MIRAGES OF MISALIGNMENT: HOW SUPERPOSITION DISTORTS NEURAL REPRESENTATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks trained on the same tasks achieve similar performance but often show surprisingly low representational alignment. We argue this is a measurement artifact—a *mirage of misalignment*—caused by superposition, where individual neurons represent mixtures of features. Consequently, two networks representing identical feature sets can appear dissimilar if their neurons mix those features differently. To formalize this intuition, we derive an analytic theory that predicts this apparent misalignment for common linear metrics like representational similarity analysis and linear regression. We validate our theory in settings of increasing complexity. It perfectly predicts misalignment between random projections of identical features. On real data, we use sparse autoencoders to find underlying disentangled features, showing their latent codes are often far more aligned than the raw neural representations. This work reveals that linear alignment metrics, when applied to raw neural activations, can be systematically misleading due to superposition. Our findings suggest that neural networks are more aligned than previously believed and that the common practice of comparing raw neural activations with linear probing may systematically underestimate model similarity.

1 INTRODUCTION

The development of deep neural networks capable of human-level performance on tasks such as object recognition and natural language has prompted a fundamental question: do different neural systems learn to represent the same information? (**?**Goldstein et al., 2022; Peterson et al., 2018; Sucholutsky et al., 2023; Huh et al., 2024; Reizinger et al., 2024). Answering this requires comparing representations across models with varied architectures, training data, and objectives, a challenge central to ideas like the platonic representation hypothesis (Huh et al., 2024; Reizinger et al., 2025). To measure these similarities, researchers turn to alignment metrics such as Representational Similarity Analysis (RSA) (Kriegeskorte & Wei, 2021) which abstract away from individual neurons to compare the geometry of population-level activity. Alternatively, Linear Regression is also used which learns a linear map to predict one network's activity from another. Both metrics have become powerful alignment tools, yielding remarkable insights into shared structure (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cadena et al., 2019; Khosla et al., 2021; Schrimpf et al., 2021; Conwell et al., 2024; Prince et al., 2024). However, even when models are trained on identical tasks and data, comparisons consistently reveal a persistent "alignment ceiling," (Schrimpf et al., 2018; Ahlert et al., 2024; Chen & Bonner, 2025), suggesting that these systems do not converge to identical representational solutions.

We propose this alignment ceiling is caused by the phenomenon of *superposition*, where neural networks *linearly* represent more features than they have neurons (Smolensky, 1990; Elhage et al., 2022; Klindt et al., 2025). Compressed sensing theory states that this is a viable strategy if the features that the neurons represent are sparse. In that case, sparse dictionary learning is able to recover the features from the representation in superposition (Donoho, 2006; Candes et al., 2006). However, this highly efficient compression strategy comes at a cost for direct comparisons: *to extract the representations, we need a nonlinear decoding mechanism* (O'Neill et al., 2024).

Consequently, two networks could learn the *exact same* set of underlying features, but if they represent them in different arrangements, then linear comparison metrics like RSA and Linear Regression will erroneously measure their representations as dissimilar. Geometric distortions (RSA) and

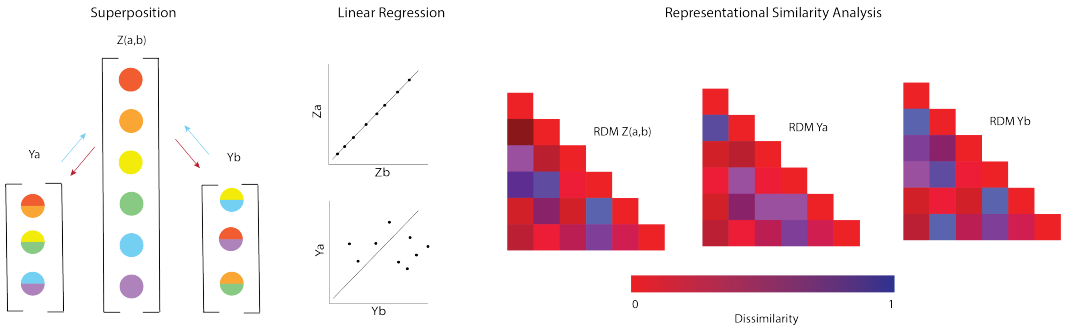

Figure 1: **Illustration of Core Idea. Superposition:** Two neural networks share an identical set of latent features ($Z_a = Z_b$), but compress them (red arrows) in different ways $Y_a \neq Y_b$. Thus, computing alignment over the raw neural activations of network A (Y_a) and B (Y_b) will underestimate the representational similarity of these networks. We propose using sparse dictionary learning to recover (blue arrows) the shared features of networks from their raw activations prior to using alignment metrics (Donoho, 2006). **Linear regression:** Assuming perfect latent recovery, the maximum pairwise correlation between latent activations is 1.0, and will be greater than the correlation between raw neural activations. **Representational similarity analysis:** Rather than directly correlating neural (or latent) activation, RSA first computes pairwise (dis)similarity matrices of neural responses to features. Depicted are representational similarity matrices (or their dissimilarity counterparts), which are correlated to produce an alignment score. As with linear regression, the RSA score for perfectly recovered latents is 1.0, and greater than the RSA score over neural activations.

predictive performance (Linear Regression) occur not because the underlying representations are different, but because the superposition arrangement, i.e., projection from features down to neurons in each system is unique. This could explain why models with higher dimensionality, i.e., presumably less superposition and closer to linear decodability, seem to lead to better linear regression (Elmoznino & Bonner, 2024).

One approximate way of lifting features out of superposition is using sparse autoencoders (SAEs) (Ng et al., 2011; Cunningham et al., 2023; Rao et al., 2024; Lan et al., 2024). This is a form of sparse dictionary learning (Olshausen & Field, 1997), where the *sparse inference* (given a dictionary) is amortized using a perceptron (i.e., linear-relu encoder) (O'Neill et al., 2024; **?**). A common implementation of SAEs finds an overcomplete (i.e., higher-dimensional) basis for representing neural activations, allowing for observation of more disentangled, interpretable features than the original neural basis (Bricken et al., 2023). This should partially relieve low alignment due to the networks' idiosyncratic feature mixtures.

In this work, we formalize this intuition and investigate its consequences for representational alignment. Specifically, we develop an analytic theory that precisely quantifies how idiosyncratic feature mixtures in superposition lead to misalignment under RSA and Linear Regression. We validate this theory in a series of simulation studies. On real neural activity from brains and models, we use SAEs to learn a disentangled feature basis; showing that RSA and Linear Regression performed with these recovered features, rather than the raw activations, often reveal significantly higher alignment across systems. Taken together, our results highlight a critical limitation of standard similarity metrics and demonstrate that accounting for superposition is a necessary step toward a more complete understanding of representational alignment.

2 Related Work

Comparing Neural Representations. A central goal in neuroscience and machine learning is to compare learned representations across different systems, be they biological or artificial (Goldstein et al., 2022; Peterson et al., 2018; Sucholutsky et al., 2023). A primary tool for this is Representational Similarity Analysis (RSA), which abstracts away from the activity of individual neurons to compare the geometric structure of population-level responses (Kriegeskorte & Wei, 2021). Another common approach is linear regression, which assesses alignment by training a linear map to predict

the activations of one system from another. These methods have been instrumental in revealing shared representational structure between brains and models (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cadena et al., 2019), across different models (Khosla et al., 2021; Schrimpf et al., 2021; Conwell et al., 2024), and under varying training objectives (Prince et al., 2024). However, despite their power, these linear methods consistently encounter an "alignment ceiling," where even models trained on identical tasks fail to converge to perfectly aligned solutions (Schrimpf et al., 2018; Ahlert et al., 2024; Chen & Bonner, 2025). This limitation has spurred the development of more sophisticated techniques, such as generalized shape metrics, to capture more complex representational transformations (Williams et al., 2021). Our work posits that this ceiling is not necessarily a failure of the models to learn similar features, but a failure of linear metrics to account for non-linear encoding schemes like superposition.

Superposition in Neural Networks. The concept of *superposition* was formally introduced by Smolensky (1990) as the principle that networks represent concepts through linear combination. This idea has strong theoretical grounding in identifiability theory, which explores the conditions under which underlying latent variables can be recovered from their mixtures, with theoretical guarantees now extending from classic approaches (Hyvarinen & Morioka, 2016; Hyvärinen et al., 2023; Park et al., 2023; Arora et al., 2016; 2018) to modern supervised (Reizinger et al., 2024) and self-supervised paradigms (Zimmermann et al., 2021). However, the universality of linear feature encoding has been debated (Pfau et al., 2020; Higgins et al., 2018; Bouchacourt et al., 2021; Engels et al., 2024). More recently, the term has been used to describe the specific case where a network linearly represents more features than it has available neurons (Elhage et al., 2022), reviewed in Klindt et al. (2025). When features are sparse, representing them in superposition becomes a standard compressed sensing problem, where a small number of features can be reliably recovered from a compressed representation (Donoho, 2006; Candes et al., 2006). This insight has fueled significant progress in AI interpretability, enabling the use of tools like sparse autoencoders to uncover meaningful, disentangled features from within modern architectures like Transformers (Yun et al., 2021; Bricken et al., 2023; Templeton et al., 2024) and other large foundation models (Simon & Zou, 2024). Beyond artificial networks, these principles have deep roots in theoretical neuroscience. Foundational work demonstrated that sparse coding applied to natural scenes yields features resembling V1 receptive fields (Olshausen & Field, 1996), while superposition has been explored as a coding strategy in biological circuits (Fyshe et al., 2014; Klindt et al., 2023) and as an explanation for mixed selectivity, where single neurons respond to conjunctions of features (Rigotti et al., 2013).

3 THEORY

Let $z \in \mathbb{R}^n$ be *latent variables* and $y \in \mathbb{R}^m$ be neural *representations*, which are functions of these latent variables, i.e., $y = f(z)$.

Definition 3.1 (**Superposition**). *We say that a representation $f : \mathbb{R}^n \to \mathbb{R}^m$ is in superposition if it is a linear map and a low-dimensional projection, i.e., $m < n$.*

3.1 ASSUMPTIONS

Throughout our analysis, we make the following assumptions:

1. **Linearity:** The neural representations are in superposition and are thus linear, described by a matrix $A \in \mathbb{R}^{m \times n}$:
$$y = Az \tag{1}$$
The condition $m < n$ implies that the columns of A are not all orthogonal, aligning with the common assumption of having fewer neurons than latent variables.
2. **Sparsity of Latent Variables:** The latent variables are sparse, e.g., $\|z\|_0 \leq K$ for some $K \ll n$.
3. **Restricted Isometry Property (RIP):** The matrix A satisfies the RIP, which allows for the theoretical possibility of recovering z from observations of y via compressed sensing.
4. **Distribution of Latent Variables:** For a dataset of d inputs, the latent vectors $z_1, \ldots, z_d$ are treated as independent and identically distributed (i.i.d.) random variables satisfying:

- Zero mean: $\mathbb{E}[z_i] = \mathbf{0}$ for all i.
- White distribution (Identity covariance): $\mathbb{E}[z_i z_i^\mathsf{T}] = I_n$ for all i.

If these assumptions do not fully hold, we incur an irreducible reconstruction error when retrieving the sparse codes. This error would lower the ceiling of RSA alignment, correctly reflecting that if two features cannot be separated in one system, it should count as a representational misalignment.

3.2 Representational Similarity Matrix (RSM)

For a dataset of neural responses $Y = (y_1, ..., y_d)$, the *representational similarity matrix* (RSM) is defined as:

$$M(Y)_{i,j} = \langle y_i, y_j \rangle \quad \forall i, j \in \{1, ..., d\}. \tag{2}$$

Given the linearity assumption equation 1, we can rewrite the RSM in terms of the latent variables:

$$M(Y)_{i,j} = \langle y_i, y_j \rangle = \langle Az_i, Az_j \rangle = z_i^\mathsf{T} A^\mathsf{T} A z_j \tag{3}$$

This shows that the similarity between latent variables z_i, z_j is measured by a semi-inner product $\langle \cdot, \cdot \rangle_G$ induced by the positive semi-definite Gram matrix $G := A^\mathsf{T} A$.

4 Alignment Under Superposition

Consider two neural representations in superposition, with matrices A_a, A_b, generating responses $Y_a = (A_a z_1, ..., A_a z_d)$ and $Y_b = (A_b z_1, ..., A_b z_d)$ to the same set of latent variables $Z = (z_1, ..., z_d)$. While the underlying latent variables are identical, the observed neural representations Y_a and Y_b may differ. We now analyze how standard alignment metrics behave in this scenario.

The key insight of our work is that while these two neural representations Y_a, Y_b originate from the same latent variables, any direct linear measure of alignment will be confounded by the differing projection matrices.

4.1 Representational Similarity Analysis (RSA)

The RSA metric is the Pearson correlation between the vectorized upper-triangular elements of two RSMs, $\vec{m}_a$ and $\vec{m}_b$.

$$\rho(Y_a, Y_b) = \frac{\text{Cov}(\vec{m}_a, \vec{m}_b)}{\sqrt{\text{Var}(\vec{m}_a)\text{Var}(\vec{m}_b)}} \tag{4}$$

Under the assumptions outlined previously, we arrive at the following result in the limit of large datasets.

Theorem 4.1 (Asymptotic RSA Alignment)**.** *The RSA correlation between two representations Y_a and Y_b in superposition is approximately the cosine similarity between their respective Gram matrices, $G_a = A_a^\mathsf{T} A_a$ and $G_b = A_b^\mathsf{T} A_b$.*

$$\boxed{\rho(Y_a, Y_b) \approx \frac{Tr(G_a G_b)}{\sqrt{Tr(G_a^2) Tr(G_b^2)}} = \frac{\langle G_a, G_b \rangle_F}{\|G_a\|_F \|G_b\|_F}} \tag{5}$$

where $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ are the Frobenius inner product and norm, respectively.

This result shows that RSA is fundamentally sensitive to the similarity of the metric tensors induced by the representations on the latent space.

4.2 Linear Regression

Alternatively, we can measure alignment by determining how well one representation can be linearly predicted from the other using a multivariate linear model $Y_b = W Y_a + E$. The Ordinary Least Squares (OLS) estimator $\hat{W}$ minimizes the squared Frobenius norm of the residuals, $\|Y_b - W Y_a\|_F^2$.

Theorem 4.2 (Asymptotic Linear Regression)**.** *In the asymptotic limit and under the stated assumptions, the OLS estimator $\hat{W}$ and the resulting model performance are given by:*

1. ***Optimal Weights:*** *The weight matrix $\hat{W}$ converges to:*

$$\boxed{\hat{W} \approx A_b A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1}} \tag{6}$$

2. ***Mean-Squared Error (MSE):***

$$\boxed{\mathit{MSE}(Y_b|Y_a) \approx \frac{1}{m_b} \left\| A_b - \hat{W} A_a \right\|_F^2} \tag{7}$$

3. ***Explained Variance (R^2):***

$$\boxed{R^2 = 1 - \frac{\mathit{Tr}\left((A_b - \hat{W}A_a)^\mathsf{T}(A_b - \hat{W}A_a)\right)}{\mathit{Tr}(A_b^\mathsf{T} A_b)}} \tag{8}$$

4. ***Pearson Correlation ($\rho(\hat{Y_b}, Y_b)_{ij}$):***

$$\boxed{\rho(\hat{Y_b}, Y_b)_{ij} = \frac{(\hat{W} A_a A_b^\mathsf{T})_{ij}}{\sqrt{(\hat{W} A_a A_b^\mathsf{T})_{ii} (A_b A_b^\mathsf{T})_{jj}}}} \tag{9}$$

5 Simulating Superposition's Impact on Alignment

5.1 Experimental Setup

In this section, we test our theoretical prediction that superposition is sufficient to reduce alignment in cases where two networks use an identical set of features. We generate a single feature set Z of $d \times N$ dimensional features (i.e. $Z \in \mathbb{R}^{d \times N}$), which are random uniform values between 0 and 1, i.e., $Z_{i,j} \sim \mathcal{U}(0, 1)$. To simulate the sparsity condition, we then zero mask all but the top K activating latent variables within each generated sample (i.e. individual row in Z). Next, we generate two projection matrices, each $N \times M$ dimensional, with elements drawn from a standard normal distribution, i.e., $A_0, A_1 \in \mathbb{R}^{N \times M}$ where $A_{i,j} \sim \mathcal{N}(0, 1)$. These matrices are used to produce two random linear projections of a shared set of features. We manipulate the degrees of superposition by varying M from $0.2K \log(N/K)$ to $50K \log(N/K)$. Next, we measured alignment of the random linear projections using RSA (Experiment 1) and Linear Regression (Experiment 2). To test the effect of sparsity, we repeated these experiments across different numbers of active latents (K). We also calculate and show the minimum dimensionality of M required for accurate latent recovery under compressed sensing as $M = K \log(N/K)$ (Candes et al., 2006).

5.2 Results

Consistent with our analytical predictions, we observe that RSA alignment decreases as superposition—i.e., compression—increases (Fig. 2, left). The relative rate of this decrease is similar across varying sparsity levels (K active components). Notably, this decay in alignment persists even below the critical compression threshold ($M = K \log(N/K)$), the regime where compressed sensing theory guarantees that feature recovery is, in principle, possible (Donoho, 2006).

Analogously, the variance explained by linear regression also declines with greater compression, exhibiting a similar relative drop across all sparsity levels (Fig. 2, right). Nevertheless, in regimes with low compression (high M) and low sparsity (high K), performance remains high, with explained variance exceeding 0.8. Together, these simulations validate our theory, confirming that superposition systematically decreases linear alignment measures like RSA and linear regression performance.

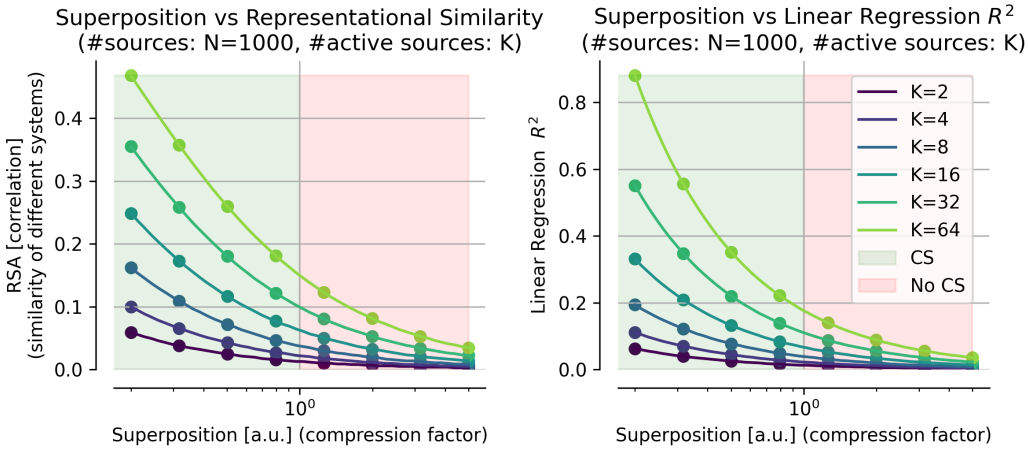

Figure 2: **Neural Network Alignment Decreases with Superposition.** Alignment measured with RSA (**Left**) as well as with Linear Regression (**Right**) as a function of compression (N/M). This experiment is repeated across multiple sparsity levels (K). Analytical predictions are represented by solid curves, while empirical results from simulation across different superposition compressions is represented by the dots. We note where accurate latent recovery from compressed representations is (CS; green shading) or is not (No CS; red shading) possible Donoho (2006).

6 Superposition's Impact on Alignment in Real Networks

6.1 Experimental Setup

We now test whether superposition disentanglement increases alignment in real neural networks. We measure model-model (Fig. 3), model-brain (Fig. 4), and brain-brain (Fig. 5) alignment using RSA and Linear Regression. To begin, we measure alignment on raw neural activations to obtain a baseline. Next, we train SAEs on models and brains to recover latent features and use them in place of neurons for computing alignment. For RSA, both networks have their neurons replaced with SAE latents for alignment, whereas with Linear Regression, only the source neurons are replaced with their SAE latents. This is done to keep the targets the same as in the base comparison (i.e., predicting neurons). It is technically sound because Linear Regression is capable of remixing the source latents back into the target's superposition arrangement. Finally, we report the difference between alignment over latent activations and alignment over raw neural activations to quantify the relative increase in alignment provided by disentangling features from superposition.

6.2 Data

We obtained neural activations from both biological and artificial neural networks. Biological data is from the publicly available Natural Scenes Dataset (NSD) (Allen et al., 2022), which uses fMRI to record human neural responses to subsets of the COCO natural images dataset (Lin et al., 2014). We analyzed three brain areas along the visual processing hierarchy: the ventral portion of the primary visual cortex (V1v), fusiform face area 1 (FFA-1) and the parahippocampal place area (PPA).

For model-model alignments, we obtained activations from the penultimate layers of ResNet50 (layer4.2) (He et al., 2016), ViT-B/16 (encoder.layers.encoder_layer_11) (Dosovitskiy et al., 2021), and CLIP-ViT-B/32 (visual.transformer.resblocks.11) (Radford et al., 2021), with ResNet50 and ViT-B/16 being trained on ImageNet classification (Deng et al., 2009). To be consistent with the brain alignment experiments, we use the 10,000 images for NSD Subject 1.

For model-brain alignments, we used preprocessed (the result of Step 5 described in (Allen et al., 2022)) neural activations from NSD Subject 1 in response to 10,000 unique images. Each neural

response was averaged over 3 image presentations and z-scored. The same images were used to obtain activations from ResNet-50 layers 1-4 (He et al., 2016).

For brain-brain alignments, we use Subjects 1 and 2 of the NSD. NSD subjects largely viewed non-overlapping images, so neural activations in response to 1,000 images viewed by both study participants were used. Given the limited data, we trained sparse autoencoders on each subjects' neural responses to the 10,000 images they viewed, before latent activations in response to the shared 1,000 images were extracted for further analysis.

6.3 SAE TRAINING

We train sparse autoencoders with an L1 sparsity penalty (L1-SAEs) to learn disentangled latent features (z). The SAE has an encoder and a decoder. Encoding is given by:

$$z = \text{ReLU}(W_{\text{enc}}x + b_{\text{enc}})$$

where x represents the raw neural activations, and learned parameters W_{enc} and b_{enc} are the encoder weights and bias respectively. Decoding is given by:

$$\hat{x} = W_{\text{dec}}z + b_{\text{dec}}$$

where $\hat{x}$ are reconstructed neural activations, and learned parameters W_{dec} and b_{dec} are the decoder weights and bias respectively. The model is trained using a combined loss function, which is the sum of a reconstruction loss

$$\mathcal{L}_{\text{reconstruction}} = \frac{1}{d \cdot M}\sum_{i=1}^{d}(x_i - \hat{x}_i)^2$$

and sparsity loss

$$\mathcal{L}_{\text{sparsity}} = \frac{\lambda}{d \cdot N}\sum_{i=1}^{d}\sum_{j=1}^{N}|(W_{\text{dec}})_{:,j}| \cdot |z_i^j|$$

which is the L1 norm of latent activations scaled by the decoder norm (to avoid collapse with vanishing latents and exploding decoder norms) and weighted by the hyperparameter λ. We varied λ from 10^{-3} to 20 and tested expansion factors of 1, 2 and 4 times the dimensionality of the base neural activations.

For the model-model analysis, we train SAEs on the three models using their layer activations to the ImageNet training set, with a batch size of 1024 for 300 epochs. Alignment is taken over each SAE architecture and we report the SAE with the highest mean alignment increase for each metric. For model-brain alignment analysis, to ensure fair comparison of the latents across model and brain, both the model and brain voxel's SAEs were trained on their respective activations towards the same set of 10,000 unique images that Subject 1 was exposed to. For analysis of brain-brain alignment between Subject 1 and Subject 2, Subject 1's and Subject 2's SAEs were trained on their respective voxel recordings (in response to their respective sets of images). For the alignment calculations, only the activations/sparse codes that correspond to the overlapping subset of 1000 images were used.

6.4 RESULTS

Model to Model. Alignment results between the three models are presented in Figure 3, where each number in the grid shows the increase in alignment obtained over the respective SAE latents relative to the raw neural activations. In all model pairs and across both RSA and Linear Regression, we see a positive value, reflecting SAE latents have at least partially recovered the true alignment obscured in the raw neurons. The SAE with the largest RSA increase has an expansion of 4 and $\lambda = 0.1$, and the largest Linear Regression increase has an expansion of 4 and $\lambda = 0.01$. We also plot the raw alignment values beside the grid, where each point is the Neuron-versus-Latent alignment for a given model pair, with all points lying above the diagonal. It is noteworthy that the relative alignment gain seems more pronounced in RSA versus Linear Regression, which may indicate how different metrics interact with superposition deflation and/or sparse matrices. We leave this inquiry for future work.

Model to Brain. Alignment results between model and brain are presented in Figure 4. Across most model-brain pairs, we observe an increase in alignment with both RSA and Linear Regression, when transitioning from neuron/voxel to SAE latent representation. In the case of alignment with Linear Regression, we find that the increase in alignment is much more noticeable when using the brain latents as the source mapping to model neurons, as opposed to using model latents as the source mapping to brain voxels.

Brain to Model. Alignment results between two different brains are presented in Figure 5. We find relatively stable or modest decreases in alignment with RSA transitioning from voxel-voxel to latent-latent across all pairs, while no significant differences in alignment with Linear Regression transitioning from voxel-voxel to latent-voxel, regardless of whether subject 1 or subject 2 is used as the source.

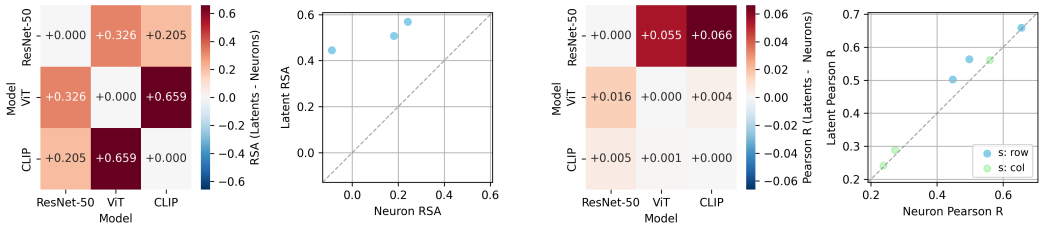

Figure 3: **Model-Model Comparison. Left Plots:** Increase in RSA over features compared to neurons shown as a difference (heatmap) and over the identity line (scatterplot). **Right Plots:** The same results for Linear Regression. The scatterplot legend indicates which set of comparisons served as a source for Linear Regression mapping.

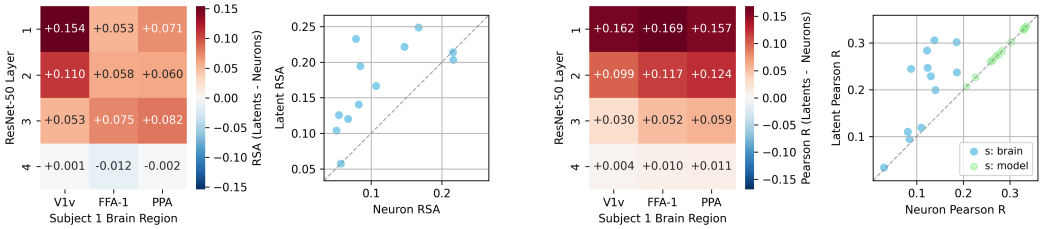

Figure 4: **Model-Brain Comparison. Left Plots:** Increase in RSA over features compared to neurons shown as a difference (heatmap) and over the identity line (scatterplot). **Right Plots:** The same results for Linear Regression. The scatterplot legend indicates which set of comparisons served as a source for Linear Regression mapping.

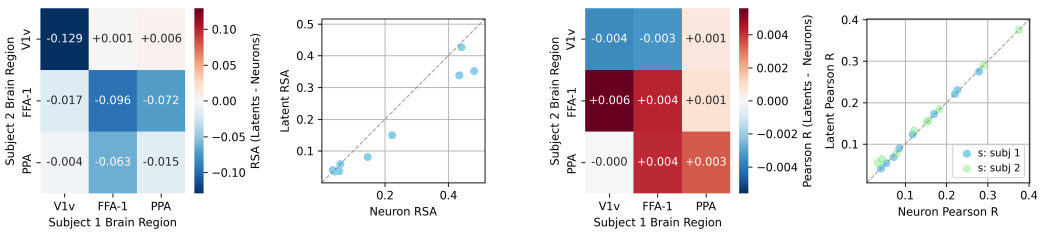

Figure 5: **Brain-Brain Comparison. Left Plots:** Increase in RSA over features compared to neurons shown as a difference (heatmap) and over the identity line (scatterplot). **Right Plots:** The same results for Linear Regression. The scatterplot legend indicates which set of comparisons served as a source for Linear Regression mapping.

7 LIMITATIONS

There are several limitations in our study. The first is our practical assumptions that 1) projections from the latent to neural basis are random and 2) that all features are shared. At certain scales and in certain areas, biological neural networks have a bias towards privileged, rather than random, projections (Khosla et al., 2024; Posani et al., 2025). The impact of this on alignment is likely complex (e.g., higher alignment over neurons if biases are shared, lower alignment if biases differ between systems) and worth further exploration. It is also unlikely that all of the real networks in our study represent the exact same feature set.

The second limitation stems from our use of SAEs, known to suffer various problems such as an amortization gap O'Neill et al. (2024), inconsistent latents across training seeds (Paulo & Belrose, 2025) and the sensitivity of discovered latents to dictionary dimensionality (Leask et al., 2025; Chanin et al., 2024). Further work could explore recent efforts to alleviate such problems (Fel et al., 2025), but we stress that our theory does not depend on SAEs. We pragmatically adopt SAEs as the current best method to disentangle features in superposition, and our experiments should be revisited if improved approaches are designed. Finally, we have only focused on the visual domain, but our hypothesis is modality-agnostic and should be further tested in other domains such as language.

8 DISCUSSION

In this work, we derive analytic predictions and contribute simulation experiments showing that representational alignment decreases as a function of distinct superposition arrangements of the same underlying features (i.e., compression via random projections). These experiments suggested that alignment computed over disentangled features would be higher. Based on this prediction, we used SAEs to extract approximations of the features in real neural networks, showing that alignment over the SAE latent activations is often significantly higher for the commonly used metrics of RSA and Linear Regression. These findings suggest that the representational similarities between models and across biological and artificial networks is greater than previously estimated. A notable exception is brain-brain alignment. Given the relatively low number of shared images in this comparison (1,000 shared NSD images), it is difficult to attribute this result to data and related processing limits, resolution limits (e.g., voxels instead of neurons), or to biological phenomena. Applying our methods to single- or multi-unit datasets and datasets with more shared stimuli is a critical next step.

Our findings have implications for model selection criteria. If superposition masks similarity between two systems that represent even identical features, then computing RSA or Linear Regression between models and brains over base activations of models with variable dimensionality puts small models at a systematic disadvantage. This may explain why scaling models often produces more reliable alignment boosts than producing models that have more apparent alignment with human perception (Schaeffer et al., 2022; 2024).

As we seek to understand whether models and brains represent the same information, it is important to consider the best uses of common representational alignment metrics. In this work, we demonstrate that alignment metrics systematically underestimate the similarity of neural networks because of the manner in which they arrange features across their neurons. We offer superposition disentanglement via SAEs as a practical and effective solution to address the alignment ceiling currently facing neural network comparisons with otherwise similar behavior.

REFERENCES

Jannis Ahlert, Thomas Klein, Felix Wichmann, and Robert Geirhos. How aligned are different alignment metrics? *arXiv preprint arXiv:2407.07530*, 2024.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Compu-*

tational Linguistics, 4:385–399, dec 2016. ISSN 2307-387X. doi: 10.1162/tacl_a_00106. URL https://direct.mit.edu/tacl/article/43373.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, dec 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00034. URL https://direct.mit.edu/tacl/article/43451.

Diane Bouchacourt, Mark Ibrahim, and Stéphane Deny. Addressing the topological defects of disentanglement via distributed operators. *arXiv preprint arXiv:2102.05623*, 2021.

Trenton Bricken, Andy Chen, and et al. Anthropic. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Zirui Chen and Michael F Bonner. Universal dimensions of visual representation. *Science Advances*, 11(27):eadw7697, 2025.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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 Computer Society, 2009.

David L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. URL https://ieeexplore.ieee.org/abstract/document/1614066/?casa_token=vtpGjU5mzFcAAAAA:rU2N5NCWY2K9IaaU0GHdJEuOj8P0dFk39KnF-rchFhrMrAe9T0XiWvCPGgJ5pszVR4-UWxvhvg. Publisher: IEEE.

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. In *International Conference on Learning Representations*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS computational biology*, 20(1):e101a1792, 2024.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9v1eW8HgMU.

Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning. In Kristina Toutanova and Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 489–499, Baltimore, Maryland, jun 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1046. URL https://aclanthology.org/P14-1046.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

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.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*, 2023.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, 2021.

Meenakshi Khosla, Alex H Williams, Josh McDermott, and Nancy Kanwisher. Privileged representational axes in biological and artificial neural networks. *bioRxiv*, pp. 2024–06, 2024.

David Klindt, Sophia Sanborn, Francisco Acosta, Frédéric Poitevin, and Nina Miolane. Identifying Interpretable Visual Features in Artificial and Biological Neural Systems, oct 2023. URL http://arxiv.org/abs/2310.11431. arXiv:2310.11431 [cs, stat].

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. From superposition to sparse codes: interpretable representations in neural networks. *arXiv preprint arXiv:2503.01824*, 2025.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, jun 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL https://www.nature.com/articles/381607a0. Publisher: Nature Publishing Group.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders. *arXiv preprint arXiv:2411.13117*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42 (8):2648–2669, 2018.

David Pfau, Irina Higgins, Alex Botev, and Sébastien Racanière. Disentangling by subspace diffusion. *Advances in Neural Information Processing Systems*, 33:17403–17415, 2020.

Lorenzo Posani, Shuqi Wang, Samuel P Muscinelli, Liam Paninski, and Stefano Fusi. Rarely categorical, always high-dimensional: how the neural code changes along the cortical hierarchy. *bioRxiv*, pp. 2024–11, 2025.

Jacob S Prince, George A Alvarez, and Talia Konkle. Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances*, 10(39):eadl1776, 2024.

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

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pp. 444–461. Springer, 2024.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. *arXiv preprint arXiv:2410.21869*, 2024.

Patrik Reizinger, Randall Balestriero, David Klindt, and Wieland Brendel. An empirically grounded identifiability theory will accelerate self-supervised learning research. *bioRxiv*, 2025.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451):585–590, 2013.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, 35:16052–16067, 2022.

Rylan Schaeffer, Mikail Khona, Sarthak Chandra, Mitchell Ostrow, Brando Miranda, and Sanmi Koyejo. Position: maximizing neural regression scores may not identify good models of the brain. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990. URL https://www.sciencedirect.com/science/article/pii/000437029090007M. Publisher: Elsevier.

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.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, and et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

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.

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

Zeyu Yun, Yubei Chen, Bruno A. Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors, 2021. URL http://arxiv.org/abs/2103.15949. arXiv:2103.15949 [cs].

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, feb 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html. arXiv: 2102.08850.

A Appendix

A.1 Derivation of analytical RSA

To derive an analytic expression for the RSA under superposition, we first express the RSMs in terms of the Gram matrices $G_a = A_a^\intercal A_a$ and $G_b = A_b^\intercal A_b$. These matrices act as metric tensors, defining the geometry of the representations.

$$M(Y_a) = (A_a Z)^\intercal (A_a Z) = Z^\intercal G_a Z \tag{10}$$

$$M(Y_b) = (A_b Z)^\intercal (A_b Z) = Z^\intercal G_b Z \tag{11}$$

An individual element of these matrices is the quadratic form $M(Y_a)_{ij} = z_i^\intercal G_a z_j$. Our derivation relies on the following standard assumptions about the distribution of the latent variable vectors z_i:

1. The latent vectors $z_1, \ldots, z_d$ are independent and identically distributed (i.i.d.).
2. The distribution has a mean of zero: $\mathbb{E}[z_i] = \mathbf{0}$.
3. The distribution is white, with an identity covariance matrix: $\mathbb{E}[z_i z_j^T] = \delta_{ij} I_n$.

Expectation of RSM Elements We first derive the empirical mean of all RSM matrix elements μ_Y in asymptotic limit, then derive the empirical mean of only the off-diagonal upper triangular RSM matrix elements μ_Y^{UT}, and show that in the asymptotic limit the two empirical quantities are equivalent and converge to zero:

$$\mu_Y \equiv \frac{1}{d^2} \sum_{i,j} M(Y)_{ij} = \frac{1}{d^2} \sum_{i,j} z_i^T G z_j \tag{12}$$

$$= \frac{1}{d} \sum_i z_i^T G \left[\frac{1}{d} \sum_j z_j \right] \tag{13}$$

$$\approx \frac{1}{d} \sum_i z_i^T G \mathbb{E}[z_j] = z_i^T G \mathbf{0} \tag{14}$$

$$= 0 \tag{15}$$

$$\mu_Y^{UT} \equiv \frac{1}{d(d-1)/2} \sum_{i<j} M(Y)_{ij} = \frac{1}{d(d-1)} \sum_{i \neq j} M(Y)_{ij} \tag{16}$$

$$= \frac{1}{d(d-1)} \left\{ \left[\sum_{i,j} M(Y)_{ij} \right] - \left[\sum_i M(Y)_{ii} \right] \right\} \tag{17}$$

$$= \frac{d^2}{d(d-1)} \mu_Y - \frac{1}{d-1} \mu_Y^{\text{diag}} \tag{18}$$

$$\approx \mu_Y \tag{19}$$

$$= 0 \tag{20}$$

Covariance and Variance Since the mean of the off-diagonal elements is zero, their covariance for $i \neq j$ is the empirical mean of their product: The Covariance of the off-diagonal elements of two

RSMs can then be shown as:

$$\begin{aligned}
\text{Cov}(\vec{m}_a, \vec{m}_b) &= \text{Cov}(M(Y_a)^{\text{UT}}, M(Y_b)^{\text{UT}}) && (21)\\
&= \frac{1}{d(d-1)/2} \sum_{i<j} \{M(Y_a)_{ij} - \mu_a^{\text{UT}}\}\{M(Y_b)_{ij} - \mu_b^{\text{UT}}\} && (22)\\
&\approx \frac{1}{d(d-1)/2} \sum_{i<j} M(Y_a)_{ij} M(Y_b)_{ij} = \frac{1}{d(d-1)} \sum_{i \neq j} M(Y_a)_{ij} M(Y_b)_{ij} && (23)\\
&= \frac{1}{d(d-1)} \left\{ \left[\sum_{i,j} M(Y_a)_{ij} M(Y_b)_{ij} \right] - \left[\sum_{i} M(Y_a)_{ii} M(Y_b)_{ii} \right] \right\} && (24)\\
&\approx \frac{1}{d(d-1)} \sum_{i,j} M(Y_a)_{ij} M(Y_b)_{ij} && (25)\\
&= \frac{1}{d(d-1)} \sum_{i,j} (z_i^{\mathsf{T}} G_a z_j)(z_i^T G_b z_j) && (26)\\
&= \frac{1}{d(d-1)} \sum_{i,j} (z_i^{\mathsf{T}} G_a z_j)(z_j^T G_b^{\mathsf{T}} z_i) && (27)\\
&= \frac{1}{d-1} \sum_{i} z_i^{\mathsf{T}} G_a \left[\frac{1}{d} \sum_{j} z_j z_j^{\mathsf{T}} \right] G_b z_i && (28)\\
&\approx \frac{1}{d-1} \sum_{i} z_i^{\mathsf{T}} G_a \mathbb{E}[z_j z_j^{\mathsf{T}}] G_b z_i && (29)\\
&= \frac{1}{d-1} \sum_{i} z_i^T G_a G_b z_i && (30)\\
&= \frac{d}{d-1} \text{Tr} \left[G_a G_b \left(\frac{1}{d} \sum_{i} z_i z_i^{\mathsf{T}} \right) \right] && (31)\\
&\approx \text{Tr} \left[G_a G_b \mathbb{E}[z z^{\mathsf{T}}] \right] && (32)\\
&= \text{Tr} \left[G_a G_b \right] && (33)
\end{aligned}$$

The variance of the elements is found by setting $G_a = G_b$, and can be related to the Frobenius norm ($\|X\|_F^2 = \text{Tr}(X^T X)$):

$$\text{Var}(\vec{m}_a) = \text{Var}(M(Y_a)^{\text{UT}}) = \text{Tr}(G_a G_a) = \text{Tr}(G_a^{\mathsf{T}} G_a) = \|G_a\|_F^2 \quad (34)$$

$$\text{Var}(\vec{m}_b) = \text{Var}(M(Y_b)^{\text{UT}}) = \text{Tr}(G_b G_b) = \text{Tr}(G_b^{\mathsf{T}} G_b) = \|G_b\|_F^2 \quad (35)$$

For a large number of data points d, the correlation of the vectorized RSMs is well-approximated by the correlation of their constituent elements. Substituting the covariance and variance into the Pearson formula yields our main result:

$$\boxed{\rho(Y_a, Y_b) \approx \frac{\text{Tr}(G_a G_b)}{\sqrt{\|G_a\|_F^2 \|G_b\|_F^2}} = \frac{\langle G_a, G_b \rangle_F}{\|G_a\|_F \|G_b\|_F}} \quad (36)$$

A.2 Derivation of analytical Linear Regression results

We consider a multivariate linear regression model to predict the activity of representation Y_b from Y_a:

$$Y_b = W Y_a + E \quad (37)$$

where $W \in \mathbb{R}^{m_b \times m_a}$ is the weight matrix and E is the matrix of residuals. The Ordinary Least Squares (OLS) method finds the estimator $\hat{W}$ that minimizes the sum of squared errors, given by the squared Frobenius norm $\|Y_b - W Y_a\|_F^2$.

OLS Estimator and Asymptotic Simplification The standard OLS solution for the weight matrix is:

$$\hat{W} = Y_b Y_a^\mathsf{T} (Y_a Y_a^\mathsf{T})^{-1} \tag{38}$$

To find an analytic expression in terms of the underlying superposition matrices, we substitute $Y_a = A_a Z$ and $Y_b = A_b Z$. We then leverage the same statistical properties of the latent variables Z used in the RSA derivation. For a large number of i.i.d. samples d, the sample covariance of the latent variables converges to a scaled identity matrix:

$$\frac{1}{d} Z Z^\mathsf{T} = \frac{1}{d} \sum_{i=1}^{d} z_i z_i^\mathsf{T} \to \mathbb{E}[z z^\mathsf{T}] = I_n \quad \implies \quad Z Z^\mathsf{T} \approx d I_n$$

Using this approximation, the terms in the OLS estimator simplify:

$$Y_b Y_a^\mathsf{T} = (A_b Z)(A_a Z)^\mathsf{T} = A_b (Z Z^\mathsf{T}) A_a^\mathsf{T} \approx d (A_b A_a^\mathsf{T}) \tag{39}$$

$$Y_a Y_a^\mathsf{T} = (A_a Z)(A_a Z)^\mathsf{T} = A_a (Z Z^\mathsf{T}) A_a^\mathsf{T} \approx d (A_a A_a^\mathsf{T}) \tag{40}$$

Substituting these into the formula for $\hat{W}$ gives the ideal "population" level regression coefficient, which is free from the sampling noise of a specific Z:

$$\hat{W} \approx d (A_b A_a^\mathsf{T}) \left(d (A_a A_a^\mathsf{T}) \right)^{-1} = A_b A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1} \tag{41}$$

Derivation of the Mean Squared Error The Mean Squared Error (MSE) is the total squared error divided by the total number of predicted elements, $m_b d$. The prediction error matrix is $E = Y_b - \hat{W} Y_a$.

$$E \approx A_b Z - \left(A_b A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1} \right) A_a Z \tag{42}$$

$$= \left(A_b - A_b A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1} A_a \right) Z \tag{43}$$

The total squared error is the squared Frobenius norm of E.

$$\begin{aligned} \|E\|_F^2 = \text{Tr}(E^\mathsf{T} E) &\approx \text{Tr}\left(Z^\mathsf{T} (\dots)^\mathsf{T} (\dots) Z \right) \\ &= \text{Tr}\left((\dots)^\mathsf{T} (\dots) (Z Z^\mathsf{T}) \right) \quad \text{(using cyclic property of trace)} \\ &\approx d \cdot \text{Tr}\left((\dots)^\mathsf{T} (\dots) \right) = d \left\| A_b - A_b A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1} A_a \right\|_F^2 \end{aligned} \tag{44}$$

Dividing the total squared error by $m_b d$ yields the final MSE expression:

$$\boxed{\text{MSE}(Y_b | Y_a) \approx \frac{1}{m_b} \left\| A_b - A_b \left(A_a^\mathsf{T} (A_a A_a^\mathsf{T})^{-1} A_a \right) \right\|_F^2} \tag{45}$$

Notation:

$$\hat{Y_b} = (\hat{y}_{b,(1)}, ..., \hat{y}_{b,(d)}) \tag{46}$$

$$\begin{aligned} \text{E}[\hat{Y_b}^i] \equiv \frac{1}{d} \sum_{k=1}^{d} \hat{y}^i_{b,(k)} &= \frac{1}{d} \sum_{k=1}^{d} \sum_m \hat{W}^{im} y^m_{a,(k)} \\ &= \frac{1}{d} \sum_{k=1}^{d} \sum_{m,n} \hat{W}^{im} A_a^{mn} z^n_{(k)} = \sum_{m,n} \hat{W}^{im} A_a^{mn} \frac{1}{d} \sum_{k=1}^{d} z^n_{(k)} \\ &\approx \sum_{m,n} \hat{W}^{im} A_a^{mn} \text{E}[z^n] \\ &= 0 \end{aligned}$$

$$\text{E}[y^i y^j] = \sum_{m,n} A^{im} A^{jn} \text{E}[z^m z^n] = \sum_{m,n} A^{im} A^{jn} \delta_{mn} = \sum_m A^{im} A^{jm} = (A A^\mathsf{T})_{ij}$$

Derivation of the Explained Variance R^2 The Explained Variance R^2 is defined by:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{47}$$

where

$$SS_{\text{res}} = \sum_{k=1}^{d} ||y_{b,(k)} - \hat{y}_{b,(k)}||^2 \tag{48}$$

$$SS_{\text{tot}} = \sum_{k=1}^{d} ||y_{b,(k)} - \bar{y}_b||^2 \tag{49}$$

$$\bar{y}_b = \frac{1}{d}\sum_{k=1}^{d} y_{b,(k)} \tag{50}$$

We can derive an analytical expression of SS_{res}, SS_{tot}, and $\bar{y}_b$ in terms of the projection matrices A_a and A_b:

$$\bar{y}_b = \frac{1}{d}\sum_{k=1}^{d} y_{b,(k)} = A_b \frac{1}{d}\sum_{k=1}^{d} z_k \approx A_b \mathrm{E}[z] \tag{51}$$

$$= 0 \tag{52}$$

$$SS_{\text{res}} = \sum_{k=1}^{d} ||y_{b,(k)} - \hat{y}_{b,(k)}||^2 = \text{Tr}[(Y_b - \hat{Y}_b)^\mathsf{T}(Y_b - \hat{Y}_b)] = \text{Tr}[Z^\mathsf{T}(A_b - \hat{W}A_a)^\mathsf{T}(A_b - \hat{W}A_a)Z] \tag{53}$$

$$= \text{Tr}[(A_b - \hat{W}A_a)^\mathsf{T}(A_b - \hat{W}A_a)ZZ^\mathsf{T}] \approx d \cdot \text{Tr}[(A_b - \hat{W}A_a)^\mathsf{T}(A_b - \hat{W}A_a)] \tag{54}$$

$$SS_{\text{tot}} = \sum_{k=1}^{d} ||y_{b,(k)} - \bar{y}_b||^2 \approx \sum_{k=1}^{d} ||y_{b,(k)}||^2 = \text{Tr}[Y_b^\mathsf{T} Y_b] \tag{55}$$

$$= \text{Tr}[Z^\mathsf{T} A_b^\mathsf{T} A_b Z] = \text{Tr}[A_b^\mathsf{T} A_b Z Z^T] \tag{56}$$

$$\approx d \cdot \text{Tr}[A_b^\mathsf{T} A_b] \tag{57}$$

Thus the analytical expression of R^2 can be expressed as:

$$\boxed{R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\text{Tr}[(A_b - \hat{W}A_a)^\mathsf{T}(A_b - \hat{W}A_a)]}{\text{Tr}[A_b^\mathsf{T} A_b]}} \tag{58}$$

Derivation of the Pearson Correlation The prediction is $\hat{Y}_b = \hat{W}Y_a$

The Pearson Correlation matrix between the prediction and the ground truth is given by:

$$\rho(\hat{Y}_b, Y_b)_{ij} \equiv \rho(\hat{Y}_b{}^i, Y_b{}^j) = \frac{\text{Cov}(\hat{Y}_b{}^i, Y_b{}^j)}{\sqrt{\text{Var}(\hat{Y}_b{}^i)\text{Var}(Y_b{}^j)}} \tag{59}$$

Where indices i and j correspond to system dimensions. The Covariances can be expressed as:

$$
\begin{aligned}
\text{Cov}(\hat{Y_b}^i, {Y_b}^j) &= \frac{1}{d-1}\sum_{k=1}^{d} \hat{y}^i_{b,(k)} y^j_{b,(k)} = \frac{1}{d-1}\sum_{k=1}^{d}\sum_m \hat{W}^{im} y^m_{a,(k)} y^j_{b,(k)} \\
&= \frac{1}{d-1}\sum_{k=1}^{d}\sum_{m,n,l} \hat{W}^{im} A^{mn}_a z^n_{(k)} A^{jl}_b z^l_{(k)} = \frac{1}{d-1}\sum_{m,n,l} \hat{W}^{im} A^{mn}_a A^{jl}_b \sum_{k=1}^{d} z^n_{(k)} z^l_{(k)} \\
&\approx \frac{1}{d-1}\sum_{m,n,l} \hat{W}^{im} A^{mn}_a A^{jl}_b d \cdot \text{E}[z^n z^l] = \frac{d}{d-1}\sum_{m,n,l} \hat{W}^{im} A^{mn}_a A^{jl}_b \delta_{nl} \\
&\approx \sum_{m,n} \hat{W}^{im} A^{mn}_a A^{jn}_b = (\hat{W} A_a A_b^\intercal)_{ij} = (A_b A_a^\intercal (A_a A_a^\intercal)^{-1} A_a A_b^\intercal)_{ij}
\end{aligned}
$$

And the Variances:

$$
\begin{aligned}
\text{Var}(\hat{Y_b}^i) &= \frac{1}{d-1}\sum_{k=1}^{d} \hat{y}^i_{b,(k)} \hat{y}^i_{b,(k)} = \frac{1}{d-1}\sum_{k=1}^{d}\sum_{m,n} \hat{W}^{im} y^m_{a,(k)} \hat{W}^{in} y^n_{a,(k)} \\
&= \frac{1}{d-1}\sum_{m,n} \hat{W}^{im}\hat{W}^{in} \sum_{k=1}^{d} y^m_{a,(k)} y^n_{a,(k)} \\
&\approx \frac{1}{d-1}\sum_{m,n} \hat{W}^{im}\hat{W}^{in} d \cdot \text{E}[y^m_a y^n_a] \\
&= \frac{d}{d-1}\sum_{m,n} \hat{W}^{im}\hat{W}^{in} (A_a A_a^T)_{mn} \\
&\approx (\hat{W}(A_a A_a^\intercal)\hat{W}^\intercal)_{ii} \\
&= (A_b A_a^\intercal (A_a A_a^\intercal)^{-1} (A_a A_a^\intercal)\hat{W}^\intercal)_{ii} \\
&= (A_b A_a^\intercal (A_a A_a^\intercal)^{-1} A_a A_b^\intercal)_{ii}
\end{aligned}
$$

$$
\begin{aligned}
\text{Var}(Y^j_b) &= \frac{1}{d-1}\sum_{k=1}^{d} y^j_{b,(k)} y^j_{b,(k)} \approx \frac{d}{d-1}\text{E}[y^j_b y^j_b] \\
&\approx (A_b A_b^\intercal)_{jj}
\end{aligned}
$$

Expressed in A_a and A_b, the Pearson Correlation matrix becomes:

$$
\boxed{\rho(\hat{Y_b}, Y_b)_{ij} \approx \frac{(A_b A_a^\intercal (A_a A_a^\intercal)^{-1} A_a A_b^\intercal)_{ij}}{\sqrt{(A_b A_a^\intercal (A_a A_a^\intercal)^{-1} A_a A_b^\intercal)_{ii}(A_b A_b^\intercal)_{jj}}} = \frac{(\hat{W} A_a A_b^\intercal)_{ij}}{\sqrt{(\hat{W} A_a A_b^\intercal)_{ii}(A_b A_b^\intercal)_{jj}}}} \tag{60}
$$