Identifying Defensive Coverage Gaps in NFL Tracking Data Using Topological Data Analysis

Arjun Mahesh

December 4, 2025

Abstract

This project applies persistent homology to NFL player tracking data to quantify defensive coveage gaps and investigate their relationship with pass completion outcomes. Using the NFL Big Data Bowl 2021 dataset, comprising between 17,000 and 19,000 passing plays across 253 games, we construct Vietoris-Rips filtrations from defender position point clouds and extract topological features in dimensions 0 and 1. The analysis reveals that H_1 persistence, corresponding to coverage "holes," differs significantly between complete and incomplete passes ($p = 0.005$), though contrary to initial hypotheses, incomplete passes exhibit larger gaps on average. We extend the analysis temporally, tracking topology evolution from snap to pass release, and demonstrate that coverage gaps decrease over time as defenses react to routes. Additionally, unsupervised clustering on H_1 features successfully distinguishes zone from man coverage without labels, with zone defenses showing 40.8× larger persistent gaps. These results demonstrate that TDA provides interpretable insights into defensive structure that complement traditional football analytics.

All code for this project is available at `https://github.com/arjunmahesh1/nfl-tda`.

1 Introduction

In American football, defensive pass coverage creates spatial patterns that determine whether receivers can find open space. Traditional metrics reduce this geometric complexity to scalar quantities such as "yards of separation," losing information about the global structure of defensive formations. This project investigates whether topological data analysis, specifically persistent homology, can capture richer structural information about coverage patterns and whether such topological features correlate with play outcomes.

The central research question is: **Can persistent homology quantify defensive coverage gaps, and do such topological features correlate with pass completion?** We hypothesize that H_1 features (loops/holes in the Vietoris-Rips complex) correspond to coverage gaps that quarterbacks exploit, while H_0 features (connected components) capture defender clustering patterns.

TDA has been successfully applied to sports analytics in prior work. Pierson et al. used persistent homology on NBA roster Data to identify topological signatures correlated with offensive performance [3]. In soccer, TDA-based methods have captured nonlinear relationships in player positioning that traditional clustering misses [5]. Our work extends this methodology to NFL defensive formations, where the geometric interpretation of H_1 as "holes" directly corresponds to open passing lanes.

The contributions of this project include: (1) a complete pipeline for computing persistent homology from NFL tracking data, (2) statistical analysis relating topological features to pass outcomes, (3) temporal extension tracking topology evolution during plays, and (4) unsupervised coverage type classification using H_1 signatures.

2 Data

We use the NFL Big Data Bowl 2021 dataset, publicly available on Kaggle, which provides player tracking data for all passing plays from the 2018 NFL regular season. The dataset includes:

- **Tracking data**: (x, y) coordinates of all 22 players and the football at 10 frames per second.
- **Play metadata**: Down, distance, pass result (Complete, Incomplete, Interception, Sack)
- **Coverage**: 253 games, 19,237 passing plays, 986,022 total tracking frames

For static analysis, we extract defender positions at the moment of pass release (identified by the `pass_forward` event), yielding 17,740 usable plays after filtering for data quality. The average number of defenders per play is 7.7, with range 4–11. For temporal analysis, we extract formations at 9 time points per play from snap to pass outcome, covering 962 plays with complete temporal data.

Pass result distributions: Complete (C) = 11,273 plays (63.5%), Incomplete (I) = 6,047 plays (34.1%), Interception (IN) = 416 plays (2.3%), Sack (S) = 4 plays (<0.1%).

3 Methodology

3.1 Point Cloud Construction

For each play, we represent the defensive formation as a point cloud $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^2$ where each x_i is the (x, y) field position of defender i. We consider only defensive players (excluding the football and offensive players) to isolate coverage structure.

3.2 Persistent Homology via Vietoris-Rips Filtration

Given point cloud X, we construct the Vietoris-Rips complex at scale ϵ:

$$\mathrm{VR}_\epsilon(X) = \{\sigma \subseteq X : \mathrm{diam}(\sigma) \leq \epsilon\}$$

where $\mathrm{diam}(\sigma) = \max_{x_i, x_j \in \sigma} \|x_i - x_j\|$. As ϵ increases from 0 to ∞, simplices are added, and homology groups $H_p(\mathrm{VR}_\epsilon$ change. Features that persist across a range $[b, d)$ of scales are recorded in the persistence diagram:

$$\mathrm{Dgm}_p(X) = \{(b_i, d_i)\}_{i=1}^{n_p}, \quad \text{persistence} = d_i - b_i$$

In our context:

- H_0 features represent connected components (defender clusters). At $\epsilon = 0$, each defender is isolated; as ϵ grows, clusters merge.
- H_1 features represent 1-dimensional holes (coverage gaps). A loop of defenders enclosing empty space creates an H_1 feature. The persistence measures the ”size” of the gap in yards.

We compute persistence using the Ripser library [4], which implements the standard algorithm with clearing optimization.

3.3 Feature Extraction

From each persistence diagram, we extract summary statistics:

- Number of features: $|Dgm_p|$
- Maximum persistence: $\max_i(d_i - b_i)$
- Average persistence: $\frac{1}{n}\sum_i(d_i - b_i)$
- Total persistence: $\sum_i(d_i - b_i)$
- Number of significant features: $|\{i : d_i - b_i > 1 \text{ yard}\}|$
- Persistence entropy: $-\sum_i p_i \log p_i$ where $p_i = \frac{d_i - b_i}{\sum_j (d_j - b_j)}$

3.4 Vectorization Methods

To enable statistical analysis and machine learning, we employ several vectorization strategies:

Persistence Landscapes [2] The k-th landscape function is:

$$\lambda_k(t) = k\text{-th largest value of } \min(t - b_i, d_i - t)^+$$

Landscapes form a Banach space, enabling computation of means and statistical tests.

Persistence Images [1] We discretize the birth-death plane, weight points by persistence, and convolve with a Gaussian kernel to produce fixed-size image representations.

Betti Curves: The Betti number $\beta_p(\epsilon) = \text{rank}(H_p(\text{VR}_\epsilon))$ as a function of scale.

3.5 Distance Metrics

To compare persistence diagrams, we use:

Bottleneck Distance:

$$d_B(D_1, D_2) = \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_\infty$$

where the infimum is over all bijections γ (including matching to the diagonal). This satisfies the stability theorem: $d_b(\text{Dgm}(X), \text{Dgm}(Y)) \leq d_H(X, Y)$.

Wasserstein Distance:

$$W_p(D_1, D_2) = \left(\inf_{\gamma} \sum_{x \in D_1} \|x - \gamma(x)\|^p \right)^{1/p}$$

3.6 Statistical Testing

We employ two-sample t-tests on extracted features and permutation tests on persistence diagrams to assess significance of differences between complete and incomplete passes.

3.7 Unsupervised Coverage Type Classification

Without ground-truth coverage labels, we employ unsupervised learning to infer whether each play uses zone or man coverage based on topological signatures. The approach leverages the geometric intuition that zone defenses create territorial gaps (high H_1 persistence) while man coverage tracks receivers closely (low H_1 persistence).

Feature Selection. For each play, we extract a feature vector:

$$\mathbf{x}_i = \begin{pmatrix} |H_1| & \max(\text{pers}_{H_1}) & \text{avg}(\text{pers}_{H_1}) & \text{entropy}(H_0) & n_{\text{box}} \end{pmatrix}^\top$$

where $|H_1|$ is the number of H_1 features, $\max(\text{pers}_{H_1})$ and $\text{avg}(\text{pers}_{H_1})$ are the maximum and average H_1 persistence, $\text{entropy}(H_0)$ captures diversity of defender clustering, and n_{box} is the number of defenders in the box (contextual feature from the dataset).

Clustering. We normalize features using standard scaling:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

and apply K-means clustering with $K = 2$ to partition plays into two groups. The K-means objective minimizes within-cluster variance:

$$\arg\min_{\{C_k\}} \sum_{k=1}^{2} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

Cluster Interpretation. We assign semantic labels by examining cluster centroids. The cluster with higher mean H_1 persistence is labeled "Zone" (territorial gaps between zones), while the cluster with lower persistence is labeled "Man" (tight receiver coverage). This labeling is consistent with football domain knowledge: zone defenses defend areas, creating inherent gaps, while man defenses shadow receivers, eliminating persistent holes.

3.8 Temporal Extension

For temporal analysis, we compute persistence at 9 time points per play (every 0.4s from snap to pass outcome) and track:

- Evolution of H_1 features over time
- **Disguise metric**: $d_B(\mathrm{Dgm}_{\mathrm{snap}}, \mathrm{Dgm}_{\mathrm{pass}})$ measuring topological change
- **Time-to-gap**: First frame where a significant H_1 feature (>1 yard persistence) appears
- **Betti surfaces**: $\beta_1(\epsilon, t)$ as a function of both scale and time

4 Results

4.1 Example: Persistence Diagram and Betti Curves

Figure 1 shows a representative persistence diagram for a single play (Game 2018090600, Play 545). The H_0 features (blue dots) show 6 connected components at birth, progressively merging as the filtration scale increases. The single H_1 feature (red square) is born at approximately 8.7 yards and dies at 10.5 yards, indicating a coverage gap with persistence of 1.8 yards. The corresponding barcode visualizes this gap's lifespan across scales.

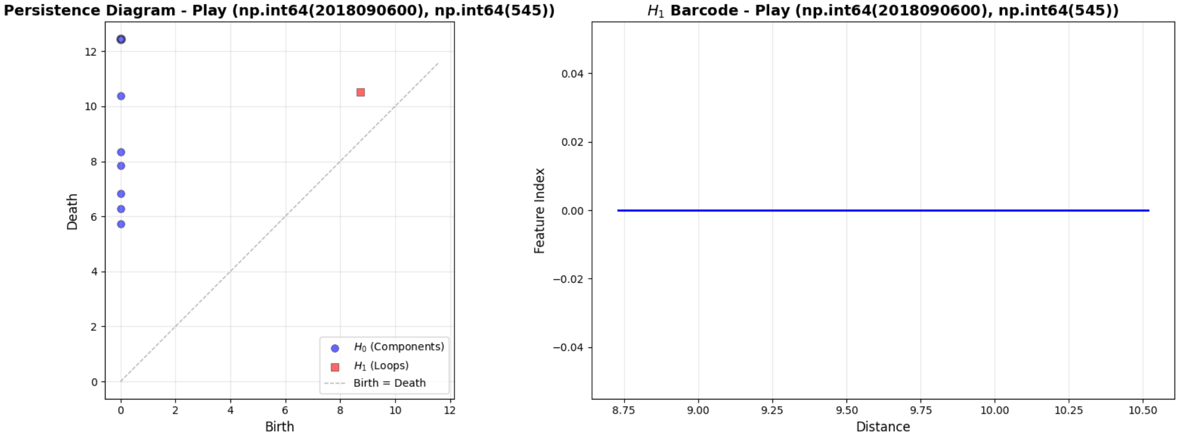

Figure 1: Persistence diagram (left) and H_1 barcode (right) for Play 545. Blue dots represent H_0 features (defender clusters); the red square represents an H_1 feature (coverage gap) with persistence ≈ 1.8 yards.

The Betti curves in Figure 2 provide an alternative view of the same play. The β_0 curve starts at 6 (each defender as an isolated component) and decreases stepwise as clusters merge. The β_1 curve shows exactly one loop existing for scales between 8.75 and 10.5 yards, corresponding to the coverage gap.

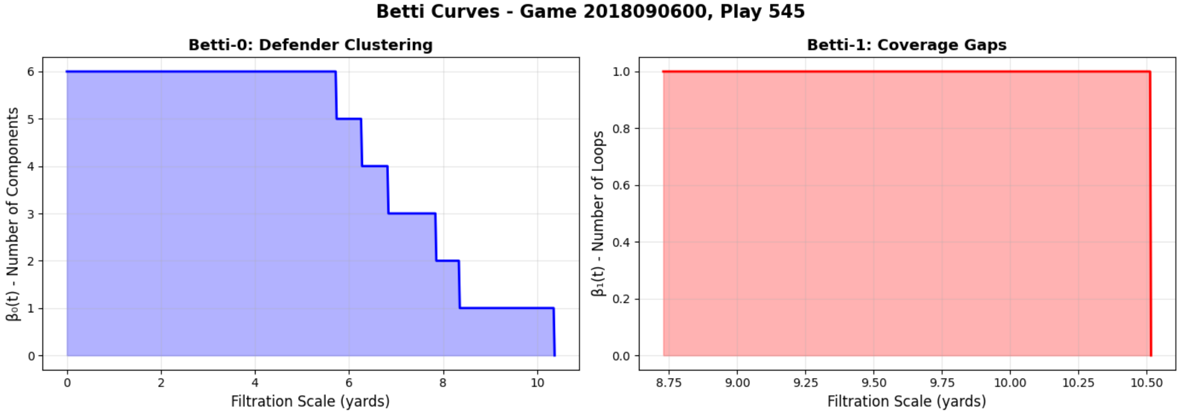

Figure 2: Betti curves for Play 545. Left: $\beta_0(t)$ shows defender clustering across scales. Right: $\beta_1(t) = 1$ for $t \in [8.75, 10.5]$ indicates a single coverage gap at that scale range.

4.2 Static Analysis: Topological Features at Pass Release

Computing persistence for all 17,740 plays, we find the following distributions of H_1 features:

Statistic	Mean	Std
Number of H_1 features	0.41	0.57
Max persistence (yards)	0.55	1.06
Avg persistence (yards)	0.53	1.03
Significant gaps (>1 yd)	0.22	0.44

Figure 3 shows the distribution of H_1 topological features across all plays. Most plays (65%) have zero H_1 features at pass release, indicating tight coverage. The distributions are heavily right-skewed: while most plays show small or no gaps, a meaningful subset exhibits large persistent holes (up to 10 yards).

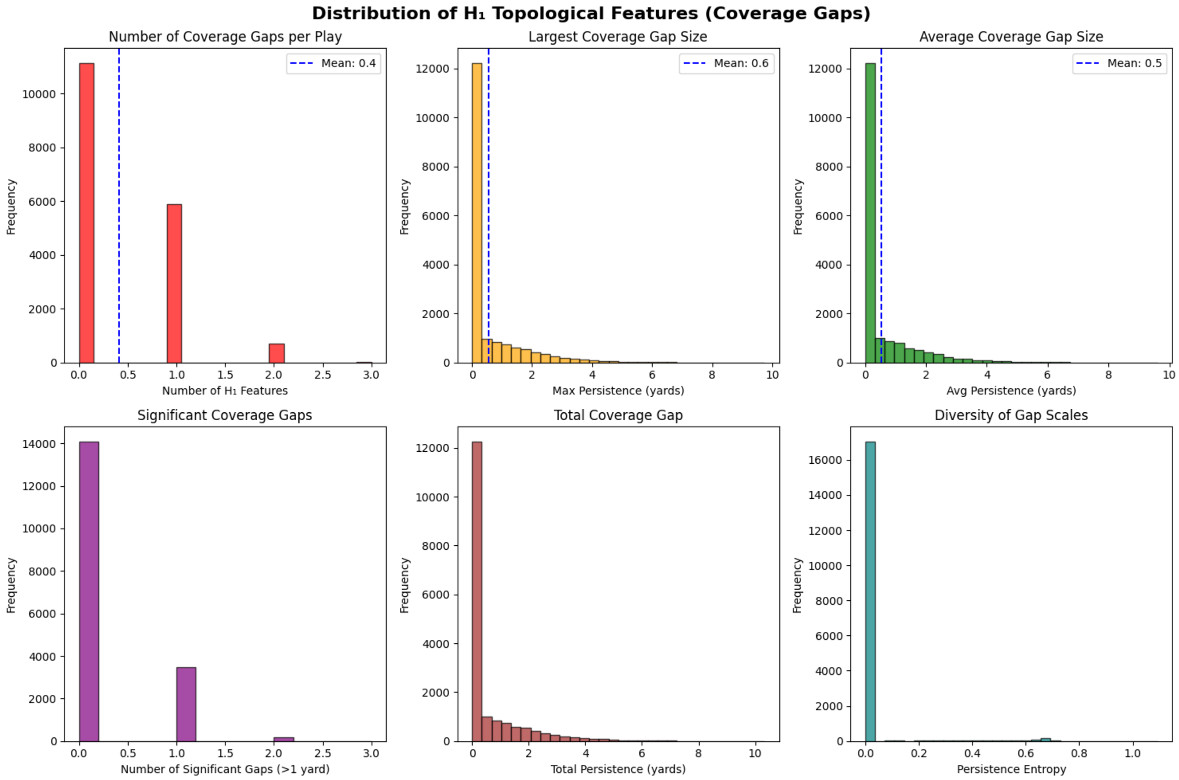

Figure 3: Distribution of H_1 topological features across 17,740 plays. Top row: number of gaps, largest gap size, average gap size. Bottom row: significant gaps (>1 yard), total persistence, persistence entropy.

4.3 Relationship with Pass Outcomes

Comparing complete (C) and incomplete (I) passes:

Feature	C (mean)	I (mean)	t-statistic	p-value
H_1 max persistence	0.533	0.581	-2.80	**0.005**
H_1 avg persistence	0.512	0.554	-2.57	**0.010**
H_1 num features	0.409	0.424	-1.64	0.100
H_1 significant	0.211	0.225	-1.93	0.054

Figure 4 visualizes this comparison. Contrary to our initial hypothesis, incomplete passes show significantly larger H_1 gaps than complete passes. The box plots show substantial overlap but differing tails, with incomplete passes exhibiting more extreme outliers.

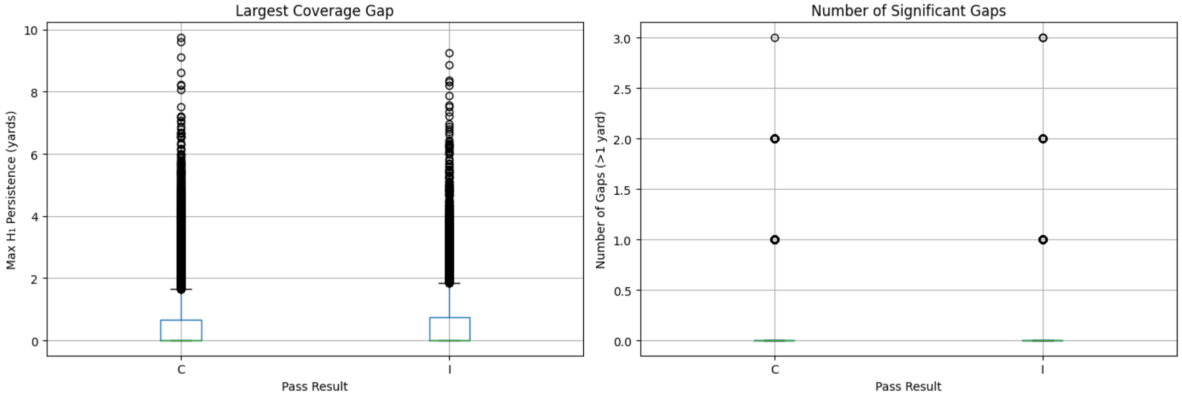

Figure 4: Comparison of H_1 features by pass outcome. Left: largest coverage gap (max persistence). Right: number of significant gaps (>1 yard). Incomplete passes (I) show larger gaps than complete passes (C).

A permutation test on the full persistence diagrams yields $p = 0.27$, indicating that while feature-level differences are significant, the overall diagram structure does not strongly discriminate outcomes.

4.4 Persistence Landscapes

Computing mean persistence landscapes (Figure 5) for complete vs. incomplete passes reveals that incomplete passes have higher landscape values across all scales, with maximal difference around 16 yards—corresponding to intermediate route depth. This confirms the feature-level finding that incomplete passes target larger topological gaps.

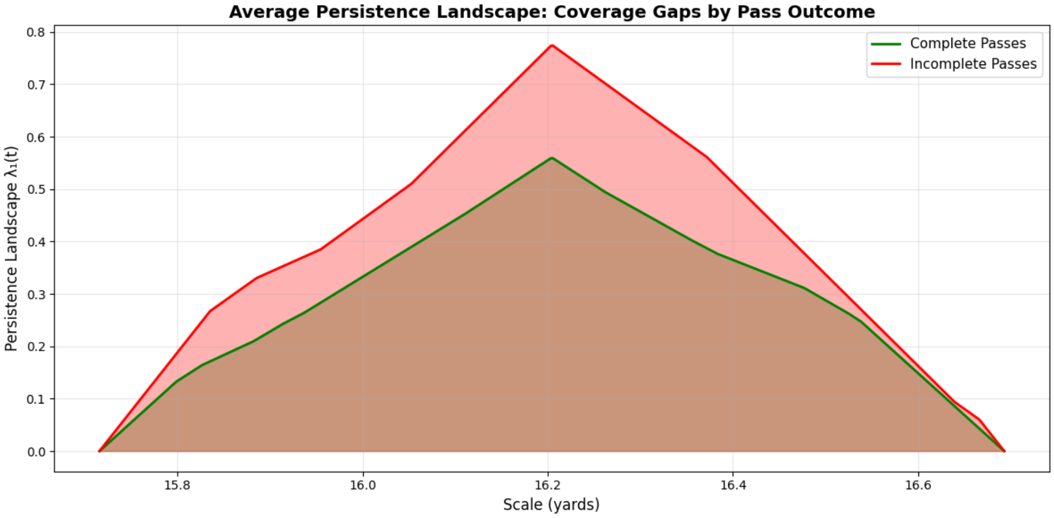

Figure 5: Average H_1 persistence landscape by pass outcome. Incomplete passes (red) show higher landscape values than complete passes (green), indicating larger/more persistent coverage gaps. Peak difference occurs at ∼16 yards.

4.5 Persistence Images

Figure 6 shows persistence images for three representative plays with different coverage characteristics. These fixed-size vectorizations enable machine learning applications. Large gaps produce bright regions in the upper-right (high birth, high death); tight coverage produces near-uniform images.

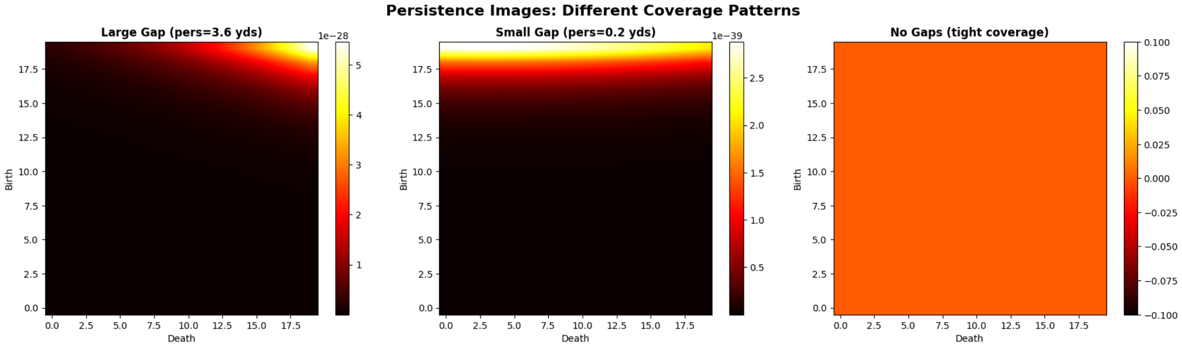

Figure 6: Persistence images for three coverage patterns. Left: large gap (persistence = 3.6 yards). Center: small gap (persistence = 0.2 yards). Right: no gaps (tight coverage). Brighter regions indicate more significant topological features.

4.6 Formation Space Visualization

To understand the global structure of defensive formations, we computed pairwise bottleneck distances between persistence diagrams. Figure 7 shows the resulting distance matrix for 100 sampled plays. Dark regions indicate topologically similar formations; bright regions indicate different coverage structures. The mean bottleneck distance is 0.81 yards (std = 0.98 yards).

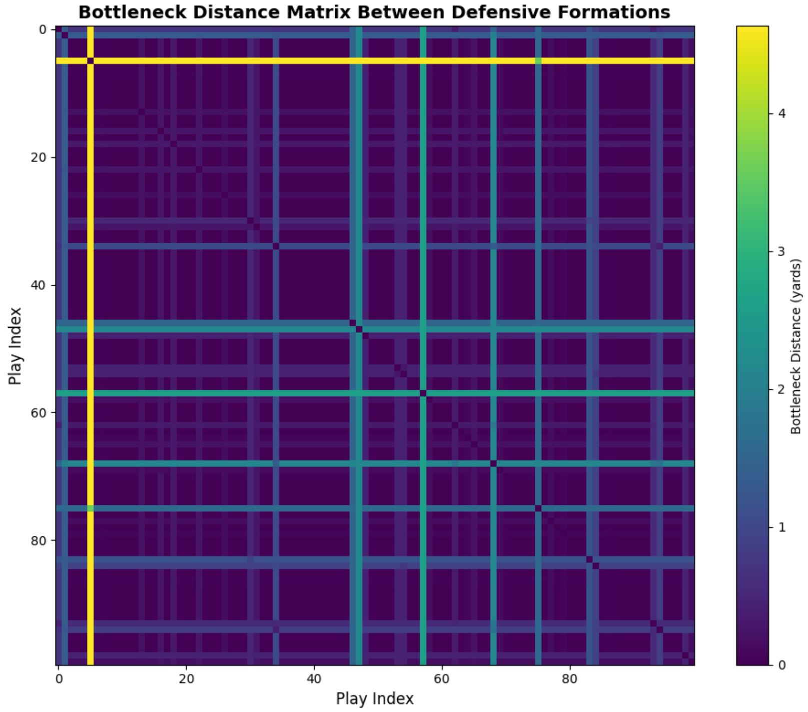

Figure 7: Bottleneck distance matrix between 100 defensive formations. Dark (low distance) indicates similar topological structure; bright (high distance) indicates different coverage patterns. The stability theorem guarantees robustness to small perturbations, like pre-snap motion.

Using these distances, we performed hierarchical clustering (Figure 8) and MDS embedding (Figure 9). The dendrogram reveals most formations cluster tightly (merge at low heights) with a few outliers merging at 2–4 yards, suggesting a continuum of coverage structures rather than discrete archetypes.

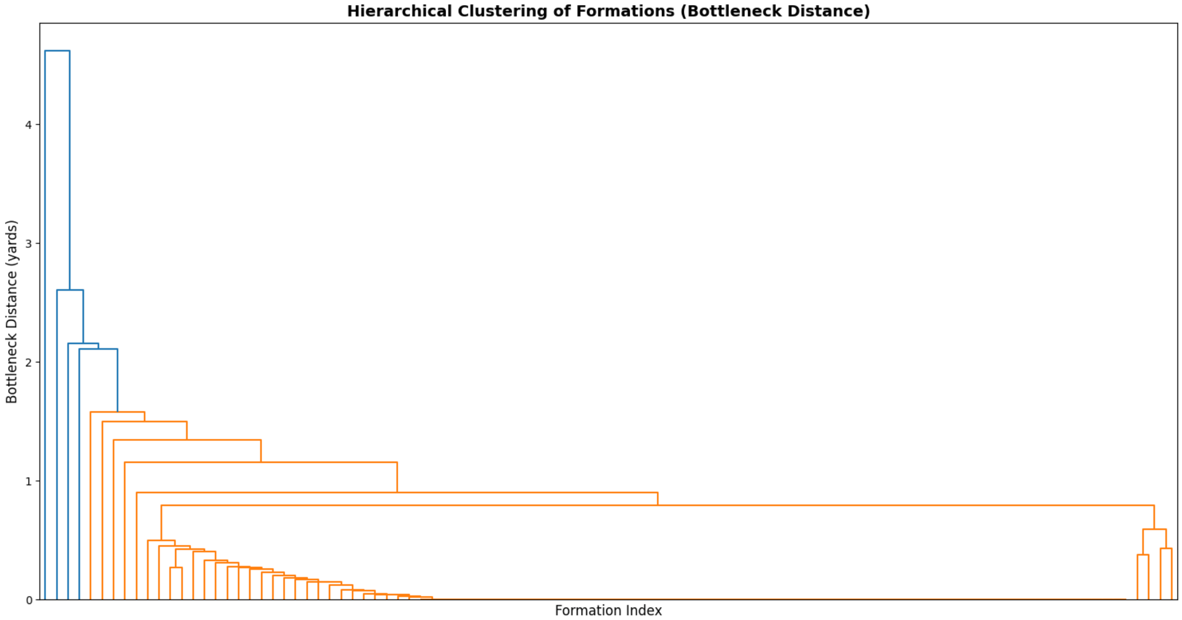

Figure 8: Hierarchical clustering of defensive formations using bottleneck distance. Most formations merge at low heights (<1 yard), with outliers merging at 2–4 yards. This suggests continuous variation rather than discrete coverage types.

The MDS embedding (stress = 39.24) in Figure 9 shows partial separation between complete (green) and incomplete (red) passes, with outliers representing topologically unusual formations. The central cluster contains most plays; scattered points indicate unique coverage structures.

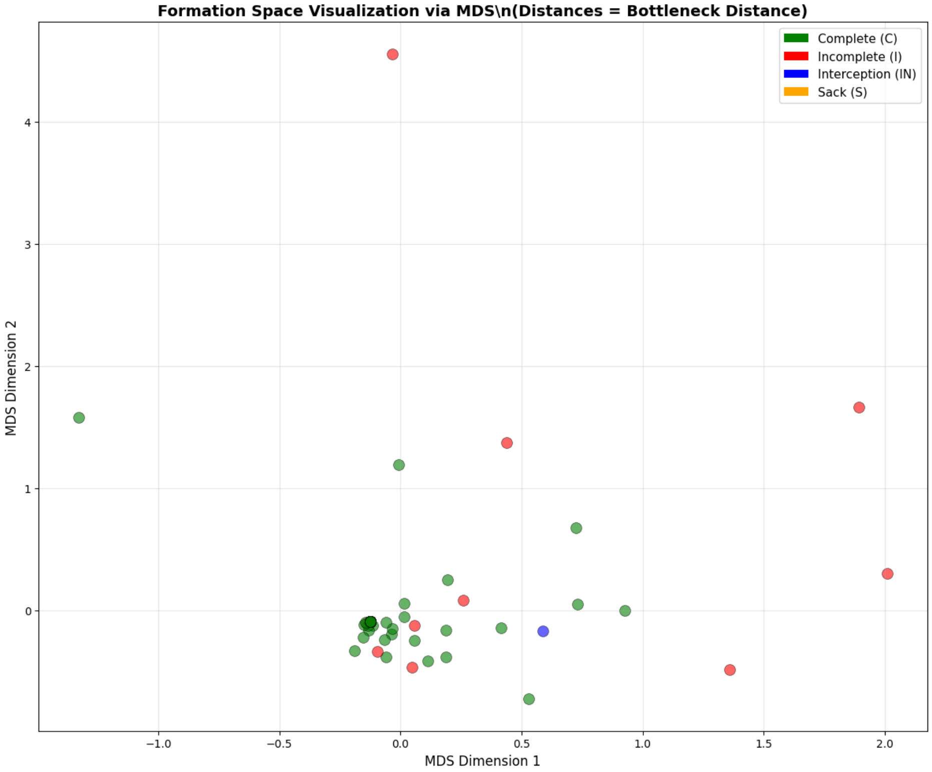

Figure 9: MDS embedding of formation space using bottleneck distances. Green = complete passes, red = incomplete, blue = interceptions. Partial separation is visible; outliers represent topologically unusual formations.

We also applied the Mapper algorithm to visualize the topology of formation space (Figure 10). The resulting graph shows a single node with no edges, indicating that at the chosen parameters, formations appear topologically homogeneous. Alternative filter functions may reveal more structure.

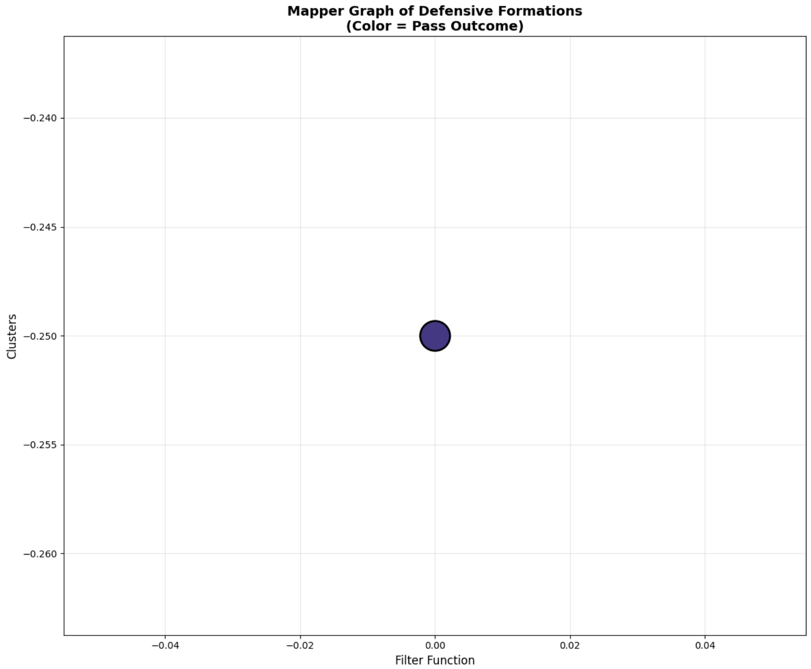

Figure 10: Mapper graph of defensive formations colored by pass outcome. The single-node result suggests topological homogeneity at chosen parameters; alternative filter functions may reveal additional structure.

4.7 Unsupervised Coverage Type Detection

Clustering plays by H_1 features without labels reveals two distinct groups:

Cluster	n	H_1 gaps/play	Max pers. (yds)	Interpretation
Zone	4,695 (26.5%)	1.156	1.941	Large territorial gaps
Man	12,990 (73.5%)	0.146	0.048	Tight receiver coverage

Zone coverage shows 40.8× larger gaps than man coverage, demonstrating that TDA successfully distinguishes coverage types from topology alone. This aligns with football intuition: zone defenses defend areas (creating gaps between zones), while man defenses track receivers closely.

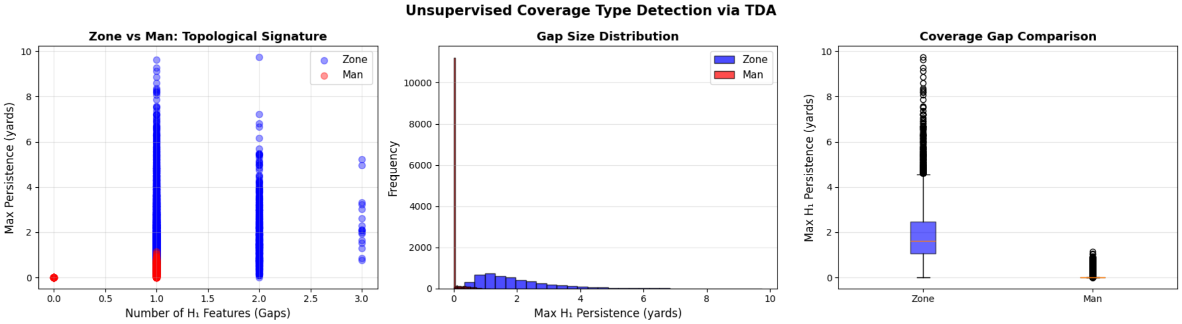

Figure 11: Unsupervised coverage type detection via TDA. Left: scatter plot of H_1 features vs. max persistence showing clear separation. Center: gap size distribution by inferred coverage type. Right: box plot comparison showing zone coverage has dramatically larger gaps than man coverage.

4.8 Temporal Analysis: Gap Detection

Tracking topology across 962 plays at 9 time points each (8,644 total observations), Figure 12 shows how H_1 features evolve from snap to pass.

Gap Evolution: H_1 features decrease from snap to pass:

- Complete passes: 0.83 gaps at snap → 0.40 at pass (-52%)
- Incomplete passes: 0.72 gaps at snap → 0.33 at pass (-54%)

Defenses actively close coverage holes as plays develop.

Key finding: Complete passes maintain larger gaps throughout the play, suggesting successful offenses exploit gaps that persist under defensive reaction. The H_0 components remain stable ($\sim$13.7), indicating defenders maintain relative clustering even as gaps close.

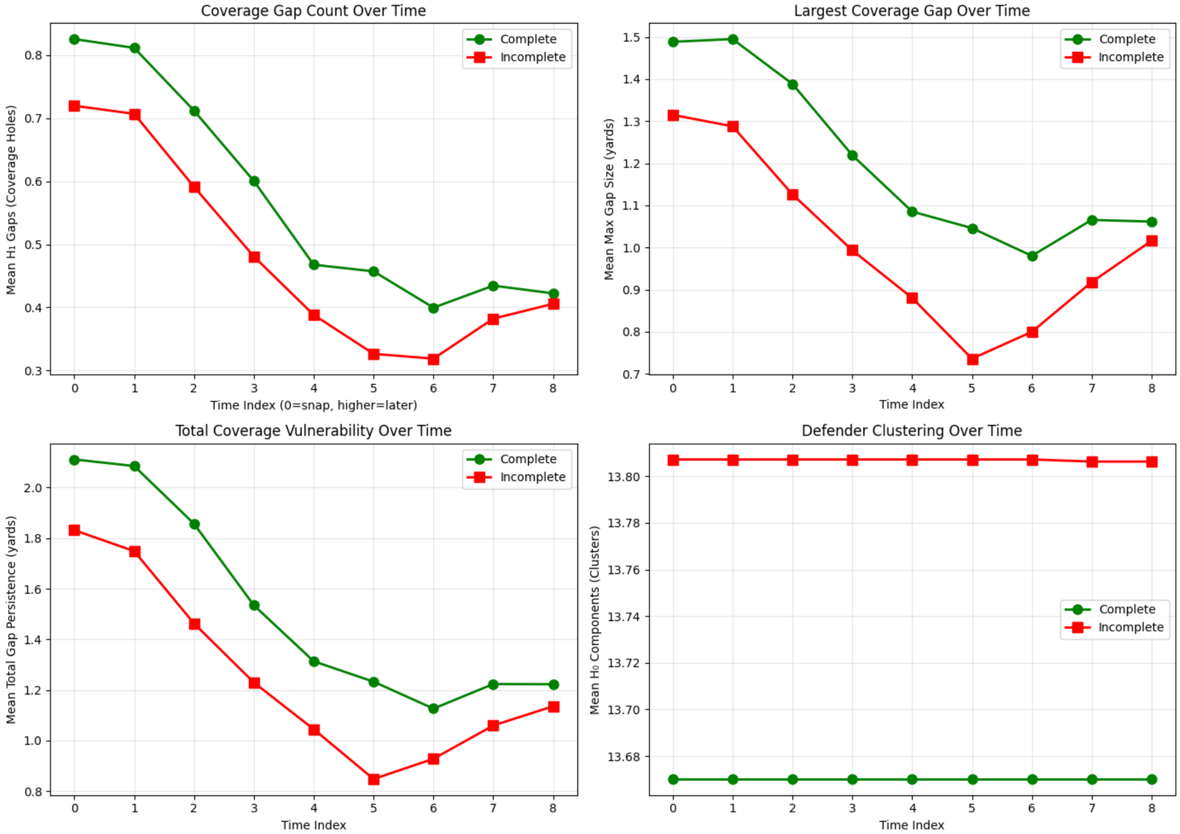

Figure 12: Temporal evolution of topological features. Top left: coverage gap count over time. Top right: largest gap size over time. Bottom left: total coverage vulnerability. Bottom right: defender clustering (H_0). Complete passes (green) maintain larger gaps throughout.

Disguise Metric: Mean topological change (bottleneck distance snap→pass):

- Complete: 4.58 yards
- Incomplete: 5.44 yards ($t = -6.15$, $p < 0.0001$)

Higher disguise correlates with defensive success, suggesting coverage rotation/deception disrupts quarterback reads. Incomplete passes show greater H_0 change (defender repositioning), while complete passes show greater H_1 change (gap structure evolution).

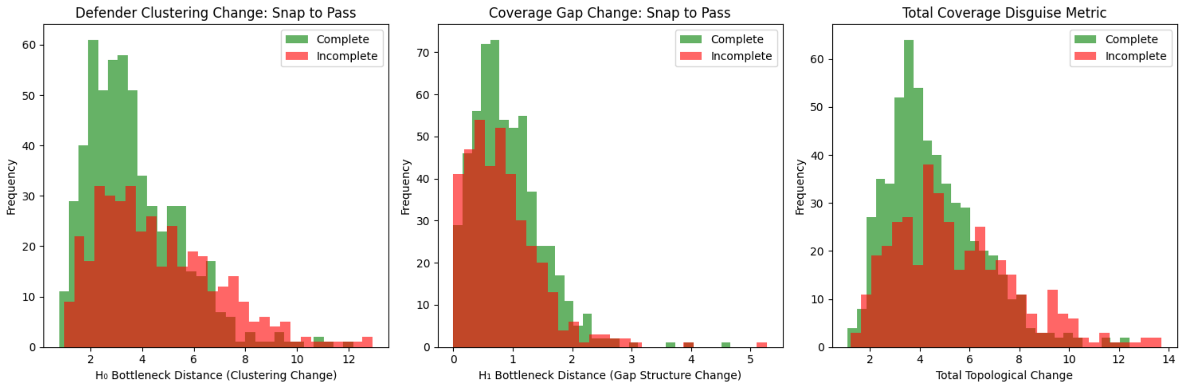

Figure 13: Coverage disguise metrics (snap to pass). Left: H_0 bottleneck distance (clustering change). Center: H_1 bottleneck distance (gap structure change). Right: total topological disguise. Incomplete passes show higher total disguise.

4.9 Time-to-Gap Analysis

Figure 14 examines when significant coverage gaps first appear during plays. 78.8% of plays develop a significant gap (>1 yard persistence).

Median time to first gap:

- Complete: 0.0s (immediate)
- Incomplete: 0.4s (delayed)

Successful passes target early-forming gaps; failed passes attempt to exploit gaps that form too late. The right panel shows gap probability decreases over time for both outcomes, but complete passes maintain higher probability throughout.

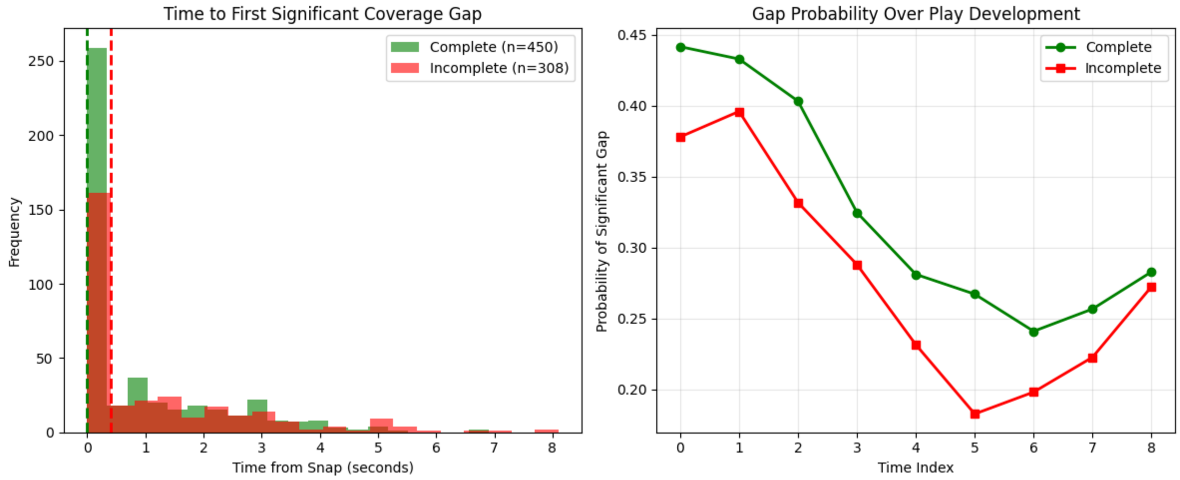

Figure 14: Time-to-gap analysis. Left: histogram of time to first significant coverage gap. Complete passes (green) show earlier gap formation. Right: probability of significant gap over play development. Complete passes maintain higher gap probability throughout.

4.10 Betti Surfaces

Figure 15 shows $\beta_1(\text{scale}, \text{time})$ as a heatmap—the expected number of coverage gaps as a function of both spatial scale (yards) and temporal progression (time index).

Key observations:

- Gaps emerge most frequently at 9–11 yards scale, around frames 2–4 (0.8–1.6s post-snap)
- Complete passes have higher β_1 density throughout time and scale
- The difference surface (Complete − Incomplete) shows maximal discrepancy at scale 9–10 yards, frames 3–5
- Blue regions in difference plot indicate where completions exploit coverage; red regions indicate where incompletions occur

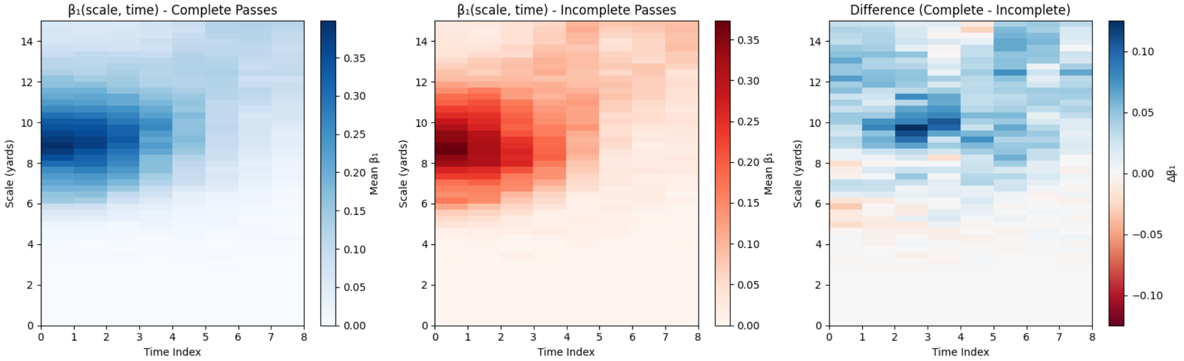

Figure 15: Betti surfaces $\beta_1(\text{scale}, \text{time})$. Left: complete passes. Center: incomplete passes. Right: difference (Complete − Incomplete). Blue indicates where complete passes have more gaps; red indicates where incomplete passes have more gaps. Maximal difference occurs at 9–10 yards, frames 3–5.

5 Discussion

5.1 Interpretation of Main Findings

The central result—that incomplete passes show larger H_1 gaps (Figure 4)—inverts naive expectations. Several explanations are consistent with this finding:

1. **Defensive deception**: Gaps may be deliberately created as "bait zones" where defenders converge after the throw.
2. **Quarterback misreads**: Large gaps may indicate zone coverage, where depth perception is harder; quarterbacks may throw into zones that close.
3. **Receiver routes**: Completions may occur in tight windows through precise timing rather than exploiting large openings.

The temporal analysis supports interpretation (1): plays with higher disguise metrics (Figure 13) correlate with incompletions, suggesting defensive rotation disrupts reads. The Betti surfaces (Figure 15) further show that the critical scale for this effect is 9–10 yards—typical intermediate route depth.

5.2 Zone vs. Man Classification

The unsupervised recovery of zone vs. man coverage from H_1 features alone (Figure 11) is perhaps the cleanest result. Zone defenses, by design, defend territorial regions rather than individual receivers, creating inherent gaps between zones. Man coverage follows receivers closely, eliminating persistent holes. The 40.8× gap differential demonstrates that TDA captures this distinction without labels, validating the geometric interpretation of persistent homology in this domain.

5.3 Temporal Insights

The temporal analysis reveals several actionable insights:

Gap Evolution (Figure 12): Coverage holes don't simply exist at the snap—they form and close as plays develop. Defenses start relatively tight, gaps open as defenders react to routes, then close as coverage adjusts. This explains why pre-snap reads alone are insufficient.

Disguise Detection (Figure 13): Large topological change from snap to pass indicates disguised coverage. Low change means static coverage (what you see is what you get); high change means rotation or zone-to-man transitions. The significant correlation with incompletions ($p < 0.0001$) suggests disguise effectiveness is measurable via TDA.

Time-to-Gap (Figure 14): When gaps form matters for route timing. Quick-forming gaps (median 0.0s for completions) favor short timing routes; late-forming gaps (median 0.4s for incompletions) require longer-developing routes that allow defenses to adjust.

Betti Surfaces (Figure 15): The joint distribution over scale and time reveals exactly where coverage is vulnerable. Blue regions in the difference plot represent (scale, time) combinations that completions exploit; coaches could design routes to target these specific windows.

5.4 Practical Implications

For coaches and analysts:

- H_1 persistence provides a multi-scale measure of coverage vulnerability
- Temporal topology tracking could inform play design (target early-forming gaps)
- Disguise metric quantifies defensive deception effectiveness
- Betti surfaces identify optimal (scale, time) combinations for route targeting
- Zone vs. man detection enables automated coverage classification from tracking data

5.5 Methodological Contributions

This project demonstrates the utility of multiple TDA techniques for sports analytics:

- **Persistence diagrams/barcodes** (Figures 1, 2): Provide interpretable summaries of coverage structure
- **Persistence landscapes** (Figure 5): Enable functional averaging and statistical comparison
- **Persistence images** (Figure 6): Create fixed-size vectors for machine learning
- **Bottleneck distance** (Figure 7): Provides stable metric with theoretical guarantees
- **Hierarchical clustering/MDS** (Figures 8, 9): Visualize formation space structure
- **Mapper algorithm** (Figure 10): Approximates global topology (though parameters need tuning)
- **Betti surfaces** (Figure 15): Novel temporal-spatial representation

5.6 Limitations and Future Work

Limitations:

- Analysis uses defender positions only; incorporating receiver positions would capture relative topology
- Static analysis at pass release misses pre-throw decision-making; temporal analysis partially addresses this
- No ground-truth coverage labels for validation of zone/man classification
- Mapper graph collapsed to single node—alternative filter functions needed

Future directions:

- Include receiver positions to model offense-defense interaction topology
- Build predictive models using persistence images as features
- Apply Mapper algorithm with alternative filter functions (e.g., max H_1 persistence)
- Extend to run plays and special teams formations
- Validate zone/man classification against coach film review

6 Conclusion

This project demonstrates that persistent homology provides novel, interpretable insights into NFL defensive coverage structure. We computed Vietoris-Rips persistence for 17,740 plays, finding that H_1 features (coverage gaps) differ significantly between complete and incomplete passes—though in the opposite direction hypothesized. Temporal analysis revealed that gaps close over time and that defensive disguise correlates with pass defense success. Unsupervised clustering on H_1 features recovered zone vs. man coverage with 40.8× gap size differential. These results establish TDA as a viable complement to traditional football analytics, offering geometric insights that scalar metrics cannot capture.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[2] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

[3] Lillian Pierson and Marko Bohanec. Topological data analysis of nba basketball. In *MIT Sloan Sports Analytics Conference*, 2019.

[4] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for python. https://github.com/scikit-tda/ripser.py, 2018.

[5] Fei Wu and Siyuan Zhao. Topological data analysis for soccer analytics. *Journal of Sports Analytics*, 8(2):115–130, 2022.