COMPASS: COVARIANCE-MAPPED PROTOTYPES WITH ANISOTROPIC SPHERICAL SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Case-based reasoning networks like ProtoPNet ground their predictions in visual similarities between image patches and learned concepts that can be visually audited by users, offering an alternative to post-hoc explanation methods. Recent advances represent prototypes as mixtures of Gaussian distributions to capture richer similarity structures, but the nature of these representations breaks the canonical projection onto image patches that makes prototypes interpretable. Separately, architectures using cosine similarity over Euclidean distance have demonstrated improved accuracy on fine-grained recognition, but their probabilistic relaxation remains restricted to isotropic von Mises-Fisher distributions. In this work, we propose a family of anisotropic spherical distributions – including the Kent, Anisotropic Spherical Gaussian (ASG), Angular Central Gaussian (ACG), and Symmetric Mahalanobis Cosine (SMC) distributions – as prototype similarity measures for cosine-based prototype networks, alongside principled volume regularizers that prevent geometric collapse during training. Across fine-grained image classification benchmarks, our anisotropic models consistently outperform isotropic baselines, improving accuracy by up to 4.7% on Stanford Dogs, while preserving the interpretability of prototype-based reasoning. Ultimately, our findings demonstrate that expressive covariance-aware metrics are crucial for closing the gap between interpretable-by-design and black-box models.

1 INTRODUCTION

As artificial intelligence systems' adoption rapidly increases, and as their capabilities quickly improve, the need for trustworthy AI systems becomes vital. Black-box machine learning models are highly expressive and display remarkable performance but are untrustworthy in high-stakes domains like banking, medical applications, and criminal justice (Rudin, 2019). Furthermore, this social need had been translated into law; the "right to an explanation" (European Parliament and Council of the European Union, 2024).

As such, the research community has dedicated efforts to post-hoc explanations (e.g., gradients, random smoothing, and feature attribution). However, these methods often focus on salient objects to appear credible to humans, remaining ultimately unfaithful to the model's true internal logic (Adebayo et al., 2018; Hedström et al., 2024). Therefore, the community has pushed forward transparent models, interpretable-by-design, which intrinsically explain their reasoning as they take their decisions. While these interpretable models somewhat solve this fidelity issue (Nauta et al., 2021a; Xu-Darme et al., 2023), they often incur an accuracy penalty (Bell et al., 2022; Crook et al., 2023).

Prototypical-Parts Networks [ProtoPNet] (Chen et al., 2019) is the root of the family of interpretable-by-design models which rely on the "*this* looks like *that*" case-based reasoning. Its core principle is to rely on an instance's similarity to learned prototypes, anchored to patches of samples from the training dataset. The similarity score between a prototype and a sample provides a justification for the decision, enabling faithfulness.

For example, as pictured in Figure 1, a prototype may represent a specific part of the body of the car. ProtoPNet's interpretability hinges entirely on this exact patch projections: the prototype is anchored to a part of a real image. However, it relies on ℓ_2 distance—or cosine similarity in a variant called TesNet (Wang et al., 2021)—which severely limits the expressivity of the prototypes:

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.

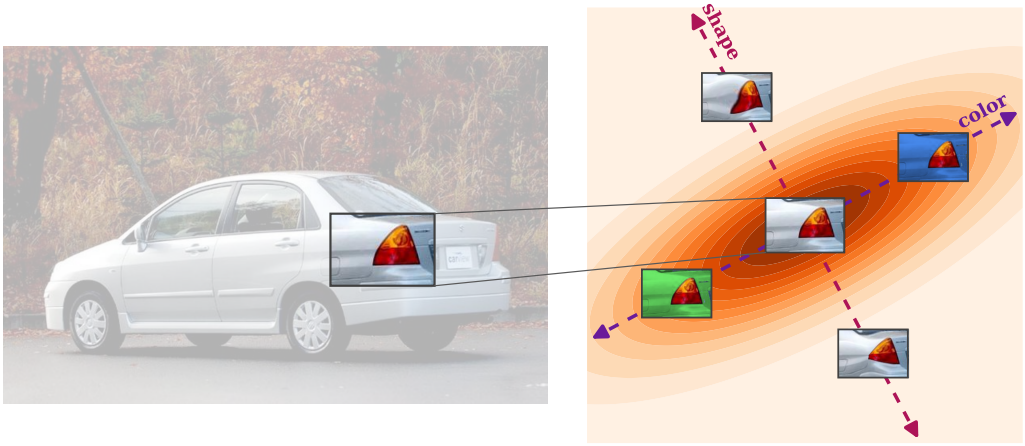

Figure 1: Conceptual illustration of an anisotropic similarity landscape centered on a real image patch from Stanford Cars (Krause et al., 2013). Equal-magnitude edits along the low-precision axis (changing the colour of the car) barely change similarity to the prototype, while edits along the high-precision axis (warping the taillight) collapse it. Isotropic metrics enforce a rigid circle, failing to capture this decoupling.

they are represented by an image patch, but do not express what is meaningful (e.g., in our toy example, the shape), nor what is meaningless (e.g., in our toy example, the colour).

Some glass-box models have addressed this problem by adopting probabilistic prototype distributions, using Gaussian mixtures (Wang et al., 2025; Moradinasab et al., 2024). While these ℓ_2 variants have partially closed the gap with state-of-the-art opaque models, they have to some extent traded the inherent interpretability of these prototype-based networks for accuracy: it is mathematically impossible to properly project a prototype onto a real image patch, the displayed patch is only a reflection of the prototype, not the prototype itself.

Other probabilistic prototype-based models (Li et al., 2024) have shifted the gaussian on the hypersphere directly, but stay stuck at isotropic von Mises-Fisher distributions. As we illustrate in Section 3.1, the natural workaround, applying a precision matrix, harbors a mathematical trap: a prototype does not activate itself.

To resolve this tension, our objective is to establish a mathematically rigorous framework that grants prototype networks the geometric expressivity of anisotropic covariance, while strictly preserving the foundational axiom of exact patch projection. Towards this end, our contributions are as follows:

- We diagnose existing probabilistic prototype-based approaches and identify two flaws: impossible projection for Gaussian mixtures, and peak-displacement for (anisotropic) vMF.
- We propose four desiderata for prospective anisotropic spherical prototype distributions.
- We introduce COvariance-Mapped Prototypes with Anisotropic Spherical Similarity (COMPASS), a family of four mathematically sound spherical distributions that warp the similarity space, and motivate them around the proposed desiderata.
- We show how to train COMPASS networks with these anisotropic spherical distributions, by using Cholesky decomposition to represent the precision matrix, and regularizing with Kullback-Leibler divergence.
- We thoroughly and empirically evaluate the performance of baseline vs isotropic vs anisotropic models, both with ℓ_2 distance and cosine similarity, on a variety of datasets and backbones. To avoid hyperparameter sensitivity to the benchmark (Musgrave et al., 2020), we employ a Hyperband search (Li et al., 2018) to select the most favourable hyperparameters for the baselines and novel networks.

2 Related Works

Case-Based Reasoning in Vision Prototype-based networks have been introduced as a family of self-explainable image classification models. ProtoPNet (Chen et al., 2019) is the foundational architecture and it operates as follows: the image is passed through an encoder – usually composed of a traditional neural network for its "backbone", and extra convolutional blocks as its "neck" – from there, the latent representation is decomposed into multiple patches, each of which is compared to a set of learned prototypes, which are grounded in representations of real image patches, and based on these (max) similarities, a weighted sum provides class logits. The interpretability of this architecture hinges on the faithful one-to-one mapping between a learned prototype and the latent representation of a single training patch during a projection step. The similarity scores are originally derived from ℓ_2 distance, and subsequent structural improvements to the paradigm like Hierarchical ProtoPNet (Hase et al., 2019), ProtoPShare (Rymarczyk et al., 2021), ProtoTree (Nauta et al., 2021b) and ProtoPool (Rymarczyk et al., 2022), while improving the accuracy and expressivity of the models, continue to rely on rigid Euclidean geometry.

Cosine and Spherical ProtoPNets A critical issue with ℓ_2 distance is that it is unsuited for neural networks, which prefer to encode semantic information directionally rather than positionally (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2023). In ProtoPNets, Euclidean prototypes have been shown to be prone to feature collapse or weak to background magnitude (Choukali et al., 2024; Sinhamahapatra et al., 2024; Parise & Namee, 2025). In the broader deep learning community, approaches have favoured reducing vectors to the unit sphere (Wang & Isola, 2020; Mohan et al., 2023), and relying on similarity as angles rather than Euclidean distance. As such, TesNet (Wang et al., 2021) fixes this by relying on cosine similarity instead, directly addressing the geometric limitations of ℓ_2 distance, and that switch has led to significant improvements over fine-grained classification (Moffett et al., 2026). Subsequent architectures have since then relied on this similarity function (Donnelly et al., 2022; Wang et al., 2023; Jia et al., 2026; Carretero et al., 2026).

Probabilistic ProtoPNets A different route taken to improve the performance of prototype-based models has been to make prototypes more expressive, and take the form of a probability distribution rather than a point (Li et al., 2024). Originally, the idea was to learn per-class covariances and apply them to few-shot classification (Fort, 2017). Then, this probabilistic approach was carried over to cross-domain generalization (Khanday & Sofi, 2022), incremental learning (Hu et al., 2026), and 3D affordance detection (Li et al., 2025). Gerstenberger et al. (2023) introduce a differentiable Gaussian prototype layer for explainable segmentation; they still rely on single Gaussians[1], so projection to a single patch remains well-defined. MGProto (Wang et al., 2025) replaces each prototype with a Gaussian mixture; ProtoGMM (Moradinasab et al., 2024) applies the same idea to segmentation. As argued later in Section 3.1, a multi-centroid prototype has no single representative point to refer to when projecting onto a patch; MGProto acknowledges this, and reverts to displaying the nearest patch, while ProtoGMM drops it. ProtoConcepts (Ma et al., 2023) also treat a prototype as a region of the latent space rather than a point, and visualize it with several training patches; COMPASS instead keeps a single, exactly projected center and encodes the shape of the region in Λ_j. Among probabilistic prototypes, only Li et al. (2024; 2025) move to the hypersphere, bringing von Mises–Fisher prototypes (Fisher, 1953) into case-based reasoning networks; this combines the angular geometry of TesNet with probabilistic prototypes, but only in isotropic form.

Anisotropic distributions on the sphere Directional statistics offers several anisotropic generalizations of the von Mises–Fisher distribution: the Fisher–Bingham family and its Kent submodel (Kent, 1982), the angular central Gaussian (Tyler, 1987), the elliptically symmetric angular Gaussian (ESAG; Paine et al., 2018), and the scaled von Mises–Fisher distribution (Scealy & Wood, 2019); tangent-space normals are obtained by pushing a Gaussian through the logarithmic map (Pennec, 2006; Hauberg, 2018; Papillon et al., 2025), and anisotropic spherical Gaussians are widely used in graphics (Xu et al., 2013). A recurring device in these constructions is to make the mean direction a principal axis of the dispersion: the major and minor axes of the Kent distribution are orthogonal to its mean direction, and ESAG constrains its covariance V so that $V\mu = \mu$. COMPASS applies this device to prototype similarities, where it is precisely what preserves exact projection (Section 4).

[1] The paper only extends to diagonal anisotropy in practice.

3 THE GEOMETRY OF INTERPRETABLE PROTOTYPES

The quest for higher accuracy in case-based reasoning networks has driven a shift from rigid, isotropic distance metrics toward highly expressive anisotropic spaces. As illustrated in Figure 1, a robust visual prototype must differentiate between core semantic features (which require strict similarity penalties) and nuisance features like background context (which require high variance tolerance). While learning a covariance structure elegantly solves this problem, previous attempts to integrate it into prototype networks have inadvertently broken their foundational interpretability guarantees.

3.1 THE BREAKDOWN OF CANONICAL PROJECTION

Case-based reasoning networks rely on a strict geometric assumption: a prototype must correspond to a specific, localized patch from the training set. Recent probabilistic relaxations violate this premise, sacrificing true interpretability for classification performance.

The Gaussian Mixture Problem. Recent architectures model prototypes as mixtures of Gaussians to capture diverse visual appearances. However, a multi-centroid prototype cannot physically correspond to a single image patch. When projecting the prototype back into pixel space, finding the "nearest" patch requires aggregating across multiple disjoint modes, breaking the exact one-to-one conceptual mapping that makes traditional prototype networks interpretable.

The vMF Displaced Vector Problem. In section 2, we mentioned isotropic hyperspherical prototype-based networks using a von Mises-Fisher (vMF) distribution. A natural anisotropic extension of this existing distribution would relax the precision matrix $\Lambda_j \in \mathbb{S}^D_{++}$. Given $z \in \mathbb{S}^{D-1}$ a unit-norm latent feature vector and $\mu_j \in \mathbb{S}^{D-1}$ the base prototype direction, the similarity evaluates the unnormalized inner product: $S_{\text{vMF}}(z, \mu_j) = z^T \Lambda_j \mu_j$. However, we show in Proposition 1 that this similarity is not maximized at the prototype μ_j but at another unit-norm vector $w_j = \Lambda_j \mu_j$, which we call the *effective prototype*.

Proposition 1. *Let $\Lambda_j \in \mathbb{S}^D_{++}$ and $S_{vMF}(z, \mu_j) = z^T \Lambda_j \mu_j$ for $z \in \mathbb{S}^{D-1}$. Then* $\operatorname{argmax}_z S_{vMF} = \Lambda_j \mu_j / \|\Lambda_j \mu_j\|$*, which equals μ_j if and only if μ_j is an eigenvector of Λ_j.*

Proof. By Cauchy-Schwarz, the inner product $z^T(\Lambda_j \mu_j)$ on the unit sphere is strictly maximized when z is parallel to the effective prototype $w_j = \Lambda_j \mu_j$. Therefore, $z = w_j / \|w_j\|$. This maximum aligns with the original prototype μ_j only when $\Lambda_j \mu_j \propto \mu_j$, which is the definition of an eigenvector. □

This is illustrated on Figure 2. Consequently, if we evaluate the prototype against its own base patch ($z = \mu_j$), it does *not* yield the maximum possible similarity unless the learned precision matrix perfectly aligns with the prototype vector (e.g., an isotropic $\Lambda_j = \lambda_j I$). Applying a general precision matrix geometrically displaces the peak away from the image patch, destroying the core definition of a visual prototype.

3.2 DESIDERATA FOR PROTOTYPE SIMILARITY

To preserve interpretability while expanding geometric expressivity, we claim that an anisotropic spherical similarity metric should satisfy four conditions:

1. **Maximized at Self-Identity:** Evaluating a feature vector identical to the prototype's peak must yield the global maximum similarity ($S(\mu_j, \mu_j) = \max_z S(z, \mu_j)$).
2. **Stable and Bounded:** The metric must not collapse or produce unbounded activations, which destabilize heavily regularized prototype learning objectives ($-\infty < \min_z S(z, \mu_j) < \max_z S(z, \mu_j) < \infty$).
3. **Unambiguous Projection (Unimodality):** The metric must possess a unique global maximum on the unit hypersphere $\mathbb{S}^{D-1}$. For directional concepts, the strict global maximizer must be the prototype itself ($\operatorname{argmax}_z S(z, \mu_j) = \{\mu_j\}$). For axial metrics, the maximizers

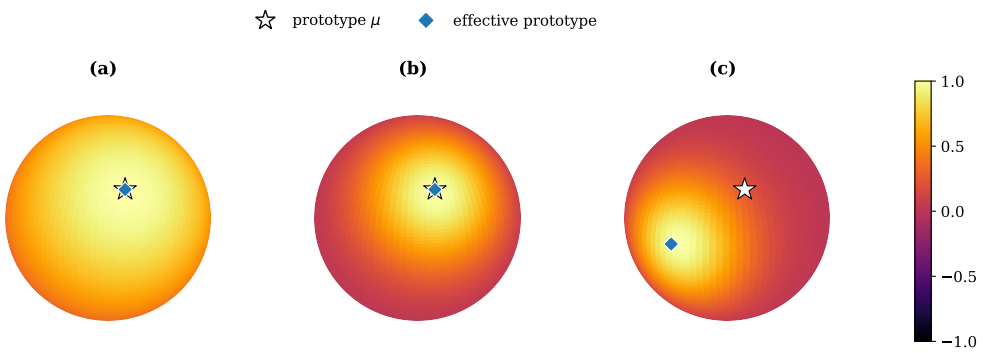

Figure 2: Cosine similarity (a), isotropic vMF (b), and anisotropic vMF (c) on the unit sphere, with the prototype μ (white star) and the location of maximal similarity (blue diamond). In (a)–(b), the precision matrix is a scalar multiple of the identity, so the two coincide: the isotropic parameters only sharpen the similarity, they never move it. In (c), an anisotropic precision matrix rotates the effective prototype $\Lambda\mu$ away from μ, so the point of maximal similarity no longer corresponds to the prototype itself ; the prototype sits in a near-zero similarity region of its own similarity function.

must be strictly restricted to the antipodal pair ($\{\mu_j, -\mu_j\}$). This guarantees a deterministic projection to a single image patch.

4. **Isotropic Cosine Reduction:** To ensure backward compatibility with standard hyperspherical networks (Wang et al., 2021), the similarity function must reduce to (or become monotonically equivalent to) standard cosine similarity when the learned covariance/precision matrix is the identity matrix. For directional metrics, $S(z_1, \mu_j) > S(z_2, \mu_j) \iff z_1^T \mu_j > z_2^T \mu_j$. For axial metrics, $S(z_1, \mu_j) > S(z_2, \mu_j) \iff |z_1^T \mu_j| > |z_2^T \mu_j|$.

4 The COMPASS Framework

We propose four anisotropic spherical similarity functions, as illustrated on Figure 3, which satisfy the previously defined desiderata. In order to do so, we must refer to the projection of the feature z onto the tangent plane of the prototype, defined as $z_\perp = z - (z^T \mu_j)\mu_j$. Therefore, the precision matrix Λ_j (the inverse of the covariance matrix) acts exclusively on the orthogonal complement $z_\perp$ (which is identically zero at $z = \pm\mu_j$). By reasoning on $z_\perp$, we set the global maximum precisely at μ_j, and we avoid the displaced-peak issue identified in Section 3.1. These similarities are parametrized by a learnable precision matrix and a mean vector for each prototype; the training phase of prototype-based models relying on these similarities remains unchanged. In particular, projection moves only the mode μ_j onto the selected patch and keeps Λ_j unchanged; since D1–D3 hold for any mode and any $\Lambda_j \in \mathbb{S}^D_{++}$, the projected prototype satisfies the same desiderata as before projection. We present the four similarities in Section 4.1, we show how to represent the precision matrix to make them learnable in Section 4.2, and we derive a regularization term explained in Section 4.3. An overview of COMPASS is illustrated by Table 1, which briefly summarizes the compliance of existing and proposed prototype similarities with the claimed desiderata in Section 3.2.

4.1 Anisotropic Spherical Similarities

Symmetric Mahalanobis Cosine (SMC). SMC computes a warped geometric ratio strictly on the tangent plane:

$$S_{\text{SMC}}(z, \mu_j; \Lambda_j) = \frac{z^T \mu_j}{\sqrt{(z^T \mu_j)^2 + z_\perp^T \Lambda_j z_\perp}} \tag{1}$$

Unlike standard Mahalanobis Cosine similarity used classically in pattern recognition (Bolme et al., 2003; Perlibakas, 2004; Nguyen & Bai, 2010), which applies Λ_j to both vectors and can therefore skew the similarity landscape asymmetrically around μ_j, SMC restricts the precision matrix to

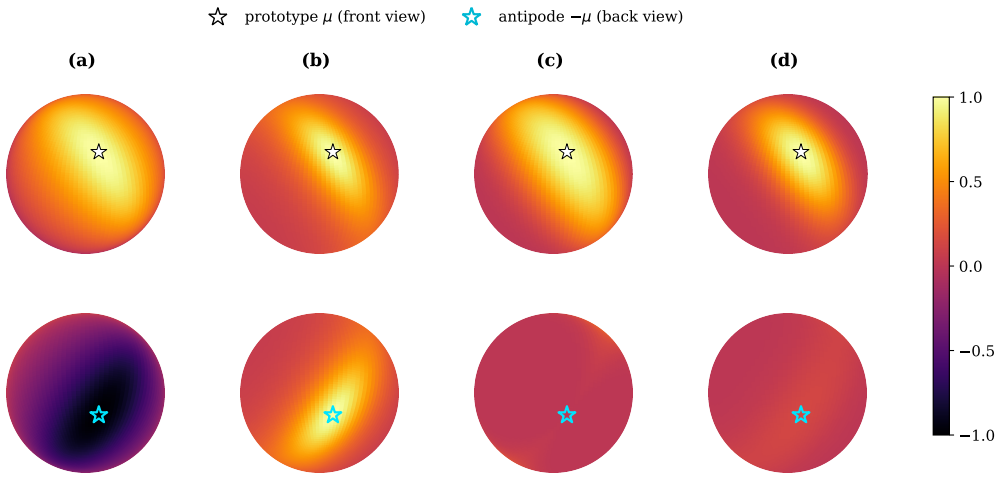

Figure 3: SMC (a), ACG (b), ASG (c), and Kent (d) on the unit sphere under the same prototype μ and anisotropic precision Λ, viewed from μ (top) and from the antipode $-\mu$ (bottom), on a shared color scale. All four share the same anisotropic elongation around μ, but differ sharply on the back hemisphere: SMC is signed and reaches -1 at $-\mu$; ACG is axial and reproduces its peak at $-\mu$; ASG nearly vanishes (e.g., $e^{-\pi^2/2} \approx 0.007$ for $\Lambda = I$); Kent takes the fixed value e^{-2} at the antipode.

the orthogonal complement. Note that scaling Λ_j alters the sharpness of the similarity. Thus, an isotropic SMC ($\Lambda_j = \lambda_j I$) acts as a learnable concentration parameter, whereas standard Cosine Similarity (baseline TesNet) is strictly fixed to the Identity matrix (I).

Angular Central Gaussian (ACG). To model axial concepts (Watson, 1965), we build on the Angular Central Gaussian (ACG) distribution (Tyler, 1987; Mardia & Jupp, 2000). Classical ACG is invariant to isotropic scaling of Λ_j, but its density is unbounded: the peak value is proportional to $\det(\Lambda_j)^{1/2}\lambda_{\min}^{-D/2} = \prod_i (\lambda_i/\lambda_{\min})^{1/2}$, which grows without bound as Λ_j becomes ill-conditioned. As with Gaussian likelihoods under learned covariance (Bishop, 2006), similarity scores can thus be inflated simply by collapsing the distribution onto a single axis, a failure mode known to make learned-covariance objectives poorly conditioned (Seitzer et al., 2022). To satisfy our boundedness desideratum (D2), in line with the bounded similarities used by ProtoPNet (Chen et al., 2019) and TesNet (Wang et al., 2021), we define:

$$S_{\text{ACG}}(z, \mu_j; \Lambda_j, \nu) = (1 + z_\perp^T \Lambda_j z_\perp)^{-\nu}, \tag{2}$$

where $\nu > 0$ is a hyperparameter. This is exactly a tempered, peak-normalized ACG density whose smallest precision eigenvalue, 1, lies along $\pm\mu_j$ (Appendix A.4). Hence $\pm\mu_j$ are the unique maximizers (D3, axial clause) and the peak value is exactly 1 regardless of the conditioning of Λ_j (D2). We use $\nu = 3/2$ in all experiments. Since sharpness is governed jointly by ν and Λ_j, and the KL term anchors Λ_j near I, the exact probabilistic value $\nu = D/2$ would make the kernel far sharper at initialization than the other similarities.

Anisotropic Spherical Gaussian (ASG). Unlike the previous similarities that use the chordal projection for efficiency, ASG (Xu et al., 2013) projects features using the Riemannian logarithmic map. Let $\theta = \arccos(z^T \mu_j)$. The tangent vector is scaled by the geodesic distance: $\text{Log}_{\mu_j}(z) = \frac{\theta}{\sin(\theta)} z_\perp$. The similarity is evaluated as a standard Gaussian density in this exact tangent space:

$$S_{\text{ASG}}(z, \mu_j; \Lambda_j) = \exp\left(-\frac{1}{2}\text{Log}_{\mu_j}(z)^T \Lambda_j \text{Log}_{\mu_j}(z)\right) \tag{3}$$

Since the logarithmic map is undefined at the antipode $-\mu_j$, we evaluate ASG with gating only, which restricts it to the hemisphere where the map is well-behaved.

Table 1: Desiderata compliance of prototype similarities (Section 3.2). D1–D3 must hold for every learned $\Lambda_j \in \mathbb{S}^D_{++}$; D4 is evaluated on the un-gated similarity at $\Lambda_j = I$. *Aniso.* indicates whether the resulting similarity can be anisotropic.

Similarity	Aniso.	D1 Self-identity	D2 Bounded	D3 Unique max.	D4 Cosine red.
Cosine / isotropic vMF	✗	✓	✓	✓	✓
"Anisotropic" vMF	✗$^\sharp$	✗	✓	✗	✓
Gaussian mixture	✓	✗	–	✗	–
SMC	✓	✓	✓	✓	✓
ACG	✓	✓	✓	✓$^\diamond$	✓$^\diamond$
ASG	✓	✓	✓	✓$^\circ$	✓
Kent	✓	✓	✓	✓	✗*

$^\sharp$ the resulting similarity is isotropic about the effective prototype $\Lambda_j \mu_j$ (Section 3).
$^\diamond$ Axial form: the maximizers are $\{\pm\mu_j\}$, and the ranking follows $|z^T \mu_j|$.
$^\circ$ On $\mathbb{S}^{D-1} \setminus \{-\mu_j\}$, where the log map is defined.
* Holds if and only if $\kappa \geq 2$; our experiments use $\kappa = 1$.
– Not applicable.

Kent Distribution. The Kent distribution (Kent, 1982) evaluates a localized density combining a fixed radial decay with a learned anisotropic tangent penalty:

$$S_{\text{Kent}}(z, \mu_j; \Lambda_j) = \exp\left(\kappa\left(z^T \mu_j - 1\right) - z_\perp^T \Lambda_j z_\perp\right) \tag{4}$$

Because scaling κ and Λ_j jointly by $t > 0$ only raises the similarity to the power t, the temperature κ need not be learned. Satisfying D4 at the prior $\Lambda_j = I$ requires $\kappa \geq 2$. Interestingly, because $z_\perp$ vanishes exactly at the antipode ($z = -\mu_j$), this single measure-zero point remains fixed at a similarity of $e^{-2\kappa}$ regardless of Λ_j (≈ 0.135 for $\kappa = 1$).

4.2 Parameterization and Numerical Stability

To study the trade-off between geometric expressivity and parameter count, we parameterize Λ_j in three capacities, each guaranteeing $\Lambda_j \in \mathbb{S}^D_{++}$ through unconstrained log-space parameters: *isotropic*, $\Lambda_j = e^{-v_j} I$ with $v_j \in \mathbb{R}$ (1 parameter per prototype; a spherical cluster); *diagonal*, $\Lambda_j = \operatorname{diag}(e^{-v_{j,1}}, \ldots, e^{-v_{j,D}})$ with $v_j \in \mathbb{R}^D$ (D parameters; axis-aligned ellipsoids); and *general*, $\Lambda_j = L_j L_j^T$ with L_j lower-triangular, $[L_j]_{dd} = e^{v_{j,d}}$, and unconstrained off-diagonal entries ($D(D+1)/2$ parameters; arbitrarily rotated ellipsoids). Since our four similarities are intentionally not scale-invariant (Appendix A.4 for ACG), even the isotropic case acts as a learnable concentration. Evaluating the tangent penalty through an explicit Λ_j can return small negative values in floating point when its true value is near zero (e.g., for the prototype's own patch after projection), producing NaNs downstream; we instead compute $z_\perp^T \Lambda_j z_\perp = \|L_j^T z_\perp\|_2^2$ (with $L_j = \Lambda_j^{1/2}$ in the isotropic and diagonal cases), which is non-negative by construction. Finally, a small change in a log-space parameter multiplies the precision, so diagonal and off-diagonal parameters have very different gradient scales; we optimize them as separate parameter groups with their own learning rates.

4.3 Regularization via KL Divergence

Left unconstrained, the precision can grow without bound along some directions (over-sharpening) or vanish along others. We anchor Λ_j to the identity I with the Gaussian divergence $\mathrm{KL}\big(\mathcal{N}(0, I) \,\|\, \mathcal{N}(0, \Lambda_j^{-1})\big)$:

$$\mathcal{L}_{\mathrm{KL}}(\Lambda_j) = \tfrac{1}{2}\big(\operatorname{Tr}(\Lambda_j) - \log\det\Lambda_j - D\big), \tag{5}$$

which vanishes if and only if $\Lambda_j = I$: the trace penalizes large precisions and $-\log\det\Lambda_j$ vanishing ones. Its weight λ_{KL} is tuned in Section 5, and allows the optimization process to smoothly trade off between strict isotropic rigidity (ProtoPNet in ℓ_2, TesNet in cosine similarity) and unconstrained anisotropic expressivity.

4.4 HEMISPHERE GATING

On the sphere, a prototype also responds to its back hemisphere, where cosine similarity is negative; whether this evidence should count against a class or be ignored is an open question, and cosine-based networks typically apply a ReLU (Moffett et al., 2026). We formalize this as a gate:

$$S_{\text{gated}}(z, \mu_j; \Lambda_j) = S(z, \mu_j; \Lambda_j) \cdot \mathbb{1}(z^T \mu_j \geq 0). \tag{6}$$

For SMC, this is exactly a ReLU and remains continuous. ACG, ASG, and Kent are positive everywhere: the gate removes their back-hemisphere evidence but makes them discontinuous at the equator (at $\Lambda_j = I$, jumps of $2^{-\nu} \approx 0.35$, $e^{-\pi^2/8} \approx 0.29$, and $e^{-2} \approx 0.14$, respectively). The continuous hemispherical factor $\max(z^T \mu_j, 0)$ of the original ASG (Xu et al., 2013) is a natural alternative, left to future work.

5 EVALUATION AND RESULTS

5.1 EXPERIMENTAL SETUP

To ensure performance gains stem from geometric expressivity rather than hyperparameter advantages, we conducted a Hyperband search (Li et al., 2018) across learning rates, prototype dimension D and spatial pooling configurations for all baselines (standard ProtoPNet & TesNet). We then injected our COMPASS similarities directly into the winning configurations to strictly isolate the effect of the similarity functions, and ran another Hyperband search on the additional hyperparameters: the learning rate of the (log-)diagonal precision parameters, that of the off-diagonal Cholesky entries (general precision only), and the KL coefficient λ_{KL}, including $\lambda_{\text{KL}} = 0$. We trained models on three fine-grained image classification datasets: CUB-200 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), and Stanford Dogs (Khosla et al., 2011). Hyperparameters are selected on a validation set holding out 20% of the training data. The Mahalanobis ℓ_2 variant replaces $\|z - \mu_j\|^2$ by $(z - \mu_j)^T \Lambda_j (z - \mu_j)$ in ProtoPNet's similarity. All prototype-based models are trained using the CaBRNet framework (Xu-Darme et al., 2024).

5.2 ANISOTROPIC PERFORMANCE EVALUATION

Table 2 reports our experimental results with a ResNet-50 (He et al., 2016) backbone architecture for the prototype-based networks (Appendix A.3 shows similar findings on additional backbones). We observe that a learnable concentration alone often improves accuracy: isotropic precision, which adds a single parameter per prototype, yields gains of up to 3.6% over the baseline (gated SMC on Stanford Dogs), in line with the isotropic vMF baseline. Anisotropy provides further expressive power and performance on top of this, reaching up to 4.7% over the baseline and 1.9% over vMF (gated Kent with general precision on Stanford Dogs). Furthermore, diagonal and general precision perform comparably: excluding failed runs (†/‡), the diagonal models match or outperform the general ones in 8 of 18 comparisons across both geometries, with a mean difference of less than 0.1%. The diagonal parameterization thus emerges as the empirical sweet spot: it adds only $\mathcal{O}(DP)$ learnable parameters (0.3–4.3% of the backbone across our settings, with D from 64 to 512), whereas the general setting adds $D(D+1)/2$ parameters per prototype, from 11% of the backbone ($D = 64$) to about 11 times its size ($D = 512$ on Stanford Cars). Across all three datasets and both gating settings, the best-performing spherical model is a COMPASS similarity, outperforming both TesNet and the isotropic vMF baseline; Kent reaches the best spherical accuracy on all three datasets, improving over TesNet with the same gating by 2.5%, 4.4%, and 4.7% on CUB-200 (ungated), Stanford Cars (ungated), and Stanford Dogs (gated).

6 CONCLUSION AND FUTURE WORK

Conclusion. In this work, we identified a tension between the expressivity sought by recent probabilistic prototype-based models and the grounding that makes them interpretable, and diagnosed two failure modes: Gaussian-mixture prototypes can only be *reflected* onto a training patch, not *projected*, and anisotropic vMF prototypes are not maximally similar to themselves (Proposition 1). From these, we derived four desiderata for interpretable anisotropic similarities, which discriminate

Table 2: Test accuracy (%) of baselines and COMPASS similarities on CUB-200, Stanford Cars, and Stanford Dogs with a ResNet-50 backbone. Euclidean rows use ProtoPNet; spherical rows use TesNet. *Gated* sets the similarity to zero on the back hemisphere ($z^T\mu_j < 0$); *Ungated* keeps the full similarity. *Best* Δ is the largest gain over the corresponding baseline (ProtoPNet, or TesNet with the same gating), across covariances and datasets. Within each geometry, the best result per column is in **bold** and the second best is underlined.

			CUB-200		Stanford Cars		Stanford Dogs		
Geometry	**Similarity**	**Covariance**	Gated	Ungated	Gated	Ungated	Gated	Ungated	**Best Δ**
Euclidean	ℓ_2 (ProtoPNet)	Identity	87.2		86.9		80.3		
	Mahalanobis ℓ_2	Isotropic	<u>88.4</u>		<u>87.5</u>		**81.5**		
		Diagonal	87.8		**87.7**		<u>81.2</u>		+1.9 (CUB)
		General	**89.1**		86.9		80.2		
Spherical	Cosine (TesNet)	Identity	85.7	85.4	82.7	83.1	78.8	79.4	
	vMF	Isotropic	<u>85.8</u>	85.8	81.7	83.1	81.6	80.0	
	SMC	Isotropic	85.7	85.8	83.5	84.0	82.4	76.5‡	
		Diagonal	<u>85.8</u>	84.2‡	**84.9**	85.8	76.0‡	80.3	+3.8 (Dogs)
		General	**86.1**	87.4	<u>84.7</u>	86.5	82.6	**80.7**	
	ACG	Isotropic	82.2	82.4‡	0.7†	80.2	81.9	80.3	
		Diagonal	84.0	86.7	14.5†	<u>86.9</u>	82.0	**80.7**	<u>+4.6 (Dogs)</u>
		General	29.8‡	<u>87.6</u>	74.5	84.8	<u>83.4</u>	79.9	
	ASG	Isotropic	81.9	–	0.8†	–	81.1	–	
		Diagonal	84.3	–	0.8†	–	82.5	–	+3.7 (Dogs)
		General	85.5	–	0.8†	–	80.8	–	
	Kent	Isotropic	83.5	85.9	3.1†	85.0	82.1	80.5	
		Diagonal	84.9	87.4	7.2†	**87.5**	82.8	<u>80.6</u>	**+4.7 (Dogs)**
		General	85.7	**87.9**	80.2	86.0	**83.5**	80.1	

† Failed to learn (near-chance accuracy). ‡ Training diverged. – Not applicable.

between existing methods (Table 1). COMPASS answers them with a single construction: by letting the precision matrix act only on the component orthogonal to the prototype, it reshapes the similarity without moving its peak, so that projection remains exact. This yields four similarities with distinct back-hemisphere behavior, which satisfy the desiderata up to the caveats of Table 1. Empirically, a COMPASS similarity is the best spherical model across different dataset, gating, and backbone settings (Table 2 and Appendix A.3), with gains of up to 4.7% over TesNet and 1.9% over the isotropic vMF baseline. Diagonal precision captures most of this benefit, whereas general precision brings no systematic further gain for its $\mathcal{O}(D^2P)$ additional parameters. Overall, we show that anisotropy is a principled and inexpensive inductive bias for prototype-based networks: the model learns which feature directions matter for each prototype, without giving up exact projection.

Future work. Throughout, "preserving interpretability" refers to a structural property: each prototype is exactly a training patch located at the center of its similarity function; we did not evaluate explanation quality and leave this as open directions for further research. Perhaps the most promising direction opened by COMPASS is that the learned covariance carries geometric meaning beyond classification: the principal axes of Λ_j encode which feature directions a prototype tolerates and which it insists on, offering a new form of explanation—"this looks like that, *regardless of color*"—that traditional point prototypes cannot express (Figure 1). Visualizing this covariance structure in a readable way for users is worth pursuing. Concerning formal explainability, the boundedness guaranteed by D2 makes COMPASS similarities natural candidates for provable robustness certificates: abductive latent explanations (Soria et al., 2026a;b), which currently assume a single shared metric, could be extended to prototypes with distinct precision matrices via cross-metric bounds and ellipsoid-intersection approximations.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Gemini) to provide critical ingredients for proving mathematical claims (the tempered ACG density interpretation, the scale-equivalence argument for Kent's κ), assist in the writing of proofs (the peak-normalized ACG derivation and the Kent monotonicity condition), implement methods (drafting PyTorch similarity layers and regularization terms).

We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, propose or refine hypotheses, help develop theoretical models or conceptual frameworks, assist with translation, clean or reformat datasets, or support qualitative and thematic data analysis, as these tasks were either performed entirely by the authors or are not applicable to this work.

Additionally, we used generative AI tools for several recommended-disclosure tasks, including drafting and editing parts of the research paper to improve readability and meet page limits, identifying and summarizing relevant literature, brainstorming and structuring arguments, and creating scientific figures and visualizations (Figures 1–3).

We have carefully reviewed all AI-assisted work. All mathematical derivations were independently verified step-by-step by the authors. All AI-generated code was reviewed, tested, and debugged; several numerical-stability issues were identified and corrected during this process. All cited papers were manually surveyed to confirm accuracy and relevance; in multiple instances, AI-provided characterizations of prior work were corrected by the authors. The empirical analysis is strictly grounded in the raw outputs of our own experiments. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to advance trustworthy AI by improving the interpretability of image classification models. By strictly grounding predictions in single, projectable prototypes, COMPASS preserves the ability of domain experts to audit model reasoning—a crucial property in high-stakes applications where opaque models pose accountability risks.

However, we emphasize that interpretability is not a guarantee of correctness or fairness. A model that explains its reasoning transparently can still learn spurious correlations or reflect biases present in the training data. While prototype networks make these biased representations explicitly auditable (serving as a diagnostic tool rather than hiding the bias in uninterpretable weights), human users must guard against automation bias. Users should not treat visually plausible prototype explanations as sufficient evidence for consequential decisions without independent validation.

Finally, all datasets used in this work (CUB-200, Stanford Cars, Stanford Dogs, Oxford Flowers-102, Oxford-IIIT Pet) are publicly available benchmarks with no known privacy or consent concerns, and no human subjects were involved. The image used in Figure 1 is drawn from the Stanford Cars dataset and is used under its original license.

REPRODUCIBILITY STATEMENT

As part of the ICLR submission, the authors provide `similarities.py`, a modified Python script compatible with minimal code changes to the library CaBRNet (Xu-Darme et al., 2024). The file shows precisely how we implement each similarity function, each KL-divergence regularization function associated with it, and how we represent the precision matrix for each setting: isotropic, anisotropic-diagonal, and anisotropic-general (see Section 4.2).

We further provide:

- A script to generate the Hyperband model propositions (see Appendix A.1)
- A script to run the experiments and iteratively prune the models (see Appendix A.1).
- A script to visualize all COMPASS functions.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 248–266, 2022.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

David S Bolme, J Ross Beveridge, Marcio Teixeira, and Bruce A Draper. The csu face identification evaluation system: its purpose, features, and structure. In *International Conference on Computer Vision Systems*, pp. 304–313. Springer, 2003.

Ilán Carretero, Gustavo Jesús Angulo, Rocío del Amor, and Valery Naranjo. Opal: Orthonormal prototype alignment learning for interpretable image classification. *arXiv preprint arXiv:2608.30003*, 2026.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mohammad Amin Choukali, Mehdi Chehel Amirani, Morteza Valizadeh, Ata Abbasi, and Majid Komeili. Pseudo-class part prototype networks for interpretable breast cancer classification. *Scientific Reports*, 14(1):10341, 2024.

Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pp. 316–324. IEEE, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Jon Donnelly, Alina J Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using spatial deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`.

Ronald Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A*, 217 (1130):295–305, 1953.

Stanislav Fort. Gaussian prototypical networks for few-shot learning on omniglot. *arXiv preprint arXiv:1708.02735*, 2017.

Moritz Gerstenberger, Sebastian Maaß, Peter Eisert, and Sebastian Bosse. A differentiable gaussian prototype layer for explainable segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023.

Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pp. 32–40, 2019.

Søren Hauberg. Directional statistics with the spherical normal distribution. In *2018 21st international conference on information fusion (FUSION)*, pp. 704–711. IEEE, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. A Fresh Look at Sanity Checks for Saliency Maps, 2024.

Chenxi Hu, Yifan Hu, Yunxiang Zhao, Tao Wu, Meng Zhang, Chunsheng Liu, and Yangyi Hu. Gaussian-augmented prototypical network for class-incremental few-shot relation classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2026.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

Junhao Jia, Jiaqi Wang, Yunyou Liu, Haodong Jing, Yueyi Wu, Xian Wu, and Yefeng Zheng. This looks distinctly like that: Grounding interpretable recognition in stiefel geometry against neural collapse. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

John T Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

Nuaman Khanday and S. A. Sofi. Learned gaussian protonet for improved cross-domain few-shot classification and generalization. *Neural Computing and Applications*, 35(10):7481–7495, 2022.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Maximilian X. Li, K. F. Rudolf, N. Blank, and R. Lioutikov. An overview of prototype formulations for interpretable deep learning. *arXiv preprint arXiv:2410.08925*, 2024.

Maximilian X. Li, K. Rudolf, N. Blank, and R. Lioutikov. Interpretable affordance detection on 3d point clouds with probabilistic prototypes. *arXiv preprint arXiv:2504.18355*, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36:39212–39235, 2023.

Kanti V Mardia and Peter E Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

Luke Moffett, Frank Willard, Maximillian Machado, Emmanuel Mokel, Jon Donnelly, Zhicheng Guo, Adam Costarino, Julia Yang, Giyoung Kim, Alina Jade Barnett, et al. Cosine similarity is almost all you need (for prototypical-part models). In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1586–1596. IEEE, 2026.

Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Deep metric learning for computer vision: A brief overview. *Handbook of Statistics*, 48:59–79, 2023.

Nazanin Moradinasab, Laura S Shankman, Rebecca A Deaton, Gary K Owens, and Donald E Brown. Protogmm: Multi-prototype gaussian-mixture-based domain adaptation model for semantic segmentation. In *International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2024.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 441–456. Springer, 2021a.

Meike Nauta, Ron van Bree, and Philipp Seegerer. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pp. 709–720. Springer, 2010.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of intra-class varieties. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, pp. 722–729. IEEE, 2008.

Phillip J Paine, Simon P Preston, Michail Tsagris, and Andrew TA Wood. An elliptically symmetric angular gaussian distribution. *Statistics and Computing*, 28(3):689–697, 2018.

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, and Nina Miolane. Beyond euclid: An illustrated guide to modern machine learning with geometric, topological, and algebraic structures. *Machine Learning: Science and Technology*, 6(3):031002, 2025.

Alec Parise and Brian Mac Namee. This looks like that and that and that: Multi-objective optimization for diverse prototype learning. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 49–62. Springer, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Vytautas Perlibakas. Distance measures for pca-based face recognition. *Pattern recognition letters*, 25(6):711–724, 2004.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

Dawid Rymarczyk, Łukasz Struski, Michał Górska, Jacek Tabor, and Bartosz Zieliński. Protopool: Interpretable image classification with pooled prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

JL Scealy and Andrew TA Wood. Scaled von mises–fisher distributions and regression models for paleomagnetic directional data. *Journal of the American Statistical Association*, 2019.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022.

Poulami Sinhamahapatra, Suprosanna Shit, Anjany Sekuboyina, Malek Husseini, David Schinz, Nicolas Lenhart, Joern Menze, Jan Kirschke, Karsten Roscher, and Stephan Guennemann. Enhancing interpretability of vertebrae fracture grading using human-interpretable prototypes. *arXiv preprint arXiv:2404.02830*, 2024.

Jules Soria, Zakaria Chihani, Julien Girard-Satabin, Alban Grastien, Romain Xu-Darme, and Daniela Cancila. Formal abductive latent explanations for prototype-based networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a.

Jules Soria, Alban Grastien, Romain Xu-Darme, Julien Girard-Satabin, Zakaria Chihani, and Daniela Cancila. Beyond L_2: Generalizing abductive latent explanations to diverse prototype-based architectures. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 116–133. Springer, 2026b.

David E Tyler. Statistical analysis for the angular central gaussian distribution on the sphere. *Biometrika*, 74(3):579–589, 1987.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Helen Frazer, and Gustavo Carneiro. Learning support and trivial prototypes for interpretable image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2062–2072. IEEE, 2023.

Chong Wang, Yuanhong Chen, Fengbei Liu, Yuyuan Liu, Davis James McCarthy, Helen Frazer, and Gustavo Carneiro. Mixture of gaussian-distributed prototypes with generative modelling for interpretable and trustworthy image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6974–6989, 2025.

Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Geoffrey S Watson. Equatorial distributions on a sphere. *Biometrika*, 52(1/2):193–201, 1965.

Kun Xu, Wei-Lun Sun, Zhao Dong, Dan-Yong Zhao, Run-Dong Wu, and Shi-Min Hu. Anisotropic spherical gaussians. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013.

Romain Xu-Darme, Georges Quénot, Zakaria Chihani, and Marie-Christine Rousset. Sanity checks for patch visualization in prototype-based image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023.

Romain Xu-Darme, Aymeric Varasse, Alban Grastien, Julien Girard-Satabin, and Zakaria Chihani. Cabrnet: A case-based reasoning network infrastructure. In *Proceedings of the World Conference on eXplainable Artificial Intelligence*. Springer, 2024.

A Appendix

A.1 Hyperparameter Search

All models are tuned in two stages with Hyperband. Stage 1 tunes the baselines (ProtoPNet with ℓ_2, TesNet with cosine similarity) for each dataset, architecture, and backbone. Stage 2 starts from the best Stage 1 configuration and tunes only the hyperparameters introduced by the covariance (Table 4). All other settings are inherited unchanged.

Table 3: Stage 1 search space (baselines). 16 configurations are sampled uniformly without replacement per dataset, architecture, and backbone.

Hyperparameter	Values
Latent grid (adaptive average pooling)	1×1, 2×2, 3×3, 4×4
Prototype dimension D	64, 128, 256, 512
Backbone learning rate	10^{-5}, $5{\cdot}10^{-5}$, 10^{-4}, $3{\cdot}10^{-4}$, $5{\cdot}10^{-4}$
Add-on and prototype learning rate	$10\times$ backbone learning rate

Table 4: Stage 2 search space (covariance). The search is stratified by covariance type, with 8 configurations sampled without replacement per type; each sampled configuration is evaluated with every similarity of the corresponding architecture. Off-diagonal hyperparameters apply to the general covariance only.

Hyperparameter	Values
Covariance type	isotropic, diagonal, general
(Log-)diagonal learning rate	10^{-4}, $5{\cdot}10^{-4}$, 10^{-3}, $5{\cdot}10^{-3}$, 10^{-2}
(Log-)diagonal weight decay	0, 10^{-4}, 10^{-3}
Off-diagonal learning rate	$\{0.1, 0.2, 0.5, 1.0\}\times$ diagonal learning rate
Off-diagonal weight decay	0, 10^{-3}, 10^{-2}, $5{\cdot}10^{-2}$
KL coefficient λ_{KL}	0, 10^{-4}, 10^{-3}, $5{\cdot}10^{-3}$, 10^{-2}
Similarities (ProtoPNet)	Mahalanobis ℓ_2
Similarities (TesNet, gated)	vMF (isotropic only), SMC, ACG, ASG, Kent
Similarities (TesNet, ungated)	vMF (isotropic only), SMC, ACG, Kent

Table 5: Training schedules. At each Hyperband rung, the best half of the configurations is promoted.

Datasets	Warm-up	Learning-rate schedule	Hyperband rungs (epochs)
Flowers-102, Oxford-IIIT Pet	5 epochs	Cosine ($T_{\max} = 20$)	8, 15, 25, 30
Stanford Cars, Stanford Dogs	5 epochs	Step (every 10 epochs, $\times0.1$)	15, 30, 45, 60
CUB-200	10 epochs	Cosine ($T_{\max} = 50$)	15, 30, 45, 60

Fixed settings. All models use AdamW (Loshchilov & Hutter, 2019) with 10 prototypes per class and a batch size of 32 (16 for DenseNet-161, with learning rates scaled linearly). Weight decay is 0.05 for ConvNeXt backbones and 10^{-3} for others, $3{\cdot}10^{-4}$ for the add-on layers (10^{-3} during warm-up), and 0 for prototypes; the last layer uses a learning rate of $3{\cdot}10^{-3}$. Loss coefficients are 0.8 (clustering), -0.08 (separation), and 0.01 (last-layer L_1), plus -10^{-7} (subspace separation) for TesNet. Prototypes are projected from epoch 20 onward, each projection followed by 10 epochs of last-layer fine-tuning. Following Nauta et al. (2021b), ResNet-50 is initialized from weights pretrained on iNaturalist (Van Horn et al., 2018) for CUB-200 and Flowers-102; all other models are initialized from ImageNet (Deng et al., 2009).

A.2 Selected Hyperparameters

Table 6 reports the Stage 1 settings selected for each baseline in Table 2; all COMPASS models for a given dataset and architecture inherit these settings. Table 7 reports the covariance hyperparameters

selected in Stage 2. Two observations follow from these tables. First, the search allowed the latent feature map to be pooled down to a single location, and the 1×1 grid was selected in 7 of the 9 baseline settings; in these settings, each image is represented by a single latent vector, and prototypes are projected onto entire training images rather than image regions. Second, the KL coefficient $\lambda_{\mathrm{KL}} = 0$ was selected in 18 of the 78 configurations, including some of our best-performing models, indicating that the KL anchor is not required to reach the reported accuracy, although it was retained by the search in most settings.

Table 6: Hyperparameters selected for the ResNet-50 baselines of Table 2 (Stage 1). These settings are kept fixed for all COMPASS models of the same dataset and architecture; only the covariance hyperparameters are tuned in Stage 2.

Dataset	Architecture	Latent grid	D	Backbone LR
CUB-200	ProtoPNet	1×1	64	$5{\cdot}10^{-5}$
	TesNet (gated)	2×2	128	10^{-5}
	TesNet (ungated)	1×1	256	10^{-4}
Stanford Cars	ProtoPNet	1×1	128	10^{-4}
	TesNet (gated)	1×1	256	$5{\cdot}10^{-4}$
	TesNet (ungated)	1×1	512	$3{\cdot}10^{-4}$
Stanford Dogs	ProtoPNet	1×1	64	10^{-5}
	TesNet (gated)	1×1	128	10^{-5}
	TesNet (ungated)	3×3	256	$5{\cdot}10^{-5}$

Table 7: Stage 2 hyperparameters selected for the ResNet-50 models of Table 2, which inherit the baseline settings of Table 6. η_d: learning rate of the (log-)diagonal precision parameters; ρ: ratio of the off-diagonal to the diagonal learning rate; λ_{KL}: KL coefficient. G/U: gated/ungated.

			CUB-200			Stanford Cars			Stanford Dogs		
Similarity	Covariance	Gate	η_d	ρ	λ_{KL}	η_d	ρ	λ_{KL}	η_d	ρ	λ_{KL}
Mahalanobis ℓ_2	Isotropic	–	1e-2	–	0	5e-3	–	5e-3	1e-3	–	1e-2
	Diagonal	–	5e-3	–	1e-3	1e-2	–	1e-3	1e-3	–	1e-3
	General	–	1e-2	0.1	0	1e-2	0.1	0	1e-3	0.5	5e-3
vMF	Isotropic	G	1e-2	–	1e-4	1e-3	–	0	1e-2	–	5e-3
	Isotropic	U	5e-3	–	1e-3	1e-3	–	5e-3	1e-3	–	5e-3
SMC	Isotropic	G	5e-3	–	1e-3	1e-2	–	1e-4	1e-2	–	1e-4
	Isotropic	U	1e-3	–	5e-3	1e-2	–	0	1e-3	–	1e-3
	Diagonal	G	1e-2	–	1e-4	1e-2	–	1e-3	1e-3	–	1e-3
	Diagonal	U	1e-3	–	0	1e-2	–	0	1e-2	–	0
	General	G	1e-2	1.0	5e-3	1e-3	0.1	1e-2	1e-2	0.2	1e-3
	General	U	1e-2	0.5	1e-4	5e-3	0.2	1e-2	5e-3	0.1	5e-3
ACG	Isotropic	G	1e-3	–	0	5e-3	–	0	1e-2	–	1e-2
	Isotropic	U	1e-2	–	1e-2	1e-2	–	0	5e-3	–	1e-3
	Diagonal	G	1e-2	–	0	1e-2	–	0	1e-2	–	1e-3
	Diagonal	U	5e-3	–	1e-2	1e-2	–	1e-2	5e-3	–	1e-3
	General	G	1e-3	0.5	1e-4	1e-2	0.2	1e-3	1e-3	0.2	1e-3
	General	U	1e-2	1.0	1e-2	5e-3	0.1	5e-3	1e-2	0.5	1e-3
ASG	Isotropic	G	5e-3	–	1e-3	5e-3	–	0	1e-2	–	1e-3
	Diagonal	G	1e-2	–	1e-4	1e-3	–	1e-3	1e-2	–	1e-3
	General	G	1e-2	1.0	5e-3	1e-2	0.5	5e-3	1e-2	0.2	1e-3
Kent	Isotropic	G	1e-2	–	1e-4	1e-2	–	1e-4	5e-3	–	0
	Isotropic	U	1e-3	–	5e-3	5e-3	–	1e-3	1e-3	–	1e-2
	Diagonal	G	1e-2	–	1e-3	1e-2	–	0	5e-3	–	1e-3
	Diagonal	U	1e-2	–	0	5e-3	–	0	5e-3	–	1e-3
	General	G	1e-2	0.2	1e-3	1e-3	0.5	1e-4	1e-3	0.1	1e-2
	General	U	5e-3	0.1	5e-3	5e-3	0.1	5e-3	5e-3	0.1	5e-3

A.3 ADDITIONAL BACKBONES

Table 9 and Table 8 show similar findings as Table 2 in Section 5, for networks trained using ConvNeXt-Small (Liu et al., 2022) and DenseNet161 (Huang et al., 2017) as their backbone architecture. Table 10 exhibits the results of a more exhaustive evaluation of the backbones, as it takes into account network families of different sizes, though restricted to the ℓ_2 baseline (not cosine similarity), over three datasets: Stanford Dogs (Khosla et al., 2011), Oxford Flowers-102 (Nilsback & Zisserman, 2008), and Oxford-IIIT Pet (Parkhi et al., 2012).

Table 8: Same setting as Table 2, with a DenseNet-161 backbone.

			CUB-200		Stanford Cars		Stanford Dogs		
Geometry	**Similarity**	**Covariance**	Gated	Ungated	Gated	Ungated	Gated	Ungated	**Best Δ**
Euclidean	ℓ_2 (ProtoPNet)	Identity	84.2		89.9		80.7		
	Mahalanobis ℓ_2	Isotropic	84.7		**90.3**		81.9		
		Diagonal	84.7		89.9		**82.2**		+1.7 (CUB)
		General	**85.9**		89.5		81.5		
Spherical	Cosine (TesNet)	Identity	82.2	85.0	89.8	90.6	80.5	81.2	
	vMF	Isotropic	82.3	84.4	89.6	90.1	81.9	82.3	
	SMC	Isotropic	81.8	85.0	90.2	90.5	81.7	82.3	
		Diagonal	81.6	84.7	**90.3**	90.4	81.9	82.8	**+2.5 (Dogs)**
		General	**82.7**	85.0	89.4	90.4	**82.6**	**83.7**	
	ACG	Isotropic	80.7	83.5	88.4	90.3	81.1	82.7	
		Diagonal	80.5	84.0	81.5	90.8	81.3	82.7	+2.0 (Dogs)
		General	70.4‡	84.4	6.4†	90.5	81.8	83.2	
	ASG	Isotropic	82.1	–	2.4†	–	81.0	–	
		Diagonal	81.4	–	0.6†	–	81.2	–	+1.1 (Dogs)
		General	81.7	–	75.4	–	81.6	–	
	Kent	Isotropic	82.0	85.1	89.7	90.9	81.7	82.7	
		Diagonal	82.6	**85.6**	**90.3**	**91.1**	82.3	83.2	+2.0 (Dogs)
		General	82.0	85.5	78.7	90.8	81.9	82.8	

† Failed to learn (near-chance accuracy). ‡ Training diverged. – Not applicable.

Table 9: Same setting as Table 2, with a ConvNeXt-Small backbone.

			CUB-200		Stanford Cars		Stanford Dogs		
Geometry	**Similarity**	**Covariance**	Gated	Ungated	Gated	Ungated	Gated	Ungated	**Best Δ**
Euclidean	ℓ_2 (ProtoPNet)	Identity	×		×		93.7		
	Mahalanobis ℓ_2	Isotropic	×		×		**94.0**		
		Diagonal	×		×		93.9		+0.3 (Dogs)
		General	×		×		93.3		
Spherical	Cosine (TesNet)	Identity	85.8	85.6	88.9	89.0	94.1	94.3	
	vMF	Isotropic	87.0	86.6	87.6	88.7	95.0	**95.0**	
	SMC	Isotropic	86.9	86.8	88.8	89.1	**95.2**	94.9	
		Diagonal	86.9	64.8‡	**89.4**	**89.5**	95.0	94.8	+1.2 (CUB)
		General	87.0	86.3	88.5	86.7	95.0	94.4	
	ACG	Isotropic	15.7‡	85.6	77.9	86.2	93.5‡	93.3‡	
		Diagonal	16.6‡	86.4	83.6	87.6	94.8	93.4‡	+1.1 (CUB)
		General	16.4‡	86.7	88.8	85.8	94.7	94.4	
	ASG	Isotropic	16.8‡	–	8.5‡	–	94.7	–	
		Diagonal	16.3‡	–	84.8	–	95.0	–	+0.9 (Dogs)
		General	45.9	–	88.5	–	94.8	–	
	Kent	Isotropic	16.3‡	87.2	80.9	88.6	94.8	94.8	
		Diagonal	86.0	87.1	85.6	89.0	**95.2**	94.9	**+2.1 (CUB)**
		General	**87.4**	**87.7**	**89.4**	87.4	95.0	94.6	

† Failed to learn (near-chance accuracy). ‡ Training diverged. – Not applicable. × Not run.

Table 10: Test accuracy (%) of ProtoPNet across eight backbones on Stanford Dogs (Khosla et al., 2011), Oxford Flowers-102 (Nilsback & Zisserman, 2008), and Oxford-IIIT Pet (Parkhi et al., 2012). *Id.* is the standard ℓ_2 baseline; *Iso.*, *Diag.*, and *Gen.* are Mahalanobis ℓ_2 with isotropic, diagonal, and general covariance. For each dataset and backbone, the best result is in **bold** and the second best is underlined. The last row averages over backbones.

	Stanford Dogs				Oxford Flowers-102				Oxford-IIIT Pet			
Backbone	Id.	Iso.	Diag.	Gen.	Id.	Iso.	Diag.	Gen.	Id.	Iso.	Diag.	Gen.
ResNet-18	71.3	70.7	**71.6**	70.1	67.6	70.0	69.7	**70.7**	90.9	91.4	91.3	**92.4**
ResNet-34	75.4	75.2	76.4	**76.5**	65.2	67.9	69.8	**71.7**	91.8	**92.9**	92.7	92.4
ResNet-50	80.3	**81.5**	81.2	80.2	86.8	87.5	87.6	**88.0**	93.2	92.9	92.8	**93.8**
ResNeXt-50	80.3	81.2	**81.9**	81.5	78.8	**81.9**	81.5	81.6	93.9	94.4	**94.7**	94.6
DenseNet-121	78.6	78.7	**79.2**	78.5	79.3	78.4	79.3	**81.0**	92.0	**92.1**	91.2	91.2
DenseNet-161	80.7	81.9	**82.2**	81.5	79.2	85.4	**85.6**	84.3	92.5	**94.2**	92.5	93.5
ConvNeXt-T	91.6	**92.2**	91.9	91.9	86.2	88.9	88.1	**90.2**	**96.5**	96.3	95.9	96.2
ConvNeXt-S	93.7	**94.0**	93.9	93.3	88.6	88.2	88.1	**89.1**	96.7	96.9	**97.3**	96.6
Mean	81.5	81.9	**82.3**	81.7	79.0	81.0	81.2	**82.1**	93.4	**93.9**	93.6	93.8

A.4 Tempered Peak-Normalized ACG Derivation

Let $P = I - \mu_j \mu_j^T$ be the projection matrix onto the tangent plane of the prototype μ_j. We define the constructed precision matrix as:

$$\tilde{\Lambda}_j = \mu_j \mu_j^T + P(I + \Lambda_j)P \tag{7}$$

By construction, μ_j is an eigenvector of $\tilde{\Lambda}_j$ with eigenvalue 1 ($\tilde{\Lambda}_j \mu_j = \mu_j$). The tangent plane is invariant under $\tilde{\Lambda}_j$, and the eigenvalues of $P\Lambda_j P$ restricted to the tangent plane are bounded below by $\lambda_{\min}(\Lambda_j)$. Because we constrain the learned parameters such that Λ_j is strictly positive definite ($\Lambda_j \in \mathbb{S}^D_{++}$), every tangent eigenvalue of $\tilde{\Lambda}_j$ is strictly greater than 1.

For any latent vector $z \in S^{D-1}$, we decompose it as $z = c\mu_j + z_\perp$, where $c = z^T \mu_j$ and $c^2 + \|z_\perp\|^2 = 1$. The quadratic form evaluates to:

$$z^T \tilde{\Lambda}_j z = c^2 + \|z_\perp\|^2 + z_\perp^T \Lambda_j z_\perp = 1 + z_\perp^T \Lambda_j z_\perp \tag{8}$$

The classical ACG density is given by $p_{\text{ACG}}(z; \tilde{\Lambda}_j) = \frac{\Gamma(D/2)}{2\pi^{D/2}} \det(\tilde{\Lambda}_j)^{1/2} (z^T \tilde{\Lambda}_j z)^{-D/2}$. Because all tangent eigenvalues are strictly > 1 and the eigenvalue along μ_j is exactly 1, the minimum of the quadratic form $z^T \tilde{\Lambda}_j z$ on the hypersphere is uniquely achieved at $z = \pm\mu_j$. Therefore, the maximum of the classical density occurs precisely at the antipodal pair.

Dividing the density by its peak value gives:

$$\frac{p_{\text{ACG}}(z; \tilde{\Lambda}_j)}{\max_z p_{\text{ACG}}(z; \tilde{\Lambda}_j)} = (z^T \tilde{\Lambda}_j z)^{-D/2} = (1 + z_\perp^T \Lambda_j z_\perp)^{-D/2} \tag{9}$$

Raising both sides to the power $2\nu/D$ yields our similarity function:

$$S_{\text{ACG}}(z, \mu_j; \Lambda_j, \nu) = (1 + z_\perp^T \Lambda_j z_\perp)^{-\nu} = \left(\frac{p_{\text{ACG}}(z; \tilde{\Lambda}_j)}{\max_z p_{\text{ACG}}(z; \tilde{\Lambda}_j)} \right)^{2\nu/D} \tag{10}$$

Consequently, for any $\nu > 0$, our similarity is exactly a tempered peak-normalized ACG density, with ν acting as an inverse temperature. Since raising to a positive power is strictly monotonic, for fixed μ_j and Λ_j every ν induces the same level sets, the same ranking of latent patches (hence the same projected patch), and the same desiderata compliance (D1–D4). The exponent ν only rescales similarity values along these level sets; because class logits aggregate these values, it does affect training dynamics and hence the learned parameters.

Specifically, this formulation yields two distinct theoretical benefits:

1. Because the peak value is exactly 1 for every valid Λ_j, the similarity score cannot be inflated by ill-conditioning, satisfying the boundedness desideratum (D2).
2. Separately, scaling $\Lambda_j \to c\Lambda_j$ does not scale $\tilde{\Lambda}_j$ isotropically, since its eigenvalue along μ_j is pinned at 1. Unlike the classical ACG distribution, S_{ACG} is therefore not scale-invariant, allowing an isotropic Λ_j to successfully act as a learnable concentration parameter.