POPULATION-LEVEL REPRESENTATION VARIABILITY AND TRIPLET-DERIVED MODEL-LEVEL REPRESENTATION DESCRIPTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Independently trained neural networks can develop different representations under the same learning conditions, motivating a population-level view of learned representations. Yet, most analyses focus on pairwise comparisons, limiting characterization of population variability and individual representations. We address both levels with Representational Dispersion (RD), a measure of population-level variability, and the Triplet-Derived Representation Descriptor (TRD), which characterizes each model's representation through relative distances over shared observations. We show that RD captures consistent variability patterns across learning conditions and training stochasticity, while TRD preserves Centered Kernel Alignment (CKA)-based relationships between models despite differences in their representation spaces. We further introduce Compact TRD and Surrogate-Guided Subset Optimization (SGSO), which efficiently optimizes an observation subset under a fixed budget to construct a compact TRD that preserves model-level representation geometry. Together, these tools extend representation analysis from pairwise similarity to population structure and provide a basis for large-scale model characterization, organization, and selection.

1 INTRODUCTION

Learned representations are shaped by a neural network's architecture, learning objective, training data, and training configuration (Hermann & Lampinen, 2020; Kornblith et al., 2021). Yet, even under matched learning conditions, independently trained models can develop different representations due to stochastic variation across training runs (Mehrer et al., 2020; Zhang & Bottou, 2023). If such variation exhibits reproducible statistical regularities, independently learned representations may be viewed not merely as isolated training outcomes, but as realizations from a population of representations under shared learning conditions. This raises the question of whether representational variability itself constitutes a measurable and systematic property of neural learning.

Most methods for comparing learned representations measure similarity between pairs of representations. Centered Kernel Alignment (CKA) (Kornblith et al., 2019), for example, compares two representations but does not directly quantify variability across a population of independently trained models. We extend this pairwise perspective by defining a population-level measure of representational dispersion from CKA similarities and examining its consistency across learning conditions and sources of training stochasticity. Our results show that representational variability exhibits reproducible patterns associated with learning conditions.

We next introduce the Triplet-Derived Representation Descriptor (TRD), which characterizes each model's representation by vectorizing the triplet relations it induces over a shared observation set. By capturing relational geometry rather than internal activations, TRD provides a common observational space for comparing model representations. To address the combinatorial growth of triplets, we further optimize a compact observation subset whose induced TRD preserves model-level representation geometry, enabling efficient model-level characterization. Taken together, we present a framework for analyzing independently learned neural representations as a structured population under shared learning conditions and quantitatively characterizing individual model representations within that population in a shared observational coordinate system. Our main contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1. **Population-level variability.** We introduce Representational Dispersion (RD) to quantify population-level variability and show that it varies systematically across learning conditions while exhibiting reproducible patterns across training stochasticity.
2. **Model-level descriptor.** We introduce TRD to characterize individual representations in a shared observational space and show that it preserves CKA-based relationships.
3. **Scalable characterization.** We introduce Compact TRD and a Surrogate-Guided Subset Optimization (SGSO) method that identifies compact observation sets preserving the model-level geometry of the full descriptor and generalizing to unseen model populations.

2 Related Work

2.1 Variability in Learned Representations

Machine learning models can admit multiple solutions with comparable predictive performance, commonly described by the Rashomon effect and formalized by the Rashomon set (Hsu et al., 2026). Relatedly, underspecification suggests that models with similar task performance may develop different internal representations, further influenced by training stochasticity such as parameter initialization, data sampling, and optimization. This perspective has been extended to representation learning, where geometrically distinct embeddings can preserve similar underlying data structure. For example, Sun et al. (2026) formalized a Rashomon set of embeddings from dimensionality reduction and examined structures shared across alternative embeddings. These findings motivate viewing a learned representation as one of multiple plausible realizations. Various methods quantify representational similarity across neural networks (Kriegeskorte et al., 2008; Raghu et al., 2017; Morcos et al., 2018). CKA, in particular, is widely used to compare representations across independently trained models, architectures, layers, and datasets. However, these methods primarily focus on pairwise relationships, leaving the population-level structure of independently learned representations relatively underexplored. We extend this perspective by quantifying representational variability and its statistical structure at the population level.

2.2 Model-Level Characterization of Neural Representations

Common representation comparison methods characterize relationships between pairs of representations rather than providing model-level descriptors in a shared observational space (Klabunde et al., 2025; Harvey et al., 2024). Pairwise similarity matrices can be embedded using dimensionality-reduction methods such as multidimensional scaling (Borg & Groenen, 2005), but the resulting coordinates depend on the compared model population rather than directly describing each model from shared observations. In contrast, models with different internal spaces can be characterized by the relational geometry they induce over common observations. Metric learning and ordinal embedding use relational information to represent data points (Schroff et al., 2015; Anand et al., 2026). We instead apply common triplets across models and vectorize their relations to characterize each model in a shared observational space. Because the number of triplets grows as $O(N^3)$, prior work uses triplet sampling, landmarks, and coresets to reduce cost (Wilber et al., 2014; Anderton & Aslam, 2019; Ghosh et al., 2019; Mirzasoleiman et al., 2020), typically while preserving data geometry. We instead seek a compact observation subset that preserves model-level relational information.

3 Methods

3.1 Problem Formulation: Representations as Population Realizations

A learned representation depends on the learning domain, model architecture, objective, training data, and hyperparameters, while stochasticity in initialization, data sampling, and optimization can produce different representations under otherwise fixed conditions. Let C denote a fixed learning condition and ξ the stochastic factors in training, such that $R = f(C; \xi)$. Independently trained representations can then be viewed as realizations $R_m \sim P(R \mid C)$, where $P(R \mid C)$ denotes the **population of learned representations** under condition C, without assuming a particular parametric form. This formulation motivates characterization at two levels: population-level variability

Table 1: Semantic Category Pairs Used to Construct the D-Datasets. Pair types are abbreviated as A = Animal, O = Object, F = Food, and I = Instrument.

ID	Pair	Type	ID	Pair	Type	ID	Pair	Type
D1	Wild boar – Hognose snake	AA	D9	Rock crab – Moped	AI	D17	Rotisserie – Slide rule	II
D2	Vacuum cleaner – Foxhound	AI	D10	Garter snake – Slide rule	AI	D18	Gyromitra – Bannister	FO
D3	Toy terrier – Chocolate sauce	AF	D11	Window screen – Car wheel	OI	D19	Little blue heron – Lampshade	AO
D4	Standard poodle – Shih-Tzu	AA	D12	Pineapple – Walker hound	AF	D20	Throne – Hare	AO
D5	Doberman – Tripod	AI	D13	Garden spider – Fiddler crab	AA	D21	Honeycomb – Hare	AO
D6	Great Dane – Standard poodle	AA	D14	Cocktail shaker – Walker hound	AI	D22	Rocking chair – Wing	OI
D7	Carbonara – Pineapple	FF	D15	Garden spider – Red fox	AA			
D8	Saluki – Tabby cat	AA	D16	Gyromitra – American coot	AF			

across independently learned representations and model-level characterization of individual representations within the population. We therefore introduce **Representational Dispersion (RD)** to quantify population-level variability and the **Triplet-Derived Representation Descriptor (TRD)** to characterize the relational geometry of individual model representations.

3.2 Population-Level Representational Dispersion

Under a fixed learning condition C, let $R_1, \ldots, R_M \overset{\text{i.i.d.}}{\sim} P(R \mid C)$ denote representations from M independently trained models, with pairwise similarity $S_{ij}^{\text{CKA}} = \text{CKA}(R_i, R_j)$. We define **CKA-based Representational Dispersion (RD)** as $\text{RD}_C^{\text{CKA}} = \mathbb{E}_{R,R' \sim P(R|C)}[1 - \text{CKA}(R, R')]$, the expected pairwise CKA dissimilarity within the population. Given M models, we estimate it as

$$\widehat{\text{RD}}_C^{\text{CKA}} = \frac{2}{M(M-1)} \sum_{i<j} \left(1 - S_{ij}^{\text{CKA}}\right) = 1 - \overline{\text{CKA}}_C, \tag{1}$$

where $\overline{\text{CKA}}_C$ is the mean pairwise CKA among the sampled models. More formally, $\widehat{\text{RD}}_C^{\text{CKA}}$ is a second-order U-statistic with kernel $h^{\text{CKA}}(R_i, R_j) = 1 - \text{CKA}(R_i, R_j)$ (Hoeffding, 1992), providing a population-level interpretation of RD. RD measures **population-level pairwise dispersion**, rather than covariance-based variance. Low RD indicates that independent training runs tend to produce similar representational structures, whereas high RD indicates greater representational heterogeneity. When only the learning domain varies while other components of C are fixed, we denote the corresponding quantity by RD_d^{CKA}. In Section 4.2, we examine whether representational dispersion differs systematically across domains and how its domain-wise pattern changes across sources of randomness, architectures, and learning endpoints (See Appendix A.1).

3.3 Triplet-Derived Model-Level Representation Descriptor

RD describes the variability of a representation population, but does not describe individual models within that population. We therefore introduce the **Triplet-Derived Representation Descriptor (TRD)**, which describes each model using the relative distance relations it induces among observations. Given an observation set $X = \{x_1, \ldots, x_N\}$, let $\mathcal{T} = \{t_1, \ldots, t_K\}$ denote all possible triplets constructed from X, shared across all models, where $K = N\binom{N-1}{2}$. For each triplet $t = (a, b, c) \in \mathcal{T}$, we define $r_m(t) = d_m(a, c) - d_m(a, b)$, where d_m denotes distance in the representation space of model m. The sign of $r_m(t)$ indicates which of b and c is closer to the anchor a, while its magnitude gives the relational margin. We define the TRD of model m as

$$\text{TRD}_m = [r_m(t_1), \ldots, r_m(t_K)]^\top. \tag{2}$$

Because the same triplets are used for every model, each TRD dimension corresponds to the same relational observation. TRD therefore provides a **shared observational space** in which models with different internal representation spaces can be described and compared.

Pairwise correspondence. To test whether TRD provides a meaningful model-level representation descriptor, we compare the geometry induced by TRD with that of the original representations. For M models, let $S_{ij}^{\text{CKA}} = \text{CKA}(R_i, R_j)$ and $S_{ij}^{\text{TRD}} = \text{sim}(\text{TRD}_i, \text{TRD}_j)$. We measure their correspondence by $\rho_{\text{TRD}} = \text{corr}(\{S_{ij}^{\text{CKA}}\}_{i<j}, \{S_{ij}^{\text{TRD}}\}_{i<j})$. A high ρ_{TRD} indicates that TRD

Algorithm 1 Surrogate-Guided Subset Optimization for Compact TRD

Require: Observations X, representations $\{R_m\}_{m=1}^{M}$, subset size n, CKA similarities $\mathbf{s}^{\mathrm{CKA}}$
Ensure: Optimized subset S^* and compact descriptors $\{\mathrm{cTRD}_m\}_{m=1}^{M}$
1: Sample initial subsets $\mathcal{D} = \{S_l : S_l \subset X,\ |S_l| = n\}$
2: Evaluate $Q(S)$ for each $S \in \mathcal{D}$
3: Represent each S by its observation indices and train $g_\theta(S)$ using $\mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{MSE}}$
4: **repeat**
5: Generate candidates by random sampling and perturbing high-scoring subsets
6: Rank candidates with g_θ and retain the top candidates
7: Generate local neighbors by replacing a small number of observations
8: Evaluate true $Q(S)$ for selected neighbors and random subsets
9: Update $\mathcal{D}$ with elite and newly evaluated subsets
10: Retrain g_θ on $\{(S, Q(S)) : S \in \mathcal{D}\}$
11: **until** the stopping criterion is reached
12: $S^* \leftarrow \arg\max_{S \in \mathcal{D}} Q(S)$
13: $\mathrm{cTRD}_m \leftarrow \mathrm{TRD}_m(S^*)$ for each model m
14: **return** S^*, $\{\mathrm{cTRD}_m\}_{m=1}^{M}$

preserves the pairwise geometry of model representations measured by CKA, providing empirical support for its use as a model-level representation descriptor.

Population characterization. We also define TRD-based representational dispersion in parallel with the CKA-based RD in Section 3.2. For two independent TRDs $\mathbf{r}$ and $\mathbf{r}'$ from the population under condition C, we define $\mathrm{RD}_C^{\mathrm{TRD}} = \mathbb{E}[1 - \cos(\mathbf{r}, \mathbf{r}')]$. Given M models, its sample estimate is

$$\widehat{\mathrm{RD}}_C^{\mathrm{TRD}} = \frac{2}{M(M-1)} \sum_{i<j} \left[1 - \cos\left(\mathrm{TRD}_i, \mathrm{TRD}_j\right)\right]. \tag{3}$$

To evaluate finite-sample estimation of TRD population statistics, we examine four statistics: $\phi \in \mathrm{RD}^{\mathrm{TRD}}, \mathrm{Tr}, H, \mu$, denoting TRD-based dispersion, total variance, normalized spectral entropy, and the population mean, respectively. For a centered TRD matrix with singular values $\sigma_1, \ldots, \sigma_L$, we define $\lambda_k = \sigma_k^2/(M-1)$ for $k = 1, \ldots, L$, $\mathrm{Tr} = \sum_{k=1}^{L} \lambda_k$, $p_k = \lambda_k / \sum_{j=1}^{L} \lambda_j$, and $H = -\sum_{k=1}^{L} p_k \log p_k / \log L$, where L is the number of singular values (Roy & Vetterli, 2007); μ is the mean TRD across models. For each statistic, we compare its estimate $\widehat{\phi}_M$ from M sampled models with the full-population value ϕ_{pop} using the statistic-specific estimation error defined in Appendix A.2, and examine how the error changes with M. Together, pairwise correspondence with CKA and population-level estimation assess whether TRD captures relational structure at both the individual-model and population levels. TRD is not intended to reconstruct the full internal representation, but to provide an observable model-level description of its relational geometry.

3.4 Compact TRD via Surrogate-Guided Subset Optimization

The number of triplets grows as $O(N^3)$ with the number of observations, making the full TRD expensive for large observation sets. We therefore seek a compact subset $S \subset X$, with $|S| = n \ll N$, whose triplets preserve the model-level geometry captured by the full observation set. We refer to the resulting descriptor as the **Compact TRD (cTRD)**. Following Section 3.3, let $\mathbf{s}^{\mathrm{CKA}} = [\mathrm{CKA}(R_i, R_j)]_{i<j}$ denote the pairwise model similarities computed from the full observation set. For a candidate subset S, we compute the candidate compact descriptor $\mathbf{r}_m^S = \mathrm{TRD}_m(S)$ for each model and the corresponding pairwise similarity vector $\mathbf{s}^S = [\cos(\mathbf{r}_i^S, \mathbf{r}_j^S)]_{i<j}$. We measure correspondence with the reference CKA geometry by $Q(S) = \tau(\mathbf{s}^{\mathrm{CKA}}, \mathbf{s}^S)$, where τ denotes Kendall's rank correlation (Kendall et al., 1938). The optimal subset of size n is then $S^* = \arg\max_{S \subset X,\, |S|=n} Q(S)$. The goal is therefore not to select observations that are representative of the data distribution, but to find a small observation set that preserves the **relative geometry among models**. Since exhaustive search over all subsets is intractable, we propose **Surrogate-Guided Subset Optimization (SGSO)**, an iterative search procedure that combines surrogate-

guided candidate selection with global exploration and local search. For surrogate prediction, each subset S is represented only by its observation indices, without image content or representation features as inputs. A permutation-invariant DeepSets (Zaheer et al., 2017) surrogate $g_\theta(S)$ is trained on evaluated subset–score pairs $(S, Q(S))$ using a pairwise ranking loss with an auxiliary regression loss, $\mathcal{L}_{\text{sur}} = \mathcal{L}_{\text{rank}} + \lambda_{\text{reg}}\mathcal{L}_{\text{MSE}}$. At each iteration, we generate candidate subsets through random sampling and perturbations of high-scoring subsets. The surrogate ranks these candidates, and promising candidates and their neighbors are evaluated using the true score $Q(S)$. The evaluated subsets are then added to the training pool and the surrogate is updated. This process combines global exploration with local search around promising subsets. The full optimization procedure is described in Algorithm 1. The highest-scoring subset found during the search is selected as S^*. The resulting compact descriptor for model m is defined as

$$\text{cTRD}_m = \text{TRD}_m(S^*). \tag{4}$$

4 Experiments

4.1 Experimental Setup

We use two types of datasets: multiclass datasets (M-datasets) and domain datasets (D-datasets). The M-datasets comprise Caltech-UCSD Birds-200-2011 (CUB; 200 classes) (Wah et al., 2011), Stanford Cars196 (CAR; 196 classes) (Krause et al., 2013), and ImageNet-100 (IMG; 100 classes) (Deng et al., 2009). For comparability, we constructed a class-balanced training pool of 5,864 samples for each M-dataset by random subsampling, matching the size of the smallest dataset. Training samples for each task were then drawn from the fixed pool. The D-datasets comprise 22 binary classification domains constructed from random class pairs in ImageNet-100 and are denoted D1–D22 (Table 1). The indices were assigned retrospectively according to their $\widehat{\text{RD}}^{\text{CKA}}$ in Fig. 1a and do not indicate an inherent ranking. D-datasets are used for population-level analyses across domains, whereas M-datasets are used for large-population convergence and ensemble experiments. We use ResNet-50 and ViT-Small for M-datasets and ResNet-18 and DeiT-Tiny for D-datasets (He et al., 2016; Touvron et al., 2021; Dosovitskiy et al., 2020). Representations are extracted from the backbone output before the classification head, with dimensions of 2,048, 384, 512, and 192, respectively. Specific settings are provided in the corresponding sections and Appendix A.3.

4.2 Population-Level Representation Variability across Learning Conditions

We examined whether independently trained models exhibit reproducible population-level representation variability under a fixed learning condition C, defined by the semantic domain, model architecture, and training hyperparameters including the learning objective. We considered three sources of randomness: data sampling, weight initialization, and training stochasticity from dropout and data shuffling. For each domain, we sampled 1,000 training and 100 representation-measurement examples with balanced classes, and independently trained 30 classifiers per condition. Dropout ($p = 0.2$) was applied only to the classifier head, while the validation and representation-measurement sets were fixed within each domain. Population-level variability was quantified by $\widehat{\text{RD}}^{\text{CKA}}$ and estimated over 20 independent trials. As shown in Fig. 1, $\widehat{\text{RD}}^{\text{CKA}}$ varied across domains, architectures, and early-stopping thresholds, but remained stable across repeated trials within each condition. Its domain-wise pattern was also highly consistent across sources of randomness: at threshold 0.6, correlations of data randomness with weight and training randomness were $r = .983$ and $r = .894$ for ResNet-18, and $r = .957$ and $r = .966$ for DeiT-Tiny, respectively (Table 2). R@1 is defined as the proportion of representation-measurement samples whose nearest neighbor in the representation space belongs to the same class. Across domains, lower $\widehat{\text{RD}}^{\text{CKA}}$ generally coincided with higher R@1, suggesting that lower population-level representation variability tends to accompany greater local class consistency in the representation space. In contrast, the domain-wise correlation between ResNet-18 and DeiT-Tiny was weak ($r = .258$, $p = .246$), whereas the pattern remained largely consistent across early-stopping thresholds ($r = .871$ between 0.6 and 0.7 for ResNet-18). These results suggest that a learning condition C constrains a distribution of representation realizations with reproducible population-level variability. This variability is robust across sources of random-

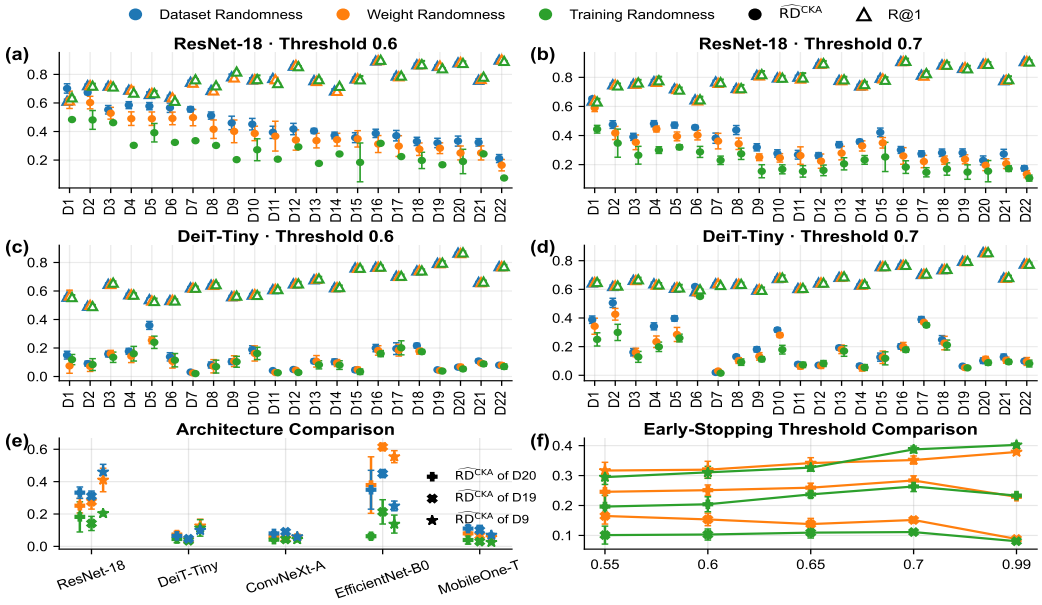


Figure 1: Population-level representation variability across learning conditions. (a–d) Domain-wise $\widehat{\mathrm{RD}}^{\mathrm{CKA}}$ and R@1 across randomness sources; (e) architecture (Liu et al., 2022; Tan & Le, 2019; Vasu et al., 2023) and (f) early-stopping threshold comparisons. Each randomness condition varies only the indicated factor; data-sampling randomness is excluded in (f).

Table 2: Domain-wise consistency of $\widehat{\mathrm{RD}}^{\mathrm{CKA}}$ across selected experimental conditions (Pearson r, 22 D-datasets; reference: data randomness at $\tau = .6$). $^{\dagger}p = .246$; all other $p < .001$.

Reference	Comparison	Factor	r	Reference	Comparison	Factor	r
ResNet-18, $\tau = .6$	Weight randomness	Randomness	.983	DeiT-Tiny, $\tau = .6$	Weight randomness	Randomness	.957
ResNet-18, $\tau = .6$	Training randomness	Randomness	.894	DeiT-Tiny, $\tau = .6$	Training randomness	Randomness	.966
ResNet-18, $\tau = .6$	DeiT-Tiny, $\tau = .6$	Architecture	$.258^{\dagger}$	ResNet-18, $\tau = .6$	ResNet-18, $\tau = .7$	Threshold	.871

ness while being systematically shaped by the learning condition, including the semantic domain and architecture (See Appendix A.2).

4.3 TRD as a Statistical Descriptor of Representation Variability

In the previous section, we showed using CKA that independently trained models exhibit population-level representation variability with reproducible statistical structure. However, CKA requires direct access to internal representations for the same inputs, limiting its applicability to models or observers whose internal states are inaccessible, such as humans. We therefore examined whether TRD can characterize both pairwise representation relationships and population-level variability. We first computed $\widehat{\mathrm{RD}}^{\mathrm{TRD}}$ from the TRDs of each model population in Fig. 2(a–b) and compared it with $\widehat{\mathrm{RD}}^{\mathrm{CKA}}$. The two measures showed strong domain-level correspondence, which was largely preserved when only 50% or 10% of the TRD dimensions were randomly retained (HD and 10D). This suggests substantial redundancy in TRD, allowing population-level variability to be captured with a subset of triplets. Binarizing the distance-based TRD moderately reduced this correspondence, indicating that continuous distance information provides additional detail about representation geometry. At the model-pair level, TRD similarity was also strongly correlated with CKA similarity (Fig. 2(c–d)). These results support TRD as a representation descriptor that captures pairwise similarity at the model-pair level and representation dispersion (RD) at the population level.

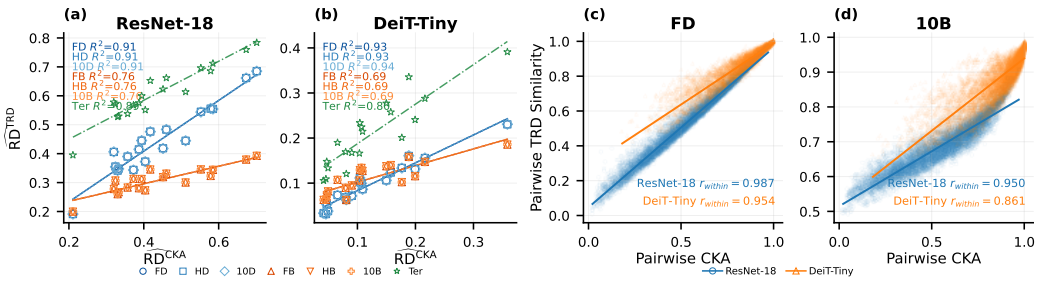


Figure 2: Correspondence between CKA- and TRD-based representation geometry. (a–b) Domain-level $\widehat{\mathrm{RD}}^{\mathrm{TRD}}$ versus $\widehat{\mathrm{RD}}^{\mathrm{CKA}}$ for ResNet-18 and DeiT-Tiny, with linear fits and R^2. (c–d) Model-pair-level CKA–TRD similarity correspondence for FD and 10B; r_{within} denotes the Pearson correlation after within-domain mean-centering. TRD variants are FD/HD/10D (distance-based using 100/50/10% of triplets), FB/HB/10B (binary using 100/50/10%), and Ter (ternary).

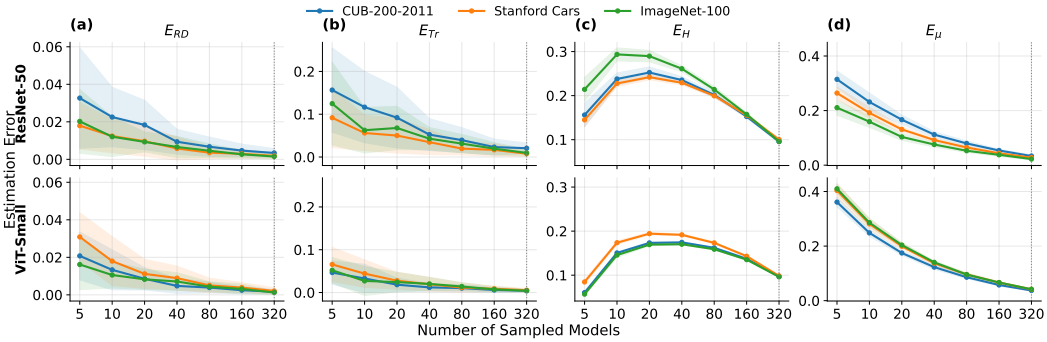


Figure 3: Convergence of population-statistic estimation with population size. (a–d) Estimation errors for TRD-based dispersion ($\mathrm{RD}^{\mathrm{TRD}}$), total variance (Tr), normalized spectral entropy (H), and population mean (μ) across M-datasets. The dotted line indicates the full population ($N = 320$).

Finally, we examined whether TRD-derived population statistics can be reliably estimated from a finite sample of models. For CUB, CAR, and IMG, each model was trained on an independently sampled set of 15 images per class, and its TRD was computed from 30 observations, yielding 12,180 triplets (See Appendix A.2). For each dataset and backbone, statistics computed from all available models served as the reference, against which we evaluated estimates of $\mathrm{RD}^{\mathrm{TRD}}$, total variance Tr, normalized spectral entropy H, and population mean μ from randomly sampled model subsets. Estimation errors decreased with the number of sampled models across datasets and backbones (Fig. 3). E_H initially increased at small sample sizes before decreasing, likely because covariance rank is constrained when few models are available. Together, these results show that TRD-derived population statistics can be reliably estimated from finite model samples.

4.4 Preserving Representation Geometry with Compact TRD

We evaluated whether a cTRD constructed from a small observation subset can preserve the representation geometry and population structure of the full TRD. Experiments were conducted on nine selected D-datasets. For each domain, we constructed a population of 50 models, each trained on an independently sampled, class-balanced set of 1,000 training examples. A common set of 100 observations was then used for CKA and TRD measurement. From these observations, we selected subsets of size $n \in 3, 6, 9$ and compared SGSO with K-Means, Farthest-Point Sampling (FPS), Facility Location, autoencoder bottleneck representation (AE Latent), PCA Extreme, Landmark Ordinal Selection (LOS) (Anderton & Aslam, 2019), and Local Ordinal (See Appendix A.4). SGSO

Table 3: Pearson correlation (r) between CKA- and cTRD-derived model-pair similarities across nine selected D-datasets (mean ± SD). Full TRD is included as a full-observation reference.

Method	ResNet-18			DeiT-Tiny		
	$n = 3$	$n = 6$	$n = 9$	$n = 3$	$n = 6$	$n = 9$
Full TRD	0.979 ± 0.016	0.979 ± 0.016	0.979 ± 0.016	0.956 ± 0.040	0.956 ± 0.040	0.956 ± 0.040
Random	0.202 ± 0.192	0.602 ± 0.196	0.666 ± 0.162	0.558 ± 0.237	0.677 ± 0.210	0.752 ± 0.171
K-Means	0.335 ± 0.161	0.575 ± 0.082	0.694 ± 0.091	0.404 ± 0.277	0.850 ± 0.064	0.884 ± 0.052
FPS	0.336 ± 0.095	0.657 ± 0.090	0.751 ± 0.077	0.613 ± 0.182	0.876 ± 0.085	0.922 ± 0.039
AE Latent	0.030 ± 0.122	0.019 ± 0.105	0.088 ± 0.107	-0.012 ± 0.048	0.011 ± 0.041	-0.008 ± 0.030
PCA Extreme	0.312 ± 0.188	0.709 ± 0.080	0.762 ± 0.069	0.557 ± 0.216	0.885 ± 0.067	0.925 ± 0.032
Facility Location	0.283 ± 0.089	0.628 ± 0.134	0.715 ± 0.132	0.284 ± 0.200	0.786 ± 0.134	0.876 ± 0.088
LOS	0.318 ± 0.226	0.619 ± 0.155	0.672 ± 0.149	0.501 ± 0.210	0.845 ± 0.090	0.848 ± 0.096
Local Ordinal	0.060 ± 0.062	0.324 ± 0.267	0.548 ± 0.172	0.271 ± 0.215	0.715 ± 0.232	0.823 ± 0.128
SGSO	**0.708 ± 0.096**	**0.851 ± 0.048**	**0.898 ± 0.037**	**0.864 ± 0.064**	**0.961 ± 0.026**	**0.970 ± 0.024**

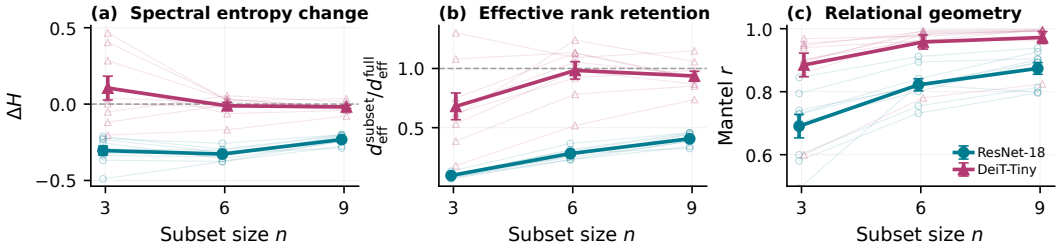


Figure 4: Population-structure preservation by cTRDs from SGSO-selected subsets. (a) Normalized spectral entropy change ($\Delta H = H_{\text{subset}} - H_{\text{full}}$). (b) Effective dimensionality retention ($d_{\text{eff}}^{\text{subset}}/d_{\text{eff}}^{\text{full}}$). (c) Relational geometry preservation measured by the Mantel correlation between full- and cTRD distance matrices. Thin lines show domains; thick lines show mean ± SEM.

used Kendall's τ between subset-induced cTRD similarities and reference CKA similarities as its subset-quality objective. Final correspondence was evaluated using Pearson correlation (r) between the corresponding model-pair similarities. SGSO achieved the highest performance across all subset sizes and both backbones (Table 3 and Appendix A.6), reaching $r = .708, .851, .898$ for ResNet-18 and $r = .864, .961, .970$ for DeiT-Tiny at $n = 3, 6, 9$, respectively. Performance increased with n, showing that a small set of informative observations selected by SGSO can preserve pairwise representation geometry. We further examined preservation of the full-TRD population structure (Fig. 4). Spectral preservation was backbone-dependent: DeiT-Tiny largely preserved spectral entropy and effective dimensionality at larger n, whereas ResNet-18 retained a lower-dimensional spectral structure. Despite this difference, relational geometry was consistently preserved, with Mantel correlations (Mantel, 1967) increasing with n and approaching .9 or higher. Thus, cTRDs constructed from SGSO-selected subsets can preserve population-level relational geometry without fully reproducing the spectral structure of the full TRD. Finally, we examined the optimization process (Fig. 5). A permutation-invariant DeepSets surrogate was initialized from randomly evaluated subsets and iteratively refined by combining random exploration with local search around high-scoring subsets. Surrogate prediction accuracy increased over iterations, as measured by prediction τ and top-30 precision (Fig. 5a). Surrogate-guided neighborhoods also achieved higher best and mean subset quality than random search (Fig. 5b–c), indicating that the surrogate progressively concentrated the search on high-quality regions. Overall, SGSO identifies compact observation sets that preserve both model-pair representation geometry and population-level relational structure.

4.5 Generalization of Optimized Observation Subsets to Unseen Models

We tested whether observation subsets generalize to models not used for subset selection and how this depends on subset size (See Appendix A.7). We used M-datasets with ResNet-50 and ViT-Small, each trained on 4,000 sampled, class-balanced examples. For each dataset–backbone pair, SGSO used 50 models and 200 observations to select $n \in \{10, 15\}$ observations, with $\lambda_{\text{reg}} = 0.1$. Subsets were evaluated on the 50 optimization models (Seen) and 100 independent models (Unseen). Table 4 reports the Pearson correlation between pairwise CKA and descriptor-derived model relationships.

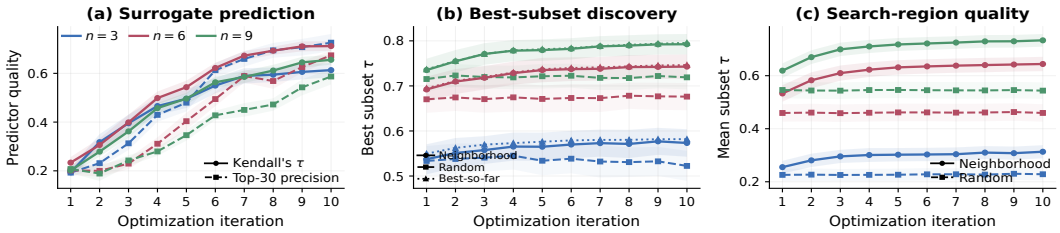


Figure 5: Optimization dynamics of surrogate-guided TRD subset selection. (a) Surrogate performance: Kendall's τ between predicted and true subset quality and Top-30 precision. (b) Best quality from local search (Neighborhood best), random search (Random best), and all evaluated subsets (Best-so-far). (c) Mean quality of local-search and random subsets. Subset quality is Kendall's τ between observer-pair similarities from subset TRDs and reference CKA similarities.

Table 4: Pearson correlation between cTRD- and CKA-derived model-pair similarities on M-datasets. Full denotes the full TRD; Seen and Unseen denote the optimization and independent evaluation populations, respectively. PCA-Ext. and Facility denote PCA Extreme and Facility Location, respectively; methods requiring the full analysis set were excluded.

				$n = 10$						$n = 15$					
Dataset	Backbone	Eval.	Full	Random	K-means	FPS	PCA-Ext.	Facility	SGSO	Random	K-means	FPS	PCA-Ext.	Facility	SGSO
CAR	ResNet-50	Seen	.980	.516	.626	.733	.810	.778	**.925**	.572	.761	.847	.785	.858	**.940**
		Unseen	.980	.459	.729	.754	.840	.692	**.896**	.469	.806	.848	.829	.849	**.929**
	ViT-Small	Seen	.970	.837	.884	.673	.630	.818	**.954**	.866	.930	.841	.794	.887	**.963**
		Unseen	.975	.843	.850	.582	.621	.824	**.900**	.892	.894	.748	.762	.880	**.924**
CUB	ResNet-50	Seen	.957	.462	.551	.847	.854	.787	**.944**	.482	.701	.871	.866	.774	**.953**
		Unseen	.952	.515	.559	.809	.812	.721	**.909**	.534	.674	.835	.821	.727	**.924**
	ViT-Small	Seen	.966	.733	.722	.545	.772	.794	**.885**	.811	.873	.684	.810	.847	**.928**
		Unseen	.968	.720	.710	.535	.724	.747	**.783**	.813	.843	.687	.774	.854	**.878**
IMG	ResNet-50	Seen	.916	.396	.513	.640	.618	.590	**.839**	.394	.659	.753	.762	.717	**.866**
		Unseen	.917	.459	.497	.631	.597	.622	**.808**	.482	.663	.747	.742	.714	**.822**
	ViT-Small	Seen	.986	.861	.838	.725	.780	.848	**.908**	.896	.881	.796	.844	.901	**.929**
		Unseen	.986	.861	.848	.740	.815	.846	**.884**	.895	.876	.830	.880	.908	**.918**

SGSO achieved the highest correlation among subset-based methods across dataset–backbone pairs in the Seen population for both subset sizes, and this advantage largely transferred to Unseen models. Increasing n from 10 to 15 improved correspondence; for example, Unseen correlation increased from .896 to .929 for CAR with ResNet-50 and from .884 to .918 for IMG with ViT-Small. Despite some Seen-to-Unseen degradation, subsets optimized on 50 models retained strong correspondence with CKA relationships on 100 independent models, indicating that SGSO preserves model-level representation geometry beyond the population used for selection.

5 Discussion and Conclusion

This work studied representation variability as a population-level property of independently trained models and introduced TRD as a model-level descriptor. Our results show that variability under shared learning conditions exhibits measurable and reproducible statistical structure. High dispersion suggests that a single model may provide a limited view of a learning condition, motivating repeated training and population-aware analysis. Such variability may also inform diversity-aware model selection, ensemble construction, and analyses of how supervision constrains representations. TRD characterizes individual models in a shared observational space while preserving model-to-model relationships that closely correspond to CKA. Unlike pairwise similarity measures, it assigns each model an explicit descriptor in this shared space. To address the dimensional growth of full TRD with the number of observations, we introduced Surrogate-Guided Subset Optimization (SGSO) for constructing compact TRDs. Optimized subsets preserved representation geometry and generalized to unseen model populations. Together, population-level dispersion and model-level descriptors provide complementary views of learned representations, enabling characterization of individual representations and their population-level variability across learning conditions.

AI USE STATEMENT

Large language models (LLMs) were used as supporting tools during manuscript preparation and code implementation. For writing, their use was limited to paragraph-level assistance with language editing, condensation, and LaTeX formatting, rather than generation or revision of the manuscript as a whole. LLMs were also used to identify potentially relevant literature; all suggested references and their relevance to the study were subsequently checked by the authors. For experimental code, LLM assistance was restricted to implementing individual modules based on algorithms and specifications defined by the authors. All AI-assisted code was reviewed and validated before use. We did not employ agentic AI systems to autonomously develop or modify the project from high-level instructions, but instead used LLMs interactively for narrowly defined tasks with explicit instructions. The authors remain fully responsible for the final manuscript, implementation, and experimental results.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our work. The code used in our experiments is provided in the supplementary material, together with documentation describing the data preprocessing procedures required to reproduce the experiments. We do not redistribute the benchmark datasets used in our experiments, as they are widely available public datasets; instead, the accompanying documentation provides detailed instructions for processing the datasets following the procedures used in our experiments. In addition to the provided code, the Appendix describes the key experimental parameters and comparison methods in detail.

REFERENCES

Vivek Anand, Alec Helbling, Mark Davenport, Gordon Berman, Sankaraleengam Alagapan, and Christopher Rozell. Lore: Jointly learning the intrinsic dimensionality and relative similarity structure from ordinal data. In *International Conference on Learning Representations*, volume 2026, pp. 3622–3655, 2026.

Jesse Anderton and Javed Aslam. Scaling up ordinal embedding: A landmark approach. In *International Conference on Machine Learning*, pp. 282–290. PMLR, 2019.

Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, NY, 2 edition, 2005.

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

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

Nikhil Ghosh, Yuxin Chen, and Yisong Yue. Landmark ordinal embedding. *Advances in neural information processing systems*, 32, 2019.

Sarah E Harvey, Brett W Larsen, and Alex H Williams. Duality of bures and shape distances with implications for comparing neural representations. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pp. 11–26. PMLR, 2024.

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.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006, 2020.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Ethan Hsu, Tony Cao, Lesia Semenova, and Chudi Zhong. The rashomon set has it all: Analyzing trustworthiness of trees under multiplicity. *Advances in Neural Information Processing Systems*, 38, 2026.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34: 28648–28662, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

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

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2_Part_1):209–220, 1967.

Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Yiyang Sun, Haiyang Huang, Gaurav Rajesh Parikh, and Cynthia Rudin. The rashomon effect for visualizing high-dimensional data. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Mobileone: An improved one millisecond mobile backbone. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7907–7917. IEEE, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Michael Wilber, Iljung Kwak, and Serge Belongie. Cost-effective hits for relative similarity comparisons. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pp. 227–233, 2014.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Jianyu Zhang and Léon Bottou. Learning useful representations for shifting tasks and distributions. In *International Conference on Machine Learning*, pp. 40830–40850. PMLR, 2023.

A Appendix

A.1 Statistical Interpretation of CKA-Based Representation Dispersion

We provide a more detailed statistical interpretation of the CKA-based representation dispersion (RD) estimator introduced in the main text. Let R denote a learned representation obtained under a fixed learning condition C, and suppose that independently trained models induce representations

$$R_1, \ldots, R_M \overset{\text{i.i.d.}}{\sim} \mathcal{P}(R \mid C).$$

Here, the condition C specifies the factors held fixed when defining a model population, such as the dataset, architecture, and learning configuration, while the remaining source of randomness determines variation among individual learned representations.

For two independently sampled representations R and R' from this population, we define their CKA-based dissimilarity as

$$h^{\mathrm{CKA}}(R, R') = 1 - \mathrm{CKA}(R, R').$$

The corresponding population-level representation dispersion is therefore

$$\mathrm{RD}_C^{\mathrm{CKA}} = \mathbb{E}_{R,R' \sim \mathcal{P}(R|C)} \left[1 - \mathrm{CKA}(R, R')\right], \tag{5}$$

where R and R' are independent draws from the same representation population. Thus, $\mathrm{RD}_C^{\mathrm{CKA}}$ measures the expected pairwise representational dissimilarity between two models learned under the same condition C. A small value indicates that independently sampled models tend to converge to similar representational geometries, whereas a larger value indicates greater variability within the representation population.

Given a finite sample of M learned representations, the population quantity in Eq. 5 is estimated by averaging the kernel over all distinct unordered model pairs:

$$\widehat{\mathrm{RD}}_C^{\mathrm{CKA}} = \binom{M}{2}^{-1} \sum_{i<j} h^{\mathrm{CKA}}(R_i, R_j) = \frac{2}{M(M-1)} \sum_{i<j} \left(1 - S_{ij}^{\mathrm{CKA}}\right), \tag{6}$$

where

$$S_{ij}^{\mathrm{CKA}} = \mathrm{CKA}(R_i, R_j).$$

Because there are $\binom{M}{2} = M(M-1)/2$ unique unordered pairs, the normalization in Eq. 6 gives equal weight to every distinct pair of sampled representations.

Equivalently, defining the mean pairwise CKA as

$$\overline{\mathrm{CKA}}_C = \frac{2}{M(M-1)} \sum_{i<j} S_{ij}^{\mathrm{CKA}}, \tag{7}$$

the estimator can be written as

$$\widehat{\mathrm{RD}}_C^{\mathrm{CKA}} = 1 - \overline{\mathrm{CKA}}_C. \tag{8}$$

This form provides an intuitive interpretation: RD is the complement of the average representational similarity within the sampled model population. Importantly, the quantity is computed from all model pairs rather than relative to a particular reference model. It therefore characterizes dispersion of the population as a whole rather than the deviation of individual models from an arbitrarily selected reference.

The estimator in Eq. 6 is a second-order U-statistic with symmetric kernel

$$h^{\mathrm{CKA}}(R_i, R_j) = 1 - \mathrm{CKA}(R_i, R_j). \tag{9}$$

In the standard U-statistic formulation,

$$U_M = \binom{M}{2}^{-1} \sum_{i<j} h(R_i, R_j), \tag{10}$$

which is identical to $\widehat{\mathrm{RD}}_C^{\mathrm{CKA}}$ when $h = h^{\mathrm{CKA}}$. Under the assumption that $R_1, \ldots, R_M$ are independent samples from $\mathcal{P}(R \mid C)$, its expectation is

$$\mathbb{E}\left[\widehat{\mathrm{RD}}_C^{\mathrm{CKA}}\right] = \mathbb{E}\left[\binom{M}{2}^{-1} \sum_{i<j} h^{\mathrm{CKA}}(R_i, R_j)\right] \tag{11}$$

$$= \mathbb{E}_{R,R' \sim \mathcal{P}(R|C)}\left[h^{\mathrm{CKA}}(R, R')\right] \tag{12}$$

$$= \mathrm{RD}_C^{\mathrm{CKA}}. \tag{13}$$

Hence, $\widehat{\mathrm{RD}}_C^{\mathrm{CKA}}$ is an unbiased estimator of the expected pairwise CKA dissimilarity under the assumed sampling model. The U-statistic interpretation also clarifies why all $\binom{M}{2}$ pairwise comparisons can be used without treating them as independent observations: although individual pairwise similarities share models and are therefore statistically dependent, the U-statistic is defined directly through this collection of overlapping pairs (Hoeffding, 1992).

As M increases, the sampled model population provides an increasingly stable estimate of the underlying expectation in Eq. 5. This motivates the sample-size analysis reported in our experiments, where we examine how rapidly the empirical RD estimate approaches its value obtained from a larger model population. In this sense, increasing M improves estimation of population-level representation variability without changing the definition of the underlying quantity.

Finally, RD should be distinguished from the diversity of a particular model relative to the remaining population. The latter can be written as

$$D_m^{\mathrm{CKA}} = \frac{1}{M-1} \sum_{j \neq m} \left(1 - S_{mj}^{\mathrm{CKA}}\right), \tag{14}$$

which measures how atypical representation R_m is relative to the other sampled representations. By contrast, $\widehat{\mathrm{RD}}_C^{\mathrm{CKA}}$ summarizes the entire population. Indeed,

$$\widehat{\mathrm{RD}}_C^{\mathrm{CKA}} = \frac{1}{M} \sum_{m=1}^{M} D_m^{\mathrm{CKA}}, \tag{15}$$

showing that population-level RD is exactly the average model-level diversity over the sampled population. This relationship connects the population variability analysis to the observer-specific diversity measure used in our subsequent experiments.

A.2 Finite-Sample Estimation of TRD Population Statistics

We evaluated how reliably population statistics derived from TRD can be estimated from a finite number of models. Experiments were conducted separately for each dataset–backbone combination (CUB, CAR, and IMG; ResNet-50 and ViT-Small) under a fixed learning condition. The full available model population was used as the reference population.

For each population, we computed reference values for representational dispersion $\mathrm{RD}_{\mathrm{ref}}$, total variance $\mathrm{Tr}_{\mathrm{ref}}$, normalized spectral entropy H_{ref}, and population mean μ_{ref}. For a sample of M models, the TRD vectors were stacked into a matrix and centered by their sample mean. Singular value decomposition (SVD) was then applied to the centered matrix. Given singular values $\{\sigma_k\}_{k=1}^{L}$, where L is the number of singular values, the covariance eigenvalues were computed as

$$\lambda_k = \frac{\sigma_k^2}{M-1}, \qquad k = 1, \ldots, L. \tag{16}$$

Total variance and normalized spectral entropy were computed as

$$\mathrm{Tr} = \sum_{k=1}^{L} \lambda_k, \qquad H = -\frac{\sum_{k=1}^{L} p_k \log p_k}{\log L}, \qquad p_k = \frac{\lambda_k}{\sum_{j=1}^{L} \lambda_j}. \tag{17}$$

We randomly sampled $M \in \{5, 10, 20, 40, 80, 160, 320\}$ models without replacement from each population. Sampling was repeated 30 times for each sample size. The population mean was computed from the uncentered TRD vectors as

$$\mu = \frac{1}{M} \sum_{m=1}^{M} \mathrm{TRD}_m. \tag{18}$$

The TRD matrix was then centered by this sample mean for computing Tr and H. For every sampled population, we recomputed the four statistics and measured their deviation from the corresponding full-population reference values as

$$E_{\mathrm{RD}} = |\mathrm{RD} - \mathrm{RD}_{\mathrm{ref}}|, \qquad E_{\mathrm{Tr}} = \frac{|\mathrm{Tr} - \mathrm{Tr}_{\mathrm{ref}}|}{\mathrm{Tr}_{\mathrm{ref}}}, \tag{19}$$

$$E_H = |H - H_{\mathrm{ref}}|, \qquad E_\mu = \frac{\|\mu - \mu_{\mathrm{ref}}\|_2}{\|\mu_{\mathrm{ref}}\|_2}. \tag{20}$$

For each sample size, we report the mean and standard deviation of each estimation error across the 30 random samples.

A.3 Detailed Experimental Settings for Population-Level Representation Variability

D-datasets and data splits. The population-level representation experiments were conducted on binary domains constructed from class pairs in ImageNet-100 (D-datasets). Within each domain, samples from the two classes were balanced. Unless otherwise specified, each classifier was trained on 1,000 examples (500 per class). Validation and representation-measurement samples were selected separately from the training samples and fixed within a domain so that representations from independently trained models were evaluated on the same observations. For each experimental condition, we constructed a population of 30 independently trained classifiers.

Model training. Models were trained from random initialization using cross-entropy loss and AdamW. We used a batch size of 128, a learning rate of 1×10^{-4} for the backbone and 5×10^{-4} for the classification head, and weight decay of 1×10^{-4}. Training was performed for at most 50

epochs. The classifier head contained a nonlinear hidden layer followed by dropout ($p = 0.2$) and the binary output layer; dropout was applied only within the classifier head.

Training was terminated when validation accuracy reached a predefined endpoint. If the endpoint was not reached, training was stopped when validation accuracy failed to improve for four consecutive epochs, and the parameters corresponding to the best validation accuracy were retained. Only models that reached the specified validation-accuracy endpoint were included in the resulting model population.

Sources of randomness. To examine whether population-level representation variability depends on the particular source of stochasticity, we separately constructed model populations under three randomness conditions: *data sampling*, *weight initialization*, and *training stochasticity*. The experimental factors not under investigation were held fixed as far as possible within each condition.

For *data-sampling randomness*, the validation and representation-measurement sets were fixed, while a new class-balanced training subset was sampled for each model. The initialization and training configuration were otherwise controlled across the population.

For *weight-initialization randomness*, the training, validation, and representation-measurement sets were fixed across models, while the random seed used for parameter initialization was varied. The remaining training configuration, including the random state governing stochastic training operations, was held fixed.

For *training randomness*, the training, validation, and representation-measurement sets and the parameter initialization were fixed, while the random state governing stochastic operations during optimization was varied across models. This condition captures variation arising from operations such as minibatch ordering and dropout while holding the initial model parameters and sampled training data constant.

Architecture comparison. To examine the dependence of population-level variability on model architecture, we additionally constructed model populations using ResNet-18, DeiT-Tiny, ConvNeXtV2-Atto, EfficientNet-B0, and MambaOut-Tiny. All architectures were trained from random initialization under the same binary classification framework and optimization settings described above. Each architecture used a lightweight classification head with dropout ($p = 0.2$). The hidden dimensions of the heads were 256 for ResNet-18, 192 for DeiT-Tiny, 320 for ConvNeXtV2-Atto, and 256 for EfficientNet-B0 and MambaOut-Tiny. ReLU, GELU, SiLU, and GELU activations were used for ResNet-18, DeiT-Tiny, EfficientNet-B0, and MambaOut-Tiny, respectively, with GELU also used for ConvNeXtV2-Atto.

For each architecture and domain, 30 independently trained models were constructed under the same experimental protocol. The training data, validation data, representation-measurement observations, optimization hyperparameters, and stopping criterion were controlled when comparing architectures, so that the architecture itself constituted the primary difference between conditions.

Learning-endpoint comparison. We further examined whether population-level variability changes with the stage at which learning is terminated. The validation-accuracy endpoint was therefore varied while keeping the domain, architecture, data construction, and remaining optimization settings unchanged. A model was retained when its validation accuracy first reached the specified endpoint. Comparing populations constructed at different endpoints allowed us to examine whether the magnitude and domain-wise pattern of representation variability remained consistent across different stages of learning.

Representation extraction and similarity measurement. Representations were extracted from the backbone output immediately before the classification head. For ResNet-18 and DeiT-Tiny, this corresponds to 512- and 192-dimensional representations, respectively. For architectures with spatial backbone outputs, global spatial averaging was applied before constructing the representation matrix. All models within a condition were evaluated on the same fixed representation-measurement set.

Given representation matrices R_i and R_j from two independently trained models, their similarity was measured using linear CKA. For a population of $M = 30$ models, CKA was computed for all

$\binom{M}{2}$ unordered model pairs, and population-level representational dispersion was estimated as

$$\widehat{\mathrm{RD}}_C^{\mathrm{CKA}} = \frac{2}{M(M-1)} \sum_{i<j} \left[1 - \mathrm{CKA}(R_i, R_j)\right].$$

We additionally measured R@1 for each model, defined as the proportion of representation-measurement samples whose nearest neighbor in representation space belonged to the same class.

Repeated population-level analysis. Population-level statistics were estimated over repeated trials to assess their stability rather than relying on a single sampled model population. For each learning condition, we evaluated both the magnitude of $\widehat{\mathrm{RD}}^{\mathrm{CKA}}$ and its domain-wise pattern. Correspondence of these patterns across randomness sources, architectures, and learning endpoints was assessed using correlations between domain-wise RD estimates. These comparisons test whether representation variability reflects a reproducible property of the learning condition rather than a pattern specific to one realization of training randomness.

A.4 Comparison Baselines for Compact TRD

We compared SGSO with several baselines covering random, representative, coverage-based, and ordinal selection strategies. For feature-based methods, model features were first averaged across the model population to obtain a common observation-level feature space. Unless otherwise noted, each method selected n observations, and all triplets induced by the selected observations were used to construct the corresponding compact TRD.

Random. n observations were sampled uniformly without replacement. This provides a selection-free reference for the expected quality of an arbitrary observation subset.

K-Means. K-Means with n clusters was applied to the mean observation features. For each cluster, the observation closest to its centroid was selected, providing a representative subset that covers the major feature-space clusters.

Farthest-Point Sampling (FPS). Starting from a random observation, FPS iteratively selected the observation with the largest minimum Euclidean distance to the current subset. This produces a geometrically dispersed subset that favors feature-space coverage.

PCA Extreme. PCA was applied to the mean observation features. Observations with the largest positive and negative projections along successive principal components were selected, favoring samples that span the major directions of feature variation.

AE Latent. As a dimensionality-matched representation baseline, we trained a Transformer-based autoencoder to reconstruct the full set of model features and used its bottleneck representation as the model descriptor. The latent dimension was matched to the dimensionality of the TRD induced by n observations. Unlike the other baselines, AE Latent does not explicitly select an observation subset.

Facility Location. Observations were selected greedily using a facility-location objective based on cosine similarity. At each step, the observation giving the largest increase in coverage of the full observation set was added:

$$S^* = \arg\max_{|S|=n} \sum_{i \in X} \max_{j \in S} \mathrm{sim}(x_i, x_j). \tag{21}$$

Landmark Ordinal Selection (LOS). We implemented a landmark-based ordinal selection baseline inspired by landmark approaches to ordinal embedding (Anderton & Aslam, 2019). A small set of landmarks was sampled, and ordinal triplets were constructed from both landmark-to-observation and observation-to-landmark distance orderings. Adjacent rank constraints were prioritized, followed by wider-rank constraints when needed. The number of triplets was matched to the TRD dimension induced by n observations.

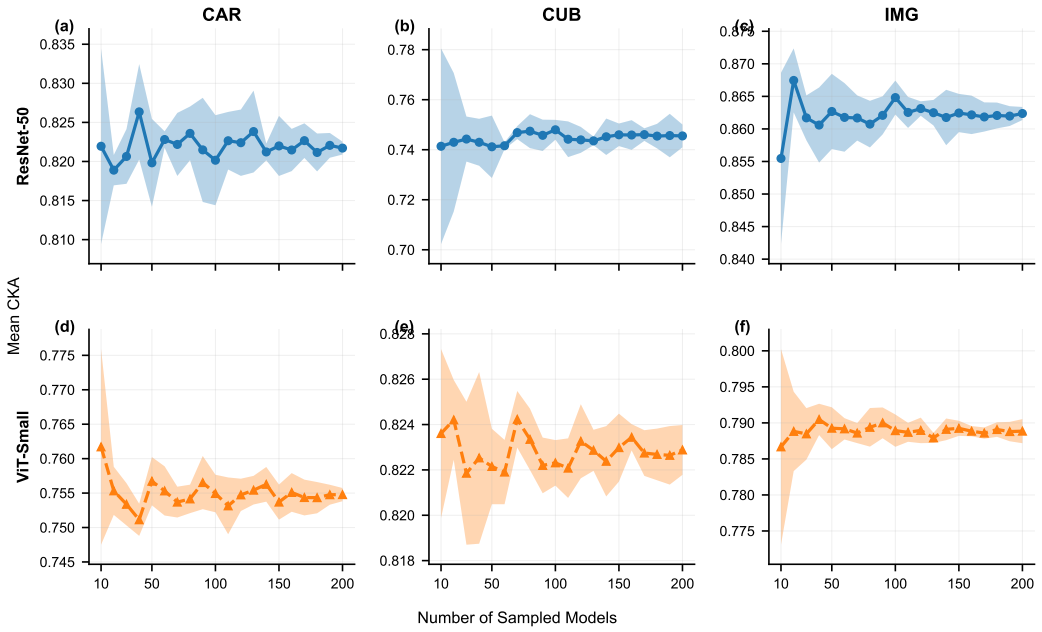


Figure 6: Sampling stability of population-level CKA estimates with increasing model sample size. Columns show CAR, CUB, and IMG, and rows show ResNet-50 and ViT-Small. For each M, population sampling was independently repeated five times without overlap among sampled models. Each replicate reports mean pairwise CKA, equivalent to $1 - \widehat{\mathrm{RD}}_C^{\mathrm{CKA}}$. Curves and shaded regions show the mean and ± 1 SD, respectively, across the five independently sampled populations, not across model pairs within a population.

Local Ordinal Selection. We also implemented a local ordinal baseline based on k-nearest-neighbor structure. For each anchor, triplets preferentially compared a local neighbor with a non-local observation; local–local comparisons were added when needed. The total number of triplets was again matched to that of an n-observation TRD.

A.5 Sampling Stability of Population-Level CKA

We additionally examined the sampling stability of population-level representation dispersion as a function of the number of sampled models. For each M-dataset (CAR, CUB, and IMG) and backbone (ResNet-50 and ViT-Small), models were sampled from a population of 1,000 independently trained models. For each sample size $M \in \{10, 20, \ldots, 200\}$, population sampling was independently repeated five times, with no overlap among the models sampled across the five repetitions. For each sampled population, CKA was computed for all $\binom{M}{2}$ model pairs and averaged to obtain $\overline{\mathrm{CKA}}_C = 1 - \widehat{\mathrm{RD}}_C^{\mathrm{CKA}}$.

Figure 6 reports the mean and standard deviation of these five independently sampled population estimates at each M, rather than the variability of pairwise CKA values within a single population. Across all dataset–backbone combinations, between-sample variability was larger at small M and decreased as more models were sampled, while the estimated population mean became increasingly stable. Although absolute CKA levels differed across datasets and backbones, this stabilization pattern was observed for both ResNet-50 and ViT-Small.

These results indicate that population-level representational dispersion can be estimated with increasing stability as the number of independently sampled models increases, providing empirical support for finite-model estimation of population-level representation statistics.

Table 5: Domain-wise subset-selection performance for ResNet-18. Each entry reports Pearson correlation (r) for $n = 3/6/9$, respectively. Only D-datasets with available results are shown. Bold values indicate the best subset-selection method for each domain and subset size. Full TRD is shown as a reference.

Method	D4	D5	D8	D10	D11	D12	D13	D18	D21
Full TRD	.973/.973/.973	.958/.958/.958	.949/.949/.949	.990/.990/.990	.993/.993/.993	.993/.993/.993	.991/.991/.991	.990/.990/.990	.976/.976/.976
Random	.168/.294/.395	.520/.638/.646	.190/.691/.719	-.068/.770/.802	.190/.851/.853	.481/.556/.587	.201/.690/.840	.122/.646/.696	.017/.283/.459
K-Means	.249/.458/.561	.485/.539/.625	.424/.589/.651	.428/.465/.710	.353/.644/.679	.111/.606/.753	.570/.692/.828	.103/.645/.814	.295/.536/.621
FPS	.329/.711/.771	.291/.691/.733	.414/.521/.590	.157/.762/.803	.331/.797/.827	.366/.630/.828	.497/.615/.786	.369/.607/.679	.269/.578/.743
AE Latent	-.077/.149/.118	.009/.022/-.003	.030/-.030/.043	-.151/-.129/.237	.243/.215/.248	-.081/-.015/.060	.140/-.043/.119	.076/.037/.056	.081/-.035/-.084
PCA Extreme	.061/.636/.745	.275/.588/.647	.571/.692/.672	.603/.791/.813	.097/.819/.844	.341/.663/.725	.201/.748/.807	.262/.788/.796	.392/.658/.809
Facility Location	.355/.557/.711	.217/.461/.671	.312/.687/.469	.195/.667/.788	.405/.767/.914	.284/.546/.731	.393/.837/.854	.159/.679/.699	.228/.448/.602
LOS	.038/.668/.674	.118/.594/.698	.330/.512/.477	.564/.401/.530	.554/.904/.946	.561/.783/.805	.230/.591/.597	.447/.638/.761	.018/.479/.558
Local Ordinal	.152/.205/.396	.030/.120/.483	.128/.274/.508	.028/.039/.629	.032/.745/.674	.090/.294/.392	.013/.522/.743	.101/.683/.798	-.040/.036/.304
SGSO	**.639/.799/.849**	**.650/.787/.857**	**.682/.859/.890**	**.755/.888/.929**	**.870/.923/.954**	**.740/.873/.904**	**.746/.886/.930**	**.756/.848/.908**	**.531/.792/.858**

Table 6: Domain-wise subset-selection performance for DeiT-Tiny. Each entry reports Pearson correlation (r) for $n = 3/6/9$, respectively. Only D-datasets with available results are shown. Bold values indicate the best subset-selection method for each domain and subset size. Full TRD is shown as a reference.

Method	D4	D5	D8	D10	D11	D12	D13	D18	D21
Full TRD	.891/.891/.891	.896/.896/.896	.936/.936/.936	.991/.991/.991	.997/.997/.997	.982/.982/.982	.986/.986/.986	.961/.961/.961	.968/.968/.968
Random	.281/.365/.380	.625/.659/.764	.696/.881/.867	.546/.827/.852	.786/.509/.781	.126/.410/.581	.867/.935/.938	.464/.672/.776	.631/.839/.829
K-Means	.135/.745/.834	.644/.908/.815	.159/.925/.874	.571/.768/.819	.766/.803/.903	.123/.899/.953	.167/.866/.937	.756/.883/.932	.318/.858/.887
FPS	.646/.817/.900	.795/.929/.934	.800/.889/.869	.665/.884/.912	.271/.948/.986	.571/.922/.941	.801/.890/.932	.541/.675/.868	.425/.932/.955
AE Latent	.017/-.008/-.004	-.069/-.010/-.002	-.018/-.008/-.027	.053/.011/.000	-.000/-.028/-.019	-.005/.072/.005	.051/-.015/-.020	-.065/-.004/-.061	-.069/.088/.053
PCA Extreme	.646/.845/.902	.504/.924/.935	.857/.814/.878	.507/.784/**.975**	.171/.955/.935	.648/.932/.943	.773/.951/.958	.304/.936/.900	.606/.822/.901
Facility Location	.010/.771/.855	.418/.884/.954	.170/.909/.850	.394/.609/.748	.239/.726/.904	.123/.927/.948	.229/.900/.944	.697/.785/.950	.272/.560/.729
LOS	.207/.776/.808	.650/.853/.880	.131/.707/.671	.679/.953/.918	.671/.968/.978	.378/.905/.936	.583/.883/.861	.644/.780/.827	.568/.776/.749
Local Ordinal	.090/.260/.841	.078/.732/.801	.050/.659/.628	.542/.913/**.975**	.656/.947/.918	.195/.854/.896	.151/.893/.939	.343/.732/.782	.335/.442/.623
SGSO	**.790/.902/.920**	**.908/.950/.963**	**.926/.977/.983**	**.814/.956**/.948	**.878/.980/.990**	**.906/.987/.991**	**.950/.977/.986**	**.767/.975/.983**	**.840/.947/.972**

A.6 Domain-Wise Results of Compact TRD Subset Selection in D-datasets

Table 5 and Table 6 provide the domain-wise results underlying the aggregate comparison reported in Table 3. Results are shown for the nine selected D-datasets with available experiments (D4, D5, D8, D10, D11, D12, D13, D18, and D21), separately for ResNet-18 and DeiT-Tiny.

For each domain, we report the Pearson correlation between CKA-derived and cTRD-derived model-pair similarities for observation subset sizes $n \in \{3, 6, 9\}$. Full TRD is included as a reference, and SGSO is compared with Random, K-Means, FPS, AE Latent, PCA Extreme, Facility Location, LOS, and Local Ordinal. Each table entry reports the results for $n = 3/6/9$ in that order. Bold values indicate the highest correlation among subset-selection methods for each domain and subset size.

The domain-wise results show that SGSO generally maintains its advantage across individual domains and subset sizes rather than deriving its aggregate performance from a small number of domains. Performance also generally improves as the observation subset size increases from $n = 3$ to $n = 9$, consistent with the aggregate results in the main text. An exception occurs for DeiT-Tiny on D10 at $n = 9$, where PCA Extreme and Local Ordinal achieve higher correlation than SGSO. These results provide the individual-domain breakdown of the compact-TRD evaluation summarized in the main text.

A.7 Subset Optimization for M-datasets

For each M-dataset (CUB, CAR, and IMG) and backbone (ResNet-50 and ViT-Small), we constructed a population of independently trained classification models. Each model was trained on an independently sampled, class-balanced set of 4,000 training examples under the same training configuration. For the generalization experiment, 50 models were used for subset optimization (Seen), and a separate population of 100 independently sampled models was used for evaluation (Unseen). A common set of 200 observations was used as the candidate pool. Using all observations yields

$$T(200) = 200\binom{199}{2} = 3{,}940{,}200, \tag{22}$$

corresponding to a 3,940,200-dimensional full TRD. Despite this high dimensionality, full-TRD similarities can be computed in practical time by vectorizing triplet construction and evaluating

model-pair similarities with parallel matrix operations. We therefore used the full descriptor as a reference geometry while seeking much smaller observation subsets that preserve this geometry.

For a candidate subset S of size n, we constructed all triplets consisting of one anchor and an unordered pair of the remaining observations, yielding

$$T(n) = n\binom{n-1}{2} = \frac{n(n-1)(n-2)}{2}. \tag{23}$$

Thus, $n = 10$ and $n = 15$ produce 360 and 1,365 triplets, respectively, reducing the descriptor from 3,940,200 dimensions to 360 or 1,365 dimensions. For model m, each triplet (a, b, c) was represented by the relative distance

$$r_m(a, b, c) = d_m(a, c) - d_m(a, b), \tag{24}$$

where d_m is the Euclidean distance in the representation space of model m. Vectorizing these values over all triplets defines the cTRD induced by S.

To evaluate a candidate subset, cTRDs were ℓ_2-normalized and their pairwise cosine similarities were computed across the 50 Seen models. Let

$$\mathbf{s}^{\mathrm{CKA}} = [\mathrm{CKA}(R_i, R_j)]_{i<j} \tag{25}$$

denote the reference pairwise similarities computed from the full observation set, and let

$$\mathbf{s}^{S} = [\cos(\mathrm{cTRD}_i^S, \mathrm{cTRD}_j^S)]_{i<j} \tag{26}$$

denote the corresponding similarities induced by candidate subset S. Subset quality was defined as

$$Q(S) = \tau\left(\mathbf{s}^{\mathrm{CKA}}, \mathbf{s}^{S}\right), \qquad S^* = \arg\max_{\substack{S \subset X \\ |S|=n}} Q(S), \tag{27}$$

where τ denotes Kendall's rank correlation. Thus, Subset Optimization seeks a compact observation set whose induced cTRD preserves the CKA-based geometry of the model population.

For efficient subset evaluation, pairwise Euclidean distances among all 200 observations were precomputed separately for each model. Given a candidate subset, its triplet relations could then be obtained directly from these distance matrices without recomputing model representations or pairwise distances. Candidate subsets were evaluated in parallel, and all model-pair cTRD similarities for a subset were computed simultaneously by matrix multiplication. This allowed exact evaluation of many candidate subsets while avoiding repeated high-dimensional feature computations.

Because exhaustive search over all $\binom{200}{n}$ subsets remains infeasible, we used a surrogate-guided iterative search. We first generated 300 random subsets and evaluated their exact objective values $Q(S)$. These evaluated subset–score pairs were used to train a permutation-invariant DeepSets surrogate $g_\theta(S)$ that predicts subset quality from observation indices (Zaheer et al., 2017). Each observation index was represented by a learnable 64-dimensional embedding, followed by two 256-dimensional hidden layers and mean pooling over the subset. The pooled representation was mapped to a scalar subset score by a second network. The surrogate was optimized using Adam with a learning rate of 10^{-3} and a batch size of 64. Its objective combined a pairwise logistic ranking loss with an auxiliary regression loss,

$$\mathcal{L}_{\mathrm{sur}} = \mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{MSE}}, \qquad \lambda_{\mathrm{reg}} = 0.1. \tag{28}$$

The initial surrogate was trained for 30 epochs.

Subset search was then performed for 15 iterations, gradually shifting from exploration to exploitation. In iterations 1, 2, and 3–15, respectively, 700/300, 500/500, and 200/800 random/exploitation candidates were generated. Exploitation candidates were produced by perturbing one of the 30 highest-scoring previously evaluated subsets. The surrogate ranked the candidate subsets, after which the top 30 candidates were retained for local refinement. For each retained candidate, neighboring subsets were generated by replacing a small number of observations. The number of local neighbors was gradually reduced from 10 to 6 over the search.

At each iteration, the true objective $Q(S)$ was evaluated for selected local neighbors together with additional randomly sampled subsets. The surrogate training set was then updated using high-scoring previously evaluated subsets, newly evaluated neighbors, and random subsets, after which

g_θ was retrained. This iterative procedure combines global exploration with local refinement around promising regions of the subset space. The best subset encountered during the entire search was selected as

$$S^* = \arg\max_{S \in \mathcal{D}} Q(S), \tag{29}$$

where $\mathcal{D}$ denotes the set of subsets evaluated with the true objective.

For the generalization experiment, S^* was selected using only the 50 Seen models. The same observation subset was then applied without further optimization to the independent population of 100 Unseen models. For each population, we constructed cTRDs using S^* and evaluated their model-pair similarities against the corresponding CKA similarities using Pearson correlation. This separates subset optimization from evaluation and tests whether an observation subset selected from one model population preserves representation geometry in independently trained models.