0.00 ± 0.00**
+PRE(S)	68.83 ± 3.62	45.57 ± 2.33	68.83 ± 3.62	75.66 ± 0.17	60.95 ± 0.45	0.20 ± 0.04
+E	63.53 ± 0.44	47.73 ± 0.23	63.54 ± 0.44	**75.90 ± 0.09**	**62.60 ± 0.48**	**0.00 ± 0.00**
+PRE(S)+C	70.07 ± 3.48	47.28 ± 2.05	70.06 ± 3.48	75.69 ± 0.50	60.99 ± 0.18	0.12 ± 0.03
+PRE(S)+C(S)	76.89 ± 0.82	52.27 ± 0.32	76.89 ± 0.82	75.59 ± 0.09	61.28 ± 0.36	**0.00 ± 0.00**
+PRE(S)+E	68.23 ± 0.19	51.30 ± 0.13	68.23 ± 0.18	75.34 ± 0.10	61.50 ± 0.58	**0.00 ± 0.00**
+C+E	67.38 ± 0.57	49.96 ± 0.36	67.38 ± 0.57	75.70 ± 0.15	62.59 ± 0.46	0.01 ± 0.01
+C(S)+E	66.68 ± 0.53	50.50 ± 0.18	66.67 ± 0.52	75.49 ± 0.11	62.44 ± 0.24	**0.00 ± 0.00**
+PRE(S)+C+E	69.94 ± 1.06	52.41 ± 0.52	69.94 ± 1.06	75.59 ± 0.15	61.87 ± 0.70	**0.00 ± 0.01**
+PRE(S)+C(S)+E	68.37 ± 0.28	51.46 ± 0.17	68.38 ± 0.28	75.36 ± 0.10	61.58 ± 0.53	**0.00 ± 0.00**
+PNet	72.52 ± 1.61	54.81 ± 0.93	72.52 ± 1.61	58.41 ± 2.61	43.80 ± 1.79	**0.00 ± 0.00**
+ResNet-PNet	50.16 ± 5.98	39.52 ± 4.95	50.15 ± 5.98	47.40 ± 0.13	25.88 ± 0.25	0.07 ± 0.08
+MOSAIC	**89.13 ± 0.12**	**57.59 ± 1.29**	**89.12 ± 0.12**	62.45 ± 3.02	41.42 ± 3.17	**0.00 ± 0.00**

Table 23: **BDD-OIA, CCN+ baseline, support-set size $S = 20$ per concept.** Comparison of MOSAIC ($\upsilon = 0.75$) against reasoning-shortcut mitigation strategies (C, C(S), PRE(S), E, and their combinations) and prototypical-network baselines under the CCN+ (PiShield) NeSy layer. For each of the 21 concepts, $S = 20$ positive support examples are annotated, with negatives sampled at the natural per-concept imbalance ratio of the dataset (no capping). This yields 7828 annotated (image, concept) pairs, corresponding to 1.63% of the full BDD-OIA dataset (22,924 images × 21 concepts = 481,404 possible annotations).

Model/Mitigation	Acc(C)%↑	F1(C)%↑	WRec(C)%↑	Acc(Y)%↑	F1(Y)%↑	Cls(C)%↓
CCN+	52.96 ± 11.68	31.98 ± 6.15	52.96 ± 11.68	75.84 ± 1.02	58.07 ± 5.14	0.34 ± 0.08
+C	54.54 ± 4.74	35.54 ± 2.57	54.54 ± 4.75	75.44 ± 0.42	58.71 ± 5.27	0.22 ± 0.07
+C(S)	85.45 ± 0.34	53.00 ± 0.49	85.45 ± 0.34	75.50 ± 0.15	60.76 ± 0.23	0.02 ± 0.04
+PRE(S)	85.09 ± 0.68	50.49 ± 1.64	85.09 ± 0.68	76.66 ± 0.26	60.85 ± 0.83	0.24 ± 0.08
+E	56.91 ± 1.08	43.93 ± 0.42	56.90 ± 1.08	76.60 ± 0.22	**63.03 ± 0.76**	**0.00 ± 0.00**
+PRE(S)+C	85.37 ± 0.43	52.62 ± 0.76	85.37 ± 0.43	**76.67 ± 0.15**	60.77 ± 0.78	0.22 ± 0.07
+PRE(S)+C(S)	85.51 ± 0.07	53.73 ± 0.66	85.51 ± 0.07	75.32 ± 0.18	60.72 ± 0.33	0.02 ± 0.02
+PRE(S)+E	61.35 ± 0.64	48.59 ± 0.46	61.35 ± 0.64	76.27 ± 0.30	62.62 ± 1.22	**0.00 ± 0.01**
+C+E	62.55 ± 1.45	46.54 ± 0.64	62.55 ± 1.44	75.97 ± 1.31	62.95 ± 0.56	0.03 ± 0.03
+C(S)+E	63.36 ± 1.13	48.56 ± 0.39	63.33 ± 1.13	76.17 ± 0.41	62.89 ± 0.52	0.01 ± 0.01
+PRE(S)+C+E	66.33 ± 0.60	50.79 ± 0.68	66.32 ± 0.59	76.20 ± 0.36	62.73 ± 1.04	0.01 ± 0.02
+PRE(S)+C(S)+E	63.31 ± 4.72	49.31 ± 1.82	63.30 ± 4.72	76.09 ± 0.51	62.48 ± 0.41	**0.00 ± 0.01**
+PNet	73.05 ± 1.01	55.17 ± 0.51	73.04 ± 1.02	61.58 ± 1.22	45.60 ± 0.62	**0.00 ± 0.00**
+ResNet-PNet	47.43 ± 13.79	37.29 ± 10.04	47.42 ± 13.79	57.42 ± 8.06	37.99 ± 3.33	0.10 ± 0.14
+MOSAIC	**89.17 ± 0.12**	**57.17 ± 0.89**	**89.16 ± 0.12**	67.83 ± 1.80	25.35 ± 4.66	**0.00 ± 0.01**

Table 24: **Omniglot-EvenOdd, MOSAIC vs. PNet, CCN+, 10-alphabet split — mode purity and IRS-rate.** Held-out query-level mode diagnostics under [**A3**]: $\gamma_i(x) = \arg\min_j \|f_\theta^i(x) - c_j\|^2$ is the nearest-prototype map from true character identity to prototype. **Mode purity (global)** is plurality-vote cluster purity over all character-level modes; **mode purity (within-EC)** restricts pooling to prototypes of the correct digit-level EC, so a global error landing outside the correct EC is counted as an ERS (Definition 1.1) rather than a pure mode-level error. **IRS-rate** = $1-|\mathrm{im}(\hat{\gamma}_i)|/m$ is 0 iff $\hat{\gamma}_i$ is injective (Theorem 4.6); for PNet, which maintains exactly one prototype per EC.

Model (coord.)	Purity (global)%↑	Purity (within-EC)%↑	IRS-rate%↓
MOSAIC	**77.69 ± 1.27**	**80.98 ± 1.08**	**2.94 ± 0.51**
PNet	3.02 ± 0.05	2.93 ± 0.00	97.14 ± 0.00

Table 25: **Omniglot-EvenOdd, MOSAIC (IMP) vs. PNet, CCN+, 40-alphabet split — mode purity and IRS-rate.** Same protocol as Table 24, on the full 40-alphabet (40 alphabets) split.

Model (coord.)	Purity (global)%↑	Purity (within-EC)%↑	IRS-rate%↓
MOSAIC	**61.45 ± 0.69**	**69.32 ± 0.58**	**7.14 ± 1.12**
PNet	1.03 ± 0.02	1.00 ± 0.00	99.03 ± 0.00

Table 26: **BDD-OIA, MOSAIC vs. PNet, CCN+ — mode usage.** BDD-OIA carries no ground-truth mode/character identity, so mode purity and IRS-rate are not computable. We instead report **top-1 share** as the mass on the most-used prototype and **entropy ratio** that normalizes usage entropy by the logarithm over the number of maintained prototypes, so 0 means all mass on one prototype and 1 means uniform usage. All entries are computed wrt. a support size of 20 annotations.

Model	Top-1 share%↓	Entropy ratio%↑
MOSAIC	**32.29 ± 7.35**	**73.68 ± 6.89**
PNet	100.00 ± 0.00	0.00 ± 0.00

Moving from the 10 to the 40-alphabet split multiplies the number of IC-level modes realising each EC. The best mitigation loses roughly half of its F1(C) — DPL C+E+PRE 48.19% → 23.73%, Scallop 52.92% → 23.20%, CTSketch 54.24% → 23.07% — whereas MOSAIC retains ≈ 81% of its score (87.00% → 70.71%, 86.98% → 70.51%, 87.22% → 70.51%). The strategies of Marconato et al. (2023b) were designed for benchmarks in which each EC has one appearance; their efficacy decays exactly in the regime our setting stress tests.

Classical mitigation strategies also showed unreliability across various seeds. CTSketch +C+E reports F1(C) 18.44% ± 14.68%; CCN+ 40-alphabet +C+PRE 16.32% ± 9.58% and +E+PRE 7.19% ± 6.53%. MOSAIC's does not exceed 0.59% on `Omniglot-EvenOdd` and 1.29% on `BDDOIA`. In the latter, the strongest single mitigation is C(S), i.e. cross-entropy on exactly the support pool MOSAIC consumes: on CCN+ it lifts Acc(C) 52.96% → 85.45% and F1(C) 31.98% → 53.00%, and PRE(S)+C(S) gives 85.51% / 53.73%. MOSAIC improves on top of it with 89.17% and 57.17%, together with WRec(C) 89.16% against 85.51%. On DPL the same mitigation is markedly weaker (77.15% and 52.33%), so mitigation efficacy is backbone-dependent. MOSAIC lands within 0.42% F1(C) of itself across the two layers (57.59% vs. 57.17%). Finally, E is the only strategy that reliably zeroes Cls(C) on `BDDOIA` but caps Acc(C) at ≈ 63.5%: entropy spreads the EC predictions without making them correct.

D.3 RQ.3: Can greater encoder capacity avoid implicit Reasoning Shortcuts?

The ResNet-PNet rows of Tables 9–18 and 19–23 give the full encoder-capacity ablation; MOSAIC retains the four-block convolutional extractor of Snell et al. (2017) throughout.

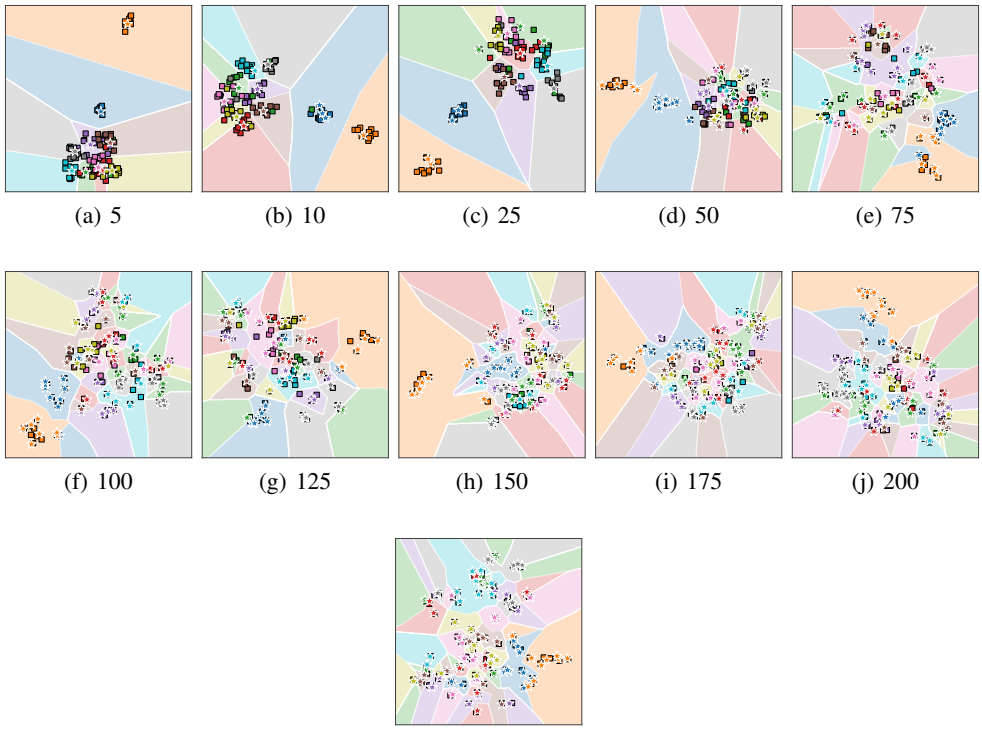
On the 10-alphabet split the ResNet12 encoder buys +8% to +10% F1(C) over conv4 PNet on SL (74.71%), Scallop (75.34%), CTSketch (75.16%) and CCN+ (75.14%), but *costs* 7.12% on DPL (68.00% → 60.88%): the effect of capacity is not even consistently signed across inference layers. The best of these remains 11.64 points below MOSAIC with the small extractor. On the 40-alphabet split, where there is more room, capacity improves every layer (+9.92% to +20.44% F1(C)) and still trails MOSAIC by 6.60% to 27.99% points.

Crucially, we found that capacity cannot change what is structurally unidentifiable. ResNet-PNet maintains one prototype per EC, so $|\mathrm{im}(\hat{\gamma}_i)| \leq 10$ regardless of the encoder and the IRS-rate floor $1 - h_i/m$ of Corollary 4.7 is unaffected: 97.14% in the 10-alphabet and 99.03% in the 40-alphabet split. A larger encoder can tighten each EC's embedding cloud but it cannot make one centroid represent several visually disjoint modes. This is why the gains above saturate well below MOSAIC.

Interestingly, on `BDDOIA` capacity becomes actively harmful. With a support set of size 20, ResNet-PNet scores Acc(C) 50.16% ± 5.98% and F1(C) 39.52% ± 4.95% on DPL, and 47.43% ± 13.79% / 37.29% ± 10.04% on CCN+, against PNet's 72.52% / 54.81% and 73.05% / 55.17% and MOSAIC's 89.13% / 57.59% and 89.17% / 57.17%. Notably, it falls below even the unmitigated DPL backbone on Acc(C) (59.21%) and Cls(C) rises from 0.00% to 0.07%–0.10%, i.e. capacity *re-introduces* the concept collapse that prototypical inference had removed; we also observe that Acc(Y) pins at 47.40% ± 0.13% at every support size. It also fails to benefit from additional annotation: its F1(C) moves 36.78% → 38.80% → 39.52% → 39.52% as S goes from 5 to 20, while MOSAIC moves 54.71% → 57.59% over the same range. With 1.63% of the EC labels available, the overparameterised encoder memorises the support pool; the mode-aware mixture, on a far smaller encoder, converts the same labels into structure.

D.4 RQ.4: How effectively does MOSAIC discover modes from the data?

Figure 6 extends Figure 4 to the full sweep $5, 10, 25, 50, 75, 100, 125, 150, 175, 200, 500$ supports per EC. The transition from a single collapsed region per EC to one prototype per character-level mode is continuous in the support size. We stress that the CRP concentration is held at $\upsilon = 0.1$ across the whole experimental test: the number of prototypes is not tuned per setting but is the posterior's response to the data, so what the figure exhibits is the support becoming sufficient to *expose* the modes (c.f. Definition 4.2). At 10 supports per EC on `Omniglot-EvenOdd` concept accuracy falls back to mode-blind levels. This is the expected and, we argue, the desirable behaviour: complete mode coverage (Definition 4.2) is an assumption on the *data*, not on the model, and when it is violated MOSAIC has no evidence for the unobserved modes. What matters is how it fails —



(a) 5 (b) 10 (c) 25 (d) 50 (e) 75

(f) 100 (g) 125 (h) 150 (i) 175 (j) 200

(k) 500

Figure 6: Prototypes and query assignments for increasing numbers of support datapoints per class.

the CRP posterior keeps a single component per EC and MOSAIC degrades continuously onto PNet, rather than instantiating spurious prototypes or failing unpredictably. The same limit is reached from the model side in RQ5 by taking $\upsilon = 0$.

For the case of `BDDOIA`, Tables 19–22 move the support set size $S \in \{5, 10, 15, 20\}$, i.e. 0.42%, 0.82%, 1.21% and 1.63% of the possible EC annotations. MOSAIC is monotone in S on both concept metrics — Acc(C) $87.42\% \rightarrow 88.60\% \rightarrow 89.00\% \rightarrow 89.13\%$, F1(C) $54.71\% \rightarrow 55.77\% \rightarrow 57.13\% \rightarrow 57.59\%$ — with diminishing but non negative returns and with the standard deviation tightening ($\pm 0.37\%$ to $\pm 0.12\%$ on Acc(C)). Conversely PNet has Acc(C) $72.74\% \rightarrow 76.19\% \rightarrow 73.93\% \rightarrow 72.52\%$ and F1(C) $51.86\% \rightarrow 55.11\% \rightarrow 54.95\% \rightarrow 54.81\%$, i.e. non-monotone and flat past $S = 10$, at a standard deviation up to 4.87%. This another empirical confirmation of Corollary 4.7: a single centroid is an average over all the modes present in the support, so supports drawn from new modes make it *less* representative of any of them, while a mixture instantiates a new component instead. The practical consequence is an annotation argument: MOSAIC at $S = 5$ already matches PNet's F1(C) at $S = 20$ (54.71% vs. 54.81%) while adding 14.90 points of Acc(C) and weighted recall — a fourfold reduction of the annotation budget at strictly better concept identification.

D.5 RQ.5: Can MOSAIC recover mode-blind prototypical predictors under complete mode coverage?

Tables 27 and 28 compare MOSAIC at $\upsilon = 0$ against PNet on `Omniglot-EvenOdd`. On SL, MOSAIC scores Acc(C) 70.59% and F1(C) 66.69% against 71.00% and 67.00% ($\Delta = -0.41\%$, -0.31%); on CCN+, 69.45% and 65.51% against 69.12% and 65.12% ($\Delta = +0.33\%$, $+0.39\%$). All four deltas are below half a point, smaller than MOSAIC's own seed standard deviation (0.78%–1.10%), and of opposite sign on the two layers: there is no systematic difference in either direction. since at $\upsilon = 0$ the CRP never opens a second component and the update of Algorithm 1 reduces to the mode-blind one. Cls(C) remains 0.00% throughout.

Tables 29 and 30 test the complementary condition, with MOSAIC run at its default υ and therefore free to split: SL 97.64% and 97.54% against PNet's 97.07% and 96.87%, Scallop 97.68% and 97.56% against 96.46% and 96.22%. The CRP prior is self-regularising so mode-awareness adapts on the benchmarks where mode-blindness is correct. As two further observations, MOSAIC is marginally *above* PNet on all four non-degenerate metrics on both layers (+0.57% and +1.22% Acc(C), at Cls(C) = 0.00% for both models), and it roughly halves the seed variance (±0.54% vs. ±1.31% on SL, ±0.61% vs. ±1.54% on Scallop): even in the regime where the mixture is redundant, it is the better-conditioned estimator of the two.

`MNIST-EvenOdd` also calibrates the F1(Y) discussion of Appendix D.1. At Acc(C) 97.68% and Acc(Y) 95.40% — nineteen sums in twenty exactly right — MOSAIC still records F1(Y) 20.76%, i.e., $\approx 4/19\%$, the fraction of sum values the benchmark's restricted digit pairs actually realise.

Taken together, RQ5 shows that MOSAIC is a strict generalisation of the mode-blind predictor of Andolfi & Giunchiglia (2025): it matches PNet, to within seed noise, in both regimes where PNet's single-prototype assumption is correct, and separates from it by 19% to 38% F1(C) points on `Omniglot-EvenOdd` (Table 1) when it is not. Mode-awareness is therefore not a trade-off the practitioner has to make in advance, but a property recovered from the data.

Table 27: **Omniglot-EvenOdd, MOSAIC (single-mode ablation, υ=0) vs. PNet, SL, 10-alphabet split.** Comparison of MOSAIC with extreme λ threshold at υ=0 against the single-prototype PNet baseline, under Semantic Loss as the NeSy layer. Both models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset (1,623 characters / 32,460 images).

Model	Acc(C)↑	F1(C)↑	Acc(Y)↑	F1(Y)↑	Cls(C)↓
MOSAIC (υ=0)	70.59 ± 0.78	66.69 ± 0.79	**50.37 ± 1.24**	9.78 ± 0.22	**0.00 ± 0.00**
PNet	**71.00 ± 0.01**	**67.00 ± 0.01**	50.00 ± 0.01	**10.00 ± 0.00**	**0.00 ± 0.00**

Table 28: **Omniglot-EvenOdd, MOSAIC (unimodal ablation, υ=0) vs. PNet, CCN+, 10-alphabet split.** Comparison of MOSAIC with extreme λ threshold at υ=0 against the single-prototype PNet baseline, under CCN+ as the NeSy layer. Both models share a fixed annotation budget of 2 labeled images per character, corresponding to 2.16% of the full Omniglot dataset (1,623 characters / 32,460 images).

Model	Acc(C)↑	F1(C)↑	Acc(Y)↑	F1(Y)↑	Cls(C)↓
MOSAIC (υ=0)	**69.45 ± 1.02**	**65.51 ± 1.03**	**48.89 ± 1.47**	9.59 ± 0.31	**0.00 ± 0.00**
PNet	69.12 ± 0.92	65.12 ± 1.10	48.38 ± 1.21	**9.65 ± 0.33**	**0.00 ± 0.00**

Table 29: **MNIST-EvenOdd, MOSAIC vs. PNet, SL.** In-distribution comparison of MOSAIC against the single-prototype PNet baseline, under Semantic Loss as the NeSy layer. Models are trained with a fixed annotation budget of 2 labeled digit images per digit.

Model	Acc(C)↑	F1(C)↑	Acc(Y)↑	F1(Y)↑	Cls(C)↓
MOSAIC	**97.64 ± 0.54**	**97.54 ± 0.58**	**95.30 ± 1.06**	**20.41 ± 2.08**	**0.00 ± 0.00**
PNet	97.07 ± 1.31	96.87 ± 1.53	94.20 ± 2.53	20.07 ± 1.99	**0.00 ± 0.00**

Table 30: **MNIST-EvenOdd, MOSAIC vs. PNet, Scallop.** In-distribution comparison of MOSAIC against the single-prototype PNet baseline, under Scallop as the NeSy layer, both with the standard conv4 backbone. As in Omniglot, models are trained with a fixed annotation budget of 2 labeled digit images per digit.

Model	Acc(C)↑	F1(C)↑	Acc(Y)↑	F1(Y)↑	Cls(C)↓
MOSAIC	**97.68 ± 0.61**	**97.56 ± 0.68**	**95.40 ± 1.14**	**20.76 ± 1.50**	**0.00 ± 0.00**
PNet	96.46 ± 1.54	96.22 ± 1.81	93.08 ± 2.90	19.55 ± 1.46	**0.00 ± 0.00**