PROCRUSTES SIMILARITY WITH THE RECONSTRUCTED SPACE: QUANTIFYING REPRESENTATIONAL SIMILARITY THROUGH RELATIVE INFORMATION CONTENT

Anonymous authors
Paper under double-blind review

ABSTRACT

With the growing reliance on high-dimensional embeddings in many areas including machine learning, neuroscience, and computational biology, the need for rigorous methods to compare and assess these representations and their information content is critical. Quantifying the similarity between two variables or distributions is a longstanding challenge in statistics and information theory, and this challenge becomes even more pronounced in high-dimensional spaces. Established methods, such as Mutual Information (MI), Canonical Correlation Analysis (CCA), and Centered Kernel Alignment (CKA), struggle with accurately capturing complex dependencies, handling noise, and modeling nonlinear relationships in high-dimensional settings, limiting their practical usability. To address this, we introduce Procrustes Similarity with the Reconstructed Space (PSRS), a novel approach that quantifies the Relative Information Content (RIC) between two embedding spaces by reconstructing one from another using deep learning techniques and measuring the reconstruction fidelity via Procrustes similarity. Through a combination of synthetic tests and real-world case studies, we demonstrate that PSRS significantly outperforms state-of-the-art methodologies in capturing both linear and nonlinear relationships while exhibiting superior robustness to noise and extraneous information.

1 INTRODUCTION

The concept of embedding spaces has evolved from early dimensionality reduction techniques like PCA (Pearson, 1901; Jolliffe, 2002) and LDA (Fisher, 1936) to highly complex, high-dimensional representations in modern deep learning models. Advances such as autoencoders (Hinton & Salakhutdinov, 2006) and Word2Vec (Mikolov, 2013) have established embeddings as a powerful way to encapsulate rich, context-specific information in a continuous space, where algebraic operations capture meaningful relationships.

Systematic comparison of different embeddings has proven to be a valuable source of insight and an effective tool for scientific discoveries. Whether comparing different embedding approaches applied to the same set of objects, or comparing the same embedding approach across different datasets, such comparative analysis can reveal structures and relationships that are not apparent when examining a single embedding space (Sivaraman et al., 2022; Boggust et al., 2019).

For example, Procrustes-based methods have long been a cornerstone in evolutionary biology and morphology, where they are widely used for aligning and comparing geometric structures (Bookstein, 1997). In morphometrics, these techniques help infer evolutionary relationships that may not be apparent from raw measurements (Pillay et al., 2024), while in neuroscience (Tong et al., 2022), linguistics (Enggaard et al., 2024), and genomics, embedding comparison and alignment methods have been applied to uncover structural similarities across diverse datasets.

Additionally, techniques such as Canonical Correlation Analysis (CCA) (Hardoon et al., 2004), Centered Kernel Alignment (CKA) (Cortes et al., 2012; Kornblith et al., 2019), and Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) have been widely adopted in deep learning, representation learning, and generative models. In multimodal learning, for instance, Deep CCA (DCCA) plays a crucial role in aligning different modalities, such as learning joint embeddings

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.

between images and text in vision-language models (Yan & Mikolajczyk, 2015). CKA has been extensively applied to compare representations in deep neural networks, providing insights into how layers of different architectures develop similar or distinct feature spaces (Phang et al., 2021). Meanwhile, MINE has emerged as a tool in self-supervised learning and feature selection, enabling the measurement of information flow between different network layers (Hjelm et al., 2018).

There is an increasing body of work that emphasizes the centrality of representational alignment as a research problem, both conceptually and empirically. Recent studies (Sucholutsky et al., 2023; Klabunde et al., 2024; Ding et al., 2021) highlight that despite the growing reliance on high-dimensional embeddings, no existing similarity measure achieves generality, robustness, and interpretability across domains.

In this work, we introduce a novel framework for evaluating and comparing embedding spaces, offering a robust methodology with a broad applicability.

2 Related Works

From a broad perspective, embedding comparison approaches can be categorized into two main groups, each with several subcategories. The two primary categories are intrinsic methods, which evaluate the internal structure and properties of embeddings independently of any external tasks, and extrinsic methods, which assess the performance of embeddings based on their effectiveness in specific tasks and their ability to generalize across different tasks or domains. In this section, we explore various approaches within each category in more details.

Intrinsic methods analyze the internal structure and properties of embeddings without relying on external tasks. These methods can be further divided into three subcategories: (1) structural evaluation approaches, (2) alignment and transformation methods, and (3) statistical and information-theoretic measures.

Structural evaluation assesses clustering quality and neighborhood preservation between embeddings. For example, quantitative metrics such as Trustworthiness and Continuity (Van Der Maaten et al., 2009) are commonly used to evaluate the quality of different dimensionality reduction or embedding techniques. Meanwhile visualization techniques like t-SNE (Van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018) have long been used as qualitative tools to assess how well embeddings capture the natural groupings in data.

More recently, Centered Kernel Alignment (CKA) has emerged as a powerful method for comparing embedding representations (Kornblith et al., 2019). It builds on the Hilbert-Schmidt Independence Criterion (HSIC), a statistical measure of dependency between two sets of variables (Gretton et al., 2005; Wang et al., 2021), by first centering the kernel (Gram) matrices of each representation and then computing their normalized alignment using HSIC. The Gram matrices can be constructed using both linear kernels, which capture dot-product relationships, and nonlinear kernels (e.g., RBF kernel), which capture more complex feature interactions in high-dimensional spaces.

Alignment and transformation methods are used to directly compare different embedding spaces. For instance, Procrustes based methods (Williams et al., 2021; Pospisil et al., 2023; Gower & Dijksterhuis, 2004), align embeddings through scaling, translation, and rotation, providing a measure of disparity between two spaces. Other cross-domain mapping techniques, such as Canonical Correlation Analysis (CCA) (Hotelling, 1992), can reveal common structures shared across different embeddings.

Several extensions of CCA, such as Deep CCA (Andrew et al., 2013), Kernel CCA (Akaho, 2006), and SVCCA (Raghu et al., 2017), have been developed to improve CCA's applicability, for example, by capturing nonlinear dependencies and improve robustness. Though, they often struggle to generalize across diverse scenarios (Yang et al., 2019; Uurtio et al., 2017). This underscores the need for a unified approach that captures shared structures across diverse datasets while maintaining robustness to varying dimensionalities, nonlinearities, and noise.

A more theoretical approach to embedding comparison involves **statistical and information-theoretic measures**, which provide quantitative insights into the relationships and dependencies between embedding spaces. However, extending these methods to higher dimensions remains challenging. Traditional measures such as Mutual Information (Shannon, 1948; Kinney & Atwal, 2014;

Zhang, 2020) and KL Divergence (Kullback & Leibler, 1951) have been used to assess the statistical dependencies between variables. To address scalability issues, approaches like Mutual Information Neural Estimation (MINE) leverage neural networks to estimate MI, enabling more effective computation in high-dimensional spaces and with large sample sizes (Belghazi et al., 2018).

Unlike intrinsic methods, extrinsic approaches evaluate embedding quality based on their practical utility in downstream tasks, such as classification, ranking, or recommendation. For general-purpose models, predefined benchmark datasets such as MMLU (Hendrycks et al., 2020), GPQA (Rein et al., 2023) and MATH (Mishra et al., 2022) are often used to compare the quality of embeddings.

The approach we are offering in this paper is an intrinsic approach that compare embedding spaces by quantifying the Relative Information Content (RIC) between them.

3 METHODS

We begin this section by defining the problem more precisely. Consider a set of N objects, each represented as a vector in two distinct embedding spaces: Embedding Space A (ES-A) and Embedding Space B (ES-B). Let ES-A be p-dimensional and ES-B q-dimensional, where p and q may differ. The goal of this study is to quantify RIC between ES-A and ES-B, measuring the extent to which the information in ES-B is present in ES-A, and vice versa. Later in this section, we will clearly define "information" in this context. To evaluate RIC, we implement a two-step methodology: (1) Reconstructing ES-B from ES-A, and (2) Measuring how well the reconstructed ES-B (denoted as Rec-B) captured the information in the original ES-B. A key aspect of this analysis is establishing a precise definition of "information" in this context.

While information can be defined through various theoretical lenses (Shannon, 1948; Fisher, 1922), in this paper, we adopt a definition grounded in the concept of variance, analogous to R^2 in regression and explained variance in PCA (see Appendix A.3 for a more detailed discussion).

It is also worth noting that, while the examples we use in this paper to demonstrate the effectiveness of this methodology are based on different embeddings for the same set of objects, the methodology is general and can also be applied to compare embeddings of different sets of objects as in morphometrics and shape analysis problems.

Step 1: Reconstruction of ES-B from ES-A. Reconstructing ES-B from ES-A requires a method that can effectively translate between spaces of differing dimensionalities while capturing both linear and nonlinear relationships. To achieve this, we use deep learning, which is is well-suited for this task due to its flexibility and ability to model complex functions. However, the performance of deep learning models can vary significantly due to factors such as network architecture, training hyperparameters, and random initialization of parameters. This variability is a concerning issue when attempting to consistently assess RIC between embedding spaces.

To address this issue, we develop a specific deep learning architecture named DeepShallowNet, inspired by the Wide and Deep architecture (Cheng et al., 2016) and residual network (He et al., 2016). DeepShallowNet (depicted in Appendix A.6) consists of two independent branches: a deep branch, which is a multi-layer feedforward network, and a shallow branch, which is a single linear layer. Both branches transform the input to match the dimensionality of the target space, and their outputs are summed up to construct the final representation. This architecture is designed to effectively capture both linear and nonlinear transformations in high-dimensional spaces, and exhibits consistent performance across different training instances and hyperparameters. Although empirical validation of the architecture's stability in different scenarios is valuable, detailed analysis of this aspect is beyond the scope of this paper and is therefore not discussed here. Nevertheless, in Appendix A.6, we demonstrate the stability of DeepShallowNet under various architectural and hyperparameter configurations.

As for the loss function, it is natural to choose Mean Squared Error (MSE) since we use variance as a proxy for information. MSE inherently minimizes the difference between reconstructed and target embeddings in a way that aligns with variance preservation, ensuring that the model optimally captures the structural relationships between the two spaces. This makes MSE well-suited for our methodology, as it directly encourages the reconstruction process to retain the same variance

distribution as the target embeddings. The model training details used in this paper are provided in Appendix A.

Additionally, to ensure that the reconstruction process does not overfit or falsely detect information patterns that do not inherently exist in the embeddings, we use a train-test separation strategy. The model is trained on a portion of the data (training set), but the next step is applied to a separate test set, which is transformed by the model but never seen during training. This ensures that the evaluation measures the actual shared information between embedding spaces rather than artifacts introduced by the training process.

Step 2: Measuring the reconstruction fidelity. Following the reconstruction of ES-B from ES-A, the next step is to evaluate how accurately the reconstructed space (Rec-B) captures the information in ES-B. To quantify this, we use Procrustes Similarity, a metric derived from Procrustes analysis. Procrustes analysis is a method that finds the optimal linear transformation (translation, scaling, and rotation) that minimizes the differences between corresponding points in two datasets. To ensure translation and scale invariance, both datasets are first centered and normalized such that their total variance (the Frobenius norm squared) equals 1. One of the spaces is then orthogonally rotated to best align with the other.

The Procrustes discrepancy, D, is then calculated as the Sum of Squared Residuals (SSR) between the aligned datasets, which represents the proportion of unexplained variance, due to the normalization. Finally, we define Procrustes Similarity as $1 - D$. If we denote the original spaces X and Y, then:

$$\begin{gathered} D = \sum_{i,j} \left(x_{ij}^{\text{aligned}} - y_{ij}^{\text{aligned}} \right)^2 = SSR \\ \text{Procrustes Similarity} = 1 - D = 1 - SSR \end{gathered} \tag{1}$$

where i indexes the data points, and j indexes the dimensions. For one-dimensional spaces, this is directly analogous to R2, which measures the explained variance ratio. In higher-dimensional spaces, Procrustes Similarity provides a similar interpretation by quantifying the proportion of explained variance across multiple dimensions.

Thus, in the context of this study, the Procrustes similarity between Rec-B and ES-B represents the proportion of variance–and hence information–in ES-B that is captured in Rec-B, and by extension, how much of ES-B's information is present in ES-A. We refer to this two-step approach–reconstructing one embedding space from another and evaluating the quality of reconstruction using Procrustes Similarity–as Procrustes Similarity with the Reconstructed Space (PSRS). The notation PSRS(A $\rightarrow$ B) reads as PSRS from A to B.

Theoretical properties of PSRS. Before presenting empirical evaluations, it is important to highlight the theoretical properties of PSRS. First, the upper bound of PSRS is 1, which reflects the non-negative nature of Procrustes discrepancy. An upper bound of 1 indicates perfect alignment between the reconstructed space and the original space, meaning the full variance–and thus the entire information content– of one space exists in the other. Similar to the R2, the lower bound of PSRS could theoretically be negative, however, we have not ever observed such cases in practice. For unrelated embedding spaces, PSRS values close to zero are typically expected.

A second fundamental property of PSRS is its asymmetry, meaning that, $\text{PSRS}(A \rightarrow B) \neq \text{PSRS}(B \rightarrow A)$ in general. This asymmetry reflects the inherent imbalance in information content between two representations; that is, the amount of information from ES-A that is present in ES-B may differ from the amount of information from ES-B that is present in ES-A. A simple example demonstrating this inherent asymmetry is dimensionality reduction, where a high-dimensional embedding may contain all the information present in its lower-dimensional version, but the reverse is not true. While this directionality is useful for studying asymmetric information retention, some analyses require a single symmetric score summarizing the similarity between two representations. For this purpose, we define symmetric PSRS as:

$$\text{sPSRS}(A, B) = \min(\text{PSRS}(A \rightarrow B), \text{PSRS}(B \rightarrow A)) \tag{2}$$

A third important characteristic is that PSRS is invariant to scaling and translation, as Procrustes analysis inherently accounts for these transformations. This invariance ensures that the metric is robust to differences in the magnitudes and offsets of the embeddings, focusing solely on the

underlying structure and relationships between two spaces. Finally, it is essential to emphasize that PSRS measures only RIC between embedding spaces, and does not provide any direct insights into the quality or usefulness of that information.

These theoretical properties provide a foundation for the subsequent empirical analysis, where PSRS is systematically evaluated across various scenarios to assess its effectiveness in quantifying RIC.

4 RESULTS

In this section, we present several examples to demonstrate the effectiveness and accuracy of the PSRS approach in measuring RIC and representational similarity. These examples are carefully chosen to highlight three essential capabilities of PSRS: (1) its ability to capture both linear and nonlinear relationships between embedding spaces, (2) its applicability to compare spaces with different dimensionalities, and (3) its robustness to noise and extraneous information. First, we apply PSRS to quantify the RIC between two completely independent embedding spaces. Then, we evaluate PSRS on two linearly correlated spaces with differing dimensions, followed by an analysis of two spaces with highly nonlinear correlations. Next, we examine the robustness of PSRS against noise and extraneous information. All of these analyses are performed on synthetic datasets. We create these embedding spaces from randomly generated datasets, which allows us to carefully curate them to serve specific objectives while ensuring that a ground truth is available to assess PSRS's accuracy. To ensure the generalizability of results, we introduce random linear correlations between dimensions in the generated embeddings and repeat each analysis on five datasets with different random seeds. We report the mean and standard deviation of the resulting PSRS values. Each space consists of 10,000 data points, each represented as a 100-dimensional vector.

In all these cases, we compare the PSRS results with the theoretical expectations and with two other cutting-edge methods, DCCA and CKA with a nonlinear kernel (RBF kernel). Since these methods are designed to measure similarity between two embedding spaces using related but different metrics, direct comparison of their absolute values is only meaningful in boundary cases where the embedding spaces are either completely similar or independent. Specifically, PSRS measures explained variance, while DCCA measures canonical correlation and CKA evaluates structural alignment in high-dimensional spaces. Thus in non-boundary cases, we focus primarily on the qualitative trends observed across different conditions.

Finally, we apply this framework to evaluate information retention in Matryoshka Representation Learning (MRL), quantifying how much information from full-dimensional embeddings is preserved in their lower-dimensional prefixes.

Across all examples, we keep the hyperparameters of all methods fixed, using their recommended default values. While it would have been possible to tune hyperparameters to achieve closer alignment with the ground truth in some cases, we purposefully avoided such tuning. In real-world applications, the ground truth is typically unknown, and hyperparameter optimization based on it would not be feasible. Therefore, we opted to simulate a more realistic scenario where models are evaluated under default settings without privileged information. The detailed list of all hyperparameters are listed in the appendix.

4.1 TWO INDEPENDENT EMBEDDING SPACES

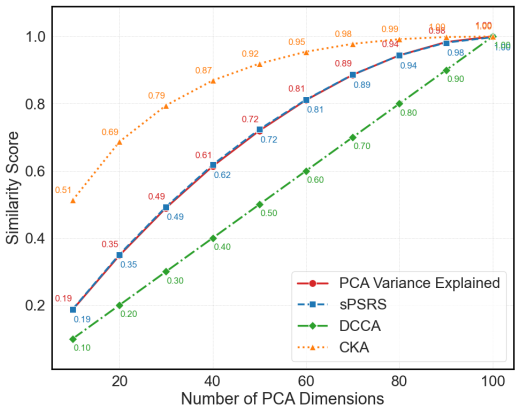
In this example, we generate two completely independent embedding spaces by randomly creating two sets of 10,000 100-dimensional vectors, referred to as ES-A and ES-B. Since these spaces are generated independently, they are theoretically uncorrelated, meaning that the RIC between them should be close to zero. In other words, ES-B should not be reconstructible from ES-A, and vice versa. Calculations using sPSRS confirm this expectation (see Table 1) and yield a value of 0.016 (± 0.000), indicating close to zero overlap between the two spaces, as anticipated. Similarly, CKA reports a low similarity score of 0.008 (± 0.000), consistent with the independence of the spaces, whereas DCCA outputs an inflated score of 0.19 (± 0.002), likely due to its tendency to overfit and extract spurious correlations from unstructured noise in high-dimensional data.

Table 1: Comparison of similarity scores across different embedding space relationships for sPSRS, nonlinear CKA, and DCCA.

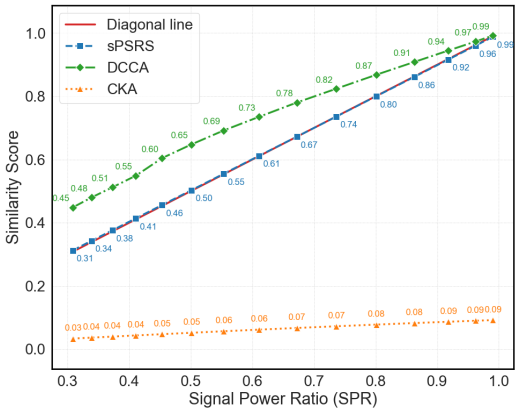
Methods \ Spaces	Completely Independent	Perfect Linear Correlation	Perfect Nonlinear Correlation
Ground Truth	0	1	1
sPSRS	0.016 ($\pm$0.000)	1.00	0.986 ($\pm$0.007)
Nonlinear CKA	0.008 ($\pm$0.000)	1.00	0.157 ($\pm$0.025)
DCCA	0.19 ($\pm$0.002)	1.00	0.999 ($\pm$0.003)

4.2 Two linearly correlated spaces with different dimensions

Next, we examine the performance of the PSRS approach on two linearly connected spaces with different dimensionalities. To generate synthetic data, we first randomly create a space, ES-A, identical in shape to the previous example. We apply PCA to transform ES-A and truncate it at various dimensions to construct several ES-Bs. The retained variance in a chosen ES-B is the sum of retained variance in all kept PCA dimensions, and is known as the "explained variance".



(a) Different similarity scores vs. PCA dimensions. sPSRS closely matches theoretical explained variance. The corresponding DCCA and CKA values are also shown for comparison.



(b) Different similarity scores vs. Signal Power Ratio (SPR) under Gaussian noise. sPSRS(ES-B, ES-A*) closely follow the SPR, indicating the robustness of PSRS to noise. The corresponding DCCA and CKA values are also shown for comparison.

Figure 1: Behavior of sPSRS, CKA, and DCCA across two synthetic scenarios.

Since ES-B is simply a truncated linear transformation of ES-A, it should be fully reconstructable by ES-A, regardless of its dimensionality. However, ES-B can only partially reconstruct ES-A as some information is lost during dimensionality reduction. Thus, the RIC from ES-B to ES-A should be equal to the amount of theoretical explained variance.

PSRS analysis aligns well with these theoretical expectations. In this example, for all ES-Bs with different dimensionality, PSRS(ES-A $\rightarrow$ ES-B) consistently returns a value of 1.00 ($\pm$0.00), indicating perfect reconstruction. Since CKA and DCCA are symmetric and have not been designed for directional information flow analysis, this part of the evaluation is not possible with them, except for the case where we kept all PCA dimensions. In that case, CKA and DCCA also return a value of 1.00 ($\pm$0.00), indicating perfect reconstruction.

Additionally, PSRS(ES-B $\rightarrow$ ES-A), which is the same as sPSRS in this case, closely matches the theoretical PCA explained variance for all tested dimensionalities of ES-B. Figure 1 (a) compares the sPSRS score with the PCA explained variance across different dimensions of ES-B, demonstrating a strong agreement between the two measures. The figure also includes the performance of CKA and DCCA in this setting. CKA exhibits a similar exponential growth pattern to PSRS, where the similarity between ES-B and ES-A increases rapidly with the number of PCA dimensions. DCCA

however, behaves differently and only reflects the number of PCA kept dimensions rather than information content (see the appendix for more details).

These results validate PSRS as an accurate tool for capturing linear relationships in embedding spaces.

4.3 Two nonlinearly correlated spaces

In order to fully assess the capabilities and generalizability of the PSRS approach, it is essential to evaluate its performance on spaces with nonlinear relationships. To create this scenario, once again, we begin with a randomly generated dataset, similar to the earlier examples. We assume this dataset to represent points in a 100-dimensional Cartesian coordinate system. We then transform the dataset into hyperspherical coordinates. This transformation is inherently nonlinear, but crucially, it does not result in any information loss. In other words, either dataset (Cartesian or hyperspherical) should be fully reconstructible from the other.

We apply a standard generalized hyperspherical coordinate transformation (detailed in Appendix A.5) which is inherently nonlinear and fully preserves information. This transformation, except for some edge cases, is unique and therefore the amount of information lost is minimal, and PSRS should return a near-perfect score. Here, we refer to the Cartesian coordinate space as ES-A and the hyperspherical coordinate space as ES-B. As shown in the Table 1, the outcome of the analysis shows PSRS(ES-A $\rightarrow$ ES-B) = 0.992 (± 0.003) and PSRS(ES-B $\rightarrow$ ES-A) = 0.986 (± 0.007), indicating its ability to retrieve highly nonlinear relationships. The few percentage points mismatch with the ground truth could be attributed to the aforementioned edge cases that makes the transformation deviate from a one-to-one mapping slightly. DCCS is also very effective in this case, yielding a score of 0.999 (± 0.005) indicating a perfect correlation between two spaces, while CKA performs poorly, with a score of 0.157 (± 0.025), despite its nonlinear kernel. The poor performance of CKA could be due to the choice of kernel, but again in real world applications, we usually do not know the nature of the relationship between two spaces to change the kernel accordingly.

4.4 Robustness to noise and extraneous information

Robustness to noise and irrelevant information is a critical requirement for any embedding similarity approach. To evaluate this property for PSRS, we examine how it responds when controlled noise and extraneous information are introduced. To that end, we will use the Cartesian (ES-A) and its corresponding hyperspherical (ES-B) coordinate datasets from the previous example, as they already represent embedding spaces with a nonlinear relationship.

First, we evaluate the impact of extraneous information. We extend ES-A by concatenating it with an additional set of randomly generated vectors with the same dimensionality, creating ES-A+. ES-B has no access to the appended portion of ES-A+, and the PSRS analysis should be able to reflect this fact.

As expected, PSRS(ES-A+ $\rightarrow$ ES-B) returns a value of 0.99 (± 0.003), confirming that ES-A+ still contains all the relevant information from ES-B. However, PSRS(ES-B $\rightarrow$ ES-A+) drops to 0.502, correctly indicating that ES-B cannot reconstruct the extraneous information in ES-A+. These results demonstrate that PSRS can robustly distinguish between relevant and extraneous information. Because CKA and DCCA cannot perform directional analysis, their results in this setting are limited to symmetric similarity values. Applied here, CKA and DCCA return the similarity score of 0.178 and 0.999 between ES-B and ES-A+ respectively. Evidently, DCCA fails to reflect that ES-A+ contains substantially more information than ES-A, while the CKA score is very close to its score between ES-B and ES-A (0.157), highlighting its insensitivity to extraneous information.

Next, we evaluate PSRS's response to noise. We systematically add dimension-wise Gaussian noise to ES-A, scaled by an increasing noise factor, to create several noisy datasets ES-A* (see Appendix A.4 for formulation). Then, we evaluate how well PSRS captures the relative information content in noisy settings. To this end, we compare PSRS(ES-B $\rightarrow$ ES-A*) to the the Signal Power Ratio (SPR) of ES-A*, which we define as:

$$\text{Signal Power Ratio (SPR)} = \frac{\text{Variance(ES-A)}}{\text{Variance(ES-A*)}}. \tag{3}$$

A robust method for evaluating RIC should be able to quantify the SPR in ES-A*, as the variance ratio directly reflects the proportion of the original signal preserved despite the added noise. Figure 1 (b) illustrates the relationship between sPSRS(ES-B, ES-A*), DCCA, and CKA and SPR across different noise levels.

As expected, sPSRS closely follows SPR, confirming its ability to quantify the proportion of signal retained within a noisy dataset, even in high-noise scenarios. CKA, which already exhibited poor performance in capturing the nonlinear relationship between ES-A and ES-B in the absence of noise, continues to perform poorly as noise increases. In contrast, DCCA follows the correct qualitative trend, with its similarity score decreasing as the noise level increases, although, due to its correlation-based nature, the absolute values may not directly correspond to those of sPSRS.

The synthetic experiments above show that PSRS is highly effective in accurately capturing the RIC between embedding spaces, in different scenarios, including linear and nonlinear relationships, unequal dimensionalities, and presence of noise and extraneous information. Comparing it with other state-of-the-art methods, such as DCCA and CKA, we showed that PSRS consistently outperforms them in terms of accuracy and robustness.

4.5 A REAL WORLD CASE STUDY

Next, we apply PSRS to investigate information retention across prefixes of different dimensionalities in embeddings learned using Matryoshka Representation Learning (MRL) (Kusupati et al., 2022). Rather than optimizing only the full-dimensional representation, MRL jointly trains nested lower-dimensional prefixes to remain useful for downstream tasks. This allows a single embedding to be shortened through truncation, providing a trade-off between representation size and task performance. However, strong performance on selected tasks does not necessarily establish how much information from the full representation is stored in prefixes of different lengths. This makes MRL a natural setting for PSRS: we complement existing task-specific evaluations (OpenAI, 2024) with an intrinsic, reconstruction-based assessment of information retention. Specifically, we use directional PSRS to quantify RIC and symmetric PSRS to measure representational similarity between each prefix and the full-dimensional representation, under our variance-based definition of information.

For this analysis, we use the publicly available `dbpedia-openai-3-large-1M` dataset (Thakur et al., 2021a), which contains one million 3,072-dimensional embeddings generated using OpenAI's `text-embedding-3-large` model. The dataset was constructed by embedding the concatenated title and text of each of the first one million entries in the DBpedia-Entity corpus (Hasibi et al., 2017), distributed through the BEIR benchmark (Thakur et al., 2021b). The model's native support for shortening embeddings through prefix truncation (OpenAI, 2024) provides a practical setting for examining how information retention varies with embedding dimensionality.

We randomly sampled 50,000 embeddings from the dataset and evaluated symmetric PSRS between the full 3,072-dimensional representation and its 256- and 1,024-dimensional prefixes. The 256-dimensional prefix achieved an sPSRS score of 0.465 with the full representation, while the 1,024-dimensional prefix achieved a substantially higher sPSRS score of 0.796.

These results show that, as expected, the lower-dimensional prefixes preserve a substantial amount of the information present in the full embedding. Almost half of the encoded information is concentrated in the first 256 dimensions, corresponding to only about 8.3% of the full embedding size. However, the reduction in representational information is larger than what might be suggested by downstream task performance (OpenAI, 2024). This highlights an important distinction between extrinsic measures, such as downstream task performance, and intrinsic approaches, such as PSRS.

5 CONCLUSION

In this work, we introduced Procrustes Similarity with the Reconstructed Space (PSRS) as a novel approach for quantifying the Relative Information Content (RIC) between high-dimensional embedding spaces. By leveraging deep learning for space reconstruction and Procrustes similarity for evaluation, PSRS effectively captures both linear and nonlinear relationships while demonstrating robustness to noise and extraneous information. Through synthetic scenarios and real-world case studies, we have shown that PSRS provides a reliable and interpretable measure of RIC, offering a more accurate

and reliable alternative to existing statistical and information-theoretic methods, such as DCCA and CKA.

Beyond embedding space comparisons, PSRS has the potential for broader applications across various domains. In neural network interpretability, PSRS could be used to assess how different layers or architectures encode information, shedding light on feature transformations in deep learning models. In tasks such as text summarization or image captioning generated by multimodal models, PSRS could help evaluate model performance by quantifying how well the output preserves the information present in the input data.

In other domains like comparative genomics and neuroscience, where different datasets or modalities represent the same underlying biological structures, PSRS could reveal nonlinear structural similarities or disparities that traditional methods might overlook. Additionally, in dimensionality reduction and representation learning, PSRS provides a robust means to evaluate how well different techniques preserve meaningful information when reducing the complexity of high-dimensional data.

Furthermore, PSRS provides a valuable tool for explainability, transparency, and bias detection in machine learning. As models become increasingly complex, it is crucial to understand how they encode and transform information. PSRS can help identify whether different representations of the same or similar data preserve biased information that contribute to unfair or misleading predictions, helping to create fairer and more interpretable AI systems. Additionally, integrating PSRS with techniques in causal inference could allow for deeper insights into which features drive predictive accuracy and whether certain representations introduce spurious correlations that affect decision-making in critical applications, such as medical diagnosis.

Overall, PSRS presents a versatile and principled framework for embedding space comparisons, offering insights that are crucial for scientific discovery, machine learning model evaluation, and ethical AI development.

6 AI Use Disclosure

Generative AI tools were used in this work for research ideation and execution, including providing feedback on research methodology and assisting with the implementation of research code. Generative AI tools were also used to aid and polish the writing of the manuscript, including copyediting for clarity and improving readability. All AI-assisted code and manuscript content were reviewed and verified by the authors, who take full responsibility for the final content of this work.

References

Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Angie Boggust, Brandon Carter, and Arvind Satyanarayan. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. arxiv (2019). *arXiv preprint arXiv:1912.04853*, 2019.

Fred L Bookstein. *Morphometric tools for landmark data*. Cambridge Univeristy Press, 1997.

James Chapman and Hao-Ting Wang. Cca-zoo: A collection of regularized, deep learning based, kernel, and probabilistic cca methods in a scikit-learn style framework. *Journal of Open Source Software*, 6(68):3823, 2021.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

Thyge Enggaard, August Lohse, Morten Axel Pedersen, and Sune Lehmann. Analyzing differences between discursive communities using dialectograms. *Scientific Reports*, 14(1):22497, 2024.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugenics*, 7: 179–188, 1936.

John C Gower and Garmt B Dijksterhuis. *Procrustes problems*, volume 30. OUP Oxford, 2004.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 1265–1268, 2017.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. Resi: A comprehensive benchmark for representational similarity measures. *arXiv preprint arXiv:2408.00531*, 2024.

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.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022.

OpenAI. New embedding models and API updates. OpenAI blog, January 2024. URL `https://openai.com/index/new-embedding-models-and-api-updates/`. Published January 25, 2024. Accessed September 25, 2026.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Jason Phang, Haokun Liu, and Samuel R Bowman. Fine-tuned transformers show clusters of similar representations across layers. *arXiv preprint arXiv:2109.08406*, 2021.

Aneesha Balachandran Pillay, Dharini Pathmanathan, Sophie Dabo-Niang, Arpah Abu, and Hasmahzaiti Omar. Functional data geometric morphometrics with machine learning for craniodental shape classification in shrews. *Scientific Reports*, 14(1):15579, 2024.

Dean A Pospisil, Brett W Larsen, Sarah E Harvey, and Alex H Williams. Estimating shape distances on neural representations with limited samples. *arXiv preprint arXiv:2310.05742*, 2023.

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.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Jay Roxis. Cka-similarity: A numpy and pytorch implementation of cka-similarity with cuda support. `https://github.com/jayroxis/CKA-similarity`, 2025. Accessed: 2025-05-15.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Venkatesh Sivaraman, Yiwei Wu, and Adam Perer. Emblaze: Illuminating machine learning representations through interactive comparison of embedding spaces. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 418–432, 2022.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021a. URL `https://openreview.net/forum?id=wCu6T5xFjeJ`.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021b.

Chuanjun Tong, Cirong Liu, Kaiwei Zhang, Binshi Bo, Ying Xia, Hao Yang, Yanqiu Feng, and Zhifeng Liang. Multimodal analysis demonstrating the shaping of functional gradients in the marmoset brain. *Nature Communications*, 13(1):6584, 2022.

Viivi Uurtio, João M Monteiro, Jaz Kandola, John Shawe-Taylor, Delmiro Fernandez-Reyes, and Juho Rousu. A tutorial on canonical correlation methods. *ACM Computing Surveys (CSUR)*, 50 (6):1–33, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. Dimensionality reduction: A comparative review. *Journal of machine learning research*, 10(66-71):13, 2009.

Tinghua Wang, Xiaolu Dai, and Yuze Liu. Learning with hilbert-schmidt independence criterion: A review and new perspectives. *Knowledge-Based Systems*, 234:107567, 2021. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2021.107567. URL https://www.sciencedirect.com/science/article/pii/S0950705121008297.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3441–3450, 2015. URL https://api.semanticscholar.org/CorpusID:14932020.

Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2019.

Zhiyi Zhang. Generalized mutual information. *Stats*, 3(2):158–165, 2020.

A Supporting Information

In this section we will provide more details about concepts, implementation details, and hyperparameters we used for different models. All the computations in this paper can be easily run on CPUs on a personal computer.

A.1 Data and code availability

This study only uses synthetic datasets generated during the analysis and two publicly available datasets (MNIST and Fashion-MNIST), which can be freely accessed through sklearn datasets.

A.2 Limitations

While PSRS provides a robust measure of relative information content, it does introduce a higher computational cost compared to other statistical methods like CCA or CKA, as it requires training a neural network for space reconstruction. However, with the continuous increase in available computing power, this overhead is unlikely to be prohibitive for most modern applications.

Additionally, because PSRS relies on deep learning and requires a train-test split to prevent overfitting, it demands sufficient data to train effectively specially for high dimensional embedding spaces. While all statistical methods struggle with small sample sizes, deep learning approaches can be particularly sensitive to this, making the dataset size an important consideration when applying PSRS to smaller datasets.

A.3 Definition of Information

Information has been defined in various ways across different disciplines. In information theory, it is often quantified by entropy, which captures the unpredictability or uncertainty in a dataset (Shannon, 1948). In statistics, information is sometimes defined using Fisher information, which

quantifies the amount of information that an observable random variable provides about an unknown parameter (Fisher, 1922). Both of these definitions are closely related to the concept of variance. This relationship is even more evident in methods like the Coefficient of Determination (R^2) and PCA, where variance is used more directly as a proxy for information. R^2, commonly used in regression analysis, measures the proportion of variance in the dependent variable (i.e. the target) that is predictable from the independent variables (i.e. model input). A higher R^2 value indicates that more variance and hence more information about the dependent variable is captured by the regression model. Similarly in PCA, data is transformed into a new coordinate system where the axes (principal components) are chosen to align with the directions of maximum variance. Each principal component is associated with an eigenvalue that represents the amount of variance in the data captured by that component. The sum of the explained variances across the selected components represents the total information content retained after dimensionality reduction.

A.4 Synthetic Noise Generation

To evaluate PSRS's response to noise, we systematically add Gaussian noise to every element in ES-A, gradually increasing its magnitude. We generate a random perturbation scaled to a specific ratio of the signal power in each dimension. Namely, the noise power for each dimension is defined as:

$$\sigma_{\text{noise}} = \sigma_{\text{signal}} \times \alpha, \tag{4}$$

where σ is the standard deviation, and α is the noise factor that controls the noise intensity. The noise for a specific dimension is then sampled from a normal distribution with zero mean and a standard deviation of σ_{noise}. Once generated, the noise is added to ES-A element-wise, creating a perturbed dataset, ES-A*.

A.5 Hyperspherical Transformation

There are different methods for defining a hyperspherical coordinate system, and all involve complex nonlinear transformations from Cartesian coordinates. The specific transformation we used follows:

$$\begin{aligned}
r &= \sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_2^2 + x_1^2}, \\
\varphi_1 &= \operatorname{atan2}\left(\sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_2^2}, x_1\right), \\
\varphi_2 &= \operatorname{atan2}\left(\sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_3^2}, x_2\right), \\
&\vdots \\
\varphi_{n-2} &= \operatorname{atan2}\left(\sqrt{x_n^2 + x_{n-1}^2}, x_{n-2}\right), \\
\varphi_{n-1} &= \operatorname{atan2}(x_n, x_{n-1}).
\end{aligned} \tag{5}$$

where r is the Euclidean norm or the radial distance from the origin, x_i is the i-th Cartesian coordinate of a given point, and φ_is are the generalized spherical coordinates, which define the angular position of the point in an n-dimensional hyperspherical system. Each angle is recursively computed using the two-argument arctangent function $(\operatorname{atan2})$, which determines the orientation of the projection of the remaining Cartesian components onto a subspace of decreasing dimensionality. The final angle, φ_{n-1}, is computed using the last two coordinates x_n and x_{n-1}, akin to the azimuthal angle in conventional three-dimensional spherical coordinates.

A.6 DeepShallowNet Architecture and Robustness

The DeepShallowNet architecture consists of a deep multi-layer feedforward branch and a shallow single linear layer branch, whose outputs are summed up to construct the final representation. This architecture effectively captures both linear and nonlinear relationships between two embedding spaces.

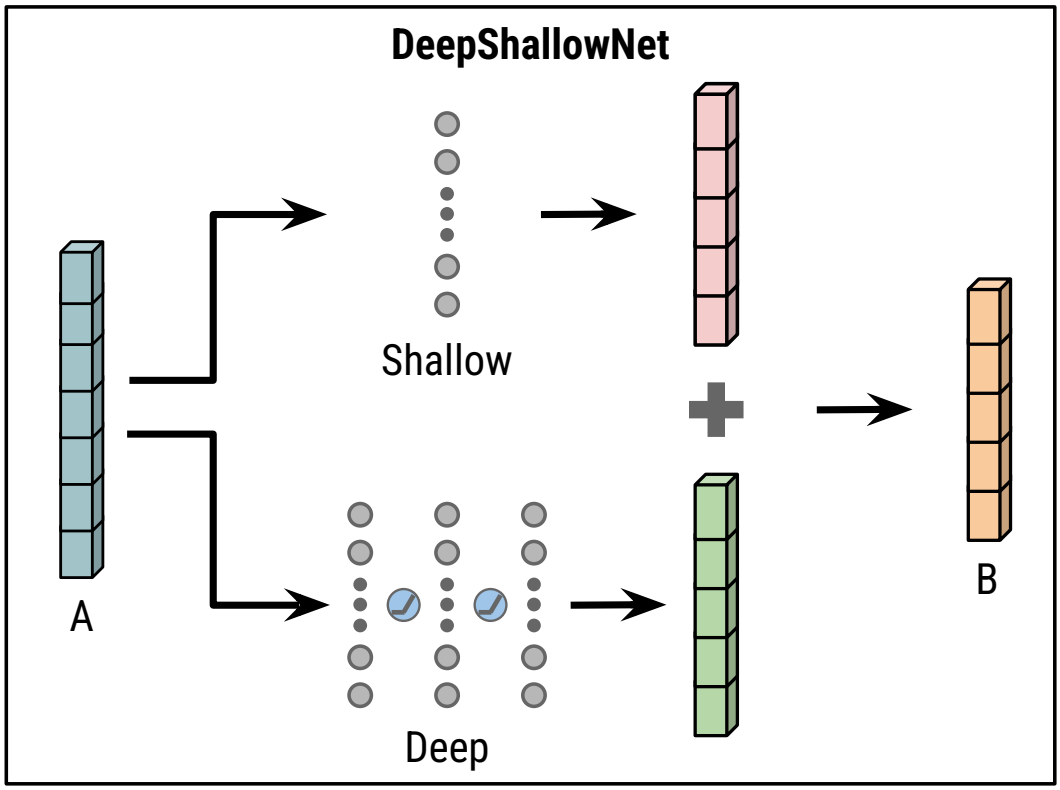


Figure 2: DeepShallowNet architecture, which consists of deep and shallow network components designed to capture nonlinear and linear relationships between two embedding spaces.

For a method to be trustworthy in practice, it is essential that its performance be rather robust to hyperparameter changes. For example, even in simple synthetic examples we discussed, one could achieve widely different results by altering hyperparameters, such as RBF Bandwidth in CKA, or latent dimension in DCCA. In most real-world applications, where ground truth is unknown, it is often impossible to tune hyperparameters to achieve optimal performance. This makes robustness to such choices a critical requirement.

To evaluate the robustness of PSRS, we systematically vary the hyperparameters of DeepShallowNet, and rerun the analysis on the most complicated dataset: PSRS(Cartesian $\rightarrow$ hyperspherical).

First, we varied the learning rate from 0.002 to 0.01 in increments of 0.002. Across this range, the model achieved a mean PSRS value of 0.983 with a standard deviation of 0.005, indicating a high degree of stability.

Next, we tested the sensitivity to architectural changes by modifying the number and size of hidden layers in the deep branch of DeepShallowNet. For equi-dimensional spaces of 100 dimensions, the default architecture has three hidden layers with the widths of $[200, 100, 100]$. We additionally evaluated alternative configurations including $[100, 50]$, $[20, 20, 20]$, and $[100]$. Across all these architectures, the model remained highly stable, yielding a mean PSRS value of 0.982 with a standard deviation of 0.003.

Finally, we examined convergence behavior and potential overfitting by training the model for up to 500 epochs. We found that PSRS values on the test set saturated around 100 epochs, and no evidence of overfitting was observed beyond that point.

These results demonstrate that DeepShallowNet is a stable architecture that performs robustly across a wide range of hyperparameter settings.

A.7 Baseline Selection: MINE Exclusion

While we initially considered including Mutual Information Neural Estimation (MINE) as an information-theoretic baseline, we ultimately excluded it for two primary reasons. First, MINE estimates mutual information in unbounded form and does not yield a normalized similarity measure within a fixed range, making it less suitable for direct comparisons even in theoretical limits. Second, MINE's output is highly sensitive to training hyperparameters such as number of optimization iterations. These characteristics made it a less stable reference point in the context of this evaluation.

A.8 DCCA

For Deep Canonical Correlation Analysis (DCCA), we used the implementation from the CCA-Zoo (Chapman & Wang, 2021) library (version 1.15.2), which provides a modular framework for CCA and its variants. DCCA learns nonlinear transformations of two input spaces into a latent space that maximizes their cononical correlations. For this study, we used the following settings:

- Latent dimension: Equal to the smaller of the two input embedding dimensions. - Optimizer: Adam with a learning rate of 0.001. - Batch size: 1000 - Epochs: 100 - Deep neural network architecture: The deep neural network has 3 layers of [feature size (100 in synthetic examples), 128, and latent dimension. The layers are activated by leaky ReLU.

Note that DCCA similarity scores are highly sensitive to the choice of latent dimensionality. A clear example is provided in subsection 4.2, where embedding spaces of unequal dimensions are compared. It is standard practice to set the intermediate projection dimension in DCCA to match the smaller of the two spaces. However, in this particular case—-where one space is a truncated transformation of the other-—the linear growth observed in the DCCA similarity curve reflects the relative size of the intermediate space with respect to ES-B, rather than the actual similarity between the embeddings.

As evidence, when comparing ES-A with a 10-dimensional ES-B using a DCCA intermediate dimension of 10, the resulting similarity score is approximately 0.1. Yet, increasing the intermediate dimension to 100 while keeping ES-B fixed, produces a similarity score close to 1, despite no change in the true relationship between the spaces. This example illustrates how DCCA's sensitivity to architectural choices can lead to misleading results and limit its trustworthiness as a reliable similarity measure.

A.9 CKA

For Centered Kernel Alignment (CKA), we used the implementation by Jay Roxis (Roxis, 2025). Here are the hyperparameters we used in CKA calculations.

- Kernel: Radial Basis Function (RBF) kernel which is a nonlinear kernel. - RBF Bandwidth (σ): Median of the pairwise distances between examples. - Centering: Gram matrices were centered using the standard double-centering formula.

A.10 PSRS

For the PSRS part the hyperparameters are as follows

- Reconstruction network: For this part, we used DeepShallowNet as described in the paper. The deep branch consists of a three-layer fully connected architecture. The hidden layer dimensions are dynamically set based on the dimensionalities of the source and target embedding spaces: The first hidden layer has a width of 2 x the input space dimension, the second hidden layer is sized to the maximum of the input and target embedding dimension, followed by two layers with the target embedding dimension. Each layer is followed by a ReLU activation, and no dropout or batch normalization was applied unless otherwise noted. - Loss function: Mean Squared Error (MSE). - Optimizer: Adam with learning rate of 0.005. - Number of epochs: 100 - Train-test split: 70/30, where reconstruction is learned on the training set and evaluated on the unseen test set to prevent overfitting.