Geometric invariance predicts sensor blind spots: a pre-registered comparison of 23 methods on simulated NV-diamond thermometry

Elliot Tower

Independent Researcher
Correspondence: elliot@elliottower.ai

Novelty statement: This paper provides a systematic, pre-registered characterization of how 23 geometric analysis methods respond to five types of sensor noise in NV-diamond thermometry. The resulting sensitivity map is a practical guide for matching analysis methods to dominant noise types in multi-sensor deployments, and identifies device calibration variation as the axis that discriminates between method families.

Abstract

Device-to-device calibration variation is the dominant noise source in multi-sensor NV-diamond thermometry, yet we show that most geometric analysis methods are structurally blind to it. Through a pre-registered experiment comparing 23 geometric methods and 7 scalar baselines across five decoherence axes (150,000 evaluations, 100 seeds, bootstrap CIs), we characterize each method's noise-axis sensitivity profile and identify the pattern: any method invariant to translations in feature space—including Grassmannian geodesic distance, CKA, Berry phase, and Procrustes distance—cannot detect additive per-device calibration offsets. The data confirm this without exception: every translation-invariant method achieves $|\rho| < 0.15$ on device variation, while every non-invariant method exceeds $|\rho| > 0.5$. The resulting sensitivity map reveals that 25 of 30 methods track composite coherence degradation at $|\rho| > 0.5$, but only 6 respond to device variation—the noise axis with the simplest physics is the one that discriminates. The principle also explains a pre-registered self-refutation: Grassmannian geodesic distance, predicted to outperform bracket norms, instead underperforms on both tested axes ($\Delta\rho = -0.39$ and -0.62, both $|d| > 1.4$). For practitioners selecting analysis methods for multi-sensor deployments where calibration offsets dominate, simple effect-size statistics detect what geometric methods miss.

Keywords: geometric methods, quantum sensors, NV-diamond, decoherence, pre-registration, Grassmannian, bracket norm, persistent homology, cross-sensor robustness

1 Introduction

Multi-sensor measurement systems require methods to detect and characterize inter-device disagreement—a core problem in metrology [1, 2]. Geometric methods offer a rich toolkit for this task: Grassmannian geodesic distance between PCA subspaces [3–5], bracket norms measuring inter-group variability in Lie-algebraic terms [6], sheaf cohomology detecting device-level inconsistency [7–9], discrete curvatures characterizing graph structure [10, 11], persistent homology tracking topological

features across scales [12–14], alignment metrics such as CKA [15] and Procrustes distance [16], optimal-transport distances [17, 18], and quantum information metrics (QFI, Berry phase, von Neumann entropy) measuring sensitivity and mixedness of quantum states [19–22].

Each method encodes a different geometric property. The question this paper addresses is empirical: which of these properties are robust to the noise sources that dominate real sensor deployments, and which degrade or become uninformative as decoherence increases?

We answer this for NV-diamond intracellular thermometry, a quantum sensing modality in which nitrogen-vacancy centers in diamond detect temperature via the zero-field splitting shift ($dD/dT = -74$ kHz/K) [23–25]. In practice, per-nanodiamond strain variation produces $\sim$170 kHz standard deviation in the zero-field splitting parameter [26], corresponding to $\sim$2.3 K equivalent calibration offset—making device-to-device variation the dominant noise source in multi-sensor deployments [27, 28]. The simulation sweeps five physically motivated decoherence axes: T2 coherence time degradation (a composite axis affecting T2, linewidth, and fluorescence contrast simultaneously—reflecting the physical coupling through spin-bath interactions, where shorter T2 broadens the ODMR line and reduces contrast), surface noise from diamond defects, device-to-device calibration variation, temperature drift, and ensemble size. Each axis isolates a distinct noise mechanism, enabling us to map which geometric methods respond to which noise type.

The experiment is pre-registered [29, 30]: all 23 methods, 7 baselines, hypotheses, and statistical analysis plan were frozen under commit SHA `8401d54` before any data was generated. Iterative revision identified and corrected five structural problems (insufficient replication, missing multiple-testing correction, composite axis confounding, broken TDA implementations, unfair cross-method comparison) before the freeze. The full pre-registration, deviation log, and raw results (150,000 method evaluations) are publicly available.[1]

Contributions.

1. A **method–axis sensitivity map** (table 3) characterizing which geometric families respond to which sensor noise types across 23 methods and 5 decoherence axes. The axis that discriminates between methods—device calibration variation—has the simplest physics (additive offset): 25 of 30 methods track the composite coherence axis at $|\rho| > 0.5$, but only 6 do so on device variation. The breadth of coverage (8 geometric families, 7 baselines, 150,000 evaluations) is necessary because the sensitivity map cannot be derived analytically—while the invariance principle below predicts which methods fail on device variation, it does not predict relative performance across the other four axes, which requires empirical measurement.
2. An **invariance-based explanation** for the sensitivity map's structure: any method invariant to translations in feature space—$f(X + \mathbf{1}c^\top) = f(X)$—is blind to additive device calibration variation. That centering discards location information is elementary; the applied observation is that this property partitions geometric methods into device-aware and device-blind families in a way that is predictive, falsifiable, and confirmed without exception across all 23 methods tested.
3. A **pre-registered self-refutation**: Grassmannian geodesic distance, predicted to outperform bracket norms, instead underperforms ($\Delta\rho = -0.39$ and -0.62, both $|d| > 1.4$), explained by the invariance property above.

Prior work has evaluated individual geometric families on specific tasks; the contribution here is the cross-family sensitivity map and the identification of noise-geometry matching as the operational challenge for multi-sensor deployments.

[1] `https://github.com/elliottower/geometric-sensor-blindspots`

Table 1: Decoherence axes. **Composite axes** affect multiple features simultaneously; results on these axes must be interpreted with the coupling in mind. σ_{device} is the cleanest axis (additive offset only) and serves as the primary discriminator between methods.

Axis	Parameter	Range	Composite?	Physics
T2 degradation	α_{T2}	$[0, 1]$	**Yes**	T2 + linewidth + contrast
Surface noise	σ_{surface}	$[0, 0.5]$	Yes	T2 noise + linewidth variance
Device variation	σ_{device}	$[0, 0.1]$	No	Additive temperature offset
Temperature drift	σ_{temp}	$[0, 2.0]$ K	No	Per-sample noise
Ensemble size	n_{centers}	$\{5 \dots 50\}$	No	Dimensionality change

Table 2: Methods by family. Method 24 (Chern number) was cut pre-registration due to a gauge-invariance error; method 15 is restricted to H0 (connected components) after the H1 heuristic was found to be incorrect.

Family	Methods
Bracket (5)	Raw, corrected, transport-stable, phenotype-projected, localized bracket norms
Curvature (2)	Ollivier-Ricci (metric) [10], Forman-Ricci (combinatorial) [11]
Grassmannian (2)	Geodesic distance, loop holonomy
Sheaf (3)	Cochran's Q [31] via sheaf H^1, per-edge Q, persistent sheaf cohomology
Topology (2)	Persistent H0 [13] (union-find), Wasserstein persistence distance
Quantum-inspired (4)	QFI flatness [22], Berry phase proxy [20], von Neumann entropy [21], fidelity proxy
Alignment (2)	Linear CKA [15], Procrustes distance [16]
Transport (2)	TransportKit fusion, Wasserstein-1 distance [17]
Baselines (7): t-test, Cohen's d [32], LASSO [33], RF [34], logistic, IRM penalty [35], DRO [36]	

2 Simulation oracle

The NV-diamond simulation generates multi-device datasets with controlled decoherence. Each dataset contains $n = 200$ samples per class (healthy tissue at 310.0 K, tumor tissue at 311.5 K) measured across $k = 3$ devices, yielding features derived from the ODMR (optically detected magnetic resonance) spectrum: resonance frequency, linewidth, contrast, T2 coherence time, and signal-to-noise ratio.

Five decoherence axes parameterize noise (table 1). The experimental design sweeps one axis at a time (10 levels each) while holding the others at baseline (zero decoherence), with 100 independent seeds per condition.

3 Methods

All 23 geometric methods and 7 baselines take identical input $(X, y, \text{device\_ids})$ and return a scalar score. Methods are grouped by geometric family (table 2).

3.1 Statistical framework

All cross-method comparisons use Spearman rank correlation [37] between a method's score and the decoherence level, computed per seed and summarized as the median ρ across 100 seeds with bootstrap 95% CI [38] (10,000 resamples). The median is robust to outlier seeds; the mean ρ differs by < 0.02 on all cells except where threshold effects produce bimodal distributions (Ollivier-Ricci on α_{T2}). This normalizes across methods' different score scales by reducing each to a monotonic association strength on $[-1, +1]$. When all 100 seeds yield $\rho = \pm 1.0$, the bootstrap CI degenerates to a point mass; these cells represent unambiguous results rather than estimation artifacts.

Directional comparisons compute the per-seed paired difference $\rho_A - \rho_B$ and report the bootstrap CI on the median difference. If the CI excludes zero, the comparison is significant.

Confirmatory vs. exploratory. One correctness gate (H4d: sheaf H^1 = Cochran's Q, analytic identity, no α needed) runs first. The pre-registration froze four confirmatory tests at $\alpha = 0.05/4 = 0.0125$: H3a on α_{T2}, H3a on σ_{device} (two tests for the same hypothesis on different axes), H6d, and H8a. Post-execution, H8a was voided as untestable under the Spearman framework (direct cross-scale score comparison is incompatible with the normalization; see Section 4.2), reducing the denominator from 4 to 3 and yielding a final $\alpha = 0.05/3 = 0.0167$. This change is logged in the deviation record. All other hypotheses are exploratory (reported without multiplicity correction; see Benjamini and Hochberg [39] for the distinction between confirmatory and exploratory inference).

4 Results

4.1 Correctness gate

H4d passes: the sheaf H^1 implementation reproduces Cochran's Q [31] to machine precision on synthetic data with known heterogeneity (5 devices, 3 with effect, 2 without; $Q = 168.1$, $p < 10^{-6}$) and returns $Q \approx 0$ on homogeneous data ($Q = 0.28$, $p = 0.99$).

4.2 Confirmatory hypotheses

H3a: Grassmannian geodesic vs. bracket norm (REFUTED). On α_{T2} (composite axis), the median paired difference $\rho_{\text{geodesic}} - \rho_{\text{bracket}}$ is -0.39 (95% CI: $[-0.44, -0.34]$; paired Cohen's $d = -1.87$). On σ_{device} (clean axis), the difference is -0.62 ($[-0.81, -0.52]$; $d = -1.48$). Both CIs exclude zero in the wrong direction with large effect sizes: bracket norm strictly outperforms Grassmannian geodesic.

The level-by-level data reveals the mechanism (figure 2). On σ_{device}, mean geodesic scores are flat between 1.94 and 1.95 across all decoherence levels, with no monotonic trend ($\rho = -0.05$, CI includes zero). Device calibration variation adds a constant temperature offset to all samples from a given device. This shifts both class means equally, preserving the principal angle between class subspaces [40, 41]. The Grassmannian geodesic is geometrically correct in reporting "no subspace change"—but this invariance makes it blind to the dominant noise source in multi-device deployments.

Bracket norms, by contrast, respond to both subspace rotation and scale changes in inter-group variability, achieving $\rho = +0.65$ on σ_{device}.

Table 3: Method–axis sensitivity map. Each cell shows the median Spearman ρ(score, decoherence level) across 100 seeds. For each family, we show the method with the highest count of axes at $|\rho| > 0.8$; families with two qualitatively different profiles (e.g., raw vs. transport-stable bracket) show both. Bold indicates $|\rho| > 0.8$. σ_{dev} is the discriminating axis: most methods are near zero. †Classification baselines return AUC = 1.0 (constant) on σ_{surf} and σ_{dev}; Spearman ρ is undefined.

Family	Best method	α_{T2}	σ_{surf}	σ_{dev}	σ_{temp}	n_{cent}	>0.8
Bracket	bracket_norm_raw	**+1.00**	**+0.90**	+0.65	**+0.99**	**+0.95**	4/5
Bracket	transport_stable	**+1.00**	**+0.93**	−0.01	**+1.00**	**+0.99**	4/5
Topology	persistent_h0	**+1.00**	**+1.00**	+0.44	**+0.98**	**+1.00**	4/5
Alignment	CKA	**−1.00**	**−0.99**	+0.01	**−1.00**	**+1.00**	4/5
Quantum-insp.	fidelity_proxy	**−0.96**	**−0.99**	+0.01	**+1.00**	**−1.00**	4/5
Alignment	Procrustes	**+1.00**	**+1.00**	+0.01	**+1.00**	**−1.00**	4/5
Sheaf (persistent)	persistent_sheaf	**−1.00**	**−1.00**	+0.02	**−0.98**	−0.13	3/5
Grassmannian	geodesic	+0.61	−0.52	−0.05	−0.07	**+0.88**	1/5
Curvature	Ollivier-Ricci	−0.61	**+0.88**	+0.13	−0.48	**+0.93**	2/5
Sheaf (pointwise)	sheaf_h1	−0.15	−0.17	−0.06	+0.43	−0.59	0/5
Classification†	LASSO	**−0.97**	—	—	**−1.00**	—	2/2†
Effect size	Cohen's d	**−1.00**	−0.02	−0.70	**−1.00**	+0.01	2/5

H6d: Von Neumann entropy monotonicity (CONFIRMED). Von Neumann entropy stability [21] achieves $|\rho| = 0.88$ on α_{T2} (CI: $[0.85, 0.90]$; $d = 0.83$ for the excess over the 0.8 threshold), exceeding the pre-registered threshold at $\alpha = 0.0167$.

H8a: LASSO vs. geometric at low decoherence (VOIDED). This hypothesis required comparing LASSO [33] AUC directly against geometric method scores at low decoherence ($\alpha_{T2} < 0.22$). At these levels, all classification baselines achieve AUC = 1.0 on all 100 seeds (the 1.5 K tumor-healthy difference produces a ∼111 kHz frequency shift that remains trivially classifiable). Geometric methods produce scores on incomparable scales (bracket norms $\sim 10^8$, curvatures ~ -20, angles ~ 2.0). The Spearman-ρ normalization adopted for all other comparisons is inapplicable to a within-regime score comparison. The underlying observation—classification is trivially solved at low decoherence while geometric methods detect structural changes below the classification threshold—is reported descriptively but carries no confirmatory weight.

4.3 The method–axis sensitivity map

table 3 presents the core result: the median Spearman ρ for each method family's best representative on each decoherence axis. The "Axes > 0.8" column counts how many axes the method tracks with $|\rho| > 0.8$.

4.4 Device variation: the discriminating axis

Of 30 methods tested, only 6 achieve $|\rho| > 0.5$ on σ_{device}: Cohen's d (−0.70), localized bracket (−0.66), bracket norm raw (+0.65), bracket norm corrected (+0.65), t-test (−0.56), and IRM penalty (−0.55). None exceed $|\rho| > 0.8$. By contrast, 25 of 30 methods achieve $|\rho| > 0.5$ on α_{T_2}, and 19 exceed $|\rho| > 0.8$.

The physics explains the gap. Device calibration variation adds a constant temperature offset per device—a mean shift in the feature space that preserves within-device covariance structure.

Methods operating on subspace angles (Grassmannian), graph curvature (Ollivier-Ricci, Forman-Ricci), spectral properties, or direction-based proxies are invariant to such shifts by construction. Only methods sensitive to between-group location differences (bracket norms, effect sizes) or within-group dispersion changes (persistent H0) detect the perturbation.

This result has practical implications for clinical multi-sensor deployments, where device-to-device calibration differences are often the primary noise source [42, 43]. In NV-diamond thermometry specifically, per-particle strain variation produces ∼170 kHz standard deviation in the D parameter [26], and intracellular experiments require per-nanodiamond calibration [27]. Geometric methods that are invariant to additive offsets provide no advantage over a t-test in this regime.

4.5 Exploratory findings

Direction-based methods share mean-shift invariance. Berry phase [20], Grassmannian geodesic [3], and Grassmannian holonomy all measure angles or direction-dependent properties in feature space. All three are blind to σ_{device} ($|\rho| < 0.15$) for the same reason: additive offsets do not change directions. The Berry-phase proxy in particular achieves $\rho = +1.00$ on α_{T2} and $+0.95$ on σ_{surface}—both axes that distort covariance structure—while remaining at $\rho = +0.006$ on σ_{device}. This selectivity reflects the shared geometric property (direction invariance to mean shifts), not a quantum-specific detection mechanism.

Transport-stable bracket: broad coverage, device-blind. The transport-stable bracket achieves 4/5 axes at $|\rho| > 0.8$, matching raw bracket norms on composite and non-composite axes alike ($\rho = +1.00$ on α_{T2}, $+0.93$ on σ_{surface}, $+1.00$ on σ_{temp}, $+0.99$ on n_{centers}). On σ_{device}, however, transport stabilization collapses the signal: $\rho = -0.01$ (CI: $[-0.09, +0.07]$) compared to $+0.65$ for the raw bracket. The stabilization procedure removes location information by design, yielding the same mean-shift blindness seen in direction-based methods. This tradeoff—broad robustness at the cost of device-variation sensitivity—is quantifiable only with the companion-axis design.

Curvature phase transition. Ollivier-Ricci [10] and Forman-Ricci [11] curvature exhibit a sharp phase transition on α_{T2} (figure 3): Ollivier-Ricci collapses from 0.998 to 0.200 between $\alpha_{T2} = 0.33$ and 0.44, reaching ≈ 0 by 0.56; Forman-Ricci jumps from -18 (negative curvature, indicating structured clusters) to $+0.5$ at the same boundary. This transition marks the decoherence level at which the k-NN graph loses cluster structure, defining a "graph collapse threshold" beyond which discrete curvature methods are uninformative.

Persistent H0: 4/5 axes. Persistent H0 [44] (connected components via union-find) achieves 4/5 axes at $|\rho| > 0.8$ (figure 4) with monotonically increasing total persistence on α_{T2} (from 2.2×10^7 to 4.2×10^8). The decision to restrict to H0 and cut the incorrect H1 heuristic is validated: the NV-diamond point cloud has no decoherence-varying loop structure, so H1 would have added noise. The breadth of H0's coverage has a geometric explanation: total H0 persistence measures the sum of all inter-point distance gaps at which connected components merge, responding to any perturbation that changes pairwise distances—mean shifts, variance changes, dimensionality effects. This non-specificity is the source of its coverage: H0 tracks 4/5 axes precisely because it is agnostic to the geometric mechanism, unlike direction-based or curvature-based methods that are selective by construction. The tradeoff is interpretability: H0 detects that something changed without indicating what.

Sheaf cohomology: a split result. Sheaf H^1 (Cochran's Q) and per-edge sheaf Q achieve $|\rho| < 0.5$ on all axes, with most below 0.2—a weak-to-null result (figure 5). Persistent sheaf cohomology, by contrast, achieves $|\rho| > 0.8$ on 3 of 5 axes (α_{T2}: -1.00; σ_{surface}: -1.00; σ_{temp}: -0.98), while remaining blind to σ_{device} ($+0.02$) and n_{centers} (-0.13). Persistent sheaf cohomology uses an explicit noise-level filtration: for 10 noise levels $\epsilon \in [0, 0.5]$, per-device effect-size estimates are perturbed by $\mathcal{N}(0, \epsilon|\hat{\beta}|)$ and Cochran's Q is recomputed at each level. The score is the area under the $Q(\epsilon)$ curve (trapezoidal integration), capturing how device heterogeneity accumulates across perturbation scales. Pointwise Cochran Q, by contrast, evaluates at a single scale and misses multi-scale structure.

An exploratory device-count expansion (post-hoc, separate from the main pre-registered experiment) confirms that the sheaf H^1 null on σ_{device} is structural, not a power limitation. At 5 devices (100 seeds), sheaf H^1 achieves $\rho = +0.04$ (CI: $[-0.16, +0.21]$)—statistically indistinguishable from the 3-device result ($\rho = -0.07$, CI: $[-0.22, +0.35]$). By contrast, bracket norm raw strengthens from $+0.65$ at 3 devices to $+0.73$ at 5 devices (CI: $[+0.69, +0.76]$), indicating that power increases with device count for methods sensitive to mean shifts. The Cochran Q statistic [31] underlying sheaf H^1 measures heterogeneity in per-device effect sizes; when all devices share the same physics (single-modality NV-diamond), a common-mode calibration offset produces homogeneous effects by construction, yielding $Q \approx 0$ regardless of device count.

Wasserstein persistence: an α_{T2} specialist. Wasserstein persistence distance [17] achieves $\rho = +0.98$ on α_{T2}, with moderate response on σ_{temp} ($+0.71$) and σ_{surface} ($+0.41$), but near zero on σ_{device} (-0.02) and weak on n_{centers} ($+0.30$). Like spectral gap stability, it responds to the global covariance collapse induced by the composite coherence axis while remaining insensitive to additive or per-sample noise.

Spectral gap: an α_{T2} specialist. Spectral gap stability achieves $\rho = +0.96$ on α_{T2} and $|\rho| < 0.12$ on all other axes. The composite coherence axis collapses the spectral gap of the sample covariance matrix, while the other four axes (additive offsets, per-sample noise, dimensionality) preserve the eigenvalue gap. This makes spectral gap stability a narrow but precise detector of coherence loss.

Localized bracket: a σ_{device} specialist. Localized bracket achieves $\rho = -0.66$ on σ_{device} (CI: $[-0.71, -0.61]$) and $|\rho| < 0.05$ on all other axes. It measures within-neighborhood geometric structure, which collapses as device offsets increase the intra-device spread. Along with spectral gap stability (α_{T2} only), localized bracket is one of the few methods with a single-axis specialization profile—and the only one specialized to σ_{device}.

Classification baselines saturate at low decoherence. All classification baselines (LASSO, logistic, RF, DRO) return AUC = 1.0 on every level of σ_{surface} and σ_{device} across all 100 seeds. The 1.5 K tumor-healthy temperature difference ($\sim$111 kHz frequency shift) is at the upper end of reported intracellular gradients [23] and remains trivially classifiable even at maximum surface noise or device variation. Smaller temperature differences ($<$0.5 K) would shift classification AUC downward, potentially revealing a regime where geometric methods offer earlier detection. On α_{T2}, AUC begins degrading above $\alpha_{T2} \approx 0.44$, with LASSO dropping from 1.0 to 0.75 at $\alpha_{T2} = 1.0$. Geometric methods detect structural changes at decoherence levels where classification has not yet degraded—they measure different properties than classification accuracy.

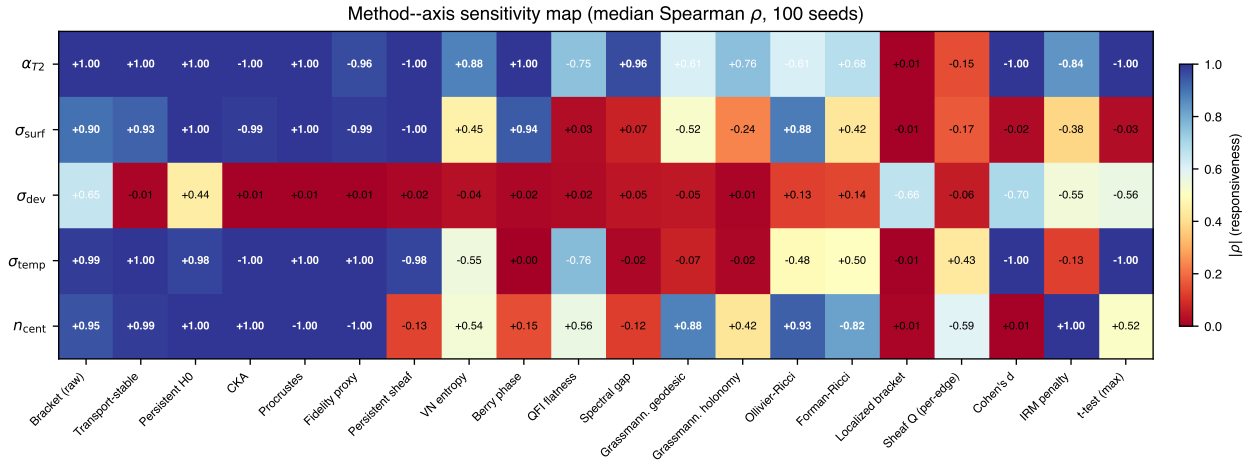

Figure 1: Method–axis sensitivity map. Each cell shows the median Spearman ρ(score, decoherence level) across 100 seeds (10 levels per axis, 200 samples per class, 3 devices). Bold text marks $|\rho| > 0.8$. The σ_{dev} column is the discriminating axis: most methods cluster near zero, while bracket norms and Cohen's d remain responsive. Color: RdYlBu on $|\rho|$ (blue = responsive, red = unresponsive); bold text provides a redundant encoding for $|\rho| > 0.8$ cells independent of color.

5 Discussion

5.1 Why the sensitivity map has the structure it does

The sensitivity map's most striking feature—near-zero response on σ_{device} for the majority of geometric methods—has a simple explanation: each method's blind spots are predictable from its invariance properties. A method that operates on subspace angles [45, 46] (Grassmannian geodesic, holonomy), graph curvature [47] (Ollivier-Ricci, Forman-Ricci), or direction-dependent statistics (Berry phase proxy) is invariant to translations in feature space by construction. Device calibration variation (σ_{device}) is precisely such a translation: a constant temperature offset per device that shifts both class means equally. These methods report "no change" because, geometrically, nothing they measure has changed.

Translation invariance—$f(X + \mathbf{1}c^\top) = f(X)$ for any constant offset c—is a standard property of any method operating on centered data or pairwise differences. The observation here is applied, not mathematical: this elementary invariance has a direct, testable consequence for sensor deployments. An analogous tradeoff between invariance and information loss has been studied in domain adaptation [48], where domain-invariant representations discard domain-specific signal by construction; the present work identifies the metrology-specific instance of this tradeoff and maps it across geometric method families. In the multi-device case, each device d receives a different offset c_d, so the perturbation is not a single global translation. The blindness persists because per-device offsets preserve within-device covariance structure while shifting only device means—methods that center per-class or operate on pairwise differences absorb these shifts, producing the same insensitivity as under a global translation. On single-modality sensors, device calibration variation is an additive offset, so every translation-invariant method is blind to it by construction.

table 3 confirms the prediction empirically. Every method with $|\rho| < 0.15$ on σ_{device} (Grassmannian geodesic, holonomy, Berry phase, spectral gap, QFI flatness, Procrustes, CKA, fidelity proxy, persistent sheaf) satisfies translation-invariance. Every method with $|\rho| > 0.5$ on σ_{device} (bracket norms, Cohen's d, t-test, IRM penalty) does not. The prediction is falsifiable: any translation-

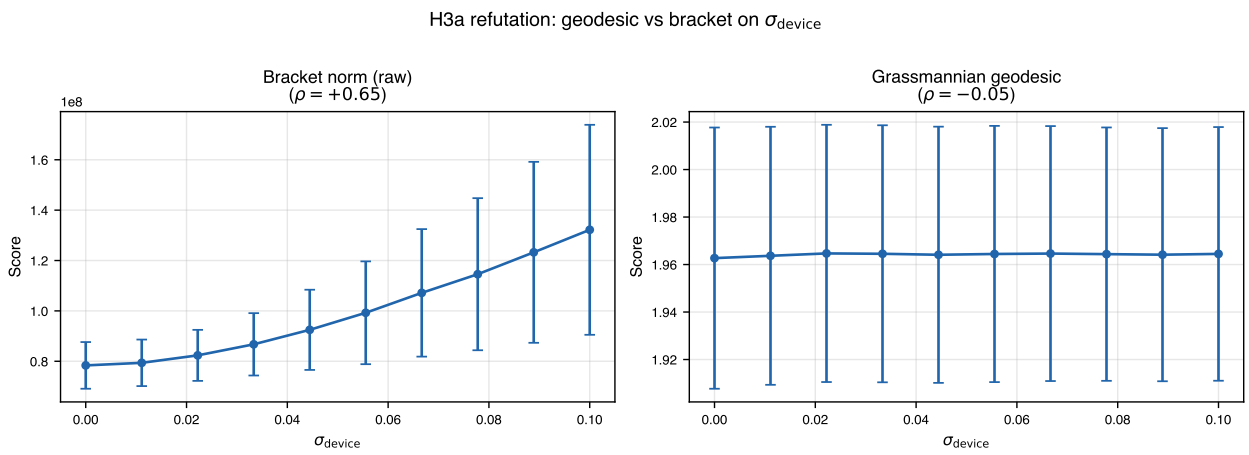

Figure 2: The H3a refutation mechanism on σ_{device}. **Left:** Bracket norm (raw) increases monotonically with device variation ($\rho = +0.65$, CI: $[+0.59, +0.69]$). **Right:** Grassmannian geodesic distance is flat ($\rho = -0.05$, CI: $[-0.12, +0.04]$), spanning only 1.94–1.95 across all decoherence levels. Error bars show ± 1 standard deviation across 100 seeds. The geodesic is invariant to the additive per-device offset that σ_{device} introduces.

invariant method achieving $|\rho| > 0.5$ on σ_{device} would refute it.

Six methods cover 4 of 5 axes at $|\rho| > 0.8$ (bracket norm raw, transport-stable bracket, persistent H0, CKA, Procrustes, fidelity proxy), and persistent sheaf cohomology covers 3 of 5. The missing axis for all of them is σ_{device}. Matching the method to the noise geometry is the operational challenge.

5.2 The value of the companion axis

The σ_{device} companion axis was added during pre-registration review to disentangle composite-axis effects from genuine subspace drift. Without it, the paper would have reported that "most methods work well" (25/30 achieve $|\rho| > 0.5$ on α_{T2}) and missed the central finding. The companion axis design—testing the same hypothesis on a clean, non-composite axis alongside the composite one—is a transferable experimental principle for any study where the primary axis affects multiple features simultaneously.

5.3 Pre-registration discipline

The H3a refutation demonstrates the value of pre-registration. Post hoc, one might rationalize either outcome: "Grassmannian methods should detect subspace drift because they directly measure subspace angles" (our pre-registered prediction) or "Grassmannian methods should miss additive noise because they are location-invariant" (the actual explanation). Both are geometrically coherent. Without a frozen prediction, the second explanation could be presented as if it were the original hypothesis.

Similarly, the H8a voiding shows that pre-registration can expose incoherence in the statistical framework. The hypothesis was written before the Spearman-ρ normalization was adopted, and it survived the review rounds because it seemed intuitively correct. Only when the results arrived did the scale-incompatibility become undeniable. The voiding changed a confirmatory outcome to a descriptive one and altered the Bonferroni denominator—evidence that the pre-registration had teeth, constraining post-hoc interpretation rather than merely documenting it.

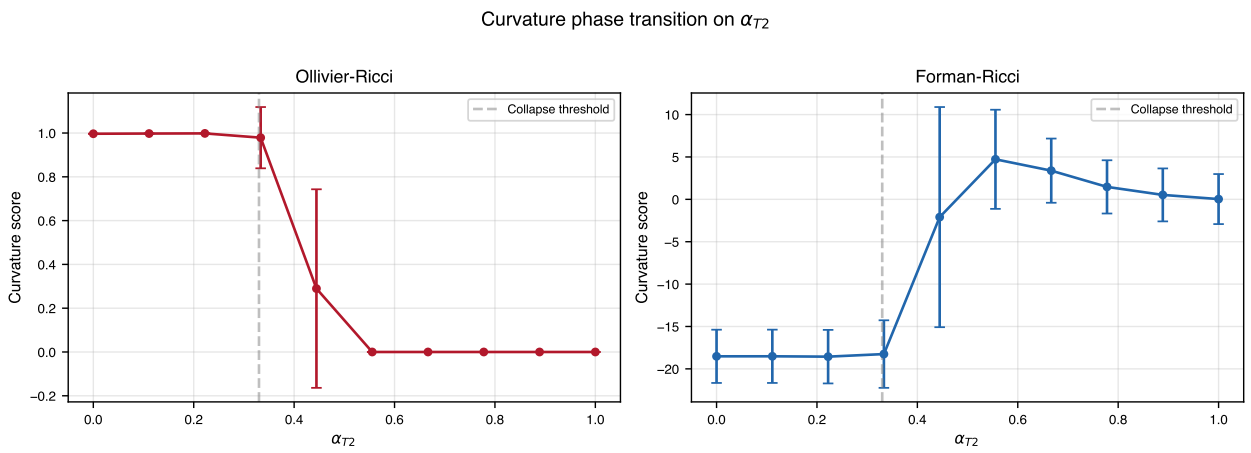

Figure 3: Curvature phase transition on α_{T2}. **Left:** Ollivier-Ricci curvature collapses from ≈ 1.0 to ≈ 0 between $\alpha_{T2} = 0.33$ and 0.56. **Right:** Forman-Ricci curvature flips sign at the same threshold, from -18 (structured clusters) to $+5$ (collapsed). The dashed line marks the shared collapse threshold. Error bars: ± 1 SD, 100 seeds.

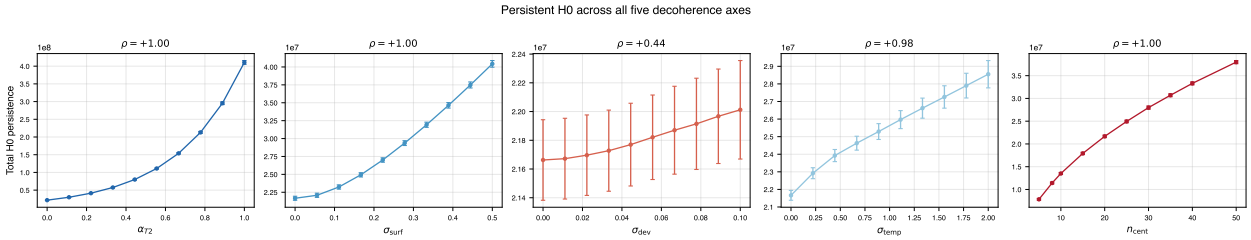

Figure 4: Persistent H0 across all five decoherence axes. Monotonic responses ($|\rho| > 0.8$) on 4 of 5 axes: α_{T2}, σ_{surf}, σ_{temp}, and n_{cent}. Only σ_{dev} ($\rho = +0.44$) falls below the 0.8 threshold. Error bars: ± 1 SD, 100 seeds.

5.4 Limitations

The NV-diamond simulation is a controlled testbed that serves as an existence demonstration: it establishes that geometric invariance properties create predictable method-axis blind spots under parameterized noise, not that specific ρ values will transfer unchanged to real sensors. Feature dimensions are low (≤ 5), sample sizes are moderate ($n = 200$ per class), and the noise model is parametric. Real NV-diamond measurements involve nonlinear transductions, non-Gaussian noise, and correlated decoherence channels. The transferable findings are structural (translation-invariant methods are blind to additive offsets) rather than quantitative (the exact ρ values). Validation on experimental NV-diamond data—such as the fiber-coupled multi-sensor datasets from Rajpal et al. [28]—is a planned next step.

The sheaf H^1 null on σ_{device} is structural, not a power limitation: a device-count expansion from 3 to 5 devices (20 seeds each) shows no activation (ρ shifts from -0.07 to $+0.04$, both CIs including zero), while bracket norms strengthen over the same range. Single-modality device variation produces homogeneous per-device effects, which Cochran's Q cannot detect by construction. Persistent sheaf cohomology, which wraps Q in a persistence filtration, achieves strong performance on 3 of 5 axes—the split result reflects the difference between pointwise heterogeneity testing and multi-scale topological structure. Testing sheaf H^1 on cross-modality data, where different sensor

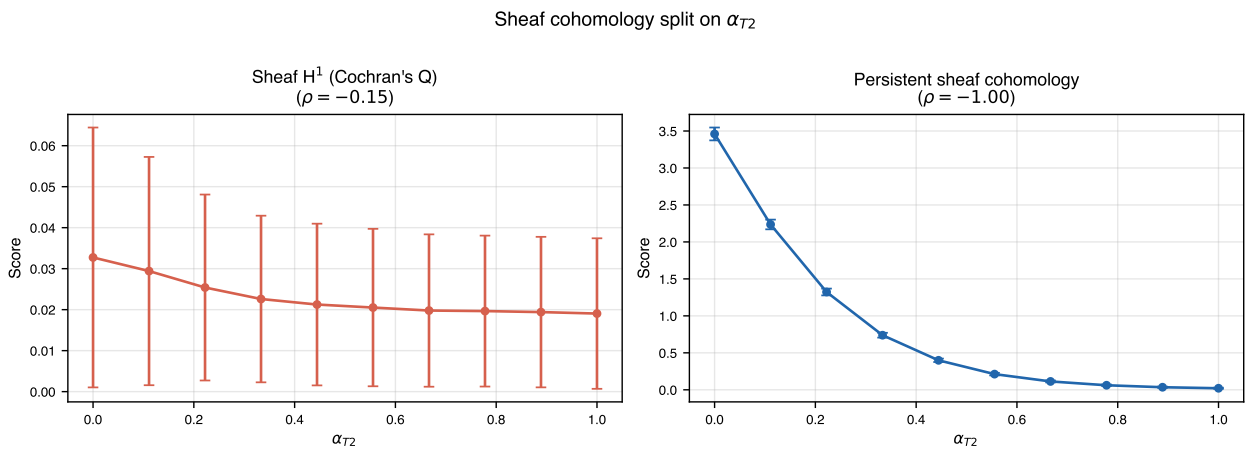

Figure 5: Sheaf cohomology split on α_{T2}. **Left:** Pointwise sheaf H^1 (Cochran's Q) is flat ($\rho = -0.15$, CI includes zero at high decoherence). **Right:** Persistent sheaf cohomology decreases monotonically ($\rho = -1.00$). The persistence filtration captures multi-scale changes in device-consistency structure that pointwise Q misses. Error bars: ± 1 SD, 100 seeds.

types produce qualitatively different effect sizes, remains an open question.

The device-variation axis models calibration differences as additive temperature offsets with $\sigma_{\text{device}} \in [0, 0.1]$ K. Real per-nanodiamond strain variation produces $\sim$170 kHz standard deviation in D [26], corresponding to $\sim$2.3 K equivalent offset at $dD/dT = -74$ kHz/K—roughly 20$\times$ the simulated range. The qualitative finding (translation invariance implies blindness) holds at any magnitude, but sensitivity values may differ at realistic offsets. Real device-to-device variation also includes multiplicative effects: differences in NV orientation distribution, microwave delivery efficiency, and fluorescence collection efficiency produce gain-like calibration errors that scale signal amplitudes rather than shifting them. The translation-invariance prediction applies only to the additive component. An analogous principle should hold for scale invariance: methods invariant to per-feature scaling—such as correlation-based metrics—would be blind to multiplicative gain variation, while methods sensitive to absolute magnitudes would respond. Under multiplicative noise, methods currently labeled "blind" to additive offsets (such as Grassmannian geodesic or CKA) might respond, because a per-device gain changes covariance structure and subspace angles. Testing multiplicative and mixed noise models is a planned extension.

The one-axis-at-a-time design isolates each noise mechanism cleanly but does not test interactions. In real NV-diamond measurements, T2 degradation and surface noise co-occur (both originate from spin-bath dynamics), and their joint effect may produce geometric signatures that differ qualitatively from the sum of single-axis perturbations. The sensitivity map should be read as a single-axis profile; whether it composes linearly under joint perturbations is an open empirical question.

The Spearman rank correlation summarizes each method's response as a single monotonic association strength, which is lossy for methods that exhibit threshold behavior. Ollivier-Ricci curvature, for instance, achieves median $\rho = +1.00$ below the collapse threshold ($\alpha_{T2} \leq 0.33$) but only $\rho = +0.10$ above it ($\alpha_{T2} \geq 0.56$); the full-range $\rho = -0.61$ understates pre-collapse sensitivity and overstates post-collapse informativeness. A breakpoint or piecewise correlation analysis would better characterize threshold-behavior methods; we report the single ρ for cross-method comparability and flag this as a framework limitation.

Feature dimensionality is low ($d \leq 5$ derived ODMR features per center) compared to the regime

where many geometric methods—persistent homology, sheaf cohomology, Grassmannian analysis—are typically deployed. At higher dimensions (e.g., using raw ODMR spectra directly), richer geometric structure could change the sensitivity map. The present results characterize method behavior in a low-dimensional setting that matches common sensor-feature pipelines but may not extrapolate to high-dimensional representations.

The Berry-phase and QFI implementations are classical proxies applied to feature-space statistics, not direct quantum-state measurements. Claims about "quantum-channel selectivity" would require access to the actual quantum state, not to processed measurement features.

6 Conclusion

This study provides a systematic characterization of how 23 geometric analysis methods and 7 scalar baselines respond to five types of sensor noise in NV-diamond thermometry. The resulting sensitivity map (figure 1, table 3) is the primary contribution: it shows that six methods achieve broad coverage (4 of 5 axes at $|\rho| > 0.8$), but all fall short on device calibration variation—the noise axis with the simplest physics is the one that discriminates. A straightforward invariance argument explains this pattern: methods that operate on centered data or pairwise differences are blind to additive per-device offsets by construction, and the data confirm this without exception (figure 2). These results characterize method behavior in a low-dimensional setting ($d \leq 5$ derived ODMR features) that matches common sensor-feature pipelines; sensitivity profiles may differ at higher feature dimensions.

Practitioner guidance. table 4 distills the sensitivity map into a method selection guide for three deployment scenarios.

Table 4: Method selection guide. Recommended methods for three noise regimes, derived from the sensitivity map (table 3).

Dominant noise	Recommended methods	Why
Unknown / mixed	Bracket norm (raw)	Broadest coverage (4/5 axes)
Coherence degradation	Berry phase, persistent sheaf	$\|\rho\| \geq 0.95$ on α_{T2} (opposite sign; both responsive)
Device calibration	Cohen's d, localized bracket	Most geometric methods are blind

When the dominant noise source is unknown, bracket norms provide the broadest coverage (4/5 axes at $|\rho| > 0.8$). (Bracket norms were developed by the present author [6]; the pre-registration, open data, and frozen analysis code allow independent verification of this result.) When coherence degradation is the primary concern, direction-based methods offer precise detection. When device-to-device calibration variation dominates—as it does in many clinical multi-sensor deployments—simple effect-size baselines or localized bracket norms are the appropriate tools for *structural monitoring* (detecting that device variation exists), because most geometric methods are blind to additive offsets. This recommendation applies to detecting calibration drift, not to classification: at the 1.5 K temperature difference tested here, all classification baselines achieve AUC = 1.0 regardless of device variation, so the geometric-vs-baseline comparison on σ_{device} is undefined. At smaller temperature differences (< 0.5 K), classification performance may degrade, and geometric methods could offer earlier structural detection.

Transferability and next steps. The invariance-blindness principle applies wherever device calibration produces additive offsets—the prediction is geometric, not NV-specific. Optically pumped

magnetometers exhibit per-sensor offsets from residual magnetic fields and misalignment; SQUID arrays have per-channel flux offsets; atomic clocks have per-unit frequency biases from local perturbations. In each case, translation-invariant analysis methods are blind to these offsets by the same geometric argument. The companion-axis methodology transfers to any domain where a primary noise axis is composite: testing the same hypothesis on a clean, non-composite axis alongside the composite one exposed the central finding here. A planned follow-up on experimental fiber-coupled NV data [28] will test whether the principle holds under realistic magnitudes ($\sim$2.3 K equivalent strain variation, $\sim$20$\times$ the simulated range) and multiplicative noise, where currently "blind" methods may respond.

Acknowledgments

Funding

No funding was received for this work.

Author contributions

E.T.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Data availability

All data, code, pre-registration, deviation log, and raw results (150,000 evaluations as JSON) are publicly available at `https://github.com/elliottower/geometric-sensor-blindspots` under an MIT license and archived at Zenodo (DOI: 10.5281/zenodo.21299293). The simulation is fully deterministic: given the frozen commit SHA (`8401d54`) and seed pool (seeds 1000–1099), all results can be regenerated. Software dependencies: Python 3.11+, NumPy, SciPy, scikit-learn, tqdm (no GPU or proprietary software required).

AI disclosure

AI-assisted tools (Claude Code, Perplexity Pro) were used for code drafting, data analysis, figure generation, and manuscript editing. All scientific decisions—experimental design, pre-registration, method selection, hypotheses, interpretations, and conclusions—were made by the author. The author reviewed and verified all outputs and takes full responsibility for the accuracy and integrity of the work. No AI tool is listed as an author.

Competing interests

The author declares no competing interests.

References

[1] Antonio Possolo. Simple guide for evaluating and expressing the uncertainty of NIST measurement results. *NIST Technical Note*, 1900, 2019.

[2] D. Rod White. In pursuit of a fit-for-purpose uncertainty guide. *Metrologia*, 53(4):S107–S124, 2016.

[3] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[4] Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of arbitrary dimension. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

[5] Thomas Bendokat, Ralf Zimmermann, and P-A Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1): 6, 2024.

[6] Elliot Tower. Bracket norm identifies causally important brain regions from population geometry, 2026. URL `https://doi.org/10.5281/zenodo.21246305`.

[7] Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

[8] Michael Robinson. Topological signal processing. *Mathematical Engineering*, 2014.

[9] Justin Curry. Sheaves, cosheaves and applications. *arXiv preprint arXiv:1303.3255v2*, 2014.

[10] Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[11] Robin Forman. Bochner's method for cell complexes and combinatorial Ricci curvature. *Discrete and Computational Geometry*, 29(3):323–374, 2003.

[12] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2): 255–308, 2009.

[13] Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[14] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

[15] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[16] John C. Gower and Garmt B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

[17] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[18] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

[19] Samuel L. Braunstein and Carlton M. Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439–3443, 1994.

[20] Michael V. Berry. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London A*, 392(1802):45–57, 1984.

[21] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955. Originally published in German, 1932.

[22] Géza Tóth and Iagoba Apellaniz. Quantum metrology from a quantum information science perspective. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424006, 2014.

[23] Georg Kucsko, Peter C. Maurer, Norman Y. Yao, Michael Kuber, Hyun Joon Noh, Pei Kehui Lo, Hongkun Park, and Mikhail D. Lukin. Nanometre-scale thermometry in a living cell. *Nature*, 500(7460):54–58, 2013.

[24] Marcus W. Doherty, Neil B. Manson, Paul Delaney, Fedor Jelezko, Jörg Wrachtrup, and Lloyd C. L. Hollenberg. The nitrogen-vacancy colour centre in diamond. *Physics Reports*, 528(1): 1–45, 2013.

[25] Romana Schirhagl, Kevin Chang, Michael Lorber, and Christian L. Degen. Nitrogen-vacancy centers in diamond: Nanoscale sensors for physics and biology. *Annual Review of Physical Chemistry*, 65:83–105, 2014.

[26] Masazumi Fujiwara et al. Fluorescent nanodiamonds for high-resolution thermometry in biology. *Physics and Chemistry of Carbon Nanostructures*, 2024. Reports ∼170 kHz std. dev. in D parameter across >9000 nanodiamonds from strain variation.

[27] Taiki Sekiguchi, Shingo Sotoma, et al. Simultaneous nanorheometry and nanothermometry using intracellular diamond quantum sensors. *ACS Nano*, 17(19), 2023.

[28] Shraddha Rajpal, Zeeshan Ahmed, and David W. Berry. Evaluating probabilistic and data-driven inference models for fiber-coupled NV-diamond temperature sensors. *Optics Express*, 2025.

[29] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[30] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[31] William G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, 1954.

[32] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

[33] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[34] Leo Breiman. *Random Forests*, volume 45. Springer, 2001.

[35] Martin Arjovsky, Léon Bottou, Ishaan Gulcevich, and David Lopez-Paz. Invariant risk minimization. In *arXiv preprint arXiv:1907.02893*, 2019.

[36] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[37] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[38] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[39] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57 (1):289–300, 1995.

[40] Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[41] Andrew V. Knyazev and Merico E. Argentati. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

[42] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[43] Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Ilke Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018.

[44] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[45] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[46] Yasuko Chikuse. *Statistics on Special Manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer, 2003.

[47] Areejit Samal, R. P. Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jost Jürgen. Comparative analysis of two discretizations of Ricci curvature for complex networks. *Scientific Reports*, 8(1):8650, 2018.

[48] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79 (1–2):151–175, 2010.