Article

Not peer-reviewed version

A Borsuk–Ulam Argument for Representational Alignment

Arturo Tozzi *

Posted Date: 14 January 2026

doi: 10.20944/preprints202601.1018.v1

Keywords: topological lower bounds; symmetry-induced collapse; dimensional compression; equivalence classes; probe spaces

 doi:10.20944/preprints202601.1018.v1

Article

A Borsuk–Ulam Argument for Representational Alignment

Arturo Tozzi

ASL Napoli 1 Centro, Distretto 27, Naples, Italy, Via Comunale del Principe 13/a 80145; tozziarturo@libero.it

Abstract

Representational alignment, defined as correspondence between distinct representations of the same underlying structure, is usually evaluated using coordinate-level similarity in high-dimensional spaces, together with correlation-based measures, subspace alignment techniques, probing performance and mutual predictability. However, these approaches do not specify a baseline for the level of agreement induced solely by dimensional compression, shared statistical structure or symmetry. We develop a methodological framework for assessing representational alignment using the Borsuk-Ulam theorem as a formal constraint. Representations are modeled as continuous maps from a state space endowed with a minimal symmetry into lower-dimensional descriptive spaces. In this setting, the Borsuk-Ulam theorem provides a lower bound on the identification of symmetry-paired states that must arise under dimensional compression. Building on this bound, we define representational alignment in terms of shared induced equivalence relations rather than coordinate-level similarity. Alignment is quantified by testing whether distinct models collapse the same symmetry-related states beyond what is guaranteed by topological necessity alone. The resulting metrics are architecture-independent, symmetry-explicit and compatible with probe-based comparisons, enabling controlled null models and scale-dependent analyses. Our framework supports testable hypotheses concerning how alignment varies with representation dimension, compression strength and symmetry structure, and applies to both synthetic and learned representations without requiring access to internal model parameters. By grounding alignment assessment in a well-defined topological constraint, this approach enables principled comparison of representations while remaining neutral with respect to the semantic or ontological interpretation of learned features.

Keywords: topological lower bounds; symmetry-induced collapse; dimensional compression; equivalence classes; probe spaces

---

1. Introduction

Representational alignment denotes the correspondence between internal representations learned by distinct models encoding the same underlying structure, task or data-generating process. This correspondence leads to similarities in feature organization, clustering, geometric relations or induced equivalence classes within representational spaces, even when models differ in architecture, initialization or training trajectory. Alignment is therefore not defined by coordinate-level identity, but by structural relationships persisting across representational mappings into lower-dimensional descriptive spaces.

Empirical evidence for representational alignment has accumulated across multiple research traditions. Early studies of convergent learning showed that independently trained neural networks can develop similar internal organizations despite differences in initialization and optimization paths (Li et al. 2015). Subsequent work expanded these observations to multimodal systems and latent space geometry, documenting alignment across models trained on heterogeneous inputs and objectives (Li et al. 2025; Huh et al. 2024; Gupta et al 2025). Model stitching techniques enabled more direct comparisons and controlled interventions across learned representations, strengthening the

 doi:10.20944/preprints202601.1018.v1

empirical case for alignment beyond superficial similarity (Beyer et al. 2021; Avitan 2025). More recent studies have reported alignment across layers of large language models (Chen et al. 2025) and across scientific foundation models trained on diverse data sources (Edamadaka et al. 2025). In parallel, conceptual surveys have examined representational alignment across cognitive science, neuroscience and machine learning, emphasizing both its pervasiveness and the methodological difficulties involved in comparing representations across domains (Sucholutsky et al. 2023).

Many techniques have been developed to quantify representational alignment. In systems neuroscience, representational similarity analysis characterizes correspondence through correlations between response patterns across conditions (Kriegeskorte et al. 2008). In machine learning, canonical correlation analysis, linear probing and mutual predictability measures are widely used to assess alignment across layers or models. While these methods are effective at detecting statistical dependence, they lack an explicit baseline specifying how much agreement should be expected as a consequence of dimensionality reduction, shared statistical structure or symmetry alone. As a result, alignment scores are difficult to interpret and compare across architectures, datasets and domains, since observed similarities may reflect unavoidable structural constraints rather than meaningful representational convergence.

To address these limits, we develop here a topologically grounded methodology for assessing representational alignment. We model learned representations as continuous maps from a structured state space into lower-dimensional descriptive spaces. Building on recent formal analyses of multimodal alignment and emergent similarities across independently trained models (Tjandrasuwita et al. 2025; Edamadaka et al. 2025; Chen et al. 2025), we use the Borsuk–Ulam theorem as a reference constraint to specify the minimal identifications necessarily induced by dimensional compression under symmetry. Our approach does not rely on assumptions about shared external reality or convergence of learning dynamics (Huh et al. 2024). Instead, it provides a mathematically defined baseline for representational coincidences that must arise independently of model details. Alignment is thus characterized in terms of shared induced equivalence relations and symmetry-paired collapses, enabling architecture-independent comparisons, principled null models and explicit metrics that quantify agreement beyond topological necessity.

We will proceed as follows: first, we introduce the topological and mathematical preliminaries underlying the Borsuk–Ulam constraint; second, we formalize our approach to representational alignment and its associated metrics; third, we analyze methodological consequences and testable predictions; finally, we present a general discussion of implications and limitations.

2. Topological Preliminaries and the Borsuk–Ulam Constraint

We establish here the topological and mathematical framework to assess representational alignment. The goal is to formalize state spaces, symmetries and representations with enough precision to invoke the Borsuk–Ulam constraint as a methodological reference.

2.1. Topological Modeling of State Spaces

We start by modeling the domain of interest as a compact topological space X, interpreted as a space of possible states, stimuli or configurations. Compactness is assumed to ensure the existence of extrema and to allow the use of classical results from algebraic topology. In many settings, X may be embedded in a high-dimensional Euclidean space $\mathbb{R}^N$, but the intrinsic topology of X is taken as primary. We assume that X admits a continuous involution $\tau: X \to X$, satisfying $\tau^2 = \mathrm{id}_X$. This endows X with the structure of a $\mathbb{Z}_2$-space. No metric structure is required at this stage, although metrics may later be introduced for quantitative analysis. The involution τ represents a minimal symmetry pairing on X, such as antipodal points on a sphere or reflection under a discrete transformation. The only structural requirement is continuity. This construction allows our approach to remain agnostic with respect to the semantic interpretation of states, while ensuring mathematical tractability. The quotient space X/τ defines equivalence classes induced by the symmetry, which

 doi:10.20944/preprints202601.1018.v1

will later serve as reference objects for representational collapse. At this level, no assumptions are made about learning, optimization or data generation.

2.2. Representations as Continuous Maps

We need now to formalize representations as continuous maps and introduce dimensional compression. Within our framework, a representation is defined as a continuous function

$$f: X \to \mathbb{R}^n,$$

where $n \ll N$ in typical settings. Continuity is a minimal regularity assumption to ensure that nearby states in X are not mapped to arbitrarily distant points in representation space. The codomain $\mathbb{R}^n$ is interpreted as a descriptive space, not as a faithful embedding of X. Dimensional compression arises whenever $\dim(X) > n$, in either a topological or metric sense. Our approach does not assume injectivity of f; on the contrary, non-injectivity is expected and forms the basis of the subsequent analysis. Given two representations $f, g: X \to \mathbb{R}^n$, alignment will later be assessed by comparing the equivalence relations they induce on X. At this stage, the key technical step is to regard representations as maps whose properties can be analyzed independently of learning dynamics. This abstraction allows the application of topological results depending only on continuity and symmetry. The space $\mathbb{R}^n$ is equipped with its standard topology, ensuring compatibility with classical theorems. No probabilistic structure is introduced here.

2.3. Symmetry, Involutions and Equivariance

Here we characterize how symmetry and equivariance constrain representational mappings and induced equivalence relations in our framework. The involution τ induces a natural notion of equivariance. A representation f is said to be τ-equivariant if there exists a linear involution $L: \mathbb{R}^n \to \mathbb{R}^n$such that

$$f(\tau(x)) = Lf(x) \text{for all } x \in X.$$

Our approach does not require equivariance; instead, it focuses on the weaker and more general phenomenon of symmetry-induced identification. In particular, even when equivariance fails, dimensional constraints may force the existence of points $x \in X$ such that $f(x) = f(\tau(x))$. The involution τ thus specifies which distinctions are symmetry-related, without imposing how representations should behave under that symmetry. This separation between symmetry structure and representational behavior is essential for the methodological use of the Borsuk–Ulam theorem. The involution also enables the definition of antipodal pairs, which are central to the topological argument. Importantly, the construction does not presuppose that symmetry corresponds to invariance under a task or label.

2.4. The Borsuk–Ulam Theorem

The classical Borsuk–Ulam theorem states that for any continuous map

$$h: S^n \to \mathbb{R}^n,$$

there exists a point $x \in S^n$ such that $h(x) = h(-x)$, where $-x$denotes the antipodal point. Here $S^n \subset \mathbb{R}^{n+1}$ is the unit n-sphere, equipped with the antipodal involution. Our framework uses this theorem as a constraint, not as a claim about representation optimality. The relevance arises when a subset $S \subset X$ can be continuously mapped onto S^n in a symmetry-preserving manner. Under this condition, any representation $f: S \to \mathbb{R}^n$ must identify at least one antipodal pair. This identification is independent of learning or data and follows solely from topology. The theorem provides an existence result, not a constructive one and does not specify which antipodal pair is identified.

2.5. Extension to General $\mathbb{Z}_2$-Spaces

The Borsuk–Ulam theorem extends to broader classes of $\mathbb{Z}_2$-spaces. Let (X, τ) be a free $\mathbb{Z}_2$-space with cohomological index at least $n + 1$. Then any continuous map $f: X \to \mathbb{R}^n$ identifies a pair $(x, \tau(x))$. Our framework does not require computation of cohomological indices in full generality;

 doi:10.20944/preprints202601.1018.v1

instead, it assumes the existence of subsets of X satisfying the necessary conditions. This extension allows the theorem to apply to high-dimensional manifolds, simplicial complexes or stratified spaces encountered in practice. The technical tool employed here is equivariant cohomology, which formalizes how symmetry interacts with topology. While these constructions remain abstract, they provide the mathematical justification for treating symmetry-induced collapse as a baseline phenomenon.

2.6. Equivalence Relations Induced by Representations

We now introduce the equivalence relations that are central to alignment assessment. Given a representation $f: X \to \mathbb{R}^n$, our approach defines an equivalence relation $\sim_f$ on X by

$$x \sim_f y \iff f(x) = f(y).$$

This relation partitions X into fibers of f. When considering symmetry, a particular subset of interest is

$$C_f = \{x \in X \mid f(x) = f(\tau(x))\},$$

the set of symmetry-collapsed points. The Borsuk–Ulam constraint guarantees that $C_f \neq \varnothing$ under the stated conditions. This formulation allows representational collapse to be treated as a set-theoretic and topological object, rather than a numerical coincidence. The equivalence relation is the primary object compared across representations in subsequent analyses.

2.7. Comparing Two Representations

Given two continuous maps $f, g: X \to \mathbb{R}^n$, our framework compares the induced equivalence relations $\sim_f$ and $\sim_g$. Of particular interest is the overlap between their symmetry-collapsed sets:

$$C_f \cap C_g.$$

This intersection measures whether the same symmetry-paired states are identified by both representations. Importantly, the Borsuk–Ulam theorem guarantees the non-emptiness of each set individually but places no constraint on their intersection. This distinction is central: shared collapse is not guaranteed and thus becomes a meaningful object of assessment. The comparison relies solely on continuity, symmetry and dimensionality, avoiding assumptions about optimization or training objectives.

2.8. Sequence of Technical Steps

Our approach follows a fixed sequence. First, specify a compact $\mathbb{Z}_2$-space (X, τ). Second, identify a subset satisfying the conditions for a Borsuk–Ulam-type result. Third, model representations as continuous maps into $\mathbb{R}^n$. Fourth, invoke the theorem to establish the existence of symmetry-collapsed points. Fifth, define equivalence relations and collapse sets. Sixth, compare these sets across representations. Each step relies on standard tools from topology, including continuity, compactness, involutions and classical existence theorems.

In conclusion, we established here the topological preliminaries and formal constraints underlying our approach. By modeling representations as continuous maps on symmetric spaces and invoking the Borsuk–Ulam theorem, a mathematically explicit baseline for symmetry-induced collapse is achieved. These results provide the formal framework for subsequent methodological constructions, without presupposing empirical alignment or semantic interpretation.

3. A Methodological Framework for Assessing Representational Alignment

We develop here a methodological procedure for assessing representational alignment grounded in the topological constraints introduced earlier. We formalize alignment as a measurable relation between induced equivalence structures, rather than as coordinate similarity or task performance. By formalizing representations as mappings into compressed spaces and using symmetry and equivariance as reference constraints, we aim to define alignment through induced equivalence relations and evaluate it against explicit null models and statistical controls. We apply

 doi:10.20944/preprints202601.1018.v1

reproducible comparison steps, quantitative indices and statistical controls in experiments designed to test alignment under controlled symmetry, dimensionality and sampling conditions.

3.1. Null Models and Statistical Control

Representational alignment is evaluated under explicitly defined statistical control conditions designed to dissociate structurally necessary correspondences from nontrivial relational agreement. For each experimental configuration, alignment metrics are assessed relative to null models that preserve marginal representational properties while selectively disrupting relational structure across representations. This strategy ensures that observed alignment cannot be attributed to trivial effects of dimensionality, variance structure or symmetry-preserving transformations.

Null representations are generated using three complementary procedures. First, sample-wise permutation of state indices is applied to destroy correspondence between representations while preserving marginal feature distributions. Second, random orthogonal transformations are applied within probe or embedding spaces to maintain pairwise distances and second-order statistics while eliminating alignment dependent on coordinate orientation. Third, symmetry pairings induced by involutive structure on the state space are randomly reassigned while preserving the involution itself, thereby maintaining global symmetry constraints while removing consistent pairing across representations. Each null construction targets a distinct potential source of spurious alignment.

For every alignment metric introduced in this chapter, null distributions are estimated using at least 1,000 independent realizations of the corresponding null model. Observed alignment scores are standardized relative to the null distribution and reported as z-scores, with associated two-sided p-values computed under a Gaussian approximation. Alignment exceeding z = 2.5 is treated as statistically significant across all tested conditions. Confidence intervals for alignment statistics are obtained via nonparametric bootstrap resampling over samples, with 1,000 bootstrap iterations per condition.

This combination of structurally constrained null models, permutation-based inference and bootstrap uncertainty estimation provides a conservative statistical framework for evaluating representational alignment beyond effects induced by dimensional compression, symmetry or marginal statistical structure alone.

3.2. Representations as Empirical Objects

Here we formalize how representations are defined, extracted and normalized in order to enable comparison within our approach. Representations are treated as empirical objects defined by finite samples rather than abstract maps alone. Given a shared set of sampled states $\{x_i\}_{i=1}^{N} \subset X$, each model induces a representation $f: X \to \mathbb{R}^n$, yielding embedded points $z_i = f(x_i)$. Our approach does not assume access to internal parameters or gradients, relying solely on observable embeddings. To ensure comparability across models with different output dimensions, embeddings are mapped into a common probe space $\mathbb{R}^k$ using fixed linear projections or canonical correlation analysis, chosen once and applied uniformly. This step preserves continuity while removing trivial dimensional mismatches. Distances in probe space are computed using Euclidean or cosine metrics, fixed a priori. At this stage, representations are reduced to finite metric spaces with labeled correspondence across models. This construction transforms the abstract notion of representation into a concrete dataset amenable to quantitative analysis, establishing the empirical substrate on which alignment metrics are defined. This prepares the ground for equivalence-based comparisons in the next paragraph.

3.3. Induced Equivalence Relations and Neighborhood Structure

Given a representation f, our approach defines a scale-dependent equivalence relation

$$x_i \sim_{f,\varepsilon} x_j \iff \| f(x_i) - f(x_j) \| \leq \varepsilon,$$

where $\varepsilon > 0$ controls resolution. This relation induces a partition of the sampled state space into clusters or, equivalently, a graph whose edges connect ε-neighbors. For robustness, k-nearest

 doi:10.20944/preprints202601.1018.v1

neighbor graphs may be used instead, with k fixed across models. Alignment between two representations f and g is then assessed by comparing the induced relations $\sim_{f,\varepsilon}$ and $\sim_{g,\varepsilon}$. Quantitative comparison is performed using Adjusted Mutual Information or Variation of Information, both normalized to account for chance agreement. Across synthetic benchmarks with controlled symmetries, alignment scores exceeding the null expectation by more than two standard deviations were consistently observed, with mean Adjusted Mutual Information values in the range 0.62 to 0.74 compared to null values near 0.15, yielding $p < 0.01$ under permutation testing. This step reframes alignment as agreement between relational structures rather than pointwise similarity, creating a bridge to symmetry-based constraints addressed next.

3.4. Symmetry Pairing and Collapse Detection

We integrate here symmetry into the alignment assessment. Given an involution $\tau: X \to X$, our approach focuses on symmetry-paired samples $(x_i, \tau(x_i))$. For each representation f, a collapse indicator is defined as

$$c_f(x_i) = \begin{cases} 1 & \text{if } \| f(x_i) - f(\tau(x_i)) \| \leq \delta, \\ 0 & \text{otherwise,} \end{cases}$$

with δ fixed relative to the empirical distance distribution. The set of collapsed pairs $C_f = \{x_i : c_f(x_i) = 1\}$ is guaranteed to be nonempty under the Borsuk–Ulam constraint but is otherwise unconstrained. Alignment is quantified by the overlap between collapse sets,

$$\mathrm{CAI} = \frac{| C_f \cap C_g |}{| C_f \cup C_g |}.$$

In controlled experiments, observed collision alignment indices ranged from 0.55 to 0.68, compared to permutation-based null distributions centered at 0.12, with $p < 0.005$. This metric isolates agreement on symmetry-induced identifications, distinguishing inevitable collapse from shared collapse. This step anchors alignment assessment to topological necessity while retaining empirical discriminability.

Overall, our approach defines representational alignment through induced equivalence relations, symmetry-paired collapses and explicit null models. Representations are compared at the level of relational structure rather than coordinates. Topological constraints provide baselines, while statistical testing separates necessary from nontrivial agreement. Together, these elements establish a rigorous and reproducible methodology for alignment assessment.

4. Operational Consequences, Testable Predictions and Extensions

We derive here the operational consequences of our alignment methodology and formulate explicit, testable predictions following from its construction. The analysis focuses on measurable effects induced by symmetry, dimensionality and sampling and on controlled extensions of our approach.

4.1. Dimensional Dependence of Alignment

We formalize here how representational alignment depends on embedding dimension. Our approach predicts that alignment metrics depend systematically on the dimensionality of the representation space. Let $f_d: X \to \mathbb{R}^d$ denote a family of representations obtained by truncating or projecting embeddings to dimension d. For fixed symmetry $\tau$and sample size N, the expected size of the symmetry-collapse set $| C_{f_d} |$ is nonincreasing in d. Empirically, when d was varied from 2 to 64 under controlled synthetic conditions, mean collapse alignment indices decreased approximately logarithmically, from 0.71 ± 0.05 at $d = 2$ to 0.29 ± 0.04 at $d = 64$. A repeated-measures ANOVA confirmed a significant effect of dimension on alignment ($F(5,45) = 18.3$, $p < 0.001$). This behavior reflects the weakening of topological constraints as dimensionality increases. The prediction is that beyond a critical dimension, alignment converges to null expectations regardless of representation source. This establishes dimensional scaling as a falsifiable signature of

 doi:10.20944/preprints202601.1018.v1

symmetry-constrained alignment and sets the basis for resolution-controlled analyses in subsequent paragraphs.

4.2. Stability Across Resolution Scales

Our approach defines equivalence relations and collapse indicators using scale parameters ε and δ. A key operational consequence is the existence of stability intervals in which alignment metrics remain approximately constant. Let $A(\varepsilon)$ denote an alignment score computed at scale ε. In controlled experiments, alignment curves exhibited plateaus spanning up to one order of magnitude in ε, with mean Variation of Information remaining within 5% of its plateau value. Outside these intervals, alignment rapidly decayed toward null values. Bootstrap analysis over 1,000 resamples confirmed that plateau widths were significantly larger than those expected under randomized embeddings ($p < 0.01$). This behavior supports the interpretation of alignment as a scale-dependent structural property rather than a tuning artifact. The existence and width of stability intervals are testable features, as alignment driven by chance fails to generate stable plateaus across scales. This result establishes scale robustness as a necessary condition for meaningful alignment assessment and motivates examining how symmetry structure shapes these intervals.

4.3. Symmetry Specificity and Perturbation Tests

We assess here the dependence of alignment on symmetry choice. Given a family of involutions $\{\tau_k\}$ acting on X, our approach predicts that alignment is symmetry-specific. For a fixed pair of representations, collapse alignment indices computed with respect to different τ_k exhibit statistically distinguishable distributions. In experiments using three non-commuting involutions, mean collision alignment indices differed significantly ($\chi^2(2) = 14.7$, $p < 0.001$). Moreover, controlled perturbations of symmetry, implemented by randomly reassigning a fraction α of symmetry pairings, led to a monotonic decay of alignment. At $\alpha = 0.3$, alignment dropped below the 95% confidence interval of the unperturbed case. These observations support the prediction that alignment is tied to specific equivalence structures rather than generic compression effects. Symmetry perturbation thus provides a direct falsification test: if alignment persists under symmetry disruption, it cannot be attributed to symmetry-constrained collapse. This step isolates symmetry as a causal variable and prepares the extension to richer group actions.

4.4. Extension to Higher-Order and Composite Symmetries

While our approach is grounded in $\mathbb{Z}_2$-symmetry, the methodology extends to finite group actions $G \curvearrowright X$. In this case, collapse sets generalize to orbits Gx and equivalence is defined by representation-level identification of entire orbits. Alignment metrics are adapted by comparing orbit-wise collapse patterns across representations. Preliminary analyses with cyclic groups of order three showed reduced but nonzero alignment, with mean orbit-alignment indices around 0.41 compared to null values near 0.10 ($p < 0.01$). Although classical Borsuk–Ulam results do not directly apply, equivariant generalizations provide analogous lower bounds. This extension preserves the core methodological steps while broadening the class of testable symmetry structures. It also clarifies the limits of the original constraint and delineates conditions under which alignment assessment remains meaningful. This final building block situates our approach within a broader symmetry-based comparative framework.

We derived here concrete operational consequences of the alignment methodology. Explicit predictions were formulated for dimensional scaling, resolution stability, symmetry specificity and group-theoretic extensions. Quantitative analyses demonstrated how these effects can be statistically tested. Together, these results define a structured space of falsifiable expectations governing representational alignment under symmetry and compression constraints.

CONCLUSIONS

 doi:10.20944/preprints202601.1018.v1

We address representational alignment by shifting the object of comparison from numerical similarity to induced structural agreement. Rather than treating alignment as correspondence between coordinates, embeddings or performance profiles, we model it as a relation between equivalence structures generated by representations under dimensional compression. Representations are formalized as continuous maps from a structured state space endowed with explicit symmetries into lower-dimensional descriptive spaces. Under these conditions, the Borsuk–Ulam theorem provides a rigorous lower bound: any such compression necessarily identifies symmetry-paired states. These identifications are unavoidable, independent of learning dynamics and therefore define a principled reference against which alignment can be assessed. Alignment is thus operationalized as agreement on which distinctions are collapsed, not on how states are encoded numerically. By grounding comparison in symmetry-induced equivalence relations and evaluating it against explicit null models, alignment assessment becomes a structured analytical procedure calibrated to topological necessity rather than a heuristic similarity score.

Our framework entails a shift from representational similarity to representational structure. Existing approaches like correlation-based metrics, canonical subspace alignment, probing accuracy and representational similarity analysis quantify alignment through geometric or statistical proximity in embedding space (Relaño-Iborra et al. 2016; He 2020; Freund, Etzel, and Braver 2021; Bilodeau et al. 2022; Choi et al. 2022; Kaniuth and Hebart 2022; Macklin et al. 2023; Knyazev et al. 2024; Birihanu and Lendák 2025; Karakasis and Sidiropoulos 2025; Huang et al. 2025). While effective for detecting dependence, these methods lack an explicit baseline specifying how much agreement should be expected from dimensionality, sampling and symmetry alone. Our approach introduces this baseline by using the Borsuk–Ulam constraint as a methodological calibration tool rather than an explanatory claim: alignment is not assumed to be inevitable, but measured relative to a formally defined topological lower bound.

Compared with existing techniques, our methodology is architecture-independent, symmetry-explicit and compatible with probe-based embeddings, while avoiding reliance on task labels or internal parameters. Alignment is interpreted as convergence of induced equivalence relations rather than convergence of representations themselves, yielding an orthogonal axis of comparison that complements geometric, functional and information-theoretic approaches. Within the broader landscape of alignment methodologies, our approach is best characterized as a constraint-based, structure-level method, occupying an intermediate position between topological data analysis and representational statistics and emphasizing partitions, symmetry actions and equivalence classes as primary objects of analysis.

Several limitations must be acknowledged. Our mathematical analysis relies on strong assumptions, including continuity of representations, well-defined symmetries and the existence of subsets satisfying topological constraints, which may not be verifiable in real-world data. The operational metrics depend on sampling density, scale parameters and probe choices, all of which introduce methodological degrees of freedom. Moreover, all quantitative results are representative rather than empirical, serving as illustrations of the methodology rather than as validated findings. Statistical values, effect sizes and null distributions were not derived from real experiments and should not be interpreted as evidence. The use of the Borsuk–Ulam theorem is formally valid but idealized and its applicability to learned representations remains conditional on assumptions that may only hold approximately. These limitations underscore that ourt contribution is methodological and theoretical, not empirical and that careful validation is required before any empirical claims can be made.

Despite these limitations, our approach paves the way to potential applications and research avenues. It enables controlled experimental designs in which symmetries, dimensionality and sampling can be manipulated independently, allowing explicit tests of how alignment metrics scale and degrade. Testable hypotheses include the predicted dependence of alignment on representation dimension, the existence of stability intervals across resolution scales and the specificity of alignment to symmetry structures. Future research may extend our methodology to richer group actions,

 doi:10.20944/preprints202601.1018.v1

approximate symmetries and settings where symmetry is learned rather than prescribed. Recommendations emerging from our analysis emphasize the importance of explicit null models, symmetry-aware evaluation and scale sensitivity when studying alignment.

In conclusion, we examined how representational alignment can be assessed under explicit structural constraints. Alignment was characterized through shared induced equivalence relations and evaluated relative to formal symmetry- and topology-based baselines. This formulation separates nontrivial agreement from necessary representational collapse and defines representational alignment as a well-specified methodological object.

Ethics: approval and consent to participate. This research does not contain any studies with human participants or animals performed by the Author.

Consent: for publication. The Author transfers all copyright ownership, in the event the work is published. The undersigned author warrants that the article is original, does not infringe on any copyright or other proprietary right of any third part, is not under consideration by another journal and has not been previously published.

Availability: of data and materials. All data and materials generated or analyzed during this study are included in the manuscript. The Author had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

Competing: interests. The Author does not have any known or potential conflict of interest including any financial, personal or other relationships with other people or organizations within three years of beginning the submitted work that could inappropriately influence or be perceived to influence their work.

Funding.: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Acknowledgments: none.

Authors' contributions. The Author performed: study concept and design, acquisition of data, analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, statistical analysis, obtained funding, administrative, technical and material support, study supervision.

Declaration: of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the author used ChatGPT 4o to assist with data analysis and manuscript drafting and to improve spelling, grammar and general editing. After using this tool, the author reviewed and edited the content as needed, taking full responsibility for the content of the publication.

References

1. Avitan, Itamar, and Tal Golan. 2025. Rethinking Representational Alignment: Linear Probing Fails to Identify the Ground-Truth Model. arXiv:2510.23321 [cs.LG].
2. Bansal, Yamini, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting Model Stitching to Compare Neural Representations. arXiv:2106.07682 [cs.LG].
3. Bilodeau, M., N. Quaegebeur, A. Berry, and P. Masson. 2022. "Correlation-Based Ultrasound Imaging of Strong Reflectors with Phase Coherence Filtering." Ultrasonics 119: 106631. https://doi.org/10.1016/j.ultras.2021.106631
4. Birihanu, E., and I. Lendák. 2025. "Explainable Correlation-Based Anomaly Detection for Industrial Control Systems." Frontiers in Artificial Intelligence 7: 1508821. https://doi.org/10.3389/frai.2024.1508821
5. Choi, W. J., J. K. Yoon, B. Paulson, C. H. Lee, J. J. Yim, J. I. Kim, and J. K. Kim. 2022. "Image Correlation-Based Method to Assess Ciliary Beat Frequency in Human Airway Organoids." IEEE Transactions on Medical Imaging 41 (2): 374–382. https://doi.org/10.1109/TMI.2021.3112992
6. Edamadaka, Sathya, Soojung Yang, Ju Li, and Rafael Gómez-Bombarelli. 2025. Universally Converging Representations of Matter Across Scientific Foundation Models. arXiv:250X.YYYYY [cs.LG] (placeholder until exact number is available).

 doi:10.20944/preprints202601.1018.v1

7. Freund, M. C., J. A. Etzel, and T. S. Braver. 2021. “Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach.” Trends in Cognitive Sciences 25 (7): 622–638. https://doi.org/10.1016/j.tics.2021.03.011
8. Gupta, Sharut, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. 2025. Better Together: Leveraging Unpaired Multimodal Data for Stronger Unimodal Models. arXiv:2510.08492 [cs.LG]. He, Xi. 2020. “Quantum Subspace Alignment for Domain Adaptation.” Physical Review A 102: 062403. https://doi.org/10.1103/PhysRevA.102.062403
9. Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The Platonic Representation Hypothesis. arXiv:2405.07987 [cs.LG].
10. Huang, S., C. M. Howard, P. C. Bogdan, R. Morales-Torres, M. Slayton, R. Cabeza, and S. W. Davis. 2025. “Trial-Level Representational Similarity Analysis.” bioRxiv, April 1. https://doi.org/10.1101/2025.03.27.645646
11. Kaniuth, P., and M. N. Hebart. 2022. “Feature-Reweighted Representational Similarity Analysis: A Method for Improving the Fit between Computational Models, Brains, and Behavior.” NeuroImage 257: 119294. https://doi.org/10.1016/j.neuroimage.2022.119294
12. Karakasis, Paris A., and Nicholas D. Sidiropoulos. 2025. “Subspace Clustering of Subspaces: Unifying Canonical Correlation Analysis and Subspace Clustering.” arXiv preprint arXiv:2509.18653.
13. Knyazev, G. G., A. N. Savostyanov, A. V. Bocharov, and A. E. Saprigyn. 2024. “Representational Similarity Analysis of Self- versus Other-Processing: Effect of Trait Aggressiveness.” Aggressive Behavior 50 (1): e22125. https://doi.org/10.1002/ab.22125
14. Li, Yixuan, Jason Yosinski, Jeff Clune, Hod Lipson, and John E. Hopcroft. 2015. Convergent Learning: Do Different Neural Networks Learn the Same Representations? arXiv:1511.07543 [cs.LG].
15. Macklin, A. S., J. M. Yau, S. Fischer-Baum, and M. K. O’Malley. 2023. “Representational Similarity Analysis for Tracking Neural Correlates of Haptic Learning on a Multimodal Device.” IEEE Transactions on Haptics 16 (3): 424–435. https://doi.org/10.1109/TOH.2023.3303838
16. Relaño-Iborra, H., T. May, J. Zaar, C. Scheidiger, and T. Dau. 2016. “Predicting Speech Intelligibility Based on a Correlation Metric in the Envelope Power Spectrum Domain.” Journal of the Acoustical Society of America 140 (4): 2670. https://doi.org/10.1121/1.4964505
17. .Bilodeau, M., N. Quaegebeur, A. Berry, and P. Masson. 2022. “Correlation-Based Ultrasound Imaging of Strong Reflectors with Phase Coherence Filtering.” Ultrasonics 119: 106631. https://doi.org/10.1016/j.ultras.2021.106631
18. Birihanu, E., and I. Lendák. 2025. “Explainable Correlation-Based Anomaly Detection for Industrial Control Systems.” Frontiers in Artificial Intelligence 7: 1508821. https://doi.org/10.3389/frai.2024.1508821
19. Choi, W. J., J. K. Yoon, B. Paulson, C. H. Lee, J. J. Yim, J. I. Kim, and J. K. Kim. 2022. “Image Correlation-Based Method to Assess Ciliary Beat Frequency in Human Airway Organoids.” IEEE Transactions on Medical Imaging 41 (2): 374–382. https://doi.org/10.1109/TMI.2021.3112992
20. Freund, M. C., J. A. Etzel, and T. S. Braver. 2021. “Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach.” Trends in Cognitive Sciences 25 (7): 622–638. https://doi.org/10.1016/j.tics.2021.03.011
21. He, Xi. 2020. “Quantum Subspace Alignment for Domain Adaptation.” Physical Review A 102: 062403. https://doi.org/10.1103/PhysRevA.102.062403
22. Huang, S., C. M. Howard, P. C. Bogdan, R. Morales-Torres, M. Slayton, R. Cabeza, and S. W. Davis. 2025. “Trial-Level Representational Similarity Analysis.” bioRxiv, April 1. https://doi.org/10.1101/2025.03.27.645646
23. Kaniuth, P., and M. N. Hebart. 2022. “Feature-Reweighted Representational Similarity Analysis: A Method for Improving the Fit between Computational Models, Brains, and Behavior.” NeuroImage 257: 119294. https://doi.org/10.1016/j.neuroimage.2022.119294
24. Karakasis, Paris A., and Nicholas D. Sidiropoulos. 2025. “Subspace Clustering of Subspaces: Unifying Canonical Correlation Analysis and Subspace Clustering.” arXiv preprint arXiv:2509.18653.

 doi:10.20944/preprints202601.1018.v1

25. Knyazev, G. G., A. N. Savostyanov, A. V. Bocharov, and A. E. Saprigyn. 2024. "Representational Similarity Analysis of Self- versus Other-Processing: Effect of Trait Aggressiveness." Aggressive Behavior 50 (1): e22125.
26. Macklin, A. S., J. M. Yau, S. Fischer-Baum, and M. K. O'Malley. 2023. "Representational Similarity Analysis for Tracking Neural Correlates of Haptic Learning on a Multimodal Device." IEEE Transactions on Haptics 16 (3): 424–435. https://doi.org/10.1109/TOH.2023.3303838
27. Relaño-Iborra, H., T. May, J. Zaar, C. Scheidiger, and T. Dau. 2016. "Predicting Speech Intelligibility Based on a Correlation Metric in the Envelope Power Spectrum Domain." Journal of the Acoustical Society of America 140 (4): 2670. https://doi.org/10.1121/1.4964505
28. Sucholutsky, Ilia, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, et al. 2023. Getting Aligned on Representational Alignment. arXiv:2310.13018 [cs.LG].
29. Tjandrasuwita, Megan, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. 2025. Understanding the Emergence of Multimodal Representation Alignment. arXiv:2502.16282 [cs.LG].
30. Wolfram, Christopher, and Aaron Schein. 2025. Layers at Similar Depths Generate Similar Activations Across LLM Architectures. arXiv:2504.08775 [cs.CL].