# Manifold geometry underlies a unified code for category and category-independent features

Lorenzo Tiberi[1] and Haim Sompolinsky[1,2]

[1] *Center for Brain Science, Harvard University, Cambridge, MA 02138*

[2] *Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 9190401, Israel*

In everyday vision, animals routinely extract from the same visual stimulus both object identity and continuous identity-independent variables such as position and size. It has been shown that linear decoding performance of both kinds of information increases along the ventral stream, suggesting that inferior temporal cortex may be implementing a joint code for object category and category-independent features. A central open question is whether such a code can indeed exist within a single representation and, if so, what geometric properties enable it. Here, we show that convolutional neural networks can develop such a code. We then derive a theory of regression on category manifolds, identifying the key manifold-geometry measures that enable accurate readout of category-independent features, and showing how they can be optimized while preserving manifold properties known to support classification performance. We further characterize how common experimental constraints, such as subsampling neural units and using a limited number of categories, affect the empirical estimation of regression performance. Our findings thus provide a principled understanding of the geometry underlying joint codes and yield testable predictions for future neural recordings probing the joint-code hypothesis in the ventral stream.

## I. INTRODUCTION

Natural intelligence often relies on extracting multiple kinds of information from the same sensory stimulus. In everyday vision, for example, animals routinely extract both object identity and continuous, identity-independent attributes such as position and size from the same visual stimulus. A central question in neuroscience is how the brain implements this ability.

Neural recordings in higher-order areas, such as macaque prefrontal cortex, anterior cingulate cortex, and hippocampus [1, 2], show that multiple task-relevant variables can be decoded from the same population activity using simple (often linear) readouts. This has led to the hypothesis that a single high-level population code can jointly encode multiple stimulus variables in a linearly accessible form, raising the question of which representational principles make such multi-variable encoding possible [3]. Existing theoretical accounts have largely addressed highly controlled regimes with very small stimulus sets (on the order of 4 cues) [1, 2], leaving open the question of how multi-variable decodability is achieved in regimes with many stimulus categories and multiple category-invariant features.

A natural setting to study this regime is vision, where single-object images within a category can vary widely under a range of category-preserving transformations. A seminal study [4] showed that along the macaque ventral stream (V1→V4→inferior temporal cortex, IT)—and similarly across the layers of an artificial convolutional neural network (CNN)—linear decodability of object category *and* linear regressibility of category-independent features (e.g., object position, size, and pose within the image) both increase. Yet, despite these improvements along the visual hierarchy, our analysis shows that even in IT cortex the absolute regression performance as estimated from the macaque recordings remains limited, raising the question of whether this reflects experimental limitations—such as population subsampling, limited image sets, and neural noise—or instead a genuine constraint on how ventral-stream populations can jointly encode object category and category-independent features.

To address this question, CNNs provide a useful testbed because they align well with ventral-stream neural responses [5] and can be systematically investigated without the experimental limitations of neural recordings. Recent work [6] suggested that CNNs can indeed be jointly optimized for object classification and regression of category-independent features, indicating that a joint population code may at least be possible in artificial systems. Here we address the following fundamental questions: What properties of a population representation enable effective joint encoding of category and category-independent information? What learning objectives give rise to such representations?

Identifying these distinctive properties requires a theoretical framework that relates linear classification and linear regression performance to properties of neural representations. For object classification, such a framework already exists in terms of object manifolds, defined as the set of population responses elicited by images from the same category. Theoretical work relates aspects of manifold geometry, such as radius and dimensionality, to proxy measures of linear classification performance, including storage capacity [7] and few-shot generalization [8]. These approaches have been used to interpret trends in classification performance along the hierarchy of both the macaque ventral stream and artificial CNNs [8, 9]. In contrast, an analogous theory linking manifold geometry to the regression performance of category-independent

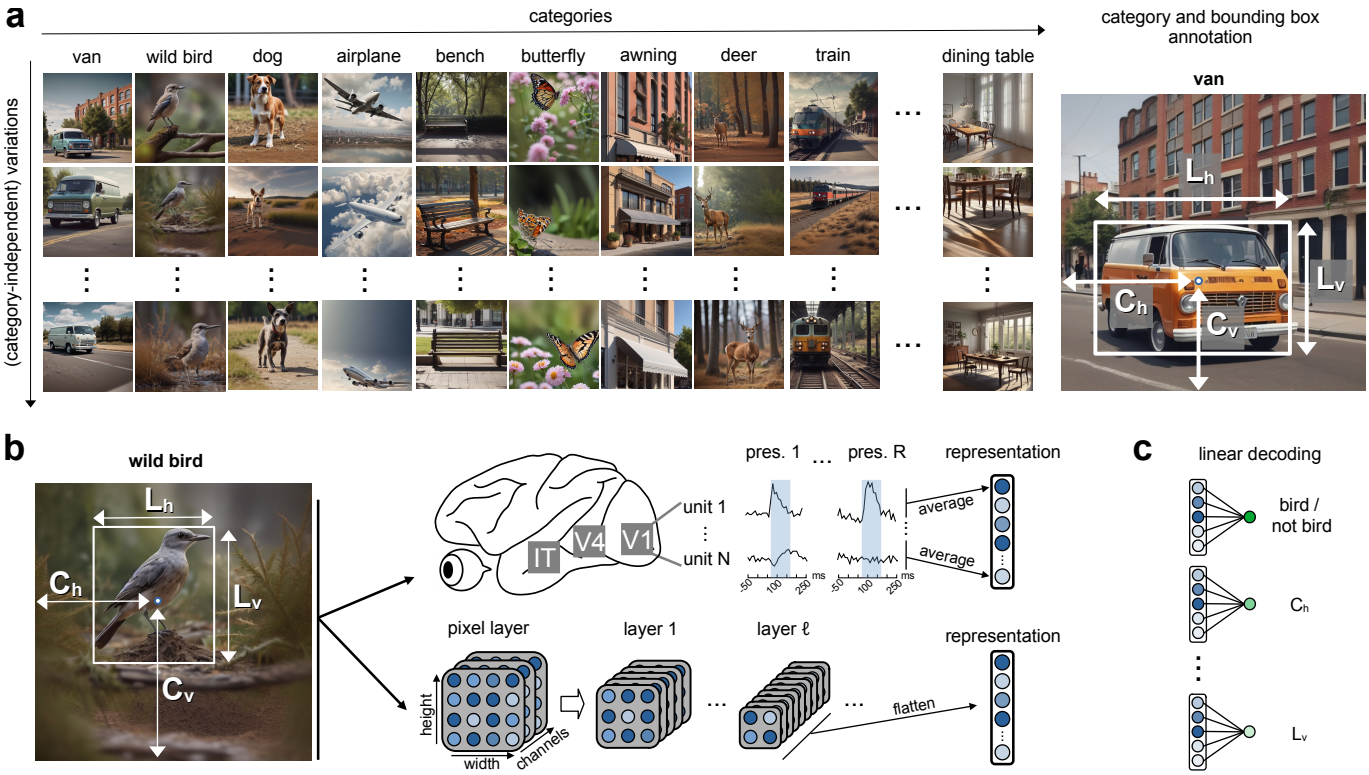


Figure 1. Framework to quantify how well neural representations encode category and category-independent information. (**a**) Overview of the main image dataset used in this work. The dataset contains single-object images annotated with object category and four bounding box coordinates $C_h$, $C_v$, $L_h$, $L_v$. (**b**) From image to neural representation. An image is presented to the ventral visual stream (top), modeled in this work by a convolutional neural network (CNN; bottom), and the resulting neural representations are analyzed. Top: Representations from neural recordings are obtained by averaging spike counts over a post-timulus time window (shaded region) and across stimulus repetitions. Bottom: Representations from a given CNN layer are obtained by flattening that layer's activations into a one-dimensional vector. (**c**) We evaluate the ability of linear decoders to extract category and bounding-box coordinates from these representations.

features is currently lacking. Crucially, because our goal is a theory of joint coding, it is not sufficient to develop a regression theory in a setting disconnected from categorization. Instead, the theory should be embedded within the category manifold framework, so that it can address how geometric properties relevant for regression of category-independent features relate to—and potentially trade off with—those supporting object classification. Here we develop such a theory and test it in CNNs trained under different objectives, by analyzing the resulting representations and task performance.

In this work, we first show that a ventral-stream–like CNN can be trained so that the feature layer representations support high-accuracy linear readout of both object category and category-independent features, such as object position and size. We then develop a manifold-geometry theory for linear regression of category-independent features that decomposes the global regression error into a "local" error, which quantifies how well a feature is linearly encoded within individual category manifolds, and a "local–global" error gap, which captures the additional error incurred by requiring a single category-independent linear readout shared across categories. We then identify the across-category differences in manifold geometry that contribute to this gap. We further investigate how this regression-relevant manifold geometry can coexist with the manifold geometry that supports linear object classification. Finally, we use this framework to identify signatures that distinguish a true joint code from a category-only code and to quantify how common experimental constraints (limited categories and neural-unit subsampling) can obscure those signatures.

## II. RESULTS

### 1. *Linear decoding framework*

To assess how well neural representations encode both an image's object category and features that vary independently of category ("category-independent" features), we adopt the linear-decoding framework of [4] (**Fig. 1**).

We consider image datasets consisting of single-object images labeled by category and by four continuous bounding-box parameters describing object position and size: center coordinates $C_h$, $C_v$ and side lengths $L_h$, $L_v$ along the image's horizontal (h) and vertical (v) axes (**Fig. 1a**).

For each image, an $N$-dimensional neural representation vector is extracted from the activations of neurons in the brain visual hierarchy or from the activations of a convolutional neural network (CNN) that mimics this hierarchy (**Fig. 1b**).

Finally, we evaluate the performance of linear decoders in extracting category and category-independent information from these representations, treating it as a proxy for the decoding performance achieved by hypothetical downstream neurons (**Fig. 1c**). For category, we train one-vs-rest classifiers and evaluate their balanced classification accuracy, averaged across category. For the bounding box parameters, we train one linear regressor per parameter and evaluate it by normalized mean squared error (nMSE), i.e., MSE divided by the parameter variance. All decoding metrics are cross-validated over multiple train–test splits (see the Online Methods for further details).

### 2. *One code supports both classification and regression*

We have applied this decoding framework to the publicly available dataset of macaque recordings [10] studied in [4], reproducing their qualitative results (see SI): decoding performance of both category and category-independent features improves along the visual hierarchy. However, we also find that absolute regression performance remains limited: root nMSE remains above 0.5, meaning that the root-mean-square prediction error is still larger than half the across-image standard deviation of the decoded feature.

To investigate what it takes to accurately perform both categorization and regression tasks we apply the above decoding framework to CNNs trained on a newly constructed large-scale dataset of single-object images with category and bounding-box annotations (**Fig. 1a**, left). Images span many object categories (265) as well as many images per category (20,000 per category, split into 10,000 images for training the CNN and 10,000 for generating neural representations and evaluating linear decoders). The large number of images and categories serves several purposes: it enables effective training, a more precise statistical characterization of neural representations, and it places us in a regime closer to the limit of infinite number of categories, which is valuable for testing theoretical predictions and serves as a clean baseline from which we can later introduce controlled category subsampling. Images were generated using a multi-stage pipeline that combines text-to-image generation (Stable Diffusion XL, SDXL [11, 12]), object detection (CerberusDet [13]), and image outpainting (Stable Diffusion v1.5, SD1.5 [14–16]) to produce single-object images with controlled object placement and a uniform distribution of bounding-box coordinates. The overall generation procedure is summarized in **Fig. 2a** and described in detail in the Online Methods.

On this dataset, we fine-tuned an ImageNet-1k–pretrained ResNet-50 (**Fig. 2b**), a state-of-the-art CNN with good alignment to ventral stream responses [5]. We treated the final feature layer as the shared representation and attached two types of readout: (i) the standard linear classification head for predicting object category, and (ii) four additional linear regression heads, each predicting one bounding-box parameter ($C_h, C_v, L_h, L_v$). The network was optimized end-to-end using stochastic gradient descent on a combined loss: cross-entropy loss on the classification head and Smooth L1 loss on each regression head (see Online Methods). In addition to this model, which we call network CR, we also trained the same architecture exclusively on the classification task (network C) and exclusively on the regression task (network R). The C and R networks provide upper bounds on the performance achievable on this dataset by this architecture, when it is allowed to specialize in a single task.

**Fig. 2c** shows the performance of linear decoders trained on the feature-layer representations of these three networks. Network CR matches the performance of network C on classification and that of network R on regression. In particular, the regression error of the CR network is substantially lower than that of the classification-only network C. We thus conclude that the CR network implements a single population code that jointly supports optimal linear readout of object category and category-independent features.

### 3. *Theory*

What properties of the population code of network CR, or of any population code in general, allow this optimal performance in both classification and regression of category-independent features? For classification, theories already exist that relate neural representation geometry to proxy measures of linear classification performance, such as storage capacity [7], or few-shot learning accuracy [8]. Such theories are formulated within the framework of object manifolds, defined as the collection of all neural representations of images belonging to the same category (**Fig. 3a**). It has been shown that along the hierarchy of both classification-trained CNNs [9] and the ventral visual stream [8], such object manifolds change their geometric properties, such as their dimensionality, radius, and centroid separation, so as to become more and more disentangled, thereby improving linear classification performance (**Fig. 3a**). In a system optimized also for regression, one could imagine a similar scenario whereby, along the visual hierarchy, the encoding of category-independent features becomes both

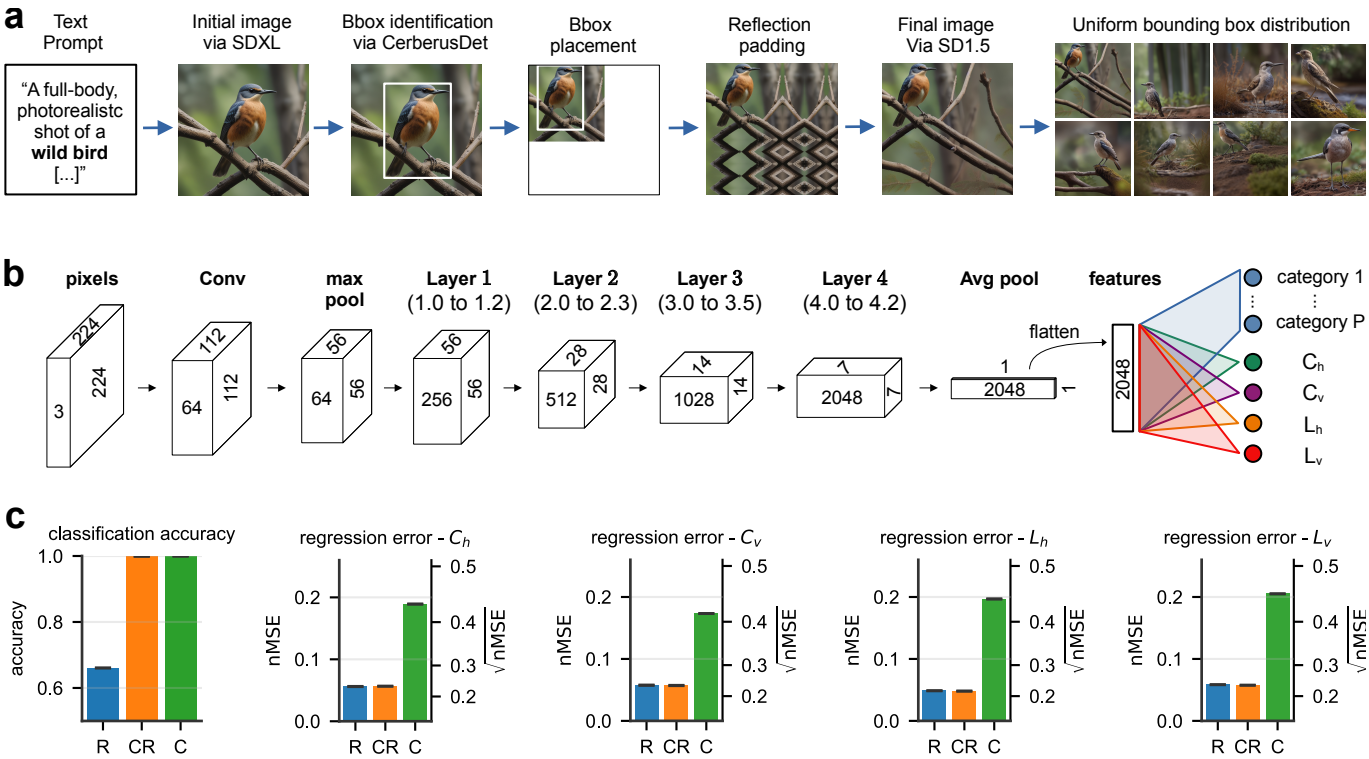


Figure 2. One code supports both object classification and regression of category-independent features. (**a**) Generation process for an image in our dataset. From left to right: an initial image of an object in the desired category (e.g. "wild bird") is generated from a text prompt using Stable Diffusion XL, with categories drawn from the taxonomy of Objects365 [17], an image dataset for object detection. The object's bounding box is then estimated by CerberusDet, an Objects365-trained detector. The image is then isotropically rescaled and pasted onto a white canvas so that the detected box matches a target box whose parameters $(C_h, C_v, L_h, L_v)$ have been drawn uniformly (conditional on preserving aspect ratio and remaining inside the canvas). The remaining white area is filled by reflection-padding the border region around the pasted image, and the resulting canvas is used as the conditioning image for a second round of generation with a Stable Diffusion 1.5 inpainting model, yielding the final sample. The method allows generation of images with a controlled, uniform distribution of bounding boxes. For example images from other categories, see Fig. 1a. (**b**) Schematic representation of our model's architecture, consisting of a ResNet-50 backbone with additional regression heads. Left: The main layers of the ResNet-50 backbone. Size in channel-width-height format and number of sublayers ($\ell.0$, $\ell.1$, ..., with $\ell$ the layer number) are reported. Right: In addition to the standard classification head, we also attach four regression heads to the final feature layer—one per bounding-box coordinate. (**c**) Results of linear decoding from the feature layer of networks R, CR and C. Error bars indicate the SEM across cross-validation splits.

more linear within each manifold and more consistently aligned across manifolds, such that it can be regressed accurately by a single category-independent regressor (**Fig. 3a**). Unlike for classification, however, a theory identifying and linking the relevant manifold geometry to linear regression performance is currently lacking. We now develop such a theory.

We begin by distinguishing two additive contributions, of very different nature, that determine linear regression performance for category-independent features (**Fig. 3b-c**). Let $E$ denote the test nMSE obtained when linearly regressing a given category-independent feature with a single category-independent readout. We decompose it as

$$E = E_{\text{loc}} + \Delta E \tag{1}$$

where $E_{\text{loc}} = \text{Avg}_{\mu}\left[E^{\mu}\right]$ and $E^{\mu}$ is the test nMSE achieved by a regressor trained and tested exclusively on the $\mu$-th category. The "local error" $E_{\text{loc}}$ quantifies the quality of the feature's linear encoding at the *local* level of a single category, i.e. within a single object manifold (**Fig. 3b**). It provides a lower bound on the global error $E$, corresponding to the regression error of an oracle that is allowed to use separate regressors for each category. The residual $\Delta E$, termed the "local-global error gap", captures the additional error incurred by requiring a single category-independent readout, and reflects the global geometric arrangement of object manifolds across categories in the $N$-dimensional space of neural units (**Fig. 3c**).

**Fig. 3d** shows the decomposition $E = E_{\text{loc}} + \Delta E$ for the nMSE of regressing object bounding box coordinates from the feature layer of networks C and CR. Relative to network C, network CR improves both the within-

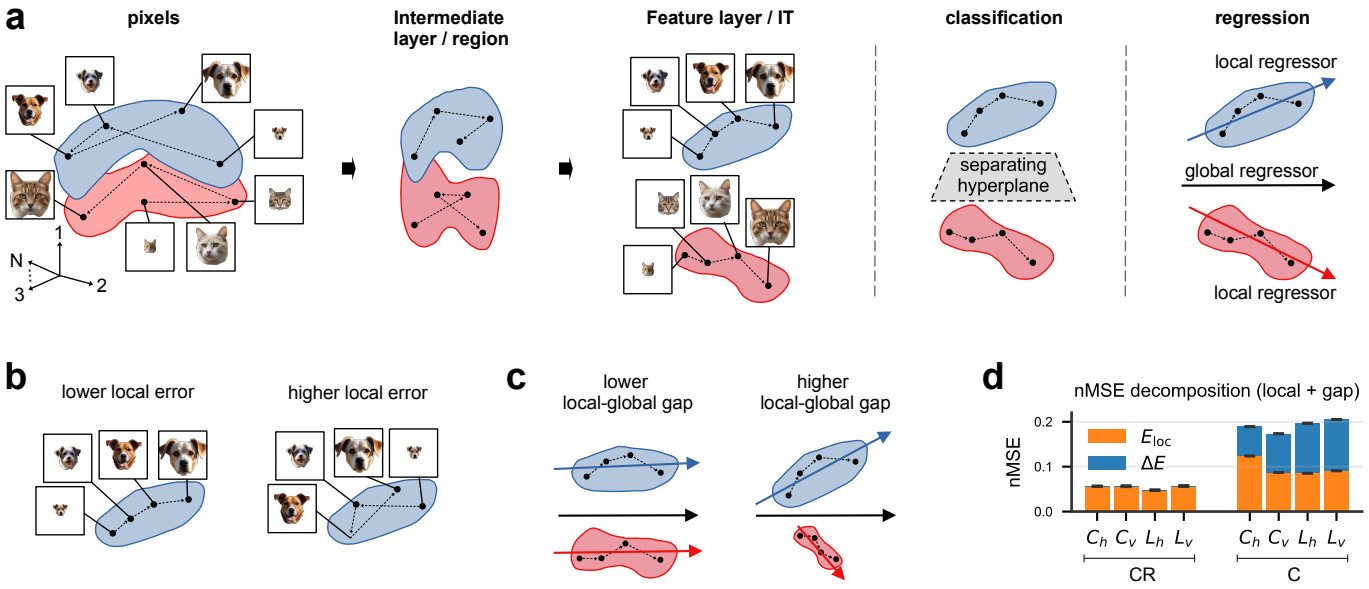


Figure 3. Theoretical framework. (**a**) An object manifold is defined as the collection of all neural representations of images belonging to the same category. We illustrate a "dog" manifold (blue) and a "cat" manifold (red). Each image is mapped to a point (black dot) in the $N$-dimensional space of neural units. Variability within a manifold arises from both category-dependent changes (e.g., different dog or cat exemplars) and category-independent changes, here schematically represented by variation in object size (arrows connect representations in order of increasing size). Along the hierarchy of a system optimized for both object classification and regression of category-independent features, manifolds become more geometrically disentangled (improving linear classification) and category-independent features become more linearly encoded (improving linear regression). (**b**) The local regression error $E_{\mathrm{loc}}$ depends on how linearly the feature is encoded within individual object manifolds. In the left manifold, object size is encoded more linearly than in the right manifold, leading to a lower $E_{\mathrm{loc}}$. (**c**) The local-global error gap $\Delta E$ depends on the geometric arrangement of object manifolds across categories. Along the regressed feature, the manifolds on the left are both more aligned and more similar in scale than the manifolds on the right, resulting in a lower $\Delta E$. (**d**) Decomposition of the nMSE into local error $E_{\mathrm{loc}}$ (orange) and local–global gap $\Delta E$ (blue) for regression of the four bounding box coordinates from the feature layer of networks C and CR. Note that $\Delta E$ is very small and barely visible for network CR.

manifold linear encoding of the feature, captured by $E_{\mathrm{loc}}$, and the across-manifold organization of this encoding, reflected by $\Delta E$. The dominant change is in $\Delta E$, which drops by orders of magnitude in network CR, indicating a coordinated reorganization of manifold geometry across categories that makes regression by a single category-independent readout markedly more effective. Notably, in the SI we show that $\Delta E$ is largely reduced not only in network CR, which is optimized for *category-independent* regression and classification, but also in a network that we term CRloc, which we optimized for *category-specific* regression and classification. In CRloc, category-specific regression is implemented by attaching multiple regression heads to the feature layer—one for each category—rather than a single head shared across categories. This result indicates that optimization for regression alone is sufficient to substantially reduce the local-global gap, even in the absence of the additional constraint imposed by a shared readout across categories.

To gain insights into the manifold-geometry reorganization reflected by $\Delta E$, we develop a theory that predicts $\Delta E$ from readily interpretable geometric measures. Under fairly general assumptions on manifold statistics, in the limit of infinitely many images and categories, and up to order $\mathcal{O}\left(\frac{1}{N}\right)$ in the number $N$ of neurons, we show (see Online Methods and SI) that

$$\Delta E = E_{\mathrm{c}} + \frac{\hat{\sigma}^2}{\sigma^2}\bar{\Delta E}, \qquad \bar{\Delta E} = 1 - \frac{1}{(1+E_{\mathrm{s}})\,(1+E_{\mathrm{o}})} \quad (2)$$

where $E_{\mathrm{s}}$, $E_{\mathrm{o}}$ and $E_{\mathrm{c}}$ represent three distinct sources of error:

1. **Centroid error ($E_{\mathrm{c}}$).** The term $E_{\mathrm{c}}$ reflects the error in fitting the manifold centroids (**Fig. 4a**, left). Each category $\mu$ has a centroid (mean representation) $\bar{x}^{\mu} \in \mathbb{R}^N$ and a mean feature value $\bar{y}^{\mu} \in \mathbb{R}$. A single category-independent regressor must be able to fit the mapping $\bar{x}^{\mu} \to \bar{y}^{\mu}$. Operationally, $E_{\mathrm{c}}$ is the part of $\Delta E$ that would be eliminated by centering each category manifold (and the corresponding feature labels) around its category mean before fitting and evaluating the category-independent regressor (Online Methods). The remaining term $\bar{\Delta E}$ represents the error due to fitting manifold variability around the centroids.

2. **Scale error ($E_{\mathrm{s}}$).** The scale error $E_{\mathrm{s}}$ reflects an

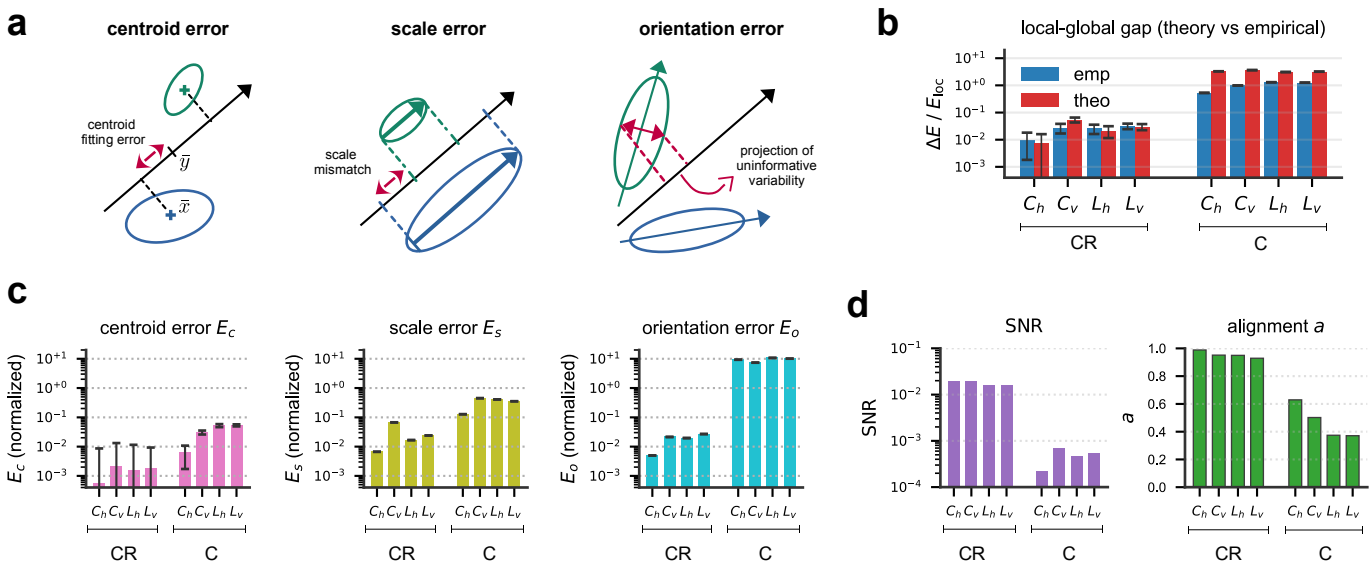


Figure 4. Theory for the local-global error gap. (**a**) Schematic illustration of the three sources of error ($E_\mathrm{c}$, $E_\mathrm{s}$, $E_\mathrm{o}$) contributing to $\Delta E$. Left: The centroid error $E_\mathrm{c}$ reflects the error in fitting the manifold centroids. Center: The scale error $E_\mathrm{s}$ arises from variability in the scale of the local feature encoding across categories. Right: The orientation error $E_\mathrm{o}$ arises from misalignment of the local feature-encoding direction across categories, causing the projection of uninformative manifold variability onto the global regressor. (**b**) Measured relative local-global gap $\Delta E/E_\mathrm{loc}$ (blue) compared to our theory prediction (red). (**c**) Decomposition of the relative gap $\Delta E/E_\mathrm{loc}$ into the contributions from the centroid error $E_\mathrm{c}$ (pink) scale error $E_\mathrm{s}$ (olive) and orientation error $E_\mathrm{o}$ (cyan). All three errors are normalized by dividing by $E_\mathrm{loc}$, and scale and orientation error are further rescaled by $\frac{\hat{\sigma}^2}{\sigma^2}$. This rescaling allows comparability of the three terms, as one can note that, to first order expansion, $\Delta E \sim E_\mathrm{c} + \frac{\hat{\sigma}^2}{\sigma^2}E_\mathrm{s} + \frac{\hat{\sigma}^2}{\sigma^2}E_\mathrm{o}$ (**d**) The two elements contributing to the orientation error: SNR, and alignment $a$ of the local feature-encoding directions.

across-category mismatch in the scale at which the feature is encoded within each manifold (**Fig. 4a**, center). Let $W^\mu \in \mathbb{R}^N$ denote the optimal local readout vector regressing the feature in category $\mu$, and define the scale of the local encoding as $\|W^\mu\|$. We then define

$$E_\mathrm{s} = \frac{\mathrm{Var}_\mu\left[\|W^\mu\|\right]}{\mathrm{Avg}_\mu\left[\|W^\mu\|\right]^2}\,, \tag{3}$$

which is the squared coefficient of variation of $\|W^\mu\|$ across categories (and vanishes only when all categories encode the feature with the same scale).

3. **Orientation error ($E_\mathrm{o}$).** The orientation error $E_\mathrm{o}$ reflects across-category misalignment of the local encoding directions $\hat{w}^\mu = W^\mu/\|W^\mu\|$. Misalignment causes within-manifold variability that is uninformative for the feature to project onto the global category-independent regressor, increasing error (**Fig. 4a**, right). We obtain

$$E_\mathrm{o} = \frac{1-a}{N\cdot a\cdot \mathrm{SNR}} \tag{4}$$

where $a = \left\|\mathrm{Avg}_\mu\left[\hat{w}^\mu\right]\right\|^2$ quantifies the alignment of the local encoding directions ($a = 0$: no net alignment; $a = 1$: perfect alignment) and $\mathrm{SNR} = \mathrm{Avg}_\mu\left[\hat{w}^{\mu\top}\Sigma^\mu\hat{w}^\mu\right]/\mathrm{Avg}_\mu\left[\mathrm{Tr}\Sigma^\mu\right]$ measures the signal-to-noise ratio between the (informative) variance along the local encoding direction and the total (uninformative) manifold variance. Here $\Sigma^\mu \in \mathbb{R}^{N\times N}$ is the centered covariance matrix of representations belonging to the $\mu$-th manifold. The $N^{-1}$ scaling reflects a high-dimensional advantage: in larger representational spaces, uninformative within-manifold variability is increasingly likely to be nearly orthogonal to the global regressor and thus to contribute less readout noise, allowing a fixed $E_\mathrm{o}$ to be attained with smaller alignment $a$ or SNR.

Finally, the prefactor $\hat{\sigma}^2/\sigma^2$ converts the dimensionless gap $\Delta E$ into nMSE units: $\hat{\sigma}^2$ is the variance of the feature explained by the optimal local linear readout (averaged across categories), whereas $\sigma^2$ is the total feature variance (Online Methods).

**Fig. 4b** compares our theoretical prediction with the measured local–global error gap $\Delta E$. The theory captures well the orders-of-magnitude reduction in $\Delta E$ achieved by network CR compared to network C. **Fig. 4c** further decomposes the theoretical prediction of $\Delta E$ into contributions from the centroid error $E_\mathrm{c}$, scale error $E_s$, and orientation error $E_\mathrm{o}$. In both networks, the centroid error is always negligible compared to the other errors. In network CR, both $E_\mathrm{s}$ and $E_\mathrm{o}$ are substantially reduced to

comparably small values, with neither term dominating. In contrast, the poor regression performance of network C is primarily driven by a large orientation error $E_o$. Finally, **Fig. 4d** shows the two factors contributing to the orientation error $E_o$: the SNR and the alignment index $a$. Both are improved in network CR.

### 4. *Manifold geometry reveals optimization strategy*

Since network CR optimally supports both classification of object category and regression of category-independent features, it must have a manifold geometry optimized for both tasks. We therefore ask how optimization of manifold geometry relevant for regression can affect, or coexist with, the geometry relevant for classification, such as manifold radius, dimensionality, and centroid separation [7, 8]. To address this, we consider the manifold geometry of network C, which is optimized for classification, and examine how it changes in network CR, which is additionally optimized for regression. We find that, because regression-relevant geometry is mostly related to a particular direction within the manifold—the feature-encoding direction—rather than the manifold as a whole, it can be optimized with high flexibility while largely preserving classification-relevant geometry.

We begin by examining how regression optimization affects manifold shape, which we estimate from the principal components (PCs) of the manifold covariance. Manifold shape is directly related to classification-relevant geometry such as manifold radius and dimensionality [7, 8]. One example of a regression-relevant measure whose optimization may affect manifold shape is SNR. As mentioned above, optimizing this quantity is flexible, as it can proceed via two complementary strategies (**Fig. 5a**): (i) keep the regressed feature's local encoding direction fixed and reshape the manifold by increasing variance along this direction (and/or reducing variance in directions orthogonal to it); or (ii) keep the manifold shape fixed and instead re-encode the feature along more dominant PC directions, increasing variance along the encoding direction without changing the manifold shape. We find that network CR implements strategy (ii) almost exclusively, thereby preserving the manifold shapes found in the classification-optimized network C: manifold PC spectra are essentially unchanged relative to network C (**Fig. 5b**), while the overlap of the local feature-encoding direction with the top PCs is substantially increased (**Fig. 5c**).

Similarly to manifold shapes, we also find that manifold centroids—the other key geometric property relevant for classification performance—show only modest changes in norm and relative distance compared to network C (see SI).

More generally, it is interesting to ask how regression training affects direction-agnostic measures of manifold geometry, i.e., quantities that can be defined without reference to the feature-encoding direction or explicit identification of the underlying category-independent features. In addition to the manifold shapes and centroids discussed above, another such measure is global manifold alignment. We summarize it by an alignment matrix $A$, whose entry $A_{\mu\nu} \in [0, 1]$ is large when manifolds $\mu$ and $\nu$ are aligned, i.e., their axes of variability (PCs) point in similar directions and carry comparable variance (see **Fig. 5** caption for the formal definition). We have discussed how regression optimization increases a form of direction-specific alignment, namely the alignment $a$ of the local feature-encoding directions. Its effect on global manifold alignment $A$ is, however, nontrivial, because $a$ could be increased via two complementary strategies (**Fig. 5d**) : (1) rotating (in the sense of re-encoding) the feature-encoding direction *within* each manifold, without changing manifold orientation and thus global manifold alignment; or (2) rotating the manifolds themselves while keeping the feature-encoding direction fixed relative to their associated manifold, thereby increasing global manifold alignment. While we have observed a mechanism compatible with option (1) in our analysis of manifold shapes (**Fig. 5c**), we find that network CR also implements option (2), as reflected by an average increase in the entries of $A$ compared to network C (**Fig. 5e**). Interestingly, while in network C the alignment matrix $A$ exhibits a block structure that mirrors the semantic organization of categories, in network CR this structure is largely obscured by an overall increase of alignment across all categories, likely driven by the aligned encoding of category-independent features (**Fig. 5f**).

In sum, these results show that even explicit optimization for regression of category-independent features can leave only a subtle signature on manifold geometry: manifold shapes and centroid separation see only modest relative changes, while manifold alignment increases appreciably, yet remains limited in absolute terms (mean alignment $\ll 1$). In the SI, we show that this picture changes dramatically when the network is optimized exclusively for regression (network R). Without the constraint imposed by the classification objective, manifold dimensionality drops to values close to four—the number of encoded features—and manifold alignment increases drastically (mean alignment $\sim 0.82$).

### 5. *Decoding performance along the visual hierarchy*

To understand how the linear encoding of category and category-independent features emerges along the visual hierarchy, we next examine linear classification and bounding-box regression performance as a function of layer depth (**Fig. 6a**). We make three observations: (1) regression performance improves with layer depth even in network C, which had no explicit regression optimization (consistent with what observed in [4]); (2) The relative local-global gap remains mostly constant in network C,

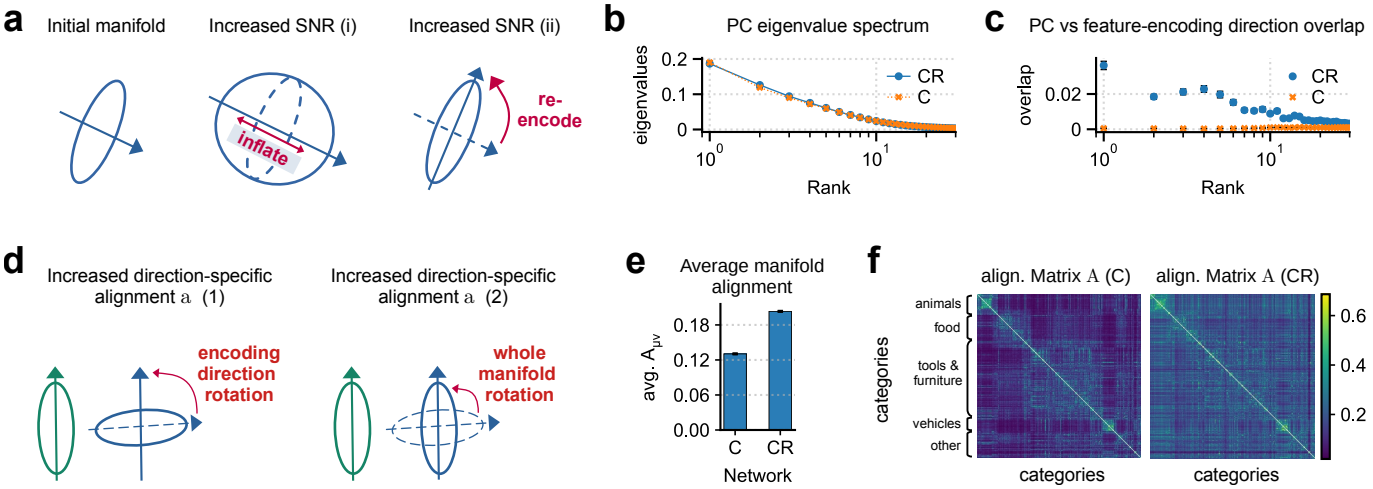


Figure 5. Geometry optimization strategy. (**a**) Two complementary strategies to increase SNR: (i) Inflate manifold variance along the feature-encoding direction (and/or deflate variance along orthogonal directions); or (ii) rotate (in the sense of re-encoding) the feature-encoding direction towards more dominant manifold PCs. (**b**) PC eigenvalue spectrum (eigenvalue vs rank) for category manifolds in network C (orange) and CR (blue), averaged across categories (error bars are SEM). For each manifold $\mu$, the eigenvalues were normalized by the manifold total variance $\mathrm{Tr}\Sigma^{\mu}$, averaged across networks C and CR. Per-category comparisons of the PC spectra are provided in the SI and lead to the same conclusion: manifold shapes are largely unchanged in network CR compared to C. (**c**) Squared cosine angle between the feature-encoding direction $\hat{w}^{\mu}$ and the PC directions of manifold $\mu$ (ordered by rank) for networks C (orange) and CR (blue), averaged across categories (error bars are SEM). We show the case of feature $L_{\mathrm{h}}$ (see SI for other features). (**d**) Two complementary ways of increasing the alignment $a$ between feature-encoding directions: (1) rotating the feature encoding direction *within* the manifold, without rotating the manifold, or (2) rotating the manifold itself, thereby rotating the feature-encoding direction with it. (**e**) Average of the off-diagonal entries of $A$, for networks C and CR. An entry of the global alignment matrix $A$ is defined as $A_{\mu\nu} = \mathcal{N}^{-\frac{1}{2}}\,\mathrm{Tr}\left(\Sigma^{\mu}\Sigma^{\nu}\right)$, where $\Sigma^{\mu}$ is the covariance of manifold $\mu$ and $\mathcal{N} = \mathrm{Tr}\left(\Sigma^{\mu 2}\right)\mathrm{Tr}\left(\Sigma^{\nu 2}\right)$. (**f**) Manifold alignment matrix $A$, for network C (left) and network CR (right). Each matrix entry $A_{\mu\nu}$ quantifies the alignment between a different pair of category manifolds. Categories are ordered (and grouped) by semantic similarity; macro-groups are indicated on the left.

while it improves by almost two-orders-of-magnitude in network CR; and (3) sharp improvements in regression performance, as well as a clear separation between CR and C, emerge only in late layers (from layer 4.0 onward), coincident with a sharp rise in classification accuracy in both networks.

Taken together, these observations motivate two comments. First, although a monotonic improvement in global regression error is consistent with the emergence of a joint code, it is not by itself specific to such a code, since it is also observed in network C. By contrast, an improvement in the local–global gap appears to provide a more informative distinction between networks C and CR. Experimental studies of the joint-coding hypothesis in the brain have so far emphasized changes in global regression performance across the visual hierarchy [4]. Our observations suggest that future experiments could probe the joint-code hypothesis more decisively by complementing this approach with direct tests of how the local–global gap evolves across cortical areas.

The second comment is that, although not in itself evidence for a joint code, it is nonetheless noteworthy that regression performance improves across the layers even in a network not explicitly optimized for this task, such as network C. This trend is somewhat counter to a simple expectation. One might have expected early layers, with their localized, oriented, Gabor-like receptive fields, to provide relatively direct linear access to the spatial information summarized by the bounding-box parameters. Conversely, one might have expected this information to become less linearly accessible in later layers, whose larger receptive fields and classification objective could favor invariance to bounding-box variations that are irrelevant for categorization. One possible interpretation is that bounding-box regression for complex natural objects is not purely a low-level spatial computation. Estimating an object's bounding box may also require recognizing and parsing the object from the rest of the image, thereby linking the problem to object recognition. From this perspective, as categorization improves along the hierarchy, category-independent information may also become more linearly accessible. Consistent with this possibility, network CR, which is explicitly trained for regression, shows its largest improvements in regression performance only in late layers, where classification accuracy also rises sharply. We do not view this as conclusive, but rather as a suggestive pattern.

Finally, we also report how the manifold-geometry

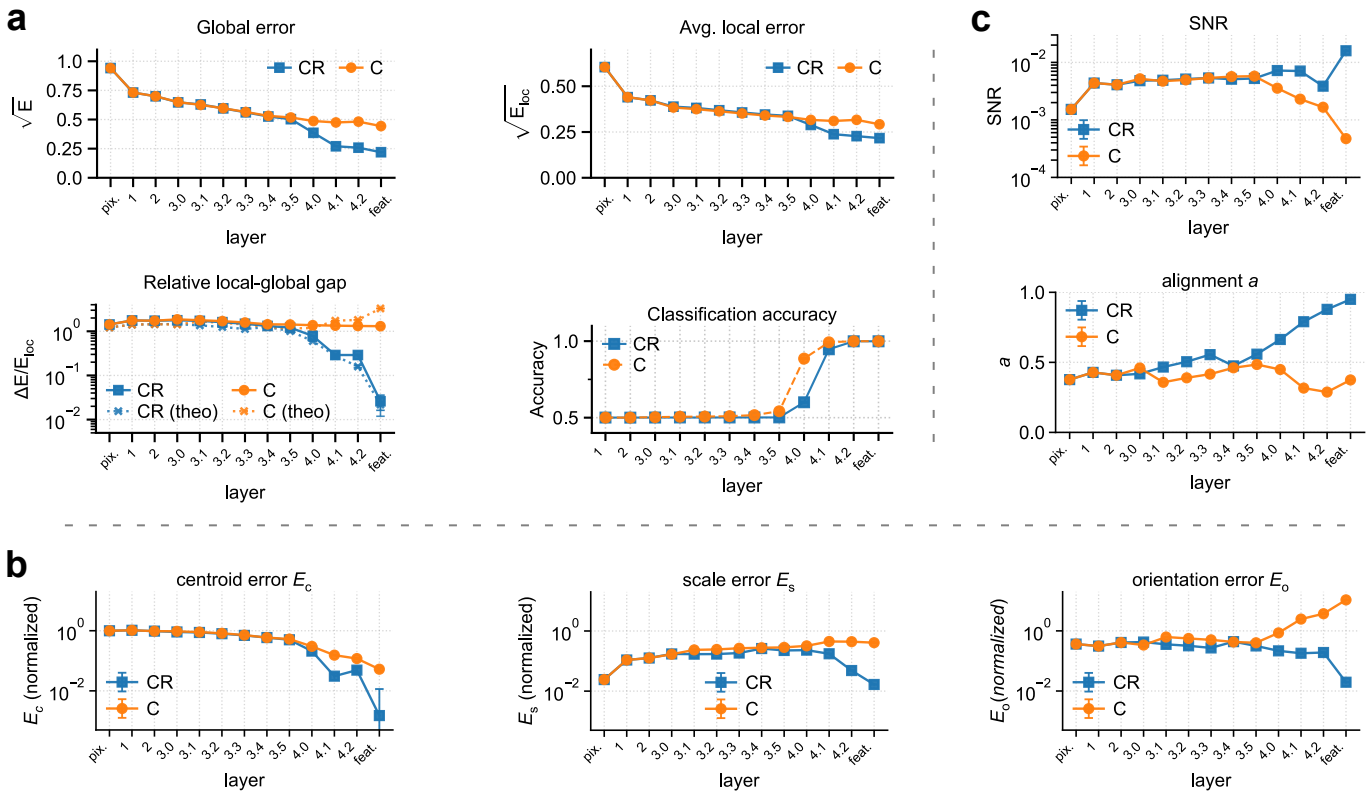


Figure 6. Decoding performance along the visual hierarchy. (**a**) Regression performance for the bounding-box parameter $L_{\mathrm{h}}$ and classification performance, using linear readouts from progressively deeper layers of networks C (orange circles, full line) and CR (blue squares, full line). At each layer, linear readouts were trained on activations randomly projected to 2048 dimensions, to match the dimensionality of the final feature layer. For regression, we report the global regression error, as well as its decomposition into local error $E_{\mathrm{loc}}$ and relative local-global gap $\Delta E/E_{\mathrm{loc}}$. For the gap, we also report our theory prediction (x markers, dotted lines). Note that global and local error are reported as root nMSE. (**b**) The three contributions to the relative local-global gap, as identified by our theory: centroid error $E_{\mathrm{c}}$, orientation error $E_{\mathrm{o}}$, scale error $E_{\mathrm{s}}$. These quantities are normalized in the same way as explained in **Fig. 4**. (**c**) The two elements contributing to the orientation error: SNR, and alignment $a$ of the local feature-encoding directions.

measures contributing to the local–global gap evolve across layers—namely the centroid, scale, and orientation errors (**Fig. 6b**). In both networks C and CR, the centroid error is only relevant in early layers, while in later layers it becomes negligible compared to the other sources of error. The scale error remains approximately constant in network C, whereas network CR tends to reduce it in the final layers. In contrast, the orientation error increases sharply in network C. This behavior can be understood by inspecting the two components of the orientation error: SNR and alignment $a$ (**Fig. 6c**). In network C, alignment $a$ remains approximately constant, while SNR decreases drastically. This is consistent with the classification-only objective of network C, which incentivizes suppression of class-irrelevant variability such as bounding-box coordinates, thereby reducing their SNR. By contrast, network CR appears to counteract this increase in orientation error by improving both SNR and alignment $a$ across layers.

### 6. *Subsampling effects*

We next characterize how typical experimental constraints, namely the subsampling of object categories and recorded units, affect bounding-box regression error in its two components: average local error $E_{\mathrm{loc}}$ and local-global error gap $\Delta E$. We have seen that the most distinctive feature of network CR compared to C is a small relative local–global gap ($\Delta E/E_{\mathrm{loc}} \ll 1$), therefore our analysis focuses especially on this quantity.

We first analyze the effect of subsampling the number of object categories. Since the local error $E_{\mathrm{loc}}$ quantifies regression performance within individual categories (averaged across categories), it is not systematically affected by this form of subsampling, except for an increase in estimation error due to the reduced number of sample categories. We thus focus on the local-global error gap: **Fig. 7a** (blue curve) shows the relative gap $\Delta E/E_{\mathrm{loc}}$ for network C, as a function of the number of categories

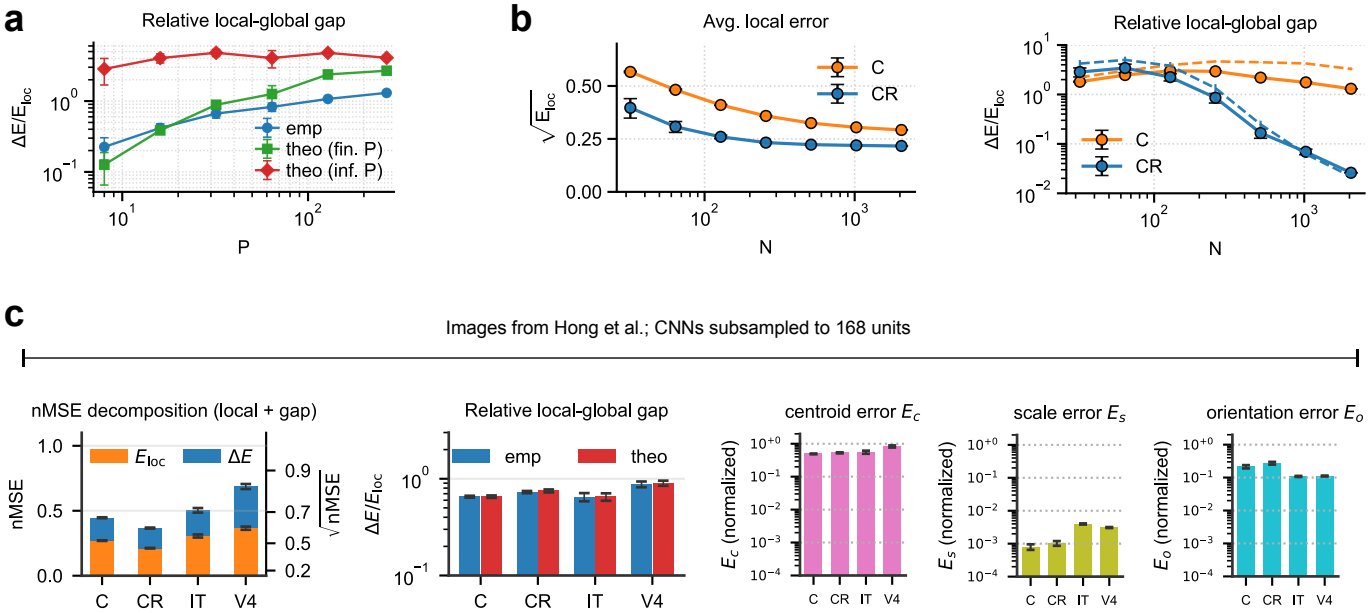


Figure 7. Subsampling effects. (**a**) relative gap $\Delta E/E_{\text{loc}}$ for regressing the bounding-box feature $L_h$ from network C, as a function of the number $P$ of randomly subsampled categories (out of the 265 in our dataset). Blue dots: measured gap; red diamonds: theoretical prediction in the infinite-categories limit; Green squares: theoretical prediction for finite $P$. Error bars denote standard deviation across 4 independent category subsamples. (**b**) $E_{\text{loc}}$ and $\Delta E/E_{\text{loc}}$ for regressing the bounding-box feature $L_h$ from networks C (orange) and CR (blue) on our dataset, as a function of the number $N$ of randomly subsampled units (out of 2048 feature-layer units). (**c**) Analysis of the regression performance of feature $L_h$, using the images and corresponding neural recordings found in [10]. We report the analysis of regression from the IT and V4 units recorded in [10], as well as 168 randomly projected units from the feature layer of networks C and CR. Error bars denote SEM across cross-validation splits and, where applicable, random unit subsamples.

$P$ used in the estimation of regression performance. As $P \to 0$, the relative gap is largely underestimated, risking to confound network C, which is only optimized for category classification, with a network that is also optimized for bounding-box regression. We extended our theory (see Online Methods and SI) to capture this finite-$P$ behavior (green curve). Intuitively, the apparent reduction in $\Delta E$ at small $P$ can be viewed as a form of overfitting: with fewer category manifolds constraining the global regressor, the latter can more easily orient in a direction that avoids projecting onto uninformative manifold variability and thus reduces $\Delta E$, even if this direction would be suboptimal once many categories are included. Crucially, our theory allows one to avoid mistaken conclusions due to this finite-$P$ overfitting, as it can extrapolate the local-global gap to the limit of infinite number of categories (**Fig. 7a**, red curve). In particular, even when the theory is fit using manifold statistics computed from as few as $P = 8$ categories (the same number of categories found, for example, in macaque recordings [10] used to probe the joint-code hypothesis [4]) it predicts a local–global gap consistent with the value predicted using all of the $P = 265$ categories (red curve).

Second, we analyze the effect of subsampling the number of neural units. **Fig. 7b** shows the local error $E_{\text{loc}}$ and the relative gap $\Delta E/E_{\text{loc}}$ in networks C and CR, as a function of the number $N$ of randomly subsampled units drawn from the 2048 units in the feature layer. Note that the effect of subsampling becomes strongly destructive for sufficiently small $N$: when $N$ falls to the order of a few hundred units, the local–global gap in network CR increases to the same level as in network C, rendering the two networks indistinguishable by this key metric. In the SI we show that, under this form of subsampling, each of the three error sources identified by our theory converge to similar values in networks C and CR, with the orientation error $E_{\text{o}}$ and centroid error $E_{\text{c}}$ becoming the dominant error contributions. We find a similar phenomenology when analyzing the images and corresponding neural recordings in [10] (**Fig. 7c**): IT, V4, as well as networks C and CR subsampled to 168 units, matching the number recorded in IT, all show comparable values of $\Delta E/E_{\text{loc}}$, as well as comparable contributions from each of the three error sources, with $E_{\text{o}}$ and $E_{\text{c}}$ emerging as the dominant terms. Although not conclusive, the similarity of the local–global gap and its error sources across all networks and neural recordings, together with the same effect observed in our subsampling analysis of CNNs on our image dataset, suggests that this collapse may be driven by the limited number of recorded units and that future experiments may benefit from larger-scale recordings. Unfortunately, the same subsampling analy-

sis shown in **Fig. 7b** cannot be carried out with our CNNs on the image set used in [10], because that dataset contains only 400 images per category. When the number of neurons becomes comparable to, or exceeds, 400, the local regressors overfit, artificially inflating the estimated local error and thereby underestimating the local–global gap (see SI).

## III. DISCUSSION

In this work, we addressed the questions of what properties of a population code enable an effective joint encoding of object category and category-independent features, and what learning objectives lead to such a code. Using ventral-stream–like CNNs [5], we first showed that such a joint code is possible: a CNN can be trained so that the same feature-layer representation supports accurate linear readout of both category and bounding-box parameters. We established this on a large-scale single-object image dataset developed for this study, which features photo-realistic images with controlled bounding-box distributions. We then developed a theory within the framework of category manifolds [7, 8] to understand which properties of these manifolds enable not only linear separability of categories, but also regression of category-independent features.

While previous manifold-based theories linked category-manifold geometry to classification performance [7, 8, 18], here we extended that framework to provide an analogous account for regression of category-independent features. At the basis of our theoretical framework is a decomposition of regression error into a local error and a local-global error gap, which was not considered in previous studies of the joint-coding hypothesis [4, 6], yet provides an essential element for understanding regression performance. The local error captures how well a feature is linearly encoded within individual category manifolds, whereas the local-global gap captures whether this encoding is organized consistently enough across manifolds to be accurately read out by a single category-independent regressor. In our empirical observations, the local-global gap emerges as a key distinctive signature of a joint code, being reduced by orders of magnitude relative to a population code optimized exclusively for object classification. Notably, the local-global gap is strongly reduced not only when networks are optimized for *category-independent* regression, but also when they are optimized for *category-specific* regression, i.e., regression from each category separately. This result indicates that optimization for regression alone is sufficient to produce a small local-global gap, even in the absence of the additional constraint imposed by a shared readout across categories.

Our theory explains the local-global error gap in terms of interpretable measures of geometric mismatch across category manifolds: variability in the scale at which a feature is encoded within each manifold, and misalignment of the directions along which that feature is encoded. In theory, these measures can be optimized without affecting the separability of category manifolds. Consistent with this observation, when we compare a CNN implementing a joint code with one optimized exclusively for object classification, we find that these measures are improved while largely preserving manifold shapes and locations, and thus manifold properties tied to classification performance such as manifold radius, dimensionality and centroid separation. An important open question is whether this picture persists when the code must support many more category-independent variables. At present, large-scale datasets for studying the joint-code hypothesis [6], including ours, contain annotations for only a few such variables. Developing large-scale datasets that can probe the regime of many more category-independent features remains an important challenge for future work.

It is important to emphasize that the question discussed above—how regression-relevant geometry can be optimized while preserving classification-relevant geometry—is directly addressable in our framework because our theory of regression is formulated at the level of category manifolds. This differs from a recent theory of manifold regression [19], in which manifolds are defined by discretized values of the regressed feature rather than by category.

Examining linear decoding performance across CNN layers, we found, consistent with [4], that regression performance improves even across the layers of a CNN trained exclusively for classification. Importantly, here we make an additional observation: in the classification-only CNN, this improvement is confined to the local error, while the local-global gap remains essentially unchanged. By contrast, in a CNN optimized to support the two tasks, the local-global gap is reduced by orders of magnitude across the final layers, confirming our theory that a small local-global gap is a particularly distinctive signature of a joint population code. Experimental studies probing the joint-coding hypothesis in the brain have so far focused primarily on improvements in global regression performance across the visual hierarchy [4]. Our results suggest that future experiments could go further and probe the joint-code hypothesis more decisively by directly testing how the local-global gap evolves across cortical areas.

To elucidate the effect of experimental constraints, we characterized the behavior of the local-global gap in our CNNs under two common constraints: subsampling of neural units, and limited number of categories in the image dataset. We find that the local-global gap becomes severely degraded once the population is reduced to roughly 200 units, suggesting that future experiments aimed at testing the local-global gap might require recordings from a large number of units. Another important issue is the finite number of categories, denoted here by $P$, in experimental setups. Previous manifold theo-

ries [7, 18] focused on the regime of a finite number of categories per neuron, namely a fixed ratio of $\alpha = P/N$ where $N$ is the number of neurons used for the readout. Other theories, such as few-shot generalization of object categories consider tasks involving a small number of categories per task [8]. In contrast, for a representation to support global regression of category-independent features, it must be able to do so for images associated with any category from the same statistical ensemble, hence the relevant limit is $P \to \infty$ for large but fixed $N$. Indeed our theory predicts that using regression with small $P$, the local-global gap can be strongly underestimated. Fortunately, the theory shows how to extrapolate results from finite $P$ to the large $P$ limit (cf. **Fig. 7a**). In this sense, our work goes beyond previous empirical analyses of the local-global gap in pairs of manifolds [20] not only because it provides a theoretical account of the gap in terms of interpretable manifold-geometry measures, but also because it treats this gap in the $P \to \infty$ regime most relevant for cross-category generalization, while still providing accurate finite-$P$ corrections.

Regarding empirical studies of artificial networks, several directions remain open. Here we focused on CNNs as a ventral-stream model [5, 6], and on supervised training to cleanly compare representations explicitly optimized for a joint code with those optimized only for classification. Recent work in self-supervised vision has emphasized the learning of general-purpose visual features that transfer across a broad range of downstream tasks [21]. An important next step would be to ask whether a joint code can arise in self-supervised training [22–27], and more specifically how the image transformations used during training preserve or weaken the linear accessibility of category-independent variables. We also did not analyze state-of-the-art object detectors [28–33], because our goal was to study direct linear readout from a single shared representation, rather than the more elaborate decoding pipelines implemented in these systems. Among current detectors, DETR [33] is arguably the closest to implementing a joint population code, since both category and bounding-box information are read out from a common representation space (although nonlinearly for bounding-box); because this readout is applied repeatedly across queries to recover multiple objects in a scene, DETR may offer a useful paradigm for thinking about how category and category-independent features of multiple objects could be decoded from a shared population representation.

## ACKNOWLEDGMENTS

This research is supported by the Swartz Foundation, the Gatsby Charitable Foundation, the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University, the grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics and Office of Naval Research grant No. N0014-23-1-2051.

[1] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, Nature **497**, 585 (2013).

[2] S. Bernardi, M. K. Benna, M. Rigotti, J. Munuera, S. Fusi, and C. D. Salzman, Cell **183**, 954 (2020).

[3] S. Fusi, E. K. Miller, and M. Rigotti, Current opinion in neurobiology **37**, 66 (2016).

[4] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo, Nature neuroscience **19**, 613 (2016).

[5] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, *et al.*, BioRxiv , 407007 (2018).

[6] Y. Xie, W. Huang, E. Alter, J. Schwartz, J. B. Tenenbaum, and J. J. DiCarlo, in *The Thirteenth International Conference on Learning Representations*.

[7] S. Chung, D. D. Lee, and H. Sompolinsky, Physical Review X **8**, 031003 (2018).

[8] B. Sorscher, S. Ganguli, and H. Sompolinsky, Proceedings of the National Academy of Sciences **119**, e2200800119 (2022).

[9] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, Nature communications **11**, 746 (2020).

[10] N. J. Majaj, H. Hong, E. A. Solomon, and J. J. DiCarlo, Journal of Neuroscience **35**, 13402 (2015).

[11] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, arXiv preprint arXiv:2307.01952 (2023).

[12] S. Lin, A. Wang, and X. Yang, arXiv preprint arXiv:2402.13929 (2024).

[13] I. Tolstykh, M. Chernyshov, and M. Kuprashevich, in *International Conference on Image Processing and Vision Engineering* (Springer, 2025) pp. 95–110.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022) pp. 10684–10695.

[15] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, arXiv preprint arXiv:2310.04378 (2023).

[16] S. Luo, Y. Tan, S. Patil, D. Gu, P. Von Platen, A. Passos, L. Huang, J. Li, and H. Zhao, arXiv preprint arXiv:2311.05556 (2023).

[17] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, in *Proceedings of the IEEE/CVF international conference on computer vision* (2019) pp. 8430–8439.

[18] A. J. Wakhloo, T. J. Sussman, and S. Chung, Physical Review Letters **131**, 027301 (2023).

[19] W. Slatton, C.-N. Chou, and S. Chung, arXiv preprint arXiv:2603.10956 (2026).

[20] M. Kuoch, C.-N. Chou, N. Parthasarathy, J. Dapello, J. J. DiCarlo, H. Sompolinsky, and S. Chung, in *Conference on Parsimony and Learning* (PMLR, 2024) pp. 395–418.

[21] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, arXiv preprint arXiv:2304.07193 (2023).

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, in *International conference on machine learning* (PmLR, 2020) pp. 1597–1607.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020) pp. 9729–9738.

[24] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, Advances in neural information processing systems **33**, 21271 (2020).

[25] A. Bardes, J. Ponce, and Y. LeCun, arXiv preprint arXiv:2105.04906 (2021).

[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, in *Proceedings of the IEEE/CVF international conference on computer vision* (2021) pp. 9650–9660.

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022) pp. 16000–16009.

[28] S. Ren, K. He, R. Girshick, and J. Sun, Advances in neural information processing systems **28** (2015).

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, in *European conference on computer vision* (Springer, 2016) pp. 21–37.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, in *Proceedings of the IEEE international conference on computer vision* (2017) pp. 2980–2988.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 779–788.

[32] Z. Tian, C. Shen, H. Chen, and T. He, in *Proceedings of the IEEE/CVF international conference on computer vision* (2019) pp. 9627–9636.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, in *European conference on computer vision* (Springer, 2020) pp. 213–229.

# Methods

### A. Linear decoding

To measure the performance of linear readouts in classifying object category and regressing category-independent features, we used the following methods.

***Classification.*** We decoded object category using linear one-vs-rest classifiers trained with ridge regression. For each category, we fit a separate linear readout with binary label ($y = 1$ for images in the category, $y = 0$ otherwise) and converted readout scores to class predictions using a fixed threshold of 0.5. The ridge regularization strength was chosen by nested cross-validation, and reported performance was computed on held-out data across outer splits. Classification performance is reported as balanced accuracy, defined for each one-vs-rest task as

$$\mathrm{BalAcc} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{Pos}} + \frac{\mathrm{TN}}{\mathrm{Neg}}\right) \tag{5}$$

where TP and TN are the numbers of true positives and true negatives on the test set, and Pos and Neg are the total numbers of positive and negative test samples, respectively. We report the final performance as balanced accuracy averaged across categories.

***Global regression.*** To decode bounding-box features at the global, across-category level, we fit linear ridge regressors from the neural representation to each scalar bounding-box feature separately. The ridge regularization strength was selected by nested cross-validation, and reported performance was computed on held-out data across outer splits. Outer and inner splits used a k-fold procedure. Regression performance is reported as normalized mean squared error (nMSE), or, equivalently, its root $\sqrt{\mathrm{nMSE}}$. We first compute the test-set MSE in each outer fold and average it across folds. We then normalize this average MSE by the variance of the target feature across the full dataset. Because k-fold cross-validation uses non-overlapping test folds that collectively cover the full dataset, normalizing by the variance computed over the full dataset provides a natural and consistent scale for nMSE.

***Local regression.*** To estimate the local linear encoding of bounding-box features within each category (used to compute $E_{\mathrm{loc}}$ and the geometry statistics entering our theory), we fit category-specific linear regressors $(W^\mu, b^\mu)$ by jointly optimizing all categories with a multi-task objective. Specifically, we minimize the average within-category MSE plus a shared regularizer that penalizes dispersion of the regressors around their mean:

$$E_{\mathrm{common}} = \frac{1}{P}\sum_{\mu=1}^{P}\left\langle \left(W^{\mu\top}x^\mu + b^\mu - y^\mu\right)^2\right\rangle_\mu + \gamma L\left(\{W^\mu\}\right) \tag{6}$$

with

$$L\left(\{W^\mu\}\right) = \frac{1}{P}\sum_{\mu=1}^{P}\left\| W^\mu - \frac{1}{P}\sum_{\nu=1}^{P} W^\nu \right\|^2 \tag{7}$$

Here $\langle\rangle_\mu$ denotes averaging over samples of category $\mu$ in the training split, and $\gamma$ controls the strength of the shared regularizer. The regularization strength $\gamma$ was selected by nested cross-validation. Performance was evaluated on held-out data across outer splits by computing $E_{\mathrm{loc}}$ as the average test nMSE across categories. In practice, this procedure yielded $E_{\mathrm{loc}}$ values statistically compatible with those obtained by fitting separate ridge regressors per category. Compared to fitting independent per-category ridge regressors, the joint objective above has the advantage of yielding more stable estimates of the local regressors $\{W^\mu\}$, which enter our theory through their across-category variability in norm and direction (i.e., the scale error $E_{\mathrm{s}}$ and orientation error $E_{\mathrm{o}}$). When fitting each category independently with its own ridge penalty selected by cross-validation, we found that the cross-validation curves are often relatively flat, so that a broad range of ridge penalties yield statistically indistinguishable performance. This leads to high variability in the selected penalties across categories and, consequently, unnecessary variability in the estimated $\{W^\mu\}$, which can inflate the estimates of norm and direction variability. The coupled objective above reduces this estimation variance by shrinking each $W^\mu$ toward the mean regressor $\bar{W} = \frac{1}{P}\sum_{\nu=1}^{P} W^\nu$.

***Data splits.*** When applied to artificial networks, all decoding analyses explained above were performed on images not used to train the networks. Specifically, our image dataset consists of 10,000 images per category, which we used for training networks C, R, and CR, and a separate set of another 10,000 images per category, unseen during training, which we used to evaluate linear decoding performance (applying further train/test splitting as described in the paragraphs above).

### B. Dataset generation

Our dataset consists of images representing a single main object, placed in a contextually relevant scene. Each image is labeled with both the object identity (out of 265 categories) and the parameters of the bounding-box surrounding the main object, namely center coordinates $C_h$, $C_v$ and side lengths $L_h$, $L_v$ along the image's horizontal (h) and vertical (v) axes. The 265 categories are a subset of the 365 categories in the taxonomy of the Object365 image dataset, chosen because they could be generated reliably with our pipeline (described below). Each category consists of 10,000 training images (used to train networks C, R, and CR) and a disjoint set of 10,000 test images, unseen during training, used to evaluate linear decoding performance.

Images in our dataset were generated using a multi-stage pipeline that combines text-to-image generation (Stable Diffusion XL, SDXL [11, 12]), object detection (CerberusDet [13]), and image outpainting (Stable Diffusion v1.5, SD1.5 [14–16]) to produce single-object images with controlled object placement and a uniform distribution of bounding-box coordinates. The pipeline to generate one image in our dataset consists of the following three steps (cf. **Fig. 2a**).

***Step 1: Object generation.*** We first generate a "seed" image of a single object in the desired category using Stable Diffusion XL. The goal of this step is simply to obtain a clear, unobstructed, single-instance depiction of the target object in a contextually appropriate scene, without yet enforcing any control over object position or scale (i.e., control over the object's bounding-box parameters). To make large-scale generation feasible, we use SDXL-Lightning weights (model release identifier: ByteDance/SDXL-Lightning) with the SDXL base checkpoint (stabilityai/stable-diffusion-xl-base-1.0), enabling generation in only 4 denoising steps. This allowed us to synthesize the full dataset within a few days on a 10×A100 GPU cluster. For a given category named <category>, we use the following prompt: "A full-body, photorealistic shot of a single <category>. The <category> is the clear main subject, fully in the scene, and unobstructed. The background is detailed, contextually appropriate, but does not distract.". We also use the negative prompt: "clutter, multiple objects, occlusion, low quality, blurry, distorted, unrealistic, extra limbs, cropped, artifacts, persons, people, humans, mannequin".

***Step 2: Object placement.*** The goal of this step is to enforce controlled object location and size by mapping the seed image from step 1 onto a new canvas so that the main object occupies a newly sampled bounding box. We first obtain the object's bounding box in the seed image using CerberusDet, and treat this detected box as the object's initial bounding box. If CerberusDet does not detect any instance of the object, we discard the seed image and return to step 1. We then sample a target bounding box on a fixed 512×512 canvas as follows. First, we sample a target maximum side length $L_{\max} \sim \mathrm{Unif}\,[200, 500]$ (in pixel units) and isotropically rescale the seed image such that the detected bounding-box's maximum side length matches the newly sampled one, that is we ensure $\max(L_h, L_v) = L_{\max}$. Next, conditioned on the resulting $(L_h, L_v)$, we sample the target bounding-box center uniformly over all centers that would keep the box inside the canvas: $C_h \sim \mathrm{Unif}\,[L_h/2, 512 - L_h/2]$ and $C_v \sim \mathrm{Unif}\,[L_v/2, 512 - L_v/2]$. In sum, object size (via $L_{\max}$) is sampled uniformly over a prescribed range, and object position is sampled uniformly conditional on size, subject to the newly sampled bounding-box staying fully within the image. To realize this target bounding-box, we create a blank 512×512 canvas and paste the rescaled seed image onto it with a translation chosen so that the rescaled detected box is centered at $(C_h, C_v)$. This produces an image in which the object occupies the desired bounding box, while regions of the canvas not covered by the pasted image are left undefined; these regions are filled in Step 3 via outpainting.

***Step 3: Outpainting.*** The goal of this step is to fill in the undefined regions introduced in Step 2, producing a natural-looking full 512×512 image while preserving the placed object and its newly sampled bounding box. We achieve this by outpainting the undefined regions using an outpainting variant of SD1.5. Starting from the 512×512 canvas produced in Step 2, we construct an outpainting mask that marks the undefined regions (i.e., pixels not covered by the pasted, rescaled seed image). To reduce visible seams at the boundary between original and synthesized content, we blur the mask edges with a Gaussian blur. Before outpainting, we initialize the undefined regions with reflection padding to provide the generative model with plausible color and texture statistics and reduce sharp discontinuities at the boundary between the pasted seed image and the area to be synthesized. Technically, we fill each blank side by repeatedly tiling a mirror-reflected strip of pixels taken only from the portion of the seed image between the pasted-image border and the object's bounding-box edge (never from inside the bounding box), so that the object itself is not reflected—avoiding to encourage the outpainting model to generate additional instances of the object. We finally outpaint the masked regions using SD1.5 inpainting (stable-diffusion-v1-5/stable-diffusion-inpainting) with an LCM scheduler and the latent-consistency/lcm-lora-sdv1-5 LoRA, using 4 denoising steps, guidance scale 0, and strength 0.95. The outpainting prompt is the same category prompt from Step 1 with the suffix: 'The background extends naturally from the existing scene. Seamless transition between regions, no visible boundaries. Consistent lighting and focus throughout the entire image." The negative prompt is identical to Step 1. Finally, we rerun CerberusDet on the outpainted 512×512 image and accept the sample only if the detector returns exactly one instance of the desired category, so that images with multiple instances of the object are not accepted; for accepted images, we save the outpainted image together with its category and the detected bounding-box coordinates.

***Categories selection.*** As mentioned above, we generated images in 265 categories out of the 365 categories in the Objects365 taxonomy, selected among those that could be generated reliably with our pipeline. To select these categories, we ran a diagnostic pass generating 50 images per category and measured how often the pipeline failed due to CerberusDet either not detecting the target object or detecting multiple instances. We excluded categories with high failure rates, since their low yield would have substantially increased the runtime and compute required to produce 20,000 accepted images per category. Furthermore, we visually inspected the 50 images in each category and excluded categories that systematically produced implausible or low-quality images, such

as ambiguous or inconsistent object identity, severe artifacts, or the presence of multiple salient objects. For a few categories, SDXL introduced characteristic artifacts that were *consistent* across images within that category (e.g., "knife" images often contained forked or unusually curved blades). Since our goal is a dataset in which images within each category share a consistent semantic label, rather than strict realism, we did not exclude such categories. During this inspection we also observed that, even when SDXL produced an unambiguous instance of the intended object, CerberusDet's predicted class name could differ from the prompt category (e.g., "Handbag/Satchel" or "Luggage" for "Backpack," or the non-intuitive confusion of "Parking meter" for "Microphone"). Importantly, in these cases CerberusDet almost never failed to localize the object: the predicted bounding box typically tightly enclosed the intended object, even when the class label was wrong. Because in our pipeline we use CerberusDet exclusively as an object localizer (bounding-box extractor) rather than as a semantic classifier, we treated such label confusions as acceptable when the detected box clearly corresponded to the intended object. Concretely, for each category we defined a small set of allowed detector labels and considered an image successful if CerberusDet detected exactly one instance of any allowed label. This increased pipeline robustness without changing the semantic identity imposed by the text prompt and without affecting bounding-box localization.

### C. CNN architecture and training

***Architecture.*** All CNN models were based on an ImageNet-pretrained ResNet-50. We removed the original classification head and used the final 2048-dimensional feature layer as the shared representation. On top of this feature layer we attached either one or a combination of the following heads: a linear classification head for object category (265 classes) and four linear regression heads predicting the bounding-box parameters $(C_h, C_v, L_h, L_v)$. These heads were randomly initialized with Xavier-uniform weights and zero biases. We trained four variants of this architecture. Network C included only the classification head. Network R included only the four regression heads. Network CR included both the classification head and the four regression heads. Network CRloc had the same backbone and classification head as CR, but replaced each category-independent regression head with a category-specific one: each bounding-box head produced one output per class, and during training the output corresponding to the ground-truth class was selected.

***Training.*** Images were resized to $224 \times 224$ and normalized with the standard ImageNet channel statistics. Bounding-box coordinates were standardized using the training-set mean and standard deviation before regression. Training used shuffled mini-batches from the training split (8000 images per category) and performance was monitored on a held-out validation split (2000 images per category). Models were trained with batch size 208 using stochastic gradient descent (learning rate: 0.02; momentum: 0.9; weight decay: $10^{-4}$; learning-rate schedule: cosine annealing over 40 epochs), Smooth L1 loss for each regression coordinate, and cross-entropy loss for classification. The total loss was the sum of the regression and classification losses. In models without the classification head or without regression heads, the corresponding loss terms were simply omitted. Models were trained until performance on the validation set plateaued. We found that gradually unfreezing the weights of the ResNet-50 layers through the epochs, from the feature layer downstream to layer 1, yielded models with slightly better regression performance. In particular, the models reported in the main text were trained by unfreezing a new layer each epoch, in sequence: the feature layer, layer 4, and layer 3, while layers 2 and 1 were kept frozen throughout the training. In the SI, we show that the trends across layers reported in the main text (cf. **Fig.** 6) are virtually unchanged in networks trained end to end with all layers unfrozen from the very first epoch. In particular, we still observe that a sharp rise in regression performance emerges only from layer 4, as well as an appreciable distinction in regression performance between networks C and CR.

### D. Theoretical framework and results

***Regression errors definition.*** Consider a set $C$ consisting of $P$ categories. Call $x^\mu \in \mathbb{R}^N$ the neural representation of an image (or, more generally, an input stimulus) in the $\mu$-th category, where $N$ is the number of neural units. Let $y^\mu$ be a continuous label associated to the image (e.g. one of the bounding-box parameters). Without loss of generality, we consider globally centered data and labels, that is $\left\langle \langle x^\mu \rangle_\mu \right\rangle_C = 0$ and $\left\langle \langle y^\mu \rangle_\mu \right\rangle_C = 0$, where $\langle \ldots \rangle_\mu$ denotes averaging over all images within the $\mu$-th category, and $\langle \ldots \rangle_C = \mathrm{Avg}_\mu [\ldots]$ denotes averaging over categories.

Define the global regression MSE as

$$E = \min_W \left[ \left\langle \left\langle \left( W^\top x^\mu - y^\mu \right)^2 \right\rangle_\mu \right\rangle_C \right] \tag{8}$$

where $W \in \mathbb{R}^N$ is the global regression vector. We decompose the global regression MSE as $E = E_{\mathrm{loc}} + \Delta E$. Here, $E_{\mathrm{loc}} = \langle E^\mu \rangle_C$ and $E^\mu$ is the local regression MSE for the $\mu$-th category, defined as

$$E^\mu = \min_{(W^\mu, b^\mu)} \left[ \left\langle \left( W^{\mu\top} x^\mu + b^\mu - y^\mu \right)^2 \right\rangle_\mu \right] \tag{9}$$

where $W^\mu \in \mathbb{R}^{\mathbb{N}}$, $b^\mu \in \mathbb{R}$ are the local regression vector and bias for category $\mu$. It can be shown (see SI) that

$\Delta E = E - E_{\rm loc}$ is given by

$$\Delta E = \min_W \left[ \left\langle \left\langle \left( W^\top x^\mu - \hat{y}^\mu \right)^2 \right\rangle_\mu \right\rangle_C \right] \tag{10}$$

where $\hat{y}^\mu = W_*^{\mu\top} x^\mu + b_*^\mu$ is the label predicted by the local regressor for category $\mu$, with the local regressor defined as $(W_*^\mu, b_*) = \mathrm{argmin}_{(W^\mu, b^\mu)} \left[ \left\langle \left( W^{\mu\top} x^\mu + b^\mu - y^\mu \right)^2 \right\rangle_\mu \right]$. In other words, the local-global error gap is itself a global regression error in which the target is the locally linearized label $\hat{y}^\mu$ rather than the original label $y^\mu$.

***Generative model.*** We derive $\Delta E$ for the following generative model of manifolds. We first consider the case in which each manifold and associated labels are locally centered, that is $\langle x^\mu \rangle_\mu = 0$ and $\langle y^\mu \rangle_\mu = 0$, $\forall \mu$. Without loss of generality, we parameterize a point in the $\mu$-th manifold as

$$x^\mu = \sum_{d=1}^{N} s_d^\mu u_d^\mu \tag{11}$$

where $\left\{ u_d^\mu \in \mathbb{R}^N \right\}_{d=1}^N$ is an orthonormal basis and $s_d^\mu \in \mathbb{R}$ is the coordinate of the point along direction $d$. We choose $u_1^\mu$ to correspond to the feature-encoding direction, that is $u_1^\mu = W_*^\mu / \|W_*^\mu\|$. We choose the remaining directions $u_{d>1}^\mu$ to be the principal-component directions of the manifold in the subspace orthogonal to $u_1^\mu$, so that the covariance matrix of the coordinates has the form

$$\left\langle s^\mu s^{\mu\top} \right\rangle_\mu = \begin{matrix} \sigma_1^{\mu 2} & \Sigma_{12}^\mu & \dots & \Sigma_{1N} \\ \Sigma_{21}^\mu & \sigma_2^{\mu 2} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ \Sigma_{N1} & 0 & 0 & \sigma_N^{\mu 2} \end{matrix} \tag{12}$$

where $s \in \mathbb{R}$ is the vector of coordinates whose $d$-th component is $s_d^\mu$. Note that local centering of the manifolds implies $\langle s_d^\mu \rangle = 0$. This choice of parameterization can always be made without loss of generality. We then make the following assumptions on the across-manifold statistics:

1. All manifold directions with $d > 1$ are mutually uncorrelated and isotropically distributed in neuronal space, conditioned on being orthogonal to $u_1^\mu$ (as by their definition as the elements of an orthonormal basis). Concretely,

$$\langle u_d^\mu \rangle_C = 0$$

$$\left\langle u_d^\mu u_{d'}^{\mu\top} \right\rangle_C = \frac{1}{N} \delta_{d,d'} \left( \mathbb{I} - \left\langle u_1^\mu u_1^{\mu\top} \right\rangle_C \right) . \tag{13}$$

2. The feature-encoding directions $u_1^\mu$ are isotropically distributed around a common mean direction

$$u_1^\mu = \sqrt{a} u_1 + \sqrt{1-a} \eta^\mu , \tag{14}$$

where $a \in [0,1]$ is the alignment strength, $u_1 \in \mathbb{R}^N$ is a unit-norm vector representing the common mean direction, and $\eta^\mu \in \mathbb{R}^N$ is a random i.i.d. Gaussian vector with $\langle \eta^\mu \rangle_C = 0$ and $\left\langle \eta^\mu \eta^{\mu\top} \right\rangle_C = \frac{1}{N}\mathbb{I}$. Under this definition, $u_1^\mu$ has unit norm on average, i.e. $\left\langle u_1^{\mu\top} u_1^\mu \right\rangle_C = 1$, with fluctuations that vanish as powers of $1/N$, which is sufficient for our derivations in the limit $N \to \infty$.

3. Aside from the above correlations, all manifold parameters are otherwise uncorrelated.

***Theory results.*** Under this generative model, in the limit of infinitely many categories ($P \to \infty$) and up to $\mathcal{O}\left(N^{-1}\right)$ corrections, we obtain (see SI for the derivation).

$$\Delta E = \hat{\sigma}^2 \left[ 1 - \frac{1}{(1+E_{\rm s})(1+E_{\rm o})} \right] \tag{15}$$

$E_{\rm s} = \frac{\mathrm{Var}_\mu[\|W_*^\mu\|]}{\mathrm{Avg}_\mu[\|W_*^\mu\|]^2}$, $E_{\rm o} = \frac{1-a}{N \cdot a \cdot \mathrm{SNR}}$, $\mathrm{SNR} = \frac{\left\langle \sigma_1^{\mu 2} \right\rangle_C}{\sum_{d=1}^N \left\langle \sigma_d^{\mu 2} \right\rangle_C}$, and $\hat{\sigma}^2 = \left\langle \left\langle \hat{y}^{\mu 2} \right\rangle_\mu \right\rangle_C$. Note that the results in the main text are reported in terms of the nMSE rather than the MSE, i.e. after normalization by the label variance $\sigma^2 = \left\langle \left\langle y^{\mu 2} \right\rangle_\mu \right\rangle_C$. In the case in which manifolds and labels are not locally centered, we quantify the contribution of category centroids through the centroid error, defined as

$$E_{\rm c} = E - E_{\rm fl}, \tag{16}$$

where

$$E_{\rm fl} = \min_W \left[ \left\langle \left\langle \left( W^\top \delta x^\mu - \delta y^\mu \right)^2 \right\rangle_\mu \right\rangle_C \right], \tag{17}$$

with $\delta x^\mu = x^\mu - \langle x^\mu \rangle_\mu$ and $\delta y^\mu = y^\mu - \langle y^\mu \rangle_\mu$ denoting the locally centered representations and labels. That is, $E_{\rm c}$ measures the additional regression error induced by the presence of category-dependent centroids. The remainder of the theory is then derived exactly as above, with $E_{\rm fl}$ taking the role of $E$.

***Theory for finite number of categories.*** We also derive the local-global gap $\Delta E$ for finite number of categories. Formally, the result is derived in the regime of finite category density, by studying the thermodynamic limit $P, N \to \infty$ at fixed $\alpha = \frac{P}{N} \in (0, \infty)$. The result for infinitely many categories presented above is recovered as the limiting case $\alpha \to \infty$. The expression for generic $\alpha$ is lengthy and is provided in the SI, alongside its derivation.

***Fitting theory to data.*** The theory derived above applies to the training regression error. When comparing it with test errors on real data, we assume that the number of images is sufficiently large and the regularization of the fitted regressor is appropriate, so that the train-test generalization gap is small and the theoretical prediction

remains informative also for the test error. For the purpose of estimating the theory parameters from data, the local regressors are taken to be the regularized empirical solutions used in practice, as determined by the procedure described in the "Linear Decoding" section of the Methods, rather than the unregularized minimizers of the local MSE. Accordingly, all quantities entering the theory that depend on the local regressors are evaluated using these regularized estimates.

# Supplementary information

## IV. SUPPLEMENTARY FIGURES

### A. Linear decoding performance in macaque recordings

We have applied the linear decoding framework presented in the main text to the publicly available dataset of macaque recordings [10] studied in [4], as well as to an ImageNet-pretrained ResNet-50 CNN. We reproduce their qualitative results (**Fig. 8**): decoding performance of both category and category-independent features improves along the visual hierarchy. However, we also find that absolute regression performance remains limited: in macaque IT, as well as in the CNN for most features, root nMSE remains above 0.5, meaning that the root-mean-square prediction error is still larger than half the across-image standard deviation of the decoded feature.

We note that this last point may be less apparent in the original presentation of [4], where regression performance was reported using the Pearson correlation coefficient. Here we instead report $\sqrt{\text{nMSE}}$, which we find more directly interpretable, as it measures the typical prediction error (root-mean-squared error) in units of the standard deviation of the decoded feature. In particular, because Pearson correlation is a nonlinear function of nMSE, it tends to compress even appreciable differences in nMSE into a narrow interval near 1. Indeed, for ordinary least-squares (no ridge) evaluated on the training set, Pearson correlation is related to nMSE by $\text{Pears} = \sqrt{1 - \text{nMSE}}$. For ridge regression on held-out data, this relation is only approximate but still useful as an intuition. We can see that Pearson correlation compresses a broad range of $\sqrt{\text{nMSE}}$ values into a narrow range near 1. For example, $\sqrt{\text{nMSE}} \in [0, 0.5]$—from perfect prediction to errors on the order of half a standard deviation—maps to $\text{Pears} \in [0.87, 1]$, making differences in regression accuracy harder to visualize and interpret.

### B. Regression performance for network CRloc

Here we show that optimization for regression alone is sufficient to substantially reduce the local-global gap, even in the absence of the additional constraint imposed by a shared readout across categories. Indeed, we show that $\Delta E$ is largely reduced not only in network CR, which is optimized for *category-independent* regression and classification, but also in a network that we term CRloc, which we optimized for *category-specific* regression and classification (**Fig. 9a-b**). In CRloc, category-specific regression is implemented by attaching multiple regression heads to the feature layer—one for each category—rather than a single head shared across categories. We observed that, in network CRloc, $\Delta E$ is largely reduced compared to network C, down to values only slightly higher than network CR. The same observation holds for the three sources of error $E_\text{c}$, $E_\text{s}$, $E_\text{o}$ (**Fig. 9c**), and the two factors contributing to $E_\text{o}$, namely alignment $a$ of the feature-encoding directions, and SNR (**Fig. 9d**).

### C. Changes in manifold geometry due to regression optimization

*Changes in manifold shapes*

In the main text, we showed that, on average, manifold shapes are only weakly altered in network CR relative to network C (cf. **Fig. 5b**). Here we examine this effect on a category-by-category basis. For a given category $\mu$, let $\{\lambda_d^\mu\}_{d=1}^N$ and $\{\gamma_d^\mu\}_{d=1}^N$ denote the PC eigenvalues of the $\mu$-th category manifold in networks CR and C, respectively. We quantify the spectral difference between these two manifolds as

$$d^\mu = \mathcal{N}^{-1} \left[ \sum_{d=1}^{N} \left( \sqrt{\lambda_d^\mu} - \sqrt{\gamma_d^\mu} \right)^2 \right]^{1/2}, \qquad \mathcal{N} = \left[ \sum_{d=1}^{N} (\lambda_d^\mu + \gamma_d^\mu) \right]^{1/2}. \tag{18}$$

Without the normalization factor $\mathcal{N}$, this quantity corresponds to the Wasserstein-2 distance between the covariance matrices of category $\mu$ in networks C and CR, after aligning their principal axes, which is appropriate here since our goal is to compare manifold shapes independent of their orientations. Intuitively, the Wasserstein-2 distance measures how much one distribution must be rearranged to match the other. Here we normalize this distance by the combined root variance of the two manifolds, $\mathcal{N}$, so that the resulting quantity is dimensionless, bounded in the range $d^\mu \in [0, 1]$,

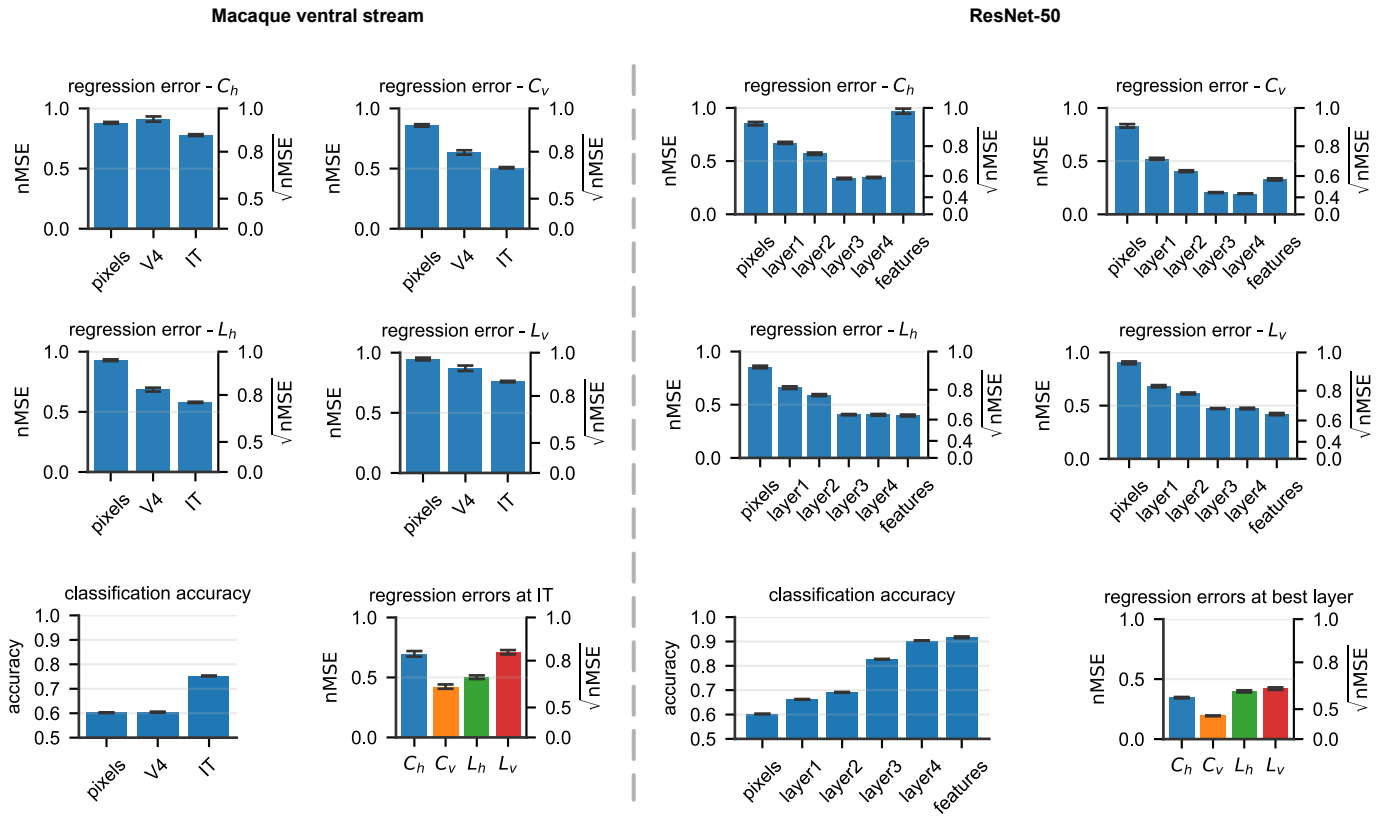


Figure 8. Linear decoding performance along the visual hierarchy, for macaque neural recordings (left column) and ResNet-50 activations (right column). For a fair comparison across regions, the 168 IT units are randomly subsampled to match the 88 units in V4. The regression error in IT using all 168 units is also reported for all four bounding-box features (bottom right of the left column). Similarly, units in layers 1–4 of ResNet-50, as well as the pixels layer, are randomly projected to match the 2048 units in the final feature layer. The regression error at the best performing layer (feature layer or layer 4) is also reported for all four bounding-box features (bottom right of the right column). Error bars indicate the standard error of the mean (SEM) across cross-validation splits and, where applicable, across random subsamples of units.

and comparable across categories. **Fig. 10** reports the histogram of $d^\mu$ across categories, together with example PC spectra for selected categories in networks C and CR. These results show that manifold shapes undergo only small relative changes between networks C and CR, with $\mathrm{Avg}_\mu[d^\mu] \sim 0.083$.

*Changes in centroid norm and separation*

For each category manifold $\mu$, we denote by $\bar{x}_\mu \in \mathbb{R}^N$ its centroid. We compare between networks C and CR two centroid measures: the centroid norm $\|\bar{x}_\mu\|$, and the pairwise relative centroid distance $\|\bar{x}_\mu - \bar{x}_\nu\|/\langle\|\bar{x}_\rho\|\rangle_\rho$ between two manifolds $\mu$ and $\nu$. Centroid norms see only modest changes in network CR relative to network C, while centroid distances are virtually unchanged in distribution (**Fig. 11**).

*Changes in manifold geometry for network R*

In the main text, we analyzed changes in manifold geometry in network CR compared to C. We observed that even explicit optimization for regression of category-independent features can leave only a subtle signature on manifold geometry: manifold shapes and centroid separation see only modest relative changes, while manifold alignment increases appreciably, yet remains limited in absolute terms. In **Fig. 12**, we show that this picture changes dramatically when the network is optimized exclusively for regression (network R). Without the constraint imposed by the classification objective, manifold dimensionality drops to values close to four—the number of encoded features— (panels a-b) and

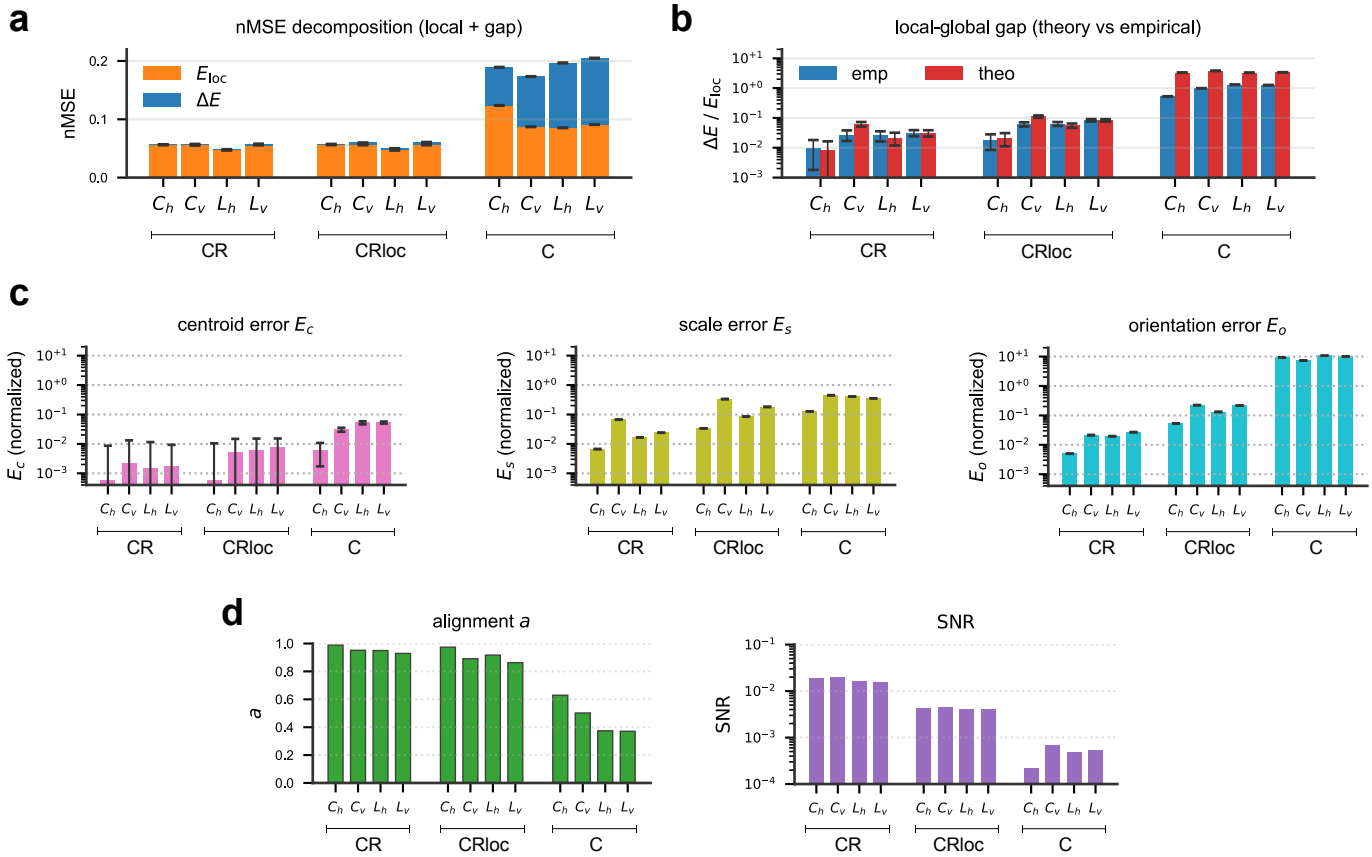


Figure 9. Regression performance in networks CR, CRloc and C. (**a**) Decomposition of the nMSE into local error $E_{\text{loc}}$ (orange) and local–global gap $\Delta E$ (blue) for regression of the four bounding box coordinates. Note that $\Delta E$ is very small and barely visible for networks CR and CRloc. (**b**) Measured relative local-global gap $\Delta E/E_{\text{loc}}$ (blue) compared to our theory prediction (red). (**c**) Decomposition of the relative gap $\Delta E/E_{\text{loc}}$ into the contributions from the centroid error $E_{\text{c}}$ (pink) scale error $E_{\text{s}}$ (olive) and orientation error $E_{\text{o}}$ (cyan). The three errors are normalized as explained in the caption of **Fig. 4**. (**d**) The two elements contributing to the orientation error: SNR, and alignment $a$ of the local feature-encoding directions.

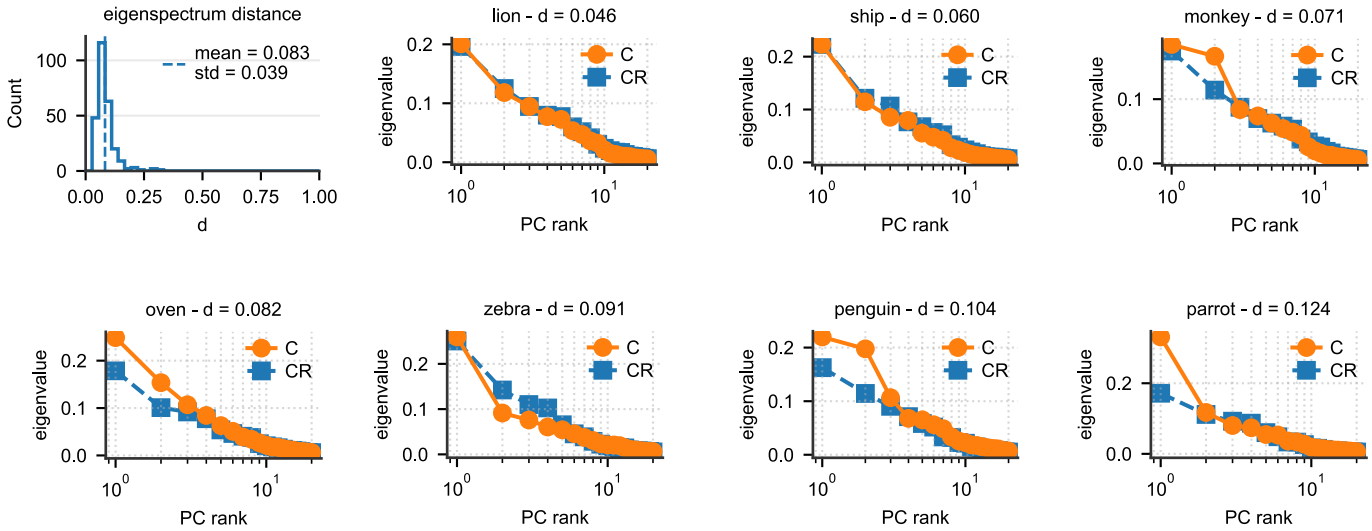


Figure 10. Change in manifold shapes between networks C and CR. Top left: histogram across categories of the spectral distance $d^{\mu}$. Other panels: PC spectra for selected categories in networks C and CR. Categories were sampled randomly, with sampling probability weighted by a Gaussian distribution with mean and standard deviation equal to those of the empirical distribution of $d^{\mu}$. For each category, the PC spectra are normalized by the category total variance, averaged between networks C and CR.

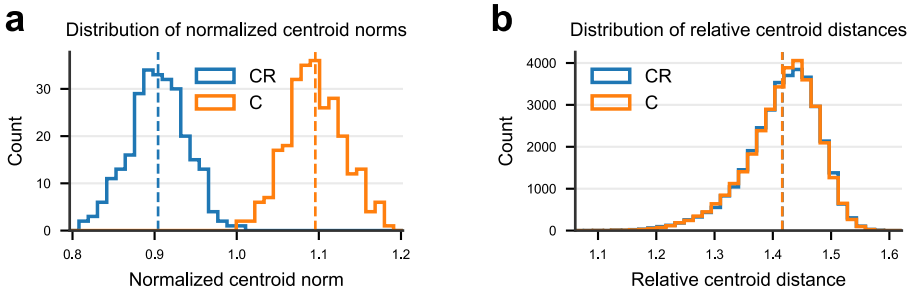


Figure 11. Manifold centroid norm and separation, in networks C (orange) and CR (blue). (a) Histogram of centroid norms $\|\bar{x}_\mu\|$. Centroid norms are normalized by their average across categories and across the two networks. (b) Histogram of relative centroid distances $\|\bar{x}_\mu - \bar{x}_\nu\|/\langle\|\bar{x}_\rho\|\rangle_\rho$. Vertical dashed lines indicate the corresponding averages.

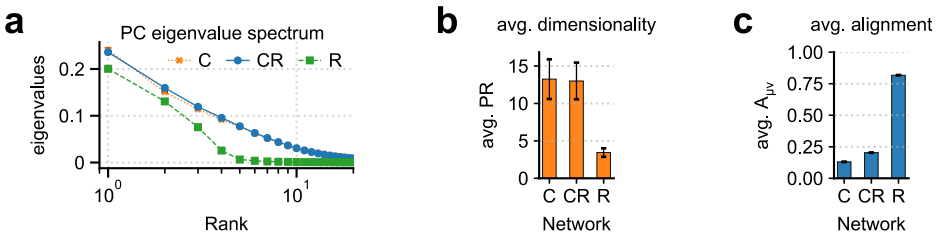


Figure 12. Differences in manifold geometry between networks C, CR and R. (**a**) PC eigenvalue spectrum (eigenvalue vs rank) for category manifolds in network C (orange), CR (blue), and R (green), averaged across categories (error bars are SEM). For each manifold $\mu$, the eigenvalues were normalized by the manifold total variance $\mathrm{Tr}\Sigma^\mu$, averaged across networks C, CR and R. (**b**) Average manifold dimensionality across categories. Error bars are standard deviation across categories. We define dimensionality by the participation ratio (PR). Denoting by $\lambda_d^\mu$, with $d = 1, \ldots, N$, the principal-component eigenvalues of the covariance matrix of the $\mu$-th category manifold, its participation ratio is defined as $\mathrm{PR}^\mu = \frac{\left(\sum_{d=1}^N \lambda_d^\mu\right)^2}{\sum_{d=1}^N \left(\lambda_d^\mu\right)^2}$. (**c**) Average of the off-diagonal entries of the global alignment matrix $A$.

manifold alignment increases drastically (panel c).

### D. Across-layer trends do not depend on the layer-freezing training schedule

Here we show that the regression-performance trends across layers reported in the main text (cf. **Fig. 6**) are virtually unchanged when network CR is trained end to end with all layers unfrozen from the very first epoch, rather than with the layer-unfreezing schedule described in the Methods. We call this network CR*. **Fig. 13** shows regression performance as a function of layer depth for networks CR, C and CR*. We still observe the trends reported in the main text: a sharp rise in regression performance emerges only from layer 4, as well as an appreciable distinction in regression performance between networks C and CR/CR*.

### E. Further details on the effect of subsampling neural units

*Effect of subsampling on the error sources*

We show here in **Fig. 14** what we mentioned in the main text: as the number of subsampled units $N \to 0$, the three error sources identified by our theory converge to similar values in networks C and CR, with the orientation error $E_\mathrm{o}$ and centroid error $E_\mathrm{c}$ becoming the dominant error contributions.

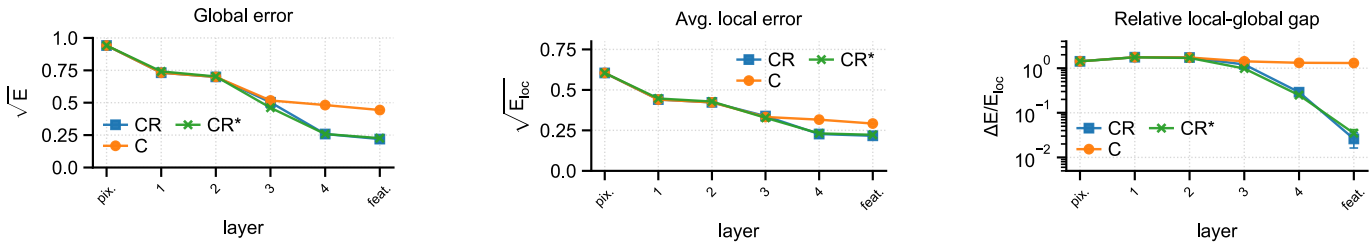


Figure 13. Regression performance along the visual hierarchy. Regression performance for the bounding-box parameter $L_{\mathrm{h}}$ using linear readouts from progressively deeper layers of networks C (orange circles, full line) and CR (blue squares, full line) and CR* (green cross, full line). At each layer, linear readouts were trained on activations randomly projected to 2048 dimensions, to match the dimensionality of the final feature layer. Note that global and local error are reported as root nMSE.

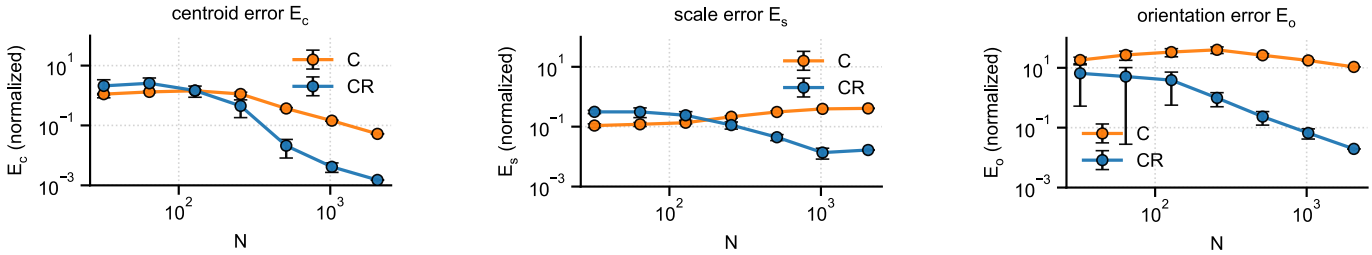


Figure 14. Subsampling effect on theory measures. Theory contributions $E_{\mathrm{c}}$, $E_{\mathrm{s}}$ and $E_{\mathrm{o}}$ to the relative local-global gap $\Delta E/E_{\mathrm{loc}}$ when regressing the bounding-box feature $L_v$ from networks C (orange) and CR (blue) on our dataset, as a function of the number $N$ of randomly subsampled units (out of 2048 feature-layer units).

*Effect of limited images per category on neuron-subsampling analysis*

In the main text we mention that, unfortunately, the same subsampling analysis shown in **Fig. 7b** cannot be carried out with our CNNs on the image set used in [4, 10], because that dataset contains only 400 images per category. When the number of neurons becomes comparable to, or exceeds, 400, the local regressors overfit, artificially inflating the estimated local error and thereby underestimating the local–global gap. This can be clearly seen in **Fig. 15:** for $N$ large enough, the local error inverts its decreasing trend and starts growing with $N$. As a consequence, the local-global gap gets largely underestimated, until it even becomes negative.

## F. Main text results for other features

Here we report the same kind of results presented in the main text, for features that were not shown in the main text.

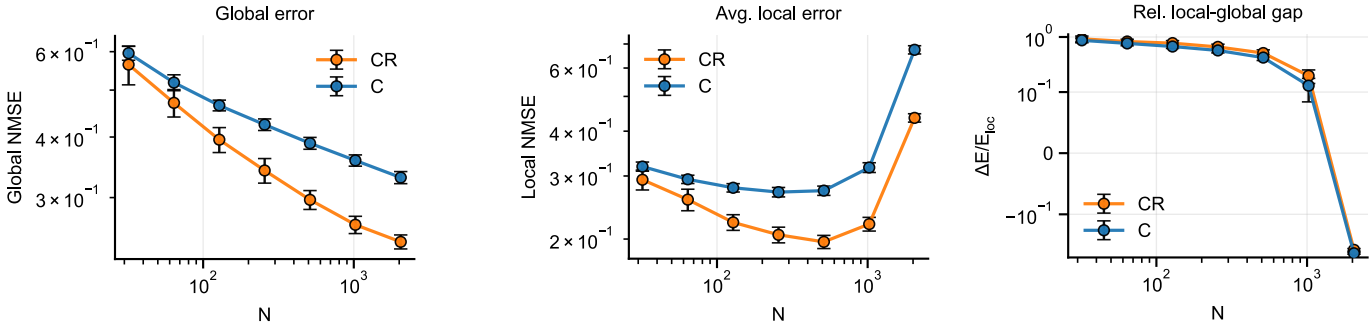


Figure 15. Global error, local error, and relative local–global gap in the regression of bounding-box parameter $L_h$, shown as a function of the number $N$ of subsampled units, for networks C and CR on the image set used in [4, 10].

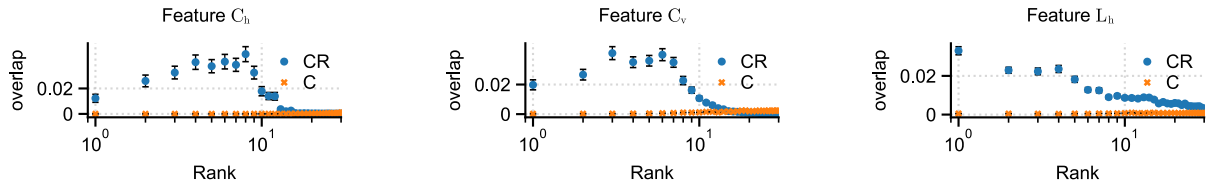


Figure 16. Feature-encoding direction vs PCs overlap. Squared cosine angle between the feature-encoding direction $\hat{w}^{\mu}$ and the PC directions of manifold $\mu$ (ordered by rank) for networks C (orange) and CR (blue), averaged across categories (error bars are SEM).

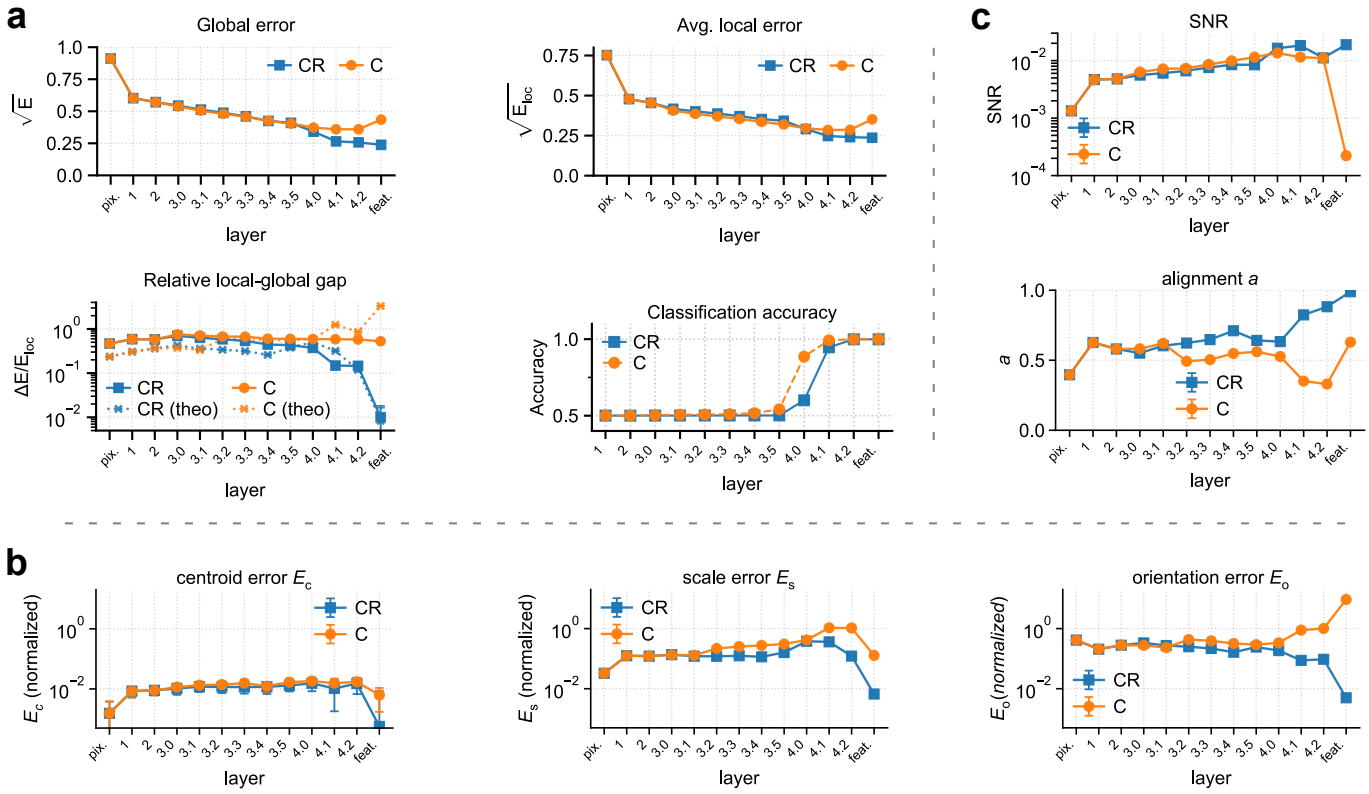


Figure 17. Decoding performance along the visual hierarchy. (**a**) Regression performance for the bounding-box parameter $C_h$ and classification performance, using linear readouts from progressively deeper layers of networks C (orange circles, full line) and CR (blue squares, full line). For regression, we report the global regression error, as well as its decomposition into local error $E_{\text{loc}}$ and relative local-global gap $\Delta E/E_{\text{loc}}$. For the gap, we also report our theory prediction (x markers, dotted lines). Note that global and local error are reported as root nMSE. (**b**) The three contributions to the relative local-global gap, as identified by our theory: centroid error $E_c$, orientation error $E_o$, scale error $E_s$. These quantities are normalized in the same way as explained in **Fig. 4**. (**c**) The two elements contributing to the orientation error: SNR, and alignment $a$ of the local feature-encoding directions.

*Feature-encoding direction vs PCs overlaps*

**Fig. 16** reports the overlap of the local feature-encoding direction with the top PCs of the corresponding category-manifold, averaged across categories. The trend is the same as discussed in the main text (cf. **Fig. 5**): in network CR compared to C, the feature-encoding directions are much more aligned with top manifold PCs.

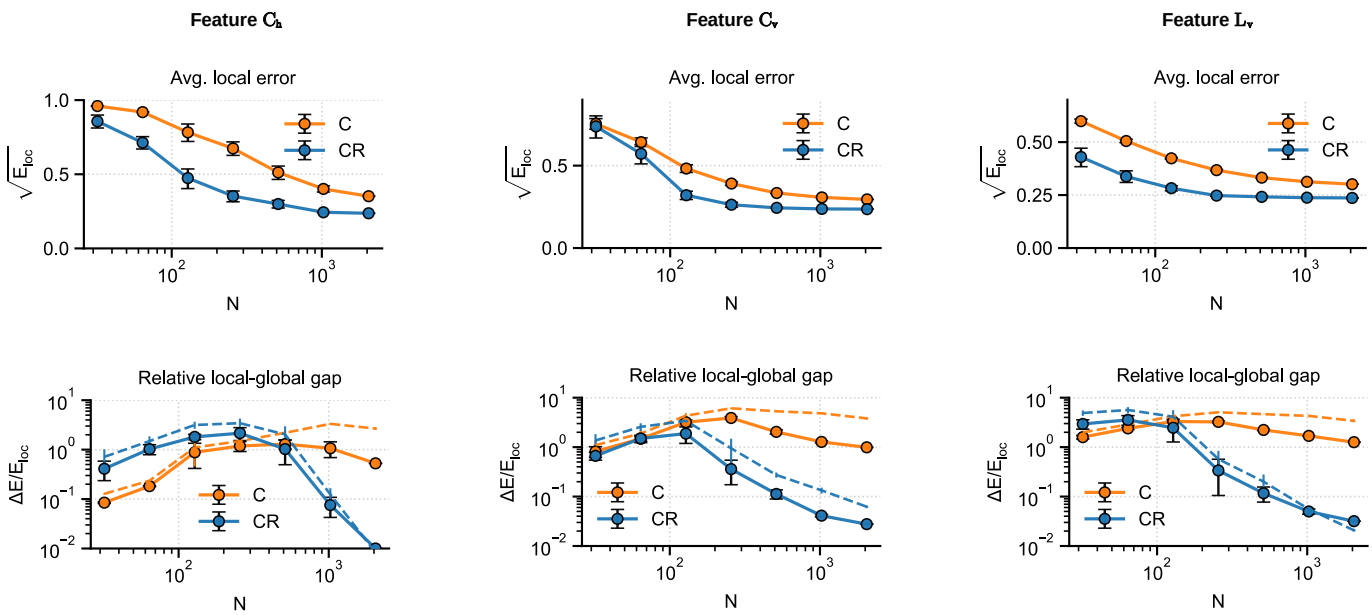


Figure 18. Effect of subsampling neural units. $E_{\mathrm{loc}}$ and $\Delta E/E_{\mathrm{loc}}$ for regressing the bounding-box features $(C_h, C_v, L_h)$ from networks C (orange) and CR (blue) on our dataset, as a function of the number $N$ of randomly subsampled units (out of 2048 feature-layer units). For the gap, we also report our theory prediction (x markers, dotted lines). Error bars denote SEM across cross-validation splits and random unit subsamples.

*Decoding performance along the visual hierarchy*

**Fig. 17** reports regression performance and associated theoretical measures for the bounding-box feature $C_{\mathrm{h}}$. We focus on $C_{\mathrm{h}}$ because $C_{\mathrm{v}}$ shows the same qualitative behavior; similarly, in the main text we report $L_{\mathrm{h}}$, while $L_{\mathrm{v}}$ follows an analogous trend. Overall, the across-layer behavior of center coordinates is similar to that of box lengths described in the main text (cf. **Fig. 6**), with one minor difference: in network C, the regression error for the center coordinates increases at the final transition from the last convolutional layer to the feature layer. A plausible explanation is that center coordinates rely more directly on the spatial arrangement of activity across the convolutional feature maps, because they encode *where* the object is located in the image. Global average pooling removes much of this spatial information by collapsing each feature map into a single scalar, and therefore can impair linear decoding of object position. By contrast, box lengths are invariant to object translation and thus may be less sensitive to the pooling operation. In network C, which is optimized only for classification, there is no pressure to recode positional information into a form that survives the final pooling step. In network CR, by contrast, explicit regression supervision can promote a reorganization of the last convolutional representation such that center-coordinate information remains linearly accessible after average pooling.

*Subsampling effects*

**Fig. 18** reports the effect of subsampling neural units on the two components of global regression error—local error $E_{\mathrm{loc}}$ and local-global gap $\Delta E$—for the bounding-box features not shown in the main text. The trend is the same as observed in the main text (cf. **Fig. 7b**): for low enough number of subsampled units $N$, the local-global gap $\Delta E$ in networks C and CR becomes indistinguishable. **Fig. 19** reports the analysis of regression performance on the images and corresponding neural recordings in [10], for the other bounding-box features not shown in the main text. The trend is the same as observed in the main text (**cf. Fig. 7c**): IT, V4, as well as networks C and CR subsampled to 168 units, matching the number recorded in IT, all show comparable values of $\Delta E/E_{\mathrm{loc}}$, as well as comparable contributions from each of the three error sources, with $E_{\mathrm{o}}$ and $E_{\mathrm{c}}$ emerging as the dominant terms.

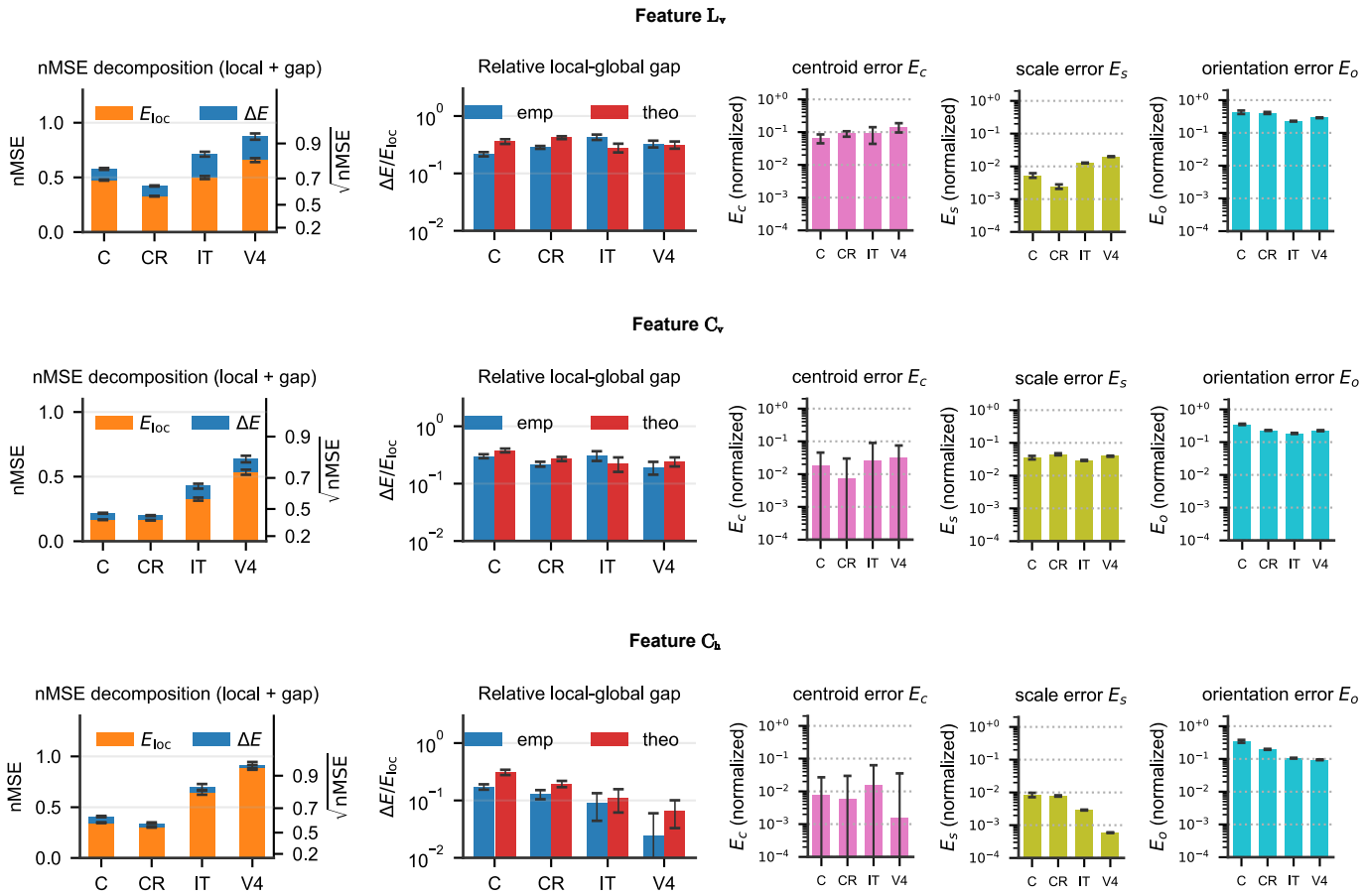


Figure 19. Analysis of regression performance on the images and corresponding neural recordings found in [10]. We report the analysis of regression from the IT and V4 units recorded in [10], as well as 168 randomly projected units from the best performing layer of networks C and CR. For features $C_{\mathrm{h}}$ and $C_{\mathrm{v}}$, the best performing layer is layer 4, for feature $L_v$ it is the feature layer. Error bars denote SEM across cross-validation splits and, where applicable, random unit subsamples.

## V. DETAILED THEORY DERIVATIONS

### A. Rewriting of the local-global gap

***Definitions.*** For convenience, we first recall the definitions of the global and local regression errors introduced in the Methods.

Without loss of generality, we consider globally centered data and labels, that is $\left\langle \langle x^\mu \rangle_\mu \right\rangle_C = 0$ and $\left\langle \langle y^\mu \rangle_\mu \right\rangle_C = 0$, where $\langle \ldots \rangle_\mu$ denotes averaging over all images within the $\mu$-th category, and $\langle \ldots \rangle_C = \mathrm{Avg}_\mu [\ldots]$ denotes averaging over categories.

The global regression MSE is

$$E = \min_W \left\langle \left\langle \left( W^\top x^\mu - y^\mu \right)^2 \right\rangle_\mu \right\rangle_C , \tag{19}$$

while the local regression MSE for category $\mu$ is

$$E^\mu = \min_{(W^\mu, b^\mu)} \left[ \left\langle \left( W^{\mu\top} x^\mu + b^\mu - y^\mu \right)^2 \right\rangle_\mu \right] \tag{20}$$

The average local error is $E_{\mathrm{loc}} = \langle E^\mu \rangle_C$, and the local-global error gap is defined as $\Delta E = E - E_{\mathrm{loc}}$. We denote by $\hat{y}^\mu = W_*^{\mu\top} x^\mu + b_*^\mu$ the prediction of the optimal local regressor for category $\mu$, where $(W_*^\mu, b_*^\mu) = \mathrm{argmin}_{(W^\mu, b^\mu)} \left[ \left\langle \left( W^{\mu\top} x^\mu + b^\mu - y^\mu \right)^2 \right\rangle_\mu \right]$.

***Result.*** We now show that $\Delta E$ can be rewritten as a global regression error on the locally linearized labels $\hat{y}^\mu$, that is

$$\Delta E = \min_W \left[ \left\langle \left\langle \left( W^\top x^\mu - \hat{y}^\mu \right)^2 \right\rangle_\mu \right\rangle_C \right] \tag{21}$$

***Derivation.*** For each category $\mu$, define the local residual

$$\epsilon^\mu = y^\mu - \hat{y}^\mu . \tag{22}$$

We can show that the residual satisfies

$$\langle \epsilon^\mu \rangle_\mu = 0, \qquad \langle x^\mu \epsilon^\mu \rangle_\mu = 0. \tag{23}$$

Indeed, the optimal local regressor is given by

$$W_*^\mu = K^{\mu +} d^\mu, \qquad b_*^\mu = \langle y^\mu \rangle_\mu - W_*^{\mu\top} \langle x^\mu \rangle_\mu , \tag{24}$$

where $K^\mu = \left\langle \delta x^\mu \delta x^{\mu\top} \right\rangle_\mu$, $d^\mu = \langle \delta x^\mu \delta y^\mu \rangle_\mu$, $\delta x^\mu = x^\mu - \langle x^\mu \rangle_\mu$, $\delta y^\mu = y^\mu - \langle y^\mu \rangle_\mu$, and $K^{\mu +}$ denotes the Moore-Penrose pseudoinverse. Therefore,

$$\epsilon^\mu = y^\mu - \hat{y}^\mu = \delta y^\mu - W_*^{\mu\top} \delta x^\mu . \tag{25}$$

Averaging immediately gives $\langle \epsilon^\mu \rangle_\mu = 0$. Moreover,

$$\langle x^\mu \epsilon^\mu \rangle_\mu = \langle \delta x^\mu \epsilon^\mu \rangle_\mu = d^\mu - K^\mu W_*^\mu = d^\mu - K^\mu K^{\mu +} d^\mu = 0, \tag{26}$$

which completes the derivation of Eq. (23).

Since $\hat{y}^\mu$ is itself a linear function of $x^\mu$, from Eq. (23) it follows that

$$\langle \hat{y}^\mu \epsilon^\mu \rangle_\mu = 0. \tag{27}$$

We can then write, for any global regressor $W$,

$$\begin{aligned} E\left(W\right) &= \left\langle\left\langle\left(W^{\top}x^{\mu}-y^{\mu}\right)^{2}\right\rangle_{\mu}\right\rangle_{C} = \left\langle\left\langle\left(W^{\top}x^{\mu}-\hat{y}^{\mu}-\epsilon^{\mu}\right)^{2}\right\rangle_{\mu}\right\rangle_{C} \\ &= \left\langle\left\langle\left(W^{\top}x^{\mu}-\hat{y}^{\mu}\right)^{2}\right\rangle_{\mu}\right\rangle_{C} - 2\left\langle\left\langle\left(W^{\top}x^{\mu}-\hat{y}^{\mu}\right)\epsilon^{\mu}\right\rangle_{\mu}\right\rangle_{C} + \left\langle\left\langle\epsilon^{\mu 2}\right\rangle_{\mu}\right\rangle_{C} \end{aligned} \tag{28}$$

where in the last step we expanded the square. The cross term vanishes, since

$$\left\langle\left(W^{\top}x^{\mu}-\hat{y}^{\mu}\right)\epsilon^{\mu}\right\rangle_{\mu} = W^{\top}\left\langle x^{\mu}\epsilon^{\mu}\right\rangle_{\mu} - \left\langle\hat{y}^{\mu}\epsilon^{\mu}\right\rangle_{\mu} = 0. \tag{29}$$

By construction, $\left\langle\epsilon^{\mu 2}\right\rangle_{\mu} = E^{\mu}$, and therefore $\left\langle\left\langle\epsilon^{\mu 2}\right\rangle_{\mu}\right\rangle_{C} = E_{\text{loc}}$. Minimizing $E\left(W\right)$ with respect to $W$, we thus obtain

$$E = \min_{W}\left\langle\left\langle\left(W^{\top}x^{\mu}-\hat{y}^{\mu}\right)^{2}\right\rangle_{\mu}\right\rangle_{C} + E_{\text{loc}}, \tag{30}$$

which immediately implies Eq. (21).

### B. Derivation of the local-global gap in terms of manifold geometry

***Generative model.*** For convenience, we first recall the definitions and assumptions of the manifold generative model introduced in the Methods. As explained in the Methods, we derive the local-global gap in the locally-centered case $\left\langle x^{\mu}\right\rangle_{\mu} = 0, \left\langle y^{\mu}\right\rangle_{\mu} = 0, \forall\mu$, and then quantify the effect of manifolds centroids by the centroid error. Without loss of generality, each point in manifold $\mu$ is parameterized as

$$x^{\mu} = \sum_{d=1}^{N} s_{d}^{\mu}u_{d}^{\mu}, \tag{31}$$

where $\{u_{d}^{\mu}\}_{d=1}^{N}$ is an orthonormal basis of $\mathbb{R}^{N}$, and $s_{d}^{\mu}$ is the coordinate along direction $d$. We choose

$$u_{1}^{\mu} = \frac{W_{*}^{\mu}}{\|W_{*}^{\mu}\|} \tag{32}$$

to be the feature-encoding direction, while the remaining directions $u_{d>1}^{\mu}$ are chosen to diagonalize the covariance in the subspace orthogonal to $u_{1}^{\mu}$. With this choice, the covariance of the manifold coordinates takes the form

$$\left\langle s^{\mu}s^{\mu\top}\right\rangle_{\mu} = \Sigma^{\mu} = \begin{pmatrix} \sigma_{1}^{\mu 2} & \Sigma_{12}^{\mu} & \cdots & \Sigma_{1N}^{\mu} \\ \Sigma_{21}^{\mu} & \sigma_{2}^{\mu 2} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ \Sigma_{N1}^{\mu} & 0 & 0 & \sigma_{N}^{\mu 2} \end{pmatrix}. \tag{33}$$

Note that local centering of the manifolds implies $\langle s_{d}^{\mu}\rangle = 0$. Let us also define

$$r^{\mu} = \|W_{*}^{\mu}\| \tag{34}$$

as the scale of the feature encoding, and note that

$$\hat{y}^{\mu} = r^{\mu}s_{1}^{\mu} \tag{35}$$

We assume the following across-manifold statistics:

1. The feature-encoding directions $u_1^\mu$ are isotropically distributed around a common mean direction

$$u_1^\mu = \sqrt{a} u_1 + \sqrt{1-a}\eta^\mu \,, \tag{36}$$

where $a \in [0,1]$ is the alignment strength, $u_1 \in \mathbb{R}^N$ is a unit-norm vector representing the common mean direction, and $\eta^\mu \in \mathbb{R}^N$ is a random i.i.d. Gaussian vector with $\langle \eta^\mu \rangle_C = 0$ and $\left\langle \eta^\mu \eta^{\mu\top} \right\rangle_C = \frac{1}{N}\mathbb{I}$. Under this definition, $u_1^\mu$ has unit norm on average, i.e. $\left\langle u_1^{\mu\top} u_1^\mu \right\rangle_C = 1$, with fluctuations that vanish as powers of $1/N$, which is sufficient for our derivations in the limit $N \to \infty$. For the coming derivations, it is useful to explicitly write the statistics of $u_1^\mu$ up to second order

$$\langle u_1^\mu \rangle_C = \sqrt{a} u_1 \tag{37}$$

$$\left\langle u_1^\mu u_1^{\mu\top} \right\rangle_C = a u_1 u_1^\top + (1-a)\frac{1}{N}\mathbb{I} \tag{38}$$

2. All manifold directions with $d > 1$ are mutually uncorrelated and isotropically distributed in neuronal space, conditioned on being orthogonal to $u_1^\mu$ (as by their definition as the elements of an orthonormal basis). Concretely,

$$\langle u_d^\mu \rangle_C = 0 \tag{39}$$

$$\left\langle u_d^\mu u_{d'}^{\mu\top} \right\rangle_C = \frac{1}{N}\delta_{d,d'}\left(\mathbb{I} - \left\langle u_1^\mu u_1^{\mu\top} \right\rangle_C\right) = \frac{1}{N}\delta_{d,d'}\left(\mathbb{I} - a\, u_1 u_1^\top\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{40}$$

3. Aside from the above correlations, all manifold parameters are otherwise uncorrelated.

***Result.*** We are going to show that, in the limit of infinitely many categories ($P \to \infty$) and up to $\mathcal{O}\left(N^{-1}\right)$ corrections,

$$\Delta E = \hat{\sigma}^2 \left[1 - \frac{1}{(1+E_\mathrm{s})(1+E_\mathrm{o})}\right] \tag{41}$$

$$E_\mathrm{s} = \frac{\mathrm{Var}_\mu\left[r^\mu\right]}{\mathrm{Avg}_\mu\left[r^\mu\right]^2} \tag{42}$$

$$E_\mathrm{o} = \frac{1-a}{N \cdot a \cdot \mathrm{SNR}} \tag{43}$$

where $\mathrm{SNR} = \frac{\langle \sigma_1^{\mu 2} \rangle_C}{\sum_{d=1}^N \langle \sigma_d^{\mu 2} \rangle_C}$, and $\hat{\sigma}^2 = \left\langle \left\langle \delta\hat{y}^{\mu 2} \right\rangle_\mu \right\rangle_C$ with $\delta\hat{y}^\mu = \hat{y}^\mu - \langle \hat{y}^\mu \rangle_\mu$.

***Derivation.*** Recall that

$$\Delta E = \min_W \left[\left\langle \left\langle \left(W^\top x^\mu - \hat{y}^\mu\right)^2 \right\rangle_\mu \right\rangle_C\right] \tag{44}$$

minimizing w.r.t. $W$ we obtain

$$\Delta E = \hat{\sigma}^2 - d^\top K^+ d \tag{45}$$

where

$$K = \left\langle \left\langle x^\mu x^{\mu\top} \right\rangle_\mu \right\rangle_C \tag{46}$$

$$d = \left\langle \left\langle x^\mu \hat{y}^\mu \right\rangle_\mu \right\rangle_C \tag{47}$$

$$\hat{\sigma}^2 = \left\langle \left\langle \hat{y}^{\mu 2} \right\rangle_\mu \right\rangle_C \tag{48}$$

Under our assumptions of manifold statistics, we have, neglecting $\mathcal{O}\left(N^{-1}\right)$ corrections,

$$K = \left\langle \sigma_1^{\mu 2} \right\rangle_C \left( a u_1 u_1^\top + \frac{1-a}{N} \mathbb{I} \right) + \frac{1}{N} \sum_{d>1} \left\langle \sigma_d^{\mu 2} \right\rangle_C \left( \mathbb{I} - a u_1 u_1^\top \right) \tag{49}$$

$$d = \left\langle r^\mu \right\rangle_C \left\langle \sigma_1^{\mu 2} \right\rangle_C \sqrt{a} u_1 \tag{50}$$

$$\hat{\sigma}^2 = \left\langle r^{\mu 2} \right\rangle_C \left\langle \sigma_1^{\mu 2} \right\rangle_C \tag{51}$$

We note that $d$ is an eigenvector of $K$, therefore we can readily derive

$$d^\top K^+ d = \frac{a \left\langle r^\mu \right\rangle_C^2 \left\langle \sigma_1^{\mu 2} \right\rangle_C^2}{a \left\langle \sigma_1^{\mu 2} \right\rangle_C + \frac{1-a}{N} \sum_d \left\langle \sigma_d^{\mu 2} \right\rangle_C} \tag{52}$$

Normalizing by $\hat{\sigma}^2 = \left\langle r^{\mu 2} \right\rangle_C \left\langle \sigma_1^{\mu 2} \right\rangle_C$ we obtain

$$\frac{d^\top K^+ d}{\hat{\sigma}^2} = \frac{\left\langle r^\mu \right\rangle_C^2}{\left\langle r^{\mu 2} \right\rangle_C} - \frac{1}{1 + \frac{1}{N} \frac{1-a}{a} \frac{\sum_d \left\langle \sigma_d^{\mu 2} \right\rangle_C}{\left\langle \sigma_1^{\mu 2} \right\rangle_C}} \tag{53}$$

We immediately identify the second term to be $\frac{1}{1+E_\mathrm{o}}$. Defining $\delta r^\mu = r^\mu - \left\langle r^\mu \right\rangle_C$, we can rewrite the first term as

$$\frac{\left\langle r^\mu \right\rangle_C^2}{\left\langle r^\mu \right\rangle_C^2 + \left\langle \delta r^{\mu 2} \right\rangle_C} = \frac{1}{1 + \frac{\left\langle \delta r^{\mu 2} \right\rangle_C}{\left\langle r^\mu \right\rangle_C^2}} = \frac{1}{1 + E_\mathrm{s}} \tag{54}$$

Therefore

$$\frac{d^\top K^+ d}{\hat{\sigma}^2} = \frac{1}{\left(1 + E_\mathrm{s}\right)\left(1 + E_\mathrm{o}\right)} \tag{55}$$

which plugged into Eq. (45) gives exactly Eq. (41), as desired.

### C. Derivation of the local-global gap in the regime of finite number of categories

Here we derive the local-global gap $\Delta E$ in the regime of finite category density, by studying the thermodynamic limit $P, N \to \infty$ at fixed $\alpha = \frac{P}{N} \in (0, \infty)$. Our previous result for infinitely many categories is recovered as the limiting case $\alpha \to \infty$.

In order to properly recover our result for infinitely many categories, we consider an asymptotic regime in which the total manifold variance is extensive, $\sum_d \left\langle \sigma_d^{\mu 2} \right\rangle_C = O(N)$. This is important to keep the orientation error

$$E_\mathrm{o} = \frac{1-a}{Na} \frac{\sum_d \left\langle \sigma_d^{\mu 2} \right\rangle_C}{\left\langle \sigma_1^{\mu 2} \right\rangle_C} \tag{56}$$

finite under the thermodynamic limit, as it would otherwise vanish proportionally to $\frac{1}{N} \to 0$. This choice is motivated by the empirical observation that the orientation and scale errors are of similar magnitude in the finite-dimensional regime relevant to our data. We emphasize that this assumption is used as a controlled scaling prescription for the asymptotic analysis. It should not be interpreted as a literal claim about the behavior of variance in arbitrarily large biological networks. Rather, it provides a convenient large-$N$ continuation of the finite-$N$ regime in which the orientation error remains finite and the $\alpha \to \infty$ solution consistently reproduces the previously derived infinite-category result.

### 1. Gibbs distribution

To derive the finite-size corrections to the local-global gap, we use the replica method. We start from the training objective

$$\Delta E(W) = \left\langle \left\langle \left(W^\top x^\mu - \hat{y}^\mu\right)^2 \right\rangle_\mu \right\rangle_C , \tag{57}$$

whose minimum is $\Delta E$. We introduce the Gibbs distribution

$$p(W) = \frac{1}{Z} \exp\left(-\beta \tilde{\Delta E}(W)\right), \tag{58}$$

with $\tilde{\Delta E}(W) = \frac{P}{2}\Delta E(W)$ and inverse temperature $\beta \in [0, \infty)$. The corresponding partition function is

$$Z = \int dW \, \exp\left(-\beta \tilde{\Delta E}(W)\right). \tag{59}$$

The average training error under this measure is

$$\langle \Delta E(W) \rangle_{p(W)} = -\frac{2}{P} \frac{\partial \ln Z}{\partial \beta}. \tag{60}$$

In the zero-temperature limit $\beta \to \infty$, the measure $p(W)$ concentrates on the minimizer $W_*$ of $\Delta E(W)$ and therefore

$$\lim_{\beta \to \infty} \langle \Delta E(W) \rangle_{p(W)} = \Delta E(W_*) = \Delta E. \tag{61}$$

The quenched disorder is given by the random manifold parameters entering $x^\mu$, namely the manifold directions. We assume that the relevant intensive quantities are self-averaging in the thermodynamic limit $N, P \to \infty$ with $P/N = \alpha \in [0, \infty)$ fixed. In particular, the typical value of the training error coincides with its disorder average, so that $\Delta E$ can be obtained from the disorder-averaged free energy computed by the replica method.

### 2. Auxiliary fields

We write explicitly the shift term appearing in $\Delta E(W)$ (Eq. (57)) as

$$W^\top x^\mu - \hat{y}^\mu = s_1^\mu \left(\sqrt{a}\, W^\top u_1 + \sqrt{1-a}\, W^\top \eta^\mu - r^\mu\right) + \sum_{d=2}^{N} s_d^\mu W^\top u_d^\mu. \tag{62}$$

This suggests introducing the order parameter

$$m = W^\top u_1, \tag{63}$$

and auxiliary fields

$$v_1^\mu = W^\top \eta^\mu, \qquad v_d^\mu = W^\top u_d^\mu, \quad d = 2, \ldots, N. \tag{64}$$

In terms of these fields, the energy becomes

$$\tilde{\Delta E}(v, m) = \frac{1}{2} \sum_{\mu=1}^{P} \left\langle \left( s_1^\mu \left(\sqrt{a}\, m + \sqrt{1-a}\, v_1^\mu - r^\mu\right) + \sum_{d=2}^{N} s_d^\mu v_d^\mu \right)^2 \right\rangle_\mu . \tag{65}$$

Carrying out the average over the manifold coordinates $s^\mu$, we obtain

$$\tilde{\Delta E}(v,m) = \frac{1}{2}\sum_{\mu=1}^{P}\Bigg[a\,\sigma_1^{\mu 2}m^2 + 2\sqrt{a(1-a)}\,\sigma_1^{\mu 2}mv_1^\mu + 2\sqrt{a}\sum_{d=2}^{N}\Sigma_{1d}^\mu m v_d^\mu - 2\sqrt{a}\,\sigma_1^{\mu 2}r^\mu m$$

$$+(1-a)\sigma_1^{\mu 2}(v_1^\mu)^2 + 2\sqrt{1-a}\sum_{d=2}^{N}\Sigma_{1d}^\mu v_1^\mu v_d^\mu - 2\sqrt{1-a}\,\sigma_1^{\mu 2}r^\mu v_1^\mu + \sum_{d=2}^{N}\sigma_d^{\mu 2}(v_d^\mu)^2$$

$$-2r^\mu\sum_{d=2}^{N}\Sigma_{1d}^\mu v_d^\mu + \sigma_1^{\mu 2}(r^\mu)^2\Bigg]. \tag{66}$$

We next enforce the definitions of the auxiliary fields through Dirac delta representations. Introducing the conjugate fields $\hat{m}$, $\hat{v}_1^\mu$, and $\hat{v}_d^\mu$, the partition function can be written as

$$Z \propto \int DW\,dm\,d\hat{m}\,Dv\,D\hat{v}\,\exp\Big(A(W,m,\hat{m},v,\hat{v}) - \beta\tilde{\Delta E}(v,m)\Big), \tag{67}$$

where

$$A(W,m,\hat{m},v,\hat{v}) = -i\hat{m}\,m + i\hat{m}\,W^\top u_1 - i\sum_{\mu=1}^{P}\hat{v}_1^\mu v_1^\mu + i\sum_{\mu=1}^{P}\hat{v}_1^\mu W^\top\eta^\mu - i\sum_{\mu=1}^{P}\sum_{d=2}^{N}\hat{v}_d^\mu v_d^\mu + i\sum_{\mu=1}^{P}\sum_{d=2}^{N}\hat{v}_d^\mu W^\top u_d^\mu. \tag{68}$$

Here and throughout, for any integration variable $z$ carrying one or more discrete indices, we denote by

$$Dz \equiv \prod_{\{\text{all indices of } z\}} dz \tag{69}$$

the corresponding integration measure. For example,

$$DW = \prod_{k=1}^{N} dW_k, \qquad Dv = \prod_{\mu=1}^{P}\prod_{d=1}^{N} dv_d^\mu, \qquad D\hat{v} = \prod_{\mu=1}^{P}\prod_{d=1}^{N} d\hat{v}_d^\mu. \tag{70}$$

### 3. *Replicas*

To compute the disorder-averaged free energy, we use the replica identity

$$\langle \ln Z\rangle_C = \lim_{n\to 0}\frac{\langle Z^n\rangle_C - 1}{n}, \tag{71}$$

where $\langle\cdot\rangle_C$ denotes the average over the quenched disorder, i.e. over the random manifold parameters.

Replicating the partition function, we obtain

$$\langle Z^n\rangle_C \propto \int DW\,Dm\,D\hat{m}\,Dv\,D\hat{v}\,\left\langle \exp\left(A(\{W^\alpha,m^\alpha,\hat{m}^\alpha,v^\alpha,\hat{v}^\alpha\}) - \beta\sum_{\alpha=1}^{n}\tilde{\Delta E}(v^\alpha,m^\alpha)\right)\right\rangle_C, \tag{72}$$

with

$$A = -i\sum_{\alpha=1}^{n}\hat{m}^{\alpha}m^{\alpha} + i\sum_{\alpha=1}^{n}\hat{m}^{\alpha}W^{\alpha\top}u_1 - i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\hat{v}_1^{\mu\alpha}v_1^{\mu\alpha} + i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\hat{v}_1^{\mu\alpha}W^{\alpha\top}\eta^{\mu}$$

$$-i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\hat{v}_d^{\mu\alpha}v_d^{\mu\alpha} + i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\hat{v}_d^{\mu\alpha}W^{\alpha\top}u_d^{\mu}, \tag{73}$$

and

$$\tilde{\Delta E}(v^{\alpha}, m^{\alpha}) = \frac{1}{2}\sum_{\mu=1}^{P}\Bigg[a\,\sigma_1^{\mu 2}(m^{\alpha})^2 + 2\sqrt{a(1-a)}\,\sigma_1^{\mu 2}m^{\alpha}v_1^{\mu\alpha} + 2\sqrt{a}\sum_{d=2}^{N}\Sigma_{1d}^{\mu}m^{\alpha}v_d^{\mu\alpha} - 2\sqrt{a}\,\sigma_1^{\mu 2}r^{\mu}m^{\alpha}$$

$$+(1-a)\sigma_1^{\mu 2}(v_1^{\mu\alpha})^2 + 2\sqrt{1-a}\sum_{d=2}^{N}\Sigma_{1d}^{\mu}v_1^{\mu\alpha}v_d^{\mu\alpha} - 2\sqrt{1-a}\,\sigma_1^{\mu 2}r^{\mu}v_1^{\mu\alpha} + \sum_{d=2}^{N}\sigma_d^{\mu 2}(v_d^{\mu\alpha})^2$$

$$-2r^{\mu}\sum_{d=2}^{N}\Sigma_{1d}^{\mu}v_d^{\mu\alpha} + \sigma_1^{\mu 2}(r^{\mu})^2\Bigg]. \tag{74}$$

### *4. Average over the quenched disorder*

We now average over the quenched disorder, i.e. over the random manifold directions. Concretely, we average over the random vectors $\eta^{\mu}$ and $u_d^{\mu}$ for $d > 1$. This boils down to the calculation of standard Gaussian integrals, at the end of which we obtain

$$\langle Z^n\rangle_C \propto \int DQ\,D\hat{Q}\,DW\,Dm\,D\hat{m}\,Dv\,D\hat{v}\,\exp\left(A(Q,W,m,\hat{m},v,\hat{v}) + B(Q,\hat{Q},W) - \beta\sum_{\alpha=1}^{n}\tilde{\Delta E}(v^{\alpha},m^{\alpha})\right), \tag{75}$$

where

$$B(Q,\hat{Q},W) = \frac{N}{2}\mathrm{Tr}\,Q^{\top}\hat{Q} - \frac{1}{2}\sum_{\alpha,\beta=1}^{n}\hat{Q}^{\alpha\beta}W^{\alpha\top}W^{\beta}, \tag{76}$$

$$A = -i\sum_{\alpha=1}^{n}\hat{m}^{\alpha}m^{\alpha} + i\sum_{\alpha=1}^{n}\hat{m}^{\alpha}W^{\alpha\top}u_1 - i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\hat{v}_1^{\mu\alpha}v_1^{\mu\alpha} - i\sum_{\alpha=1}^{n}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\hat{v}_d^{\mu\alpha}v_d^{\mu\alpha}$$

$$-\frac{1-a}{2}\sum_{\mu=1}^{P}\sum_{\alpha,\beta=1}^{n}\hat{v}_1^{\mu\alpha}Q^{\alpha\beta}\hat{v}_1^{\mu\beta} - \frac{1}{2}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\sum_{\alpha,\beta=1}^{n}\hat{v}_d^{\mu\alpha}\left(Q^{\alpha\beta} - \frac{a}{N}m^{\alpha}m^{\beta}\right)\hat{v}_d^{\mu\beta}, \tag{77}$$

and, by means of Dirac delta representations, we introduced the overlap order parameters

$$Q^{\alpha\beta} = \frac{1}{N}W^{\alpha\top}W^{\beta}, \tag{78}$$

and their conjugates $\hat{Q}^{\alpha\beta}$.

### 5. Gaussian integrals in the auxiliary fields

We now perform, in this order, the Gaussian integrals over $\hat{v}$, $W$, and $\hat{m}$. Neglecting additive constants independent of $\beta$, and dropping the subleading $\mathcal{O}(1)$ log-determinant contribution generated by the $\hat{m}$ integral, we obtain

$$\langle Z^n \rangle_C \propto \int DQ\, D\hat{Q}\, Dm\, Dv\, \exp\left( A(Q,\hat{Q},m,v) + B(Q,\hat{Q},m) - \beta \sum_{\alpha=1}^{n} \tilde{\Delta E}(v^\alpha, m^\alpha) \right), \tag{79}$$

with

$$A = -\frac{1}{2}\sum_{\alpha,\beta=1}^{n} m^\alpha \hat{Q}^{\alpha\beta} m^\beta - \frac{1}{2}(1-a)^{-1} \sum_{\mu=1}^{P}\sum_{\alpha,\beta=1}^{n} v_1^{\mu\alpha} \left(Q^{-1}\right)^{\alpha\beta} v_1^{\mu\beta}$$

$$-\frac{1}{2}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\sum_{\alpha,\beta=1}^{n} v_d^{\mu\alpha} \left[ Q^{-1} + \frac{a}{N}\frac{Q^{-1}mm^\top Q^{-1}}{1-\frac{a}{N}m^\top Q^{-1} m} \right]^{\alpha\beta} v_d^{\mu\beta}, \tag{80}$$

and

$$B = \frac{N}{2}\mathrm{Tr}\, Q^\top \hat{Q} - \frac{1}{2}PN \ln\det(Q) - \frac{1}{2}(N-1)P\ln\left(1 - \frac{a}{N} m^\top Q^{-1} m\right) - \frac{1}{2}N\ln\det(\hat{Q}), \tag{81}$$

where we used the identities

$$\left(Q - \frac{a}{N}mm^\top\right)^{-1} = Q^{-1} + \frac{a}{N}\frac{Q^{-1}mm^\top Q^{-1}}{1-\frac{a}{N}m^\top Q^{-1}m}, \tag{82}$$

and

$$\det\left(Q - \frac{a}{N}mm^\top\right) = \left(1 - \frac{a}{N}m^\top Q^{-1} m\right)\det(Q). \tag{83}$$

### 6. Replica symmetric ansatz

Assuming a replica-symmetric saddle point, we make the ansatz

$$Q = (q_1 - q_0)\,\mathbb{I} + q_0 \mathbf{1}, \qquad \hat{Q} = (\hat{q}_1 - \hat{q}_0)\,\mathbb{I} + \hat{q}_0\mathbf{1}, \qquad m^\alpha = m, \tag{84}$$

Using standard identities for replica-symmetric matrices, we then obtain

$$A(Q,\hat{Q},m,v) = -\frac{1}{2}n\,(\hat{q}_1 - \hat{q}_0)\,m^2 - \frac{1}{2}\frac{(1-a)^{-1}}{q_1 - q_0}\sum_{\mu=1}^{P}\sum_{\alpha=1}^{n}(v_1^{\mu\alpha})^2 + \frac{1}{2}\frac{(1-a)^{-1}q_0}{(q_1-q_0)^2}\sum_{\mu=1}^{P}\left(\sum_{\alpha=1}^{n} v_1^{\mu\alpha}\right)^2$$

$$-\frac{1}{2}\frac{1}{q_1-q_0}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\sum_{\alpha=1}^{n}(v_d^{\mu\alpha})^2 + \frac{1}{2}\frac{q_0}{(q_1-q_0)^2}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\left(\sum_{\alpha=1}^{n} v_d^{\mu\alpha}\right)^2 - \frac{1}{2}\frac{a}{N}\frac{m^2}{(q_1-q_0)^2}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\left(\sum_{\alpha=1}^{n} v_d^{\mu\alpha}\right)^2, \tag{85}$$

and

$$B(Q,\hat{Q},m) = \frac{Nn}{2}\left[(q_1-q_0)(\hat{q}_1-\hat{q}_0)+(q_1-q_0)\hat{q}_0+(\hat{q}_1-\hat{q}_0)q_0\right]$$

$$-\frac{1}{2}PN\,n\left[\ln(q_1-q_0)+\frac{q_0}{q_1-q_0}\right]-\frac{1}{2}N\,n\left[\ln(\hat{q}_1-\hat{q}_0)+\frac{\hat{q}_0}{\hat{q}_1-\hat{q}_0}\right]+\frac{1}{2}(N-1)P\frac{a}{N}\,n\,\frac{m^2}{q_1-q_0}. \tag{86}$$

where we retained only terms up to linear order in $n$.

### 7. *Remaining Gaussian integrals*

Note that Eq. (85) contains replica-coupling terms of the form $(\sum_\alpha v_1^{\mu\alpha})^2$ and $(\sum_\alpha v_d^{\mu\alpha})^2$. We decouple these terms by standard Hubbard-Stratonovich transformations, introducing Gaussian auxiliary fields $t_1^\mu$ and $t_d^\mu$ for $d=2,\dots,N$. At fixed $t$, the integrals over the replicated variables factorize over $\alpha$. Since the replica limit $n\to 0$ is taken at the end, we retain only the linear term in $n$, using

$$\int Dt^\mu\left[Z_\mu(t^\mu)\right]^n = 1+n\int Dt^\mu \ln Z_\mu(t^\mu)+\mathcal{O}(n^2), \tag{87}$$

where $Dt^\mu$ denotes the standard Gaussian measure over the auxiliary fields associated with manifold $\mu$. We then perform sequentially the Gaussian integrals over $v_d^\mu$ for $d=2,\dots,N$, over $v_1^\mu$, and finally over the auxiliary fields $t^\mu$. This yields

$$\langle Z^n\rangle_C \propto \int dm\,dq_0\,d\delta q\,d\hat{q}_0\,d\delta\hat{q}\,\exp\left(\frac{Pn}{2}S_{\rm RS}(m,q_0,\delta q,\hat{q}_0,\delta\hat{q})\right), \tag{88}$$

where

$$\frac{2}{Pn}S_{\rm RS}(m,q_0,\delta q,\hat{q}_0,\delta\hat{q}) = -\beta\,\frac{1}{P}\sum_{\mu=1}^{P}\sigma_1^{\mu 2}r^{\mu 2}-\beta a\left\langle\sigma_1^{\mu 2}\right\rangle_C m^2+2\beta\sqrt{a}\,\frac{1}{P}\sum_{\mu=1}^{P}r^\mu\sigma_1^{\mu 2}\,m-\frac{\beta}{P}\delta\hat{q}\,m^2$$

$$+\alpha^{-1}\left[\delta q\,\delta\hat{q}+\beta^{-1}\delta q\,\hat{q}_0+\beta\,\delta\hat{q}\,q_0\right]-\alpha^{-1}\left[\ln(\beta\delta\hat{q})+\frac{\hat{q}_0}{\beta\delta\hat{q}}\right]$$

$$-N\left[\ln(\beta^{-1}\delta q)+\frac{\beta q_0}{\delta q}\right]+\frac{N-1}{N}\frac{\beta a\,m^2}{\delta q}-\frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\ln\beta U_d^\mu+\frac{1}{P}\beta\sum_{\mu=1}^{P}\sum_{d=2}^{N}S_d^\mu\,(\delta_m^\mu)^2+\frac{1}{P}\beta\sum_{\mu=1}^{P}\sum_{d=2}^{N}(U_d^\mu)^{-1}V_2$$

$$-\frac{1}{P}\sum_{\mu=1}^{P}\ln\left(\beta U_1^\mu-\beta\sum_{d=2}^{N}S_d^\mu\right)+\frac{1}{P}\beta\sum_{\mu=1}^{P}\left(U_1^\mu-\sum_{d=2}^{N}S_d^\mu\right)^{-1}\left(\sigma_1^{\mu 2}-\sum_{d=2}^{N}S_d^\mu\right)^2(\delta_m^\mu)^2$$

$$+\frac{1}{P}\beta\sum_{\mu=1}^{P}\left(U_1^\mu-\sum_{d=2}^{N}S_d^\mu\right)^{-1}V_1+\frac{1}{P}\beta\sum_{\mu=1}^{P}\left(U_1^\mu-\sum_{d=2}^{N}S_d^\mu\right)^{-1}V_2\sum_{d=2}^{N}\left((U_d^\mu)^{-1}\Sigma_{1d}^\mu\right)^2, \tag{89}$$

with

$$U_1^\mu=\sigma_1^{\mu 2}+\frac{(1-a)^{-1}}{\delta q},\qquad U_d^\mu=\sigma_d^{\mu 2}+\frac{1}{\delta q},\qquad d=2,\dots,N, \tag{90}$$

$$V_1 = \frac{(1-a)^{-1} q_0}{\delta q^2}, \qquad V_2 = \frac{q_0}{\delta q^2} - \frac{a}{N}\frac{m^2}{\delta q^2}, \tag{91}$$

$$S_d^\mu = (U_d^\mu)^{-1} (\Sigma_{1d}^\mu)^2, \qquad \delta_m^\mu = r^\mu - \sqrt{a}\, m. \tag{92}$$

Note that we performed the change of variables

$$\beta^{-1} \delta q = q_1 - q_0 \tag{93}$$

$$\beta \delta \hat{q} = \hat{q}_1 - \hat{q}_0 \tag{94}$$

while keeping $q_0$ and $\hat{q}_0$ unchanged.

### 8. *Saddle point*

In the limit $N, P \to \infty$, with $\frac{P}{N} = \alpha \in [0, \infty]$, the integral in (88) can be solved via saddle point. We note that $S_{\mathrm{RS}}$ is linear in $q_0$ and $\hat{q}_0$, so we do not need to compute the saddle point solution for $q_0$ and $\hat{q}_0$ explicitly, as any term of $S_{\mathrm{RS}}$ that is linear in these variables vanishes under the saddle point condition. The saddle point equation w.r.t. $\hat{q}_0$ gives

$$\delta q = \delta \hat{q}^{-1} \tag{95}$$

The saddle point equation w.r.t. $q_0$ gives the self-consistent equation for $\delta q$

$$\boxed{\delta \hat{q} \frac{1}{N} \frac{1}{P} \sum_\mu \left\{ \sum_{d=2}^N (U_d^\mu)^{-1} + (1-a)^{-1} \left[ U_1^\mu - \sum_{d=2}^N S_d^\mu \right]^{-1} + \left[ U_1^\mu - \sum_{d=2}^N S_d^\mu \right]^{-1} \sum_{d=2}^N \left( (U_d^\mu)^{-1} \Sigma_{1d}^\mu \right)^2 \right\} = 1 - \frac{1}{N} \alpha^{-1}} \tag{96}$$

### 9. *Order parameter in the $\alpha \to \infty$ limit*

It is useful to study the order parameter in the limit $\alpha \to \infty$. We rewrite the self-consistent equation Eq. (96)

$$\frac{1}{N} \frac{1}{P} \sum_\mu \left\{ \sum_{d=2}^N \frac{\delta \hat{q}}{\sigma_d^{\mu 2} + \delta \hat{q}} + \frac{(1-a)^{-1} \delta \hat{q}}{\sigma_1^{\mu 2} + (1-a)^{-1} \delta \hat{q} - \sum_{d>1} \frac{\left(\Sigma_{1d}^\mu\right)^2}{\sigma_d^{\mu 2} + \delta \hat{q}}} \right.$$
$$\left. + \frac{\delta \hat{q}}{\sigma_1^{\mu 2} + (1-a)^{-1} \delta \hat{q} - \sum_{d>1} \frac{\left(\Sigma_{1d}^\mu\right)^2}{\sigma_d^{\mu 2} + \delta \hat{q}}} \sum_{d=2}^D \left( \frac{\Sigma_{1d}^\mu}{\sigma_d^{\mu 2} + \delta \hat{q}} \right)^2 \right\} = 1 - \frac{1}{N} \alpha^{-1} \tag{97}$$

It is readily verified that $\delta \hat{q} \overset{\alpha \to \infty}{\to} \infty$. Indeed, taking this limit we get the identity

$$\frac{1}{P} \frac{1}{N} \sum_\mu \left\{ \sum_{d=2}^N 1 + 1 + \mathcal{O}\left( \frac{1}{\delta \hat{q}} \right) \right\} = 1 + \mathcal{O}\left( \alpha^{-1} \right) \tag{98}$$

To determine the asymptotic behavior of $\delta \hat{q}$, we expand the l.h.s of Eq. (97) for $\delta \hat{q} \to \infty$. Term by term, we have

$$\frac{1}{\frac{\sigma_d^{\mu 2}}{\delta \hat{q}} + 1} \sim 1 - \frac{1}{\delta \hat{q}} \sigma_d^{\mu 2} \tag{99}$$

$$\frac{1}{1+\frac{\sigma_1^{\mu 2}}{(1-a)^{-1}\delta\hat{q}}+\mathcal{O}\left(\frac{1}{\delta\hat{q}^2}\right)} \sim 1-\frac{1}{\delta\hat{q}}\left(1-a\right)\sigma_1^{\mu 2} \tag{100}$$

while the last term is of subleading order $\mathcal{O}\left(\frac{1}{\delta\hat{q}^2}\right)$. Plugging this back into Eq. (97) we get

$$\boxed{\delta\hat{q} \overset{\alpha\to\infty}{\sim} \alpha\left[\left(1-a\right)\left\langle\sigma_1^{\mu 2}\right\rangle_C+\sum_{d>1}\left\langle\sigma_d^{\mu 2}\right\rangle_C\right]} \tag{101}$$

Note that, at least for large $\alpha$, the order parameter is proportional to $\sum_{d>1}\left\langle\sigma_d^{\mu 2}\right\rangle_C=\mathcal{O}\left(N\right)$. This will be useful to determine which terms can be neglected in the final expression for $\Delta E$, because of subleading order in $N$, and which should be kept.

### *10. Local-global gap at the saddle point*

At the saddle point, the local-global gap is given by

$$\Delta E=\min_m\left[\Delta E\left(m\right)\right] \tag{102}$$

where

$$\Delta E\left(m\right)=-\frac{1}{n}\frac{2}{P}\frac{1}{\beta}S_{\mathrm{RS}}=\frac{1}{P}\sum_{\mu=1}^{P}\sigma_1^{\mu 2}r^{\mu 2}+a\left\langle\sigma_1^{\mu 2}\right\rangle_C m^2-2\sqrt{a}m\frac{1}{P}\sum_{\mu=1}^{P}r^{\mu}\sigma_1^{\mu 2}+\frac{1}{P}\delta\hat{q}\,m^2 \tag{103}$$

$$-\left(N-1\right)\frac{1}{N}\delta\hat{q}\,a\,m^2-\frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N}S_d^{\mu}\left(r^{\mu}-\sqrt{a}\,m\right)^2+\frac{1}{N}\delta\hat{q}^2am^2\frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N}(U_d^{\mu})^{-1} \tag{104}$$

$$-\frac{1}{P}\sum_{\mu=1}^{P}\left(U_1^{\mu}-\sum_{d=2}^{N}S_d^{\mu}\right)^{-1}\left(\sigma_1^{\mu 2}-\sum_{d=2}^{N}S_d^{\mu}\right)^2\left(r^{\mu}-\sqrt{a}\,m\right)^2 \tag{105}$$

$\delta\hat{q}$ is determined by the self-consistent equation Eq. (96), and we have kept only those terms that are of leading order in both $\beta$ as $\beta\to\infty$, and in $N$ or $P$, in the limit $N,P\to\infty$, with $\frac{P}{N}=\alpha\in[0,\infty]$. Note that some terms in $\Delta E\left(m\right)$, such as $\frac{1}{P}\delta\hat{q}\,m^2$, may appear to be of subleading order in this limit, but are in fact to be kept because, at least for large $\alpha$, the order parameter $\delta\hat{q}$ is of $\mathcal{O}\left(N\right)$.

Observing that Eq. (105) is quadratic in $m$, it's minimization is readily obtained as

$$\Delta E = -\frac{k_1^2}{4k_2} + k_0 \tag{106}$$

with

$$k_2 = a\left\langle \sigma_1^{\mu 2}\right\rangle_C + \frac{1}{P}\delta\hat{q} - \frac{(N-1)}{N}a\delta\hat{q} - a\frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N} S_d^\mu + \frac{a}{N}\left(\delta\hat{q}\right)^2 \frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N}\left(U_d^\mu\right)^{-1}$$
$$- a\frac{1}{P}\sum_\mu \left(U_1^\mu - \sum_{d=2}^{N} S_d^\mu\right)^{-1}\left(\sigma_1^{\mu 2} - \sum_{d=2}^{N} S_d^\mu\right)^2 \tag{107}$$

$$k_1 = -2\sqrt{a}m\,\frac{1}{P}\sum_{\mu=1}^{P} r^\mu \sigma_1^{\mu 2} + 2\sqrt{a}\frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{D} S_d^\mu r^\mu + 2\sqrt{a}\frac{1}{P}\sum_\mu \left(U_1^\mu - \sum_{d=2}^{N} S_d^\mu\right)^{-1}\left(\sigma_1^{\mu 2} - \sum_{d=2}^{N} S_d^\mu\right)^2 r^\mu \tag{108}$$

$$k_0 = \frac{1}{P}\sum_{\mu=1}^{P}\sigma_1^{\mu 2} r^{\mu 2} - \frac{1}{P}\sum_{\mu=1}^{P}\sum_{d=2}^{N} S_d^\mu r^{\mu 2} - \frac{1}{P}\sum_{\mu=1}^{P}\left(U_1^\mu - \sum_{d=2}^{N} S_d^\mu\right)^{-1}\left(\sigma_1^{\mu 2} - \sum_{d=2}^{N} S_d^\mu\right)^2 r^{\mu 2} \tag{109}$$

### *11. Local-global gap in the $\alpha \to \infty$ limit*

It is instructive to check that our expression for the local-global gap in the case of infinite categories, i.e. $\alpha \to \infty$, is recovered. As $\delta\hat{q} \to \infty$, most terms in Eqs. (107-109) that depend on $\delta\hat{q}$ vanish, except for the following three terms in Eqs. (107)

$$+\frac{1}{P}\delta\hat{q} \sim \frac{\alpha}{P}\left[(1-a)\left\langle \sigma_1^{\mu 2}\right\rangle_C + \sum_{d>1}\left\langle \sigma_d^{\mu 2}\right\rangle_C\right] = \frac{1}{N}\left[(1-a)\left\langle \sigma_1^{\mu 2}\right\rangle_C + \sum_{d>1}\left\langle \sigma_d^{\mu 2}\right\rangle_C\right] \tag{110}$$

$$-\frac{(N-1)}{N}a\delta\hat{q} \tag{111}$$

$$+\frac{a}{N}\left(\delta\hat{q}\right)^2\frac{1}{P}\sum_\mu\sum_{d>1}\left(U_d^\mu\right)^{-1} = \frac{a}{N}\frac{1}{P}\sum_\mu\sum_{d>1}\delta\hat{q}\left[\frac{1}{1+\frac{\sigma_d^{\mu 2}}{\delta\hat{q}}}\right] \sim \frac{a}{N}\left(N-1\right)\delta\hat{q} - \frac{a}{N}\sum_{d>1}\left\langle \sigma_d^{\mu 2}\right\rangle_C \tag{112}$$

The second term, which is diverging, cancels with part of the third, and we have (assuming $r^\mu$ and $\sigma_1^\mu$ to be uncorrelated as in our infinite-$P$ theory)

$$k_2 = a\left\langle \sigma_1^{\mu 2}\right\rangle_C + (1-a)\frac{1}{N}\sum_{d>1}\left\langle \sigma_d^{\mu 2}\right\rangle_C \tag{113}$$

$$k_1 = -2\sqrt{a}\,\left\langle r^\mu\right\rangle_C \left\langle \sigma_1^{\mu 2}\right\rangle_C \tag{114}$$

$$k_0 = \left\langle \sigma_1^{\mu 2}\right\rangle_C \left\langle r^{\mu 2}\right\rangle_C \tag{115}$$

which plugged into Eq. (106) give exactly our prediction for $\Delta E$ in the limit $\alpha \to \infty$, Eq. (41).