SPARSE PROTOTYPE CODE UNDERLIES CLASSIFICATION AND PREDICTION ACROSS MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural representations have become a central tool for studying the internal mechanisms of modern AI models, yet their complex high-dimensional structure makes it difficult to identify the geometric features that determine predictive performance. We identify a universal representational geometry shared across state-of-the-art models in vision, audio, and language, spanning tasks from object recognition to next-token prediction. Its defining feature is that within-class variability is strongly structured: its task-relevant component is aligned with the class's own centroid and the centroids of competing classes. Building on this observation, we derive an analytical mean-field theory governed by the statistics of projections along true-class and rival-class centroid directions, together with a global renormalization of the class radius that compensates for the non-Gaussianity of real representations. The resulting theory accurately predicts per-class accuracy across tasks, architectures, and modalities. The relevant geometric quantities improve systematically with model scale, revealing how larger models achieve higher accuracy even though their overall within-class variability often increases. A striking feature of this framework is its sparsity: accurate predictions require only a small set of centroid coordinates associated with the true class and its strongest rivals. These rivals are often semantically related to the true class, suggesting a connection to sparse-feature extraction approaches such as sparse autoencoders. Together, these results provide a parsimonious predictive theory of neural representations and suggest that classification and prediction in deep networks are governed by a sparse, centroid-aligned structure embedded within the full high-dimensional representation space.

1 INTRODUCTION

One of the central challenges in modern Deep Neural Network research is to open the "black box" and understand the internal mechanisms by which networks transform their inputs into useful predictions. Such understanding is essential for building safer, more reliable AI systems (Bommasani et al., 2021; Hendrycks et al., 2021). A primary tool toward this goal is the analysis of representations - the hidden-layer activations induced by specific inputs. However, extracting interpretable principles from the complex high-dimensional geometry of representations remains a significant challenge.

A growing body of work in mechanistic interpretability addresses this challenge by extracting sparse, human-interpretable features from network activations (Elhage et al., 2022; Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2026). Sparse autoencoders have shown that semantically meaningful concepts are often encoded along specific directions in activation space and can be isolated within otherwise opaque representations. These approaches are powerful tools for identifying the essential features of the network; they are typically not aimed, however, at providing a quantitative theory linking the geometric structure of a representation to how well it supports a downstream task. Such a theory would help address fundamental questions, such as why some networks generalize better than others.

Predictive theories of representational geometry - manifold capacity (Chung et al., 2018; Cohen et al., 2020; Stephenson et al., 2021; Wakhloo et al., 2023) and neural collapse (Papyan et al., 2020; Han et al., 2021; Galanti et al., 2021) - have made important progress in connecting geometry to performance within their respective regimes. Capacity theory studies the linear separability of object manifolds under random binary dichotomies, relating global capacity to geometric quantities such as manifold radius and effective dimension. Neural collapse describes the final phase of supervised

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.

training, in which within-class variability collapses and class means converge to a tight frame. Modern foundation models, however, typically operate in a different regime: they are evaluated using multiclass accuracy and are often trained with early stopping or self-supervised objectives, leaving substantial within-class variability. Recent work connects pairwise class geometry to performance and generalization bounds in few-shot learning (Sorscher et al., 2022; Luthra et al., 2026). A predictive geometric theory of multiclass accuracy in the finite-radius regime that accounts for both competition among multiple classes and the rich within-class statistics of many examples is currently missing.

Our analytical lens is prototype-based classification. A long line of work - from nearest-class-mean classifiers (Mensink et al., 2013) to prototypical networks (Snell et al., 2017) - has established that classifying inputs by proximity to class centroids is a strong strategy whenever representations are well-structured. Adopting the centroid-based classifier has a clear advantage: the centroids are geometric objects, allowing classification accuracy to be related directly to geometric measures of the representation - while remaining relevant for realistic classification settings. Thus, we leverage this framework to close the theoretical gap, deriving and empirically validating a quantitative geometric theory of multi-class classification accuracy, and apply it across architectures and modalities. Our main contributions are:

Universal centroid-aligned structure: We identify a universal representational geometry shared across vision, language, and audio backbones. Its defining property is that within-class variability is strongly correlated with both the true-class centroid and the centroids of competing classes (see Fig. 1a). We explicitly demonstrate this by surgically removing only these centroid-aligned correlations (see Fig. 1b), and find that doing so drives per-class accuracy to essentially 100%. The centroid-aligned correlations are therefore precisely the geometric source of classification errors.

Sparsity of competition: These correlations do not spread evenly across the other classes; the competition concentrates on a small set of rivals, which is sufficient to reproduce the predicted per-class accuracy. These rivals also appear to carry semantically meaningful relations to the true class (see Fig. 1a, Table 2). The geometry of a class is therefore faithfully captured by a sparse, semantically meaningful centroid-aligned basis embedded in the full high-dimensional representation - an analog of the sparse-feature picture of mechanistic interpretability.

Compact predictive mean-field theory: We derive an analytical mean-field theory in the sparse centroid basis (Sec. 5) of prototype classification accuracy. Its inputs are the variability along the true- and rival-class centroid directions and a single global radius renormalization that absorbs the heavy-tailed statistics of real representations (see Fig. 3b). Crucially, this factor is tightly correlated with a direct measure of tail heaviness (Fig. 4e). The theory predicts per-class accuracy across classes, architectures, and modalities (see Fig. 4(a-d), Appendix I).

Geometry across model scales: Tracking the theory's geometric ingredients across model families, we find that they systematically improve as the models scale up, mirroring the gains in accuracy (Table 3). Importantly, we show that the naive overall within-class variance does not predict these gains. Our theory identifies the specific geometric measures that drive this improvement - taking a step toward understanding the key geometrical ingredients required for good generalization in real state-of-the-art models. We hypothesize that these measures are practically actionable: targeting them during training could improve generalization, and matching them between teacher and student could guide more geometry-aware distillation (Hinton et al., 2015).

2 Setup and Notation

We study how the geometry of neural representations determines classification accuracy. For each input we consider a representation vector, and the task is to assign it to one of P classes.

Prototype classification: We use a prototype classifier (Snell et al., 2017), in which the class-specific readout vector is not a learned weight but the empirical mean of the class in representation space (i.e., centroid, see Def. 2.1), normalized to a unit vector. Centroids are estimated from a labeled set, and at inference, the learner classifies one example against all P centroids ("Leave one out" method (Lachenbruch & Mickey, 1968)). This has a direct geometric definition while remaining relevant for realistic classification, and empirically tracks trained linear readouts closely (see Appendix B.1), making it a useful proxy for the standard linear head.

Experimental setup: We evaluate frozen z-scored representations from pretrained backbones spanning vision (ImageNet-1K (Deng et al., 2009)), language (masked and next-token prediction

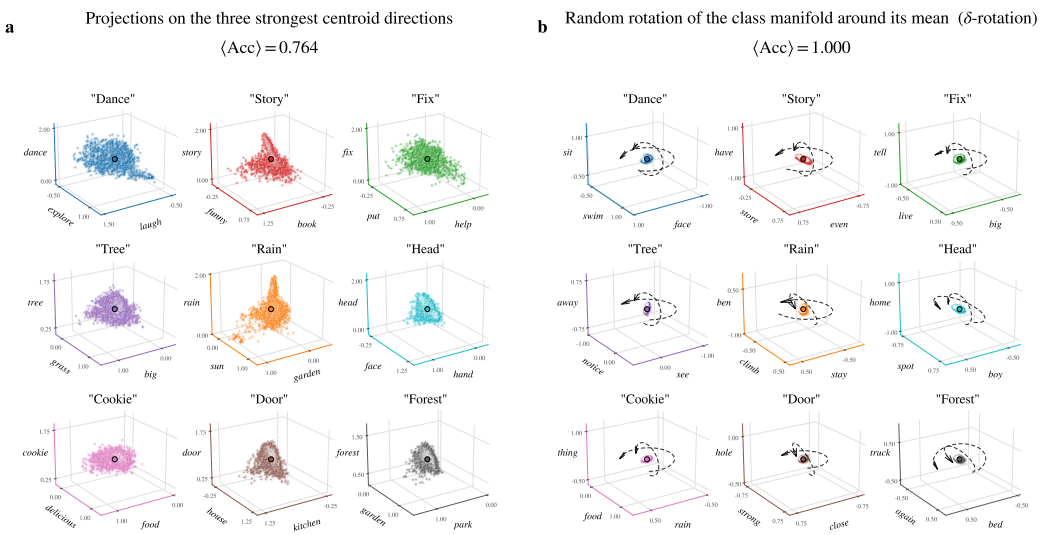

Figure 1: **Centroid-aligned variability**. The variability of each class is highly correlated with rival centroids. **(a)** Qwen3-32B representations (Yang et al., 2025) on TinyStories (Eldan & Li, 2023). Examples are projected onto the three centroids with the largest correlations among the class variability. The true class centroid is among these top 3 in 68% of the classes, including all nine shown. The rivals carry semantic meaning (e.g., "cookie" $\mapsto$ "food", "delicious"; "head" $\mapsto$ "face", "hand"). **(b)** Random rotation of the variability around the class centroid removes all classification errors, raising mean prototype accuracy for the nine shown classes from 76.4% to 100%. The manipulation breaks residual–centroid correlations while preserving class radii and centroid distances, identifying these correlations as the main source of classification errors.

on TinyStories (Eldan & Li, 2023)), and audio (VGGSound (Chen et al., 2020)), across transformer and convolutional architectures, supervised, self-supervised, and contrastive objectives. Full lists of datasets, models, and extraction details are in Appendix C. For language, the task is to predict a content word from its representation. In autoregressive models this is the embedding of the token preceding the word. We group each word's inflections under a single lemma label (e.g., say/says/said, friend/friends), and keep only lemmas with at least 1000 examples, so their statistics are well estimated. Full details needed to reproduce all results are given in Appendix C; all datasets and models are publicly available.

Definition 2.1 (*prototype classification*): Let $\mathbf{x} \in \mathbb{R}^N$ be the representation of an input at a given layer of a neural network, and consider a P-class task with labels $y \in \{1, \dots, P\}$ and dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. For each class μ, define the centroid $\mathbf{c}_\mu = \mathbb{E}[\mathbf{x} \mid y = \mu]$ and its unit direction $\hat{\mathbf{c}}_\mu = \mathbf{c}_\mu / \|\mathbf{c}_\mu\|$, which serves as the prototype class readout. Throughout the paper, we use the terms centroid and prototype interchangeably. The prototype logits and accuracy are defined as

$$t_\mu(\mathbf{x}) = \hat{\mathbf{c}}_\mu \cdot \mathbf{x}, \qquad \text{Acc} = \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}[\mathbf{1}\{\arg\max_\mu t_\mu = y\}]. \tag{1}$$

Definition 2.2 (*manifold geometry*): The class manifold is defined as the point cloud of a particular class k: $\mathcal{M}_k = \{\mathbf{x}_i : y_i = k\}$. Each $\mathbf{x} \in \mathcal{M}_k$ decomposes as $\mathbf{x} = \mathbf{c}_k + \delta\mathbf{x}$ with $\mathbb{E}_{\mathbf{x}\in\mathcal{M}_k}[\delta\mathbf{x}] = \mathbf{0}$. We name $\delta\mathbf{x}$ the **residual** vector. The class radius R measures the overall within-class variance, and $\sigma_\mu^2 \in [0, 1]$ is the fraction of the residual variance along a particular centroid direction $\hat{\mathbf{c}}_\mu$:

$$R^2 = \frac{\mathbb{E}_k[\|\delta\mathbf{x}\|^2]}{\|\mathbf{c}_k\|^2}, \qquad \sigma_\mu^2 = \frac{\mathbb{E}_k[(\delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu)^2]}{\mathbb{E}_k[\|\delta\mathbf{x}\|^2]}. \tag{2}$$

Throughout, geometric quantities depend on a fixed true class k. We treat k as fixed and drop the subscript to lighten notation: $R \equiv R_k$, $\sigma_{k\mu} \equiv \sigma_\mu$, and σ_k^2 is the fraction of residual variance along the true centroid $\hat{\mathbf{c}}_k$. In addition, $\mathbb{E}_k$ denotes the average over all the points in the manifold k.

3 Isotropic Uncorrelated Model

We first examine a baseline of an uncorrelated model, in which the centroids $\mathbf{c}_\mu$ and the residuals $\delta\mathbf{x}$ are drawn independently from an isotropic Gaussian distribution. The model yields testable

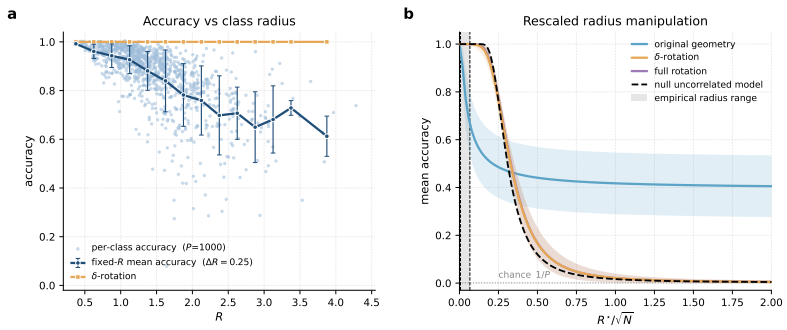

Figure 2: **Geometric signatures of multiclass classification. (a)** Per-class accuracy vs. empirical radius R: sharp decrease across the narrow empirical R range and a wide spread at fixed R. After random rotation of within-class variability $\delta\mathbf{x}$ (δ-rotation), the accuracy is perfect. **(b)** We artificially scale the variability $\delta\mathbf{x} \to (R^\star/R)\delta\mathbf{x}$. The accuracy decreases fast but stabilizes at 39% accuracy for $R^\star \to \infty$ due to strong true centroid correlations (blue line). After random rotation of the variability δ-rotation, the accuracy is well described by the isotropic model ($\mathrm{MSE} = 1.26 \times 10^{-3}$, orange line). Additional breaking of centroid-centroid correlations (full rotation) has a very small effect; lines overlap.

predictions for how geometry shapes accuracy, which we evaluate in Sec.4 on real data: deviations from these predictions quantify the role of correlations in the classification task.

Assumption 3.1 (*uncorrelated statistics*). For each class $\mu = 1, \dots, P$, the centroid is drawn as $\mathbf{c}_\mu \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, I_N)$. Conditional on the centroids, examples of class k are generated by sampling $\mathbf{x} = \mathbf{c}_k + \delta\mathbf{x}$, with residuals drawn as $\delta\mathbf{x} \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, R^2 I_N)$.

Assumption 3.2 (*high-dimensional geometry*). We work in the high-dimensional limit $N \to \infty$, with $P = N^{\mathcal{O}(1)}$ classes and a radius $R = R(N)$ that may scale with N. We refer to Assumptions 3.1-3.2 jointly as the isotropic model.

Theorem 3.3 (*isotropic model accuracy*). Under the isotropic model,

$$\mathrm{Acc}^{\text{iso.}}(R, N, P) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\Phi^{P-1}\left(\frac{Rz + \sqrt{N}}{\sqrt{R^2+1}}\right)\right], \tag{3}$$

where Φ is the standard Gaussian CDF.

Predictions. Several consequences follow directly from Theorem 3.3:
(i) For any $R = \mathcal{O}(1)$, the classes are well separated, and classification is trivial: $\mathrm{Acc}^{\text{iso.}} \to 1$.
(ii) In the infinite radii limit, $R/\sqrt{N} \to \infty$, the isotropic residual retains no information about the true class, and accuracy collapses to chance, $\mathrm{Acc}^{\text{iso.}} \to 1/P$.
(iii) Non-perfect accuracy requires large radii, $\mathrm{Acc}^{\text{iso.}} < 1 \Rightarrow R = \Omega(\sqrt{N/\ln P})$.
(iv) In the high-dimensional regime, the accuracy is self-averaging: its variance across centroid realizations is $\mathcal{O}((PR^2)^{-1})$.
The theorem and predictions are rigorously derived in Appendix D.1, and verified numerically with synthetic data in Appendix Fig.6.

4 Empirical Results

In this section we characterize the geometric structure of trained representations that governs prototype classification. We begin from the predictions of the isotropic model (Sec.3) and show that they are not satisfied by real representations. This identifies which correlations dominate the geometry responsible for accuracy, and isolates the components needed to build the analytical framework of Sec.5. Results are presented for DINOv3 7B (Siméoni et al., 2025) representations on ImageNet-1K (Deng et al., 2009), and replicated in many models and modalities (Appendix A).

Accuracy changes on an $\mathcal{O}(1)$ radius scale: The empirical class radii lie in a narrow $R \in [0.4, 4.3]$ range. In Sec.3, we saw that an uncorrelated model predicts that accuracy should decline only in $R = \Omega(\sqrt{N/\ln P}) \sim 20$. However, we observe a dramatic change in accuracy, ranging from 100% to 61% averaged over a small R window (see Fig.2a). The sharp accuracy decline therefore indicates structure that is not described by the isotropic model.

Large class-to-class variability at fixed radius: The isotropic model predicts that class-to-class fluctuations are of order $1/R\sqrt{P}$ ($\sim 1\% - 3\%$). Empirically, however, the spread of accuracies at fixed R is large. For example, the 89 classes in the range $R \in [1.2, 1.3]$ have a mean accuracy of 90.4%, with a minimum accuracy of 61.8%, and a maximum of 99.6% (see Fig.2a). The radius alone therefore cannot explain classification performance: additional class-dependent geometric structure must determine which classes are easy to classify and which are difficult.

4.1 Geometric manipulations

To causally test which components cause the deviation from the isotropic model, we use geometric manipulations to probe specific structures of the representations. We apply two geometric manipulations:
(i) **Variability inflation**, where we set each class radius to a chosen $R^\star$ artificially by scaling its deviations, $\mathbf{x}(R^\star) = \mathbf{c}_k + (R^\star/R)\delta\mathbf{x}$, where R is the empirical radius as defined in Sec.2.
(ii) **Geometric rotations**, where we surgically destroy specific correlations in the data while preserving other aspects of the representational geometry.

Strong residual–true-centroid correlations: We first examine the representations under variability inflation. Naively, as $R^\star \to \infty$ one expects chance accuracy, as we saw in the isotropic model. Empirically, accuracy approaches 39.2%, over two orders of magnitude above $1/P = 0.1\%$ (Fig.2b). The within-class variability is thus far more correlated with the true centroid than with the rival centroids.

Residual-centroid correlations are the main drivers of errors: We manipulate the representations by rotating the residuals $\delta\mathbf{x}$ around their own class centroid (δ-rotations): for each class k we replace $\mathbf{x}$ by $\mathbf{x}^{\delta\text{rot}} = \mathbf{c}_k + O_k\delta\mathbf{x}$, with O_k a class-specific random orthogonal matrix (see Fig. 1b for illustration). This preserves the distance between classes and the class radius and dimensions, while breaking the correlations between residuals and the centroid directions. At the empirical radius, the class-averaged accuracy rises from the baseline of 85.5% to 100%, completely eliminating classification errors (see Fig.2a, orange line). This is achieved *without* breaking centroid-centroid correlations. We stress that this manipulation uses the true class label and therefore serves as a diagnostic tool rather than a procedure applicable to unlabeled examples.

Breaking residual–centroid correlations recovers the isotropic model: To look for any further role of residual–centroid correlations, we go outside the empirical radius range, where δ-rotation has already saturated the accuracy. We set each class radius to a chosen $R^\star$ by variability inflation. Strikingly, after removing only the residual-centroid correlations, the data is already well described by the isotropic model accuracy (Eq.3), with $\mathrm{MSE} = 1.26 \times 10^{-3}$. The accuracy declines only at $R = \Omega(\sqrt{N/\ln P})$, far beyond the empirical R range (see Fig.2b).

Table 1: Definitions of the geometric measures of the theory.

Notation	Name	Definition
$\mathbf{c}_k$	Centroid	$\mathbb{E}_k[\mathbf{x}]$
$\hat{\mathbf{c}}_\mu$	Unit centroid	$\mathbf{c}_\mu/\lVert\mathbf{c}_\mu\rVert$
$\delta\mathbf{x}$	Residual	$\mathbf{x} - \mathbf{c}_k$
R^2	Radius^2	$\mathbb{E}_k\big[\lVert\delta\mathbf{x}\rVert^2\big]\big/\lVert\mathbf{c}_k\rVert^2$
σ_μ^2	Variance fraction	$\mathbb{E}_k\big[(\delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu)^2\big]\big/\mathbb{E}_k\big[\lVert\delta\mathbf{x}\rVert^2\big]$
s_μ	Std. projection	$(\delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu)\,/\mathrm{std}_k\,[\delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu]$
ρ_μ	Proj. correlation	$\mathrm{Corr}_k\,(\delta\mathbf{x}\cdot\hat{\mathbf{c}}_k, \delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu)$
g_μ	Centroid overlap	$\hat{\mathbf{c}}_k\cdot\hat{\mathbf{c}}_\mu$
$\mathcal{R}_K$	Rival set	$\{\,\mu \neq k:\ K$ largest $\sigma_\mu\,\}$

Centroid-centroid correlations play a minor role: To test the role of centroid-centroid correlations, we apply a stronger manipulation: rotation of the whole class manifold around the origin by another random orthogonal matrix after δ-rotations, $\mathbf{x}^{\text{full-rot}} = U_k\big(\mathbf{c}_k + O_k\delta\mathbf{x}\big)$, which breaks centroid–centroid correlations on top of the δ-rotation. Even at these larger radii the extra effect is small: the full-rotation curve differs from the δ-rotation curve by only $\mathrm{MSE} = 2.0\times10^{-5}$ (see Fig. 2b). We infer that centroid–centroid correlations play a minor role in accuracy. This echoes the picture from neural collapse (Papyan et al., 2020), where the centroids are as spread as possible.

5 Centroid-Aligned Variability Model

The empirical results of Sec. 4 identify the key geometric structure of trained representations: the class-relevant geometry is governed by strong correlations between within-class variability and a sparse set of centroid directions—dominated by the true-class centroid.

We introduce the centroid-aligned model that incorporates this geometry.

The resulting theory takes as input geometric quantities measurable from the representation, together with a single global renormalization of the class radius (Sec. 5.2), and predicts per-class accuracy.

Full proofs are deferred to Appendix D.2; the main text retains only the core assumptions and results. We stress that the following assumptions define an analytically solvable model and are not intended as exact claims about the statistics of the data. Nevertheless, the model is motivated by empirical observations and yields accurate predictions across models spanning different architectures and modalities. We therefore believe that it captures important aspects of real representations.

Definition 5.1 (*rival support*). Fix a class k and an integer K. Define its rival support $\mathcal{R}_K$ as the set of indices $\mu \neq k$ corresponding to the K largest values of σ_μ (Def. 2.2), and set $\mathcal{R}_K^+ \equiv \{k\} \cup \mathcal{R}_K$. For each $\mu \in \mathcal{R}_K$, define the true-rival centroid correlation by $g_\mu = \hat{\mathbf{c}}_k \cdot \hat{\mathbf{c}}_\mu$.

Definition 5.2 (*standardized centroid projections*). Conditioned on class k, each example is represented by a set of projection coefficients $s_\mu = (\delta\mathbf{x} \cdot \hat{\mathbf{c}}_\mu)/\mathrm{std}_k\left[\delta\mathbf{x} \cdot \hat{\mathbf{c}}_\mu\right]$.

Assumption 5.3 (*centroid-aligned variability model*). Let $G_{\mu\nu} = \hat{\mathbf{c}}_\mu \cdot \hat{\mathbf{c}}_\nu$ be the centroids Gram matrix. Assume that the restricted centroid Gram matrix of the rival support $(G_{\mu\nu})_{\mu,\nu\in\mathcal{R}_K^+}$ is invertible. We assume the following generative model of examples of class k

$$\mathbf{x}(\mathbf{s}) = \|\mathbf{c}_k\| \left[\hat{\mathbf{c}}_k + R\left(\sum_{\mu,\nu\in\mathcal{R}_K^+} \sigma_\mu s_\mu G^{-1}_{\mu\nu}\hat{\mathbf{c}}_\nu + \sigma_\perp s_\perp \hat{\mathbf{e}}_\perp\right)\right] \tag{4}$$

where $\hat{\mathbf{e}}_\perp$ is perpendicular to the $\mathrm{span}(\hat{\mathbf{c}}_{\mu\in\mathcal{R}_K^+})$, σ_μ^2 is the variance fraction along the centroid μ direction, and R is the empirical class radius (see Def. 2.2). Eq. 4 encodes the geometry identified in Sec. 4 by placing the residuals $\delta\mathbf{x}$ along rival centroid directions in the rival support.

Assumption 5.4 (*Gaussian statistics*). We assume the projection coefficients s_μ are jointly Gaussian with zero mean and unit variance, and that their covariance has the following form

$$\mathrm{Cov}\left(s_\mu, s_\nu\right) = \delta_{\mu\nu}\left(1-\rho_\mu^2\right) + \rho_\mu\rho_\nu. \tag{5}$$

Where we define $\rho_\mu = \mathbb{E}_k\left[s_k s_\mu\right]$. Equivalently, the assumption is that all dependence among the rival projections is captured by s_k: conditioned on s_k, they are mutually independent.

Theorem 5.5 (*centroid-aligned model accuracy*). Under assumptions 5.3-5.4

$$\mathrm{Acc}_k^{\mathrm{th}}(R) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\prod_{\mu\in\mathcal{R}_K} \Phi\left(\frac{\sigma_k - \rho_\mu\sigma_\mu}{\sigma_\mu\sqrt{1-\rho_\mu^2}}z + \frac{1-g_\mu}{R\sigma_\mu\sqrt{1-\rho_\mu^2}}\right)\right]. \tag{6}$$

The inputs to Eq. 6 are entirely geometric and class-specific: the empirical radius R, the rival projected variances $\{\sigma_\mu\}$, the centroid-centroid overlaps $\{g_\mu\}$, and the correlation terms ρ_μ are all measured directly from the representations. For convenience, we summarize our notations and how we measure them in Table 1.

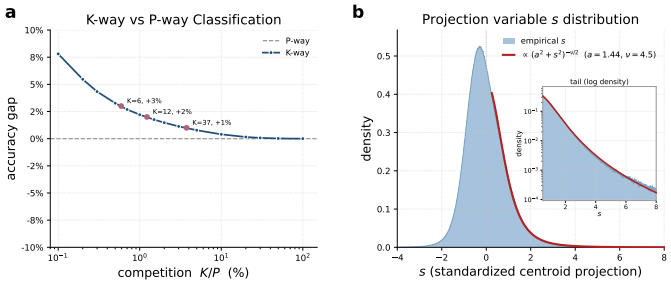

Figure 3: **Sparsity and heavy tail statistics. (a)** K-way classification of rivals chosen by centroid projection variance σ_μ converges quickly to the full P-way classification. **(b)** The standardized centroid projections s_μ (Def. 5.2) have heavy tail power law statistics, which inflate the extreme values that control classification accuracy. The magnified panel shows the log density of the far tail.

5.1 Sparsity

The rival count K is a hyperparameter of the theoretical model controlling how many rival directions enter the prediction in Eq. 6. Because Φ approaches one quickly as its argument grows, small

Table 2: Rival classes selected by the projection's variance σ_μ are semantically related words (Qwen3-32B, TinyStories) or visually similar and co-occurring objects (DINOv3 ViT-7B, ImageNet).

model	class	top-5 rivals (by σ_μ)
Qwen3-32B	*cake*	delicious, cookie, food, ice, big
	bad	sad, sorry, scare, angry, tired
	forest	park, garden, grass, world, sky
DINOv3 ViT-7B	*leopard*	snow leopard, jaguar, cheetah, impala, tusker
	king crab	Dungeness crab, rock crab, hermit crab, plate, American lobster
	flute	panpipe, oboe, ocarina, bassoon, stage

projections σ_μ have very little effect on the overall accuracy. Thus, when ranking rivals by their projection magnitudes σ_μ, predictive power plateaus rapidly once K exceeds the head of the spectrum.

We assess the effective number of rivals both theoretically and empirically. We first replace the full P-way task with a $K+1$-way task restricted to the top rivals ranked by σ_μ and true class. We find that only 6 rivals are needed to come within 3% of the full P-way classification accuracy and 37 for 1% (see Fig.3a). The competition is thus not about the full set of P (1000) classes, but rather a sparse set of strongly correlated rivals. In Appendix F, we compare several rival-selection methods and show that ranking rivals by σ_μ yields results comparable to more complex selection rules, while remaining simple, interpretable, and easy to compute. In particular, it outperforms selection based on centroid proximity, providing additional evidence that centroid correlations play only a secondary role.

The theory's predictions also saturate quickly as the number of rivals K increases. We find that $K \in [10, 20]$ is already enough in most models to achieve accurate predictions, with very small gain from increasing K further (see Appendix F). The results in the main text are shown for $K = 20$ for all models.

We also note that in many cases, the rivals chosen by σ_μ carry semantic meaning. In the Qwen3-32B, many rival words are similar in both grammatical role ("smile"↦"reply", "nod", "say") and semantic meaning ("forest"↦"park", "garden", "grass"). In vision (DINOv3 ViT-7B), we observe similar objects ("leopard"↦"jaguar", "cheetah") or frequently co-occurring objects ("king crab"↦"plate", "flute"↦"stage"). For more examples, see Table 2 and Fig. 1a.

5.2 Non-Gaussian tails and a global radius renormalization

The theoretical accuracy predicted by Eq.6 alone consistently overestimates the accuracy on all the models we checked (see Appendix G). Our main simplifying assumption (Assumption 5.4) is that the statistics of the standardized centroid projections s_μ are Gaussian. Empirically, the projections have markedly heavy-tailed statistics (see Fig.3b).

Although the variance of s_μ is normalized, classification depends on extreme statistics ($\max_\mu t_\mu$), and heavier tails inflate these extremes relative to a Gaussian of equal variance. We find that the right tail of the variables s_μ that controls the max projection on a certain centroid is well fitted by a Student's-t power-law distribution $p(s) \propto (a^2 + s^2)^{-\nu/2}$ (see Fig.3b), where the exponent ν shapes the decay of the tail.

We adopt the simplest solution: a uniform global variance inflation of all manifolds $R \mapsto \lambda R$ in Theorem 5.5, which adjusts the scale of the Gaussian theory such that its extreme matches the empirical heavy-tailed distribution.

The global parameter λ is fitted once per model by

$$\lambda^\star = \arg\min_\lambda \left(\sum_{k=1}^{P} \Big[\mathrm{Acc}_k^{\mathrm{emp}}(R) - \mathrm{Acc}_k^{\mathrm{th}}(\lambda R) \Big]^2 \right) \tag{7}$$

Thus λ is a single global calibration of non-Gaussian statistics of a specific model, not a per-class fitting parameter. Empirically $\lambda^\star$ is well correlated with the inverse tail exponent $1/\nu$ of the model, $r = 0.90$ (see Fig. 4e). This relation is global across 36 models spanning vision, audio, and language, supporting our interpretation of λ as a compensation for non-Gaussian statistics of the theory. Across 36 backbones, the calibrated theory achieves per-class correlations above $r = 0.93$, with mean MSE of 4.06×10^{-3} (Table 12, Appendix I).

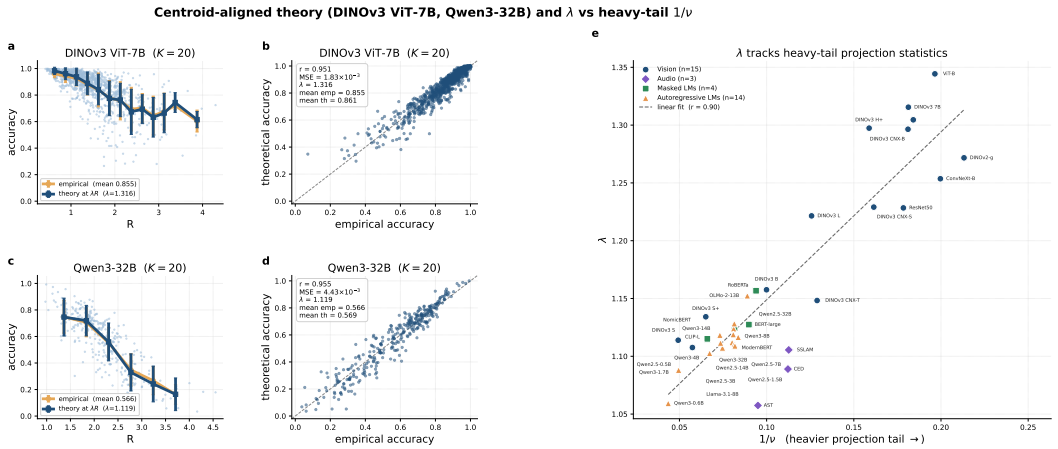

Figure 4: **Centroid-aligned theory matches per-class accuracy across modalities. (a, c)** Per-class accuracy vs. empirical radius R for DINOv3 ViT-7B on ImageNet (P=1000, K=20) and Qwen3-32B on TinyStories next-token prediction (P=320, K=20). Per-class scatter is presented with empirical binned mean. The λ-rescaled theory ($\lambda^\star$=1.316, 1.119 resp.) accounts for the empirical accuracy accurately. **(b, d)** Accurate per-class prediction of the accuracy. MSE=1.83×10^{-3} (vision) and 4.43×10^{-3} (language), Pearson r=0.951 and 0.955. **(e)** Optimal $\lambda^\star$ vs. heavy tail measurement ($1/\nu$) on representations for 15 vision backbones, 14 autoregressive language models, 4 masked language models, and 3 audio models. Overall Pearson correlations between λ and $1/\nu$ are r=0.90, confirming that λ is a byproduct of the heavy-tail shape of the projection variables.

6 Geometry and Accuracy Across Model Scales

The theory of Sec. 5 reduces accuracy to a handful of ingredients of the centroid-aligned geometry. We use these quantities as a diagnostic lens across the DINOv3 and Qwen3 model families at different scales ($S(21M)/S_+(29M)/B(86M)/L(300M)/H_+(840M)/7B$, $0.6B/1.7B/4B/8B/14B/32B$ resp.). The theory reveals which geometric reorganizations drive the improvements in classification performance of larger models. For autoregressive models, we report geometry at the layer maximizing prototype accuracy, consistently 1–4 layers before the final layer. Depth-dependent results appear in Appendix H.

A natural argument is that the accuracy gains of larger models are a trivial variability-reduction effect — that bigger models simply produce smaller manifolds. The Qwen3 family rules this out directly: the mean radius actually grows with scale ($\langle R \rangle = 1.98 \to 2.20$ from 0.6B to 32B), yet accuracy rises monotonically from 41% to 57%, the opposite of what a naive approach predicts. In the DINOv3 family, the radius does initially shrink with scale, but the trend breaks at the largest model: ViT-7B carries a mean radius 0.21 larger than ViT-H$^+$ yet matches its accuracy. Reduced variability, therefore, cannot be the mechanism behind improved generalization. Our other geometric measures are independent of R, so they isolate genuine reorganizations of the geometry across scale, and correctly identify the components that make the larger models' representations better.

We specifically look at the worst rival μ^*, measured by the maximum σ_μ (most correlated centroid among the variability) for each class, and average its geometric statistics over all classes. The theory exposes a coherent set of reorganizations, shared among the two families. Within-class variability becomes more aligned with the true centroid and less aligned with the rival (the ratio σ_k/σ_μ^* increases, σ_μ^* decreases in both families). The centroids decorrelate - g_μ^* drops - and the projections of the residual on the centroids also become less correlated (ρ^* decreases). In language, we see similar trends to vision but with less favorable values, which explains the overall lower accuracy. Finally, representations become more heavy-tailed with scale (ν decreases, $19.3 \mapsto 4.5$ in vision and $21.9 \mapsto 11.4$ in language). While we cannot directly link this trend to improved accuracy, it is consistent.

7 Discussion

This work identifies geometric organization principles for multiclass classification: prototype accuracy is controlled by a sparse, centroid-aligned code embedded in the full representation space. Across vision, audio, and language models, and across architectures and training objectives, the same

Table 3: **Geometric measures across model scales:** DINOv3 ViT family and Qwen3 family. The geometric quantities are defined in Table 1 and Sec.5; the superscript * flags the hardest competitor of a class, identified by max σ_μ. Best measure within each family marked in bold.

Model	$\langle R \rangle$	σ_k/σ_μ^*	σ_μ^*	g_μ^*	ρ_μ^*	ν	Acc
DINOv3 ViT-S	1.68	1.06	0.172	0.574	0.558	19.3	0.706
DINOv3 ViT-S+	1.60	1.09	0.172	0.555	0.535	14.4	0.738
DINOv3 ViT-B	1.52	1.19	0.163	0.542	0.529	9.0	0.794
DINOv3 ViT-L	1.39	1.36	0.147	0.447	0.322	7.0	0.845
DINOv3 ViT-H+	**1.29**	1.59	0.164	**0.236**	**-0.011**	5.3	**0.855**
DINOv3 ViT-7B	1.50	**2.01**	**0.120**	0.289	0.038	**4.5**	**0.855**
Qwen3-0.6B (L27, -1)	**1.98**	0.85	0.241	0.576	0.515	21.9	0.406
Qwen3-1.7B (L26, -2)	2.09	0.95	0.212	0.542	0.494	13.9	0.474
Qwen3-4B (L34, -2)	2.05	0.98	0.202	0.526	0.463	12.7	0.519
Qwen3-8B (L34, -2)	2.12	0.99	0.188	0.514	0.453	**11.3**	0.540
Qwen3-14B (L37, -3)	2.10	1.00	0.186	0.504	**0.447**	**11.3**	0.556
Qwen3-32B (L61, -3)	2.20	**1.01**	**0.175**	**0.503**	0.451	11.4	**0.566**

geometry recurs. Within-class variability is not isotropic around each class center; its classifier-relevant components are arranged along the true centroid and a small set of rival centroids. The evidence is causal as well as descriptive: breaking this alignment, while preserving the remaining parts of the geometry, nearly eliminates prototype errors. Thus the full N-dimensional representation can be replaced, for the purpose of predicting class accuracy, by a much smaller set of task-aligned coordinates.

The theory makes this reduction quantitative. Per-class accuracy is predicted from the geometric measures of centroid-projection statistics, and a single global scale λ that captures the effect of heavy tails in the marginal projection distributions. We show that as models become larger, the improved accuracy is not due to reduced class variance but because of reorganization of the geometry that the theory reveals: variability shifts toward the true-class coordinate, correlations with the strongest rivals decrease, and centroid directions become more decorrelated. Importantly, per-class prototype accuracy also tracks trained linear-readout accuracy (Appendix B.1), supporting the relevance of this framework to standard linear classification.

This perspective complements existing theories of representational geometry. In manifold-capacity theory (Chung et al., 2018; Cohen et al., 2020) the primary component of the geometry is the random structure of the whole P manifolds, and correlations enter as corrections on top of the random theory (Wakhloo et al., 2023). We showed that in our task the centroid-aligned component of within-class variability is the dominant part, while the random fluctuations of inter-centroid overlaps have little to no effect. The two frameworks are complementary regimes of the same underlying geometry, and it would be interesting to further investigate where the two descriptions meet.

The sparse code uncovered here also suggests links to interpretability and neuroscience. Sparse autoencoders and dictionary-learning methods seek low-dimensional, semantically meaningful directions in activations (Makhzani & Frey, 2013; Huben et al., 2024; Templeton et al., 2026; Gao et al., 2025); our results identify a complementary sparse class-supervised set of directions whose functional role is directly tied to classification. In parallel, deep networks have been shown to predict responses in primate ventral-stream areas, including IT cortex (Yamins et al., 2014; Schrimpf et al., 2020), and object-manifold theory has been applied to biological populations (Froudarakis et al., 2020; Sorscher et al., 2022; Kuoch et al., 2024). It would be interesting to examine whether the universal geometry we identify across many artificial models also applies to representations in the brain.

Finally, the theory points to practical applications. Because its ingredients are compact and inexpensive to estimate, they can serve as diagnostics for representation quality, to select which pretrained model is most suitable for fine-tuning on a new task (Kornblith et al., 2018). They could also suggest geometry-aware objectives: training could explicitly encourage weak rival alignment and sparse competition, while distillation could match the teacher and student geometry rather than the entire N-dimensional representation or the logit distribution (Hinton et al., 2015). Following these directions would turn the centroid-aligned description from a predictive theory of existing representations into a constructive principle for designing better ones.

8 AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic; Codex, OpenAI) to implement methods numerically, data analysis code, figure-generation code, and to edit the text. We have not used generative AI tools to formulate mathematical claims, to prove them, or to develop the theoretical model, and no data were created using generative AI. Translation is not applicable to this work. All AI-assisted text and code were thoroughly reviewed, and the reported results were checked for correctness and reproducibility. The authors take full responsibility for the final manuscript.

REFERENCES

Tony Alex, Sara Atito, Armin Mustafa, Muhammad Awais, and Philip Jackson. Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In *International Conference on Learning Representations*, volume 2025, pp. 22608–22626, 2025.

David Arps, Laura Kallmeyer, Younes Samih, and Hassan Sajjad. Multilingual nonce dependency treebanks: Understanding how language models represent and process syntactic structure. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7822–7844, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. Universal dependencies. *Computational linguistics*, 47(2):255–308, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. Ced: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295. IEEE, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

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.

Emmanouil Froudarakis, Uri Cohen, Maria Diamantaki, Saumil Patel, Zheng Tan, Taliah Muhammad, Edgar Y Walker, Jacob Reimer, Philipp Berens, Haim Sompolinsky, et al. Object manifold geometry across the mouse cortical visual hierarchy. *BioRxiv*, pp. 2020–08, 2020.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

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*, pp. 770–778, 2016.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018.

Michael Kuoch, Chi-Ning Chou, Nikhil Parthasarathy, Joel Dapello, James J DiCarlo, Haim Sompolinsky, and SueYeon Chung. Probing biological and artificial neural networks with task-dependent neural manifolds. In *Conference on Parsimony and Learning*, pp. 395–418. PMLR, 2024.

Peter A Lachenbruch and M Ray Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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.

Achleshwar Luthra, Yash Salunkhe, and Tomer Galanti. Directional neural collapse explains few-shot transfer in self-supervised learning. *arXiv preprint arXiv:2603.03530*, 2026.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022.

Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. *arXiv preprint arXiv:2105.14602*, 2021.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Albert J Wakhloo, Tamara J Sussman, and SueYeon Chung. Linear classification of neural manifolds with correlated variability. *Physical Review Letters*, 131(2):027301, 2023.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A SHARED GEOMETRY ACROSS ARCHITECTURES AND MODALITIES

The empirical structure of Sections 4-5 was characterized on DINOv3 ViT-7B. Table 4 reports the same measurements across a diverse set of vision, audio, and language backbones: Every observation persists: R is in an $\mathcal{O}(1)$ range for any N, with large accuracy fluctuations for fixed R (independent of N). δ-rotation drives the class-averaged accuracy to near-perfect on all backbones; the saturation accuracy ($R^\star \to \infty$) is consistently orders of magnitude above chance, indicating true-class dominance among the residual correlations. To check sparsity, for each example $\mathbf{x}$ of class k, we record its strongest rival, $\mu^*(\mathbf{x}) = \arg\max_{\mu \neq k} t_\mu(\mathbf{x})$, and count the number of distinct rivals that appear across the ~ 1300 examples of that class. Define $K_{90\%}$ as the number of rivals that are the strongest competitor of 90% of the class's examples. This rival set is sparse, $K_{90\%} \ll P$ in all cases. The projection s_μ right tail is heavy on every backbone. The geometric characterizations required by the theory of Section 5 are therefore present and consistent across different architectures, modalities, and training procedures. Appendix I and Figs. 10-13 report the theory-empirical comparison for all 36 backbones.

Table 4: Universality of the centroid-aligned geometry. For each backbone, we report: the number of classes P; feature dimension N; per-class radius range; per-class mean accuracy $\langle \mathrm{Acc} \rangle$ and its fluctuation around the mean accuracy at a fixed R; mean accuracy after δ-rotation; saturation accuracy at $R^\star \to \infty$ (controlled only by the residual $\delta\mathbf{x}$); $K_{90\%}$ (mean rivals covering 90% of examples); right-tail Student-t tail index ν (K=20).

Backbone	P	N	R range	$\langle \mathrm{Acc} \rangle$	Fixed R fluctuations	δ-rot	$R^\star \to \infty$	$\langle K_{90\%} \rangle$	ν
Vision – ImageNet-1K									
DINOv3 ViT-7B/16	1000	4096	[0.4, 4.3]	0.855	0.099	1.000	0.392	14.4	4.5
DINOv3 ConvNeXt-L	1000	1536	[0.2, 3.1]	0.814	0.100	1.000	0.312	14.2	4.4
CLIP ViT-L/14	1000	768	[0.6, 2.7]	0.792	0.124	0.999	0.077	13.9	16.4
ViT-B/16 (IN21K)	1000	768	[0.5, 2.7]	0.853	0.088	1.000	0.265	14.1	4.1
Audio – VGGSound									
SSLAM	310	768	[0.1, 3.7]	0.628	0.155	0.999	0.067	11.0	7.9
CED	310	768	[0.1, 4.0]	0.628	0.160	1.000	0.062	11.3	7.9
Language - TinyStories masked-token									
RoBERTa-large	320	1024	[0.6, 2.6]	0.767	0.108	1.000	0.120	13.3	9.6
ModernBERT-large	320	1024	[0.6, 2.7]	0.734	0.112	1.000	0.118	12.4	11.3
BERT-large	320	1024	[0.6, 2.7]	0.710	0.112	1.000	0.120	15.8	10.1
Language - TinyStories next-token									
Qwen3-8B (L34, -2)	320	4096	[0.9, 4.2]	0.540	0.118	1.000	0.107	33.7	11.3
OLMo2-13B (L36, -4)	320	5120	[1.0, 4.8]	0.563	0.116	1.000	0.112	35.8	10.2
Llama-3.1-8B (L29, -3)	320	4096	[0.8, 4.1]	0.550	0.119	1.000	0.101	31.7	12.4

B PROTOTYPE CLASSIFICATION

B.1 PROTOTYPE CLASSIFICATION AND LINEAR READOUT COMPARISON

One natural concern regarding our work is that it is applied only to prototype classification, with no relation to the more common linear readout. We test this by splitting the data into 90% train and 10% validation, training a max-margin SVM (using the Crammer–Singer method (Crammer & Singer, 2001)) on the train data, and evaluating on the validation set. We compare it with computing the centroids over the 90% train representations, and computing the prototype scores on the 10%. The gap between the two methods is small (see Fig.5 and Table 5), especially in vision (0.047) and autoregressive models (0.063). We note that in the bidirectional language models, the gap is substantially larger than in the other modalities (0.121), which is also reflected in a lower per-class correlation between the two classifiers ($r = 0.58$, against 0.87 in vision and 0.76 in the autoregressive models). The additional structure that allows the linear readout to separate the data better than the centroids in these models remains an open question.

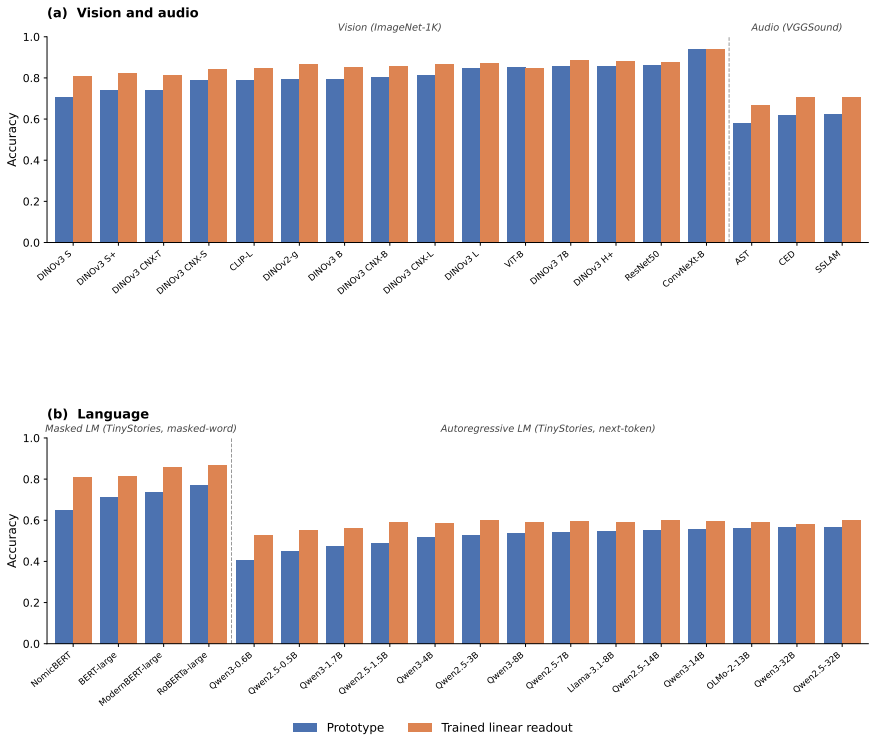

Figure 5: Centroid vs trained linear-readout accuracy, per backbone. The gap between methods is small (see also Table 5).

C Experimental Setup

The main text evaluates frozen representations from pretrained backbones spanning vision, language, and audio, across transformer and convolutional architectures, supervised, self-supervised, and contrastive objectives. This appendix gives the full list of datasets, models, and extraction details. No backbone is fine-tuned or changed for our analysis.

C.1 Datasets

Vision-ImageNet-1K: We use ImageNet-1K (ILSVRC 2012) (Deng et al., 2009), the standard 1000-class image benchmark with $\approx$1.28M images.

Language-TinyStories: For language, the task is token prediction on the validation dataset of TinyStories (Eldan & Li, 2023), a corpus of 20K short synthetic children's stories. We treat token prediction as a multi-class classification problem.

Content words and lemmatizing: In order to get semantically meaningful representations, we keep only data points of *content words*. Following Arps et al. (2024), content words are those whose part-of-speech is an open lexical class of the Universal Dependencies scheme (De Marneffe et al., 2021): NOUN, PROPN, VERB, ADJ, or ADV. We assign part-of-speech tags and lemmas with the spaCy pipeline (Honnibal et al., 2020), and unify all tokens sharing a lemma into a single class (for example, "say, said, says" or "friend, friends"). We keep only classes with at least 1000 occurrences, so that the manifold geometry is well estimated. Finally, to obtain an identical class set across models with different tokenizers, we discard the 11 lemmas that are not realized consistently in every tokenizer - fragmented or unified into/from sub-words differently in at least one model ("benny, bunny, fire, frog, grandma, grow, mia, mommy, sweet, window, yummy", for example one tokenizer treats "fireman" $\mapsto$ "fire+man"). This leaves 320 classes shared by all backbones.

Table 5: Per-class Pearson correlation r between trained linear-readout and prototype-classifier accuracy, together with the mean accuracy gap $\Delta\mathrm{Acc}$ (readout minus prototype) on the held-out split. The two classifiers align closely, with the exception of masked language models (see Sec. B.1).

Backbone	r	ΔAcc
Vision (ImageNet-1K)		
ViT-B/16 (IN21K)	0.937	−0.005
ConvNeXt-B (IN1K)	0.893	−0.001
DINOv3 ViT-L/16	0.885	+0.028
DINOv3 ViT-H$^+$/16	0.873	+0.028
DINOv3 ConvNeXt-S	0.872	+0.058
CLIP ViT-L/14	0.869	+0.057
DINOv3 ConvNeXt-T	0.865	+0.076
DINOv3 ViT-S$^+$/16	0.864	+0.084
ResNet-50	0.861	+0.016
DINOv3 ViT-S/16	0.859	+0.100
DINOv3 ViT-B/16	0.857	+0.058
DINOv3 ConvNeXt-B	0.856	+0.054
DINOv3 ConvNeXt-L	0.854	+0.052
DINOv3 ViT-7B/16	0.839	+0.030
DINOv2 ViT-g/14	0.791	+0.074
mean	**0.865**	**+0.047**
Audio (VGGSound)		
SSLAM	0.821	+0.084
AST	0.815	+0.091
CED	0.808	+0.088
mean	**0.815**	**+0.088**
Masked LM (TinyStories, masked-word)		
ModernBERT-large	0.584	+0.126
NomicBERT	0.581	+0.159
BERT-large	0.568	+0.102
RoBERTa-large	0.565	+0.096
mean	**0.575**	**+0.121**
Autoregressive LM (TinyStories, next-token)		
Qwen3-4B	0.785	+0.070
Qwen3-1.7B	0.777	+0.089
Qwen3-8B	0.772	+0.050
Qwen2.5-0.5B	0.769	+0.102
Qwen3-0.6B	0.765	+0.122
Llama-3.1-8B	0.764	+0.041
Qwen2.5-3B	0.762	+0.076
Qwen3-32B	0.761	+0.019
Qwen3-14B	0.754	+0.038
OLMo-2-13B	0.750	+0.029
Qwen2.5-1.5B	0.749	+0.102
Qwen2.5-32B	0.747	+0.036
Qwen2.5-7B	0.740	+0.052
Qwen2.5-14B	0.736	+0.050
mean	**0.759**	**+0.063**
Overall	**0.787**	**+0.065**

Masked-word task (bidirectional encoders). We mask a single content-word position at a time, run a full bidirectional forward pass, and read out the encoder's representation at the masked position. The representation is the final hidden state feeding the model's masked-LM head. The class label is the *lemma* of the masked word, restricted to the same open-class content-word construction as above.

Next-token task (autoregressive models): From a standard causal forward pass, the representation is the hidden state (at a specified layer) at the last token *preceding* the predicted word, and the class label is the lemma of the actual next word, restricted to the same open-class content-word construction as above.

Audio - VGGSound. For audio we use VGGSound (Chen et al., 2020), a large-scale dataset of $\sim$10-second clips sourced from web video. We evaluate on the train split: 310 single-label classes and $183,630$ clips (~ 592 for each class), decoded to 16 kHz mono.

C.2 Models

The canonical roster comprises 36 pretrained backbones across the four modality groups (Table 6). Vision contains the self-supervised DINOv3 family (Siméoni et al., 2025): DINOv3 ViT-S/16 through ViT-7B/16 and DINOv3 ConvNeXt-T through ConvNeXt-L, together with the self-supervised DINOv2 ViT-g/14 (Oquab et al., 2023), contrastive CLIP ViT-L/14 (Radford et al., 2021), the supervised ImageNet-21K ViT-B/16 (Dosovitskiy et al., 2020), the supervised ImageNet-1K ConvNeXt-B (Liu et al., 2022), and the supervised ResNet-50 (He et al., 2016).

The language backbones are evaluated on the TinyStories dataset (Eldan & Li, 2023). The masked-word encoders are RoBERTa-large (Liu et al., 2019), BERT-large (Devlin et al., 2019), ModernBERT-large (Warner et al., 2025), and NomicBERT (Nussbaum et al., 2024). The next-token autoregressive models are the Qwen2.5-Instruct family (0.5B–32B) (Qwen et al., 2025), the Qwen3 family (0.6B–32B) (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and OLMo-2-13B (OLMo et al., 2024).

Table 6: Model roster (36 backbones). N is the feature dimension of the extracted representation; P is the number of classes in the task.

Family	Model	N	Data / task	P	Objective
Vision — ImageNet-1K					
Vision	DINOv3 ViT-S/16	384	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ViT-S+/16	384	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ViT-B/16	768	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ViT-L/16	1024	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ViT-H+/16	1280	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ViT-7B/16	4096	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ConvNeXt-T	768	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ConvNeXt-S	768	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ConvNeXt-B	1024	ImageNet-1K	1000	Self-supervised
Vision	DINOv3 ConvNeXt-L	1536	ImageNet-1K	1000	Self-supervised
Vision	DINOv2 ViT-g/14	1536	ImageNet-1K	1000	Self-supervised
Vision	CLIP ViT-L/14	768	ImageNet-1K	1000	Contrastive (image-text)
Vision	ViT-B/16 (IN21K)	768	ImageNet-1K	1000	Supervised (IN21K)
Vision	ConvNeXt-B (IN1K)	1024	ImageNet-1K	1000	Supervised (IN1K)
Vision	ResNet-50	2048	ImageNet-1K	1000	Supervised (IN1K)
Audio — VGGSound					
Audio	SSLAM	768	VGGSound	310	Self-supervised + AudioSet-2M FT
Audio	CED-base	768	VGGSound	310	Supervised distillation (AudioSet)
Audio	AST	768	VGGSound	310	Supervised (AudioSet)
Masked LM — TinyStories masked-word					
Masked LM	RoBERTa-large	1024	TinyStories (masked)	320	Masked language modeling
Masked LM	BERT-large	1024	TinyStories (masked)	320	Masked language modeling
Masked LM	ModernBERT-large	1024	TinyStories (masked)	320	Masked language modeling
Masked LM	NomicBERT	768	TinyStories (masked)	320	Masked language modeling
Autoregressive LM — TinyStories next-token					
Autoregressive LM	Qwen2.5-0.5B	896	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen2.5-1.5B	1536	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen2.5-3B	2048	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen2.5-7B	3584	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen2.5-14B	5120	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen2.5-32B	5120	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-0.6B	1024	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-1.7B	2048	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-4B	2560	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-8B	4096	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-14B	5120	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Qwen3-32B	5120	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	Llama-3.1-8B	4096	TinyStories (next-token)	320	Causal language modeling
Autoregressive LM	OLMo-2-13B	5120	TinyStories (next-token)	320	Causal language modeling

The audio backbones are SSLAM (Alex et al., 2025), an EAT-style self-supervised model fine-tuned on AudioSet-2M; CED-base (Dinkel et al., 2024), an AudioSet-supervised distilled transformer; and AST (Gong et al., 2021), an AudioSet-supervised spectrogram transformer.

C.3 Representation extraction

For vision transformers, we take the class token of the final layer as the representation. For convolutional architectures (ConvNeXt, ResNet), we take the global-average-pooled feature vector of the final stage. For audio backbones, we take the model's pre-classifier pooled embedding: SSLAM its fc-norm of the class token, CED-base the mean over its final-layer hidden states, AST its standard pre-classifier pooled outputs. For masked language models the representation is the final hidden state at the masked position — the input to the masked-LM head. For autoregressive language models the representation is the hidden state (in a specific layer, see Sec. H) of the last token preceding the predicted word. In every case, the feature dimensions N reported in Table 6 are those of the stored arrays.

Before any analysis, we z-score each representation dimension-wise, using the empirical mean and standard deviation of the data.

C.4 RESOURCES

Feature extraction ran on GPU (A100 and H100); the numerical integration of the theory and the fitting of the rescaling factor λ ran on CPU. The largest extractions were Qwen-14B/32B and OLMo on language, and DINOv3 ViT-7B on vision.

D THEORETICAL PROOFS

This appendix contains all material deferred from Sec. 3 and Sec. 5. Appendix D.1 works out the isotropic uncorrelated model and derives the corollaries that show how the model fails qualitatively on real data. Appendix D.2 gives the full proofs of the centroid-aligned theory used in the main text.

D.1 ISOTROPIC MODEL

We analyze the maximally unstructured baseline, in which class centroids and within-class fluctuations are both isotropic Gaussian. It retains only the ambient dimension N, the number of classes P, and the noise-to-signal scale R, discarding every correlation between centroids and fluctuations. We derive its exact large-N accuracy, show that accuracy is self-averaging, and extract corollaries delineating its regimes. In Section 4 these predictions are shown to match real representations only after the residual-centroid correlations are removed.

Assumption D.1.1 *(Isotropic uncorrelated model).* For each class $\mu = 1, \dots, P$ the centroid is drawn $\mathbf{c}_\mu \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, I_N)$ and held fixed (quenched). Conditioned on $\mathbf{c}_k$, an example of class k is $\mathbf{x} = \mathbf{c}_k + \delta\mathbf{x}$ with $\delta\mathbf{x} \sim \mathcal{N}(\mathbf{0}, R^2 I_N)$, sampled independently across examples.

Assumption D.1.2 *(High-dimensional regime).* $N \to \infty$, the number of classes grows at most polynomially, $P = N^{\mathcal{O}(1)}$, and the radius $R = R(N)$ may scale with N.

We refer to Assumptions D.1.1–D.1.2 jointly as the *isotropic model*. The classifier is the unit-centroid (prototype) rule $\hat{y}(\mathbf{x}) = \arg\max_\mu \hat{\mathbf{c}}_\mu \cdot \mathbf{x}$, with $\hat{\mathbf{c}}_\mu := \mathbf{c}_\mu / \|\mathbf{c}_\mu\|$, and its logits are $t_\mu := \hat{\mathbf{c}}_\mu \cdot \mathbf{x}$.

Lemma D.1.3 *(Overlaps).* Let $\mathbf{u}, \mathbf{v} \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, I_N)$. Then, as $N \to \infty$, $\|\mathbf{u}\|/\sqrt{N} \xrightarrow{p} 1$ and $\sqrt{N}\,(\hat{\mathbf{u}} \cdot \hat{\mathbf{v}}) \xrightarrow{d} \mathcal{N}(0, 1)$. Consequently, the dot product of the target centroid with the true and rival directions obeys

$$\hat{\mathbf{c}}_\mu \cdot \mathbf{c}_k = \varepsilon_\mu + \delta_{\mu k}\big(\sqrt{N} - \varepsilon_\mu\big), \qquad \varepsilon_\mu \sim_{\text{iid}} \mathcal{N}(0, 1). \tag{8}$$

Proof. $\mathbf{u} \cdot \mathbf{v} = \sum_i u_i v_i$ is a sum of N i.i.d. centred unit-variance variables, so $\mathbf{u} \cdot \mathbf{v}/\sqrt{N} \xrightarrow{d} \mathcal{N}(0, 1)$ by the CLT, while $\|\mathbf{u}\|/\sqrt{N} \xrightarrow{p} 1$ by the law of large numbers. □

Theorem D.1.4 (Mean accuracy). *Under the isotropic model,*

$$\overline{\mathbb{E}_{\delta\mathbf{x}}[\mathrm{Acc}]} = \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\Phi^{P-1} \left(\frac{\sqrt{N} + Rz}{\sqrt{1 + R^2}} \right) \right]. \tag{9}$$

where $\overline{\mathbb{E}_{\delta\mathbf{x}}[\mathrm{Acc}]}$ denotes the average over the quenched disorder $\{\mathbf{c}_\mu\}$ after averaging over the thermal $\delta\mathbf{x}$.

Proof. Conditional on the example and centroids, the accuracy is the probability that the correct logit is the largest,

$$\mathrm{Acc} \;\big|\; \delta\mathbf{x}, \{\mathbf{c}_\nu\} = \int_{-\infty}^{\infty} dt_k \, P(t_k) \prod_{\mu \neq k} \int_{-\infty}^{t_k} dt_\mu \, P(t_\mu), \qquad P(t_\mu) = \delta\big(t_\mu - \hat{\mathbf{c}}_\mu \cdot \mathbf{x}\big). \tag{10}$$

Writing each δ through its Fourier representation introduces conjugate variables $\hat{t}_\mu$:

$$\mathrm{Acc} \;\big|\; \delta\mathbf{x}, \{\mathbf{c}_\nu\} = \int_{-\infty}^{\infty} \frac{dt_k \, d\hat{t}_k}{2\pi} \prod_{\mu \neq k} \int_{-\infty}^{t_k} \frac{dt_\mu \, d\hat{t}_\mu}{2\pi} \exp\Big(i \sum_{\mu=1}^{P} \hat{t}_\mu \big(t_\mu - \hat{\mathbf{c}}_\mu \cdot (\mathbf{c}_k + \delta\mathbf{x})\big)\Big). \tag{11}$$

Carrying out the Gaussian $\delta\mathbf{x}$ and $\hat{t}_\mu$ integrals,

$$\mathbb{E}_{\delta x}\big[\,\mathrm{Acc}\;\big|\;\{\mathbf{c}_\mu\}\big] = \int_{-\infty}^{\infty}\frac{dt_k}{\sqrt{2\pi R^2}}\prod_{\mu\neq k}\int_{-\infty}^{t_k}\frac{dt_\mu}{\sqrt{2\pi R^2}}\exp\Big(-\frac{1}{2R^2}\sum_{\mu=1}^{P}\big(t_\mu-\hat{\mathbf{c}}_\mu\cdot\mathbf{c}_k\big)^2\Big). \tag{12}$$

Lemma D.1.3 sets $\hat{\mathbf{c}}_\mu\cdot\mathbf{c}_k = \varepsilon_\mu + \delta_{\mu k}(\sqrt{N}-\varepsilon_\mu)$. The target mean concentrates at $\sqrt{N}$, while integrating each rival overlap $\varepsilon_\mu\sim\mathcal{N}(0,1)$ inflates the rival variance $R^2\to 1+R^2$:

$$\overline{\mathrm{Acc}} = \int_{-\infty}^{\infty}\frac{dt_k}{\sqrt{2\pi R^2}}\prod_{\mu\neq k}\int_{-\infty}^{t_k}\frac{dt_\mu}{\sqrt{2\pi(1+R^2)}}\exp\Big(-\frac{(t_k-\sqrt{N})^2}{2R^2}-\frac{1}{2}\sum_{\mu\neq k}\frac{t_\mu^2}{1+R^2}\Big). \tag{13}$$

The rival integrals factorize into $\Phi\big(t_k/\sqrt{1+R^2}\big)$; substituting $t_k=\sqrt{N}+Rz$ yields Eq. 9. □

The average Eq. 9 is taken over both variability $\delta\mathbf{x}$ and centroids; the next result shows that the accuracy is self-averaging, such that taking the annealed average over the centroids disorder is valid.

Proposition D.1.5 *(Self-averaging).* Under the isotropic model, the accuracy is self-averaging. $\mathrm{Var}(\mathrm{Acc}) = \mathcal{O}\big(1/(PR^2)\big)\to 0$.

Proof. We bound the variance of the disorder-averaged accuracy. Introduce two independent replicas $\mathbf{x},\mathbf{x}'$ of class k sharing the centroids. As in Eq.12, we get,

$$\begin{aligned}\mathbb{E}_{\delta x}\big[\,\mathrm{Acc}^2\;\big|\;\{\mathbf{c}_\mu\}\big] = \int_{-\infty}^{\infty}\frac{dt_k\,dt_k'}{2\pi R^2}\prod_{\mu\neq k}\int_{-\infty}^{t_k}\frac{dt_\mu}{\sqrt{2\pi R^2}}\int_{-\infty}^{t_k'}\frac{dt_\mu'}{\sqrt{2\pi R^2}}\\ \exp\Big(-\frac{1}{2R^2}\sum_{\mu}\big[(t_\mu-\hat{\mathbf{c}}_\mu\cdot\mathbf{c}_k)^2+(t_\mu'-\hat{\mathbf{c}}_\mu\cdot\mathbf{c}_k)^2\big]\Big).\end{aligned} \tag{14}$$

Averaging over the centroids couples the replicas through the shared mean. The integral becomes bivariate normal with variance $1+R^2$ and covariance 1,

$$\begin{aligned}\overline{\mathrm{Acc}^2} = \int\frac{dt_k\,dt_k'}{2\pi R^2}\prod_{\mu\neq k}\frac{1}{\mathcal{Z}}\int_{-\infty}^{t_k+\sqrt{N}}dt_\mu\int_{-\infty}^{t_k'+\sqrt{N}}dt_\mu'\\ \exp\Big(-\frac{t_k^2+t_k'^2}{2R^2}-\frac{(1+R^2)(t_\mu^2+t_\mu'^2)-2t_\mu t_\mu'}{2R^2(R^2+2)}\Big),\end{aligned} \tag{15}$$

with the normalization $\mathcal{Z}=2\pi R\sqrt{R^2+2}$.

The regime $R=\mathcal{O}(1)$: The target shift $\sqrt{N}\to\infty$ pushes the rival limits to $+\infty$, the rival integrals yield 1 up to an exponentially small correction whenever $\ln P = o(N)$, and $\overline{\mathrm{Acc}^2}=\overline{\mathrm{Acc}}^2=1$, so $\mathrm{Var}\to 0$.

The regime $R\sim\sqrt{N}$: Set $r=R/\sqrt{N}$ and rescale. The inter-replica coupling is weak, of order $1/(Nr^2)$, and we can expand in the coupling. Solving the uncoupled Gaussian integrals yields:

$$\begin{aligned}\overline{\mathrm{Acc}^2} &= \mathbb{E}_{z,z'}\Big[\Big(\Phi(z+\tfrac{1}{r})\Phi(z'+\tfrac{1}{r})+\tfrac{1}{Nr^2}\phi(z+\tfrac{1}{r})\phi(z'+\tfrac{1}{r})+\mathcal{O}(N^{-2})\Big)^{P-1}\Big]\\ &= \overline{\mathrm{Acc}}^2+\frac{P-1}{Nr^2}\Big(\mathbb{E}_z\big[\Phi^{P-2}(z+\tfrac{1}{r})\phi(z+\tfrac{1}{r})\big]\Big)^2+\mathcal{O}\Big(\tfrac{\ln P}{N^2P^2}\Big).\end{aligned} \tag{16}$$

where $\phi(x)$ is a standard normal PDF, and $\Phi(x)$ its CDF. The expectation can be performed using the Laplace method, yielding the classic extreme-value extremum $z^*+1/r=\sqrt{2\ln P}$, $\phi(z^*+1/r)=\sqrt{2\ln P}/P$, and the width of the Laplace region $\Delta z = 1/\Big(2\sqrt{\ln P}\Big)$. We note that the higher binomial terms are suppressed not by the $1/N$ prefactors but by the Gaussian PDF densities ϕ, each one of the order $\sqrt{2\ln P}/P$, which cancel the binomial growth.

$$\mathrm{Var}(\mathrm{Acc}) = \frac{P-1}{Nr^2}\Big(\mathbb{E}_z\big[\Phi^{P-2}(z+\tfrac{1}{r})\phi(z+\tfrac{1}{r})\big]\Big)^2 = \mathcal{O}(\frac{1}{PNr^2}) = \mathcal{O}(\frac{1}{PR^2})\to 0 \tag{17}$$

In both regimes $\mathrm{Var}(\mathrm{Acc}) \to 0$. □

The mean accuracy Eq.9 has four immediate consequences, delineating the regimes of the isotropic model.

Corollary D.1.6 *(Separable regime).* For $R = \mathcal{O}(1)$ the classes are well separated and classification is trivial: $\mathrm{Acc} \to 1$.

Proof. The scale of the maximum of P i.i.d. Gaussians with variance $\sigma^2 = 1 + R^2$ converges to $\sigma\sqrt{2\ln P}$ for large P. The signal is $\mathbb{E}[t_k] = \sqrt{N}$. From Assumption D.1.2, $\ln P = \mathcal{O}(\ln N)$ thus for any $R = \mathcal{O}(1)$, $\sigma\sqrt{2\ln P} \ll \sqrt{N}$ which yields $\mathrm{Acc} \to 1$. □

Corollary D.1.7 *(Saturation regime).* In the infinite-radius limit $R/\sqrt{N} \to \infty$ the isotropic residual dominates the input, which retains no information about the true class: all P classes are equally likely and accuracy collapses to chance, $\mathrm{Acc} \to 1/P$.

Proof. Dividing through by R, the argument of Φ in Eq.9 tends to z pointwise, so bounded convergence gives $\mathrm{Acc} \to \mathbb{E}_z[\Phi^{P-1}(z)] = \mathbb{P}(\xi_k > \max_{\mu\neq k}\xi_\mu) = 1/P$ for $\xi_1, \ldots, \xi_P \sim_{\mathrm{iid}} \mathcal{N}(0,1)$, by permutation symmetry. □

Corollary D.1.8 *(Critical radius).* Non-perfect accuracy requires large radii:
$\mathrm{Acc} < 1 \Rightarrow R = \Omega(\sqrt{N/\ln P})$.

Proof. Contrapositive. If $R = o(\sqrt{N/\ln P})$ then $1 + R^2 = o(N/\ln P)$, so the argument of Φ in Eq.9 is $(\sqrt{N} + Rz)/\sqrt{1+R^2} = \omega(\sqrt{\ln P}) \gg \sqrt{2\ln P}$, and Corollary D.1.6 forces $\mathrm{Acc} \to 1$. □

Corollary D.1.9 *(Per-class fluctuations).* Per-class accuracy fluctuations are of order $1/R\sqrt{P}$.

Proof. Calculation of the variance in Proposition D.1.5 yields $\mathrm{Var}(\mathrm{Acc}) = \mathcal{O}(1/PR^2)$. Different classes are equivalent to a resampling of the centroids, so per-class fluctuations are of the same order. □

D.2 Centroid-aligned model

Definition D.2.1 *(Centroid geometry).* Fix the true class k. For $\mu, \nu \in \{1, \ldots, P\}$ let $G_{\mu\nu} := \hat{\mathbf{c}}_\mu \cdot \hat{\mathbf{c}}_\nu$ be the Gram matrix of the unit centroids, and let $g_\mu := \hat{\mathbf{c}}_k \cdot \hat{\mathbf{c}}_\mu$ be the centroid overlap of class μ with the true class, so that $g_k = 1$. The rival support is $\mathcal{R}_K$ and $\mathcal{R}_K^+ := \{k\} \cup \mathcal{R}_K$, as defined in Sec. 5.

Assumption D.2.2 *(Deterministic centroids and Gaussian projections).* We assume that the centroids $\{\mathbf{c}_\mu\}_{\mu=1}^P$ are deterministic and held fixed, and that the restricted Gram matrix of $\mathcal{R}_K^+$ is invertible. The source of stochasticity in the system is the sampling of Gaussian projection coefficients $\{s_\mu\}_{\mu=1}^P$ and $s_\perp$.

Definition D.2.3 *(Centroid-aligned generative model).* An example of class k is generated by

$$\mathbf{x}(\mathbf{s}) = \|\mathbf{c}_k\| \left[\hat{\mathbf{c}}_k + R \left(\sum_{\mu,\nu \in \mathcal{R}_K^+} \sigma_\mu s_\mu G^{-1}_{\mu\nu} \hat{\mathbf{c}}_\nu + \sigma_\perp s_\perp \hat{\mathbf{e}}_\perp \right) \right], \tag{18}$$

where $\hat{\mathbf{e}}_\perp$ is orthogonal to $\mathrm{span}(\hat{\mathbf{c}}_\mu)_{\mu\in\mathcal{R}_K^+}$, R is the empirical class radius, and σ_μ is the projected standard deviation along $\hat{\mathbf{c}}_\mu$.

Definition D.2.4 *(Normalized logits).* The prototype logits of (Def 2.1) are defined as $t_\mu = \hat{\mathbf{c}}_\mu \cdot \mathbf{x}$. Throughout this section it is convenient to divide all of them by the common factor $\|\mathbf{c}_k\|$, and we set

$$t_\mu := \frac{\mathbf{x}\cdot\hat{\mathbf{c}}_\mu}{\|\mathbf{c}_k\|}, \qquad \mu \in \mathcal{R}_K^+. \tag{19}$$

Since $\|\mathbf{c}_k\|$ is a common factor independent of μ, the argmax properties of the logits do not change. The classifier and its accuracy are therefore those of Definition 2.1, and no generality is lost. We restrict the logits only to the sparse set of rivals $\mathcal{R}_K$, and neglect weak competitors.

Lemma D.2.5 *(Logit representation).* Under Assumption D.2.2 and Definitions D.2.3–D.2.4, $t_\mu = g_\mu + R\,\sigma_\mu s_\mu$ for every $\mu \in \mathcal{R}_K^+$.

Proof. Taking the inner product of Eq. 18 with $\hat{\mathbf{c}}_\mu$ and dividing by $\|\mathbf{c}_k\|$ gives $t_\mu = g_\mu + R\big(\sum_{\alpha,\nu\in\mathcal{R}_K^+} \sigma_\alpha s_\alpha G^{-1}_{\alpha\nu} G_{\nu\mu} + \sigma_\perp s_\perp\, \hat{\mathbf{e}}_\perp \cdot \hat{\mathbf{c}}_\mu\big)$. The last term vanishes since $\hat{\mathbf{e}}_\perp \perp \hat{\mathbf{c}}_\mu$ for $\mu \in \mathcal{R}_K^+$, and $\sum_\nu G^{-1}_{\alpha\nu} G_{\nu\mu} = \delta_{\alpha\mu}$ by Assumption D.2.2. □

Definition D.2.6 *(Projection covariance).* The coefficients are centered and standardized by definition, $\mathbb{E}_k[s_\mu] = 0$ and $\mathbb{E}_k[s_\mu^2] = 1$. We write

$$\mathbb{E}_k[s_\mu s_\nu] = \frac{\Sigma_{\mu\nu}}{R^2\sigma_\mu\sigma_\nu}, \qquad \mu,\nu \in \mathcal{R}_K^+. \tag{20}$$

By Lemma D.2.5 this makes $\Sigma_{\mu\nu} = \mathrm{Cov}_k[t_\mu, t_\nu]$ the covariance matrix of the logit vector itself. No structure is imposed on Σ beyond that of a well-conditioned covariance matrix.

Proposition D.2.7 *(Accuracy as an orthant probability).* Under Lemma D.2.5, Definition D.2.6 and Assumption D.2.2,

$$\mathrm{Acc}^{\mathrm{th}}(R) = \frac{1}{\sqrt{\det\Sigma}} \int_{-\infty}^{\infty} \frac{dt_k}{\sqrt{2\pi}} \prod_{\mu\in\mathcal{R}_K} \int_{-\infty}^{t_k} \frac{dt_\mu}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_{\mu,\nu\in\mathcal{R}_K^+} (t_\mu - g_\mu)\, \Sigma^{-1}_{\mu\nu}\, (t_\nu - g_\nu) \right). \tag{21}$$

Proof. By Lemma D.2.5, Definition D.2.6 and Assumption D.2.2 the vector $(t_\mu)_{\mu\in\mathcal{R}_K^+}$ is Gaussian with mean (g_μ) and covariance Σ. An example is classified correctly iff $t_\mu < t_k$ for all $\mu \in \mathcal{R}_K$. □

Eq. 21 is a multivariate orthant probability, which in general can be evaluated only numerically (for example by Monte-Carlo simulation). A specific structure admits an analytical solution.

Lemma D.2.8 *(Gaussian conditioning).* Under Assumption D.2.2, set $\rho_\mu := \mathbb{E}_k[s_\mu s_k]$ for $\mu \in \mathcal{R}_K$. Then for all $\mu,\nu \in \mathcal{R}_K$,

$$\mathbb{E}_k[t_\mu \mid t_k] = g_\mu + \rho_\mu\sigma_\mu \frac{t_k - 1}{\sigma_k}, \qquad \mathrm{Cov}[t_\mu, t_\nu \mid t_k] = R^2\sigma_\mu\sigma_\nu\,(\mathbb{E}_k[s_\mu s_\nu] - \rho_\mu\rho_\nu)\,, \tag{22}$$

and in particular the conditional covariance does not depend on t_k.

Proof. By Definition D.2.6, $\mathrm{Cov}(t_\mu, t_\nu) = \Sigma_{\mu\nu}$, $\Sigma_{\mu k} = R^2\sigma_\mu\sigma_k\rho_\mu$ and $\Sigma_{kk} = R^2\sigma_k^2$, while $\mathbb{E}_k[t_\mu] = g_\mu$ and $\mathbb{E}_k[t_k] = 1$. For a jointly Gaussian vector, $\mathbb{E}[t_\mu \mid t_k] = \mathbb{E}[t_\mu] + \Sigma_{\mu k}\Sigma_{kk}^{-1}(t_k - \mathbb{E}[t_k])$ and $\mathrm{Cov}[t_\mu, t_\nu \mid t_k] = \Sigma_{\mu\nu} - \Sigma_{\mu k}\Sigma_{\nu k}\Sigma_{kk}^{-1}$. Substituting gives Eq. 22. □

Assumption D.2.9 *(Conditional independence).* Conditionally on t_k the rival logits are uncorrelated: $\mathrm{Cov}[t_\mu, t_\nu \mid t_k] = 0$ for all $\mu \neq \nu$ in $\mathcal{R}_K$. Equivalently, by Lemma D.2.8,

$$\mathbb{E}_k[s_\mu s_\nu] = \delta_{\mu\nu}\left(1 - \rho_\mu^2\right) + \rho_\mu\rho_\nu, \qquad \text{so that} \qquad \mathrm{Cov}[t_\mu, t_\nu \mid t_k] = \delta_{\mu\nu} R^2\sigma_\mu^2\left(1 - \rho_\mu^2\right). \tag{23}$$

Theorem D.2.10 *(Conditional factorization).* Under Assumption D.2.9,

$$\mathrm{Acc}^{\mathrm{th}}(R) = \int_{-\infty}^{\infty} \frac{dt_k}{R\sigma_k\sqrt{2\pi}} \exp\!\left(-\frac{(t_k - 1)^2}{2R^2\sigma_k^2} \right) \prod_{\mu\in\mathcal{R}_K} \Phi\left(\frac{t_k\,(1 - \rho_\mu\sigma_\mu/\sigma_k) + \rho_\mu\sigma_\mu/\sigma_k - g_\mu}{R\sigma_\mu\sqrt{1 - \rho_\mu^2}} \right). \tag{24}$$

Proof. Condition on t_k. By Lemma D.2.8 and Assumption D.2.9 the variables $(t_\mu)_{\mu\in\mathcal{R}_K}$ are then independent Gaussians with mean $g_\mu + \rho_\mu\sigma_\mu(t_k - 1)/\sigma_k$ and variance $R^2\sigma_\mu^2(1 - \rho_\mu^2)$, so the orthant probability factorizes into $\prod_\mu \Pr[t_\mu < t_k \mid t_k]$, which yields Eq. 24. □

Corollary D.2.11 *(Standardized form).* Substituting $t_k = R\sigma_k z + 1$ in Eq. 24 yields

$$\mathrm{Acc}_k^{\mathrm{th}}(R) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\prod_{\mu\in\mathcal{R}_K}\Phi\left(\frac{\sigma_k-\rho_\mu\sigma_\mu}{\sigma_\mu\sqrt{1-\rho_\mu^2}}\,z+\frac{1-g_\mu}{R\sigma_\mu\sqrt{1-\rho_\mu^2}}\right)\right], \tag{25}$$

which is Theorem 5.5 of the main text.

Proposition D.2.12 *(Unconditional moments).* Under Assumption D.2.9, for $\mu,\nu\in\mathcal{R}_K$,

$$\mathbb{E}_k[t_\mu] = g_\mu, \qquad \Sigma_{\mu\nu} = \mathrm{Cov}_k[t_\mu,t_\nu] = R^2\left(\delta_{\mu\nu}\sigma_\mu^2\left(1-\rho_\mu^2\right)+\sigma_\mu\sigma_\nu\rho_\mu\rho_\nu\right), \tag{26}$$

which is a rank-one correction to a diagonal covariance. This is the covariance structure of $\mathbf{s}$ stated in Assumption 5.4 of the main text.

Proof. From Lemma D.2.5, $\mathbb{E}_k[t_\mu] = g_\mu$ and $\Sigma_{\mu\nu} = R^2\sigma_\mu\sigma_\nu\mathbb{E}_k[s_\mu s_\nu]$. Inserting the covariance structure from Eq. 23 yields the result. □

Remark D.2.13 *(Quenched versus deterministic centroids).* Assumption D.2.2 treats the centroids as deterministic. Drawing them instead from a Gaussian ensemble whose second moments reproduce the overlap structure G does not change the result for large N: the induced fluctuations of the coefficients g_μ, σ_μ and ρ_μ entering Eq. 24 are subleading corrections to their $\mathcal{O}(1)$ values, and the accuracy can be shown to be self-averaging by the same argument as in Proposition D.1.5.

E Synthetic Data

We validate the two generative models of the main text on synthetic data, where the geometry is controlled exactly and the analytic predictions can be checked without any of the confounds of real representations. Figure 6 tests the isotropic uncorrelated model of Section 3, and Figure 7 tests the centroid-aligned model of Section 5. For each model we verify three claims: (i) the analytic accuracy matches the simulation; (ii) the accuracy is self-averaging, i.e. a single quenched centroid bank gives the same per-class accuracy as a fresh, independently drawn set of rivals for every example; and (iii) the fluctuation of the accuracy across realizations shrinks with the number of classes P and for the isotropic model also with the class radius R as predicted by the previous section.

F Rival Selection

For a given example, only a small number of the P classes are genuine competitors. This is the reason that the theory of Section 5 evaluates the accuracy over a small rival set $\mathcal{R}_K$. This appendix (i) compares four possible rival-selection rules on convergence to the full P-way classification (ii) shows that the theory's error saturates by $K\approx$ 10-20, so $K=20$ is a principled choice.

F.1 Rival-selection rules

Fix a class k. We define a ranking method for every candidate rival $\mu\neq k$:
Centroid correlations: $g_\mu = \hat{\mathbf{c}}_k\cdot\hat{\mathbf{c}}_\mu$, (descending).
Projection variance: $\sigma_\mu = \sqrt{\mathbb{E}_k\big[(\delta\mathbf{x}\cdot\hat{\mathbf{c}}_\mu)^2\big]/\mathbb{E}_k\big[\|\delta\mathbf{x}\|^2\big]}$, (descending).
Theory margin: $(1-g_\mu)/\left(R\sigma_\mu\sqrt{1-\rho_\mu^2}\right)$, (ascending). The z-independent argument in Eq.6.
Gaussian margin: $(1-g_\mu)/\sqrt{\sigma_k^2+\sigma_\mu^2-2\,\sigma_k\sigma_\mu\rho_\mu}$, (ascending). The exact z-expectation over Eq.6 for one rival.

F.2 Convergence of the K-way classification accuracy

Restricting the classification to a smaller rival set can only improve accuracy, so the K-way accuracy exceeds the full P-way accuracy and decreases monotonically to it as K grows. Table 7 reports, for each selector, the median number of rivals a backbone needs before its K-way accuracy is within a

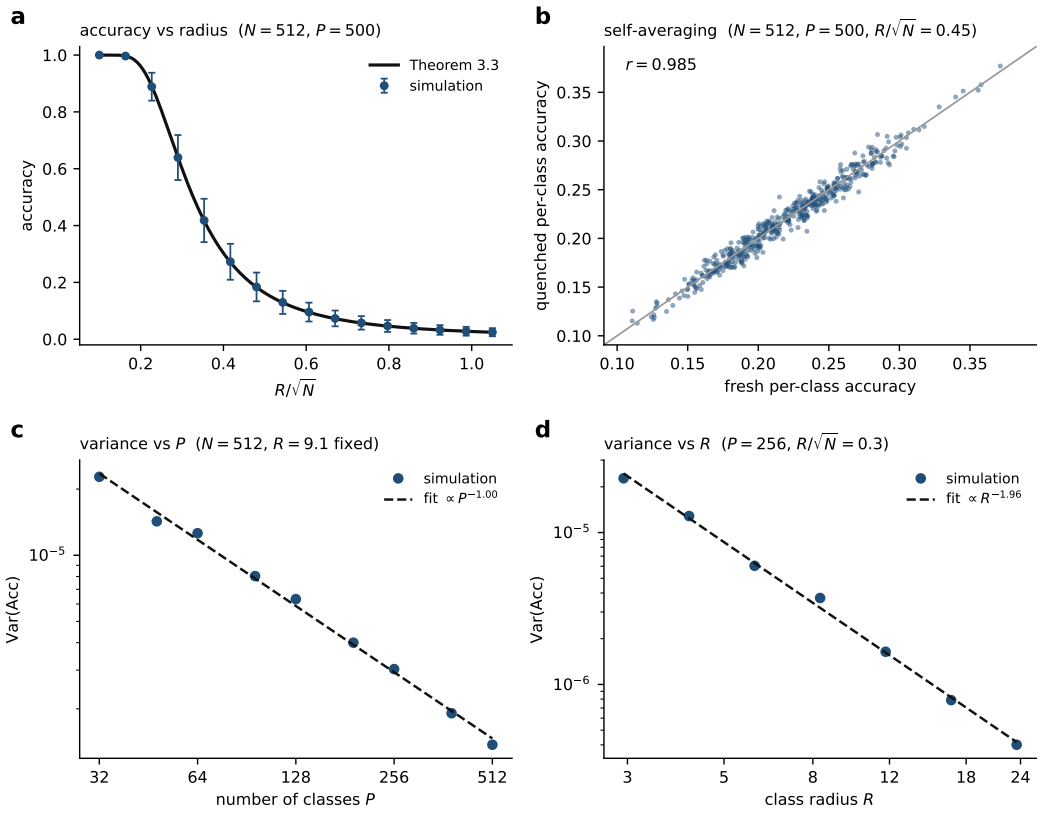

Figure 6: **Validation of the isotropic model's theory on synthetic data.** **(a)** Class-averaged accuracy versus $R/\sqrt{N}$: simulation (markers, mean $\pm$ between-class std) against Theorem 3.3 (line), $N = 512$, $P = 500$. **(b)** Self-averaging: quenched per-class accuracy (one shared centroid bank) versus the fresh-rival accuracy (rivals redrawn per example), the line is $y = x$. $r = 0.985$, confirming self-averaging - the thermal average is equivalent to the quenched average. **(c)** Across-realization variance of the accuracy versus P at fixed radius $R = 9.05$, fitted power law $\propto P^{-1.00}$, predicted by Proposition C.1.5. **(d)** Variance versus radius R (swept via the dimension at fixed $R/\sqrt{N}$), fitted power law $\propto R^{-1.96}$, predicted by Proposition C.1.5.

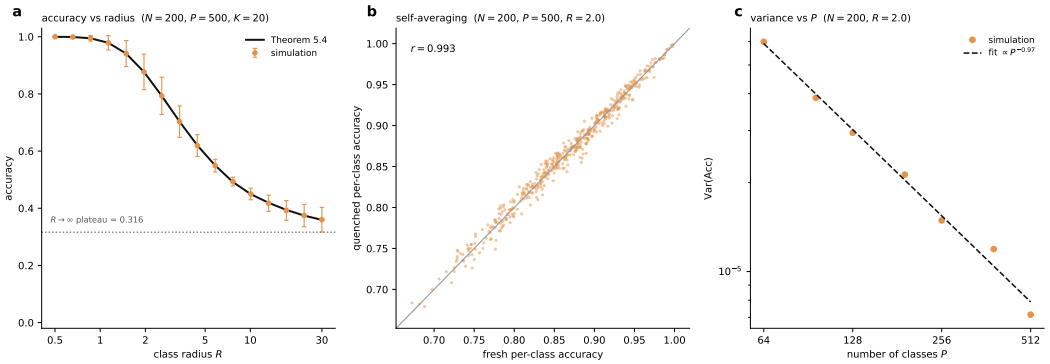

Figure 7: **Synthetic validation of the centroid-aligned model.** Examples are generated according to the centroid-aligned variability model (Sec. 5 and Sec. D.2) with $K = 20$ rivals, $N = 200$, $P = 500$. The geometry is calibrated to DINOv3 ViT-7B: own-class projected std $\sigma_k = 0.24$, rival stds $\sigma_\mu = 0.12\,\mu^{-0.4}$. **(a)** Class-averaged accuracy versus class radius R: simulation (markers) against Theorem 5.5 (line). The dotted line is the $R \to \infty$ plateau 0.316, which is set by σ_k/σ_μ alone. **(b)** Quenched centroid versus fresh per-example sampling of the centroids, $y = x$ line, $r = 0.993$, validating self-averaging. **(c)** Across-realization variance versus P at fixed radius, fitted $\propto P^{-0.97}$, similar to the isotropic model (Proposition C.1.5).

fixed band of the full accuracy, aggregated over all 36 backbones. The margin and variance selectors predict the P-way classification at the smallest K; the centroid-correlation and Gaussian-margin selectors need substantially more rivals at the tight 1% band, confirming that ranking by projected variance (or by the conditional margin) is the most efficient way to choose the true competitors. The conclusion is consistent for all modalities. Since the projection variance and the conditional margin perform comparably (within five rivals at every band), we adopt the projection variance in the main text, as it is a single measured quantity and is therefore cheaper to compute and easier to interpret.

F.3 Optimal K for theory predictions

Increasing K enlarges the rival set the theory integrates over. Table 9 tracks the root-mean-square error between the predicted and empirical per-class accuracies as a function of K. The error drops steeply up to $K \approx 10$ and is essentially flat thereafter: at $K = 20$ the mean error is within 0.003 of its $K \to 300$ floor for 31 of the 36 backbones, and the median number of rivals needed to come within 0.001 of the floor is 20, the value we chose for K in the theory throughout.

Table 7: **Rivals needed to reach the full accuracy.** Median number of rivals K (over 36 backbones) at which the restricted K-way accuracy comes within $5/3/2/1\%$ of the full P-way accuracy, per selector. Smaller is better.

Selector	within 5%	within 3%	within 2%	within 1%
Centroid correlations (g_μ)	14	24	36	87
Projection variance (σ_μ)	14	20	30	54
Theory margin	11	18	25	49
Gaussian margin	10	19	28	73

Table 8: **Rivals needed within 1%, by modality.** Median K to reach within 1% of the full accuracy, per selector and modality (n backbones).

Modality	n	Centroids corr.	Proj. variance	Theory margin	Gaussian margin
Vision	15	76	47	45	56
Autoregressive LM	14	102	89	82	108
Masked LM	4	48	42	35	43
Audio	3	60	39	37	49

Table 9: **The theory's error saturates by $K \approx 10$–20.** Mean per-class accuracy RMSE of the theory at $K = 10$ and $K = 20$, its large-K floor, and the number of backbones already within 0.003 of that floor (large K) at $K = 20$, by modality.

Modality	n	RMSE K=10	RMSE K=20	floor	within 0.003 by K=20
Vision	15	0.049	0.047	0.045	13/15
Audio	3	0.081	0.076	0.064	0/3
Masked LM	4	0.064	0.062	0.061	4/4
Autoregressive LM	14	0.080	0.075	0.075	14/14
All	36	0.065	0.062	0.060	31/36

G HEAVY TAIL STATISTICS AND RENORMALIZATION

The theory of Sect. 5 treats the standardized centroid projections s_μ as jointly Gaussian. Empirically, the variables s_μ (see Table. 1 for the definition) have a pronounced right tail. Since a prediction is decided by the largest rival score $\max_\mu t_\mu$, this positive tail controls the error rate, and it is exactly where a Gaussian underestimates the extreme value. Fitting the pooled positive projections gives a power-law

$$p(s) \;\propto\; \left(a^2 + s^2\right)^{-\nu/2} \tag{27}$$

obtained by a maximum-likelihood fit of a Student-t distribution of parameters a, ν to the positive s values. Here ν is the tail (power-law) exponent, a smaller ν means a heavier tail, and $\nu \to \infty$ recovers a Gaussian. A heavier tail inflates the extreme rival projections that produce errors, so the Gaussian theory is over-optimistic; the single global rescaling $R \to \lambda R$ used in Section 5 (fit per model) compensates for this by enlarging the effective within-class radius, and the fitted λ tracks the tail heaviness $1/\nu$ (see Fig.4 and Table 11).

Figure 8 contrasts the density of s across six backbones spanning three modalities, ordered from the lightest to the heaviest tail, each with its fitted curve (27).

In Sec. 5.2 we claim that without the additional rescaling of the variability by λ, the theory is overoptimistic. This is demonstrated explicitly in Table 10. The theory with $\lambda = 1$ consistently predicts accuracy high by 6% on average across all backbones, compared to only 0.5% with λ.

Table 11 lists the tail exponent ν for all 36 backbones, grouped by modality and sorted from heaviest to lightest tail, alongside the rescaling λ. The exponent varies substantially, from $\nu \approx 3.7$ (DINOv2 ViT-g, the heaviest tail on the roster) to $\nu \approx 22$ (Qwen3-0.6B, nearly Gaussian); the vision backbones span the widest range, while the audio and language backbones have larger ν (lighter tail) and less variability between models. Across the roster the heaviness $1/\nu$ correlates strongly with the fitted rescaling λ (Pearson $r = 0.90$ over the 36 backbones): the heaviest-tail models require the largest correction (λ up to 1.34) and the near-Gaussian ones need almost none. The main text shows this λ-versus-$1/\nu$ relationship as a scatter (Fig. 4).

Table 10: **Without global radius rescaling, the theory is over-optimistic.** Per-class accuracy predicted by the centroid-aligned theory of Sec. 5.2, evaluated at $\lambda = 1$ (no rescaling of the variability) and at the fitted λ, compared with the empirical accuracy. The theoretical accuracy with $\lambda = 1$ is higher than the empirical one, while the theory with λ is within 1% of the empirical one, with significantly better RMSE.

			$\lambda = 1$		fitted λ		
Modality	n	$\langle \mathrm{Acc} \rangle$	$\langle \mathrm{Acc}^{\mathrm{th}} \rangle$	RMSE	$\langle \mathrm{Acc}^{\mathrm{th}} \rangle$	RMSE	$\langle \lambda \rangle$
Vision	15	0.812	0.876	0.080	0.819	0.047	1.228
Audio	3	0.613	0.657	0.087	0.617	0.076	1.084
Masked LM	4	0.715	0.785	0.090	0.724	0.062	1.131
Autoregressive LM	14	0.522	0.583	0.096	0.524	0.075	1.111
All	36	0.672	0.734	0.088	0.677	0.062	1.160

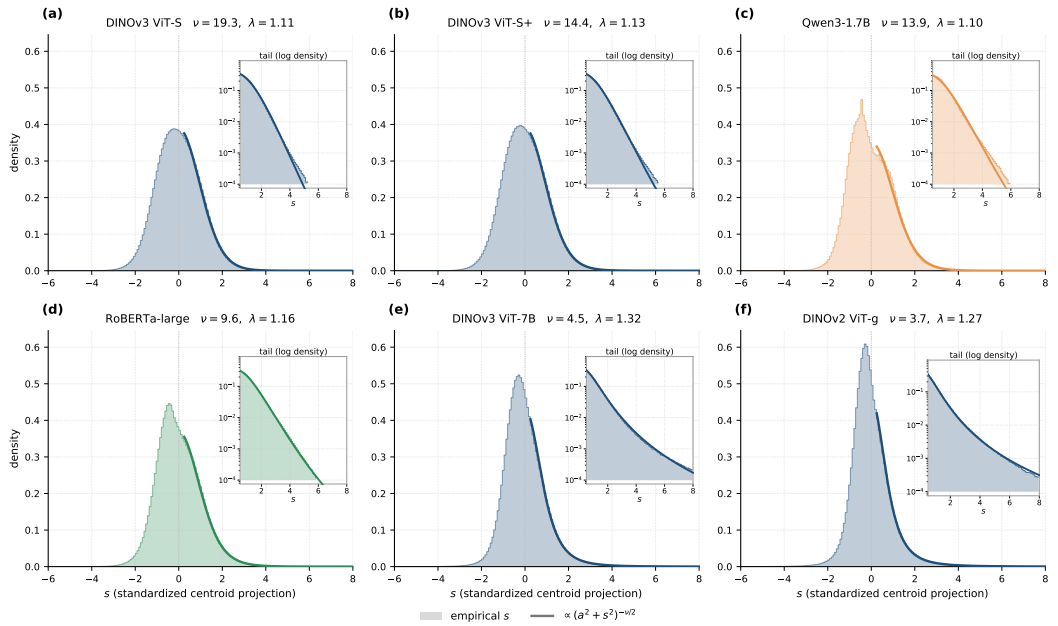

Figure 8: **Standardized centroid projections have a heavy right tail that varies across models.** Density of the standardized projection s (filled) with its fitted half-Student-t right tail $p(s) \propto (a^2 + s^2)^{-\nu/2}$ (solid), $K = 20$. Panels run light $\rightarrow$ heavy tail (left to right, top to bottom), with the fitted exponent ν and model in each title. Color encodes modality (vision blue, autoregressive LM orange, masked LM green). As ν decreases, the peak sharpens, and the right tail fattens.

Table 11: **Right-tail exponent ν across all backbones.** ν is the power-law exponent of the fitted right tail $p(s) \propto (a^2 + s^2)^{-\nu/2}$; a smaller ν means a heavier tail. λ is the global radius rescaling of the theory. Grouped by modality, sorted from heaviest to lightest tail. Pearson correlations between λ and $1/\nu$ are high in every family ($r = 0.89, 0.95, 0.81, 0.89$ in vision, audio, masked language models and autoregressive models, respectively) and globally over all models ($r = 0.90$).

Backbone	ν	λ
Vision (ImageNet-1K)		
DINOv2 ViT-g	3.7	1.272
ConvNeXt-B	4.0	1.254
ViT-B	4.1	1.344
DINOv3 CNX-L	4.4	1.305
DINOv3 ViT-7B	4.5	1.316
DINOv3 CNX-B	4.5	1.296
ResNet50	4.6	1.228
DINOv3 CNX-S	5.2	1.229
DINOv3 ViT-H$^+$	5.3	1.297
DINOv3 CNX-T	6.8	1.148
DINOv3 ViT-L	7.0	1.221
DINOv3 ViT-B	9.0	1.158
DINOv3 ViT-S$^+$	14.4	1.134
CLIP-L	16.4	1.108
DINOv3 ViT-S	19.3	1.114
$r(\lambda, 1/\nu)$	**0.888**	
Audio (VGGSound)		
SSLAM	7.9	1.106
CED	7.9	1.089
AST	9.5	1.057
$r(\lambda, 1/\nu)$	**0.948**	

Backbone	ν	λ
Masked LM (TinyStories, masked-word)		
RoBERTa-large	9.6	1.157
BERT-large	10.1	1.127
ModernBERT-large	11.3	1.124
NomicBERT	14.2	1.115
$r(\lambda, 1/\nu)$	**0.805**	
Autoregressive LM (TinyStories, next-token)		
OLMo-2-13B	10.2	1.152
Qwen2.5-32B	10.9	1.117
Qwen2.5-7B	11.2	1.109
Qwen3-8B	11.3	1.128
Qwen2.5-1.5B	11.3	1.110
Qwen3-14B	11.3	1.124
Qwen3-32B	11.4	1.119
Qwen2.5-14B	11.5	1.112
Llama-3.1-8B	12.4	1.107
Qwen2.5-3B	12.6	1.112
Qwen3-4B	12.7	1.118
Qwen3-1.7B	13.9	1.103
Qwen2.5-0.5B	19.2	1.088
Qwen3-0.6B	21.9	1.059
$r(\lambda, 1/\nu)$	**0.887**	
Overall	**0.901**	

Table 12: **Centroid-aligned theory**: The theory was tested on all 36 models. For each backbone, we report: number of classes P, feature dimension N, empirical prototype accuracy, per-class Pearson r between predicted and empirical accuracy, the fitted global rescaling λ, and the per-class prediction error RMSE. The theory consistently predicts the accuracy well across all models ($r > 0.93$).

Backbone	P	N	Acc	$\mathrm{Acc}^{\mathrm{th}}$	r	λ	RMSE
Vision (ImageNet-1K)							
ConvNeXt-B	1000	1024	0.943	0.952	0.954	1.254	0.0256
ResNet50	1000	2048	0.861	0.871	0.936	1.228	0.0454
DINOv3 ViT-H$^+$	1000	1280	0.856	0.862	0.960	1.297	0.0390
DINOv3 ViT-7B	1000	4096	0.855	0.861	0.951	1.316	0.0427
ViT-B	1000	768	0.853	0.861	0.959	1.344	0.0378
DINOv3 ViT-L	1000	1024	0.845	0.852	0.964	1.221	0.0384
DINOv3 CNX-L	1000	1536	0.814	0.819	0.961	1.305	0.0440
DINOv3 CNX-B	1000	1024	0.804	0.808	0.958	1.296	0.0455
DINOv2 ViT-g	1000	1536	0.795	0.804	0.941	1.272	0.0631
DINOv3 ViT-B	1000	768	0.794	0.800	0.954	1.158	0.0460
CLIP-L	1000	768	0.792	0.804	0.936	1.108	0.0596
DINOv3 CNX-S	1000	768	0.786	0.793	0.956	1.229	0.0481
DINOv3 CNX-T	1000	768	0.740	0.749	0.937	1.148	0.0602
DINOv3 ViT-S$^+$	1000	384	0.738	0.742	0.951	1.134	0.0504
DINOv3 ViT-S	1000	384	0.706	0.710	0.944	1.114	0.0555
mean			**0.812**	**0.819**	**0.951**	**1.228**	**0.0468**
Audio (VGGSound)							
SSLAM	310	768	0.628	0.631	0.952	1.106	0.0720
CED	310	768	0.628	0.632	0.947	1.089	0.0739
AST	310	768	0.583	0.587	0.944	1.057	0.0819
mean			**0.613**	**0.617**	**0.948**	**1.084**	**0.0759**
Masked LM (TinyStories, masked-word)							
RoBERTa-large	320	1024	0.767	0.776	0.944	1.157	0.0542
ModernBERT-large	320	1024	0.734	0.743	0.942	1.124	0.0614
BERT-large	320	1024	0.710	0.720	0.942	1.127	0.0630
NomicBERT	320	768	0.649	0.658	0.945	1.115	0.0683
mean			**0.715**	**0.724**	**0.943**	**1.131**	**0.0617**
Autoregressive LM (TinyStories, next-token)							
Qwen2.5-32B	320	5120	0.567	0.572	0.945	1.117	0.0722
Qwen3-32B	320	5120	0.566	0.569	0.955	1.119	0.0665
OLMo-2-13B	320	5120	0.563	0.567	0.948	1.152	0.0701
Qwen3-14B	320	5120	0.556	0.558	0.954	1.124	0.0694
Qwen2.5-14B	320	5120	0.552	0.556	0.942	1.112	0.0758
Llama-3.1-8B	320	4096	0.550	0.553	0.943	1.107	0.0772
Qwen2.5-7B	320	3584	0.544	0.550	0.940	1.109	0.0756
Qwen3-8B	320	4096	0.540	0.542	0.955	1.128	0.0693
Qwen2.5-3B	320	2048	0.528	0.531	0.946	1.112	0.0785
Qwen3-4B	320	2560	0.519	0.520	0.953	1.118	0.0697
Qwen2.5-1.5B	320	1536	0.491	0.494	0.931	1.110	0.0822
Qwen3-1.7B	320	2048	0.474	0.474	0.949	1.103	0.0747
Qwen2.5-0.5B	320	896	0.452	0.450	0.944	1.088	0.0853
Qwen3-0.6B	320	1024	0.406	0.403	0.939	1.059	0.0839
mean			**0.522**	**0.524**	**0.946**	**1.111**	**0.0750**
Overall			**0.672**	**0.677**	**0.948**	**1.160**	**0.0618**

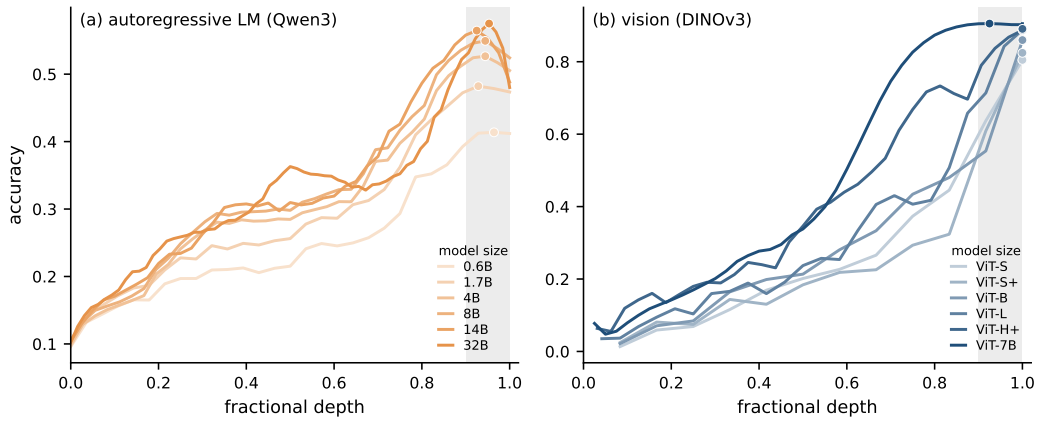

Figure 9: **The best prototype layer is a few blocks before the final one for language, but is the final block for vision.** Per-layer prototype classification accuracy versus fractional depth. Color runs from light (small) to dark (large) within each modality, and the peak performance is marked. **(a)** The six Qwen3 autoregressive language models; the optimal accuracy is a few blocks before the final layer, with a big drop at the final layer in most models. **(b)** the six DINOv3 vision transformers. The final layer is the best, with a plateau on DINOv3 7B.

H Depth analysis

The prototype classifier of Sec. 5 reads out a single hidden layer. Here we analyze how prototype classification accuracy and geometry change across depth, and identify the best layer for this task. For the autoregressive language models the best layer for content-word prediction is consistently a few blocks before the final one, and accuracy dips at the last block; for vision the final block is best (in DINOv3 7B, there is a plateau). We subsequently analyze the geometry of the largest language model, Qwen3-32B across depth, similar to the analysis in Sec. 6 for the Qwen3 family across scale. Similar to the scale effect, we see that the best layer has large true centroid correlation compared to rivals σ_k/σ_μ^*, decorrelates centroids (small g_μ^*), and has small correlations between projections ρ_μ^*. We note that, as observed before with scale, the class radius R does not predict the overall best accuracy.

Table 13: **Qwen3-32B across depth.** Class-averaged geometric measures for the seven last blocks of Qwen3-32B (64 blocks). * marks each class's hardest competitor $\mu^\star$, measured by σ_μ, the projection standard deviation. $\langle R\rangle$ is the mean class radius; σ_k/σ_μ^* is the ratio between the standard deviation of the projection onto the true-class and hardest-rival directions. σ_μ^* is the hardest rival projected std, $g_\mu^* = \hat{\mathbf{c}}_k \cdot \hat{\mathbf{c}}_{\mu^\star}$ the centroid overlap and ρ_μ^* the correlation between projections. ν is the right-tail exponent. Best measure highlighted in bold.

Layer	$\langle R\rangle$	σ_k/σ_μ^*	σ_μ^*	g_μ^*	ρ_μ^*	ν	Acc
L58 (0.91)	2.40	0.993	**0.172**	0.514	0.496	**10.9**	0.521
L59 (0.92)	2.32	0.996	0.175	0.511	0.502	11.0	0.542
L60 (0.94)	2.24	0.999	0.175	0.510	0.474	**10.9**	0.557
L61 (0.95)	2.20	**1.007**	0.175	**0.503**	0.451	11.4	**0.566**
L62 (0.97)	2.17	0.957	0.181	0.517	0.463	13.7	0.553
L63 (0.98)	2.16	0.894	0.191	0.521	**0.448**	18.1	0.528
L64 (final)	**2.10**	0.754	0.224	0.533	0.471	98.5	0.470

I Theoretical Predictions Across Models and Modalities

The main text validates the centroid-aligned variability theory (Sec. 5) on a handful of representative backbones. Here we apply it, unchanged, to a comprehensive list of 36 state-of-the-art models, spanning four modalities: 15 vision encoders (ImageNet), 14 autoregressive language models and 4 masked language models (TinyStories), and 3 audio encoders (VGGSound). The models have

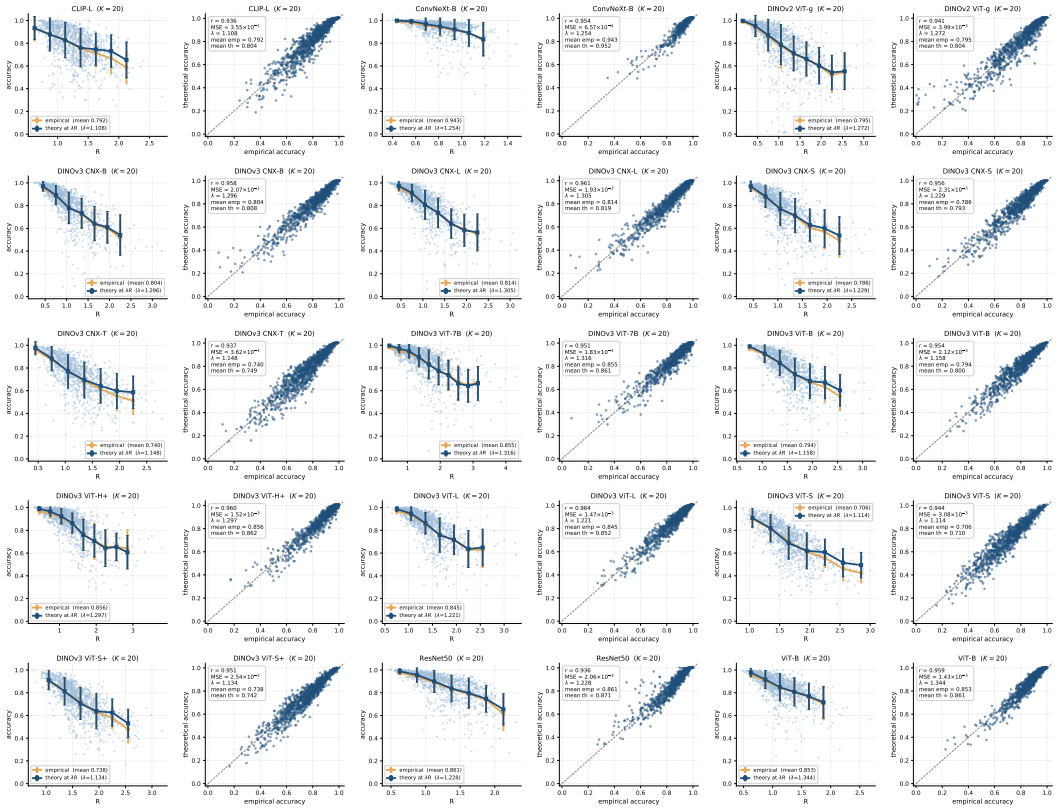

Figure 10: 15 vision models. For each backbone: (left) per-class accuracy versus empirical radius R, empirical binned mean and theory binned mean nearly overlap in most models; (right) per-class predicted versus empirical accuracy with r, RMSE and λ. The dashed line is $y = x$. Mean Pearson $r = 0.951$, mean RMSE $= 0.047$.

different architectures (transformers and convolutional neural networks), training methods (supervised, self-supervised, and contrastive), and training datasets. For every model, we compute the per-class prediction with a single fitted global rescaling λ and the measured quantities the theory predicts, and compare it to the empirical prototype accuracy class by class. No per-model tuning beyond the scalar λ is used.

Table 12 reports, for each backbone, the number of classes P, the feature dimension N, the empirical prototype accuracy, the per-class Pearson correlation r between predicted and empirical accuracy, the fitted λ, and RMSE.

Across all four modalities the per-class correlation between the theory and the empirical accuracy is high: $r = 0.951$ (vision), 0.946 (autoregressive), 0.943 (masked) and 0.948 (audio). Per-class errors have nearly zero bias (mean per modality bias $0.002 - 0.009$) and only a few percent mean deviation (mean per modality RMSE 0.047–0.076). Thus our theory predicts prototype accuracy class by class across vision, language and audio representations.

For each model we also reproduce the two per-class panels of the theory vs the empirical accuracy as in Fig. 4. The four figures show all backbones grouped by modality. The theory not only tracks the general R trend, but often closely follows the individual pattern of the model.

J VARIANTS OF THE THEORY

The prototype-accuracy theory of the main text (Theorem 5.5) is one point in a small family of predictors, obtained by making different assumptions about the covariance of the standardized centroid projections s_μ. Here we compare the main theory against a simpler baseline and one richer refinement. All three predictors are evaluated at a fixed rival $K = 20$, on the same 36 backbones

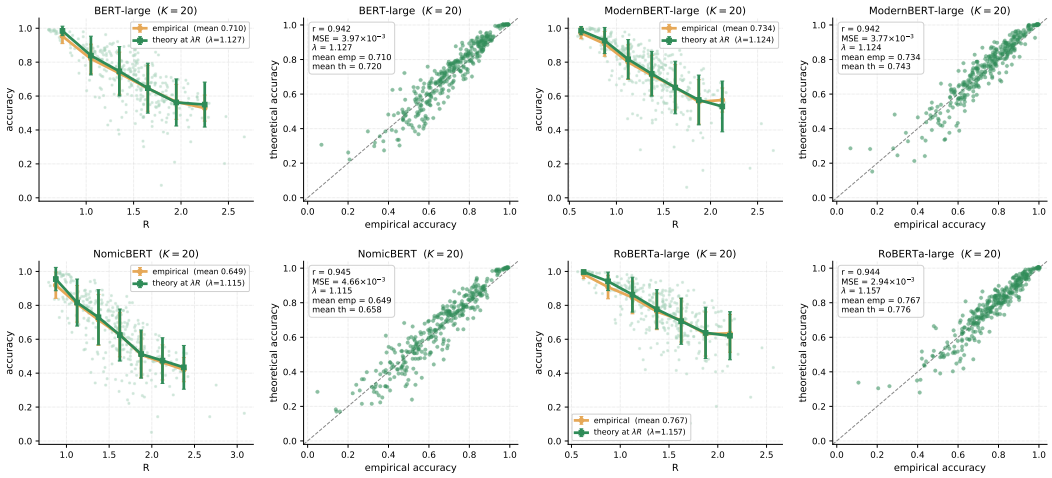

Figure 11: Masked language models (4 models, TinyStories content-word prediction, best layer). Panels as in Fig. 10. Mean $r = 0.943$, mean RMSE $= 0.062$.

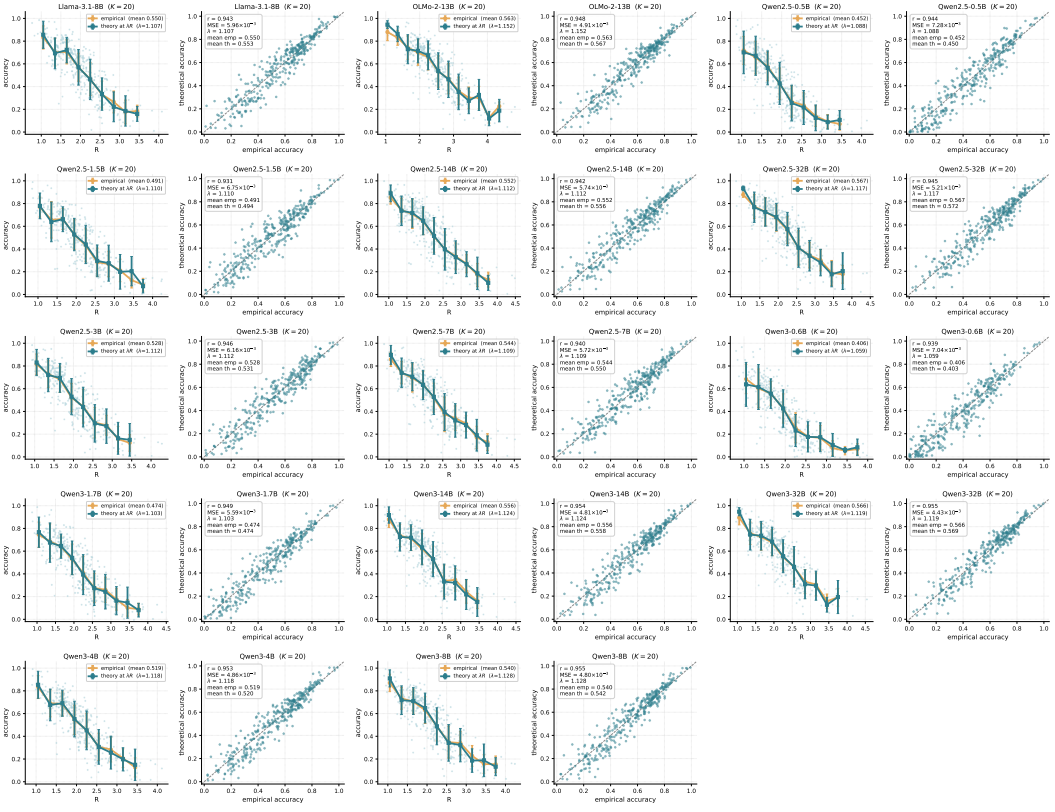

Figure 12: Autoregressive language models (14 models, TinyStories content-word prediction, best layer). Panels as in Fig. 10; equally spaced R bins fixed for this modality. Mean $r = 0.946$, mean RMSE $= 0.075$.

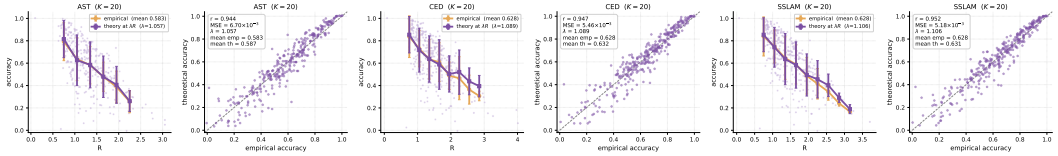

Figure 13: Audio (3 encoders, VGGSound training split, final layer). Panels as in Fig. 10. Mean $r = 0.948$, mean RMSE $= 0.076$.

and the same per-class accuracy targets. The global rescaling factor λ was fitted separately for each variant.

Uncorrelated s: The model is the same as the one presented in Sec. 5, with $\rho_{\mu \neq k} = 0$, so that different s_μ are uncorrelated. The uncorrelated theory fits the data significantly worse than the theory of Sec. 5, where we consider a simple off-diagonal structure, and thus it is not used in the main text.

Cluster expansion: We start from Eq. 21. We assume that $\Sigma_{\mu\nu}$ has the structure identified in Sec. 5, plus a small correction on the terms outside the diagonal, $\Sigma = \Sigma_0 + \delta\Sigma, \;\; \delta\Sigma_{\mu\mu} = 0$. We shift the logits to have zero mean for convenience, and write the orthant probability conditioned on t_k, which has a diagonal element from the original structure plus the small correction

$$\mathrm{Acc}^{\mathrm{th}}(R) \propto \prod_{\mu \in \mathcal{R}_K} \int_{-\infty}^{t_k - f_\mu(t_k) - g_\mu} \frac{dt_\mu}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_{\mu,\nu \in \mathcal{R}_K} t_\mu \left(\delta_{\mu\nu} \Lambda_\mu^2 + \delta\Sigma_{\mu\nu} \right)^{-1} t_\nu \right) \tag{28}$$

,

where $f_\mu(t_k) = \rho_\mu \sigma_\mu (t_k - 1)/\sigma_k$ is the conditional mean similar to Sec. D.2, and Λ_μ is the conditional standard deviation $\Lambda_\mu = R\sigma_\mu \sqrt{1 - \rho_\mu^2}$. Expanding to linear order in $\delta\Sigma_{\mu\nu}$:

$$\mathrm{Acc}^{\mathrm{th}}(R) = \prod_{\mu \in \mathcal{R}_K} \int_{-\infty}^{t_k - f_\mu(t_k) - g_\mu} \frac{dt_\mu}{\sqrt{2\pi\Lambda_\mu^2}} \exp\left(-\frac{1}{2} \sum_{\mu \in \mathcal{R}_K} \frac{t_\mu^2}{\Lambda_\mu^2} \right) \left(1 + \frac{1}{2} \sum_{\mu \neq \nu} \frac{\delta\Sigma_{\mu\nu} t_\mu t_\nu}{\Lambda_\mu^2 \Lambda_\nu^2} \right) \tag{29}$$

Define the conditional margin

$$F_\mu(t_k) = \frac{t_k - f_\mu(t_k) - g_\mu}{\Lambda_\mu}, \tag{30}$$

and perform the truncated Gaussian integrals

$$\mathrm{Acc}^{\mathrm{th}}(R) = \prod_{\mu \in \mathcal{R}_K} \Phi\left(F_\mu(t_k)\right) \left(1 + \frac{1}{2} \sum_{\mu \neq \nu} \frac{\delta\Sigma_{\mu\nu}}{\Lambda_\mu \Lambda_\nu} m\left(F_\mu(t_k)\right) m\left(F_\nu(t_k)\right) \right) \tag{31}$$

We used the common inverse Mills ratio $m(x) = \phi(x)/\Phi(x)$, where $\phi(x)$ is a standard normal PDF and $\Phi(x)$ is a standard normal CDF. Substitute $t_k = Rz\sigma_k + 1$ and averaging over z we get the final result:

$$\mathrm{Acc}^{\mathrm{th}}(R) = \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\prod_{\mu \in \mathcal{R}_K} \Phi\left(F_\mu(z)\right) \left(1 + \frac{1}{2} \sum_{\mu \neq \nu} \frac{\delta\Sigma_{\mu\nu}}{\Lambda_\mu \Lambda_\nu} m\left(F_\mu(z)\right) m\left(F_\nu(z)\right) \right) \right] \tag{32}$$

With

$$F_\mu(z) = \frac{\sigma_k - \rho_\mu \sigma_\mu}{\Lambda_\mu / R} z + \frac{1 - g_\mu}{\Lambda_\mu} \tag{33}$$

Although the cluster expansion theory remains analytic, and fits the data slightly better than the original theory derived in Sec. 5, it adds another $K(K-1)/2$ parameters $\delta\Sigma_{\mu\nu}$ that need to be measured, and it is less transparent than the original simpler theory. Thus, we bring it here as an additional theoretical result with its performance on the 36 backbones we tested (Table. 14), but do not use it in the main text.

Table 14: Fitted global rescaling λ and per-class prediction error RMSE for three predictors at $K=20$, on all 36 backbones: the uncorrelated-projections baseline ($\rho_\mu=0$), the main theory (rank-one true–rival correlation), and the first-order cluster expansion. λ is refitted separately for each predictor.

	Uncorrelated		**Main theory**		**Cluster expansion**	
Backbone	λ	RMSE	λ	RMSE	λ	RMSE
Vision (ImageNet-1K)						
ConvNeXt-B	0.979	0.0493	1.254	0.0256	1.281	0.0234
ResNet50	1.114	0.0589	1.228	0.0454	1.320	0.0434
DINOv3 ViT-H$^+$	1.212	0.0578	1.297	0.0390	1.317	0.0377
DINOv3 ViT-7B	1.248	0.0574	1.316	0.0427	1.337	0.0402
ViT-B	1.196	0.0593	1.344	0.0378	1.393	0.0339
DINOv3 ViT-L	0.996	0.0733	1.221	0.0384	1.248	0.0366
DINOv3 CNX-L	1.160	0.0706	1.305	0.0440	1.360	0.0417
DINOv3 CNX-B	1.087	0.0799	1.296	0.0455	1.337	0.0420
DINOv2 ViT-g	1.114	0.1166	1.272	0.0631	1.311	0.0561
DINOv3 ViT-B	0.864	0.0929	1.158	0.0460	1.196	0.0416
CLIP-L	0.671	0.1188	1.108	0.0596	1.199	0.0471
DINOv3 CNX-S	0.966	0.0935	1.229	0.0481	1.275	0.0450
DINOv3 CNX-T	0.855	0.1224	1.148	0.0602	1.209	0.0535
DINOv3 ViT-S$^+$	0.861	0.0981	1.134	0.0504	1.179	0.0438
DINOv3 ViT-S	0.838	0.1061	1.114	0.0555	1.173	0.0464
mean	**1.011**	**0.0837**	**1.228**	**0.0468**	**1.276**	**0.0422**
Audio (VGGSound)						
SSLAM	0.920	0.1194	1.106	0.0720	1.179	0.0555
CED	0.894	0.1164	1.089	0.0739	1.183	0.0599
AST	0.861	0.1256	1.057	0.0819	1.173	0.0617
mean	**0.892**	**0.1205**	**1.084**	**0.0759**	**1.178**	**0.0590**
Masked LM (TinyStories, masked-word)						
RoBERTa-large	0.943	0.0856	1.157	0.0542	1.235	0.0477
ModernBERT-large	0.884	0.0985	1.124	0.0614	1.225	0.0511
BERT-large	0.943	0.0930	1.127	0.0630	1.235	0.0492
NomicBERT	0.920	0.1048	1.115	0.0683	1.238	0.0565
mean	**0.923**	**0.0955**	**1.131**	**0.0617**	**1.234**	**0.0511**
Autoregressive LM (TinyStories, next-token)						
Qwen2.5-32B	0.937	0.1069	1.117	0.0722	1.363	0.0755
Qwen3-32B	0.940	0.1004	1.119	0.0665	1.376	0.0707
OLMo-2-13B	0.986	0.0922	1.152	0.0701	1.393	0.0726
Qwen3-14B	0.953	0.0984	1.124	0.0694	1.383	0.0716
Qwen2.5-14B	0.933	0.1093	1.112	0.0758	1.370	0.0773
Llama-3.1-8B	0.923	0.1061	1.107	0.0772	1.373	0.0750
Qwen2.5-7B	0.946	0.1114	1.109	0.0756	1.350	0.0780
Qwen3-8B	0.963	0.0947	1.128	0.0693	1.363	0.0703
Qwen2.5-3B	0.956	0.1058	1.112	0.0785	1.357	0.0780
Qwen3-4B	0.950	0.0993	1.118	0.0697	1.347	0.0685
Qwen2.5-1.5B	0.953	0.1166	1.110	0.0822	1.330	0.0865
Qwen3-1.7B	0.940	0.1003	1.103	0.0747	1.337	0.0718
Qwen2.5-0.5B	0.937	0.1072	1.088	0.0853	1.350	0.0808
Qwen3-0.6B	0.900	0.1096	1.059	0.0839	1.311	0.0772
mean	**0.944**	**0.1042**	**1.111**	**0.0750**	**1.357**	**0.0753**
Overall	**0.965**	**0.0960**	**1.160**	**0.0618**	**1.295**	**0.0574**