Measuring Fracture and Entanglement in Neural Networks

Gabriel Kahen, Shyam Sharma | Stony Brook University | 18 November 2025

Abstract

Modern deep learning models achieve remarkable external performance, yet often rely on brittle and disorganized internal representations. The *Fractured Entangled Representation (FER) hypothesis* argues that conventional training produces networks whose internal structures are redundant (fractured) and interfering (entangled) rather than unified and modular (Kumar et al., 2025). Despite growing qualitative evidence for this phenomenon, no widely accepted quantitative framework currently exists to measure it. This paper introduces two complementary metrics, Fracture (*F*) and Entanglement (*E*), that formalize the FER hypothesis and enable systematic evaluation of representation quality. We define both metrics mathematically, validate them across multiple varied tests, and analyze where they succeed and where they fail.

1 Introduction

In recent years especially, deep learning models have demonstrated extraordinary task performance. Believers in the current paradigm of deep learning preach that by simply scaling our current approaches, we will be able to achieve ‘artificial general intelligence’. However, the current success of these models shrouds the underlying weaknesses in their internal representational structure that skeptics believe are a roadblock to true general/superintelligence.

In the field, an internal representational structure refers to the hidden way a model organizes and stores concepts beneath the surface to make sense of the world enough to produce its outputs.

Well, why does representation even matter? Don't the results just speak for themselves? Think of it like two students who both score well on a test. Knowing that the test problems would come from the textbook, one student simply memorized all the answers, while the other studied the actual methods behind solving them. They both ace the exam, but only the second student can handle new problems. In the same way, AI models need good internal representations to truly generalize, adapt, and be trusted beyond the narrow tasks they were trained on. Scaling up a model without improving its representation is like trying to model a sphere by gluing on more strands of hair to a hairball. From far away, it might look like a sphere, but up close it's still messy, uneven, and fragile. True progress comes not from piling on more hairs, but from rethinking the structure itself.

A growing body of work is questioning the current dogma of "representational optimism," the assumption that good outcomes imply well-formed internal organizations. The *Fractured Entangled Representation (FER)* hypothesis put forth by Kumar et al. puts forth a concrete argument against this dogma. The paper argues that "underlying representation is not about capturing a particular pattern or output behavior. Rather, it is about whether the system can build on the regularities of that representation to learn and generate new behaviors" (9). It then theorizes that standard trained modern models, particularly those trained via stochastic gradient descent (SGD), are hairballs that encourage two big representational issues: fracture and entanglement.

Fracture (F) is when a model splits a single factor into many scattered and redundant features. This means that changing one thing requires adjusting multiple parts of the model instead of a single, unified control. Hence, the opposite of a fractured representation is a *unified* one.

Entanglement (E) is when a model mixes together different factors that should be independent, so that changing one thing unintentionally alters other unrelated parts as well. The opposite of an entangled representation is a modular or *factored* one.

Because of this, we call an aspirational representation that doesn't exhibit much fracture or entanglement a *Unified Factored Representation (UFR)***.** An ideal UFR would represent each capability once, cleanly separated from others, and would be able to generalize on each of these capabilities.

To more clearly understand a FER compared to a UFR, let's use the human body as an example. In our body, each organ has a clear role: the heart pumps blood, the lungs process oxygen, and the stomach digests food. This is like a UFR, where each factor is handled once and cleanly separated from the others. A FER, by contrast, would be like a body where breathing requires the cooperation of several scattered mini-organs, each doing part of the job (fracture), and where the stomach and lungs constantly interfere with each other so that adjusting one throws the other off balance (entanglement). Both "bodies" might stay alive, but only the UFR-like one has the clarity and efficiency to adapt and stay healthy. We can even see this idea in our genes: our DNA doesn't redundantly encode separate blueprints for the right and left arms, but instead encodes a single "arm plan" that is flexibly expressed on both sides. That kind of unified, factored representation is what gives biological systems their robustness and

adaptability. However, Kumar et al.'s paper only gives qualitative examples of fracture and entanglement, and never provides a quantitative measure for either of them.

This paper attempts to fill that gap by creating and validating simple metrics for fracture and entanglement. These metrics will allow researchers to move beyond intuitive demonstrations and systematically compare representations across models. The paper proceeds by outlining the current state of representational research, introducing quantitative definitions for both *F* and *E*, applying these metrics on two targeted tests, then examining where our approach succeeds and fails.

2 Background

The concept that internal representations matter has long been discussed in deep learning research. Pioneers of deep learning understood the representation issue as one of the main barriers to success. Bengio et al. (2013) articulated a unifying vision for representational learning, arguing that the key to intelligence lies in discovering internal representations that make useful abstractions of the world. They wrote: “explicitly dealing with representations is interesting … [because] they can be convenient to express many general priors about the world around us … [that] are not task-specific but would be likely to be useful for a learning machine to solve AI-tasks” (3). In this view, representations are not just tedious intermediate encodings, but the very medium through which generalization and reasoning become possible. This framework redefined the goal of machine learning: not simply to map inputs to outputs, but to discover the internal organization that makes learning itself more efficient, transferable, and generalizable.

Following that groundbreaking research, the field began to place more importance on representations. However, Locatello et al. (2019) challenged the field's confidence in its understanding of what actually makes a representation "good," showing that even when models appear to learn structured internal features, these structures may be arbitrary and uninformative without explicit inductive biases or supervision. Their analysis specifically targeted the goal of *disentanglement*, demonstrating that this goal cannot be reliably achieved or even identified in an unsupervised setting. This ambiguity revealed how easily models can give the illusion of internal coherence while relying on arbitrary latent organizations. In doing so, Locatello et al. reframed the field's central challenge from figuring out how to make representations work to how to rigorously define and measure their quality.

Working on this uncertainty, Chen et al. (2018) proposed a more rigorous way to evaluate the structure of learned representations. Their work focused on the variational autoencoder (VAE), a generative model that learns to compress data into a low-dimensional latent space and then reconstruct it, effectively forcing the model to discover a structured internal representation of the input. They introduce the β-TCVAE, a refinement of the variational autoencoder (VAE) that isolates *total correlation*, a statistical measure of dependence among latent variables, as the key quantity governing disentanglement. Their approach involved penalizing correlations between latent dimensions, which encouraged models to represent independent factors of variation more cleanly, providing a concrete step forward in disentanglement. Their work shifted the discussion from whether disentanglement was even possible to how it could be effectively operationalized.

As the theoretical debate over representational quality deepened, Dziri et al. (2023) brought some empirical clarity by examining how large language models, specifically GPT-3 and

GPT-4, handle compositional reasoning tasks. Their paper tested whether these models could break problems into intermediate steps and combine those steps into coherent solutions, arguing that this requires internally consistent representations so that the partial results can be meaningfully reused across reasoning steps. Despite strong performance on familiar examples, the authors found that transformers often relied on shallow pattern matching rather than genuine multistep reasoning. They write that "while models can memorize single-step operations, they fail to compose them into correct reasoning paths, suggesting that they mostly make predictions based on shallow, rote learning rather than a deep holistic task understanding (2). These findings suggest that even highly capable systems organize information in ways that are brittle, exemplifying the persistence of representational weakness in today's models.

This issue is tackled by Kumar et al. (2025) with their *Fractured Entangled Representation (FER)* hypothesis discussed earlier. However, the paper only provided qualitative demonstrations, relying on visual and conceptual examples of fracture and entanglement without a formal method to measure them. As a result, the FER hypothesis remains an important but largely theoretical contribution. The following sections, *Measuring Fracture (F)* and *Measuring Entanglement (E)*, attempt to introduce a framework for operationalizing these concepts through quantitative metrics that evaluate the presence of Fracture and Entanglement.

3 Measuring Fracture (*F*)

As discussed previously, fracture is essentially just how a single concept gets split across multiple disconnected pieces inside a model's representation. At a mechanistic level, *fracture* quantifies how many separate representation fragments the network uses to encode what should be a single underlying concept. If a model learns one coherent "chunk" of representation for each

concept, it has low fracture. If it learns several partially redundant chunks, each capturing the same concept in a slightly different context, it has high fracture.

Mathematically, we can think of the representation as an n x d matrix R, where each of the d neurons (or features) contributes to encoding various concepts. As Raghu et al. (2017) establish, "with this interpretation of neurons as vectors (and layers as subspaces, spanned by neurons)... we introduce a powerful method for analyzing deep representation" (1). This conceptualization underlies our treatment of each concept's representation as a subspace within the broader network. For any given concept z_k, we can look at how activations change when that concept is present. If all variance associated with z_k lies along one dominant direction in activation space, it means that when the concept z_k changes, the network's activation changes in a consistent, coordinated way, meaning that the representation is unified.

Conversely, if the variance is scattered across several independent directions, it is fractured. Thus, we can quantify fracture by counting how many distinct directions are needed to reconstruct most of the concept's activity. To make this intuition precise, we can quantify how widely the variance associated with each concept spreads across independent directions in activation space.

To operationalize this idea, let $R \in \mathbb{R}^{n\,x\,d}$ denote the activation matrix, where n is the number of samples (e.g. datapoints or stimuli) and d is the number of representational units (neurons, dimensions, or features) in the layer being analyzed. Each row of R thus represents the activation pattern of all d units for one sample. Before analysis, we must normalize this activation matrix to remove arbitrary scaling differences between neurons and layers. We do this by z-scoring each feature across samples so that every neuron has zero mean and unit variance.

Since fracture is defined relative to particular underlying concepts, we define $\{z_k\}_{k=1}^{K}$, or Z, be the set of latent factors or concepts we are probing. Each z_k corresponds to a distinct underlying concept (for example, color, shape, or orientation) that we want to measure fracture for. For each possible value v of that concept z_k, we compute its mean activation vector: $\mu_k(v) = \mathbb{E}[R \mid z_k = v]$. These means represent the average activation pattern of the model when that concept takes value v, revealing how the network's internal state systematically shifts as the concept changes. Subtracting the global mean $\overline{R} = \mathbb{E}[R]$ yields center means which isolate the component of activity attributable to that concept alone:

$$\vec{u}_k(v) = \mu_k(v) - \overline{R}. \quad (1)$$

The structure of these shifts can be described by the between-concept covariance matrix:

$$S_{B,k} = \sum_v p(v)\, \vec{u}_k(v)\, \vec{u}_k(v)^\top. \quad (2)$$

Here, $p(v)$ is the empirical probability of each value of z_k. This matrix captures how the average activation moves through representation space as the concept changes, disentangling systematic concept-dependent variance from random within-group noise.

Performing an eigendecomposition of $S_{B,k}$ gives eigenvalues $\lambda_{k,1} \geq \lambda_{k,2} \geq \ldots \geq \lambda_{k,d}$, each corresponding to an independent direction along which the concept causes the representation to vary.

We then define the *effective dimensionality* of concept z_k as a measure of how many independent directions in activation space carry substantial variation related to that concept. To

estimate this, we use the participation ratio of the eigenvalue spectrum, which quantifies how evenly variance is distributed among the eigen-directions of $S_{B,\,k}$. The participation ratio was introduced in *A Theory of Multineuronal Dimensionality, Dynamic and Measurement* (Gao et al., 2017) as a rigorous way to quantify the effective dimensionality of a covariance spectrum. This is the exact tool we need for assessing how a concept's representation spreads across different eigen-directions, and is defined as follows:

$$m_k = \frac{(\sum_{i=1}^{d} \lambda_{k,i})^2}{\sum_{i=1}^{d} \lambda_{k,i}^2}. \quad (3)$$

This value behaves like an "effective number" of directions that the concept occupies in activation space. When almost all variance lies along few dominant axes, we get a small m_k; when variance is spread across several orthogonal directions, m_k increases, indicating greater fracture. Since it depends only on the relative distribution of eigenvalues, this formulation is scale-free and robust to small, noise-driven components that might otherwise inflate dimensionality estimates.

Finally, to summarize the overall degree of representational fracture across all K concepts, we normalize and average these dimensionalities:

$$F = \frac{1}{K} \sum_{k=1}^{K} \frac{m_k - 1}{d - 1}. \quad (4)$$

Here, K is the total number of concepts, and $F \in [0, 1]$. A lower F indicates a more unified representation, while higher values indicate greater fracture across the representation.

4 Measuring Entanglement (*E*)

As discussed previously, entanglement captures how different underlying concepts become mixed together inside a model's representation. At a mechanistic level, entanglement measures how much distinct representational functions overlap or interfere with one another. A representation is entangled when changes in one concept's activation patterns also cause correlated changes in others, indicating that their internal subspaces are not independent. Conversely, a representation is factored when each concept occupies its own orthogonal direction in activation space, varying independently from the rest. This notion of orthogonality as independence is supported by Flesch et al. (2022), who show that separating task representations into orthogonal subspaces prevents interference across contexts. This empirically demonstrates that orthogonality yields modular, non-interacting representations, precisely what we define as low entanglement. To make this intuition precise, we can quantify entanglement by measuring the degree to which these subspaces cohabit or bleed into one another, indicating representational interference and a loss of modularity.

To express this formally, we use the same notation introduced in in Section 3.1: R is the activation matrix of hidden representations, $Z = [z_1, \ldots, z_K]$ the set of explanatory factors, and $S_{B,k}$ is the between-factor covariance as defined in Equation (2). Each $S_{B,k}$ describes how the model's internal activity changes when factor z_k varies. We then again compute its eigendecomposition:

$$S_{B,k} = U_k \Lambda_k U_K^{\top}. \quad (5)$$

Here, U_k is the matrix of orthonormal eigenvectors and Λ_k is the diagonal matrix of corresponding eigenvalues. Each column of U_k represents a principal direction in representation

space through which changes in z_k are expressed, and each eigenvalue in Λ_k indicates how much variance occurs along that direction.

To define the subspace that the model uses to represent z_k, we let $U_k^{(m_k)} \in \mathbb{R}^{d \times m_k}$ denote the matrix containing the top m_k eigenvectors corresponding to the largest eigenvalues of $S_{B,\,k}$ where m_k is again the effective dimensionality defined in Formula (3).

To measure how much the two subspaces overlap, we define a *projection operator*:

$$P_k = U_k^{(m_k)} (U_k^{(m_k)})^{\top} . \quad (6)$$

This operator maps any vector in the representation space onto the subspace corresponding to factor z_k. Intuitively, P_k acts as a filter that preserves only the component of a representation associated with z_k. This formulation follows prior work treating concept-subspaces via orthogonal projection operators (Moreira et al., 2025), which view each semantic factor as a distinct linear manifold in hidden-state space. If two factors are represented independently (that is, if each concept occupies its own orthogonal direction in activation space) then projecting onto one and then the other removes nearly all information. However, if they are entangled, the two projections recover much of the same content, because their internal subspaces cohabit or bleed into one another, indicating representational interference and a loss of modularity.

We must next introduce the *Frobenius norm*, which measures the total magnitude of a matrix by summing the squares of all its entries. Intuitively, this norm captures how much the two projection matrices "see" of each other when multiplied together. For two projection

matrices P_k and P_l, the quantity $\left|\left|P_k P_l\right|\right|_{fro}^2$ equals the sum of squared cosines of the principal angles between the subspaces they represent. More simply, it measures the total shared volume between the two regions of representation space, which is exactly what we need to calculate E.

With this, we can define the pairwise entanglement between two factors as:

$$E_{kl} = \frac{\left|\left|P_k P_l\right|\right|_{fro}^2}{min(m_k, m_l)}. \quad (7)$$

Here, the numerator measures how strongly the subspaces for z_k and z_l overlap, while the denominator normalizes the size of each subspace. E_{kl} ranges from 0 (perfectly independent) to 1 (perfectly overlapping).

The global level of Entanglement is summarized as:

$$E = \frac{1}{K(K-1)} \sum_{k \neq l} E_{kl}. \quad (8).$$

Here, K is the total number of concepts, and $E \in [0, 1]$. A lower E indicates a more factored representation, while higher values indicate greater entanglement across the representation.

5 Experimental Validation

In order to evaluate whether the proposed metrics actually behave in line with our theoretical expectations, some experimental validation is necessary. In this section, we apply the metrics in two controlled settings designed to probe different aspects of representational structure. The goal of these tests is to examine how well the metrics actually track changes in

representational structure and align with the qualitative behavior suggested by prior research. Taken together, they provide a basic checker on whether F and E capture the kind of differences that theory of fracture and entanglement should detect.

5.1 Test 1: Toy Model of Superposition

The first test for our metrics for F and E replicates the core findings of *Toy Models of Superposition* (Elhage et al., 2022). This section demonstrates that neural networks can encode more features than they have neurons by storing them in partially overlapping directions, a phenomenon termed *superposition*. In their toy setup, small ReLU networks trained on sparse data transitioned from disentangled, orthogonal feature representations to mixed ones as sparsity increased, revealing that networks trade off representational capacity for interference when limited by dimensional constraints.

This test essentially reverses that setup by fixing the number of features and instead varying the network's width, which determines how much representational space the model has. When the width is small, the model cannot assign each feature its own direction. This means that the model is forced to have different features share dimensions, making the representations more fractured and entangled. As width increases, we expect F and E to decrease since there is more space for the model to represent features more independently, reducing redundancy and overlap.

The implementation of this test is publicly available on GitHub (Kahen, 2025), yielding the results shown below.

Width	F	E
5	0.5296	0.7423
10	0.5421	0.5697
20	0.4457	0.5964
50	0.1126	0.2909
100	0.0492	0.2482

The results of this experiment results in the exact pattern that we expect: as width increases, F and E both decrease. This indicates that the metrics correctly capture the transition from high-overlap superposed features to more orthogonal, factored representations.

5.2 Test 2: Picbreeder vs. SGD

The second test compares representations learned through open-ended evolutionary search with those learned through conventional SGD, following the experimental design of *Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis* (Kumar et al., 2025). This paper argues that networks trained via SGD tend to develop representations that exhibit high fracture and entanglement, while networks evolved through open-ended processes such as *Picbreeder* tend to form more unified and factored representations. Kumar et al. evolved CPPNs to generate images such as apples, butterflies, and skulls, then retrained equivalent networks with SGD to produce the same images. Although the outputs were visually identical, the paper argues that the SGD models are structurally inferior due to their high fracture and entanglement.

This test applies our metrics for F and E on the six models provided by the paper. The implementation of this test is publicly available on GitHub (Kahen, 2025), yielding the results shown below.

Model	***F***	***E***
Picbreeder Apple	0.1151	0.1951
SGD Apple	0.1212	0.1093
Picbreeder Skull	0.2041	0.6923
SGD Skull	0.3197	0.3480
Picbreeder Butterfly	0.1927	0.5230
SGD Butterfly	0.2081	0.1744

The results confirm the expected behavior for F, but not for E. Across all images, the Picbreeder networks generally exhibit lower F values, indicating the expected trend that evolved networks exhibit less fracture than SGD ones. However, E is consistently *higher* for the Picbreeder networks, contrary to the theoretical expectation that evolution-based training should produce less entanglement.

6 Discussion

The empirical results can be explained via a deeper conceptual tension that runs through this entire paper: fracture and entanglement, as described in the *FER Hypothesis*, are fundamentally *functional* properties, while the metrics developed here measure *geometric structure*. Here, I use "functional" to refer to how the model's structure manifests in behavior and

outputs, with “geometric” referring purely to its spatial arrangement in activation space. To a large extent, this paper relies on the assumption that functional fracture and entanglement can be understood directly through their geometric signatures. However, it may very well be the case that this assumption is inherently flawed, and that functionality cannot be understood as reducible to geometry.

This distinction clarifies both the strengths of the proposed frameworks and the anomalies it produces, particularly in the Picbreeder comparison (Section 5.2). Understanding this distinction allows us to reinterpret those anomalies not as mere failures of the metrics, but as evidence of a systematic mismatch between functionality and geometric structure, and to frame the rest of the discussion around when geometric proxies succeed or fail.

6.1 Interpreting the Picbreeder Anomaly

The most direct illustration of this mismatch appears in the Picbreeder test in Section 5.2. Again, this test diverged from theoretical expectations; entanglement was consistently higher for picbreeder models than their SGD counterparts. According to the FER framework that much of this paper is based on, evolutionary systems should, in principle, favor unified and modular representations, yet our metric E suggests otherwise. This apparent contradiction dissolves with the recognition that E measures geometric overlap between subspaces, not actual functionality.

The *FER hypothesis* characterizes entanglement in terms of interference between capabilities, or how changes in one factor disrupts the expression of another. An evolutionary process might be able to avoid entanglement and be functionally factored without producing clean geometric separation. In particular, I suspect that evolutionary algorithms can arrive at representations that reuse directions in activation space in non-linear or hierarchical ways. Such

representations could be functionally factored even though their activations occupy overlapping regions in a geometric sense. In this scenario, its capabilities do not interfere with one another, yet *E* is artificially large due to the apparent entanglement.

This suggests that an attempt to calculate *functional* entanglement would require a fundamentally different approach. This approach may involve capturing how activations interact during computation rather than how they are arranged in space. Such a metric would complement *E* by assessing what representations *do* rather than where they reside, offering a fuller picture of how modularity and interference emerge in both evolved and trained systems. Pessimistically, it may also be the case that functional entanglement is an inherently qualitative attribute, one whose expression depends too largely on context and subjective interpretation to be assigned a number. If so, any attempt to quantify may be inevitably reductive and unable to generalize in any meaningful way.

Interestingly, this distinction between structural and functional arrangement may also explain why fracture behaved as expected even when entanglement did not. *F* measures the dispersion of variance *within* a single subspace, while *E* depends on the relationship *between* subspaces. Since *E* is confounded by the emergent geometry of the whole system while *F* isn't, the latter is able to remain stable even when representations are reorganized globally. Evolutionary processes like those in Picbreeder may preserve local coherence even as they warp global structure, allowing *F* to hold while *E* drifts away from theoretical expectations.

6.2 Independence of *F* and *E*

The same functional-geometric mismatch helps explain why *F* and *E*, despite measuring conceptually orthogonal properties, are not mathematically independent in our metrics. In theory,

fracture and entanglement describe independent properties of representations; fracture is internal, while entanglement is relational, meaning that we should be able to change one without the other. However, our implementation of the two metrics *are not* fully independent as they both derive from the same covariance structure; excessive dispersion within a subspace can blur its boundaries and inflate apparent overlap with others. This coupling means that while fracture and entanglement are conceptually independent, our metrics F and E aren't.

Thus, our geometric understanding of functional properties forces them to depend on one another in a way that the underlying theory does not. In the functional picture, a model could be highly fractured yet clearly factored, or conversely, unified yet super entangled. However, in our geometric picture, F and E are constrained by the same activation manifold, so any change that affects one can easily affect the other. Viewed through this lens, the partial dependence between F and E is not an accident of implementation but a direct consequence of treating functional notions as geometric proxies.

This partial dependence complicates how we interpret changes in F and E. When both rise or fall together, it may become difficult to determine whether a model's behavior reflects genuine shifts in representational organization or simply artifacts of shared geometry. This ambiguity limits how confidently we can attribute trends in either metric to underlying structure, meaning that the metrics are context-dependent and must be considered mutually.

6.3 Future Research Directions

The limitations that come from trying to understand fracture and entanglement through pure geometry point toward several natural extensions of this work. Since fracture and entanglement are fundamentally functional notions, future attempts at measuring them should

aim to understand how they manifest *during computation* (rather than in their static structure). This is a difficult problem considering how multivariate and context-dependent neutral networks are, yet it suggests some avenues worth exploring. One approach may be to study how representations behave under controlled perturbations to the model. This means selectively altering one factor to see how that affects the model's performance. By examining how these changes propagate through the network, dependencies or separations that may be invisible to the static geometry might reveal themselves.

Pessimistically, it may also be the case that fracture and entanglement are inherently qualitative attributes whose expression depends too largely on context and subjective interpretation to be assigned a number. If so, any attempt to quantify may be inevitably reductive and unable to generalize in any meaningful way.

6.4 The Optimization Pitfall

Although F and E provide insight into representational quality, directly optimizing for them would defeat their purpose. As Kenneth Stanely, one of the co-authors of *Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis* put it in his incredible book *Why Greatness Cannot Be Planned: The Myth of the Objective,* "the greatest achievements become less likely when they are made objectives" (10). In the same spirit, representations that genuinely capture structure and meaning often emerge only when models are free from rigid geometric targets. If a model were trained to minimize F or E, it could learn to manipulate its geometry in trivial ways, such as compressing variance or orthogonalizing subspaces. However, these representations would not be meaningful in any

genuine way and totally contradict our search for intelligently well-factored and disentangled representations. Any attempt to optimize a model towards a lower *F* or *E* is foolish.

7 Conclusion

This paper introduced quantitative definitions for Fracture (*F)* and Entanglement (*E*), two complementary metrics designed to formalize the Fractured Entangled Representation (FER) hypothesis. Through theoretical grounding and empirical validation, it demonstrated that these measures capture distinct aspects of representational organization. While the metrics generally align with theoretical expectations, the divergence observed in the Picbreeder test and their partial interdependence together highlight the challenges of observing these structural factors in complex representational spaces. Future work should focus on extending these measures to better quantitatively capture functional and geometric organization, providing a more comprehensive understanding of how representations shape learning dynamics in modern neural networks.

8 Works Cited

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. “Representation Learning: A Review and New Perspectives.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, Aug. 2013, pp. 1798–1828. IEEE Xplore, doi:10.1109/TPAMI.2013.50.

Chen, Ricky T. Q., Xuechen Li, Roger Grosse, and David Duvenaud. “Isolating Sources of Disentanglement in Variational Autoencoders.” *Advances in Neural Information Processing Systems* (NeurIPS), vol. 32, 2018. arXiv:1802.04942.

Dziri, Nouha, et al. “Faith and Fate: Limits of Transformers on Compositionality.” *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Elhage, Nelson, et al. “Toy Models of Superposition.” *Transformer Circuits Thread*, 2022, https://transformer-circuits.pub/2022/toy_model/index.html.

Flesch, Timo, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. “Orthogonal Representations for Robust Context-Dependent Task Performance in Brains and Neural Networks.” Neuron, vol. 110, no. 8, 2022, pp. 1258–1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005.

Gao, Peiran, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy, and Surya Ganguli. “A Theory of Multineuronal Dimensionality, Dynamics and Measurement.” bioRxiv, 5 Nov. 2017, doi:10.1101/214262

Kahen, Gabriel. Measuring Fracture and Entanglement in Neural Networks. GitHub repository, 2025. https://github.com/Gabriel-Kahen/fer.

Kumar, A., Jeff Clune, Joel Lehman, and Kenneth O. Stanley. *Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis*. 2025. *arXiv*, arXiv:2505.11581.

Locatello, Francesco, et al. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations.” *arXiv preprint arXiv:1811.12359*, 2019.

Moreira, Gabriel & Marinho, Zita & Marques, Manuel & Costeira, Joao & Xiong, Chenyan. (2025). *Native Logical and Hierarchical Representations with Subspace Embeddings*. 10.48550/arXiv.2508.16687.

Raghu, Maithra, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. "SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability." arXiv preprint arXiv:1706.05806, 2017. https://arxiv.org/abs/1706.05806.

Secretan, Jimmy, Nicholas Beato, David B. D'Ambrosio, Adelein Rodriguez, Adam Campbell, Jeremiah T. Folsom-Kovarik, and Kenneth O. Stanley. "Picbreeder: A Case Study in Collaborative Evolutionary Exploration of Design Space." Evolutionary Computation, vol. 19, no. 3, MIT Press, 2011, pp. 373–403. https://doi.org/10.1162/EVCO_a_00030.

Stanley, K. O., & Lehman, J. A. (2015). Why Greatness Cannot Be Planned: The Myth of the Objective. Springer. ISBN 978-3-319-15524-1.

Yosinski, Jason, et al. "Understanding Neural Networks through Deep Visualization." *Deep Learning Workshop, 31st International Conference on Machine Learning (ICML)*, 2015.