A Comparative Study of Label-free Representation Quality Metrics in Deep Learning

Anonymous authors
Paper under double-blind review

Abstract

We present a comparative study of label-free metrics for assessing the quality of representations in deep neural networks to understand their reliability under a wide variety of configurations. We group existing label-free metrics into three families based on their construction and analytically establish connections between metrics within the same family. We then characterise the sensitivity of spectral metrics through controlled synthetic experiments. Finally, all label-free metrics are evaluated against downstream task accuracy across a diverse set of 260 vision models on CIFAR-10, CIFAR-100, and ImageNet-1K, stratifying results by architecture class, training objective, and representation dimension. We find that intrinsic dimensionality (ID) is the most reliable predictor in all settings, while the reliability of other metrics is strongly moderated by architecture class and training objective. Our results provide a clearer understanding of what label-free representation quality metrics measure, when they are reliable, and how to interpret them in practice.

1 Introduction

The popularity of deep learning research has largely been driven by supervised learning, where models are trained to perform a sequence of transformations from input data to corresponding labels using large annotated datasets (Krizhevsky et al., 2012; LeCun et al., 2002). The intermediate representations learnt by these models were shown to retain sufficient structure useful for new datasets and tasks (Yosinski et al., 2014; Donahue et al., 2014). The transferability of these models to other datasets inspired research on building general-purpose feature extractors that do not depend on task-specific supervision. Self-supervised learning (SSL) has emerged as the dominant paradigm for this purpose, with methods based on contrastive learning (Chen et al., 2020; He et al., 2020; Bardes et al., 2022), self-distillation (Caron et al., 2021), and masked reconstruction (He et al., 2022), demonstrating that models trained without labels can produce representations that match or exceed those of their supervised counterparts across a wide range of downstream tasks.

With the increasing availability of such models through public repositories such as HuggingFace (Wolf et al., 2020), selecting a model to use as a feature extractor for a given task has become a challenge. In practice, the suitability of a model for a task is evaluated using simple algorithms, including k-nearest neighbours (Wu et al., 2018), linear probing (Kornblith et al., 2019), or full fine-tuning (Kornblith et al., 2019). However, as the number of publicly available models continues to grow, an exhaustive search over all candidates becomes computationally expensive. A parallel line of work has sought to identify the characteristics that a good representation space should exhibit for any downstream task. These studies examine the covariance structure (Agrawal et al., 2022; Garrido et al., 2023) and intrinsic dimensionality (Ansuini et al., 2019) of the representation manifold. Crucially, the proposed metrics that capture these properties are computed solely from the features themselves. This has motivated their use as label-free metrics that require no training, offering a computationally efficient alternative for model selection.

The increasing usage and number of label-free metrics raise important questions about when to use which metric as well as how reliable they are across different architectures, training objectives, and datasets. Existing evaluations (Agrawal et al., 2022; Garrido et al., 2023) tend to validate a single metric against a narrow set of architectural and training configurations, making it difficult to draw general conclusions

about their reliability. Furthermore, several spectral metrics are constructed from a common underlying eigenspectrum, yet it remains unclear how they differ in their sensitivity to spectral shape and representation dimensionality. Since these metrics are computed solely from a geometric perspective, without any knowledge of the downstream task, some degree of unreliability is to be expected a priori. Thus, examining the conditions under which these metrics serve as reliable proxies for downstream performance is of considerable value.

We perform a comparative analysis of representation quality metrics to evaluate their reliability with respect to training configurations as well as their relation to downstream performance. Our main contributions are:

- **A taxonomy for label-free metrics** that groups them into three families based on their construction: (i) Covariance spectrum-based, (ii) Pairwise relation-based, and (iii) Manifold-based. We further bring their formulations into a common framework to highlight similarities between metrics within each group.
- **Sensitivity analysis of spectral metrics** through controlled synthetic experiments that characterise how covariance spectrum-based metrics respond to changes in the shape of the eigenspectrum and in representation dimensionality.
- **A systematic study** demonstrating the correlation between the metrics and test accuracy from an Ordinary Least Squares fit, as well as their consistency when varying architecture class, training objective, and representation dimension.

Broadly, our results suggest that existing label-free metrics should be used with caution, as their reliability is influenced by model architecture, training objective, and representation dimensionality. We hope our insights are useful in developing more robust and generalisable representation quality metrics.

2 Related Work

Representation learning aims to extract informative features from raw data that facilitate efficient transfer to downstream tasks. Self-supervised learning (SSL) emerged as the dominant paradigm for this purpose, with methods based on contrastive learning (Chen et al., 2020; He et al., 2020; Bardes et al., 2022), self-distillation (Caron et al., 2021; Oquab et al., 2024), and masked reconstruction (He et al., 2022) producing representations that are competitive with supervised counterparts across a range of downstream tasks, especially under fine-tuning. To measure the quality of extracted representations for downstream tasks, a variety of evaluation methods have been used. Broadly, these methods can be classified into label-dependent metrics (which require labels) and label-free metrics (which do not).

Label-dependent metrics: Linear probing, in which a linear classifier is trained on frozen features to assess how much task-relevant information is encoded, has been shown to often be predictive of transfer accuracy (Zhang et al., 2016). A common alternative is to use K-Nearest-Neighbour (KNN) classification or clustering of the same frozen features as proxy for representation quality (Balestriero et al., 2023). Fine-tuning the entire model using the target task can also be used as proxy (He et al., 2022), but is substantially more computationally costly. More efficient transferability metrics, which do not require fine-tuning, have been proposed. These include LEEP, which estimates a joint distribution over source predictions and target labels (Nguyen et al., 2020); LogME, which computes the marginal likelihood of target labels given features (You et al., 2021); and PACTran, which computes Probably Approximately Correct Bayesian bounds on the labelled empirical risk (Ding et al., 2022). However, all of these methods still require access to target labels. This continued dependence on labelled data motivates the development of completely label-free evaluation metrics.

Label-free metrics: Renggli et al. (2022) showed that commonly used task-aware metrics, such as fine-tuning, linear probing, and KNN, may produce inconsistent model rankings across tasks and datasets. To avoid dependence on labels, several label-free metrics (which we detail in Section 3) have been proposed, motivated by different theoretical perspectives. These include metrics based on the spectral properties of the representation covariance matrix (Garrido et al., 2023; Agrawal et al., 2022; He & Ozay, 2022),

relational metrics that work with inter-point relations (Tsitsulin et al., 2023; Liao et al., 2024), and the local neighbourhood structure of the representation manifold (Ansuini et al., 2019). However, these metrics have largely been proposed and validated in isolation, often as by-products of new SSL methods, making it difficult to assess their reliability. A further practical limitation is that some metrics are computed on projector outputs rather than on backbone representations. For example, RankMe (Garrido et al., 2023) was originally computed this way, but projector embeddings are discarded after SSL training and are not available in publicly released models. This constraint has not been consistently addressed in prior work, limiting the applicability of such metrics to comparing models from public repositories.

Direct comparisons of label-free representation quality metrics are rare. Tsitsulin et al. (2023) compile prior label-free representation quality measures (Agrawal et al., 2022; Garrido et al., 2023; He & Ozay, 2022) and introduce additional metrics inspired by numerical linear algebra. They report aggregate results without systematically examining the influence of the model's architecture class, its representation dimension, or its training objective. More broadly, different metrics tend to favour different representation geometries. Spectral measures such as RankMe (Garrido et al., 2023) and NE Sum (He & Ozay, 2022) reward representations with high effective rank and more uniform eigenvalue distributions, while α-Req (Agrawal et al., 2022) favour representations with a decaying covariance eigenspectrum, and relational and manifold-based metrics (Tsitsulin et al., 2023; Liao et al., 2024; Ansuini et al., 2019) favour locally structured geometries. Understanding which metrics are reliable under which conditions, and whether they generalise across architecture classes, training objectives, and representation dimensions, requires a systematic evaluation.

3 A Taxonomy for Label-Free Metrics

In this section, we review a range of representation quality metrics. We highlight the connections between them by reformulating them where applicable. We then propose a taxonomy that groups these metrics according to their underlying theory.

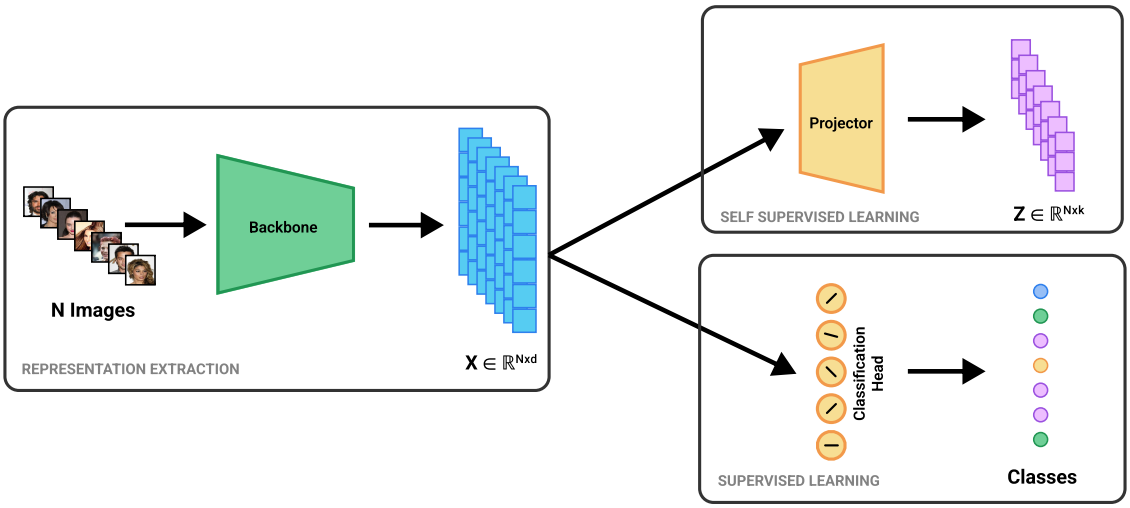


Figure 1: Illustration of the representation extraction pipeline for self-supervised (SSL) and supervised models. In SSL models, the projector MLP is discarded after training, and the backbone output $X \in \mathbb{R}^{N \times d}$ is used as the representation. For supervised models, the penultimate layer activations are used as the representation, also denoted as X. In both cases, representation quality metrics are computed directly on X, without access to the projector embeddings Z or classification heads.

Preliminaries We consider a model $f_\theta : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \to \mathbb{R}^d$ that extracts features from inputs, mapping them to a d-dimensional representation space. For image inputs, this reduces to $f_\theta : \mathbb{R}^{W \times H \times 3} \to \mathbb{R}^d$, where W and H denote the width and height of an RGB image. We refer to this model as the encoder or backbone model, and let $X \in \mathbb{R}^{N \times d}$ denote the corresponding extracted representations, where N is the number of

samples. In most self-supervised models, a Multilayer Perceptron (MLP) is usually applied to the backbone output to produce embeddings, which we denote by $Z \in \mathbb{R}^{N \times k}$. SSL loss functions are usually applied on these embeddings. Once the SSL model is fully trained, the MLP is discarded. Since our motivation comes from identifying which model to use for feature extraction, we consider the backbone output as the representation and compute the metrics on this output. For the supervised models, we consider the output of the penultimate layer as the representation (see Fig.1).

A central underlying question for metrics aiming to estimate representation quality is *which structural properties of a representation correlate with strong downstream performance?* Different metrics provide complementary perspectives on this question by emphasising distinct aspects such as global variance structure, cluster separability, and local neighbourhood consistency.

Table 1 groups and summarises the representation quality metrics that we consider in our analysis. We group them into three broad categories on the basis of how they are constructed: (i) **Spectral metrics** that capture the global structure of representations through the eigenvalues of the representation covariance matrix or on the singular values of the representation matrix X (Section 3.1). (ii) **Relational metrics** consider the pairwise relation between these points in the representation space as a basis for measuring representation quality (Section 3.2). (iii) **Manifold-based metric** which works with the manifold geometry of the representation (Section 3.3).

Table 1: Summary of representation quality metrics considered in this work, grouped by construction. Metrics take the representation matrix, $X \in \mathbb{R}^{N \times d}$, as input, where N is the number of samples and d is the representation dimension (we assume $N \gg d$).

Metric	Category	Complexity	Description
α-ReQ (Agrawal et al., 2022)	Spectral	$\mathcal{O}(Nd^2)$	Fits power law $\lambda_i \propto i^{-\alpha}$ to the eigenspectrum. Sensitive to fitting range in log-log space (Clauset et al., 2009).
RankMe (Garrido et al., 2023)	Spectral	$\mathcal{O}(Nd^2)$	Entropy of normalised singular values. Originally computed on projector output; we compute on backbone output.
NE Sum (He & Ozay, 2022)	Spectral	$\mathcal{O}(Nd^2)$	Measures degree of feature whitening. Equivalent to Stable Rank (Tsitsulin et al., 2023) for mean-centred representations (Section 3.1).
κ (Tsitsulin et al., 2023)	Spectral	$\mathcal{O}(Nd^2)$	Ratio of largest to smallest non-zero singular value. Measures sensitivity of linear system to perturbations.
Self-Cluster (Tsitsulin et al., 2023)	Relational	$\mathcal{O}(Nd^2)^\dagger$	Measures deviation of ℓ_2-normalised representations from a uniform spherical distribution. Reduced from $\mathcal{O}(N^2d)$ via batched reformulation (Section 3.2).
DSE (Liao et al., 2024)	Relational	$\mathcal{O}(N'^2d)$	Spectral entropy of anisotropic diffusion kernel matrix. Subsampled to $N' = 10{,}000$ for scalability. Analytically related to Self-Cluster (Section 3.2).
ID (Ansuini et al., 2019)	Manifold	$\mathcal{O}(N_b^2d)$	MLE estimator based on ratio of first and second nearest-neighbour distances. Computed on batches of size N_b and averaged. Moderately underestimates true ID when ID > 20.

† Reduced from $\mathcal{O}(N^2d)$ via the batched reformulation described in Section 3.2.

3.1 Spectral Metrics

Spectral metrics are defined over the eigenspectrum of the representation covariance matrix or the singular values of the representation matrix X. Intuitively, an ideal representation space should avoid the issue of *rank collapse*, i.e., variance concentrated in a small number of directions. Spectral metrics aim to capture properties such as effective rank, spectral decay, and the dimensionality of the representation space. In this section, we restate and, where possible, reformulate them to highlight their similarities.

α-ReQ Building on the idea that the eigenspectrum of neural population responses in mammals follows a power law, $\lambda_i \propto i^{-\alpha}$, with $\alpha \approx 1$, Agrawal et al. (2022) proposed the α-ReQ metric. α is estimated by fitting a line in the log-log space of the eigenspectrum computed from the covariance matrix and taking its slope. However, Clauset et al. (2009) caution that such procedures can lead to inaccurate estimates of the power-law exponent. In particular, the slope estimate is sensitive to the choice of range in the log-log plot, which may induce a bias. In our experiments, we fit the power law to eigenvalues with indices $i \in [10, \lfloor 0.9 \times d \rfloor]$ as a consistent heuristic to ensure comparability across models.

RankMe Garrido et al. (2023) propose RankMe as an entropy-based metric of the effective rank of an embedding matrix. It is defined by

$$\text{RankMe} = \exp\left(- \sum_{i=1}^{\min(N,d)} p_i \log p_i \right),$$

where $p_i = \frac{\sigma_i(X)}{\|\sigma(X)\|_1}$, and σ_i denotes the singular values of the representation matrix $X \in \mathbb{R}^{N \times d}$ in decreasing order. Here, $\|\sigma(X)\|_1$ represents the ℓ_1 norm of the singular value vector. **Note:** RankMe was originally computed on embeddings (i.e., the output of the projector, $Z \in \mathbb{R}^{N \times k}$ in Fig. 1). In this work, we compute it directly on the representations, i.e., the backbone output. While Garrido et al. (2023) assume access to both the pre-trained backbone and the projector, in practice it is not realistic for model selection using existing sources, where popular hubs like Hugging Face typically provide only the backbone.

NE Sum He & Ozay (2022) introduced this metric to measure the degree of whitening of the representation matrix X, i.e., how evenly the eigenvalues of the covariance matrix are distributed. It is defined by

$$\text{NE Sum} = \frac{\sum_{i=1}^{d} \lambda_i}{\max_i \lambda_i},$$

where λ_i are the eigenvalues of the covariance matrix. NE Sum attains its maximum when the eigenvalues are equal. We note that the metric **Stable Rank**, used by Tsitsulin et al. (2023) is equivalent to NE Sum under the assumption that the representation is centred (i.e., the columns are mean-subtracted).

Stable rank quantifies the difficulty of estimating a matrix from sub-sampled rows. In our analysis, this translates to the difficulty of estimating X. It is defined as follows:

$$\begin{aligned} \text{StableRank(X)} = \frac{\|X\|_F^2}{\|X\|_2^2} &= \frac{\sum_{i=1}^{d} \sigma_i^2}{\max_i \sigma_i^2} \\ &= \frac{\sum_{i=1}^{d} \lambda_i}{\max_i \lambda_i}, \end{aligned}$$

where σ_i denotes the ith singular value of X, and, since X is centred, $\sigma_i^2 = (N-1)\lambda_i$.

Condition Number To assess the stability of the learnt representations, Tsitsulin et al. (2023) consider the pseudo-condition number. In numerical linear algebra, the condition number of the matrix X is defined by

$$\kappa(X) = \frac{\sigma_1}{\sigma_n},$$

where σ_1 and σ_n are the largest and smallest singular values of X, respectively. This quantifies the sensitivity of the solution W to perturbations in the linear system $WX = y$: a higher condition number indicates that small errors in y or X can lead to large errors in W. Tsitsulin et al. (2023) modify this metric by considering only the smallest non-zero singular value, thereby avoiding numerical instability from near-zero values.

3.2 Relational Metrics

Rather than using the covariance matrix to assess the geometry of the representation space, relational metrics characterise it through pairwise interactions among samples. This is typically done by constructing an $N \times N$ relational matrix. These metrics differ in how they quantify pairwise relationships among samples. We restate their construction and, where possible, reformulate them to improve memory efficiency.

Self-clustering metric Tsitsulin et al. (2023) introduce this metric on the basis of the observation that contrastive learning algorithms tend to distribute representations uniformly in the representation space (Assran et al., 2023). The metric estimates how much the embeddings are clustered in the embedding space compared to a random distribution on the sphere. Let $\tilde{X} \in \mathbb{R}^{N \times d}$ denote the ℓ_2-normalised representations of X (applied to each row of X). Assuming that the representations are uniformly distributed on the unit hypersphere, the expected and maximum values of $\|\tilde{X}\tilde{X}^\top\|_F$ are $\mathbb{E}[\|\tilde{X}\tilde{X}^\top\|_F] = N + \frac{N(N-1)}{d}$ and $\max_{\tilde{X}} \|\tilde{X}\tilde{X}^\top\|_F = N^2$ (Tsitsulin et al., 2023), respectively, which gives:

$$
\begin{aligned}
\text{Self-Cluster}(\tilde{X}) &= \frac{\|\tilde{X}\tilde{X}^\top\|_F - \mathbb{E}[\|\tilde{X}\tilde{X}^\top\|_F]}{\max_{\tilde{X}}(\|\tilde{X}\tilde{X}^\top\|_F) - \mathbb{E}[\|\tilde{X}\tilde{X}^\top\|_F]} \\
&= \frac{\|\tilde{X}\tilde{X}^\top\|_F - (N + \frac{N(N-1)}{d})}{N^2 - (N + \frac{N(N-1)}{d})} \\
&= \frac{d\|\tilde{X}\tilde{X}^\top\|_F - N(d+N-1)}{(d-1)(N-1)N}.
\end{aligned}
$$

Since $N \gg d$, in practice, computing $\tilde{X}\tilde{X}^\top \in \mathbb{R}^{N \times N}$ directly is infeasible for large datasets. Consequently, Tsitsulin et al. (2023) omit this computation for ImageNet-scale datasets. We note that $|\tilde{X}\tilde{X}^\top|_F = |\tilde{X}^\top\tilde{X}|_F$ holds, and instead compute $\tilde{X}^\top\tilde{X} \in \mathbb{R}^{d \times d}$. This reformulation allows $\tilde{X}$ to be partitioned into batches $\{\tilde{X}_{b,i}\}$ and compute $\tilde{X}^\top\tilde{X}$ via aggregation, i.e., $\sum_i \tilde{X}_{b,i}^\top \tilde{X}_{b,i}$, thereby reducing the memory complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(d^2)$.

Diffusion Spectral Entropy (DSE) Liao et al. (2024) compute the spectral entropy of a diffusion matrix constructed from pairwise similarity measures. The implementation uses an anisotropic Gaussian kernel as follows:

$$
G_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),
$$

where x_i denotes the representation, i.e., the i-th row of X. G is then symmetrically normalised to remove the effect of local sampling density and capture only the intrinsic geometry, yielding K, which is then used to compute the DSE,

$$
\begin{aligned}
K = DGD, \quad D = \text{diag}(1/\sqrt{\sum_k G_{ik}}), \\
\text{DSE} = -\sum_i p_i^t \log p_i^t, \quad p_i = \frac{\hat{\lambda}_i}{\sum_j \hat{\lambda}_j},
\end{aligned}
$$

where $\hat{\lambda}_i$ are the eigenvalues of K and t is the diffusion time parameter, which we set to 1 in our experiments. For small σ, G_{ij} decays rapidly with distance, producing a nearly diagonal K regardless of global geometry; thus, the metric loses sensitivity to global structure. For large σ, all entries of G saturate towards a constant, again collapsing the eigenspectrum and making the metric uninformative. High DSE indicates a uniform eigenspectrum corresponding to representations with a rich, distributed structure; low DSE indicates a few

dominant eigenvalues, corresponding to representations with a simple or highly concentrated structure. Since this metric involves pairwise evaluations, we follow the authors' approach and randomly sample 10,000 points. We exclude the Diffusion Mutual Information (DMI) metric introduced alongside DSE, as it requires label information.

For ℓ_2-normalised representations with sufficiently large σ, the Gaussian kernel may be approximated as (see Appendix A):

$$G \approx C\left(\mathbf{1}_{N\times N} + \frac{1}{\sigma^2}\tilde{X}\tilde{X}^\top\right).$$

After anisotropic normalisation, K depends on $\tilde{X}\tilde{X}^\top$, up to scaling by the degree matrix. Since Self-Cluster is derived from $\|\tilde{X}\tilde{X}^\top\|_F$, both DSE and Self-Cluster depend on $\tilde{X}\tilde{X}^\top$. A clustered representation space produces a few dominant eigenvalues in $\tilde{X}\tilde{X}^\top$, which increases $\|\tilde{X}\tilde{X}^\top\|_F$ (i.e., Self-Cluster increases) while reducing spectral entropy (DSE decreases), and vice versa for a uniform eigenspectrum. We empirically confirm this negative correlation in Section 4.2.

3.3 Manifold Metrics

The representations are often observed to lie on a lower-dimensional manifold. Manifold-based metrics aim to capture geometric properties such as dimensionality and the local structure of the underlying manifold.

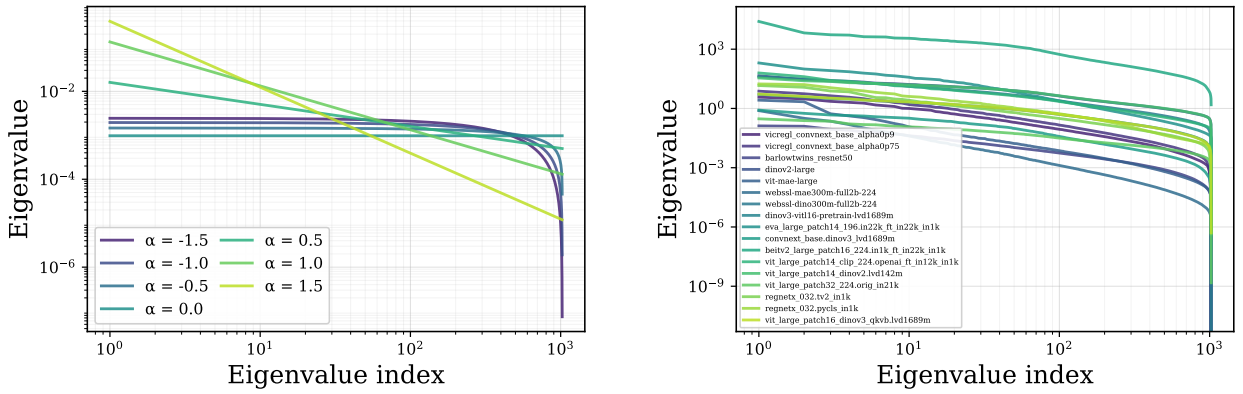
Intrinsic dimension (ID) Ansuini et al. (2019) propose a metric for estimating the minimal number of parameters required to describe the representations. It is based on the observation that the ratio of distances to the first and second nearest neighbours follows a known distribution under specific assumptions. The corresponding distribution parameter, which serves as the metric, is computed using maximum likelihood estimation. Lower ID suggests that the learnt representation lies on a lower-dimensional manifold, which the authors show is associated with better generalisation performance. In particular, the ID of the last hidden layer predicts test accuracy. Lu et al. (2023) studied the **Cluster Learnability and Intrinsic Dimension (CLID)**, which can be thought of as an extension of ID. CLID combines the Intrinsic dimension and clusterability (KNN performance, which requires label information). We ignore this in our evaluation as it requires labels. We take the mean of IDs computed for each batch as the final ID.

4 Comparative Analysis

The metrics in Section 3 have been proposed as proxies for downstream performance while operating only on the feature geometry without requiring label information. However, it remains unclear whether these metrics consistently and reliably predict downstream task performance across a wide range of configurations. In particular, while several spectral metrics are constructed from the same underlying eigenspectrum, it is unclear how they differ from one another when the underlying spectral shape and representation dimensionality remain the same. In this section, (i) we decouple confounders arising from the choice of backbone model and datasets by considering a simulated eigenspectrum and analyse how spectral metrics respond to changes in spectral shape and representation dimensionality (4.1), and (ii) we assess how well all the metrics correlate with OLS fit accuracy across a diverse set of backbone models, datasets, and training objectives (4.2).

4.1 Sensitivity of Spectral Metrics

The eigenspectrum of the representation covariance matrix has been shown to exhibit an approximately power-law decay, with a small number of dominant directions that capture most of the variance (Nassar et al., 2020; Xie et al., 2022). A power law of the form of $\lambda_i \propto i^{-\alpha}$ provides a natural parametric family to characterise this decay. In addition, (Agrawal et al., 2022) used this formulation to study representation quality. Thus, we synthetically generate eigenspectra under the power law assumption, setting $\sum_{i=1}^{d} \lambda_i = 1$ without loss of generality, since all three metrics are scale-invariant with respect to the eigenvalues. While Agrawal et al. (2022) restrict their attention to $\alpha > 0$, we additionally consider $\alpha < 0$ to examine the spectral shape. For a consistent basis of comparison, we sort the eigenvalues in decreasing order. As α becomes increasingly negative, most of the eigenspectrum is largely flat, followed by a sharp decline (see the



(a) $\alpha \in [-1.5, 1.5]$

(b) CIFAR-100 Eigenspectrum

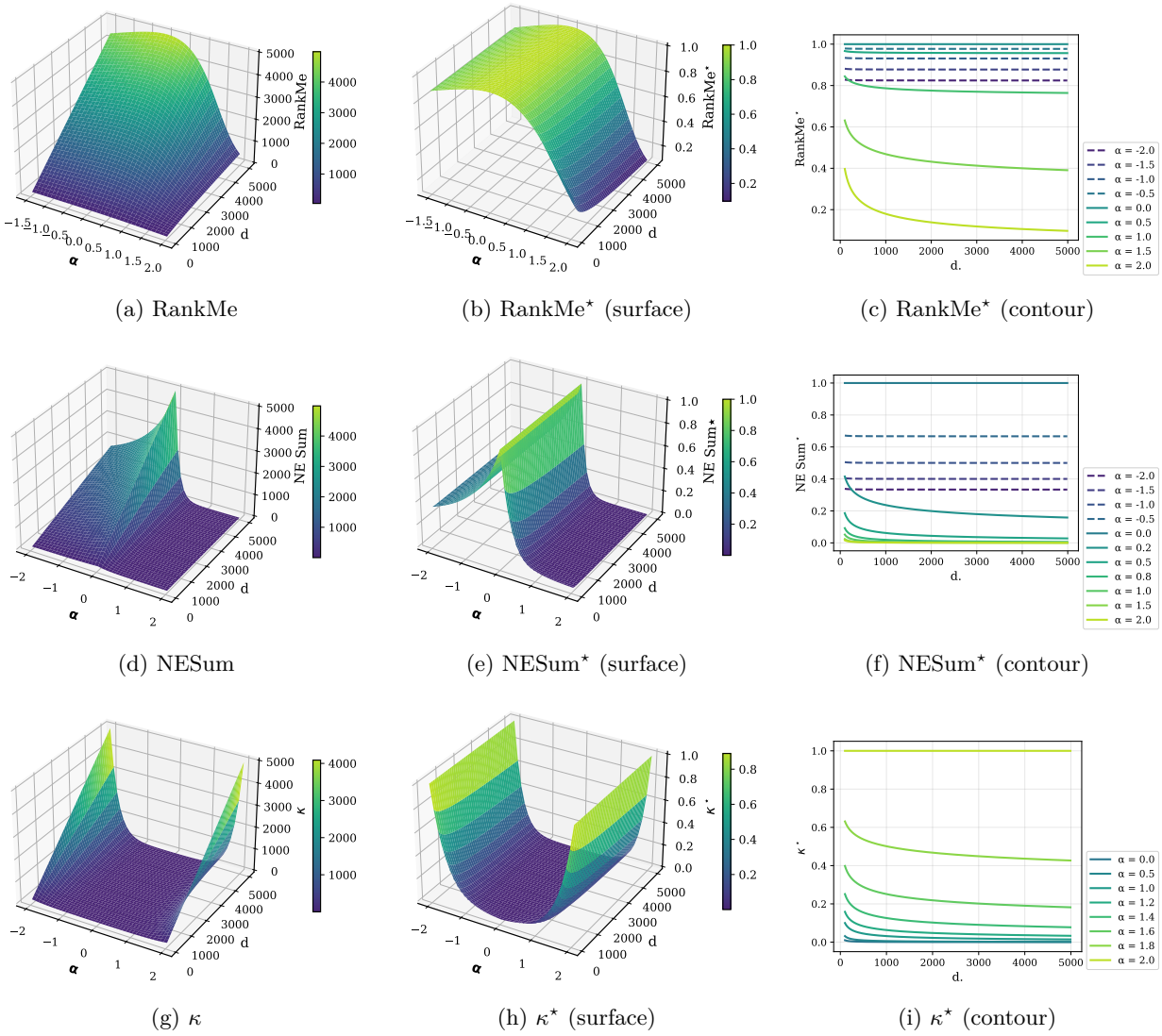
Figure 2: Comparison of real and synthetic eigenspectra in log-log scale. Real backbone representations with dimension $d \in [1000, 1024]$ (Fig. 2b) exhibit spectral shapes consistent with the synthetically generated power-law family $\lambda_i \propto i^{-\alpha}$, $\sum_{i=1}^{d} \lambda_i = 1$, under negative values of α (Fig. 2a). This similarity justifies including $\alpha < 0$ in the synthetic sensitivity analysis

curves corresponding to $\alpha < 0$ in Fig. 2a). We note that this closely resembles the eigenspectra observed in real representations (Fig. 2b), justifying including $\alpha < 0$ in our analysis. We vary $\alpha \in [-2, 2]$ and $d \in [100, 5000]$, where the latter range spans the representation dimensions of the model families considered in Section 4.2.

We first examine the behaviour of these metrics across this parameter space. Figs. 3a, 3d show that both RankMe and NE Sum are maximised at $\alpha = 0$ for all values of d. This is consistent with their construction, as RankMe measures spectral entropy which is maximised when the spectrum is uniform ($\alpha = 0$ generates a uniform eigenspectrum as observed in Fig. 2a) and NE Sum, by construction measures features whitening which is synonymous with $\alpha = 0$. However, these metrics are asymmetric about $\alpha = 0$, which is explained by the fact that the spectrum is nearly flat for $\alpha < 0$, compared to the rapid decay for $\alpha > 0$. On the other hand, κ is symmetric by construction, as it depends only on the ratio of the largest to the smallest singular value, which is identical for $\pm\alpha$ under the power law after sorting. κ attains its maximum at the extremities and is unbounded (Fig. 3g); as $|\alpha|$ increases, the smallest eigenvalue becomes vanishingly small. Both RankMe and NE Sum scale with d, while κ displays similar behaviour, but this effect is more pronounced as $|\alpha|$ increases. To aid comparison between models with different d, we normalise the metrics by dividing by d.

The normalised surfaces (Figs. 3b, 3e) still reveal qualitative asymmetry around $\alpha = 0$. The contour lines in Figs. 3c and 3f reveal that for the same value of α, RankMe* consistently show higher values than NE Sum* in both regimes. For $\alpha > 0$, both metrics decay with increasing d even after normalisation. This indicates sub-linear scaling in d in this regime. For $\alpha \leq 0$, both RankMe* and NE Sum* are insensitive to changes in d, consistent with linear scaling in that regime. NE Sum* exhibits steeper variation near $\alpha = 0$ than RankMe*, as seen from the more tightly spaced contour lines in that region (Figs. 3c and 3f). Since κ is symmetric about $\alpha = 0$, we consider only positive values of α in Fig. 3i. On the other hand, $\kappa^\star$ is minimised near $\alpha = 0$ and grows as $|\alpha|$ increases (Fig. 3i).

These results reveal that the three metrics exhibit different sensitivities to the shape of eigenspectra. NE Sum* exhibits high sensitivity to α around $\alpha = 0$ and RankMe* exhibits low sensitivity in the same region. On the other hand, $\kappa^\star$ exhibits complementary behaviour by being most sensitive at the spectral extremes and largely uninformative near $\alpha = 0$. This suggests that no single spectral metric provides uniformly consistent sensitivity across the full range of spectral shapes encountered in practice, and care is needed when interpreting these metric values.



(a) RankMe (b) RankMe⋆ (surface) (c) RankMe⋆ (contour)

(d) NESum (e) NESum⋆ (surface) (f) NESum⋆ (contour)

(g) κ (h) $\kappa^\star$ (surface) (i) $\kappa^\star$ (contour)

Figure 3: Sensitivity of spectral metrics RankMe, NE Sum, and κ to spectral shape (α) and representation dimension (d), where eigenspectra are synthetically generated as $\lambda_i \propto i^{-\alpha}$ with $\sum_{i=1}^{d} \lambda_i = 1$. Each row shows the un-normalised metric (col. 1), the dimension-normalised surface (col. 2), and the normalised metric as a function of d for fixed values of α (col. 3). Normalisation is performed by dividing by d, following the normalisation proposed in Tsitsulin et al. (2023) for RankMe, which we extend to all three metrics. In col. 3, dashed lines correspond to $\alpha < 0$ and solid lines to $\alpha \geq 0$. Since κ is symmetric about $\alpha = 0$ (Fig. 3h), only $\alpha \geq 0$ is shown in Fig. 3i.

4.2 Correlation of Metrics with Test Accuracy

The synthetic analysis in the previous section, by construction, decoupled the effects of architecture, training objective, and dataset from spectral shape. However, these factors also influence the geometry of real-world representations. In this subsection, we assess how well the metrics listed in Table 1 correlate with downstream performance by grouping backbone models according to architecture class, training objective, and the data set used to compute representations.

Experimental Setup We consider (i) supervised models and (ii) Self-Supervised models. In total, we consider 260 backbone models for our analysis (details of the models, training data, and training methods are provided in Appendix B). Across all experiments, we extract features from the final layer of the backbone. We consider Vision Transformers (ViT) (Dosovitskiy et al., 2021), Residual Networks (He et al., 2016), ConvNeXT (Liu et al., 2022), and RegNet (Radosavovic et al. (2020)) model architectures. Following standard practice, we use the [CLS] token as the representation for ViT models, and for CNN-based architectures (ResNet, ConvNeXT, RegNet), we apply global average pooling over the final convolutional feature maps (He et al., 2016).

To enable faster and more stable evaluation, we compute a closed-form regularised least squares solution, to predict class labels from one-hot encoded targets on ℓ_2-normalised features, with final accuracy determined by the argmax of the predicted outputs. This is partly motivated by the linear regression framework used in the α-ReQ metric (Agrawal et al., 2022; Bartlett et al., 2020). We use Spearman's correlation between the metrics considered and the test accuracy for our comparative analysis. Spearman's correlation is preferred over Pearson's as it does not assume a linear relationship between metric and accuracy.

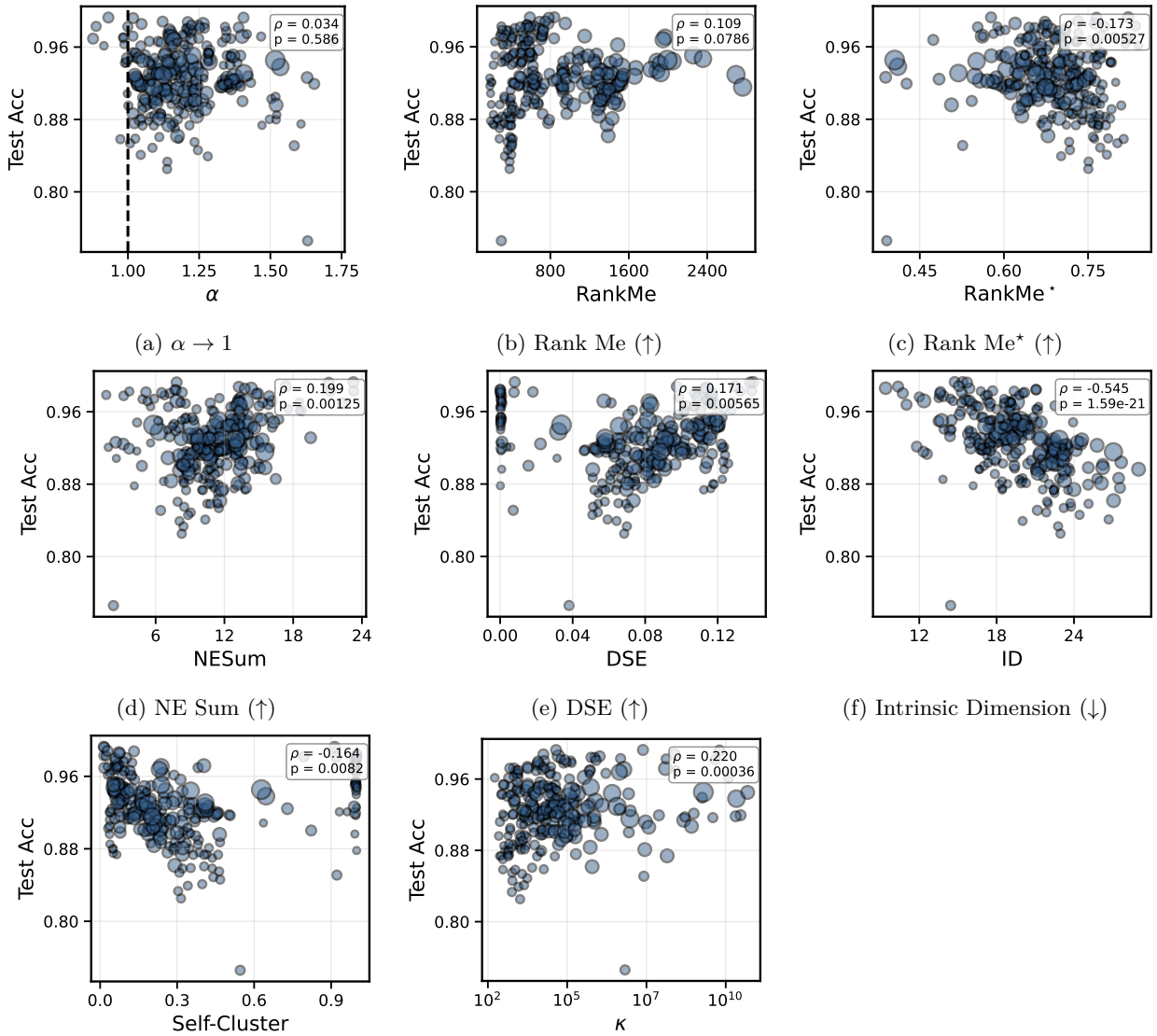
Table 2: Spearman correlation (ρ) between representation quality metrics and test accuracy on CIFAR-10, CIFAR-100, and ImageNet-1k. Metrics are grouped by construction: spectral (top), relation-based (middle), and manifold-based (bottom). Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), → 1 (closer to 1 is better). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

		CIFAR-10	**CIFAR-100**	**ImageNet**
$\alpha < 1.0$	↑	-0.3242	-0.5381^{***}	-0.347^{***}
$\alpha \geq 1.0$	↑	0.1156	0.06995	0.1688
α-Req	→ **1**	0.03393	-0.01968	-0.1169
RankMe	↑	0.1093	0.2443^{***}	0.2166^{***}
RankMe*	↑	-0.1725^{**}	-0.1362^{*}	-0.05487
NE-Sum	↑	0.1991^{**}	0.2309^{***}	0.1048
κ	↑	0.2196^{***}	0.3331^{***}	0.1599^{**}
Self-Cluster	↓	-0.1636^{**}	-0.1803^{**}	-0.09359
DSE	↑	0.1712^{**}	0.1967^{**}	0.1096
ID	↓	-0.5451^{***}	-0.5981^{***}	-0.6254^{***}

Observation In Table 2, ID is the most reliable predictor, showing a strong, and consistent negative correlation with test accuracy across all datasets ($\rho \in [-0.63, -0.55]$, $p \ll 0.05$). Self-Cluster, DSE, and NE Sum show mild but statistically significant correlations on the CIFAR datasets, though none of these are significant on ImageNet. κ shows a consistent and statistically significant positive correlation in all three datasets. Surprisingly, RankMe* shows a statistically significant mild negative correlation on the CIFAR datasets. However, Figs. 4c and 5c suggest that the statistical significance does not translate to practical significance. Models with high RankMe* scores exhibit no discernible trend. RankMe, on the other hand, exhibits positive correlations that are statistically significant in CIFAR-100 and ImageNet.

Interestingly, Figs. 4g, 5g, and 6g reveal a small subset of models with high accuracy and Self-Cluster values close to 1.0, and similarly near-zero values are observed for DSE (Figs. 4e, 5e, 6e). We also observe an almost mirror image characteristic for these two metrics in all datasets, and indeed, the correlation scores between the two metrics (CIFAR-10: -0.999, CIFAR-100: -0.999, and ImageNet: -0.998) suggest that they are nearly the same metric in practice. In Section 3.2, we illustrated the relationship between these two metrics and how they capture the same phenomenon but in opposite directions. This near perfect negative correlation is expected since the DSE measures the entropy of the diffusion process over the representation graph, which is maximised when representations are spread uniformly. On the other hand, Self-Cluster explicitly measures deviation from a uniform spherical distribution.

Since Agrawal et al. (2022) hypothesised that $\alpha \approx 1$ is indicative of high-quality representations, we split the correlation analysis into two regimes: $\alpha \geq 1.0$ and $\alpha < 1.0$. Interestingly, the α-ReQ metric shows negative



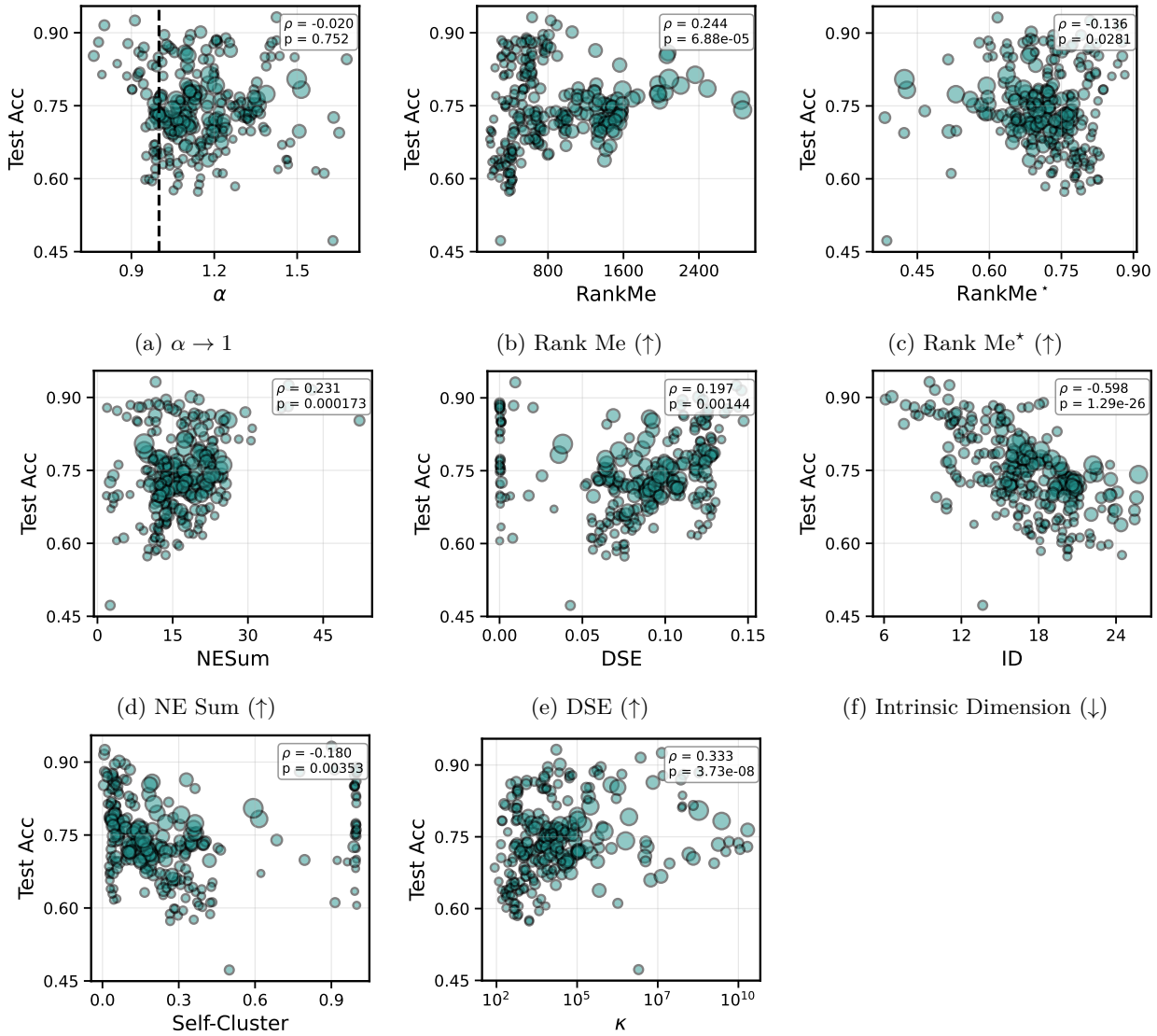
(a) $\alpha \to 1$ (b) Rank Me (↑) (c) Rank Me* (↑)

(d) NE Sum (↑) (e) DSE (↑) (f) Intrinsic Dimension (↓)

(g) Self-Cluster (↓) (h) κ(↓)

Figure 4: Comparison of representation metrics and CIFAR-10 test accuracy using an OLS fit. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or → 1 (closer to 1 is better). We mark the $\alpha = 1$ line in Fig. 4a in accordance with Agrawal et al. (2022). Fig. 4h uses a $\log_{10}$ scale on the x-axis to visualise κ, motivated by the exponential like behaviour observed in the synthetic analysis (Sec. 4.1).

correlations with the test accuracy (with $p < 0.05$) in the $\alpha < 1.0$ region, which means that higher α values below 1 correspond to lower accuracy, which directly contradicts the hypothesis. It is also worth noting that the region $\alpha \geq 1.0$ yields unreliable correlation scores in all datasets. For the CIFAR datasets, as is evident in Figs. 4a and 5a, the representations predominantly have $\alpha \geq 1.0$, leaving very few models in the $\alpha < 1.0$ regime and making correlation estimates unreliable for these datasets. For ImageNet, the wider spread of α on both sides of 1.0 (Fig. 6a) provides a more balanced split, leading to a credible observation of a negative correlation in the $\alpha < 1.0$ regime. The overall mild to moderate levels of correlation overall suggest that these metrics are insufficient as stand-alone indicators of downstream performance. Since architecture class and training objective jointly influence representation geometry, we conduct a more fine-grained analysis by grouping models accordingly.



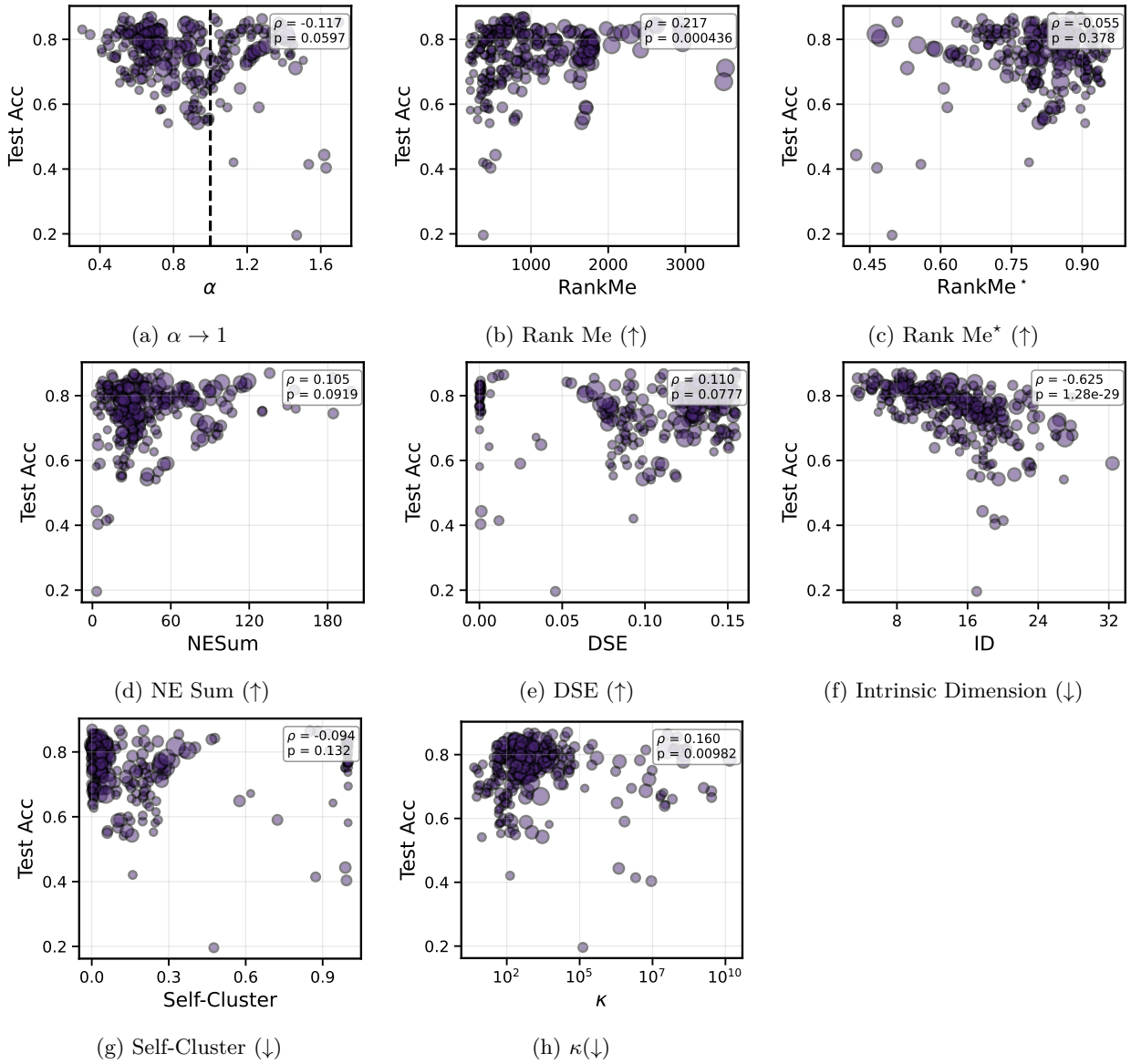
(a) $\alpha \to 1$ (b) Rank Me (↑) (c) Rank Me* (↑)

(d) NE Sum (↑) (e) DSE (↑) (f) Intrinsic Dimension (↓)

(g) Self-Cluster (↓) (h) κ(↓)

Figure 5: Comparison of representation metrics and CIFAR-100 test accuracy using an OLS fit. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or → 1 (closer to 1 is better). We mark the $\alpha = 1$ line in Fig. 5a in accordance with Agrawal et al. (2022). Fig. 5h uses a $\log_{10}$ scale on the x-axis to visualise κ, motivated by the exponential like behaviour observed in the synthetic analysis (Sec. 4.1).

Model architectures' impact on metrics Grouping models by architecture reveals that metric reliability varies considerably across architecture classes. In total, we consider 234 models for this analysis (ConvNeXT: 52, RegNet: 63, ResNet: 64, ViT: 55 models). In particular, Tab. 3 indicates that RankMe is a strong predictor at the architecture level, showing significant positive correlations across almost all architecture classes in CIFAR-10 and CIFAR-100 ($\rho \in [0.56, 0.90]$), a finding that contrasts with Table 2 where it shows a mild correlation. This suggests that architecture class is a significant confounding factor in the aggregate results. Interestingly, ID is the most consistent metric overall, showing significant correlations in all architecture classes ($\rho \in [-0.44, -0.64]$) and for most classes in the CIFAR datasets, except for the ResNet class and the ViT class for CIFAR-10.



(a) $\alpha \to 1$ (b) Rank Me (↑) (c) Rank Me* (↑)

(d) NE Sum (↑) (e) DSE (↑) (f) Intrinsic Dimension (↓)

(g) Self-Cluster (↓) (h) κ(↓)

Figure 6: Comparison of representation metrics and CIFAR-100 test accuracy using an OLS fit. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or → 1 (closer to 1 is better). We mark the $\alpha = 1$ line in Fig. 6a in accordance with Agrawal et al. (2022). Fig. 6h uses a $\log_{10}$ scale on the x-axis to visualise κ, motivated by the exponential like behaviour observed in the synthetic analysis (Sec. 4.1).

The insignificant correlations for Self-Cluster and DSE for ViT across all datasets partly explains the near-zero DSE and near-one Self-Cluster values in Figs. 4g, 5g, 6g, 4e, 5e, and 6e. These collapsed values originate predominantly in ViT models, as illustrated in the top row of Fig. 7. Furthermore, the bottom row of Fig. 7 reveals that within ViT models, the near-unit Self-Cluster values arise predominantly from supervised and SSL+Sup models rather than SSL models, consistent with the finding in Section 4.2 that Self-Cluster is unreliable for supervised and SSL+Sup models, with the same pattern observed for DSE (Appendix D).

κ shows significant positive correlations across the CIFAR datasets, though its reliability weakens for the ConvNeXT and ViT classes on ImageNet. Self-Cluster and DSE are reliable for RegNet and ResNet across

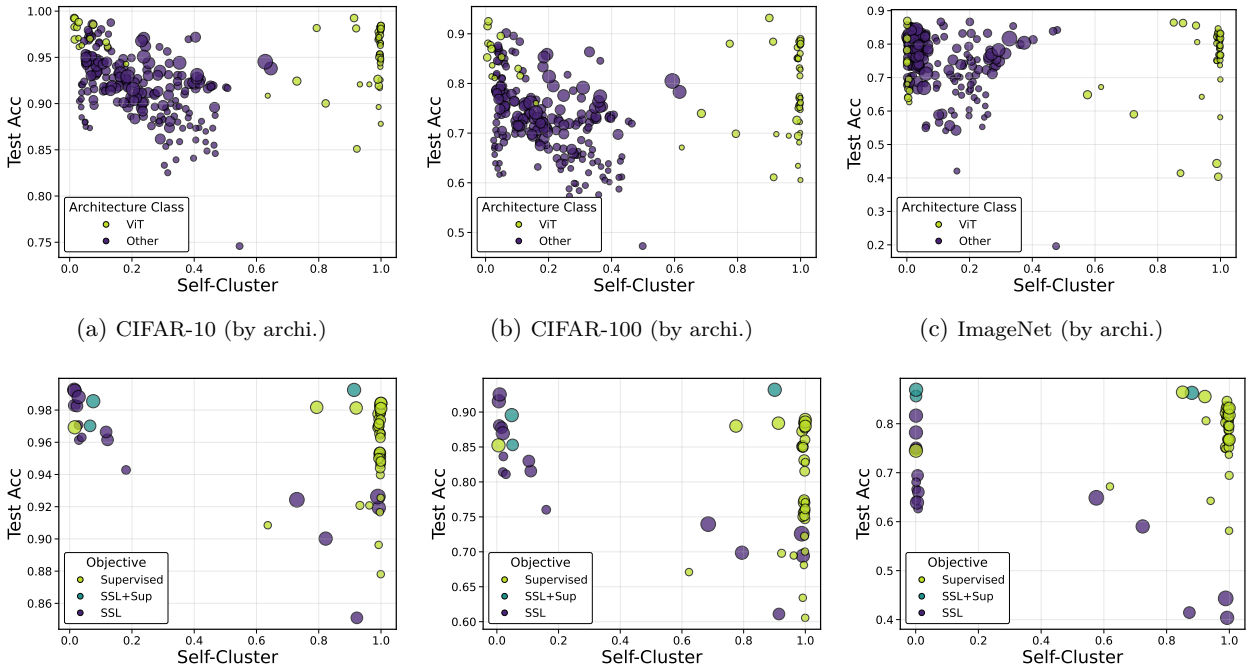
Table 3: Spearman correlation (ρ) between representation metrics and test accuracy grouped by architecture across CIFAR-10, CIFAR-100, and ImageNet-1k. Metrics are grouped by construction: spectral (top), relation based (middle), and manifold-based (bottom). Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), → 1 (closer to 1 is better). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey values indicate statistically insignificant correlations ($p > 0.05$).

			ConvNeXT	**RegNet**	**ResNet**	**ViT**
CIFAR-10	**α-Req**	→ 1	-0.469^{***}	-0.08933	-0.1038	-0.2193
	RankMe	↑	0.7983^{***}	0.779^{***}	0.7668^{***}	0.5649^{***}
	RankMe*	↑	0.1358	-0.3692^{**}	-0.3669^{**}	0.1175
	NE-Sum	↑	0.2707	0.4884^{***}	0.3144^{*}	0.3697^{**}
	κ	↑	0.3292^{*}	0.6438^{***}	0.6204^{***}	0.409^{**}
	Self-Cluster	↓	0.181	-0.5331^{***}	-0.7515^{***}	-0.1731
	DSE	↑	-0.1278	0.5481^{***}	0.7579^{***}	0.1714
	ID	↓	-0.5375^{***}	-0.5716^{***}	-0.1554	-0.2191
CIFAR-100	**α-Req**	→ 1	-0.5233^{***}	0.08324	-0.02837	-0.1791
	RankMe	↑	0.8189^{***}	0.8952^{***}	0.8425^{***}	0.5985^{***}
	RankMe*	↑	0.2661	-0.5361^{***}	-0.4239^{***}	0.1086
	NE-Sum	↑	-0.2015	0.4731^{***}	0.5276^{***}	0.4809^{***}
	κ	↑	0.3774^{**}	0.7608^{***}	0.675^{***}	0.3554^{**}
	Self-Cluster	↓	0.1316	-0.3331^{**}	-0.7623^{***}	-0.143
	DSE	↑	-0.000555	0.3578^{**}	0.7743^{***}	0.1401
	ID	↓	-0.8087^{***}	-0.5637^{***}	-0.06056	-0.5633^{***}
ImageNet	**α-Req**	→ 1	-0.233	-0.3788^{**}	-0.4998^{***}	0.1701
	RankMe	↑	0.3897^{**}	0.7538^{***}	0.6047^{***}	0.1835
	RankMe*	↑	0.1452	0.1151	0.136	-0.1315
	NE-Sum	↑	-0.03091	0.5665^{***}	0.2135	0.1395
	κ	↑	0.006318	0.3854^{**}	0.4237^{***}	-0.238
	Self-Cluster	↓	0.542^{***}	-0.5665^{***}	-0.5677^{***}	0.1932
	DSE	↑	-0.5304^{***}	0.5717^{***}	0.5864^{***}	-0.1783
	ID	↓	-0.4361^{**}	-0.64^{***}	-0.4489^{***}	-0.5089^{***}

datasets, and interestingly they show a reversal in the direction of correlation for ConvNeXT on ImageNet. Specifically, for ConvNeXT on ImageNet specifically, Self-Cluster shows a positive correlation, and DSE shows a negative correlation, reversing the pattern observed in all other architectures. Upon examining Fig. 8, a majority of the models exhibit varying test accuracy for the same metric value: near zero for Self-Cluster and ~ 0.15 for DSE, leading to the reversal in correlation scores. Surprisingly, ViT, which generally produces high-quality transferable representations (Dosovitskiy et al., 2021), only shows strong correlations on the CIFAR datasets but fewer significant metrics on ImageNet, suggesting that its representations may be less distinguishable by these metrics on more complex tasks. α-ReQ remains largely unreliable across all architecture classes, with the exception of ConvNeXT on the CIFAR datasets.

Influence of training objective on the metrics Another factor that influences how representations relate to each other is the training objective. We consider two main groups: models trained in a supervised fashion (Supervised, $n = 209$) and models trained with SSL objectives (SSL, $n = 32$). A third group of models pre-trained with SSL and subsequently fine-tuned for supervised tasks (SSL+Sup, $n = 19$) is included for completeness but contains too few models to draw reliable conclusions (results for this group are provided in Appendix C.1)

From Table 4, ID shows significant correlations across all datasets for both supervised models and SSL models. Self-Cluster ($\rho \in [-0.82, -0.73]$) and DSE ($\rho \in [0.75, 0.82]$) show strong and significant correlations for SSL models across all datasets, but are insignificant for supervised models. This suggests that these metrics



(a) CIFAR-10 (by archi.) (b) CIFAR-100 (by archi.) (c) ImageNet (by archi.)

(d) CIFAR-10 (by objective) (e) CIFAR-100 (by objective) (f) ImageNet (by objective)

Figure 7: Self-Cluster metric values plotted against test accuracy across datasets. The top row shows all models coloured by architecture class, revealing that the models with extreme values predominantly belong to the ViT architecture class. The bottom row shows only ViT models coloured by training objective, illustrating that the near unit Self-Cluster values originate predominantly from Supervised and SSL+Sup models rather than SSL models. This suggests that the Self-Cluster metric may be a reliable predictor for SSL models (though the sample size is small), but is unreliable for Supervised and SSL+Sup models.

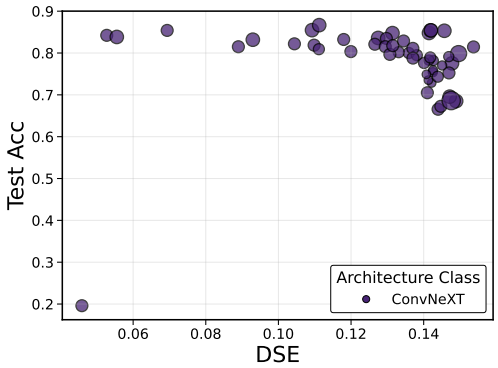
capture properties relevant to representations trained with SSL objectives, where semantically similar data are encouraged to lie close in representation space.

RankMe* ($\rho \in [0.49, 0.61]$) shows a significant positive correlation for SSL models in all three datasets, which is opposite to the trend observed in Table 2. This suggests that the relationship between the metric with test accuracy is dependent on the training objective. However, consistent with Table 2, it displays statistically significant negative correlation scores for Supervised models. RankMe is largely insignificant for SSL models but mildly significant for supervised models ($\rho \in [0.15, 0.30]$).

κ shows significant positive correlations only for supervised models in all datasets. NE Sum is significant for SSL but unreliable for Supervised models in all three datasets. α-ReQ shows significant negative correlations for SSL models across all datasets. For Supervised models, it shows a weak but statistically significant positive correlation on CIFAR datasets, however, Fig. 9 illustrates that no clear visual trend is present, suggesting that there is no meaningful relationship.

From the above observations, we infer that the training objective plays a strong role in determining the reliability of these metrics. Consistent with the observations in Table 2, ID is the only consistent metric that can reliably predict performance across both SSL and the supervised objectives.

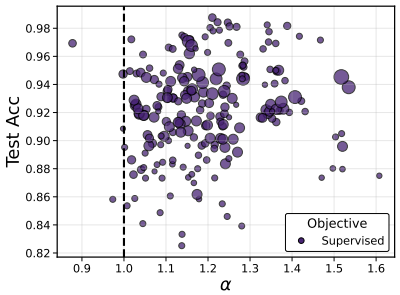
Reliability of metrics across representation dimensionality Motivated by the synthetic analysis in Section 4.1, which showed that normalised spectral metrics exhibit different sensitivities to d, we group models by their representation dimension based on quartiles ($d \leq 520$, $n = 63$), ($520 < d \leq 960$, $n = 70$), ($960 < d \leq 1792$, $n = 61$) and ($d > 1792$, $n = 66$).



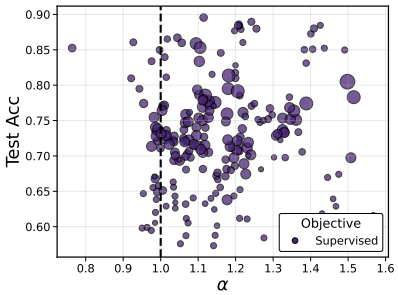
(a) DSE metric for ConvNeXT models against Test Accuracy

(b) Self-Cluster metric for ConvNeXT models against Test Accuracy

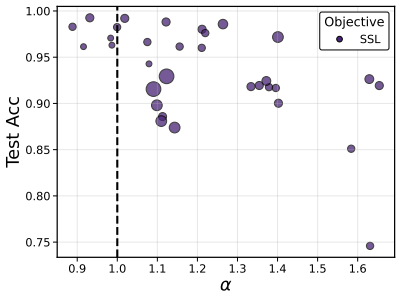
Figure 8: DSE and Self-Cluster plotted against ImageNet test accuracy for ConvNeXT models. A small number of models exhibit near-zero Self-Cluster and ~ 0.15 DSE values but varying levels of test accuracy. These outliers drive the sign reversal observed in Table 3, where ConvNeXT shows a positive correlation for Self-Cluster and a negative correlation for DSE, reversing the pattern observed across all other architecture classes.



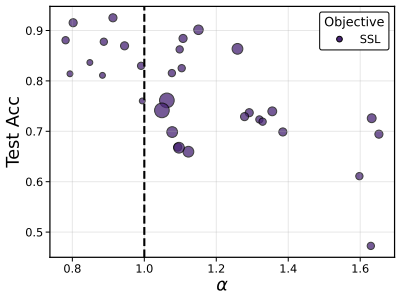
(a) CIFAR-10



(b) CIFAR-100



(c) CIFAR-10



(d) CIFAR-100

Figure 9: α-Req metric values plotted against the test accuracy for the CIFAR datasets. The top row shows supervised models, while the bottom row shows SSL models. Although Table 4 indicates a statistically significant positive correlation for Supervised models, visual inspection reveals no discernible trend in either dataset. In contrast, SSL models exhibit the expected negative trend.

Consistent with observations across architecture classes and training objectives, Table 5 shows that ID shows statistically significant correlations in all settings except CIFAR-10 in $520 < d \leq 960$. Relation based metrics such as Self-Cluster and DSE display significant correlation scores in $d \leq 520$ and $960 < d \leq 1792$ on the CIFAR datasets. However, for the ImageNet dataset, they display significant scores in $d \leq 520$ and

Table 4: Spearman correlation (ρ) between representation metrics and test accuracy grouped by training objective across CIFAR-10, CIFAR-100, and ImageNet-1k. Metrics are grouped by construction: spectral (top), relation based (middle), and manifold-based (bottom). Arrows indicate the preferred direction: $\uparrow$ (higher is better), $\downarrow$ (lower is better), $\rightarrow 1$ (closer to 1 is better). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey values indicate statistically insignificant correlations ($p > 0.05$).

			SSL	**Supervised**
CIFAR-10	**α-Req**	$\rightarrow$ **1**	-0.4703^{**}	0.1877^{**}
	RankMe	$\uparrow$	0.09054	0.1503^{*}
	RankMe*	$\uparrow$	0.4985^{**}	-0.3683^{***}
	NE-Sum	$\uparrow$	0.5942^{***}	0.0295
	κ	$\uparrow$	-0.06452	0.294^{***}
	Self-Cluster	$\downarrow$	-0.7885^{***}	0.003576
	DSE	$\uparrow$	0.7903^{***}	0.006315
	ID	$\downarrow$	-0.5249^{**}	-0.5387^{***}
CIFAR-100	**α-Req**	$\rightarrow$ **1**	-0.6162^{***}	0.1634^{*}
	RankMe	$\uparrow$	0.2258	0.3044^{***}
	RankMe*	$\uparrow$	0.6103^{***}	-0.3593^{***}
	NE-Sum	$\uparrow$	0.6617^{***}	0.1204
	κ	$\uparrow$	-0.07808	0.3992^{***}
	Self-Cluster	$\downarrow$	-0.8134^{***}	-0.02742
	DSE	$\uparrow$	0.8167^{***}	0.04236
	ID	$\downarrow$	-0.6855^{***}	-0.5586^{***}
ImageNet	**α-Req**	$\rightarrow$ **1**	-0.7038^{***}	-0.05376
	RankMe	$\uparrow$	0.331	0.2411^{***}
	RankMe*	$\uparrow$	0.5033^{**}	-0.1564^{*}
	NE-Sum	$\uparrow$	0.4296^{*}	0.08246
	κ	$\uparrow$	0.1554	0.4054^{***}
	Self-Cluster	$\downarrow$	-0.7302^{***}	-0.1091
	DSE	$\uparrow$	0.7551^{***}	0.1267
	ID	$\downarrow$	-0.4567^{**}	-0.5948^{***}

$d > 1792$. The correlation strength is highest in the $960 < d \leq 1792$ bin for CIFAR datasets, suggesting that relation-based metrics are reliable predictors in higher-dimensional settings.

The spectral metrics do not exhibit a discernible pattern. α-ReQ is largely an unreliable metric for predicting performance on CIFAR datasets. On the ImageNet dataset, it displays a positive correlation score for $d \leq 520$, a negative correlation for $d > 1792$, and insignificant correlation scores for $520 < d \leq 1792$. Interestingly, RankMe exhibits negative correlation scores in $d \leq 520$ for CIFAR datasets, but displays the expected direction of correlation for $d > 1792$. However, Fig. 10 indicates that this result is influenced by a concentration of models with high RankMe values and low test accuracy. RankMe* displays a significant negative correlation for $d \leq 520$ across all three datasets, suggesting that in the low-dimensional setting, models that concentrate information into fewer dimensions outperform those that spread variance more uniformly. It also shows no significant correlation for $d > 550$, with CIFAR-10 being the only exception, where a significant negative correlation re-emerges in $d > 1792$. κ is only reliable for the CIFAR datasets and only when $d \leq 520$.

Table 5: Spearman correlation (ρ) between representation metrics and test accuracy grouped by training objective across CIFAR-10, CIFAR-100, and ImageNet-1k. Metrics are grouped by construction: spectral (top), relation based (middle), and manifold-based (bottom). Arrows indicate the preferred direction: $\uparrow$ (higher is better), $\downarrow$ (lower is better), $\rightarrow 1$ (closer to 1 is better). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Grey values indicate statistically insignificant correlations ($p > 0.05$).

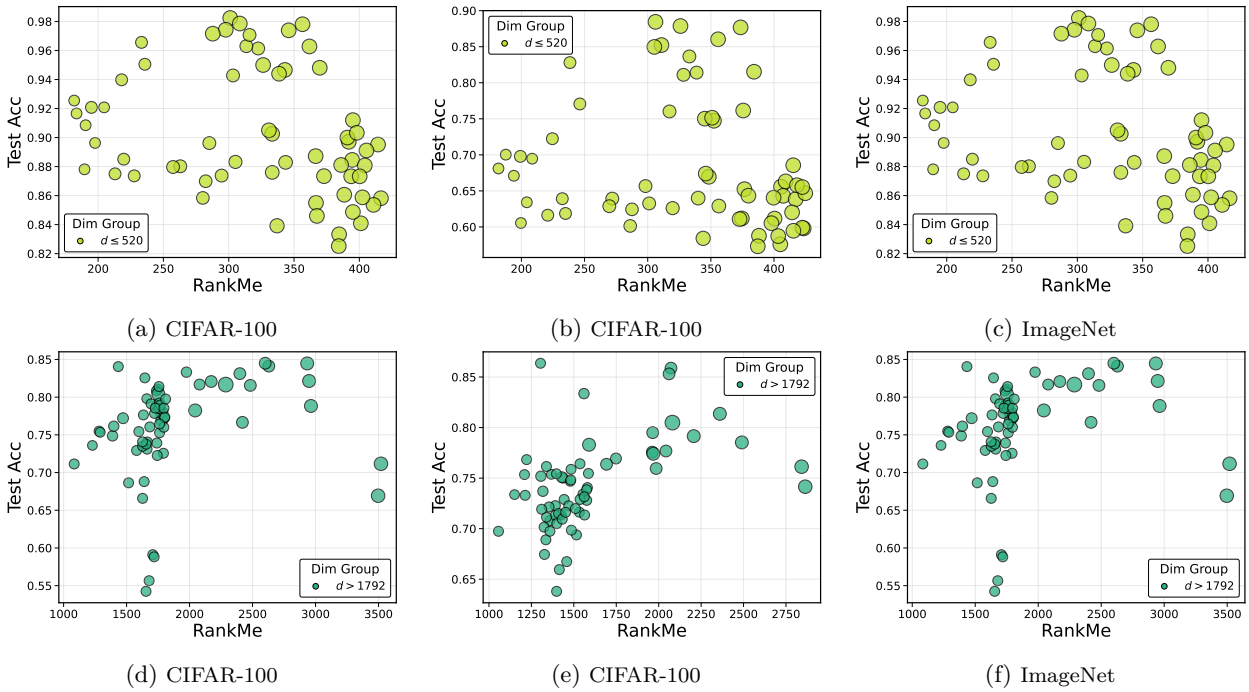
			$d \leq 520$	$520 < d \leq 960$	$960 < d \leq 1792$	$d > 1792$
CIFAR-10	**α-Req**	$\rightarrow 1$	0.1734	-0.1571	0.04738	0.1315
	RankMe	$\uparrow$	-0.3555^{**}	-0.02149	-0.1327	0.4376^{***}
	RankMe*	$\uparrow$	-0.3226^{**}	0.08077	-0.03157	-0.2487^{*}
	NE-Sum	$\uparrow$	-0.2765^{*}	0.1083	0.4189^{***}	0.1896
	κ	$\uparrow$	0.4829^{***}	0.1335	0.04447	0.04695
	Self-Cluster	$\downarrow$	0.3405^{**}	0.03767	-0.583^{***}	-0.1373
	DSE	$\uparrow$	-0.3379^{**}	-0.0316	0.5723^{***}	0.1593
	ID	$\downarrow$	-0.578^{***}	-0.1895	-0.8023^{***}	-0.6649^{***}
CIFAR-100	**α-Req**	$\rightarrow 1$	0.1731	-0.09751	-0.008382	0.1117
	RankMe	$\uparrow$	-0.2754^{*}	-0.03186	-0.05085	0.5141^{***}
	RankMe*	$\uparrow$	-0.2871^{*}	0.0889	0.04651	-0.2368
	NE-Sum	$\uparrow$	-0.101	0.1313	0.1457	0.1815
	κ	$\uparrow$	0.5506^{***}	0.133	0.1513	0.03697
	Self-Cluster	$\downarrow$	0.3401^{**}	0.01835	-0.5816^{***}	-0.0354
	DSE	$\uparrow$	-0.3344^{**}	-0.01253	0.587^{***}	0.06931
	ID	$\downarrow$	-0.6787^{***}	-0.4224^{***}	-0.8049^{***}	-0.6618^{***}
ImageNet	**α-Req**	$\rightarrow 1$	0.3707^{**}	-0.06594	-0.1894	-0.4712^{***}
	RankMe	$\uparrow$	-0.1344	-0.001592	0.0307	0.4758^{***}
	RankMe*	$\uparrow$	-0.436^{***}	0.02231	0.05674	0.1557
	NE-Sum	$\uparrow$	-0.2037	-0.01923	0.08408	0.3476^{**}
	κ	$\uparrow$	0.234	0.04222	0.02922	-0.01129
	Self-Cluster	$\downarrow$	0.2987^{*}	0.1644	-0.1078	-0.4651^{***}
	DSE	$\uparrow$	-0.2758^{*}	-0.1461	0.1191	0.4824^{***}
	ID	$\downarrow$	-0.5015^{***}	-0.4952^{***}	-0.8368^{***}	-0.6808^{***}

5 Discussion

Based on the taxonomy in Section 3 and the analysis in Section 4, we observe how current label-free metrics follow some notable patterns in behaviour, which we further discuss in this section, along with some of the important limitations of our study.

Spectral metrics are generally not reliable The spectral metrics exhibit an indiscernible relationship with test accuracy and, in certain cases, exhibit a correlation sign that is opposite to what is expected. For example, RankMe* exhibits a significant negative correlation in all cases except for the SSL objective, where it displays a positive correlation. This indicates that RankMe* is not a reliable indicator of performance and only captures the extent of dimensional utilisation, and may not be sufficient to capture downstream performance. Although the aggregate results in Table 2, indicate a weak correlation for RankMe, architecture grouping revealed strong correlation scores (Table 2). This indicates that the relationship is more nuanced, and that it depends to a large extent on the architecture class.

The hypothesis of Agrawal et al. (2022) that $\alpha \approx 1$ is indicative of high-quality representations does not generalise in our evaluation. In the $\alpha < 1$ regime, α-ReQ shows a statistically significant negative correlation with accuracy in CIFAR-100 and ImageNet, directly contradicting this hypothesis. α-ReQ does not exhibit a meaningful correlation in most groupings except when grouped by training objective. When visually inspecting the correlations, Fig. 9 indicates a negative correlation for SSL models. However, beyond



(a) CIFAR-100 (b) CIFAR-100 (c) ImageNet

(d) CIFAR-100 (e) CIFAR-100 (f) ImageNet

Figure 10: RankMe metric value plotted against the test accuracy for all datasets for $d \leq 520$ in the top row and $d > 1792$ in the bottom row. Although Table 5 indicated a negative correlation for RankMe in $d \leq 520$, Figs. 10a, 10b and 10c illustrates that visually this correlation is due to a concentration of points around RankMe value 400.

predicting downstream performance, the α values provide insight into underlying representation geometry. Representations for ImageNet have lower α values than the CIFAR datasets, indicating that more complex tasks induce representations that make better use of feature space.

Relation-based metrics are effective for SSL-based representations The near perfect negative correlation between DSE and Self-Cluster ($\rho \approx -0.999$) confirms the theoretical prediction established in Section 3: under ℓ_2-normalisation and the default σ, both metrics are functions of $\tilde{X}\tilde{X}^\top$. Both metrics are strong predictors for SSL models but are uninformative for supervised models, suggesting that they capture properties specific to geometry induced by contrastive and self-distillation objectives. Although they are reliable predictors in the $d \leq 520$ bin for all datasets, significant scores also appear in the high dimensional range, but this range appears to increase with dataset complexity (see Section 4.2). Specifically, in the CIFAR datasets, reliable scores appear in the range $960 < d \leq 1792$, whereas for ImageNet they emerge only for $d > 1792$. However, it is not immediately clear whether this is purely an artefact of ImageNet and requires further analysis. At the architecture level, they are reliable predictors only for the RegNet and ResNet models, and occupy extreme values for the ViT models. ViT models displayed Self-Cluster values near zero for models trained with SSL objective and near one for supervised models. This is understandable, since SSL representations are encouraged to be uniformly spread (Wang & Isola, 2020) and supervised models cluster representation for classification. This suggests that these metrics are sensitive to interactions between architecture and training objective.

Manifold-based metrics show the best overall generalisation Ansuini et al. (2019) used this metric to validate the claim that the intrinsic dimensionality of the representation in convolutional models progressively decreases through the layers. Since lower ID at the final layer was found to correlate with better generalisation (Ansuini et al., 2019), ID has been used as a proxy for representation quality. In general, Tables. 3, 4, and 5 suggest that ID remains a reliable predictor across all architecture classes, training objectives, and representation dimension groups, with correlation scores in the range $\rho \in [-0.84, -0.42]$

indicating a moderate to strong negative relationship with test accuracy. Although their analysis focused on supervised models, Table 4 shows that ID is also a reliable predictor of performance on SSL models. The robustness of ID indicates that it measures a property of the representation manifold that is relatively invariant to architectural choices and training procedures. This is in contrast to spectral and relational metrics, whose reliability was shown in Section 4.2 to be strongly moderated by architecture class and training objective.

Limitations We use an OLS fit to circumvent the long training time that linear probes take. Although this removes the stochasticity in test accuracies, converting a classification task into a regression task is a strong assumption. Also, OLS assesses only linear separability of representations, and thus the relationship between these metrics and performance with a non-linear head remains to be analysed. We consider only 19 SSL+Sup models, which makes conclusions about this group unreliable. A detailed study with a larger and more balanced model pool would be needed to better characterise their relationship with both SSL and supervised models. Finally, our experiments are conducted solely on vision models and image classification tasks, and their correlation with other modalities or task types needs to be explored.

6 Conclusion

Our work provides a systematic comparative evaluation of label-free representation quality metrics. We catalogue and group the metrics based on their construction into (i) spectral-based, (ii) pairwise relation-based, and (iii) manifold-based families and reformulate them where possible to highlight similarities. We further investigate the sensitivity of spectral metrics to spectral shape and representation dimension. Our analysis of 260 models shows that the reliability of these metrics is strongly influenced by architecture class, training objective, and representation dimension d, which were obscured in the aggregate results (see Table 2). Across most groupings, **ID** is the only metric that displays a significant correlation with test accuracy. Since the different types of metrics capture different aspects of the representation space, their complementary nature could potentially be explored by combining different metrics, such as in CLID (Lu et al., 2023), and we consider this as an interesting direction for future work.

References

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. \alpha-req : Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/70596d70542c51c8d9b4e423f4bf2736-Paper-Conference.pdf.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cfcce0621b49c983991ead4c3d4d3b6b-Paper.pdf.

Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=04K3PMtMckp.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson,

Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907378117.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Nan Ding, Xi Chen, Tomer Levinboim, Soravit Changpinyo, and Radu Soricut. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *European Conference on Computer Vision*, pp. 252–268, 2022.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 647–655, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/donahue14.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank, 2023. URL https://openreview.net/forum?id=uGEBxC8dnEh.

Bobby He and Mete Ozay. Exploring the gap between collapsed &; whitened features in self-supervised learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8613–8634. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/he22c.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Danqi Liao, Chen Liu, Benjamin W Christensen, Alexander Tong, Guillaume Huguet, Guy Wolf, Maximilian Nickel, Ian Adelstein, and Smita Krishnaswamy. Assessing neural network representations during training using noise-resilient diffusion spectral entropy. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2024.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Yuchen Lu, Zhen Liu, Aristide Baratin, Romain Laroche, Aaron Courville, and Alessandro Sordoni. Using representation expressiveness and learnability to evaluate self-supervised learning methods. *Transactions on Machine Learning Research*, 2023.

Josue Nassar, Piotr Sokol, Sueyeon Chung, Kenneth D Harris, and Il Memming Park. On 1/n neural representation and robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6211–6222. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/44bf89b63173d40fb39f9842e308b3f9-Paper.pdf`.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International conference on machine learning*, pp. 7294–7305. PMLR, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

Cedric Renggli, André Susano Pinto, Luka Rimanic, Joan Puigcerver, Carlos Riquelme, Ce Zhang, and Mario Lučić. Which model to transfer? finding the needle in the growing haystack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9205–9214, 2022.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In Timothy Doster, Tegan Emerson, Henry Kvinge, Nina Miolane, Mathilde Papillon, Bastian Rieck, and Sophia Sanborn (eds.), *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 169–188. PMLR, 28 Jul 2023. URL `https://proceedings.mlr.press/v221/tsitsulin23a.html`.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/wang20k.html`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Zeke Xie, Qian-Yuan Tang, Yunfeng Cai, Mingming Sun, and Ping Li. On the power-law spectrum in deep learning: A bridge to protein science. *arXiv preprint arXiv:2201.13011*, 2, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. PMLR, 2021.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Appendix

A Approximation of DSE

For ℓ_2-normalised representations, $\|x_i - x_j\|^2 = 2 - 2x_i \cdot x_j$, so:

$$G_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1 - x_i \cdot x_j}{\sigma^2}\right) = C \cdot \exp\left(\frac{x_i \cdot x_j}{\sigma^2}\right)$$

where $C = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{\sigma^2}\right)$ is a constant. We work with a default setting of $\sigma = 10$ for which the first-order Taylor expansion is $\exp\left(\frac{x_i \cdot x_j}{\sigma^2}\right) \approx 1 + \frac{x_i \cdot x_j}{\sigma^2}$, which is a good approximation, since $\frac{|x_i \cdot x_j|}{\sigma^2} \leq 0.01$ and $|x_i \cdot x_j| \leq 1$.

$$G \approx C\left(\mathbf{1}_{N\times N} + \frac{1}{\sigma^2}\tilde{X}\tilde{X}^\top\right)$$

B Models used

Table 6: Summary of torch hub models and their characteristics

GitHub Repo	Model Name	d	Arch.	Objective
facebookresearch/barlowtwins:main	barlowtwins	1000	ResNet	SSL
facebookresearch/vicregl:main	resnet50_alpha0p9	2048	ResNet	SSL
facebookresearch/vicregl:main	resnet50_alpha0p75	2048	ResNet	SSL
facebookresearch/vicregl:main	convnext_small_alpha0p9	768	ConvNeXT	SSL
facebookresearch/vicregl:main	convnext_small_alpha0p75	768	ConvNeXT	SSL
facebookresearch/vicregl:main	convnext_base_alpha0p9	1024	ConvNeXT	SSL
facebookresearch/vicregl:main	convnext_base_alpha0p75	1024	ConvNeXT	SSL
facebookresearch/vicreg:main	resnet50	2048	ResNet	SSL
facebookresearch/vicreg:main	resnet50x2	4096	ResNet	SSL
facebookresearch/vicreg:main	resnet200x2	4096	ResNet	SSL

Table 7: Summary of Hugging face (Wolf et al., 2020) models and their characteristics

Model Name	d	Arch.	Objective
facebook/dinov2-small	384	ViT	SSL
timm/vit_small_patch16_224.dino	384	ViT	SSL

Continued on next page

Model Name	d	Arch.	Objective
timm/vit_small_patch16_dinov3_qkvb.lvd1689m	384	ViT	SSL
timm/vit_small_plus_patch16_dinov3_qkvb.lvd1689m	384	ViT	SSL
facebook/dinov2-base	768	ViT	SSL
timm/vit_base_patch16_224.dino	768	ViT	SSL
timm/vit_base_patch16_dinov3_qkvb.lvd1689m	768	ViT	SSL
timm/vit_base_patch8_224.dino	768	ViT	SSL
facebook/vit-mae-base	768	ViT	SSL
timm/vit_so150m_patch16_reg4_gap_256.sbb_e250_in12k_ft_in1k	896	ViT	SSL
timm/vit_so150m_patch16_reg4_gap_384.sbb_e250_in12k_ft_in1k	896	ViT	SSL
facebook/dinov2-large	1024	ViT	SSL
facebook/webssl-dino300m-full2b-224	1024	ViT	SSL
timm/vit_large_patch16_dinov3_qkvb.lvd1689m	1024	ViT	SSL
facebook/vit-mae-large	1024	ViT	SSL
facebook/webssl-mae300m-full2b-224	1024	ViT	SSL
facebook/vit-mae-huge	1280	ViT	SSL
facebook/webssl-mae700m-full2b-224	1280	ViT	SSL
timm/eva_giant_patch14_224.clip_ft_in1k	1408	ViT	SSL
timm/convnext_small.dinov3_lvd1689m	768	ConvNeXT	SSL
timm/convnext_tiny.dinov3_lvd1689m	768	ConvNeXT	SSL
timm/convnextv2_tiny.fcmae	768	ConvNeXT	SSL
timm/convnext_base.dinov3_lvd1689m	1024	ConvNeXT	SSL
timm/convnext_large.dinov3_lvd1689m	1536	ConvNeXT	SSL
timm/vicregl_convnext_xlarge_alpha0p75	2048	ConvNeXT	SSL
timm/vit_base_patch16_clip_384.laion2b_ft_in12k_in1k	768	ViT	SSL+Sup
timm/eva_large_patch14_196.in22k_ft_in22k_in1k	1024	ViT	SSL+Sup
timm/vit_large_patch14_clip_224.openai_ft_in12k_in1k	1024	ViT	SSL+Sup
timm/convnextv2_atto.fcmae_ft_in1k	320	ConvNeXT	SSL+Sup
timm/convnextv2_femto.fcmae_ft_in1k	384	ConvNeXT	SSL+Sup
timm/convnextv2_pico.fcmae_ft_in1k	512	ConvNeXT	SSL+Sup
timm/convnextv2_nano.fcmae_ft_in1k	640	ConvNeXT	SSL+Sup
timm/convnextv2_nano.fcmae_ft_in22k_in1k	640	ConvNeXT	SSL+Sup
timm/convnextv2_nano.fcmae_ft_in22k_in1k_384	640	ConvNeXT	SSL+Sup
timm/convnextv2_tiny.fcmae_ft_in1k	768	ConvNeXT	SSL+Sup
timm/convnextv2_tiny.fcmae_ft_in22k_in1k	768	ConvNeXT	SSL+Sup
timm/convnextv2_tiny.fcmae_ft_in22k_in1k_384	768	ConvNeXT	SSL+Sup
timm/convnextv2_base.fcmae_ft_in1k	1024	ConvNeXT	SSL+Sup
timm/convnextv2_base.fcmae_ft_in22k_in1k	1024	ConvNeXT	SSL+Sup
timm/convnextv2_base.fcmae_ft_in22k_in1k_384	1024	ConvNeXT	SSL+Sup
timm/convnext_base.clip_laion2b_augreg_ft_in12k_in1k	1024	ConvNeXT	SSL+Sup
timm/convnext_base.clip_laion2b_augreg_ft_in12k_in1k_384	1024	ConvNeXT	SSL+Sup
timm/convnext_base.clip_laion2b_augreg_ft_in1k	1024	ConvNeXT	SSL+Sup
timm/convnext_base.clip_laiona_augreg_ft_in1k_384	1024	ConvNeXT	SSL+Sup
timm/vit_dpwee_patch16_reg1_gap_256.sbb_in1k	256	ViT	Supervised
timm/vit_dpwee_patch16_reg1_gap_256.sbb_nadamuon_in1k	256	ViT	Supervised
timm/vit_dwee_patch16_reg1_gap_256.sbb_in1k	256	ViT	Supervised
timm/vit_dwee_patch16_reg1_gap_256.sbb_nadamuon_in1k	256	ViT	Supervised
timm/vit_pwee_patch16_reg1_gap_256.sbb_in1k	256	ViT	Supervised
timm/vit_wee_patch16_reg1_gap_256.sbb_in1k	256	ViT	Supervised
timm/vit_dlittle_patch16_reg1_gap_256.sbb_nadamuon_in1k	320	ViT	Supervised
timm/vit_little_patch16_reg1_gap_256.sbb_in12k	320	ViT	Supervised
timm/vit_little_patch16_reg1_gap_256.sbb_in12k_ft_in1k	320	ViT	Supervised
timm/vit_little_patch16_reg4_gap_256.sbb_in1k	320	ViT	Supervised
timm/vit_medium_patch16_reg1_gap_256.sbb_in1k	512	ViT	Supervised
timm/vit_medium_patch16_reg4_gap_256.sbb_in12k	512	ViT	Supervised
timm/vit_medium_patch16_reg4_gap_256.sbb_in12k_ft_in1k	512	ViT	Supervised
timm/vit_medium_patch16_reg4_gap_256.sbb_in1k	512	ViT	Supervised
timm/vit_medium_patch16_rope_reg1_gap_256.sbb_in1k	512	ViT	Supervised
timm/vit_mediumd_patch16_reg4_gap_256.sbb2_e200_in12k	512	ViT	Supervised
timm/vit_mediumd_patch16_reg4_gap_256.sbb2_e200_in12k_ft_in1k	512	ViT	Supervised
timm/vit_mediumd_patch16_reg4_gap_256.sbb_in12k	512	ViT	Supervised
timm/vit_mediumd_patch16_reg4_gap_256.sbb_in12k_ft_in1k	512	ViT	Supervised
timm/vit_mediumd_patch16_reg4_gap_384.sbb2_e200_in12k_ft_in1k	512	ViT	Supervised
timm/vit_mediumd_patch16_rope_reg1_gap_256.sbb_in1k	512	ViT	Supervised
timm/vit_betwixt_patch16_reg1_gap_256.sbb_in1k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_256.sbb2_e200_in12k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_256.sbb2_e200_in12k_ft_in1k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_256.sbb_in12k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_256.sbb_in12k_ft_in1k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_256.sbb_in1k	640	ViT	Supervised
timm/vit_betwixt_patch16_reg4_gap_384.sbb2_e200_in12k_ft_in1k	640	ViT	Supervised
timm/vit_betwixt_patch16_rope_reg4_gap_256.sbb_in1k	640	ViT	Supervised
timm/vit_base_patch16_rope_reg1_gap_256.sbb_in1k	768	ViT	Supervised
timm/vit_so150m2_patch16_reg1_gap_256.sbb_e200_in12k	832	ViT	Supervised
timm/vit_so150m2_patch16_reg1_gap_256.sbb_e200_in12k_ft_in1k	832	ViT	Supervised
timm/vit_so150m2_patch16_reg1_gap_384.sbb_e200_in12k_ft_in1k	832	ViT	Supervised
timm/vit_large_patch32_224.orig_in21k	1024	ViT	Supervised
timm/resnet10t.c3_in1k	512	ResNet	Supervised
timm/resnet18.a1_in1k	512	ResNet	Supervised
timm/resnet18.a2_in1k	512	ResNet	Supervised
timm/resnet18.a3_in1k	512	ResNet	Supervised
timm/resnet18.fb_ssl_yfcc100m_ft_in1k	512	ResNet	Supervised
timm/resnet18.fb_swsl_ig1b_ft_in1k	512	ResNet	Supervised
timm/resnet18.gluon_in1k	512	ResNet	Supervised
timm/resnet18.tv_in1k	512	ResNet	Supervised
timm/resnet18d.ra2_in1k	512	ResNet	Supervised
timm/resnet18d.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnet34.a1_in1k	512	ResNet	Supervised
timm/resnet34.a2_in1k	512	ResNet	Supervised
timm/resnet34.a3_in1k	512	ResNet	Supervised
timm/resnet34.bt_in1k	512	ResNet	Supervised
timm/resnet34.gluon_in1k	512	ResNet	Supervised
timm/resnet34.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnet34.tv_in1k	512	ResNet	Supervised
timm/resnet34d.ra2_in1k	512	ResNet	Supervised
timm/resnetv2_18.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnetv2_18d.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnetv2_34.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnetv2_34d.ra4_e3600_r224_in1k	512	ResNet	Supervised
timm/resnet33ts.ra2_in1k	1280	ResNet	Supervised
timm/resnet32ts.ra2_in1k	1536	ResNet	Supervised
timm/resnet101.a1h_in1k	2048	ResNet	Supervised
timm/resnet101.a3_in1k	2048	ResNet	Supervised
timm/resnet101.tv2_in1k	2048	ResNet	Supervised
timm/resnet14t.c3_in1k	2048	ResNet	Supervised

Continued on next page

Model Name	d	Arch.	Objective
timm/resnet26.bt_in1k	2048	ResNet	Supervised
timm/resnet26d.bt_in1k	2048	ResNet	Supervised
timm/resnet26t.ra2_in1k	2048	ResNet	Supervised
timm/resnet50.a1_in1k	2048	ResNet	Supervised
timm/resnet50.a1h_in1k	2048	ResNet	Supervised
timm/resnet50.a2_in1k	2048	ResNet	Supervised
timm/resnet50.a3_in1k	2048	ResNet	Supervised
timm/resnet50.am_in1k	2048	ResNet	Supervised
timm/resnet50.b1k_in1k	2048	ResNet	Supervised
timm/resnet50.b2k_in1k	2048	ResNet	Supervised
timm/resnet50.bt_in1k	2048	ResNet	Supervised
timm/resnet50.c1_in1k	2048	ResNet	Supervised
timm/resnet50.c2_in1k	2048	ResNet	Supervised
timm/resnet50.d_in1k	2048	ResNet	Supervised
timm/resnet50.fb_ssl_yfcc100m_ft_in1k	2048	ResNet	Supervised
timm/resnet50.fb_swsl_ig1b_ft_in1k	2048	ResNet	Supervised
timm/resnet50.gluon_in1k	2048	ResNet	Supervised
timm/resnet50.ra_in1k	2048	ResNet	Supervised
timm/resnet50.ram_in1k	2048	ResNet	Supervised
timm/resnet50.tv2_in1k	2048	ResNet	Supervised
timm/resnet50.tv_in1k	2048	ResNet	Supervised
timm/resnet50_gn.a1h_in1k	2048	ResNet	Supervised
timm/resnet50c.gluon_in1k	2048	ResNet	Supervised
timm/resnet50d.a1_in1k	2048	ResNet	Supervised
timm/resnet50d.a2_in1k	2048	ResNet	Supervised
timm/resnet50d.a3_in1k	2048	ResNet	Supervised
timm/resnet50d.gluon_in1k	2048	ResNet	Supervised
timm/resnet50d.ra2_in1k	2048	ResNet	Supervised
timm/resnet50d.ra4_e3600_r224_in1k	2048	ResNet	Supervised
timm/resnetrs50.tf_in1k	2048	ResNet	Supervised
timm/regnetx_002.pycls_in1k	368	RegNet	Supervised
timm/regnety_002.pycls_in1k	368	RegNet	Supervised
timm/regnetx_004.pycls_in1k	384	RegNet	Supervised
timm/regnetx_004_tv.tv2_in1k	400	RegNet	Supervised
timm/regnety_004.pycls_in1k	440	RegNet	Supervised
timm/regnety_004.tv2_in1k	440	RegNet	Supervised
timm/regnetx_006.pycls_in1k	528	RegNet	Supervised
timm/regnetz_040.ra3_in1k	528	RegNet	Supervised
timm/regnety_006.pycls_in1k	608	RegNet	Supervised
timm/regnetx_008.pycls_in1k	672	RegNet	Supervised
timm/regnetx_008.tv2_in1k	672	RegNet	Supervised
timm/regnety_008.pycls_in1k	768	RegNet	Supervised
timm/regnety_008_tv.tv2_in1k	784	RegNet	Supervised
timm/regnety_016.pycls_in1k	888	RegNet	Supervised
timm/regnety_016.tv2_in1k	888	RegNet	Supervised
timm/regnetx_016.pycls_in1k	912	RegNet	Supervised
timm/regnetx_016.tv2_in1k	912	RegNet	Supervised
timm/regnetx_032.pycls_in1k	1008	RegNet	Supervised
timm/regnetx_032.tv2_in1k	1008	RegNet	Supervised
timm/regnetv_040.ra3_in1k	1088	RegNet	Supervised
timm/regnety_040.pycls_in1k	1088	RegNet	Supervised
timm/regnety_040.ra3_in1k	1088	RegNet	Supervised
timm/regnetv_064.ra3_in1k	1296	RegNet	Supervised
timm/regnety_064.pycls_in1k	1296	RegNet	Supervised
timm/regnety_064.ra3_in1k	1296	RegNet	Supervised
timm/regnetx_040.pycls_in1k	1360	RegNet	Supervised
timm/regnety_032.pycls_in1k	1512	RegNet	Supervised
timm/regnety_032.ra_in1k	1512	RegNet	Supervised
timm/regnety_032.tv2_in1k	1512	RegNet	Supervised
timm/regnetz_040_h.ra3_in1k	1536	RegNet	Supervised
timm/regnetz_b16.ra3_in1k	1536	RegNet	Supervised
timm/regnetz_c16.ra3_in1k	1536	RegNet	Supervised
timm/regnetz_c16_evos.ch_in1k	1536	RegNet	Supervised
timm/regnetx_064.pycls_in1k	1624	RegNet	Supervised
timm/regnetz_d32.ra3_in1k	1792	RegNet	Supervised
timm/regnetz_d8.ra3_in1k	1792	RegNet	Supervised
timm/regnetz_d8_evos.ch_in1k	1792	RegNet	Supervised
timm/regnetx_080.pycls_in1k	1920	RegNet	Supervised
timm/regnetx_080.tv2_in1k	1920	RegNet	Supervised
timm/regnety_080.pycls_in1k	2016	RegNet	Supervised
timm/regnety_080.ra3_in1k	2016	RegNet	Supervised
timm/regnety_080_tv.tv2_in1k	2016	RegNet	Supervised
timm/regnetx_160.pycls_in1k	2048	RegNet	Supervised
timm/regnetx_160.tv2_in1k	2048	RegNet	Supervised
timm/regnetz_e8.ra3_in1k	2048	RegNet	Supervised
timm/regnetx_120.pycls_in1k	2240	RegNet	Supervised
timm/regnety_120.pycls_in1k	2240	RegNet	Supervised
timm/regnety_120.sw_in12k_ft_in1k	2240	RegNet	Supervised
timm/regnetx_320.pycls_in1k	2520	RegNet	Supervised
timm/regnetx_320.tv2_in1k	2520	RegNet	Supervised
timm/regnety_160.deit_in1k	3024	RegNet	Supervised
timm/regnety_160.lion_in12k_ft_in1k	3024	RegNet	Supervised
timm/regnety_160.pycls_in1k	3024	RegNet	Supervised
timm/regnety_160.sw_in12k_ft_in1k	3024	RegNet	Supervised
timm/regnety_160.swag_ft_in1k	3024	RegNet	Supervised
timm/regnety_160.swag_lc_in1k	3024	RegNet	Supervised
timm/regnety_160.tv2_in1k	3024	RegNet	Supervised
timm/regnety_320.pycls_in1k	3712	RegNet	Supervised
timm/regnety_320.seer_ft_in1k	3712	RegNet	Supervised
timm/regnety_320.swag_ft_in1k	3712	RegNet	Supervised
timm/regnety_320.swag_lc_in1k	3712	RegNet	Supervised
timm/regnety_320.tv2_in1k	3712	RegNet	Supervised
timm/regnety_640.seer_ft_in1k	4920	RegNet	Supervised
timm/mobilenetv4_conv_small_050.e3000_r224_in1k	480	others	Supervised
timm/convformer_m36.sail_in22k_ft_in1k	576	others	Supervised
timm/mobilenetv4_conv_aa_large.e230_r384_in12k_ft_in1k	960	others	Supervised
timm/mobilenetv4_conv_aa_large.e230_r448_in12k_ft_in1k	960	others	Supervised
timm/mobilenetv4_conv_aa_large.e600_r384_in1k	960	others	Supervised
timm/mobilenetv4_conv_blur_medium.e500_r224_in1k	960	others	Supervised
timm/mobilenetv4_conv_large.e500_r256_in1k	960	others	Supervised
timm/mobilenetv4_conv_large.e600_r384_in1k	960	others	Supervised
timm/mobilenetv4_conv_medium.e500_r224_in1k	960	others	Supervised
timm/mobilenetv4_conv_medium.e500_r256_in1k	960	others	Supervised
timm/mobilenetv4_conv_small.e1200_r224_in1k	960	others	Supervised

Continued on next page

Model Name	d	Arch.	Objective
timm/mobilenetv4_conv_small.e2400_r224_in1k	960	others	Supervised
timm/beitv2_large_patch16_224.in1k_ft_in22k_in1k	1024	others	Supervised
timm/mobilenetv1_100.ra4_e3600_r224_in1k	1024	others	Supervised
timm/mobilenetv1_100h.ra4_e3600_r224_in1k	1024	others	Supervised
timm/mobilenetv1_125.ra4_e3600_r224_in1k	1280	others	Supervised
timm/mobilenetv3_large_100.ra4_e3600_r224_in1k	1280	others	Supervised
timm/mobilenetv3_large_100.ra_in1k	1280	others	Supervised
timm/mobilenetv3_large_150d.ra4_e3600_r256_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_large.e600_r384_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_large.ix_e600_r384_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_medium.e200_r256_in12k_ft_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_medium.e500_r224_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_medium.ix_e550_r256_in1k	1280	others	Supervised
timm/mobilenetv4_hybrid_medium.ix_e550_r384_in1k	1280	others	Supervised
timm/mobilenet_edgetpu_v2_m.ra4_e3600_r224_in1k	1344	others	Supervised
timm/convnext_atto.d2_in1k	320	ConvNeXT	Supervised
timm/convnext_atto_ols.a2_in1k	320	ConvNeXT	Supervised
timm/convnext_femto.d1_in1k	384	ConvNeXT	Supervised
timm/convnext_femto_ols.d1_in1k	384	ConvNeXT	Supervised
timm/convnext_pico.d1_in1k	512	ConvNeXT	Supervised
timm/convnext_pico_ols.d1_in1k	512	ConvNeXT	Supervised
timm/convnext_nano.d1h_in1k	640	ConvNeXT	Supervised
timm/convnext_nano.in12k	640	ConvNeXT	Supervised
timm/convnext_nano.in12k_ft_in1k	640	ConvNeXT	Supervised
timm/convnext_nano_ols.d1h_in1k	640	ConvNeXT	Supervised
timm/convnext_small.fb_in1k	768	ConvNeXT	Supervised
timm/convnext_small.fb_in22k_ft_in1k	768	ConvNeXT	Supervised
timm/convnext_small.fb_in22k_ft_in1k_384	768	ConvNeXT	Supervised
timm/convnext_small.in12k	768	ConvNeXT	Supervised
timm/convnext_small.in12k_ft_in1k	768	ConvNeXT	Supervised
timm/convnext_small.in12k_ft_in1k_384	768	ConvNeXT	Supervised
timm/convnext_tiny.fb_in1k	768	ConvNeXT	Supervised
timm/convnext_tiny.fb_in22k_ft_in1k	768	ConvNeXT	Supervised
timm/convnext_tiny.fb_in22k_ft_in1k_384	768	ConvNeXT	Supervised
timm/convnext_tiny.in12k	768	ConvNeXT	Supervised
timm/convnext_tiny.in12k_ft_in1k	768	ConvNeXT	Supervised
timm/convnext_tiny.in12k_ft_in1k_384	768	ConvNeXT	Supervised
timm/convnext_tiny_hnf.a2h_in1k	768	ConvNeXT	Supervised
timm/convnext_base.fb_in1k	1024	ConvNeXT	Supervised
timm/convnext_base.fb_in22k_ft_in1k	1024	ConvNeXT	Supervised
timm/convnext_base.fb_in22k_ft_in1k_384	1024	ConvNeXT	Supervised

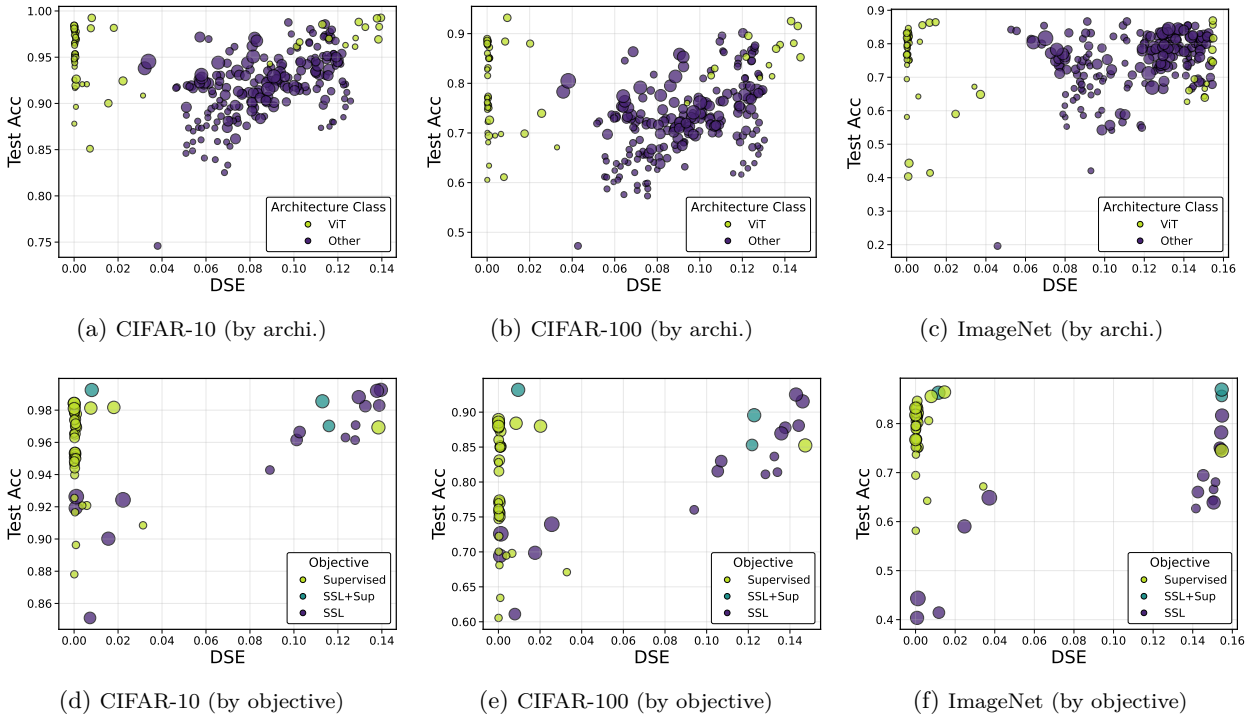
C OLS Fit

C.1 Results with SSL+Supervised models

Table 8: Spearman correlation (ρ) between representation metrics and test accuracy grouped by training objective across CIFAR-10, CIFAR-100, and ImageNet-1k. Metrics are grouped by construction: spectral (top), relation based (middle), and manifold-based (bottom). Arrows indicate the preferred direction: $\uparrow$ (higher is better), $\downarrow$ (lower is better), $\rightarrow 1$ (closer to 1 is better). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

			SSL	SSL+Sup	Supervised
CIFAR-10	**α-Req**	$\rightarrow$ **1**	-0.4703^{**}	0.05614	0.1877^{**}
	RankMe	$\uparrow$	0.09054	0.8105^{***}	0.1503^{*}
	RankMe*	$\uparrow$	0.4985^{**}	-0.2667	-0.3683^{***}
	NE-Sum	$\uparrow$	0.5942^{***}	0.3614	0.0295
	κ	$\uparrow$	-0.06452	0.7579^{***}	0.294^{***}
	Self-Cluster	$\downarrow$	-0.7885^{***}	0.5596^{*}	0.003576
	DSE	$\uparrow$	0.7903^{***}	-0.5439^{*}	0.006315
	ID	$\downarrow$	-0.5249^{**}	-0.3351	-0.5387^{***}
CIFAR-100	**α-Req**	$\rightarrow$ **1**	-0.6162^{***}	0.05789	0.1634^{*}
	RankMe	$\uparrow$	0.2258	0.8158^{***}	0.3044^{***}
	RankMe*	$\uparrow$	0.6103^{***}	-0.2404	-0.3593^{***}
	NE-Sum	$\uparrow$	0.6617^{***}	-0.2947	0.1204
	κ	$\uparrow$	-0.07808	0.8649^{***}	0.3992^{***}
	Self-Cluster	$\downarrow$	-0.8134^{***}	0.4579^{*}	-0.02742
	DSE	$\uparrow$	0.8167^{***}	-0.2719	0.04236
	ID	$\downarrow$	-0.6855^{***}	-0.8158^{***}	-0.5586^{***}
ImageNet	**α-Req**	$\rightarrow$ **1**	-0.7038^{***}	-0.005263	-0.05376
	RankMe	$\uparrow$	0.331	0.6965^{***}	0.2411^{***}
	RankMe*	$\uparrow$	0.5033^{**}	-0.2386	-0.1564^{*}
	NE-Sum	$\uparrow$	0.4296^{*}	0.1561	0.08246
	κ	$\uparrow$	0.1554	0.8649^{***}	0.4054^{***}
	Self-Cluster	$\downarrow$	-0.7302^{***}	-0.003509	-0.1091
	DSE	$\uparrow$	0.7551^{***}	0.01228	0.1267
	ID	$\downarrow$	-0.4567^{**}	0.08772	-0.5948^{***}

D The curious case of ViT models



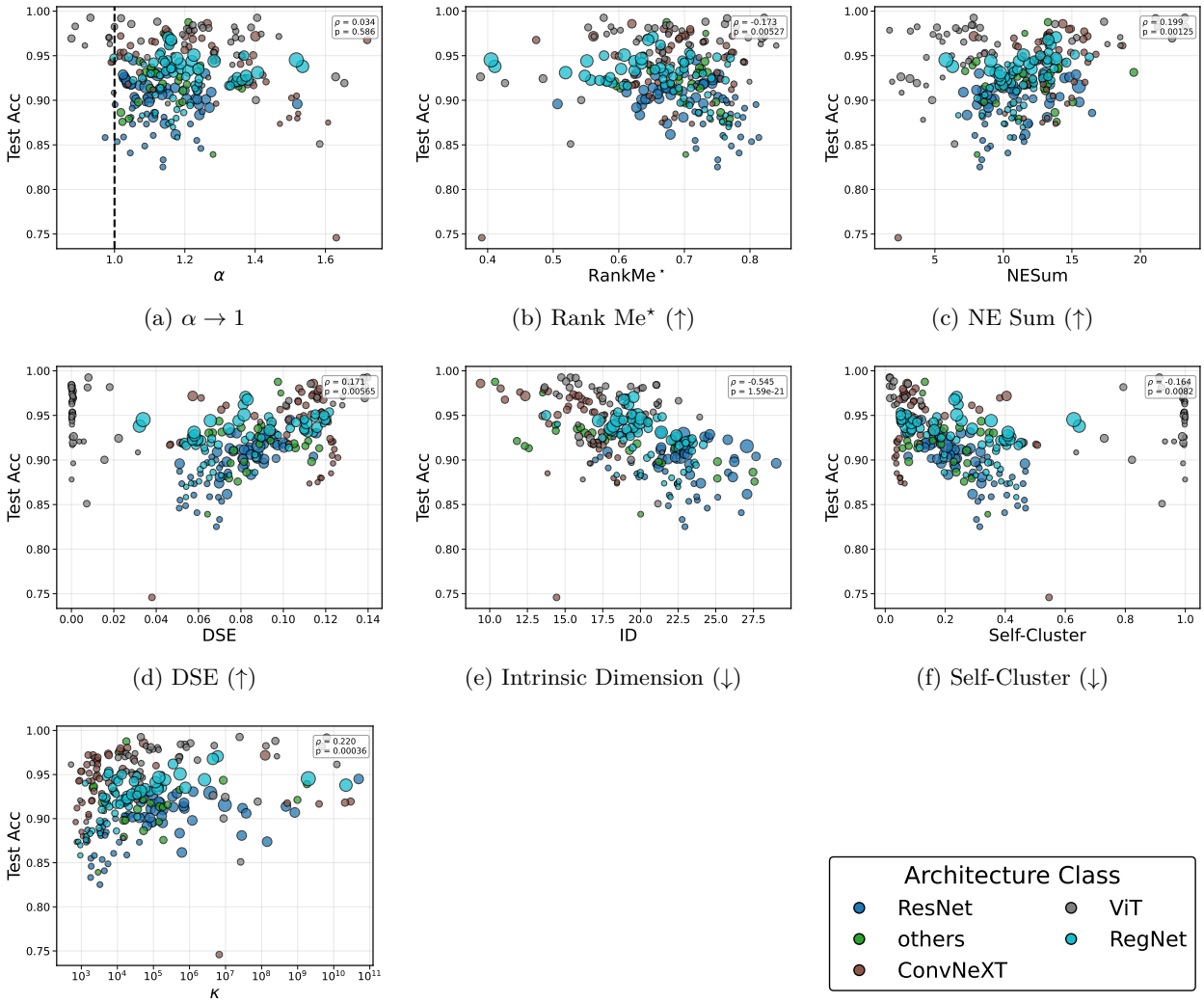
(a) CIFAR-10 (by archi.) (b) CIFAR-100 (by archi.) (c) ImageNet (by archi.)

(d) CIFAR-10 (by objective) (e) CIFAR-100 (by objective) (f) ImageNet (by objective)

Figure 11: DSE metric values plotted against test accuracy across datasets. The top row shows all models coloured by architecture class, revealing that the models with extreme values predominantly belong to the ViT architecture class. The bottom row shows only ViT models coloured by training objective, illustrating that the near zero DSE values originate predominantly from Supervised and SSL+Sup models rather than SSL models. This suggests that the DSE metric may be a reliable predictor for SSL models (though the sample size is small), but is unreliable for Supervised and SSL+Sup models.

D.1 By architecture

We visualise the results, grouping models by their architecture class

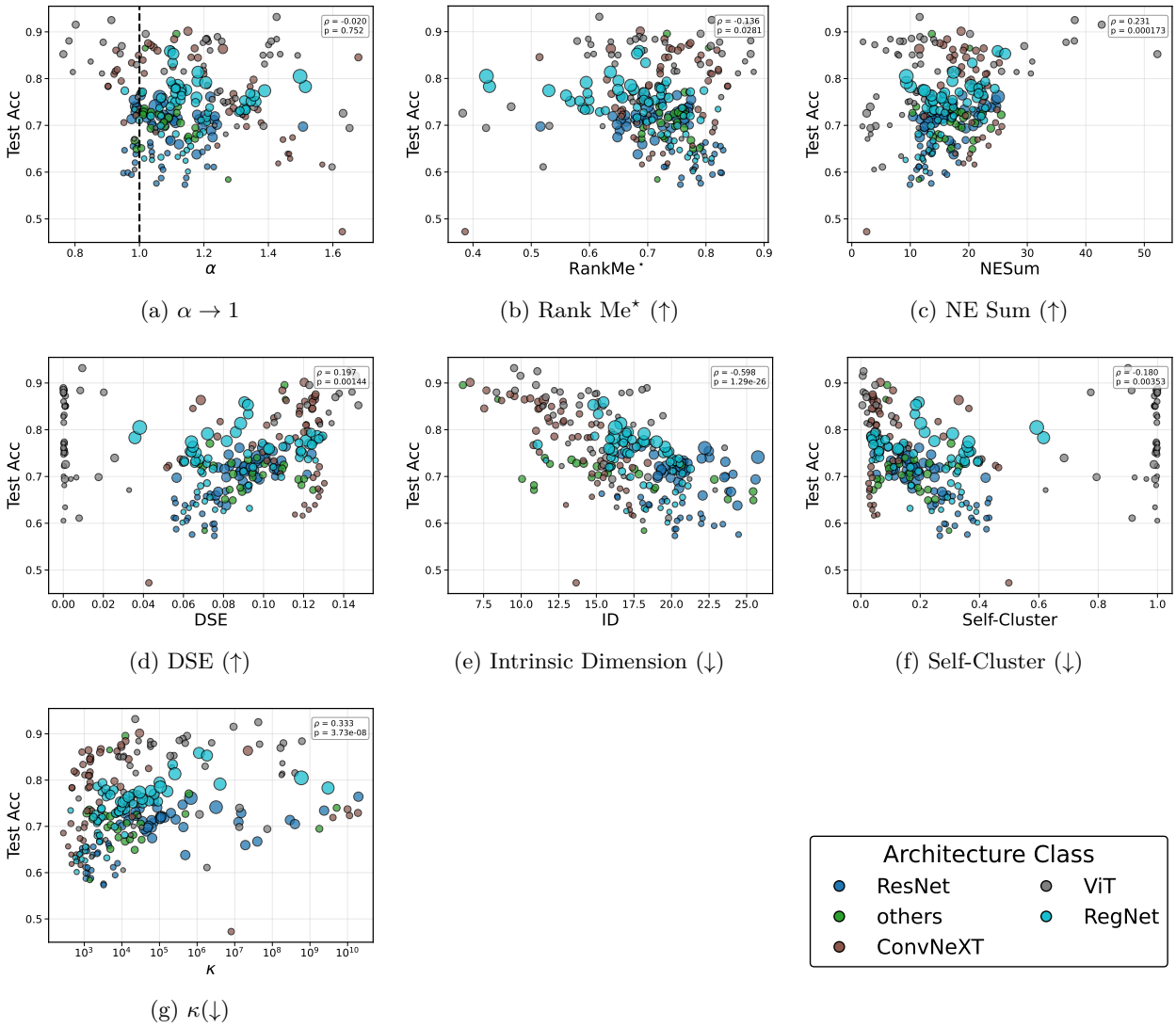


(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 12: Comparison of representation metrics and CIFAR-10 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 12a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.

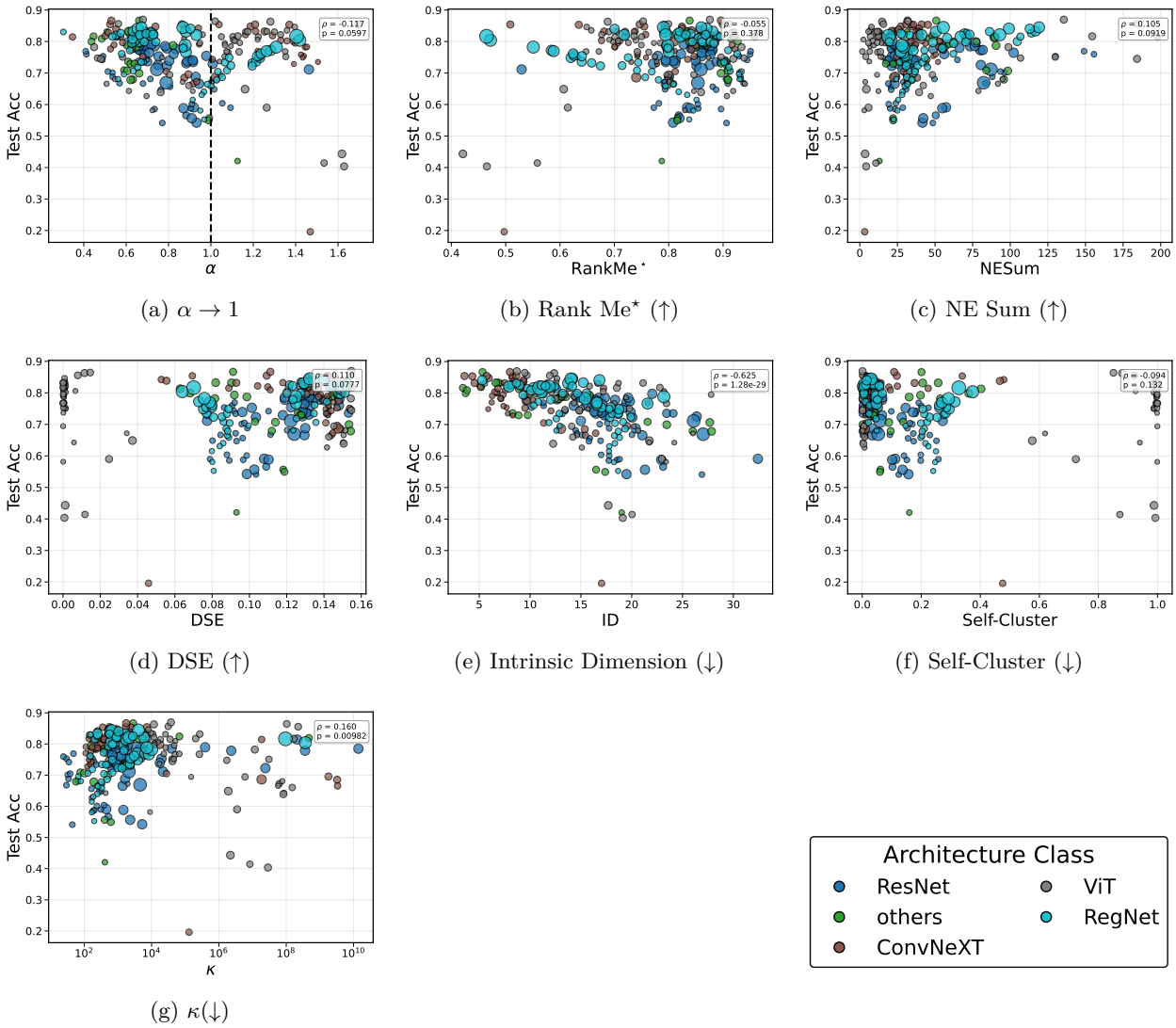


(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 13: Comparison of representation metrics and CIFAR-100 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 13a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.



(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

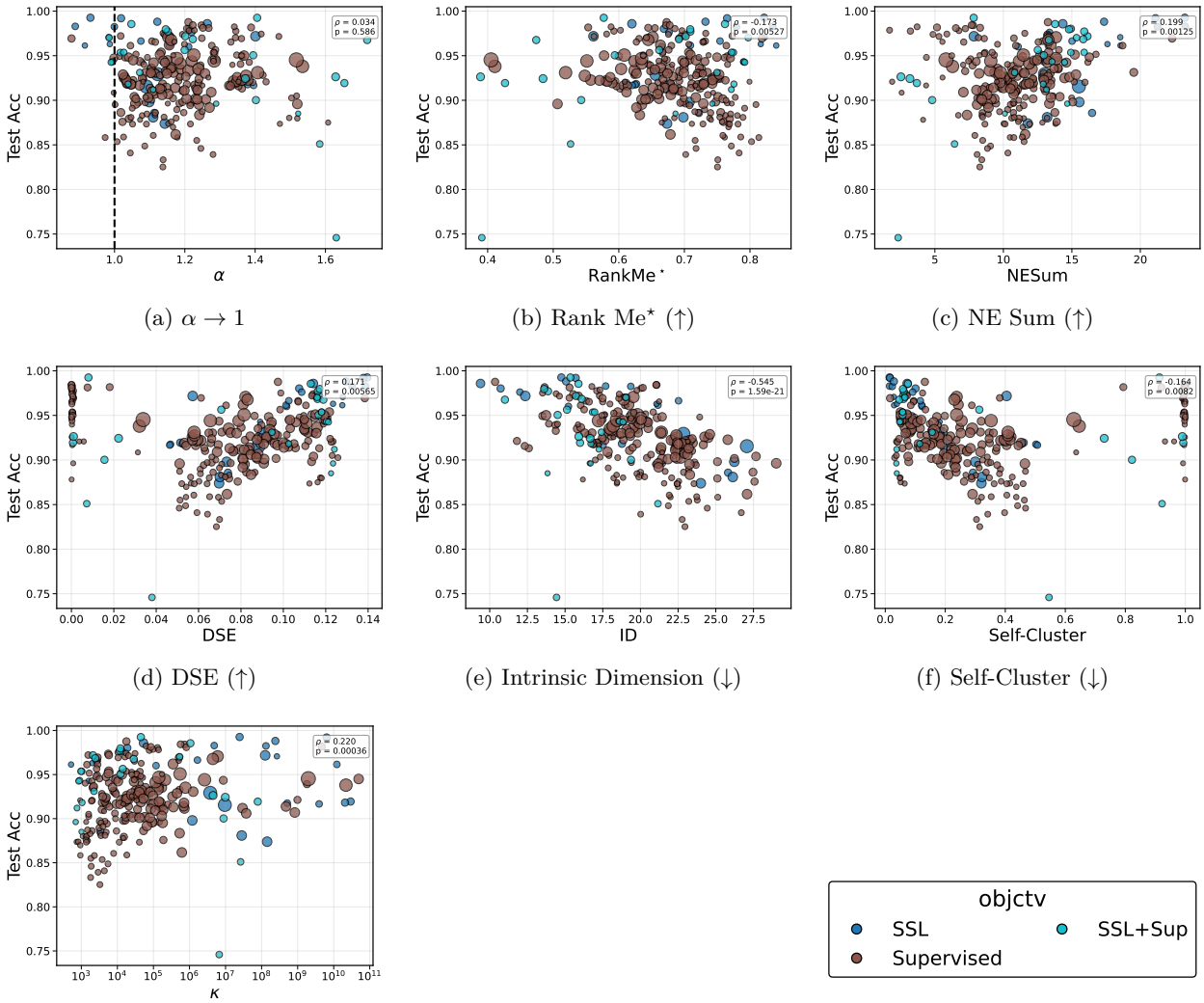
(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 14: Comparison of representation metrics and ImageNet test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 14a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.

D.2 By objective

We visualise the results, grouping models by their training objective

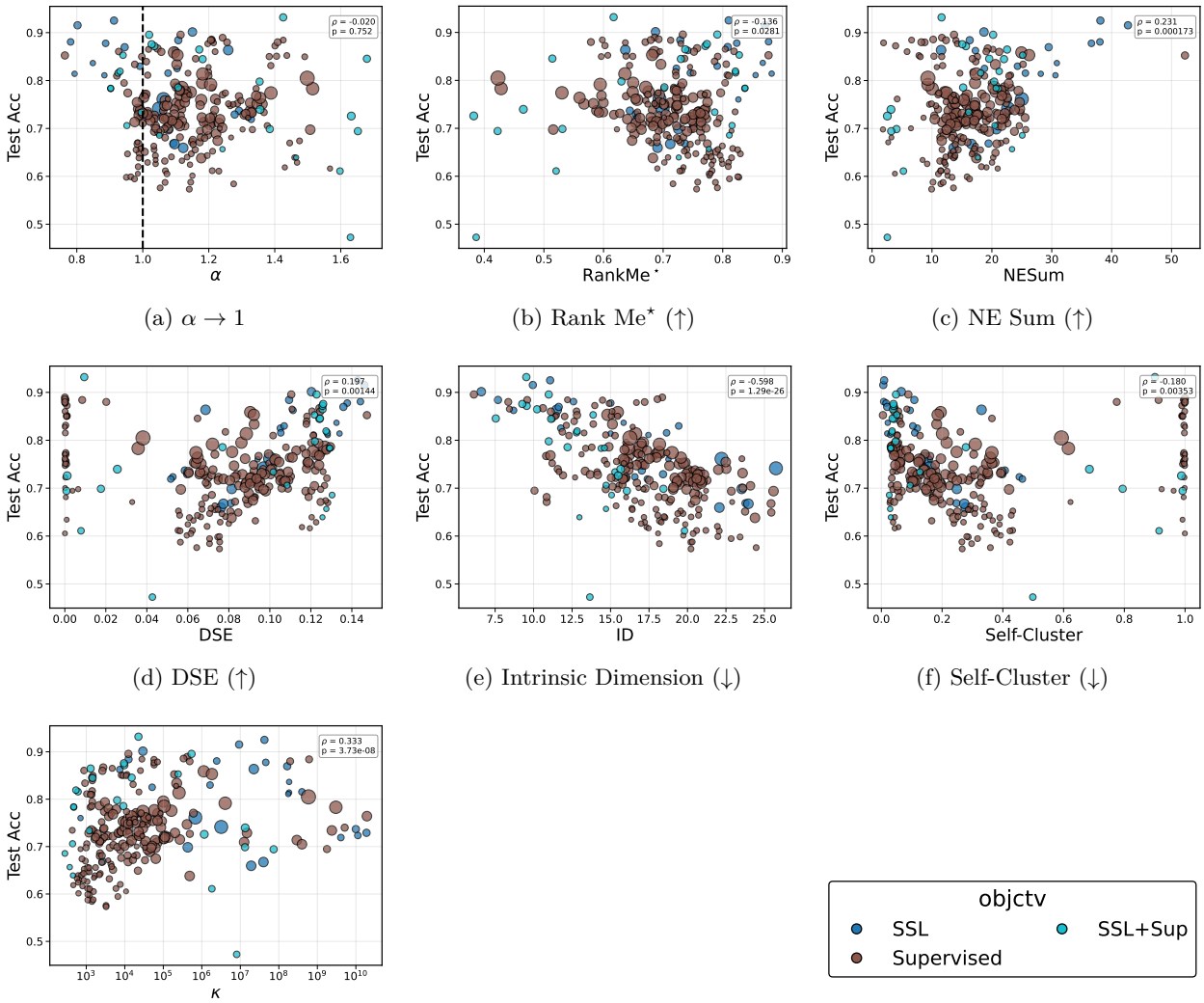


(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 15: Comparison of representation metrics and CIFAR-10 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 15a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.

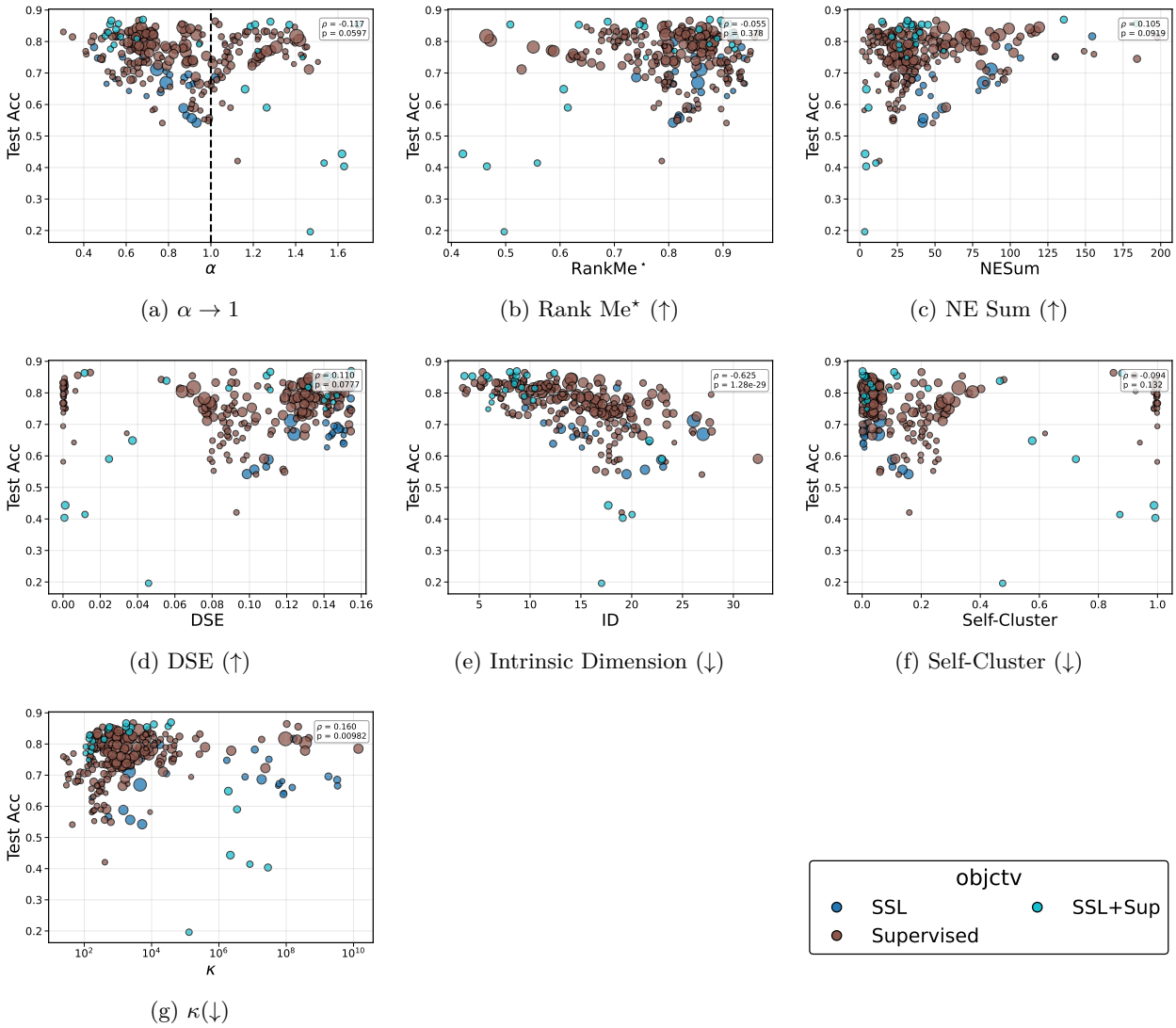


(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 16: Comparison of representation metrics and CIFAR-100 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 16a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.



(a) $\alpha \to 1$ (b) Rank Me* (↑) (c) NE Sum (↑)

(d) DSE (↑) (e) Intrinsic Dimension (↓) (f) Self-Cluster (↓)

(g) κ(↓)

Figure 17: Comparison of representation metrics and ImageNet test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 17a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.

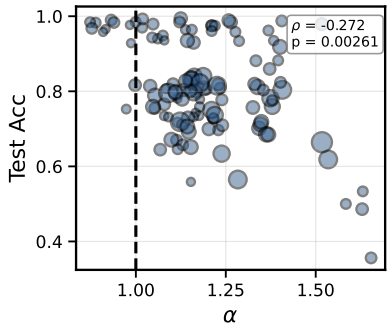
E SGD results

We freeze the backbone and train only the linear probe using SGD with a momentum of 0.9 and without weight decay, following standard practice for linear evaluation. We use Optuna Akiba et al. (2019) to find the optimal learning rate in the range $[1 \times 10^{-4}, 3 \times 10^{-1}]$, performing three rounds of hyper-parameter search to determine the final learning rate. We schedule the learning rate using a Cosine Scheduler and train for a total of 100 epochs. We perform ℓ_2 normalisation on the features as is commonly done in foundation models (Balestriero et al. (2023)). We train only ~ 120 models, a subset of models in Tab. B for this analysis.

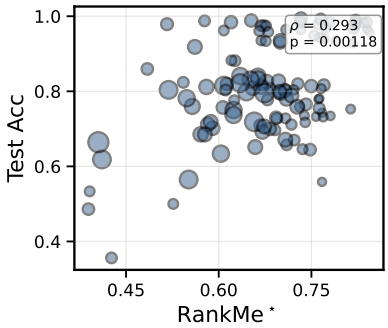
E.1 Summary of the results

Table 9: Spearman correlation (ρ) between representation metrics and test accuracy on Linear Probe trained using SGD, across CIFAR-10, CIFAR-100, and ImageNet-1k. We analyze three regimes: $\alpha < 1$, $\alpha > 1$, and unconstrained (all α). Bold underlined values indicate statistical significance ($p < 0.05$).()

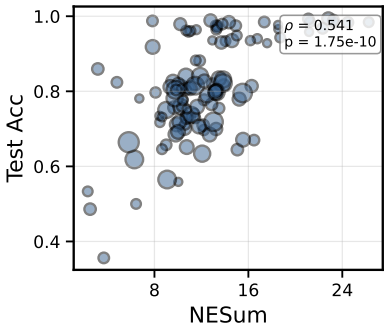
	CIFAR-10		CIFAR-100		ImageNet-1k	
Metric	ρ	p	ρ	p	ρ	p
$\alpha >= 1.0$	<u>0.1357</u>	<u>3.2360×10^{-2}</u>	0.0817	0.2298	0.0675	0.5651
$\alpha < 1.0$	-0.4794	0.0602	<u>-0.5287</u>	<u>1.5860×10^{-4}</u>	<u>-0.2734</u>	<u>1.6470×10^{-3}</u>
α-Req	0.0259	0.6752	-0.0146	0.8129	-0.0970	0.1666
RankMe	0.1138	0.0644	<u>0.2472</u>	<u>4.8930×10^{-5}</u>	<u>0.2009</u>	<u>3.8680×10^{-3}</u>
RankMe*	<u>-0.1565</u>	<u>1.0720×10^{-2}</u>	<u>-0.1378</u>	<u>2.5140×10^{-2}</u>	-0.0225	0.7493
NE-Sum	<u>0.2169</u>	<u>3.7580×10^{-4}</u>	<u>0.2162</u>	<u>4.0230×10^{-4}</u>	0.0588	0.4026
κ	<u>0.2663</u>	<u>1.1140×10^{-5}</u>	<u>0.3458</u>	<u>7.9080×10^{-9}</u>	0.0903	0.1980
Self-Cluster	<u>-0.1787</u>	<u>3.5210×10^{-3}</u>	<u>-0.1797</u>	<u>3.4000×10^{-3}</u>	-0.0585	0.4048
DSE	<u>0.1856</u>	<u>2.4150×10^{-3}</u>	<u>0.1960</u>	<u>1.3720×10^{-3}</u>	0.0740	0.2916
ID	<u>-0.5608</u>	<u>2.3610×10^{-23}</u>	<u>-0.6085</u>	<u>3.9740×10^{-28}</u>	<u>-0.6189</u>	<u>4.6090×10^{-23}</u>



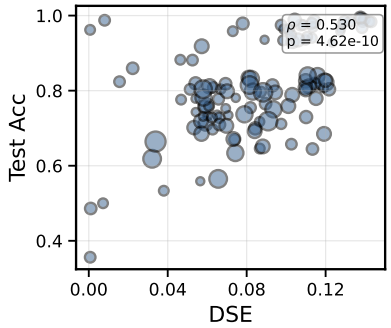
(a) $\alpha \to 1$



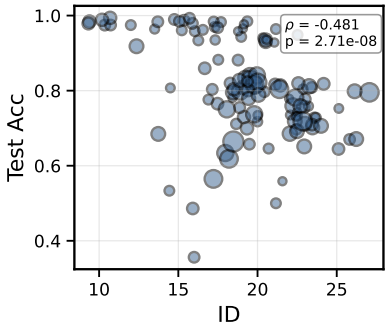
(b) Rank Me* (↑)



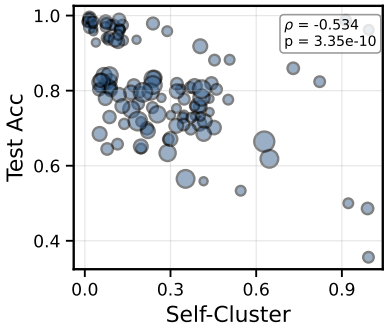
(c) NE Sum (↑)



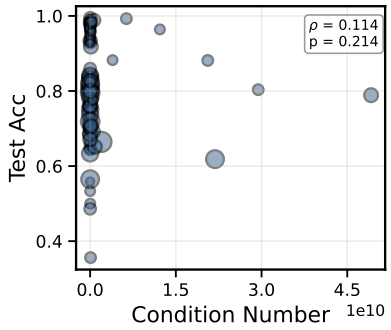
(d) DSE (↑)



(e) Intrinsic Dimension (↓)

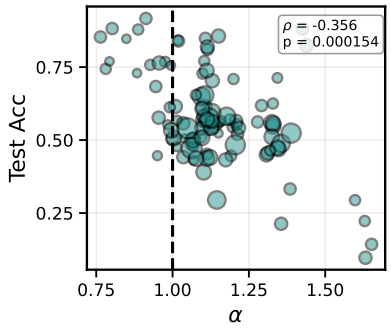


(f) Self-Cluster (↓)

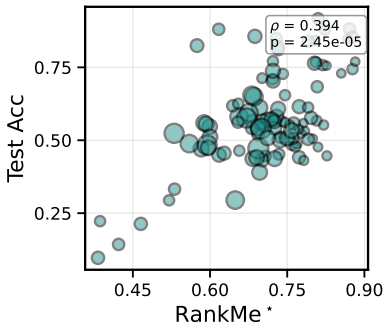


(g) $\kappa(\downarrow)$

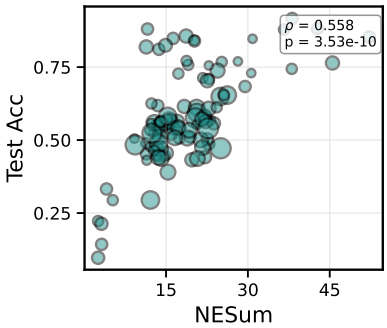
Figure 18: Comparison of representation metrics and CIFAR-10 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 18a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.



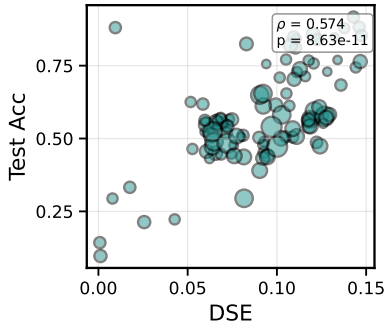
(a) $\alpha \to 1$



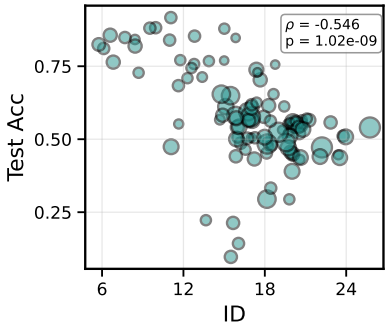
(b) Rank Me* (↑)



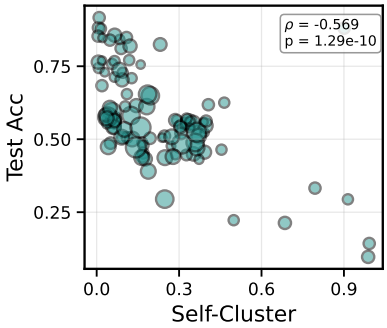
(c) NE Sum (↑)



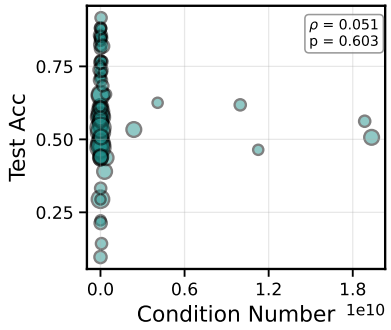
(d) DSE (↑)



(e) Intrinsic Dimension (↓)

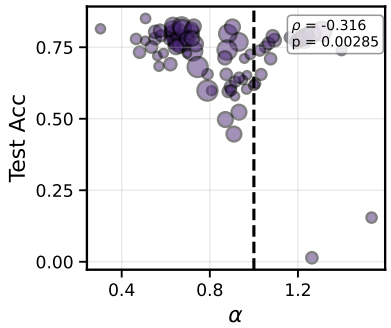


(f) Self-Cluster (↓)

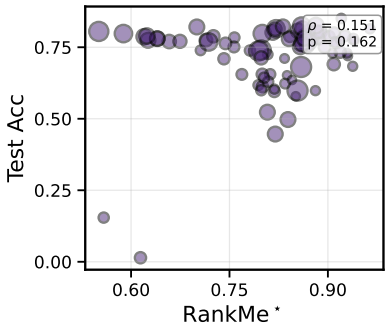


(g) κ(↓)

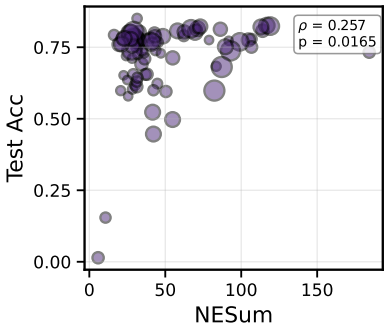
Figure 19: Comparison of representation metrics and CIFAR-100 test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 19a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.



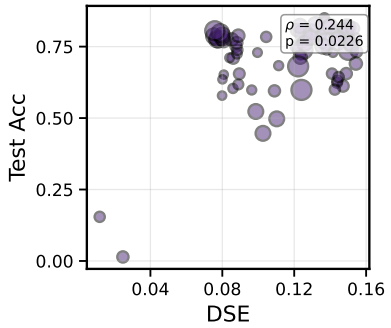
(a) $\alpha \to 1$



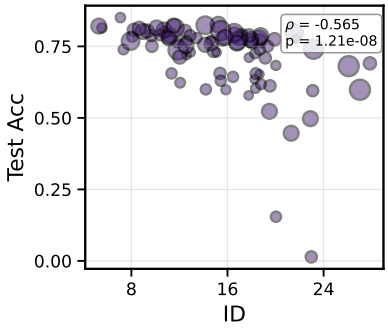
(b) Rank Me* (↑)



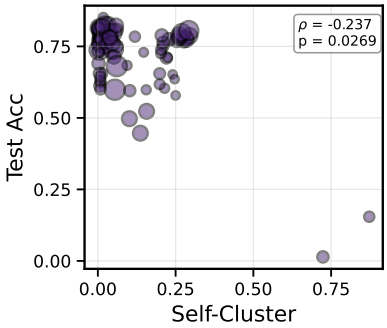
(c) NE Sum (↑)



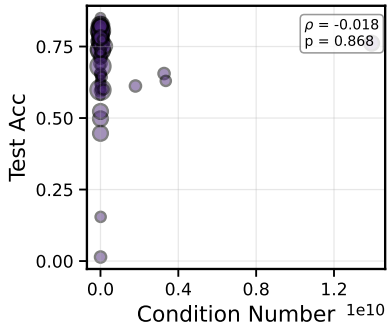
(d) DSE (↑)



(e) Intrinsic Dimension (↓)



(f) Self-Cluster (↓)



(g) κ(↓)

Figure 20: Comparison of representation metrics and ImageNet1K test accuracy. The ρ denotes the Spearman correlation and the corresponding p-value is annotated in each subplot. Arrows indicate the preferred direction: ↑ (higher is better), ↓ (lower is better), or $\to 1$ (closer to 1 is better). We mark the $\alpha = 1$ line in 20a accordance with the recommendation from Agrawal et al. (2022) that the ideal α is 1.

E.1.1 Architecture

SGD results, grouped models by their architecture class

Table 10: Spearman correlation (ρ) between representation metrics and test accuracy on CIFAR-10 across architectures. p denotes the p-value; correlations with $p < 0.05$ are underlined.

	ConvNeXT		RegNet		ResNet		ViT	
Metric	ρ	p	ρ	p	ρ	p	ρ	p
α-Req	-0.7324	1.2550×10^{-3}	-0.2886	2.1780×10^{-2}	-0.6573	2.0190×10^{-2}	-0.3853	0.0519
RankMe	0.4059	0.1188	0.2462	0.0518	0.4336	0.1591	0.7251	2.7820×10^{-5}
RankMe*	0.5412	3.0400×10^{-2}	0.0741	0.5640	0.2937	0.3541	0.3662	0.0658
NE-Sum	0.5765	1.9420×10^{-2}	0.4188	6.3470×10^{-4}	0.0769	0.8122	0.6198	7.3230×10^{-4}
κ	-0.5618	2.3540×10^{-2}	0.0938	0.4648	0.0140	0.9656	0.2752	0.1736
Self-Cluster	-0.8000	1.9860×10^{-4}	-0.4407	3.0090×10^{-4}	0.0070	0.9828	-0.6602	2.4270×10^{-4}
DSE	0.8147	1.2130×10^{-4}	0.4517	2.0230×10^{-4}	0.1399	0.6646	0.6663	2.0210×10^{-4}
ID	-0.3882	0.1373	0.0119	0.9264	0.0210	0.9484	-0.5843	1.7230×10^{-3}

Table 11: Spearman correlation (ρ) between representation metrics and test accuracy on CIFAR-100 across architectures. p denotes the p-value; correlations with $p < 0.05$ are underlined.

	ConvNeXT		RegNet		ResNet		ViT	
Metric	ρ	p	ρ	p	ρ	p	ρ	p
α-Req	-0.5560	3.8940×10^{-2}	-0.1563	0.2456	-0.4476	0.1446	-0.5325	1.0740×10^{-2}
RankMe	0.5648	3.5330×10^{-2}	0.2840	3.2300×10^{-2}	0.5175	0.0849	0.4715	2.6750×10^{-2}
RankMe*	0.5429	4.4860×10^{-2}	0.0130	0.9238	0.1608	0.6175	0.5844	4.2870×10^{-3}
NE-Sum	0.4637	0.0949	0.4727	2.0500×10^{-4}	0.5385	0.0709	0.6420	1.2770×10^{-3}
κ	-0.4725	0.0880	0.1297	0.3364	0.1189	0.7129	-0.0909	0.6874
Self-Cluster	-0.5297	0.0514	-0.3238	1.4020×10^{-2}	-0.4685	0.1245	-0.6002	3.1440×10^{-3}
DSE	0.5824	2.8850×10^{-2}	0.3317	1.1720×10^{-2}	0.4196	0.1745	0.6172	2.2150×10^{-3}
ID	-0.7582	1.6730×10^{-3}	-0.1495	0.2670	0.5385	0.0709	-0.6059	2.8030×10^{-3}

Table 12: Spearman correlation (ρ) between representation metrics and test accuracy on ImageNet across architectures. p denotes the p-value; correlations with $p < 0.05$ are underlined.(ResNet: 11 models, ConvNeXT: 10 models, ViT: 8 models, RegNet: 54 models)

	ConvNeXT		RegNet		ResNet		ViT	
Metric	ρ	p	ρ	p	ρ	p	ρ	p
α-Req	-0.3212	0.3655	-0.2468	0.0668	-0.8727	4.5460×10^{-4}	-0.8810	3.8500×10^{-3}
RankMe	0.1030	0.7770	0.5874	1.9430×10^{-6}	0.4091	0.2115	-0.0238	0.9554
RankMe*	0.2000	0.5796	0.0740	0.5880	0.6909	1.8570×10^{-2}	0.8571	6.5300×10^{-3}
NE-Sum	-0.4061	0.2443	0.4988	9.1350×10^{-5}	0.0546	0.8734	0.8333	1.0180×10^{-2}
κ	-0.3939	0.2600	0.3234	1.5050×10^{-2}	0.4364	0.1797	-0.0238	0.9554
Self-Cluster	0.5152	0.1276	-0.3661	5.5220×10^{-3}	-0.8182	2.0830×10^{-3}	-0.8095	1.4900×10^{-2}
DSE	-0.4182	0.2291	0.3638	5.8520×10^{-3}	0.8000	3.1100×10^{-3}	0.7857	2.0820×10^{-2}
ID	-0.4303	0.2145	-0.5306	2.5990×10^{-5}	-0.4727	0.1420	-0.5000	0.2070

E.1.2 Dimension Group

SGD results grouped models by their training objective

Table 13: Spearman correlation (ρ) between representation metrics and CIFAR-10 test accuracy across architectures. Underlined values indicate statistical significance ($p < 0.05$). Correlations are computed per objective, including only objectives with at least 5 models (Classification: 81 models, DINO: 15 models, VicReg: 10 models, MAE: 5 models).

	SSL		SSL+Sup		Supervised	
Metric	ρ	p	ρ	p	ρ	p
α-Req	-0.3091	0.0906	$\underline{-0.7333}$	$\underline{1.5800 \times 10^{-2}}$	-0.2014	0.0751
RankMe	-0.0797	0.6702	$\underline{0.7091}$	$\underline{2.1670 \times 10^{-2}}$	-0.0073	0.9490
RankMe*	$\underline{0.4059}$	$\underline{2.3480 \times 10^{-2}}$	$\underline{0.7333}$	$\underline{1.5800 \times 10^{-2}}$	0.1479	0.1932
NE-Sum	$\underline{0.5708}$	$\underline{7.9820 \times 10^{-4}}$	$\underline{0.6970}$	$\underline{2.5100 \times 10^{-2}}$	$\underline{0.4961}$	$\underline{3.3350 \times 10^{-6}}$
κ	0.0919	0.6228	$\underline{-0.7697}$	$\underline{9.2220 \times 10^{-3}}$	-0.1140	0.3170
Self-Cluster	$\underline{-0.8200}$	$\underline{1.6510 \times 10^{-8}}$	-0.4909	0.1497	$\underline{-0.4715}$	$\underline{1.1520 \times 10^{-5}}$
DSE	$\underline{0.8253}$	$\underline{1.1150 \times 10^{-8}}$	0.4909	0.1497	$\underline{0.4690}$	$\underline{1.3040 \times 10^{-5}}$
ID	$\underline{-0.6700}$	$\underline{3.7370 \times 10^{-5}}$	-0.3939	0.2600	$\underline{-0.2967}$	$\underline{7.9230 \times 10^{-3}}$

Table 14: Spearman correlation (ρ) between representation metrics and CIFAR-100 test accuracy across architectures. Underlined values indicate statistical significance ($p < 0.05$). Correlations are computed per objective, including only objectives with at least 5 models (Classification: 81 models, DINO: 15 models, VicReg: 10 models, MAE: 5 models).

	SSL		SSL+Sup		Supervised	
Metric	ρ	p	ρ	p	ρ	p
α-Req	$\underline{-0.4084}$	$\underline{3.4430 \times 10^{-2}}$	-0.6333	0.0671	-0.1330	0.2654
RankMe	-0.2424	0.2232	0.6333	0.0671	0.0517	0.6660
RankMe*	$\underline{0.5501}$	$\underline{2.9540 \times 10^{-3}}$	$\underline{0.9333}$	$\underline{2.3590 \times 10^{-4}}$	0.1191	0.3189
NE-Sum	$\underline{0.4438}$	$\underline{2.0390 \times 10^{-2}}$	$\underline{0.8167}$	$\underline{7.2250 \times 10^{-3}}$	$\underline{0.5344}$	$\underline{1.3310 \times 10^{-6}}$
κ	-0.0720	0.7210	-0.6500	0.0581	-0.0533	0.6566
Self-Cluster	$\underline{-0.7222}$	$\underline{2.1060 \times 10^{-5}}$	-0.5667	0.1116	$\underline{-0.4394}$	$\underline{1.1260 \times 10^{-4}}$
DSE	$\underline{0.7411}$	$\underline{9.7710 \times 10^{-6}}$	0.5667	0.1116	$\underline{0.4403}$	$\underline{1.0850 \times 10^{-4}}$
ID	$\underline{-0.7766}$	$\underline{1.9170 \times 10^{-6}}$	-0.5000	0.1705	$\underline{-0.4250}$	$\underline{1.9850 \times 10^{-4}}$

Table 15: Spearman correlation (ρ) between representation metrics and Imagenet test accuracy across architectures. Underlined values indicate statistical significance ($p < 0.05$). Correlations are computed per objective, including only objectives with at least 5 models (Classification: 66 models, DINO: 7 models, VicReg: 9 models, MIM: 2 models).

	SSL		Supervised	
Metric	ρ	p	ρ	p
α-Req	$\underline{-0.7377}$	$\underline{7.2380 \times 10^{-4}}$	-0.2007	0.1007
RankMe	-0.4559	0.0659	$\underline{0.5346}$	$\underline{2.6480 \times 10^{-6}}$
RankMe*	$\underline{0.5172}$	$\underline{3.3510 \times 10^{-2}}$	0.0176	0.8867
NE-Sum	0.1005	0.7012	$\underline{0.3245}$	$\underline{6.9310 \times 10^{-3}}$
κ	0.2230	0.3895	$\underline{0.3502}$	$\underline{3.4200 \times 10^{-3}}$
Self-Cluster	$\underline{-0.7941}$	$\underline{1.4090 \times 10^{-4}}$	$\underline{-0.2648}$	$\underline{2.9120 \times 10^{-2}}$
DSE	$\underline{0.7794}$	$\underline{2.2570 \times 10^{-4}}$	$\underline{0.2698}$	$\underline{2.6080 \times 10^{-2}}$
ID	$\underline{-0.5490}$	$\underline{2.2460 \times 10^{-2}}$	$\underline{-0.5002}$	$\underline{1.4030 \times 10^{-5}}$